VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

POROVNÁVÁNÍ MODELŮ PRO DOLOVÁNÍ Z DAT

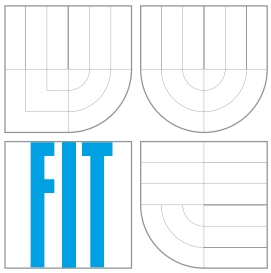
BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

JAN POSPÍŠIL

BRNO 2007

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

POROVNÁVÁNÍ MODELŮ PRO DOLOVÁNÍ Z DAT

DATAMINING MODELS COMPARISON

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — JAN POSPÍŠIL
AUTHOR

VEDOUCÍ PRÁCE — Ing. ROMAN LUKÁŠ, Ph.D.
SUPERVISOR

BRNO 2007

Abstrakt

Tato práce se zabývá základním porovnáváním vlastností dataminingových modelů vzhledem k různých povahám dat. Důraz byl kladen především na nalezení klíčových vlastnotí, které ovlivňují přesnost klasifikace dat. Práce je členěna do několika částí tak, aby i neodborný čtenář nebo dokonce uplný laik porozuměl tématu a mohl ze závěrů této práce profitovat. V první fázi je čtenář seznámen s problematikou dataminingu, potřebných modelů a algoritmů, druhá část se zabývá porovnáváním modelů a zhodnocením výsledků.

Klíčová slova

Dolování z dat, klasifikace, porovnávání modelů, SAS Enterprise miner, získávání znalostí z dat, datamining

Abstract

This thesis focuses on comparing of the datamining models features depending on the different databazis topology. The objekt was to find key features that at most involve the accuracy of classification. Thesis is composed from chapters in a way that even a non-professional or even a complete laik could understand the object and could find theese thesis results useful. In the beginning the reader is beeing made familiar with all the background information about datamining and its models and algorithms, the second part denotes about the model comparison and discusses its results.

Keywords

Datamining, classification, datamining models, SAS Enterprise miner

Citace

Jan Pospíšil: Porovnávání modelů pro dolování z dat, bakalářská práce, Brno, FIT VUT v Brně, 2007

Porovnávání modelů pro dolování z dat

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana Ing. Romana Lukáše, Ph.D .

. . . . . .
Jan Pospíšil
14. května 2007

Zadání

Porovnání modelů pro dolování dat z dat

- **Vedoucí:** Lukáš Roman, Ing., Ph.D., UIFS FIT VUT
- **Oponent:** Bartík Vladimír, Ing., Ph.D., UIFS FIT VUT

1. Seznamte se podrobně s metodami pro dolování dat z databází a s prostředím programu SAS Enterprise Miner.
2. Z různých zdrojů nalezněte data vhodná pro dolování.
3. Pro dolování z těchto dat pomocí různých modelů (např. regresní analýza, rozhodovací strom, neuronová síť BP) použijte program SAS Enterprise Miner.
4. Z předchozího výsledku udělejte analýzu, který model je vhodný pro jaký druh dat a proč.
5. Zhodnoťte dosažené výsledky.

Licenční smlouva

Licenční smlouva je uložena v archívu Fakulty informačních technologií Vysokého učení technického v Brně.

Obsah

1 Získávání znalostí z dat **3**
- 1.1 Úvod . . . 3
- 1.2 Definice . . . 4
- 1.3 Potencionální aplikace . . . 4
- 1.4 Datové zdroje . . . 5
  - 1.4.1 Multidimenzionální datový model . . . 5
  - 1.4.2 Operace nad multidimenzionální kostkou . . . 5
- 1.5 Proces získávání znalostí . . . 7
  - 1.5.1 Fáze . . . 7
  - 1.5.2 Metodiky . . . 7
- 1.6 Typické dataminingové úlohy . . . 8
  - 1.6.1 Dolování asociačních pravidel . . . 8
  - 1.6.2 Shlukovací analýza . . . 10
  - 1.6.3 Predikce . . . 11
  - 1.6.4 Klasifikace . . . 11

2 Klasifikace **12**
- 2.1 Úvod . . . 12
  - 2.1.1 Proces klasifikace . . . 12
  - 2.1.2 Vlastnosti klasifikátoru . . . 12
- 2.2 Rozhodovací strom . . . 13
  - 2.2.1 Algoritmus . . . 13
  - 2.2.2 Výběr atributu . . . 14
  - 2.2.3 Optimalizace . . . 15
- 2.3 Neuronová síť . . . 15
  - 2.3.1 Neuron . . . 15
  - 2.3.2 Neuronová síť Backpropagation . . . 16

3 Porovnávání modelů pro klasifikaci **19**
- 3.1 Výzkumný záměr . . . 19
- 3.2 Testovací data . . . 19
- 3.3 Prostředí SAS Enterprise miner . . . 20
- 3.4 Porovnávání modelů . . . 24
  - 3.4.1 Výběr klasifikátoru s nejlepšími výsledky . . . 24
  - 3.4.2 Vliv počtu atributů na klasifikaci . . . 26
  - 3.4.3 Vydolovatelnost informace . . . 29
  - 3.4.4 Zašumělá třída . . . 30

4 Závěr **33**

Seznam použitých zdrojů **33**

A Příloha **35**

Kapitola 1

Získávání znalostí z dat

1.1 Úvod

Historický vývoj

Získávání znalostí dat je považováno za jeden z hlavních směrů vývoje databázových technologií dneška. Je ale nutno dodat, že snahy o analytický přístup k ukládaným datům zde byl již od samého začátku 60.let, v době vzniku prvních datově intenzivních aplikací. Na data uložená v hieratických či síťových databázích byly aplikovány jednoduché statistické metody logické regrese a rozhodovacích stromů. Výsledky obsahovaly chyby vlivem náhodných korelací, se kterými se tento přístup nebyl schopen vypořádat. Též vzhledem k výkonové spolehlivostní úrovni tehdejší výpočetní techniky, zůstaly pokusy tohoto typu pouze na bázi akademických projektů.

Prekvizity

Dalším impulzem se ukázal v 80. letech až příchod nových systémů řízení báze dat založených buď na relačním nebo později i objektovém přístupu. Takto řízené databáze zaznemenaly v krátké době velké rozšíření v globálním měřítku a prosadily se jako plnohodnotný způsob uchovávání informací.

Přístup k databázi už nebyl doménou programátorů. Do hry přicházejí i lidé, kterým se počítač dostal na kancelářský stůl, a kteří s databází, pomocí různých přátelských uživatelských rozhraní, každodenně pracují. Z pohledu dataminingu cílovou skupinou jsou pak lidé, kteří využívají přístupu k databázím k analýze či jinému způsobu získavání poznatků o datech v nich uložených.

Motivace

Vlastníci databází již kromě operační databáze, kde uchovávají svá aktuální data, disponují i mechanismy na ukládání a archivaci těchto dat ve velkých objemech, viz. [3].

> Topíme se v datech, ale trpíme nedostatkem informací.

Vlivem rozvoje informační společnosti a technického pokroku vznikají nové zdroje dat:

- Multimediální zdroje dat, datové proudy.
- Web a strukturované formy dokumentů.

Explozivně rostou objemy informací uchovávaných v databázích. Objemy dat začínají uživatelům přerůstat přes hlavu. Úměrně tomu tedy roste zájem o nástroje, které dovedou velké objemy dat efektivně zpracovávat a transformovat na znalosti pro podporu rozhodování.

1.2 Definice

Jako **získávání znalostí z dat** lze definovat činnost, která vede k získání zajímavých netriviálních zjištění z velkého množství dat. Předmětem zájmu jsou souvislosti mezi daty, nikoliv hodnoty, které jsou v databázi explicitně uvedeny. Nejde tedy o koncept typický v operačních databázích. Pouze se nedotazujeme na požadované hodnoty. Aby měla tato činnost smysl, musí získané znalosti přinášet nějaký užitek, například v podobě informací, které pomáhají při strategických rozhodnutích, nějakým způsobem obohacují naše povědomí o datech, ze kterých dolujeme právě ony skryté souvislosti. Lidé na vedoucích pozicích se rozhodují podle své vlastní intuice a dobrá intuice patří bezesporu ke kvalitám správného manažera. Typickým úkolem získávání znalostí z dat je poskytnout lepší výchozí podmínky při strategickém rozhodnutí tím ,že pomůžeme pochopit některé zákonitosti, které se nachází v historických datech k danému tématu.

Pro termín **“získávání znalostí z databází”** se v literatuře můžeme setkat s řadou alternativních názvů a definic, v angličtině například Information harvesting, Data distilery a další. Český překlad se s těmito názvy někdy až kuriozně popral jako: Rýpáni se v datech, Datokopectví. Tento fakt je zapříčiněn tím, že tato vědní disciplína prošla prudkým vývojem a definitivní koncept se ustálil až v poslední době. Z anglického **Datamining**, které bylo původně pouze jednou z fází získávání znalostí, se tato hornická metafora stala přeneseným významem a je tedy možné ji brát jako synonymum pro celý proces.

1.3 Potencionální aplikace

Intuitivně lze oblasti aplikace získávání znalostí hledat tam, kde víme, že se hromadí data s nějakou vypovídající schopností a zároveň je informace v nich skrytá atraktivní pro danou oblast. Obecně jsou to **trh a marketing, pojišťovnictví a analýza rizik, medicína, bezpečnost.**

- **Analýza nákupního košíku -** Úloha se zabývá nalezením společně prodávaného zboží, **frekventovaných vzorů**. Pokud například zjistíme, že na jedné účtence se významně často společně vyskytuje pivo a dětské pleny, vyplývá z toho například zjištění, že pro těžký balík plen je vyslán tatínek, který si s oblibou po cestě přibere i pro sebe pivo. Obchodník má pak k dispozici podklady pro rozhodnutí, jak má uspořádat jednotlivá oddělení, aby tatínkovo rozhodnutí podpořil. Z frekventovaných vzorů se pak generují asociační pravidla.

- **Finanční a riziková analýza -** Zabývá se dolováním znalostí ohledně profilu zákazníka, kdy je třeba jistá forma predikce chování zákazníka. Například jak bude reagovat na různé formy reklamní kampaně, nebo naopak zda je riziko, že nebude splácet hypotéku.

- **Biologická a klinická data -** Dolování z biologických dat je velmi široká oblast, která zahrnuje případy od testování hypotéz, analýzy klinických případů až po molekulární biologii a genetiku. Touto oblastí dolování dat se zabýva obor **Bioinformatiky**, který je zde s dataminingem úzce spojen.

- **Dolování v datových proudech a webu -** Zcela speciální oblastí pro dolování z dat jsou datové proudy. Největší výzva zde spočívá v tom, že veškeré dolování se musí odehrávat pouze v okamžiku jednoho průchodu daty. Jde typicky o úlohy jako analýzy telefonních hovorů či záznamů z dohledových kamer. V případě dolování z webu a z textu obecně jde o vyhledávání a analýzu klíčových slov a následnou klasifikaci dokumentů, například podle tématu o kterém pojednávají.

