INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Analýza a klasifikace dat

Jiří Holčík

Březen 2012

Příprava a vydání této publikace byly podporovány projektem ESF č. CZ.1.07/2.2.00/07.0318 **„Víceoborová inovace studia Matematické biologie“** a státním rozpočtem České republiky.

Předmluva

Těžko se hledá přiléhavý název, který by výstižně pojmenoval náplň této publikace. Za celou dobu, po kterou se metody analýzy a klasifikace dat rozvíjejí, dostala tato disciplina mnohá jména. Ty různé názvy ani tak nesouvisí s vlastní podstatou tohoto způsobu práce s daty, jako spíše s účelem zpracování. Asi nejobecnější název zní „rozpoznávání obrazů“, v angličtině „*pattern recognition*“. Ve skutečnosti ale nejde o žádné obrazy ve smyslu děl Leonarda da Vinci či jiných velikánů výtvarného umění (jak by bylo možné vyvozovat z českého překladu), nýbrž o pouhý matematický popis vlastností reálného objektu, jehož stav chceme hodnotit, nějakým abstraktním způsobem – např. vektorem hodnot, grafem, apod. Proto, navzdory skutečnosti, že se někdy (v souladu se rčením, že čeština pro nějaké specifické anglické názvy nemá slov) anglické názvosloví považuje za fetiš, ani anglické slovo „*pattern*“ znamenající především „vzor, schéma, předloha, šablona“, možná i „systém“ není úplně to pravé a ideální. Jiný obecný název disciplíny – „vícerozměrné statistické metody“ zase navozuje představu, že dané metody využívají pouze pravděpodobnostní principy a že tyto metody jsou součástí statistiky. Ano, z velké části je to správný názor, ale ne zas až tak úplně. Mnohé publikace zabývající se touto problematikou nesou i anglický originální název „*data mining*“ a v české kotlině máme hned problém, zda tento název překládat jako „dolování dat“, nebo spíše „vytěžování dat“. V podstatě je cílem těchto aktivit v daných datech odhalit nějaké skryté jevy, souvislosti, závislosti. Takže vlastně analýza. Ale toto slovo v průběhu všech těch tisíciletí, po která se používá, nabylo tak generálního, obecného významu, že je ho snad i stydno použít ve spojení s daty. I označení „strojové učení“ („*machine learning*“) se používá. To je ale zase spíše důsledek toho, že ta vlastní analýza, klasifikace nebo predikce není zas až tak velká věda, ale to, jak přimět počítače, aby to udělaly za nás, to je teprve ta správná disciplína, ve které lze psát dizertace. No, a když už se zabýváme tímto výčtem, ten by určitě nebyl úplný bez zmínky o „umělé inteligenci“. Tato disciplína, pokud nebudeme komentovat tu hrůzu, kterou ve všech důsledcích její jméno vyvolává, se ale z gruntu věnuje mimetickým modelům lidského rozhodování. Umělé neuronové sítě jsou krásným příkladem metod umělé inteligence. Všechny ostatní matematicky založené metody, které sice slouží k témuž účelu, ale nejsou tak zcela inspirovány činností lidského ducha, do této přihrádky jaksi nepatří.

Kdyby se tato publikace chtěla pokusit shrnout, co lidstvo v této disciplíně vytvořilo, musel by být její předpokládaný rozsah překročen mnoha mnohanásobně. To skromně naznačují i některé knihy uvedené na konci tohoto textu v doporučené literatuře. Proto, i s ohledem na další souvislosti vzniku této publikace (vznikla v souvislosti s řešením projektu ESF č. CZ.1.07/2.2.00/07.0318 „Víceoborová inovace studia Matematické biologie“, který si dal do vínku inovovat náplň a zlepšit provázanost povinných předmětů studijního oboru Matematická biologie na PřF MU), je to spíše publikace typu „obrázky z analýzy a klasifikace dat“. Obsahuje metody, které umožní čtenářům, jimž je určena, rozšířit si obecné povědomí o některých snad typických metodách, postupech a algoritmech sloužících ke zpracování, analýze i klasifikaci údajů o objektech sice všeliké obecné, hlavně však biologické a medicínské podstaty. Byli bychom určitě potěšeni, kdyby se tak i stalo.

V Brně 23. března 2012

Jiří Holčík

ISBN 978-80-7204-793-2

1 Kapitola úvodní aneb o čem to tady bude

1.1 Zpracování dat – základní principy

Reálný život nás dennodenně staví před různá rozhodnutí všelijaké úrovně a kvality – zda si právě koupit dva nebo tři nebo čtyři rohlíky, či zda si vzít za manžela toho fešného mladého muže od vedle, příp. jaké bude vzájemné soužití s ním po 20 letech. Tyto úlohy zvládáme zpravidla intuitivně, aniž si uvědomujeme, že i tato rozhodnutí jsou podložena sběrem a analýzou potřebných informací tak, aby závěrečné rozhodnutí bylo co nejlepší. Např. jak velký máme hlad, kolik máme peněz, zda vůbec máme chuť na rohlíky i to, zda náš zdravotní stav umožní konzumovat jiné pochutiny než rohlíky, apod.

Složitější úlohy samozřejmě potřebují více informace i složitější způsob uvažování, který už většinou není možné zvládnout intuitivně, ale je potřeba rozhodovací postupy zformalizovat, následně algoritmizovat a vytvořené algoritmy implementovat zpravidla v počítačovém prostředí, především ale připravit vstupní informaci/data ve formalizované podobě, vhodné pro strojové zpracování.

Zpracováním dat se obecně snažíme zkoumat vztahy mezi ***stavy***, ***jevy***[1] a ***procesy***[2], které charakterizují určitý objekt a jsou charakterizovány naměřenými daty.

Základní výchozí představa spočívá v tom, že existuje nějaký reálný objekt (skupina pacientů, kůň, vodní tok, doprava ve městě Brně), který poskytuje informaci o svém stavu (věk, váha, krevní tlak, datum diagnózy, okamžitý průtok vody a její chemické složení, výskyt živočišných druhů v určité tůni na potoce, počet nasazených autobusů ve městě, okamžitý počet přepravovaných cestujících, apod). Ta informace je ukryta v datech, která daný objekt generuje a my jsme schopni je přiměřeně přesně změřit. Data jsou obecně mnohorozměrná (stav objektu je popsán mnoha proměnnými) a dynamická (v čase proměnná). Počet relevantních proměnných popisujících dostatečně přesně stav objektu z hlediska účelu, kvůli kterému objekt sledujeme, udává řád tohoto objektu, resp. jeho modelu, jehož konstrukce může být jednou z cest jak zákonitosti vyplývající z chování objektu analyzovat.

Data obsahují jednak ***deterministickou***, jednak ***nedeterministickou složku***. Deterministická část dat umožňuje najít příčinný vztah mezi stavem objektu a výrokem, kterým objekt, resp. jeho stav hodnotíme, popřípadě klasifikujeme. Toto hodnocení komplikuje nedeterministická složka dat, která může vznikat jako důsledek nějakých, z hlediska dané úlohy nežádoucích skutečností, jevů či procesů.

Často se vnímá, že zpracování dat zahrnuje pouze metody založené na statistických základech a principech. Skutečnost je ale taková, že stejně významnou skupinu tvoří metody vycházející z postupů a přístupů, které se primárně snaží postihnout deterministickou podstatu zkoumané skutečnosti. Každý z obou přístupů si vytvořil svou specifickou terminologii, která často používá k vyjádření různých skutečností stejných pojmů, nebo naopak různých pojmů

---

[1] Obecně chápeme ***jev*** jako souhrn vnějších, smyslům bezprostředně či zprostředkovaně (např. měřením) přístupných vlastností a vztahů daného objektu. Statistika dále specificky definuje ***náhodný jev*** jako výsledek náhodného pokusu, o kterém lze po provedení pokusu rozhodnout, zda nastal nebo nenastal. Náhodný jev tedy představuje událost, která za určitých podmínek buď nastane, nebo nenastane.

[2] ***Proces*** vnímáme jako postupné, vnitřně navzájem svázané transformace jevů, objektů, systémů v jiné jevy, objekty nebo systémy. Zatímco stav, resp. jev považujeme spíš za statický fenomén, proces má charakter dynamický.

k popisu téhož, což logicky může vést v lepším případě k nedorozumění, v horším případě k chybným interpretacím dosažených výsledků[3].

1.2 Cíl zpracování dat

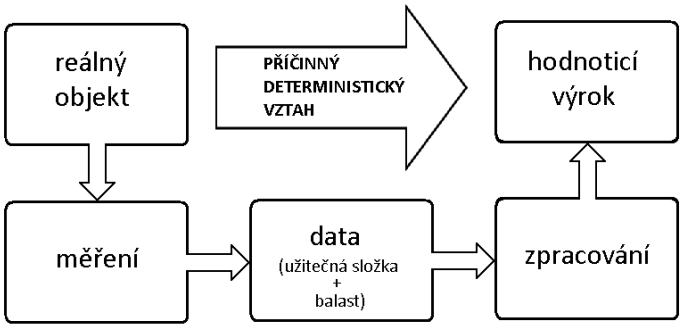
Cílem jakéhokoliv zpracování (analýzy) dat je zpravidla posouzení zkoumaného reálného objektu (živého či neživého), který je zdrojem analyzovaných dat, příp. jeho stavu.

Toto posouzení může nejčastěji vyústit:

- v rozhodnutí o typu či charakteru objektu – např. že daná rostlina je pomněnka lesní (*Myosotis sylvatica*), zvíře že je medvěd hnědý (*Ursus arctos*) nebo že daná budova je vystavěna v renesančním slohu – ***klasifikační,*** resp. ***rozpoznávací úloha***;
- v posouzení kvality stavu analyzovaného objektu, např. zda je pacient v pořádku nebo má infarkt myokardu, cirhózu jater, apod. – opět ***klasifikační,*** resp. ***rozpoznávací úloha***;
- v rozhodnutí o budoucnosti objektu – např. zda lze pacienta léčit a vyléčit, zda les po 20 letech odumře, jaké bude sociální složení obyvatelstva na daném území a v daném čase – ***klasifikační*** nebo také ***predikční úloha***[4].

Formálně tedy hledáme cestu od skutečného reálného objektu k formálnímu výroku o jeho kvalitě, stavu, příp. budoucnosti (obr.1.1).

Hovoříme-li o zpracování či analýze dat, pak v zobrazeném řetězci potřebných operací většinou pomíjíme blok měření, který je většinou vázán s řešením technických, nikoliv matematických problémů. Přesto všechno je tento blok, kromě samotného zkoumaného objektu, nejvíce spojen se vznikem různých rušivých složek, které naměřené údaje obsahují. Tyto rušivé složky vznikají jak přímo ve zdroji (měřeném objektu), tak při vlastním měření. V měřeném objektu vznikají vlivem neovlivnitelných změn podmínek existence daného objektu v čase (***intraindividuální variabilita***), vlivem odlišnosti jednotlivých, zdánlivě ekvivalentních částí celku (***inte-***



Obr.1.1 *Cíl zpracování dat a kroky k jeho dosažení*

[3] Základní případ, kdy se terminologie v oblasti statistického a nestatistického zpracování poněkud liší, je vnímání pojmů ***zpracování*** a ***analýza***. Zatímco statistické pojetí dává přednost označovat veškeré výpočty nad daty analýzou, oblast nestatistická (deterministická) dává přednost použití slova analýza pro vyjádření specifičtějších operací, více odpovídajících definici uvedené v poznámce 5) a globální proces označuje spíše pojmem zpracování.

[4] ***Klasifikace*** a ***predikce*** jsou opět dva pojmy, jejichž použití v odborné literatuře často splývá. Pojem predikce (z lat. *prae-*, před, a *dicere*, říkat) zjevně nese časové (příp. prostorové) hledisko, když jej používáme ve významu předpovědi či prognózy, jako soud o tom, co se stane nebo nestane v budoucnosti. V tomto významu je používán např. v analýze či zpracování časových řad. Zmatení, které vyplývá ze zaměňování pojmů klasifikace a predikce, se některá odborná literatura snaží ne zcela přesvědčivě rozmotat konstatováním, že pojem klasifikace je používán, použije-li se klasifikačního algoritmu pro známá data. Pokud jsou data nová, pro která apriori neznáme klasifikační třídu, pak hovoříme o predikci klasifikační třídy (to by znamenalo, že za klasifikaci považujeme pouze procesy spojené s návrhem klasifikátoru, vlastní činnost klasifikátoru by pak měla být nazývána predikcí – s takovým vysvětlením se lze smířit pouze velice obtížně). Za příjemnější rozlišení obou pojmů považujeme výklad, který říká, že pojem klasifikace používáme, pokud vybíráme identifikátor klasifikační třídy z určitého diskrétního konečného počtu možných identifikátorů. Pokud určujeme (predikujeme) spojitou hodnotu, např. pomocí regrese, pak hovoříme o predikci, i když tento pojem nezbytně časovou dimenzi nemá.

rindividuální variabilita) i vlivem skutečnosti, že na objektu měříme veličinu, která je ovlivněna jinými ději téže povahy (např. signál EKG matky při snímání fetálního EKG). Při vlastním měření se rušivé složky objevují v datech nejčastěji vlivem špatného uspořádání měřicího experimentu (měříme něco, co jsme ani měřit nechtěli a co nenese požadovanou informaci, použili jsme nevhodný měřicí přístroj, přístroj používá nevhodnou metodu měření, resp. nevhodný algoritmus primárního předzpracování dat, rušení z vnějšího prostředí proniká k původním, šumu prostým datům při přenosu dat v čase i prostoru).

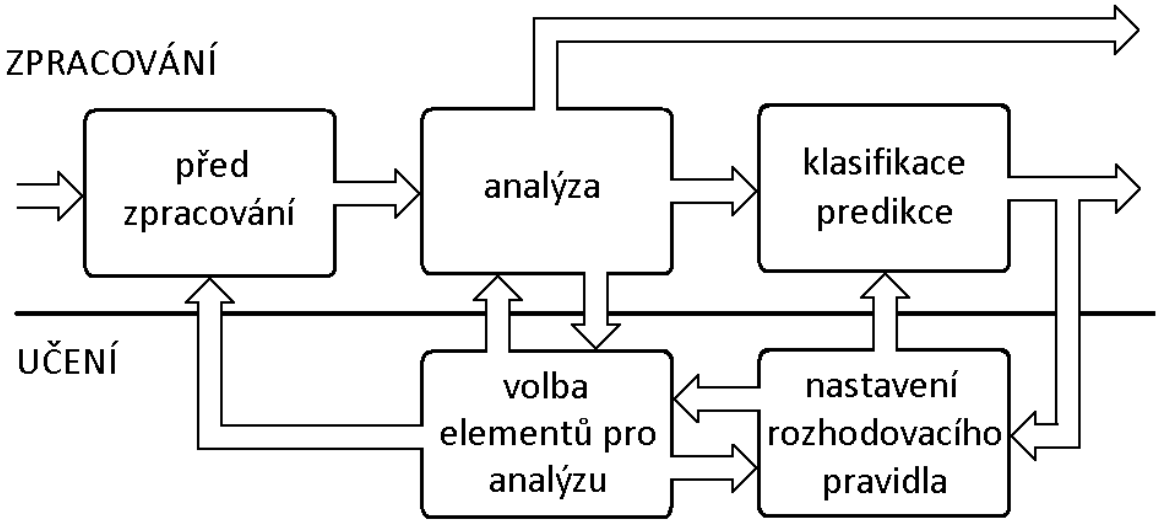
Ve statistice rušivé komponenty dat zpravidla označujeme jako ***variabilitu dat***, kterou je potřeba odstranit, potlačit či dodatečně vysvětlit. Pokud jsou data zpracovávána některými nestochastickými postupy, je zvykem nazývat je ***rušením***, ***šumem***, nebo ***poruchami***. Opět se snažíme rušení z dat eliminovat, tentokrát zpravidla na základě apriorní znalosti jeho charakteristik, příčin vzniku, apod.

Navzdory různým rušivým složkám, které se v datech vyskytují, musí být mezi finálním hodnotícím výrokem (diagnostickým, klasifikačním, predikčním), který vyřkneme na základě znalosti dat o daném objektu, a tímto objektem jasný deterministický příčinný vztah. Pokud by tento příčinný vztah neexistoval, pak ani data nemohou obsahovat využitelnou informaci a je zbytečné se o jeho hledání snažit, a to jakýmikoliv prostředky – ať statistickými či založenými na jiných principech.

Chceme-li upřesnit cíl zpracování (analýzy) dat, definovaný na začátku této kapitoly, pak je to právě odhalení toho příčinného deterministického vztahu, navzdory všemu tomu, co to odhalení kazí.

1.3 Blokové schéma zpracování dat

Blok zpracování v obr.1.1 lze podrobněji vyjádřit schématem na obr. 1.2.



Obr.1.2 *Blokové schéma zpracování dat*

Procesně se blok zpracování skládá ze tří následných, podstatou odlišných operací – ***předzpracování***, ***analýzy*** a ***klasifikace***, resp. ***predikce***. První z těchto bloků reprezentuje postupy vázané na odstranění rušivé, resp. zvýraznění užitečné složky originální reprezentace dat, rekonstrukci a doplnění chybějících údajů, příp. redukci dat, např. odstraněním jejich redundantní (nadbytečné) nebo irelevantní (neužitečné) části, na převod do formy, v níž je podstatná informace lépe patrná (např. časová vs. frekvenční reprezentace časových řad), příp. i

na převod zpravidla spojité proměnné, jejíž hodnoty měříme, na diskrétní hodnoty, tzv. ***vzorkování***, resp. ***kvantování***. Zatímco prvním pojmem zpravidla rozumíme diskretizaci s ohledem na nezávisle proměnou, druhý pojem používáme pro diskretizaci funkčních hodnot.

Blok analýzy[5] je zde vnímán v zúženém smyslu jako blok operací, které vedou k nalezení hodnot proměnných, resp. jiných elementů (např. určitých geometrických tvarů v obrazu), které představují významnou složku zpracovávané informace, případně nalezení vazeb mezi nimi. Konečně, poslední blok představuje blok zařazení dat do stanovených klasifikačních, např. diagnostických kategorií nebo odhadu budoucího stavu objektu. Při řešení určitých úloh nemusí být řetězec úplný, úloha může např. skončit analýzou, některé bloky mohou rekurzívně obsahovat, pro vyřešení dílčího problému, řetězec úloh, který opět obsahuje všechny tři fáze zpracování, apod.

Řetězec zpracování dat je podporován bloky fáze učení, které jsou s bloky zpracování spojeny mnohými vazbami informačních toků. Fáze učení může předcházet fázi zpracování (algoritmus je navržen před vlastním zpracováním dat), nebo pracovat paralelně (charakteristiky algoritmu jsou dolaďovány během vlastního zpracování).

Rozeberme nyní účel a obsah jednotlivých dílčích bloků podrobněji.

1.3.1 Blok předzpracování

Jak bylo výše uvedeno, blok předzpracování reprezentuje některé operace nad zpracovávanými daty, které jednak zajišťují čitelnost dat, jednak zvyšují jejich kvalitu. Jedním ze základních dominantních cílů předzpracování je tzv. ***čištění*** či ***filtrace dat***, což představuje operace vedoucí k potlačení parazitní variability dat či odstranění rušení, resp. zvýraznění užitečné (z hlediska cíle zpracování) složky dat.

Základní formou reprezentace dat vstupujících do bloku předzpracování je množina (příp. uspořádaná) vektorů, obsahujících hodnoty veličin, které o zpracovávaných objektech získáváme. Tyto veličiny mohou být jakéhokoliv běžného typu, tj.

- ***kvantitativní*** (***numerické*** – ***spojité***, ***diskrétní*** a ve speciálním případě ***binární***, ***logické***, které nabývají dvou diskrétních hodnot, ale lze určit, která z nich znamená méně, která více a je možné s nimi provádět matematické operace);
- ***ordinální*** – typ ***kategoriální*** proměnné, její hodnoty ale lze vzájemně seřadit, je však obtížné kvantifikovat jejich hodnoty (např. bolest zanedbatelná, malá, střední, velká, nesnesitelná);
- ***nominální*** – opět typ ***kategoriální*** proměnné, v tomto případě ale nelze jejich hodnoty seřadit podle velikosti (např. příslušnost studenta MU k určité fakultě), speciálním typem nominální proměnné je tzv. ***dichotomická*** proměnná, která podobně jako proměnná binární, či logická nabývá dvou hodnot, které se navzájem vylučují, ale v tomto případě nelze určit jejich velikost, např. pohlaví muž/žena.

Kvůli možnosti srovnání hodnot různých veličin se často hodnoty veličin upravují např. ***centrováním*** (odečtení střední hodnoty), ***normalizací*** (vztažení hodnoty proměnné k nějak definované normě), resp. ***standardizací*** (centrovaná hodnota je vztažena k určité specifické hodnotě, často např. k směrodatné odchylce). To bývá nezbytné např. v situaci, kdy z hodnot jednotlivých veličin jako jsou výška v centimetrech (chceme ji měřit v centimetrech?), nebo metrech (je lepší ji měřit v metrech?), hmotnost v kilogramech nebo gramech, věk udaný

---

[5] ***Analýza*** (z řec. *analyó*, rozvazovat, rozebírat) znamená rozbor, metodu zkoumání složitějších skutečností rozkladem na jednodušší. Je založena na dekompozici celku na elementární části. Cílem analýzy je identifikovat podstatné a nutné vlastnosti elementárních částí celku, poznat jejich podstatu a zákonitosti. Používá se v mnoha vědách, ve filosofii i v běžném životě, pokud chceme dospět k výsledku na základě detailního poznání podrobností.

v letech, počet bílých krvinek vztažený na nějakou objemovou jednotku mm^3 nebo litr (kdy je lépe použít tu či onu jednotku?), apod, máme určit nějakou hodnotu, která globálně charakterizuje všechna potřebná data, např. normu vektoru $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ určenou podle vztahu

$$\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{n} x_i^2}, \tag{1.1}$$

přičemž jednotlivé složky jsou reprezentovány výše uvedenými hodnotami charakterizujícími pacienta. V tom případě normalizace i standardizace udělá z jednotlivých hodnot hodnoty bezrozměrné, které už můžeme použít k výpočtu nějakého sumarizujícího, globalizujícího parametru. Je ovšem třeba si uvědomit, že např. standardizací vůči směrodatné odchylce jsme se v datech zbavili informace o střední hodnotě i rozptylu. Obě tyto hodnoty mohou být důležité z hlediska následného zpracování, které se touto operací může výrazně zkomplikovat, nebo i zcela znemožnit. Proto musíme vždy pečlivě zvážit, zda operace, které určitou fázi zpracování zjednoduší, jsou přípustné z hlediska dalších cílů zpracování.

Předzpracování dat se principiálně může lišit podle charakteru dat – zda jsou ***statická*** (vyjadřují stav zdrojového objektu bez potřeby popsat jeho dynamiku – zdroj buď dynamický není, nebo jeho dynamika není důležitá – očekávané změny probíhají pomaleji než důsledky provedeného zpracování), nebo ***dynamická*** (vyjadřují změny stavu zdroje – data musí být uspořádána, nejčastěji jako funkce času nebo prostoru). Typickými představiteli dynamických dat jsou signály, obrazy, časové řady (což jsou v podstatě matematické modely diskrétních signálů). Přes některé rozdíly v přístupu ke zpracování dat toho či onoho typu má zpracování v obou případech mnoho společného. Pokusme se vystihnout tento společný základ.

Čištění (filtrace) dat - samozřejmě ideální případ nastává, když data žádné parazitní nežádoucí komponenty neobsahují. Jejich obsah lze účinně minimalizovat vhodným uspořádáním experimentu, při kterém údaje o sledovaném objektu nebo objektech měříme. Uspořádáním experimentu můžeme různé zdroje rušení vyloučit. V případě, že to není možné, lze experiment uspořádat tak, že se snažíme průběžně stanovovat vlastnosti rušení, když ne přímo jeho hodnoty.

Vliv variability při statistickém zpracování omezujeme ***náhodným výběrem*** (např. [18], [19]) zkoumaných subjektů, odhad charakteristik základní variability dat může poskytnout tzv. ***kontrolní skupina***. Variabilita dat může být potlačena identifikací a odstraněním ***odlehlých (vybočujících) vzorků*** (např. [18], nebo [19]).

Typickým případem, jak odstranit parazitní složky dat v případě signálů či časových řad je využití případné frekvenční disjunktnosti užitečné i parazitní komponenty, tj. případu, kdy se užitečná i parazitní složka dat skládají z harmonických průběhů odlišných frekvencí. (Podrobnější informaci o frekvenčních vlastnostech signálů a časových řad a jejich frekvenčních spektrech může čtenář tohoto textu najít např. v [2].) V tom případě je velice účinným a relativně jednoduchým nástrojem pro filtraci nežádoucího rušení použití ***lineárních časově invariantních systémů*** (***filtrů***), které budou podrobněji rozebrány v kapitole „Lineární systémy a modely časových řad“.

Pokud obě složky dat (užitečná i parazitní) nejsou frekvenčně separabilní, pak je třeba k vzájemnému oddělení použít jakoukoliv jinou dostupnou informaci, např. míru ***korelace*** obou složek. V případě, že obě složky nejsou korelovány a užitečná složka dat má repetiční charakter, lze pro potlačení rušení použít kumulačních technik (***zprůměrování),*** např. [5]. Naopak, pokud neznáme přesný průběh rušení, pouze nějaká data s rušením korelovaná, nebo naopak data korelovaná s užitečnou složkou analyzované posloupnosti, pak možným řešením je použití ***adaptivního filtru*** [5].

Řešíme-li situaci, kdy se snažíme odhadnout okamžik výskytu nějakého známého průběhu v posloupnosti silně ovlivněné náhodným rušením, pak lze použít ***souhlasného filtru***, což v podstatě reprezentuje průběžný výpočet vzájemné korelace mezi datovou posloupností a

posloupností, která reprezentuje hledaný průběh. Pro filtraci posloupnosti dat vhodně ovlivněné náhodným rušením očekávaných vlastností lze použít např. ***optimální Wienerův filtr***[6] nebo adaptivní ***filtr Kálmánův***[7].

Redukce dat se zpravidla dopouštíme, když chceme umožnit nebo alespoň usnadnit přenos či zajistit efektivní uchovávání dat, nebo požadujeme-li zrychlení zpracování dat. Zatímco v prvních dvou případech očekáváme, že budeme schopni data opět zcela přesně rekonstruovat do původní podoby (***vratná redukce dat***), v posledním případě obnovu původní datové reprezentace apriori nepředpokládáme (***nevratná redukce dat***). (Konec konců samotné klasifikační vyvrcholení zpracování dat je demonstrativním případem datové redukce, protože veškerá data popisující analyzovaný objekt reprezentujeme pouze identifikátorem klasifikační třídy.)

Algoritmy vratné redukce dat využívají pravděpodobnostních charakteristik dat a odstraňují pouze redundantní (nadbytečnou, známou) složku dat. Jsou založeny na reprezentaci redundantní složky pomocí různých aproximačních, interpolačních či extrapolačních algoritmů a uchovávají se parametry modelů redundantní složky spolu s odchylkami reálných dat od hodnot určených modelem, které jsou zpravidla podstatně menší než hodnoty původních, což posléze vede k možnému omezení rozsahu dat.

Naopak algoritmy nevratné redukce dat jsou založeny pouze na určení a uchování parametrů modelů dat za předpokladu, že rozdíl mezi skutečnou hodnotou a hodnotou vyplývající z modelu je menší než předpokládaná prahová hodnota. Modely dat bývají určeny pomocí po částech polynomiální aproximace, či jinými formami vyjádření dat – parametry harmonických složek nezbytných pro dostatečně přesné vyjádření signálu, parametry složek kosinové transformace. V poslední době jsou pro nevratnou redukci dat velice často používány parametry složek určených pomocí vlnkové transformace.

Rekonstrukce a doplnění chybějících údajů - je vždy ke zvážení, zda neúplnou informaci ze zpracování vyloučit, či zda se pokusit o její odhadnutí.

Základní algoritmy doplnění dat jsou dominantně založeny na interpolaci, ať již polynomiální či na základě rozkladu do řady. Využíváme-li statistických vlastností dat, pak zpravidla hovoříme o regresi, jednorozměrné či vícerozměrné. Často používaný přístup vychází z modelu zdroje dat, který konstruujeme na základě apriorních informací o sledovaném procesu, tj. o chování zdrojového objektu.

Součástí předzpracování může být i převod hodnot kategoriálních proměnných do hodnot, se kterými lze následně provádět výpočty.

Dalšími dvěma bloky se bude tato publikace v dalších kapitolách zabývat poměrně podrobně, proto na tomto místě uveďme pouze nejdůležitější fakta, zásady a principy.

1.3.2 Blok analýzy dat a blok volby elementů pro analýzu

Jak je uvedeno v pozn.5), obecným cílem analýzy je identifikovat podstatné a nutné vlastnosti elementárních částí celku, poznat jejich podstatu a zákonitosti. To v praxi znamená nalezení zákonitostí v rozložení hodnot použitých proměnných, stanovení míry korelace, příp. závislosti mezi hodnotami použitých proměnných. V některých případech přímo nalezení

6 http://en.wikipedia.org/wiki/Wiener_filter (12.12.2011)

7 http://en.wikipedia.org/wiki/Kalman_filter (12.12.2011)

V různých komunitách se citace informací převzatých z Wikipedie vnímá jako cosi nežádoucího, nepěkného, protože konec konců informace zde uváděné nejsou recenzované a verifikované. Domníváme se však, že tvářit se, že tento zdroj informace neexistuje, je druhým právě tak škodlivým extrémem. Pokud nebereme wikipedické informace za jedině správné dogma, nýbrž zdravě kriticky (což konec konců je nezbytné i u mnohokrát recenzovaných zdrojů), pak je určitě Wikipedie velice užitečný zdroj poučení, což platí i pro tento případ.

vhodného matematického vztahu vyjadřujícího funkční závislost mezi použitými proměnnými. To vše pro usnadnění práce s daty, která zpracováváme.

Výsledky analytických výpočtů pak mohou být použity k transformaci předzpracovaných dat do vstupního formátu klasifikačního bloku, nebo mohou být i finálním výsledkem zpracování dat (bez navazující klasifikace).

Druhým nezbytným cílem tohoto bloku by měla být redukce počtu proměnných, což může usnadnit následné klasifikační operace. V našem případě jde o nalezení těch proměnných[8], jejichž hodnoty nesou správnou informaci pro kvalitní funkci posledního bloku zpracování, tj. klasifikaci nebo predikci.

Obecně způsob, jak primárně určit příznakové veličiny nesoucí nejvíce informace pro klasifikaci, není teoreticky formalizován ([1], [6]), tzn. že neexistuje teorie, podle níž by bylo možné předem stanovit veličiny, jejichž hodnoty poskytují užitečnou informaci, nebo naopak ty, které jsou pro následné zpracování nepodstatné. Současná teorie nabízí pouze dílčí, suboptimální řešení, spočívající ve výběru nezbytného počtu veličin z předem zvolené množiny proměnných, příp. ve vyjádření původních příznakových veličin pomocí menšího počtu skrytých (tzv. latentních) nezávislých proměnných, které nelze přímo měřit, ale mohou, ale i nemusí mít určitou věcnou interpretaci.

V žádné z obou možností však není specifikováno, jak určit výchozí množinu příznakových proměnných. V tomto směru nezbývá než se spolehnout na empirickou znalost analyzovaného problému u těch, kteří se daným konkrétním problémem zabývají, a na technických možnostech dokázat změřit hodnoty takto vybraných veličin. Není proto jisté, že zvolená výchozí množina bude obsahovat právě i ty veličiny, jejichž hodnoty jsou pro klasifikaci ty nejužitečnější.

1.3.3 Blok klasifikace

Klasifikací[9] rozumíme rozdělení dané (konkrétní či teoretické) skupiny (množiny) objektů, jevů či procesů na konečný počet dílčích skupin (podmnožin), v nichž všechny objekty, jevy či procesy mají dostatečně podobné společné vlastnosti. Vlastnosti, podle nichž lze klasifikaci zadat či provádět, určují klasifikační kritéria. Objekty, které mají podobné vlastnosti, tvoří klasifikační třídu. Každá klasifikace musí být úplná, tzn., že každý předmět musí patřit do nějaké třídy a nemůže být současně ve dvou či více třídách.

Klasifikaci provádíme pomocí klasifikátoru (obr.1.3), což je algoritmus se vstupem, odpovídajícím charakteru dat, popisujících analyzovaný objekt a jedním diskrétním výstupem, jehož hodnota je identifikátorem klasifikační třídy, do které klasifikátor zařadí vstupní reprezentaci dat. Tedy platí

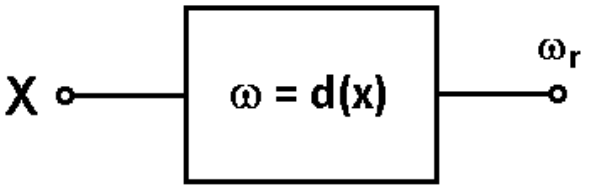
---

[8] I v tomto případě panuje velká diverzita pojmů, používaných pro označení popisných proměnných a jejich hodnot. Zatímco statistické metody analýzy dat rády používají pojmu znak, pozorování, nebo i diskriminátor [3], publikace zabývající se specificky problémy klasifikace dat, používají pojmu příznaková proměnná, resp. příznak pro její konkrétní hodnotu. V tomto případě je daný babylon kupodivu pouze v češtině, anglická literatura dominantně používá pojem „*feature*“, který bohužel čeští odborníci z různých oblastí začali překládat různě.

[9] S pojmem klasifikace se často zaměňuje pojem diskriminační analýza, která hledá vztah mezi kategoriální proměnnou a množinou vzájemně vázaných příznakových proměnných. Předpokládáme-li, že existuje konečný počet R různých a priori známých populací, kategorií, tříd nebo skupin, které označujeme ω_r, r=1,…,R, je úkolem diskriminační analýzy nalézt vztah (diskriminační, rozhodovací pravidlo), na základě kterého pro daný vektor příznaků popisujících konkrétní objekt tomuto vektoru přiřadíme hodnotu ω_r [10]. Porovnáme-li tuto definici s dalším vysvětlujícím textem v této kapitole, určitě si uvědomíme, že diskriminační analýza je, za předpokladu příznakového popisu klasifikovaného objektu či procesu, jednou z možných, specifických náplní bloku nastavení rozhodovacího pravidla.

$$\omega_r = d(X) \qquad (1.2)$$

kde d(X) je funkce argumentu X představujícího reprezentaci vstupních dat, kterou nazýváme ***rozhodovací pravidlo klasifikátoru*** a ω_r, r = 1, …, R je ***identifikátor klasifikační třídy***. Rozhodovací pravidlo se stanoví v odpovídajícím bloku učební fáze.

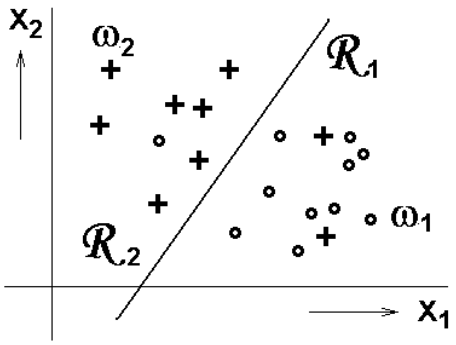


Obr.1.3 *Schéma klasifikátoru*

Proměnnou X, formálně popisující klasifikovaný objekt, obecně nazýváme ***obrazem***[10].

Na tomto místě je dobré zmínit a uvědomit si i některé skutečnosti, týkající se značení identifikátoru klasifikační třídy, které budeme v tomto textu respektovat. Na obr.1.4 je znázorněna situace, kdy jsou ve dvourozměrném prostoru zobrazeny vektory vyjadřující zástupce dvou skutečných tříd – koleček, popsaných např. identifikátorem ω_1 a křížků, popsaných např. identifikátorem ω_2. Výsledkem učení klasifikátoru bylo, že klasifikační třídy můžeme vyjádřit pomocí hraniční přímky ve vektorovém prostoru, která rozdělila celý prostor na dvě poloroviny – polorovinu $\mathcal{R}_1$, v níž se vyskytují především zástupci třídy koleček a polorovinu $\mathcal{R}_2$, v níž leží spíše zástupci klasifikační třídy křížků. Nicméně rozdělení není dokonalé, v polorovině $\mathcal{R}_1$ leží i dva křížky, v polorovině $\mathcal{R}_2$ jedno kolečko. Na to konto bude kolečko ležící v polorovině $\mathcal{R}_2$ přiřazeno chybně do třídy křížků, označené identifikátorem ω_2. Nelze tedy směšovat identifikátor skutečné klasifikační třídy, byť je i přiřazena chybně (ω_i) a identifikátor části prostoru ($\mathcal{R}_i$), který se používá při konstrukci rozhodovacího pravidla.

Rozhodovací pravidla pracují na základě vzdálenosti či podobnosti mezi vstupní datovou reprezentací a vzorem klasifikační třídy (speciálním případem této formy klasifikace je možnost ztotožnění vstupních dat s etalonem klasifikační třídy), hranic rozdělujících obrazový prostor dat, pomocí funkcí, které určují míru příslušnosti k dané klasifikační třídě, tzv. ***diskriminačních funkcí***, případně doplňkových logických pravidel.



Obr.1.4 *Označení klasifikačních tříd a částí obrazového prostoru*

Všechny tyto zmíněné postupy mohou být deterministické či pracovat na pravděpodobnostních principech, přičemž deterministické pravidlo může vycházet i z pravděpodobnostních charakteristik zpracovávaných dat. Proto za deterministický klasifikátor považujeme takový, který daná vstupní data zpracuje vždy se stejným jednoznačným výsledkem.

[10] Z tohoto označení vyplývá i jeden z obecných názvů této discipliny – rozpoznávání obrazů, což ovšem nemá nic společného ani s malířstvím, ani s rentgenovými snímky, nýbrž s výsledkem zobrazení originálního reálného objektu do jeho určité abstraktní reprezentace. To může být např. vektor hodnot popisujících daný objekt – tzv. vektor pozorování, nebo jak posléze uvidíme, vektor příznaků. Jinou formou může být i graf či obecně nějaká relační struktura. V anglickém originálu je rozpoznávání obrazů vyjádřeno jako „*pattern recognition*“, což by spíš odpovídalo překladu rozpoznávání vzorků, resp. vzorů, spíše evokující zpracování nějakých etalonů dané skutečnosti. Často se můžeme v této oblasti setkat i s jiným názvem, v anglickém originálu „*data mining*“, překládaným jako „dolování dat“, nebo „vytěžování dat“. To už ale není klasická klasifikační disciplina, nýbrž postup (byť má s analýzou a klasifikací mnoho společného), kterým se v datech snažíme nalézt nějaké skryté závislosti a skutečnosti.

Na druhé straně, nedeterministický klasifikátor může táž data při opakovaném zpracování klasifikovat různě. Nedeterministické klasifikátory nemusí být jen a pouze pravděpodobnostní. Existují i jiné další matematické disciplíny, které pracují s neurčitostí, uveďme zde jako příklad fuzzy logiku, příp. fuzzy algebru.

Při práci s algoritmy nejen tohoto typu se často vyskytuje členění na ***parametrické*** a ***neparametrické*** algoritmy, metody, modely. Parametrický algoritmus pracuje na základě nějaké dané funkce, jejíž konkrétní vlastnosti jsou určeny a mohou se měnit s hodnotami konečného počtu stanovených parametrů. Ve statistice rozumíme parametrickým odhadem hustoty pravděpodobnosti postup, kdy na základě určité apriorní informace předpokládáme určitý typ rozložení pravděpodobnosti a do formule, která toto rozložení popisuje, určujeme jen konkrétní hodnoty jejich parametrů, jako např. střední hodnotu a směrodatnou odchylku pro normální rozložení. Příkladem parametrického klasifikačního algoritmu je prahová klasifikace, která např. zařadí vstupní obraz do určité klasifikační třídy, pokud hodnota, charakterizující daný obraz, překračuje nebo nepřekračuje danou prahovou úroveň. Tuto prahovou hodnotu určujeme v učební fázi algoritmu. Typickým představitelem parametrických klasifikačních algoritmů jsou proto rozhodovací stromy. Příkladem neparametrického klasifikačního algoritmu je například klasifikace podle minimální vzdálenosti od etalonu klasifikační třídy. V tom případě určíme vzdálenost vstupního obrazu od všech etalonů klasifikačních tříd a obraz zařadíme do té třídy, jejíž etalon má ke vstupnímu obrazu nejblíže. I když pro stanovení vzdálenosti používáme různé metriky, které formálně také mají různé parametry (Euklidova metrika používá druhou odmocninu součtu čtverců rozdílů dílčích souřadnic daných vektorů, resp. bodů v prostoru), tyto hodnoty už ve fázi učení neurčujeme, jsou pevně svázány s daným typem metriky.

Parametričnost či neparametričnost vlastního klasifikačního algoritmu ale nic nepředurčuje, pokud jde o charakter algoritmů učení klasifikátorů. Existuje na příklad velká třída trénovacích algoritmů pro klasifikační stromy, které si nekladou žádné požadavky na způsob učení, učící postup není závislý na žádných partikulárních parametrech, jsou to tedy algoritmy neparametrické. Často je vlastní klasifikace poměrně jednoduchý postup a to zajímavé, co se týká zvoleného klasifikátoru, je způsob jeho návrhu, resp. učení a proto charakter učícího algoritmu v odborné literatuře často předurčuje náhled na typ klasifikátoru. Je proto potřeba rozlišovat.

1.3.4 Blok nastavení rozhodovacího pravidla

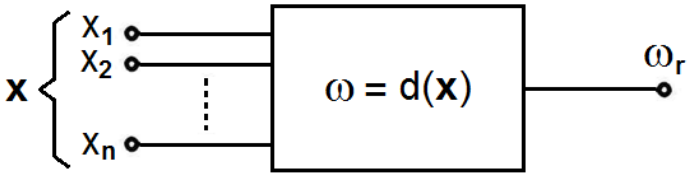
Tento blok je jedním ze dvou základních bloků učící fáze zpracování dat. Výsledkem tohoto bloku je návrh obecného tvaru rozhodovacího pravidla, případně určení jeho parametrů, v případě, že rozhodovací pravidlo je parametrické. Zatímco návrh obecného tvaru rozhodovacího pravidla není formalizován a závisí především na zkušenostech konstruktéra buď s danou reálnou úlohou, nebo s charakterem naměřených a zobrazených dat. Návrh parametrů rozhodovacího pravidla pak standardně vede na použití nějaké optimalizační úlohy. Děje se to na základě tzv. učební nebo trénovací množiny, která obsahuje vstupní obrazy spojené s informací o předpokládané správné klasifikaci (uspořádané dvojice datového popisu a identifikátoru klasifikační třídy). V tom případě hovoříme o ***učení s učitelem***, a podle míry spolehlivosti údaje o předpokládané klasifikaci rozlišujeme algoritmy ***učení s dokonalým*** či ***nedokonalým učitelem***. V případě, že trénovací množina není k dispozici, pak blok nastavení rozhodovacího pravidla obsahuje pouze návrh jeho obecného tvaru a případné nastavování parametrů rozhodovacího pravidla probíhá současně s klasifikací. Tento postup označujeme jako ***učení bez učitele***. Typickým příkladem je shlukování.

2 Příznakové metody klasifikace dat

2.1 Základní pojmy a principy

Příznakový obraz **x** hodnoceného objektu je formálně vyjádřen n-rozměrným (sloupcovým) vektorem x_i, i = 1, …, n ***příznakových proměnných*** (***veličin***) charakterizujících daný objekt, tj. platí $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$. Příznakové proměnné mohou popisovat kvantitativní i kvalitativní vlastnosti objektu. Jejich hodnoty nazýváme příznaky. Vrchol každého příznakového vektoru (obrazu) představuje bod n-rozměrného prostoru $\mathcal{X}^n$, který nazýváme obrazovým prostorem. Obrazový prostor je definován pomocí kartézského součinu definičních oborů všech příznakových proměnných, tzn. že jej tvoří všechny možné obrazy zpracovávaného objektu. V případě, že příznaky vyjadřují kvantitativní vlastnosti objektu, může být obrazový prostor euklidovský.

Je-li klasifikovaný objekt popsán vektorem příznaků, představuje klasifikátor algoritmus (stroj – podle v angličtině v tomto případě standardně používaného slova *machine*) s tolika vstupy, kolik je použito příznaků (jinými slovy jaký je rozměr příznakového vektoru, popisujícího klasifikovaný obraz) a s jedním diskrétním výstupem, jehož hodnoty určují třídu, do které klasifikátor zařadil rozpoznávaný obraz. Klasifikátor si tedy lze představit jako zařízení (obr.2.1), které realizuje matematickou operaci (rozhodovací pravidlo)



Obr.2.1 *Schéma příznakového klasifikátoru*

$$\omega_r = d(\mathbf{x}), \tag{2.1}$$

kde d(**x**) je skalární funkce vektorového argumentu **x**.

Příznakové klasifikátory se v principu mohou lišit časovým sledem použití jednotlivých příznaků – buď lze zpracovávat celý vektor jako celek, nebo lze nejen zpracovávat, nýbrž především i pořizovat (měřit) jednotlivé příznaky postupně, což umožňuje minimalizovat počet potřebných příznaků při požadované kvalitě rozhodnutí a v praktickém důsledku náklady na pořízení nezbytné informace pro dostatečně kvalitní klasifikaci. První z uvedených způsobů klasifikace nazýváme ***paralelní klasifikací***, zatímco druhý označujeme jako ***klasifikace sekvenční***. Základní principy paralelní klasifikace budou dále popsány a rozvinuty v kap.2.2 až 2.5, základní principy sekvenční klasifikace budou jen stručně naznačeny v kap.2.6 a dále v určitých směrech rozvinuty v navazujících učebních textech [4].

Klasifikační třídy jsou vymezeny, jak již bylo uvedeno v předchozích kapitolách, určitými nepřekrývajícími se částmi obrazového prostoru, které představují obrazy klasifikovaných objektů s dostatečně podobnými vlastnostmi. Formálně můžeme předpokládat, že obrazový prostor je rozdělen na R disjunktních prostorů $\mathcal{R}_r$, r = 1, …, R, přičemž každá podmnožina $\mathcal{R}_r$ obsahuje ty obrazy **x**, pro které $\omega_r = d(\mathbf{x})$.

Klasifikační třídy lze v obrazovém příznakovém prostoru vymezit několika následujícími způsoby:

a) pomocí tzv. ***diskriminačních funkcí***;
b) pomocí ***etalonů*** klasifikačních tříd – počet etalonů klasifikační třídy může být různý – od jednoho reprezentativního vzorku, který exkluzivně představuje danou klasifikační třídu až po úplný výčet všech vektorů (obrazů) patřících do dané klasifikační třídy; s tímto

způsobem popisu je pak nejčastěji vázána ***klasifikace podle minimální vzdálenosti***, resp. maximální podobnosti;

c) vymezením ***hraničních ploch***.

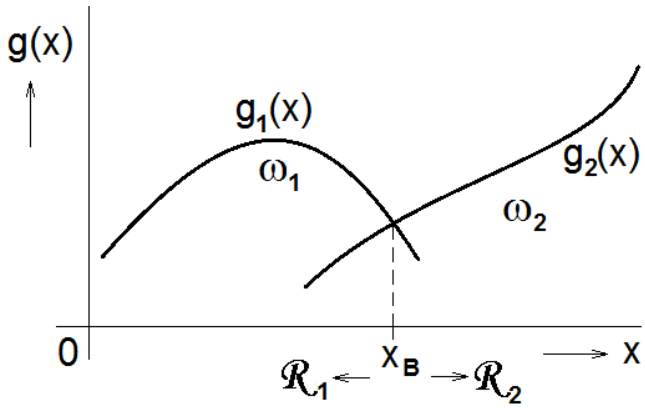
2.2 Klasifikace podle diskriminačních funkcí

2.2.1 Základní principy

Příslušnost do jednotlivých klasifikačních tříd v tomto případě vyjadřujeme pomocí R skalárních funkcí $g_1(\mathbf{x})$, $g_2(\mathbf{x})$, …, $g_R(\mathbf{x})$ takových, že pro obraz **x** z podmnožiny $\mathcal{R}_r$, o níž předpokládáme, že reprezentuje obrazy ze třídy ω_r, pro všechna r platí

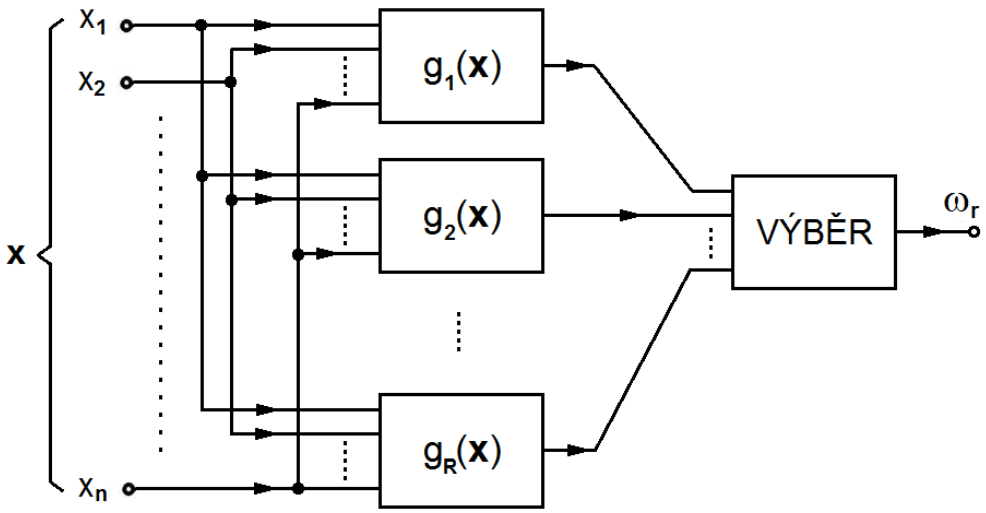
$$g_r(\mathbf{x}) > g_s(\mathbf{x}), \quad \text{pro } s = 1,2,...,R \text{ a } r \neq s. \tag{2.2}$$

Funkce $g_r(\mathbf{x})$ mohou vyjadřovat např. míru výskytu obrazu **x** patřícího do r-té klasifikační třídy v odpovídajícím místě obrazového prostoru. Nazýváme je diskriminační funkce a z analyticky geometrického hlediska definují plochy nad obrazovým prostorem. Pro jednorozměrný příznakový prostor a dvě klasifikační třídy je princip klasifikace pomocí diskriminačních funkcí zobrazen na obr.2.2.



Obr.2.2 *Princip metody pomocí diskriminačních funkcí*

Blokové schéma klasifikátoru založeného na metodě pomocí diskriminačních funkcí je na obr.2.3. Všechny příznaky x_1, x_2, …, x_n jsou současně přivedeny do R bloků, ve kterých se vyčíslí hodnoty diskriminačních funkcí $g_r(\mathbf{x})$, r = 1, 2, …, R. Na výstupu výběrového bloku se objeví identifikátor vybrané klasifikační třídy. Protože diskriminační funkce jsou definovány vztahem (2.2), je v případě deterministických klasifikátorů výběrový algoritmus definován jednoznačně výběrem maxima, zatímco u nedeterministických klasifikátorů je výběr určen nějakým nejednoznačným algoritmem. Buď se např. zvolí hodnota vý-



Obr.2.3 *Blokové schéma klasifikátoru pomocí diskriminačních funkcí*

stupní proměnné na základě některé varianty náhodného výběru, nebo mohou být jednotlivé výstupní hodnoty oceněny patřičnou mírou příslušnosti k daným klasifikačním třídám (např. velikostí pravděpodobnosti).

V případě klasifikace do dvou klasifikačních tříd, tzv. ***dichotomie***, pracuje klasifikátor pouze se dvěma diskriminačními funkcemi. Určujeme-li, která z obou funkcí má pro obraz **x** větší hodnotu, stačí zjistit znaménko funkce

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) \tag{2.3}$$

a výběrový blok pak reprezentuje nelineární příkaz

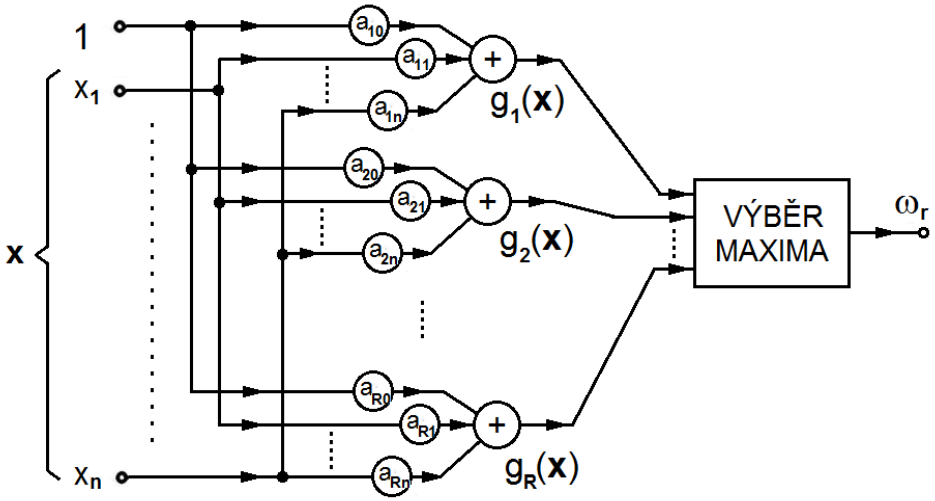
$$\omega_r = \text{sign}\big(g(\mathbf{x})\big), \tag{2.4}$$

pro který je $\omega_1 = 1$, když $g(x) \geq 0$, tj. předpokládáme, že $\mathbf{x} \in \mathcal{R}_1$ a $\omega_2 = -1$, když $g(x) < 0$, tj předpokládáme, že $\mathbf{x} \in \mathcal{R}_2$.

Nejjednodušším tvarem diskriminační funkce je funkce lineární, která má tvar

$$g_r(\mathbf{x}) = a_{r0} + a_{r1}x_1 + a_{r2}x_2 + \ldots + a_{rn}x_n, \tag{2.5}$$

kde a_{r0} je práh diskriminační funkce posouvající počátek souřadnicového systému a a_{ri}, i = 1, 2, … n jsou váhové koeficienty i-tého příznaku x_i. Schéma lineárního klasifikátoru je na obr.2.4.



Obr.2.4 *Schéma deterministického klasifikátoru s lineárními diskriminačními funkcemi*

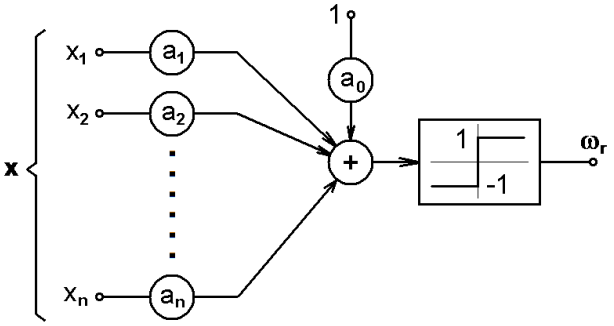
Diskriminační funkce lineárního dichotomického klasifikátoru má tvar (ze vztahů (2.3) a (2.5))

$$g(\mathbf{x}) = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n = \\ = a_0 + \mathbf{a}^T.\mathbf{x}, \tag{2.6}$$

kde vektor **a** je sloupcový vektor váhových koeficientů, pro jehož souřadnice a pro koeficient a_0 platí

$$a_i = a_{1i} - a_{2i}, \text{ pro } i = 0, 1, \ldots, n. \tag{2.7}$$

Blokové schéma takového klasifikátoru je na obr.2.5.



Obr.2.5 *Schéma dichotomického lineárního klasifikátoru*

2.2.2 Určení diskriminačních funkcí na základě statistických vlastností množiny obrazů – Bayesův klasifikátor

Při řešení praktických klasifikačních úloh je nutné předpokládat, že hodnoty příznaků jsou ovlivněny víceméně náhodnými fluktuacemi různého původu. Poloha příznakového obrazu (vektoru) je tedy ovlivněna všelijakým náhodným rušením, které způsobuje zvýšený rozptyl obrazů nejen v prostoru určité klasifikační třídy, nýbrž i vně tohoto prostoru, takže dochází k překrývání množin obrazů z různých klasifikačních tříd. Je zřejmé, že díky tomuto překrývání nebude klasifikace vždy bezchybná a chybná klasifikace může způsobit určitou ztrátu.

Možnost vyjádření ztráty při chybné klasifikaci nabízí tzv. ***ztrátová funkce*** $\lambda(\omega_r|\omega_s)$, udávající ztrátu vzniklou chybným zařazením obrazu do třídy ω_r, když ve skutečnosti patří do třídy ω_s. Pro celou klasifikační úlohu vyjádříme ztráty všech možných chybných klasifikací pomocí ***matice ztrátových funkcí*** (*cost matrix*)

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda(\omega_1|\omega_1) & \lambda(\omega_1|\omega_2) & \cdots & \lambda(\omega_1|\omega_R) \\ \lambda(\omega_2|\omega_1) & \lambda(\omega_2|\omega_2) & \cdots & \lambda(\omega_2|\omega_R) \\ \vdots & \vdots & \ddots & \vdots \\ \lambda(\omega_R|\omega_1) & \lambda(\omega_R|\omega_2) & \cdots & \lambda(\omega_R|\omega_R) \end{bmatrix}. \tag{2.8}$$

Tato matice je obecně nesymetrická. Např. pokud není u pacienta správně diagnostikován infarkt myokardu, může to pro dotyčného pacienta mít fatální důsledky. Naopak, ztráta bude naprosto jiná, bude-li u zdravého pacienta infarkt myokardu diagnostikován chybně.

