Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra informačních technologii

Bakalářská práce

Dolování dat z informačních systémů

Ladislav Karas

Čestní prohlášení:

Prohlašuji, že jsem následující bakalářskou práci na téma „Dolování dat z informačních systémů“ vypracoval samostatně, pouze za odborné asistence Ing. Martina Havránka, a že všechny prameny, ze kterých jsem čerpal, jsou uvedeny ve zdrojích použité literatury.

Dne …….……. Podpis ……

Poděkování autora

Chtěl bych poděkovat vedoucímu bakalářské práce Ing. Martinu Havránkovy za veškerý čas strávený odbornou pomocí, kterou mi při psaní této práce věnoval, dále děkuji Lence Kozlové za gramatickou kontrolu bakalářské práce.

Dolování dat z informačních systémů

Souhrn

Bakalářská práce se věnuje tématu dolování dat z databázových systémů. Teoretická část popisuje obecné znalosti, potřebné pro pochopení dolování dat, algoritmy využívající se k dolování dat z databází a proces přípravy dat. Praktická část práce se zaměřuje na analýzu a sestavení predikčního modelu pro data získané z internetových stránek zaznamenávající různé události (chyby). K analýze a sestavení predikčního modelu byla použita metoda nevyvážených rozhodovacích stromů. Vzhledem k složité struktuře vzniklé dolováním dat byly popsány postupy prořezávání stromů.

Klíčová slova

Dolování dat, Získávání znalostí z databází, Vzory, Informace, Databázové systémy, CRISP-DM, MS SQL Server 2008, Nevyvážené rozhodovací stromy, Prořezávání stromů.

Data mining

Summary

This Bachelor Thesis addresses the subject of data mining from database systems. The theoretical part describes the general knowledge needed for understanding the data mining, algorithms used for data mining from the databases and data preparation process. The practical part focuses on the analysis and compiling of the prediction model used for the data obtained from the internet pages which record various events (mistakes). Method of unbalanced decision trees was used for the analysis and compiling the prediction model. Procedures of pruning was described because of the complicated structure formed by data mining.

Keywords

Data mining, Knowledge discovery in database, Patterns, Information, Database systems, CRISP-DM, MS SQL Server 2008, Unbalanced decision trees, Trees pruning.

Obsah

1. Úvod

O dolování dat z databází jako samostatném oboru se začalo mluvit počátkem 90. let minulého století, kdy se začala řešit problematika zpracování velkého množství dat uchovaných v databázových systémech. Dolování dat (DD) definoval Usama M. Fayyad v roce 1996 jako: *„netriviální získávání implicitních, dříve neznámých a potenciálně užitečných informací z dat.“*

DD je více-disciplinární obor, který se skládá především z oblastí databázových technologií, statistiky, strojového učení, rozpoznávání vzorů, vyhledávání informací, neuronových sítí, vizualizace dat a mnoha dalších oblastí. Téměř všechny tyto obory byly popsány před DD, ale šly si vlastní cestou. DD jakožto nový obor propojil poznatky z oborů statistiky, strojového učení a neuronových sítí s oborem databázových technologií, čímž nechal vzniknout novým postupům pro rozpoznávání vzorů, vyhledávání skrytých informací a novým možnostem vizualizace dat. V současnosti je DD implementováno ve všech komerčně nejúspěšnějších databázových serverech a mnoha dalších aplikacích pro podporu rozhodování.

Z počátku 90. let se DD využívalo především k různým vědeckým účelům a dále v sektorech s vysokou přidanou hodnotou kupříkladu v bankovnictví, energetice či pojišťovnictví. Postupně se uplatnění DD rozšiřovalo a v dnešní době nachází nejrůznější uplatnění v pestré škále oborů.

2. Cíl práce a metodika

Bakalářská práce má za cíl shrnout dosavadní poznatky z oblasti dolování dat z databázových systémů a na základě získaných znalostí provést analýzu dat a sestavit predikční model použitím libovolné metody a softwaru pro dolování dat. Výsledky analýzy dat a predikce ověřit na testovacích datech, zhodnotit výsledky predikce a popsat nejdůležitější části tohoto modelu.

Na začátku práce bude provedena rešerše odborné literatury, která bude rozdělena do tří kapitol: *Přehled řešené problematiky* shrne obecný popis DD a dále základní znalosti metodiky procesu DD, zdroje dat pro DD, základní statické minimum a strojové učení. *Metody využívané pro data mining* vysvětlí postup, funkci, výhody a nevýhody jednotlivých algoritmů využívaných DD. *Příprava dat* se bude zabývat postupem transformace výchozího souboru na vstupní data pro samotný algoritmus dolování dat.

Na základě získaných znalostí bude vybrán jeden z vhodných algoritmů DD pro splnění zadání popsaného v cíli. Postup procesu DD bude vycházet z metodiky CRIPS-DM. Dále se stanoví software, který bude použit v jednotlivých krocích procesu DD. Výsledky DD budou upraveny do srozumitelné podoby a dopočítá se z nich úspěšnost sestaveného predikčního modelu. Na závěr budou popsány nejdůležitější poznatky získané DD.

3. Přehled řešené problematiky

Databázové systémy v dnešní době uchovávají ohromné množství dat. Problém je, jak z těchto dat získat cenné informace. Velké databázové tabulky obsahují miliony až miliardy řádků, které jsou různě strukturovány a uloženy různým způsobem. Taková data v sobě uchovávají cenné informace, problém je jak tyto informace získat.

Dolování dat[1] slouží k vyhledávání vzorů v datech. Na základě nasbíraných dat se pokouší zjišťovat a odkrývat různé závislosti. Takto vytvořené znalosti potom využívá při prediktivní analýze. Vzory v datech jsou hledány od začátku historie lidstva. Lovci pozorovali chování zvířat, zemědělci zkoumali procesy pěstování plodin a chovu dobytka a obchodníci sledovali chování zákazníka. Tyto informace si lidé předávali z generace na generaci.[2] Pomocí dolování dat z informačních systémů je možno tyto závislosti získat prakticky v reálném čase. Toho může být dosaženo za předpokladu kvalitně navrženého systému zpracování a uchování dat.

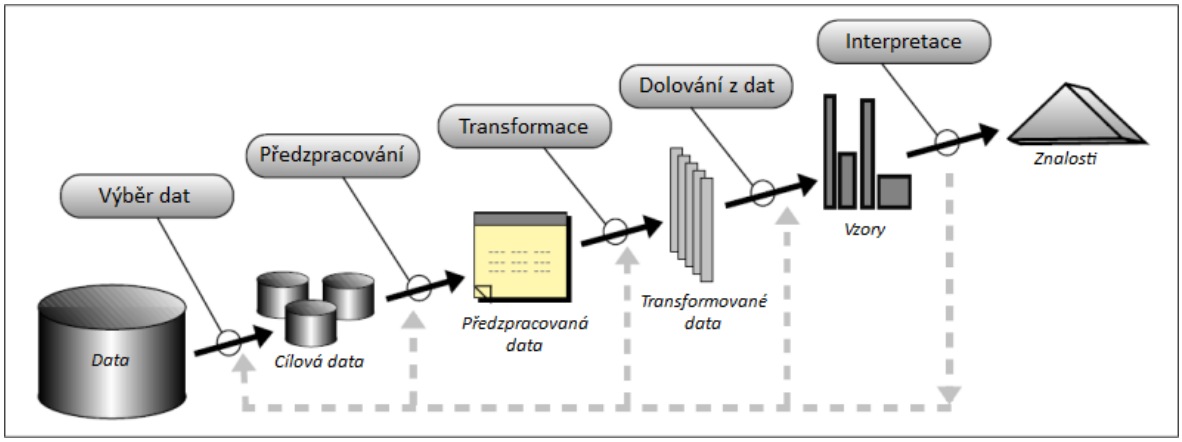
„Data mining je proces analýzy dat z různých perspektiv a jejich přeměna na užitečné informace. Z matematického a statistického hlediska jde o hledání korelací, tedy vzájemných vztahů nebo vzorů v datech. Data mining je proces, jehož cílem je těžba informací v databázích. Využívá statistické metody a další metody hraničící s oblastí umělé inteligence.“[3]

Dolování dat je v současnosti implementováno do většiny komerčně úspěšných databázových serverů a jde o nejrychleji rostoucí segment Business Intelligence[4]. Dolování dat nachází uplatnění v oborech bankovnictví, pojišťovnictví a finančnictví, telekomunikací, energetiky, marketingu, různých vědeckých oborech a mnoha dalších oborech spravujících velká množství dat.

[1]Pojem „Dolování dat“ vychází z překladu anglického pojmu „Data Mining“, někteří autoři jako. Petr Berka, používají pojem „získávání znalostí“ nebo „dobývání znalostí. Toto slovíčkaření má svůj význam, jelikož samotná data nejsou v informatice chápána jako informace, nýbrž jen jako uchovávatelé informací. K získání informací je potřeba jejich správná interpretace. Při procesu data miningu se získávají znalosti, a proto je přesnější název „dobývání znalostí“.
[2]LACKO, Luboslav. Business Intelligence v SQL Serveru 2008, str. 264
[3]LACKO, Luboslav. Business Intelligence v SQL Serveru 2008, str. 265
[4]Business Intelligence je souhrnné označené analýz pro podporu rozhodování zahrnující: OLAP, dolování dat, reporting, podporu analýz, přehledové zobrazení, podnikové řízení výkonnosti a prediktivní analýzy.

3.1 Metodiky dolování dat

Metodiku dolování dat z technického pohledu popsal Usama M. Fayyad v roce 1996. Tato metodologie představovala 5 činností (od výběru dat až po interpretaci) a 6 stavů (od dat až ke znalostem). Takto sestavený proces opomíjel velmi důležitou fázi zadání projektu, a proto byly vyvinuty další metodiky postupu, z nichž dnes je nejrozšířenější metodika CRISP-DM.



Obrázek 1: Proces při dolování dat podle Usama M. Fayyad v roce 1996[5]

3.1.1 „5A“

Metodiku 5A používá firma SPSS v produktu Clementine.

- Assess – posouzení potřeb projektu
- Access – shromáždění potřebných dat
- Analyze – provedení analýz
- Akt – přeměna znalostí na akční znalosti
- Automate – převedeni výsledků do praxe

3.1.2 SEMMA

Metodika používaná firmou SAS ve svém softwarovém produktu Enterprise Miner.

- Sample – výběr vhodného objektu
- Explore – vizuální explorace a redukce dat
- Modify – seskupení objektu a atributů, datové transformace
- Model – použití algoritmu pro dolování dat
- Access – porovnání modelu a interpretace

[5] Zdroj obrázku www.data-mining-blog.com/wp-content/uploads/2010/01/fayyad1996.png, překlad vlastní

3.1.3 CRISP-DM

Metodologie CRISP-DM (CRoss-Industry Standard Process for Data Mining) vznikla v rámci výzkumného projektu Evropské komise a je zastupována největší skupinou firem zabývajících se dolováním dat. Cílem projektu bylo vytvoření jednotného postupu použitelného v nejrůznějších komerčních aplikacích. V roce 2006 byl oznámen záměr na vypracování CRISP-DM 2.0, ale od roku 2007 není známa žádná zpráva o jeho vývoji. Pravděpodobně byl zastaven.

Metodologie CRISP-DM 1.0 obsahuje 6 následujících fází.

Porozumění problematice (Business Understanding)

Počáteční fáze se zaměřuje na pochopení cíle projektu a požadavků z obchodního hlediska a převedení těchto poznatků na zadání úlohy pro dolování dat. Hodnotí se možná rizika, náklady a přínos metody dolování dat.

Porozumění datům (Data Understanding)

Tato fáze začíná počátečním sběrem dat. Následují činnosti, které mají za hlavní cíl zjistit kvalitu dat a získat základní charakteristiky souboru dat.

Příprava dat (Data Preparation)

Fáze přípravy dat zahrnuje všechny činnosti na vytvoření finálního datového souboru ze základního souboru dat. Finální datový soubor musí obsahovat relevantní údaje k řešené úloze a musí mít podobu, která je vyžadována vlastními analytickými metodami.

