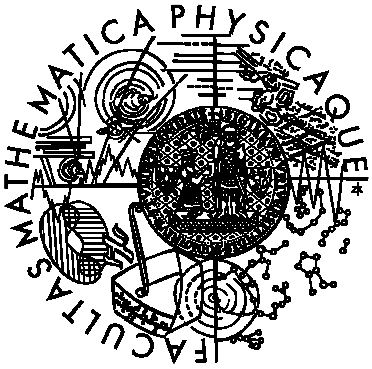
UNIVERZITA KARLOVA V PRAZE

MATEMATICKO-FYZIKÁLNÍ FAKULTA

BAKALÁŘSKÁ PRÁCE



MICHAL ZACHAR

Grafické modely v analýze diskrétních finančních dat

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce : **RNDr. Jitka Zichová, Dr.**

Studijní program : **Matematika**

Studijní obor : **Finanční matematika**

2006

Na tomto mieste by som sa chcel poďakovať predovšetkým vedúcej mojej bakalárskej práce, RNDr. Jitke Zichovej, Dr. za voľbu zaujímavej témy, cenné pripomienky a rady, ochotu ku konzultáciám a pomoc s problémami a otázkami, ktoré vznikali počas písania tejto práce, ako aj za poskytnutie potrebnej literatúry.

Ďalej by som chcel poďakovať svojím rodičom, pretože bez ich podpory počas celej doby môjho štúdia by táto práca nemohla vzniknúť.

Prehlasujem, že som svoju bakalársku prácu napísal sám a výhradne s použitím citovaných prameňov. Súhlasím so zapožičiavaním práce a jej zverejňovaním.

V PRAHE DŇA 25.5.2006 .....

Michal Zachar

Obsah

Názov práce : **Grafické modely v analýze diskrétních finančních dat**

Autor : **Michal Zachar**

Katedra : **Katedra pravděpodobnosti a matematické statistiky**

Vedúci bakalárskej práce : **RNDr. Jitka Zichová, Dr.**

E-mail vedúceho : zichova@karlin.mff.cuni.cz

Abstrakt : Táto práca sa zaoberá grafickými modelmi použitými pri analýze diskrétnych finančných dat. V praxi sa táto metóda uplatňuje pri tzv. credit scoringu, hodnotení bonity žiadateľov o úver. Práca popisuje uvedenú metódu teoreticky, v druhej časti ju potom aplikujeme na konkrétne napozorované údaje. Na priloženej diskete sa nachádza program *Selekcia.nb* spolu s databázou údajov, na ktorých zvolenú metódu testujem. Na analýzu som použil matematický program Mathematica 4.1.

Kľúčové slová : credit scoring, IPF algoritmus, grafický model, deviancia

Title : **Graphical Models in Discrete Financial Data Analysis**

Author : **Michal Zachar**

Department : **Department of probability and mathematical statistics**

Supervisor : **RNDr. Jitka Zichová, Dr.**

Supervisor's e-mail address : zichova@karlin.mff.cuni.cz

Abstract : This bachelor thesis is dedicated to the selection of graphical models for analysis of discrete financial data. Method can be used for example in "credit scoring", to rate credibility of credit applicants. There is theoretical description of this technique, as well as application on real data. We can find program *Selekcia.nb* for analysis and also tested database file on enclosed floppy disc. For my work, I've been using mathematical software Mathematica 4.1.

Keywords : credit scoring, IPF algorithm, graphical model, deviance

Kapitola 1
Úvod

V poslednej dobe rapídne vzrástol počet žiadateľov o úvery. A tento fakt sa odzrkadlil najmä v nutnosti bánk a iných finančných spoločností klásť väčší dôraz na individuálne posudzovanie dôveryhodnosti týchto žiadateľov a tým chrániť seba aj svojich klientov. Je v záujme týchto organizácií svoje metódy stále zdokonaľovať, chrániť tým samých seba a taktiež byť vo výhode oproti konkurencii na trhu. Na tento účel sa dá použiť mnoho štatistických metód. Jedným z príkladov je tzv. *Credit Scoring*, metóda, ktorá rozdelí klientov do dvoch rizikových skupín podľa bonity – schopnosti splácať poskytnutý úver.

Moja práca je zameraná na jeden prístup z tejto oblasti, analýzu aplikáciou grafických modelov na už známe údaje. Grafické modely sú vhodnou voľbou pokiaľ chceme vyšetriť vzájomnú závislosť veľkého počtu premenných alebo zistiť, do akej miery sú jednotlivé premenné ovplyvňované tými ostatnými. Môžeme tiež napríklad určiť charakteristiku jednotlivcov z dostupných údajov, či naopak predpovedať pravdepodobné chovanie klientov v budúcnosti.