Obecně mohou hodnoty ztrátové funkce záviset na obrazu $\mathbf{x}$, čehož formálně snadno dosáhneme dosazením za ω_r podle vztahu (2.1), tj. $\lambda(d(\mathbf{x})|\omega_s)$, nebo zahrneme-li mezi parametry rozhodovacího pravidla i nastavení parametrů klasifikátoru, vyjádřené vektorem $\mathbf{a}$, je ztrátová funkce vyjádřena jako $\lambda(d(\mathbf{x},\mathbf{a})|\omega_s)$. Často je výhodnější, než použití celé matice, vyjádřit kvalitu rozhodování klasifikátoru jedním jednoduchým zobecňujícím parametrem. K tomu účelu se používá tzv. ***střední ztráta*** $J(\mathbf{a})$, udávající průměrnou ztrátu při chybné klasifikaci obrazu $\mathbf{x}$.

Než se pustíme do podrobnějšího popisu jednotlivých metod, připomeňme pro méně zkušené čtenáře vztah, se kterým tyto metody pracují, tj. Bayesův vzorec ve tvaru a s interpretací, které jsou užitečné z pohledu výkladů v této kapitole. Tedy

$$P(\omega_r|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_r).P(\omega_r)}{p(\mathbf{x})}, \tag{2.9}$$

kde $P(\omega_r|\mathbf{x})$ je ***aposteriorní*** podmíněná ***pravděpodobnost*** zatřídění obrazového vektoru $\mathbf{x}$ do třídy ω_r, $p(\mathbf{x}|\omega_s)$ je podmíněná hustota pravděpodobnosti výskytu obrazů $\mathbf{x}$ ve třídě ω_r, $P(\omega_s)$ je ***apriorní pravděpodobnost*** třídy ω_r a konečně $p(\mathbf{x})$ je celková hustota pravděpodobnosti rozložení všech obrazů $\mathbf{x}$ v celém obrazovém prostoru.

Kritérium minimální střední ztráty

Pokud bychom se soustředili pouze na obrazy ze třídy ω_s, je střední ztráta dána průměrnou hodnotou z $\lambda(d(\mathbf{x},\mathbf{a})|\omega_s)$ vzhledem ke všem obrazům ze třídy ω_s, tj.

$$J_s(\mathbf{a}) = \int_X \lambda(d(\mathbf{x},\mathbf{a})|\omega_s).p(\mathbf{x}|\omega_s)d\mathbf{x}, \tag{2.10}$$

kde zopakujme, že $p(\mathbf{x}|\omega_s)$ je podmíněná hustota pravděpodobnosti výskytu obrazu $\mathbf{x}$ ve třídě ω_s.

Ve skutečnosti na vstup klasifikátoru přicházejí obrazy ze všech tříd, proto musíme celkovou střední ztrátu $J(\mathbf{a})$ stanovit jako průměrnou hodnotu ze ztrát $J_s(\mathbf{a})$. Tedy

$$J(\mathbf{a}) = \sum_{s=1}^{R} J_s(\mathbf{a}).P(\omega_s) = \int_{\mathcal{X}} \sum_{s=1}^{R} \lambda(d(\mathbf{x},\mathbf{a})|\omega_s).p(\mathbf{x}|\omega_s).P(\omega_s)d\mathbf{x}, \tag{2.11}$$

kde opět $P(\omega_s)$ je apriorní pravděpodobnost výskytu třídy ω_s, nebo podle Bayesova vzorce

$$J(\mathbf{a}) = \int_{\mathcal{X}} \sum_{s=1}^{R} \lambda(d(\mathbf{x},\mathbf{a})|\omega_s).p(\mathbf{x}).P(\omega_s|\mathbf{x}).d\mathbf{x}, \tag{2.12}$$

kde, jak již víme, je $p(\mathbf{x})$ hustota pravděpodobnosti výskytu obrazu $\mathbf{x}$ v celém obrazovém prostoru $\mathcal{X}$ a $P(\omega_s|\mathbf{x})$ je podmíněná pravděpodobnost toho, že daný obraz $\mathbf{x}$ patří do třídy ω_s (tzv. aposteriorní pravděpodobnost třídy ω_s). Celková střední ztráta $J(\mathbf{a})$ je tedy již pouze funkcí nastavení klasifikátoru. Návrh optimálního klasifikátoru, který by minimalizoval celkovou střední ztrátu, spočívá v nalezení takové množiny parametrů rozhodovacího pravidla $\mathbf{a}^*$, že platí

$$J(\mathbf{a}^*) = \min_{\forall \mathbf{a}} J(\mathbf{a}), \tag{2.13}$$

tj. $\mathbf{a}^*$ je vektor takových hodnot parametrů rozhodovacího pravidla, pro který je střední ztráta nejmenší. Dosadíme-li do (2.13) podle (2.11), dostaneme

$$J(\mathbf{a}^*) = \min_{\mathbf{a}} \int_{\mathcal{X}} \sum_{s=1}^{R} \lambda(d(\mathbf{x},\mathbf{a})|\omega_s).p(\mathbf{x}|\omega_s).P(\omega_s)d\mathbf{x} \tag{2.14}$$

a za předpokladu, že ztrátová funkce $\lambda(\omega_r|\omega_s)$ je konstantní pro všechny obrazy $\mathbf{x}$ ze třídy ω_s, platí dále

$$J(\mathbf{a}^*) = \int_{\mathcal{X}} \min_{r} \sum_{s=1}^{R} \lambda(\omega_r|\omega_s).p(\mathbf{x}|\omega_s).P(\omega_s)d\mathbf{x}\,. \tag{2.15}$$

Když označíme ztrátu při klasifikaci obrazu $\mathbf{x}$ do třídy ω_r

$$L_{\mathbf{x}}(\omega_r) = \sum_{s=1}^{R} \lambda(\omega_r|\omega_s).p(\mathbf{x}|\omega_s).P(\omega_s) \tag{2.16}$$

a dosadíme-li podle (2.16) do (2.15), dostaneme vztah

$$J(\mathbf{a}^*) = \int_{\mathcal{X}} \min_{r} L_{\mathbf{x}}(\omega_r).d\mathbf{x}\,. \tag{2.17}$$

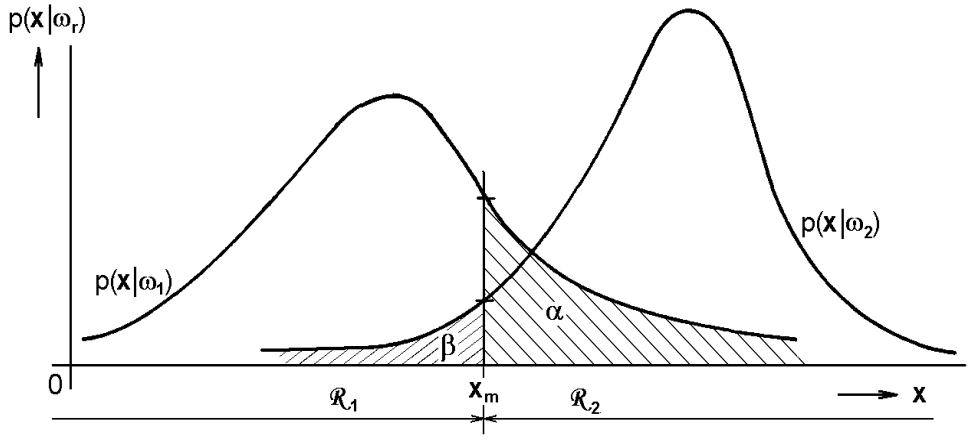
Úloha nalezení minima celkové střední ztráty se tímto postupem transformovala na minimalizaci funkce $L_{\mathbf{x}}(\omega_r)$. Optimální rozhodovací pravidlo $d(\mathbf{x},\mathbf{a}^*)$ podle ***kritéria minimální celkové střední ztráty*** (nebo podle v literatuře používaného názvu ***kritéria minimální chyby***, resp. ***Bayesova kritéria***) je určeno vztahem

$$L_{\mathbf{x}}(d_{ME}(\mathbf{x},\mathbf{a}^*)) = \min_{r} L_{\mathbf{x}}(\omega_r)\,. \tag{2.18}$$

Pokud chceme, tak jak se v této kapitole očekává, realizovat klasifikátor na principu diskriminačních funkcí, vyjdeme ze vztahu

$$\min L_{\mathbf{x}}(\omega_r) = \max(-L_{\mathbf{x}}(\omega_r))\,. \tag{2.19}$$

Diskriminační funkci optimálního klasifikátoru podle kritéria minimální chyby pak můžeme určit podle vztahu



Obr.2.6 *Pravděpodobnosti chybného zatřídění*

$$g_r(\mathbf{x}) = -L_{\mathbf{x}}(\omega_r) = -\sum_{s=1}^{R} \lambda(\omega_r|\omega_s).p(\mathbf{x}|\omega_s).P(\omega_s)\,. \tag{2.20}$$

Důsledky návrhu optimálního rozhodovacího pravidla podle kritéria minimální chyby si nyní demonstrujme na jednoduchém příkladu dichotomického klasifikátoru.

Celková střední ztráta při klasifikaci do dvou tříd bude

$$\begin{aligned} J(\mathbf{a}) &= \int_{\mathcal{R}_1} \sum_{s=1}^{2} \lambda(\omega_1|\omega_s).p(\mathbf{x}|\omega_s).P(\omega_s).d\mathbf{x} + \int_{\mathcal{R}_2} \sum_{s=1}^{2} \lambda(\omega_2|\omega_s).p(\mathbf{x}|\omega_s).P(\omega_s).d\mathbf{x} = \\ &= \lambda(\omega_1|\omega_1).P(\omega_1).\int_{\mathcal{R}_1} p(\mathbf{x}|\omega_1).dx + \lambda(\omega_1|\omega_2).P(\omega_2).\int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2).dx + \\ &\quad + \lambda(\omega_2|\omega_1).P(\omega_1).\int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1).dx + \lambda(\omega_2|\omega_2).P(\omega_2).\int_{\mathcal{R}_2} p(\mathbf{x}|\omega_2).dx = \\ &= \lambda(\omega_1|\omega_1).P(\omega_1).(1-\alpha) + \lambda(\omega_1|\omega_2).P(\omega_2).\beta + \\ &\quad + \lambda(\omega_2|\omega_1).P(\omega_1).\alpha + \lambda(\omega_2|\omega_2).P(\omega_2).(1-\beta)\,, \end{aligned} \tag{2.21}$$

kde α a β jsou pravděpodobnosti chybného rozhodnutí odpovídající vyšrafovaným plochám na obr.2.6 a vyjádřené vztahy

$$\alpha = \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)d\mathbf{x} \text{ a } \beta = \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)d\mathbf{x}. \tag{2.22}$$

Diskriminační funkce pro dichotomický klasifikátor podle kritéria minimální chyby bude s pomocí vztahů (2.3) a (2.20)

$$\begin{aligned} g(\mathbf{x}) &= g_1(\mathbf{x}) - g_2(\mathbf{x}) = -L_{\mathbf{x}}(\omega_1) + L_{\mathbf{x}}(\omega_2) = \\ &= -\lambda(\omega_1|\omega_1).p(\mathbf{x}|\omega_1).P(\omega_1) - \lambda(\omega_1|\omega_2).p(\mathbf{x}|\omega_2).P(\omega_2) + \\ &\quad + \lambda(\omega_2|\omega_1).p(\mathbf{x}|\omega_1).P(\omega_1) + \lambda(\omega_2|\omega_2).p(\mathbf{x}|\omega_2).P(\omega_2) = \\ &= \left[\lambda(\omega_2|\omega_1) - \lambda(\omega_1|\omega_1)\right]p(x|\omega_1).P(\omega_1) + \left[\lambda(\omega_2|\omega_2) - \lambda(\omega_1|\omega_2)\right]p(x|\omega_2).P(\omega_2)\,. \end{aligned} \tag{2.23}$$

Položíme-li výsledný výraz ve vztahu (2.23) roven nule, dostaneme výraz pro hraniční plochu dichotomického klasifikátoru, ze kterého můžeme určit poměr hustot pravděpodobnosti obrazu **x** v každé z obou klasifikačních tříd, jenž nazýváme ***věrohodnostní poměr*** [11]

$$\Lambda_{12} = \frac{p(x|\omega_1)}{p(x|\omega_2)} = \frac{\left[\lambda(\omega_1|\omega_2) - \lambda(\omega_2|\omega_2)\right]P(\omega_2)}{\left[\lambda(\omega_2|\omega_1) - \lambda(\omega_1|\omega_1)\right]P(\omega_1)}. \tag{2.24}$$

Podle vztahu (2.24) zařadíme obraz **x** do třídy ω_1, když je věrohodnostní poměr větší než výraz na pravé straně; je-li menší, pak obraz **x** zařadíme do třídy ω_2. Hraniční plocha prochází právě těmi body **x** obrazového prostoru, pro které platí rovnost definovaná vztahem (2.24).

Výše uvedené odvození kritéria minimální střední ztráty předpokládalo relativně obecné podmínky. Jediné zjednodušení vycházelo z předpokladu, že ztrátová funkce $\lambda(\omega_r|\omega_s)$ je konstantní pro všechny obrazy **x** ze třídy ω_s. Při řešení praktických úloh se však často setkáváme se situacemi, kdy je velice obtížné zjistit všechny informace, potřebné k realizaci tohoto kritéria. Proto se nyní seznamme, jak se kritérium mění, nejsou-li k dispozici všechny požadované údaje.

Kritérium minimální pravděpodobnosti chybného rozhodnutí

Vzhledem k obtížnosti stanovení hodnot ztrátových funkcí $\lambda(\omega_r|\omega_s)$ se kritérium minimální chyby zjednodušuje použitím jednotkových ztrátových funkcí definovaných

$$\lambda(\omega_r|\omega_s) = \begin{cases} 0; & \text{pro } r = s; \\ 1; & \text{pro } r \neq s. \end{cases} \tag{2.25}$$

Matice jednotkových ztrátových funkcí má pak tvar

$$\boldsymbol{\lambda} = \begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{bmatrix} \tag{2.26}$$

a celková střední ztráta s pomocí vztahů (2.11) a (2.25) je

$$J(\mathbf{a}) = \sum_{\substack{s=1 \\ s\neq r}}^{R} \int_{\mathcal{X}-\mathcal{R}_s} p(\mathbf{x}|\omega_s).P(\omega_s)d\mathbf{x}, \tag{2.27}$$

což udává hodnotu pravděpodobnosti chybného rozhodnutí klasifikátoru. Minimalizací celkové střední ztráty v tomto případě určíme parametry **a*** klasifikátoru rozhodujícího s nejmenší pravděpodobností chybného rozhodnutí.

Dosadíme-li (2.25) do (2.16), dostaneme

$$L_{\mathbf{x}}(\omega_r) = \sum_{\substack{s=1 \\ s\neq r}}^{R} p(\mathbf{x}|\omega_s).P(\omega_s) = \sum_{s=1}^{R} p(\mathbf{x}|\omega_s).P(\omega_s) - p(\mathbf{x}|\omega_r).P(\omega_r) \tag{2.28}$$

a s využitím Bayesova vztahu je dále

---

[11] ***Věrohodnostní poměr*** *(likelihood ratio) LR* udává podíl pravděpodobnosti, že se vyskytne nějaký jev *A* za určité podmínky (jev *B*), k pravděpodobnosti, že se jev *A* vyskytne, když podmínka neplatí (jev nonB). Má-li například pacient náhlou ztrátu paměti (jev *A*), chceme znát věrohodnostní poměr výskytu jevu *A* v případě, že má mozkový nádor (jev *B*), tj. podíl pravděpodobnosti, s jakou ztráta paměti vzniká při nádoru mozku, k pravděpodobnosti, s jakou vzniká v ostatních případech. Věrohodnostní poměr je tedy podíl podmíněných pravděpodobností LR=P(A | B)/P(A | nonB).

$$L_{\mathbf{x}}(\omega_r) = p(\mathbf{x}).\sum_{s=1}^{R} P(\omega_s|\mathbf{x}) - p(\mathbf{x}|\omega_r).P(\omega_r) = p(\mathbf{x}) - p(\mathbf{x}|\omega_r).P(\omega_r)\,. \tag{2.29}$$

Hustota pravděpodobnosti $p(\mathbf{x})$ nezávisí na klasifikační třídě, pro daný obraz je konstantní pro všechna $L_{\mathbf{x}}(\omega_r)$ a tedy neovlivňuje výběr minima. Proto můžeme stanovit diskriminační funkci jako

$$g_r(\mathbf{x}) = p(\mathbf{x}|\omega_r).P(\omega_r)\,. \tag{2.30}$$

V případě dichotomického klasifikátoru je diskriminační funkce

$$g(\mathbf{x}) = p(\mathbf{x}|\omega_1).P(\omega_1) - p(\mathbf{x}|\omega_2).P(\omega_2)\,. \tag{2.31}$$

Z tohoto vztahu můžeme určit věrohodnostní poměr Λ_{12}, který určuje hranici mezi dichotomickými klasifikačními třídami podle kritéria minimální pravděpodobnosti chybného rozhodnutí

$$\Lambda_{12} = \frac{p(x|\omega_1)}{p(x|\omega_2)} = \frac{P(\omega_2)}{P(\omega_1)}\,. \tag{2.32}$$

Kritérium maximální aposteriorní pravděpodobnosti

Modifikujeme-li vztah (2.16) pro ztrátu při klasifikaci obrazu $\mathbf{x}$ do třídy ω_r podle Bayesova vztahu (tj. $P(\omega_s|\mathbf{x}).p(\mathbf{x}) = p(\mathbf{x}|\omega_s).P(\omega_s)$), platí, že

$$L_{\mathbf{x}}(\omega_r) = \sum_{s=1}^{R} \lambda(\omega_r|\omega_s).p(\mathbf{x}).P(\omega_s|\mathbf{x}) = p(\mathbf{x}).\sum_{s=1}^{R} \lambda(\omega_r|\omega_s).P(\omega_s|\mathbf{x})\,. \tag{2.33}$$

Když opět využijeme toho, že hustota pravděpodobnosti $p(\mathbf{x})$ výskytu obrazu $\mathbf{x}$ v celém obrazovém prostoru nezávisí na klasifikační třídě, její odstranění neovlivní konstrukci rozhodovacího pravidla. Lze tedy místo $L_{\mathbf{x}}(\omega_r)$ použít proměnnou $L'_{\mathbf{x}}(\omega_r)$ určenou vztahem

$$L'_x(\omega_r) = \frac{L_x(\omega_r)}{p(\mathbf{x})} = \sum_{s=1}^{R} \lambda(\omega_r|\omega_s).P(\omega_s|\mathbf{x})\,. \tag{2.34}$$

Uvažujeme-li znovu nejjednodušší volbu hodnot ztrátových funkcí, tj. jednotkové ztrátové funkce, je

$$L'_x(\omega_r) = \sum_{\substack{s=1 \\ s\neq r}}^{R} P(\omega_s|\mathbf{x}) = \sum_{s=1}^{R} P(\omega_s|\mathbf{x}) - P(\omega_r|\mathbf{x}) = 1 - P(\omega_r|\mathbf{x}). \tag{2.35}$$

Minimum ztráty $L'_{\mathbf{x}}(\omega_r)$ nalezneme právě tehdy, když $P(\omega_r|\mathbf{x})$ bude maximální. To znamená, že jako diskriminační funkci můžeme volit právě hodnotu aposteriorní pravděpodobnosti třídy ω_r, tj.

$$g_r(\mathbf{x}) = P(\omega_r|\mathbf{x})\,. \tag{2.36}$$

Konečně se opět zabývejme případem klasifikace do dvou klasifikačních tříd. Diskriminační funkce je při tom definována

$$g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x}) = 0\,. \tag{2.37}$$

Z toho dále platí, že hranici mezi dvěma třídami určuje vztah

$$P(\omega_1|\mathbf{x}) = P(\omega_2|\mathbf{x}), \tag{2.38}$$

nebo

$$\frac{P(\omega_1|\mathbf{x})}{P(\omega_2|\mathbf{x})} = 1. \tag{2.39}$$

Podle tohoto kritéria, tzv. kritéria maximální aposteriorní pravděpodobnosti zatřídíme obraz **x** logicky do té třídy, jejíž pravděpodobnost je při výskytu obrazu **x** větší.

Z odvození tohoto kritéria a kritéria minimální pravděpodobnosti chyby vyplývá, že jsou obě kritéria rovnocenná.

Kritérium maximální pravděpodobnosti

Všechna dosud uvedená optimální kritéria vycházela ze znalosti hustoty pravděpodobnosti výskytu obrazů **x** ve všech klasifikačních třídách $p(\mathbf{x}|\omega_s)$ a apriorních pravděpodobností všech tříd $P(\omega_s)$. Pokud nemáme informaci o výskytu klasifikačních tříd, předpokládáme rovnoměrné rozložením, tedy všechny třídy jsou stejně pravděpodobné ($P(\omega_s) = P(\omega) = 1/R$). Potom celková střední ztráta

$$J(\mathbf{a}) = \frac{1}{R}\sum_{s=1}^{R}\int_{\mathcal{X}} \lambda(\omega_r|\omega_s).p(\mathbf{x}|\omega_s).d\mathbf{x} \tag{2.40}$$

dosáhne minima, když

$$J(\mathbf{a}^*) = \frac{1}{R}\min_{\forall \mathbf{a}} \int_{\mathcal{X}} \sum_{s=1}^{R} \lambda(\omega_r|\omega_s).p(\mathbf{x}|\omega_s).d\mathbf{x}\,. \tag{2.41}$$

Diskriminační funkci pomocí ztráty při klasifikaci obrazu x do třídy ω_r můžeme podobně jako v (2.20) určit ze vztahu

$$g_r(\mathbf{x}) = -L_{\mathbf{x}}(\omega_r) = -\sum_{s=1}^{R}\lambda(\omega_r|\omega_s).p(\mathbf{x}|\omega_s)\,. \tag{2.42}$$

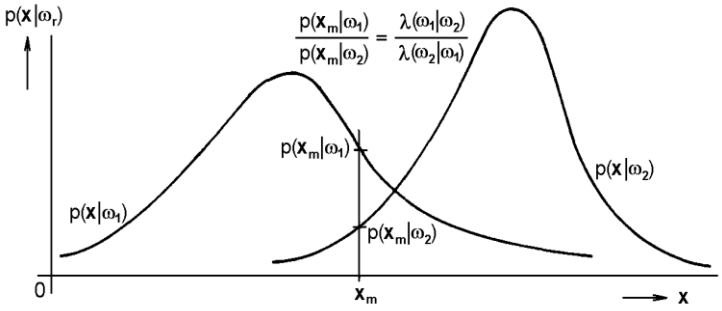
V případě dichotomie je věrohodnostní poměr

$$\Lambda_{12} = \frac{p(x|\omega_1)}{p(x|\omega_2)} = \frac{\left[\lambda(\omega_1|\omega_2) - \lambda(\omega_2|\omega_2)\right]}{\left[\lambda(\omega_2|\omega_1) - \lambda(\omega_1|\omega_1)\right]}. \tag{2.43}$$

Když jsou ceny správného rozhodnutí nulové, tj. $\lambda(\omega_1|\omega_1) = \lambda(\omega_2|\omega_2) = 0$, je

$$\Lambda_{12} = \frac{p(x|\omega_1)}{p(x|\omega_2)} = \frac{\lambda(\omega_1|\omega_2)}{\lambda(\omega_2|\omega_1)}. \tag{2.44}$$

Obraz **x** je zařazen do třídy ω_1, když je věrohodnostní poměr větší než poměr cen ztrát chybných zatřídění (obr.2.7). Jsou-li obě ceny stejné, tj. i jednotkové, je obraz **x** zařazen do té třídy, pro kterou je hodnota hustoty pravděpodobnosti $p(\mathbf{x}|\omega_s)$ větší.



Obr.2.7 *Stanovení klasifikační hranice pro dichotomický klasifikátor podle kritéria maximální pravděpodobnosti*

2.3 Klasifikace podle minimální vzdálenosti

2.3.1 Základní principy

Jak již bylo uvedeno, prostor odpovídající jednotlivým klasifikačním třídám můžeme v obrazovém prostoru vymezit kromě diskriminačních funkcí a hraničních ploch rovněž polohou reprezentativních obrazů – etalonů. Je-li v obrazovém prostoru zadáno R etalonů vektory $\mathbf{x}_{1E}$, $\mathbf{x}_{2E}$, …, $\mathbf{x}_{RE}$, zařadí klasifikátor podle minimální vzdálenosti klasifikovaný obraz $\mathbf{x}$ do té třídy, jejíž etalon má od bodu $\mathbf{x}$ nejmenší vzdálenost. Rozhodovací pravidlo je tedy určeno vztahem

$$\omega_r = d(\mathbf{x}) = \|\mathbf{x}_{rE} - \mathbf{x}\| = \min_{\forall s} \|\mathbf{x}_{sE} - \mathbf{x}\|. \tag{2.45}$$

Pokud je i nejmenší možná vzdálenost příliš velká, lze rozhodovací pravidlo upravit do tvaru

$$\text{když } d(\mathbf{x}) = \|\mathbf{x}_{rE} - \mathbf{x}\| = \min_{\forall s} \|\mathbf{x}_{sE} - \mathbf{x}\| \le T, \text{ pak } \omega_r, \text{ jinak } \omega_{R+1}, \tag{2.46}$$

kde T je určená prahová hodnota a ω_{R+1} reprezentuje klasifikační třídu, která obsahuje vzorky, které klasifikátor neumí zatřídit.

Vzdálenost je hodnota, kterou můžeme považovat za míru nepodobnosti. Čím je vzdálenost mezi dvěma objekty větší, tím méně jsou si podobny. Duální mírou ke vzdálenosti je ***podobnost*** – čím větší je podobnost dvou objektů, tím bližší si tyto objekty jsou.

Klasifikace podle minimální vzdálenosti bývá někdy až příliš automaticky a samozřejmě spojována pouze se shlukováním. Shlukovací algoritmy, které jsou detailněji popsány např. v [3], představují typický klasifikační ***učící se*** algoritmus, který samočinně modifikuje vlastnosti klasifikační třídy s klasifikací každého nového obrazu. Pokud ale definici klasifikační třídy (její etalon) během klasifikačního procesu neměníme, pak lze klasifikaci podle minimální vzdálenosti považovat za klasický neučící se algoritmus, byť se zde používá pojmu „etalon“, jehož výskyt se v některých odborných zdrojích považuje za pevně vázaný se sebeučícími se algoritmy.

2.3.2 Metrika, vzdálenost, podobnost

Abychom uměli obrazy podle vzdálenosti, resp. podobnosti klasifikovat, je potřeba umět vzdálenost/podobnost spočítat. Jinými slovy je třeba znát předpis (algoritmus, funkci), na základě kterého vzdálenost/podobnost počítáme. Způsob výpočtu vzdálenosti, resp. podobnosti bude záležet na mnoha okolnostech – na způsobu, jakým matematicky popíšeme analyzovaný objekt, na charakteru (typu) dat, samozřejmě i na vlastnostech předpisu, podle kterého vzdálenost, resp. podobnost počítáme.

Zabývejme se nejdříve vzdáleností.

Metrikou ρ na $\mathcal{X}$ nazýváme takovou funkci $\rho: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, kde $\mathcal{X}$ je n-rozměrný obrazový prostor a $\mathbb{R}$ je množina reálných čísel, splňující následující předpoklady:

$$\exists \rho_0 \in \mathbb{R}: -\infty < \rho_0 \le \rho(\mathbf{x},\mathbf{y}) < +\infty, \forall \mathbf{x},\mathbf{y} \in \mathcal{X}; \tag{2.47}$$ [12]

[12] Konstanta ρ_0 je tomto případě definována velice obecně. V praktických případech je rovna nule, což znamená, že vzdálenost nabývá nezáporných hodnot s tím, že pokud jsou oba obrazy (vektory) totožné, je vzdálenost nulová.

$$\rho(\mathbf{x},\mathbf{x}) = \rho_0,\ \forall \mathbf{x} \in \mathcal{X}$$

a má následující vlastnosti

$$\begin{gathered}\rho(\mathbf{x},\mathbf{y}) = \rho(\mathbf{y},\mathbf{x}),\ \forall\ \mathbf{x},\mathbf{y} \in \mathcal{X}\ \text{(symetrie)};\\ \rho(\mathbf{x},\mathbf{y}) = \rho_0\ \text{když a jen když}\ \mathbf{x} = \mathbf{y}\ \text{(totožnost)}.\end{gathered} \tag{2.48}$$

Pokud navíc platí i trojúhelníková nerovnost

$$\rho(\mathbf{x},\mathbf{z}) \le \rho(\mathbf{x},\mathbf{y}) + \rho(\mathbf{y},\mathbf{z}),\ \forall\ \mathbf{x},\mathbf{y},\mathbf{z} \in \mathcal{X}, \tag{2.49}$$

nazýváme metriku ***pravou metrikou***. Prostor $\mathcal{X}$, ve kterém je metrika ρ definována, označujeme jako ***metrický prostor***. ***Vzdálenost*** je pak hodnota určená podle metriky.

Pokud se týče podobností, pak ***metrikou podobnosti*** σ na $\mathcal{X}$ je taková funkce σ: $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$, kde $\mathcal{X}$ je opět n-rozměrný obrazový prostor a $\mathbb{R}$ je množina reálných čísel, splňující následující předpoklady:

$$\begin{gathered}\exists \sigma_0 \in \mathrm{R}\colon -\infty < \sigma(\mathbf{x},\mathbf{y}) \le \sigma_0 < +\infty,\ \forall\ \mathbf{x},\mathbf{y} \in \mathcal{X};\\ \sigma\,(\mathbf{x},\mathbf{x}) = \sigma_0,\ \forall \mathbf{x} \in \mathcal{X}\end{gathered} \tag{2.50}$$

a má stejně jako v předešlém případě následující vlastnosti

$$\begin{gathered}\sigma(\mathbf{x},\mathbf{y}) = \sigma(\mathbf{y},\mathbf{x}),\ \forall\ \mathbf{x},\mathbf{y} \in \mathrm{X}\ \text{(symetrie)}\\ \text{a}\\ \sigma(\mathbf{x},\mathbf{y}) = \sigma_0\ \text{když a jen když}\ \mathbf{x} = \mathbf{y}\ \text{(totožnost)}.\end{gathered} \tag{2.51}$$

V případě trojúhelníkové nerovnosti je situace poněkud složitější. Vztahy primárně vycházejí ze vztahu pro trojúhelníkovou nerovnost definovanou pro metriku vzdálenosti a její konkrétní tvar pro metriku podobnosti souvisí se základním vztahem mezi podobností a vzdáleností. Tento vztah může být vyjádřen např. pomocí následujících formulí:

$$\sigma = \frac{1}{\rho}; \tag{2.52}$$

$$\sigma = \frac{1}{1+\rho}, \tag{2.53}$$

nebo

$$\sigma = \mathrm{c} - \rho,\ \text{když}\ \mathrm{c} \ge \rho_{\max}. \tag{2.54}$$

Pokud se hodnoty vzdáleností pohybují v intervalu $\langle 0, \infty)$, pak v případě vztahu (2.52) se hodnoty podobnosti nacházejí také v tomto intervalu a výraz pro trojúhelníkovou nerovnost (2.49) se transformuje do tvaru

$$\sigma(\mathbf{x},\mathbf{y}).\ \sigma(\mathbf{y},\mathbf{z}) \le [\sigma(\mathbf{x},\mathbf{y}) + \sigma(\mathbf{y},\mathbf{z})].\ \sigma(\mathbf{x},\mathbf{z}),\ \forall\ \mathbf{x},\mathbf{y},\mathbf{z} \in \mathcal{X}. \tag{2.55}$$

Řídí-li se relace mezi vzdáleností a podobností vztahem (2.53), pak se hodnoty podobnosti vyskytují v intervalu $\langle 0,1\rangle$ a trojúhelníková nerovnost má tvar

$$\sigma(\mathbf{x},\mathbf{y}).\ \sigma(\mathbf{y},\mathbf{z}) \le [\sigma(\mathbf{x},\mathbf{y}) + \sigma(\mathbf{y},\mathbf{z}) - \sigma(\mathbf{x},\mathbf{y}).\ \sigma(\mathbf{y},\mathbf{z})].\ \sigma(\mathbf{x},\mathbf{z}),\ \forall\ \mathbf{x},\mathbf{y},\mathbf{z} \in \mathcal{X}. \tag{2.56}$$

Konečně, v případě formule (2.54) spadají hodnoty podobnosti do intervalu $\langle \mathrm{c} - \rho_{\max}, \mathrm{c}\rangle$ a trojúhelníková nerovnost se změní na

$$\sigma(\mathbf{x},\mathbf{z}) \ge \sigma(\mathbf{x},\mathbf{y}) + \sigma(\mathbf{y},\mathbf{z}) - \mathrm{c},\ \forall\ \mathbf{x},\mathbf{y},\mathbf{z} \in \mathcal{X}. \tag{2.57}$$

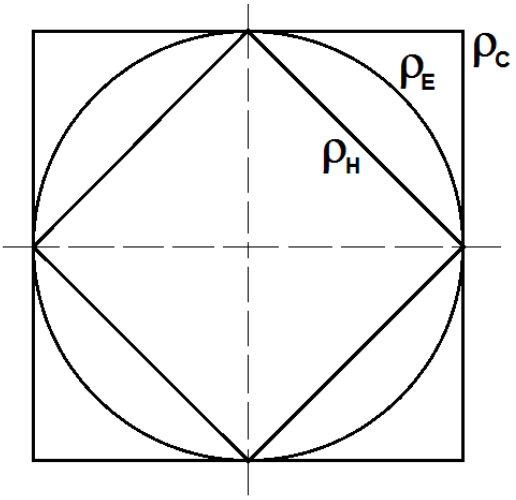
2.3.3 Metriky pro určení vzdálenosti mezi dvěma obrazy s kvantitativními příznaky

Použití konkrétní metriky závisí vždy na řešené úloze, a pokud se používá klasifikace podle minimální vzdálenosti, pak rozhodujícím kritériem pro posouzení vhodnosti té které metriky musí být kvalita výsledků klasifikace. Kromě tohoto základního kritéria, lze při výběru možné metriky použít i další dílčí kritéria, jako např. výpočetní nároky, charakter rozložení dat, apod. Obecně nelze doporučit vhodný postup pro výběr optimální metriky ani pro úlohy určitých standardních typů.

Euklidova metrika je definována vztahem

$$\rho_E(\mathbf{x}_1,\mathbf{x}_2)=\left[\sum_{i=1}^{n}(x_{1i}-x_{2i})^2\right]^{1/2}. \qquad (2.58)$$

Je to metrika zřejmě s nejnázornější geometrickou interpretací, geometrickým místem bodů s toutéž Euklidovou vzdáleností od daného bodu je koule (kruh ve dvourozměrném prostoru – viz obr.2.8). Kvadrát rozdílů souřadnic znamená, že klade větší důraz na větší rozdíly mezi souřadnicemi než v lineárním případě (což je třeba v každém konkrétním případě posoudit, zda je to stav žádoucí či nežádoucí). Pokud bychom počítali vzdálenost podle vztahu (2.58), ovšem bez použití odmocniny, tzv. kvadratická Euklidova vzdálenost, pak je výpočet určitě méně náročný, ale vztah nesplňuje trojúhelníkovou nerovnost. Vypočtené hodnoty lze považovat za míry nepodobnosti, ale výpočetní vztah není pravou metrikou. Kvadratickou euklidovskou vzdálenost lze tedy používat tehdy, kdy je rozhodující relativní porovnávání dvou hodnot (což klasifikace podle minimální vzdálenosti je), nikoliv absolutní hodnoty jako takové, což už by byl i případ klasifikace podle vztahu (2.46).



Obr.2.8 *Geometrická místa bodů se stejnou vzdáleností od souřadnicového počátku ve dvourozměrném příznakovém prostoru: ρ_E - Euklidova metrika, ρ_C - Čebyševova metrika, ρ_H – Hammingova metrika*

Příklad

Určete hodnoty euklidovské vzdálenosti a kvadratické euklidovské vzdálenosti pro dvourozměrné body **x** = (0,0), **y** = (5,0) a **z** = (6,2).

Euklidovská vzdálenost:

$$d_E(\mathbf{x},\mathbf{y})=\sqrt{5^2+0^2}=5;\ d_E(\mathbf{y},\mathbf{z})=\sqrt{1^2+2^2}=\sqrt{5}\cong 2{,}27;\ d_E(\mathbf{x},\mathbf{z})=\sqrt{6^2+2^2}=\sqrt{40}\cong 6{,}32.$$

Kvadratická euklidovská vzdálenost:

$$d_{E2}(\mathbf{x},\mathbf{y})=5^2+0^2=25;\quad d_{E2}(\mathbf{y},\mathbf{z})=1^2+2^2=5;\quad d_{E2}(\mathbf{x},\mathbf{z})=6^2+2^2=40.$$

Zatímco pro libovolné dvě hodnoty Euklidovy vzdálenosti platí, že jejich součet je větší než zbývající hodnota, v případě kvadratické Euklidovy vzdálenosti je $d_{E2}(\mathbf{x},\mathbf{y}) + d_{E2}(\mathbf{y},\mathbf{z})$ není větší nebo rovno než $d_{E2}(\mathbf{x},\mathbf{z})$. □□□

Hammingova metrika, také nazývána manhattanská metrika, nebo v angličtině *city-block* metrika, resp. *taxi driver* metrika, protože svým výpočtem ve dvourozměrném prostoru navozuje představu vzdálenosti, kterou urazí automobil jedoucí z jednoho místa do druhého v pravoúhle zastavěném městském prostředí. Je definována vztahem

$$\rho_H(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{n} |x_{1i} - x_{2i}| . \qquad (2.59)$$

Hammingova metrika je vytvořena linearizací Euklidovy metriky, což má za následek jednak snížení významu členů s větším rozdílem mezi dílčími souřadnicemi obou vektorů, jednak snížení výpočetní pracnosti vůči Euklidově metrice. Absolutní hodnota je nezbytná pro zachování kladné výsledné hodnoty vzdálenosti. Geometrickým místem bodů s toutéž Hammingovou vzdáleností od počátku v dvourozměrném prostoru je čtverec uvnitř Euklidovy kružnice (viz obr.2.8). Jak posléze uvidíme (kap.2.3.5 a 4.3.3) má Hammingova metrika použití i při posuzování vzdáleností dvou binárních vektorů, resp. řetězců stejné délky. Uplatňuje se i při hodnocení podobnosti dvou objektů, příp. jevů pomocí asociačních koeficientů (kap.2.3.6).

Minkovského metrika je definována vztahem

$$\rho_M(\mathbf{x}_1, \mathbf{x}_2) = \left[\sum_{i=1}^{n} |x_{1i} - x_{2i}|^m \right]^{1/m} . \qquad (2.60)$$

Zobecňuje Euklidovu nebo v podstatě i Hammingovu metriku. Místo druhé mocniny, příp. odmocniny, je použita mocnina i odmocnina obecná. To znamená, že zvyšuje váhu vlivu členů s větším rozdílem dílčích souřadnic obou obrazů. Čím větší mocnina, tím větší důraz na velké rozdíly mezi příznaky.

Čebyševova metrika je definována vztahem

$$\rho_C(\mathbf{x}_1, \mathbf{x}_2) = \max_{\forall i} |x_{1i} - x_{2i}| . \qquad (2.61)$$

Je limitním případem Minkovského metriky, protože platí

$$\rho_C(\mathbf{x}_1, \mathbf{x}_2) = \lim_{m \to \infty} \rho_M(\mathbf{x}_1, \mathbf{x}_2) . \qquad (2.62)$$

Používá se ve výpočetně kriticky náročných případech, kdy je pracnost výpočtu dle euklidovsky orientovaných metrik nepřijatelná. Geometrickým místem bodů s toutéž čebyševovskou vzdáleností od daného bodu je krychle, čtverec ve dvourozměrném prostoru (obr.2.8). Prostor mezi kružnicí euklidovské metriky ρ_E a čtvercem Čebyševovy metriky ρ_C vyplňují křivky Minkovského metriky pro různé hodnoty parametru $m > 2$.

Pokud je potřeba použít „euklidovskou“ metriku, ale s nižší výpočetní náročností, používá se v první řadě Hammingova nebo Čebyševova metrika. Možným přiblížením může být také kombinace obou metrik.

Vzdálenost určenou podle Hammingovy metriky lze považovat za dolní odhad vzdálenosti podle Euklidovy metriky a vzdálenost podle Čebyševovy metriky za její horní odhad.

Všechny uvedené metriky mají mnohé společné nevýhody. Jednak to, že je fyzikálně nesmyslné vytvářet součet rozdílů veličin s různým fyzikálním rozměrem, jednak to, že jsou-li začleněny příznakové veličiny do výsledné vzdálenosti se stejnými vahami, zvyšuje to vliv korelovaných veličin na celkový výsledek.

Tyto nevýhody mohou být odstraněny vhodnou transformací proměnných. Vliv různých fyzikálních veličin lze odstranit vztažením jejich hodnot k nějakému vyrovnávacímu faktoru, např. střední hodnotě $\bar{x}$, směrodatné odchylce σ_x, normě daného obrazu definované pro obraz $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ jako

$$\|\mathbf{x}\| = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}, \tag{2.63}$$

rozpětí $\Delta_i = \max_j x_{ij} - \min_j x_{ij}$, resp. standardizací podle vztahu (někdy také nazývaného z-skóre)

$$u_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}, \; i = 1, \ldots, n; \; j = 1, \ldots, K. \tag{2.64}$$

Můžeme také buď čistě subjektivně, nebo lépe na základě nějaké objektivní apriorní informace přiřadit každé příznakové proměnné koeficient udávající váhu této proměnné při výpočtu vzdálenosti. Např. vztah pro Minkovského metriku se váhováním mění na

$$\rho_{WM}(\mathbf{x}_1, \mathbf{x}_2) = \left[\sum_{i=1}^{n} a_i . \left| x_{1i} - x_{2i} \right|^m \right]^{1/m}. \tag{2.65}$$

Transformaci pomocí váhových koeficientů lze vyjádřit maticovým zápisem

$$\mathbf{u} = \mathbf{C}^T . \mathbf{x}, \tag{2.66}$$

kde koeficienty transformační matice **C** jsou dány

$$\begin{aligned} c_{ii} &= a_i, \text{ pro } i = 1, \ldots, n; \\ c_{ij} &= 0, \text{ pro } i \neq j. \end{aligned} \tag{2.67}$$

S takovým vyjádřením transformace příznakových proměnných je váhovaná Euklidova metrika definována vztahem

$$\rho_{WE}(\mathbf{x}_1, \mathbf{x}_2) = \left[(\mathbf{x}_1 - \mathbf{x}_2)^T . \mathbf{C} . \mathbf{C}^T . (\mathbf{x}_1 - \mathbf{x}_2) \right]^{1/2}. \tag{2.68}$$

Pokud jsou složky transformovaného obrazu dány lineární kombinací více složek původního obrazu, není ani matice **C**, ani matice $\mathbf{C}.\mathbf{C}^T$ čistě diagonální . Použijeme-li místo matice $\mathbf{C}.\mathbf{C}^T$ inverzní kovarianční (disperzní) matici $\mathbf{K}^{-1}$ je vztah (2.68) definičním vztahem tzv. ***Mahalanobisovy metriky***

$$\rho_E(\mathbf{u}_1, \mathbf{u}_2) = \rho_{MA}(\mathbf{x}_1, \mathbf{x}_2) = \left[(\mathbf{x}_1 - \mathbf{x}_2)^T . \mathbf{K}^{-1} . (\mathbf{x}_1 - \mathbf{x}_2) \right]^{1/2}. \tag{2.69}$$

Kovarianční matice dvou sloupcových vektorů $\mathbf{x} = (x_1, \ldots, x_m)$ a $\mathbf{y} = (y_1, \ldots, y_n)$ je určena podle vztahu

$$\mathbf{K}(\mathbf{x}, \mathbf{y}) = E((\mathbf{x} - E\mathbf{x}).(\mathbf{y} - E\mathbf{y})^T) = [\mathrm{cov}(x_i, y_j)]_{m,n}. \tag{2.70}$$

Přestože použití kovarianční matice je pro Mahalanobisovu metriku naprosto dominantní, lze nalézt definice této metriky i s korelační maticí $E(\mathbf{x}.\mathbf{y}^T) = [\mathrm{corr}(x_i, y_j)]_{m,n}$. V tomto případě je to opět situace, kdy je potřeba posoudit, zda je pro řešenou úlohu více informace v datech obsahujících i jejich střední hodnotu či zda střední hodnota pouze překrývá důležitou informaci obsaženou pouze ve variabilitě dat.

Využívá-li výpočet vzdálenosti hodnot příznakových proměnných vztažených vůči rozdílům maximální a minimální hodnoty dané proměnné, pak na příklad Hammingova normovaná metrika je v tomto případě definovaná vztahem

$$\rho_{NHnx}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{n} \frac{\left| x_{1i} - x_{2i} \right|}{\max x_i - \min x_i}. \tag{2.71}$$

kde max x_i a min x_i jsou maximální a minimální hodnoty dané souřadnice. Pro rozšíření intervalu, ve kterém se hodnoty vzdálenosti vyskytují, existuje i její logaritmická varianta definovaná jako

$$\rho_G(\mathbf{x}_1, \mathbf{x}_2) = -\log_{10}\left(1 - \frac{1}{n}\sum_{i=1}^{n}\frac{|x_{1i} - x_{2i}|}{\max x_i - \min x_i}\right). \qquad (2.72)$$

Ve všech těchto případech je třeba pečlivě zvážit, zda transformací dat nepřicházíme o významnou část informace, potřebné při navazujícím zpracování dat. Např. při použití Mahalanobisovy metriky, tak i při použití proměnných vztažených ke směrodatné odchylce, je potlačen vliv rozptylů příznakových proměnných na výslednou hodnotu vzdálenosti, což může mít na jedné straně příznivý, na druhé i nepříznivý vliv na získané výsledky a jejich interpretaci. Je potřeba si i uvědomit, že hodnota např. Mahalanobisovy metriky nebo normované Hammingovy metriky ρ_{NHnx} definované vztahem (2.71), příp. i metriky ρ_G nezávisí pouze na poloze vektorů $\mathbf{x}_1$ a $\mathbf{x}_2$, nýbrž i na vlastnostech prostoru vektorů $\mathcal{X}$. To znamená, že nabývá-li na příklad vzdálenost určená metrikou ρ_G hodnoty $d_G(\mathbf{x}_1,\mathbf{x}_2)$ v prostoru $\mathcal{X}$ a hodnoty $d'_G(\mathbf{x}_1,\mathbf{x}_2)$ v prostoru $\mathcal{X}'$, pak obecně $d_G(\mathbf{x}_1,\mathbf{x}_2) \neq d'_G(\mathbf{x}_1,\mathbf{x}_2)$.

Příklad

Mějme dva třírozměrné vektory $\mathbf{x}_1 = (0,1,2)^T$ a $\mathbf{x}_2 = (4,3,2)^T$. Pak za předpokladu neváhovaných metrik je $d_H(\mathbf{x}_1,\mathbf{x}_2) = 6$, $d_E(\mathbf{x}_1,\mathbf{x}_2) = 2.\sqrt{5}$ a $d_C(\mathbf{x}_1,\mathbf{x}_2) = 4$. Všimněme si, že $d_H(\mathbf{x}_1,\mathbf{x}_2) > d_E(\mathbf{x}_1,\mathbf{x}_2) > d_C(\mathbf{x}_1,\mathbf{x}_2)$.

Nyní předpokládejme, že tyto vektory patří do prostoru $\mathcal{X}$, který obsahuje vektory s maximálními hodnotami jednotlivých příznakových proměnných $\mathbf{x}_{max} = (10; 12; 13)^T$ a minimálními hodnotami příznaků $\mathbf{x}_{min} = (0; 0{,}5; 1)^T$. Pak $d_G(\mathbf{x}_1,\mathbf{x}_2) = 0{,}0922$. Pokud ale vektory $\mathbf{x}_1$ a $\mathbf{x}_2$ patří do prostoru vektorů $\mathcal{X}'$ = s maximálními hodnotami příznakových proměnných $\mathbf{x}'_{max} = (20; 22; 23)^T$ minimálními hodnotami $\mathbf{x}'_{max} = (-10; -9{,}5; -9)^T$, pak $d'_G(\mathbf{x}_1,\mathbf{x}_2) = 0{,}0295$. □□□

Relativizovanou variantou Hammingovy metriky je i tzv. ***canberrská metrika*** daná formulí[13]

$$\rho_{CA}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{n}\frac{|x_{1i} - x_{2i}|}{|x_{1i}| + |x_{2i}|}. \qquad (2.73)$$

Jednotlivé zlomky jsou z intervalu $\langle 0; 1\rangle$, celkový součet ale může být větší než 1. Je-li hodnota jednoho příznaku nulová, je dílčí zlomek roven jedné bez ohledu na druhou hodnotu. Jedničce se rovná dílčí zlomek i v případě, kdy obě souřadnice mají tutéž hodnotu, ale s opačným znaménkem. Jsou-li hodnoty obou příznaků ve zlomku nulové, pak předpokládáme, že i hodnota zlomku je nulová (někdy se z praktických výpočetních důvodů nahrazují nulové hodnoty velmi malými hodnotami – menšími než nejmenšími možnými naměřenými hodnotami). Canberrská metrika je velice citlivá na malé změny souřadnic, pokud se oba obrazy nacházejí v blízkosti počátku souřadnicové soustavy. Naopak je méně citlivá na změny hodnot příznaků, pokud jsou tyto hodnoty velké.

Příklad

Jsou dány dva vektory $\mathbf{x}_1 = (0{,}001; 0{,}001)^T$ a $\mathbf{x}_2 = (0{,}01; 0{,}01)^T$. Předpokládejme, že souřadnice prvního z vektorů se změní na $\mathbf{x}'_1$ $(0{,}002; 0{,}001)^T$. Jaká je Hammingova a canberrská vzdálenost v obou případech a jaká je relativní změna vzdáleností, vyvolaná uvedenou modifikací?

[13] V literatuře lze najít i verzi bez absolutních hodnot ve jmenovateli (tak, jak byl vzorec původně navržen), samozřejmě s dovětkem, že vztah je vhodný pouze pro kladné hodnoty příznaků.

$$d_H(\mathbf{x}_1,\mathbf{x}_2) = |0{,}001\text{-}0{,}01| + |0{,}001\text{-}0{,}01| = 0{,}009 + 0{,}009 = 0{,}018;$$

$$d_H(\mathbf{x}'_1,\mathbf{x}_2) = |0{,}002\text{-}0{,}01| + |0{,}001\text{-}0{,}01| = 0{,}008 + 0{,}009 = 0{,}017;$$

$$d_{CA}(\mathbf{x}_1,\mathbf{x}_2) = \frac{|0{,}001-0{,}01|}{0{,}001+0{,}01} + \frac{|0{,}001-0{,}01|}{0{,}001+0{,}01} = \frac{0{,}009}{0{,}011} + \frac{0{,}009}{0{,}011} = 0{,}8182 + 0{,}8182 = 1{,}6364;$$

$$d_{CA}(\mathbf{x}'_1,\mathbf{x}_2) = \frac{|0{,}002-0{,}01|}{0{,}002+0{,}01} + \frac{|0{,}001-0{,}01|}{0{,}001+0{,}01} = \frac{0{,}008}{0{,}012} + \frac{0{,}009}{0{,}011} = 0{,}6667 + 0{,}8182 = 1{,}4849.$$

Relativní změny vzdáleností, určující citlivost té které metriky, které jsou způsobeny změnou hodnoty první souřadnice, jsou

$$\Delta d_H = \frac{|d_H(\mathbf{x}_1,\mathbf{x}_2) - d_H(\mathbf{x}'_1,\mathbf{x}_2)|}{d_H(\mathbf{x}_1,\mathbf{x}_2)} = \frac{|0{,}018-0{,}017|}{0{,}018} = \frac{0{,}001}{0{,}018} = 0{,}056;$$

$$\Delta d_{CA} = \frac{|d_{CA}(\mathbf{x}_1,\mathbf{x}_2) - d_{CA}(\mathbf{x}'_1,\mathbf{x}_2)|}{d_{CA}(\mathbf{x}_1,\mathbf{x}_2)} = \frac{|1{,}6364-1{,}4849|}{1{,}6364} = 0{,}093.$$

Ze získaných výsledků je zřejmé, že relativní změna vzdáleností je v případě canberrské metriky pro toto zadání o poznání větší.

Nyní mějme dány vektory $\mathbf{x}_1 = (1000; 1000)^T$ a $\mathbf{x}_2 = (100; 100)^T$ a předpokládejme, že dojde ke změně první souřadnice vektoru $\mathbf{x}_1$ na $\mathbf{x}'_1 = (1002; 1000)^T$. Jaká je Hammingova a canberrská vzdálenost pro tyto vektory a jaká je relativní změna vzdáleností, vyvolaná uvedenou modifikací?

$$d_H(\mathbf{x}_1,\mathbf{x}_2) = |1000\text{-}100| + |1000\text{-}100| = 900 + 900 = 1800;$$

$$d_H(\mathbf{x}'_1,\mathbf{x}_2) = |1002\text{-}100| + |1000\text{-}100| = 902 + 900 = 1802;$$

$$d_{CA}(\mathbf{x}_1,\mathbf{x}_2) = \frac{|1000-100|}{1000+100} + \frac{|1000-100|}{1000+100} = \frac{900}{1100} + \frac{900}{1100} = 0{,}8182 + 0{,}8182 = 1{,}6364;$$

$$d_{CA}(\mathbf{x}'_1,\mathbf{x}_2) = \frac{|1002-100|}{1102+100} + \frac{|1000-100|}{1000+100} = \frac{902}{1102} + \frac{900}{1100} = 0{,}8185 + 0{,}8182 = 1{,}6367.$$

Relativní změny vzdáleností způsobených změnou hodnoty první souřadnice pak v tomto případě jsou

$$\Delta d_H = \frac{|d_H(\mathbf{x}_1,\mathbf{x}_2) - d_H(\mathbf{x}'_1,\mathbf{x}_2)|}{d_H(\mathbf{x}_1,\mathbf{x}_2)} = \frac{|1800-1802|}{1800} = \frac{2}{1800} = 0{,}0011;$$

$$\Delta d_{CA} = \frac{|d_{CA}(\mathbf{x}_1,\mathbf{x}_2) - d_{CA}(\mathbf{x}'_1,\mathbf{x}_2)|}{d_{CA}(\mathbf{x}_1,\mathbf{x}_2)} = \frac{|1{,}6364-1{,}6367|}{1{,}6364} = \frac{0{,}0003}{1{,}6364} = 0{,}00018.$$

Jak je zřejmé, citlivost canberrské metriky je v tomto případě řádově menší. □□□

Kromě uvedených metrik s poměrně obecným použitím existuje řada dalších způsobů výpočtu nepodobnosti dvou vektorů odvozených pro speciální účely. Z nichž uveďme alespoň tzv. ***nelineární metriku*** definovanou vztahem

$$\rho_N(\mathbf{x}_1,\mathbf{x}_2) = \begin{cases} 0, & \text{pro } \rho_E(\mathbf{x}_1,\mathbf{x}_2) < D; \\ H, & \text{pro } \rho_E(\mathbf{x}_1,\mathbf{x}_2) \geq D, \end{cases} \tag{2.74}$$

kde D je prahová hodnota a H je nějaká konstanta. I když existují doporučení, jak volit obě hodnoty na základě statistických vlastností vektorového prostoru, výhodnější se zdá volit obě

konstanty na základě expertní analýzy řešeného problému. I když ve vztahu (2.74) je použita jako základní Euklidova metrika, teoreticky nic nebrání použití jakékoliv jiné metriky vzdálenosti.

2.3.4 Metriky pro určení podobnosti dvou obrazů s kvantitativními příznaky

V praxi se pro obrazy s kvantitativními příznaky (spojitými i diskrétními) používají především následující míry podobnosti.

Skalární součin je pro dva sloupcové vektory $\mathbf{x}_1$ a $\mathbf{x}_2$ definován v euklidovském prostoru vztahem

$$\sigma_{ss}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T . \mathbf{x}_2 = \sum_{i=1}^{n} x_{1i} x_{2i} \,. \tag{2.75}$$

Ve většině případů je skalární součin jako míra podobnosti použit pro vektory $\mathbf{x}_1$ a $\mathbf{x}_2$ o stejné délce, např. a. V těch případech jsou horní, resp. dolní mez skalárního součinu a^2, resp. $-a^2$ a hodnoty skalárního součinu v tom případě závisí výhradně na úhlu, který oba vektory svírají. Hodnoty a^2 nabývá, pokud oba vektory svírají nulový úhel, hodnoty $-a^2$, pokud úhel mezi nimi je 180° a nulové hodnoty, pokud jsou oba vektory na sebe kolmé. Z dosud uvedeného plyne, že skalární součin je invariantní vůči rotaci (jejich absolutní orientace není podstatná, důležitý je pouze úhel mezi nimi), nikoliv však vůči lineární transformaci (závisí na délce vektorů).

Ze skalárního součinu vektorů o délce a je možné odvodit i metriku vzdálenosti podle vztahu

$$\rho_{ss}(\mathbf{x}_1, \mathbf{x}_2) = a^2 - \sigma_{ss}(\mathbf{x}_1, \mathbf{x}_2) \,. \tag{2.76}$$

Na výpočtu skalárního součinu je založena i ***metrika kosinové podobnosti***, která předpokládá, že oba vektory jsou normovány, tedy mají jednotkovou délku. Platí

$$\sigma_{cos}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_1^T . \mathbf{x}_2}{\|\mathbf{x}_1\| . \|\mathbf{x}_2\|} \,, \tag{2.77}$$

kde $\|x_i\|$ je norma (délka) vektoru $\mathbf{x}_i$, určená podle vztahu (2.63). To znamená, že platí vše výše uvedené pro skalární součin s tím, že délka obou vektorů je jednotková, tj. a = 1. Hodnoty $\sigma_{cos}(\mathbf{x}_1, \mathbf{x}_2)$ jsou pak rovny kosinu úhlu mezi oběma vektory.