Samotná příprava dat zahrnuje selekci vytváření, integrování a formátování dat. Tyto úkony jsou obvykle prováděny opakovaně a v nejrůznějším pořadí. Fáze přípravy dat bývá obvykle nejdelší z celého řešení úlohy.

Modelování (Modeling)

Fáze modelování obsahuje samotné využití metod pro dolování dat. Ve většině případů je možné použít několik algoritmů a jejich výsledky zkombinovat. V případě využití více algoritmů je někdy potřeba vracet se zpět do fáze přípravy dat vzhledem ke specifickým požadavkům některých metod. Z výsledků je sestaven výsledný model.

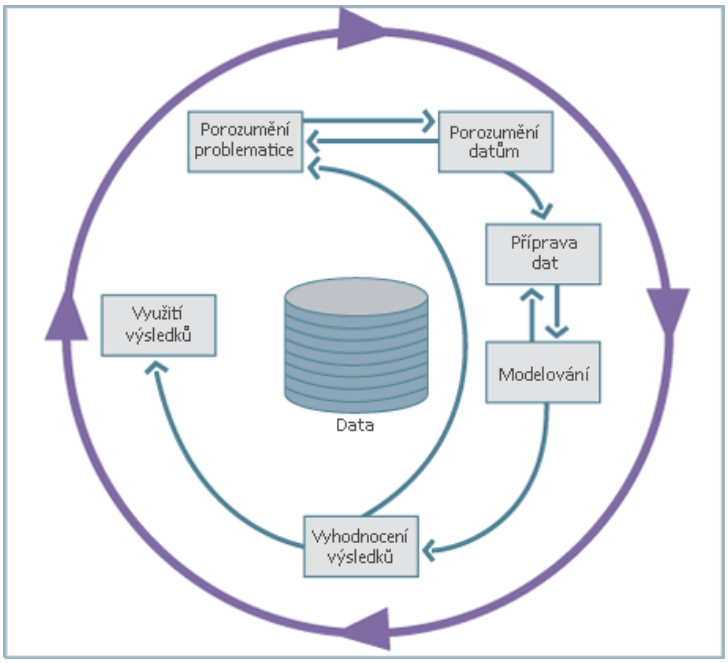
Vyhodnocení výsledků (Evaluation)

V této fázi je model či modely hotový a z pohledu analýzy dat má vysokou kvalitu. Před konečným přijetím modelu je nezbytné více vyhodnotit model a přezkoumat kroky k sestavení tohoto modelu z pohledu splnění cílů, zadaných v prvním kroku. Na konci této fáze by se mělo rozhodnout o využití znalostí, kterých bylo dolováním dat dosaženo.

Využití výsledků (Deployment)

Vytvoření modelu většinou mění konec projektu. Je potřeba získané znalosti prezentovat tak, aby byly pro zadavatele plně využitelné. V závislosti na požadavcích zákazníka může být zavádění modelu jednoduché sepsáním výsledné zprávy, nebo také zavedením komplexního systému pro automatickou klasifikaci nových případů. V mnoha případech to bude zákazník, kdo provede využití modelu v praxi. Je důležité pochopit, co je potřebné učinit v praxi, aby mohly být výsledky efektivně využity.[6]

Celkové schéma metodiky CRISP-DM je znázorněno na obr. 2. Je třeba si všimnout obousměrných propojení mezi přípravou dat a modelováním. Takřka vždy je nutné se několikrát vracet a upravovat vstupní data pro modelování a to samé platí i mezi porozuměním problematice a porozuměním datům.



Obrázek 2: fáze postupu pro metodiku CRISP-DM[7]

[6] Originální znění a další informace lze nalézt na www.crisp-dm.org
[7] Zdroj obrázku www.crisp-dm.org/Images/Crisp-dmchartnew.gif, překlad vlastní

3.2 Zdroje dat

Zdrojem dat mohou být data uložené v nejrůznějších strukturách od flat souborů až po datové sklady či OLAP krychle. Vzhledem k začlenění nástrojů pro dolování dat databázových serveru jsou nejvyužívanější datové sklady, OLAP krychle a relační databáze.

3.2.1 Relační databáze

Relační databáze znamenala velký pokrok - nahrazení archaického ukládání dat do flat souborů za ukládání dat do tabulek vzájemně propojených pomocí identifikačních klíčů. Ukládání dat, jejich selekce, projekce a spojení se provádí jazykem SQL. Pro analytiky, kteří neuměli SQL, nezbývala jiná možnost, než napsat dotaz ve slovním znění a zadat programátorovi, aby jej přepsal do SQL, což značně zpomalovalo práci analytika.

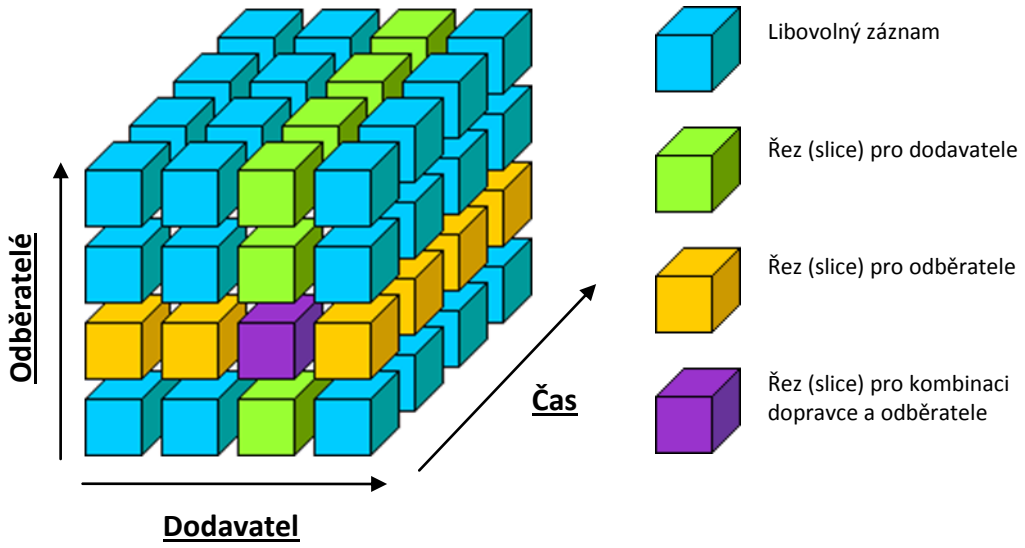
3.2.2 EIS

EIS (Executed Information System) je mezičlánek mezi relační databází a uživatelem. Umožňuje uživatelům přístup k databázi bez znalosti SQL a vnitřní struktury databáze za použití předdefinovaných dotazů. Nevýhodou je omezená množina předdefinovaná dotazů. Pro rozšíření této množiny je potřeba opět programátora.

3.2.3 OLAP

Zkratka OLAP vychází z anglického Online Analytical Processing. OLAP představuje multidimenzionální uložení dat do datových krychlí (data cube). Každá dimenze krychle představuje určitý atribut a jednotlivé buňky krychle představují záznam v databázi. Takovéto uložení dat umožňuje velmi rychlé provádění analytických operací a to použitím operací jako jsou různá natáčení (pivot), provádění řezů (slice), výběry určitých částí (dice) a zobrazování různých agregovaných hodnot.

Uložení dat do krychlí přináší i nevýhody v podobě zvýšených nároků na datové servery, neboť krychle obsahují prázdné buňky a existuje i redundance dat. Z tohoto důvodu se používají dva hlavní druhy OLAP a to MOLAP (multidimenzionální OLAP) a ROLAP (relační OLAP) a jejich kombinace v podobě HOLAP (hybridní OLAP). ROLAP funguje na principu převádění dotazů do SQL dotazů, neboť data jsou uložena v relační databázi, zatímco MOLAP pracuje s daty v krychlích, které se ukládají do úložišť pro vícedimenzionální pole.



Obrázek 3: příklad krychle OLAP a ukázka výběrů dat

3.2.4 Datové sklady a datové trhy

Datové sklady

Definice datového skladu od Billa Inmona: *„Datový sklad je podnikově strukturovaný depozitář subjektově orientovaných, integrovaných, časově proměnných, historických dat použitých pro získání informací a podporu rozhodování. V datovém skladu jsou uložena atomická a sumární data.“*[8]

V datových skladech se neuplatňují normálové formy a existuje redundance dat. Vzhledem k neměnnosti nezpůsobuje redundance dat žádné jiné problémy, než větší rozměry. Každá tabulka představuje určitý podnikatelský předmět např. zákazníka, dodavatele nebo výrobek. K uložení dat do datových skladů je potřeba značná transformace daných dat[9], jelikož do skladu mohou být ukládány data z různých zdrojů. Takto uložená data se poté dají velmi rychle dotazovat a slouží především k různým analitickým operacím, které by u normálních databází nebyly vůbec možné.

Datové trhy

Jsou v podstatě menší datové sklady, které obsahují jen některá data potřebná pro určitou divizi podniku. Dají se budovat již z hotového datového skladu nebo opačně nejdřív sestavit datový trh a až z nich postavit celý datový sklad.

[8] Překlad citace převzat z: LACKO, Luboslav. Business Intelligence v SQL Serveru 2008, str. 38

[9] Transformace dat do datových skladů se nazývá ETL – Extract (výběr dat z databáze), Transform (transformace dat), Load (zavedení dat do datového skladu).

3.3 Statistické minimum

3.3.1 Kontingenční tabulka

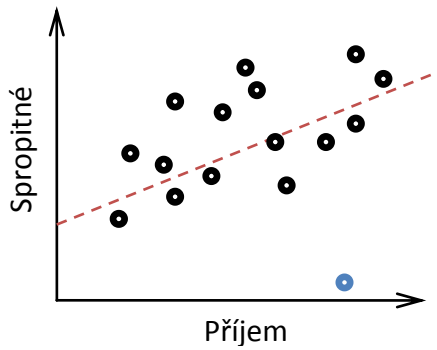
Kontingenční tabulka předpokládá, že každá jednotka může být kvalifikována podle dvou kvalitativních (kategoriálních) proměnných. Tabulka tedy představuje počty jednotek zařazených pod různé kombinace popisujících proměnných. Z takto sestavené tabulky lze stanovit, zda existuje závislost mezi dvěma kvalitativními proměnnými a poté spočítat sílu této závislosti. Kontingenční tabulka o rozměrech 2x2 se nazývá asociační tabulka a zkoumá závislost dvou kvalitativních alternativních znaků.

Tabulka 1: ukázková kontingenční tabulka spotřebitelských úvěrů

	Vysoký příjem	Střední příjem	Nízký příjem	Σ
Úvěr ano	5	24	36	65
Úvěr ne	32	84	47	163
Σ	37	108	83	228

3.3.2 Regresní analýza

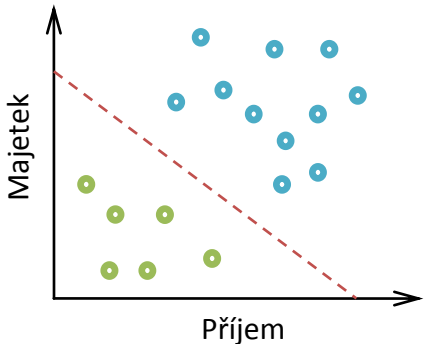
Regresní analýza slouží k určení, zda existuje závislost mezi jednou proměnou závislou a jednou či více proměnnými nezávislými, k určení síly dané závislosti a ke stanovení regresní funkce. Z regresní funkce jde predikovat hodnotu závisle proměnné v závislosti na hodnotách jejích nezávislých proměnných, zjistit koeficienty pro nezávislé proměnné a stanovit hodnotu náhodné veličiny. Regresní diagnostika slouží ke stanovení kvality dat, tedy zjištění vlivných a odlehlých hodnot v analyzovaném souboru. Podle výsledků regresní diagnostiky lze upravit soubor a tím zvýšit sílu závislosti.