1.4 Datové zdroje

Než se budeme věnovat samotnému procesu dolování dat, je třeba se zabývat jistými prekvizitami a nástroji, které s efektivním dolováním úzce souvisejí. Především pak systémy, které jsou schopny poskytovat kvalitní datovou podporu pro typické dataminingové operace. V praxi se velmi často doluje z dat z různých zdrojů (operační databáze, údaje od zákazníků ...) a ty mohou trpět různou formou šumu a nekonzistentnosti. Lze sice říci, že dolování lze provádět nad jakýmikoliv daty, nicméně úspěch je úzce spjat s kvalitou dat. Odpovědí na naše nároky jsou nástroje pro kolekci, zpracovávání a následnou strukturovanou prezentaci dat vhodnou pro zpracování dolovacími algoritmy - **datové sklady**. Při ukládání do datového skladiště prochází data čištěním a transformacemi a obecně lze říci, že mají vyšší kvalitu a větší vypovídací schopnost.
Data v datových skladech mají následující vlastnosti:

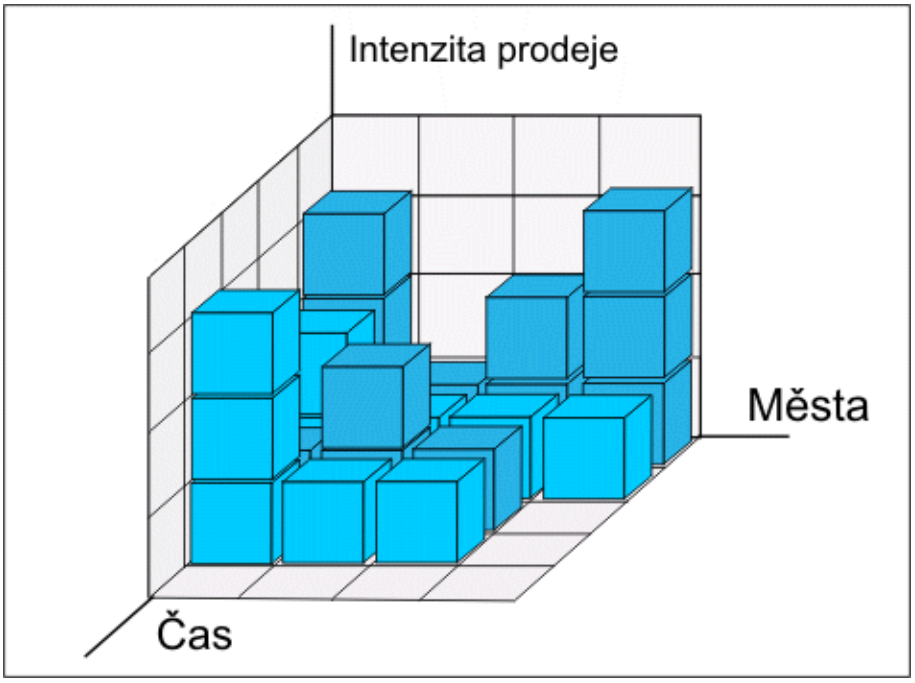
- **Integrovaný** - Extrakce a integrace dat do heterogenního formátu, i když data původně pocházejí z různých zdrojů.
- **Dlouhodobá data**- Archivace dat za větší časové periody, řádově roky.
- **Perzistence** - Po převedení dat do skladiště, se už data needitují, jsou konečná.
- **Subjektová orientace** - Data jsou organizována podle konkrétních konceptů jako zákazníci, prodeje, výrobky.

1.4.1 Multidimenzionální datový model

Od tabulek či objektů v OTLP (operační databáze) se zde dostáváme k jinému modelu pro uchovávání dat - **multidimenzionální kostce**. Dimenze reprezentují aspekty, ze kterých mohou být nahlížena data, typickým příkladem je čas, geograficka poloha a typ výrobku jehož prodej sledujeme. V tomto případě máme pro jednoduchost pouze tři dimenze a výsledek si tedy lze představit jako trojrozměrnou kostku, viz obr. 1.1.

1.4.2 Operace nad multidimenzionální kostkou

V takto organizované datové struktuře jsou třeba operace, které nám umožní lepší pohled na data. Respektive se stávajícím pohledem, určeným multidimenzionální kostkou, pracovat a přízpůsobovat ho pro potřeby analýzy.



Obrázek 1.1: Ukázka uložení prodejů v jednotlivých městech v průběhu času

- **ROLL UP** - Procházení konceptuální hierarchie v dimenzi směrem nahoru. Zvyšování agregace hodnot. Pokud se přesuneme z úrovně například prodejů za týden na úroveň prodejů za měsíc.
- **DRILL DOWN** - Opačný postup než u ROLL UP, zavrtáváme se do detailů dat v dimenzi a snižujeme agregaci.
- **SLIDE & DICE** - Operace, která je synonymum pro výběr. Výsledkem je ořezání dat až na ty, která nás zajímají. Vznikne podkostka, která v dimenzích obsahuje pouze data vyhovující podmínkám. Například prodeje pouze z určitého regionu v určitém měsíci.
- **PIVOT** - Geometrická rotace ve smyslu změny orientace dat na osách. Nijak do dat nezasahuje. Používá se pro vizuální úpravu například 2D grafů.

1.5 Proces získávání znalostí

1.5.1 Fáze

Z hlediska práce s daty rozeznáváme z procesu dolování následující fáze, mezi kterými se iterativně prochází. V závislosti na dílčích výsledcích se fáze mohou opakovat nebo naopak vypouštět.

1. **Porozumění problému** - Analýza situace, definování problému, nalezení aktérů, stanovení zisků a případných rizik, vypracování projektového plánu.
2. **Porozumění datům**- Vytvoření konceptu, záměru ohledně toho, jaké konkrétní zjištění by pro náš projekt byla přínosná. Obstarání vhodných dat, o kterých si myslíme, že by mohly obsahovat odpovědi na naše otázky a porozumět jejich struktuře a povaze.
3. **Příprava dat** - Transformace dat do vhodné struktury, čištění dat od vychýlených a chybějících hodnot, selekce cílových dat pro samotný proces.
4. **Datamining** - Jádro procesu, výběr a aplikace nejvhodnější techniky, sestavení modelu.
5. **Vyhodnocení** - Transformace a vyhodnocení výsledků, učinění patřičných závěrů.
6. **Nasazení** - Promítnutí závěrů to praxe, monitoring odezvy a následné shrnutí úspěšnosti celého projektu.

Posloupnost je řazena z demonstračních důvodů tak, aby na sebe její prvky logicky navazovaly. V praxi je však obvyklé, že se některé fáze slučují.Například při čištění dat je třeba upravená data někam ukládat a je tedy praktičtější provádět rovnou i jejich integraci. Operace pro přípravu dat se fyzicky realizují již při ukládání dat do datových skladů.

1.5.2 Metodiky

Snahy o podporu efektivity tohoto procesu daly vzniknout řadě metodik, které navíc poskytují uživateli pevný rámec pro usnadnění řešení dolovacích úloh. Za některými metodikami stojí přední firmy na poli softwarových řešení v této oblasti a každá uplatňuje trochu jiný pohled na problematiku.

- **Metodika 5A** - Vznikla na půdě firmy SPSS, a svůj název získala podle pěti fází ze kterých se skladá.
  - Assess - posouzení potřeb projektu
  - Access - shromáždění dat
  - Analyze - provedení analýz
  - Akt - přeměna znalostí na plán potřebných změn
  - Automate - zavedení znalostí do praxe

- **Metodika SEMMA** - Používaná softwarovými produkty firmy SAS, název je opět zkratkou jednotlivých kroků procesu.
  - Sample - výběr vhodných objektů
  - Explore - prozkoumání struktury dat
  - Modify - datové transformace
  - Model - analýza dat pomocí umělé inteligence, metod strojového učení
  - Assess - zhodnocení modelu a interpretace závěrů

1.6 Typické dataminingové úlohy

1.6.1 Dolování asociačních pravidel

Úloha se soustředí na získávání asociací - souvislotí mezi daty. Typicky se pak jedná o již zmíněnou analýzu nákupního košíku. Prodejce tak získá informace, podle kterých si může udělat lepší představu o tom, jaké výrobky jsou nakupovány dohromady a tak například lépe uspořádat zboží v prodejně nebo vytipovat jednotlivé skupiny nakupujících a lépe se přizpůsobit jejich potřebám. Pokud máme k dispozici účtenky, které reprezentují seznamy nakupovaných věcí. Můžeme stanovit, podle toho zda se výrobek v košíku vyskytuje nebo ne, například následující asociační pravidlo:

$$MOUKA \wedge VEJCE \Rightarrow KVASNICE$$

Jestliže si zákazník koupí mouku a vejce, s velkou pravděpodobností si koupí také kvasnice. Bude se asi jednat o nákup surovin na pečení.V praxi se nicméně nemusí jednat jenom o nákupy a košíky, ale lze dolovat souvislosti mezi událostmi, hodnotami v různých procesech a podobně. Nyní se zběžně podíváme na algoritmy a principy, jakými se z dat získávají frekventované vzory (často opakující se výrobky), ze kterých se pak asociační pravidla tvoří.

Algoritmus Apriori

Princip činnosti spočívá ve dvou krocích.

1. Krok - **generování kandidátů** spojením, na principu spojení podmnožim. V prvním kole jsou kandidáti všechny prvky databáze.
2. Krok - **eliminací kandidátů**, které se nevyskytují požadovaně často

Tyto dva kroky se opakují v každém kole. Algoritmus končí, pokud pro kandidáty aktuálního kola již nemá podporu, výsledkem jsou pak podporovaní kandidáti předchozího kola.

Nevyhnutelnými úkony jsou pak procházení celé databáze při počítání výskytu (podpory) kandidátů a zátěž systému při samotném generování kandidátů. Oba fakty jsou o to závažnější, že počty kandidátů můžou být obrovské, větší než rozsah původní databáze. Proto se tento alg. dočkal úprav do mnoha jiných verzí, které jeho nejbolavější místa odstraňují, podrobněji zde [1].

FP strom

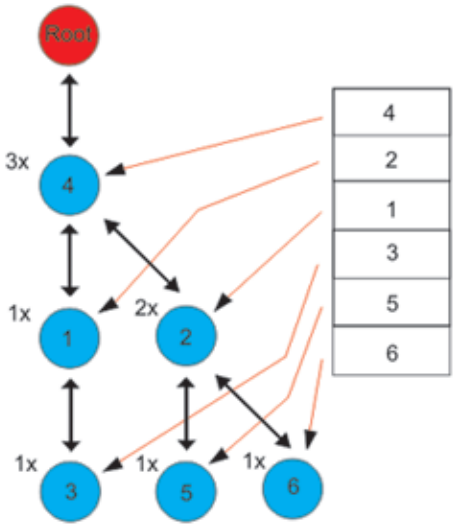
Výrazně efektivnější algoritmus. Uspořádání frekventovaných množin do struktury stromu při jednom průchodu databáze.

1. Krok - stejný jako u Apriori alg. - získání požadovaně podporovaných prvků
2. Krok - uspořádání podle podpory, konstrukce FP stromu

Pro každou frekventovanou množinu je založena větev. Některé větve mohou mít společné prefixy (počátky), v případě, že obsahují společné frekventované podmnožiny. Výsledkem je pak samotný strom, respektive jeho nejdelší větve, viz obr. 1.2. Pro rychlejší průchod uchováváme tabulku odkazů na počáteční uzly.