Pearsonův korelační koeficient, známá statistická míra definovaná výrazem

$$\sigma_{PC}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_{d1}^T . \mathbf{x}_{d2}}{\|\mathbf{x}_{d1}\| . \|\mathbf{x}_{d2}\|} \,, \tag{2.78}$$

kde $\mathbf{x}_{di} = (x_{i1} - \bar{x}_i \,,\, x_{i2} - \bar{x}_i \,,\, \ldots,\, x_{in} - \bar{x}_i)^T$, x_{ij} představují j-tou souřadnici vektoru $\mathbf{x}_i$ a $\bar{x}_i$ je střední hodnota určená ze souřadnic vektoru $\mathbf{x}_i$ ($\bar{x}_i = \sum_{j=1}^{n} x_{ij} / n$). Vektory $\mathbf{x}_{di}$ se obvykle nazývají diferenční vektory. Podobně jako v případě kosinové podobnosti, nabývá Pearsonův korelační koeficient hodnot z intervalu $\langle -1;1 \rangle$, rozdíl vůči kosinové míře podobnosti je ten, že určuje vztah nikoliv vektorů $\mathbf{x}_1$ a $\mathbf{x}_2$, nýbrž jejich diferenčních variant.

I z hodnot Pearsonova korelačního koeficientu lze určit vzdálenost obou vektorů pomocí metriky

$$\rho_{PC}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1 - \sigma_{PC}(\mathbf{x}_1, \mathbf{x}_2)}{2}, \tag{2.79}$$

jejíž hodnoty se, díky dělení dvěma, vyskytují v intervalu ⟨0;1⟩. Tato metrika se používá např. při analýze dat genové exprese.

Tanimotova metrika podobnosti je další, celkem běžně používaná metrika podobnosti, definovaná vztahem

$$\sigma_T(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_1^T \mathbf{x}_2}{\|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2 - \mathbf{x}_1^T \mathbf{x}_2}. \tag{2.80}$$

Přičteme-li a odečteme-li ve jmenovateli výraz $\mathbf{x}_1{}^T\mathbf{x}_2$ a podělíme-li čitatele i jmenovatele zlomku toutéž hodnotou, dostaneme

$$\sigma_T(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{1 + \frac{(\mathbf{x}_1 - \mathbf{x}_2)^T(\mathbf{x}_1 - \mathbf{x}_2)}{\mathbf{x}_1^T \mathbf{x}_2}}. \tag{2.81}$$

Tanimotova podobnost vektorů $\mathbf{x}_1$ a $\mathbf{x}_2$ je tedy nepřímo úměrná kvadrátu Euklidovy vzdálenosti vektorů $\mathbf{x}_1$ a $\mathbf{x}_2$ vztažené k jejich skalárnímu součinu. Pokud skalární součin považujeme za míru korelace obou vektorů, můžeme formulovat výše uvedenou formulaci tak, že $s_T(\mathbf{x}_1, \mathbf{x}_2)$ je nepřímo úměrný kvadrátu Euklidovy vzdálenosti podělenému velikostí jejich korelace, což znamená, že je korelaci, jako míře podobnosti přímo úměrný.

Konečně poslední z prakticky užitečných metrik podobnosti je metrika definovaná vztahem

$$\sigma_C(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{\rho_E(\mathbf{x}_1, \mathbf{x}_2)}{\|\mathbf{x}_1\| + \|\mathbf{x}_2\|}. \tag{2.82}$$

Vzdálenost podle metriky $\sigma_C(\mathbf{x}_1, \mathbf{x}_2)$ je rovna jedné, když $\mathbf{x}_1 = \mathbf{x}_2$ a svého minima, tj. $\sigma_C(\mathbf{x}_1, \mathbf{x}_2) = -1$, když $\mathbf{x}_1 = -\mathbf{x}_2$.

2.3.5 Metriky pro určení vzdálenosti mezi dvěma obrazy s kvalitativními příznaky

Tyto metriky dominantně vycházejí z pojmu kontingenční matice, resp. tabulka.

Předpokládejme, že hodnoty uvažovaných vektorů patří do konečné k-prvkové množiny $\mathcal{F}$ kategoriálních, nebo případně diskrétně kvantitativních hodnot. Dále předpokládejme, že máme vektory $\mathbf{x},\mathbf{y} \in \mathcal{F}^n$, kde n je jejich délka a nechť $\mathbf{A}(\mathbf{x},\mathbf{y}) = |a_{ij}|$, $i,j \in \mathcal{F}$, je matice o rozměru k × k, a její prvky a_{ij} jsou určeny počtem případů, kdy se hodnota i nachází na určité pozici ve vektoru **x** a současně se na téže pozici nachází hodnota j ve vektoru **y**. Matici **A** nazýváme ***kontingenční tabulkou*** (***maticí***). Pokud je kontingenční tabulka rozměru 2 x 2, tj. k = 2, nazýváme ji ***čtyřpolní tabulkou***, slouží ke srovnání dichotomických znaků.

Kromě prostého popisu četností kombinací hodnot dvou znaků a výpočtu vzdáleností, či podobností dvou vektorů hodnot uvedených vlastností, nabízí kontingenční tabulka také možnost testování vztahu mezi oběma hodnotami.

Příklad:

Předpokládejme, že množina $\mathcal{F}$ obsahuje symboly {0, 1, 2}, tj. k = 3 a vektory **x** a **y** jsou $\mathbf{x} = (0, 1, 2, 1, 2, 1)^T$ a $\mathbf{y} = (1, 0, 2, 1, 0, 1)^T$, n = 6. Potom kontingenční matice $\mathbf{A}(\mathbf{x},\mathbf{y})$ je

$$\mathbf{A}(\mathbf{x},\mathbf{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}. \tag{2.83}$$

Lze snadno ukázat, že součet hodnot všech prvků matice **A**(**x**,**y**) je roven délce n obou vektorů, tj. v našem případě

$$\sum_{i=0}^{2}\sum_{j=0}^{2} a_{ij} = 6\,. \tag{2.84}$$

□□□

Hammingova metrika (jak lze usoudit z dále uvedené definice, určitě není náhodná shoda jména s metrikou uvedenou v kap.2.3.3.) je definována počtem pozic, v nichž se oba vektory liší, tj.

$$\rho_{HQ}(\mathbf{x},\mathbf{y}) = \sum_{i=0}^{k-1}\sum_{\substack{j=0 \\ i\neq j}}^{k-1} a_{ij}\,. \tag{2.85}$$

tj. je dána součtem všech prvků matice **A**, které leží mimo hlavní diagonálu.

Pro k = 2, kdy jsou hodnoty obou vektorů binární, se definiční vztah Hammingovy vzdálenosti transformuje na

$$\rho_{HQB}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}(x_i + y_i - 2x_i y_i)\,, \tag{2.86}$$

kde třetí člen v závorce kompenzuje případ, kdy jsou hodnoty x_i i y_i rovny jedné a součet prvních členů v závorce je tím pádem roven dvěma, nicméně nastává shoda hodnot, která k celkové vzdálenosti nemůže přispět. Protože x_i a y_i nabývají hodnot pouze 0 a 1, můžeme také psát

$$\rho_{HQB}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}(x_i^2 + y_i^2 - 2x_i y_i) = \sum_{i=1}^{n}(x_i - y_i)^2 \tag{2.87}$$

a díky speciálnímu případu hodnot x_i a y_i je možná i nejjednodušší forma

$$\rho_{HQB}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}|x_i - y_i|\,. \tag{2.88}$$

V případě bipolárních vektorů, kdy jednotlivé složky vektorů nabývají hodnot +1 a -1, je Hammingova vzdálenost určena vztahem

$$\rho_{HQP}(\mathbf{x},\mathbf{y}) = \frac{\left(n - \sum_{i=1}^{n} x_i y_i\right)}{2}\,. \tag{2.89}$$

Příklad:

Určete Hammingovu vzdálenost vektorů z předchozího příkladu, tj. $\mathbf{x} = (0, 1, 2, 1, 2, 1)^T$ a $\mathbf{y} = (1, 0, 2, 1, 0, 1)^T$.

Vzájemným porovnáním obou vektorů lze určit, že oba vektory se liší v první, druhé a páté souřadnici, to znamená, že se oba vektory liší ve třech pozicích, což definuje hodnotu Hammingovy vzdálenosti obou vektorů, tj . $d_{HQ}(\mathbf{x},\mathbf{y}) = 3$.

Chceme-li určit tuto vzdálenost z kontingenční matice $\mathbf{A}(\mathbf{x},\mathbf{y})$ podle vztahu (2.83), pak je vzdálenost podle vztahu (2.85) určena součtem všech prvků matice $\mathbf{A}(\mathbf{x},\mathbf{y})$ mimo hlavní diagonálu. Tedy $d_{HQ}(\mathbf{x},\mathbf{y}) = a_{12} + a_{21} + a_{31} = 1 + 1 + 1 = 3$. □□□

Příklad:

Určete Hammingovu vzdálenost binárních vektorů $\mathbf{x} = (0, 1, 1, 0, 1)^T$ a $\mathbf{y} = (1, 0, 0, 0, 1)^T$.

Podle definičního principu je vzdálenost obou vektorů dána počtem pozic, ve kterých se oba vektory liší, tj. $d_{HQB}(\mathbf{x},\mathbf{y}) = 3$.

Použijeme-li vztah (2.84), $d_{HQB}(\mathbf{x},\mathbf{y})$ rovna

$$d_{HQB}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}(x_i + y_i - 2x_i y_i) =$$

$$= (0+1-2.0.1) + (1+0-2.1.0) + (1+0-2.1.0) + (0+0-2.0.0) + (1+1-2.1.1) = 3.$$

Podle vztahu (2.85) je

$$d_{HQB}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}(x_i - y_i)^2 =$$

$$= (0-1)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2 + (1-1)^2 = 1+1+1+0+0 = 3.$$

Konečně, využijeme-li vztah (2.88), je

$$d_{HQB}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}|x_i - y_i| = |0-1| + |1-0| + |1-0| + |0-0| + |1-1| = 1+1+1+0+0 = 3.$$

□□□

Příklad:

Určete Hammingovu vzdálenost dvou bipolárních vektorů $\mathbf{x} = (1, 1, 1, -1, 1)^T$ a $\mathbf{y} = (1, -1, 1, -1, -1)^T$.

Podle definičního principu se oba vektory liší ve dvou pozicích, tj. $d_{HQP}(\mathbf{x},\mathbf{y}) = 2$.

Z kontingenční matice, která je pro tento případ rovna

$$\mathbf{A}(\mathbf{x},\mathbf{y}) = \begin{bmatrix} 2 & 2 \\ 0 & 1 \end{bmatrix}$$

je $d_{HQP}(\mathbf{x},\mathbf{y})$ rovna součtu hodnot prvků ležících mimo hlavní diagonálu, tj. $d_{HQP}(\mathbf{x},\mathbf{y}) = 2$.

Použijeme-li vztah (2.89), je také

$$d_{HQP}(\mathbf{x},\mathbf{y}) = \frac{\left(n - \sum_{i=1}^{n} x_i y_i\right)}{2} = \frac{5 - \left((1.1) + (1.(-1)) + (1.1) + ((-1).(-1)) + (1.(-1))\right)}{2} =$$

$$= \frac{5 - (1-1+1+1-1)}{2} = \frac{5-1}{2} = 2.$$

□□□

2.3.6 Metriky pro určení podobnosti mezi dvěma obrazy s kvalitativními příznaky

Metriky podobnosti pro vektory kvalitativních příznaků, resp. vektorů s diskrétními hodnotami příznaků je vhodné rozdělit na případy obecné a případy s dichotomickými příznaky, pro které je definována celá řady tzv. ***asociačních koeficientů***. Asociační koeficienty až na výjimky nabývají hodnot z intervalu $\langle 0, 1\rangle$, hodnoty 1 v případě shody vektorů, 0 pro případ nepodobnosti.

První možností, jak definovat metriku podobnosti pro nedichotomické příznaky, je odvodit ji z Hammingovy metriky

$$\sigma_{HQ}(\mathbf{x},\mathbf{y}) = b_{max} - \rho_{HQ}(\mathbf{x},\mathbf{y}). \qquad (2.90)$$

Zřejmě nejrozšířenější metrikou podobnosti dvou vektorů je ale tzv. ***Tanimotova metrika podobnosti*** (název opět není jen náhodnou podobností s názvem metriky uvedené v kap.2.3.4). Zhusta se používá na příklad v chemii při porovnávání vzorců chemických sloučenin. Její princip vychází z postupu srovnání dvou množin.

Předpokládejme, že máme dvě množiny $\mathcal{X}$ a $\mathcal{Y}$ a $n_{\mathcal{X}}$, $n_{\mathcal{Y}}$ a $n_{\mathcal{X}\cap\mathcal{Y}}$ jsou kardinality (počty prvků) množin $\mathcal{X}$, $\mathcal{Y}$ a $\mathcal{X}\cap\mathcal{Y}$. V tom případě je Tanimotova míra podobnosti dvou množin určena podle vztahu

$$\sigma_T(\mathcal{X},\mathcal{Y}) = \frac{n_{\mathcal{X}\cap\mathcal{Y}}}{n_{\mathcal{X}} + n_{\mathcal{Y}} - n_{\mathcal{X}\cap\mathcal{Y}}}. \qquad (2.91)$$

Jinými slovy, Tanimotova podobnost dvou množin je určena počtem společných prvků obou množin vztaženým k počtu všech rozdílných prvků.

Využijme nyní tohoto principu pro stanovení podobnosti dvou obrazových vektorů s kvalitativními, resp. diskrétními hodnotami příznaků. Pro výpočet Tanimotovy podobnosti pak jsou použity všechny páry složek srovnávaných vektorů, kromě těch, jejichž hodnoty jsou obě nulové.[14]

Nyní definujme pro porovnávané vektory **x** a **y** hodnoty

$$n_x = \sum_{i=1}^{k-1}\sum_{j=0}^{k-1} a_{ij} \text{ a } n_y = \sum_{i=0}^{k-1}\sum_{j=1}^{k-1} a_{ij}, \qquad (2.92)$$

kde k je počet hodnot souřadnic obou vektorů a a_{ij} jsou prvky kontingenční matice $\mathbf{A}(\mathbf{x},\mathbf{y})$, tzn. že n_x, resp. n_y udává počet nenulových položek vektoru **x**, resp. **y**. Pak je Tanimotova metrika podobnosti dvou vektorů definována vztahem

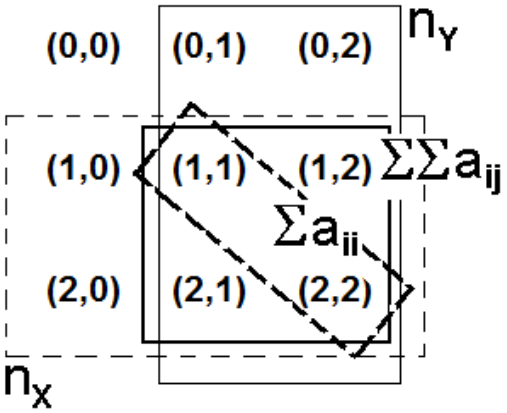
$$\sigma_{TQ}(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{n_x + n_y - \sum_{i=1}^{k-1}\sum_{i=1}^{k-1} a_{ij}} \qquad (2.93)$$

Hodnoty Tanimotovy podobnosti se vyskytují v intervalu od 0 při nepodobnosti do 1 při úplné shodě obou vektorů.

Příklad

Určete hodnoty Tanimotových podobností $s_{TQ}(\mathbf{x},\mathbf{x})$, $s_{TQ}(\mathbf{x},\mathbf{y})$ a $s_{TQ}(\mathbf{x},\mathbf{z})$, jsou-li vektory $\mathbf{x} = (0, 1, 2, 1, 2, 1)^T$ a $\mathbf{y} = (1, 0, 2, 1, 0, 1)^T$ a $\mathbf{z} = (2, 0, 0, 0, 0, 2)^T$.

Ze zadání vyplývá, že množina symbolů $\mathcal{F} = \{0, 1, 2\}$, k = 3, n = 6.



Obr.2.9 *Prvky kontingenční matice použité pro výpočet Tanimotovy podobnosti dvou vektorů*

[14] Tuto volbu se pokusme zdůvodnit případem, kdy analyzujeme vektory ordinálních kvalitativních příznaků, přičemž hodnotu i-tého příznaku daného vektoru považujeme za míru výskytu určitého jevu popisovaného i-tým příznakem. Podle této interpretace jsou páry složek vektorů s oběma hodnotami nulovými méně důležité než ostatní. Tento problém úzce souvisí i s tzv. problémem „***dvojité nuly***“, který se vyskytuje při analýze environmentálních dat (to, že se např. sledovaný druh na dvou sledovaných lokalitách nevyskytuje, není pro posouzení kvality obou lokalit tak důležité, jako společný výskyt některých druhů). Při řešení některých úloh může být stanovení absence nějakého sledovaného rysu i principiálně nemožné – detekce určitých signálových prvků.

Kontingenční tabulky jsou

$$\mathbf{A}(\mathbf{x},\mathbf{x})=\begin{bmatrix}1&0&0\\0&3&0\\0&0&2\end{bmatrix};\ \mathbf{A}(\mathbf{x},\mathbf{y})=\begin{bmatrix}0&1&0\\1&2&0\\1&0&1\end{bmatrix};\ \mathbf{A}(\mathbf{x},\mathbf{z})=\begin{bmatrix}0&0&1\\2&0&1\\2&0&0\end{bmatrix}.$$

V prvním případě při maximální podobnosti jsou nenulové prvky kontingenční tabulky pouze na hlavní diagonále, v případě nejmenší podobnosti jsou naopak na hlavní diagonále jen nulové prvky.

V případě první podobnosti $s_{TQ}(\mathbf{x},\mathbf{x})$ je $n_x = 5$, $n_Y = 5$, součet prvků na hlavní diagonále Σa_{ii} také 5 a konečně součet $\Sigma\Sigma a_{ij}$ opět 5. Hodnota podobnosti pak po dosazení je

$$s_{TQ}(\mathbf{x},\mathbf{x})=\frac{5}{5+5-5}=1.$$

Pro podobnost $s_{TQ}(\mathbf{x},\mathbf{y})$ je $n_x = 5$, $n_y = 4$, součet prvků na hlavní diagonále $\Sigma a_{ii} = 3$ a konečně součet $\Sigma\Sigma a_{ij} = 3$. Hodnota podobnosti pak po dosazení je

$$s_{TQ}(\mathbf{x},\mathbf{y})=\frac{3}{5+4-3}=0{,}5.$$

Konečně, pro podobnost $s_{TQ}(\mathbf{x},\mathbf{z})$, což představuje srovnání dvou nejméně podobných vektorů, je $n_x = 5$, $n_y = 2$, součet prvků na hlavní diagonále $\Sigma a_{ii} = 0$ a konečně součet $\Sigma\Sigma a_{ij} = 1$. Hodnota podobnosti pak po dosazení je

$$\sigma_{TQ}(\mathbf{x},\mathbf{y})=\frac{0}{5+2-1}=0.$$

□□□

Další míry podobnosti vektorů $\mathbf{x},\mathbf{y} \in \mathcal{T}^n$ jsou definovány pomocí různých prvků kontingenční matice $\mathbf{A}(\mathbf{x},\mathbf{y})$. Některé z nich používají pouze počet shodných pozic v obou vektorech (ovšem s nenulovými hodnotami), jiné míry používají i shodu s nulovými hodnotami. Příkladem metriky podobnosti z první uvedené kategorie může být např. metrika definovaná vztahem

$$\sigma_1(\mathbf{x},\mathbf{y})=\frac{\sum_{i=1}^{k-1} a_{ii}}{n}, \tag{2.94}$$

nebo i metrika

$$\sigma_2(\mathbf{x},\mathbf{y})=\frac{\sum_{i=1}^{k-1} a_{ii}}{n-a_{00}}. \tag{2.95}$$

Příkladem metriky druhé uvedené skupiny je např.

$$\sigma_3(\mathbf{x},\mathbf{y})=\frac{\sum_{i=0}^{k-1} a_{ii}}{n}. \tag{2.96}$$

Při řešení dichotomických problémů, tj. když $k = 2$, nabývá kontingenční tabulka tvar podle obr.2.10, který vyjadřuje čtyři možné situace:

A. hodnota k-té souřadnice obou vektorů signalizuje, že u obou obrazů sledovaný jev nastal (oba odpovídající si příznaky

		$\mathbf{x}_j$	
		false/0	true/1
$\mathbf{x}_i$	false/0	D	C
	true/1	B	A

Obr.2.10 *Kontingenční tabulka pro dichotomické hodnoty*

mají hodnotu true) – ***pozitivní shoda***;

B. ve vektoru $\mathbf{x}_i$ jev nastal (x_{ik} = true), zatímco ve vektoru $\mathbf{x}_j$ nikoliv (x_{jk} = false);

C. u obrazu $\mathbf{x}_i$ jev nenastal (k-tá souřadnice má hodnotu x_{ik} = false), zatímco u obrazu $\mathbf{x}_j$ ano (x_{jk} = true);

D. sledovaný jev nenastal (oba odpovídající si příznaky mají hodnotu false) – ***negativní shoda***.

Při výpočtu podobnosti dvou vektorů sledujeme kolikrát pro všechny souřadnice obou vektorů $\mathbf{x}_j$ a $\mathbf{x}_j$ nastaly případy shody či neshody – A+D určuje celkový počet shod, B+C celkový počet neshod a A+B+C+D = n, tj. celkový počet souřadnic obou vektorů (obrazů).

Pokud budeme pokračovat v popisu Tanimotovy metriky podobnosti, pak pro dichotomické proměnné se výpočet, s ohledem na symboliku podle obr.2.10, transformuje do vztahu (často je též označován jako ***Jaccardův-Tanimotův asociační koeficient***)

$$\sigma_{JT}(\mathbf{x},\mathbf{y}) = \frac{A}{A+B+C}, \tag{2.97}$$

což je díky zjednodušení i dichotomická varianta metriky podle vztahu (2.93). Tento vztah se dominantně používá v ekologických studiích.

Dichotomická varianta vztahu (2.94) je tzv. ***Russelův - Raoův asociační koeficient***.

$$\sigma_{RR}(\mathbf{x},\mathbf{y}) = \frac{A}{A+B+C+D}, \tag{2.98}$$

Vztah (2.96) modifikovaný pro dichotomické aplikace

$$\sigma_{SM}(\mathbf{x},\mathbf{y}) = \frac{A+D}{A+B+C+D} \tag{2.99}$$

se označuje jako ***Sokalův - Michenerův asociační koeficient***.

Kromě uvedených koeficientů se v odborné literatuře vyskytují i ***Dicův*** (***Czekanowského***) ***koeficient***

$$\sigma_{DC}(\mathbf{x},\mathbf{y}) = \frac{2A}{2A+B+C} = \frac{2A}{(A+B)+(A+C)} \tag{2.100}$$

a ***Rogersův - Tanimotův koeficient***

$$\sigma_{RT}(\mathbf{x},\mathbf{y}) = \frac{A+D}{A+D+2.(B+C)} = \frac{A+D}{(B+C)+(A+B+C+D)}, \tag{2.101}$$

které zvyšují význam shod v datech – v případě Dicova koeficientu zvýšením váhy počtu pozitivních shod v čitateli i jmenovateli, v druhém případě zvýšením váhy počtu neshod ve jmenovateli.

Hamanův koeficient

$$\sigma_{HA}(\mathbf{x},\mathbf{y}) = \frac{A+D-(B+C)}{A+B+C+D} \tag{2.102}$$

nabývá na rozdíl od všech dříve uvedených koeficientů hodnot z intervalu ⟨-1, 1⟩. Hodnoty -1 nabývá, pokud se příznaky pouze neshodují, je roven nule, když je počet shod a neshod v rovnováze a +1 v případě úplné shody všech příznaků.

V případě Jaccardova a Dicova koeficientu je třeba vyřešit (pokud jsou používány v situacích, kdy může nastat úplná negativní shoda) jejich hodnotu, když A = B = C =0. Pak zpravidla předpokládáme, že $\sigma_{JT}(\mathbf{x},\mathbf{y}) = \sigma_{DC}(\mathbf{x},\mathbf{y}) = 1$

Z asociačních koeficientů, které vyjadřují míru podobnosti, lze jednoduše odvodit i míry nepodobnosti (vzdálenosti) pomocí formule

$$\rho_X(\mathbf{x},\mathbf{y}) = 1 - \sigma_X(\mathbf{x},\mathbf{y}). \tag{2.103}$$

Na základě četností A až D lze pro případ binárních příznaků vytvářet i zajímavé vztahy pro již dříve uvedené míry:

Hammingova metrika

$$\rho_H(\mathbf{x},\mathbf{y}) = B + C; \tag{2.104}$$

Euklidova metrika

$$\rho_H(\mathbf{x},\mathbf{y}) = \sqrt{B + C}; \tag{2.105}$$

Pearsonův korelační koeficient

$$\sigma_{PC}(\mathbf{x},\mathbf{y}) = \frac{A.D - B.C}{\sqrt{(A+B).(C+D).(A+C).(B+D)}} \tag{2.106}$$

i jiné.

2.3.7 Deterministické metriky pro určení vzdálenosti mezi dvěma množinami obrazů

Při klasifikaci podle minimální vzdálenosti, stejně jako i v jiných klasifikačních disciplínách (např. při shlukování), je třeba pro posouzení vzdálenosti či podobnosti dvou obrazů, umět určit i vzdálenost mezi obrazem a množinou obrazů, představujících určitou klasifikační třídu, případně vzdálenost mezi dvěma různými množinami obrazů. Oba problémy lze vyřešit zavedením metrik pro dvě množiny – za předpokladu, že samotný obraz považujeme za jednoprvkovou množinu. Tyto metriky, samozřejmě splňující podmínky uvedené v kap.2.3.2, ke každé dvojici množin (C_i, C_j) obrazů z rozkladu $S = (C_1, C_2, \ldots, C_m)$ přiřazují hodnotu znamenající vzdálenost či podobnost obou množin C_i a C_j.

Způsoby výpočtu vzdáleností tohoto typu záleží na způsobu reprezentace množiny obrazů - zda je vyjádřena úplným výčtem obrazů, nebo zda je reprezentována nějakým významným obrazem či obrazy.

Následující metriky předpokládají reprezentaci množiny úplným výčtem jejích položek.

Metoda nejbližšího souseda

Je-li ρ libovolná metrika vzdálenosti dvou obrazů a C_i a C_j jsou libovolné množiny rozkladu množin (klasifikačních tříd) obrazů $\{x_i\}$, i = 1, …, K, potom metoda nejbližšího souseda definuje vzdálenost mezi množinami C_i a C_j

$$\rho_{NN}(C_i, C_j) = \min_{\substack{\mathbf{x}_p \in C_i \\ \mathbf{x}_q \in C_j}} \rho(\mathbf{x}_p, \mathbf{x}_q). \tag{2.107}$$

Při použití tohoto způsobu výpočtu vzdálenosti se mohou vyskytovat v jedné množině často i poměrně vzdálené obrazy, pokud se mezi nimi vyskytují obrazy další. To znamená, že metoda nejbližšího souseda může vytvářet klasifikační třídy protáhlého tvaru.

Metoda k nejbližších sousedů

Tento postup je zobecněním předcházející metody. Je definován vztahem

$$\rho_{NNk}(C_i, C_j) = \min_{\substack{\mathbf{x}_p \in C_i \\ \mathbf{x}_q \in C_j}} \sum^{k} \rho(\mathbf{x}_p, \mathbf{x}_q), \tag{2.108}$$

tj. vzdálenost dvou množin obrazů je v tomto případě definována součtem k nejkratších vzdáleností mezi obrazy obou množin. Metoda částečně potlačuje výše uvedenou tendenci ke generování protáhlých řetězcových struktur.

Metoda nejvzdálenějšího souseda

Je založena na přesně opačném principu než obě předcházející metody. Platí, že

$$\rho_{FN}(C_i, C_j) = \max_{\substack{\mathbf{x}_p \in C_i \\ \mathbf{x}_q \in C_j}} \rho(\mathbf{x}_p, \mathbf{x}_q). \tag{2.109}$$

Generování protáhlých struktur tato metoda potlačuje, naopak vede k tvorbě nevelkých kompaktních množin.

Tak jako v předcházejícím případě je možné i zobecnění použitím k nejvzdálenějších obrazů z obou shluků, pak platí

$$\rho_{FNk}(C_i, C_j) = \max_{\substack{\mathbf{x}_p \in C_i \\ \mathbf{x}_q \in C_j}} \sum^{k} \rho(\mathbf{x}_p, \mathbf{x}_q). \tag{2.110}$$

Metoda průměrné vazby

Metoda definuje vzdálenost dvou množin C_i a C_j pomocí průměrné vzdálenosti mezi všemi obrazy obou množin. Obsahuje-li množina C_i P obrazů a množina C_j Q obrazů, pak jejich vzdálenost podle metody průměrné vazby je určena vztahem

$$\rho_{GA}(C_i, C_j) = \frac{1}{P.Q} \sum_{p=1}^{P} \sum_{q=1}^{Q} \rho(\mathbf{x}_p, \mathbf{x}_q). \tag{2.111}$$

Tento způsob výpočtu často vede k podobným výsledkům jako metoda nejvzdálenějšího souseda.

Centroidní metoda

Je představitelkou metod, které určují vzdálenost mezi množinami pomocí vzdálenosti jejich reprezentativních obrazů. Takovým obrazem může být tzv. ***centroid***, což je obraz, který je určený průměrem, mediánem, resp. jinou významnou charakteristikou, vyjadřující nějakou souhrnnou vlastnost všech obrazů dané množiny. Zatímco centroid je nový, uměle spočítaný obraz reprezentující množinu, tzv. ***medoid*** je jeden z obrazů dané množiny, který má optimální vlastnost z hlediska nějaké souhrnné charakteristiky, např. jehož vzdálenost od všech ostatních obrazů množiny je minimální.

V případě centroidu v euklidovském n-rozměrném prostoru je vzdálenost dvou shluků určena euklidovskou vzdáleností mezi centroidy, reprezentujícími obě množiny.

Je-li např. centroid definován pomocí středních hodnot souřadnic všech obrazových vektorů patřících do dané množiny, tj. představuje-li

$$\bar{\mathbf{x}}_i = (\bar{x}_{i1}, \bar{x}_{i2}, ..., \bar{x}_{in}) \tag{2.112}$$

reprezentativní vektor (centroid) množiny C_i, kde

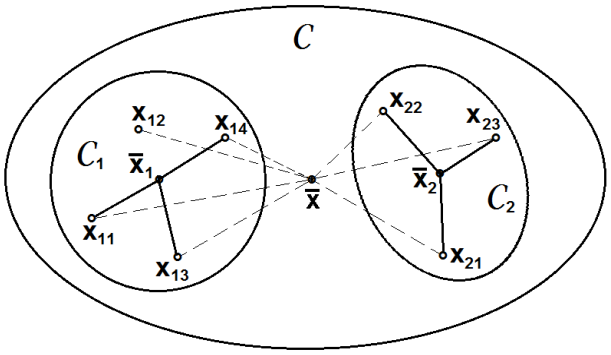
$$\bar{x}_{is} = \frac{1}{K_i} \sum_{k=1}^{K_i} x_{isk}, \; s = 1, ..., n, \tag{2.113}$$

pak

$$\rho_{CE}(C_i, C_j) = \rho_E(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j) = \sqrt{\left(\sum_{s=1}^{n} (\bar{\mathbf{x}}_{is} - \bar{\mathbf{x}}_{js})^2 \right)}. \tag{2.114}$$

Wardova metoda

Je kombinovaný postup, který potřebuje jak znalost všech obrazů obou uvažovaných množin, tak i znalost reprezentativních obrazů.



Obr.2.11 *Princip výpočtu vzdálenosti podle Wardovy metody*

Vzdálenost mezi množinami je podle této metody definována přírůstkem součtu čtverců odchylek mezi centroidem a obrazy množiny vytvořené z obou vstupních množin C_i a C_j oproti součtu čtverců odchylek mezi obrazy a centroidy v obou množinách C_i a C_j.

Jsou-li $\bar{\mathbf{x}}_i$ a $\bar{\mathbf{x}}_j$ n-rozměrné centroidy množin C_i a C_j a $\bar{\mathbf{x}}$ centroid sjednocené množiny, pak je Wardova metrika definována výrazem (viz obr.2.11)

$$\rho_W(C_i, C_j) = \sum_{\mathbf{x}_i \in C_i \cup C_j} \sum_{s=1}^{n} (x_{is} - \bar{x}_s)^2 - \left(\sum_{\mathbf{x}_i \in C_i} \sum_{s=1}^{n} (x_{is} - \bar{x}_{1s})^2 + \sum_{\mathbf{x}_i \in C_j} \sum_{s=1}^{n} (x_{is} - \bar{x}_{2s})^2 \right). \qquad (2.115)$$

Wardova metoda má tendenci vytvářet kompaktní, poměrně malé množiny, zhruba stejné velikosti.

2.3.8 Metriky pro určení vzdálenosti mezi dvěma množinami obrazů používající jejich pravděpodobnostní charakteristiky

Klasifikační třídy (množiny obrazů se společnými charakteristikami) nemusí být definovány jen výčtem obrazů, nýbrž vymezením obecnějších vlastností, jak ostatně tento text zmiňuje velice často – definicí hranic oddělujících část obrazového prostoru, která náleží dané klasifikační třídě, diskriminační funkcí, pravděpodobnostními charakteristikami výskytu obrazů v dané třídě, atd. Jestliže jsme v předchozí kapitole využívali znalosti vlastností dané množiny, které byly určeny polohou jednotlivých konkrétních obrazů, patřících do té které klasifikační třídy, dále popíšeme způsoby stanovení vzdálenosti mezi množinami, které používají pravděpodobnostní charakteristiky rozložení obrazů v dané množině.

Pokud si na metriky klademe určité požadavky, i metriky pro stanovení vzdálenosti dvou množin, pro něž využíváme rozložení pravděpodobnosti výskytu obrazů, by měly vyhovovat standardním požadavkům. Logicky tyto metriky splňují následující vlastnosti (protože jejich výpočet je založen na poněkud jiném přístupu a protože i dále uvedené vlastnosti nesplňují vše, co od metrik očekáváme, bývá zvykem je značit jiným písmenem, zpravidla J):

1. J = 0, pokud jsou hustoty pravděpodobnosti obou množin identické, tj. když
$$p(\mathbf{x}|\omega_1) = p(\mathbf{x}|\omega_2);$$
2. $J \geq 0$;
3. J nabývá maxima, pokud jsou obě množiny disjunktní, tj. když
$$\int_{-\infty}^{\infty} p(\mathbf{x}|\omega_1).p(\mathbf{x}|\omega_2)d\mathbf{x} = 0\,.$$

(Jak vidíme, není mezi vlastnostmi pravděpodobnostních metrik uvedena trojúhelníková nerovnost, jejíž splnění by se zajišťovalo vskutku jen velmi obtížně.)

Základní myšlenkou, na které jsou pravděpodobnostní metriky založeny, je podobně, jak bylo popsáno pro bayesovský klasifikátor v kap.2.2, využití pravděpodobnosti způsobené chyby. Čím více se hustoty pravděpodobnosti výskytu obrazů **x** v jednotlivých množinách překrývají, tím je větší pravděpodobnost chyby.

Pokusme se nyní tuto myšlenku zformalizovat. Pravděpodobnost P_e chybného zařazení je s pomocí vztahů (2.13), (2.17) a(2.28), resp.(2.29) rovna

$$\begin{aligned} P_e = J(\mathbf{a}^*) = \min_{\forall \mathbf{a}} J(\mathbf{a}) = \int_{\mathcal{X}} \min_{\forall r} L_{\mathcal{X}}(\mathbf{a}) d\mathbf{x} = \int_{\mathcal{X}} \left[p(\mathbf{x}) - p(\mathbf{x}|\omega_r).P(\omega_r) \right] d\mathbf{x} = \\ = \int_{\mathcal{X}} p(\mathbf{x}) d\mathbf{x} - \int_{\mathcal{X}} \max_{\forall r} p(\mathbf{x}|\omega_r).P(\omega_r) d\mathbf{x} = 1 - \int_{\mathcal{X}} \max_{\forall r} p(\mathbf{x}|\omega_r).P(\omega_r) d\mathbf{x} . \end{aligned} \tag{2.116}$$

Pro dichotomický případ (R = 2) je celková pravděpodobnost chybného rozhodnutí určena vztahem

$$P_e = 1 - \int_{\mathcal{X}} \left| p(\mathbf{x}|\omega_1).P(\omega_1) - p(\mathbf{x}|\omega_2).P(\omega_2) \right| d\mathbf{x}, \tag{2.117}$$

což lze podle Bayesova vzorce upravit i do tvaru

$$P_e = 1 - \int_{\mathcal{X}} \left| P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x}) \right| .p(\mathbf{x}).d\mathbf{x}. \tag{2.118}$$

Integrál ve vztahu (2.118) nazýváme ***Kolmogorovova variační vzdálenost*** a jeho hodnota přímo souvisí s pravděpodobností chybného rozhodnutí. Ostatní dále uvedené pravděpodobnostní míry vzdálenosti, odvozené z obecné formule

$$J(\mathbf{x}) = \int f\left[p(\mathbf{x}|\omega_i), P(\omega_i), i = 1,2 \right] d\mathbf{x} \tag{2.119}$$

už tuto přímou souvislost nemají, ale mohou být použity k určení mezí odhadu chyby.

Jednou z hlavních nevýhod pravděpodobnostních metrik je potřeba odhadnout průběh hustot pravděpodobnosti a poté je numericky integrovat, což může způsobit problémy, které znemožní použití tohoto přístupu v mnoha různých aplikacích. Situace se výrazně zjednoduší, pokud lze předpokládat určitý charakter rozložení pravděpodobnosti. V tom případě je možné provést mnohé výpočty analyticky.

Mezi nejpoužívanější míry pravděpodobnostní vzdálenosti dvou množin patří

Chernoffova metrika

$$J_C(\omega_1, \omega_2) = -\ln \int p^s(\mathbf{x}|\omega_1).p^{1-s}(\mathbf{x}|\omega_2).d\mathbf{x}, \; s \in \langle 0; 1 \rangle; \tag{2.120}$$

Bhattacharyyova metrika

$$J_B(\omega_1, \omega_2) = -\ln \int [p(\mathbf{x}|\omega_1).p(\mathbf{x}|\omega_2)]^{0,5}.d\mathbf{x}. \tag{2.121}$$

(Jak lze snadno rozpoznat, Bhattacharyyova metrika je speciální případ Chernoffovy metriky pro s = 0,5).

Divergence

$$J_D(\omega_1, \omega_2) = \int [p(\mathbf{x}|\omega_1) - p(\mathbf{x}|\omega_2)].\ln\left(\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} \right) .d\mathbf{x} ; \tag{2.122}$$

nebo ***Patrickova -Fisherova metrika***

$$J_{PF}(\omega_1, \omega_2) = \left\{ \int [p(\mathbf{x}|\omega_1) - p(\mathbf{x}|\omega_2)]^2 .d\mathbf{x} \right\}^{0,5}. \tag{2.123}$$

Alternativou mohou být jejich zprůměrněné verze, které zahrnují i apriorní pravděpodobnost jednotlivých množin:

zprůměrněná Chernoffova metrika

$$J_{AC}(\omega_1,\omega_2) = -\ln \int [p(\mathbf{x}|\omega_1).P(\omega_1)]^s.[p(\mathbf{x}|\omega_2).P(\omega_2)]^{1-s1}.d\mathbf{x},.\ s \in \langle 0;1\rangle; \tag{2.124}$$

zprůměrněná Bhattacharyyova metrika

$$J_{AB}(\omega_1,\omega_2) = -\ln \int [p(\mathbf{x}|\omega_1).P(\omega_1).p(\mathbf{x}|\omega_2).P(\omega_2)]^{0,5}.d\mathbf{x}; \tag{2.125}$$

zprůměrněná divergence

$$J_{AD}(\omega_1,\omega_2) = \int [p(\mathbf{x}|\omega_1).P(\omega_1) - p(\mathbf{x}|\omega_2).P(\omega_2)].\ln\left(\frac{p(\mathbf{x}|\omega_1).P(\omega_1)}{p(\mathbf{x}|\omega_2).P(\omega_2)}\right).d\mathbf{x}; \tag{2.126}$$

nebo ***zprůměrněná Patrickova -Fisherova metrika***

$$J_{PF}(\omega_1,\omega_2) = \left\{\int [p(\mathbf{x}|\omega_1).P(\omega_1) - p(\mathbf{x}|\omega_2).P(\omega_2)]^2.d\mathbf{x}\right\}^{0,5}. \tag{2.127}$$

Pro R množin byl odvozen vztah pro ***Bayesovu metriku***

$$J_{BA}(\omega_1,...,\omega_R) = \int\left(\sum_{r=1}^{R} P^2(\omega_r|\mathbf{x})\right).p(\mathbf{x}).d\mathbf{x}. \tag{2.128}$$

Hodnoty vzdálenosti určené podle tohoto předpisu se pohybují v intervalu $(0;\ 1\rangle$. Jednotkové hodnoty nabývá v případě, že aposteriorní pravděpodobnost $P(\omega_r|\mathbf{x})$ jedné množiny je rovna jedné, zatímco pro zbývající množiny jsou jejich aposteriorní pravděpodobnosti nulové. Nejmenší hodnoty, které Bayesova vzdálenost nabývá je 1/R, to v případě, že jsou všechny aposteriorní pravděpodobnosti stejné. Když $R\to\infty$, pak hodnota vzdálenosti se limitně blíží k nule.

Uvedené vztahy se liší zejména pracností výpočtu a vazbou k hodnotám pravděpodobnosti chyby. Tato vazba je vyjádřena hodnotami dolního $D(\mathbf{x})$ a horního $H(\mathbf{x})$ odhadu pravděpodobnosti chyby, z nichž především horní odhad má praktický význam.

Pro některé z uvedených pravděpodobnostních měr jsou hodnoty horního odhadu

$$\begin{aligned} H_C(\mathbf{x}) &= \min_{s\in\langle 0;1\rangle} J_C(s); \\ H_B(\mathbf{x}) &= J_B; \\ H_{BA}(\mathbf{x}) &= 1 - J_{BA}. \end{aligned} \tag{2.129}$$

V případě, že známe dichotomické pravděpodobnostní míry a je třeba řešit problém klasifikace do více tříd, lze definovat metriku podle vztahu

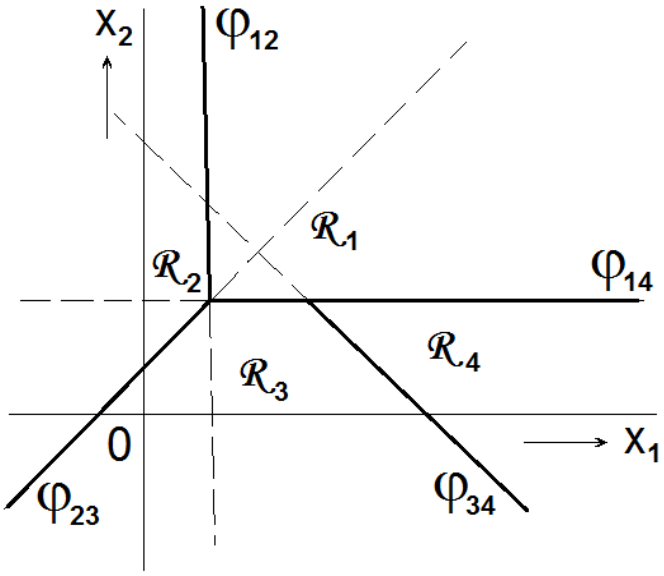
$$J(\omega_1,...,\omega_R) = \sum_{r=1}^{R-1}\sum_{q=r+1}^{R} P(\omega_r).P(\omega_q).J(\omega_r,\omega_s). \tag{2.130}$$

V tom případě ale neplatí těsný vztah k hodnotě pravděpodobnosti chyby, jako ve výše uvedených vztazích.

2.4 Klasifikace pomocí hranic v obrazovém prostoru

2.4.1 Základní principy

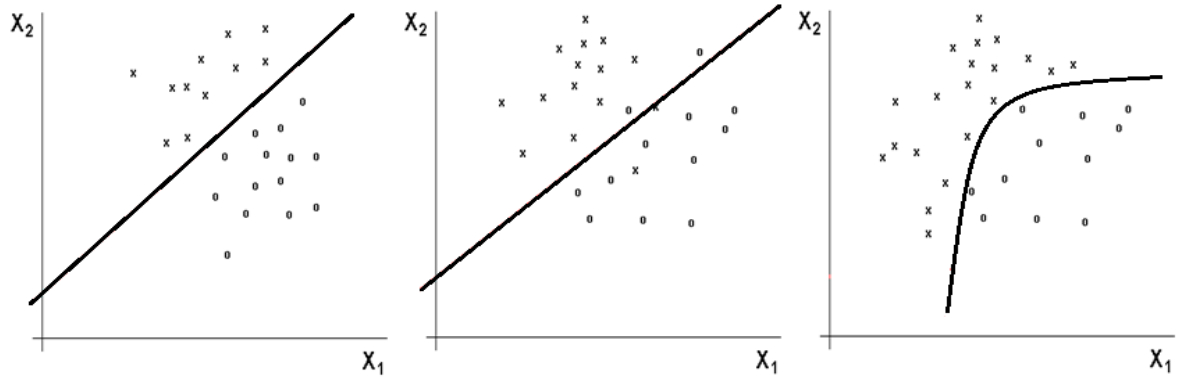
Rozdělení příznakového prostoru do vzájemně disjunktních dílčích prostorů, odpovídajících jednotlivým klasifikačním třídám, pravděpodobně odpovídá nejjednodušší představě o reprezentaci klasifikačních tříd (obr.2.12). Hranice jsou tvořeny obecně nadplochami o rozměru o jednotku menší než je rozměr příznakového prostoru – v dvoupříznakovém prostoru je to obecně křivka, ve speciálním lineárním případě přímka, v trojrozměrném prostoru plocha, resp. rovina, atd. Způsoby určení oddělujících hranic závisí jednak na vlastnostech klasifikačních tříd (zda se jejich obrazy vyskytují v navzájem překrývajících se oblastech, či nikoliv – v tom případě hovoříme o ***separabilních*** či ***neseparabilních množinách***; zda je možné množiny obrazů oddělit lineární hraniční plochou, či zda je vhodnější použít plochu nelineární, …), jednak na kritériích, která použijeme pro optimalizaci polohy hranic.



Obr.2.12 *Dvoupříznakový prostor s přímkami oddělujícími množiny obrazů* $\mathcal{R}_i$

V dalším textu se budeme zabývat výhradně metodami pro stanovení lineárních hranic mezi klasifikačními třídami. V případě, že jsou klasifikační třídy lineárně neseparabilní, používají se dva principiálně odlišné přístupy:

a) zachováme původní obrazový prostor a zvolíme nelineární hraniční funkci
   aa) definovanou obecně;
   ab) složenou po částech z lineárních úseků;
b) zobrazíme původní n rozměrný obrazový prostor $\mathcal{X}^n$ nějakou nelineární transformací



a) b) c)

Obr.2.13 *Případy separability klasifikačních tříd - a) lineárně separabilní úloha; b)lineárně neseparabilní úloha, ovšem s lineárně separovanými třídami; c) nelineárně separabilní klasifikační úloha.*

Φ: $\mathcal{X}^n \rightarrow \mathcal{X}^m$ do nového m rozměrného prostoru $\mathcal{X}^m$, obecně je $n \neq m$, tak, aby v novém prostoru byly klasifikační třídy lineárně separabilní a v novém prostoru použijeme lineární klasifikátor.

ad aa)

Abychom byli schopni analyticky specifikovat nelineární hranici mezi dvěma klasifikačními třídami, je potřeba pro každou hranici určit obecný tvar funkce, které je možné pro daný účel použít (např. $h_1(\mathbf{x}) = a.x_1.x_2^3$, nebo $h_2(\mathbf{x}) = [(a.x_1)^3 + (b.x_2)^2]^c$) a stanovit jejich parametry (v našem případě a, b a c). První problém se zpravidla řeší heuristicky pomocí apriorní informace o klasifikační úloze, stanovení parametrů hraniční funkce vede na obtížně řešitelné nelineární optimalizační úlohy, proto se tomuto způsobu popisu klasifikačních tříd snažíme co nejvíce vyhýbat.

ad ab)

Náhrada nelineární hraniční nadplochy po částech nadrovinami je další možností, jak zjednodušit stanovení parametrů hraniční funkce tím, že optimalizační úlohu parciálně linearizujeme, i když za cenu násobné realizace.

ad b)

Transformaci Φ do nového prostoru provedeme pomocí m funkcí $\Phi_1(\mathbf{x})$, $\Phi_2(\mathbf{x})$, …, $\Phi_m(\mathbf{x})$, kde $\Phi_1(\mathbf{x}) \equiv 1$ a transformované hraniční nadroviny pak mohou mít tvar

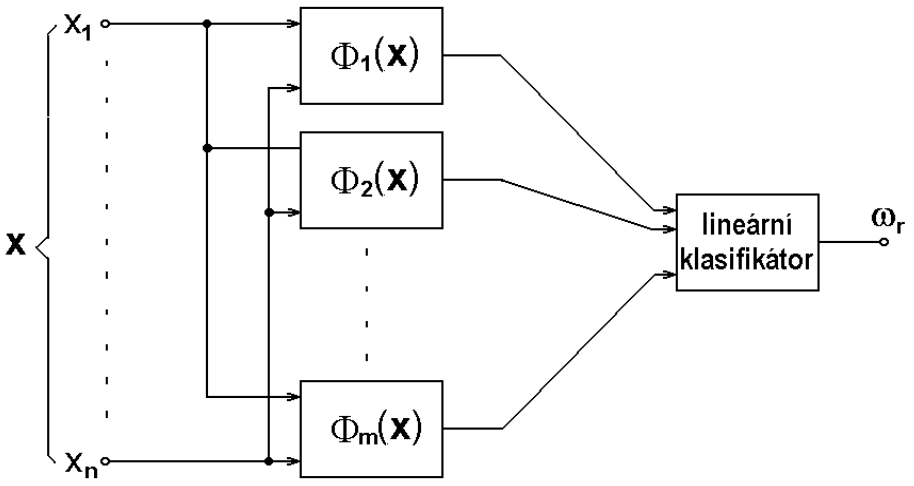
$$h_r(\mathbf{x}) = a_{r1}\Phi_1(\mathbf{x}) + a_{r2}\Phi_2(\mathbf{x}) + \ldots + a_{rm}\Phi_m(\mathbf{x}), \tag{2.131}$$

nebo při vektorovém zápisu

$$h_r(\mathbf{x}) = \mathbf{a}_r^T.\mathbf{\Phi}(\mathbf{x}), \tag{2.132}$$

kde $\mathbf{\Phi}(\mathbf{x}) = (\Phi_1(\mathbf{x}), \Phi_2(\mathbf{x}), \ldots, \Phi_m(\mathbf{x}))^T$ a $\mathbf{a}_r^T = (a_{r1}, a_{r2}, \ldots, a_{rm})$. Protože $\Phi_1(\mathbf{x}) \equiv 1$, představuje a_{r1} prahový koeficient, posouvající počátek souřadnicového systému v obrazovém prostoru $\mathcal{X}^m$ oproti počátku $\mathcal{X}^n$.

Tato metoda klasifikace se obvykle nazývá metodou Φ funkcí, resp. Φ převodník. V zásadě se neliší od klasifikátoru s obecnými nelineárními hraničními funkcemi, protože nelineární funkci lze rozvést v řadu a tak získat vztah (2.131). Blokové schéma metody Φ funkcí je na obr.2.14, z něhož plyne, že tuto metodu lze také interpretovat jako lineární klasifikátor s předřazeným blokem funkčních převodníků.



Obr.2.14 *Blokové schéma Φ převodníku*

Z lineárních metod určení hraničních funkcí mezi klasifikačními třídami se budeme dále zabývat metodou nejmenších čtverců, Fisherovou diskriminační metodou, metodou jednovrstvého perceptronu a algoritmem podpůrných vektorů.

Než se začneme věnovat jednotlivým klasifikačním algoritmům, uveďme několik základních skutečností. K tomu použijme nejjednodušší klasifikační, tzv. ***dichotomickou***

úlohu[15], která se zabývá klasifikací do dvou vzájemně se vylučujících kategorií.

Předpokládejme, že je dána funkce

$$b(\mathbf{x}) = \mathbf{w}^T.\mathbf{x} + w_0, \tag{2.133}$$

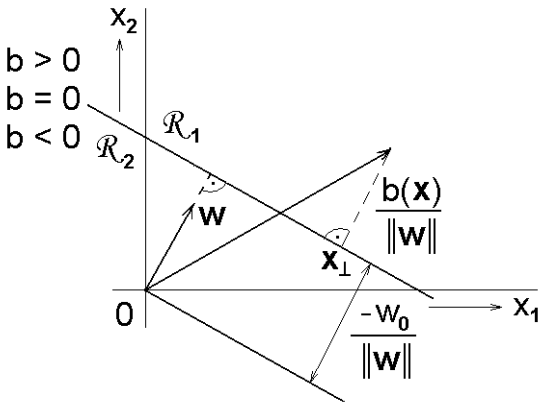
kde $\mathbf{w}^T = (w_1, w_2, \ldots, w_n)$ je tzv. váhový vektor a $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ je příznakový vektor, popisující klasifikovaný objekt. Konečně absolutní člen w_0 můžeme chápat jako prahovou hodnotu. Abychom formálně splnili podmínku systémové linearity, tj. aby funkce $b(\mathbf{x})$ procházela počátkem souřadnicové soustavy, zavádíme novou souřadnici $x_0 = 1$ a pak lze psát

$$b(\tilde{\mathbf{x}}) = \tilde{\mathbf{w}}^T.\tilde{\mathbf{x}}, \tag{2.134}$$

kde $\tilde{\mathbf{w}}^T = (w_0, \mathbf{w})$ a $\tilde{\mathbf{x}}^T = (x_0, \mathbf{x}^T) = (1, \mathbf{x}^T)$. Předpokládejme dále, že vektor $\mathbf{x}$ zařadíme do třídy $\mathcal{R}_1$, pokud $b(\mathbf{x}) \geq 0$ a do třídy $\mathcal{R}_2$, když platí $b(\mathbf{x}) < 0$. Odpovídající hraniční funkce (rovina) je tedy jeho kolmý průmět na hraniční rovinu tak, že

$$b(\mathbf{x}) = \mathbf{w}^T.\mathbf{x} + w_0 = 0\ , \tag{2.135}$$

tj. průsečík funkce $b(\mathbf{x})$ s obrazovým prostorem. Nyní uvažme dva body $\mathbf{x}_A$ a $\mathbf{x}_B$, které leží na hraniční rovině. Protože pro oba platí, že $b(\mathbf{x}_A) = b(\mathbf{x}_B) = 0$, platí také $\mathbf{w}^T.(\mathbf{x}_A - \mathbf{x}_B) = 0$ a proto můžeme říci, že vektor $\mathbf{w}$ je kolmý k libovolnému vektoru ležícímu na hraniční rovině a tím také určuje směr hraniční plochy.



Obr.2.15 *Základní vztahy pro lineární hraniční plochu ve dvourozměrném prostoru*

Podobně, leží-li bod $\mathbf{x}_A$ na hraniční ploše a tedy je $b(\mathbf{x}_A) = 0$, pak normálová vzdálenost počátku souřadnicové soustavy od hraniční plochy je

$$\frac{\tilde{\mathbf{w}}^T.\tilde{\mathbf{x}}}{\|\mathbf{w}\|} = -\frac{w_0}{\|\mathbf{w}\|}. \tag{2.136}$$

Z toho plyne, že prahový parametr w_0 určuje i polohu hraniční roviny. Dále hodnota $b(\mathbf{x})$ určuje i kolmou vzdálenost d bodu $\mathbf{x}$ od hraniční plochy. Uvažme libovolný bod $\mathbf{x}$ a nechť $\mathbf{x}_\perp$ je jeho kolmý průmět na hraniční rovinu tak, že

$$\mathbf{x} = \mathbf{x}_\perp + d\frac{\mathbf{w}}{\|\mathbf{w}\|}. \tag{2.137}$$

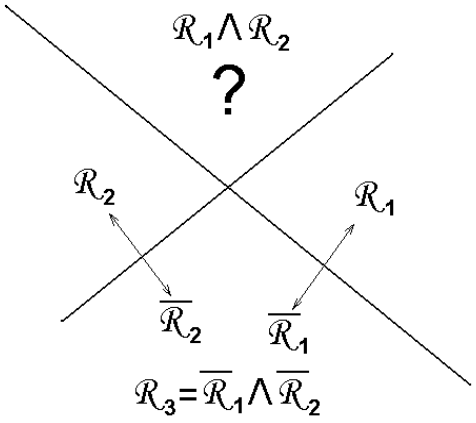
Vynásobíme-li obě strany tohoto vztahu $\mathbf{w}^T$ a přičteme w_0, pak dostaneme s použitím vztahů $b(\mathbf{x}) = \mathbf{w}^T.\mathbf{x}+w_0$ a $b(\mathbf{x}_\perp) = \mathbf{w}^T.\mathbf{x}_\perp+w_0 = 0$

[15]**Dichotomie** (z řec. *dicha,* oddělený, na dvakrát a *tome*, řez) označuje jakékoli rozdělení celku do dvou, navzájem disjunktních podmnožin. U některých kategorií je toto dělení naprosto přirozené, např. pohlaví nabývající dvou hodnot muž/žena, nebo polarita celých čísel – jsou kladná a záporná s jasnou hranicí. Na druhé straně jsou kategorie, které jsou sporné, např. kuřák/nekuřák – je ten, kdo vykouří jednu cigaretu za rok kuřákem, či nekuřákem, resp. patří ryzí nekuřák, žijící ve společnosti silného kuřáka, tedy pasivní kuřák, do kategorie kuřák nebo nekuřák? Konečně kategorie typu malý/velký. Z hlediska klasifikačních úloh je třeba vždy rozhodnout, zda daný objekt patří do jedné z obou kategorií, ovšem zařazení do těchto kategorií může být opatřeno velkou chybou danou skutečností, že většina z objektů se zpravidla nachází právě mezi extrémy obou kategorií.

$$d = \frac{|b(\mathbf{x})|}{\|\mathbf{w}\|}. \qquad (2.138)$$

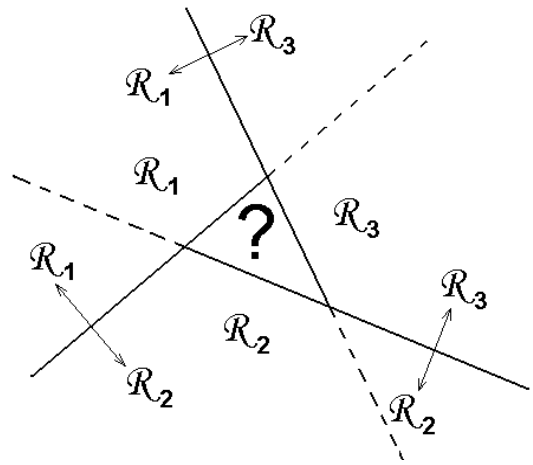
Uvažujme nyní případ více klasifikačních tříd, tj. R > 2. V tom případě se můžeme snažit vytvořit hraniční plochu kombinací několika dichotomických hraničních ploch. To nám ale může způsobit určité těžkosti.

