TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

ASOCIAČNÍ METODY V DATAMININGOVÝCH ÚLOHÁCH

Bakalářská práce

Studijní program: B2612 – Elektrotechnika a informatika
Studijní obor: 2612R011 – Elektronické informační a řídicí systémy

Autor práce: **Markéta Malá**
Vedoucí práce: RNDr. Klára Císařová, Ph.D.

Liberec 2015

TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky a mezioborových studií
Akademický rok: **2014/2015**

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Markéta Malá**
Osobní číslo:	**M12000309**
Studijní program:	**B2612 Elektrotechnika a informatika**
Studijní obor:	**Elektronické informační a řídicí systémy**
Název tématu:	**Asociační pravidla v dataminingových úlohách**
Zadávající katedra:	**Ústav mechatroniky a technické informatiky**

Zásady pro vypracování:

1. Seznamte se s dataminingem a nástrojem IBM SPSS Modeler.
2. Prostudujte problém asociačních pravidel a jejich užití v DM.
3. Naprogramujte v libovolném programovacím jazyce vybraný algoritmus pro hledání asociací.
4. Použijte svoji implementaci pro daná vstupní data, vizualizujte výsledek pro studenty se zrakovým postižením.
5. Porovnejte komfort práce studenta se zrakovým postižením při práci v komerčním IBM SPSS Modeleru se stejným zadáním a prací ve vzniklé aplikaci.

Rozsah grafických prací: **dle potřeby dokumentace**

Rozsah pracovní zprávy: **30–40 stran**

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam odborné literatury:

[1] **RUD, Olivia Parr. Datamining, Vyd.1. Praha: Computer Press, 2006, XVII, 329 s. ISBN 80-722-6577-6.**

[2] **SKALSKÁ, Hana. Datamining a klasifikační modely, Vyd. 1. Hradec Králové: GAUDEAMUS, 2010, 154 s. ISBN 978-80-7435-088-7**

[3] **BERKA, Petr. Dobávání znalostí z databází. Praha: Academia, 2003, s.18. ISBN 80-200-1062-9**

[4] **IBM CORPORATION. IBM SPSS Modeler 14.2. Algorithms Guide 2011.**

[5] **MAYER-SCHONBERGER, Viktor; CUKIER, Kenneth. Big Data. Vyd. 1. Praha: Computer Press, 2014, 256 s. ISBN 978-80-251-4119-9**

Vedoucí bakalářské práce: **RNDr. Klára Císařová, Ph.D.**
Ústav mechatroniky a technické informatiky

Datum zadání bakalářské práce: **10. října 2014**

Termín odevzdání bakalářské práce: **15. května 2015**

prof. Ing. Václav Kopecký, CSc.
děkan

L.S.

doc. Ing. Milan Kolář, CSc.
vedoucí ústavu

V Liberci dne 10. října 2014

Prohlášení

Byla jsem seznámena s tím, že na mou bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

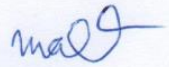
Beru na vědomí, že Technická univerzita v Liberci (TUL) nezasahuje do mých autorských práv užitím mé bakalářské práce pro vnitřní potřebu TUL.

Užiji-li bakalářskou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědoma povinnosti informovat o této skutečnosti TUL; v tomto případě má TUL právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Bakalářskou práci jsem vypracovala samostatně s použitím uvedené literatury a na základě konzultací s vedoucím mé bakalářské práce a konzultantem.

Současně čestně prohlašuji, že tištěná verze práce se shoduje s elektronickou verzí, vloženou do IS STAG.

Datum: 13.5.2015

Podpis: 

Poděkování

Ráda bych na tomto místě poděkovala všem, kteří se podíleli na tvorbě mé bakalářské práce. V prví řadě patří mé velké díky vedoucí mé práce RNDr. Kláře Císařové, Ph.D. za její ochotu a trpělivost, za cenné rady a připomínky, které mi pomohly při vypracování bakalářské práce. Dále bych ráda poděkovala studentům navazujícího studia IT a svým přátelům, kteří testovali vzhled a funkčnost aplikace vytvářené v rámci této práce a poskytli mi řadu námětů na její zdokonalení.

Abstrakt

Tato práce se zabývá asociačními metodami v data miningu a vizualizací dat pro vybranou data miningovou úlohu pro studenty se zrakovým handicapem.

Cílem práce bylo seznámit se s problematikou data miningu, detailněji pak s užitím asociačních metod v data miningových úlohách, a na základě získaných znalostí vytvořit aplikaci, která bude pracovat s rozsáhlým datovým souborem a těžit z něho informace. Aplikace zpracovává data obsažená v textovém souboru, analyzuje je a třídí, dle požadovaných kritérií vyhledává v souboru pouze určitá data a výsledky zprostředkovává uživateli vizuálně prostřednictvím přehledů a grafů s respektem k zrakovému postižení. Dále aplikace umí hlubší analýzou a použitím složitějších algoritmů najít v datech asociační pravidla a měla by být užitečnou pomůckou studentům při studiu asociačních metod.

Aplikace je vytvořena ve vývojovém prostředí Delphi 10 s užitím programovacího jazyka Pascal.

Klíčová slova: data mining, vizualizace, asociační pravidla

Abstract

This thesis deals with association rules in data mining and with data visualization for the selected data mining task for students with visual handicap.

The aim of this work was to become familiar with the problematics of data mining, thoroughly with using the association methods in data mining tasks, and on the basis to this knowledge create an application that will work with large data set and obtain certain information of it. The application process data contained in a text file, analyses and categorizes them, according to the required criteria it searches only certain data in the file and convey these information to the user visually through a variety of reports and graphs with respect to visual handicap. Furthermore, the application can retrieve association rules of the data set using deeper analysis and complex algorithms and it should be a useful tool for students studying association methods.

Application is programmed in Delphi 10 development environment using programming language Pascal.

Key words: data mining, visualization, association rules

Obsah

Seznam obrázků

Seznam tabulek

1 Pojem data mining

Data mining, neboli v češtině dolování či vytěžování znalostí z dat, je soubor matematických metod sloužících k získávání doposud neznámých, potenciálně užitečných a určitým způsobem zajímavých a významných informací z dat či k hledání souvislostí, vzorů a vztahů ukrytých v datech.

Takto získané informace a odhalené vztahy mezi daty se později dále využívají v mnohých sférách a oblastech: od těch, kde je to přirozené, jimiž jsou obchod, bankovnictví, medicína či oblast bezpečnosti, až po specializované oblasti genomiky nebo astrofyziky.

Úkolem data miningu je pomoci při rozhodování jakéhokoli typu.

2 Historie data miningu

Vznik data miningu souvisí se zavedením elektronického sběru dat, kdy začaly vznikat veliké datové soubory, které bylo nutné zpracovávat, aby je následně bylo možné lépe vyhodnocovat a čerpat z nich informace. Pro práci s obrovskými objemy dat se však klasické, již dříve známé a používané, statistické metody ukázaly jako ne příliš vhodné, a bylo nutné přijít s novými metodami, které dokáží nalézt i složité nelineární vztahy, a to navíc bez omezujících předpokladů.

První náznaky aktivit, které dnes označujeme jako data mining, se objevily v 60. letech 20. století, jednalo se například o využívání regresní analýzy s automatickým výběrem proměnných a prvních rozhodovacích stromů. Šlo však zpravidla o ojedinělé, většinou akademické záležitosti.

V 70. a 80. letech 20 století byly podmínky pro rozvoj data miningu více než příznivé. Rozvíjely se databázové aplikace i umělá inteligence, zvětšovala se paměť počítačů a zvyšovala se jejich rychlost, a data miningové postupy mohly být konečně opravdu reálně využívány v praxi. Nicméně ve společnosti stále spíše přetrvávaly nedůvěra a pochyby o důvěryhodnosti výsledků data miningu. Toto slovní spojení se označovalo jako „vyzobávání rozinek“ z dat a lidé byli přesvědčeni, že hledání korelací

ve velkých datových souborech s sebou nese příliš velké riziko, že nalezneme pouze nahodilé fluktuace bez možnosti zobecnění a následného praktického využití.

Obrat nastal v 90. letech 20. století, kdy již byly vybudovány metody, jak se výše zmíněnému nebezpečí falešných korelací úspěšně vyhýbat. V té době rostla poptávka mnohých komerčních společností po data miningových nástrojích. Jednalo se o takové organizace, které disponovaly velikými objemy dat, z nichž už nebyly dále schopny pomocí běžných tabelačních metod efektivně čerpat informace a získávat potřebné podklady pro rozhodování. Data mining se tak velice rychle rozšířil zejména v komerční sféře (analýza nákupního košíku, segmentace zákazníků, předpověď odchodu klientů ke konkurenci, řízení zaměstnanců apod.), ale i v jiných oblastech jako je vědecký výzkum – analýza genetické informace; bezpečnost – monitorování aktivit na internetu s cílem odhalit případné „škůdce“ nebo teroristy.

V roce 1991 definoval William J. Frawley data mining takto: „Data mining je netriviální získávání předtím neznámé a potenciálně užitečné informace ukryté v datech.“

Na začátku nového tisíciletí se pak data mining osamostatnil jako nové odvětví statistiky.

V současné době je data mining běžnou součástí podpory fungování organizací a v případě špičkových organizací se již bez něho neobejde žádná plošnější obchodní aktivita.

3 Data miningové úlohy

Data mining řeší, jak již bylo výše zmíněno, problémy z mnoha různých oborů od marketingu a bankovnictví, přes medicínu či oblast bezpečnosti až po specializované oblasti genomiky nebo astrofyziky.

Pro zlepšení efektivity řešení a zkvalitnění jeho výsledku je každé úloze, která má být řešena data miningovými nástroji, třeba nejprve přiřadit typ (skupinu), do níž spadá. Jednotná podoba rozdělení typů úloh přitom neexistuje, nejčastěji se uvádějí čtyři typy úloh: klasifikace, predikce, deskripce a hledání nuggetů.

Pro řešení úloh každého typu se využívají jiné data miningové metody, přičemž pro každý typ lze zpravidla využít více metod a získat tak několik odlišných výsledků k porovnání.

3.1 Typy data miningových úloh

Klasifikace

Jedná se o nejčastější typ data miningové úlohy, jehož podstatou je třídění objektů a jejich zařazování do určitých tříd na základě dříve nalezených znalostí. Klasifikační metody mají široké využití zejména v oblastech, kde se shromažďuje větší množství dat.

Podstatou klasifikačních úloh je výběr jednoho cílového atributu, zkoumání vlivu ostatních atributů na tento atribut a získávání znalostí, které bude možné použít pro hodnocení nových případů a jejich následnému zařazení do konkrétní skupiny.

Příkladem klasifikace může být:

Tabulka 1. Příklad klasifikace

Objekt:	**Třída**
Úvěr	ano/ne
E-mail	SPAM/ne SPAM
Pacient	zdravý/nemocný

Predikce

Predikce znamená předpověď, odhad vývoje nějakého ukazatele v čase pomocí netriviálních statistických technik. Na základě analýzy hodnot již známých z minulosti jsou odvozovány hodnoty, které je pravděpodobné očekávat v budoucnu.

Deskripce

Deskripce neboli popis je proces hledání dominantní struktury či vazby skryté v datech a charakterizující tato data jako celek. Na rozdíl od klasifikace či predikce, kde je kladen důraz především na přesnost výsledků, nikoli tolik na jejich srozumitelnost, u deskripce je tomu naopak a je zde upřednostňováno získání menšího množství méně přesných, ovšem srozumitelných informací pokrývajících celý problém.

Hledání nuggetů

Hledání nuggetů se blízce podobá deskripci, a proto bývá občas také pod deskripci zařazováno. Je to proces vyhledávání nových srozumitelných informací charakterizujících určená data, přičemž tyto informace nemusí pokrývat daný problém jako celek, důraz je ovšem kladen na jejich zajímavost a překvapivost.

3.2 Metody řešení data miningových úloh

Pro řešení data miningových úloh je možné využít různé metody, kdy jejich výběr závisí na typu dané úlohy. Tato volba přitom není jednoznačná, na každý typ úlohy lze aplikovat více metod a získat tak několik odlišných výsledků k dalšímu porovnání.

Klasifikace

Úkolem klasifikace je třídění objektů datového souboru a jejich zařazování do předem deklarovaných skupin dle vzájemné podobnosti. V praxi bývá klasifikace používána například k rozpoznávání bonitních či naopak problémových zákazníků a klientů.

Regrese

Regrese slouží obecně pro vysvětlení a předpověď složitých proměnných za pomoci dostupných informací z historických dat. Regresní úloha se liší od klasifikační především typem výsledku: zatímco výsledkem klasifikace je odhad dané kategorie (třídy), výsledkem regrese je spojitá číselná hodnota. Příkladem této metody může být odhad ceny domu vzhledem k lokalitě.

Predikce

Predikce se zaměřuje na předpovědi vývoje nějakého ukazatele v čase, kdy jsou na základě analýzy hodnot již známých z minulosti odvozovány hodnoty budoucí. Příkladem predikce může být odhad kurzu akcií či posouzení rizikovosti žadatele o úvěr.

Shlukování – segmentace

Segmentace je nejstarším nástrojem data miningu vůbec. Jejím cílem je najít objekty, shlukovat je a zařazovat do skupin na základě jejich vzájemné podobnosti, která ovšem není na první pohled patrná. Objekty v každém takovém shluku jsou si vzájemně podobné, zatímco jednotlivé shluky by se od sebe měly lišit maximálně.

Sumarizace

Sumarizace je postup hledání uceleného popisu souboru objektů a tato technika je používána zejména k interaktivní analýze dat a automatizovanému generování reportů.

Modelování závislostí

Jedná se o metodu spočívající v hledání modelu, který popisuje významné závislosti a vztahy mezi objekty. Existují dva typy závislostí: strukturální a kvantitativní. Strukturální úroveň modelu udává, které proměnné jsou navzájem místně závislé, kvantitativní úroveň modelu potom specifikuje sílu těchto závislostí na číselné škále.

Detekce změn a odchylek

Tato metoda se zaměřuje na objevování nejvýznamnějších změn z dříve naměřených hodnot anebo na odchylky od závazných normových hodnot. Užívají se například při odhalování podvodů.

Asociace

Asociace neboli analýza vztahů je proces objevování zajímavých vztahů mezi objekty v souboru a získávání pravidel typu *IF-THEN (jestliže-pak)* pomocí asociačních algoritmů. Typickým příkladem, kdy se využívá hledání asociací, je analýza nákupního košíku.

Objevování posloupností

Objevování posloupností blízce souvisí s analýzou asociací, přičemž položky, které spolu vzájemně souvisejí, jsou obohaceny o časový údaj a cílem této metody je stanovení vývoje jednotlivých událostí v čase (uspořádání událostí v čase, délku trvání událostí či délku časových intervalů mezi událostmi).

Dalšími typickými metodami, které spadají do oblasti data miningu jsou metody strojového učení a genetické algoritmy.

3.3 Algoritmy pro řešení data miningových úloh

3.3.1 Shluková analýza (cluster analysis)

Shluková nebo také clusterová analýza je statistická metoda sloužící ke klasifikaci objektů, a to k jejich třídění do skupin (clusterů) na základě jejich vzájemné podobnosti. Objekty v každé takové skupině by si vzájemně měly být co nejvíce podobné, zatímco naopak jednotlivé skupiny by se od sebe měly maximálně lišit.

Shluková analýza se provádí na množinách předmětů či jevů, které jsou popisovány *p*-ticí stavů předem stanovených *p* znaků, přičemž znaky mohou být buď kvalitativní (konečná množina popisujících termínů, např. barva očí) anebo kvantitativní (interval reálných nebo celých čísel, např. délka, teplota). Každému stavu jsou pak přiřazovány číselné hodnoty – hodnoty daných znaků. Objektem pro shlukovou analýzu je tedy *p* rozměrný vektor čísel.

Typy metod shlukové analýzy

Shlukovací metody se rozdělují podle způsobů algoritmizace na hierarchické a nehierarchické.

Hierarchické shlukování vytváří systém podmnožin „zdola nahoru“, Na začátku každý objekt tvoří samostatný shluk. Postupně se jednotlivé shluky spojují, až skončí všechny objekty v jednom shluku.[1] Najít dva nejpodobnější shluky ke spojení je

možné několika způsoby. Používá se metoda nejbližšího či nejvzdálenějšího souseda, metoda průměrné vzdálenosti nebo centroidní metoda.

Nehierarchické shlukování vytváří systém, kde jsou shluky disjunktivní množiny, tedy platí, že průnikem každých dvou neprázdných podmnožin systému není ani jedna z těchto podmnožin. Známých je několik algoritmů, ve kterých se typicky posuzuje kvalita shluků funkcionálem kvality. Sleduje se například vnitroshlukový rozptyl, podobnost objektů ve shluku, vzdálenost objektů od těžiště.

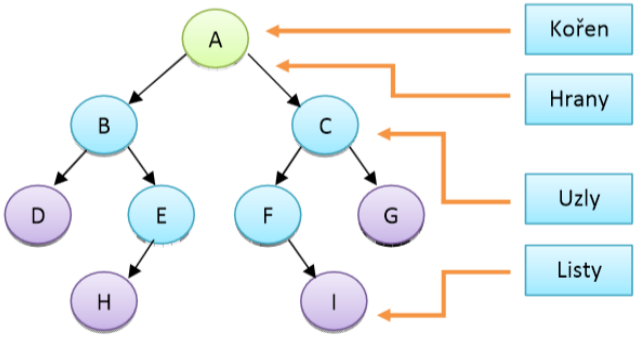
3.3.2 Rozhodovací stromy

Tato metoda je jednou z nejrozšířenějších a nejpoužívanějších data miningových technik. Rozhodovací stromy umožňují snadnou a přehlednou interpretaci dat a tedy i rychlé vyhodnocování získaných výsledků a identifikaci klíčových položek. Cílem rozhodovacích stromů je klasifikovat objekty, popsané různými atributy, a rozdělit je do tříd.

Podoba rozhodovacích stromů

Rozhodovací stromy jsou acyklické grafy skládající se z hran (spojnic uzlů) a uzlů, které představují rozhodování podle jedné vlastnosti posuzovaných objektů – takové, která dané objekty od sebe maximálně odliší. Uzly se dále dělí podle charakteru na kořen, který je vrcholem stromu a do něhož nevedou žádné hrany, vnitřní uzly, které mají své další potomky, a listy, což jsou konečné uzly nemající další potomky. Právě listy pak představují cílové klasifikační třídy.

Příklad podoby rozhodovacího stromu je uveden na následujícím obrázku:



Obrázek 1. Struktura rozhodovacího stromu.

Rozhodovací stromy se podle topologie dělí do dvou skupin:

- Binární stromy – z každého uzlu vystupují maximálně dvě větve
- Obecné (nebinární) stromy – z uzlů může vystupovat více než dvě větve

Další možné dělení rozhodovacích stromů je pak podle podoby interpretace výsledku úlohy na stromy klasifikační a regresní. Zatímco klasifikační stromy pouze zařazují data do tříd, regresní stromy odhadují hodnotu numerických atributů.

Tvorba rozhodovacích stromů

Při tvorbě rozhodovacích stromů se užívá metoda „rozděl a panuj“, kdy se nejprve trénovací data rozdělují do menších a menších podmnožin (uzlů) tak, aby v těchto podmnožinách převládaly objekty jedné třídy. Na začátku procesu tvoří data jednu množinu, která je postupně rozdělována, na konci zůstanou pouze podmnožiny tvořené objekty stejné třídy.

