**Atribútovo orientované analýzy**

**1 Všeobecnosť a špecifickosť (Generality and specifity)**

**1.1 Prezentovanie entíc**

Entice sú reprezentované ako súbor, množina:

a) ak X, Y sú entice, X = <x1, x2, x3, .....xn>, Y = <y1, y2, y3, …..yn>

b) a predpokladajme, že atribúty sú A1, A2, A3,...An

c) potom je možné entice prezentovať ako súbor atribútovo-hodnotových párov:

X = {A1 = x1, A2 = x2,...An = xn}, Y = {A1 = y1, A2 = y2,…An = yn}



**1.2 Poradie všeobecnosti rôznych typov atribútov**

a) nominálne atribúty – X je všeobecnejšie ako Y (X zahŕňa Y) ak X ⊆ Y a naopak, Y je špecifickejšie ako X (Y je zahrnuté v X) ....teda, každý prvok X je tiež prvkom Y

b) štrukturálne atribúty (atribúty vytvárajúce koncept hierarchie) – X je všeobecnejšie ako Y, ak y1 je nástupcom xi v koncepte hierarchie atribútov

c) zmena nominálneho atribútu na štrukturálny – vytvorením dvojúrovňovej koncepcie hierarchie s definovaním koreňovej premennej zahŕňajúcej všetky hodnoty atribútu; v dátových kockách sa najčastejšie využíva označenie „ALL“

**2 Zovšeobecňovanie atribútov (Attribute generalization)**

a) nominálne atribúty – „dropping condition“\_vylučovaním atribútovo-hodnotového páru z X, čo vedie k vytváraniu podmnožín X; táto operácia zodpovedá operácii „dice“ dátovej kocky (výber dvoch alebo viacerých dimenzií z dátovej kocky)

b) štrukturálne atribúty – „climbing up concept hierarchy“\_nahrádzanie pôvodnej hodnoty atribútu zovšeobecnenou hodnotou; táto operácia zodpovedá operácií „roll-up“ dátovej kocky

***3 Príklad***

***3.1 Prezentovanie entíc***

X1 = {x1 = sunny, x2 = hot, x3 = high, x4 = weak, y = no}

X2 = {x1 = sunny, x2 = hot, x3 = high, x4 = strong, y = no}

X3 = {x1 = overcast, x2 = hot, x3 = high, x4 = weak, y = yes}

...

...

...

X14 = {x1 = rainy, x2 = mild, x3 = high, x4 = strong, y = no}

***3.2 Zovšeobecnenie***

Y1 = {x2 = hot, x3 = high, x4 = weak} (vylúčili sme prvá a posledný atribút, opísali sme opakujúce sa hodnoty, ide akoby o substitúciu z matematiky ☺)

Y1 je všeobecnejšie ako X1 a X3, nakoľko Y1 ⊆ X1 a Y1 ⊆ X3 (všetky hodnoty Y1 sa nachádzajú v X1 aj v X3) ....takto je možné pokračovať v celej dátovej kocke a na základe takéhoto zovšeobecňovania je možné tvoriť aj asociačné pravidlá

**4 Opodstatnenosť, dôležitosť atribútov (Attribute relevance)**

- výber podmnožiny atribútov z celkovej sady atribútov, ktorá by najlepšie charakterizovala alebo predikovala skúmanú triedu sa realizuje 2 spôsobmi:

1. *filtrovanie*: schematicky nezávislý výber atribútov

a) minimálne množstvo atribútov, ktoré oddelia všetky entice bez ohľadu na triedu

b) minimálne množstvo atribútov, ktoré zachovajú rozdelenie v triedach (metódy založené na rozhodujúcom prvku, metódy založené na neurčitosti)