Obrázek 4: regresní přímka pro závislost spropitného na příjmu

3.3.3 Diskriminační analýza

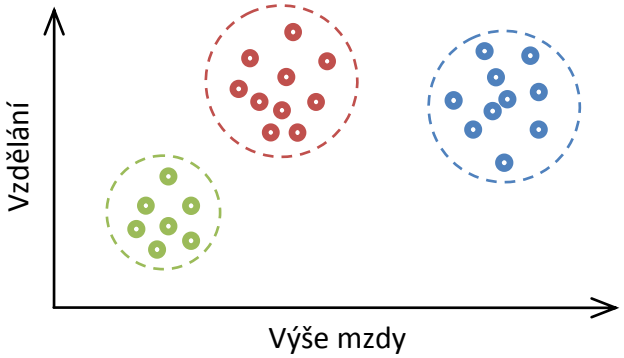
„Diskriminační analýza je vlastně úloha klasifikace příkladů do předem zadaných tříd. Z pohledu statistického tedy úloha hledání závislosti jedné nominální veličiny (určující příslušnost ke třídě) na dalších v numerických veličinách.“ [10] Stanovením závislostí mezi nominální a dalšími veličinami může následně sloužit k přiřazení dalších případů do příslušných skupin.



Obrázek 5: diskriminační analýza pro skupiny bohatých a chudých

3.3.4 Shluková analýza

Shluková analýza slouží k seskupení navzájem si podobných jednotek do shluků tak, aby si jednotky ve shluku byly podobnější než ostatní jednotky mimo daný shluk. Vzdálenost (podobnost) dvou prvků se vyjadřuje na základě numerických veličin jednotky pomocí různých měřítek např. Hemmingovi, Euklidovské či Čebyševové vzdálenosti. Algoritmus pro shlukovou analýzu může být např. „zdola nahoru“, kde se postupně shlukují prvky od stejného počtu prvků jako shluků, až do jednoho jediného shluku.

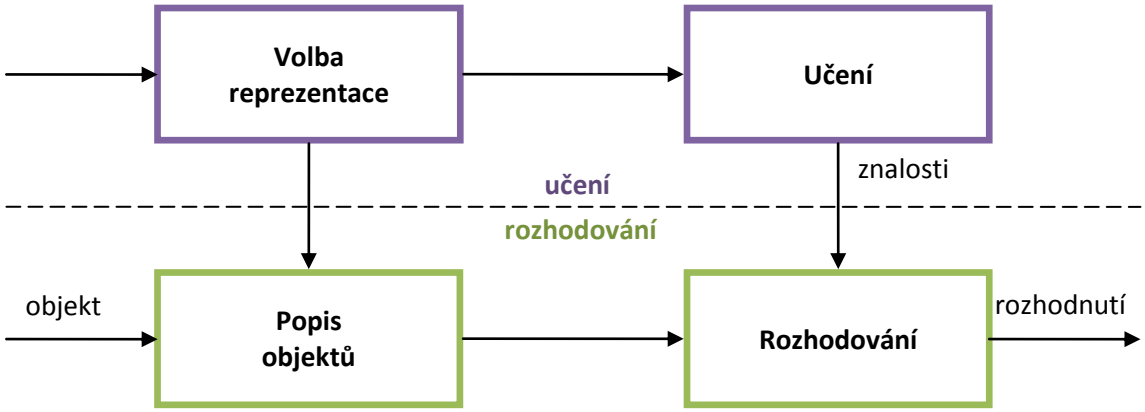


Obrázek 6: příklad shlukové analýzy

[10] BERKA, Petr. Dobývání znalostí z databáze. Str. 53.

3.4 Strojové učení

Strojové učení je podoblast umělé inteligence zabývající se algoritmy umožňující učení počítačovému systému. Prvky učení se mohou pod různými názvy nalézt v řadě vědních disciplín; ve statistice se objevuje explorační analýza dat nebo inteligentní analýza dat, v umělé inteligenci se hovoří o metodách rozpoznávání obrazů, strojového učení nebo automatizovaného získávání znalostí, v teorii řízení (kybernetice) lze nalézt adaptivní a učící se systémy. V souvislosti se získáváním znalostí z databází se používá termín dolování z dat.[11]



Obrázek 7: obecné schéma učícího se systému

Rozdělení metod učení z pohledu vynaloženého úsilí:

1. *Učení zapamatováním* – systém pouze zaznamenává data nebo dílčí znalosti dodané externím zdrojem
2. *Učení se z instrukcí* – systém získává znalosti z externího zdroje a integruje je se znalostmi již získanými
3. *Učení se z analogie* – získávání znalostí založeno na zapamatování si příkladů
4. *Učení na základě vysvětlení* – při učení se používá několik příkladů a rozsáhle znalosti z dané oblasti
5. *Učení se z příkladů* – zde se využívá velkého množství příkladů, které se má systém naučit; používá se metoda indukce
6. *Učení se z pozorování a objevování* – pracuje se s velkým množstvím dat, tyto data na rozdíl od předchozích případů, kde byli popsány učitelem, jsou získána pozorováním či objevováním

[11] BERKA, Petr. Dobývání znalostí z databáze. Str. 60.

Zásadním problémem strojového učení je, jak poznat, že proces učení probíhá správně. Rozlišuje se několik základních typů učení:

1. *Učení s učitelem* – učitel poskytuje systému explicitní informace, takovéto data si stroj pouze zapamatuje
2. *Učení bez učitele* – stroj sám hledá informace v množství dat, využívá k tomu metody používané v dolování dat
3. *Kombinace učení s učitelem a bez učitele* – při této kombinaci jsou některá data popsána učitelem, z těch se vychází i pro data ostatní
4. *Zpětnovazební učení* – vychází z Markovova rozhodovacího procesu[12] a je založeno na odměňování za správné chování a postihování za chování nesprávné

3.5 Software pro dolování dat[13]

Software pro dolování dat můžeme rozdělit do 4 skupin. Databázové servery s implementací business inteligence, jejíž součástí jsou i nástroje pro dolování dat z databází. Statistický software a software pro podporu rozhodování obsahující součásti DD. Ostatní aplikace pro DD vyvíjené komunitami vědců, většinou se jedná o volné licence.

Databázové servery s implementací business inteligence představují softwaroví giganti Microsoft se svým „*Microsoft SQL Serverem 2008“*, IBM s *„IBM DB2“* a Oracle s „Oracle Database 11g“.

Statistický software representuje společnost SAS svým produktem *„SAS Enterprise Mine“*, StatSoft s aplikací *„STATISTICA Data Miner“* a od IBM nástroj *„SPSS Modeler“*.

Software pro podporu rozhodování velmi rozšířený je SAP R/3 od německé firmy SAP v němž jsou zabudovány nástroje pro dolování dat.

Ostatní aplikace pro DD z nekomerčních jmenujme WEKA, Tanagra, YALE, Orange nebo český LISp-Miner z komerčních aplikací Bayesia, Miner 3D a v neposlední řadě Minitab.

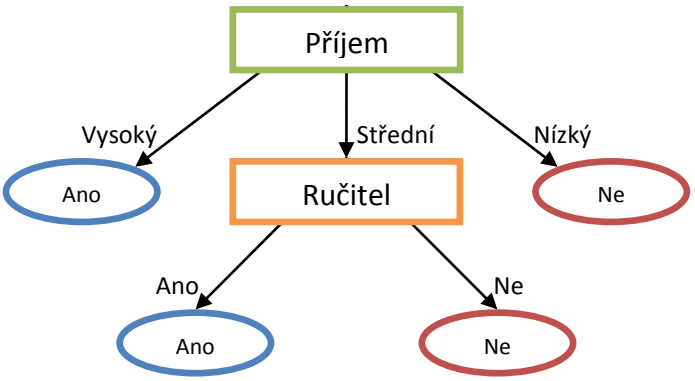
[12] Markovův rozhodovací proces je rozšířením Markovských řetězců o přidání akcí (umožňující výběr) a užitků (motivace).
[13] Seznam některých programů obsahujících algoritmy pro dolování dat se nachází v příloze 1.

4. Metody využívané pro data mining

4.1 Rozhodovací stromy

Jedná se o nejpoužívanější algoritmus pro dolování dat. Rozhodovací stromy charakterizuje jednoduchá interpretace výsledků, srozumitelnost a přehlednost. Tyto vlastnosti umožňují uživateli rychle a snadno vyhodnotit získané výsledky, identifikovat klíčové položky a vyhledávat zajímavé segmenty případů. Cílem algoritmu je identifikovat entity popsané různými atributy a rozdělit je do tříd.[14]

Rozhodovací stromy mají hierarchickou strukturu, a proto je jejich algoritmus velmi rychlý a poměrně jednoduchý. Používají se především v úlohách klasifikace a predikce, kde převažují kvalitativní data. Spojitá (numerická) data se totiž musí nejdříve agregovat do konečného počtu skupin a to tak, aby měly entropii[15] co nejbližší nule optimálně tak, aby obsahovaly pokud možno jen jednu hodnotu predikované veličiny. Proces agregace provádí algoritmy automaticky. Po převedení všech veličin na kategoriální data jsou postupně vybírány veličiny na základě jejich entropie. Veličina nejbližší nule je určena jako větev a rozděluje se na další uzel. Uzel s nulovou entropií lze nazývat listem, neboť se již dál nedělí. V případě uzlu s nenulovou entropií se jedná o větev, která se většinou dále dělí.



Obrázek 8: jednoduchý rozhodovací strom o poskytnutí úvěru

[14] LACKO, Luboslav. Business Intelligence v SQL Serveru 2008. Str. 275.

[15] Entropie představuje míru neuspořádanosti určitého systému. Výpočet entropie se provádí vzorcem: $H = -\sum_{t=1}^{T}(p_t \cdot log_2\, p_t)$. Nulová entropie znamená, že sledovaná vlastnost nabývá jen jedné hodnoty. Čím víc je entropie vzdálenější od 0 tím nižší má daná veličina vypovídající schopnost.

4.2 Rozhodovací pravidla

Tato pravidla slouží ke klasifikaci na základě rozhodovacích pravidel podobně jako u rozhodovacích stromů. Jedná se tedy opět o učení s učitelem. Rozhodovací pravidla jsou v běžném životě často používaná např.: nebude-li pršet, nezmokneme. Z tohoto jednoduchého případu jde odvodit syntaxi pravidel.

IF (předpoklad) THEN (příslušnost do třídy)

Taková pravidla lze vytvořit pomocí několika algoritmů. Jedním z nich je i odvození pravidel z rozhodovacích stromů. Na rozdíl od rozhodovacích stromů, se dá při sestavování rozhodovacích pravidel postupovat jak shora dolu, tak zdola nahoru. Výhodou rozhodovacích pravidel je velmi lehká interpretace výsledků.

4.3 Asociační pravidla

Tento algoritmus spolu s rozhodovacími stromy je nejčastěji využívaným algoritmem. Z matematicko-statistického hlediska se jedná o hledání korelací v asociačních tabulkách. Tyto korelace mohou nabývat hodnot od -1 do 1, přičemž se mluví o negativní korelaci, nulové korelaci a kladné korelaci[16]. Negativní a kladné korelace se využívají u analýzy spotřebního košíku pro stanovení souvislostí mezi dvěma produkty. Nalezení souvislostí slouží k efektivnímu rozmístění zboží v obchodě nebo efektivním akčním nabídkám atd. Při analýze spotřebního košíku je převážná část nalezených závislostí předem jasná[17] a jen malá část nalezených znalostí je nová[18].