Při tvorbě rozhodovacího stromu může dojít k dvěma zásadním problémům. Prvním je, že vytvořený strom je příliš rozsáhlý a složitý a stává se proto nesrozumitelným, druhý problém je pak přílišné přizpůsobení trénovací množině dat, což má za následek snížení schopnosti generalizace a následné selhání systému. Stromy proto bývají zjednodušovány.

Redukce rozhodovacího stromu je možná dvojím způsobem: předčasným zastavením růstu stromu či prořezáváním stromu, kdy se z hotového stromu odstraní některé málo významné větve.

Algoritmy rozhodovacích stromů

Vybrat atribut, který bude strom dále dělit, lze několika způsoby, a existuje tedy i více algoritmů pro hledání tohoto atributu. Jsou jimi například ID3, C4.5, C5.0, CART, CHAID či QUEST. Každý z těchto algoritmů má své výhody a nevýhody a volba nejvhodnějšího z nich závisí na typu řešené úlohy.

3.3.3 Asociační metody

Asociační metody patří spolu s rozhodovacími stromy k nejčastěji používaným prostředkům pro objevování zajímavých vztahů mezi velkým množstvím datových položek. Tato technika umožňuje z velkého množství záznamů stanovit pravidla vhodná pro další rozhodování.

Problematice asociačních metod a hledání asociačních pravidel je věnována čtvrtá kapitola, jako základ pro tvůrčí část této práce.

3.3.4 Neuronové sítě

Neuronová síť je dalším algoritmem užívaným pro klasifikaci a predikci. Může být použita jako náhrada rozhodovacích stromů nebo asociačních pravidel v případech, kdy primárně nezáleží na srozumitelnosti výstupu. V této technice je, na rozdíl od rozhodovacích stromů, vhodnější pracovat se spojitými číselnými daty.

Principem neuronových sítí je napodobení chování lidského mozku zejména ve třech aspektech:

- Umět uložit znalosti
- Aplikovat uložené znalosti na řešení problémů
- Získávání nových znalostí

Neuronové sítě jsou inspirovány biologickým systémem, konkrétně stavbou neuronu, tedy nervové buňky člověka, jejíž strukturu a funkčnost se snaží simulovat pomocí počítačů.

Z hlediska průchodu informací neuronovou sítí rozlišujeme dva typy neuronových sítí:

- Dopředné sítě
- Rekurzivní sítě

V případě dopředných sítí se signál šíří jedním směrem od vstupu k výstupu, zatímco v případě rekurzivních sítí se může šířit i směrem opačným – to je dáno strukturou zapojení sítě (zapojení se zpětnými vazbami).

4 Asociační metody v data miningu

Asociační metody jsou jedním z nástrojů pro dobývání znalostí z databází. Tyto metody patří (spolu s rozhodovacími stromy) k nejčastěji používaným prostředkům pro objevování zajímavých vztahů mezi velkým množstvím datových položek. Jsou určeny k identifikaci silných pravidel zjištěných v databázích za použití různých měřítek zajímavosti.

Původně se asociační metody aplikovaly na transakční data a využívaly se k analýze nákupních košíků zákazníků obchodních řetězců. V současné době se proces hledání asociačních pravidel užívá v mnoha různých odvětvích: v oblasti marketingu a v komerční sféře (analýza nákupního košíku, analýza bankovních služeb, analýza služeb mobilních operátorů aj.), ale rovněž i ve vědeckém výzkumu, v sociologii, v hutnictví a dalších.

4.1 Asociační pravidla

Získávání asociačních pravidel z dat je jedním z významných oborů data miningu.

4.1.1 Podoba asociačních pravidel

Jedná se o pravidla se syntaxí: IF-THEN (v češtině JESTLIŽE-PAK), tedy slovně formulována následovně: „*Jestliže platí předpoklad A, pak platí závěr B“*.

Obecná podoba pravidla:

IF *podmínka* ***THEN*** *výsledek*

Konkrétnější podoba pravidla:

IF *položka_i* ***THEN*** *položka_j*

Příklady pravidel:

IF *pohlaví = žena* ***THEN*** *tv_žánr = romantika*

IF *pohlaví = muž* ***THEN*** *tv_žánr = sport*

IF *kolečkové brusle* ***THEN*** *helma* ***AND*** *chrániče*

4.1.2 Charakteristiky asociačních pravidel

U pravidel vytvořených z dat nás zajímá, kolik příkladů splňuje předpoklad (antecedent) a kolik závěr (sukcedent) pravidla, resp. kolik příkladů splňuje předpoklad i závěr současně, kolik příkladů splňuje předpoklad a nesplňuje závěr, kolik příkladů naopak nesplňuje předpoklad a splňuje závěr a kolik příkladů nesplňuje ani závěr a ani předpoklad.

Zajímá nás, jak pro pravidlo

$Ant \Rightarrow Suc$

vypadá příslušná kontingenční (čtyřpolní) tabulka pro *n* prvků:

Tabulka 2. Kontingenční tabulka pro n prvků

	Suc	*¬Suc*	∑
Ant	*a*	*b*	*r*
¬Ant	*c*	*d*	*s*
∑	*k*	*l*	*n*

a... počet příkladů pokrytý současně předpokladem i závěrem

$a = n(Ant \wedge Suc)$

b... počet příkladů pokrytý předpokladem a nepokrytý závěrem

$b = n(Ant \wedge \neg Suc)$

c... počet příkladů nepokrytý předpokladem a pokrytý závěrem

$c = n(\neg Ant \wedge Suc)$

d... počet příkladů nepokrytých ani předpokladem ani závěrem

$d = n(\neg Ant \wedge \neg Suc)$

$k = n(Suc) = a + c$

$l = n(\neg Suc) = b + d$

$r = n(Ant) = a + b$

$s = n(\neg Ant) = c + d$

$n = a + b + c + d$

Z těchto čísel můžeme počítat různé ***charakteristiky pravidel*** a kvantitativně hodnotit nalezené znalosti.

Základní charakteristiky asociačních pravidel – Rakesh Agrawal:

- **Podpora** (support) je počet objektů splňující předpoklad i závěr.

  Absolutní podpora: a

  Relativní podpora: $P(Ant \wedge Suc) = \frac{a}{a+b+c+d}$

- **Spolehlivost** (platnost = validity, konsistence = consistency, správnost = accuracy) je podmíněná pravděpodobnost závěru, pokud platí předpoklad.

  $$P(Suc|Ant) = \frac{a}{a+b}$$

- Počet objektů, které splňují předpoklad:

  Absolutní: $a + b$

  Relativní: $\mathrm{P(Ant)} = \frac{a+b}{a+b+c+d}$

- Počet objektů, které splňují závěr

  Absolutní: $a + c$

  Relativní: $\mathrm{P(Suc)} = \frac{a+c}{a+b+c+d}$

- **Pokrytí** (coverage) = pravděpodobnost předpokladu, pokud platí závěr.

  $$P(Suc|Ant) = \frac{a}{a+c}$$

- **Kvalita** = vážený součet spolehlivosti a pokrytí

  $$Kvalita = w_1 \frac{a}{a+b} + w_2 \frac{a}{a+c}$$

  kde w_1 a w_2 se obvykle volí tak, aby $w_1 + w_2 = 1$

Rozšíření charakteristik asociačních pravidel – Kodratof

- **Kauzální podpora** (causal support)

  $$P(Ant \wedge Suc) + P(\neg Ant \wedge \neg Suc) = \frac{a+d}{a+b+c+d}$$

- **Kauzální spolehlivost** (causal confidence)

$$\frac{1}{2}P(Suc|Ant)+\frac{1}{2}P(\neg Ant|\neg Suc)=\frac{1}{2}\frac{a}{a+b}+\frac{1}{2}\frac{d}{b+d}$$

- **Deskriptivní potvrzení** (descriptive confirmation)

$$P(Ant\wedge Suc)-P(Ant\wedge\neg Suc)=\frac{a-b}{a+b+c+d}$$

- **Kauzální potvrzení** (causal confirmation)

$$P(Ant\wedge Suc)+P(\neg Ant\wedge\neg Suc)-2P(Ant\wedge\neg Suc)=\frac{a+d-2b}{a+b+c+d}$$

- **Ujištění** (conviction)

$$\frac{\mathrm{P(Ant)P(\neg Suc)}}{\mathrm{P(Ant\wedge\neg Suc)}}=\frac{\mathrm{(a+b)(b+d)}}{\mathrm{d(a+b+c+d)}}$$

- **Zajímavost** (interestingness)

$$\frac{\mathrm{P(Ant\wedge Suc)}}{\mathrm{P(Ant)P(Suc)}}=\frac{\mathrm{a(a+b+c+d)}}{\mathrm{(a+b)(a+c)}}$$

- **Závislost** (dependency)

$$P(Suc|Ant)-P(Suc)=\frac{a}{a+b}+\frac{a+c}{a+b+c+d}$$

Dělení asociačních pravidel

Implikace (tedy asociační pravidla) lze dělit na základě ***platnosti*** a ***pokrytí*** do těchto skupin***:***

- **Konzistentní pravidla** jsou pravidla s platností rovnou 1, kdy levá strana implikace je postačující podmínkou pro splnění pravé strany.
- **Úplná pravidla** jsou pravidla s pokrytím rovným 1, kdy levá strana implikace je nutnou podmínkou pro splnění pravé strany.
- **Deterministická pravidla** jsou pravidla s platností i pokrytím rovným 1, kde levá strana implikace je nutnou a postačující podmínkou pro splnění pravé strany.

4.1.3 Hledání asociačních pravidel

Cílem tohoto procesu je nalézt tzv. „silná pravidla“, tedy pravidla, která mají vysokou (předem určenou) hodnotu podpory a spolehlivosti.

Množina silných pravidel (strong association rules) je definována takto:

$$SAR = \{ar | c(ar) \geq minconf \wedge s(ar) \geq minsup\}$$

ar ..... asociační pravidlo tvaru $A \Rightarrow B$, kde A, B jsou konjunkce predikátů tvaru $a_1 \wedge a_2 \ldots \wedge a_n$
c(ar) ..... spolehlivost pravidla
s(ar) ..... podpora pravidla
minconf ..... minimální spolehlivost
minsup..... minimální podpora

Proces hledání asociačních pravidel probíhá ve dvou krocích:

1. Generování frekventovaných vzorů (množin)
   - hledání kandidátů, které mají vyšší podporu, než je zadaná minimální podpora
   - nalezení tzv. „silných množin“
   - pro každé asociační pravidlo $X \Rightarrow Y$ musí platit, že $X \cup Y$ je frekventovaná množina položek
   - platí, že podmnožina frekventované množiny je rovněž frekventovanou množinou
   - platí, že pro m položek existuje 2^{m-1} kandidátů
2. Generování asociačních pravidel
   - vygenerování asociačních pravidel s využitím silných množin nalezených v předchozím kroku
   - odstranění pravidel, jejichž spolehlivost nedosahuje předem určené minimální hodnoty
   - je-li L frekventovaná množina a platí-li, že $|L| = k$, pak existuje 2^k-2 kandidátních asociačních pravidel (ignoruje se $L \rightarrow \emptyset$ a $\emptyset \rightarrow L$)

Nalezená asociační pravidla se dále testují aplikací na konkrétní data – zjišťuje se, zda pravidlo splňuje požadavky na hodnoty numerických charakteristik.

Požadavky na asociační pravidla

Asociační pravidla by měla být:

- pochopitelná – je-li nalezen nějaký vztah, lze ho snadno ověřit
- použitelná – obsahují užitečné informace vedoucí k dalším intervencím

Asociační pravidla by neměla být:

- triviální – pravidla, jejichž výsledky jsou již známé
- nevysvětlitelná – neexistují k nim vysvětlení, nedávají žádné užitečné informace

4.2 Algoritmy pro hledání asociačních pravidel

4.2.1 Generování kombinací

Základem všech algoritmů pro hledání asociačních pravidel je generování kombinací (konjunkcí) hodnot atributů.

Existuje několik metod, jak tyto kombinace generovat:

- **do šířky**
  - jde o generování kombinací podle délek
  - nejprve se vygenerují všechny kombinace délky jedna, pak všechny kombinace délky dvě atd.

- **do hloubky**
  - vyjde se od první kombinace délky jedna, která se dále prodlužuje (vždy o první kategorii dalšího atributu), dokud to lze; nelze-li kombinaci prodloužit, změní se kategorie posledního atributu, pokud nelze ani to (kategorie atributu jsou vyčerpány), kombinace se zkrátí a současně se změní poslední kategorie

- **heuristická** – generování podle četnosti
  - vytváří kombinace v pořadí podle jejich výskytu v datech
  - při tomto způsobu generování se kombinace s nejvyšší četností objevují na začátku seznamu, kombinace s nulovou četností naopak na konci seznamu

Počet kombinací

Počet generovaných kombinací (konjunkcí) je exponenciálně závislý na počtu atributů.

Označíme-li $K_{A1}, K_{A2}, \ldots K_{Am}$ počet kategorií atributů $A_1, A_2 \ldots A_m$, kde m je počet atributů, z nichž vytváříme kombinace. Pak:

- počet kombinací délky jedna

$$\sum_{i=1}^{m} KAi$$

- počet kombinací délky dvě

$$\sum_{i,j=1,i\neq j}^{m} KAi.KAj$$

- počet kombinací délky tři

$$\sum_{i,j,k=1,i\neq j\neq k}^{m} KAi.KAj.KAk$$

- počet všech možných kombinací

$$\prod_{i=1}^{m}(1+KAi)-1$$

4.2.2 Struktura dat

Každý algoritmus pro hledání asociačních pravidel je vždy aplikován na konkrétní data. Tato data lze z pohledu data miningu dělit podle jejich struktury na dva typy: data ve formátu transakčním či tabulárním.

Transakční data

Transakční data mají oddělené záznamy pro každou položku transakce. V případě, že transakce obsahovala více položek, je každá uložena zvlášť vždy pod stejným *ID* zákazníka.

Příklad transakčních dat je uveden v následující tabulce:

Tabulka 3. Transakční data

ID zákazníka	*Nákup*
1	džem
2	mléko
3	džem
3	chléb
4	džem
4	chléb
4	mléko

Tabulární data

Pro všechny položky nabídky je pro každou transakci specifikována její přítomnost či absence v této transakci pomocí pravdivostních hodnot. Každá transakce má přitom v tabulce vlastní záznam. Tedy tabulární data vznikají převodem kategoriální proměnné na tzv. indikátorové proměnné.

Příklad tabulárních dat je uveden v následující tabulce:

Tabulka 4. Tabulární data

ID zákazníka	*Džem*	*Chléb*	*Mléko*
1	T	F	F
2	F	F	T
3	T	T	F
4	T	T	T

4.2.3 Algoritmus apriori

Jedná se o nejznámější algoritmus pro hledání asociačních pravidel. V souvislosti s analýzou nákupního košíku ho navrhl R. Agrawal.

Jádrem algoritmu je hledání často se opakujících množin položek (frequent itemsets), jde o kombinace kategorií dosahující předem zadané minimální četnosti. Při hledání kombinací délky *k* s vysokou četností se přitom využívá toho, že již známe kombinace délky *k-1* a jejich četnosti. Kombinace délky *k* se vytvářejí spojováním

kombinací délky *k-1* (generování kombinací do šířky), které dosahují požadované minimální četnosti.

Poté, co jsou nalezeny všechny kombinace s požadovanou četností, jsou vytvářena vlastní asociační pravidla na základě předem určeného kritéria minimální spolehlivosti.

Algoritmizace v krocích:

Krok1: Generování celé kombinace do šířky:

Algoritmus apriori

1. do L_1 přiřaď všechny hodnoty atributů, které dosahují alespoň požadované četnosti
2. polož $k=2$
3. dokud $L_{k-1} = \emptyset$
   3.1. pomocí funkce *apriori-gen* vygeneruj na základě L_{k-1} množinu kandidátů C_k
   3.2. do L_k zařaď ty kombinace z C_k, které dosáhly alespoň požadovanou četnost
   3.3. zvětši počítadlo k

Funkce apriori-gen (L_{k-1})

1. pro všechny dvojce kombinací p, q z L_{k-1}
   - pokud p a q se shodují v prvních $k-2$ položkách přidej do C_k sjednocení $p \cup q$
2. pro každou kombinaci c z C_k
   - pokud některá z jejich podkombinací délky $k-1$ není obsažena v L_{k-1} odstraň c z C_k

Krok2: Vytváření asociačních pravidel

Každá kombinace *C* se rozdělí na všechny možné dvojice podkombinací *Ant* a *Suc* takové, že $Suc = C - Ant$. Hledají se pravidla $Ant \Rightarrow Suc$ tak, že se postupně přesouvají kategorie z *Ant* do *Suc*, je-li *Ant'* podkombinací *Ant*, potom:

$$conf(Ant' \Rightarrow C - Ant') \leq conf(Ant \Rightarrow C - jAnt)$$

Algoritmus je řízen parametry *minsup* (minimální podpora) a *minconf* (minimální spolehlivost).

Algoritmus apriori není nikterak složitým algoritmem, je poměrně snadno implementovatelný a značně urychluje proces generování frekventovaných množin a následné vyhledávání asociačních pravidel, i když rozhodně nepatří k nejrychlejším vzhledem k nutnosti vícenásobného procházení datového souboru. K jeho zdokonalení proto vznikla celá řada dalších optimalizací.

4.2.4 Algoritmus CARMA

CARMA (Continuous Association Rule Mining Algorithm) je, podobně jako algoritmus apriori, dalším algoritmem pro objevování nových asociačních pravidel v datech.

Princip algoritmu

Algoritmus CARMA používá efektivní dvouprůchodovou metodu pro nalezení sekvencí v datech, k vyhledání všech kombinací položek mu tedy stačí dva průchody celého datového souboru. Proces navíc probíhá takzvaně on-line – udržuje nepřetržitou zpětnou vazbu s uživatelem, kterému umožňuje měnit požadavky na asociační pravidla, tedy požadovanou minimální podporu a minimální spolehlivost.

Algoritmus prochází data a průběžně generuje asociační pravidla, jejichž podpora a spolehlivost odpovídá zadaným parametrům.

V první fázi skenování souboru má uživatel možnost zmíněné prahy požadované minimální podpory a minimální spolehlivosti kdykoli změnit – snížit či zvýšit. Ve druhé fázi skenování už tuto možnost nemá a algoritmus nalezne všechna asociační pravidla, jejichž parametry odpovídají naposledy zadaným prahovým hodnotám. Druhý průchod datového souboru by v některých případech ani nebyl nutný, bylo by to tehdy, kdyby uživatel buď vůbec neměnil požadovanou minimální podporu a minimální spolehlivost, anebo kdyby jejich hodnoty neustále pouze zvyšoval.

Algoritmus je velice rychlý v porovnání například z výše zmíněným algoritmem apriori, který pro vygenerování všech kombinací položek a následnému nalezení asociačních pravidel potřebuje více průchodů datovým souborem, a dokáže objevovat i poměrně složitá asociační pravidla.

CARMA algoritmus urychluje proces detekování asociací a sbírá detailní informace z dat, lze pomocí něho generovat jak jednoduchá, tak i složitější asociační pravidla – pravidla s více závěry platícími současně. Nalezená asociační pravidla mohou být dále použita pro širokou sféru aplikací.

5 Data miningový projekt

První fází každého projektu je samotné pochopení projektu – jeho smyslu a cíle (na co bude používán, k čemu je určený), a provedení návrhu projektu – vytvoření plánu pro řešení daného problému.

Je nutné detailně se seznámit s daty, se kterými pracujeme, a porozumět jim. Bez dostatečné znalosti a pochopení dat by totiž mohlo dojít ke znehodnocení zdrojů dat při jejich následném zpracování a ovlivnění kvality výsledného řešení.