Příklad:

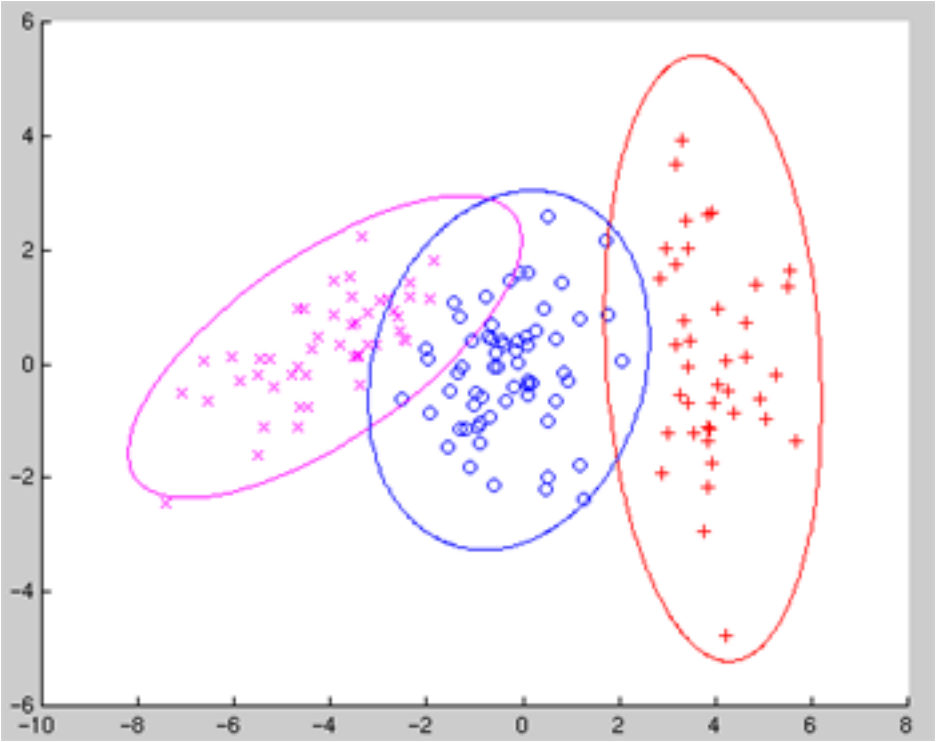
{ 1, 3, 4 }
{ 2, 4, 5 }
{ 2, 4, 6 }



Obrázek 1.2: Ukázka konstrukce FP stromu pro uvedené záznamy

1.6.2 Shlukovací analýza

Proces, který vyhledává podobné vlastnosti v datech a rozděluje je podle nich do tříd. Cílem je, aby si prvky ve stejné třídě byly co nejvíce podobné a zároveň, aby se co nejméně podobaly prvkům z jiných tříd. Měřítko podobnosti často bývá **vzájemná vzdálenost**, odtud také pochází název úlohy. Metoda pracuje pouze s množinou vzorků, kterou má zpracovávat, nepotřebuje žádné doplňující informace. Jedná se o druh metody **bez učitele**.



Obrázek 1.3: Výsledek aplikace shlukovací metody

Shlukovací analýzu využíváme všude, kde jsou v datech potřeba indetifikovat jisté podobné skupiny, například v bioinformatice, při rozpoznávání vzorů v multimédiích (objekty na fotografii) nebo jako stupeň předzpracování pro jiné dataminingové úlohy - Klasifikace. K dispozici je mnoho variant algoritmu. Ty se liší principem, na kterém pracují nebo vlastnostmi a tím pádem vhodností pro různé povahy dat, povahy aplikací či výpočetní náročností.

Metody založené na rozdělování

Lze aplikovat na N objektů pro rozdělení do M tříd. Na začátku je nutné vědět počet tříd, do kterých má algoritmus objekty rozdělovat. Principem činnosti je vyhledávání ohniska, objektu, který nejlépe vyhovuje vzdálenostním funkcím ostatních objektů.

1. Krok - Náhodné stanovení ohnisk všech tříd.
2. Krok - Iterativně vyhledávat lépe vyhovující ohniska a vhodně přeskupovat objekty.

Algoritmus končí, pokud nezbyly žádné objekty, které by bylo výhodnější přesunout.Pro tuto úlohu jsou využívany dvě hlavní heuristiky.

- **Metoda centrálního středu**- Známá též jako K-MEANS. Jako ohnisko je fiktivní bod, střed, jehož poloha se odvozuje pomocí středních hodnot atributů objektů určité třidy.
- **Metoda reprezentujícího objektu** - Metoda využívá podobného principu, s tím rozdílem, že jako ohnisko je vybrán jeden z bodů.

Obě metody tedy pracují na podobných principech a mají podobné vlastnosti. Jsou vhodné především pro menší a střední dobře ohraničené shluky kulatých tvarů. Metoda reprezentujícího bodu má větší odolnost proti šumu a odlehlým hodnotám, ale zato je kvůli výpočtům zvolení nového ohniska náročnější. Nevýhodou obou metod zůstává stanovení počtu tříd před samotnou analýzou a nevhodnost pro komplikovanější tvary. Shlukovací metody založené na jiných principech nalezneme například zde [1].

1.6.3 Predikce

Metoda, která posuzuje data podle jejich atributů a přiřazuje jim hodnoty obecně spojitého charakteru. Metoda pracuje na bázi učení s učitelem. Je třeba ji poskytnout ohodnocená trénovací data a potom ji aplikovat na jiná data s podobnou nebo stejnou strukturou. Pomocí predikce můžeme například řešit úlohy typu předpovídání úrody na určité zemědělské regiony, kdy známe místní historické hodnoty počasí (jako teploty, srážkové úhrny) a výslednou úrodu a podle letošních hodnot bychom rádi předpověděli objem té letošní. Mezi metody, se kterými se zde setkáváme nejčastěji patří lineární a vícenásobná lineární regrese, ale i jiné viz. [2].

- **Lineární regrese** - Vektor dat, atributů $x_1, x_2, \ldots, x_n$ a hodnot atributů $y_1, y_2, \ldots, y_n$, aproximuje rovnicí přímky pomocí metody nejmenších čtverců.

  $$Y = kX + q$$

  Množinu trénovacích dat metoda využije na nastavení koeficientů k, q. Predikce se provádí dosazením hodnoty atributů do naučeného klasifikátoru.

- **Vícenásobná regrese** - Využívá stejného principu predikce, pouze zobecněného pro N atributů.

1.6.4 Klasifikace

Klasifikace je rovněž jedna z velmi důležitých úloh a je tedy rovněž uvedena ve výčtu, ale z hlediska zaměření této práce se jí bude velmi podrobně zabývat následující kapitola.

Kapitola 2

Klasifikace

2.1 Úvod

2.1.1 Proces klasifikace

Klasifikace dat je rozdělování objektů do skupin, tříd podle jejich atributů. Pod pojmem N-nární klasifikace rozumíme rozdělování do N tříd, avšak $N < \infty$. Klasifikace je metoda **učení s učitelem**, klasifikátor konkrétní klasifikační metody se naučí na množině dat, která rozdělení do tříd již obsahuje, a pak je schopný již sám klasifikovat data podobné povahy.

1. **Učení klasifikátoru** na množině dat, u kterých známe třídu.
2. **Validace klasifikátoru**, ověření úspešnosti učební fáze. Klasifikátor zpracuje data, u kterých mu zatajíme, do které třídy patří, a pak porovnáme úspěšnost, respektive stanovíme chybu. Validační data nesmí být podmnožinou trénovacích dat.
3. **Klasifikace** ostrých dat, které mají stejnou povahu a strukturu jako validační a trénovací, ale nejsou identická.

Představme si lékařskou kliniku, kde jsou pacienti před vyšetřením dotazování na údaje o jejich celkovém zdravotním stavu, životním stylu nebo životosprávě a zda někdy prodělali srdeční příhodu. Z takto získané databáze je klasifikátor schopný izolovat skupinu lidí, u kterých je vyšší riziko infarktu myokardu. U každého nově příchozího pacienta může být, na základě zjištění potřebných údajů o jeho aktuálním stavu, stanoveno, jestli patří do rizikové skupiny před tím, než zaznamenal jakékoliv potíže se srdcem a tato informace může být využita lékařem například k lepšímu adresování prevence.

2.1.2 Vlastnosti klasifikátoru

Kvalita klasifikátorů se odvozuje hlavně od dosahovaných výsledků. Pokud některý model nevyhoví například ve fázi validace, může být chyba buď v datech, ve vhodnosti modelu pro danou úlohu, nebo ve formulaci celé dolovací úlohy. Kvalita klasifikátoru se dále pak hodnotí z těchto hledisek.

- **Přesnost** - Schopnost odhalit podstatné vazby v datech a na základě nich provést co nejúspěšnější klasifikaci.
- **Srozumitelnost** - Jednoduše interpretovatelné, na informace převeditelné výsledky.
- **Rychlost** - Nízká časová náročnost fází, zejména pak učení.

- **Stabilita** - Nízká chybovost výsledků vzhledem k šumu v datech nebo rozsáhlosti databáze.

2.2 Rozhodovací strom

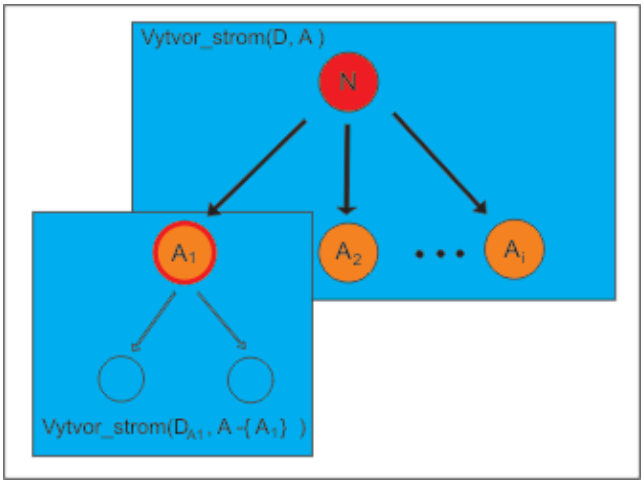
2.2.1 Algoritmus

Způsob prezentace dat formou rozhodovacího stromu je znám z mnoha oblastí lidské činnosti. Princip indukce rozhodovacích stromů byl převzat z metod strojového učení. Postupujeme stylem rozděl a panuj. Trénovací data postupně, v závislosti na hodnotách atributů, rozdělujeme na menší a menší celky tak, aby v jednotlivých uzlech převládaly data stejných znaků. V nejnižší úrovni, listech, jsou data již rozdělená do tříd. Počínaje kořenovým uzlem provádíme analýzu od shora dolů, postupnou specializaci atributů tříd.

Množina klasifikovaných dat D, Množina užitých atributů A.

VytvorStrom(*D*, *A*)

1. Vytvoř nový uzel N.
2. Skonči a vrať uzel N jako list dané třídy, pokud všechny data z množiny D jsou ve stejné třídě.
3. Skonči, pokud je seznam atributů A prázdný.
4. Vyber a odstraň **vhodný** atribut A_i z množiny A a pojmenuj podle něj uzel N.
5. Pro každou hodnotu atributu A_i opakuj:
   (a) Vytvoř větev z uzlu N.
   (b) Vytvoř podmnožinu množiny D, obsahuje pouze zvolenou hodnotu v atributu A_i.
   (c) Pokud je množina prázdná, pak spoj list s nejběžnější třídou v množině D.
   (d) Pokud je množina neprázdná, rekurzivně volej algoritmus VytvorStrom (viz. obr. 2.1) pro tuto podmnožinu dat a množinu atributů A, ve kterém budou chybět všechny A_i podle, kterých jsme již klasifikovali.



Obrázek 2.1: Výstavba rozhodovacího stromu

2.2.2 Výběr atributu

V předchozí kapitole je atribut podle kterého se pojmenuje uzel popsán jako **vhodný**. V této chvíli se podíváme, co to znamená a jak se taková vhodnost stanoví.