Mějme R - 1 dichotomických klasifikátorů, které oddělují body patřící do třídy $\mathcal{R}_r$ od bodů, které do této třídy nepatří. Takovou klasifikaci nazýváme klasifikací „jedna versus zbytek“. Na obr.2.16 je znázorněn případ, kdy je tento klasifikační princip použit pro tři klasifikační třídy, což jak znázorněno vede k vytvoření oblasti v horní střední části příznakového prostoru, pro kterou neumíme jednoznačně rozhodnout, protože do ní spadají vektory, které patří do $\mathcal{R}_1$ i do $\mathcal{R}_2$. Problém nenastává s třídou $\mathcal{R}_3$, protože do té patří všechny obrazové vektory, které nepatří ani do třídy $\mathcal{R}_1$, ani do $\mathcal{R}_2$.



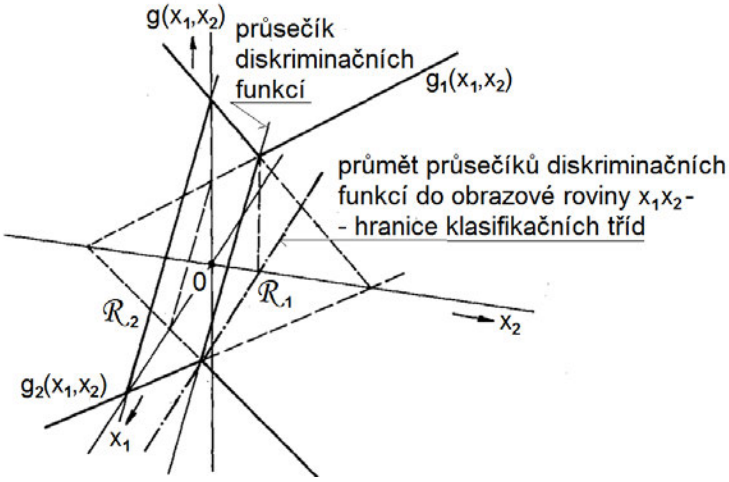
Obr.2.16 *Případ nejednoznačného rozdělení příznakového prostoru při použití klasifikace typu „jedna versus zbytek“*

Alternativou může být rozdělení příznakového prostoru pomocí R.(R-1)/2 dichotomických hraničních ploch pro oddělení každých dvou tříd. Tento způsob klasifikace nazýváme klasifikací „jedna versus jedna“. Každý vektor je pak zařazen do příslušné klasifikační třídy podle většinového pravidla. Obr.2.17 opět znázorňuje situaci pro tři klasifikační třídy, tj. R = 3, je tedy potřeba (3.2)/2 = 3 oddělující hraniční plochy. Prostor pro všechny uvažované třídy je vymezen většinovou platností dvou pravidel, v případě $\mathcal{R}_1$ platností pravidel $\mathcal{R}_1/\mathcal{R}_2$ a $\mathcal{R}_1/\mathcal{R}_3$, v případě $\mathcal{R}_2$ platností pravidel $\mathcal{R}_1/\mathcal{R}_2$ a $\mathcal{R}_2/\mathcal{R}_3$, … Ve středové oblasti nelze uplatnit většinové pravidlo, pravidlo $\mathcal{R}_1/\mathcal{R}_2$ ukazuje na výskyt oblasti $\mathcal{R}_2$, pravidlo $\mathcal{R}_1/\mathcal{R}_3$ na $\mathcal{R}_1$ a konečně pravidlo $\mathcal{R}_2/\mathcal{R}_3$ vede na $\mathcal{R}_3$. Tedy všechny tři oblasti jsou pro rozhodování zastoupeny rovnoměrně, tj. každá právě jednou a proto nelze rozhodnout.



Obr.2.17 *Případ nejednoznačného rozdělení příznakového prostoru při použití klasifikace typu „jedna versus jedna“*

Těmto problémům se lze vyhnout, použijeme-li principu diskriminačních funkcí a pro



Obr.2.18 *Konstrukce oddělovací hraniční přímky ve dvourozměrném obrazovém prostoru pomocí lineárních diskriminačních funkcí*

klasifikaci do R tříd použijeme R lineárních funkcí

$$g_r(\mathbf{x}) = \mathbf{w}_r^T.\mathbf{x} + w_{r0} \tag{2.139}$$

a poté dle definiční vlastnosti pro diskriminační funkce zařadíme vektor **x** do té klasifikační třídy $\mathcal{R}_r$, pro kterou platí $g_r(\mathbf{x}) > g_i(\mathbf{x})$ pro všechna $i \neq r$. Hraniční plocha mezi třídami je dána průmětem průsečíku obou diskriminačních rovin $g_r(\mathbf{x})$ a $g_i(\mathbf{x})$ do obrazového prostoru (obr.2.18), tedy

$$(\mathbf{w}_r - \mathbf{w}_i)^T.\mathbf{x} + (w_{r0} - w_{i0}) = 0, \tag{2.140}$$

což je funkce téhož charakteru, jako podle vztahu (2.135).

Poloha a orientace hraničních funkcí jsou jednoznačně určeny vektory svých váhových koeficientů $\mathbf{w}_r$, resp. $\widetilde{\mathbf{w}}_r$, r = 1, …, R. To znamená, že návrh klasifikátoru, resp. jeho rozhodovacího pravidla v tomto případě spočívá v určení optimálních hodnot těchto váhových koeficientů. Hovoříme-li o optimálních hodnotách, pak musí existovat nějaké kritérium, na základě kterého poznáme, že hodnoty váhových koeficientů jsou opravdu nejlepší. Dá se očekávat, že takových kritérií bude více a tato kritéria budou i tím, čím se metody popisované v následujících kapitolách budou lišit.

2.4.2 Metoda nejmenších čtverců

Princip metody nejmenších čtverců, především používaný zejména při řešení regresních úloh, je samozřejmě možné použít i pro stanovení koeficientů hraničních funkcí. Aby to bylo vskutku možné, je především třeba rozhodnout jak stanovit chyby, součet jejichž druhých mocnin určuje kriteriální funkci, jejíž hodnotu se snažíme minimalizovat.

Předpokládejme, že máme k dispozici trénovací množinu K obrazů $\{\mathbf{x}_k, \mathbf{t}_k\}$, k = 1, 2, …, K, kde tzv. cílový vektor $\mathbf{t}_k$ nese informaci o správné klasifikaci zakódovanou binárním kódem 1 z R (R je počet klasifikačních tříd). Tedy v případě, že obraz $\mathbf{x}_k$ patří do třídy ω_1, bude mít R rozměrný vektor $\mathbf{t}_k$ tvar $(1, 0, …., 0)^T$, bude-li obraz $\mathbf{x}_k$ patřit do třídy ω_R, bude cílový vektor $\mathbf{t}_k = (0, 0, …., 1)^T$. Tento způsob kódování aproximuje hodnotu podmíněné pravděpodobnosti zařazení vstupního obrazu **x** do požadované klasifikační třídy.

Každá třída ω_r bude popsána vlastní lineární diskriminační funkcí

$$g_r(\mathbf{x}) = \mathbf{w}_r^T.\mathbf{x} + w_{r0} \tag{2.141}$$

kde r = 1, …, R. Všechny funkce $g_r(\mathbf{x})$ můžeme sdružit a pomocí maticového zápisu vyjádřit jako

$$\mathbf{g}(\mathbf{x}) = \widetilde{\mathbf{W}}^T.\widetilde{\mathbf{x}}, \tag{2.142}$$

kde vektor $\mathbf{g}(\mathbf{x})$ je r-rozměrný, r-tý sloupec matice je $\widetilde{\mathbf{W}}$ tvořen (n+1)-rozměrným vektorem $\widetilde{\mathbf{w}}_r = (w_{r0}, \mathbf{w}_r^T)$ a $\tilde{\mathbf{x}}$ je rozšířený vstupní obrazový vektor $(1,\mathbf{x}^T)^T$ pro $x_0 = 1$. Rozměr matice $\widetilde{\mathbf{W}}$ je tedy (n+1)×r. Každý vstupní vektor je přiřazen do té klasifikační třídy ω_s, pro kterou je

$$g_s(\mathbf{x}) = \max_{\forall r} g_r(\mathbf{x}). \tag{2.143}$$

Úkolem metody nejmenších čtverců je nalézt pomocí trénovací množiny takové hodnoty matice parametrů $\widetilde{\mathbf{W}}$, aby byla minimalizována chybová kriteriální funkce

$$E_n(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\left[(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^T.(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})\right], \tag{2.144}$$

kde k-tý řádek trénovací obrazové matice $\widetilde{\mathbf{X}}$ je tvořen k-tým rozšířeným obrazovým vektorem trénovací množiny $\widetilde{\mathbf{x}}_k^T$ a matice **T** je matice, jejíž k-tý řádek tvoří cílový binární vektor požadovaných klasifikací $\mathbf{t}_k^T$. Položíme-li derivace funkce $E_n(\widetilde{\mathbf{W}})$ podle koeficientů $\widetilde{\mathbf{W}}$ rovné nule, dostaneme soustavu rovnic, jejíž řešení je

$$\widetilde{\mathbf{W}} = (\widetilde{\mathbf{X}}^T.\widetilde{\mathbf{X}})^{-1}.\widetilde{\mathbf{X}}^T.\mathbf{T} = \widetilde{\mathbf{X}}^+.\mathbf{T} \tag{2.145}$$

kde $\widetilde{\mathbf{X}}^+ = (\widetilde{\mathbf{X}}^T.\widetilde{\mathbf{X}})^{-1}.\widetilde{\mathbf{X}}^T$ je tzv. ***pseudoinverzní matice*** k matici $\widetilde{\mathbf{X}}$, kterou můžeme považovat za zobecnění inverzní matice pro obdélníkové matice.

Zajímavá a užitečná vlastnost řešení pomocí metody nejmenších čtverců je, že když cílové vektory trénovací množiny $\mathbf{t}_k$, k = 1, …, K, splňují lineární funkci

$$\mathbf{a}^T.\mathbf{t}_k + b = 0 \tag{2.146}$$

pro libovolné k a dané konstanty **a** a b, pak klasifikační model daný vztahem (2.141) splňuje pro libovolný obraz **x** tentýž vztah, tj.

$$\mathbf{a}^T.\mathbf{g}(\mathbf{x}) + b = 0. \tag{2.147}$$

To znamená, že pro kódovací schéma 1 z R pro R klasifikačních tříd, je součet prvků vektoru **g**(**x**) roven jedné stejně jako součet prvků vektoru $\mathbf{t}_k$ pro libovolný obrazový vektor **x**. Tento požadavek ale není postačující, protože hodnoty vektoru **g**(**x**) nejsou nutně vázány na interval ⟨0; 1⟩, což by bylo třeba, kdyby měly vyjadřovat odhady pravděpodobností zatřídění do jednotlivých klasifikačních kategorií. Tento nedostatek, kromě jiných jako je např. citlivost vůči odloučeným hodnotám, pak způsobuje, že tento algoritmus nedosahuje dostatečně spolehlivých výsledků.

Příklad:

Jednou z tradičních databází používaných pro ilustraci vlastností různých klasifikačních metod a zejména nastavení rozhodovacího pravidla klasifikátorů je tzv. ***Fisherova*** nebo také ***Andersonova databáze*** tří druhů kosatců (*Iris setosa*, *Iris virginica* a *Iris versicolor*)[16]. Obsahuje 3 x 50 vektorů, popisujících uvedené druhy kosatců pomocí čtyř příznakových proměnných – délkou a šířkou vnějších okvětních, nebo též korunních lístků (*petala*) a délkou a šířkou vnitřních okvětních, resp. kališních lístků (*sepala*).

Následující příklad prezentuje návrh lineární hraniční funkce (roviny) metodou nejmenších čtverců.

Pokud bychom chtěli použít metody nejmenších čtverců pro vytvoření všech hraničních rovin mezi třemi klasifikačními třídami ve čtyřrozměrném prostoru (pro čtyři příznakové proměnné), mají předpokládané diskriminační lineární funkce obecný tvar

$$g_r(\mathbf{x}) = w_{r0} + w_{r1}x_1 + w_{r2}x_2 + w_{r3}x_3 + w_{r4}x_4.$$

Předpokládejme, že v tomto vztahu x_1 reprezentuje délku a x_2 šířku kališních lístků a x_3 délku a x_4 šířku korunních lístků. Hraniční funkce jsou pak dány průměty průsečíků odpovídajících si diskriminačních funkcí do obrazového prostoru (viz obr.2.18).

Při použití úplné reprezentace dat metoda nejmenších čtverců vede pro jednotlivé druhy na následující diskriminační funkce:

Iris setosa

$$g_1(\mathbf{x}) = 0{,}1182 + 0{,}0660.x_1 + 0{,}2428.x_2 - 0{,}2247.x_3 - 0{,}0575.x_4;$$

Iris versicolor

$$g_2(\mathbf{x}) = 1{,}5771 - 0{,}0202.x_1 - 0{,}4456.x_2 + 0{,}2207.x_3 - 0{,}4943.x_4;$$

[16] např. http://en.wikipedia.org/wiki/Iris_flower_data_set (21.3.2012).

Iris virginica

$$g_3(\mathbf{x}) = - 0{,}6953 - 0{,}0459.x_1 + 0{,}2028.x_2 + 0{,}0040.x_3 + 0{,}5517.x_4.$$

Průsečíky jednotlivých oddělujících nadrovin potom jsou

$$b_{12}(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) = -1{,}4588 + 0{,}0862.x_1 + 0{,}6885.x_2 - 0{,}4453.x_3 + 0{,}4368.x_4 = 0;$$
$$b_{23}(\mathbf{x}) = g_2(\mathbf{x}) - g_3(\mathbf{x}) = 2{,}2720 + 0{,}0257.x_1 - 0{,}6484.x_2 + 0{,}2167.x_3 - 1{,}0461.x_4 = 0;$$
$$b_{13}(\mathbf{x}) = g_1(\mathbf{x}) - g_3(\mathbf{x}) = 0{,}8135 + 0{,}1119.x_1 + 0{,}0401.x_2 - 0{,}2286.x_3 - 0{,}6093.x_4 = 0.$$

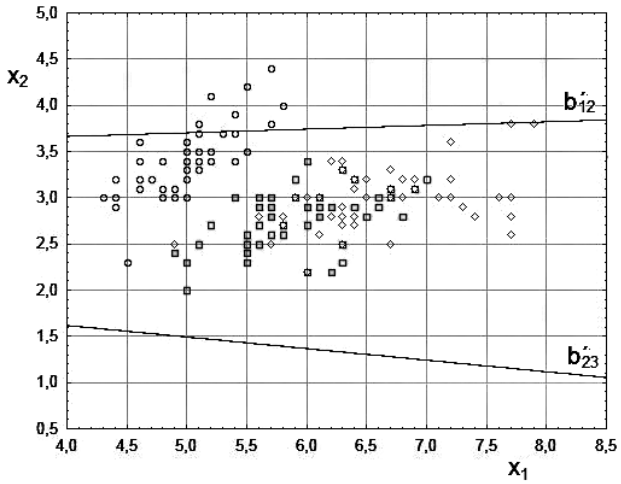
Chceme-li si ale výsledky zobrazit v rozumné, smyslově vhodně vstřebatelné reprezentaci, tj. nejlépe pomocí dvojdimenzionálních grafů, pak nezbývá, než najít průsečík oddělujících nadrovin s tou částí obrazového prostoru, kterou chceme zobrazit. Na příklad, zobrazení oddělujících ploch do roviny x_1x_2 dosáhneme tak, že souřadnice x_3 a x_4 ve výrazech pro $g_{12}(\mathbf{x})$, $g_{23}(\mathbf{x})$ a $g_{13}(\mathbf{x})$ položíme rovny nule a dostáváme

$$b'_{12}(\mathbf{x}) = g'_1(\mathbf{x}) - g'_2(\mathbf{x}) =$$
$$= -1{,}4588 + 0{,}0862.x_1 + 0{,}6885.x_2 = 0;$$

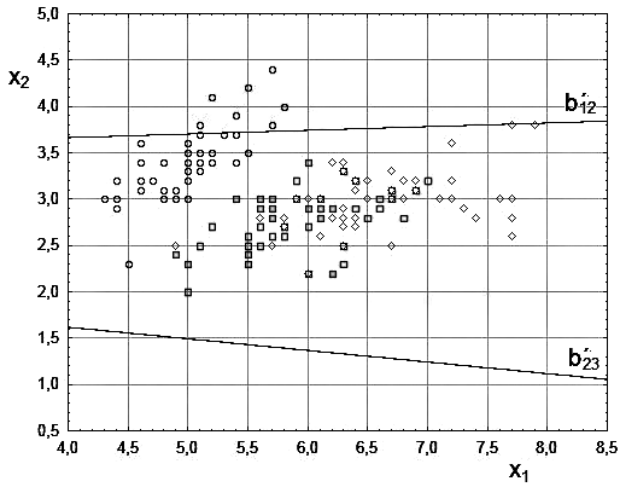
$$b'_{23}(\mathbf{x}) = g'_2(\mathbf{x}) - g'_3(\mathbf{x}) =$$
$$= 2{,}2720 + 0{,}0257.x_1 - 0{,}6484.x_2 = 0;$$

$$b'_{13}(\mathbf{x}) = g'_1(\mathbf{x}) - g'_3(\mathbf{x}) =$$
$$= 0{,}8135 + 0{,}1119.x_1 + 0{,}0401.x_2 = 0.$$

Grafické zobrazení uvedených funkcí (obr.2.19) ale není nijak přesvědčivé. Nevypadá, že bylo správné. Zde je ovšem třeba si uvědomit, že diskriminační funkce jsou určeny dle kritéria nejmenších čtverců s ohledem na všechny příznakové proměnné. Dále, že metoda nejmenších čtverců trpí jistými nepřesnostmi způsobenými charakterem rozložení obrazů v jednotlivých klasifikačních třídách. A konečně, že zobrazené dělení obrazového prostoru je jen dílčí průmět, který zjevně nevystihuje vliv dvou odstraněných proměnných.



Obr.2.19 *Příklad hraničních přímek pro klasifikaci tří druhů kosatců určených metodou nejmenších čtverců – Iris setosa (○), Iris versicolor (□) a Iris virginica(◇); x_1 – délka kališních lístků, x_2 – šířka kališních lístků*



Obr.2.20 *Příklad hraničních přímek pro klasifikaci tří druhů kosatců určených metodou nejmenších čtverců ve dvourozměrném příznakovém prostoru – Iris setosa (○), Iris versicolor (□) a Iris virginica(◇); x_1 – délka kališních lístků, x_2 – šířka kališních lístků*

Abychom dosavadní dojem napravili, pokusme se použít kritéria nejmenších čtverců na vektory obsahující pouze proměnné x1 a x2. V tom případě jsou diskriminační funkce

Iris setosa

$$g_1(x_1,x_2) = 0{,}7753 - 0{,}3744.x_1 + 0{,}5711.x_2;$$

Iris versicolor

$$g_2(x_1,x_2) = 1{,}7928 + 0{,}0141.x_1 - 0{,}5044.x_2;$$

Iris virginica

$$g_3(x_1,x_2) = - 1{,}5681 + 0{,}3603.x_1 + 0{,}0667.x_2,$$

oddělující přímky mají tvar

$$b'_{12}(x_1,x_2) = g'_1(x_1,x_2) - g'_2(x_1,x_2) = -1{,}0175 - 0{,}3886.x_1 + 1{,}0755.x_2 = 0;$$
$$b'_{23}(x_1,x_2) = g'_2(x_1,x_2) - g'_3(x_1,x_2) = 3{,}3607 - 0{,}3462.x_1 - 0{,}4377.x_2 = 0;$$
$$b'_{13}(x_1,x_2) = g'_1(x_1,x_2) - g'_3(x_1,x_2) = 2{,}3433 - 0{,}7347.x_1 + 0{,}6378.x_2 = 0$$

a jsou zobrazeny na obr.2.20. Pro rozdělení příznakového obrazového prostoru je vhodné použít pouze hraničních polopřímek. Výsledky jsou vizuálně výrazně lepší než v předcházejícím případě, nicméně i v tomto případě jsou výsledky za očekáváním. Ne zcela optimální polohy hraničních přímek vyplývají z vlastností rozložení obrazových vektorů – nejsou kompaktně unimodální, v jednotlivých množinách se vyskytují různé odlehlé hodnoty, klasifikační třídy se překrývají, nejsou lineárně separabilní, což jsou všechno nectnosti zmíněné v popisu metody. □□□

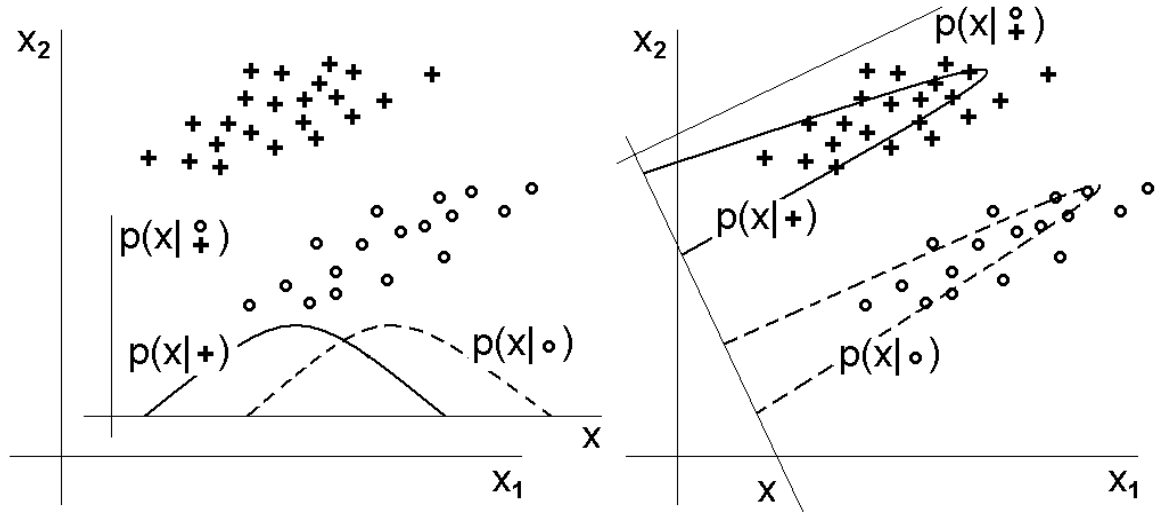
2.4.3 Fisherova lineární diskriminace

Problém lineární klasifikace lze také nahlížet z hlediska potřeb redukce dimenzionality klasifikovaných obrazových vektorů.

Pro základní vysvětlení principu předpokládejme dichotomickou klasifikační úlohu. Dále předpokládejme lineární transformaci původního n-rozměrného obrazového vektoru **x** do pouhého jednoho rozměru pomocí vztahu

$$y(\mathbf{x}) = \mathbf{w}^T.\mathbf{x} \; . \qquad (2.148)$$

Použijeme-li pro klasifikaci prahové hodnoty w_0 tak, že je-li hodnota $y(\mathbf{x}) \geq -w_0$, zařadíme obraz **x** do třídy ω_1 a v opačném případě do třídy ω_2, pak daná úloha odpovídá v dřívějších kapitolách diskutované lineární diskriminaci. Vážeme-li ale princip klasifikace na drastickou redukci obrazového prostoru (z původních n rozměrů na jediný), potom se na první pohled může zdát, že takové značné omezení informace obsažené v původním obrazovém vektoru může významně snížit kvalitu rozhodovacího procesu. Ovšem vhodným nastavením hodnot váhového vektoru **w** můžeme vytvořit takovou projekci, která maximalizuje možnou separaci obou klasifikačních tříd (viz obr.2.21). Zatímco v levé části obrázku se při průmětu do směru rovnoběžného s osou x_1 hustoty pravděpodobnosti výskytu obrazů z obou klasifikačních tříd



Obr.2.21 *Princip řízeného snížení dimenzionality obrazových vektorů*

významně překrývají, při vhodné volbě průmětu v pravé části obrázku mohou být obě klasifikační třídy pohodlně a spolehlivě odděleny.

Abychom nalezli takový výhodný směr projekce, určeme nejdříve průměrné obrazy, které budou představovat etalony obou klasifikačních tříd. Předpokládejme tedy, že se v třídě ω_1 vyskytuje K_1 obrazů a ve třídě ω_2 K_2 obrazů. Pak jsou průměrné vektory obou klasifikačních tříd dány vztahy

$$\mathbf{m_1} = \frac{1}{K_1} \sum_{k \in \omega_1} \mathbf{x}_k \quad \text{a} \quad \mathbf{m}_2 = \frac{1}{K_2} \sum_{k \in \omega_2} \mathbf{x}_k \,. \tag{2.149}$$

Nejjednodušší možnou mírou separace obou tříd při vhodné volbě koeficientů váhového vektoru **w** je vzdálenost projekcí obou průměrných obrazů a to vede na maximalizaci vzdálenosti obou průmětů

$$m_2 - m_1 = \mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_1), \tag{2.150}$$

kde

$$m_r = \mathbf{w}^T \mathbf{m}_r \,. \tag{2.151}$$

Ovšem tento výraz může být uměle zvětšován růstem modulu váhového vektoru **w**. Tuto potíž lze ale snadno odstranit zavedením požadavku na nějakou standardní velikost vektoru **w**, nejlépe s jednotkovou normou, tj. $\Sigma_i w_i^2 = 1$. Taková definice optimalizační úlohy vede na ***metodu Langrangova součinitele*** pro hledání vázaného extrému.

Tato metoda vychází z následující věty – nejdříve pro dvě proměnné:

Nechť funkce f(x,y) a g(x,y) mají v okolí bodů křivky g(x,y) = 0 totální diferenciál[17]. Dále, nechť v každém bodě křivky g(x,y) = 0 je alespoň jedna z derivací $\partial g/\partial x$, $\partial g/\partial y$ různá od nuly. Má-li pak funkce f(x,y,) v bodě (x_0,y_0) křivky g(x,y) = 0 na této křivce lokální extrém, pak existuje taková konstanta λ, že pro funkci

$$F(x,y)=f(x,y) + \lambda.g(x,y) \tag{2.152}$$

jsou v bodě (x_0,y_0) splněny rovnice (podmínky nutné)

$$\begin{aligned} \partial F(x_0,y_0)/\partial x=0; \\ \partial F(x_0,y_0)/\partial y=0 \end{aligned} \tag{2.153}$$

a samozřejmě $g(x_0,y_0)=0$.

Vázané extrémy lze tedy hledat tak, že sestrojíme funkci (2.152) a řešíme rovnice (2.153) pro neznámé x_0,y_0, λ (parametr λ nazýváme Lagrangeův součinitel (multiplikátor)).

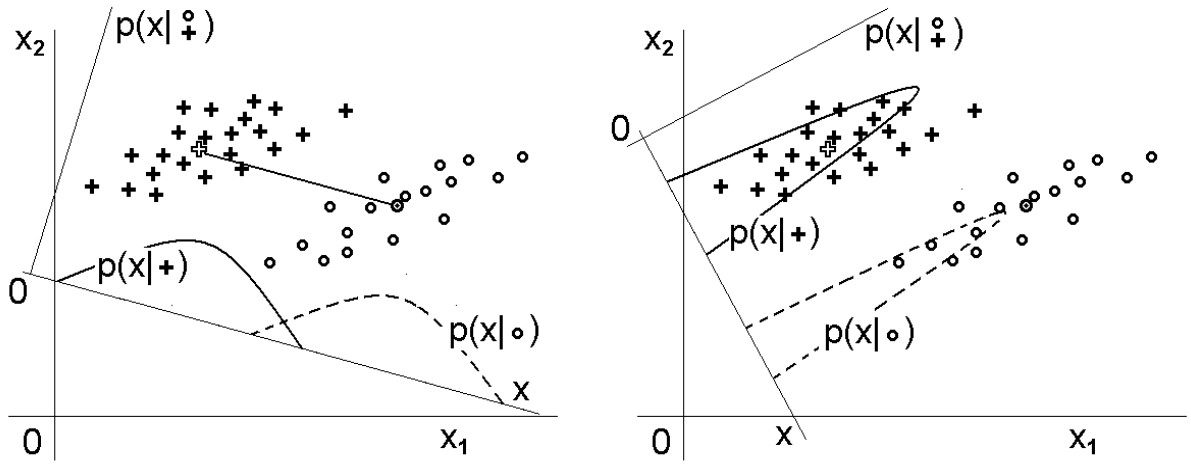
Podmínky postačující vyplývají z výpočtu druhého diferenciálu funkce (2.152) v bodě (x_0,y_0)

$$d^2F(x_0,y_0)= \partial^2F(x_0,y_0)/\partial x^2+2\partial^2F(x_0,y_0)/\partial x\, \partial y +\partial^2F(x_0,y_0)/\partial y^2. \tag{2.154}$$

Jestliže pro všechny body (x_0+dx,y_0+dy) z určitého okolí bodu (x_0,y_0) takové, že $g(x_0+dx,y_0+dy) = 0$ a současně dx a dy nejsou rovny nule, je druhý diferenciál podle (2.154) kladný, resp záporný, pak je v bodě (x_0,y_0) vázaný lokální extrém, a to minimum, resp. maximum.

Obdobně se řeší nalezení vázaných extrémů funkce několika proměnných. Např. nutná podmínka k existenci lokálního extrému funkce w=f(x,y,z,u,v) při podmínkách $F_1(x,y,z,u,v)$, $F_2(x,y,z,u,v)$ je splnění rovnic

[17] ***Totální diferenciál funkce*** z=f(x,y) se nazývá, za předpokladu, že je funkce f(x,y) v bodě (x_0,y_0) diferencovatelná, výraz $dz =(\partial f/\partial x).dx + (\partial f/\partial y).dy$.



Obr.2.22 *Řízené snížení dimenzionality obrazových vektorů se zohledněním poloh průměrných etalonů (vlevo) a se zahrnutím směrových vlastností rozptylu (vpravo)*

$$\partial G/\partial x = 0, \quad \partial G/\partial y = 0, \quad \partial G/\partial z = 0, \quad \partial G/\partial u = 0, \quad \partial G/\partial v = 0, \quad F_1 = 0 \text{ a } F_2 = 0, \qquad (2.155)$$

kde $G = f + \lambda_1 F_1 + \lambda_2 F_2$, tj. řešení úlohy spočívá v řešení soustavy sedmi rovnic pro sedm neznámých.

Na obr.2.22 vlevo je znázorněna situace, kdy je diskriminační směr určen spojnicí průměrných etalonů obou množin. Vidíme, že ani toto řešení, vyplývající ze vztahu (2.150) nereprezentuje ideální stav, v průmětu do tohoto směru se hustoty pravděpodobnosti stále překrývají. Řešení by ale mohlo být modifikováno, pokud by bylo možné zahrnout také informaci o směrových vlastnostech rozptylu v obou množinách.

Transformační vztah (2.148) převádí n-rozměrné vektory **x** do jednorozměrného prostoru. Rozptyl uvnitř jednotlivých do jednoho rozměru transformovaných množin je pak dán vztahem

$$s_r^2 = \sum_{k \in \omega_r} (y_k - m_r)^2, \qquad (2.156)$$

kde $y_k = \mathbf{w}^T.\mathbf{x}_k$. Celkový rozptyl uvnitř tříd v dichotomickém případě můžeme určit součtem dílčích rozptylů $s_1^2 + s_2^2$. Optimalizační kritérium, které zohledňuje jak vzdálenost mezi třídami (tu chceme průmětem maximalizovat), tak i rozptyl uvnitř tříd (ten naopak chceme průmětem minimalizovat) můžeme v tomto případě vyjádřit formulí

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}. \qquad (2.157)$$

S využitím vztahů (2.148), (2.151) a (2.156) můžeme Fisherovo diskriminační kritérium přepsat do obecného tvaru

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}, \qquad (2.158)$$

kde $\mathbf{S}_B$ je kovarianční matice mezi třídami definovaná vztahem

$$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1).\ (\mathbf{m}_2 - \mathbf{m}_1)^T \qquad (2.159)$$

a $\mathbf{S}_W$ je celková kovarianční matice uvnitř tříd definovaná jako

$$\mathbf{S}_W = \sum_{k\in\omega_1}(\mathbf{x}_k - \mathbf{m}_1).(\mathbf{x}_k - \mathbf{m}_1)^T + \sum_{k\in\omega_2}(\mathbf{x}_k - \mathbf{m}_2).(\mathbf{x}_k - \mathbf{m}_2)^T. \qquad (2.160)$$

Derivováním (2.158) podle váhových koeficientů vektoru **w** dostaneme, že J(**w**) nabývá maxima, když

$$(\mathbf{w}^T\mathbf{S}_B\mathbf{w})\,\mathbf{S}_W\mathbf{w} = (\mathbf{w}^T\mathbf{S}_W\mathbf{w})\,\mathbf{S}_B\mathbf{w}. \qquad (2.161)$$

Z (2.157) plyne, že $\mathbf{S}_B\mathbf{w}$ má vždy směr $\mathbf{m}_2 - \mathbf{m}_1$. Dále, protože modul vektoru **w** není důležitý, zajímá nás pouze jeho směr, můžeme ve vztahu (2.161) pominout oba skalární členy $(\mathbf{w}^T\mathbf{S}_B\mathbf{w})$ a $(\mathbf{w}^T\mathbf{S}_W\mathbf{w})$ a vynásobením obou stran tohoto výrazu zleva $\mathbf{S}_W^{-1}$ dostaneme

$$\mathbf{w} \sim \mathbf{S}_W^{-1}.(\mathbf{m}_2 - \mathbf{m}_1). \qquad (2.162)$$

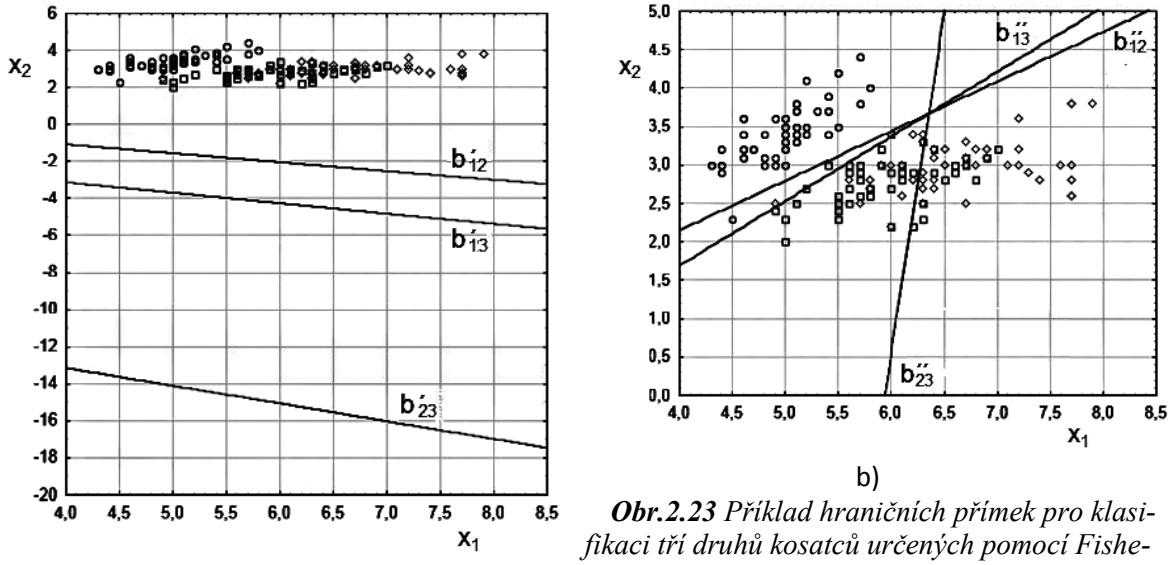
Pokud je rozptyl uvnitř obou tříd stejný ve všech směrech, tzn. $\mathbf{S}_W$ je diagonální matice se všemi prvky o stejné velikosti, jinými slovy je úměrná jednotkové matici (a také, že směrové vlastnosti rozptylu nejsou podstatné), pak je **w** úměrné pouze rozdílu průměrných etalonů obou tříd, jak jsme již dříve konstatovali na základě vztahu (2.150). Vztah (2.162) se obecně označuje jako Fisherův lineární diskriminant, i když ve skutečnosti o žádnou diskriminační funkci nejde, pouze o určení směru té jedné souřadnice, do které se promítají původní n-rozměrná data. Ovšem tento určuje směrnici hraniční plochy, protože ta je kolmá na směr vektoru **w** (viz obr.2.15) a její konkrétní pozici jednoznačně stanoví hodnota prahu w_0, se kterým vektor **x** patří do třídy ω_1, pokud $g(\mathbf{x}) \geq w_0$ a v opačném případě do třídy ω_2. Hodnotu prahu můžeme určit např. bayesovskými metodami popsanými v kap.2.2.

Příklad

Navažme na příklad řešený v kapitole o metodě nejmenších čtverců a ověřme, jak se Fisherova metoda vypořádá s nastavením hraničních funkcí pro rozdělení obrazového prostoru pro Fisherovu databázi kosatců.

Ani tentokrát v případě zobrazení průmětů čtyřrozměrných hraničních ploch do dvourozměrného obrazového prostoru není situace nijak skvělá, dokonce poloha hranic je ještě horší než v případě metody nejmenších čtverců. Rovnice oddělujících hraničních rovin jsou

$$b_{12}(\mathbf{x}) = -13.46 + 7{,}85.x_1 + 16{,}51.x_2 - 21{,}64.x_3 - 23{,}82.x_4 = 0;$$



a) b)

Obr.2.23 *Příklad hraničních přímek pro klasifikaci tří druhů kosatců určených pomocí Fisherovy lineární diskriminace – Iris setosa (○), Iris versicolor (□) a Iris virginica(◇); x_1 – délka kališních lístků, x_2 – šířka kališních lístků – a) ve čtyrozměrném prostoru; b) ve dvourozměrném prostoru*

$$b_{23}(\mathbf{x}) =: \ 31{,}51 + 3{,}25.\ x_1 + 3{,}39.\ x_2 - 7{,}55.\ x_3 - 14{,}64.\ x_4 = 0;$$
$$b_{13}(\mathbf{x}) = 18{,}05 + 11{,}10.\ x_1 + 19{,}90.\ x_2 - 29{,}19.\ x_3 - 38{,}46.\ x_4 = 0$$

a jejich průměty do souřadnic x_1 a x_2 - x_1 stejně jako v předešlém příkladu reprezentuje délku a x_2 šířku kališních lístků - tedy jsou (získáme je tak, že x_3 a x_4 ve výše uvedených vztazích položíme rovny nule)

$$b'_{12}(\mathbf{x}) = -\ 13.46 + 7{,}85.x_1 + 16{,}51.x_2 = 0;$$
$$b'_{23}(\mathbf{x}) = 31{,}51 + 3{,}25.x_1 + 3{,}39.x_2 = 0;$$
$$b'_{13}(\mathbf{x}) = 18{,}05 + 11{,}10.x_1 + 19{,}90.x_2 = 0.$$

Grafické zobrazení těchto hraničních přímek je na obr.2.23a. Zásadní důvod zůstává týž jako v případě metody nejmenších čtverců – při zobrazení není zohledněn vliv dvou proměnných, které se podílely na stanovení váhových koeficientů hraničních ploch.

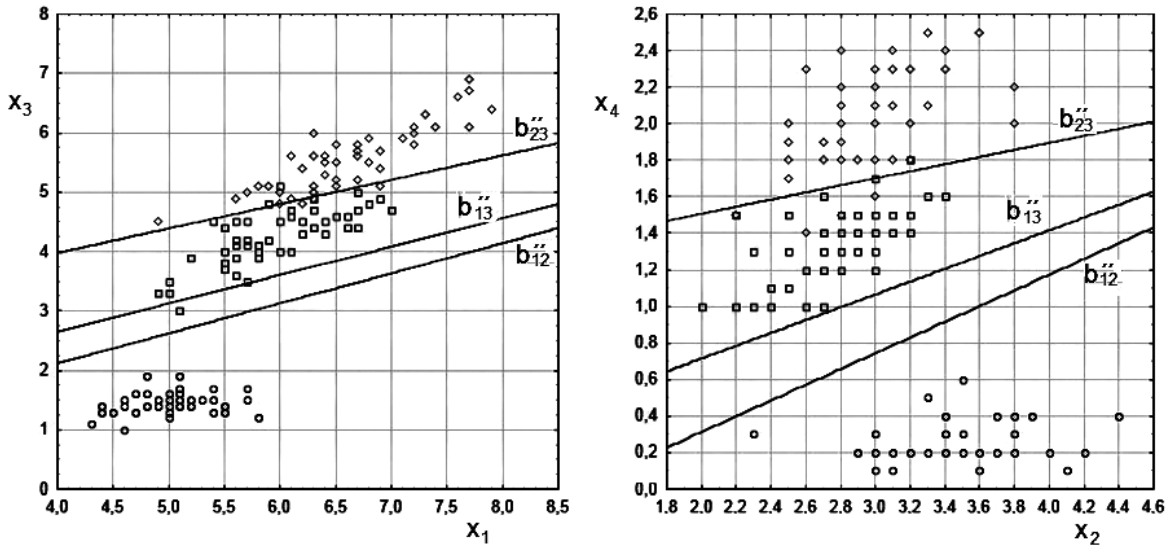
Jak proto vypadá situace při optimálním stanovení klasifikačních hranic přímo ve dvourozměrném prostoru. V tom případě jsou hraniční přímky určeny rovnicemi

$$b''_{12}(x_1,x_2) = 5{,}1528 - 7{,}6574.x_1 + 11{,}8557.x_2 = 0;$$
$$b''_{23}(x_1,x_2) = 15{,}2084 - 2{,}5620.x_1 + 0{,}2908.x_2 = 0;$$
$$b''_{13}(x_1,x_2) = 20{,}3612 - 10{,}2194.x_1 + 12{,}1465.x_2 = 0,$$

které jsou graficky zobrazeny na obr.2.23b.

Srovnáním obr.2.20 a 2.23b lze konstatovat, že hraniční přímky b''_{12} i b''_{13} pro případ Fisherovy diskriminace viditelně podstatně lépe respektují rozložení obrazových vektorů v prostoru, v případě tříd 1 a 2 nedochází k žádné chybné klasifikaci, v případě tříd 1 a 3 k výrazně menšímu počtu chybných klasifikací. Hraniční přímka b''_{23} (tj. pro výrazně se překrývající klasifikační třídy 2 a 3) má zcela jiný směr. To zda je její poloha v případě Fisherovy diskriminační metody výhodnější z hlediska počtu chybných klasifikací, by vyžadovalo podrobnější analýzu. Pro klasifikaci by bylo zřejmě možné použít hranic b''_{12} a b''_{23}., kterými je obrazový prostor rozdělen do čtyř částí, z nichž výsek vpravo nahoře na obr.2.23b neodpovídá žádné z předpokládaných klasifikačních tříd.

Pro zajímavost lze uvést hraniční přímky určené pomocí Fisherovy lineární diskriminace i pro další proměnné.



a) b)

Obr.2.24 *Příklad hraničních přímek pro klasifikaci tří druhů kosatců určených pomocí Fisherovy lineární diskriminace ve dvourozměrném prostoru – Iris setosa (○), Iris versicolor (□) a Iris virginica(◇) pro příznakové proměnné a) x_1 – délka kališních lístků, x_3 – délka korunních lístků; b) x_2 – šířka kališních lístků, x_4 – šířka korunních lístků*

Dvourozměrný obrazový prostor s příznakovými proměnnými x_1 (délka kališních lístků) a x_3 (délka korunních lístků) je rozdělen hraničními přímkami (obr.2.24a)

$$b''_{12}(x_1,x_3) = 2{,}5532 + 14{,}1079.x_1 - 27{,}8704.x_3 = 0;$$
$$b''_{23}(x_1,x_3) = 25{,}9218 + 4{,}5531.x_1 - 11{,}0954.x_3 = 0;$$
$$b''_{13}(x_1,x_3) = 28{,}4750 - 18{,}6610.x_1 - 38{,}9658.x_3 = 0,$$

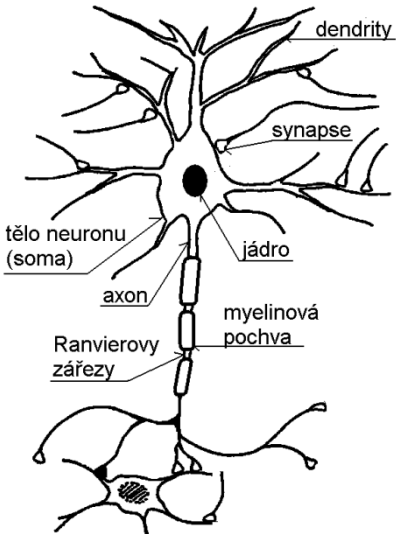
s příznakovými proměnnými x_2 (šířka kališních lístků) a x_4 (šířka korunních lístků) je rozdělen hraničními přímkami (obr.2.24b)

$$b''_{12}(x_2,x_4) = -21{,}2650 + 16.7129.x_2 - 38.8400.x_4 = 0;$$
$$b''_{23}(x_2,x_4) = 22.0499 + 3.8146.x_2 - 19.6931.x_4 = 0;$$
$$b''_{13}(x_2,x_4) = 0.7849 + 20.5275.x_2 - 58.5331.x_4 = 0,$$

Na obou obrázcích, stejně tak jako na obr.2.23 lze vidět poněkud problematické překrývání klasifikačních tříd *Iris versicolor* a *Iris virginica*, v obou případech tedy vlivem překryvu dochází k chybnému zatřídění. Na druhé straně, kosatce druhu *Iris setosa* zaujímají odlišné části obrazového prostoru bez jakéhokoliv překrytí s oběma dalšími klasifikačními třídami, hraniční přímky b''_{12} i b''_{13} nevyvolávají jakékoliv pochybnosti o správné klasifikaci. Vzhledem k rozmístění obrazů v příznakovém prostoru lze pro klasifikaci v obou zobrazených případech použít hranice b''_{12} a b''_{23}, které pro reálné hodnoty příznaků dělí obrazový příznakový prostor právě do tří segmentů, odpovídajících jednotlivým klasifikačním třídám. □□□

2.4.4 Jednovrstvý perceptron

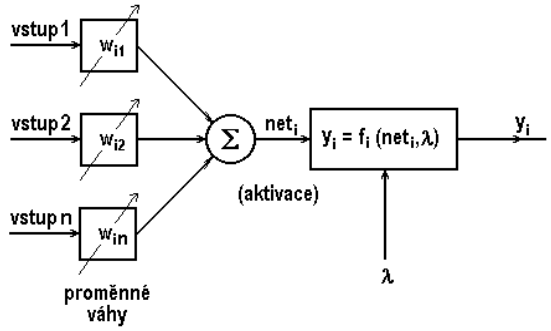
Jednovrstvý perceptron je základní strukturou vrstevnatých neuronových sítí[18]. Neuronová síť reprezentuje jeden z možných výpočetních postupů zpracování dat inspirovaný funkcí nervové soustavy. Kromě možného využití v takových matematických plinách jako je klasifikace a rozpoznávání, predikce, filtrace dat, mohou neuronové sítě sloužit i jako matematický model, sloužící ke zkoumání funkce reálné nervové soustavy.



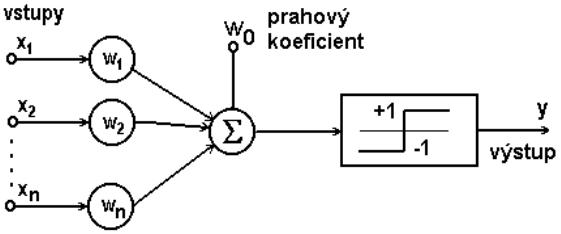
Obr.2.25 *Schématické znázornění neuronu*

Základním výpočetním prvkem neuronové sítě je umělý (formální) neuron, který je v neuronové síti zapojen podle určitých specifikovaných topologických schémat. Vzorem umělého neuronu je základní buňka nervové soustavy – ***neuron*** (obr.2.25). Neuron je strukturní i funkční jednotkou nervové soustavy.Tělo neuronu, které jako každá eukaryotická buňka obsahuje jádro, je opatřeno krátkými, několik milimetrů dlouhými, vstupními výběžky, tzv. ***dendrity***. Prostřednictvím dendritů přijímá neuron eferentní vzruchy z okolních (často tisíců) neuronů připojených na dendrity pomocí ***synapsí***. Vlastnosti sousedních neuronů i typ synapsí určují charakter informace přiváděné do těla neuronu (budicí, tlumicí). Synaptický přenos u savců probíhá téměř výhradně pomocí chemických procesů, u nižších živočichů bývá i elektrický. Synaptický přenos informace může být modifikován (excitačně i inhibičně) i jinými neurony. Neuronem zpracovaná informace se předává na další neurony prostřednictvím dlouhého výstupního vlákna - ***axonu***. Jeho délka bývá i desítky centimetrů.

[18] V těchto textech se budeme zabývat pouze nejjednodušší formou neuronové sítě, která umožňuje stejně jako ostatní metody v této kapitole lineární diskriminaci klasifikačních tříd. Podrobnějším popisem dalších sofistikovanějších struktur neuronových sítí se zabývá např. předmět Umělá inteligence.



Obr.2.26 *Základní výpočetní schéma umělého neuronu*



Obr.2.27 *Lineární výpočetní schéma umělého neuronu*

Axon je zpravidla obalen myelinovou vrstvou přerušovanou přibližně po 1,5 mm tzv. ***Ranvierovými zářezy***. Toto uspořádání umožňuje urychlit šíření elektrického vzruchu podél axonu. Na konci se axon dělí na terminální vlákna, která se prostřednictví synapsí dále připojují na okolní neurony.

Existuje celá řada matematických modelů neuronu. Nejběžnější tradiční formu má model, jehož schéma je na obr.2.26. Vstupy reprezentují dendrity a informace přivedená na vstupy je dále upravena obecně proměnnými váhovými koeficienty, které reprezentují vliv synapsí. Váhované vstupy ze všech dendritů jsou kumulovány a výsledek této kumulace je posléze podroben nelineární transformaci pomocí tzv. aktivační, resp. výstupní funkce. Nelineárně upravená informace se posléze objeví na výstupu.

Zjednodušenou lineární variantou tohoto schématu je zapojení na obr.2.27. Obsahuje pouze konstantní váhové koeficienty na vstupech a nelineární zpracování součtu všech váhovaných vstupů je reprezentováno prahovou funkcí. Matematicky je toto výpočetní schéma reprezentováno vztahem

$$y(\mathbf{x}) = \mathbf{w}^T.\mathbf{x} + w_0, \qquad (2.163)$$

což je známá formule použitá v těchto textech již několikrát (srvn. např. vztah (2.6) v kap.2.2.1, vztahy (2.133), resp. (2.134) v kap.2.4.1, atd). Znamená to, že výpočet v jednom neuronu reprezentuje jednu lineární hranici v obrazovém příznakovém prostoru a tím dělení obrazového prostoru na dvě dichotomické poloroviny. Pokud má klasifikátor třídit obrazy do více klasifikačních tříd, pak je potřeba rozdělit obrazový prostor odpovídajícím počtem hraničních přímek, každá reprezentovaná jedním neuronem, přičemž na všechny potřebné umělé neurony je přiváděna táž informace, reprezentovaná obrazovými vektory. Takové zapojení je v podstatě vyjádřeno schématem na obr.2.14 (pokud funkci (2.163) vnímáme jako diskriminační, příp. je v obrázku blok výběru maxima nahrazen blokem logických pravidel, pomocí kterých klasifikátor rozhoduje, do které části lineárně rozděleného obrazového prostoru vstupní obraz zařadí). Jednotlivé neurony jsou tedy zapojeny v jedné vrstvě a tato neuronová vrstva je následována vyhodnocovacím blokem. Protože historicky první neuronové sítě s vrstevnatou strukturou byly použity pro modelování funkce zraku a rozpoznávání geometrických obrazců, začalo se neuronovým sítí s touto topologií říkat ***perceptrony***.

Důležitou úlohou je určení hodnot váhových koeficientů na základě informací obsažených v trénovací množině tak, aby dělení obrazového prostoru a tím i funkce klasifikátoru byly optimální.

Obecné požadavky na postup nastavení hodnot váhových koeficientů perceptronu (a nejen perceptronu) jsou:

- algoritmická formulace – tj. metoda musí najít řešení pomocí konečného počtu dílčích kroků;
- konvergence – výpočet by měl být monotónní a pokud možno co nejrychlejší.

Učení perceptronu (obecně jakéhokoliv klasifikátoru) probíhá na základě následujících pravidel:

- na vstup perceptronu jsou vkládány prvky trénovací množiny a výsledek klasifikace je srovnán s očekávanou správnou klasifikací;
- pokud je rozdíl mezi klasifikátorem určenou a správnou klasifikací větší než určitá předem daná mez definující přípustnou chybu, pak se parametry klasifikátoru (váhové koeficienty) změní tak, aby se chyba mezi určenou a požadovanou klasifikací minimalizovala;
- pokud je chyba klasifikace větší než předem stanovená mez, pak učení dále pokračuje, v opačném případě se učení ukončí a klasifikátor je možné použít ke klasifikaci. Kromě zmíněného absolutního kritéria ukončení učicí se fáze se v současné době často používá pro zastavení učení i relativní kritérium založené poklesu chyby během daného časového okna.

Učení perceptronu lze tedy považovat za typickou optimalizační úlohu. Abychom ji dokázali vyřešit, je třeba použít vhodnou kriteriální (ztrátovou) funkci a přiměřený algoritmus, který dokáže najít extrém kriteriální funkce.

Pro následující výklad předpokládejme dichotomickou klasifikační úlohu, pro kterou platí

$$\begin{aligned} \tilde{\mathbf{w}}^{\mathrm{T}}\tilde{\mathbf{x}} > 0, \quad & \text{když } \mathbf{x} \in \omega_1; \\ \tilde{\mathbf{w}}^{\mathrm{T}}\tilde{\mathbf{x}} < 0, \quad & \text{když } \mathbf{x} \in \omega_2. \end{aligned} \tag{2.164}$$

Jako kriteriální funkci použijme funkci

$$\mathrm{J}(\tilde{\mathbf{w}}) = \sum_{\mathbf{x}\in\mathcal{Y}} \delta_{\mathbf{x}} \tilde{\mathbf{w}}^{\mathrm{T}}\tilde{\mathbf{x}} \tag{2.165}$$

kde $\mathcal{Y}$ je podmnožina trénovací množiny, která obsahuje chybně klasifikované obrazy pomocí hraniční roviny definované hodnotami váhového vektoru $\tilde{\mathbf{w}}$. Proměnou $\delta_{\mathbf{x}}$ volme tak, že $\delta_{\mathbf{x}} = -1$, když $\mathbf{x} \in \omega_1$ a $\delta_{\mathbf{x}} = 1$, když $\mathbf{x} \in \omega_2$. Je-li $\mathbf{x} \in \omega_1$ a je chybně klasifikován, pak je $\tilde{\mathbf{w}}^{\mathrm{T}}\tilde{\mathbf{x}} < 0$ a $\delta_{\mathbf{x}} < 0$ a součin $\delta_{\mathbf{x}}\tilde{\mathbf{w}}^{\mathrm{T}}\tilde{\mathbf{x}}$ je kladný. Podobně je součin kladný, když $\mathbf{x} \in \omega_2$ a je chybně klasifikován, pak je $\tilde{\mathbf{w}}^{\mathrm{T}}\tilde{\mathbf{x}} > 0$ i $\delta_{\mathbf{x}} > 0$. Z toho plyne, že funkce $\mathrm{J}(\tilde{\mathbf{w}})$ je nezáporná a rovna nule, pokud je množina $\mathcal{Y}$ prázdná, tj. všechny obrazy jsou klasifikovány správně.

Kriteriální funkce je spojitá a po částech lineární. Měníme-li hodnoty váhového vektoru pouze nepatrně, mění se hodnoty $\mathrm{J}(\tilde{\mathbf{w}})$ lineárně až do chvíle, kdy se změní počet chybně klasifikovaných obrazů. V tomto bodě je gradient funkce nedefinován a gradientní funkce nespojitá.