Zpracování dat je pak dalším krokem při tvorbě projektu. Data je třeba tzv. „očistit“, prvotně je předpřipravit a upravit do potřebné podoby, najít v nich nesmyslné hodnoty a ty vyřadit, připravit si nové proměnné, které pro nás budou později užitečné při následném modelování.

Dále přichází fáze modelování. To je, jak je z názvu patrné, proces vytváření nejrůznějších modelů projektu, testování vhodných metod pro řešení definovaného problému a nastavení jejich parametrů. Z vytvořených modelů posléze vybíráme ten nejlepší.

Poslední fází je nasazení modelu, tedy že výsledný vybraný model použijeme v praxi a aplikujeme ho na novou sadu dat.

Samozřejmě je stále nutné sledovat, zda je model aktuální (porovnáváním výsledků modelu a rozložením současných a historických vstupních dat), v případě velkých odchylek musíme přikročit k jeho aktualizaci na základě nově získaných poznatků. Zastaralé modely pozbývají kvality a ztrácí svou funkci.

6 Kdy je vhodné využít data mining?

Data mining je silným nástrojem, který nám pomáhá efektivně zpracovávat data. Jeho využívání se stává v moderní době stále větší nutností, přesto však nadále existují situace, kdy se bez metod data miningu lze bez problému obejít a kdy je jeho využití v podstatě zbytečné. Jedná se o situace, kdy:

- pracujeme s malým objemem dat, která jsme schopni dostatečně vyhodnotit pomocí klasických vizualizačních nástrojů – grafů, tabulek apod.
- počet atributu je malý – vyhodnocujeme-li například základní trendy objemu prodeje určitého výrobku podle času a prodejního místa (tedy na základě dvou atributů), dokážeme si vystačit s jednoduššími nástroji, určenými pro toto vyhodnocení

Použití data miningu je naopak velice efektivní, jestliže:

- výchozí objem dat je velký a jsou tím pádem nepřehledná
- máme mnoho atributů, na základě nichž bychom měli vyhodnocovat

7 Využití data miningu v praxi

Data mining se používá v oblastech, kde se shromažďuje velké množství dat. Typické příklady takovýchto datových souborů nalezneme zejména v těchto oblastech:

- Bankovnictví: informace o klientech, pohyby na účtech
- Obchod: obchodní řetězce i internetové obchody sledují zvyklosti svých zákazníků, co nakupují
- Telekomunikace: údaje o volání
- Genetika: datové informace o expresi genů
- Průmysl: záznamy průběhu provozních parametrů
- Pohyb a činnost uživatelů na internetu

Metodami data miningu se dají zpracovávat různorodé informace bez ohledu na obor a původ a možnosti jeho využití jsou různé.

7.1 Data mining a jeho využití v marketingu a komerční sféře

V této oblasti představuje data mining opravdu silný nástroj a hraje tu velmi významnou roli. Výsledky analýz data miningu pomáhají velkým společnostem: obchodním řetězcům, bankám a spořitelnám, mobilním operátorům aj. lépe přizpůsobovat nabídku svých produktů poptávce zákazníků, provádět cílené nabídky

zboží a zvyšovat tak svůj finanční zisk. Slouží ale také k tomu, aby společnosti dokázaly včas odhalit možný úmysl klienta odejít ke konkurenci a v ideálním případě jeho odchodu zabránit.

7.1.1 Analýza nákupního košíku

Slouží k zjišťování, které druhy zboží nakupovali zákazníci společně. Podle toho je pak možné provádět cílené reklamní nabídky. Prodejce například ví, že 60% zákazníků, kteří si kupovali pečivo, kupovalo zároveň i časopisy, zatímco pečivo v kombinaci s mraženými výrobky si koupilo pouze 28% zákazníků. Zdá se být tedy výhodnější, pokud si zákazník kupuje pečivo, nabídnout mu zároveň s ním časopisy, nežli mražené výrobky.

Složitějším případem je zjišťování, jaká je pravděpodobnost, že má zákazník, který si kupuje současně pečivo, mražené výrobky a časopisy, v košíku také alkohol. Víme, že ze všech zkoumaných košíků, jich pečivo, mražené výrobky i časopisy obsahovalo 13%. Alkohol obsahovalo 39% ze všech zkoumaných košíků, z košíků obsahujících také pečivo, mražené výrobky a pečivo pak alkohol obsahovalo 75%, což bylo cca 10% z celkového množství. Pravděpodobnost, že zákazníci kupují alkohol bez ohledu na ostatní zboží je 1,9 (75%/39%) krát menší než pravděpodobnost, že si koupí alkohol, pokud si zároveň koupí i pečivo, mražené výrobky a časopisy. Chytrý obchodník by měl tedy zákazníkovi, který splňuje předpoklady a nakupuje pečivo spolu s mraženými výrobky a časopisy, nabízet také alkohol.

Nákupní košíky zákazníků zkoumají rovněž internetoví prodejci. Ti poté jejich výsledkům přizpůsobují obsah a vzhled svých stránek. Například pokud si objednávám knihu v internetovém obchodě, zpravidla najdu na stránkách i informace o tom, o co dalšího měli zájem zákazníci, kteří si kupovali stejnou knihu jako já. Šance prodejce, že u mě s takto sestavenou doporučenou nabídkou uspěje, je mnohem vyšší, než kdyby další tituly do ní vybíral náhodně.

7.1.2 Segmentace zákazníků

Zákazníci jsou rozděleni do homogenních skupin podle jejich nákupního chování nebo podle demografických charakteristik. Tato segmentace slouží k lepšímu cílení

marketingu na konkrétní skupiny s určitým chováním a požadavky. Lze například určit, u jakých zákazníků nejpravděpodobněji uspějeme s určitou nabídkou zboží a následně jim poslat leták s ní, zatímco jiným nabídneme něco jiného, vhodnějšího. Tím snižujeme výdejní náklady spojené s propagací, v našem konkrétním případě náklady na tištění letáků.

7.1.3 Shluková analýza

Umožňuje určit skupiny zákazníků, kteří jsou vysoce ziskoví a zaměřit se na získávání zákazníků s podobným profilem. Z takových zákazníků pak bude mít společnost pravděpodobněji větší finanční užitek.

7.1.4 Predikce

Jedná se o „předpovídání“ budoucích událostí a trendů na základě historických dat.

Společnosti zkoumají chování svých klientů v minulosti a na základě takto zjištěných informací odhadují chování klientů s podobnými zvyklostmi do budoucna. Je možné predikovat možný budoucí zájem zákazníků o produkty a těmto je posléze nabízet, ale také odhalit například možný záměr klientů odejít ke konkurenci.

Metody predikce hojně využívají finanční společnosti (banky, spořitelny aj.) či mobilní operátoři. V dnešní době existuje jen malé množství lidí, kteří by neměli bankovní účet nebo nevlastnili mobilní telefon, a pro společnosti, nabízející tyto služby, není prakticky možné získat nové klienty jinak než na úkor konkurence. Je tedy nutné oslovovat je s nabídkami, které pro ně budou lákavé a na něž pravděpodobně zareagují. Stejně tak důležité, protože konkurence je veliká a všudypřítomná, je pak pro firmy udržet si stávající zákazníky.

Mobilní operátoři si vedou záznamy o tom, jak klienti využívají jejich služeb: jaké mají tarify, eventuálně jak často dobíjejí svůj kredit, kolik času protelefonují, jaké množství textových zpráv posílají, v jakou denní dobu volají apod. Tato data shromažďují a zkoumají a následně z nich vyvozují závěry. Operátor například usiluje o to, aby některou z jeho služeb využívalo více zákazníků. Jak toho docílí? Nejprve zaměří na zákazníky, kteří už službu využívají, a určí znaky, které skupinu těchto osob

nejlépe charakterizují. Poté vytipuje zákazníky se stejnými znaky, nicméně zmiňovanou službu dosud nevyužívající, a osloví je s nabídkou oné služby, kterou mohou získat například za zvýhodněných podmínek.

Operátoři uchovávají rovněž data o bývalých klientech. Údaje o nich porovnávají s údaji o svých stávajících zákaznících a snaží se predikovat na základě podobností a společných znaků jejich případný odchod ke konkurenci. Skupina takto vytipovaných klientů pak bývá oslovována a jsou jim nabízeny různé výhody apod.

Obdobně postupují rovněž bankovní a jiné finanční společnosti, které shromažďují například informace o účtech klientů, o jejich aktivitě, platbách, hotovostních převodech aj. a z těchto dat vyvozují různé zajímavé poznatky.

7.1.5 Risk management

Vyjadřuje pravděpodobnost výskytu sledované události, například pravděpodobnost nesplacení půjčky u konkrétního klienta.

Společnosti, poskytující půjčky, si své klienty před uzavřením smlouvy vždy pečlivě prověří, zaměřují se na jejich příjmy (tj. zda má klient dostatek finančních prostředků, aby byl schopen splácet všechny své stávající i budoucí závazky), záznamy v jejich bankovních i nebankovních registrech, věk klienta v době žádosti o úvěr i v době jeho splacení, vzdělání, počet dětí apod. Na základě získaných údajů vytipuje společnost rizikové klienty, u nichž je například vyšší pravděpodobnost, že nesplatí půjčku, a tuto půjčku jim buď odmítne poskytnout anebo ji poskytne, ovšem za méně výhodných podmínek než klientům nerizikovým (se zvýšenou sazbou úroků, s podmínkou dalšího ručitele apod.).

7.1.6 Fraud detection

Jde o odhalování podvodů, typicky pojišťovnických či úvěrových. Pojišťovny a úvěrové společnosti dokáží díky metodám data miningu rychleji a přesněji odhalovat finanční podvody, opět na základě z minulosti známých vzorců chování jejich klientů.

7.2 Data mining a jeho využití ve vědeckém výzkumu

Data mining se kromě komerční oblasti využívá také ve vědeckém výzkumu. Z DNA člověka se pomocí něj dá vyčíst, jak vysoké je riziko, že dotyčný bude trpět určitou nemocí, jestli má pro ni dědičné předpoklady.

Věda zkoumá tkáň zdravého člověka a porovnává ji s tkání jedince postiženého určitou nemocí. Obě tkáně mají stejné geny, což jsou neměnné informace, kódy, s jehož pomocí se v buňce vytvářejí nové proteiny, které se podílejí na stavbě tkání v lidském těle. Samotný gen je neměnný a liší se pouze množství proteinu, které se podle něho připraví, proces, který toto množství ovlivňuje, se nazývá genová exprese. A právě genová exprese se u tkání zdravých a nemocných lidí různí. Ve tkáni s poruchou, přestože byla vytvořena podle téhož genu, se vytvořilo jiné množství proteinu než ve tkáni zdravé.

Užitečné tedy je nalézt u nemocného jedince geny, které se podílejí na jeho poruše, a činnost těchto genů napravit.

Vzhledem k tomu, že v lidské DNA je známo zhruba 30 000 genů, změny genové exprese se hledají u každého z nich a pro věrohodnost konečných závěrů nelze spoléhat jen na jediný vzorek zdravého a jediný vzorek nemocného člověka a je naopak zapotřebí porovnávat vzorků hodně, dostáváme ve výsledku obrovské množství dat, která je nutné zpracovávat. A právě k tomu je vhodný data mining.

Mezi geny navíc existují vazby, jejich činnosti na sobě mohou vzájemně záviset (jeden gen řídí funkci jiného genu, ten pak funkci dalšího…) a tyto skryté souvislosti mezi geny se daří efektivně odhalovat právě díky data miningu. Data miningový program vezme určitý gen, nejprve zjistí, jestli byl aktivní při tvorbě zdravé nebo nemocné tkáně, a následně zkontroluje, zda s jeho aktivitou souvisí činnost i jiného genu z celkových třiceti tisíc. Celý proces zkoumání a posuzování genů byť jen u jediného člověka představuje asi milion matematických operací a bez data miningových nástrojů by byl prakticky neuskutečnitelný.

Data mining je využíván i při výrobě nových léků – pomocí něho jsou ze stovek léčebných látek vybírány ty nejvhodnější. Bere se přitom v úvahu nejen samotná účinnost látek, ale i jejich působení v kombinaci s látkami dalšími, a posuzuje se výsledný celkový dopad konečného léku na organismus člověka.

8 Možná nebezpečí a úskalí data miningu

Komerční data mining představuje masivní a inteligentní zpracovávání osobních údajů a mezi lidmi vznikají obavy ze zneužití těchto informací. Jejich únik, ať už záměrný nebo neúmyslný, může vézt k nejrůznějším problémům – od banálního spamu, který vám zkazí náladu, až po závažné případy jako je vydírání.

Za větší potenciální nebezpečí lze považovat technologie, k jejichž vzniku data mining přispívá v akademické sféře. Dekódování genomu, které je samozřejmě ve vědě velkým krokem vpřed, může například představovat riziko v případě, že bude použito k selekcím osob, nehumánním, ovšem postaveným na vědeckém základě. Pokročilé metody identifikace osob zase mohou být zneužity ke špehování občanů.

9 Softwarové nástroje pro data mining

V současné době existuje celá řada softwarových nástrojů, které slouží k řešení data miningových úloh, a vzhledem k faktu, že je již data mining běžnou součástí podpory fungování organizací a je užíván v mnoha oblastech, vyvíjejí se stále nové a nové programy zefektivňující práci s daty.

9.1 Rozdělení data miningových nástrojů

Systémy lze rozdělit do několika skupin podle jejich společných rysů. Těchto rozdělení je přitom více.

Data miningové nástroje lze klasifikovat podle jejich zaměření – k čemu se software bude primárně využívat, a to na systémy, které byly přímo vyvinuty pro data mining a na systémy umožňující data mining až v druhé řadě a jejich primární využití je jiné (např. matematické nebo statistické).

Další a pravděpodobně obvyklejší způsob, jak data miningové nástroje dělit, je podle toho, zda jsou či nejsou zdarma k dispozici veřejnosti. Existují jak systémy volně šiřitelné (open source systémy), které si může uživatel bezplatně opatřit stažením z internetu a nabízejí zpravidla základní nabídku data miningových metod, tak systémy

komerční – placené, které disponují širokou škálou funkcí a nabízejí velké množství data miningových metod.

Mezi volně šiřitelné data miningové nástroje patří:

- WEKA
- Rapid-I RapidMiner
- Orange

Mezi komerční data miningové nástroje patří:

- IBM SPSS Modeler
- SAS Enterprise Miner
- StatSoft Statistica Data Miner

9.2 IBM SPSS Modeler

Jedná se o jeden z nejrozšířenějších a nejužívanějších komerčních softwarových nástrojů pro data mining současné doby.

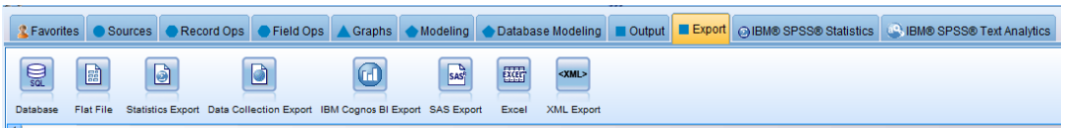
Původně byl vyvinut v polovině 90. let firmou *Integral Solutions* pod názvem *Clementine,* po sloučení firem *Integral Solutions* a *SPSS* probíhal další vývoj systému pod hlavičkou *SPSS* a poté, co tuto firmu koupila společnost *IBM*, dostal software *Clementine* název *IBM SPSS Modeler*, pod nímž ho známe v současnosti.

IBM SPSS Modeler je rozsáhlá platforma poskytující celou řadu pokročilých algoritmů a technik řešení data miningových úloh, umožňuje použití mnoha data miningových metod modelování a nabízí širokou škálu různých vizualizačních prostředků.

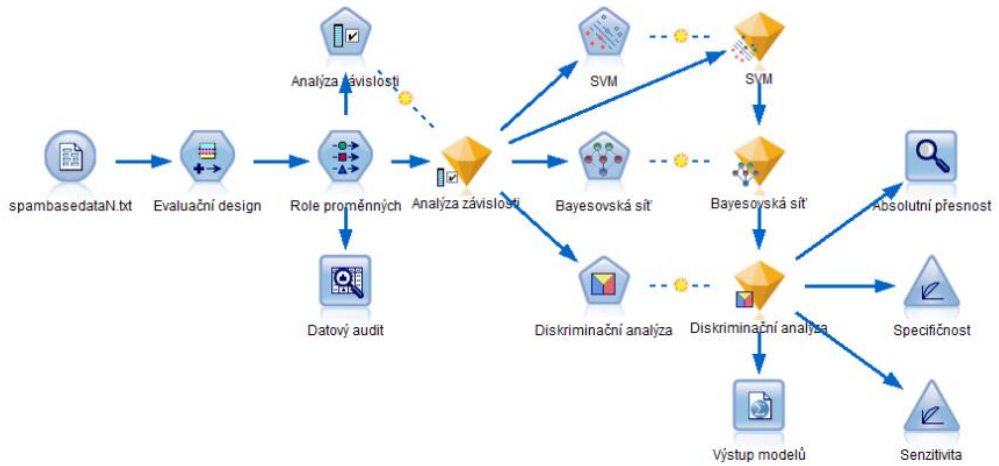
Systém vychází z metodiky CRISP-DM, která umožňuje řešit rozsáhlé data miningové úlohy rychleji, efektivněji a méně nákladně prostřednictvím osvědčených postupů a pomáhá vyhnout se potenciálním chybám.

Ovládání Modeleru je velice propracované, jedná se o tzv. vizuální programování, kdy si uživatel vybírá z nástrojů v paletách, které odpovídají jednotlivým krokům procesu dobývání znalostí z dat – předzpracování, modelování, vizualizaci a interpretaci. Z vybraných komponent je na pracovní ploše poskládán model řešení úlohy (stream), který je následně testován na konkrétních datech.

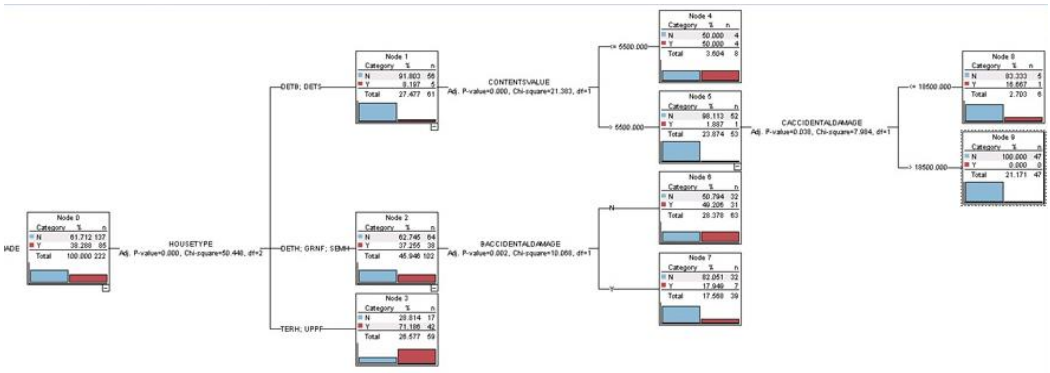
Na následujících obrázcích je vidět vizuální podobu prostředí IBM SPSS Modeleru:



Obrázek 2. IBM SPSS Modeler - paleta nástrojů



Obrázek 3. IBM SPSS Modeler – vytvořený stream



Obrázek 4. IBM SPSS Modeler - úloha vyřešená pomocí rozhodovacího stromu

Systém IBM SPSS Modeler nabízí řadu metod pro klasifikační i deskriptivní úlohy, mnoho předprogramovaných nástrojů a algoritmů k řešení data miningových úloh a nachází proto uplatnění v mnoha oblastech: v marketingu, ve vědě a výzkumu, v personalistice, užívá se pro zpracování laboratorních měření a pro sumarizace dat z velkých i menších databází různého typu.