4.4 Shlukování

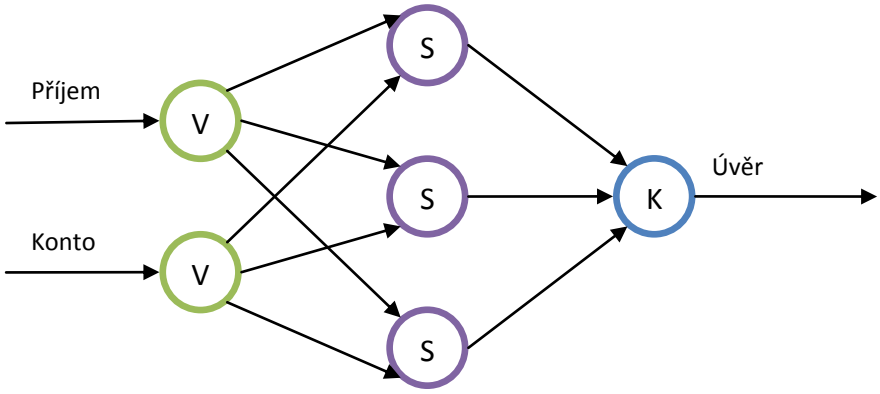
Při analýze údajů na základě shluků se seskupují údaje podle podobných charakteristik. Shlukování se využívá pro identifikaci zákaznických segmentů, které jsou založené na společných charakteristikách například demografických, sociálních, profesních a podobně. Tento algoritmus se používá pro odhalování shluků dat ve vícedimenzionálních prostorech. Pomocí něj lze rozdělit množinu případů na co nejohraničenější skupiny takzvané „ostrovy podobnosti“.

[16] Negativní korelace - substituty, nulová korelace - nezávislé produkty, kladná korelace - komplementy
[17] např.: cornflakes se prodávají s mlékem nebo hořčice se prodává s párky
[18] např.: byla nalezena závislost mezi prodejem piva a dětských plen - muži totiž chodí nakupovat pleny a s nimi si kupují i pivo, zatímco ženy doma pečují o děti.

Shluky jsou odhalovány na základě aplikování analýzy křivek pravděpodobnosti, přičemž se zkoumá, zda jednotlivé údaje nebo skupiny údajů nesplňují podmínku statistického rozložení například Gaussova normálního rozložení a podobně.[19]

4.5 Neuronové sítě

Tento algoritmus pracuje na podobném principu hledání vzorů jako lidský mozek. Každý neuron se skládá ze vstupů, těla neuronu a výstupů. Každý vstup do neuronu má přiřazenou váhu, sumu součinu vah a vstupů. Představuje celkovou velikost vstupu neuronu. Neuron dále obsahuje prahovou hodnotu a převodní funkci, která převádí hodnotu vstupu na výstupní hodnotu neuronu.



Obrázek 9: schéma neuronové sítě s jednou skrytou vrstvou

Neuronové sítě jsou tvořeny sítí neuronů uspořádaných do vrstev a vazeb mezi neurony z jednotlivých vrstev. Vrstvy se dělí na vstupní (V), konečné (K) a jednu či více vrstev skrytých (S). Před samotným procesem trénování je nutno rozdělit soubor dat na tréninkové a testovací. Během procesu trénování sleduje každá iterace velikost chyby a na základě této chyby upravuje váhy jednotlivých neuronů. Proces trénování probíhá do té doby, než je dosaženo předem stanovené velikosti maximální chyby.

Neuronové sítě odpovídají na otázky typu: „Jaká bude teplota zítra.“ nebo „O jaký předmět se jedná na základě jeho popisu.“ na co však neposkytují odpovědi: „Proč zítra bude 7 stupňů.“ respektive „Proč se jedná o předmět počítač.“ Informace „proč“ jsou uloženy ve struktuře neuronů a v nastavení jejich vah a prahových hodnot podobným způsobem, jako jsou uloženy v lidském mozku. Neuronové sítě patří mezi velmi využívané

[19] LACKO, Luboslav. Databáze: datové sklady, OLAP a dolování dat., str. 321

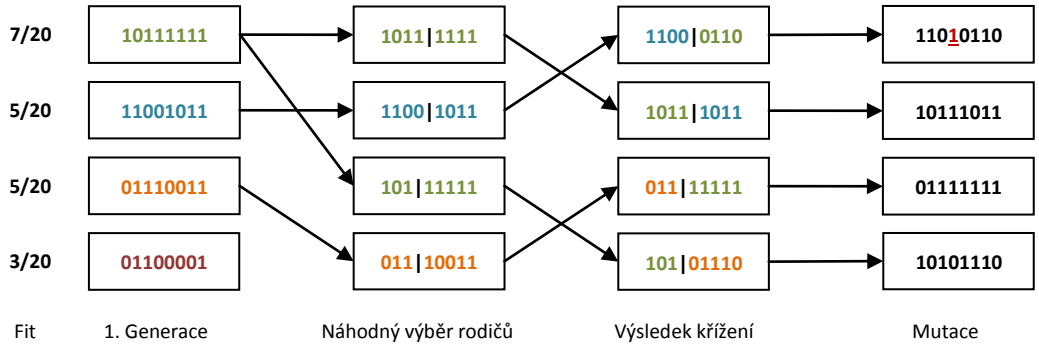
metody dolování dat z databází. Upřednostňují se především pro soubory, kde je většina veličin numerických (spojitých) a kde není požadovaná srozumitelnost nalezených závislostí. Pro kategoriální veličiny je nutné provést binarizaci dat[20]. Většina programů binarizaci dat provádí automaticky. Typickým příkladem využití neuronových sítí jsou predikce časových řád (vývoj akcií, spotřeba energie nebo meteorologická situace), další případ využití může být klasifikace.

4.6 Genetické algoritmy

Představuje druhý algoritmus nezaložený na statistickém základu, ale na biologických principech evoluce. Vychází z přirozeného výběru fungujícího v přírodě *„Genetické algoritmy nalezly uplatnění v řadě oblastí: numerická optimalizace a rozvrhování, strojové učení, tvorba modelů (ekonomických, populačních, sociálních), apod. Z hlediska dolování dat z databází je zajímavé využití genetických algoritmů přímo pro učení se konceptům, nebo použití genetických algoritmů pro optimalizaci neuronových sítí.“*[21] Pro optimalizace neuronových sítí proces nahrazuje ruční nastavování parametrů a ladění neuronové sítě. Jedinci v populaci pak odpovídají počtem neuronů, parametry neuronů a kriteriální funkcí fit[22] odpovídající chybě neuronové sítě.

Na začátku algoritmu je sestavena náhodná populace (1. generace) náhodných jedinců a postupně se použitím procesů mutace, reprodukce a selekce dochází ke zkvalitněním dalších generací až na požadovanou kvalitu. Mutace jsou tvořeny náhodnou změnou některého z bitů jedince. Reprodukce představuje proces křížení dvou jedinců nebo klonování jednoho jedince. Křížení dvou jedinců se provádí tak, že se z jednoho jedince vybere jeho část (např. počátek) a z druhého se vybere opačná část (např. konec). Spojením těchto výběrů vznikne nový jedinec v nové generaci. K selekci dochází při výběrů nejvhodnějších jedinců pro křížení nebo klonování, analogicky přirozenému výběru v přírodě.

[20] O binarizaci více v kapitole příprava dat.
[21] BERKA, Petr. Dobývání znalostí z databáze. Str. 180.
[22] K účelu zjištění kvality jedince slouží fit funkce. Fit funkce představuje jednoduchý model optimálního řešení.



Obrázek 10: ukázka selekce, reprodukce a mutace v evolučním algoritmu

4.7 Naivní Bayesova klasifikace

Bayesova klasifikace vychází ze statistické klasifikace a je postavena na Bayesově větě[23] o podmíněné pravděpodobnosti. Princip spočívá ve vypočtení pravděpodobnosti pomocí Naivní Bayesova klasifikace[24] pro různé stavy mezi popisujícími veličinami a veličinou klasifikovanou. Tyto pravděpodobnosti v případě predikce se navzájem vynásobí a vypočtou se tak aposteriorní pravděpodobnosti[25] a rozhodne se pro možnost s nejvyšší pravděpodobností. Takto spočtené pravděpodobnosti se musí ještě znormovat, tedy jednotlivé pravděpodobnosti vydělit sumou všech pravděpodobností tak, aby jejich celkový součet byl 1 nebo 100% a dal se lépe interpretovat.

Hlavní výhodou oproti rozhodovacím stromům nebo asociačním pravidlům je, že dokáže predikovat i příklady s neúplnými daty a tyto predikace pravděpodobnostně vyjádřit. Dalšími výhodami jsou velmi vysoká rychlost algoritmu a jednoduchá interpretace nalezených souvislostí.

Nevýhody vyplývají v případě nových hodnot, které neobsahovaly testovací vzorek dat. Takové hodnoty jsou ohodnoceny nulovou pravděpodobností a ostatní parametry nehrají roli. Podobně tomu je i v případě hodnot s pravděpodobností blížící se nule. Takovéto hodnoty natolik zkreslují výslednou pravděpodobnost, že se v praxi vždy používají korekce (Laplacevova korekce, m-odhad).

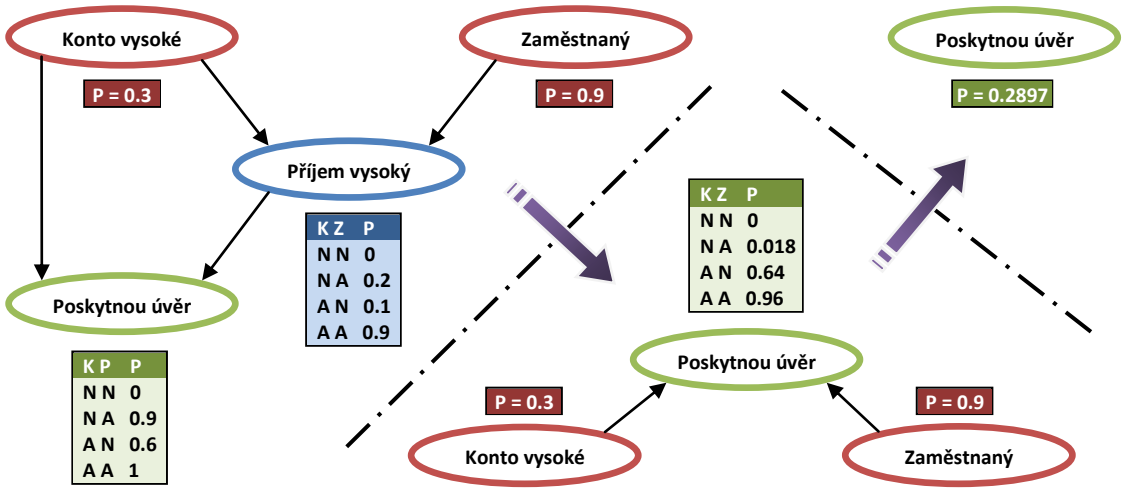
[23]Bayesova věta o podmíněné pravděpodobnosti udává, jak podmíněná pravděpodobnost jednoho jevu souvisí s opačnou podmíněnou pravděpodobností. Vzorec bayesovi věty: $P(A|B) = \frac{P(B|A)\cdot P(A)}{P(B)}$.

[24]Naivní bayesova klasifikace vychází z předpokladu, že jednotlivé evidence E1,…,EK jsou podmíněně nezávislé při platnosti hypotézy H. Vzorec pravděpodobnosti pro různé veličiny (evidence): $P(H|E_1,\dots,E_K) = \frac{P(E_1,\dots,E_K|H)\cdot P(H)}{P(E_1,\dots,E_K)}$ vynásobením tohoto vzorce $\prod_{k=1}^{K} P(E_k|H)$ získáme aposteriorní pravděpodobnost.