Pro nalezení minima funkce $\mathrm{J}(\tilde{\mathbf{w}})$ se zpravidla používá gradientní algoritmus, definovaný iterační formulí

$$\tilde{\mathbf{w}}(\mathrm{m}+1) = \tilde{\mathbf{w}}(\mathrm{m}) - \rho_{\mathrm{m}} \left.\frac{\partial \mathrm{J}(\tilde{\mathbf{w}})}{\partial \tilde{\mathbf{w}}}\right|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}(\mathrm{m})}, \tag{2.166}$$

kde $\tilde{\mathbf{w}}(\mathrm{m})$ je váhový vektor odhadnutý v m-tém iteračním kroku algoritmu a $\rho_{\mathrm{m}} \in (0,1\rangle$ je parametr algoritmu, který určuje rychlost jeho konvergence. Jak jsme ale uvedli výše, gradient kriteriální funkce není definován v místě nespojitosti. Z definičního vztahu kriteriální funkce (2.165) je v bodech, kde je gradient definován

$$\frac{\partial \mathrm{J}(\tilde{\mathbf{w}})}{\tilde{\mathbf{w}}} = \sum_{\mathbf{x}\in\mathcal{Y}} \delta_{\mathbf{x}} \tilde{\mathbf{x}}. \tag{2.167}$$

Po dosazení z (2.167) do (2.166) dostáváme

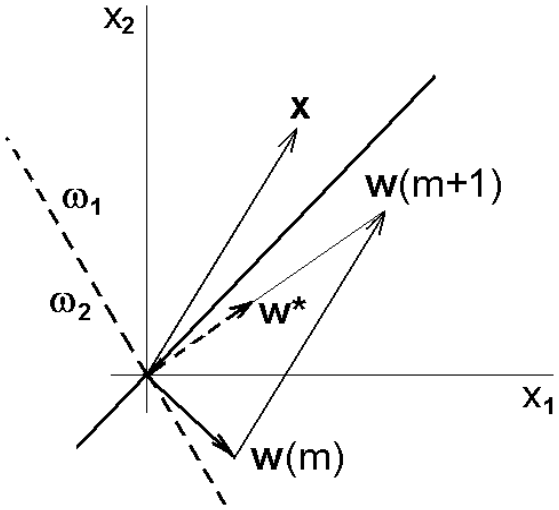
$$\tilde{\mathbf{w}}(m+1) = \tilde{\mathbf{w}}(m) - \rho_m \sum_{x \in \mathcal{Y}} \delta_x \tilde{\mathbf{x}} \tag{2.168}$$

(tato funkce je už definována pro všechny body).

Struktura vlastního algoritmu je poměrně jednoduchá, obsahuje následující kroky:

1. zvolte $\tilde{\mathbf{w}}(0)$ a ρ_0 (hodnoty váhového vektoru se zpravidla nastavují náhodně na hodnoty blízké nule);
2. m = 0;
3. ***repeat***
   $\mathcal{Y} = \emptyset$;
   for k=1 ***to*** K (K je celkový počet obrazů v trénovací množině);
   if $\delta_{xk} \tilde{\mathbf{w}}(m)^T \mathbf{x}_k \geq 0$ ***then*** $\mathcal{Y} = \mathcal{Y} \cup \{\mathbf{x}_k\}$;
   $\tilde{\mathbf{w}}(m+1) = \tilde{\mathbf{w}}(m) - \rho_m \sum_{x \in \mathcal{Y}} \delta_x \tilde{\mathbf{x}}$;
   nastavte ρ_m;
   m = m+1;
   until $\mathcal{Y} = \emptyset$.

Na obr.2.28 je zobrazena geometrická interpretace popsaného algoritmu učení jednovrstvého perceptronu. Předpokládejme, že v m-tém kroku algoritmu je chybně klasifikován pouze jeden obraz **x.** Dále předpokládejme, že parametr $\rho_m = 1$. V tom případě algoritmus koriguje váhový vektor ve směru obrazového vektoru **x**. Protože váhový vektor určuje orientaci hraniční roviny, korekcí váhového vektoru se otočí hraniční roviny tak, že je obrazový vektor klasifikován správně do třídy ω_1. Pro hodnotu $\rho_m = 1$ se uvedená změna hodnot váhového vektoru provedla v jednom iteračním kroku, pro menší hodnotu ρ_m by algoritmus proběhl ve více krocích.



Obr.2.28 *Geometrická interpretace algoritmu učení jednovrstvého perceptronu*

Vzhledem k podstatě použitého optimalizačního algoritmu a kritériu optimality, za optimální řešení je považováno první nastavení váhového vektoru, pro které je splněno kritérium optimality. Jsou-li obě klasifikační třídy lineárně separabilní, existuje pravděpodobně více možných řešení rozdělení obrazového prostoru a bylo by možné snažit se z těchto řešení vybrat opět nejlepší. To ale tento algoritmus neumožňuje.

Variantou uvedeného algoritmu je postup, který reprezentuje třídu algoritmů schématu „zisk – ztráta“. Algoritmus předpokládá, že je na vstup perceptronu postupně cyklicky přiváděno K obrazů trénovací množiny. Cykly (též epochy) zpracování všech prvků trénovací množiny se opakují, dokud nejsou správně zařazeny všechny trénovací prvky. Algoritmus se řídí následujícími vztahy:

$$\begin{aligned}
&1.\ \tilde{\mathbf{w}}(m+1) = \tilde{\mathbf{w}}(m) + \rho \mathbf{x}(m), \quad \text{když } \mathbf{x}(m) \in \omega_1 \text{ a } \tilde{\mathbf{w}}(m).\mathbf{x}(m) \leq 0; \\
&2.\ \tilde{\mathbf{w}}(m+1) = \tilde{\mathbf{w}}(m) - \rho \mathbf{x}(m), \quad \text{když } \mathbf{x}(m) \in \omega_2 \text{ a } \tilde{\mathbf{w}}(m).\mathbf{x}(m) \leq 0; \\
&3.\ \tilde{\mathbf{w}}(m+1) = \tilde{\mathbf{w}}(m), \quad \text{ve všech ostatních případech.}
\end{aligned} \tag{2.169}$$

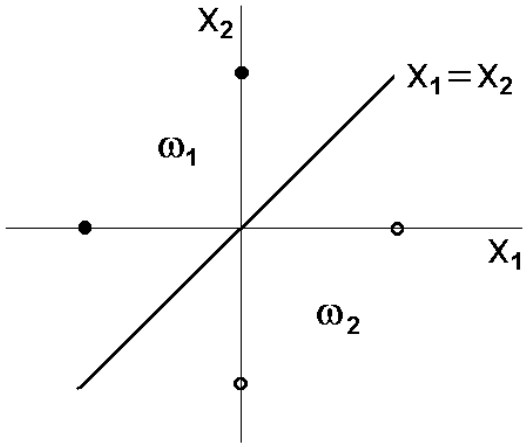
To znamená, že je-li prvek trénovací množiny klasifikován správně, hodnoty váhového vektoru se nemění. Algoritmus je odměněn tím, že není třeba nic korigovat. Je-li naopak tré-

novací obraz zatříděn špatně, k hodnotám váhového vektoru je přičtena, příp. odečtena hodnota úměrná souřadnicím vektoru $\tilde{\mathbf{x}}$. Tedy algoritmus tratí cenu korekce.

Pokud jsou obě klasifikační třídy lineárně separabilní, pak lze dokázat, že učicí algoritmus konverguje a hraniční rovina leží mezi oběma množinami. Když ale třídy nejsou lineárně separabilní, potom poloha hranice osciluje.

Příklad:

Navrhněte lineární rozdělení dvourozměrného obrazového prostoru pomocí jednovrstvého perceptronu. V obrazovém prostoru body $\mathbf{x}_1 = (-1, 0)^T$ a $\mathbf{x}_2 = (0, 1)^T$ patří do klasifikační třídy ω_1 a body $\mathbf{x}_3 = (0, -1)^T$ a $\mathbf{x}_4 = (1, 0)^T$ do třídy ω_2 (obr.2.29).



Obr.2.29 *Zadání a řešení úlohy*

Počáteční hodnoty rozšířeného váhového vektoru $\tilde{\mathbf{w}}(0)$ zvolme $(0, 0, 0)^T$ a hodnotu parametru ρ zvolme jednotkovou, $\rho = 1$. Pro první rozšířený obraz $\tilde{\mathbf{x}}_1$ je

$$\tilde{\mathbf{w}}^T(0).\tilde{\mathbf{x}}_1 = \begin{bmatrix}0 & 0 & 0\end{bmatrix}.\begin{bmatrix}1\\-1\\0\end{bmatrix} = 0\,,$$ tedy podle

pravidla 1 $$\tilde{\mathbf{w}}(1) = \begin{bmatrix}0\\0\\0\end{bmatrix} + \begin{bmatrix}1\\-1\\0\end{bmatrix} = \begin{bmatrix}1\\-1\\0\end{bmatrix}.$$

V druhém kroku pro rozšířený obraz $\tilde{\mathbf{x}}_2$ je

$$\tilde{\mathbf{w}}^T(1).\tilde{\mathbf{x}}_2 = \begin{bmatrix}1 & -1 & 0\end{bmatrix}.\begin{bmatrix}1\\0\\1\end{bmatrix} = 1 > 0\,,$$ tedy podle pravidla 3 $\tilde{\mathbf{w}}(2) = \tilde{\mathbf{w}}(1)$

V třetím kroku pro obraz $\tilde{\mathbf{x}}_3$ je

$$\tilde{\mathbf{w}}^T(2).\tilde{\mathbf{x}}_3 = \begin{bmatrix}1 & -1 & 0\end{bmatrix}.\begin{bmatrix}1\\0\\-1\end{bmatrix} = 1 > 0\,.$$ Proto podle pravidla 2 $$\tilde{\mathbf{w}}(3) = \begin{bmatrix}1\\-1\\0\end{bmatrix} - \begin{bmatrix}1\\0\\-1\end{bmatrix} = \begin{bmatrix}0\\-1\\1\end{bmatrix}.$$

V následujícím čtvrtém kroku pro obraz $\tilde{\mathbf{x}}_4$ je

$$\tilde{\mathbf{w}}^T(3).\tilde{\mathbf{x}}_4 = \begin{bmatrix}0 & -1 & 1\end{bmatrix}.\begin{bmatrix}1\\1\\0\end{bmatrix} = -1 < 0$$ a tedy podle pravidla 3 $\tilde{\mathbf{w}}(4) = \tilde{\mathbf{w}}(3)$.

Nyní přivádíme na vstup obrazy znovu od $\tilde{\mathbf{x}}_1$

$$\tilde{\mathbf{w}}^T(4).\tilde{\mathbf{x}}_1 = \begin{bmatrix}0 & -1 & 1\end{bmatrix}.\begin{bmatrix}1\\-1\\0\end{bmatrix} = 1 > 0$$ a proto dle pravidla 3 $\tilde{\mathbf{w}}(5) = \tilde{\mathbf{w}}(4)$,

dále pro obraz $\tilde{\mathbf{x}}_2$

$$\tilde{\mathbf{w}}^{\mathrm{T}}(5).\tilde{\mathbf{x}}_2 = \begin{bmatrix}0 & -1 & 1\end{bmatrix}\begin{bmatrix}1\\0\\1\end{bmatrix} = 1 > 0$$; opět podle pravidla 3 $\tilde{\mathbf{w}}(6) = \tilde{\mathbf{w}}(5)$

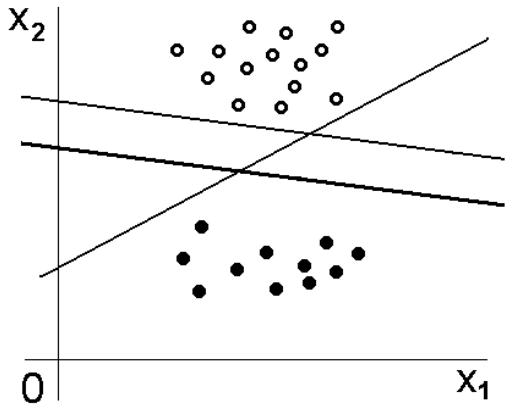
a konečně pro obraz $\tilde{\mathbf{x}}_3$

$$\tilde{\mathbf{w}}^{\mathrm{T}}(6).\tilde{\mathbf{x}}_3 = \begin{bmatrix}0 & -1 & 1\end{bmatrix}\begin{bmatrix}1\\0\\-1\end{bmatrix} = -1 < 0$$ a tedy znovu podle pravidla 3 je $\tilde{\mathbf{w}}(7) = \tilde{\mathbf{w}}(6)$.

Ve čtyřech (počet obrazů v trénovací množině) následných krocích nebylo třeba jakékoliv korekce váhového vektoru, všechny obrazy učební množiny byly klasifikovány správně, algoritmus proto dál nepokračuje, určený váhový vektor je $\tilde{\mathbf{w}}^{\mathrm{T}} = \begin{bmatrix}0 & -1 & 1\end{bmatrix}$. To znamená, že hraniční přímka je určena rovnicí $-x_1 + x_2 = 0$, resp. $x_1 = x_2$, což reprezentuje přímku procházející počátkem s jednotkovou směrnicí (obr.2.29). □□□

2.4.5 Algoritmus podpůrných vektorů

Algoritmus podpůrných vektorů (*support vector machine* – SVM) představuje velkou třídu metod s použitím v rozličných klasifikačních úlohách, v této kapitole se věnujme jeho použití především pro hledání lineární separace klasifikačních tříd. Začneme dichotomickým problémem s lineárně separabilními klasifikačními třídami, který se posléze pokusíme zobecnit.

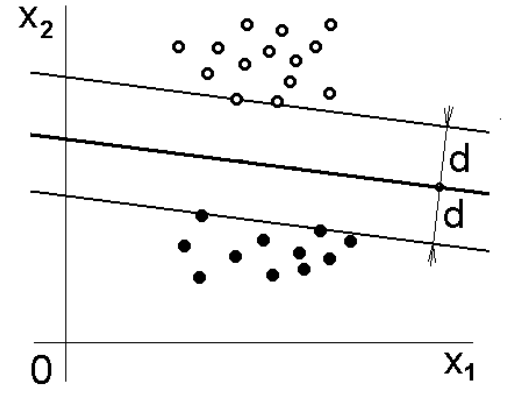


Obr.2.30 *Nejednoznačnost možné lineární separace dvou klasifikačních tříd*

Separabliní třídy

Předpokládejme, že trénovací množina obsahuje K obrazových vektorů, které patří do dvou lineárně separabilních klasifikačních tříd ω_1 a ω_2, které lze oddělit hranicí $b(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}.\mathbf{x} + w_0 = 0$ (viz též vztah (2.135)).

V předcházející kapitole jsme zmínili, že poloha takové hraniční plochy není obecně jednoznačná a použijeme-li algoritmus učení jednovrstvého perceptronu, pak její poloha nemusí být optimální. Na obr.2.30 je znázorněno několik možných řešení. Všechny třídy beze zbytku splňují požadavek na dokonalé oddělení obou množin, lze však i intuitivně odhadnout, že některé z uvedených řešení je vhodnější, jiná méně. Zřejmě nejlepší volba z dělení obrazového prostoru uvedeného na obr.2.30 představuje silná čára, která prochází v dostatečně velké vzdálenosti od vektorů obou množin. Takové řešení je nepochybně nejrobustnější pro klasifikaci nových vektorů, které nejsou součástí trénovací množiny. Poměr vzdáleností každé dílčí části trénovací množiny od hranice bude určitě záležet na očekávání, jaký prostor budou vyplňovat nově klasifikované obrazy. V případě, že je toto



Obr.2.31 *Příklad dělení obrazového prostoru pomocí algoritmu podpůrných vektorů v případě separabilních tříd*

očekávání stejné pro obě množiny, resp. nemáme-li z tohoto pohledu žádnou apriorní informaci, předpokládáme, že by tento poměr měl být roven jedné.

Na základě této úvahy lze definovat kritérium, jež umožní nalézt optimální polohu dělicí hraniční roviny. Nechť je to taková nadrovina, která vytváří největší šířku tolerančního pásma mezi hranicí a oběma částmi trénovací množiny.

Připomeňme nyní, že orientace dělicí roviny je dána souřadnicemi vektoru **w** a její poloha hodnotou prahu w_0. Dále připomeňme obr.2.15 a vztah (2.138), podle kterého je vzdálenost bodu od hraniční roviny

$$d = \frac{|b(\mathbf{x})|}{\|\mathbf{w}\|}.$$

Nyní předpokládejme, že hodnota funkce $b(\mathbf{x})$ v nejbližším bodě množiny ω_1 je rovna +1 a v nejbližším bodě množiny ω_2 je -1. V tom případě a za předpokladu stejných vah obou dílčích množin je šířka celého tolerančního pásma

$$\frac{1}{\|\mathbf{w}\|} + \frac{1}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|} \tag{2.170}$$

a pro všechny vektory trénovací množiny platí

$$\begin{aligned} \mathbf{w}^T\mathbf{x} + w_0 &\geq 1 \quad \text{pro } \forall \mathbf{x} \in \omega_1 ; \\ \mathbf{w}^T\mathbf{x} + w_0 &\leq -1 \quad \text{pro } \forall \mathbf{x} \in \omega_2 . \end{aligned} \tag{2.171}$$

Jestliže definujeme pomocnou cílovou proměnou δ_x, v tomto případě definovanou tak, že $\delta_x = 1$ pro vektory z třídy ω_1 a $\delta_x = -1$ pro vektory z třídy ω_2, pak lze definovat kriteriální funkci

$$J(\mathbf{w}, w_0) = \frac{\|\mathbf{w}\|^2}{2} \tag{2.172}$$

(kvadrát normy je použit, abychom se vyhnuli nespojitosti derivace $\|\mathbf{w}\|$ v bodě nula a koeficient 1/2 zjednoduší výrazy získané po derivování), kterou budeme minimalizovat. Ze vztahu (2.170) je nepochybně vidět, že minimalizací normy maximalizujeme šířku tolerančního pásma. Minimální hodnotu kriteriální funkce musíme hledat za podmínky, že

$$\delta_x.(\mathbf{w}^T\mathbf{x} + w_0) \geq 1, \quad \text{pro } \forall \mathbf{x} \text{ z trénovací množiny.} \tag{2.173}$$

To je opět nelineární podmíněná optimalizační úloha, která vede na využití metody Langrangova součinitele pro hledání vázaného extrému. Jejím použitím získáme soustavu rovnic (tzv. ***Karushovy – Kuhnovy – Tuckerovy podmínky***)

$$\frac{\partial}{\partial \mathbf{w}} \mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\lambda}) = \mathbf{0}; \tag{2.174}$$

$$\frac{\partial}{\partial w_0} \mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\lambda}) = \mathbf{0}; \tag{2.175}$$

$$\lambda_k[\delta_{xk}(\mathbf{w}^T\mathbf{x}_k + w_0) - 1] = 0, \; k = 1,2,...,K; \tag{2.176}$$

kde $\boldsymbol{\lambda}$ je vektor Langrangových součinitelů $\lambda_k \geq 0$, k = 1, 2, …, K a $\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\lambda})$ je Langrangova funkce definovaná vztahem

$$\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\lambda}) = \frac{\mathbf{w}^T\mathbf{w}}{2} - \sum_{k=1}^{K} \lambda_k [\delta_{xk}(\mathbf{w}^T\mathbf{x}_k + w_0) - 1] \tag{2.177}$$

Řešením uvedené soustavy rovnice dostáváme

$$\mathbf{w} = \sum_{k=1}^{K} \lambda_k \delta_{xk} \mathbf{x}_k \tag{2.178}$$

a

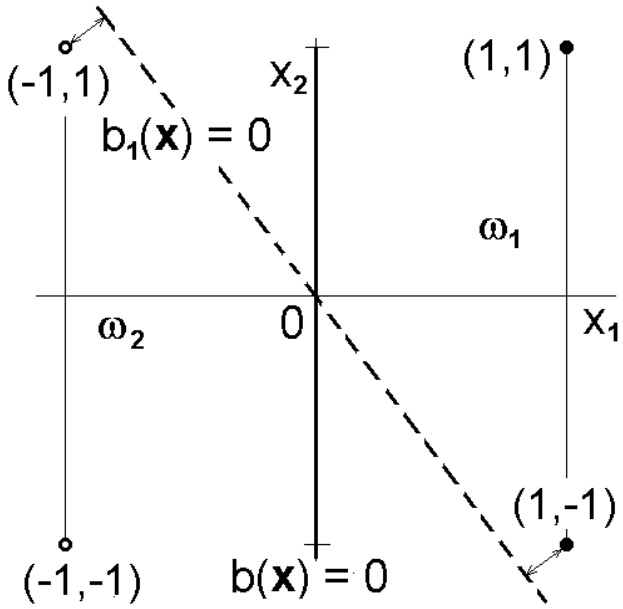
$$\sum_{k=1}^{K} \lambda_k \delta_{xk} = 0. \tag{2.179}$$

Řešením této soustavy získáváme optimální hodnoty w_0 a $\mathbf{w}$.

Vektory, pro které leží právě ve vzdálenosti $1/\|\mathbf{w}\|$, tj. vektory, které právě definují šířku tolerančního pásu a pro které platí $\mathbf{w}^T\mathbf{x} + w_0 = \pm 1$, se nazývají podpůrnými vektory a klasifikátor používající hraniční plochu optimálně určenou souřadnicemi podpůrného vektoru se označuje jako ***algoritmus podpůrných vektorů*** (viz též obr.2.27).

Příklad:

Navrhněte lineární rozdělení dvourozměrného obrazového prostoru pomocí algoritmu podpůrných vektorů, za předpokladu, že obrazové vektory $\mathbf{x}_1 = (1, 1)^T$ a $\mathbf{x}_2 = (1, -1)$ patří do třídy ω_1 a vektory $\mathbf{x}_3 = (-1, 1)^T$ a $\mathbf{x}_4 = (-1, -1)$ do třídy ω_2 (viz obr.2.32).



Obr.2.32 *Zadání a řešení příkladu*

Jednoduchá geometrie zadání patrná z obr.2.32 a znalost principu algoritmu podpůrných vektorů vede k intuitivnímu řešení úlohy, definujícímu hraniční funkci $b(\mathbf{x}) = w_1x_1 + w_2x_2 + w_0 = 0$ pro $w_1 = 1$ a $w_2 = w_0 = 0$. V tom případě jsou všechny čtyři obrazové vektory podpůrnými vektory a toleranční pás oddělující hraniční přímku od všech podpůrných vektorů je široký právě 1. Všechna ostatní možná řešení vedou k menší šířce tolerančního pásma.

Pokusme se nyní o matematickou formulaci a řešení zadané úlohy. Lagrangova funkce podle vztahu (2.177) v tomto konkrétním případě nabývá tvaru

$$\begin{aligned}\mathcal{L}(w_1, w_2, w_0, \boldsymbol{\lambda}) = \frac{w_1^2 + w_2^2}{2} &- \lambda_1(w_1 + w_2 + w_0 - 1) - \lambda_2(w_1 - w_2 + w_0 - 1) - \\ &- \lambda_3(w_1 - w_2 - w_0 - 1) - \lambda_2(w_1 + w_2 - w_0 - 1)\end{aligned} \tag{2.180}$$

Její derivací podle jednotlivých souřadnic vektoru $\mathbf{w}$ dostáváme (viz vztahy (2.174) a (2.175))

$$\frac{\partial \mathcal{L}}{\partial w_1} = 0 \Rightarrow w_1 = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4;$$
$$\frac{\partial \mathcal{L}}{\partial w_2} = 0 \Rightarrow w_2 = \lambda_1 - \lambda_2 - \lambda_3 + \lambda_4; \qquad (2.181)$$
$$\frac{\partial \mathcal{L}}{\partial w_0} = 0 \Rightarrow \lambda_1 + \lambda_2 - \lambda_3 - \lambda_4 = 0$$

a ze vztahu (2.176) plyne

$$\begin{aligned} \lambda_1(w_1 + w_2 + w_0 - 1) &= 0; \\ \lambda_2(w_1 - w_2 + w_0 - 1) &= 0; \\ \lambda_3(w_1 - w_2 - w_0 - 1) &= 0; \\ \lambda_4(w_1 + w_2 - w_0 - 1) &= 0. \end{aligned} \qquad (2.182)$$

Jednotlivé rovnice se rovnají nule, buď když $\lambda_i = 0$, nebo když jsou rovny nule výrazy v závorce, nebo obojí. Protože v této chvíli nevíme, jaké hodnoty nabudou Langrangovy součinitele λ_i, budeme se zabývat případem, kdy se budou rovnat nule výrazy v závorkách. V tom případě se soustava rovnic (2.182) transformuje do lineárního tvaru

$$\begin{aligned} w_1 + w_2 + w_0 &= 1; \\ w_1 - w_2 + w_0 &= 1; \\ w_1 - w_2 - w_0 &= 1; \\ w_1 + w_2 - w_0 &= 1, \end{aligned} \qquad (2.183)$$

která má řešení $w_1 = 1$ a $w_2 = w_0 = 0$, což odpovídá intuitivnímu řešení uvedenému na začátku řešení tohoto příkladu. Po dosazení za w_1, w_2 a w_0 do (2.181) dostáváme lineární soustavu tří rovnic pro čtyři neznámé λ_1, λ_2, λ_3 a λ_4

$$\begin{aligned} \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 &= 1; \\ \lambda_1 - \lambda_2 - \lambda_3 + \lambda_4 &= 0; \\ \lambda_1 + \lambda_2 - \lambda_3 - \lambda_4 &= 0. \end{aligned} \qquad (2.184)$$

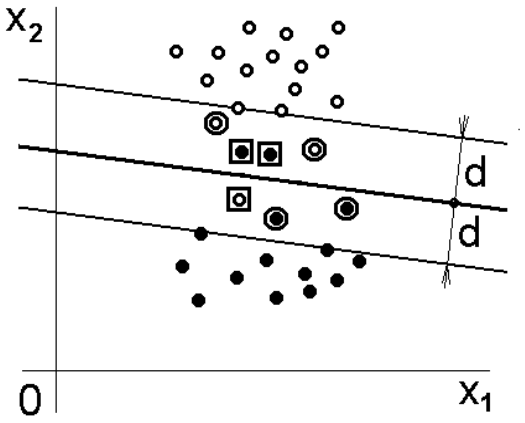
Tato soustava má nekonečně mnoho řešení, která lze parametricky popsat

$$\lambda_4 = t; \quad \lambda_3 = \frac{1-2t}{2}; \lambda_2 = \lambda_4 = t \text{ a } \lambda_1 = \lambda_3 = \frac{1-2t}{2}. \qquad (2.185)$$

Ovšem všechna řešení odpovídají optimálnímu nastavení souřadnic vektoru **w**, např. pro $t = 0{,}25$ jsou $\lambda_1 = 0{,}25$, $\lambda_2 = 0{,}25$, $\lambda_3 = 0{,}25$ a stejně $\lambda_4 = 0{,}25$.

Neseparabliní třídy

Pokud jsou obě klasifikační třídy neseparabilní, pak výše uvedené předpoklady neplatí. Předpokládejme situaci zobrazenou na obr.2.33. Kromě obrazových vektorů, které odpovídají představě rozebírané v dřívější kapitole, jsou v trénovací množině i vektory, které jsou už v tolerančním pásu, ale jsou správně klasifikovány (zakroužkované body), ovšem existují i vektory, které budou špatně klasifikovány, protože leží v opačné polorovině, než by odpovídalo správné klasifikaci (začtverečkované body).



Obr.2.33 *Příklad dělení obrazového prostoru pomocí algoritmu podpůrných vektorů v případě lineárně neseparabilních tříd*

Pro vektory ležící ve správném tolerančním pásu je

$$0 \le \delta_x.(\mathbf{w}^T\mathbf{x} + w_0) < 1 \tag{2.186}$$

pro obrazové vektory nacházející se na nesprávné straně dělicí hranice, tj. chybně klasifikované obrazy je

$$\delta_x.(\mathbf{w}^T\mathbf{x} + w_0) < 0. \tag{2.187}$$

Abychom mohli popsat všechny tři případy pomocí jediné formulace (to je třeba pro definici kriteriální funkce), zavedeme další proměnné, $\xi_k \ge 0$. Nazýváme je relaxační proměnné. Pomocí těchto proměnných můžeme psát

$$\delta_x.(\mathbf{w}^T\mathbf{x} + w_0) \ge 1 - \xi_k. \tag{2.188}$$

Obrazy první kategorie, tj. dostatečně jistě správně klasifikované odpovídají hodnotám $\xi_k = 0$, obrazy ležící na správné straně tolerančního pásu je $0 < \xi_k \le 1$ a konečně pro chybně klasifikované obrazové vektory je $\xi_k > 1$. Optimalizační úloha je teď nepoměrně komplikovanější, i když vychází ze stejných principů. Cílem optimalizace je teď, stejně jako v předchozím případě, vytvořit co nejširší toleranční pás, ale současně minimalizovat počet obrazů, pro něž je $\xi_k > 0$. Znamená to, že se v tomto případě snažíme o sloučení dvou optimalizačních úloh. Matematicky vyjádřeno, snažíme se minimalizovat kriteriální funkci

$$J(\mathbf{w}, w_0, \boldsymbol{\xi}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{k=1}^{K} I(\xi_k), \tag{2.189}$$

kde $\boldsymbol{\xi}$ je vektor relaxačních proměnných ξ_k,

$$I(\xi_k) = \begin{cases} 1, & \text{pro } \xi_k > 0; \\ 0, & \text{pro } \xi_k = 0. \end{cases} \tag{2.190}$$

a konstanta C vyjadřuje poměr vlivu obou členů kriteriální funkce podle (2.189). Bohužel úlohu značně komplikuje dle definice nespojitá funkce $I(\xi_k)$. Abychom byli schopni kriteriální funkci derivovat, používá se náhradní vyjádření kriteriální funkce ve tvaru

$$J(\mathbf{w}, w_0, \boldsymbol{\xi}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{k=1}^{K} \xi_k, \tag{2.191}$$

kterou minimalizujeme za podmínky, že

$$\delta_{xk}(\mathbf{w}^T\mathbf{x}_k + w_0) \ge 1 - \xi_k,\ k = 1,2,...,K. \tag{2.192}$$

Definice problému opět vede k řešení pomocí metody Lagrangova součinitele, tentokrát s Lagrangovou funkcí

$$\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{\mathbf{w}^T\mathbf{w}}{2} + C\sum_{k=1}^{K}\xi_k - \sum_{k=1}^{K}\mu_k\xi_k - \sum_{k=1}^{K}\lambda_k[\delta_{xk}(\mathbf{w}^T\mathbf{x}_k + w_0) - 1 + \xi_k]. \tag{2.193}$$

Tomu odpovídající Karushovy-Kuhnovy-Tuckerovy podmínky jsou

$$\frac{\partial\mathcal{L}}{\partial\mathbf{w}} = 0 \text{ nebo } \mathbf{w} = \sum_{k=1}^{K}\lambda_k\delta_{xk}\mathbf{x}_k; \tag{2.194}$$

$$\frac{\partial \mathcal{L}}{\partial w_0} = 0 \text{ nebo } \mathbf{w} = \sum_{k=1}^{K} \lambda_k \delta_{xk}; \tag{2.195}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_k} = 0 \text{ nebo } C - \mu_k - \lambda_k = 0, k = 1,2,....,K; \tag{2.196}$$

$$\lambda_k[\delta_{xk}(\mathbf{w}^T\mathbf{x}_k + w_0) - 1 + \xi_k] = 0, k = 1,2,...,K. \tag{2.197}$$

$$\mu_k \xi_k = 0, k = 1,2,...,K; \tag{2.198}$$

$$\mu_k \geq 0, \lambda_k \geq 0, k = 1,2,...,K. \tag{2.199}$$

2.5 Souvislosti jednotlivých principů klasifikace

V kap.2.4.1 a na obr.2.18 jsme si již uvedli, jak hraniční plochy souvisejí s diskriminačními funkcemi - že je tvoří průmět průsečíku diskriminační funkce do obrazového prostoru.

Vzájemné souvislosti mezi jednotlivými principy klasifikace si objasníme na jednoduchých příkladech. Začněme se srovnáním klasifikace podle minimální vzdálenosti a klasifikací podle diskriminačních funkcí.

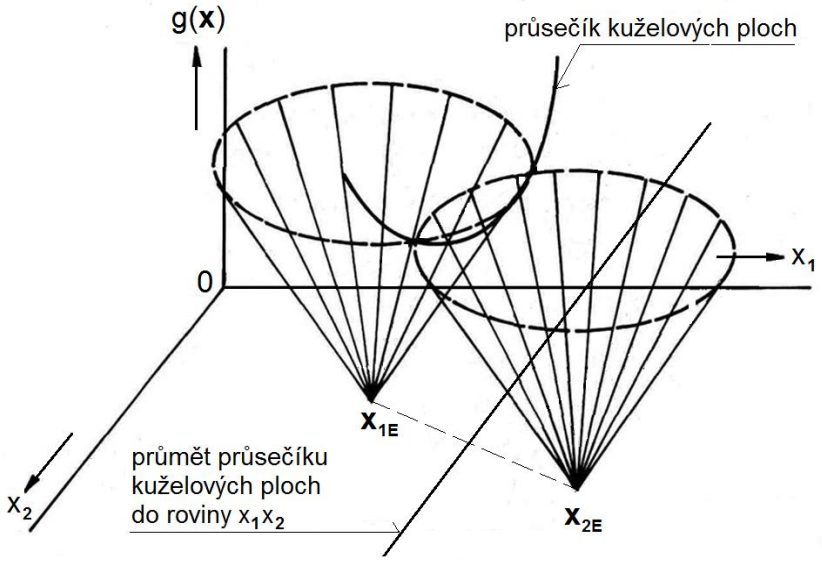
Uvažme příklad dvou tříd reprezentovaných etalony $\mathbf{x}_{1E} = (x_{11E}, x_{12E})$ a $\mathbf{x}_{2E} = (x_{21E}, x_{22E})$ v dvoupříznakovém euklidovském prostoru. Výpočet vzdálenosti mezi obrazem $\mathbf{x} = (x_1, x_2)$ a libovolným z obou etalonů je v tomto prostoru definován vztahem

$$v(\mathbf{x}_{sE}, \mathbf{x}) = \|\mathbf{x}_{sE} - \mathbf{x}\| = \min_{\forall s} \|\mathbf{x}_{sE} - \mathbf{x}\| = \sqrt{(x_{s1E} - x_1)^2 + (x_{s2E} - x_2)^2}; \quad s = 1, 2. \tag{2.200}$$

Podle definice rozhodovacího pravidla klasifikátoru podle minimální vzdálenosti hledáme menší z obou vzdáleností, tj. $\min_{s=1,2} v(\mathbf{x}_{sE}, \mathbf{x})$. Protože nám nejde o stanovení konkrétní vzdálenosti, ale o nalezení minima a rovněž díky tomu, že vzdálenost mezi dvěma body prostoru je vždy kladná, můžeme psát, že hledáme $\min_{s=1,2} v^2(\mathbf{x}_{sE}, \mathbf{x})$. To znamená, že

$$\begin{aligned} \min_{\forall s} v(\mathbf{x}_{sE}, \mathbf{x}) \sim \min_{\forall s} v^2(\mathbf{x}_{sE}, \mathbf{x}) = \|\mathbf{x}_{sE} - \mathbf{x}\| = \min_{\forall s} \left[(x_{s1E} - x_1)^2 + (x_{s2E} - x_2)^2\right] = \\ = \min_{\forall s} \left\{ x_1^2 + x_2^2 - 2\left[x_{s1E}x_1 + x_{s2E}x_2 - \frac{(x_{s1E}^2 + x_{s2E}^2)}{2} \right] \right\}. \end{aligned} \tag{2.201}$$

Pro každý etalon představuje výraz ve složených závorkách kuželovou plochu s vrcholem v etalonu (pokud je vektor $\mathbf{x}$ totožný s etalonem, je výraz ve složených závorkách roven nule) a rozšiřující se do kladných hodnot funkce $g(\mathbf{x})$ (pro souřadnice vektoru $\mathbf{x} = (x_{s1E} \pm c_1, x_{s2E} \pm c_2)$ je hodnota výrazu ve složených závorkách rovna $c_1^2 + c_2^2$) (obr.2.34). Jak je z obrázku patrné, tato orientace kuželové plochy bohužel nesplňuje podmínku pro diskriminační funkci, ovšem pro daný obraz $\mathbf{x}$ dvojčlen $x_1^2 + x_2^2$ ve složených závorkách ve výrazu (2.201) nezávisí na klasifikační třídě, můžeme jej proto považovat za aditivní konstantu, která se nepodílí na rozhodování. Poněvadž je tento člen vždy kladný, můžeme určit minimum celého výrazu právě tehdy, když najdeme ve vztahu (2.201) maximum výrazu v hranatých závorkách. Tím se orientace kuželové plochy mění a v souladu s principem klasifikace podle diskriminačních funkcí lze tento výraz považovat za definiční vztah diskriminační funkce s-té



Obr.2.34 *Klasifikace podle minimální vzdálenosti*

třídy $g_s(\mathbf{x})$. Kuželové plochy se v obou případech protínají v parabole a její průmět do obrazové roviny je přímka (obr.2.34), definovaná vztahem

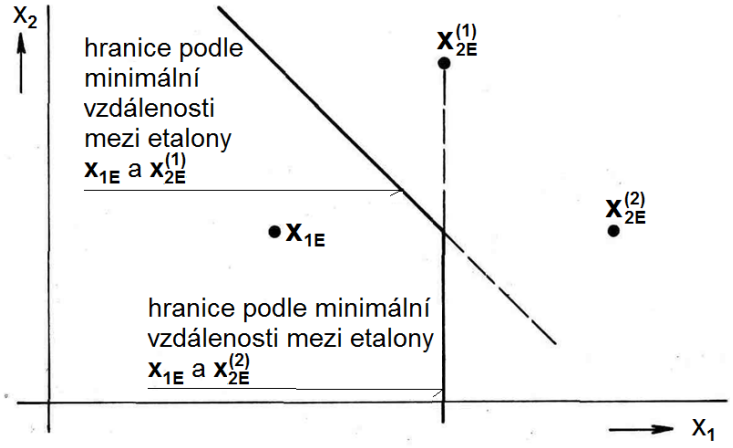
$$x_1(x_{1\,1E} - x_{2\,1E}) + x_2(x_{1\,2E} - x_{2\,2E}) - \frac{x_{1\,1E}^2 + x_{1\,2E}^2 - x_{2\,1E}^2 - x_{2\,2E}^2}{2} = 0\,. \qquad (2.202)$$

Tato hraniční přímka mezi klasifikačními třídami je vždy kolmá na spojnici obou etalonů a tuto spojnici půlí. Z uvedeného plyne, že klasifikátor pracující na základě minimální vzdálenosti je ekvivalentní lineárnímu klasifikátoru s diskriminačními funkcemi. Dále je tento příklad ukázkou toho, že i nelineární diskriminační funkce může vyústit v lineární separaci klasifikačních tříd.

Jinou možností, jak zkonstruovat diskriminační funkci na základě principu stanovení vzdálenosti, resp. podobnosti mezi klasifikovaným obrazem a etalony klasifikačních tříd je použití metriky podobnosti. Dle závislosti mezi vzdálenostní a podobnostní metrikou se mění tvar kuželové plochy, nicméně její vrchol leží vždy nad etalony klasifikačních tříd, kuželová plocha se rozšiřuje směrem k obrazovému prostoru, mění se sice tvar průsečíků kuželových ploch odpovídajících jednotlivým etalonům, ale jejich průmět do obrazové roviny zůstává lineární – za předpokladu, že metriky pro jednotlivé etalony nejsou různě váhované.

Uvažme nyní případ, kdy je třída ω_1 reprezentována etalonem $\mathbf{x}_{1E}$ a třída ω_2 dvěma etalony $\mathbf{x}_{2E}^{(1)}$ a $\mathbf{x}_{2E}^{(2)}$ a obrazový vektor $\mathbf{x}$ klasifikujeme opět pomocí kritéria nejmenší vzdálenosti.

Podle výše uvedeného vztahu pro hranici oddělující obrazy náležející jednotlivým etalonům podle kritéria minimální vzdálenosti jsou hranice mezi třemi etalony znázorněny na obr.2.35. Protože třídu ω_2 představují dva etalony, je hranice mezi oběma třídami lomená přímka půlící



Obr.2.35 *Klasifikace podle minimální vzdálenosti s víceetalonovými klasifikačními třídami*

vzdálenosti mezi etalony $\mathbf{x}_{1E}$ a $\mathbf{x}_{2E}^{(1)}$ a $\mathbf{x}_{1E}$ a $\mathbf{x}_{2E}^{(2)}$.

Klasifikace podle minimální vzdálenosti s třídami reprezentovanými více etalony je ekvivalentní klasifikaci podle diskriminační funkce s po částech lineární hraniční funkcí.

Pokusme se tedy shrnout vzájemné vztahy mezi jednotlivými principy klasifikace.

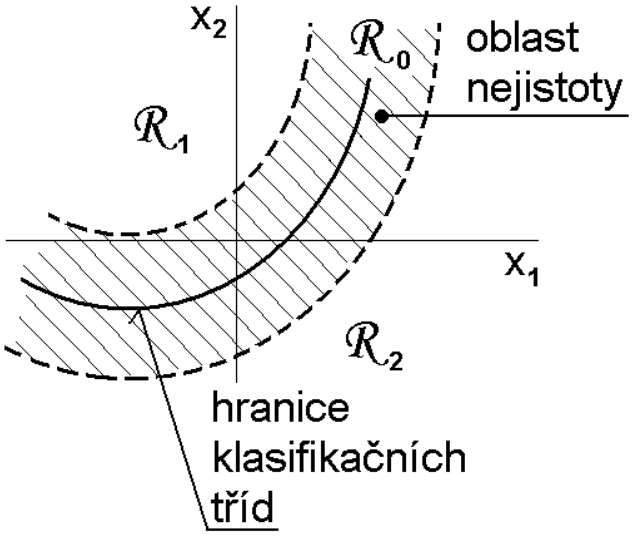
- Hranice mezi klasifikačními třídami jsou dány průmětem diskriminačních funkcí do obrazové roviny.
- Klasifikace podle minimální vzdálenosti definuje hranici, která je kolmá na spojnici etalonů klasifikačních tříd a půlí ji.
- Princip klasifikace dle minimální vzdálenosti vede buď přímo, nebo prostřednictvím využití metrik podobnosti k definici diskriminačních funkcí a ty dle prvního ze zde uvedených pravidel k určení hranic mezi klasifikačními třídami.

2.6 Sekvenční příznaková klasifikace

2.6.1 Základní úvahy

V dosud popisovaných metodách pro vymezení klasifikačních tříd jsme předpokládali, že všechny klasifikované obrazy mají konstantní počet příznaků, přičemž problém stanovení vhodného počtu příznaků jsme zatím neřešili. Je zřejmé, že nepřiměřený počet příznaků může při klasifikaci způsobit potíže. Malý počet příznaků (malé množství informace) může být příčinou nesprávné klasifikace, naopak zjišťování velkého množství dat může být z hlediska cílů klasifikace nepřiměřeně pracné, případně nákladné, zpravidla obojí. Jednou z možností jak nalézt kompromis mezi chybou klasifikace a cenou určení příznaků je sekvenční klasifikace, která spočívá v klasifikaci obrazů popisovaných stále rostoucím množstvím příznaků, přičemž okamžik ukončení klasifikace a tím celkovým počet příznaků stanoví klasifikátor sám na základě předem stanoveného kritéria posuzujícího kvalitu rozhodnutí. Algoritmus řízení sekvenční klasifikace může být jednoznačně určen předem, např. rozhodovacím (klasifikačním) stromem, nebo závislý na vlastnostech výskytu jednotlivých právě zpracovávaných obrazů. Metodám používajícím klasifikační stromy a způsobu jejich návrhu je věnována publikace [4], zabývejme se zde proto především základním principům tohoto způsobu klasifikace a druhému, alternativně uvedenému přístupu.

Předpokládejme, že n-rozměrný obrazový prostor X^n je hraničními plochami rozdělen na R disjunktních oblastí $\mathcal{R}_r$, r = 1, 2, …, R, které reprezentují představu klasifikátoru o klasifikačních třídách. Proto je obraz **x**, který se nachází v oblasti $\mathcal{R}_r$ obrazového prostoru, zařazen do třídy ω_r. Jestliže se jedná o případ neseparabilních klasifikačních tříd, může dojít k chybnému zatřídění obrazu. Pravděpodobnost chybného zatřídění je zřejmě tím větší, čím menší je vzdálenost obrazu od hranice. Máme-li zadáno kritérium ukončení klasifikačního procesu např. pomocí maximální přípustné pravděpodobnosti chybného rozhodnutí, lze si



Obr.2.36 *Princip sekvenční klasifikace*

toto kritérium znázornit graficky podle obr.2.36. Jednopříznakový obrazový prostor je hranicí rozdělen na dvě oblasti $\mathcal{R}_{11}$ a $\mathcal{R}_{21}$, které reprezentují klasifikační třídy. Okolo rozdělující hraniční plochy je oblast nejistoty $\mathcal{R}_{01}$, ve které je pravděpodobnost chyby větší než předepsaná. Nachází-li se obraz **x** v oblasti $\mathcal{R}_{01}$, je potřeba v klasifikačním procesu pokračovat přidáním a zpracováním další informace (dalšího příznaku), je-li obraz mimo tuto oblast lze klasifikaci ukončit.

Každý příznak v obraze nese určité množství informace o klasifikovaném objektu a toto množství je obecně pro jednotlivé příznaky různé. Intuitivně lze usoudit, že rozhodovací proces bude možné ukončit dříve, pokud bude obraz vyjádřen nejdříve příznaky nesoucími největší množství informace. Rychlost sekvenční klasifikace je tedy závislá na pořadí, v jakém jsou jednotlivé příznaky do obrazu klasifikovaného objektu přidávány.

Zabývejme se nyní kritérii pro řízení sekvenčního klasifikátoru, problematika výběru a uspořádání příznaků bude rozebrána později v kap.3.

2.6.2 Waldovo kritérium

Předpokládejme dichotomický klasifikátor a dále, že každý klasifikovaný obraz **x** je popsán množinou příznaků $\{x_1, x_2, \ldots\}$. Nechť $p(x_1, x_2, \ldots, x_i \mid \omega_1)$ a $p(x_1, x_2, \ldots, x_i \mid \omega_2)$ jsou i-rozměrné hustoty pravděpodobnosti výskytu obrazu $\mathbf{x} = (x_1, x_2, \ldots, x_i)$ vytvořeného v i-tém klasifikačním kroku ve třídách ω_1 a ω_2. Konečně, nechť A a B jsou konstantní parametry, pro které platí $0 < B < 1 < A < \infty$. Jestliže v i-tém klasifikačním kroku platí pro věrohodnostní poměr Λ_i, definovaný vztahem

$$\Lambda_i = \frac{p(x_1, x_2, ..., x_i|\omega_1)}{p(x_1, x_2, ..., x_i|\omega_2)} \tag{2.203}$$

že $\Lambda_i \leq B$, pak $d_w(\mathbf{x}) = \omega_2$, je-li $\Lambda_i \geq A$, pak $d_w(\mathbf{x}) = \omega_1$. Konečně, když $\Lambda_i \in (B,A)$, pak se přibere další příznak x_{i+1} a klasifikační proces se zopakuje.

Jak vyplývá z uvedeného rozhodovacího pravidla, závisí počet kroků rozhodovacího algoritmu, tj. maximální počet příznaků v obrazu x, na volbě hodnot mezních parametrů A a B a na hustotách pravděpodobnosti výskytu obrazů v obou klasifikačních třídách.

Pokud jsou dány pravděpodobnosti chybného zařazení

$$\alpha = \int_{\mathcal{R}_2} p(x|\omega_1)dx \quad \text{a} \quad \beta = \int_{\mathcal{R}_1} p(x|\omega_2)dx\,, \tag{2.204}$$

můžeme empiricky stanovit hodnoty mezí A a B např. podle vztahů

$$A \cong \frac{1-\alpha}{\beta} \quad \text{a} \quad B \cong \frac{\alpha}{1-\beta}. \tag{2.205}$$

Jsou-li příznaky $x_1, x_2, \ldots$, ze kterých jsou vytvořeny obrazy **x**, statisticky nezávislé, lze dokázat, že k přijetí rozhodnutí podle Waldova kritéria je potřeba konečný počet kroků, tj. konečný počet příznaků.

Z hlediska počtu kroků má Waldovo kritérium ve srovnání s jinými rozhodovacími pravidly optimální vlastnosti vyjádřené větami:

a) pro libovolné kritérium s pevným počtem n příznaků a s pravděpodobnostmi α a β chybných rozhodnutí platí pro n, že je větší nebo rovno střední hodnotě počtu kroků podle Waldova kritéria;
b) pro libovolné sekvenční kritérium je k rozhodnutí potřeba průměrný počet kroků větší než je průměrný počet kroků podle Waldova kritéria.

2.6.3 Reedovo kritérium

Waldovo kritérium je vázáno na výpočet věrohodnostního poměru pro dvě klasifikační třídy. Proto se pro počet klasifikačních tříd R větší než dvě využívá kritéria založeného na tzv. zobecněném věrohodnostním poměru definovaném pro každou klasifikační třídu ω_r vztahem

$$\Lambda_i(\mathbf{x}|\omega_r) = \frac{p(x_1, x_2, ..., x_i|\omega_r)}{\left(\prod_{s=1}^{R} p(x_1, x_2, ..., x_i|\omega_s)\right)^{1/R}}, \quad r = 1,2,...,R, \tag{2.206}$$

kde i, podobně jako ve vztahu (2.203), udává pořadí klasifikovaného kroku, tj. počet použitých příznaků.

Takto vypočítaný poměr $\Lambda_i(\mathbf{x} \mid \omega_r)$ se srovná s mezní hodnotou r-té třídy $A(\omega_r)$ stanovenou jako

$$A(\omega_r) = \frac{1 - P_{rr}}{\left(\prod_{s=1}^{R} (1 - P_{rs})\right)^{1/R}}, \quad r = 1,2,...,R, \tag{2.207}$$

kde P_{rs} je pravděpodobnost, že obraz $\mathbf{x}$ ze třídy ω_s je zařazen do třídy ω_r. Pokud pro třídu ω_p platí, že

$$\Lambda_i(\mathbf{x}|\omega_p) = A(\omega_p), \quad p = 1,2,...,R, \tag{2.208}$$

pak předpokládáme, že obraz $\mathbf{x}$ nepatří do třídy ω_p – třídu ω_p tedy můžeme vyloučit z množiny nadále uvažovaných klasifikačních tříd, tj. tříd do nichž lze obraz $\mathbf{x}$ zařadit. Po vyloučení všech tříd ω_p, pro které platí vztah (2.208), se spočítají nové hodnoty zobecněných věrohodnostních poměrů $\Lambda_i(\mathbf{x}|\omega_r)$ a mezních hodnot $A(\omega_r)$, nové hodnoty se srovnají a pokud pro některou třídu opět platí vztah (2.208), tato třída se vyloučí z množiny možných klasifikačních tříd, atd., pokud nezůstane poslední třída, do které je pak klasifikovaný obraz zařazen. Není-li možné vyloučit žádnou klasifikační třídu, zvýší se počet příznaků na i + 1 a klasifikační proces pokrčuje pro všechny možné klasifikační třídy.

Pro R = 2 je Reedovo kritérium ekvivalentní kritériu Waldovu a má tytéž optimální vlastnosti. Pro R > 2 nebyla optimalita Reedova kritéria ani prokázána, ani vyvrácena.

2.6.4 Modifikované Waldovo kritérium

Přes optimální vlastnosti Waldova kritéria může nastat situace, že

- počet kroků potřebných k přijetí rozhodnutí podle tohoto kritéria může být pro některé obrazy příliš velký, i když střední hodnota počtu kroků pro všechny obrazy je relativně nízká;
- i střední hodnota počtu kroků potřebných k rozhodnutí je příliš velká, jestliže požadujeme, aby byly malé pravděpodobnosti chybných rozhodnutí.

V praktických úlohách bývá proto účelné klasifikační proceduru po určitém počtu kroků přerušit a klasifikační třídu určit podle nějakého doplňkového kritéria, pracujícího s pevným počtem příznaků. To lze zajistit:

- předepsáním určitého maximálního počtu kroků;
- zavedením proměnných hranic A(i) a B(i).

Nechť A(i), příp. B(i) je nezáporná nerostoucí, resp. nekladná neklesající posloupnost počtu klasifikačních kroků. Klasifikátor v i-tém kroku zařadí obraz $\mathbf{x} = (x_1, x_2, \ldots, x_i)$ do třídy

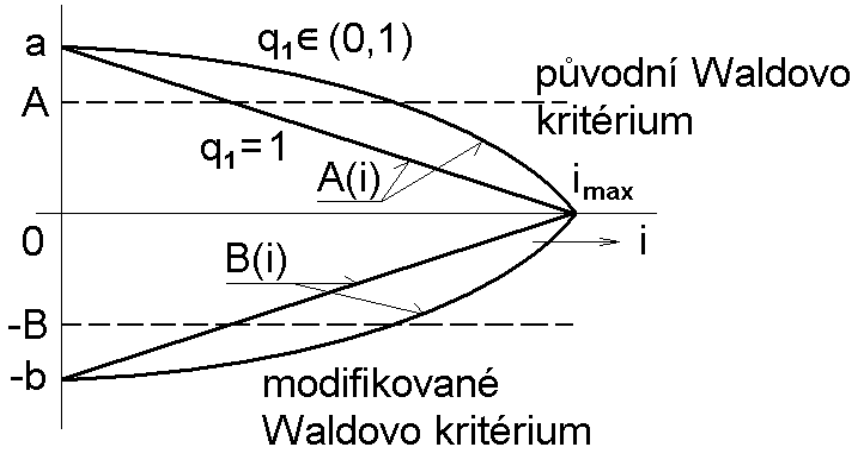
ω_1, pokud $\Lambda_i \geq e^{A(i)}$, do třídy ω_2, pokud $\Lambda_i \leq e^{B(i)}$ a přibírá další příznak, jestliže $\Lambda_i \in (e^{B(i)}, e^{A(i)})$.

Problém analytického stanovení posloupností A(i) a B(i) není obecně vyřešen, obvykle se tato úloha řeší experimentálně. Jestliže pro obě mezní posloupnosti platí, že $A(i_{max}) = B(i_{max})$ (obr.2.33), pak nejpozději pro $i = i_{max}$ je klasifikační procedura ukončena, přičemž střední počet kroků potřebných k rozhodnutí je menší než u standardního Waldova kritéria, ovšem za tu cenu, že nemusí být splněny požadavky na pravděpodobnost chybného rozhodnutí.

Poměrně širokou třídou funkcí používaných pro určení proměnných hranic modifikovaného Waldova kritéria jsou funkce

$$A(i) = a.\left(1 - \frac{i}{i_{max}}\right)^{q_1} \quad a \quad B(i) = -b.\left(1 - \frac{i}{i_{max}}\right)^{q_2}, \tag{2.209}$$

kde a, b > 0, q_1, $q_2 \in (0,1\rangle$ a i_{max} je předem stanovená hodnota maximálního počtu klasifikačních kroků klasifikátoru. Průběh funkcí A(i) a B(i) podle (2.209) s různými hodnotami exponentů q_1 a q_2 jsou na obr.2.37. Pro $i \to \infty$ jsou A(i) = a a B(i) = -b a klasifikační procedura odpovídá originálnímu Waldovu kritériu s mezemi $A = e^a$ a $B = e^{-b}$.



Obr.2.37 *Závislost klasifikačních hranic Waldova kritéria na počtu příznaků*

2.6.5 Modifikované Reedovo kritérium

Princip konstrukce proměnných mezních hodnot tak, jak byl použit u modifikovaného Waldova kritéria, lze použít i pro případ více klasifikačních tříd. V každém klasifikačním kroku může být hodnota zobecněného věrohodnostního poměru $\Lambda_i(\mathbf{x} \mid \omega_r)$ pro všechny třídy srovnávána s prahem definovaným vztahem

$$G_r(i) = g_r.\left(1 - \frac{i}{i_{max}}\right)^{q_r}, \quad r = 1,...,R. \tag{2.210}$$

Když platí, že $\Lambda_i(\mathbf{x}|\omega_r) < G_r(i)$, pak je třída ω_r vyloučena z dalšího rozhodování a počet možných klasifikačních tříd se sníží o jednu. Výpočet pokračuje způsobem ekvivalentním postupu popsaného v případě standardního Reedova kritéria, dokud nezůstane poslední jediná třída, do které se vstupní obraz přiřadí.

3 Volba a výběr příznaků

3.1 Úvod

Pro správnou činnost potřebuje klasifikátor dostatečné množství kvalitní informace. Intuitivně lze předpokládat, že čím větší množství informace data nesou, tím správnější bude rozhodování klasifikátoru, případně tím menší bude možnost, že se zmýlí. Z toho vyplývá, že čím úplnější popis klasifikovaného objektu zprostředkuje jeho matematický popis, tím kvalitnější by měla být činnost klasifikátoru. Taková úvaha v jednoduchém důsledku vede k co nejpodrobnějšímu popisu objektu pomocí velkého počtu příznaků.

Rostoucí počet příznaků ale na druhé straně komplikuje technickou realizaci klasifikátoru. Roste složitost rozhodovacího algoritmu a tím i požadavky na jeho návrh, příp. i na výpočetní čas potřebný ke klasifikaci. Z hlediska technického řešení je proto žádoucí počet příznaků v obrazu klasifikovaného objektu co nejvíce omezit.

Z těchto dvou protichůdných požadavků logicky vyplývá, že řešení každé konkrétní klasifikační úlohy spočívá v nalezení rozumného kompromisu mezi správností klasifikace a požadavky na její technickou realizaci. Abychom takový kompromis nalezli, je pro danou úlohu třeba:

- definovat přípustnou míru spolehlivosti klasifikace;
- určit ty příznakové proměnné, jejichž hodnoty nesou nejvíce informace, tj. ty proměnné, které jsou nejefektivnější pro co nejlepší separaci požadovaných klasifikačních tříd.

Definice míry spolehlivosti určuje optimalizační kritérium, podle kterého jsou příznakové proměnné hodnoceny a vybírány. V převážné většině případů se používá pravděpodobnosti chybné klasifikace, či různých dalších sofistikovanějších kritérií z pravděpodobnosti chybné klasifikace odvozených, jako jsou hodnoty senzitivity a specificity, tvaru tzv. pracovní charakteristiky klasifikátoru (ROC – *Receiver Operating Characteristic*). Vhodným kritériem může být i odchylka obrazu vytvořeného z vybraných příznaků vůči určitému referenčnímu, ve stanoveném smyslu ideálnímu obrazu.

Způsob, jak určit příznakové veličiny nesoucí nejvíce informace pro klasifikátor, není teoreticky formalizován, tj. neexistuje teoretický aparát, pomocí kterého by bylo možné předem stanovit veličiny, jejichž hodnoty poskytují užitečnou informaci, nebo naopak ty, které jsou pro klasifikaci nedůležité. Teorie nabízí pouze dílčí, suboptimální řešení, spočívající ve výběru nezbytného počtu veličin z předem zvolené množiny příznakových veličin, příp. ve vyjádření původních příznakových veličin pomocí menšího počtu skrytých (latentních) nezávislých proměnných, které nelze přímo měřit, ale mohou, ovšem také nemusí mít určitou věcnou interpretaci. První z obou postupů má přímý důsledek i na optimalizaci pořizování dat (není nadále nutné měřit ty veličiny, které neprokážou, že obsahují vhodné množství informace). Naopak, druhý postup předpokládá kompletní vstupní data, která pouze transformuje a vytváří tím možnost jejich efektivnějšího zpracování.