10 Popis a cíle vlastní práce

V bakalářské práci se zabývám vizualizací dat a jejich analýzou s využitím asociačních metod a výkladem algoritmu apriori včetně umožnění počítačového experimentu pro jeho lepší pochopení. Veškeré výstupy programu jsou vhodné nejen pro běžné uživatele, ale také pro studenty se zrakovým handicapem

Úkolem je analyzovat nákupní košíky zákazníků obchodního řetězce na základě dat získaných propojením transakčních dat, která vznikla při průchodu nakupujících pokladnou, s osobními údaji o nich.

K dispozici je soubor „školních“ dat, z něhož je možné vyčíst nejrůznější informace o zákaznících: jejich identifikační číslo, pohlaví, věk (zde se nejedná o přesné určení věku, ale zákazník je zařazen do určité věkové skupiny, např. zákazníkovi s identifikačním číslem 12 je mezi 18 a 30 lety), rodinný stav, zda mají či nemají děti, jestli pracují nebo nepracují. Dále se ze souboru dozvím, co který zákazník v obchodě kupoval. Opět, podobně jako v případě věku nakupujících, je i zboží již rozděleno a zařazeno do produktových skupin, tj. alkohol, pečivo, mražené výrobky, maso, mléčné výrobky, časopisy, občerstvení, konzervované potraviny, toaletní potřeby a zelenina. V případě zákazníka s identifikačním číslem 12 tedy nezjistím, jestli měl v nákupním košíku rohlíky nebo bochník chleba, vím pouze, že kupoval pečivo.

Pro lepší představu, jak vypadá tento datový soubor, vkládám několik řádků z něho vybraných:

8, "F","F","F","F","F","T","F","F","F","F", "Female","18 to 30","Widowed","No","No"
9, "F","F","F","F","T","T","T","T","F","F", "Female","18 to 30","Single","No","No"
10, "F","F","F","F","F","T","F","F","F","F", "Female","18 to 30","Single","No","No"
11, "F","F","F","F","F","T","F","F","F","F", "Female","18 to 30","Single","No","No"

V prvním sloupci je uvedeno identifikační číslo zákazníka, v druhém až jedenáctém informace, zda bylo či nebylo zákazníkem zakoupeno zboží z určité produktové skupiny (alkohol – zelenina) a dvanáctý až sedmnáctý sloupec podávají informace o zákaznících v pořadí: pohlaví, věková skupina, rodinný stav, má děti/nemá děti, pracující/nepracující.

Je nutné podotknout, že případ, kdy má obchodní řetězec k dispozici podrobné informace o všech svých zákaznících, je spíše učebnicový a v praxi velmi ojedinělý.

Větší supermarkety sice nabízejí zákazníkům různé slevové karty, jejichž získání je podmíněno právě tím, že obchodu poskytneme nějaké bližší údaje o sobě, nicméně vlastnit podobné karty zpravidla nebývá podmínkou a mnozí lidé, kteří v řetězcích nakupují, aniž by zde byli registrovaní, tedy průzkumům unikne.

Úkolem práce bylo vytvořit program, který dokáže hledat vazby mezi daty ve výše zmíněném souboru, a zprostředkovat je budoucímu uživateli programu vizuálně – v podobě nejrůznějších přehledů a grafů. Tyto vizualizace by přitom měly být maximálně přehledné, aby se v nich uživatel dokázal dobře zorientovat a snadno z nich získával informace.

Program například zjišťuje, co kteří zákazníci kupovali a co naopak nekupovali, jaká byla četnost nákupů jednotlivých druhů zboží, a to jak množstevní (kolik kusů zboží bylo koupeno), tak procentuální. Dále zkoumá, jaké druhy zboží byly kupovány zároveň s jinými a v jaké míře. Aplikace přiřazuje jak zboží k jednotlivým zákazníkům, eventuálně ke skupinám zákazníků, tak i obráceně, tedy zákaznické skupiny k určitému druhu zboží.

Dalším problémem, jímž se ve své práci zabývám, jsou asociační pravidla v data miningových úlohách a jejich získávání. Snažím se tedy nejen o pouhé vyhodnocování dat a jejich vizualizaci, ale také o vyvozování dalších závěrů, doposud v datech skrytých a na první pohled nepatrných, pomocí jedné z nejvyužívanějších data miningových technik – asociačních metod.

Posuzuji, kterým zákazníkům je a naopak není vhodné nabízet určité zboží (co by měl například prodejce nabídnout zákazníkovi za zvýhodněnou cenu, pokud si koupí pečivo), či které druhy zboží by měly být umístěny v regálech, případně na stránkách internetových e-shopů, blízko sebe, protože je zákazníci často kupují společně, pokouším se predikovat, co si zákazník pravděpodobně koupí při další návštěvě obchodu.

Důraz ve své práci kladu na snadnou použitelnost programu, jeho jednoduchou ovladatelnost a především přehlednost. Snažím se vytvořit takovou aplikaci, která by byla přístupná co nejširší škále uživatelů. Aby ji mohl využívat opravdu každý, navrhla jsem dva režimy zobrazení – první klasický a druhý zvětšený, který by měl být vhodný i pro zrakově handicapované osoby.

11 Struktura programu

11.1 Hlavní okno aplikace

Aplikace, kterou vytvářím, se skládá ze dvou částí:

- První část se zabývá hledáním vazeb mezi daty v datovém souboru, a jejich vizualizací – vytvářením různých přehledů a výčtů, vykreslováním grafů.
- Druhá část aplikace hlouběji analyzuje data a vyhledává v nich skryté vazby, které nejsou na první pohled zřejmé. Tato analýza je prováděna pomocí hledání asociačních pravidel s využitím algoritmu apriori.

Ke každé z výše zmíněných částí má uživatel přístup z hlavního okna této aplikace, které se zobrazí po jejím spuštění. Velikost okna je implicitně nastavena na maximální, takže se hlavní okno vždy zobrazí na celé obrazovce.

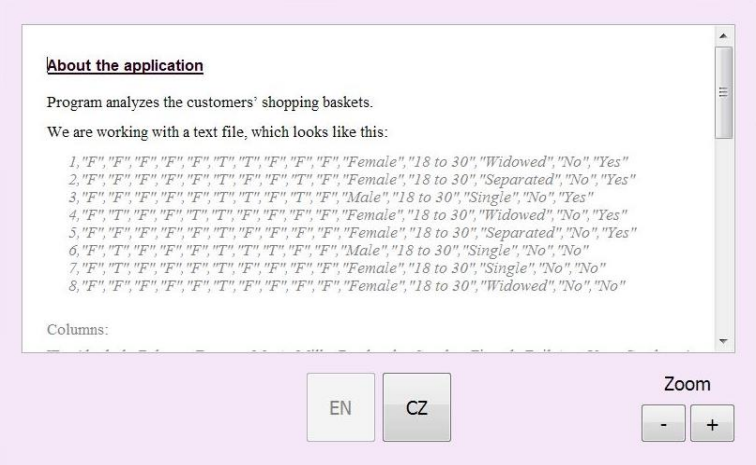
Hlavní okno aplikace vypadá takto:

Obrázek 5. Hlavní okno aplikace

Kliknutím na tlačítko *„Visualization of combinations and associations“* se uživatel dostává k vizualizačním funkcím aplikace, kliknutím na tlačítko *„Combinations and association rules“* se pak dostátá do části aplikace umožňující vyhledávání asociačních pravidel v datech, přičemž datový soubor, který je programem analyzován, je možné vybrat pomocí tlačítka *„Choose another file“*.

V levém dolním rohu okna se nachází symbol *„?“*, který se nachází také ve všech dalších podoknech aplikace. Slouží k otevření nápovědy pro dané okno programu. Uživatel programu se tak dokáže snadno v aplikaci zorientovat a rychle porozumět jejím funkcím.

Okno nápovědy vypadá takto:



Obrázek 6. Okno nápovědy

Text nápovědy je možné dle potřeby uživatele zvětšit či zmenšit díky tlačítkům *„Zoom +“* a *„Zoom -“* a je možné číst ho ve dvou jazycích: v angličtině, která je primárním jazykem celého programu, a v češtině. Volbu jazyka provádí uživatel pomocí tlačítek *„EN“* a *„CZ“*.

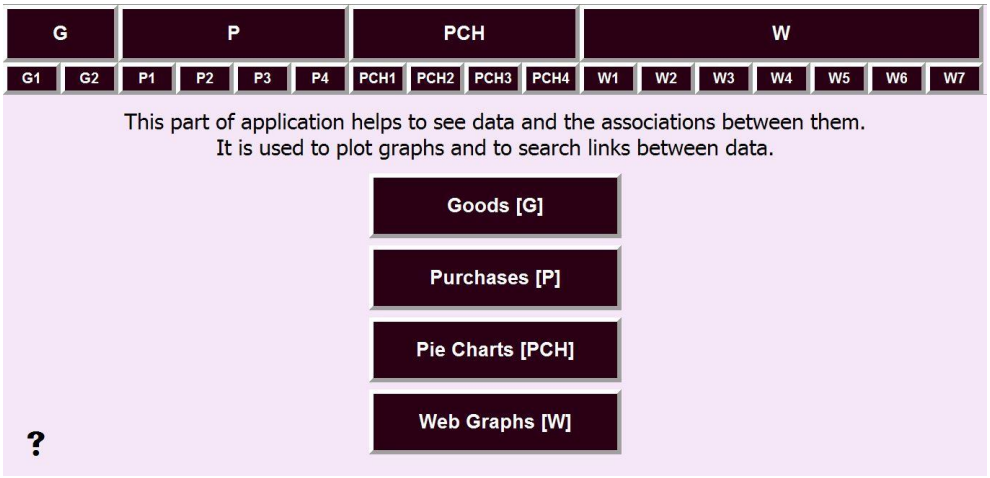
11.2 Vizualizace dat

První část aplikace se věnuje hledání základních vazeb mezi daty v datovém souboru, a jejich vizualizací – vytváření přehledů a výčtů a vykreslování grafů.

11.2.1 Hlavní okno vizualizační části aplikace

Po kliknutí na tlačítko *„Visualization of combinations and associations“* v hlavním okně programu se uživateli otevře okno, z něhož má přístup k vizualizačním funkcím aplikace. Velikost okna je, podobně jako tomu bylo i u okna hlavního, implicitně nastavena na maximální, takže se toto okno po otevření zobrazí na celé obrazovce.

Hlavní okno vizualizační části aplikace vypadá takto:



Obrázek 7. Hlavní okno vizualizační části aplikace

V záhlaví okna jsou rozmístěna tlačítka, která jsou označena pouze zkratkami, nikoli celými slovy. Tato zkratkovitá pojmenování jsou zvolena záměrně – z důvodu zachování přehlednosti. Umístěním kurzoru myši nad příslušné tlačítko pak uživatel zjistí, kterou funkci programu pod sebou skrývá, zobrazí se mu celý název okna, které se mu po kliknutí následně otevře (tedy např. pokud uživatel umístí kurzor myši nad tlačítko *„G1“*, zjistí, že toto tlačítko odkazuje na okno s názvem *„Basic list“*; po kliknutí na tlačítko se zobrazí podokno aplikace pojmenované jako *„Basic list“*, které podává informaci o tom, jaké druhy zboží nabízel zmíněný obchod).

Tlačítka označená písmeny *„G“*, *„P“*, *„PCH“* a *„W“* slouží k logickému rozčlenění tlačítek *„G1“* – *„W7“*, vztahují se k tlačítkům umístěným uprostřed okna, tedy *„G“* jako *„Goods“* (Zboží), *„P“* jako *„Purchases“* (Nákupy), *„PCH“* jako *„Pie Charts“* (Koláčové grafy) a *„W“* jako *„Web Graphs“* (Pavučinové grafy). Díky těmto tlačítkům v záhlaví má uživatel jasný přehled o tom, které funkce mu nabízejí příslušná tlačítka uprostřed (např. pod tlačítkem *„G“* jsou umístěna dvě tlačítka *„G1“* a *„G2“*, takže současně tlačítko *„Goods“* pod sebou bude skrývat okno, pomocí něhož se lze dále dostat k dvěma podoknům *„Basic list“* a *„Sorted by customer ID“*).

Tlačítka *„G1“* – *„W7“* už své funkce mají a kliknutí na ně umožňuje otevírat jednotlivá podokna aplikace. Tato podokna je možné zobrazovat na obrazovce paralelně vedle sebe, několik najednou, aplikace dovoluje libovolně překlikávat mezi nimi a porovnávat výsledky získané jejich prostřednictvím. Tato varianta otevírání a zobrazování oken ovšem neumožňuje uživateli volit mezi dvěma režimy zobrazení (klasický, zvětšený), zvětšený režim zobrazení tu není zpřístupněn.

Čtyři tlačítka umístěna uprostřed hlavního okna popsaná jako *„Goods“*, *„Purchases“*, *„Pie Charts“* a *„Web Graphs“* byla již zmíněna výše. Odkazují vždy na okno, které dále odkazuje na podokna aplikace příslušná těmto tlačítkům. Otvírání a zobrazování oken prostřednictvím těchto tlačítek se od prvního způsobu (pomocí tlačítek *„G1“* až *„W7“* liší tím, že nedovoluje mít otevřených více podoken aplikace najednou ani překlikávat kamkoli jinam, dokud nebude ukončena práce s oknem, které je momentálně aktivní. Tato metoda otevírání šetří místo v paměti počítače (protože po uzavření okna se místo v paměti uvolní) a také nabízí kromě klasického režimu zobrazování i režim zvětšený, o jehož funkcích se blíže zmíním později.

11.2.2 Tlačítko „Goods“ a podokna aplikace, k nimž se vztahuje

Po kliknutí na tlačítko *„Goods“* (Zboží) se uživateli zobrazí okno s dalšími dvěma tlačítky: *„Basic list“* (Základní přehled) a *„Sorted by customer ID“* (Rozdělení podle ID zákazníků).

Okno *„Goods“* (Zboží) vypadá takto:

Obrázek 8. „Goods“

Okna *„Basic list“* (Základní přehled) a *„Sorted by customer ID“* (Rozdělení podle ID zákazníků) vidíme na následujících obrázcích:

Obrázek 9. „Basic list“

Obrázek 10. „Sorted by customer ID“

„Basic list“ podává uživateli informaci o tom, jaké zboží měl v nabídce zmíněný obchod.

„Sorted by customer ID“ umožňuje vybírat si ze seznamu zákazníky podle jejich identifikačního čísla a zjistit, jaké druhy zboží kupovali, eventuálně nekupovali. Ze seznamu zákazníků si zvolím např. zákazníka s identifikačním číslem 13, chci vědět, které druhy zboží si koupil, takže označím „*Purchased*“ (Zakoupené) a program mi tuto informaci poskytne – druhy zboží, na něž se dotazuji, označí bíle. Nezakoupené zboží je pak označováno fuchsiovou barvou a toto barevné rozlišení platí stejně, pokud se ptám na zakoupené i nezakoupené zároveň.

11.2.3 Tlačítko *„Purchases“* a podokna aplikace, k nimž se vztahuje

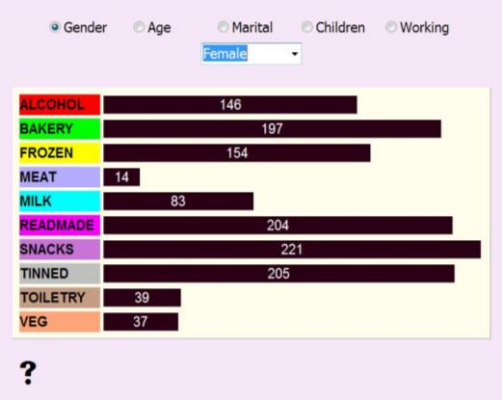
Po kliknutí na tlačítko *„Purchases“* (Nákupy) se uživateli zobrazí okno s dalšími čtyřmi tlačítky: *„Purchased goods“* (Zakoupené zboží), *„Customer Purchases – Basic survey“* (Nákupy zákazníků – jednoduchý přehled), *„Customer Purchases – Extended survey“* (Nákupy zákazníků – rozšířený přehled) a *„Details of the sale“* (Detaily o prodeji).

Okno *„Purchases“* (Nákupy) vypadá takto:

Obrázek 11. „Purchases“

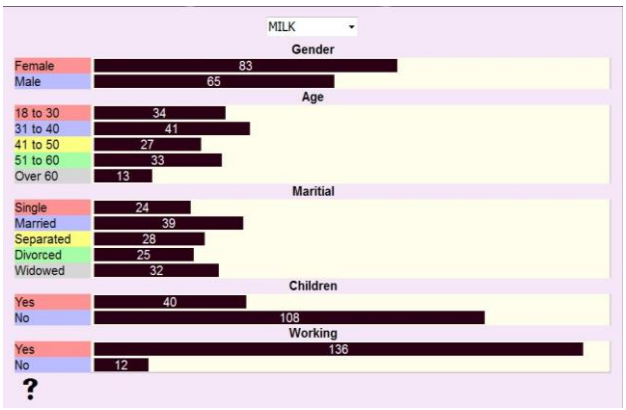
Okna *„Purchased goods“* (Zakoupené zboží), *„Customer Purchases – Basic survey“* (Nákupy zákazníků – jednoduchý přehled), *„Customer Purchases – Extended survey“* (Nákupy zákazníků – rozšířený přehled) a *„Details of the sale“* (Detaily o prodeji) vidíme na následujících obrázcích:

Obrázek 12. „Purchased goods“



Obrázek 13. „Customer purchases – Basic survey“

Obrázek 14. „Customer purchases – Extended survey“



Obrázek 15. „Details of the sale“

„Purchased goods“ uživatele informuje o tom, kolik kusů daného zboží bylo zakoupeno (tedy např. že alkohol byl v nákupních košících všech zákazníků zastoupen celkem 310 krát) a tuto informaci vykreslí pomocí sloupcového (v našem případě přesněji „řádkového“ grafu). Umístí-li uživatel kurzor myši nad příslušný graf, dozví se také, o kolik procent z celkového množství zakoupeného zboží se jednalo (v případě alkoholu šlo o necelých 13% z celkového množství zakoupeného zboží).

„Customer Purchases – Basic survey“ má podobnou funkci jako *„Purchased goods“* – opět informuje o tom, kolik kusů jednotlivých druhů zboží bylo zakoupeno, nicméně tentokrát se zaměřuje na konkrétní zákazníky, přesněji skupiny zákazníků. Uživatel např. potřebuje zjistit, kolik kusů jednotlivých druhů zboží zakoupily ženy, označí tedy rozdělení zákazníků podle pohlaví (Gender) a následně ze seznamu vybere ženy (Female). Program spočítá, že ženy kupovaly alkohol 146 krát, pečivo 197 krát atd. I zde program vypočítá, o kolik procent z celkového množství zakoupeného zboží se jednalo.

„Customer Purchases – Extended survey“ funguje prakticky stejně jako *„Customer Purchases – Basic survey“*, opět počítá zastoupení zboží, množstevní i procentuální, v košících zákazníků a tyto zákazníky opět dělí do skupin. Zatímco v *„Základním přehledu“* však bylo možné rozdělit zákazníky vždy jen do základních skupin podle pohlaví, věku, rodinného stavu atd., *„Rozšířený přehled“* nabízí i možnost dělit tyto skupiny dále do podskupin. Lze například zjistit, kolik kusů zboží nakoupily ženy ve věku od 31 do 40 let nebo nezaměstnaní muži i ženy starší 50 let, kteří nemají děti apod.