2. *schematicky špecifické metódy*

**4.1 Výber atribútov založený na rozhodujúcom prvku**

- pri stanovovaní rozhodujúceho prvku je potrebné vychádzať z mier podobnosti (resp. mier vzdialenosti):

a) pre numerické atribúty – Euklidovská vzdialenosť

 $D\left(X,Y\right)= \sqrt{(x\_{1-}y\_{1})^{2}+(x\_{2}-y\_{2})^{2}+…+(x\_{n}-y\_{n})^{2}}$

b) pre nominálne atribúty – počet rozdielov

 $D\left(X,Y\right)=\sum\_{1}^{n}d(x\_{i},y\_{i})$

c) pre zmiešané atribúty (nominálne, numerické) – normalizácia

*Postup výberu atribútov založený na podobnosti:*

- pre každú enticu nájdeme najbližších susedov (najbližšie entice v závislosti od mier vzdialenosti) rovnakej alebo odlišnej triedy – „najbližšia zhoda“ a „najbližšia nezhoda“

- ak *najbližšia zhoda* má rozdielnu hodnotu určitého atribútu potom sa tento atribút javí ako irelevantný a jeho váha bude znížená

- v prípade *najbližšie nezhody* sú atribúty s rozdielnymi hodnotami relevantné a ich váha bude zvýšená

- algoritmus výberu je nasledovný: každý atribút má na začiatku rovnakú váhu a tá je následne upravovaná podľa postupu uvedeného vyššie; tento postup umožňuje zoraďovanie atribútov podľa dôležitosti a vyberanie najlepšej podmnožiny atribútov

Najbližší susedia pre **X1** v triede „**no**“ (najbližšie zhody) sú X2 a X8 (nezohľadnením triedy *„y“ (nezapočítavaním prípadnej nezhody)*) máme:

- D(X1,X2) = 1 (v jednej hodnote sa hodnoty X2 odlišujú od hodnôt X1 a to windy = strong, inak sú všetky ostatné hodnoty entice X1 a X2 rovnaké)

- D(X1,X6) = 4 ...4 hodnoty atribútov sú rozdielne (rainy, cold, normal, strong)

- D(X1,X8) = 1

- D(X1,X4) = 3

Nakoľko atribút *x4* (windy) má rozdielnu hodnotu v entici X1 a X2, znížime jeho váhu (dôležitosť)

Atribút *x2*(temperature) má rozdielnu hodnotu v entici X1 a X8, tiež znížime jeho váhu

Najbližší sused X1 v opačnej triede, v triede „yes“ (najbližšia nezhoda) je X3 (D(X1,X3) = 1)

Atribút *x1* (outlook) má rozdielnu hodnotu v entici X1 a X3, preto zvýšime jeho váhu

***Príklad:***

**4.2 Výber atribútov založený na entropii (neurčitosti)**

- cieľom je nájdenie najrelevantnejšieho atribútu (atribútu s najvyššou diskriminačnou silou), teda atribútu s maximálnym informačným ziskom (***information gain „G“***)

$$informačný zisk\left(A\right)=I\left(S\right)-I(A)$$

 - informačný zisk je teda tvorený rozdielom medzi neurčitosťou triedy „I(S)“ a neurčitosťou atribútu „I(A)“

**neurčitosť tried *I(S)*:**

$$I\left(S\right)= -P\left(C\_{i}\right)\*log\_{2}P\left(C\_{i}\right)-P\left(C\_{2}\right)\*log\_{2}P\left(C\_{2}\right)-...-P\left(C\_{n}\right)\*log\_{2}P(C\_{n})$$

pričom $P\left(C\_{i}\right)=\frac{|S\_{i}|}{|S|}$

kde: Si – počet entíc v triede Ci, i = 1, 2, ...m

 S – celková množina entíc, ktorú je možné rozdeliť do *m* tried – C1, C2,...Cm

**neurčitosť atribútu *I(A)*:**

$$I\left(A\right)= \frac{|A\_{1}|}{|S|}\*I\left(A\_{1}\right)+\frac{\left|A\_{2}\right|}{\left|S\right|}\*I\left(A\_{2}\right)+. .+\frac{\left|A\_{k}\right|}{\left|S\right|}\*I(A\_{k})$$

kde: Ai – podmnožiny vytvorené obmenami atribútu, i = 1, 2, ...k

Postup výpočtu I(A1), I(A2), ....I(Ai) je analogický ako samotný výpočet I(S)