Z hlediska vytváření rozhodovacího stromu, jsou v hieratické struktuře nejvýše umístěné uzly, které mají na podřízené uzly největší vliv, respektive nejvíce odlišují příklady různých tříd. Nabízí se analogie z praxe, kdy v určovacím atlase rostlin, který je také vlastně jenom druh rozhodovacího stromu, též postupujeme od nejdůležitejších faktů k detailům.

Rozlišovací schopnost atributu A určíme podle hodnot funkce informačního zisku **GAIN(A)**. Čím vyšší zisk, tím větší rozlišovací schopnost. Jak vidíme ze vzorce, je složen z rozdílu dvou funkcí a říká, jak se redukuje celková entropie dat výběrem jednoho atributu.

$$GAIN(A) = H(C) - H(A)$$

Celková entropie dat:

$$H(C) = -\sum_{t=1}^{T} p_i log_2 p_i = -\sum_{t=1}^{T} \frac{n_t}{n} log_2 \frac{n_t}{n}$$

Kde p_i je pravděpodobnost, že náhodný vzorek z datové množiny bude klasifikován do i-té třídy, též se dá na veličinu pohlížet jako na relativní četnost i-té třídy. $H(C)$ je entropie pro daný atribut vzhledem k celým datům.

$$H(A(v) = -\sum_{t=1}^{T} \frac{n_t(A(v)}{n(A(v)} log_2 \frac{n_t(A(v)}{n(A(v)}$$

$$H(A)) = -\sum_{Val(A)} \frac{n(A(v)}{n} H(A(v))$$

Kde $H(A(v))$ je hodnota entropie pro každou hodnotu v atributu A nad daty, kde se vyskytuje $A(v)$. $H(A)$ je pak střední hodnota váženého součtu $H(A(v))$.

2.2.3 Optimalizace

V praxi se setkáváme s tím, že klasifikovaná data obsahují šum nebo odchýlené hodnoty, nebo prostě svou povahou (hodně blízkých hodnot atributů) stěžují indukci rozhodovacího stromu. Ať už je příčinou kvalita nebo povaha dat, může dojít k tomu, že algoritmus produkuje velké množství větví a strom je příliš košatý, nebo dokonce nevhodně postavený. Aby se předcházelo těmto nežadoucím efektům, byly do algoritmů pro indukci rozhodovacích stromů zabudovány metody pro jejich optimalizaci nebo-li takzvané **ořezávání** viz. [2].

- **Prepruning** - Analýza probíhá již při vytváření stromu. Algoritmus se rozhodne zda nepotřebný podstrom nenahradí listem, či méně časteji i naopak.
- **Postpruning** - Odstraňování nevýznamných větví až po dokončení stromu, na což se musí čekat a proces se zpomaluje. Zato má tato metoda přehled o celém (hotovém) stromu a je tedy schopná kvalifikovaněji podstromy posuzovat.V praxi se proto používá buď kombinace obou metod nebo Postpruning

2.3 Neuronová síť

Dalším z klasifikátorů, o kterém je třeba se zmínit, je neuronová síť. Jeden ze známých principů metod umělé inteligence. Využívá model fungování lidského neuronu, který ve spojení s mnoha jinými neurony tvoří síť. Z hlediska dobývání znalostí představují neuronové sítě jeden z nejmocnějších a nejpoužívanějších systémů, na kterém jsou postaveny nástroje pro klasifikaci a predikci. Silnou stránkou je práce s numerickými atributy a na tomto poli představují alternativu k rozhodovacím stromům.

2.3.1 Neuron

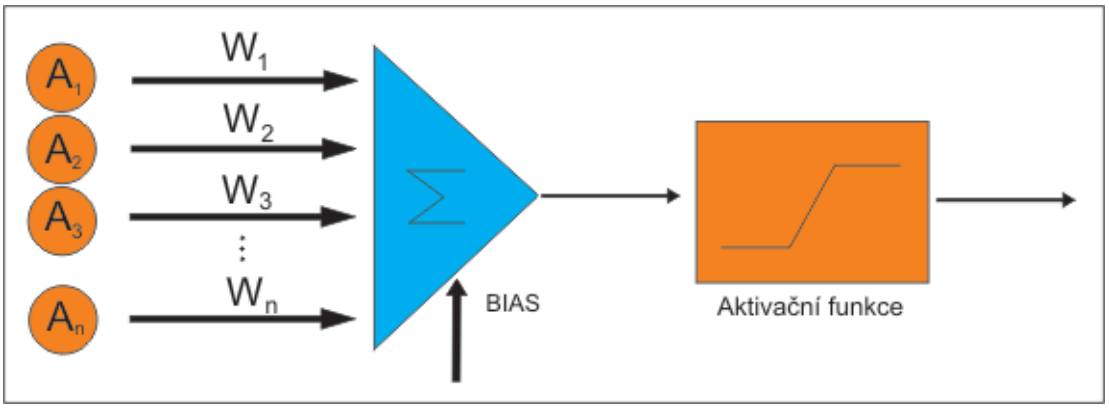
Lidský neuron se skládá z těla a výběžků, krátkých dostředivých **axonů** a jednoho dlouhého odstředivého **dentridu**. Axony jsou z jiných neuronů přiváděny informace. Pokud intenzita přenášeného vzruchu dosáhne jisté hranice, informace je po zpracování vyslána dentridem k jinému neuronu. Model používaný pro naše účely pracuje v zásadě podobně.

Na vstupy umělého neuronu jsou přivedeny hodnoty atributů dat $x_1, x_2, x_3, \ldots, x_n$. V tělě neuronu jsou zpracována tak, že informace, které nesou jednotlivé x_i, jsou ohodnocena váhami $w_1, w_2, \ldots, w_n$. Součin vah a vstupů se sčítá a k výsledné sumě je ještě přičten BIAS (konstanta) konkrétního neuronu viz. obr. 2.2.

$$\sum_{i=1}^{n} w_i x_i + BIAS$$

Aktivační funkce vhodným nelineárním způsobem transformuje součet podnětů. Výstup je potom veden do vstupu neuronu v další vrstvě nebo rovnou vyhodnocen. Nejpoužívanější aktivační funkcí v souvislosti s tímto typem neuronu je funkce:

$$y = \frac{1}{1+e^{-x}}$$



Obrázek 2.2: Schéma neuronu

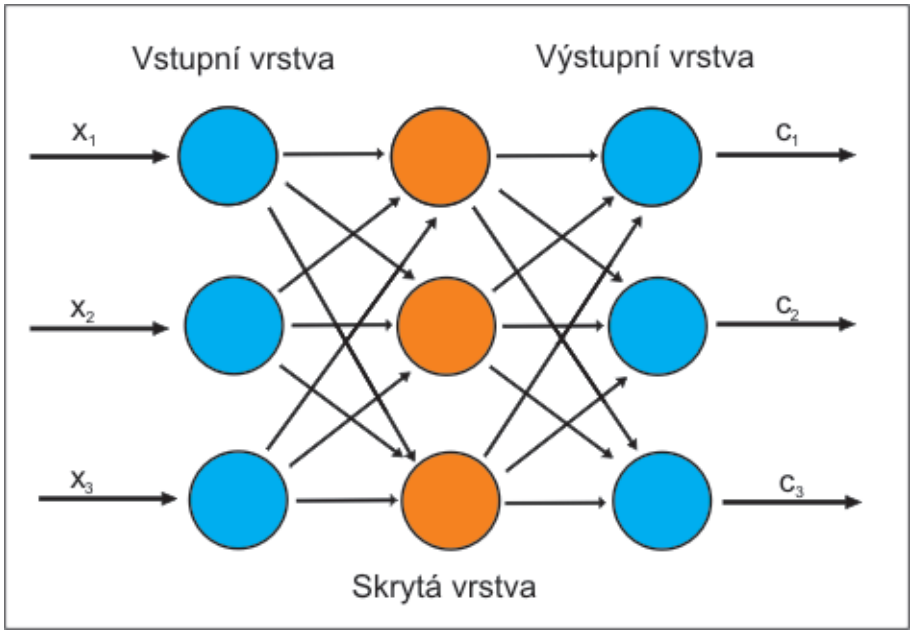
Počáteční nastavení neuronu je libovolné. Učení probíhá iterativně stanovováním chyby a následné úpravy hodnot vah a biasu.

2.3.2 Neuronová síť Backpropagation

Topologie

Spojením více samostatných neuronů tak, že výstup z neuronu *i* přivedeme opět jako vstup do neuronu $i+1$ získáváme strukturu, která sčítá rozhodovací schopnost samostatného neuronu. Všechny neurony uspořádáme do vrstev a propojíme je tak, že budou spojeny systémem každý s každým v rámci sousedních vrstev, ale neurony v rámci stejné vrstvy zůstanou nespojeny.

První vrstva, takzvaná **vstupní**, sama o sobě data nezpracovává, jenom zajišťuje distribuci mezi ostatní neurony sousední vrstvy. Klasifikace se odehravá zejména ve **skryté** vrstvě, kterých může být obecně libovolný počet, nicméně nejpoužívanejší topologie sítě backpropagation má právě jednu. Výsledek je vyhodnocen **výstupní vrstvou**. Výsledky klasifikace pro aktuální data získáváme z poslední, **výstupní** vrstvy viz. obr. 3.2.



Obrázek 2.3: Schéma neuronové sítě

Učení sítě Backpropagation

Jak už bylo zmíněno výše, učení neuronové sítě spočívá v nalezení správných hodnot vah pro neuronové vstupy a hodnot konstatnt (biasů) pro samotné neurony. V první iteraci jsou všechny hodnoty nastaveny libovolně. Pro tento jistě zmatený výstup se spočítá chyba neuronů ve výstupní vrstvě a podle ní zpětně stanovujeme chyby neuronů v ostatních vrstvách (error backpropagation). Cílem je zminimalizovat rozdíly mezi hodnotami ve výstupním vektoru a mezi požadovaným výsledkem, dokud se chyba neustálí na nějaké minimální hodnotě, nedosáhneme požadovaný počet iterací nebo již nedochází k žádným úpravám hodnot vah a biasů.

Opakujeme pro všechny vzorky X množiny S dokud síť nedosáhla jedné z výše uvedených podminek:

1. Pro každý neuron j spočítej hodnoty:

$$I_j = \sum w_i j O_i + BIAS_j$$

$$O_j = \frac{1}{1+e^{-I_j}}$$

Kde O_i je vstup do neuronu j z neuronu i nebo vstupního vektoru a $w_i j$ je váha vstupu mezi neuronem i a j. O_j je výstup aktivační funkce neuronu j.

2. Všechny neurony znají teď své výstupy, které jsou porovnány s vektorem správných výsledků T_j a nastává fáze distribuce chyby.

   Pro každý neuron j ve výstupní vrstvě spočítej:

   $$Err_j = O_j(1-O_j)(T_j-O_j)$$

   Kde O_j je hodnota z vektoru aktuálních výsledků a T_j hodnota z vektoru správných výsledků.

   Pro chybu neuronů dalších vrstev počítej:

   $$Err_j = O_j(1-O_j)\sum_k Err_k w_j k$$

3. Uprav váhy a BIASy všech neuronů:

   $$w_i j = w_i j + l(Err_j O_i)$$
   $$BIAS_j = BIAS_j + l(Err_j)$$

   Kde číslo l je koeficient učení, reálné číslo z intervalu $\langle 0,1\rangle$. Vyjadřuje do jaké míry se bude neuronová síť přizpůsobovat aktuálním vzorkům. Při vysokém koeficientu se rychle učí, zato se může stát, že se naučí klasifikovat výhradně vzorky z trénovací množiny. Kvalitní implementace klasifikátorů tento koeficient v průběhu učení dynamicky mění.