3.2 Volba příznaků

V žádném z obou přístupů však není specifikováno, jak určit výchozí množinu příznakových proměnných. To se zpravidla děje na základě empirie a expertní analýzy, vycházející ze znalosti podstaty řešeného problému. Může to být i na základě simulačních výpočtů s matematickými modely analyzovaných jevů a procesů. Neméně důležitým aspektem pro tuto počáteční volbu je i naše schopnost, daná technickými možnostmi, určité veličiny měřit.

Není proto jisté, zda zvolená výchozí množina bude obsahovat právě ty veličiny, jejichž hodnoty jsou pro danou klasifikační úlohu nejužitečnější.

Přes empirický charakter počáteční volby příznakových veličin platí některé zásady, kterými se lze při této volbě řídit. Jak záhy uvidíme, s některými principy jsme se už setkali v dřívějších kapitolách.

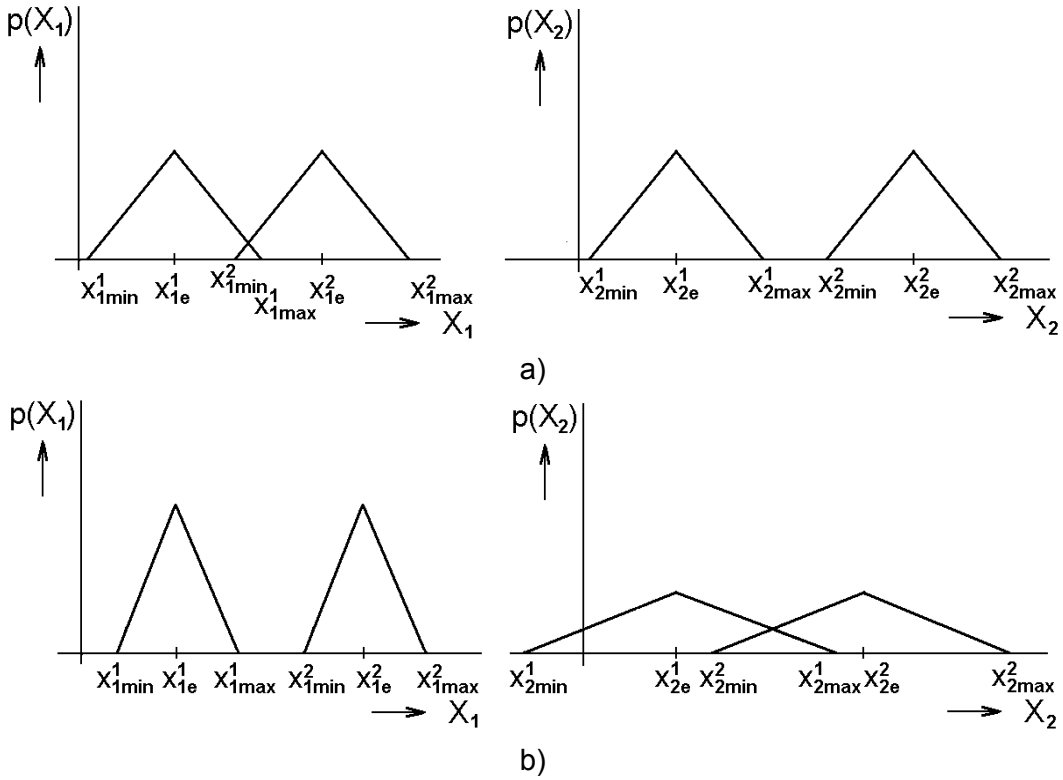
Podle první myšlenky, která napadne, to vypadá, že užitečnější (informativnější) pro klasifikaci by měla být ta příznaková veličina, pro kterou jsou dvě klasifikační třídy vzdálenější. Samotná vzdálenost ale sama o sobě není rozhodující, při zvažování, zda použít tu kterou příznakovou veličinu je třeba vzít v úvahu i rozptyl hodnot uvažovaných příznakových veličin. Lze tedy formulovat dva základní principy (viz také kap.2.4.3 o Fisherově diskriminaci):

a) ***výběr veličin s maximální vzdáleností mezi třídami***

Pokud je rozptyl příznaků ve dvou klasifikačních třídách stejný, pak jsou třídy lépe rozlišitelné pro tu příznakovou veličinu, pro kterou je vzdálenost mezi třídami větší (obr.3.1a).

b) ***výběr veličin s minimálním rozptylem uvnitř tříd***

Když je vzdálenost dvou tříd pro různé příznakové veličiny stejná, pak jsou třídy lépe rozlišitelné s tou příznakovou veličinou, jejíž hodnoty se pro každou třídu mění méně, tj. jejíž rozptyl v jednotlivých klasifikačních třídách je menší (obr.3.1b). Jinými slovy, čím menší je rozptyl příznakové veličiny uvnitř klasifikační třídy, tím více informace nese příznaková veličina o třídě, do které patří.



Obr.3.1 *Zásady pro volbu příznaků – a) preference maximální vzdálenosti mezi třídami; b) preference minimálního rozptylu uvnitř tříd*

Jestliže vyjádříme rozložení hodnot uvažovaných příznakových veličin pomocí hustoty pravděpodobnosti, tak jak je naznačeno na obr.3.1, je optimální volba charakterizována minimálním překryvem obou hustot, tj. situací, která znamená minimalizaci chybných rozhodnutí.

c) ***výběr vzájemně nekorelovaných veličin***

Pokud je možné hodnoty jedné příznakové veličiny odvodit z příznaků veličiny druhé, potom použití obou těchto veličin nepřináší žádnou další informaci o správné klasifikaci oproti použití pouze jedné z nich, jedno které.

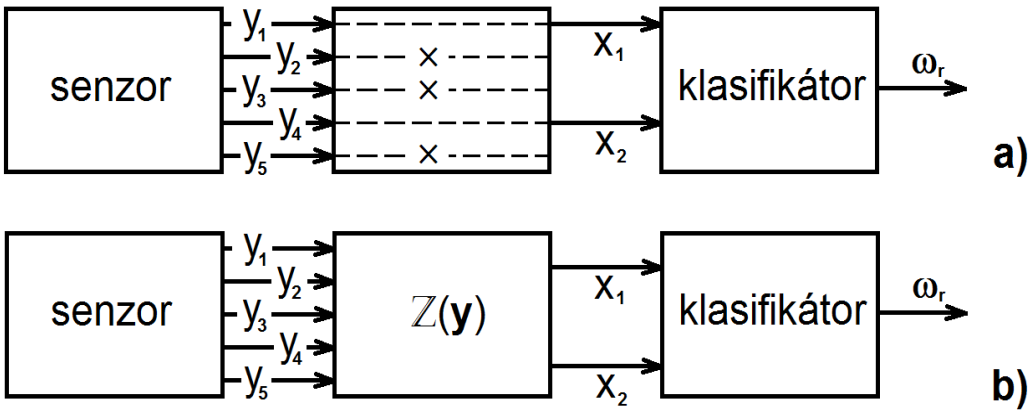
d) ***výběr veličin invariantních vůči deformacím (fluktuacím, variabilitě)***

Poslední požadavek je především praktický. Dle kap.1.3 a obr.1.2 závisí volba elementů formálního (matematického) popisu klasifikovaného objektu na jeho charakteru, charakteru původních údajů o něm, i na způsobu předzpracování. V těch případech, kdy je odstranění deformací dat příliš náročné, případně nejde vůbec realizovat, je třeba vybrat takové příznakové veličiny, které nejsou rušením ovlivněny, resp. podstatně ovlivněny. Někdy lze výběrem příznakové veličiny fázi předzpracování eliminovat a tak zjednodušit celé zpracování, i když jsou algoritmy předzpracování jednoduše realizovatelné.

3.3 Výběr příznaků

Jak bylo uvedeno dříve, nedokážeme určit nejvhodnější veličiny z hlediska klasifikace přímo, nýbrž pouze vybrat z předem dané množiny veličin. To znamená, že se obraz, reprezentovaný původně m-rozměrným příznakovým vektorem, snažíme vyjádřit vektorem n-rozměrným ($m \geq n$) tak, aby množství tzv. diskriminační informace obsažené v původním vektoru bylo v co největší míře zachováno. Výběr příznaků se tedy převádí na hledání zobrazení $\mathbb{Z}$: $\mathcal{Y}^m \rightarrow \mathcal{X}^n$, kterým se původní m-rozměrný prostor $\mathcal{Y}^m$ transformuje do nového prostoru $\mathcal{X}^n$.

Zmenšení rozměru obrazového prostoru lze dosáhnout dvěma principiálně různými způsoby (obr.3.2):



Obr.3.2 *Principy výběru příznaků – a) selekce; b) extrakce.*

a) ***selekce*** – nalezení těch příznakových funkcí, které přispívají k separabilitě klasifikačních tříd nejméně a pro klasifikaci se z původní množiny ponechá jen n nejvíce informativních proměnných. Zobrazení $\mathbb{Z}$ tedy pouze vynechává m – n příznakových proměnných.

b) ***extrakce*** – transformace původních příznakových proměnných na menší počet jiných příznakových proměnných.

K tomu, abychom dokázali realizovat libovolný z obou způsobů výběru příznaků, je třeba definovat a splnit určité podmínky optimality.

Nechť J je kriteriální funkce, jejíž pomocí vybíráme příznakové proměnné. Pak v případě selekce vybíráme vektor $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ ze všech možných n-tic χ příznaků y_i, i = 1, 2, …, m. Optimalizaci výběru příznaků tudíž můžeme formálně zapsat jako

$$\mathbb{Z}(\mathbf{y}) = \underset{\forall \chi}{\text{extr}}\, J(\chi)\,. \tag{3.1}$$

Problémy, které je nutné vyřešit, jsou stanovení kriteriální funkce, rozměru nového příznakového prostoru a optimalizačního postupu.

Při extrakční alternativě transformujeme příznakový prostor na základě výběru zobrazení $\mathbb{Z}$ z množiny všech možných zobrazení ζ prostoru $\mathcal{Y}^m$ do $\mathcal{X}^n$, tj.

$$\mathbb{Z}(\mathbf{y}) = \underset{\forall\zeta}{\text{extr}}\, J(\zeta)\,. \tag{3.2}$$

I v tomto případě je potřeba určit vhodnou kriteriální funkci, rozměr nového obrazového prostoru, zvolit požadavky na vlastnosti zobrazení i vhodný optimalizační postup (pokud nevyplývá z vlastností zobrazení).

3.3.1 Selekce příznaků

Poměr rozptylů

Jak bylo uvedeno mimo jiné i v kap.3.2, pro klasifikaci jsou výhodnější ty příznaky, pro které je menší rozptyl obrazů uvnitř klasifikačních tříd a současně co největší vzdálenost (rozptyl) mezi třídami. To znamená, že se lze při selekci příznaků řídit hodnotami poměru rozptyl mezi třídami vzhledem k rozptylu uvnitř tříd. Čím větší bude tento poměr, tím méně pravděpodobná bude chyba klasifikace a tím také bude lépe proveden výběr příznaků.

Ke stanovení zmíněného poměru je třeba charakterizovat oba použité rozptyly. Zatímco rozptyl uvnitř tříd lze charakterizovat disperzní maticí

$$\mathbf{D}(\boldsymbol{\chi}) = \sum_{r=1}^{R} P(\omega_r) \int_{\mathcal{X}} (\boldsymbol{\chi} - \boldsymbol{\mu}_r).(\boldsymbol{\chi} - \boldsymbol{\mu}_r)^T . p(\boldsymbol{\chi}|\omega_r).d\boldsymbol{\chi}\,, \tag{3.3}$$

kde

$$\boldsymbol{\mu}_r = \int_{\mathcal{X}} \boldsymbol{\chi}.p(\boldsymbol{\chi}|\omega_r).d\boldsymbol{\chi}\,. \tag{3.4}$$

Rozptyl mezi třídami může být definován např. vztahem

$$\mathbf{B}(\boldsymbol{\chi}) = \sum_{r=1}^{R-1} \sum_{s=r+1}^{R} P(\omega_r)P(\omega_s).\boldsymbol{\mu}_{rs}.\boldsymbol{\mu}_{rs}^T\,, \tag{3.5}$$

kde $\boldsymbol{\mu}_{rs} = \boldsymbol{\mu}_r - \boldsymbol{\mu}_s$.

Pokud

$$\boldsymbol{\mu}_0 = \sum_{r=1}^{R} P(\omega_r)\boldsymbol{\mu}_r = \int_{\mathcal{X}} \boldsymbol{\chi}.p(\boldsymbol{\chi}).d\boldsymbol{\chi}\,, \tag{3.6}$$

můžeme také psát

$$\mathbf{B}(\boldsymbol{\chi}) = \sum_{r=1}^{R} P(\omega_r).(\boldsymbol{\mu}_r - \boldsymbol{\mu}_0).(\boldsymbol{\mu}_r - \boldsymbol{\mu}_0)^T\,. \tag{3.7}$$

Jestliže je disperzní matice $\mathbf{D}(\boldsymbol{\chi})$ regulární, tj. jestliže má inverzní matici, pak lze vyjádřit vlastnosti výskytu obrazů v obrazovém prostoru při zvolené kombinaci příznaků, např. vztahem

$$J_{r1}(\boldsymbol{\chi}) = \text{tr}\left(\mathbf{D}^{-1}(\boldsymbol{\chi}).\mathbf{B}(\boldsymbol{\chi})\right). \tag{3.8}$$

Další možné používané způsoby popisu rozptylových vlastností obrazů jednoduchým parametrem jsou

$$J_{r2}(\boldsymbol{\chi}) = \text{tr}\left(\mathbf{B}(\boldsymbol{\chi})\right) / \text{tr}\left(\mathbf{D}(\boldsymbol{\chi})\right); \tag{3.9}$$

$$J_{r3}(\chi) = \left|\mathbf{D}^{-1}(\chi).\mathbf{B}(\chi)\right| = \left|\mathbf{B}(\chi)\right| / \left|\mathbf{D}(\chi)\right|, \tag{3.10}$$

resp. pro omezení rozsahu hodnot parametru J_{r3}

$$J_{r4}(\chi) = J_{r3}(\chi) . \tag{3.11}$$

Algoritmy selekce příznaků

Problém selekce příznaků spočívá ve výběru optimální podmnožiny obsahující n příznakových proměnných ($n \le m$). Její hledání je kombinatorický problém, přičemž celkový počet možných podmnožin, které při daných počtem proměnných m a n můžeme vytvořit je určeno výrazem m!/(m-n)!n!. To je číslo příliš veliké i pro ne příliš velké hodnoty m a n, než abychom byli schopni určit optimální řešení na základě stanovení vlastností všech možných variant. To vede k vytváření postupů, které umožňují najít alespoň kvazioptimální řešení, ovšem s přijatelnými nároky na výpočet.

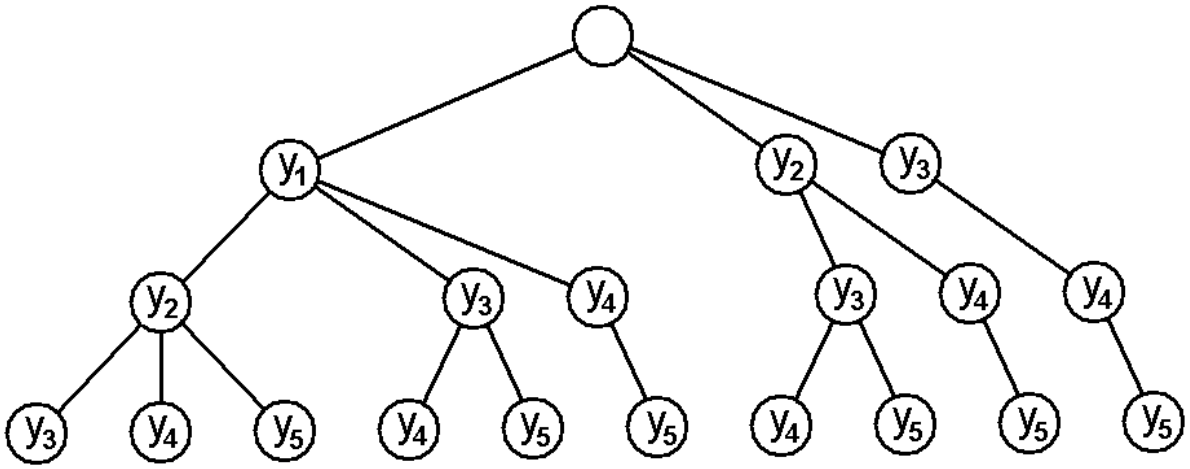
Algoritmus ohraničeného větvení umožňuje stanovit optimální množinu příznaků za předpokladu, že kriteriální funkce pro selekci příznaků je monotónní. Označíme-li X_j množinu j příznaků, pak monotónnost kriteriální funkce znamená, že pro množiny

$$X_1 \subset X_2 \subset \ldots \subset X_j \subset \ldots \subset X_m \tag{3.12}$$

splňuje selekční kriteriální funkce vztah

$$J(X_1) \le J(X_2) \le \ldots \le J(X_j) \le \ldots \le J(X_m). \tag{3.13}$$

Pro popis algoritmu uvažme případ selekce dvou příznaků z původních pěti. Všechny možné alternativy vyloučení tří příznakových proměnných z výchozí množiny ukazuje graf na obr.3.3. Každý uzel v grafu vyjadřuje eliminaci jedné označené příznakové proměnné.



Obr.3.3 *Graf algoritmu ohraničeného větvení - selekce 2 z 5*

Předpokládejme, že vyhodnocujeme hodnotu zvolené kriteriální funkce v každém uzlu stromu, přičemž postupujeme shora dolů a zleva doprava, a srovnáváme ji s dosud nejlepší dosaženou hodnotou, kterou označíme J_0. Pokud je okamžitá hodnota kriteriální funkce větší než J_0, je stále šance, že optimální řešení bude nalezeno na právě analyzované větvi grafu a proto bude hledání pokračovat po nejlevější dosud neanalyzované větvi. Jestliže dosáhneme konce větve a odpovídající hodnota selekčního kritéria je větší než J_0, pak tento uzel definuje novou optimální množinu příznaků a modifikujeme hodnotu J_0. Naopak, jestliže je v některém uzlu grafu hodnota selekčního kritéria menší než J_0, pak větve začínající v tomto uzlu nemá

smysl dále prohledávat, protože díky monotónnosti kritéria budou jeho hodnoty v dalších uzlech již jenom stále menší.

Efektivnost prohledávání se ještě zvětší, jestliže se bude výběr odstraňované veličiny na dané úrovni stromu provádět podle změny hodnoty kriteriální funkce a bude se postupovat tím směrem, kde je změna kriteriální funkce nejmenší.

Algoritmus sekvenční dopředné selekce je spolu s následujícím algoritmem nejjednodušší procedurou, která hledá suboptimální řešení. Algoritmus začíná s prázdnou množinou, do které se vloží proměnná s nejlepší hodnotou selekční kriteriální funkce. Dále, v každém následujícím kroku se přidává ta proměnná, která s dříve vybranými veličinami dosáhla nejlepší hodnoty kritéria, tj.

$$J(\mathcal{X}_{k+1}) = \max_{\forall j} J(\mathcal{X}_k \cup y_j),\ y_j \in (\mathcal{Y} - \mathcal{X}_k). \tag{3.14}$$

Algoritmus sekvenční zpětné selekce začíná, na rozdíl od předešlého, s množinou všech výchozích příznakových veličin. V každém kroku se eliminuje ta proměnná, která způsobí nejmenší pokles hodnoty kriteriální funkce, tj. po (k+1)-ním kroku platí

$$J(\mathcal{X}_{m-k+1}) = \max J(\mathcal{X}_{m-k} - y_j),\ y_j \in \mathcal{X}_{m-k}). \tag{3.15}$$

Důvodem pouhé suboptimality nalezeného řešení je v případě dopředné selekce to, že z vytvořené množiny nelze vyloučit ty veličiny, které se staly nadbytečnými po přiřazení dalších veličin. Podobně u zpětného algoritmu neexistuje možnost opravy při neoptimálním vyloučení kterékoliv proměnné. Dopředný algoritmus je výpočetně jednodušší, protože pracuje maximálně v n-rozměrném prostoru, naopak zpětný umožňuje průběžně sledovat množství ztracené informace.

Algoritmus plus p – mínus q pomáhá částečně napravit suboptimalitu obou výše uvedených algoritmů tím, že po přidání p veličin se q veličin odstraní. Proces probíhá, dokud se nedosáhne požadovaného počtu příznaků. Je-li $p > q$, pracuje algoritmus stejně jako dopředný algoritmus od prázdné množiny. Naopak, je-li $p < q$, jedná se o variantu zpětného algoritmu.

Algoritmus min – max je heuristický algoritmus, který umožňuje vybírat příznaky na základě výpočtů hodnot kriteriální funkce pouze v jedno- a dvourozměrném příznakovém prostoru.

Předpokládejme, že již bylo vybráno k příznakových veličin do množiny $\mathcal{X}_k$, pro výběr tedy zbývají veličiny z množiny $\mathcal{Y} - \mathcal{X}_k$. Výběr veličiny $y_j \in (\mathcal{Y} - \mathcal{X}_k)$ přináší novou informaci, kterou můžeme ocenit relativně k libovolné veličině $x_i \in \mathcal{X}_k$ podle vztahu

$$\Delta J(y_j,x_i) = J(y_j,x_i) - J(x_i). \tag{3.16}$$

Máme samozřejmě zájem, aby tento informační přírůstek byl co největší, nicméně musí být dostatečný vzhledem ke všem veličinám již zahrnutým do množiny $\mathcal{X}_k$. Vybíráme proto takovou veličinu y_{k+1}, pro kterou platí

$$\Delta J(y_{k+1},x_k) = \max_{\forall j} \min_{\forall i} \Delta J(y_j,x_i),\ x_i \in \mathcal{X}_k,\ y_j \in \{\mathcal{Y} - \mathcal{X}_k\}. \tag{3.17}$$

3.3.2 Extrakce příznaků

Jak bylo uvedeno v kap.3.2, spočívá extrakce příznaků v hledání optimálního zobrazení $\mathbb{Z}$, které transformuje původní m-rozměrný obraz popisující analyzovaný objekt na obraz n-

rozměrný. Prvním předpokladem pro nalezení vhodného zobrazení je stanovení kritéria optimality. V současné praxi se používá především tří následujících kritérií:

- zobrazení $\mathbb{Z}$ se určí tak, aby obrazy z nového prostoru $\mathcal{X}^n$ aproximovaly původní m-rozměrné obrazy z $\mathcal{Y}^m$ ve smyslu minimální střední kvadratické odchylky;
- zobrazení $\mathbb{Z}$ se určí tak, aby rozložení pravděpodobnosti veličin v novém prostoru splňovaly podmínky kladené na jejich pravděpodobnostní charakteristiky;
- zobrazení $\mathbb{Z}$ se určí tak, aby obrazy z $\mathcal{X}^n$ minimalizovaly odhad pravděpodobnosti chyby.

Aby byl uvedený problém teoreticky příjemně řešitelný, vybírá se zobrazení $\mathbb{Z}$ z oboru lineárních zobrazení.

Z metod extrakce příznaků se dále budeme podrobněji zabývat dvěma reprezentativními postupy – analýzou hlavních komponent a analýzou nezávislých komponent.

3.3.3 Analýza hlavních komponent

Odvození

Analýza hlavních komponent (PCA – *Principal Component Analysis*) je jednou ze základních metod extrakce příznaků. Teoreticky je založena na transformaci původního obrazového prostoru pomocí metody Karhunenova – Loevova rozvoje, který vychází z prvního z uvedených kritérií optimality.

Předpokládejme, že je dáno K m-rozměrných obrazů, které primárně nejsou rozděleny do klasifikačních tříd. Pak je k-tý obraz vyjádřen m-rozměrným sloupcovým vektorem $\mathbf{y}_k \in \mathcal{Y}^m$, k = 1, 2, …, K. Aproximujme nyní každý obraz $\mathbf{y}_k$ lineární kombinací n ortonormálních vektorů $\mathbf{e}_i$ ($n \le m$). Tedy platí

$$\mathbf{x}_k = \sum_{i=1}^{K} c_{ki}\mathbf{e}_i \,. \tag{3.18}$$

Koeficienty c_{ki} lze považovat za velikost i-té souřadnice vektoru $\mathbf{y}_k$ vyjádřeného v novém systému souřadnic s bází $\mathbf{e}_i$, i = 1, 2, …, n, tj. platí

$$c_{ki} = \mathbf{y}_k^T.\mathbf{e}_i \,. \tag{3.19}$$

Volíme-li jako kritérium optimality zobrazení, jak jsme již předeslali, kritérium minimální střední kvadratické odchylky, pak musíme stanovit vztah pro určení kvadratické odchylky ε_k^2 původního obrazu $\mathbf{y}_k$ od jeho aproximace $\mathbf{x}_k$. Nechť je

$$\varepsilon_k^2 = \left\|\mathbf{y}_k - \mathbf{x}_k\right\|^2. \tag{3.20}$$

Pak pomocí vztahů (3.18) a (3.19) je

$$\varepsilon_k^2 = \left\|\mathbf{y}_k\right\|^2 - \sum_{i=1}^{n} c_{ki}^2. \tag{3.21}$$

Střední kvadratická odchylka pro všechny obrazy $\mathbf{y}_k$, k = 1, 2, …, K je

$$\varepsilon^2 = \frac{1}{K}\sum_{k=1}^{K}\varepsilon_k^2 = \frac{1}{K}\sum_{k=1}^{K}\left\|\mathbf{y}_k\right\|^2 - \sum_{i=1}^{n}\mathbf{e}_i^T\left[\frac{1}{K}\sum_{k=1}^{K}\mathbf{y}_k\mathbf{y}_k^T\right]\mathbf{e}_i \tag{3.22}$$

a je závislá na volbě ortonormálního bázového systému $\mathbf{e}_i$, který je třeba zvolit tak, aby odchylka ε^2 byla minimální. Diskrétní konečný rozvoj podle vztahu (3.18) s bázovým systémem

$\mathbf{e}_i$ optimálním podle kritéria minimální střední kvadratické odchylky nazýváme diskrétní Karhunenův - Loevův rozvoj.

Aby byla střední kvadratická odchylka definovaná vztahem (3.22) minimální, musí druhý člen na pravé straně uvedené rovnice nabývat maximální hodnoty (vzhledem k tomu, že první člen pravé strany uvedené rovnice je pro dané zadání úlohy konstantní). Je tedy nutné maximalizovat výraz

$$\sum_{i=1}^{n} \mathbf{e}_i^T \boldsymbol{\kappa}(\mathbf{y}) \mathbf{e}_i , \tag{3.23}$$

kde

$$\boldsymbol{\kappa}(\mathbf{y}) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{y}_k \mathbf{y}_k^T . \tag{3.24}$$

Je autokorelační matice řádu m. Z jejích vlastností (symetrická, semidefinitní) vyplývá, že její vlastní čísla λ_i, i = 1, 2, …, m jsou reálná, nezáporná a jim odpovídající vlastní vektory $\mathbf{v}_i$, i = 1, 2, …, m jsou buď ortonormální, nebo je můžeme ortonormalizovat (v případě vícenásobných vlastních čísel).

Uspořádáme-li vlastní čísla sestupně podle velikosti, tj.

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0 \tag{3.25}$$

a podle tohoto seřazení očíslujeme i odpovídající vlastní vektory, pak lze dokázat, že výraz (3.24) dosahuje maxima, jestliže platí

$$\mathbf{e}_i = \mathbf{v}_i, \; i = 1, 2, \ldots, n \tag{3.26}$$

a pro velikost maxima je

$$\max \sum_{i=1}^{n} \mathbf{e}_i^T \boldsymbol{\kappa}(\mathbf{y}) \mathbf{e}_i = \sum_{i=1}^{n} \lambda_i . \tag{3.27}$$

Pro minimální střední kvadratickou odchylku tedy platí

$$\varepsilon_{min}^2 = \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{y}_k\|^2 - \sum_{i=1}^{n} \lambda_i = \mathrm{tr}\left(\boldsymbol{\kappa}(\mathbf{y})\right) - \sum_{i=1}^{n} \lambda_i = \sum_{i=n+1}^{m} \lambda_i . \tag{3.28}$$

To znamená, že je rovna součtu těch vlastních čísel, jimž odpovídající vlastní vektory nebyly použity při aproximaci obrazu podle vztahu (3.18). Pro n = m je střední kvadratická odchylka nulová.

V některých případech je vhodnější vektory $\mathbf{y}_1$, …, $\mathbf{y}_K$ před aproximací centrovat[19]. V tom případě vypočítáme střední hodnotu

$$\mu = \frac{1}{K} \sum_{k=1}^{K} \mathbf{y}_k \tag{3.29}$$

a místo s obrazem $\mathbf{y}_k$ počítáme s jeho centrovanou verzí $\bar{\mathbf{y}}_k = \mathbf{y}_k - \mu$.

[19] Zde je dobré opět připomenout, že centrováním (odečtením střední hodnoty) přicházíme o určitou informaci v datech, která charakterizuje analyzovaný objekt a kterou již nebude možné nadále využít. Je proto potřeba, abychom si byli v tomto okamžiku zcela jistí, že střední hodnota v datech nenese žádnou informaci podstatnou z hlediska řešené analytické, resp. klasifikační úlohy.

Postup výpočtu Karhunenova – Loevova rozvoje se nemění, ale místo autokorelační matice používáme matici disperzní ve tvaru

$$\mathbf{D}(\mathbf{y}) = \frac{1}{K}\sum_{k=1}^{K} \bar{\mathbf{y}}_k \bar{\mathbf{y}}_k^T . \tag{3.30}$$

Platí, že

$$\boldsymbol{\kappa}(\mathbf{y}) = \mathbf{D}(\mathbf{y}) + \boldsymbol{\mu}.\boldsymbol{\mu}^T . \tag{3.31}$$

Ortonormální systém $\mathbf{e}_1, \ldots, \mathbf{e}_n$ je v tom případě roven vlastním vektorům $\mathbf{v}_1, \ldots, \mathbf{v}_n$ disperzní matice $\mathbf{D}(\mathbf{y})$.

Podobně, v případě standardizovaných dat, tj. když jsou po odečtení střední hodnoty jednotlivé hodnoty příznakových proměnných ještě poděleny patřičnou směrodatnou odchylkou, pak místo autokorelační matice dostáváme matici hodnot Pearsonova korelačního koeficientu, které popisují vzájemné korelační vztahy mezi jednotlivými proměnnými. Závěry a důsledky vyplývající z výpočtů vlastních čísel a vektorů takovéto matice zůstávají v principu zachovány, jen je třeba si uvědomit, že se mění charakter výchozích dat.

Diskrétní Karhunenův – Loevův rozvoj a na něj navazující analýza hlavních komponent má velice názornou matematickou interpretaci (obr.3.4). Nechť je původní obrazový prostor dvourozměrný a je dán příznakovými veličinami Y_1 a Y_2 a obraz $\mathbf{y}$ má tedy v původní souřadnicové soustavě souřadnice y_1 a y_2. Po transformaci souřadnicového systému, která je primárně určena vlastnostmi autokorelační matice množiny obrazů, jsou souřadnice uvedeného obrazu transformovány do hodnot x_1 a x_2. Vzhledem k tomu, že je transformace souřadnicové soustavy lineární, jsou obě nové souřadnice určeny lineární kombinací souřadnic původních, tedy (obr.3.4a,b,c)

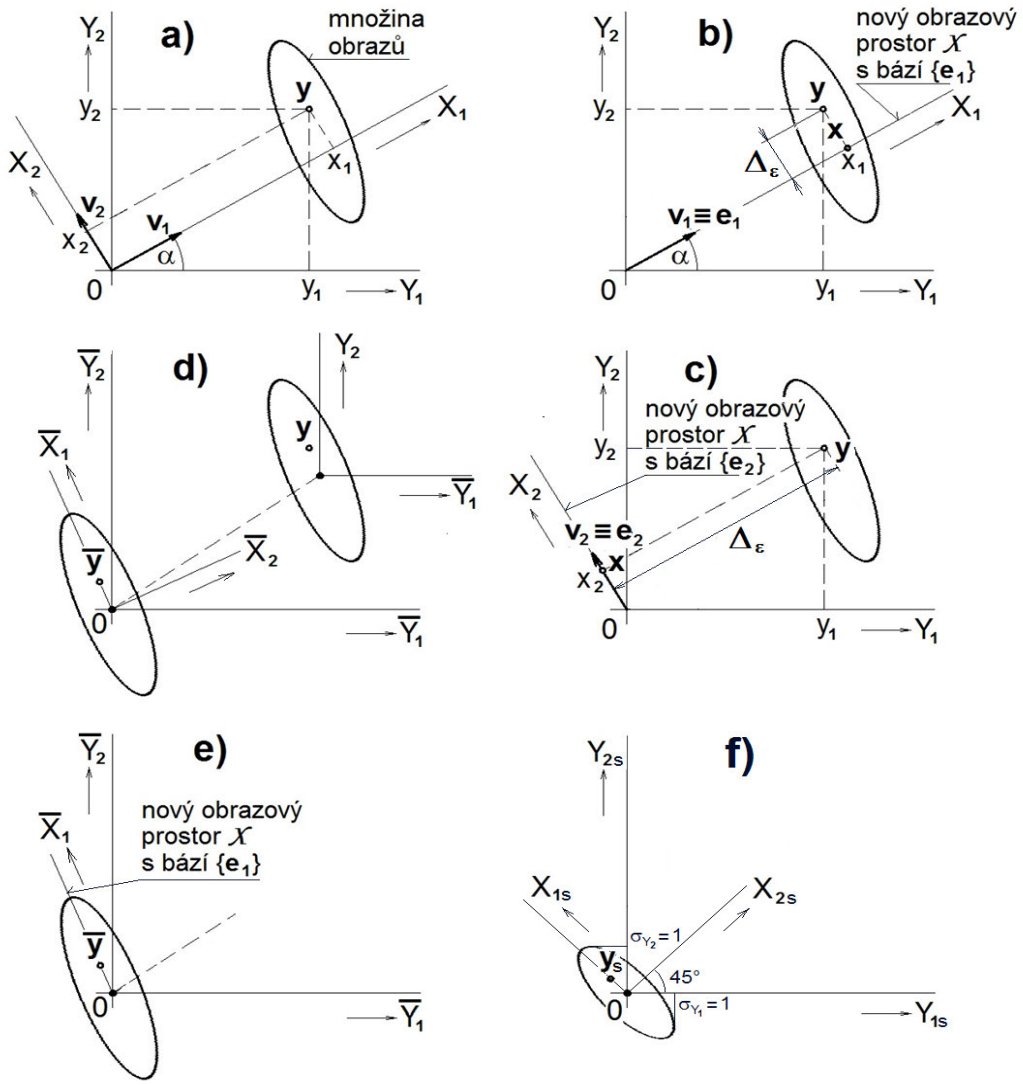
$$\begin{aligned} X_1 &= a.Y_1 + b.Y_2 = \cos\alpha.Y_1 + \sin\alpha.Y_2; \\ X_2 &= c.Y_1 + d.Y_2 = \sin\alpha.Y_1 + (\cos 2\alpha/\cos\alpha).Y_2. \end{aligned} \tag{3.32}$$

Pokud nedojde k redukci rozměru obrazového prostoru je obraz i v transformovaných souřadnicích vyjádřen zcela přesně. Omezíme-li ale počet souřadnic, vynechávají se nejdříve souřadnice, které způsobují menší střední kvadratickou chybu, jinými slovy méně přispívají k výsledné aproximaci, v zobrazeném případě souřadnice x_2. Hodnota chyby je určena právě těmito vynechanými souřadnicemi.

Při nulovém rozptylu jsou vlastní čísla autokorelační matice $\boldsymbol{\kappa}(\mathbf{y}) = \boldsymbol{\mu}.\boldsymbol{\mu}^T$ rovna $\lambda_1 = \|\boldsymbol{\mu}\|^2$ a $\lambda_2 = \ldots = \lambda_m = 0$. Vlastní vektor $\mathbf{v}_1$ prochází právě bodem, ve kterém leží všechny obrazy, a ostatní vektory $\mathbf{v}_2, \ldots, \mathbf{v}_m$ se volí tak, aby i nový souřadnicový systém byl ortonormální. Střední kvadratická odchylka je v tom případě rovna nule.

Pokud data centrujeme (obr.3.4d,e), počítáme s disperzní maticí. Pak má transformovaná bázová soustava seřazeny osy ve směrech největších rozptylů (obr.3.4d), které jsou v této nové souřadnicové soustavě číselně rovny vlastním číslům disperzní matice. Vlastní čísla a vlastní vektory disperzní matice jsou různé od vlastních čísel a vektorů autokorelační matice, proto se oba Karhunenovy – Loevovy rozvoje logicky liší.

Když originální data navíc vztáhneme ke směrodatné odchylce (standardizujeme), tj. odstraníme další možnou užitečnou informaci pro rozlišení dat, dále ztěžujeme výpočet vlastních čísel a vektorů matice korelačních koeficientů - množina obrazů získává kompaktnější, kulovitější tvar, stírá se rozdíl mezi vlivem jednotlivých nových souřadnic, z matematického hlediska autokorelační matice ztrácí dobrou podmíněnost, což v důsledku může vést i k výpočetním chybám (obr.3.4f).



Obr.3.4 *Geometrická interpretace Karhunenova – Loevova rozvoje*

Vlastnosti

Karhunenův – Loevův rozvoj má některé vlastnosti, které jej zvýhodňují před jinými typy transformací:

- při daném počtu n členů rozvoje poskytuje ze všech možných aproximací nejmenší kvadratickou odchylku;
- při použití disperzní matice jsou nové transformované příznakové proměnné nekorelované; pokud se výskyt obrazů řídí normálním rozložením, zajišťuje nekorelovanost příznaků současně i nezávislost;
- členy rozvoje nepřispívají k aproximaci rovnoměrně, vliv každého z členů uspořádané posloupnosti aproximace se zmenšuje s jeho pořadím určeným velikostí odpovídajících vlastních čísel;
- změna požadavků na velikost střední kvadratické odchylky nevyžaduje přepočítat celý rozvoj, je třeba pouze změnit počet jeho členů.

Až dosud jsme předpokládali, že množina aproximovaných obrazů je konečná a že obrazy nejsou rozděleny podle příslušnosti k jednotlivým klasifikačním třídám. Pro klasifikační úlohy je však členění obrazů základním předpokladem, proto se dále zabývejme, jak se změní podmínky, když obrazy **y** budou patřit do R klasifikačních tříd, které budou vymezeny jako

části spojitého obrazového prostoru $\mathcal{Y}^m$. Výskyt obrazů v jednotlivých klasifikačních třídách bude popsán podmíněnými hustotami pravděpodobnosti $p(\mathbf{y}|\omega_r)$ a apriorní pravděpodobnost klasifikačních tříd bude $P(\omega_r)$, r = 1, 2, …, R.

Za těchto podmínek bude autokorelační matice $\boldsymbol{\kappa}(\mathbf{y})$ definována vztahem

$$\boldsymbol{\kappa}(\mathbf{y}) = \sum_{r=1}^{R} P(\omega_r).\int_{\mathcal{Y}^m} \mathbf{y}.\mathbf{y}^T.p(\mathbf{y}|\omega_r)d\mathbf{y} = \int_{\mathcal{Y}^m} \mathbf{y}.\mathbf{y}^T.p(\mathbf{y})d\mathbf{y} \tag{3.33}$$

a disperzní matice buď podle předpisu

$$\mathbf{D}^1(\mathbf{y}) = \sum_{r=1}^{R} P(\omega_r).\int_{\mathcal{Y}^m} (\mathbf{y}-\boldsymbol{\mu}_r).(\mathbf{y}-\boldsymbol{\mu}_r)^T.p(\mathbf{y}|\omega_r)d\mathbf{y}, \tag{3.34}$$

kde

$$\boldsymbol{\mu}_r = \int_{\mathcal{Y}^m} \mathbf{y}.p(\mathbf{y}|\omega_r)d\mathbf{y}, \quad r = 1,2,...,R, \tag{3.35}$$

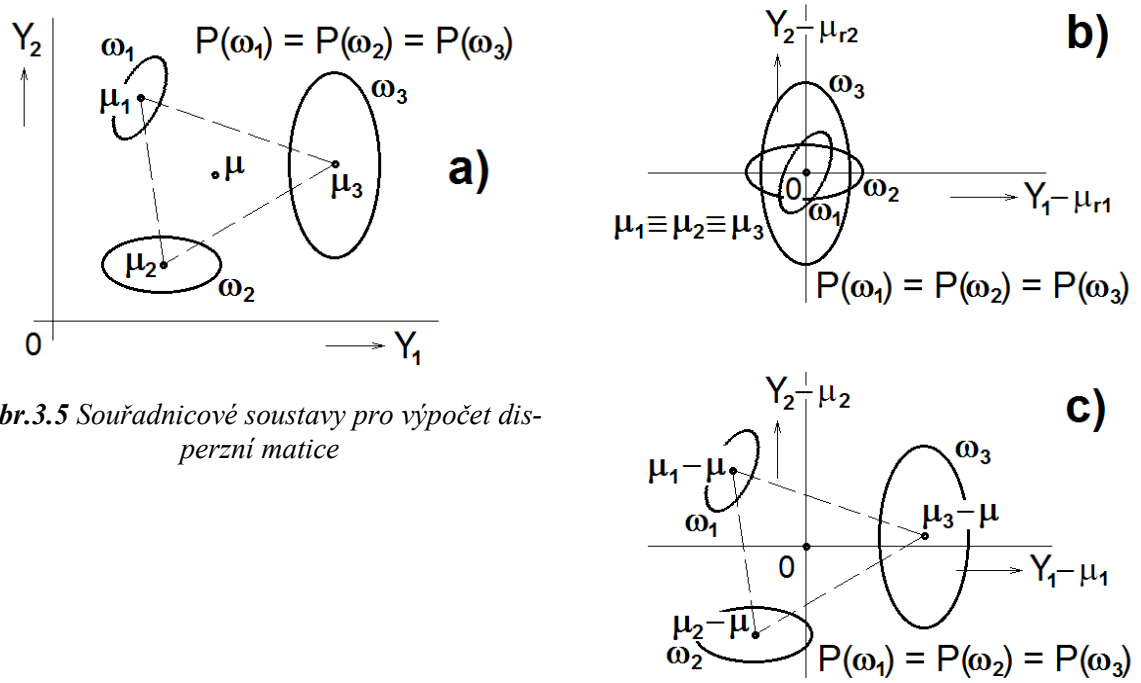
nebo vztahem

$$\mathbf{D}^0(\mathbf{y}) = \sum_{r=1}^{R} P(\omega_r).\int_{\mathcal{Y}^m} (\mathbf{y}-\boldsymbol{\mu}).(\mathbf{y}-\boldsymbol{\mu})^T.p(\mathbf{y}|\omega_r)d\mathbf{y} = \int_{\mathcal{Y}^m} (\mathbf{y}-\boldsymbol{\mu}).(\mathbf{y}-\boldsymbol{\mu})^T.p(\mathbf{y})d\mathbf{y}, \tag{3.36}$$

když střední hodnota $\boldsymbol{\mu}$ je vážený průměr středních hodnot určených podle vztahu (3.35) (obr.3.5a), tj.

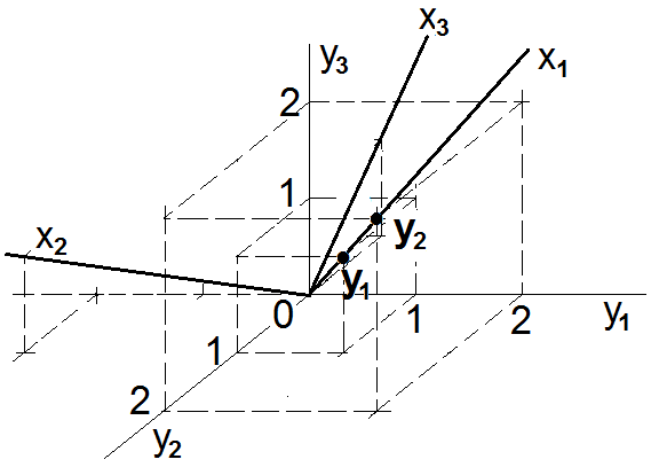
$$\boldsymbol{\mu} = \sum_{r=1}^{R} P(\omega_r).\int_{\mathcal{Y}^m} \mathbf{y}.p(\mathbf{y}|\omega_r)d\mathbf{y} = \int_{\mathcal{Y}^m} \mathbf{y}.p(\mathbf{y})d\mathbf{y}. \tag{3.37}$$

Disperze podle definice (3.34) bere ohled na střední hodnoty obrazů v jednotlivých klasifi-



Obr.3.5 *Souřadnicové soustavy pro výpočet disperzní matice*

kačních třídách, obrazy ze všech klasifikačních tříd se centrují podle středních hodnot obrazů v jednotlivých třídách (obr.3.5b). Klasifikační třídy se tedy po vycentrování mohou rozlišit pouze podle disperze ve směru jednotlivých souřadnicových os. Zato jsou transformované příznakové proměnné zcela nekorelované. Naopak disperze podle vztahu (3.36) centruje obrazy podle celkové průměrné hodnoty, neodstraňuje vliv středních hodnot obrazů v jednotlivých klasifikačních třídách (obr.3.5c) a je proto lépe použít této definice v těch případech, kdy jsou střední hodnoty výrazně odlišné a nesou tak významnou část informace o klasifikační úloze.



Obr.3.6 *Zadání a řešení příkladu*

Příklad

Předpokládejme, že množinu obrazů $\mathcal{Y}^3$ tvoří dva obrazové vektory $\mathbf{y}_1 = (1, 1, 1)^T$ a $\mathbf{y}_2 = (2, 2, 2)^T$ (viz obr.3.6). Pomocí Karhunenova – Loevova rozvoje najděme novou souřadnicovou soustavu, která umožní popsat oba vektory s minimální střední kvadratickou odchylkou.

Jak lze usoudit z elementárního znění zadání a případně i ověřit z grafického vyjádření na obr.3.6, oba zadané vektory leží přesně na přímce dané směrovým vektorem (1, 1, 1). Proto by tento vektor měl být první hlavní komponentou, další dvě souřadnice již nejsou pro vyjádření obou zadaných vektorů podstatné.

Ověřme nyní tento intuitivní závěr výpočtem. Dle definičního vztahu (3.24) pro výpočet autokorelační matice máme

$$\boldsymbol{\kappa} = \frac{1}{2}(\mathbf{y}_1.\mathbf{y}_1^T + \mathbf{y}_2.\mathbf{y}_2^T) = \frac{1}{2}\left[\begin{pmatrix}1\\1\\1\end{pmatrix}.(1\quad 1\quad 1) + \begin{pmatrix}2\\2\\2\end{pmatrix}.(2\quad 2\quad 2)\right] = \begin{bmatrix}2{,}5 & 2{,}5 & 2{,}5\\ 2{,}5 & 2{,}5 & 2{,}5\\ 2{,}5 & 2{,}5 & 2{,}5\end{bmatrix}.$$

Autokorelační matice o rozměru 3 x 3 má všechny tři řádky stejné, tj. jsou lineárně závislé. Vlastní čísla λ_i, která vypočítáme ze vztahu

$$\det\begin{bmatrix}2{,}5-\lambda & 2{,}5 & 2{,}5\\ 2{,}5 & 2{,}5-\lambda & 2{,}5\\ 2{,}5 & 2{,}5 & 2{,}5-\lambda\end{bmatrix} = 0$$

a tedy

$$(2{,}5 - \lambda)^3 + 2{,}5^3 + 2{,}5^3 - 3.2{,}5^2.(2{,}5 - \lambda) = 0$$

$$\lambda^3 - 7{,}5\lambda^2 = 0$$

jsou $\lambda_1 = 7{,}5$ a dvě násobná $\lambda_{2,3} = 0$.

Protože hodnota vlastního čísla určuje střední kvadratickou chybu vyjádření daného vektoru při odstranění vlastnímu číslu odpovídající souřadnice (dané vlastním vektorem), znamená to, že i když odstraníme souřadnice dané vlastními vektory odpovídajícími vlastním číslům λ_2 a λ_3 a použijeme pouze souřadnici definovanou vlastním vektorem náležejícím číslu λ_1, jsou oba vektory $\mathbf{y}_1$ a $\mathbf{y}_2$ vyjádřeny naprosto přesně.

Z cvičných důvodů ale spočítejme směry všech tří vlastních vektorů $\mathbf{x}_i$, i=1, 2, 3, které určíme ze vztahu

$$[\boldsymbol{\kappa} - \lambda.\mathbf{I}].\mathbf{x} = 0.$$

Pro $\lambda_1 = 7{,}5$ dostáváme lineární soustavu tří rovnic

$$\begin{aligned} -5x_1 + 2{,}5x_2 + 2{,}5x_3 &= 0; \\ 2{,}5x_1 - 5x_2 + 2{,}5x_3 &= 0; \\ 2{,}5x_1 + 2{,}5x_2 - 5x_3 &= 0, \end{aligned}$$

která obsahuje pouze dvě lineárně nezávislé rovnice a tedy její parametrické řešení je

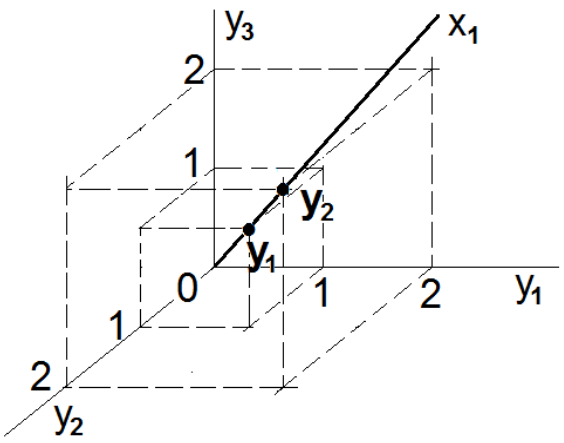
$$x_1 = \frac{x_2 + x_3}{2}; \quad x_2 = x_3 \quad \text{a} \quad x_3 = t.$$

Při volbě parametru t = 1 odpovídá vlastnímu číslu λ_1 vlastní vektor $\mathbf{x_1} = (1, 1, 1)^T$, jak jsme usoudili na základě geometrického rozboru úlohy. Pro vlastní čísla $\lambda_{2,3} = 0$ vypadá definiční soustava rovnic následovně

$$\begin{aligned} 2{,}5x_1 + 2{,}5x_2 + 2{,}5x_3 &= 0 \\ 2{,}5x_1 + 2{,}5x_2 + 2{,}5x_3 &= 0\,. \\ 2{,}5x_1 + 2{,}5x_2 + 2{,}5x_3 &= 0 \end{aligned}$$

To znamená, že dvě rovnice jsou lineárně závislé a její parametrické řešení je

$$x_1 = -x_2 - x_3; \quad x_2 = t \quad \text{a} \quad x_3 = u.$$



Obr.3.7 *Prostorová lokalizace vektorů* $\mathbf{y}_1$ a $\mathbf{y}_2$

Parametry t a u volíme tak, aby vlastní vektory byly navzájem ortogonální, pro $\mathbf{x}_2$ např. t = 1 a u = 1, pak $\mathbf{x}_2 = (-2, 1, 1)^T$ a pro $\mathbf{x}_3$ např. t = -1 a u = 1 a tedy $\mathbf{x}_3 = (0, -1, 1)^T$. V tom případě jsou všechny tři vlastní vektory navzájem ortogonální, každé jejich vzájemné skalární součty jsou rovny nule.

Jak už jsme uvedli dříve, odstraněním souřadnic daných vlastními vektory $\mathbf{x}_2$ a $\mathbf{x}_3$ a ponecháním pouze souřadnice definované vlastním vektorem $\mathbf{x}_1$ se nedopustíme žádné chyby ve vyjádření zadaných vektorů $\mathbf{y}_1$ a $\mathbf{y}_2$ (oba vektory leží na souřadnicové ose dané vektorem $\mathbf{x}_1$ a proto také obě vlastní čísla $\lambda_2 = \lambda_3 = 0$).

Jak by vypadala situace v případě, že bychom odstranili souřadnici $\mathbf{x}_1$? Protože body $\mathbf{y}_1$ a $\mathbf{y}_2$ leží na vrcholech krychlí s hranami o délce 1, resp. 2 protilehlých k počátku (obr.3.7), je jejich vzdálenost od počátku a tím i souřadnice ve směru $\mathbf{x}_1$ rovna délce prostorové úhlopříčky, tj. $d_1 = \sqrt{3}$ v případě vektoru $\mathbf{y}_1$, resp. $d_2 = \sqrt{12}$ v případě vektoru $\mathbf{y}_2$. Protože je nová souřadnicová soustava ortogonální, promítaly by se oba obrazové vektory při odstranění osy x_1 do počátku. A konečně, vzhledem k tomu, že chybu popisu obrazových vektorů ε^2 vyjadřujeme pomocí střední kvadratické odchylky, je tato chyba rovna

$$\varepsilon^2 = \frac{(d_1^2 + d_2^2)}{2} = \frac{1}{2}(3 + 12) = 7{,}5,$$

což je právě hodnota λ_1.

□□□

Metoda výběru příznaků podle Fukunagy a Koontzeho

Karhunenův – Loevův rozvoj umožňuje nalézt optimální popis obrazů s redukovaným počtem souřadnic podle kritéria střední kvadratické odchylky aproximace. Pomocí disperzní matice vede Karhunenova – Loevova transformace k příznakům s největším rozptylem, což

jak jsme naznačili v kap.3.2 není pro klasifikační úlohy zrovna to nejpříznivější řešení. Při transformaci pomocí autokorelační matice je situace příznivější, ale i v tom případě se může stát, že příznaky odpovídající velkým charakteristickým číslům jsou sice vhodné pro optimální reprezentaci dat, nikoliv však pro klasifikaci, protože u všech tříd nabývají téměř stejných hodnot. Říkáme, že takové příznaky mají malou diskriminační schopnost, resp. jsou pro klasifikační úlohu málo informativní.

Tento problém lze řešit např. tím, že se charakteristická čísla disperzní matice uspořádají vzestupně a příznaky se vybírají podle nejmenších vlastních čísel.

Jinou možnou metodou je následující postup publikovaný Fukunagou a Kootzem, který je založen na předpokladu dichotomické klasifikační úlohy. Dichotomický hendikep však lze obejít rozkladem úlohy s obecným počtem klasifikačních tříd na posloupnost dichotomií.

Metoda vychází z normalizace autokorelační matice tak, že platí

$$\boldsymbol{\kappa}(\mathbf{y}') = \mathbf{E}, \tag{3.38}$$

kde **E** je jednotková matice a **y'** reprezentuje normalizovaný obraz, pro který je

$$\mathbf{y}' = \mathbf{U}.\mathbf{y}, \tag{3.39}$$

přičemž **U** je matice normalizační transformace. Pro autokorelační matici $\boldsymbol{\kappa}(\mathbf{y}')$ můžeme psát

$$\boldsymbol{\kappa}(\mathbf{y}') = \frac{1}{K}\sum_{k=1}^{K}\mathbf{y}'_k\,\mathbf{y}'^{T}_k = \frac{1}{K}\sum_{k=1}^{K}\mathbf{U}.\mathbf{y}_k\mathbf{y}_k^T.\mathbf{U}^T = \mathbf{U}.\boldsymbol{\kappa}(\mathbf{y}).\mathbf{U}^T. \tag{3.40}$$

Pomocí (3.40) lze přepsat vztah (3.38) do tvaru

$$\mathbf{U}.\boldsymbol{\kappa}(\mathbf{y}).\mathbf{U}^T = \mathbf{E}. \tag{3.41}$$

Podle definičního vztahu (3.33) je pro dvě klasifikační třídy

$$\boldsymbol{\kappa}(\mathbf{y}) = P(\omega_1)\boldsymbol{\kappa}_{\omega 1}(\mathbf{y}) + P(\omega_2)\boldsymbol{\kappa}_{\omega 2}(\mathbf{y}), \tag{3.42}$$

kde

$$\boldsymbol{\kappa}_{\omega r}(\mathbf{y}) = \int_{\gamma^m}\mathbf{y}.\mathbf{y}^T.p(\mathbf{y}|\omega_r)d\mathbf{y}, \quad r = 1,2; \tag{3.43}$$

je autokorelační matice určená výlučně prvky r-té třídy. Vztah (3.41) pak můžeme dále přepsat do tvaru

$$\mathbf{S}_1 + \mathbf{S}_2 = \mathbf{E}, \tag{3.44}$$

kde

$$\mathbf{S}_r = P(\omega_r).\mathbf{U}.\boldsymbol{\kappa}_{\omega r}(\mathbf{y}).\mathbf{U}^T, \quad r = 1,2. \tag{3.45}$$

Pro vlastní čísla $\lambda_i^{(1)}$ a vlastní vektory $\mathbf{v}_i^{(1)}$ matice $\mathbf{S}_1$ podle definice platí

$$\mathbf{S}_1\mathbf{v}_i^{(1)} = \lambda_i^{(1)}\mathbf{v}_i^{(1)}, \quad i = 1,2,...,m. \tag{3.46}$$

Pro matici $\mathbf{S}_2$ bude obdobně, s využitím (3.44)

$$\mathbf{S}_2\mathbf{v}_i^{(2)} = (\mathbf{E} - \mathbf{S}_1)\mathbf{v}_i^{(2)} = \lambda_i^{(2)}\mathbf{v}_i^{(2)}, \quad i = 1,2,...,m, \tag{3.47}$$

odkud po úpravě dostaneme

$$\mathbf{S}_2\mathbf{v}_i^{(2)} = (1-\lambda_i^{(1)})\mathbf{v}_i^{(2)} \quad i = 1,2,...,m. \tag{3.48}$$

Srovnáním vztahů (3.46) a (3.48) vidíme, že

$$\mathbf{v}_i^{(2)} = \mathbf{v}_i^{(1)}, \quad i = 1,2,...,m \tag{3.49}$$

a

$$\lambda_i^{(1)} = (1-\lambda_i^{(2)}). \tag{3.50}$$

Protože $0 \le \lambda_i^{(r)} \le 1$ pro r = 1, 2 a i = 1,2,…, m, pak jsou-li vlastní čísla matice $\mathbf{S}_1$ uspořádána vzestupně, jsou podle téhož indexu i vlastní čísla matice $\mathbf{S}_2$ uspořádána sestupně. Tedy nejdůležitější příznaky pro popis obrazů z první třídy jsou současně nejméně důležité pro popis obrazů z třídy druhé. Výběr bázového souřadnicového systému provádíme z vektorů $\mathbf{v}_1^{(1)}$, $\mathbf{v}_2^{(1)}$, … pro třídu ω_1 a $\mathbf{v}_m^{(1)}$, $\mathbf{v}_{m-1}^{(1)}$, … pro třídu ω_2.