„Details of the sale“ funguje obráceně, než tři výše zmíněné programy, tedy nepřiřazuje zákazníkům zboží, ale naopak ke zboží přiřazuje zákazníky. Ti jsou rozděleni do 16 základních skupin: muž/žena, 5 věkových skupin, 5 skupin podle rodinného stavu, rodič/bezdětný, zaměstnaný/nezaměstnaný. Uživatel ze seznamu vybere, jaký druh zboží ho zajímá, a pomocí programu zjistí, kolik lidí z jednotlivých skupin toto zboží kupovalo. Např. mléčné výrobky mělo v nákupním košíku 83 žen a 65 mužů, 136 zaměstnaných a jen 12 nezaměstnaných lidí atd.

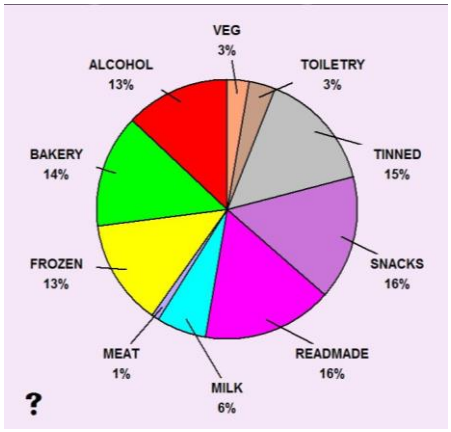
11.2.4 Tlačítko *„Pie Charts“* a podokna aplikace, k nimž se vztahuje

Po kliknutí na tlačítko *„Pie Charts“* (Koláčové grafy) se uživateli zobrazí okno s dalšími čtyřmi tlačítky: *„Global Overview“* (Celkový přehled), *„Sorted by customers – Basic survey“* (Rozdělení podle zákazníků – jednoduchý přehled), *„Sorted by customers – Extended survey“* (Rozdělení podle zákazníků – rozšířený přehled) a *„Sorted by goods“* (Rozdělení podle zboží).

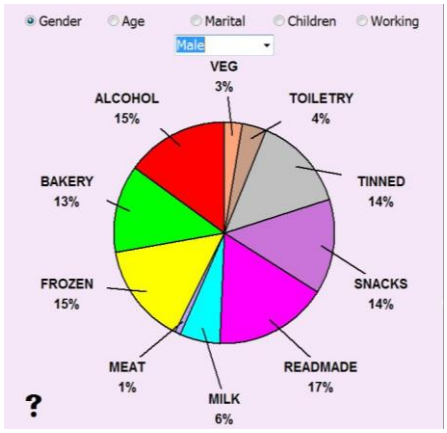
Okno *„Pie Charts“* (Koláčové grafy) vypadá takto:

Obrázek 16. „Pie Charts“

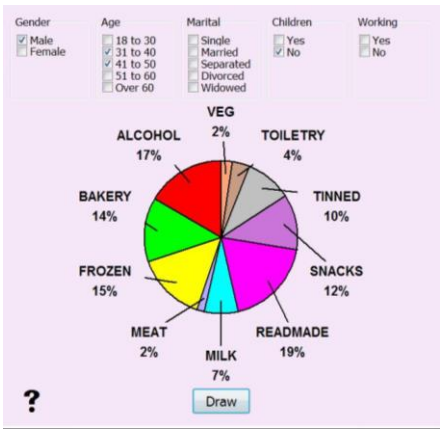
Okna *„Global Overview“* (Celkový přehled), *„Sorted by customers – Basic survey“* (Rozdělení podle zákazníků – jednoduchý přehled), *„Sorted by customers – Extended survey“* (Rozdělení podle zákazníků – rozšířený přehled) a *„Sorted by goods“* (Rozdělení podle zboží) vidíme na následujících obrázcích:



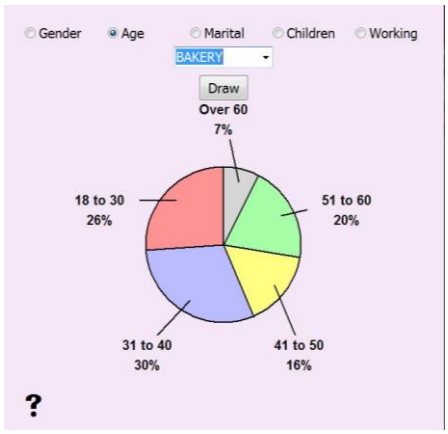
Obrázek 17. „Global Overview“



Obrázek 18. „Sorted by customers – Basic survey“



Obrázek 19. „Sorted by customers – Extended survey“



Obrázek 20. „Sorted by goods“

„Global overview“ podává uživateli celkový přehled o tom, v jaké procentuální míře byly v nákupních košících zákazníků zastoupeny jednotlivé druhy zboží a tuto informaci vykreslí pomocí koláčového (výsečového) grafu. Pokud uživatel umístí kurzor myši na políčko, kde jsou vypsaná procenta, dozví se také, o kolik kusů zboží daného druhu se jednalo. Tedy např. z celkového množství zakoupeného zboží činilo 14% pečivo, které bylo v košících zákazníků celkem obsaženo 337 krát.

„Sorted by customers – Basic survey“ funguje opět podobně jako *„Global overview“* – počítá zboží a vykresluje jeho procentuální zastoupení v košících zákazníků pomocí koláčového grafu. Zákazníci jsou však tentokrát již rozděleni do základních skupin podle pohlaví, věku atd. (podobně, jako tomu bylo v případě *„Customer Purchases – Basic survey“*) a množství zakoupeného zboží se počítá vždy

pouze pro jednotlivé skupiny. Potřebuje-li uživatel programu například vědět, jaké množství pečiva kupovali muži, vybere si třídění zákazníků podle pohlaví, v seznamu zvolí muže a zjistí, že z celkového množství zboží, které muži kupovali, bylo pečiva 13% (140 kusů).

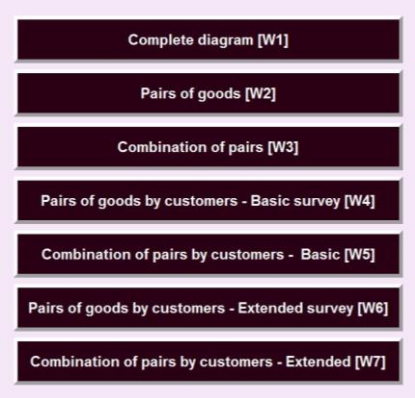
„Sorted by customers – Extended survey“, tedy rozšířený přehled o procentuálním zastoupení zboží v košících zákazníků, opět třídí zákazníky do skupin, a jak už napovídá slovo „rozšířený“ v názvu podprogramu, tyto skupiny jsou dále děleny do podskupin. V praxi program dál funguje stejně jako dva předchozí: spočítá procentuální obsažení zboží v košících vybrané skupiny zákazníků a vykreslí koláčový graf. Např. z celkového množství zboží, které koupili bezdětní muži ve věku od 31 do 50 let, činilo pečivo 14% (36 kusů).

„Sorted by goods“ počítá procentuální zastoupení vždy jednoho konkrétního druhu zboží v košících zákazníků, rozdělených do základních skupin podle pohlaví, věku, atd. Podle jakého kritéria se mají zákazníci dělit, si uživatel zvolí, stejně jako druh zboží, o který se zajímá. Chce-li např. zjistit kolik procent mužů a kolik procent žen kupovalo maso, označí v nabídce rozdělení podle pohlaví a v seznamu vybere maso. Program spočítá, že maso kupovalo 39% mužů a 61% žen (v kusech to činilo 9 a 14) z celkového množství lidí, kteří kupovali maso.

11.2.5 Tlačítko „Web Graphs“ a podokna aplikace, k nimž se vztahuje

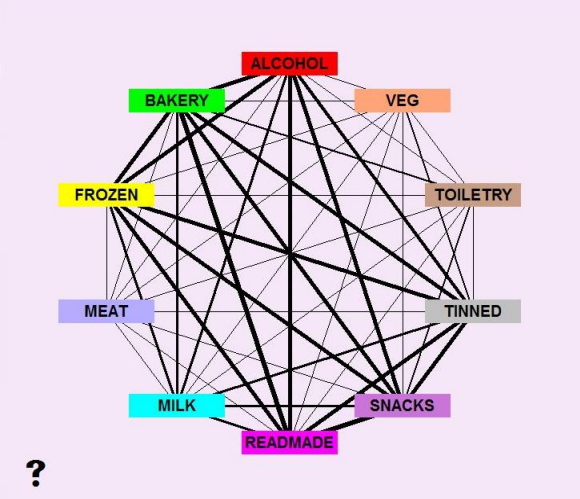
Po kliknutí na tlačítko *„Web Graphs“* (Pavučinové grafy) se uživateli zobrazí okno s dalšími sedmi tlačítky: *„Complete diagram“* (Celkový přehled), *„Pairs of goods“* (Dvojice zboží), *„Combination of pairs“* (Kombinace párů), *„Pairs of goods by customers – Basic survey“* (Páry zboží podle zákazníků – jednoduchý přehled), *„Combination of pairs by customer - Basic“* (Kombinace párů podle zákazníků – zjednodušený přehled), *„Pairs of goods by customers – Extended survey“* (Páry zboží podle zákazníků – rozšířený přehled) a *„Combination of pairs by customer – Extended“* (Kombinace párů podle zákazníků – rozšířený přehled).

Okno *„Web Graphs“* (Pavučinové grafy) vypadá takto:

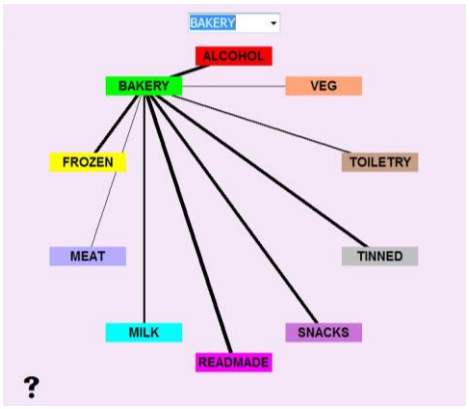


Obrázek 21. „Web Graphs“

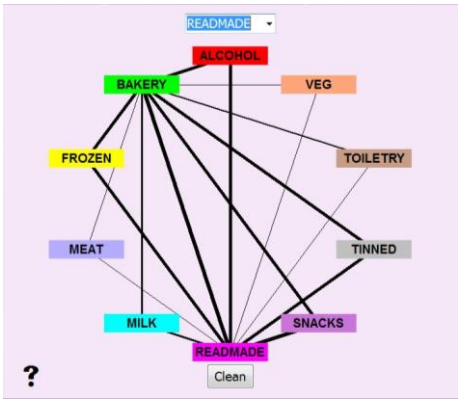
Okna „Complete diagram“ (Celkový přehled), „Pairs of goods“ (Dvojice zboží), „Combination of pairs“ (Kombinace párů), „Pairs of goods by customers – Basic survey“ (Páry zboží podle zákazníků – jednoduchý přehled), „Combination of pairs by customer - Basic“ (Kombinace párů podle zákazníků – zjednodušený přehled), „Pairs of goods by customers – Extended survey“ (Páry zboží podle zákazníků – rozšířený přehled) a „Combination of pairs by customer – Extended“ (Kombinace párů podle zákazníků – rozšířený přehled) vidíme na následujících obrázcích:



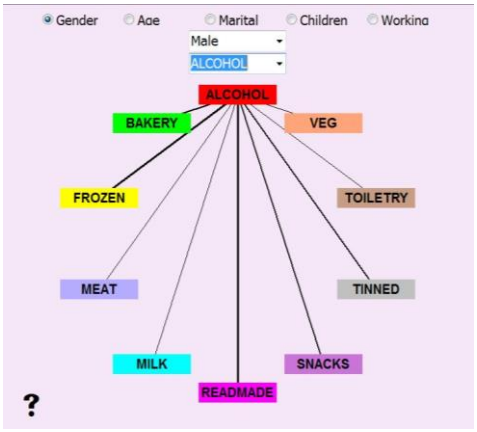
Obrázek 22. „Complete diagram“



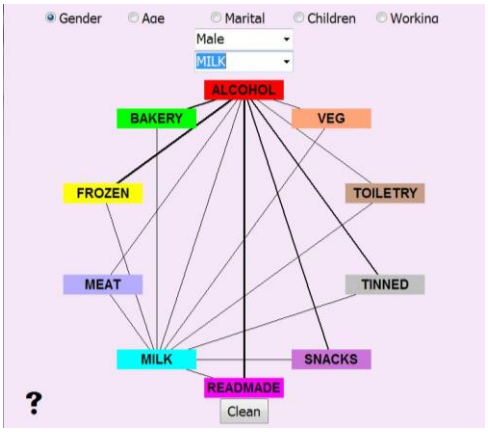
Obrázek 23. „Pairs of goods“



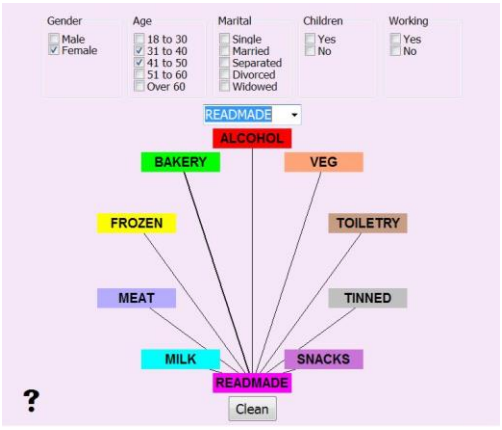
Obrázek 24. „Combination of pairs“



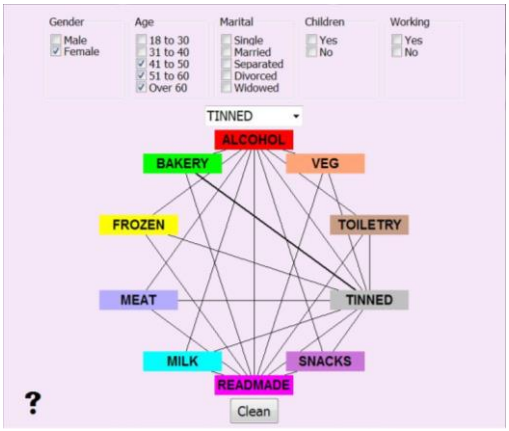
Obrázek 25. „Pairs of goods by customers – Basic survey“



Obrázek 26. „Combination of pairs by customer - Basic“



Obrázek 27. „Pairs of goods by customers – Extended survey“



Obrázek 28. „Combination of pairs by customer - Extended“

Pavučinové grafy pomáhají uživateli programu zjistit, který druh zboží byl často v nákupních košících zákazníků společně s jiným druhem. Četnost výskytu dvojic zboží udává tloušťka čáry, jíž jsou tyto druhy zboží propojeny (čím výraznější čára, tím častěji se daná dvojice v nákupních košících vyskytovala společně, pokud čára chybí, zboží se společně nekupovalo nikdy).

„Complete diagram“ dává uživateli celkový přehled o nákupech všech dvojic zboží (s čím vším a v jaké míře se kupoval alkohol, s čím pečivo atd.).

„Pairs of goods“ propojuje vždy jeden určitý vybraný druh zboží se všemi dalšími, s nimiž byl zakoupen a počítá také, v jaké míře (kusy). Zajímá-li vás například, s čím vším a v jaké míře zákazníci kupovali pečivo, vyberete z nabídky zboží v seznamu pečivo a program vykreslí žádaný graf. Chcete-li se pak dozvědět, kolik kusů pečiva bylo zakoupeno společně s časopisy, stačí umístit kurzor myši nad pole „READMADE“ a zjistíte, že pečivo se spolu s časopisy kupovalo 201 krát.

„Combination of pairs“ umožňuje, na rozdíl od předchozího *„Pairs of goods“*, propojit grafem libovolný počet položek zboží, neudává však již informaci o množství zakoupených dvojic. Umožňuje sestavovat uživateli větší skupiny spolu nejčastěji kupovaného zboží – k nejčastěji zakoupené dvojici produktů vybrat třetí, který zákazníci také často kupovali, pokud kupovali i předchozí dva, k trojici vybrat čtvrtý produkt atd. Prakticky funguje tento podprogram zhruba takto: z nabídky zboží vybereme nejprve například již výše zmíněné pečivo, které se často kupovalo s časopisy. Chceme-li ke dvojici BAKERY – READMADE vybrat třetí produkt, který byl s těmito dvěma často kupován, v nabídce zboží dále zvolíme časopisy a z grafu vyčteme, že spolu s časopisy i pečivem byl v košících zákazníků velmi často obsažen alkohol nebo mražené výrobky. Potřebujeme-li zjistit, o jaké konkrétní množství zakoupeného zboží se jednalo, stačí se vrátit k předchozímu grafu (*„Pairs of goods“*) a tuto informaci zde dohledat.

„Pairs of goods by customers – Basic survey“ funguje prakticky stejně jako *„Pairs of goods“* s jediným rozdílem: dělí zákazníky do základních skupin (opět podle pohlaví, věku atd.) a vykresluje grafy četnosti dvojic zboží v košících těchto konkrétních skupin zákazníků. Opět podává i informaci o množství zakoupeného zboží. Např. muži kupovali alkohol často s pečivem, mraženými výrobky, časopisy či občerstvením, a porovnáme-li množství, zjistíme, že z těchto čtyř se v nákupních košících mužů spolu s alkoholem opravdu nejčastěji vyskytovaly mražené výrobky (97 kusů).

„Combination of pairs by customer - Basic“ funguje obdobně jako *„Combination of pairs“* s tím, že zákazníky dělí opět (jako v případě *„Pairs of goods by customer – Basic survey“*) do základních skupin podle pohlaví, věku atd. Pro tyto skupiny pak hledá n-tice společně nejčastěji zakoupeného zboží. Například zákazníci mužského pohlaví tedy v našem obchodě kupovali nejčastěji alkohol spolu s mraženými výrobky a časopisy.

„Pairs of goods by customers – Extended survey“ člení zákazníky do skupin a ty dále do podskupin a pro tyto pak počítá a vykresluje obdobný graf jako *„Pairs of goods“* nebo *„Pairs of goods by customer – Basic survey“*. Opět udává i informaci o množství spolu zakoupených dvojic zboží. Např. ženy ve věku od 31 do 40 let, které kupovaly časopisy, měly nejčastěji společně s nimi v nákupním košíku také pečivo (37 kusů).

„Combination of pairs by customer – Extended“ je rozšířený přehled o n-ticích společně zakoupených produktů skupin, eventuálně podskupin zákazníků. Ženy ve věku 31 a 40 let, jak již z předchozího případu víme, kupovaly nejčastěji časopisy spolu s pečivem. Pečivo a časopisy se pak v nákupních košících těchto žen vyskytovaly nejčastěji v kombinaci s občerstvením.

11.2.6 Asociační pravidla v data miningu

Druhá část aplikace se věnuje hlubší analýze dat a vyhledávání skrytých vazeb a vztahů mezi nimi, odhalování závislostí, které nejsou na první pohled zřejmé. Tato analýza je prováděna pomocí asociačních metod, které jsou spolu s rozhodovacími stromy nejpoužívanější data miningovou technikou.