Kapitola 3

Porovnávání modelů pro klasifikaci

3.1 Výzkumný záměr

Dolování z dat se stává přirozenou součástí procesu rozhodování a řízení obecně. Datamingové nástroje jsou běžnou výbavou pracoviště každého analytika. Před potřebu efektivně zanalyzovat svá data jsou stavěni i lidé, kteří nemají žádné odborné znalosti, a přesto mohou být díky moderním nástrojům integrovaných do celých analytických prostředí úspěšní.

Rozhodli jsme se podívat blíže na některé vlastnosti modelů, které jsou pro úspěšné dolování z dat v běžné praxi klíčové. Hlavně jsme se zaměřili na různé aspekty metod pro klasifikaci. K dispozici je vice klasifikátorů různých vlastností (logická regrese, neuronová síť, rozhodovací strom). Běžný uživatel dataminingovýh nástrojů (manažer, analylitik) není a vlastně ani nemusí být do hloubky seznámen s principy jednotlivých klasifikátorů . Jeho prioritou zůstává užitá hodnota, kterou vydolováním určité informace získá. Našim cílem je tedy porovnat v praxi nejběžněji používané klasifikátory a z jejich vlastností stanovit, který plní, v závislosti na typu úlohy, požadavky běžného uživatele nejlépe.
Pořadavky běžného uživatele:

- **Úspěšnost, přesnost klasifikace** - správnost výsledků, ke kterým model dospěje
- Srozumitelnost modelu - použitelnost vydolované informace
- Robustnost algoritmu - schopnost se vypořádat s šumem a chybějícími hodnotami bez výrazné ztráty přesnosti, stability

3.2 Testovací data

Rozhodli jsme se porovnávat výše uvedené vlastnosti klasifikátorů na datových množinách různých vlastností. Zejména nás zajímala přesnost předpovídacích schopností klasifikátorů pro různě složité datasety. Pro naše účely bohatě dostačovaly datasety s velikostí okolo 2000 vzorků s rovnoměrným zastoupením vzorků všech tříd. Používaly jsme datasety ze zdroje:

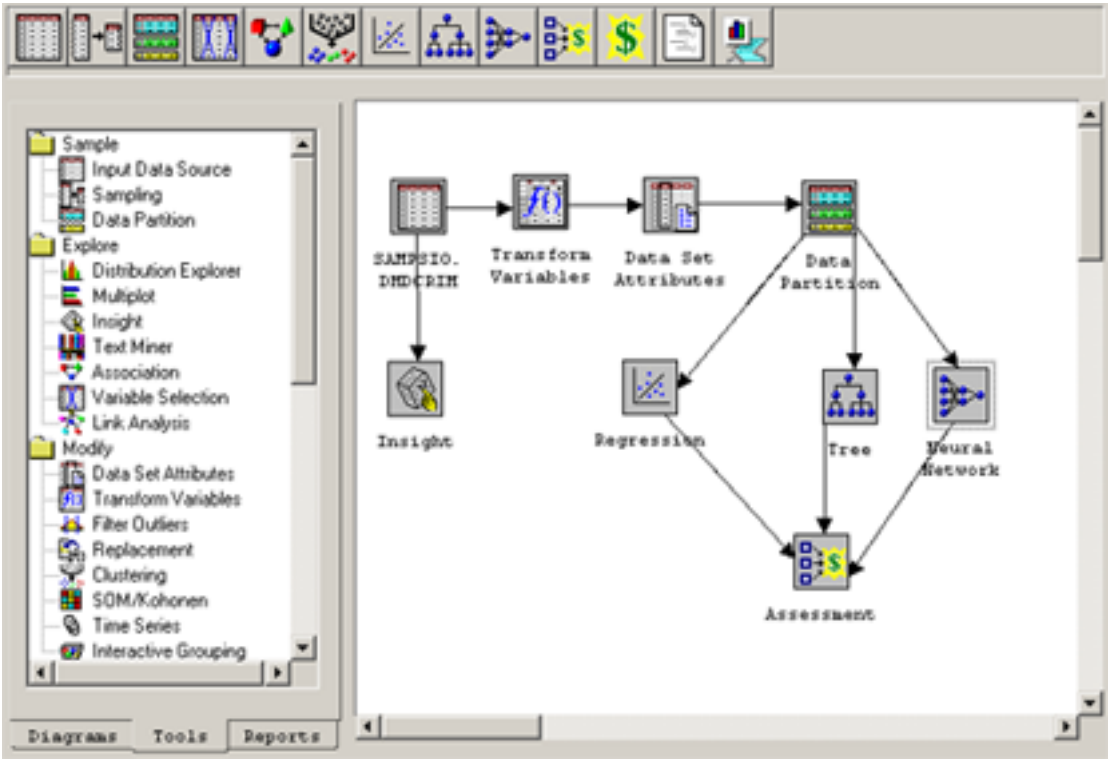
UCI Machine Learning Repository
`http://http://www.ics.uci.edu/~mlearn/MLSummary.html`)

3.3 Prostředí SAS Enterprise miner

Pro vyhodnocování jsme používali **Enterprise miner platformu SAS 9.1** od firmy **SAS Institute**. Firma SAS je jedním z hlavních producentů statistického software. Enterprise miner je obrovský balík, který obsahuje mnoho modelů pro vyhodnocování dat. Implementace je na vysoké úrovní, většinou jsou všechny potřebné datové transformace zabudované do konkrétního datového modelu. Nástroje pro práci s daty jsou členěny do skupin, které odpovídají metodice **Semma**, jejíž autorem je též firma SAS.

Uživatelské rozhraní

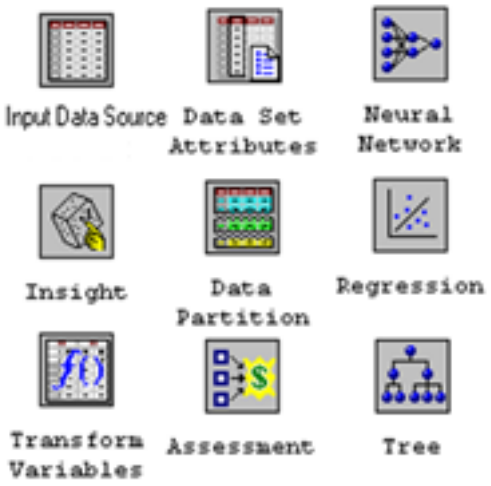
Proces dobývání se pro každou úlohu definuje pomocí **procesního diagramu**, který se skládá z uzlů a propojení mezi nimi. K dispozici je toolbar odkud jednoduše umísťujeme uzly stylem drag and drop na pracovní plochu. Uzly spojujeme šipkami, které určují jak budou data procházet zpracováním a záleží tedy na jejich orientaci.



Obrázek 3.1: Ukázka uživatelského rozhraní SAS Enteprise mineru

Procesní diagram

Každý uzel plní v dolovací úloze nějakou funkci. Při rozkliknutí může uživatel měnit a nastavovat parametry, pokud to ale neudělá, systém zpravidla za něj sám základní modelovací nastavení provede.



Obrázek 3.2: Základní nástroje

- **Input data source** - Výběr a načtení datasetu z knihoven SAS EM, editace počtu použitých vzorků. Startovní uzel.
- **Insight** - Nástroj pro prozkoumání výstupu jakéhokoliv uzlu.
- **Transform variables** - Editace datových typů, pod kterými chceme s daty pracovat. Práce s proměnými.
- **Data set attributes** - Specifikace rolí proměných, určení vstupů a cíle klasifikace. Zamítnutí proměných.
- **Data partition** - Procentuální rozdělění datasetu na trénovací, validační a testovací část.
- **Regresion** - Uzel pro nastavení parametrů klasifikátoru pro lineární regresi.
- **Neural network** - Uzel pro nastavení parametrů klasifikátoru pro neuronovou síť.
- **Decision tree** -Uzel pro nastavení parametrů klasifikátoru pro rozhodovací strom.
- **Assessment** - Vyhodnocovací uzel, zobrazuje výsledky v podobě klasifikačních grafů nebo ziskových křivek.

Vkládání dat

Data se do SAS EM importují buď přímo z datového skladu, nebo za pomocí standarního dialogu **Import Data**. K dispozici je řada formátů kompatibilních se software jako například exel, access, lotus notes. Velkým rozšířením jsou pak **uživatelsky definované formáty**, které rozpoznávají určité základní oddělovače a na základě nich je možné ručně transformovat data do podporovaného formátu.

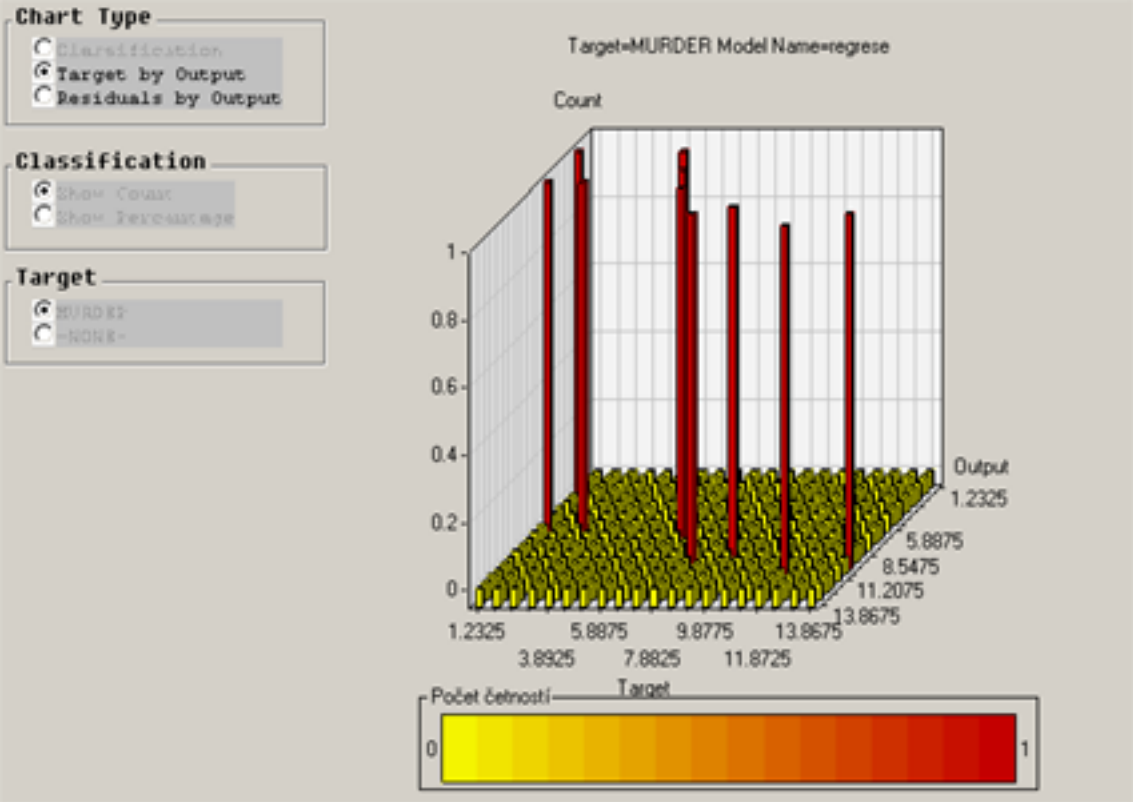
Modelování a zobrazení výsledku

Pro nastavené procesní schéma se modelování spouští kompilací Assessment uzlu, obecně posledního uzlu ve schématu. SAS EM již zajistí kompilaci všech předcházejících uzlů. Pro úspěšnost tohoto procesu je třeba:

- Mít vhodně (kompatibilně) vložená data.
- Označenou alespoň jednu proměnou jako cíl klasifikace.
- Správně zvolené a pospojované uzly.