Zbývá určit matici **U** normalizační transformace. Bez důkazu uvádíme, že

$$\mathbf{U} = \mathbf{U}_1.\mathbf{U}_2, \tag{3.51}$$

kde $\mathbf{U}_1$ představuje matici transformace autokorelační funkce $\boldsymbol{\kappa}(\mathbf{y})$ na matici diagonální $\boldsymbol{\kappa}(\mathbf{U}_1\mathbf{y})$. Uvedenou transformaci lze provést, když

$$\mathbf{U}_1 = \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_m^T \end{bmatrix}, \tag{3.52}$$

kde $\mathbf{v}_i$, i = 1, …, m, jsou charakteristické vektory autokorelační matice $\boldsymbol{\kappa}(\mathbf{y})$. Transformovaná matice má pak tvar

$$\boldsymbol{\kappa}(\mathbf{U}_1\mathbf{y}) = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix}. \tag{3.53}$$

$\mathbf{U}_2$ reprezentuje transformační matici, která převádí diagonální matici podle (3.53) na jednotkovou. To je, když

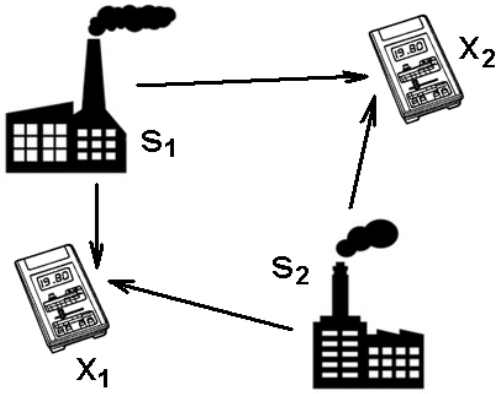
$$\mathbf{U}_2 = \begin{bmatrix} 1/\sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1/\sqrt{\lambda_m} \end{bmatrix}. \tag{3.54}$$

3.3.4 Analýza nezávislých komponent

Začínáme

Analýza nezávislých komponent (ICA – *Independent Component Analysis*) je podobně jako analýza hlavních komponent postup, který umožňuje v původních datech odhalit skryté veličiny, které nelze přímo měřit, ovšem mohou být určitým způsobem věcně interpretovány.

Zatímco analýza hlavních komponent hledá pomocí lineární transformace nové příznakové souřadnice, které nejlépe reprezentují data z hlediska střední kvadratické chyby, metoda analýzy nezávislých komponent používá k lineární separaci jednotlivých složek kritérium statistické nezávislosti. Byť je to metoda, jejímž primárním cílem není, tak jak je to v případě analýzy hlavních komponent, především redukce počtu popisných proměnných, ve svém důsledku, tj. po odhalení nezávislých skrytých zdrojů dat, může vést ke snížení rozměru dat.



Obr.3.8 *Definice metody analýzy nezávislých komponent*

Dále, zatímco metoda hlavních komponent může najít uplatnění při zpracování statických i dynamických, doménou analýzy nezávislých komponent je více zpracování dynamických dat, tj. časových řad. Nicméně, není to jediné možné využití.

Definice problému

Předpokládejme, že v daném prostoru jsou dva nezávislé zdroje znečištění (obr.3.8). Označme veličiny, které je charakterizují s_1 a s_2. Dále předpokládejme, že celková úroveň znečištění je měřena přinejmenším stejným počtem měřicích přístrojů, jejichž výstupy označme x_1 a x_2. V případě, že zanedbáme možné prostorové vlivy (např. dobu šíření znečištění od zdroje k měřícímu zařízení) a nelinearity, můžeme si naměřené veličiny vyjádřit pomocí vztahů

$$\begin{aligned} x_1 &= a_{11}s_1 + a_{12}s_2\,; \\ x_2 &= a_{21}s_1 + a_{22}s_2, \end{aligned} \tag{3.55}$$

kde parametry a_{ij} popisují přenosové vlastnosti prostředí, jímž se znečištění šíří, směrové charakteristiky, apod. Proměnné s_i nazýváme ***skryté***, nebo ***latentní proměnné*** a hodnoty x_i reprezentují ***pozorované veličiny***, které tvoří ***vektor pozorování***. Cílem analýzy je ze známých hodnot x_1 a x_2 určit hodnoty proměnných s_1 a s_2. Pokud bychom znali hodnoty ***transformačních koeficientů*** a_{ij}, pak by řešení uvedené úlohy bylo v podstatě triviální. Avšak problém je, že tyto hodnoty apriori neznáme. Znamená to, že výsledkem výpočtů vycházejících ze znalosti hodnot pozorovaných veličin musí být určení hodnot latentních veličin, ale i hodnot transformačních koeficientů. Takové řešení může vypadat jako naprosto nerealizovatelné a když tak jen v říši snů. Bez jakýchkoliv dalších podmínek by se taková úloha opravdu dost dobře řešit nedala. Je-li to ovšem třeba, zabývejme se podmínkami, za kterých dokážeme řešení nalézt.

Pokusme se proto nyní výše uvedené jednoduché konkrétní zadání úlohy formulovat obecněji. Tedy předpokládejme, že máme k dispozici n-rozměrný náhodný vektor $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, jehož jednotlivé složky představují známá naměřená data. Nechť pro jednotlivé složky x_i vektoru **x** platí

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \ldots + a_{in}s_n;\ i = 1, 2, \ldots, n, \tag{3.56}$$

nebo také pomocí maticového zápisu

$$\mathbf{x} = \mathbf{A}.\mathbf{s}, \tag{3.57}$$

kde **s** reprezentuje vektor původních, formálně skrytých zdrojových komponent a matice **A** je tzv. ***transformační matice***. Hodnoty jejích prvků, stejně jako hodnoty jednotlivých složek vektoru **s** primárně neznáme. Platí-li předpoklad, vyjádřený vztahy (3.56), resp. (3.57), můžeme také psát

$$\mathbf{s} = \mathbf{W}.\mathbf{x} = \mathbf{A}^{-1}.\mathbf{x}, \tag{3.58}$$

což je ten vztah, který umožňuje ze známých hodnot vektoru **x** určit neznámé složky vektoru latentních proměnných. Má-li být tento výpočet realizovatelný, musíme znát hodnoty prvků matice **W**, resp. **A**.

Výpočetní strategie

Pomocí lineární transformace nemůže dojít k navýšení počtu proměnných, tzn. z n naměřených veličin nemůžeme určit více než n zdrojů. Proto, chceme-li odhadnout n zdrojových proměnných, musíme mít k dispozici nejméně n pozorovaných veličin. Budou-li obě matice čtvercové o řádu n (lepší situace pro výpočet inverzní matice) a bude-li existovat pouze $m < n$ zdrojů, pak přiměřeně správný lineární algoritmus nalezne v n pozorovaných veličinách právě m zdrojových proměnných a dalších $n - m$ bude buď nulových, nebo budou obsahovat šumovou složku. Je proto vhodné navrhnout měřicí experiment tak, aby byl počet pozorovaných veličin buď právě roven počtu zdrojů, nebo případně jen o něco málo větší. Přesto, že výpočetně příjemnější je, když jsou obě matice koeficientů čtvercové, je teoreticky možné, v případě, že je naměřených pozorovaných veličin více než zdrojových, aby byly matice koeficientů obdélníkové, v případě matice **A** o rozměru $n \times m$, $n > m$.

Dalším formálním požadavkem, který významně zjednodušuje teoretické zdůvodnění výpočetního postupu i jeho realizaci, je předpoklad o nulové střední hodnotě jak pozorovaných, tak i zdrojových veličin. Pokud tomu tak při řešení praktických úloh není, lze teoretický nedostatek snadno napravit centrováním dat. Je ovšem potřeba si opět uvědomit, že centrováním data přichází o určitou informaci, které se může při následném zpracování nedostávat.

Vzhledem k tomu, jak je úloha zadána, nelze očekávat, že existuje pouze jedno její řešení, nýbrž že bude třeba volit z nekonečně mnoha možných řešení takové, které nejlépe splní určité, vhodně zvolené kritérium optimality. Protože navíc neznáme ani hodnoty skrytých proměnných, je třeba, abychom omezili prostor možných řešení tak, že budeme alespoň předpokládat nějaké jejich určité vlastnosti, které usnadní nalezení řešení.

Zásadním požadavkem na vlastnosti zdrojů a tím i latentních veličin, který dal i název algoritmu, je požadavek na jejich statistickou nezávislost. To značí, že hodnota žádné z latentních veličin neposkytuje informaci o hodnotách dalších latentních veličin. V případě vzájemné statistické nezávislosti náhodných veličin $x_1, x_2, \ldots, x_n$ platí

$$p(x_1, x_2, \ldots, x_n) = p_1(x_1).\, p_2(x_2) \ldots p_n(x_n), \tag{3.59}$$

kde $p(x_1, x_2, \ldots, x_n)$ je sdružená funkce rozložení hustoty pravděpodobnosti a $p_i(x_i)$ jsou marginální funkce rozložení hustoty pravděpodobnosti.

Pro nezávislé náhodné veličiny s určitými pravděpodobnostními rozděleními říká centrální limitní věta, že jejich součet konverguje za poměrně obecných podmínek s rostoucím počtem náhodných veličin ke Gaussovu normálnímu rozdělení bez ohledu na to, jaké je rozdělení jednotlivých náhodných veličin. Dle zadání metody nezávislých komponent jsou pozorované veličiny dány váhovaným součtem latentních proměnných. Z toho plyne, že jednotlivé pozorované veličiny x_i budou mít rozdělení o něco normálnější, než jsou rozdělení jednotlivých zdrojových komponent. Na této skutečnosti je pak založena kriteriální funkce pro optimalizační výpočet zdrojových veličin, která předpokládá, že pro jednotlivé zdrojové veličiny podle (3.58) platí $s_i = \mathbf{w}_i.\mathbf{x}$. Tedy hledáme koeficienty transformační matice **W** takové, aby pravděpodobnostní rozdělení vypočítaných zdrojových veličin bylo co nejméně normální. Aby tato myšlenka byla realizovatelná, může mít normální rozdělení maximálně jedna skrytá náhodná veličina, ostatní musí mít jiné než normální rozdělení. V současné době již existují i jiná kritéria, jak určit nezávislé nebo alespoň co nejméně závislé zdrojové veličiny, v následujícím textu se ale budeme zabývat jen tímto základním principem, založeným na centrální limitní větě.

Máme-li formalizovat výpočet kriteriální funkce, musíme toto uvedené kritérium vyjádřit matematicky. Nejčastěji používané míry statistické anormality v analýze nezávislých komponent jsou:

- koeficient špičatosti;
- negativní entropie.

Zabývejme se nyní jednotlivými mírami.

Koeficient špičatosti (angl. *kurtosis*) je klasickou mírou statistické anormality a jako kumulant 4. řádu je pro náhodnou veličinu s, za předpokladu nulové střední hodnoty, definován vztahem

$$\text{kurt}(s) = E\{s^4\} - 3(E\{s^2\})^2. \tag{3.60}$$

Protože druhý člen na pravé straně výrazu reprezentuje rozptyl náhodné veličiny s, zjednodušuje se definiční výraz pro data standardizovaná vůči směrodatné odchylce na

$$\text{kurt}(s) = E\{s^4\} - 3. \tag{3.61}$$

To znamená, že koeficient špičatosti je v podstatě dán čtvrtým momentem náhodné veličiny. Pro náhodné veličiny s normálním rozdělením je koeficient špičatosti roven nule. Pro většinu negaussovských náhodných veličin (ale ne pro všechny, což může být považováno za nevýhodu, protože tato rozdělení jsou algoritmem analýzy formálně vyloučena, protože jejich charakteristika je číselně rovna charakteristice normálního rozdělení) je různý od nuly. Může být kladný i záporný, proto se za typickou míru statistické anormality používá jeho absolutní hodnota, resp. jeho druhá mocnina.

Výhodou použití koeficientu špičatosti pro odhad zdrojových komponent je jeho relativně jednoduchý a tím i rychlý výpočet, teoretické zázemí jeho použití je rovněž příjemně jednoduché díky jeho linearitě. Platí totiž, že

$$\text{kurt}(s_1 + s_2) = \text{kurt}(s_1) + \text{kurt}(s_2) \tag{3.62}$$

a

$$\text{kurt}(\alpha.s) = \alpha^4.\text{kurt}(s)\ . \tag{3.63}$$

kde α je konstanta.

Odhad skryté zdrojové veličiny pak probíhá tak, že hledáme takové koeficienty transformačního vektoru $\mathbf{w}_i$, pro které má koeficient špičatosti veličiny $s = \mathbf{w}_i.\mathbf{x}$ maximální hodnotu. Způsob hledání extrému závisí jednak na použitém kritériu a tím i na vlastnostech a tvaru kriteriální funkce. V případě koeficientu špičatosti lze vystačit s gradientní či Newtonovou metodou.

Nevýhodou použití koeficientu špičatosti je, kromě již zmíněné diskriminace několika málo nenormálních rozdělení s nulovým koeficientem špičatosti, poměrně malá robustnost vůči odlehlým hodnotám pozorovaných veličin. To jest, pokud měření obsahují hodnoty výrazně se odlišující od běžných, potom i zdrojové veličiny budou pravděpodobně odhadnuty chybně.

Negativní entropie (***negentropie***) je parametr, vycházející z jednoho ze základních principů teoretické fyziky a Shannonovy teorie informace, tj. principu ***entropie***[20] a její míry.

Obecně pro systém S s konečným počtem možných stavů $s_1, s_2, \ldots, s_n$ a pravděpodobnostní distribucí $P(s_i)$ je informační entropie definována jako střední hodnota

$$H(S) = -\sum_{i=1}^{n} P(s_i).\log P(s_i) \tag{3.64}$$

[20] Entropie (z řec. *εντροπία*; *εν*- "k" + *τροπή* "směrem").tedy "směrem k".

(formálně pro $P(s_i) = 0$ definujeme $P(s_i).\log P(s_i) = 0$). Základ použitého logaritmu je zpravidla roven 2, v tom případě velikost entropie udáváme v bitech. Definujeme-li na intervalu $p \in \langle 0, 1\rangle$ funkci $f(p) = -p.\log p$, tato funkce nabývá nulové hodnoty v krajních bodech intervalu a jednoho maxima uvnitř definičního intervalu. Pro dvojkový logaritmus by se maximum vyskytovalo na $p = 0{,}368$.

Z informatického hlediska vnímáme entropii jako míru neurčitosti systému. Pro úzká, příp. ostrá rozdělení pravděpodobnosti je entropie nízká, naopak široká či neostrá rozdělení pravděpodobnosti mají entropii vysokou. Entropie je maximální pro rovnoměrné rozdělení, tj. pro

$$P(s_i) = \frac{1}{n} \text{ pro } \forall i\,. \tag{3.65}$$

V tom případě je

$$H(S) = -\sum_{i=1}^{n} \frac{1}{n}.\log\frac{1}{n} = -\log\frac{1}{n} = \log n\,. \tag{3.66}$$

Minimální entropie pro deterministický systém s pravděpodobnostní distribucí $P(s_k) = 1$ pro nějaké k a $P(s_k) = 0$ pro $\forall i \neq k$. Tehdy je

$$H(S) = -\log 1 = 0\,. \tag{3.67}$$

I na základě těchto konkrétních výsledků můžeme entropii interpretovat jako míru informace, kterou poskytuje daná hodnota měřené veličiny. Čím náhodnější, tj. čím méně očekávaná, či determinovaná je daná proměnná, tím je její entropie větší.

Entropie definovaná pro diskrétní náhodnou proměnnou může být zobecněna pro spojitý případ, kdy se ale spíše vžilo označení ***diferenciální entropie***. Pro náhodnou proměnnou s hustotou rozdělení pravděpodobnosti p(X) je diferenciální entropie určena vztahem

$$H(s) = -\int p_s(\xi).\log p_s(\xi)d\xi \text{ nebo obecněji } H(s) = -\int f[p_s(\xi)].d\xi\,. \tag{3.68}$$

Který může být dále zobecněn i pro vícerozměrnou proměnnou

$$H(\mathbf{s}) = -\int p_s(\xi).\log p_s(\xi)d\xi \text{ nebo obecněji } H(\mathbf{s}) = -\int f[p_s(\xi)].d\xi\,. \tag{3.69}$$

Diferenciální entropie má podobné vlastnosti jako entropie, také může být interpretována jako míra náhodnosti. Čím jsou hodnoty proměnné soustředěny v širším intervalu a čím je jejich pravděpodobnost rovnoměrnější, tím je diferenciální entropie větší. Dále platí, že entropie Gaussova normálního rozdělení má největší hodnotu ze všech rozdělení pravděpodobnosti s týmž rozptylem.

Protože hodnota diferenciální entropie pro normální rozložení není pro konkrétní případ předem známa, je hodnota kriteriální funkce použitá pro určení optimálních hodnot transformační matice definována rozdílem odhadnuté negentropie normálního rozložení

$$J(\mathbf{s}) = H(\mathbf{s}_{gauss}) - H(\mathbf{s})\,, \tag{3.70}$$

kde $\mathbf{s}_{gauss}$ je vektor hodnot náhodné veličiny s normálním rozdělením a stejným rozptylem (a tím také stejnou kovarianční maticí $\boldsymbol{\Sigma}$) jako náhodný vektor $\mathbf{s}$ a kde ze známé kovarianční matice $\boldsymbol{\Sigma}$ odhadujeme entropii normálního rozdělení podle vztahu

$$H(\mathbf{s}_{gauss}) = \frac{1}{2}\log|\det\boldsymbol{\Sigma}| + \frac{n}{2}[1 + \log 2\pi]\,, \tag{3.71}$$

kde n je rozměr vektoru $\mathbf{s}_{gauss}$. Proměnnou $J(\mathbf{s})$ zde nazýváme ***negentropií***. Nastavení transformační matice hledáme tak, aby hodnota $J(\mathbf{s})$ byla co největší. Negentropie je vždy nezápor-

ná a je rovna nule pouze tehdy a jen tehdy, pokud má **x** normální rozdělení. Dále, negentropie je invariantní vůči změně měřítka náhodné proměnné, tj. vynásobíme-li hodnoty náhodné proměnné konstantou, její negentropie se nemění.

Mezi výhody negentropie patří její jednoznačné hodnocení normality, resp. anormality. Negentropie je ve srovnání s koeficientem šikmosti odolnější vůči odlehlým hodnotám. Na druhé straně je negentropie obtížně vyčíslitelná, protože její výpočet vyžaduje vyčíslit definiční integrál podle vztahu (3.68), nebo dokonce (3.69) pro rozdělení pravděpodobnosti odpovídající zpracovávaným datům. Hodnota integrálu může být teoreticky určena analyticky (pokud to umíme) z funkčního vyjádření hustoty pravděpodobnosti, které stanovíme nějakým parametrickým odhadem. Tento způsob je ale z velké části závislý na apriorní informaci o charakteru rozložení dat, nehledě na skutečnost, že analytické výpočty nad experimentálními daty nejsou organizačně příliš praktické. Alternativním způsobem výpočtu integrálu může být numerický odhad jeho hodnoty, založený na neparametrickém odhadu hustoty pravděpodobnosti. Ať ten, či onen způsob výpočtu hodnoty entropického integrálu je zpravidla zatížen takovou chybou, která z konceptu negentropie dělá spíše teoretickou disciplínu, než praktický návod ke zpracování experimentálních dat. Zmíněné problémy proto vedou k hledání dalších způsobů, jak v daných konkrétních případech prakticky realizovat odhad negentropie.

Již klasickým postupem je aproximace negentropie pomocí kumulantů vyššího řádu vztahem (s jednorozměrnou náhodnou veličinou)

$$J(s) \approx \frac{1}{12} E\{s^3\}^2 + \frac{1}{48} \mathrm{kurt}(s)^2, \tag{3.72}$$

který lze z definičního vztahu negentropie odvodit pro data standardizovaná na směrodatnou odchylku pomocí polynomiálního rozvoje hustoty pravděpodobnosti. Nicméně pro symetrická rozdělení je první člen ve výrazu na pravé straně vztahu (3.72) nulový, což v důsledku znamená, že se pro data tohoto typu vracíme k hodnocení anormality pomocí koeficientu špičatosti se všemi negativy i pozitivy tohoto přístupu.

Případnou alternativou je zobecněná aproximace pomocí kumulantů vyššího řádu, která nahrazuje polynomiální funkce s^3, příp. s^4 jinými funkcemi G_i. Za předpokladu, že G_i jsou nekvadratické a G_1 je lichá a G_2 sudá, je možné odvodit obecný aproximační vztah

$$J(s) \approx k_1 \left(E\{G_1(s)\}\right)^2 + k_2 \left(E\{G_2(s)\} - E\{G_2(\upsilon)\}\right)^2, \tag{3.73}$$

kde k_1, $k_2 > 0$ jsou váhové konstanty vyjadřující vliv obou členů a υ je standardizovaná náhodná proměnná s normálním rozdělením, příp. pro jedinou nekvadratickou funkci G je

$$J(s) \approx \left(E\{G(s)\} - E\{G(\upsilon)\}\right)^2, \tag{3.74}$$

což je zobecněním momentové aproximace podle vztahu (3.72) za předpokladu, že náhodná proměnná má symetrické rozložení.

Až dosud jsme uvažovali poměrně obecný tvar funkcí G_i. Pro konkrétní realizaci je ale třeba pracovat s konkrétní funkcí. Praxe ukázala, že pro robustnost odhadu je užitečné pokud funkce G_i nejsou příliš rychle rostoucí. Experimentálně se prokázaly užitečné vlastnosti zejména funkcí

$$G_1(s) = \frac{1}{a_1} \log \cosh a_1 s \tag{3.75}$$

a

$$G_2(s) = -\exp\left(-\frac{s^2}{2}\right), \tag{3.76}$$

kde $a_1 \in \langle 1, 2 \rangle$ je konstanta, většinou se volí $a_1 = 1$.

Omezení metody

V dřívějším textu jsme uvedli některé vstupní předpoklady, za kterých lze analýzu nezávislých komponent provést (statistická nezávislost zdrojových veličin, maximálně jedna náhodná zdrojová veličina s normálním rozdělením, nulová střední hodnota, …), metoda má ale některé omezující důsledky, které vyplývají z teorie, která většinou přesahuje rámec tohoto textu, proto si je zde uveďme pouze víceméně bez důkazů.

Uvádí se, že metoda nezávislých komponent neumí stanovit nezávislé komponenty v originálním pořadí. V podstatě to znamená, že metoda neumí fixně stanovit pořadí jednotlivých členů v definičním vztahu (3.56). To je omezení ve většině případů formální, protože při řešení mnoha reálných úloh není pořadí proměnných důležité, ale v některých úlohách (např. snaha o odstranění artefaktů ze signálů EEG) může působit komplikace.

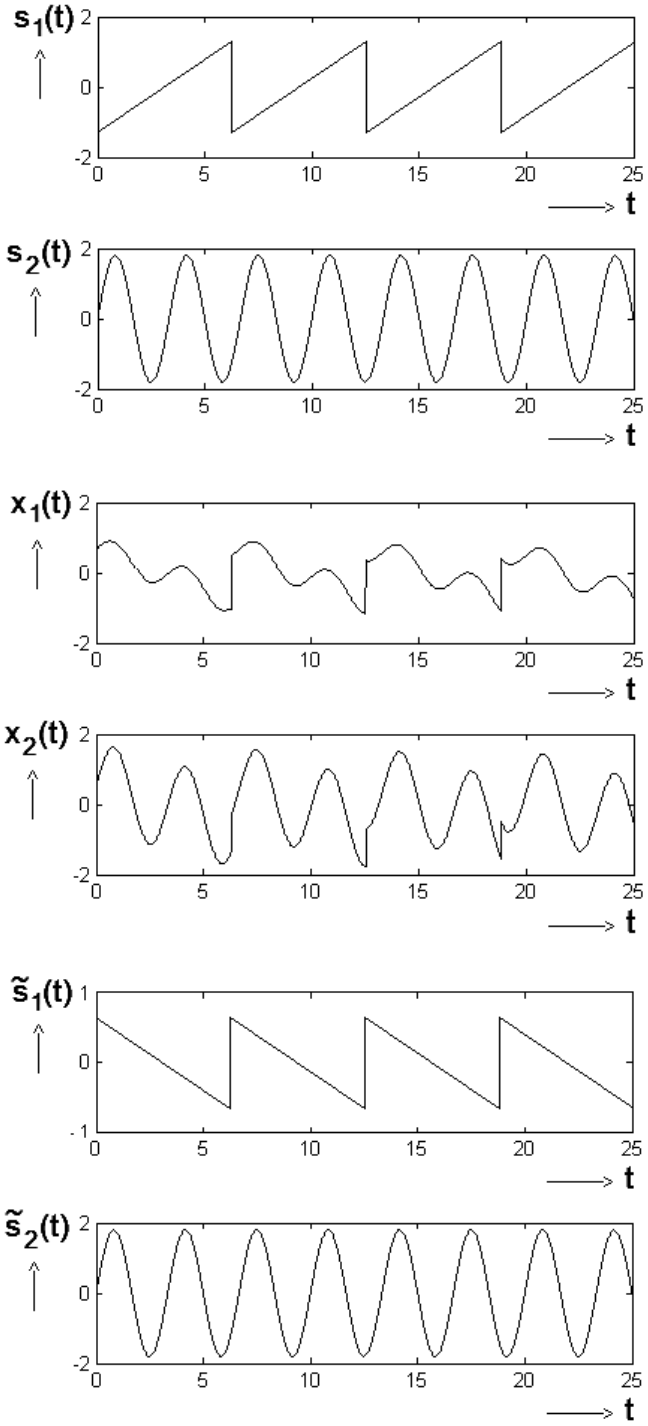
Metodou nezávislých komponent není obecně možné stanovit rozptyl originálních proměnných. To znamená, že při nulové střední hodnotě mohou být odhadnuté zdrojové veličiny vůči skutečnému stavu vynásobeny libovolnou konstantou. Pokud předpokládáme jednotkový rozptyl, jsou všechny odhady standardní.

Protože konstanta, kterou vynásobíme odhad zdrojové veličiny, může být i záporná, znamená to, že hodnoty zdrojových veličin mohou měnit i polaritu.

Příklad

Na obr.3.9 je ilustrována situace separace průběhů zdrojových veličin $s_1(t)$ a $s_2(t)$ pomocí analýzy nezávislých komponent. Průběhy známých, tzv. naměřených veličin $x_1(t)$ a $x_2(t)$ jsou dány lineární kombinací zdrojových veličin $s_1(t)$ a $s_2(t)$. V dolní části obrázku jsou uvedeny odhadnuté průběhy $\tilde{s}_1(t)$ a $\tilde{s}_2(t)$. Lze si všimnout, že zatímco $\tilde{s}_2(t)$ odpovídá zdrojovému průběhu, funkce $\tilde{s}_1(t)$ má jednak jinou velikost, jednak i jinou polaritu.

□□□



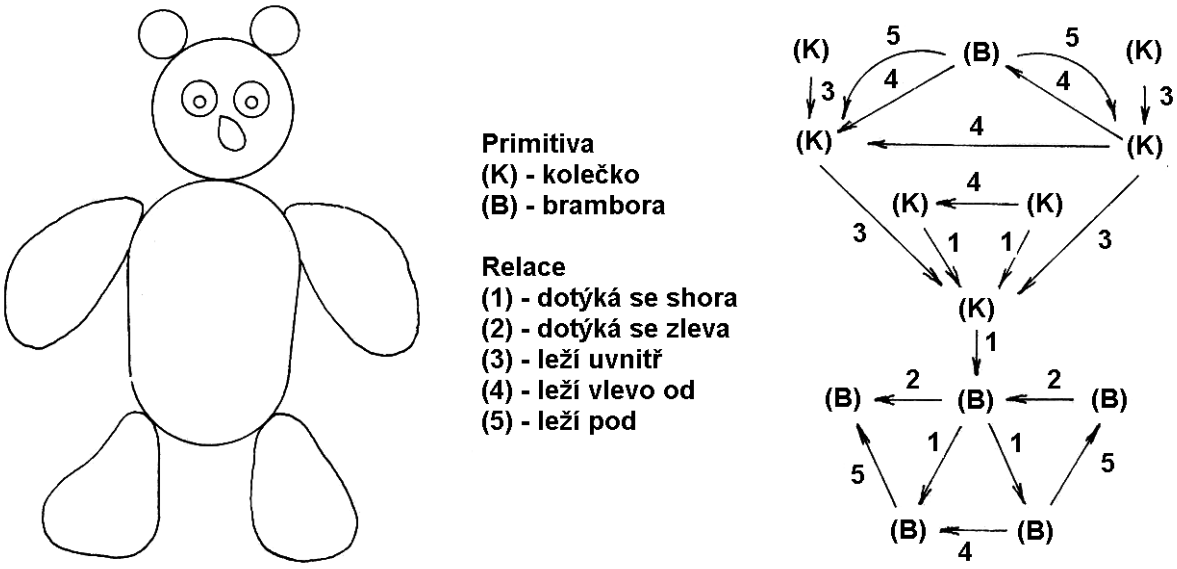
Obr.3.9 *Příklad rekonstrukce zdrojových veličin metodou nezávislých komponent*

4 Strukturální metody analýzy a klasifikace dat

4.1 Základní pojmy a principy

4.1.1 Primitiva, relace, relační struktura

Až dosud jsme se zabývali metodami analýzy a klasifikace dat, vyjádřených pomocí příznakového vektoru. V některých případech ale nestačí pouze znalost charakteristik elementárních vlastností klasifikovaných objektů vyjádřených hodnotami příznaků, ale je třeba znát i strukturu objektů a vzájemné souvislosti mezi jednotlivými elementy objektů, příp. mezi jednotlivými vlastnostmi, což pomocí příznakového popisu buď vůbec není možné, nebo je zapotřebí příliš velkého množství příznaků a výpočetního úsilí. Tento případ nastává např. při zpracování obrazů, ale i při analýze sekvencí DNA či při zpracování řetězců popisujících určité identifikační údaje – např. jméno pacienta. V těchto situacích se ukazuje výhodnější popsat zpracovávané jevy, procesy, objekty pomocí relační struktury, vytvořené z určitých elementárních popisných částí analyzovaného objektu, tzv. ***primitiv*** a vzájemných vztahů mezi nimi – tzv. ***relacemi***. Relační struktury popisující analyzovaný objekt vyjadřujeme názorně pomocí grafu (obr.4.1).



Obr.4.1 *Primitiva, relace a relační struktura čárové kresby*

Primitiva mohou obecně nést dva druhy informace. Bezpodmínečně musí obsahovat informaci ***strukturální***, která definuje jejich charakter a mnemotechnicky je vyjádřena jejich identifikátorem. Strukturální informace může být doplněna informací ***sémantickou***, podrobněji specifikující, kvantitativně nebo i kvalitativně, dílčí vlastnosti primitiva. Primitivum α tedy lze charakterizovat jako dvojici $\alpha = (s,\mathbf{x})$, kde s je název primitiva, reprezentující strukturální (syntaktickou) informaci a $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ je sémantický vektor s n numerickými, resp. kvalitativními atributy primitiva. Vektor atributů primitiva je v podstatě totéž jako vektor příznaků, jak jsme se s ním seznámili v předcházejících kapitolách. Neobsahuje-li vektor **x** žádnou položku, hovoříme o primitivu bez sémantické informace.

Určení vhodných primitiv je první etapou sestavení strukturálního modelu objektu. Obecné řešení této úlohy, podobně jako v případě volby a výběru příznaků popsaných v kap.3, neexis-

tuje. Záleží proto na konkrétních vlastnostech analyzovaných dat, aplikační oblasti a v neposlední řadě i na dostupných technických a algoritmických prostředcích pro detekci primitiv v datech.

Podobně relace, reprezentující vztahy mezi primitivy, lze definovat jako dvojici r = (u,**y**), kde u je opět název relace vystihující její podstatu a **y** = (y_1, y_2, …, y_j) je sémantický vektor atributů relace r. Podobně jako u primitiv, je-li vektor **y** prázdný, hovoříme o relaci bez sémantické informace.

Relace používané v relačních strukturách mohou být obecně k-ární, to znamená, že mohou vyjadřovat vzájemný vztah mezi k primitivy. Ovšem každou k-ární relaci je možné převést (rekurzivní dekompozicí, konjunkcí, …) na relace binární, tj. relace vyjadřující vztah mezi dvěma primitivy.

Podle rekurzivní dekompozice lze nahradit k-ární relaci R (k > 2) relacemi binárními podle předpisu

$$R(X_1, X_2, \ldots, X_k) = R_1(X_1, R_2(X_2, \ldots, R_{k-1}(X_{k-1},X_k)\ldots)) \tag{4.1}$$

a pomocí konjunkce podle vztahu

$$R(X_1, X_2, \ldots, X_k) = R_{12}(X_1, X_2) \wedge R_{13}(X_1, X_3) \wedge \ldots R_{k-1,k}(X_{k-1},X_k). \tag{4.2}$$

Vzhledem k této skutečnosti a dále vzhledem k možnosti příjemně a názorně reprezentovat relační struktury pomocí binárních relací označenými grafy se v úlohách strukturálního rozpoznávání dominantně užívá především relací binárních.

Významné postavení mezi binárními relacemi má ***relace ostrého úplného uspořádání***, pro kterou platí, že:

(1) pro každé dva prvky X a Y je buď $\langle X,Y\rangle \in R$ nebo $\langle Y, X\rangle \in R$;
(2) pro žádný prvek X neplatí $\langle X,X\rangle \in R$;
(3) když pro tři prvky X, Y, Z je $\langle X,Y\rangle \in R$ a současně $\langle Y,Z\rangle \in R$, pak musí být i $\langle X,Z\rangle \in R$.

Relaci $\langle X,Y\rangle$, která splňuje vlastnosti úplného ostrého uspořádání, interpretujeme místně jako „X je vlevo od Y“, nebo časově „X předchází Y“.

Relační strukturu složenou z n prvků $X_1, X_2, \ldots, X_n$, přičemž mezi každými dvěma prvky lze definovat právě jen relaci ostrého úplného uspořádání, nazýváme ***řetězec***. Popis řetězcem – přirozený pro popis sekvenčních struktur, např. řetězec bází DNA - se díky vytvořenému teoretickému i algoritmickému zázemí velice často používá i v těch případech, kdy tomu základní představa neodpovídá – např. pro rozpoznávání dvojdimenzionálních elementů v obraze.

Použití obou složek informace, strukturální i sémantické, vede k tomu, že algoritmy pro strukturální analýzu a klasifikaci mají obecně dvě fáze zpracování, které mohou po sobě následovat (nejdříve strukturální, posléze sémantická), nebo což je častější případ, se navzájem prolínají.

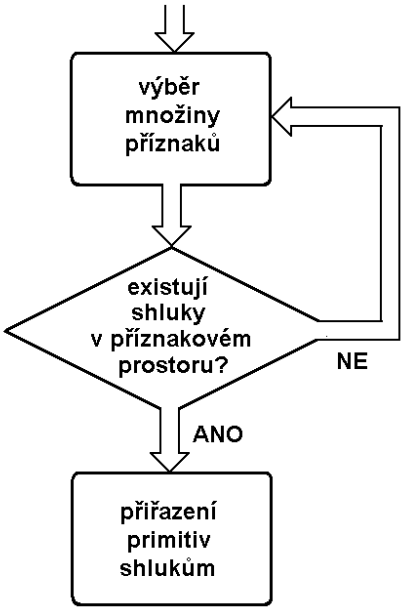
Při volbě primitiv je třeba v zásadě vyhovět třem základním požadavkům:

- primitiva musí být z hlediska řešené úlohy základními prvky struktury analyzovaného předmětu;
- zvolená primitiva a relace musí zajistit přiměřený popis dat, tj. popis vyjadřující kompromis mezi požadavky na jednoduchost primitiv na jedné straně a jednoduchost vyjádření klasifikačních tříd na straně druhé;
- je třeba, aby primitiva i relace bylo možné nalézt ve vstupních datech co nejjednodušším způsobem.

Strukturální popis lze oproti popisu příznakovému považovat za vyšší kvalitu. Z větší složitosti, která je důsledkem této změny kvality, plyne, že metody volby primitiv jsou podstatně méně formalizovatelné, než to bylo v případě volby a výběru příznaků.

Nejobvyklejší způsob algoritmizace volby primitiv je použití příznakové shlukové analýzy, pomocí které se v datech učební množiny hledají příznakové veličiny, umožňující charakterizovat dílčí segmenty dat jako dobře rozlišitelné shluky v použitém příznakovém obrazovém prostoru. Takto nalezené shluky jsou pak považovány za primitiva. Princip tohoto přístupu je znázorněn schématem na obr.4.2.

V převážné většině případů se strukturální metody rozpoznávání používají pro řešení problémů s názorným plošným či prostorovým popisem, z jejichž zadání buď přímo, či po nevelkém experimentování, vyplývá možný způsob rozkladu dat na jejich elementární části. Způsob volby popisu v takovém případě závisí na charakteru dat, zkušenosti řešitele a úspěšnosti heuristického experimentování.

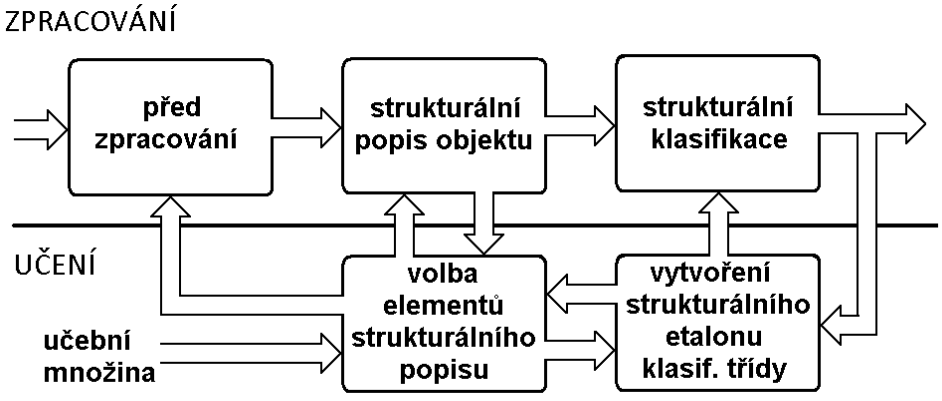


Obr.4.2 *Výběr primitiv shlukovou analýzou*

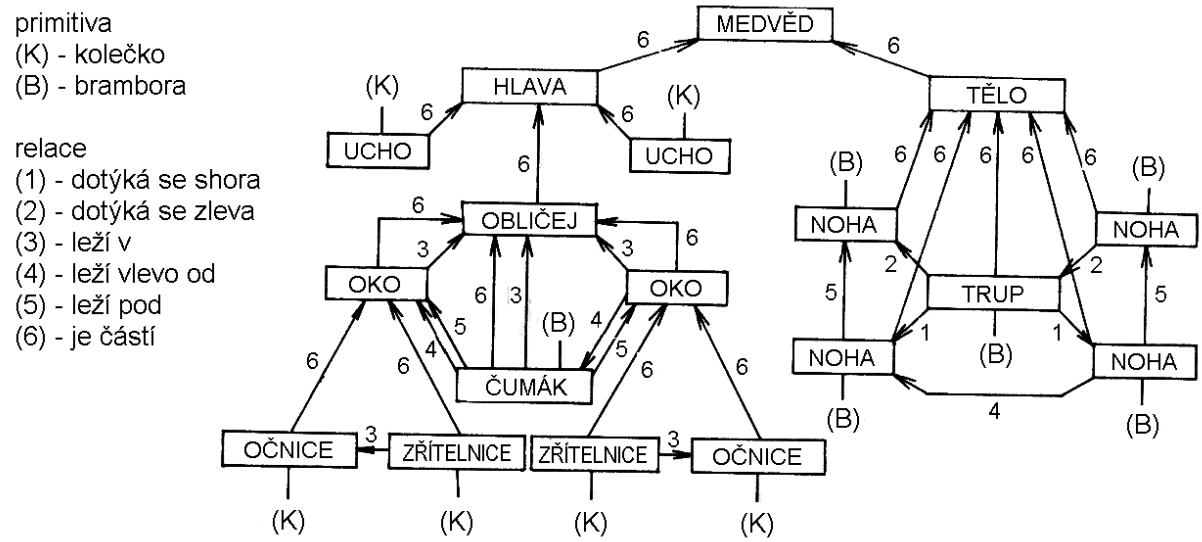
4.1.2 Blokové schéma strukturálního zpracování dat

Blokové schéma strukturálního zpracování dat bez využití sémantické informace je na obr.4.3. V principu schéma zůstává stejné jak jsme se sním seznámili v kap.1.3, jen náplň jednotlivých bloků se přizpůsobila potřebám strukturálního zpracování. První blok v učební fázi reprezentuje volbu a výběr primitiv a relací vhodných pro popis daného typu objektu. Výběr primitiv je primární, relace jsou pro danou úlohu určeny zvolenými primitivy. Klasifikační třída je v případě strukturálního klasifikátoru dána množinou všech relačních struktur požadovaných vlastností, případně s jejich povoleným chybovým okolím. Protože tato množina může být poměrně rozsáhlá, byly vytvořeny prostředky pro kompaktní matematický popis strukturální klasifikační třídy – ***gramatiky*** a ***automaty***. Návrh (inference) gramatik, resp. automatů na základě relačních struktur učební množiny je náplní druhého bloku učební fáze.

V oblasti zpracování, v bloku předzpracování zůstávají cíle činnosti tytéž jako v případě příznakových klasifikátorů – odstranění parazitních složek, zdůraznění užitečné komponenty dat, redukce redundantních složek informace, …, atd. Blok strukturálního popisu signálů vytváří na základě zvolených primitiv a relací strukturální obraz objektu, tj. relační strukturu. Klasifikátor pak rozhodne o zařazení relační struktury.



Obr.4.3 *Blokové schéma strukturálního klasifikátoru bez sémantické informace*

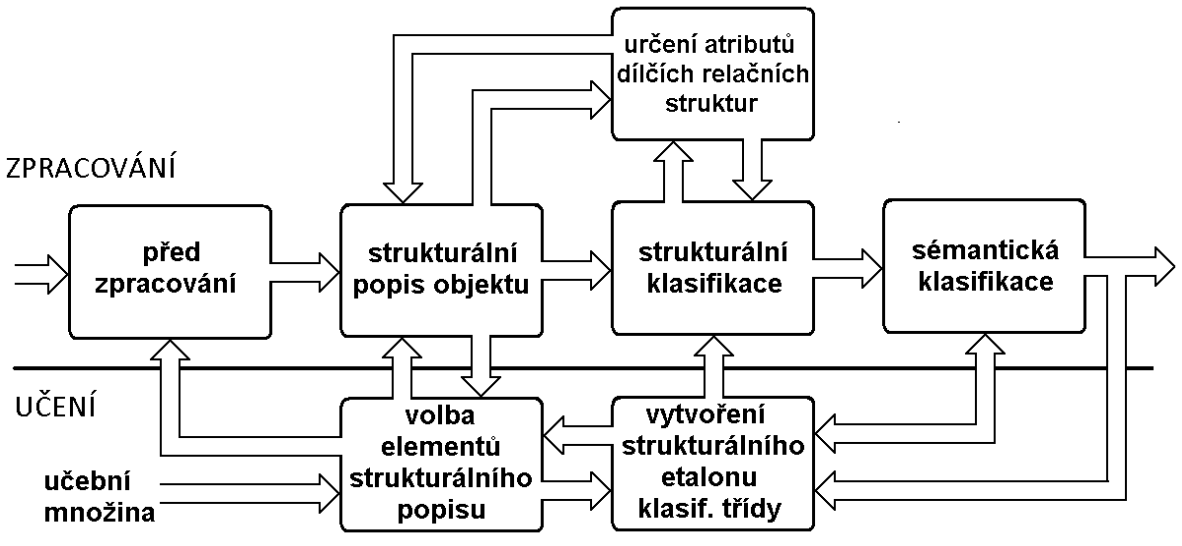


Obr.4.4 *Hierarchická relační struktura kresby z obr.4.1*

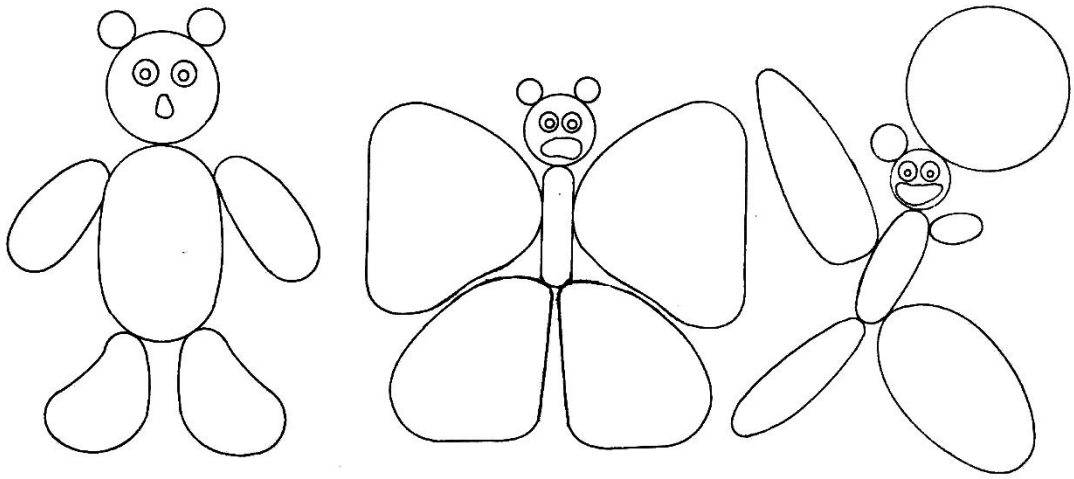
Při strukturální klasifikaci je často výhodné, když relační struktura neobsahuje pouze primitiva spojená patřičnými relacemi (vše na jedné hierarchické úrovni), nýbrž když relační struktura vyjadřuje i hierarchii skladby objektu vyznačením dílčích částí relační struktury jako mezistupňů mezi primitivy, coby reprezentanty elementů objektu a celou relační strukturou (obr.4.4). Podobně jako primitiva a relace, i zmíněné dílčí části struktury mohou být popsány vektorem atributů, které lze dílem vypočítat z atributů primitiv, relací a dílčích struktur, určených na nižší hierarchické úrovni, dílem musí být určeny z reálného objektu. Pokud by byl průměr atributem primitiv (K), ze kterých je, kromě jiného, obraz medvěda složen, můžeme spočítat velikost duhovky oka, ovšem její barvu, v případě barevné kresby, by bylo třeba zjistit z originálu.

Po zavedení hierarchické relační struktury se sémantickou informací můžeme přikročit k popisu blokového schématu strukturálně sémantické (kombinované) analýzy a klasifikace objektů (obr.4.5). V učební fázi je činnost obou bloků spojením činnosti odpovídajících bloků strukturálního i příznakového přístupu. Ve fázi zpracování zůstává i nadále táž činnost bloku předzpracování, ke změnám ale dochází v dalších fázích zpracování.

V bloku popisu signálu se nejprve vytvoří nehierarchická relační struktura, včetně vektorů



Obr.4.5 *Blokové schéma atributového strukturálního klasifikátoru*



Obr.4.6 *Čárové kresby se stejnou relační strukturou bez sémantické informace*

atributů primitiv i relací, která se podrobí v následujícím bloku strukturální klasifikaci. Strukturální klasifikace může být řízena informací o hierarchii relační struktury. Pokud se během klasifikace vyskytne potřeba manipulovat i s některými dílčími strukturami, určují se vektory jejich atributů buď výpočty z atributů již dříve stanovených, nebo opět v bloku popisu signálu z originálních dat. Po skončení strukturální klasifikace je výsledek dále upřesněn v bloku sémantické klasifikace. Sémantický klasifikátor může při vhodné volbě atributů přispět např. k rozlišení tří kreseb podle obr.4.6, jejichž prosté relační struktury bez sémantické informace jsou stejné.

Použití sémantické informace má ledajaké další praktické výhody. Zavedením sémantického vektoru lze snížit počet primitiv, resp. relací, potřebných k popisu klasifikovaného objektu, pokud lze vyjádřit rozdíly mezi nimi i pomocí sémantických atributů. Snížení počtu použitých typů primitiv a relací vede zpravidla ke zjednodušení vytvářených relačních struktur a tím se sníží i výpočetní pracnost strukturální klasifikace.

Sémantické informace lze použít i k řízení strukturálního klasifikátoru, nebo naopak strukturální analýzy pro řízení příznakové klasifikace dílčích segmentů celkové relační struktury. Na základě sémantické informace lze také rozšířit možnosti analýzy šumových složek ve strukturálním popisu dat.

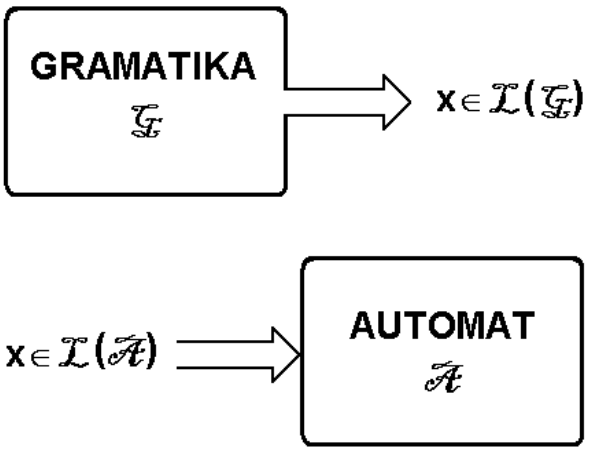
Na druhé straně, zavedením sémantické informace v případech, kdy není nutná (tato nutnost samozřejmě nemusí být na první pohled zřejmá), se zvyšuje složitost algoritmů zpracování, zvyšuje se výpočetní pracnost a tak i doba zpracování.

4.2 Popis klasifikační třídy

4.2.1 Poznámky na úvod

Strukturální etalon klasifikační třídy je reprezentován množinou relačních struktur, popisujících všechny objekty patřící do této třídy. V případě většího počtu těchto relačních struktur je efektivnější k reprezentaci třídy využít, podle účelu reprezentace, buď generátor, který na základě stanovených pravidel vytváří relační struktury, patřící právě jen do stanovené klasifikační třídy, nebo přijímač, který je na podobných principech schopen přijmout právě jen relační struktury dané klasifikační třídy (obr.4.7).

Mnohé prostředky a postupy používané v oblasti strukturálního rozpoznávání jsou inspirovány nástroji a algoritmy vytvořenými v oblasti ***teorie formálních jazyků***. Podle terminologie této teorie rozumíme formálním jazykem množinu slov (řetězců) určitých specifikovaných vlastností, přičemž ***slovem*** formálního jazyka, vytvořeným nad danou ***abecedou*** (neprázdnou množinou prvků nazývaných ***symboly abecedy***), rozumíme obecně každou konečnou posloupnost symbolů abecedy.



Obr.4.7 *Strukturální etalon klasifikační třídy*

Porovnáme-li uvedenou definici formálního jazyka a definici strukturální klasifikační třídy, zjistíme, že pojem strukturální klasifikační třídy je obecnější jen co do větší obecnosti prvků, nad kterými je klasifikační třída vytvořena. Pokud bychom se omezili jen na práci s řetězcovými relačními strukturami, pak jsou oba pojmy ekvivalentní a z toho plyne, že vše, co bylo v problematice formálních jazyků vytvořeno, lze použít i pro potřeby strukturálního rozpoznávání[21] V souladu s terminologií teorie formálních jazyků nazýváme generátor relačních struktur z dané klasifikační třídy ***gramatika*** a přijímač relačních struktur ***automat***.

4.2.2 Gramatiky

Definice gramatiky

Gramatika $\mathcal{G}$ je čtveřice $\mathcal{G} = (\mathcal{V}_n, \mathcal{V}_t, \mathcal{P}, S)$, kde $\mathcal{V}_n$ a $\mathcal{V}_t$ jsou konečné disjunktní abecedy, přičemž prvky $\mathcal{V}_n$ se nazývají ***neterminální*** (pomocné) ***symboly*** a prvky $\mathcal{V}_t$ ***terminální symboly***, $S \in \mathcal{V}_t$ je tzv. ***axiom gramatiky*** nebo také ***počáteční symbol*** a $\mathcal{P}$ je ***množina substitučních pravidel*** tvaru $\alpha \rightarrow \beta$, které definují způsob náhrady dílčí relační struktury α novou relační strukturou β.

Množinu všech relačních struktur generovaných danou gramatikou nazýváme jazykem $\mathcal{L}(\mathcal{G})$ dané gramatiky. Jazyk je tedy jednou z veličin, které gramatiku charakterizují. Gramatiky, které generují týž jazyk, se nazývají ekvivalentní.

Gramatiky můžeme dělit podle následujících kritérií:

- podle typu generovaných struktur na jednorozměrné (řetězcové) a vícerozměrné;
- podle tvaru substitučních pravidel se řetězcové gramatiky dělí dle Chomského metodiky na obecné, kontextové, bezkontextové a regulární;
- podle řízení generování řetězců můžeme řetězcové gramatiky rozdělit na standardní, programované, indexové, podmínkové, atd.;
- podle způsobu užívání substitučních pravidel na deterministické a nedeterministické (prosté, pravděpodobnostní, fuzzy).

Podle typu může být obecná definice gramatiky, jak je uvedena výše, ještě doplněna některými dalšími pomocnými strukturami.

[21] Blízkost obou problematik vede i k častému používání názvu syntaktické, resp. lingvistické metody rozpoznávání. Ovšem, jak plyne ze srovnání obou definic, název strukturální metody vyjadřuje obecnější náhled na danou problematiku, proto mu dáváme v tomto textu přednost.

Jednorozměrné (řetězcové) gramatiky

Jádrem řetězcové gramatiky jsou substituční pravidla ve tvaru $W_1 \rightarrow W_2$, určující možnou substituci řetězce W_2 na místo řetězce W_1, který představuje část generovaného řetězce a obsahuje alespoň jeden neterminální symbol.

Příklad

Řetězcová gramatika je např. $\mathcal{G} = (\{A, B\}, \{0, 1\}, \mathcal{P}, A)$, kde množina substitučních pravidel je určena následujícím způsobem $\mathcal{P} = (A \rightarrow 0B1, 0B \rightarrow 00B1, B \rightarrow$ „e“ [22]) □□□

Příklad

Mějme gramatiku $\mathcal{G} = (\{S, A\}, \{0, 1\}, \mathcal{P}, A)$ s pravidly

$\mathcal{P}$: $S \rightarrow 0A$;

$A \rightarrow 0A \mid 1A \mid 11$. [23]

Jazyk gramatiky $\mathcal{G}$ obsahuje slova $\mathcal{L}(\mathcal{G}) = \{X \mid X \in \mathcal{V}_t^*$ [24] a X začíná 0 a končí 11$\}$. □□□

Chomského kategorie řetězcových gramatik

Podle tvaru substitučních pravidel se podle Chomského dělí řetězcové gramatiky na čtyři základní typy (od nejsložitějších k nejjednodušším): ***obecné***, ***kontextové***, ***bezkontextové*** a ***regulární***.

Gramatika typu 0 (***obecná gramatika***) nepožaduje žádná omezení tvaru substitučních pravidel.

Příklad

Obecná gramatika může mít tvar $\mathcal{G} = (\{A, B\}, \{0, 1\}, \mathcal{P}, A)$ s pravidly

$\mathcal{P}$: $A \rightarrow A1B \mid 0$;

$A1B \rightarrow 10B \mid BA1B$;

$B \rightarrow 1 \mid$ „e“. □□□

Gramatika typu 1 (***kontextová nebo senzitivní gramatika***) obsahuje substituční pravidla tvaru

$$W_1AW_2 \rightarrow W_1UW_2, \qquad W_1, W_2, U \in (\mathcal{V}_t \cup \mathcal{V}_n)^*, U \neq \text{„e”}, A \in \mathcal{V}_n \tag{4.3}$$

a může obsahovat pravidlo $S \rightarrow$ „e“. To znamená, že neterminální symbol A může být nahrazen řetězcem U pouze tehdy, sousedí-li zprava s řetězcem W_1 a zleva s řetězcem W_2. Gramatika tohoto typu neobsahuje pravidla typu $W_1AW_2 \rightarrow W_1W_2$, není tedy povoleno, aby byl terminál nahrazen prázdným řetězcem. Jedinou výjimkou je pravidlo $S \rightarrow$ „e“, které umožňuje popsat příslušnost prázdného řetězce k jazyku generovanému kontextovou gramatikou. Díky těmto pravidlům nemůže dojít při generování řetězce gramatikou typu 1 k jeho zkrácení.

Příklad

Příklad senzitivní gramatiky - $\mathcal{G} = (\{A, S\}, \{0, 1, 2\}, \mathcal{P}, S)$ s pravidly $\mathcal{P}$: $S \rightarrow 0A1$; $0A \rightarrow 00A1$ ($W_1 = 0$, $W_2 =$ „e“, $U = 0A1$); $A \rightarrow 2$. □□□

Gramatika typu 2 (***bezkontextová gramatika***) obsahuje substituční pravidla tvaru

$$A \rightarrow U, \qquad U \in (\mathcal{V}_t \cup \mathcal{V}_n)^*, U \neq \text{„e”}, A \in \mathcal{V}_n \tag{4.4}$$

a může obsahovat pravidlo $S \rightarrow$ „e“. To znamená, že neterminální symbol A lze nahradit slovem U nezávisle na jeho okolí (kontextu).