[25]Aposteriorní pravděpodobnost je pravděpodobnost založená na zkušenostech

Bayesovi sítě

Bayesovi sítě (Bayesian Belief Network) jsou složeny z dvou součástí - z acyklicky orientovaného grafu a ze souboru tabulek podmíněných pravděpodobností. Každý uzel v grafu představuje náhodnou proměnou, která může nabývat jak kvalitativních, tak kvantitativních hodnot. Spojnice mezi uzly vyjadřují pravděpodobnostní závislost mezi jednotlivými proměnnými. Jestliže spojnice jde z bodu y do bodu x, pak se říká, že y je rodič nebo přímý předchůdce x a x je potomek y.[26]



Obrázek 11: ukázka Bayesovi sítě a její postupné zjednodušování

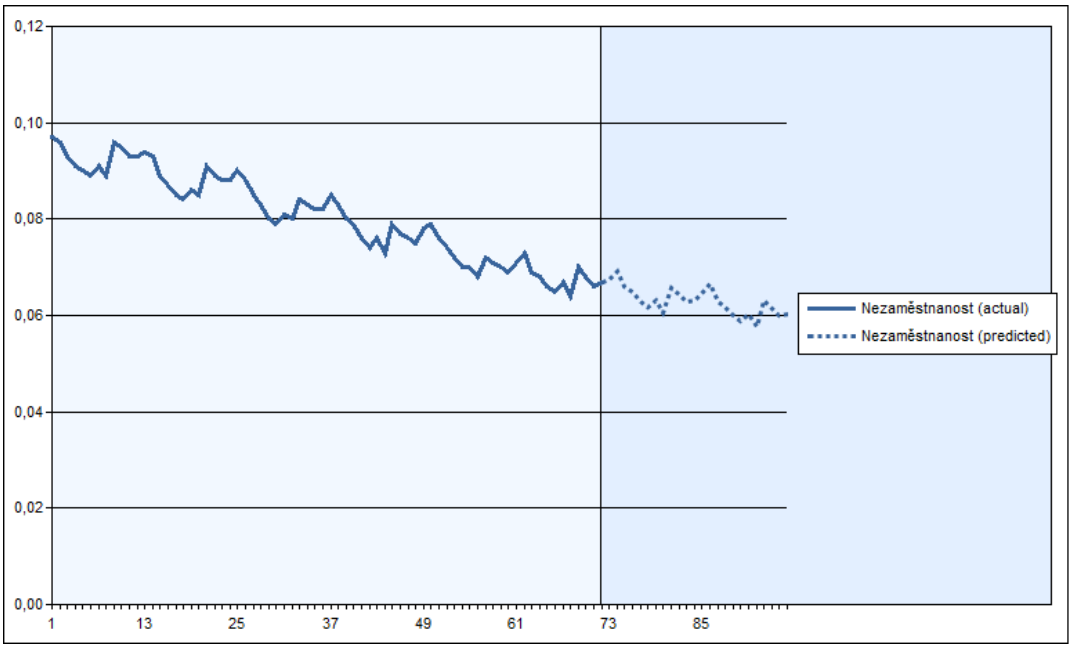
Sestavení Bayesovi sítě

- *Známá topologie sítě a všechny proměnné pozorovatelné:* pak trénink sítě je jednoduchý, neboť se jen vypočítají tabulky podmíněných pravděpodobností z trénovacího souboru dat. Výpočet se provádí stejně jako v případě Naivní Bayesovi klasifikace.
- *Neznáma topologie sítě a všechny proměnné pozorovatelné:* v tomto případě musí být odvozena i struktura samotné sítě, poté se postupuje stejně jako v předchozím případě
- *Známá topologie sítě a jen některé proměnné pozorovatelné:* řeší se podobně jako u neuronových sítí skryté vrstvy neuronů
- *Neznámá topologie a jen některé proměnné pozorovatelné:* nejsložitější varianta se dá řešit pomocí použití speciálních algoritmů

[26] Originální znění: HAN, Jiawei. KAMBER, Micheline. Data Mining: Concepts and Techniques. Str. 318

4.8 Časové řady

Analyzovaná data jsou vývoj hodnot určité veličiny v čase kupříkladu: spotřeba energie, cena akcií, nezaměstnanost, tržby a podobně. Na základě analýzy vývoje z minulosti a současnosti můžeme sestavit určitá pravidla, podle kterých můžeme predikovat vývoj veličiny. Algoritmus pro analýzy časových řad vychází ze statistického pohledu na časové řady. Model časových řad ze statistického hlediska představuje 4 základní složky: trend, sezónní, cyklickou a nahodilou složku.



Obrázek 12: predikce vývoje nezaměstnanosti (MS SQL Server)

Trend - dlouhodobá tendence změny hodnot v závislosti na čase

Sezónní složka - pravidelně se opakující složka s periodou opakování menší než 1 rok

Cyklická složka - pravidelně se opakující složka s periodou opakování větší než 1 rok

Náhodná složka - se nedá popsat žádnou funkcí času, jedná se o zbytek po odečtení ostatních složek

5. Příprava dat

Je druhou fází v metodice CRISP-DM. Vzhledem k různorodým požadavkům algoritmů na vstupní data, je zařazena tato kapitola, až po kapitole vysvětlující potřeby jednotlivých algoritmů. Přípravou dat se rozumí proces přetransformování dat z hlediska jejich struktury a jednotnosti tak, aby je dokázal příslušný algoritmus pro dolování dat správně pochopit a použít. Tento krok je zpravidla nejnáročnější z celého procesu dolování dat.

5.1 Typy atributů

5.1.1 Kategoriální atributy

Ve statistice se používá termín kvalitativní veličiny (diskrétní). Většina algoritmů pro dolování dat pracuje s kategoriálními atributy. V takovém případě se nemusí nic provádět s hodnotami s výjimkou přílišného množství hodnot. Pakliže tomu tak není, je nutno přistoupit k binarizaci hodnot.[27]

5.1.2. Numerické atributy

Ve statistice se používá termín kvantitativní veličiny (spojité). Vzhledem k využívání převážně kategoriálních atributů v dolování dat, je nutné v mnoha případech numerické atributy převést na kategoriální atributy neboli diskretizovat.

Diskretizace

K účelu diskretizace slouží velké množství různorodých postupů, jež se hodí k určitým úlohám. Jednoduché diskretizace jsou např. rozdělení na předem daný počet intervalů nebo na intervaly s přibližně stejným počtem prvků. V těchto postupech se dělí pouze na základě numerických hodnot. Vhodnějšími postupy jsou dělení na základě příslušnosti hodnot v konkrétních skupinách a tím pádem s vyšší informační hodnotou. Procesem diskretizace se přichází o určitě informace, které se mohou projevit ve sloučení různých prvků v jeden, a tím pádem se snižuje schopnost maximální přesnosti klasifikace.

[27] Binarizace kategoriálních dat se provádí tak, že pro každou možnost je přidána nová veličina např.: pro barvu očí musí být přidány veličiny barva_očí_modra, barva_očí_hnědá, barva_očí_zelená atd. do níž zapíšeme v 1. případě, že daný případ nabývá nově vzniklé veličiny a 0 pakliže nenabývá. Stejný postup pro kategoriální veličiny se využívá i v případě regresní analýzy.

5.2 Zpracování zdrojových dat

Zdrojová data mají většinou množství nedostatků, bez kterých není možné provést algoritmus dolování dat.

5.2.1 Výběr a spojení dat

Vzhledem k nejčastějšímu uložení dat v relačních tabulkách je nutné správně vybrat a navzájem propojit data nacházející se v různých tabulkách. Při této operaci se musí počítat s různými vztahy mezi jednotlivými tabulkami (1:1, 1:n, n:m). Relace 1:1 je nejlehčí, jedná se jen o vzájemné spojení atributů. Při relaci 1:n se v případě tabulky n většinou používají agregace pro numerické atributy operace sumarizace, maximum, minimum, průměr a jiné a pro kategoriální atributy počty různých hodnot, výskyt konkrétní hodnoty, majoritní hodnoty. Vztah n:m lze v podstatě převést na vztah 1:m respektive n:1 tím, že si se vybere jeden z atributů a ten se bude sledovat, tedy přibude n nebo m řádků.

5.2.2 Vytváření atributů

Některé atributy je potřeba odvodit z jiných. Často se potřebuje vybrat část/i nějakého delšího řetězce, sloučit některé sloupce nebo převést rodné číslo na věk pohlaví atd. Pro potřebu dolování dat je také potřeba některé atributy agregovat. Agregace se provádí stejně, jako tomu bylo v propojování tabulek.

5.2.3 Příliš mnoho atributů

V databázích se nachází velké množství atributů, z nichž jen část je vhodná pro zamyšlenou hypotézu. Po sjednocení a selekci tabulek lze mít v základním souboru desítky až stovky atributů. Otázka je, jak toto množství nejvhodněji redukovat. Pomoci může odborník nebo automatické metody, které provádějí redukci dvěma způsoby:

- transformací, kdy se z existujících atributů vytvoří méně nových atributů
- selekcí, kdy se z existujících atributů vyberou jen ty nejdůležitější

Při transformaci se ze skupin atributů vytvoří jejich model a na základě tohoto modelu se dopočítají nové atributy. Nevýhodou při transformaci je, že transformované atributy musí být plně popsány, aby se na základně modelu daly stanovit nové atributy a že nově vzniklé atributy jsou složitěji interpretované.

Při automatizované selekci atributů se hledají atributy, jež nejlépe rozdělují trénovací data do predikovaných skupin. Opět se využívá entropie, informačního zisku nebo χ_2 k určení nejvlivnějších atributů. Takto se nazývá metoda filtrů a je nevýhodná v tom, že sleduje nejvlivnější hodnoty jen pro jednotlivé atributy a ne pro jejich skupiny. Vhodnější je místo pro jeden atribut počítat vliv pro skupiny atributů. Tento postup je však výrazně počítačově náročnější.[28]

5.2.4 Příliš mnoho objektů

V reálných úlohách je možno se často setkat s databázemi obsahujících desítky až stovky tisíc záznamů. Výjimkami nejsou ani větší databáze. Takovéto množství dat není možné zpracovat naráz dávkovým způsobem, neboť by došlo k zahlcení operační paměti.

Využívá se především výběru určitého vzorku trénovacích dat z celého souboru. Je důležité, aby v trénovací data vystihovala (reprezentovala) původní soubor dat. Nesmějí vybírat jen určité skupiny hodnot (pakliže to není zadáním úlohy). Vhodné je vybírat data náhodným procesem kde je téměř 100% šance vybrat reprezentativní data. Další možností zpracování velkého množství objektů se nabízí použití takového uložení dat, které umožní nedávkové zpracování dat nebo vytvoření více modelů a poté tyto modely zkombinovat a vytvořit tím model jediný.

5.2.5 Chybějící hodnoty

Chybějící hodnoty představují závažný problém pro algoritmy dolování dat a jsou možné doplnit různými způsoby:

1. ignorováním objektů s některou chybějící hodnotou
2. nahrazením chybějící hodnot novou hodnotou „neznámá“
3. nahrazením chybějících hodnot některou z hodnot atributů
    a. nejčastěji se vyskytující hodnotou
    b. průměrem hodnot
    c. libovolnou hodnotou
    d. použitím predikčních modelu vytvořeného na základě známých hodnot pomocí metod dolování dat[29]

[28] BERKA, Petr. Dobývání znalostí z databáze. Str. 254.
[29] BERKA, Petr. Dobývání znalostí z databáze. Str. 267.

6. Vlastní práce

Předmětem vlastní práce je analýza dat pomocí metod dolování dat. Cílem je zjištění charakteristických znaků pro různé druhy událostí, které vznikají na internetových stránkách jedné nejmenované instituce a na základě těchto charakteristik sestavit klasifikační model jednotlivých událostí.

K provedení procesu dolování dat software Microsoft SQL Server 2008 R2 v 60ti denní trial verzi, jehož součástí je i SQL Server Business Intelligence Development Studio (BI Studio). Tato aplikace, mimo celé řady jiných služeb, poskytuje poměrně rozsáhlé nástroje a celou řadu algoritmů pro provádění dolování dat. Instalace MS SQL Serveru obsahuje také nástroj Import and Export Data, který se využil k jednoduchému převodu dat z MS Excelu na MS SQL Server.

6.1 Analýza a úprava dat

Analýza a úprava dat byla provedena z větší části v Microsoft Excelu a drobné úpravy následně v SQL Server Management Studio. Zdrojová tabulka obsahovala přes 230 tisíc řádků pro 9 atributů: datum a čas, typ chyby, jméno uživatele, id uživatele, ip adresu uživatele, informace o internetovém prohlížeči uživatele, referenční adresu a žádanou adresu. Prvních přibližně 4000 řádků neobsahovalo hodnoty pro většinu sledovaných atributů, a proto byly z tabulky odstraněny. Tab. 2 ukazuje základní soubor v „surové“ nepřipravené podobě. Ve sloupci popis bylo původně konkrétní příjmení a jméno, stejně tak i v referenční adrese byla původní stránka nahrazena slovem „neznáme“ z důvodu ochrany osobních dat.