Výsledky v uzlu assess, ke kterým model dospěl, jsou zobrazeny v podobě dvou různých druhů grafů.

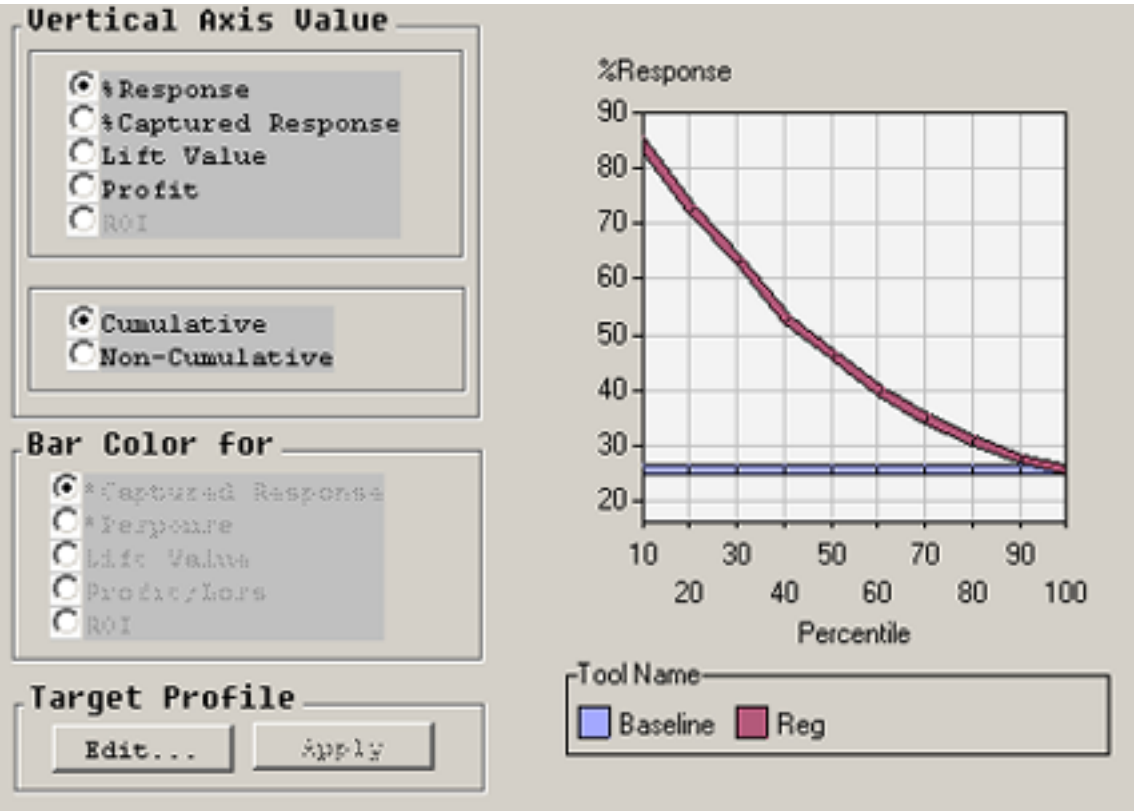
Diagnostický graf klasifikace:



Obrázek 3.3: Diagnostický graf klasifikace - Diagnostic chart

Graf má podobu krychle se základními osami. Na osu X jsou naneseny třídy, kterými jsou ohodnocena data. Na osu Y jsou pak nanášeny třídy, do kterých je zařadil klasifikátor. Na ose Z vidíme množství klasifikovaných prvků. Pokud by byla klasifikace 100% uspěšná, shodovalo by se ohodnocení vzorku z osy X s výsledkem klasifikace z osy Y a graf by měl podobu sloupců pouze v diagonále roviny X - Y. Někdy se stává, že pořadí tříd na ose Y není stejné jako na ose X a výsledná ideální diagonála je deformovaná.

Graf ziskové křivky:



Obrázek 3.4: Graf ziskové křivky - Lift chart

Pomocí grafu ziskové křivky je možné zpracovat pouze výsledky binární proměné. Křivka vyjadřuje závislost mezi procentuální úspěšností klasifikace a počtem vzorků. Například klasifikátor, jehož výsledek je na obrázku (červená křivka), by zvládl mezi 10% vzorků klasifikovat správně 77%. Modrá křivka znázorňuje celkové zastoupení dané třídy v datech. Pro 100% dat se obě křivky logicky potkají.

Zhodnocení

SAS EM je ukázkou softwaru, díky kterému jsou uživatele schopní dobývat informace, které potřebují pro své rozhodování, aniž by museli být experty na problematiku strojového učení nebo databází. Prostředí dovoluje snadno zakládat projekty a formulovat cíle úlohy, které pak prochází kvalitními implementacemi klasifikátorů. Veškeré potřebné předzpracování dat (například diskretizace pro potřeby rozhodovacího stromu) je automatické a uživatel o něm zpravidla ani neví. Na druhou stranu jsou k dispozici uzly pro záměrné předzpracování dat, čištění dat od chybějících a vychýlených hodnot. SAS EM je možné používat samostatně, ale i jako součást celé platformy
SAS business inteligence.

3.4 Porovnávání modelů

K dispozici jsme měli několik klasifikátorů s lehce odlišnými vlastnostmi a zajímalo nás, který z nich si povede lépe na několika skupinách testovacích dat. Pro srovnávání jsme používali klasifikátory: **Neuronová sít, Logická regrese, Rozhodovací strom** z prostředí SAS EM ve standartním nastavení.

- **Neuronová síť**- Je například silnější při práci s daty, které mají numerickou (spojitou) povahu. Naopak diskrétní údaje je třeba binarizovat (přiřadit diskrétním položkám numerické atributy). U neuronových sítí je nebezpečí, že se takzvaně přeučí (ve fázi učení se příliš naučí na prvky z trénovací množiny a jiné prvky pak klasifikují mylně).
- **Rozhodovací strom**- Neumí pracovat s numerickými daty, která je potřeba diskretizovat, což se ne vždy podaří ideálně. Rozhodovací strom se špatně vyrovnává s chybějícími hodnotami u atributů dat, hodnoty je třeba doplnit například nejčastěji vyskytujícím se prvkem v dané datové množině, což je zásah, který také přispívá ke snížení přesnosti výsledné klasifikace.
- **Lineární regrese**- Pracuje též s numerickými daty. Vhodnější pro predikci.

3.4.1 Výběr klasifikátoru s nejlepšími výsledky

Předpoklad

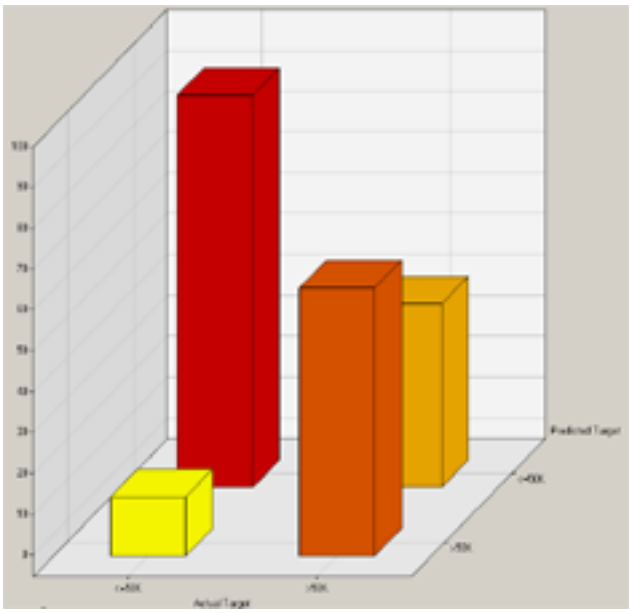
Cílem číslo jedna každého dolování je především přesnost klasifikace a ve výsledku maximalizování užitku ze získané informace. Předpokládali jsme, že **rozdíly** v povaze klasifikátorů budou mít vliv na jejich **úspěšnost** klasifikace.

Dataset

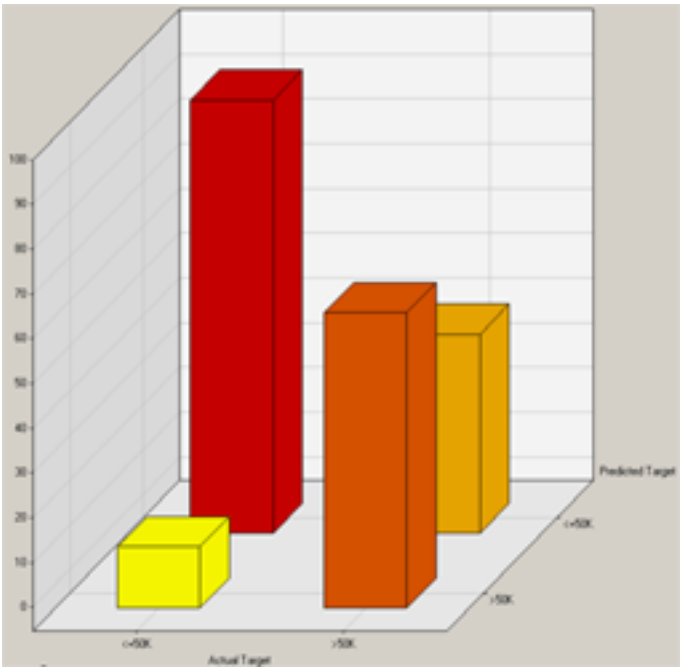
Pro testování klasifikátorů jsme používali datasety různých parametrů - rozsahů, komplikovanosti a sofistikovanosti. Pro test nejvhodnějšího klasifikátoru můžeme, vzhledem k výsledkům, vybrat jeden příklad za všechny (kompletní přehled všech výsledků v podobě grafů lze najít na přiloženém datovém nosiči). Jde o data, která byla nashromážděna mezi pracujícími lidmi v rove 1996. Atributy obsahují údaje jako vzdělání, obor ve kterém dotyčný pracuje, pozici na které pracuje, rodinný stav, zemi, pohlaví, rasu či počet odpracovaných hodin týdně a další údaje (viz příloha), které mají vliv na výši mzdy. Úkolem klasifikátoru je rozčlenit lidi, do dvou tříd podle výše platu - pod a nad 50 tisíc dolarů za časovou jednotku.