11.2.7 Hlavní okno části aplikace zabývající se asociačními metodami

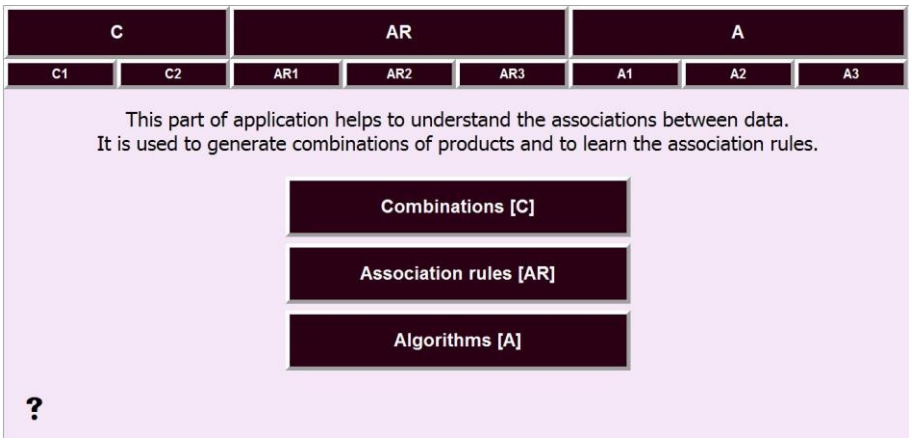
Po kliknutí na tlačítko *„Combinations and association rules“* v hlavním okně programu se uživateli otevře okno, z něhož má přístup k funkcím aplikace, které umožňují generování možných kombinací zboží zakoupeného zákazníky v obchodě a generování asociačních pravidel se syntaxí: *Jestliže platí předpoklad A, pak platí*

i závěr B (v případě analýzy nákupního košíku tedy *Jestliže si zákazník koupí zboží A, pravděpodobně si koupí i zboží B*).

Asociační pravidla jsou vyhledávána pomocí algoritmu apriori, což je pravděpodobně nejznámější algoritmus pro hledání asociačních pravidel vůbec, a ačkoli se dnes využívá i v jiných sférách, původně byl navržen a určen právě pro analýzou nákupního košíku.

Velikost okna je, podobně jak tomu bylo v případě hlavního okna celé aplikace i hlavního okna její vizualizační části, implicitně nastavena na maximální, takže se okno po otevření zobrazí na celé obrazovce.

Hlavní okno části aplikace zabývající se asociačními metodami vypadá takto:



Obrázek 29. Hlavní okno části aplikace zabývající se asociačními metodami

Struktura tohoto okna je stejná, jako tomu bylo v případě hlavního okna vizualizační části aplikace.

V záhlaví okna jsou opět rozmístěna tlačítka označená zkratkami, která umožňují otvírat jednotlivá podokna aplikace a zobrazovat je na obrazovce paralelně vedle sebe, což je výhodné, potřebuje-li uživatel porovnávat výsledky získané jejich prostřednictvím. Umístěním kurzoru myši nad příslušné tlačítko lze zjistit, kterou funkci programu pod sebou skrývá – zobrazí se celý název okna, které se po kliknutí na toto tlačítko následně otevře.

Tlačítka označená písmeny *„C“*, *„AR“* a *„A“* slouží pouze k logickému rozčlenění tlačítek *„C1“* – *„A3“*, žádnou další funkci nemají a po kliknutí na ně se nezobrazí žádné podokno aplikace. Vztahují se k tlačítkům umístěným uprostřed okna, tedy *„C“* jako *„Combinations“* (Kombinace), *„AR“* jako *„Association rules“*

(Asociační pravidla) a *„A“* jako *„Algorithms“* (Algoritmy). Díky těmto tlačítkům v záhlaví má uživatel jasný přehled o tom, které funkce mu nabízejí příslušná tlačítka uprostřed (např. pod tlačítkem *„C“* jsou umístěna dvě tlačítka *„C1“* a *„C2“*, takže současně tlačítko *Combinations* pod sebou bude skrývat okno, pomocí něhož se lze dále dostat k dvěma podoknům *„Combinations of purchased and not purchased goods“* a *„Combinations of purchased goods“*).

Pomocí tlačítek *„C1“* – *„A3“* lze otevírat jednotlivá podokna aplikace a využívat jejich funkce, přičemž tato okna je možné zobrazovat na monitoru paralelně vedle sebe a porovnávat výsledky analýz. Nevýhodou je, že takto paralelně otevřená okna neumožňují uživateli volit mezi dvěma režimy zobrazování a nabízí pouze režim klasický – nezvětšený.

Tři tlačítka umístěna uprostřed hlavního okna popsaná jako *„Combinations“*, *„Association rules“* a *„Algorithms“* byla již zmíněna výše. Odkazují vždy na okno, které dále odkazuje na podokna aplikace příslušná těmto tlačítkům. Otvírání a zobrazování oken prostřednictvím těchto tlačítek se od prvního způsobu (pomocí tlačítek *„C1“* až *„A3“* liší tím, že nedovoluje mít otevřených více podoken aplikace najednou ani překlikávat kamkoli jinam, dokud nebude ukončena práce s oknem, které je momentálně aktivní. Tato metoda otevírání šetří místo v paměti počítače (protože po uzavření okna se místo v paměti uvolní) a nabízí ještě jednu výhodu: kromě klasického režimu zobrazování i je možné aktivovat i režim zvětšený, který je vhodný i pro zrakově handicapované a o jehož funkcích se blíže zmíním na později.

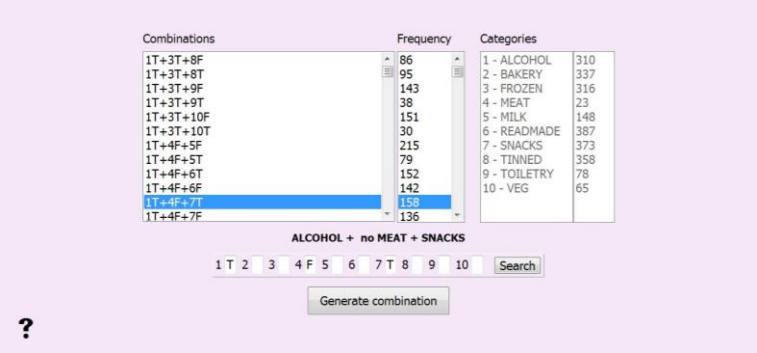
11.2.8 Tlačítko „*Combinations*“ a podokna aplikace, k nimž se vztahuje

Po kliknutí na tlačítko *„Combinations“* (Kombinace) se uživateli zobrazí okno s dalšími dvěma tlačítky: *„Combinations of purchased and not purchased goods“* (Kombinace zakoupeného a nezakoupeného zboží) a *„Combinations of purchased goods“* (Kombinace zakoupeného zboží).

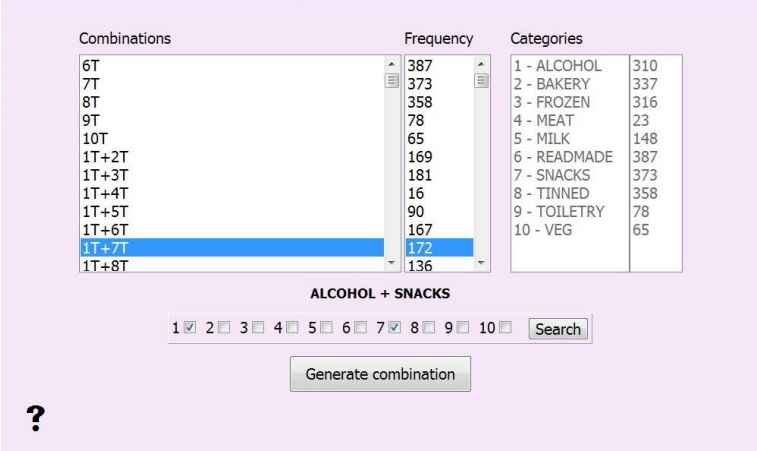
Okno „*Combinations*“ (Kombinace) vypadá takto:

Obrázek 30. „Combinations“

Okna *„Combinations of purchased and not purchased goods“* (Kombinace zakoupeného a nezakoupeného zboží) a *„Combinations of purchased goods“* (Kombinace zakoupeného zboží) vidíme na následujících obrázcích:



Obrázek 31. „Combinations of purchased and not purchased goods"



Obrázek 32. „Combinations of purchased goods"

„Combinations of purchased and not purchased goods“ slouží ke generování všech možných kombinací zakoupeného i nezakoupeného zboží a výpočtu, jak často se tato kombinace vyskytovala v nákupních košících zákazníků obchodního řetězce. Záleží přitom i na zboží, které zakoupené nebylo a v datovém souboru je reprezentováno hodnotou *„F“* (False).

„Combinations of purchased goods“ slouží ke generování všech možných kombinací zakoupeného zboží a zjišťování, jak často se tato kombinace vyskytovala v nákupních košících zákazníků obchodního řetězce. Na zboží, které zakoupeno nebylo, přitom nezáleží, berou se v úvahu pouze kombinace zakoupených položek.

První ze dvou výše zmíněných funkcí aplikace umožňuje uživateli vyhledávat kombinace zboží typu *Položka1(T) + Položka2(F),* příkladem takové kombinace zboží může být *Alkohol(T) + Maso (F) + Občerstvení(T),* která se v nákupních košících zákazníků vyskytovala 158 krát, přičemž záleží nejen na zakoupených položkách, kterými jsou v tomto případě *alkohol* a *občerstvení*, ale i na položce nezakoupené, kterou je zde *maso.*

Z praktického hlediska je pro nás ovšem zajímavější pouze zboží zakoupené, zůstaneme-li u předchozího příkladu, budou to položky *alkohol* a *občerstvení,* přičemž kombinace *alkohol + občerstvení* se v nákupních košících zákazníků vyskytovala 172 krát – je to 158 případů, kde se navíc nevyskytovalo *maso,* a 14 případů, kde se *maso* vyskytovalo. Generovat kombinace typu *Položka1 + Položka2* a zjišťovat četnost jejich zastoupení v nákupních košících zákazníků umožňuje uživateli druhá ze dvou výše zmíněných funkcí.

11.2.9 Tlačítko „*Association rules*“ a podokna aplikace, k nimž se vztahuje

Po kliknutí na tlačítko *„Association rules“* (Asociační pravidla) se uživateli zobrazí okno s dalšími třemi tlačítky: *„How to extract association rules“* (Jak získávat asociační pravidla), *„Details of the association rules“* (Detaily o asociačních pravidlech) a *„How to extract association rules – APRIORI algorithm“* (Jak získávat asociační pravidla – algoritmus APRIORI),.

Okno „*Association rules“* (Asociační pravidla) vypadá takto:

Obrázek 33. „Association rules“

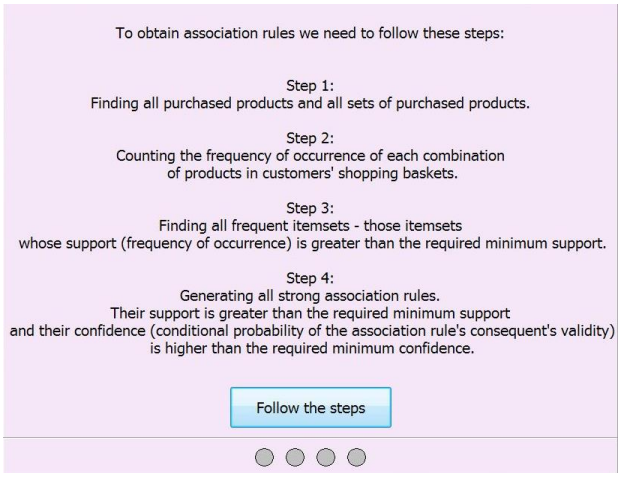
Tato část aplikace je určena k objasnění pojmu *asociační pravidlo* pomocí grafických vizualizací. Jsou zde popsány a názorně vysvětleny jednotlivé kroky získávání asociačních pravidel, uvedeny jsou dvě ukázky – první způsob klasický bez jakéhokoli urychlování a druhý způsob rychlejší s použitím algoritmu apriori, a následně také všechny důležité charakteristiky asociačních pravidel, jimiž jsou podpora (support), a to jak podpora celého asociačního pravidla, tak i pouze jeho předpokladu či pouze jeho závěru, spolehlivost (confidence), navýšení (lift) a uplatnění (deployability).

„How to extract association rules“ vysvětluje uživateli, jak se z dat získávají asociační pravidla. Celý proces je demonstrován na malém množství dat (matice 4x4 – čtyři druhy zboží zakoupené nebo nezakoupené čtyřmi zákazníky) a je vizualizován ve čtyřech krocích:

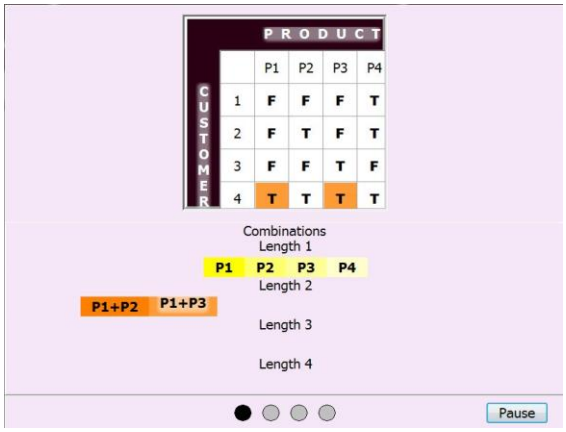
- Hledání zakoupených produktů a všech kombinací zakoupených produktů
- Výpočet frekvence výskytu každé kombinace produktů v nákupních košících zákazníků
- Hledání frekventovaných množin
- Získání silných asociačních pravidel

Každý krok je podrobně vysvětlen a poté vizualizován graficky, což by mělo vést k jeho snazšímu pochopení.

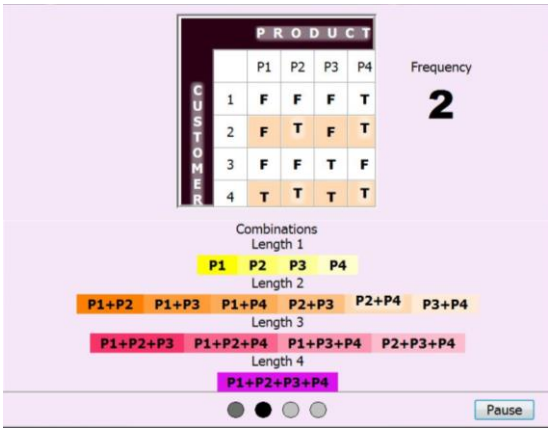
Ukázky z části programu *„How to extract association rules“* vidíme na následujících obrázcích:



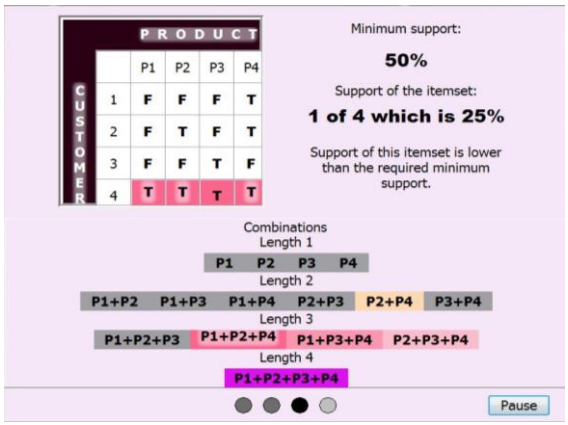
Obrázek 34. „How to extract association rules“ – popis procesu získávání asociačních pravidel



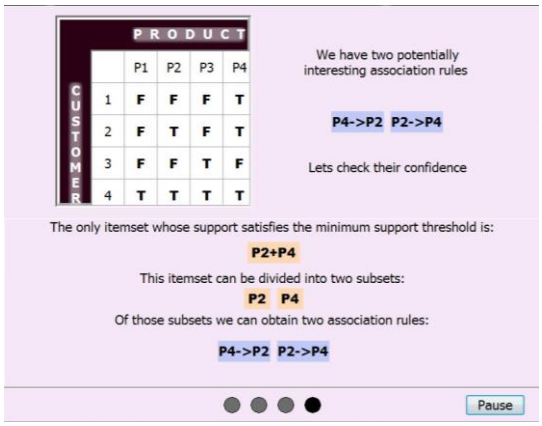
Obrázek 35. „How to extract association rules" - hledání kombinací zakoupeného zboží



Obrázek 36. „How to extract association rules" - frekvence výskytu nalezených kombinací zboží



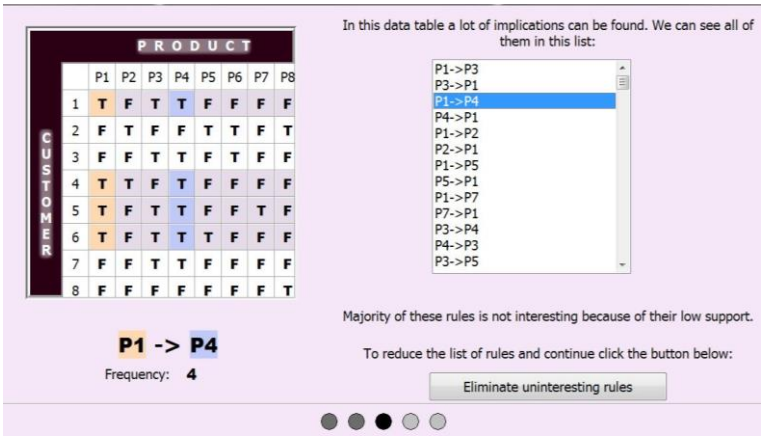
Obrázek 37. „How to extract association rules" - hledání frekventovaných množin



Obrázek 38. „How to extract association rules" - nalezení silných asociačních pravidel

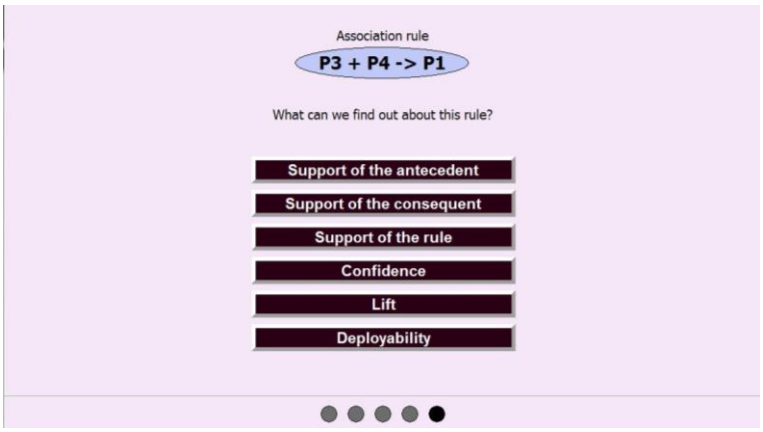
„Details of the association rules" seznamuje uživatele blíže s problematikou asociačních pravidel a s pojmy, které s nimi souvisejí – s jejich důležitými charakteristikami.

Vše je, obdobně jako v předchozím případě, demonstrováno na malém množství dat (matice 8x8 – osm druhů zboží zakoupených nebo nezakoupených osmi zákazníky). V této datové tabulce jsou nejprve nalezena všechna asociační pravidla, a ta jsou zobrazena v seznamu, jak vidíme na následujícím obrázku.



Obrázek 39. „Details of the association rules" - seznam nalezených implikací

Seznam asociačních pravidel je následně zredukován na malé množství nejzajímavějších implikací, z nichž je nakonec vybrána jedna, a na ní jsou demonstrovány všechny další výpočty.