Tabulka 2: ukázka dat před přípravou dat

id	datum	udl.	Popis	ur_id	ur_ad	user_browser	ur_referer	ur_req.
6746	8.9.2006 12:42	1	[Příjmení a jméno]	4103	162.191. 121.225	Mozilla/5.0 (Windows; U; Windows NT 5.1; cs; rv:1.8.0.6) Gecko/20060728 Firefox/1.5.0.6	https://neznáme.cz/	/i/process.php
6774	8.9.2006 15:17	2	Chybné uživatelské jméno nebo heslo….	NULL	162.191. 121.225	Mozilla/5.0 (Windows; U; Windows NT 5.1; cs; rv:1.8.0.6) Gecko/20060728 Firefox/1.5.0.6	https://neznáme.cz/	/i/process.php
6976	12.9.2006 14:36	8	Chyba databáze: Nepodařilo ze ods…	4213	162.191. 121.66	Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.7.13) Gecko/20060414	https://neznáme.cz/nastaveni.php?grant=1	/i/process.php

Predikovaná veličina událost nabývá 6 různých hodnot a to 0 (neznámá chyba), 1 (úspěšné přihlášení), 2 (neúspěšné přihlášení), 4 (nedostatečná práva), 8 (chyba databáze) a 16 (zacyklení či uvíznutí ve slepé uličce). Numerické hodnoty byly nahrazeny slovním vyjádřením jednotlivých chyb.

6.1.1 Nové atributy a jejich redukce

Popis internetového prohlížeče obsahoval mnoho dílčích informací a musel být tedy rozdělen tak, aby jej algoritmus pro dolování dat mohl zpracovat. Rozdělením bylo získáno 3 nových atributů (název a verze prohlížeče, název a verze operačního systému a jazyková sada). Bohužel v některých případech nebylo možno nové atributy vyplnit (zůstaly prázdné). Po tomto rozdělení byl původní atribut popisu internetového prohlížeče odstraněn. Dále byl odstraněn popis uživatele z důvodu ochrany osobních údajů uživatele, v ostatních případech byl vypsán slovy popis chyby.

6.1.2 Redukce počtu obměn atributů

IP adresa uživatele byla převedena na typ adresy (privátní síť nebo globální síť). Hodnoty referenční a žádané adresy byly očištěny od proměnných přenášených pomocí url a dále seskupeny do početnějších skupin. Verze prohlížečů a operačního systému byly zkráceny na maximálně jednu pod-verzi.

Spojité veličiny datum a id uživatele bylo nutné diskretizovat vzhledem k nevhodné automatické diskretizaci, jenž rozděluje soubor na stejně velké části podle hodnoty. Id uživatele se rozdělilo do tří skupin (známé, neznáme a nulové). Tímto rozdělením se vypovídací schopnost proměnné značně zvýšila. Datum byl diskretizován na pět různě dlouhých intervalů tak, aby co nejlépe rozděloval soubor podle různých druhů chyb. Tyto změny dat byly provedeny na základě dílčích analýz pomocí dolování dat.[30]

Tabulka 3: ukázka dat po přípravě dat

ID	Datum	Událost	ID_uživ.	IP	Prohlí.	OS	Jazyk	Ref.	Pož._adr.
25260	< 29.3.2007 12:00	Neúspěšné přihlášení	Neznámé	Privátní síť	Firefox 2.0	Windows NT 5.1	cs-CZ	/	/objednavky.php
20606	< 29.3.2007 12:00	Úspěšné přihlášení	Známé	Privátní síť	Firefox 2.0	Windows NT 5.1	cs	/	/wise/process.php
219868	< 29.3.2007 12:00	Nedostate-čná práva	Známé	Privátní síť	Firefox 2.0	Windows NT 5.1	en-US	/	NULL

[30] Tyto průběžné analýzy stromu jsou zdokumentovány v příloze

6.2 Provedení algoritmů pro dolování dat

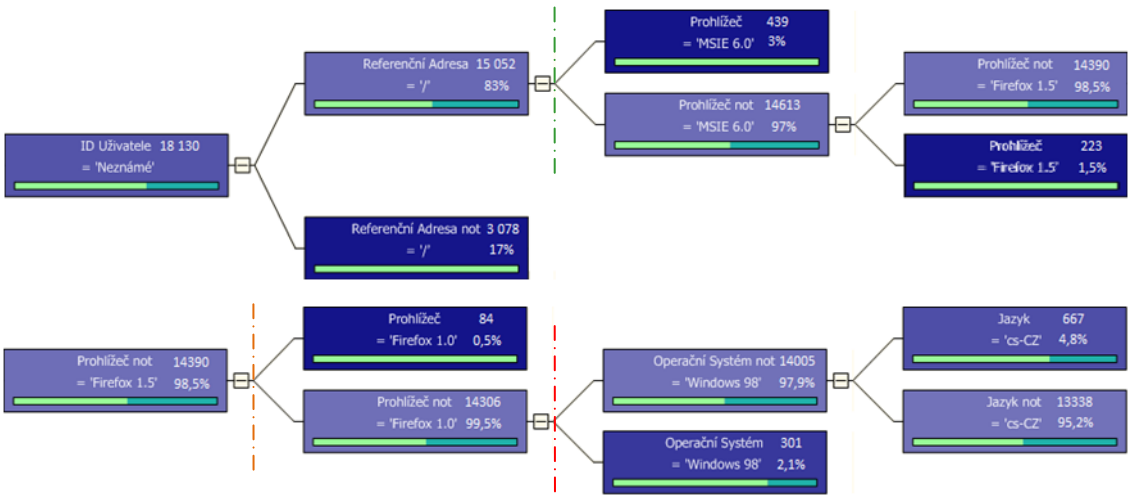
K vytvoření data mingové struktury (v BI Studiu) slouží průvodce dolováním dat, kde se postupně volí typ zdroje dat, typ algoritmu pro dolování dat, specifikace tréninkových dat, určení typů jednotlivých atributů, množství testovacích dat a název struktury. Stěžejní je specifikace tréninkových dat. V tomto kroku je zapotřebí určit zdrojová data a v těchto datech vybrat identifikační, predikovaný a vstupní atribut. K určení tréninkových dat napomůže tlačítko „Suggest“. Poté vyskočí nové okno, ve kterém jsou uvedeny predikční schopnosti jednotlivých atributů. Dalším důležitým dialogem je typ atributů, které lze také určit automaticky tlačítkem „Detect“. Pakliže jsou vyplněny veškeré dialogy, kliknutím na tlačítko „Finish“ a následně spustit „Development“ je získán „Mining Model“.

6.2.1 Prořezávání stromů

Prořezávání větví stromů má za cíl zjednodušit strukturu daného stromu tak, aby byl z pohledu uživatele co nejlépe srozumitelný za cenu mírného zhoršení predikčních schopností modelu.

Přeučení

Přeučení vzniká důsledkem snahy algoritmu zlepšovat klasifikační vlastnost modelu bez ohledu na složitost klasifikační struktury. V případě větve, která se rozděluje do dvou či více listů a jeden list je značně dominantní a má takřka totožné vlastnosti jako původní větev, s největší pravděpodobností se jedná o přeučení.

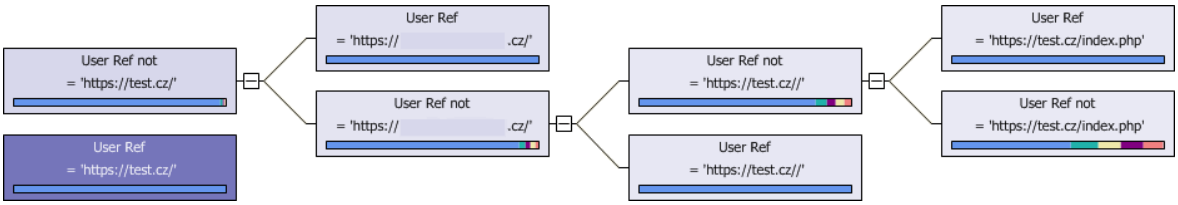


Obrázek 13: ukázka přeučení s vyznačenými možnosti řezu

Přeučení se odstraní nahrazením větve za list, neboli odstranění dalšího větveni. Obr. 11 ukazuje příklad přeučení, čerchované čáry označují možné řezy. Zelený řez vyznačuje značné zjednodušení, oranžový řez představuje střední volbu a červený řez je nejzazší, neboť předchozí větvení mají alespoň funkci filtrovací (vzniklý list obsahuje pouze jeden typ události).

Sjednocení listu[31]

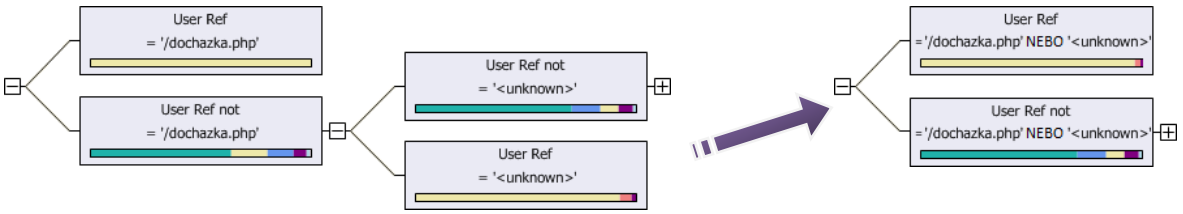
Na obr. 12 je vidět nádherný příklad růstu větvení do délky způsobený různými hodnotami se stejným významem. V našem případě se jedná stále o soubor index.php, ke kterému je možno přistupovat pomocí url více způsoby 1., 3. a 4. větev představuje naprosto stejný soubor. V 2. případě může jít o stejný soubor uložený na jiném serveru, stále však jde o index.php. Proto všechny adresy končící na „/, //, ///, nebo index.php“, byly převedeny na jednotný název „/“. Po následném generování stromu vznikne na místě dlouhého řetězce pouze jedno rozdělení. Takováto úprava je nezbytná pro zkvalitnění modelu jak po stránce zkrácení struktury, tak po stránce predikčních vlastností.



Obrázek 14: ukázka růstu stromové struktury do délky na základě

Sdružování listů

K tomuto kroku se přistupuje, jestliže je několik listů různých významů, ale s totožnou predikční vlastností. Sníží se počet listů a počet větví (za předpokladu že nebylo ještě provedeno zhušťování listů). Při sdružování listů dochází ke ztrátě informací, je třeba si být naprosto jistý tím, že toto sdružení negativně neovlivní ostatní větve stromu.



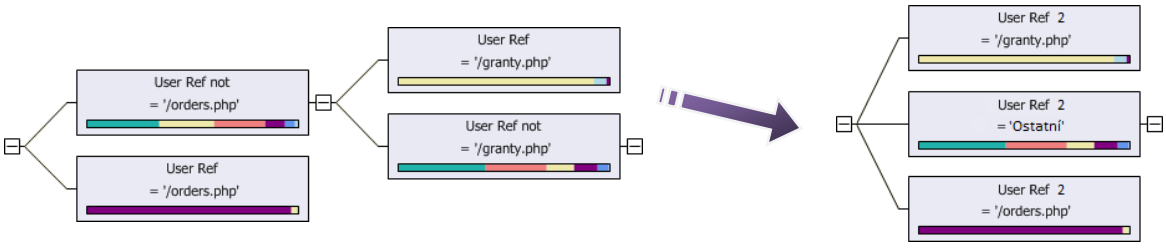
Obrázek 15: ukázka prořezávání větví sdružováním listů

[31] Názvy postupů byly odvozeny od typu úpravy, nejedná se o ustanovené termíny

Zhušťování listů

Dalším typ nevhodného větvení rozšiřující se opět do délky. Vzniká podobně jako v předchozím případě u jednoho diskrétního atributu, v němž některé hodnoty jsou pro určení velice vhodné a ostatní jsou nevhodné. V takovémto případě vzniká struktura podobná té na obr. 14 vlevo. Prořezání takovéto struktury je oproti předchozí situaci podstatně náročnější.