$$I\left(A\_{i}\right)= -P\left(C\_{i}\right)\*log\_{2}P\left(C\_{i}\right)-P\left(C\_{2}\right)\*log\_{2}P\left(C\_{2}\right)-...-P\left(C\_{n}\right)\*log\_{2}P(C\_{n})$$

pričom $P\left(C\_{i}\right)=\frac{|S\_{i}|}{|A\_{n}|}$

kde: Si – počet entíc v triede Ci, i = 1, 2, ...m

 Ai – počet entíc v konkrétnej obmene atribútu (v konkrétnej inštancii atribútu)

***Príklad:***

I(S) = -P(yes)\*log2P(yes)-P(no)\*log2P(no)

$$I\left(S\right)= -\frac{5}{14}\*log\_{2}\frac{5}{14}-\frac{9}{14}\*log\_{2}\frac{9}{14}$$

neurčitosť voľby „play“ je spôsobená početnosťou výskytu voľby „no“... „play“ sa vyskytovalo 9x a „no“ sa vyskytovalo 5x z celkových 14 záznamov

A = outlook, A1 = {1,2,8,9,11} (sunny), A2 {3,7,12,13} (overcast), A3 {4,5,6,10,14} (rainy)

$$I\left(outlook\right)= \frac{5}{14}\*I\left(A\_{1}\right)+\frac{4}{14}\*I\left(A\_{2}\right)+\frac{5}{14}\*I\left(A\_{3}\right)$$

$$I\left(A\_{1}\right)=I\left(\left\{no,no,no,yes,yes\right\}\right)=-\frac{3}{5}\*log\_{2}\frac{3}{5}-\frac{2}{5}\*log\_{2}\frac{2}{5}$$

$$I\left(A\_{2}\right)=I\left(\left\{yes,yes,yes,yes\right\}\right)=0$$

$$I\left(A\_{3}\right)=I\left(\left\{yes,yes,no,yes,no\right\}\right)=-\frac{3}{5}\*log\_{2}\frac{3}{5}-\frac{2}{5}\*log\_{2}\frac{2}{5}$$

pri

**4.3 Charakterizovanie a porovnávanie tried**

- pri zovšeobecňovaní je potrebné zohľadniť, koľko entíc (Mi) sa nám podarilo zovšeobecniť v rámci skúmaných tried Ci a koľko entíc (Ki) z celkového počtu entíc S

******

Ci = 2 (no, yes)

S = 14 (počet záznamov)

- následne sa posudzujú 2 vlastnosti zovšeobecnenej entice X:

a) ***charakterizačná sila*** „***T(X)***“ na základe vzorca $T\left(X\right)= \frac{M\_{i}}{C\_{i}}$ , ktorá definuje schopnosť zovšeobecnenej entice X charakterizovať jednotlivé triedy

Ak T(X) < 1, teda zovšeobecnená entica X nezahŕňa všetky entice skupín Ci, je potrebné definovať viac zovšeobecnených entíc, aby skupiny tried Ci mohli byť charakterizované

Ak T(X) je príliš nízke, hovoríme o prešpecializovaní

b) ***diskriminačná sila*** „***D(X)***“ na základ vzorca $D\left(X\right)= \frac{M\_{i}}{K\_{i}}$, ktorá definuje spoľahlivosť zovšeobecnenej entice X, teda spoľahlivosť vytvoreného všeobecného pravidla

Ak D(X) < 1, potom zovšeobecnená entica X zahŕňa entice nielen definovanej triedy ale aj kontrastnej triedy (napr. zovšeobecnená entica X by platila nielen pre triedu play=yes, ale aj pre triedu play=no) a v takom prípade je potrebné enticu X viac špecializovať (došlo k nadbytočnému zovšeobecneniu)

**5 Štatistické miery (Statistical measures)**

- skúmané atribúty je možné analyzovať aj na základe štatistických mier

a) miery centrálnej tendencie – aritmetický priemer

b) miery disperzie – rozptyl, smerodajná (štandardná) odchýlka