Výsledek

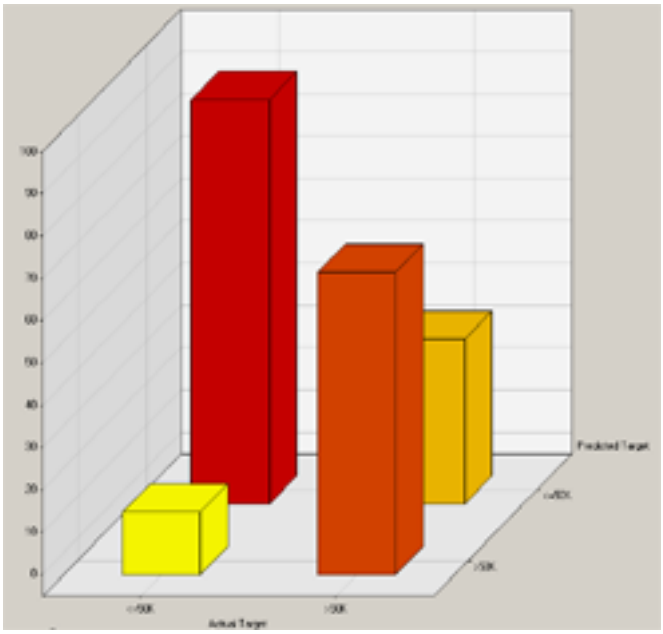
Náš předpoklad o dramatických rozdílech ve výsledcích klasifikátoru pro tento relativně komplikovaný dataset (14 atributů) se nepotvrdil viz. obr. 3.5, 3.6 a 3.7. Téměř se všemi 8-mi porovnávanými datasety měly všechny tři klasifikátory podobnou úspěšnost, respektive podobnou chybu, průměrně kolem 5%. V zásadě s podobnými výsledky jsme se setkávali u všech ostatních zkoumaných datasetů. Z pohledu relativního srovnání však nejlépe klasifikovala neuronová síť, s chybou řádově o jednotky procent nižší než rozhodovací strom a logická regrese. Tato výhoda je však vykoupena podstatně delší dobou učení, která by mohla být při větších objemech dat nevýhodou.



Obrázek 3.5: Diagnostický graf klasifikace pro neuronovou síť



Obrázek 3.6: Diagnostický graf klasifikace pro logickou regresi



Obrázek 3.7: Diagnostický graf klasifikace pro rozhodovací strom

3.4.2 Vliv počtu atributů na klasifikaci

Předpoklad

Původně jsme předpokládali, že chybovost klasifikace určitým způsobem souvisí se složitostí (počtem atributů) databáze. Pokud by se klasifikátor musel rozhodovat na základě širokého spektra atributů, musí vytvořit strukturu, ať už rozhodovacího stromu či neuronové sítě, která bude muset být složitější a tedy i náročnější na kvalitu algoritmu.

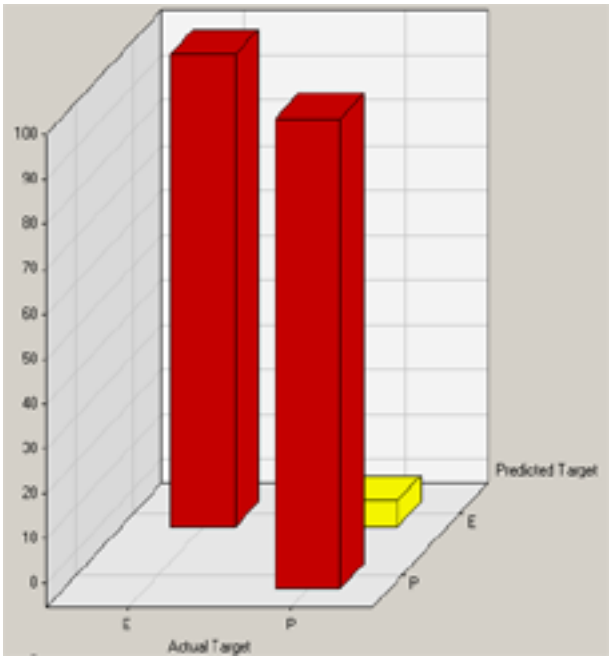
Dataset

Nechali jsme tedy vyhodnotit dva datasety, které se oba vyznačují větším počtem atributů, přitom téměř všechny jsou pro klasifikaci důležité.

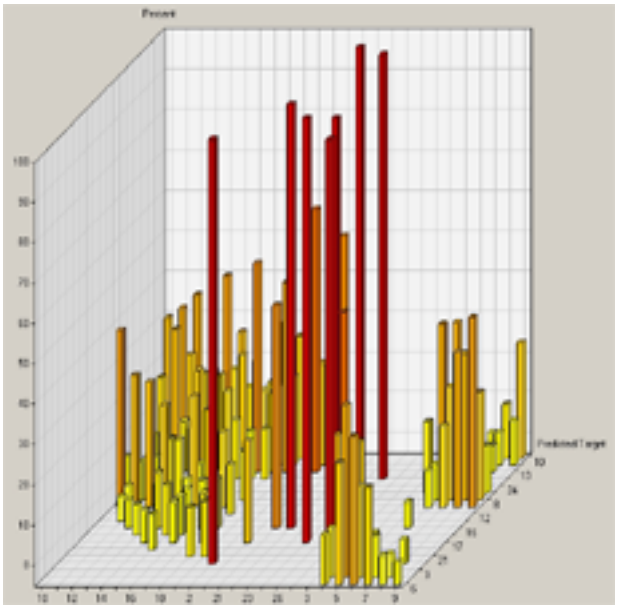
- Databáze hub - Obsahuje asi 22 atributů, které popisují typické znaky, podle kterých se určují druhy hub jako tvar a povrch kloboučku, velikost, tvar a barva nohy nebo barva podhoubí. Úkolem klasifikátoru je určit, zda je houba jedlá nebo nejedlá.
- Databáze mořských škeblí - Databáze jistého druhu mořských ušní, která obsahuje asi 8 rozměrů (atributů), které vznikly měřením různých fyzických částí těla škeble. Předmětem klasifikace je určení stáří (respektive počtu kroužků na těle, což nepřímo prozrazuje stáří u škeble) podle fyzických rozměrů jako jsou průměr, váha a výška schránky, pohlaví..

Výsledek

Porovnáním výsledků těchto dvou datasetů viz. obr. 3.9 a 3.8 jsme dospěli k závěru, že výsledek nezbytně nezávisí přímo na počtu atributů. Všechny modely podaly velmi podobné výsledky, pro jednoduchost jsou zde tedy pouze výsledky modelu logické regrese.



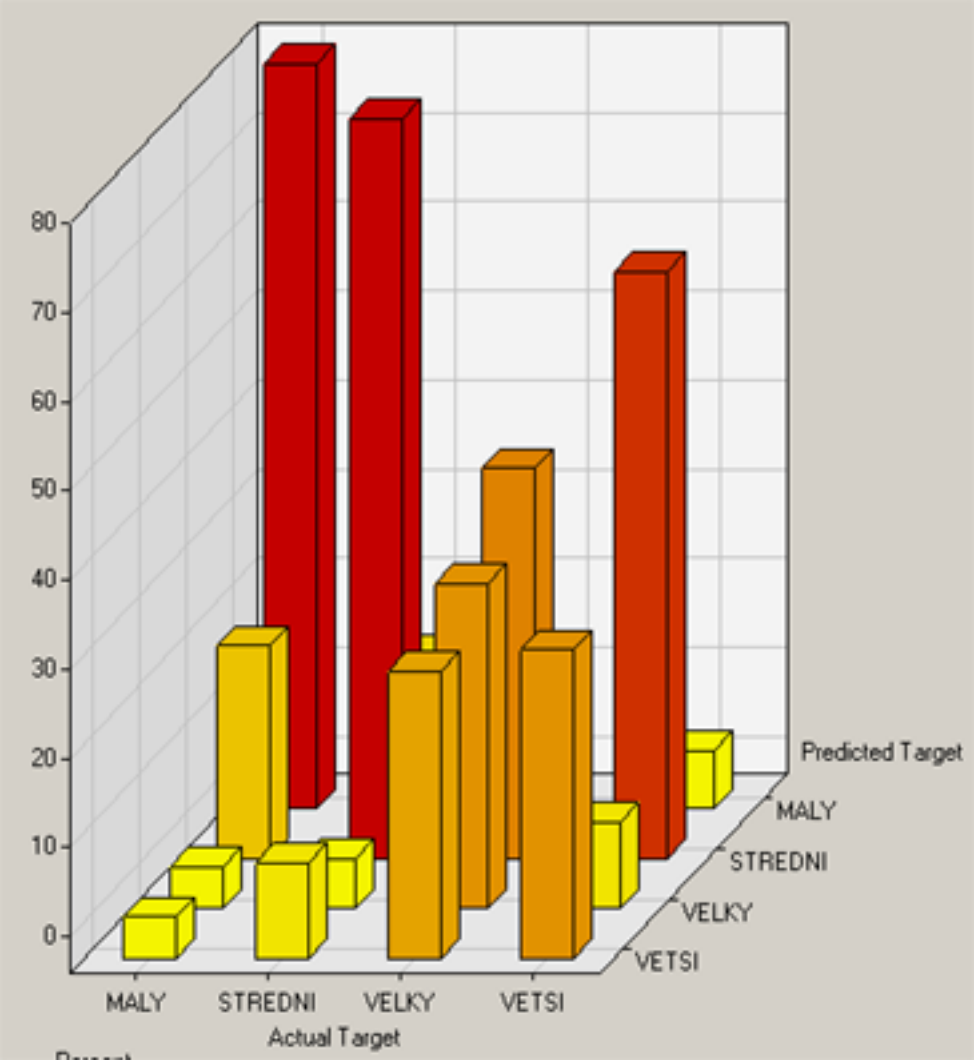
Obrázek 3.8: Diagnostický graf klasifikace databáze hub



Obrázek 3.9: Diagnostický graf klasifikace databáze mořských škeblí

Jak je vidět z diagnostického grafu klasifikace pro houbový dataset, klasifikace dopadla velmi dobře, i když má asi třikrát tolik atributů co dataset se škeblemi. Na výsledky klasifikace těchto dvou datasetů mohou bezesporu mít vliv i více různých okolností, nicméně tento pozoruhodný rozdíl nás přivedl na myšlenku, že více než počet atributů klasifikaci ztěžuje počet tříd, do kterých klasifikujeme.

Provedli jsme proto redukci počtu z 18 tříd na 4 třídy. Původní třídy členěné podle počtu kroužků od 1 až 18, jsme nahradili pouze kategoriemi malý, střední, větší, velký.



Obrázek 3.10: Diagnostický graf klasifikace pro 4 třídy

Přesto, že klasifikátor již akceptovatelně nepozná větší škebli viz. obr. 3.10, je redukce počtu tříd bezesporu prospěšná. Ropoznávací schopnost malých škeblí se zlepšila.