Nejprve se vytvoří nový atribut s podobným názvem atributu původnímu. U nově vzniklého atributu se nastaví jako výchozí hodnota např. „ostatní“, poté si pro řádky, v nichž jsou prořezávané hodnoty, vymění navzájem s novým atributem hodnoty. V nově vzniklém atributu tedy bude pro vybrané řádky hodnota z původního atributu a v původním atributu pro tytéž řádky bude „ostatní“. Takto lze odstranit několik větví při zachování stejného počtu listů. Velkou výhodou zhušťování listů představuje neovlivnění predikčních schopností modelu, v podstatě jde o „kosmetickou“ úpravu struktury a správně provedené změny je možné vrátit do původního stavu.



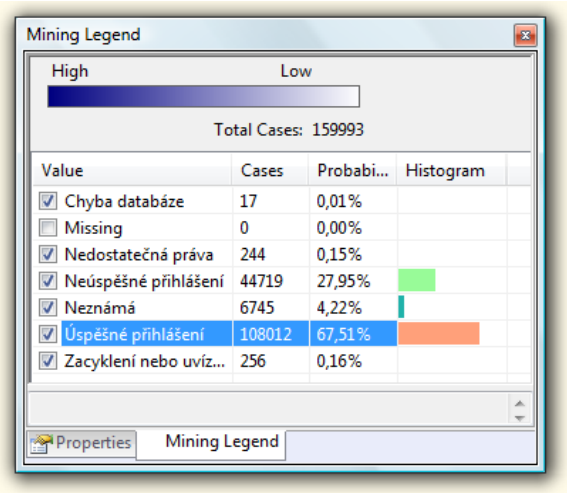
Obrázek 16: ukázka prořezávání větví zhušťováním listí

„Dvakrát měř jednou řež“

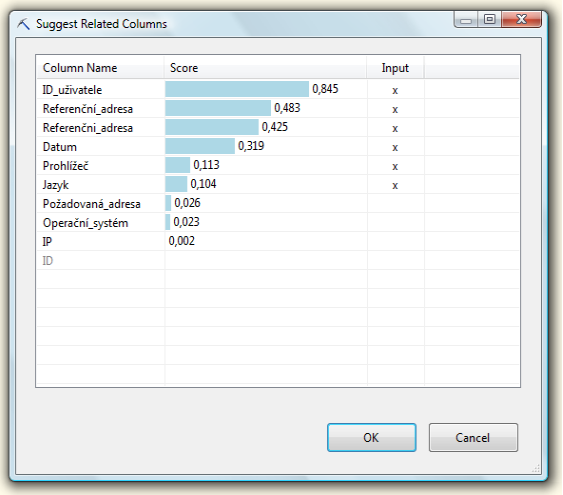
Při prořezávání stromové struktury se musí logicky uvažovat a pečlivě volit větve k odstranění. Především je třeba dbát na to, aby se prořezáváním negativně neovlivnily neprořezávané větve. V koncovém rozhodovacím stromu nebyla uříznuta větev pro referenční adresy, které nejsou ‘/’ a obsahují pouze 0,87 % případů pro známé ID uživatele. Tato větev má obrovskou vypovídající schopnost neboť z celkového počtu obsahuje 94 % chyb databáze, 87 % nedostatečných práv a 88 % zacyklení nebo uvíznutí. Uříznutím této větve by byl náš rozhodovací strom naprosto degradován z důvodu nemožnosti predikace těchto méně častých o to neméně důležitých událostí.

7. Zhodnocení výsledků

Pakliže je vytvořen konečný „Mining Model“, lze přejít k vyhodnocení výsledků a následně i k jejich zhodnocení. Nejprve bude shrnuta základní charakteristika tréninkových dat, jež byla zjištěna z kořene rozhodovacího stromu a z automatického návrhu určujícího vstupních atributy. Na obr. 11 je vidět, že nejčastější událostí je úspěšné přihlášení (67,5%) a neúspěšné přihlášení (27,95%), následuje je neznámá chyba (4,22%). Další chyby jsou výjimečné a dohromady představují jen 0,323% případů ze všech chyb. Celkově tréninková data měla bezmála 160 tis. případů a testovací data přibližně 68,5 tis. případů. Z obr. 12 jasně vyplývá, které atributy nejvíce určují typ události. Především jsou to atribut ID uživatele, Referenční adresa, Datum, dále pak Prohlížeč a Jazyk.



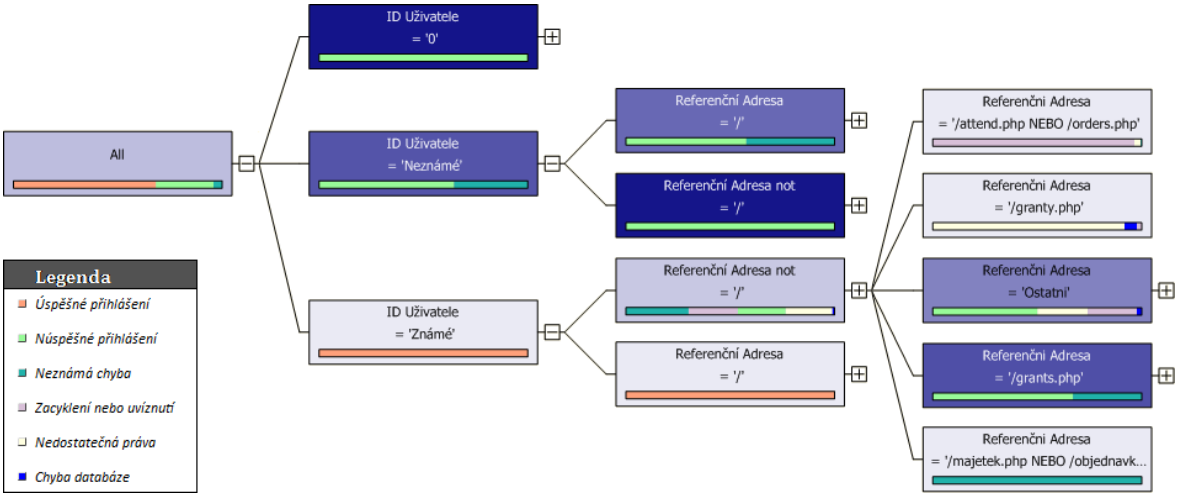
Obrázek 17: zastoupení jednotlivých událostí



Obrázek 18: určující schopnosti k určení události

Na základě výsledného rozhodovacího stromu obr. 17 [32], jde jednoduše určit, kde dochází k jaké události a určit tak klasifikační model. Klasifikačním modelem[33] v našem případě poslouží jednoduchá rozhodovací pravidla, ke kterým byla přidána i pravděpodobnost a počet případů pro jednotlivá pravidla. Pravděpodobnost a počet případů jsou získány z okna „Mining Legend“ a jsou vypočteny z tréninkových dat. Jednoduchým výpočtem se zjistí, že na základě našeho částečného klasifikačního modelu a zařazováním do listů podle největší pravděpodobnosti, došlo k zařazení 95,78% případů správně.

[32] Rozšířený rozhodovací strom přiložen jako 6. příloha
[33] Tabulka s rozhodovacími pravidly se nachází v 7. příloze



Obrázek 19: nejdůležitější části konečného rozhodovacího stromu

V BI studio bylo vybráno 30 % dat jako testovací data. Výsledky vypočtené z testovacích dat se nachází v záložce „Mining Accuracy Chart“, nejdůležitější bude pod-záložka „Classification Matrix“ jejíž výsledky jsou zobrazeny v tabulce 4. Správně predikované případy se nacházejí na hlavní diagonále, ostatní predikce jsou chybné. Sloupce určují příslušnost jednotlivých chyb tak, jak jsou popsány v testovacích datech. Řádky popisují predikce na základě predikčního modelu.

Úspěšné přihlašování bylo predikováno ve všech případech správně, neúspěšné přihlášení v 99,87%, nedostatečná práva v 92,7%, zacyklení nebo uvíznutí v 88,98%, neznámá chyba v pouhých 3,58% a chyba databáze nebyla ani jednou predikovaná správně. Neznámá chyba nebyla predikovaná správně, neboť se především vyskytuje v listech společně s majoritním neúspěšným přihlášením. Chyba databáze se vyskytuje velmi zřídka. V listech obsahujících chybu databáze, je vždy v minoritním zastoupení. Celková pravděpodobnost úspěšné predikce jednotlivých událostí dosahovala pro testovací data hodnoty 95,82 %, bohužel 2 z 6 událostí není možné predikovat.

Tabulka 4: predikce událostí z testovacích dat

Predicted	Chyba databáze ...	Nedostatečná práva ...	Neúspěšné přihlášení ...	Neznámá ...	Úspěšné přihlášení ...	Zacyklení nebo uvíznutí ...
Chyba databáze	0	0	0	0	0	0
Nedostatečná práva	5	89	1	0	0	1
Neúspěšné přihlášení	1	0	19327	2755	0	0
Neznámá	0	0	3	103	0	0
Úspěšné přihlášení	0	6	1	17	46117	12
Zacyklení nebo uvíznutí	3	1	21	0	0	105

8. Závěr

Bez znalosti zdrojových kódů bylo na základě dolování dat zjištěno, v jakých souborech a pro jaké uživatele dochází nejčastěji k různým událostem. Takto získané informace se dají využít především v příliš rozsáhlých projektech nebo tam, kde není příliš kvalitní dokumentace a dále k získání kritických míst v celém projektu.

Predikční model sestavený z rozhodovacího stromu dokáže zařadit 4 z 6 událostí s pravděpodobnostmi predikce jednotlivých událostí od 88,98 % do 100 % podle druhu chyby. Při výběru jen těchto 4 hodnot událostí byla pravděpodobnost správného zařazení 99,93 %, v případě všech událostí se snížila pravděpodobnost na 95,82 %.

Využití MS Excelu pro úpravu tabulky s 230 tis. řádky se neukázalo jako příliš vhodné, jelikož některé funkce, složené z více funkcí, trvaly řádově desítky vteřin až jednotky minut, či dokonce nemohly být zpracovány najednou pro tak velké množství dat. V závěru přípravy dat a prořezávání větví stromu byl využit SQL Server Management Studio, s jehož pomocí se práce velmi urychlila. Fáze zpracování dat představovala jednoznačně nejdelší část celého procesu dolování dat. Tento čas by se dal o dost snížit lepším uložením dat v databázi, nebo převáděním dat z této databáze do datového skladu či datového trhu s příslušnou transformací dat.

9. Seznam použitých zdrojů

9.1 Seznam použité literatury

Knižní publikace

[1] LACKO, Luboslav. Databáze: datové sklady, OLAP a dolování dat. 1. vyd. Brno: Computer Press, 2003. 486str + 1CD ROM. ISBN 80-7226-969-0

[2] BERKA, Petr. Dobývání znalostí z databází. 1. vyd. Praha: Academia, 2003. 366s + 1CD ROM. ISBN 80-200-1062-9

[3] RUD, Olivia Parr. Data mining : praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM). 1. vyd. Praha: Computer Press, 2001. 326str + 1CD ROM. ISBN 80-7226-577-6

[4] LACKO, Luboslav. Business Intelligence v SQL Serveru 2008. 1. Vyd. Brno: Computer Press, 2009. 456str. ISBN 978-80-251-2887-9 978-1-55860-901-3OO

[5] HAN, Jiawei. KAMBER, Micheline. Data Mining: Concepts and Techniques. 2. vyd. San Francisco: Morgan Kaufmann, 2006. 770str. ISBN 978-1-55860-901-3

Technická zpráva

[6] RUSSELL, Stuart. BINDER, John. KOLLER, Daphne. Adaptive probabilistic networks. Berkeley: University of California, 1994. 11str.
Přístupná na: http://www.eecs.berkeley.edu/Pubs/TechRpts/1994/CSD-94-824.pdf

Internetové zdroje

[7] Process Model, [on-line]. Přístupný na: http://www.crisp-dm.org/Process/index.htm

[8] Strojové učení, [on-line]. Přístupný na: http://en.wikipedia.org/wiki/Machine_learning

[9] SÝKORA Lukáš, Dobývání znalostí z databází pro účely business inteligence, [on-line]. Přístupný na:
http://www.lukassykora.cz/content/dob%C3%BDv%C3%A1n%C3%AD-znalost%C3%AD-z-datab%C3%A1z%C3%AD-pro-%C3%BA%C4%8Dely-business-intelligence

[10] Data mining, [on-line]. Přístupný na: http://en.wikipedia.org/wiki/Data_mining

Video přednášky

[11] DASGUPTA, Pallab. Lecture - 22 Bayesian Networks. Kharagpur: India Institut of Technology, 30. 4. 2008. 60 minut. [on-line]. Přístupný z: http://www.youtube.com/watch?v=rFQsbArQE6Y

[12] DASGUPTA, Pallab. Lecture - 23 Reasoning with Bayes Networks. Kharagpur: India Institut of Technology, 30. 4. 2008. 60 minut. [on-line]. Přístupný z: http://www.youtube.com/watch?v=cMN6ykIYF_U

[13] DASGUPTA, Pallab. Lecture - 26 Learning : Decision Trees. Kharagpur: India Institut of Technology, 30. 4. 2008. 60 minut. [on-line]. Přístupný z: http://www.youtube.com/watch?v=pMHOPezBUfU

[14] DASGUPTA, Pallab. Lecture - 27 Learning : Neural Networks. Kharagpur: India Institut of Technology, 30. 4. 2008. 60 minut. [on-line]. Přístupný z: http://www.youtube.com/watch?v=6ixqKw7uK6o

Počítačové prezentace

[15] HOŠKOVÁ, Pavla. Regresní a korelační analýza – Regresní analýza. 36 snímků.