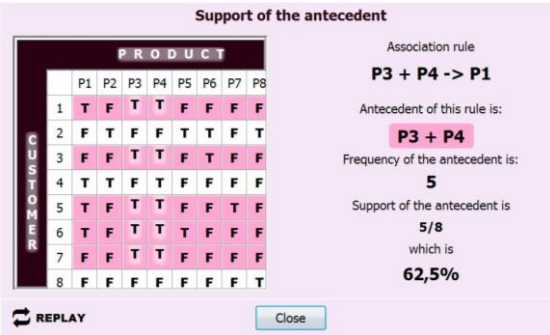
Obrázek 40. „Details of the association rules" - charakteristiky asociačních pravidel

Aplikace ukazuje, jak se vypočítá:

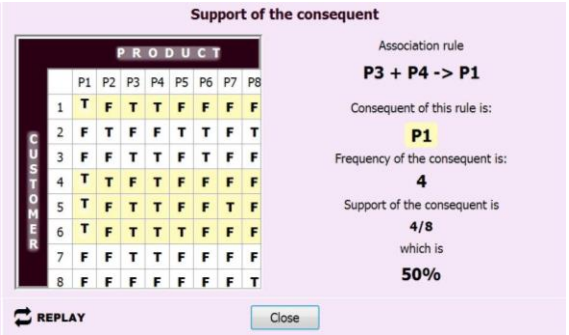
- podpora (support) předpokladu asociačního pravidla
- podpora (support) závěru asociačního pravidla
- podpora (support) asociačního pravidla
- spolehlivost (confidence) asociačního pravidla
- navýšení (lift) asociačního pravidla
- uplatnění (deployability) asociačního pravidla

Každý z těchto pojmů je popsán a jeho výpočet podrobně vysvětlen a graficky vizualizován, což by mělo usnadnit jeho pochopení.

Další ukázky z části programu *„Details of the association rules“* vidíme na následujících obrázcích:

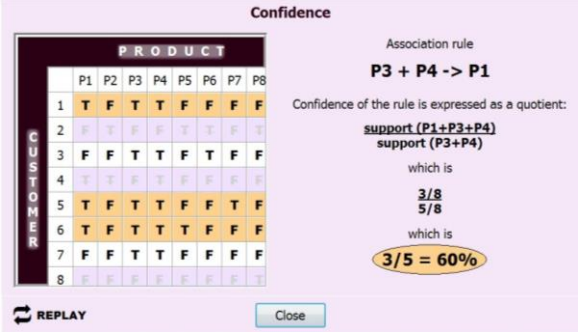


Obrázek 41. „Details of the association rules“ - podpora (support) předpokladu



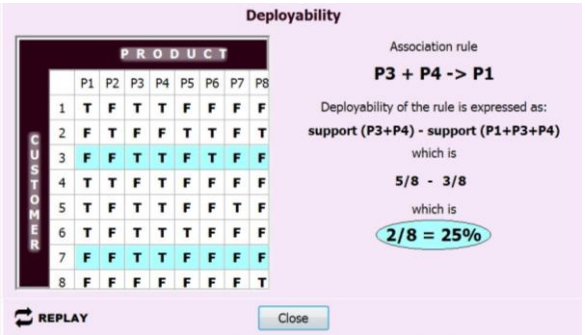
Obrázek 42. „Details of the association rules“ - podpora (support) závěru

Obrázek 43. „Details of the association rules“ - podpora (support) asociačního pravidla



Obrázek 44. „Details of the association rules“ - spolehlivost (confidence) asociačního pravidla

Obrázek 45. „Details of the association rules“ - navýšení (lift) asociačního pravidla



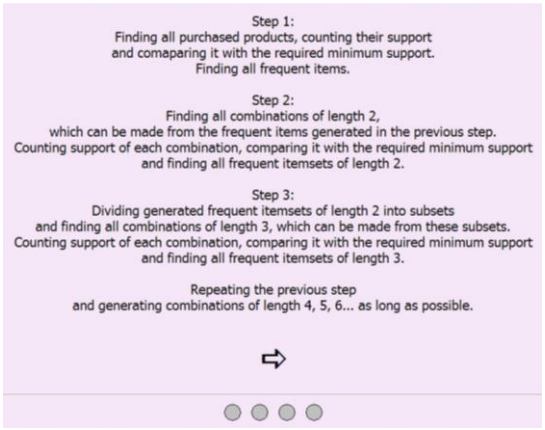
Obrázek 46. „Details of the association rules“ - uplatnění (deployability) asociačního pravidla

„How to extract association rules – APRIORI algorithm“ vysvětluje uživateli, jak se z dat získávají asociační pravidla pomocí jednoho z nejznámějších algoritmů – algoritmu apriori. Celý proces je demonstrován na malém množství dat (matice 4x4 – čtyři druhy zboží zakoupené nebo nezakoupené čtyřmi zákazníky) a je vizualizován ve třech krocích:

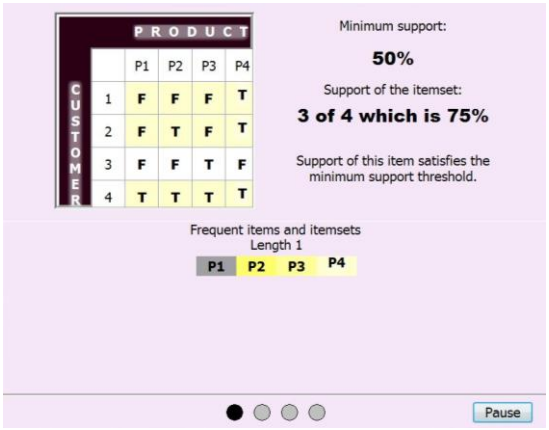
- Hledání všech frekventovaných položek
- Hledání všech frekventovaných množin délky 2 a více na základě principu algoritmu apriori
- Získání všech silných asociačních pravidel

Každý krok procesu a celý princip algoritmu apriori je podrobně vysvětlen a poté vizualizován graficky, což by mělo vést k jeho snazšímu pochopení.

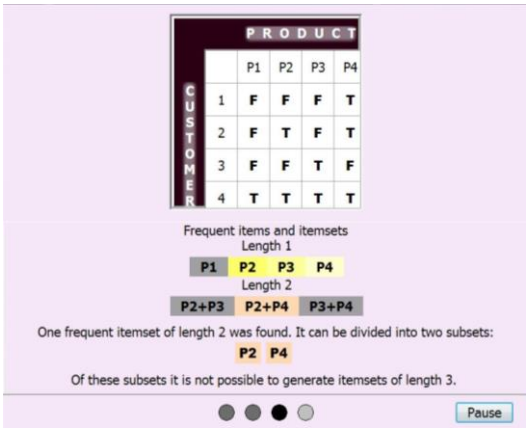
Ukázky z části programu *„How to extract association rules – APRIORI algorithm“* vidíme na následujících obrázcích:



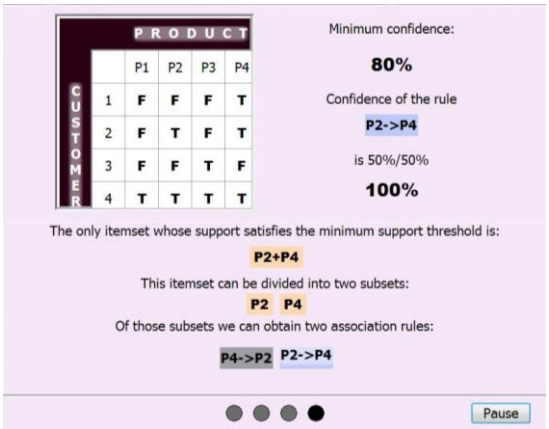
Obrázek 47. „How to extract association rules - APRIORI algorithm“ – popis



Obrázek 48. „How to extract association rules - APRIORI algorithm“ – frekventované položky



Obrázek 49. „How to extract association rules - APRIORI algorithm“ - frekventované množiny



Obrázek 50. „How to extract association rules - APRIORI algorithm“ – asociační pravidla

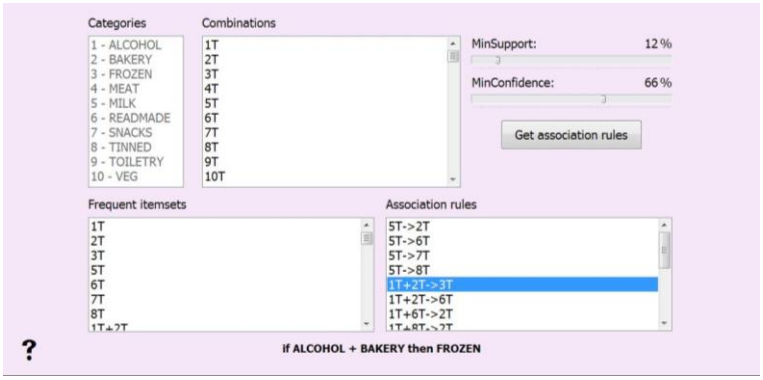
11.2.10 Tlačítko „*Algorithms*“ a podokna aplikace, k nimž se vztahuje

Po kliknutí na tlačítko *„Algorithms“* (Algoritmy) se uživateli zobrazí okno s dalšími třemi tlačítky: *„Algorithm APRIORI in steps“* (Algoritmus apriori v krocích), *„Algorithm APRIORI“* (Algoritmus apriori) a *„Algorithm APRIORI - extended“* (Algoritmus apriori - rozšíření).

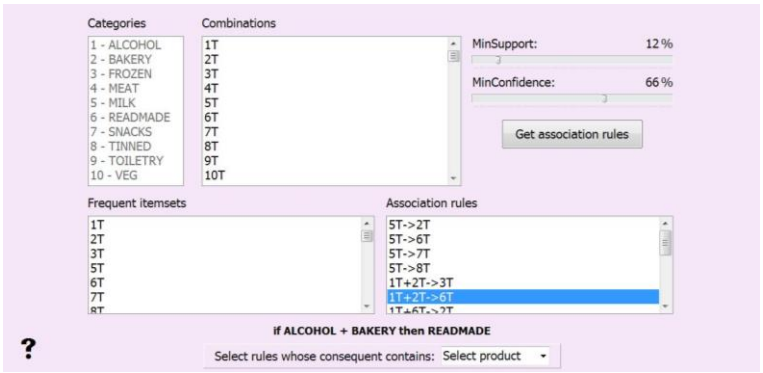
Okno *„Algorithms“* (Algoritmy) vypadá takto:

Obrázek 51. „Algorithms“

Okna *„Algorithm APRIORI“* (Algoritmus apriori) a *„Algorithm APRIORI - extended“* (Algoritmus apriori - rozšíření) vidíme na následujících obrázcích:



Obrázek 52. „Algorithm APRIORI“



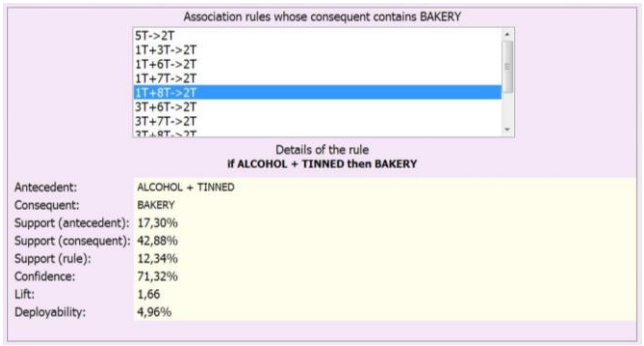
Obrázek 53. „Algorithm APRIORI - extended“

„Algorithm APRIORI“ umožňuje uživateli vyhledávat v datech asociační pravidla ve tvaru: Jestliže platí předpoklad A, pak platí i závěr B (v případě analýzy nákupního košíku tedy: Jestliže si zákazník koupí zboží A, pravděpodobně si koupí i zboží B) s použitím algoritmu apriori, který je k tomuto účelu vhodný, neboť je relativně snadno implementovatelný a naprogramovatelný a výrazně urychluje proces získávání pravidel. Princip fungování algoritmu apriori je blíže popsán v kapitole 4.2.3.

Uživatel zadává dvě kritéria: požadovanou minimální podporu a minimální spolehlivost. Program následně vyhledá všechny frekventované množiny – kombinace zakoupeného zboží, které splňují požadavek na minimální podporu, a následně všechna asociační pravidla, která odpovídají kritériu požadované minimální spolehlivosti. Výsledky jsou zobrazeny v dolní polovině okna: v levé části frekventované množiny, v pravé části pak asociační pravidla.

„Algorithm APRIORI - extended“ podobně jako *„Algorithm APRIORI“* umožňuje uživateli vyhledávat v datech asociační pravidla pomocí algoritmu apriori, přičemž tato pravidla musí splňovat kritérium požadované minimální podpory a minimální spolehlivosti, a opět zobrazuje jak nalezené frekventované množiny (splňují požadavek minimální podpory), tak všechna získaná asociační pravidla, jejichž spolehlivost je větší než požadovaná minimální spolehlivost. Tato část aplikace navíc, jak je patrné z názvu, nabízí i možnost hlubšího zkoumání asociačních pravidel a získávání více informací o nich, dále umožňuje jejich následné třídění, redukci a výběr pravidel se zaměřením na konkrétní položky v nich obsažené.

Informace o konkrétních asociačních pravidlech se zobrazují v okně, které vypadá následovně:



Obrázek 54. Detaily o vybraných asociačních pravidlech

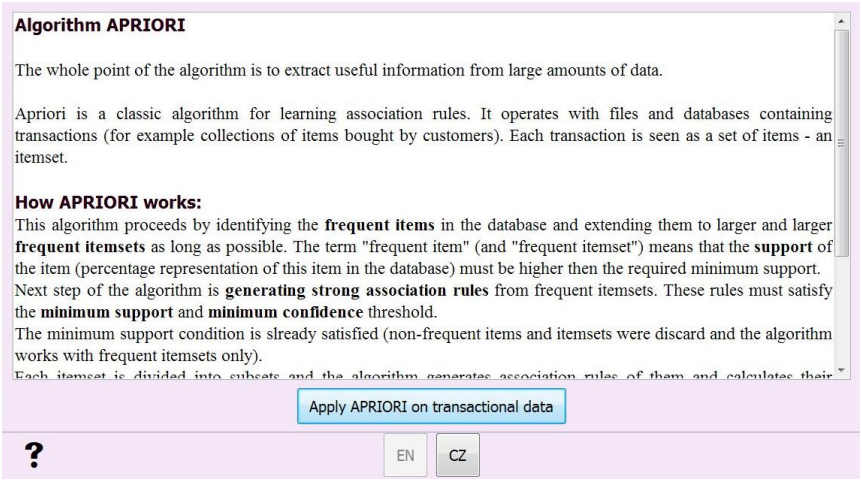
Toto okno se uživateli zobrazí poté, co zadá kritéria požadované minimální podpory a minimální spolehlivosti, na základě kterých získá asociační pravidla, a po

výběru konkrétních pravidel ze seznamu umístěném v dolní části okna *„Algorithm APRIORI – extended“* (uživatel vybírá pravidla podle druhu zboží, které je obsaženo v závěru pravidla, případně si může nechat programem zobrazit podrobné informace o všech nalezených pravidlech bez dalších omezujících kritérií).

Na obrázku 54 jsou zobrazena asociační pravidla, v jejichž závěru se vyskytovala položka *pečivo*. Kliknutím na libovolné pravidlo se lze dozvědět mnohé zajímavé informace o něm: jeho podporu (četnost, s jakou se kombinace všech položek obsažených v pravidle vyskytovala v nákupních košících zákazníků obchodu), podporu samotného předpokladu a podporu samotného závěru pravidla, spolehlivost tohoto pravidla (pravděpodobnost, s jakou platí závěr, platí-li předpoklad pravidla), lift (poměr *spolehlivost/podpora závěru*) a deployability (poměr *podpora předpokladu/podpora pravidla*).

„Algorithm APRIORI in steps“ je speciálně upravená forma okna *„Algorithm APRIORI“* nabízející všechny jeho funkce, ovšem uzpůsobené tak, aby mohl uživatel sledovat průběh algoritmu krok za krokem a dobře tak porozumět principu fungování tohoto algoritmu.

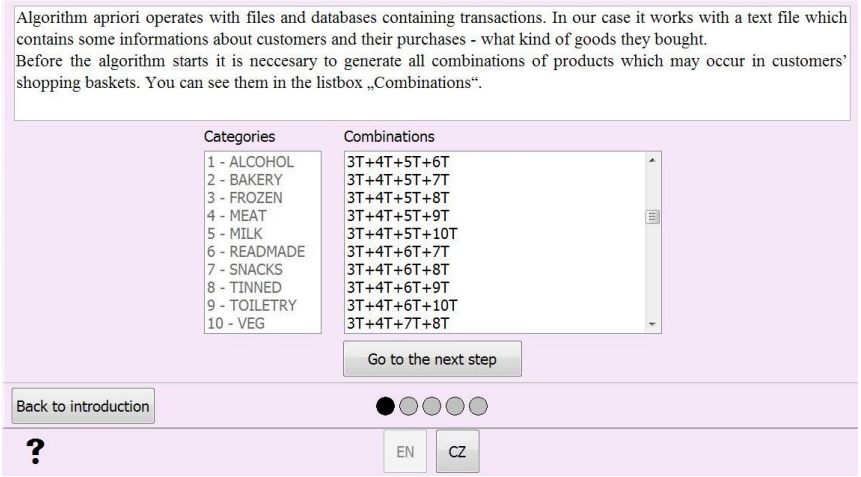
Po kliknutí na příslušné tlačítko v hlavním okně aplikace se uživateli zobrazí okno s celkovým podrobným popisem a vysvětlením, jak pracuje algoritmus apriori. Jazyk, ve kterém se text zobrazí, je možné změnit pomocí tlačítek *„EN“* a *„CZ“* a vybrat si tedy mezi angličtinou a češtinou. Toto okno vypadá takto:



Obrázek 55. „Algorithm apriori in steps" – úvodní okno

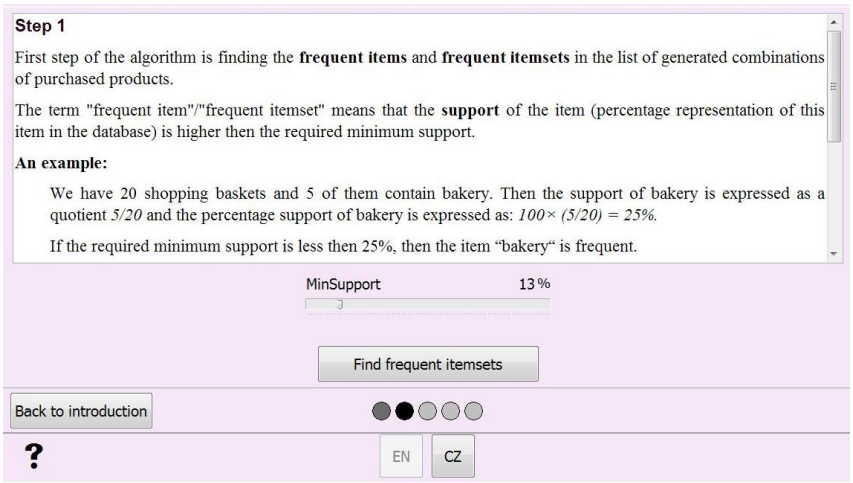
Samotný algoritmus se spustí kliknutím na tlačítko *„Apply APRIORI on transactional data“* (Aplikovat APRIORI na transakční data).