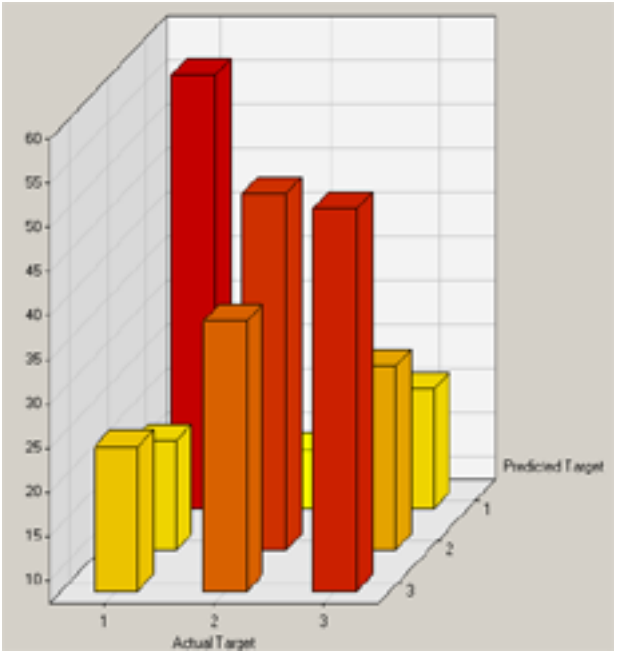
3.4.3 Vydolovatelnost informace

Předpoklad

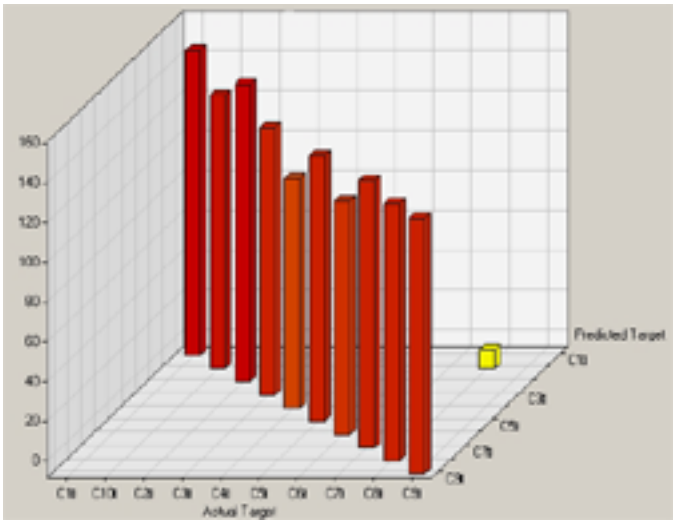
V některých databázích, jako tomu bylo například z výše uvedeným datasetem škeblí, není možné spolehlivě dolovat informace, v našem případě provádět klasifikaci. Protože o datech máme informace, které ve skutečnosti na výslednou třídu mají malý vliv. Klasifikátor pak vlastně nechybuje, jenom v datech, které mu byly poskytnuty, není požadovaná souvislost vůbec patrná.

Dataset

Pro demostraci této vlastnosti, jsme vygenerovali umělý dataset (generátor viz. web `http://www.datasetgenerator.com`), ve kterém jsou všechny třídy přesným důsledkem atributů dat. Pro srovnání je zde pak dataset, ve kterém se posuzuje používaná antikoncepční metoda v závisloti na demografických a sociálních datech, které o sobě poskytly indonézké ženy při vládním průzkumu v rove 1987. Vdané ženy byly dotazovány, aby uvedly například své vzdělání, pracovní pozici, společenské postavení, vyznání a zároveň zda používají jednorázovou, dlouhodobou nebo žádnou metodu antikoncepce. Výběr antikoncepce, zde uvádíme jako příklad věci, která závisí na velkém množství faktorů, které je těžké všechny zaznamenat.



Obrázek 3.11: Diagnostický graf klasifikace metod antikoncepce



Obrázek 3.12: Diagnostický graf klasifikace umělého datasetu

Výsledek

Jak je vidět z grafu viz. obr. 3.11 a 3.12 , klasifikátor nebyl při určení metody antikoncepce příliš úspěšný. Počet špatně klasifikovaných vzorů je o něco málo menší než počet správně klasifikovaných. Pokud se podíváme blíže na okolnosti klafikace, zjistíme, že nejpodstatnějším atributem nebyly demografické nebo sociální okolnosti, ale celkový počet dětí. Klasifikátor, prostě neměl dost relevantních informací.

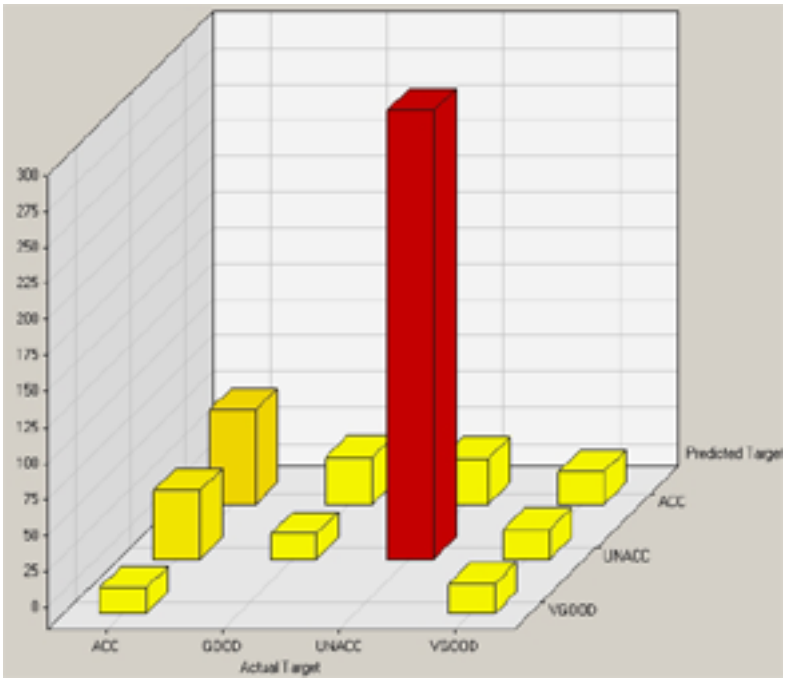
Naproti tomu syntetický dataset se skládal pouze ze vzorků, jejichž třída byla přesným odvozením z atributů založených na podobných pravidlech.

$$c1 \rightarrow B=c \;\&\; C=j \;\&\; F=b$$
$$c8 \rightarrow A=d \;\&\; D=g \;\&\; F=m$$
a podobně ...

3.4.4 Zašumělá třída

Předpoklad

V průběhu porovnávání jsme si všimli zajímavého úkazu. U některých datasetů se stávalo, že některé klasifikátory odmítaly přiřazovat vzorky do některých tříd. Po bližším prozkoumání dat v těchto třídách jsme zjistili, že se zde vyskytují vzorky, které ačkoliv mají stejné nebo téměř stejné atributy, náleží do různých tříd. Znamenalo by to, že implementace klasifikátorů v SAS EM vynechají třídu pokud narazí na rozporuplné informace.



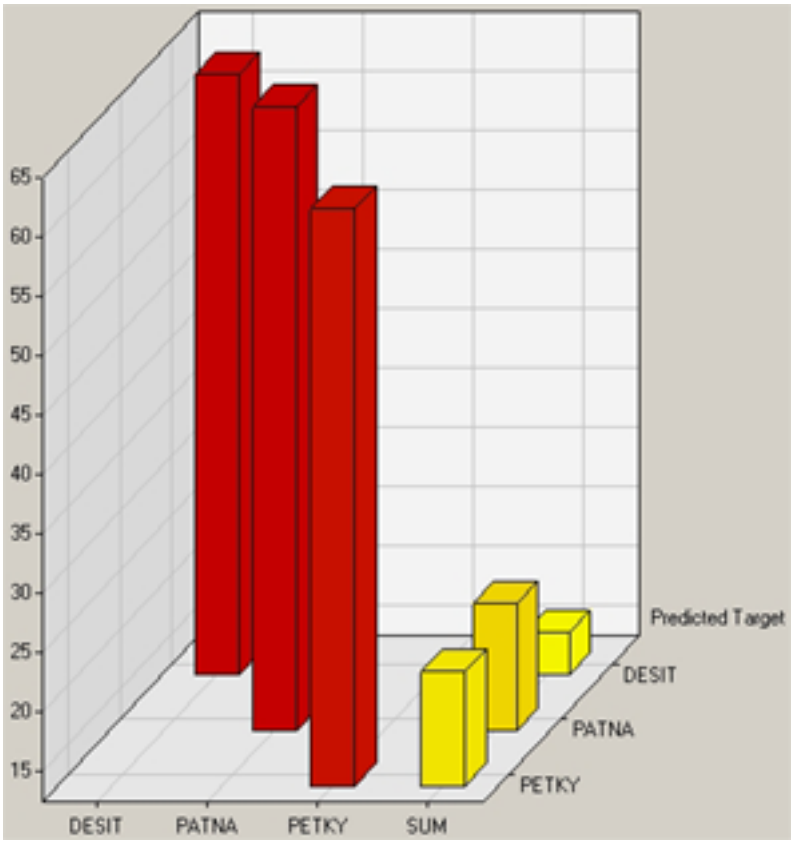
Obrázek 3.13: Diagnostický graf klasifikace datasetu se zašumělou třídou

Dataset

Graf na obr. 3.13 je výsledkem klasifikace dat automobilů s atributy s různými vlastnostmi jako pořizovací cena, náklady na udržbu, počet dveří atd. Vzorky byly rozděleny do tříd podle úspěšnosti v prodejů na trhu - nepřijatelná, přijatelná, dobrá, velmi dobrá. Ve třídě s dobrou prodejností se vyskytovalo tolik různých vzorků, že neuronová síť do této třídy odmítla klasifikovat.
Abychom si tomu doměnku potvrdili, vygenerovali jsme **umělý dataset**, který obsahoval zhruba 1000 vzorků, které byly rozděleny na tři části. Vzorky s číslem 5 a ohodnocením třídou “pětky”, číslem 10 a ohodnocením třídou “desítky”, číslem 15 a ohodnocením třídou “patnáctky”. Zhruba každý pátý vzorek byl ohodnocen třídou šum, ať jeho atribut byl jakýkoliv z výše uvedených.

Výsledek

Splnilo se očekávání, že žádný z klasifikátorů **třídu šum** to klasifikace vůbec **nezahrne**. Neobsahuje totiž žádny atribut, podle kterého by tak bylo správné učinit.



Obrázek 3.14: Diagnostický graf klasifikace datasetu s uměle zašumělou třídou

Kapitola 4

Závěr

Tato práce měla za úkol zhodnotit **vlastnosti klasifikátorů**, které ovlivňují **úspěšnost** klasifikace v různých typech databází. V první části byl čtenář na přehledové úrovni seznámen s dataminingem obecně a posléze podrobněji i s principy, na kterých pracují modely, jejichž vlastnosti jsme porovnávali. Představili jsme si prostředí SAS EM, které umožnuje rychlé a snadné řešení dolovácích úkolů. Během porovnávání výsledků klasifikace účelově zvolených datasetů jsme došli k některým poznatkům, které poněkud předčily naše očekávání. Došli jsme k některým závěrům, které pomohou i laickému uživateli, manažerovi, studentovi lépe pochopit **podstatu** dobře zformulované **dolovací úlohy** a základní hlediska její úspěšnosti. Budoucích možných rozšíření pro tuto oblast se nabízí hned několik, záleží jakým směrem se hodláme vydat, moje práce se soustředila pouze na modely klasifikace, nabízí se ostatní zmiňované modely - shlukovaní, predikce, asociační pravidla. Pokud bychom se zaměřili na spíše na úroveň detailů práce, je možné vlastnosti modelů porovnávat v podstatně větších podrobnestech, včetně stanovení určitého přenosného systému metrik pro porovnávání.

Literatura

[1] Berka, Petr: Dobývání znalostí z databází. Book. Academica, Praha, 2003.

[2] Zendulka, Jaroslav, a kol. :Získávání znalostí z databází: Studijní opora. Fakulta informačních technologií, Brno, 2006.

[3] Han,J., Kamber, M.: Concepts and techniques. Elsevier Inc.,2006.

Dodatek A

Příloha

Na přiloženém médiu (cd-rom) lze nalézt elektronickou podobu tohoto textu, všechny porovnávané datasety (včetně zde nezmiňovaných) a jejich výsledky v podobě diagnostických grafů.