[16] HOŠKOVÁ, Pavla. Analýza kategoriálních dat. 42 snímků.

[17] HOŠKOVÁ, Pavla. Analýza časových řad. 16 snímků.

[18] HOŠKOVÁ, Pavla. Modelování časových řad. 54 snímků.

9.2 Seznam obrázků

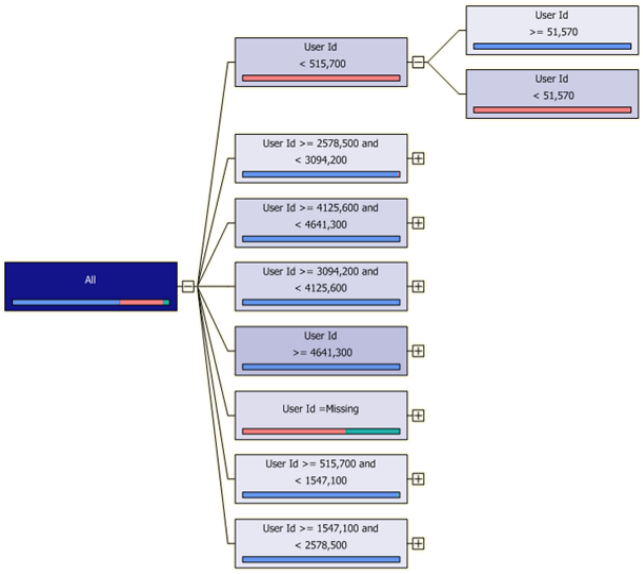
9.3 Seznam tabulek

10. Přílohy

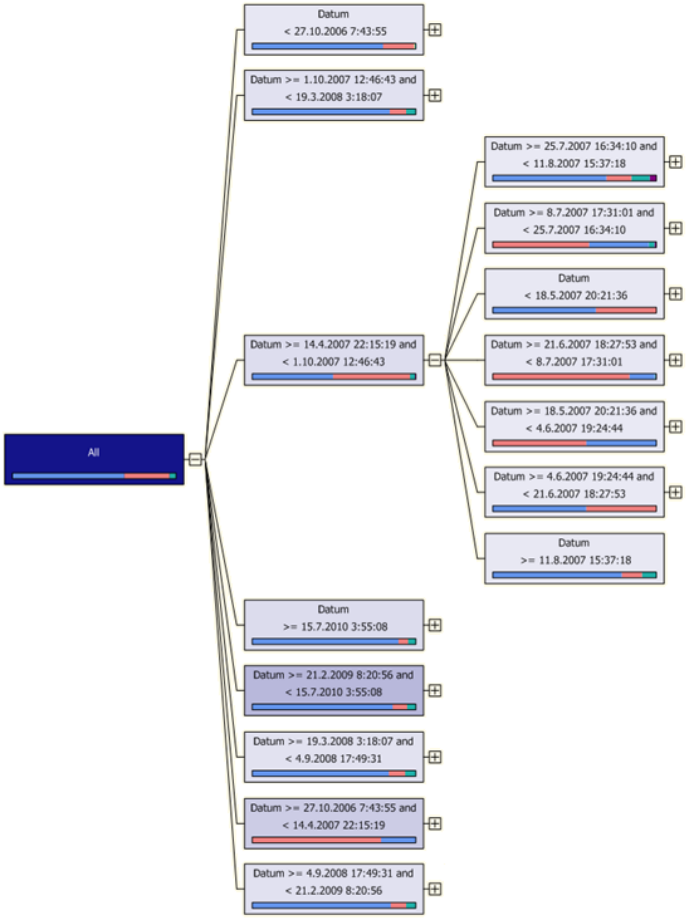
Tabulka popisuje, jaké algoritmy jsou obsažené v jednotlivých softwarech a zda se jedná o komerční či nekomerční distribuci.

Software obsahující součásti dolování dat		Rozhodovací stromy	Rozhodovací pravidla	Asociační pravidla	Shluková analýza	Neuronové sítě	Bayesova klasifikace	Časové řady	Další algoritmy	Komerční
Společnost	Produkt									
Microsoft	SQL Server 2008	☑	☑	☑	☑	☑	☑	☑	☒	☑
IBM	DB2	☑	☑	☑	☑	☑	☒	☒	☑	☑
Oracle	11g	☑	☑	☑	☑	☒	☑	☒	☑	☑
SAS	Enterprise Mine	☑	☑	☑	☑	☑	☒	☑	☑	☑
StatSoft	STATISTICA Data Miner	☑	☑	☑	☑	☑	☑	☒	☑	☑
IBM	SPSS Modeler	☑	☑	☑	☑	☑	☒	☒	☑	☑
SAP	SAP	☑	☑	☑	☑	☒	☒	☒	☑	☑
The University of Waikato	WEKA	☑	☑	☑	☑	☒	☑	☒	☑	☒
Université Lumière Lyon 2	Tanagra	☑	☑	☑	☑	☑	☑	☒	☑	☒
Rapid-I GmbH	RapidMiner (YALE)	☑	☑	☑	☑	☑	☑	☒	☑	☒
Bayesia	BayesiaLab	☒	☒	☒	☒	☒	☑	☒	☒	☑

Příloha 1: tabulka softwaru pro dolování dat

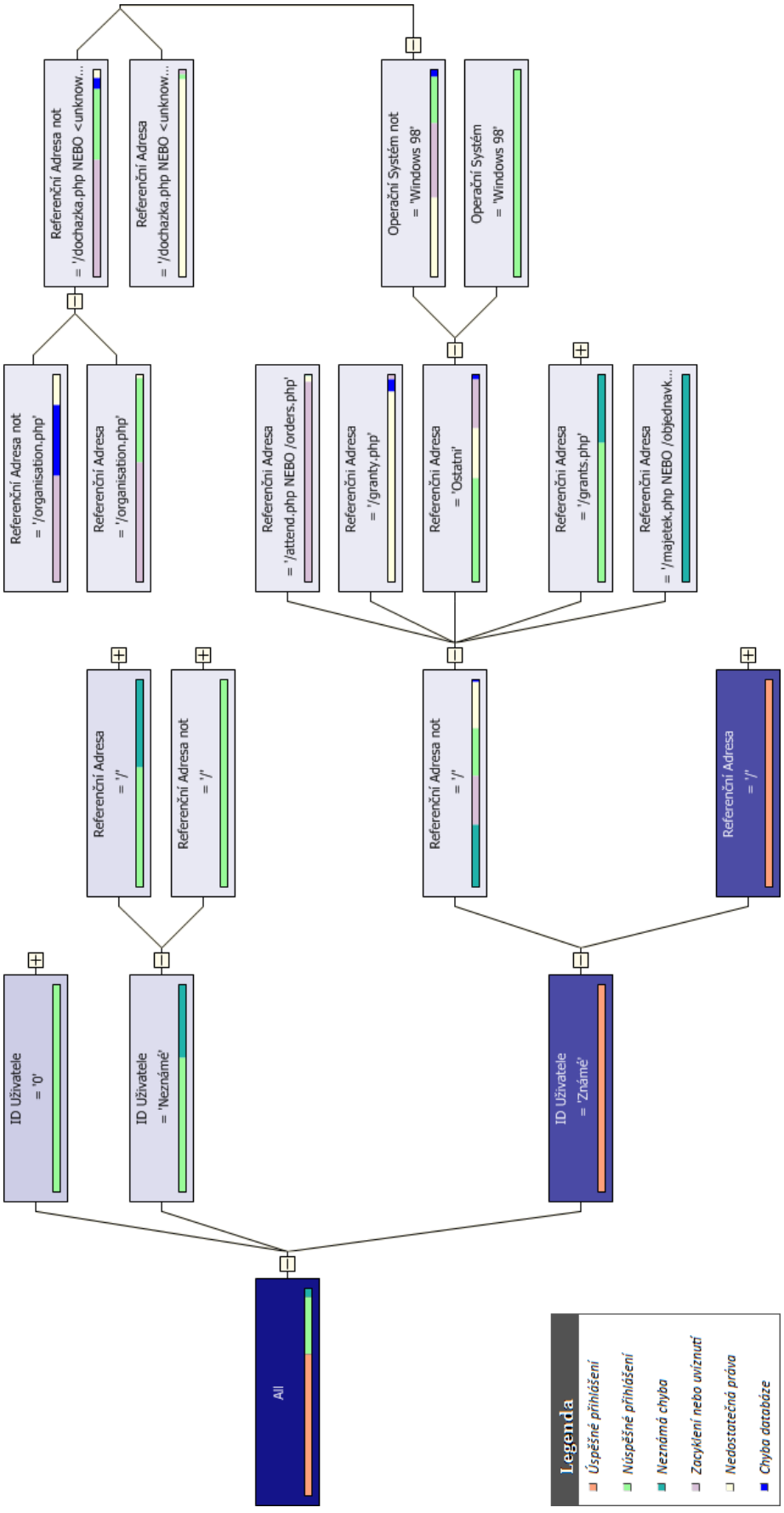


Příloha 2: obrázek části rozhodovacího stromu pro diskretizaci id uživatele



Příloha 3: obrázek části rozhodovacího stromu pro diskretizace data

Příloha 4: schéma po prvotním modelování



Příloha 5: schéma konečný nevyvážený rozhodující strom

Podmínka	Závěr	Prav.	Počet
ID Uživatele = 0	Neúspěšné přihlášení	99,98%	32797
ID Uživatele = 'Neznámé' and Referenční Adresa not = '/'	Neúspěšné přihlášení	99,58%	3078
ID Uživatele = 'Známé' and Referenční Adresa = '/'	Úspěšné přihlášení	99,90%	108121
ID Uživatele = 'Známé' and Referenční Adresa = '/majetek.php NEBO /objednavky.php'	Neznámá chyba	99,78%	230
ID Uživatele = 'Známé' and Referenční Adresa = '/attend.php NEBO /orders.php'	Zacyklení nebo uvíznutí	95,90%	158
ID Uživatele = 'Známé' and Referenční Adresa = '/granty.php'	Nedostatečná práva	90,97%	149
ID Uživatele = 'Neznámé' and Referenční Adresa = '/'	Neúspěšné přihlášení	57,37%	15052
	Neznámá chyba	42,63%	
ID Uživatele = 'Známé' and Referenční Adresa = '/grants.php'	Neúspěšné přihlášení	66,41%	117
	Neznámá chyba	33,25%	
ID Uživatele = 'Známé' and Referenční Adresa = 'Ostatní'	Neúspěšné přihlášení	49,07%	291
	Nedostatečná práva	24,38%	
	Zacyklení nebo uvíznutí	24,04%	

Příloha 6: rozhodovací pravidla sloužící k predikci události