---

[22] Symbolem „e“ značíme prázdný řetězec.

[23] Tento způsob zápisu vyjadřuje tři pravidla $A \rightarrow 0A$, $A \rightarrow 1A$, $A \rightarrow 11$ se stejnou levou stranou.

[24] Zápisem $\mathcal{V}^*$ budeme rozumět množinu všech možných řetězců vytvořených ze symbolů abecedy $\mathcal{V}$, včetně řetězce prázdného.

Příklad

Příklad bezkontextové gramatiky - $\mathcal{G} = (\{S\}, \{0, 1, 2\}, \mathcal{P}, S)$ s pravidly $\mathcal{P}$: S→ 0S1 | 2. □□□

Gramatika typu 3 (***regulární gramatika***) obsahuje substituční pravidla tvaru

$$A \rightarrow xB, \text{ nebo } A \rightarrow x; \quad x \in \mathcal{V}_t; A, B \in \mathcal{V}_n \tag{4.5}$$

a může dále obsahovat pravidlo S → „e“. Někdy se jako gramatika typu 3 definuje tzv. zprava lineární gramatika (jediný neterminál na pravé straně pravidla stojí úplně napravo), která obsahuje pravidla tvaru

$$A \rightarrow xB, \text{ nebo } A \rightarrow \text{„e“}; \quad x \in \mathcal{V}_t; A, B \in \mathcal{V}_n. \tag{4.6}$$

Dá se dokázat, že gramatiky podle vztahů (4.5) a (4.6) jsou ekvivalentní.

Příklad

Příklad regulární gramatiky - $\mathcal{G} = (\{A, B\}, \{0, 1, 2\}, \mathcal{P}, A)$ s pravidly $\mathcal{P}$: A→ 0B | 2B; B→ 1B | „e“. □□□

Pravidla gramatik podle Chomského kategorizace jsou definována tak, že všechny regulární jazyky jsou rovněž bezkontextové, všechny bezkontextové jsou současně i kontextové a kontextové jsou podmnožinou obecných jazyků. Obecné a kontextové gramatiky poskytují vhodnou základnu pro teorii formálních jazyků, pro praktické aplikace jsou důležité především gramatiky (jazyky) regulární a bezkontextové.

Deterministické a nedeterministické gramatiky

Gramatiky, které obsahují pravidla vždy s různou levou stranou, nazýváme deterministické, naopak gramatiky s více substitučními pravidly s toutéž levou stranou nazýváme nedeterministické. Větší praktický význam mají gramatiky nedeterministické, protože gramatiky jsou schopny vyprodukovat pouze jedinou relační strukturu.

Při generování struktur nedeterministickou gramatikou lze obecně vybrat libovolné z možných substitučních pravidel. Není-li tento výběr specifikován, hovoříme o prostých nedeterministických gramatikách, které generují relační struktury, aniž by některé struktury preferovaly před jinými. Nicméně informace o způsobu použití (např. jak často se má určitého substitučního pravidla používat) může generování relačních struktur zkvalitnit, protože na jejím základě můžeme např. zjistit, jak často se vytvořená relační struktura vyskytuje v množině struktur generovaných použitou gramatikou, což může být výhodné právě z hlediska klasifikačních úloh, které mají často pravděpodobnostní charakter.

Přidáme-li k substitučním pravidlům váhy, vyjadřující pravděpodobnost užití pravidel, dostaneme tzv. ***stochastické gramatiky***. Přitom tyto váhy musí být určeny tak, aby součet pravděpodobností substitučních pravidel s toutéž levou stranou byl roven jedné.

Skutečnost, že se substituční pravidlo $W_1 \rightarrow W_2$ vyskytuje s pravděpodobností P, značíme

$$W_1 \xrightarrow{P} W_2. \tag{4.7}$$

Říkáme, že slovo Y je bezprostředně odvozeno ze slova X s pravděpodobností P a značíme $X \overset{P}{\Rightarrow} Y$, když $X \Rightarrow Y$ použitím jednoho substitučního pravidla $W_1 \xrightarrow{P} W_2$. Říkáme, že slovo Y je odvozeno ze slova X s pravděpodobností $P = P_1 . P_2 . \ldots . P_k$ a značíme $X \overset{P}{\underset{+}{\Rightarrow}} Y$,

$$X = U_0, U_0 \overset{P_1}{\Rightarrow} U_1 \overset{P_2}{\Rightarrow} \ldots \overset{P_k}{\Rightarrow} U_k, U_k = Y. \tag{4.8}$$

Říkáme, že slovo je generováno gramatikou $\mathcal{G}$ s pravděpodobností výskytu

$$P_G(X) = \prod_{i=1}^{n} P_i, \tag{4.9}$$

když $S \overset{P_G}{\underset{+}{\Rightarrow}} X$, $X \in \mathcal{V}_t^*$. S ohledem na klasifikační úlohy je vhodné, aby

$$\sum_{X \in \mathcal{L}(\mathcal{G})} P_G(X) = 1. \tag{4.10}$$

Gramatika splňující tento požadavek se nazývá ***konzistentní***.

Příklad

Mějme stochastickou gramatiku se substitučními pravidly P_s: $S \xrightarrow{1} 1A$, $B \xrightarrow{0,3} 0$, $B \xrightarrow{0,7} 1S$, $A \xrightarrow{0,8} 0B$, $A \xrightarrow{0,2} 1$. Slovem generovaným touto gramatikou může být např. $S \overset{1}{\Rightarrow} 1A \overset{0,8}{\Rightarrow} 10B \overset{0,3}{\Rightarrow} 100$ s pravděpodobností $P(100) = 1.0,8.0,3 = 0,24$. Jazyk $\mathcal{L}(\mathcal{G})$ generovaný touto gramatikou:

generované slovo X	pravděpodobnost P(X)
11	0,2
100	0,24
$(101)^n11$	$0,2.(0,56)^n$
$(101)^n100$	$0,2.(0,56)^n$

Dále

$$\sum_{X \in \mathcal{L}(\mathcal{G})} P(X) = 0,2 + 0,24 + \sum_{n=1}^{\infty} (0,2 + 0,24).0,56^n = 1.$$

A tedy zadaná gramatika je konzistentní. □□□

Podle tvaru substitučních pravidel se řetězcové stochastické gramatiky rovněž dělí na obecné, kontextové, bezkontextové a regulární. Stochastické mohou být samozřejmě i jakékoliv jiné, např. vícerozměrné.

Není-li součet pravděpodobností substitučních pravidel s toutéž levou stranou roven jedné, tj. platí

$$A \xrightarrow{\rho} B,\ \rho \in \langle 0;1 \rangle \text{ a } A, B \in \{\mathcal{V}_n \cup \mathcal{V}_t\}^*, \tag{4.11}$$

kde ρ je tzv. stupeň příslušnosti řetězce B řetězci A. Takovou gramatikou nazýváme ***fuzzy gramatikou***.

4.2.3 Automaty

Relační struktury dané klasifikační třídy můžeme vyjádřit kromě pouhého výčtu a gramatikou i pomocí automatu. Zatímco gramatika jako generátor relačních struktur má význam především pro popis vlastností struktur klasifikační třídy, pro potřeby vlastní klasifikace, tj. stanovení příslušnosti relační struktury k určité klasifikační třídě, má rozhodující postavení automat.

Typ automatu záleží na typu relační struktury. Existují stromové automaty, automaty polí, atp., největšího rozšíření však dosáhly automaty pro řetězcové relační struktury, opět díky úzkým vzájemným souvislostem mezi teorií formálních jazyků a teorií strukturálního rozpoznávání, ale především díky jejich poměrné jednoduchosti.

Každému typu řetězcových gramatik podle Chomského kategorizace náleží jiný typ automatu – k regulární gramatice existuje ekvivalent ***konečný automat***, pro bezkontextovou gramatiku ***zásobníkový automat***, pro kontextovou gramatiku ***lineárně ohraničený automat*** a pro

obecné gramatiky je ekvivalentem tzv. ***Turingův stroj***. Praktické využití však dosud nalezly především konečné, příp. i zásobníkové automaty.

Protože naším cílem je především porozumění principům návrhu a použití automatů, budeme se zabývat pouze nejjednodušší třídou automatů, tj. konečnými automaty.

Konečný stavový automat $\mathcal{A}$ je pětice $\mathcal{A} = (\mathcal{X}, \mathcal{S}, s_0, \mathcal{S}_c, \delta)$, kde $\mathcal{X} = \{x_i\}$ je konečná vstupní abeceda, $\mathcal{S} = \{s_m\}$ je neprázdná konečná množina vnitřních stavů, $s_0 \in \mathcal{S}$ je počáteční stav automatu, $\mathcal{S}_c \subseteq \mathcal{S}$ je neprázdná množina cílových stavů automatu a δ: $\mathcal{X} \times \mathcal{S} \rightarrow \mathcal{D}(\mathcal{S})$ je přechodová funkce, kde $\mathcal{D}(\mathcal{S})$ je množina podmnožin $\mathcal{S}$.

Automat pracuje v diskrétních krocích k = 1, 2, … a v každém kroku setrvává po určitou dobu (takt) Δt_k v některém ze svých vnitřních stavů s_k. Po příchodu vstupního symbolu x_k se stav pro příští takt změní na

$$s_{k+1} = \delta(s_k, x_k). \tag{4.12}$$

Je-li toto přiřazení jednoznačné, tj. existuje-li pro každý vnitřní stav automatu s_m a vstupní symbol x_i pouze jediný možný nový vnitřní stav, pak takový automat nazýváme deterministický. V tom případě platí, že $\mathcal{D}(\mathcal{S}) = \mathcal{S}$. Může-li automat přejít ze stavu s_m vlivem vstupu x_i do více možných stavů, pak je automat nedeterministický.

Přivedeme-li na vstup automatu řetězec $X = x_1 x_2 \ldots x_k$, přejde automat z počátečního stavu s_0 do stavu s_k. Jazyk reprezentovaný konečným automatem $\mathcal{A}$ tvoří všechny řetězce X, jejichž vlivem přejde automat $\mathcal{A}$ z počátečního stavu s_0 do konečného stavu $s_k \in \mathcal{S}_c$.

Gramatika $\mathcal{G}$ a automat $\mathcal{A}$ jsou ekvivalentní, když $\mathcal{L}(\mathcal{G}) = \mathcal{L}(\mathcal{A})$. Platí, že ke každé regulární gramatice $\mathcal{G} = (\mathcal{V}_n, \mathcal{V}_t, \mathcal{P}, S)$ existuje konečný automat $\mathcal{A} = (\mathcal{X}, \mathcal{S}, s_0, \mathcal{S}_c, \delta)$, který je s ní ekvivalentní.

Konečný stavový automat $\mathcal{A}$, ekvivalentní gramatice $\mathcal{G}$, se konstruuje podle následujících pravidel:

1) $\mathcal{X} = \mathcal{V}_t$;
2) $\mathcal{S} = \mathcal{V}_n \cup \{s_c\}$, když $s_c \notin \mathcal{V}_n$;
3) $s_0 = S$;
4) $\mathcal{S}_c = \{s_c\} \cup \{U \mid U \in \mathcal{V}_n$ takové, že existuje pravidlo $U \rightarrow$ „e“$\}$;
5)
   a) obsahuje-li gramatika pravidla tvaru $C \rightarrow xB$ nebo $C \rightarrow x$, příp. $S \rightarrow$ „e“, $x \in \mathcal{V}_t$; $B, C, S \in \mathcal{V}_n$, pak
      - je-li $(C \rightarrow xB) \in \mathcal{P}$, pak $B = \delta(C,x)$;
      - je-li $(C \rightarrow x) \in \mathcal{P}$, pak $s_c = \delta(C,x)$;
      - je-li $(S \rightarrow$ „e“$) \in \mathcal{P}$, pak $s_0 \in \mathcal{S}_c$;
   b) obsahuje-li gramatika pravidla tvaru $C \rightarrow xB$ nebo $C \rightarrow$ „e“, $x \in \mathcal{V}_t$; $B, C \in \mathcal{V}_n$, pak
      - je-li $(C \rightarrow xB) \in \mathcal{P}$, pak $B = \delta(C,x)$;
      - je-li $(C \rightarrow$ „e“$) \in \mathcal{P}$, pak $c \in \mathcal{S}_c$.

Podobně platí i obráceně, že ke každému konečnému automatu $\mathcal{A}$ existuje ekvivalentní regulární gramatika $\mathcal{G}$, vytvořená podle následujícího postupu:

1) $\mathcal{V}_t = \mathcal{X}$;
2) $\mathcal{V}_n = \mathcal{S}$ - {stavy, ze kterých nevychází ani jeden přechod};
3) $S = s_0$;
4) množina substitučních pravidel $\mathcal{P}$ je tvořena pravidly:
   - je-li $s' = \delta(s,x)$ a $\{p \mid p \in \delta(s',x)$ pro $\forall x \in \mathcal{X}\} \neq \{0\}$, pak $(s \rightarrow xs') \in \mathcal{P}$;

- je-li $s' = \delta(s,x)$ a $s' \in \mathcal{S}_c$, pak $(s \rightarrow x) \in \mathcal{P}$;
- je-li $s_0 \in \mathcal{S}_c$, pak $(S \rightarrow$ „e") $\in \mathcal{P}$.

Poznámka: podmínku $\{p \mid p \in \delta(s',x)$ pro $\forall x \in \mathcal{X}\} \neq \{0\}$ nemusí splňovat všechny koncové stavy.

Příklad

Mějme gramatiku $\mathcal{G} = (\{A, B\}, \{0, 1, 2\}, \mathcal{P}, A)$ s pravidly $\mathcal{P}$: $A \rightarrow 00B \mid 22B$; $B \rightarrow 1B \mid$ „e". Sestavte automat, který přijímá právě slova generovaná gramatikou $\mathcal{G}$.

Gramatiku $\mathcal{G}$ je nejdříve třeba přepsat do tvaru, kdy je vlevo od neterminálu vždy pouze jeden terminální symbol, tj.

$\mathcal{P}$: $A \rightarrow 0C \mid 2D$;
$C \rightarrow 0B$;
$D \rightarrow 2B$;
$B \rightarrow 1B \mid$ „e".

Vstupní abeceda je určena abecedou terminálních symbolů, tedy $\mathcal{X} = \{0, 1, 2\}$. Abeceda vnitřních stavů automatu je dána abecedou neterminálů, příp. doplněna množinou cílových stavů, určených dalším výpočtem. Počáteční stav automatu je dán počátečním symbolem gramatiky, tj. $s_0 = A$.

Ze substitučních pravidel vyplývají následující přechody funkce δ:

$(A \rightarrow 0C) \Rightarrow C = \delta(A, 0)$;
$(A \rightarrow 2D) \Rightarrow D = \delta(A, 2)$;
$(C \rightarrow 0B) \Rightarrow B = \delta(C, 0)$;
$(D \rightarrow 2B) \Rightarrow B = \delta(D, 2)$;
$(B \rightarrow 1B) \Rightarrow B = \delta(B, 1)$;
$(B \rightarrow$ „e") $\Rightarrow B \in \mathcal{S}_c$.

Přechodová funkce δ automatu je tedy dána tabulkou

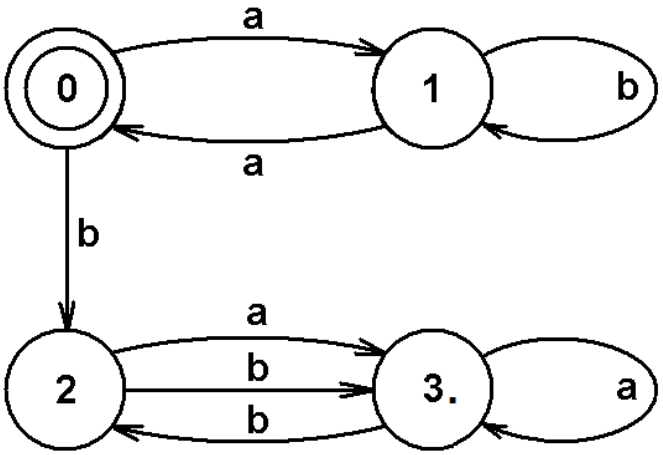
δ	(A)	B.	C	D
0	C	-	B	-
1	-	B	-	-
2	D	-	-	B

Cílovým stavem je pouze stav B, proto automat nemá jiné vnitřní stavy než dané neterminální abecedou. □□□

Příklad

Konečný automat je zadán orientovaným grafem (obr.4.8) s počátečním stavem 0 a s jedním koncovým stavem 3. Určete regulární gramatiku ekvivalentní tomuto automatu, tj. generující stejný formální jazyk, jaký automat přijímá.

Abeceda terminálních symbolů je dána vstupní abecedou automatu, tj. $\mathcal{V}_t = \{a, b\}$. Abeceda neterminálů je dána množinou vnitřních stavů automatu $\mathcal{V}_n = \{0, 1, 2, 3\}$ a axiom gramatiky určuje počáteční stav automatu S = 0.



Obr.4.8 *Zadaný konečný stavový automat*

Zbývá určit množinu substitučních pravidel gramatiky. Tabulka přechodové funkce zadaného automatu je

δ	(0)	1	2	3.
a	1	0	3	3
b	2	1	3	2

Pro všechny přechody definované přechodovou funkcí automatu platí – je-li $s' = \delta(s, x)$, pak $(s \rightarrow xs') \in \mathcal{P}$ (ze všech stavů vychází alespoň jeden přechod). Na základě tohoto pravidla určíme první část substitučních pravidel:

$$\begin{aligned}
1 = \delta(0, a) &\Rightarrow 0 \rightarrow a1;\\
2 = \delta(0, b) &\Rightarrow 0 \rightarrow b2;\\
0 = \delta(1, a) &\Rightarrow 1 \rightarrow a0;\\
1 = \delta(1, b) &\Rightarrow 1 \rightarrow b1;\\
3 = \delta(2, a) &\Rightarrow 2 \rightarrow a3;\\
3 = \delta(2, b) &\Rightarrow 2 \rightarrow b3;\\
3 = \delta(3, a) &\Rightarrow 3 \rightarrow a3;\\
2 = \delta(3, b) &\Rightarrow 3 \rightarrow b2.
\end{aligned}$$

Další substituční pravidla vyplývají z následujícího přikázání – je-li $s' = \delta(s, x)$ a přitom je $s' \in \mathcal{S}_c$, pak $(s \rightarrow x) \in \mathcal{P}$. V našem případě je koncovým stavem stav 3, uvedené pravidlo tedy platí pro všechny přechody končící ve stavu 3, tj.

$$\begin{aligned}
3 = \delta(2, a) &\Rightarrow 2 \rightarrow a;\\
3 = \delta(2, b) &\Rightarrow 2 \rightarrow b;\\
3 = \delta(3, a) &\Rightarrow 3 \rightarrow a.
\end{aligned}$$

Po přepisu jsou substituční pravidla vytvořené gramatiky

$$\begin{aligned}
&0 \rightarrow a1 \mid b2;\\
&1 \rightarrow a0 \mid b1;\\
&2 \rightarrow a3 \mid b3 \mid b3 \mid a \mid b;\\
&3 \rightarrow a3 \mid b2 \mid a.
\end{aligned}$$

□□□

Klasifikace do více klasifikačních tříd

Automatový klasifikátor skládající se z R konečných stavových automatů, tj. klasifikátor třídící vstupy do R+1 tříd (poslední třída zahrnuje řetězce, které neakceptuje žádný z použitých automatů) nechá nejdřív projít vstupní řetězec prvním automatem. Jestliže patří do jazyka reprezentovaného tímto automatem, pak vstupní řetězec zařadí do první třídy a klasifikátor ukončí svou činnost. V případě, že první automat vstupní řetězec nepřijal, vloží se na vstup druhého automatu, atd., dokud není řetězec zatříděn nebo neprojde všemi automaty klasifikátoru.

Zobecněním konečného stavového automatu je ***Moorův konečný automat***. Použití Moorova automatu zefektivňuje klasifikační proces, protože s jeho pomocí lze vstupní řetězec zatřídit do odpovídající klasifikační třídy již během jednoho průchodu automatem a nikoliv, jako v předešlém případě, po nejhůře R průchodech.

Moorův konečný automat $\mathcal{M}$ je šestice $\mathcal{M} = (\mathcal{X}, \mathcal{Y}, \mathcal{S}, s_0, \mu, \delta)$, kde $\mathcal{X} = \{x_i\}$ je konečná vstupní abeceda, $\mathcal{Y} = \{y_i\}$ je konečná výstupní abeceda, $\mathcal{S} = \{s_m\}$ je neprázdná konečná množina vnitřních stavů, $s_0 \in \mathcal{S}$ je počáteční stav automatu, $\mu: \mathcal{S} \rightarrow \mathcal{Y}$ je výstupní funkce a $\delta: \mathcal{X} \times \mathcal{S} \rightarrow \mathcal{D}(\mathcal{S})$ je přechodová funkce, kde $\mathcal{D}(\mathcal{S})$ je množina podmnožin $\mathcal{S}$. (Je-li $\mathcal{D}(\mathcal{S}) \equiv \mathcal{S}$, pak automat $\mathcal{M}$ nazýváme deterministický.)

Nechť $\mathcal{A}_r = (\mathcal{X}_r, \mathcal{S}_r, s_{0r}, \mathcal{S}_{cr}, \delta_r)$, r = 1, …, R jsou konečné stavové automaty takové, že automat $\mathcal{A}_r$ přijímá slova jazyka $\mathcal{L}_r$, představujícího klasifikační třídu ω_r. Nechť přitom platí, že

$$\bigcap_{r=1}^{R} \mathcal{S}_{cr} = \{\varnothing\}, \tag{4.13}$$

tj. že jazyky všech klasifikačních tříd jsou disjunktní. Pak lze sestavit Moorův automat $\mathcal{M} = (\mathcal{X}, \mathcal{Y}, \mathcal{S}, s_0, \mu, \delta)$ ekvivalentní klasifikátoru složenému z automatů $\mathcal{A}_r$, který ale klasifikuje pouze jedním průchodem vstupního slova automatem.

Moorův automat se vytváří podle následujících pravidel:

1) vstupní abeceda $\mathcal{X} = \bigcup_{r=1}^{R} \mathcal{X}_r$;
2) počáteční stavy s_{0r} všech automatů $\mathcal{A}_r$ ztotožníme a tento ztotožněný stav považujeme za počáteční stav automatu $\mathcal{M}$;
3) přechodovou funkci δ automatu $\mathcal{M}$ sestrojíme z dílčích přechodových funkcí δ_r opakovaným použitím následujícího pravidla, dokud nejsou zahrnuty všechny přechody původních automatů $\mathcal{A}_r$:
   - ztotožníme ty přechody, které vystupují ze ztotožněných stavů automatu $\mathcal{A}_r$ a přísluší stejným symbolům vstupní abecedy $\mathcal{X}$. Stavy, do kterých vedou ztotožněné přechody, opět ztotožníme. Ostatní přechody a stavy automatů $\mathcal{A}_r$ zachovávají původní topologii.
4) množinu vnitřních stavů $\mathcal{S}$ automatu $\mathcal{M}$ tvoří nové ztotožněné stavy a stavy původních automatů $\mathcal{A}_r$, které nelze ztotožnit;
5) výstupní abecedu $\mathcal{Y}$ tvoří identifikátory klasifikačních tříd ω_r, r = 1, …, R, spolu s identifikátorem ω_N, který označuje třídu řetězců, které nepatří do žádné klasifikační třídy ω_r, tj. $\mathcal{Y} = \{\omega_r \mid r = 1, \ldots, R\} \cup \{\omega_N\}$;
6) výstupní funkce μ automatu $\mathcal{M}$ přiřazuje hodnotu ω_r, r = 1, …, R, resp. ω_N tomu stavu automatu $\mathcal{M}$, který odpovídá koncovému stavu automatu s_{cr}, resp. stavu, kterému neodpovídá žádný koncový stav automatů $\mathcal{A}_r$.

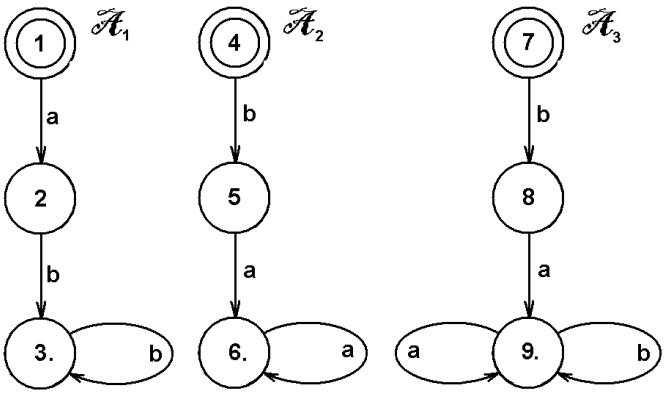
Příklad

Mějme zadány konečné automaty $\mathcal{A}_1 = (\{a,b\}, \{1,2,3\}, 1, \{3\}, \delta_1)$, $\mathcal{A}_2 = (\{a,b\}, \{4,5,6\}, 4, \{6\}, \delta_2)$ a $\mathcal{A}_3 = (\{a,b\}, \{7,8,9\}, 7, \{9\}, \delta_3)$, jejichž přenosové funkce jsou zadané tabulkami

δ_1	①	2	3.
a	2	-	-
b	-	3	3

δ_2	④	5	6.
a	-	6	6
b	5	-	-

δ_3	⑦	8	9.
a	-	-	9
b	8	9	9



Obr.4.9 *Konečné automaty podle zadání*

Zadané automaty lze znázornit pomocí orientovaných grafů na obr. 4.9. Tyto automaty přijímají formální jazyky:

$\mathcal{L}(\mathcal{A}_1) = \{X \mid X = ab\{b\}^*\}$[25]

$\mathcal{L}(\mathcal{A}_2) = \{X \mid X = ba\{a\}^*\}$

$\mathcal{L}(\mathcal{A}_3) = \{X \mid X = bb\{a \vee b\}^*\}$.

Sestavte Moorův automat, který klasifikuje vstupní řetězce stejně jako klasifikátor složený z automatů $\mathcal{A}_1$, $\mathcal{A}_2$ a $\mathcal{A}_3$.

Výsledný Moorův automat $\mathcal{M}$ bu-

[25] Hvězdička (*) v tomto případě znamená opakování výrazu v předcházejících složených závorkách.

de mít vstupní abecedu stejnou jako zadané automaty, protože $\mathcal{X} = \{a,b\} \cup \{a,b\} \cup \{a,b\} = \{a,b\}$. Počáteční stavy automatů $\mathcal{A}_1$, $\mathcal{A}_2$ a $\mathcal{A}_3$ ztotožníme do počátečního stavu automatu $\mathcal{M}$, tj. $s_0 \equiv 1 \equiv 4 \equiv 7$. Přechodová funkce automatu $\mathcal{M}$ je pak popsána tabulkou

δ	(1,4,7)	2	5,8	3	6	9
a	2	-	6	-	6	9
b	5,8	3	9	3	-	9

Po přepsání označení vnitřních stavů automatu podle pravidel: (1,4,7) → 1, 2 → 2, (5,8) → 3, 3 → 4, 6 → 5, 9 → 6 je přechodová funkce

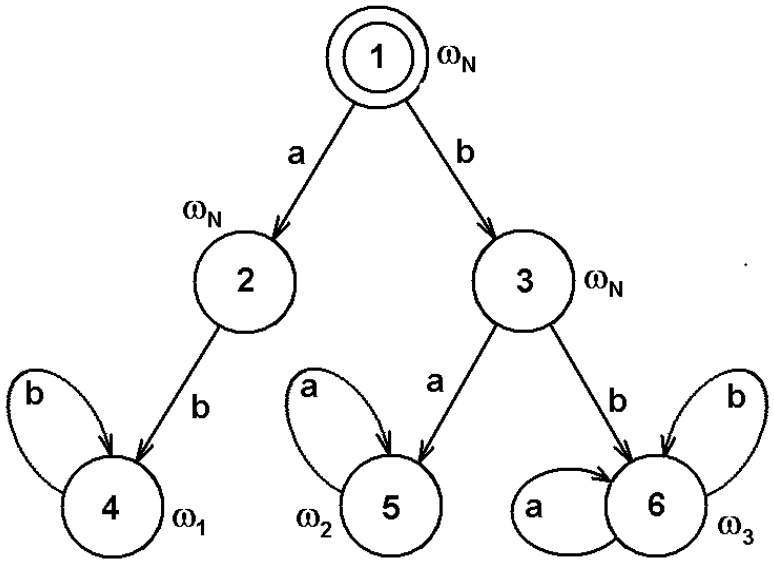
δ	(1)	2	3	4	5	6
a	2	-	5	-	5	6
b	3	4	6	4	-	6

Konečně, výstupní funkce μ automatu $\mathcal{M}$ je

μ	(1)	2	3	4	5	6
	ω_N	ω_N	ω_N	ω_1	ω_2	ω_3

protože stavy 4, 5 resp. 6 automatu $\mathcal{M}$ odpovídají koncovým stavům automatů $\mathcal{A}_r$ – 3, 6, resp. 9, zatímco stavy 1, 2, 3 automatu $\mathcal{M}$ představují stavy 1, 2, 4, 5, 7 a 8, které v zadaných automatech nejsou koncové.

Výsledný Moorův automat je možné znázornit pomocí orientovaného grafu na obr. 4.10. □□□



Obr.4.10 *Výsledný Moorův automat*

4.3 Strukturální klasifikace

4.3.1 Základní principy

Algoritmy strukturální klasifikace, tj. přiřazení identifikátoru klasifikační třídy zpracovávané relační struktuře, záleží na tom, zda etalony klasifikačních tříd respektují možnost ovlivnění klasifikované relační struktury šumovými deformacemi či nikoliv. Pokud se šumové deformace vůbec nepřipouští či zda možné deformace zahrnují již relační struktury etalonu, pak klasifikaci relační struktury provádíme ztotožněním s etalonem. Když etalon reprezentuje jen ideální nedeformované relační struktury a připustíme-li současně, že klasifikovaná relační struktura může být poruchami zdeformovaná, pak by snaha o pouhé ztotožnění s etalonem některé klasifikační třídy mohla způsobit, že by relační struktura nemusela být vůbec klasifikovatelná, protože by nemusela odpovídat etalonu žádné klasifikační třídy. V tom případě je třeba využít pravděpodobnostně definovaných etalonů – gramatik či automatů – nebo lépe principů klasifikace podle minimální vzdálenosti.

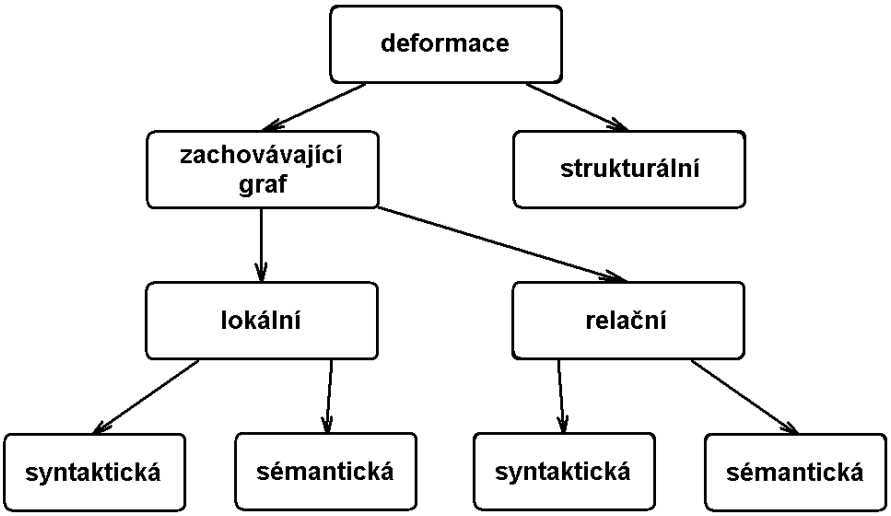
4.3.2 Klasifikace nedeformovaných struktur

Nejjednodušší jsou algoritmy klasifikace, kdy lze analyzovaná data popsat regulárním řetězcem. V tom případě může rozhodnout konečný automat, na jehož vstup vytvořený řetězec přivedeme.

V případě všech složitějších struktur je klasifikace složitější. Pro tyto struktury již není konstatování o ekvivalenci gramatik a automatů příliš užitečné, protože činnost odpovídajících automatů již není tak přímočará jako činnost konečných automatů, nýbrž při klasifikaci dochází k vytváření různých slepých cest a k návazným návratům na nižší úrovně rozhodování, proto se klasifikace provádí pomocí algoritmů, které činnost automatů pouze simulují.

4.3.3 Klasifikace deformovaných struktur

Při řešení mnoha praktických klasifikačních úloh bývá relační struktura popisující klasifikovaný objekt či jev ovlivněna působením různých poruch. Relační struktury se sémantickou

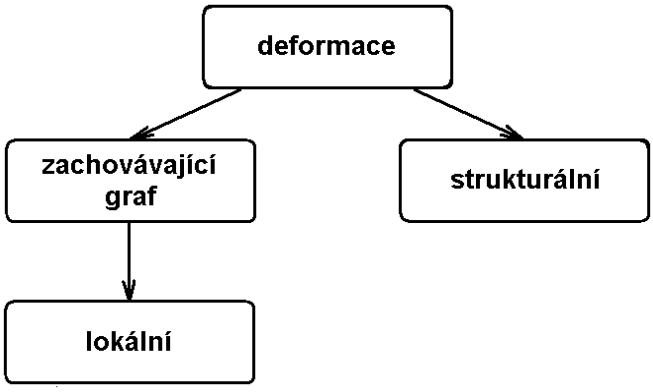


Obr.4.11 *Obecné strukturální deformační schéma*

informací mohou být deformovány podle schématu na obr.4.11:

a) strukturálně – u relační struktury je odstraněna či naopak vložena dílčí substruktura (např. u řetězců odstranění, příp. vložení jednotlivých terminálních symbolů, nebo celých jejich skupin), příp. relace struktury je zaměněna za jinou s odlišnou aritou;
b) způsobem zachovávajícím relační graf
   ba. deformací lokální – dochází k chybnému přiřazení jména primitiva k relaci, resp. ke změně hodnot atributů primitiva;
   bb. deformací relační – je způsobena použitím chybné relace při zachování arity, resp. změnou hodnot atributů relace (relační deformaci často doprovází deformace lokální).

Pro řetězec bez sémantické informace, tj. relační struktury s jediným typem relace, se deformační schéma redukuje do tvaru podle obr.4.12 – zůstává pouze možnost strukturální a lokální deformace.



Obr.4.12 *Deformační schéma pro řetězce bez sémantické informace*

Pro každou reálnou klasifikační úlohu je třeba předem určit typy přípustných deformací (dané deformačními substitučními pravidly) a klasifikační práh, který udává maximální přípustnou míru vzdálenosti (podobnosti), do které lze relační strukturu považovat za deformovaný etalon klasifikační třídy.

Strukturální vzdálenost

V případě řetězců lze deformační vlivy vyjádřit (na úrovni primitiv) trojicí tzv. ***elementárních deformačních transformací*** – eliminace, substituce a inzerce, které jsou definovány:

a) eliminační deformační transformace

$$T_E: \quad \omega_1 a \omega_2 \overset{w_E(a)}{\rightarrow} \omega_1 \omega_2; \tag{4.14}$$

b) substituční deformační transformace

$$T_S: \quad \omega_1 a \omega_2 \overset{w_E(a,b)}{\rightarrow} \omega_1 b \omega_2; \tag{4.15}$$

c) inzerční deformační transformace

$$T_I: \quad \omega_1 \omega_2 \overset{w_E(b)}{\rightarrow} \omega_1 b \omega_2, \tag{4.16}$$

kde a, b jsou libovolné terminální symboly, reprezentující primitiva řetězců, ω_1 a ω_2 jsou libovolné konečné řetězce terminálních symbolů (mohou být i prázdné) a $w_E(a)$, $w_S(a,b)$ a $w_I(b)$ jsou váhové koeficienty příslušné eliminační, substituční, resp. inzerční transformace.

Vzdálenost dvou libovolných konečných řetězců X a Y terminálních symbolů je možné určit na základě tzv. ***váhované Levenštejnovy metriky***, definované následujícím předpisem:

Je-li $\mathcal{P} = (T_1, T_2, \ldots, T_n)$, $n \geq 0$, $T_i \in (T_E, T_S, T_I)$ posloupnost elementárních deformačních transformací taková, že pro libovolná konečná slova X, Y nad abecedou $\mathcal{V}_{te} = \mathcal{V}_t \cup$ „e“ je $Y = \mathcal{P}(X)$, pak váhovaná Levenštejnova metrika je definována vztahem

$$d_{WL}(X,Y) = \min_P \left\{ \sum_{\substack{\forall a \\ T_E(a) \in P}} w_E(a) + \sum_{\substack{\forall a,b \\ T_S(a,b) \in P}} w_S(a,b) + \sum_{\substack{\forall b \\ T_I(b) \in P}} w_I(b) \right\}. \tag{4.17}$$

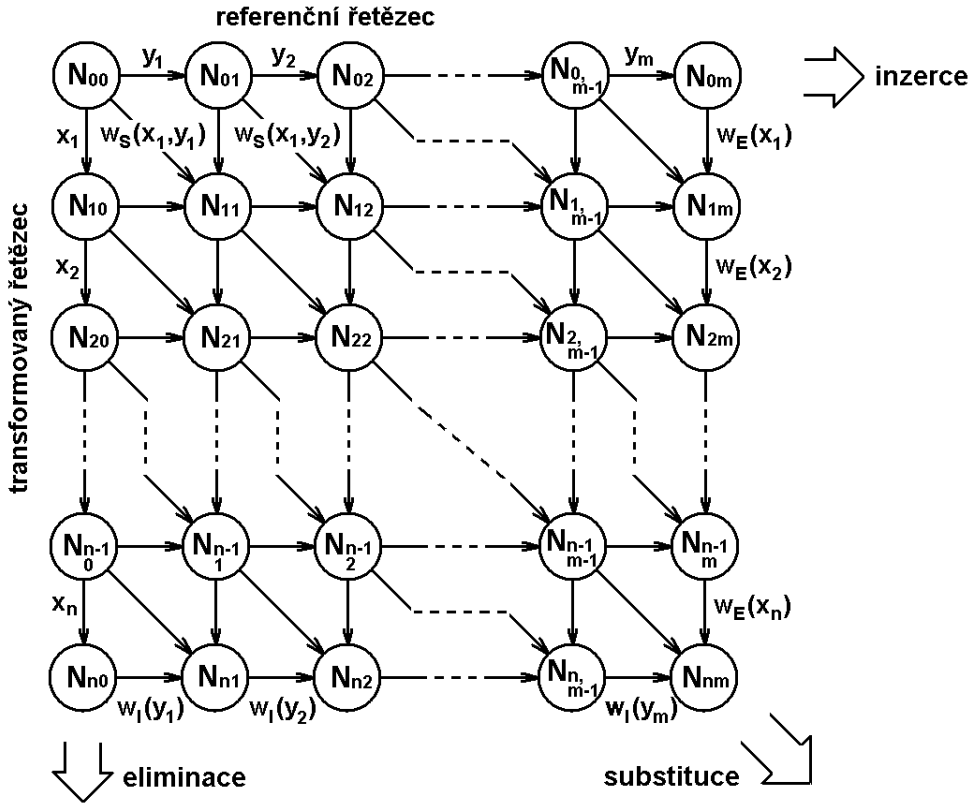
Aby byly splněny všechny tři základní axiomy metrik (axiom totožnosti, symetričnosti a trojúhelníková nerovnost) je třeba, aby platilo $w_I(a) = w_E(a)$ a $w_S(a,b) = w_S(b,a)$ pro všechny terminální symboly a, b. Váhovaná Levenštejnova metrika splňující tyto požadavky je pravá metrika.

Z váhované Levenštejnovy metriky se dají vhodnou volbou koeficientů, příp. zavedením podmínky stejné délky obou řetězců, odvodit další speciální metriky, rovněž užívané pro stanovení strukturální vzdálenosti dvou řetězců – Levenštejnova metrika, váhovaná i prostá Hammingova metrika.

Prostá (neváhovaná) Levenštejnova metrika je rovněž podle vztahu (4.17), pouze hodnoty váhových koeficientů jsou rovny $w_E(a) = w_S(a,b) = w_I(b) = 1$ pro $a \neq b$ a $w_S(a,a) = 0$ pro všechna $a, b \in \mathcal{V}_t$. To znamená, že se definiční vztah zredukuje do tvaru

$$d_L(X,Y) = \min_P \{ E_P + S_P + I_P \}, \tag{4.18}$$

kde $E_{\mathcal{P}}$, $S_{\mathcal{P}}$ a $I_{\mathcal{P}}$ je počet eliminačních, substitučních a inzerčních elementárních transformací potřebných k převedení řetězce X na řetězec Y.



Obr.4.13 *Princip výpočtu váhované Levenštejnovy vzdálenosti*

Hammingova vzdálenost je opět odvozena ze vztahu (4.10) za předpokladu, že oba řetězce jsou stejně dlouhé a $w_E(a) = w_I(a) = \infty$ a $w_S(a,a) = 0$ pro $\forall a \in \mathcal{V}_t^*$ a $w_S(a,b)$ je jistá hodnota v případě váhované Hammingovy vzdálenosti a $w_S(a,b) = 1$ v případě prosté (neváhované) Hammingovy vzdálenosti (a, b $\in \mathcal{V}_t^*$, a $\neq$ b).

Definiční vztah pro výpočet váhované Levenštejnovy vzdálenosti lze vyjádřit pomocí ohodnoceného orientovaného grafu (obr.4.13), jehož hrany reprezentují elementární deformační transformace (horizontální inzerci, vertikální eliminaci a úhlopříčné substituci) a uzly představují stavy přeměny transformovaného řetězce X na řetězec Y (počáteční uzel N_{00} původní řetězec X a uzel N_{nm} konečný stav po úplné transformaci, tj. řetězec Y). Každá cesta z uzlu N_{00} do uzlu N_{nm} odpovídá nějaké posloupnosti $\mathcal{P}$ transformací potřebných k převodu řetězce X na Y. Úkolem je tedy najít cestu mezi uzly N_{00} a N_{nm} s minimální vahou – což je standardní úloha teorie grafů.

Podobné metriky jako váhovaná Levenštejnova metrika pro řetězce byly definovány i pro složitější relační struktury, jako jsou např. pole nebo stromové relační struktury.

V případě pravděpodobnostního deformačního modelu lze podobnost dvou relačních struktur určit (za předpokladu nezávislosti jednotlivých deformačních pravidel) jako součin pravděpodobností použití příslušných deformačních pravidel.

Vlastní klasifikace deformovaných relačních struktur

Klasifikační procedura záleží na způsobu vyjádření etalonu klasifikační třídy. Pokud je klasifikační třída vyjádřena výčtem etalonů, pak lze klasifikaci provést jednoduše podle kritéria nejmenší vzdálenosti nebo na základě posouzení, zda klasifikovaný obraz patří do povoleného chybového okolí etalonových relačních struktur, definovaného opět na základě strukturální vzdálenosti. To znamená spočítat vzdálenosti zadané relační struktury ode všech etalonů a vybrat třídu, jejíž etalon je od klasifikovaného obrazu nejméně vzdálen, resp. tato vzdálenost je menší než předepsaná povolená mez.

Je-li klasifikační třída popsána gramatikou, příp. automatem, pak oba uvedené principy klasifikace (klasifikace podle minimální vzdálenosti, resp. povoleného chybového okolí) zůstávají zachovány, jen realizace klasifikačního algoritmu bude poněkud složitější.

Uvažujme opět případ nejjednodušších, tj. regulárních řetězcových relačních struktur. V tom případě je vzdálenost mezi dvěma řetězci, jak bylo dříve uvedeno, definována pomocí tří elementárních deformačních transformací – eliminace, substituce a inzerce terminálního symbolu. Je-li $\mathcal{G} = (\mathcal{V}_n, \mathcal{V}_t, \mathcal{P}, S)$ gramatika popisující klasifikační třídu, pak možný vliv deformačních transformací vyjádříme rozšířením množiny substitučních pravidel $\mathcal{P}$ přidáním pravidel reprezentujících všechny možné chybové transformace terminálních symbolů. Nově přidaná pravidla jsou opatřena nenulovými vahami, vyplývajícími z vah deformačních transformací. O původních pravidlech gramatiky předpokládáme, že mají váhu nulovou. Vzdálenost řetězce od etalonu klasifikační třídy je určena minimálním součtem vah substitučních pravidel potřebných pro vygenerování zadaného řetězce podle pravidel rozšířené gramatiky.

Podobná situace je s automaty – resp. konkrétně konečnými automaty. Tabulku přechodové funkce rozšíříme o přechody vyplývající z chybových transformací. Těmto přechodům, opět na rozdíl od původních, přisoudíme nenulové váhy podle vah chybových transformací. Vzdálenost řetězce od etalonu reprezentovaného automatem je dána minimální celkovou vahou přechodů automatu použitých při zpracování vstupního řetězce.

Příklad

Mějme opět regulární gramatiku $\mathcal{G} = (\{A, B, C, D\}, \{0, 1, 2\}, \mathcal{P}, A)$ s pravidly $\mathcal{P}$: $A \to 0C \mid 2D$; $C \to 0B$; $D \to 2B$; $B \to 1B \mid$ „e“ a dále mějme zadánu množinu substitučních elementárních transformací s vahami podle tab.4.1 a eliminačních, resp. inzerčních transformací s vahami podle tab.4.2. Určeme rozšířenou gramatiku a jí odpovídající konečný automat, umožňující klasifikaci deformovaných struktur.

Tab.4.1 *Váhy substitučních elementárních transformací*

w_S	0	1	2
0	0	1	1
1	1	0	2
2	1	2	0

Tab.4.2 *Váhy eliminačních a inzerčních elementárních transformací*

w_E, w_I	
0	1
1	1
2	1

Množina substitučních pravidel, rozšířená o pravidla substitučních a eliminačních elementárních transformací bude (nulové váhy nejsou uvedeny)

$$A \to 0C,\ A \overset{1}{\to} 1C,\ A \overset{1}{\to} 2C,\ A \to 2D,\ A \overset{2}{\to} 1D,\ A \overset{1}{\to} 0D,$$

$$C \to 0B,\ C \overset{1}{\to} 1B,\ C \overset{1}{\to} 2B,$$

$$D \to 2B,\ D \overset{1}{\to} 0B,\ D \overset{2}{\to} 1B,$$

$$B \to 1B,\ B \overset{1}{\to} 0B,\ B \overset{2}{\to} 2B,$$

$$A \overset{1}{\to} \text{"e"}C,\ A \overset{1}{\to} \text{"e"}D,\ C \overset{1}{\to} \text{"e"}B,\ D \overset{1}{\to} \text{"e"}B,\ B \overset{1}{\to} \text{"e"}B.$$

Poslední pravidlo vzniklo mechanickým uplatněním eliminační deformace na předposlední pravidlo zadané množiny $\mathcal{P}$. Protože však toto pravidlo při generování vytvářený řetězec nemění, je logické předpokládat i implicitní pravidlo $B \to B$. Protože z obou pravidel má menší váhu transformace to druhé, lze poslední nové pravidlo vypustit.

Při využití inzerčních transformací je potřeba rozhodnout, do kterého místa původního pravidla nový symbol vložíme. Předpokládejme, že to bude před terminální symbol. Pak je množina pravidel vyjadřujících inzerční transformace následující:

$A \xrightarrow{1} 00C$, $A \xrightarrow{1} 10C$, $A \xrightarrow{1} 20C$, $A \xrightarrow{1} 02D$, $A \xrightarrow{1} 12D$, $A \xrightarrow{1} 22D$,

$C \xrightarrow{1} 00B$, $C \xrightarrow{1} 10B$, $C \xrightarrow{1} 20B$,

$D \xrightarrow{1} 02B$, $D \xrightarrow{1} 12B$, $D \xrightarrow{1} 22B$,

$B \xrightarrow{1} 01B$, $B \xrightarrow{1} 11B$, $B \xrightarrow{1} 21B$, $B \xrightarrow{1} 0$, $B \xrightarrow{1} 1$, $B \xrightarrow{1} 2$.

Aby byl zachován standardní tvar pravidel regulární gramatiky, přepíšeme tato pravidla do tvaru:

$A \xrightarrow{1} 0E$, $E \rightarrow 0C$, $A \xrightarrow{1} 1E$, $A \xrightarrow{1} 2E$, $A \xrightarrow{1} 0F$, $F \rightarrow 2D$, $A \xrightarrow{1} 1F$, $A \xrightarrow{1} 2F$,

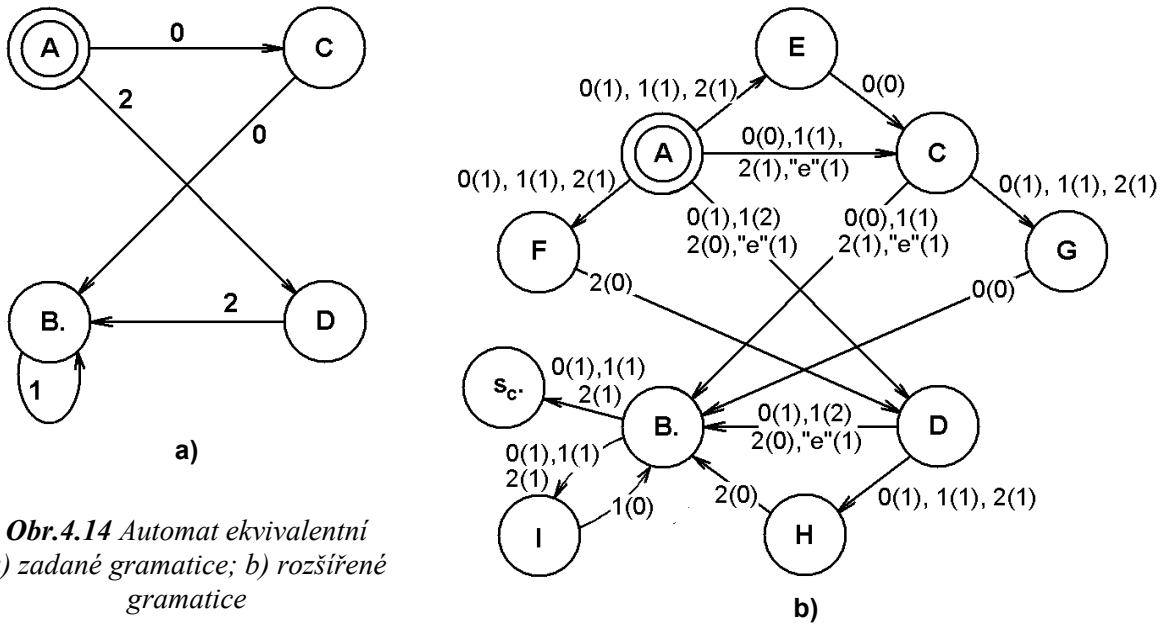
$C \xrightarrow{1} 0G$, $G \rightarrow 0B$, $C \xrightarrow{1} 1G$, $C \xrightarrow{1} 2G$,

$D \xrightarrow{1} 0H$, $H \rightarrow 2B$, $D \xrightarrow{1} 1H$, $D \xrightarrow{1} 2H$,

$B \xrightarrow{1} 0I$, $I \rightarrow 1B$, $B \xrightarrow{1} 1I$, $B \xrightarrow{1} 2I$, $B \xrightarrow{1} 0$, $B \xrightarrow{1} 1$, $B \xrightarrow{1} 2$.

To znamená, že množina neterminálních symbolů musí být dále rozšířena o symboly E, F, G, H a I. Poslední uvedená skupina substitučních pravidel spolu s první skupinou tvoří pravidla gramatiky, podle kterých jsme schopni určit vzdálenost řetězců od řetězců původního strukturálního etalonu.

Jak už bylo uvedeno, vzdálenost zadaného řetězce od etalonu klasifikační třídy je určena minimálním součtem vah substitučních pravidel potřebných pro vygenerování zadaného řetězce podle pravidel rozšířené gramatiky. Takže úloha, opět spadá mezi kombinatorické optimalizační úlohy, jejichž absolutní řešení se zpravidla hledá jen obtížně. Je třeba se spokojit jen se suboptimálním řešením, získaným např. nějakou variantou algoritmů uvedených v kapitole pojednávající o selekci.



Obr.4.14 *Automat ekvivalentní a) zadané gramatice; b) rozšířené gramatice*

Automat ekvivalentní původní gramatice lze vyjádřit orientovaným grafem na obr.4.14a. Automat, který odpovídá rozšířené gramatice je na obr.4.14b (váhy přechodů jsou uvedeny v závorkách). Je to automat nedeterministický. Výběr mezi možnými přechody je řízen optimalizační procedurou, např. suboptimalizační algoritmus, ekvivalentní algoritmu sekvenční dopředné selekce, uvedený v kapitole o selekci příznaků, vybírá ten přechod, jehož váha je nejmenší. Pokud je více přechodů s toutéž vahou, je výběr náhodný. □□□

Doporučená literatura

[1] Holčík, J., Analýza a klasifikace signálů. Nakladatelství VUT v Brně, Brno (1992)

[2] Holčík, J., Signály, časové řady a systémy. CERM, Brno (2012)

[3] Haruštiaková, D., Jarkovský, J., Littnerová, S., Dušek, L. Vícerozměrné statistické metody. CERM, Brno (2012)

[4] Komprdová, K., Rozhodovací stromy a lesy. CERM, Brno (2012)

[5] Schwarz, D., Lineární a adaptivní zpracování dat. CERM, Brno (2012)

[6] Kotek, Z., Brůha,I., Chalupa, V., Jelínek, J., Adaptivní a učící se systémy. SNTL, Praha (1980)

[7] Theodoridis, S., Koutroumbas, K., Pattern Recognition. 4th ed., Elsevier-Academic Press, Amsterdam (2009)

[8] Bishop, C.M., Pattern Recognition and Machine Learning. Springer,New York (2006)

[9] Webb, A., Statistical Pattern Recognition, 2nd ed., Wiley, Hoboken (2002)

[10] Duda, R.O., Hart, P.E., Strok, D.G., Pattern Classification, 2nd Ed., Wiley, New York (2001)

[11] Stork, D.G., Yom-Tov E., Computer Manual in Matlab® to Accompany Pattern Classification. 2nd ed., Wiley, Hoboken (2004)

[12] McLachlan, G.J., Discriminant and Statistical Pattern Recognition, Wiley Series in Probability and Statistics. Wiley, Hoboken (2004)

[13] Tan, P.-N., Steinbach, M., Kumar,V., Introduction to Data Mining. Pearson/Addison Wesley, Boston (2006)

[14] Hastie, T., Tibshirani, R., Friedman, J., The Elements of Statistical Learning. Data Mining, Inference, and Prediction. 2nd ed., Springer, New York (2009)

[15] Witten, I.H., Frank E., Data Mining. Practical Machine Learning Tools and Techniques. 2nd Ed., Elsevier, Amsterdam (2005)

[16] Han, J., Kamber, M., Data mining. Concepts and Techniques. 2nd ed., The Morgan Kaufmann Series in Data Management Systems. Elsevier, Amsterdam (2006)

[17] Mitchell, T.M., Machine Learning. Computer Science Series, McGraw-Hill, New York (1997)

[18] Jollife, I.T., Principal Component Analysis. 2nd ed., Springer Series in Statistics. Springer, New York (2002)

[19] Hyvärinen, A., Karhunen, J., Oja, E., Independent Component Analysis. Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. Wiley, New York (2001)

[20] Meloun, M., Militký, J. Statistická analýza experimentálních dat. Academia, Praha (2004)

[21] Meloun, M., Militký, J. Kompendium statistického zpracování dat. Academia, Praha (2006)

Obsah

Summary

The book “Data Analysis and Classification” links to a certain extent to the textbook “Multivariate Statistical Methods” and theoretically evolves topics described there. Both publications are intended mainly for students of the Computational Biology study programme at the Faculty of Science of the Masaryk University. They were both supported by the ESF grant no. CZ.1.07/2.2.00/07.0318 „Multidisciplinary Innovation of Study in Computational Biology“.

Chapter one introduces fundamentals and principles of data processing and analysis, defines their aims and describes individual phases of processing of static and dynamic data.

The second chapter, the longest in the publication, deals with individual feature based methods of statistical pattern recognition. The chapter begins with methods of classification by means of discrimination functions, which are demonstrated by description of Bayesian classifiers. This is followed by methods of minimum distance classification. For this purpose the terms metric and similarity metric, respectively, are defined and subsequently extended by specific metrics for determination of distance and similarity of two patterns described by quantitative and qualitative features. These metrics are further developed into deterministic and probability methods of determination of distance between two sets. Another part of the chapter is focused on algorithms of classification into classes defined by borders in feature space and, finally, methods of sequential classification.

The third chapter analyses methods for selection of attributes; the highest emphasis is put on analysis of principal components and analysis of independent components.

Chapter four, the last but not least, is focused on methods of structural analysis and classification. The chapter introduces terms such as primitive, relation, and relation structure, deals with methods of structural description of classification classes, particularly grammars and automats, and is concluded by description of methods of structural pattern recognition of both non-deformed and deformed relation structures.

Analýza a klasifikace dat
prof. Ing. Jiří Holčík, CSc.

Recenzenti:doc. RNDr. Ing. Marcel Jiřina, Ph.D., doc. Ing. Vladimír Krajča, CSc.
Obálka: Radim Šustr, DiS.
Jazyková korekce: Ing. Marie Juranová
Vydalo: AKADEMICKÉ NAKLADATELSTVÍ CERM, s.r.o., Brno
Purkyňova 95a, 612 00 Brno
www.cerm.cz
Tisk: FINAL TISK s.r.o. Olomučany
Náklad: 200 ks
Vydání: první
Vyšlo v roce 2012

ISBN 978-80-7204-793-2