Program vygeneruje všechny možné kombinace zboží, které se mohly vyskytovat v nákupních košících zákazníků a zobrazí je v seznamu, jak vidíme na obrázku 56:



Obrázek 56. „Algorithm apriori in steps" – kombinace zboží

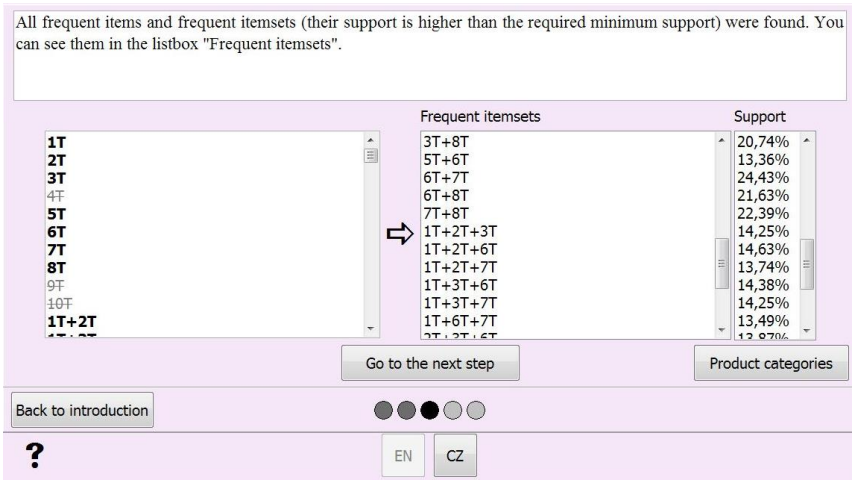
V dalším okně uživatel nastavuje požadovanou minimální podporu pro asociační pravidla. Toto okno vypadá takto:



Obrázek 57. „Algorithm apriori in steps" – nastavení požadované minimální podpory

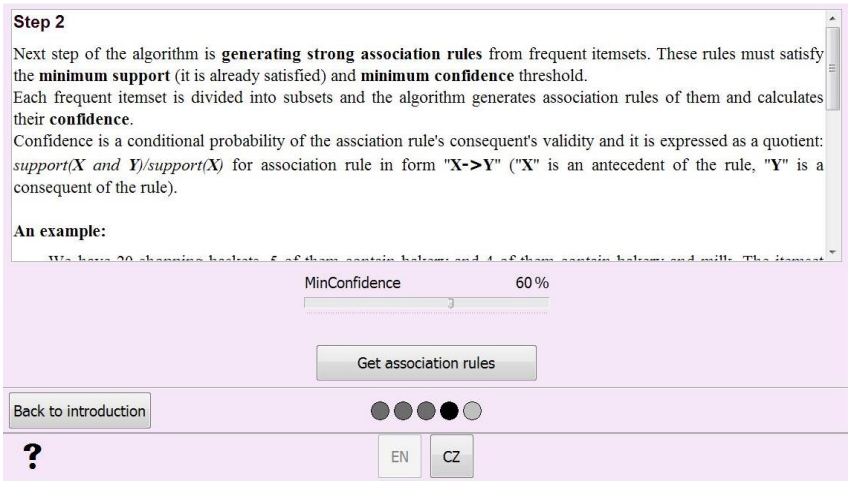
Aplikace na základě požadované minimální podpory pro asociační pravidla vyhledá všechny frekventované množiny (takové, jejichž procentuální četnost výskytu v nákupních košících zákazníků je větší nebo rovna nastavenému prahu), a to tak, že jsou nejprve identifikovány frekventované položky (množiny délky jedna) a z nich se následně vytvářejí stále delší a delší řetězce – frekventované množiny, dokud je to možné. To je princip algoritmu apriori: každá množina musí být sestavena pouze z frekventovaných položek a frekventovaných množin, položky a množiny, které nejsou frekventované, algoritmus vyřadí a dále s nimi již nepracuje.

Nalezené frekventované množiny se zobrazí v okně, jehož podobu vidíme na následujícím obrázku:



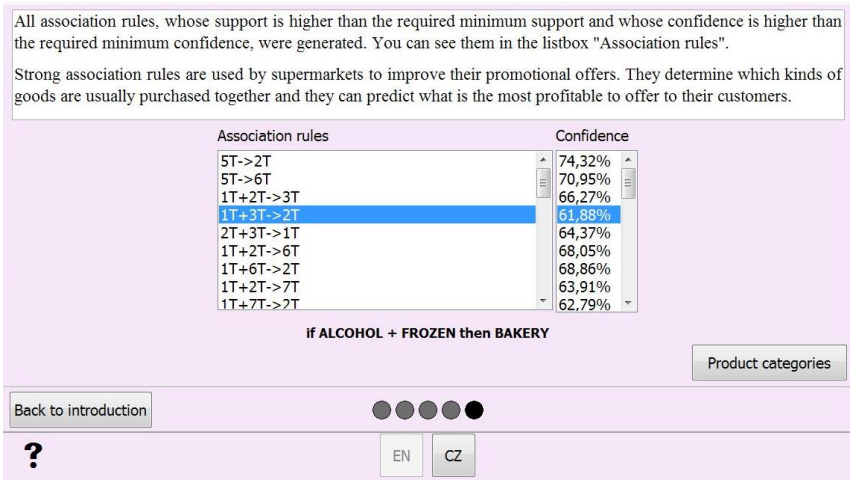
Obrázek 58. „Algorithm apriori in steps" – zobrazení frekventovaných množin

V dalším okně uživatel nastavuje požadovanou minimální spolehlivost pro asociační pravidla. Toto okno vypadá takto:



Obrázek 59. „Algorithm apriori in steps" – nastavení požadované minimální spolehlivosti

Aplikace na základě požadované minimální spolehlivosti pro asociační pravidla vygeneruje všechna tato pravidla (rozložením frekventovaných množin na podmnožiny a jejich následným slučováním a sestavováním pravidel) a výsledek zobrazí v okně, které vidíme na obrázku 60:



Obrázek 60. „Algorithm apriori in steps" – zobrazení asociačních pravidel

Každý krok algoritmu je v příslušném okně podrobně popsán a vysvětlen (včetně ukázek jednoduchých příkladů), v průběhu práce je také možné se k jednotlivým krokům, které již byly projity, vracet, a pro uživatele by tak mělo být snadné pochopit princip algoritmu apriori a jeho fungování.

Okno aplikace ***„Algorithm APRIORI in steps“*** lze popsat jako výukový program, jak pracovat s algoritmem apriori. Tato část aplikace by mohla být využívána k demonstraci algoritmu apriori během výuky data miningu a mohla by pomoci studentům snáze tomuto algoritmu porozumět.

11.3 Jazyk aplikace

Implicitním jazykem pro tento program (a původně také jediným jazykem, který měl nabízet) je angličtina – ta byla zvolena proto, že datový soubor, s nímž se pracuje, je rovněž v angličtině. Nicméně vzhledem k tomu, že aplikace je částečně výuková a měla by být přístupná a snadno ovladatelná pro co nejširší škálu uživatelů, byly nakonec některé její části (nápověda, výkladový text, vysvětlivky) přeloženy také do češtiny a mezi těmito jazyky – češtinou a angličtinou – je možné během práce s aplikací přepínat dle libosti.

11.4 Klasický a zvětšený režim zobrazení programu

Jak již bylo zmíněno několikrát výše, program nabízí uživateli dva režimy zobrazení: klasický a zvětšený, vhodný mimo jiné i pro zrakově handicapované osoby. Právě v tomto se aplikace, kterou jsem naprogramovala, odlišuje od jiných data miningových nástrojů.

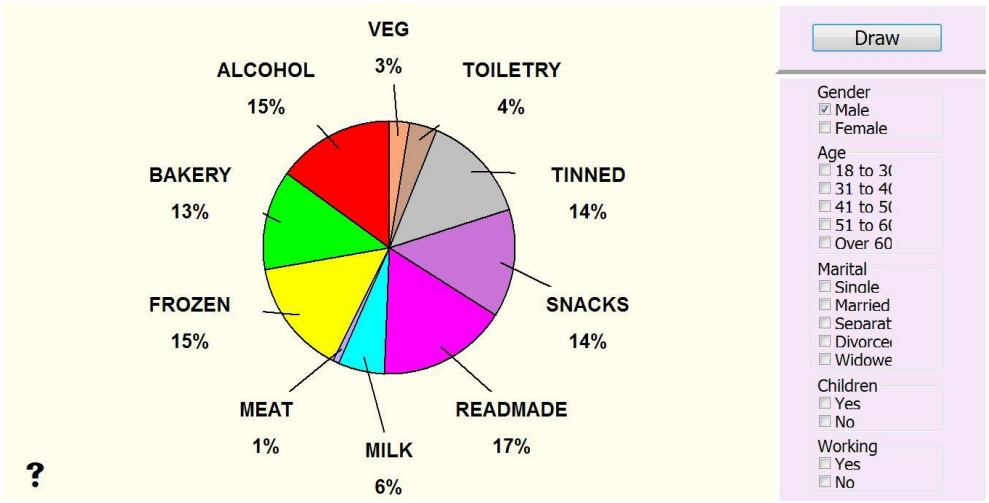
Funkce programu se pro jednotlivé režimy zobrazení nijak neliší, způsob vykreslování grafů i ovládání programu zůstávají nezměněny stejně jako struktura části programu, která se zabývá asociačními metodami, jediný rozdíl spočívá v grafické podobě oken aplikace.

Veškeré obrázky, které jsem doposud uvedla v této práci a na nichž demonstruji, jak můj program pracuje, byly pořízeny během chodu aplikace v klasickém režimu zobrazení. Font textu a jeho velikost zde byly voleny s ohledem na výslednou vizuální podobu oken a v určitých případech by tento text mohl být pro hůře vidícího člověka špatně čitelný nebo nečitelný. Program však umožňuje aktivovat zvětšený režim zobrazení (kliknutím na políčko „oko“ v pravém dolním rohu příslušného okna aplikace – každého, u něhož je tento režim zpřístupněn) a problém s čitelností snadno vyřešit.

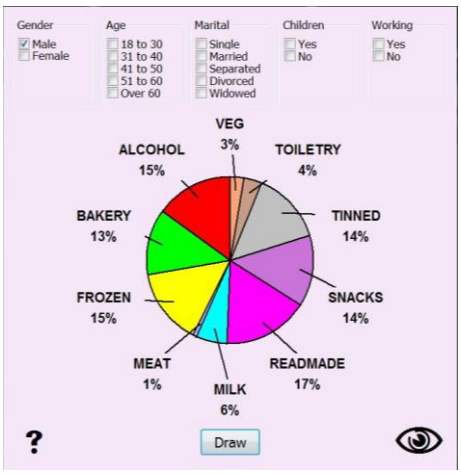
Zvětšený režim zobrazení vždy roztáhne okno přes celou obrazovku, přeskupí komponenty na něm umístěné a upraví atributy fontu písma tak, aby byl veškerý text dobře čitelný. Aktivace tohoto režimu se nijak neprojeví na funkcích programu ani na jeho ovládání.

Po ukončení práce v okně aplikace, kde byl aktivován zvětšený režim zobrazení, a jeho uzavření, se pak režim zobrazování automaticky nastaví na „klasický“.

Na obrázcích 61 a 62 uvádím příklad, jak se liší okno aplikace otevřené ve zvětšeném režimu zobrazení od okna otevřeném v režimu klasickém:



Obrázek 61. Zvětšený režim zobrazování okna



Obrázek 62. Klasický režim zobrazování okna

12 Závěr

Existuje nepřeberné množství možností, jak nahlížet na data a jak tato data zpracovávat. Vždy to závisí na tom, jaká data jsou k dispozici a jaký je základní cíl práce s daty.

Program vytvořený v rámci této bakalářské práce dokáže určitým způsobem třídit a vyhodnocovat data uložená v souboru domluveného typu a hledat mezi nimi vazby a korelace. Výsledky zprostředkovává uživateli vizuálně prostřednictvím několika typů grafů a umožňuje mu tak datům lépe porozumět. Dále umí najít složitější vztahy, které nejsou na první pohled patrné a které jsou odhalovány pomocí sofistikovanějších algoritmů – asociačních metod.

Program řeší jeden z typických data miningových problémů bez použití jiného komplexního data miningového nástroje, jehož pořízení, instalace a ovládání by mohlo představovat problém, a to zejména pro uživatele, který není primárně IT specialista a zajímá jej „jen“ posouzení transakčních dat jeho obchodu.

Aplikace je určena nejen běžným uživatelům, ale i lidem se zrakovým handicapem, což ji odlišuje od ostatních jí podobných, a činí ji v jistém smyslu výjimečnou.

Velmi důležitá je výkladová část aplikace, která má studentům oborů informačních technologií vysvětlit princip hledání asociačních pravidel algoritmem apriori. Aplikace je součástí e-learningového kurzu *Data mining*. Její vzhled a funkčnost testovalo 10 studentů navazujícího studia IT. Každý student poskytl svou zkušenost a doporučení v textové podobě. Cenné rady, byly do aplikace implementované – například grafické zobrazení pro lepší pochopení definic důležitých pojmů. Tabulka s doporučeními studentů a způsobem jejich zapracování a implementace je na v příloze této práce.

Díky své bakalářské práci jsem měla možnost blíže se seznámit s data miningem a jeho problematikou a získala jsem tak mnoho cenných a zajímavých informací z oblasti, o níž jsem dříve věděla jen velmi málo. Práce pro mě byla po všech stránkách velkým přínosem a velice mě obohatila.

Použitá literatura

[1] BERKA, Petr: Dobývání znalostí z databází, Praha: Academia, 2003, ISBN 80-200-1062-9.

[2] PLÍVA, Z., J. DRÁBKOVÁ, J. KOPRNICKÝ a L. PETRŽÍLKA: Metodika zpracování bakalářských a diplomových prací, 2. upravené vydání, Liberec: Technická univerzita v Liberci, FM, 2014, ISBN 978-80-7494-049-1.

[3] HENDL, Jan: Přehled statistických metod zpracování dat, Vyd.2. Praha: Portál, 2006, ISBN 80-7367.123.9.

[4] RUD, Olivia Parr: Datamining, Vyd.1. Praha: Computer Press, 2006, XVII, 329 s. ISBN 80-722-6577-6.

- http://www.msps.cz/data-mining/
- http://cs.wikipedia.org/wiki/Data_mining
- http://vtm.e15.cz/aktuality/data-mining-jiny-pohled-na-problem
- http://www.ceskatelevize.cz/porady/10121359557-port/tema/125-datamining-dolovani-dat/
- http://www.statsoft.cz/file1/PDF/newsletter/2014_02_26_StatSoft_Uvod_do_data_miningu.pdf
- http://si.vse.cz/archive/proceedings/2001/data-mining-v-praxi.pdf
- http://zpravy.kurzy.cz/178847-dataminingovymi-nastroji-proti-terorismu/
- http://si.vse.cz/archive/proceedings/2003/data-mining-dnesni-stav-v-cr-aktualni-novinky-a-trendy.pdf
- http://dspace.upce.cz/bitstream/10195/42661/1/PetrP_IBM_Statistics_2012.pdf
- http://www.fit.vutbr.cz/study/courses/VPD/public/0910VPD-Sebek.pdf

Přílohy

- Příloha 1: Tabulka s doporučeními studentů a způsobem jejich zapracování a implementace do aplikace
- Příloha 2: Obsah přiloženého CD

Příloha 1: Doporučení studentů a způsob jejich implementace do aplikace			
Kategorie	Název	Popis problému	Řešení problému
Algoritmus apriori	Popis algoritmu	Lépe popsat princip algoritmu apriori - jak pracuje + vysvětlit apriori vlastnost.	Popsáno detailněji, doplněn vzorový příklad a vizualizace jeho řešení pomocí algoritmu apriori.
		Přidat vzorový příklad algoritmu apriori a jeho průběhu od začátku do konce.	Vzorový příklad doplněn včetně vizualizace jeho řešení.
		Zkrátit textový popis jednotlivých kroků algoritmu apriori.	Algoritmus popsán tak, aby byl dostatečně vysvětlený. Zvýrazněny důležité pojmy.
	Minimální podpora a spolehlivost	Přidat doporučení, v jakém rozmezí by se měly pohybovat hodnoty minimální podpory aspolehlivosti pro algoritmus apriori.	Doporučení přidána ve výukové části a také v nápovědách.
	Frekventované množiny	U frekventovaných množin uvádět podporu.	Ve výukové části algoritmu apriori se zobrazují frekventované množiny + jejich podpora
		U asociačních pravidel uvádět spolehlivost.	Ve výukové části algoritmu apriori se zobrazují asociační pravidla + jejich spolehlivost
	Interpretace výsledků	Srozumitelněji interpretovat výsledky algoritmu apriori - asociační pravidla ne ve tvaru "1T + 2T -> 3T" ale "pokud si zákazník koupí zboží 1 a zboží 2, pravděpodobně si koupí také zboží 3".	Po kliknutí na konkrétní asociační pravidlo se zobrazí jeho slovní interpretace.
Výklad	Pojmy a vztahy	Zvýraznit důležité vztahy a pojmy ve výkladovém textu.	Důležité pojmy zvýrazněny, přidány vesvětlivky.
		Vysvětlit pojem "asociační pravidlo".	Přidána část vysvětlující pojem "asociační pravidlo" a jeho číselné charakteristiky.
		Vysvětlovat důležité pojmy nejen v nápovědě či výukových částech - přidat např. popisky, které by pojem připomněly, pokud je potřeba.	Doplněny vysvětlivky důležitých pojmů v angličtině a v češtině.

	Nápověda	V sekci "Help" popsat lépe používaný datový soubor a jeho strukturu - co který sloupec souboru představuje.	Upřesněn popis datového souboru.
	Vizualizace	Namísto textového popisu více věcí vizualizovat - doplnit ilustrace, animace, grafy; problematiku vysvětlit a demonstrovat na jednoduchých příkladech.	Přidány vizaualizace vysvětlující princip hledání asociačních pravidel bez použití algoritmu apriori a výpočet jejich číselných charakteristik. Dále přidána vizualizace vysvětlující princip hledání asociačních pravidel pomocí algoritmu apriori.
Vizuální podoba	Přehlednost	Zmenšit počet oken, aby byl program přehlednější.	Zpřehlednění programu díky přidání panelu zobrazujícím aktuálně otevřená okna.
		Přidat historii oken - kolik jich je otevřených, která jsou otevřená atd. Nebo přidat záložky.	Přidání panelu zobrazujícím aktuálně otevřená okna.
	Rozvržení	Zvětšit okna s popisy u výukové části algoritmu apriori.	Nastavení zobrazování uzpůsobeno tak, aby text a komponenty byly dostatečně velké při libovolném nastavení rozlišení obrazovky PC.
		Upravit rozvržení jednotlivých komponent na ploše.	
	Komponenty	Zvětšit některé komponenty nebo jejich části (posuvníky u progressbarů apod.)	
	Barvy	Zvolit jiné barvy pro aplikaci.	Zvoleny kontrastnější barvy.
Ovládání	Zadávání hodnot	Zjednodušení zadávání číselných hodnot - místo trackbarů použít tlačítka s předdefinovanými hodnotami nebo umožnit zadávání hodnot přímo pomocí klávesnice.	Možnost zadávání číselných hodnot je nyní možná pomocí trackbarů nebo přímo z klávesnice.
	Tlačítko ZPĚT	Přidat tlačítko ZPĚT	Namísto tlačítka ZPÉT přidán panel záložek zpřehledňující celý program.
Jazyk	Angličtina vs čeština	Přeložit větší část programu do češtiny, nejen delší výkladové části.	Přidány vysvětlivky důležitých pojmů v češtině.

Příloha 2: Obsah přiloženého CD

Text bakalářské práce

- Bakalarska_prace_2015_Marketa_Mala.doc
- Bakalarska_prace_2015_Marketa_Mala.pdf
- Doporuceni_studentu.xlsx
- Kopie zadání bakalářské práce: zadani_prace.pdf

Aplikace

- Soubory potřebné ke spuštění aplikace v adresářích A1files, HelpFiles a TextFiles
- Spustitelná aplikace: Application.exe