Cieľom mojej práce bolo vyšetriť závislosti medzi konkrétnymi informáciami o klientoch, z ktorých som sa hlavne sústredil na skutočnosť, či bolo v minulosti klientovej žiadosti vyhovené alebo nie. Samotný graf v našom prípade reprezentuje databáza sledovaných znakov u klientov konkrétnej banky. Jednotlivé vrcholy grafu predstavujú pozorované vlastnosti žiadateľov. Existencia hrany predstavuje vzájomnú súvislosť medzi dvoma vrcholmi.

Druhá kapitola obsahuje stručný popis metódy *credit scoring*, ako aj vysvetlenie termínov, ktoré sa v tejto oblasti môžu vyskytnúť. Základné definície a tvrdenia danej problematiky nájdeme v kapitole tretej. Pri spracovávaní údajov používame testovú štatistiku – devianciu. Pre odhad združených pravdepodobností tzv. *IPF algoritmus* a konkrétne grafické modely generujeme *jednoduchou dvojkrokovou selekčnou metódou*. Popis týchto postupov je vo štvrtej kapitole a v jej podkapitolách. V závere práce aplikujeme teoretické algoritmy na údaje z finančnej praxe. Prezentácii výsledkov sa venuje posledná, piata kapitola.

Kapitola 2
Credit Scoring

Credit Scoring je štatistická metóda, ktorá sa používa k rozdeleniu žiadateľov o úver do dvoch rizikových skupín – "dobrej" a "zlej", podľa schopnosti splácať poskytnutý úver veriteľovi. Názov vznikol z anglického "credit", čo znamená úver a "scoring" – vyhodnocovanie. Voľný preklad by teda bol vyhodnocovanie žiadateľov o úvery.

Veriteľom rozumieme väčšinou právnickú osobu, ktorá žiadateľovi poskytne úver. Žiadateľ sa v tomto okamihu stáva *dlžníkom* a zaväzuje sa poskytnutý úver splácať podľa vopred určených a oboma stranami akceptovaných podmienok.

Úver znamená určitý finančný majetok zapožičaný klientovi, ktorý musí byť splácaný väčšinou v nejakých pravidelných splátkach. Za túto pôžičku si veriteľ účtuje tzv. *úrok*, ktorý by sme mohli chápať ako odmenu za poskytnutie úveru. Doba splácania závisí na charaktere a výške poskytnutého úveru. Zvyčajne sa počíta v mesiacoch či rokoch.

Výstupom metódy Credit Scoring je rozdelenie žiadateľov do rizikových tried podľa toho, či sú alebo nie sú schopní plniť svoje záväzky plynúce zo zmluvy o poskytnutí úveru. Ich schopnosť platiť splátky načas a vo výške, ktorá bola vopred stanovená. Odhad tohto rozdelenia sa určuje na základe informácií, ktoré klient poskytol inštitúcii pri žiadaní o úver a takisto tvorí základ kladného alebo záporného rozhodnutia o pridelení úveru. Čo najpresnejší verdikt je výhodný ako pre veriteľa, tak aj pre dlžníka. Finančná spoločnosť, ktorá úver poskytne, sa vyhne zbytočným stratám, ktoré by vznikli, pokiaľ by klient úver nesplatil. Naopak klient má istotu, že by svoj záväzok nenadhodnotil.

Postupy pre rozhodovanie o pridelení alebo nepridelení úveru sa zakladajú na skúsenostiach z predchádzajúcich rozhodnutí. Údaje z databáze banky obsahujú základné informácie o klientoch ako aj informáciu, či bol klient v minulosti klasifikovaný ako *dobrý* alebo *zlý*.

Inštitúcia sa však nie stále riadi výsledkom tohto rozdelenia. Napríklad sa môže stať, že banka už v minulosti konkrétnemu klientovi poskytla úver a ten sa po čase ukázal ako zlý, to znamená, že nebol schopný tento úver splácať. Napriek tomu

však nová analýza údajov zaradila tohto klienta medzi dobrých. Banke ide pri poskytovaní úverov hlavne o zisk. Z tohto pohľadu je napríklad pre ňu ziskovejší klient z rizikovejšej triedy, ktorý síce splatí úver po termíne, no za to mu môže banka naúčtovať vyššiu úrokovú mieru, poprípade aj penále z omeškania, ako klient, ktorý pravidelne spláca úver každý mesiac a splnenie záväzku dodrží načas.

Pre analýzu dát by bolo najlepšie, aby sme mali k dispozícii čo najviac napozorovaných údajov. Čím by ich bolo viac, tým by bol výsledok vierohodnejší a presvedčivejší. No so vzrastajúcim objemom údajov by banke vzrástli náklady a takisto klientov by mohol odradiť zdĺhavý a neprehľadný formulár.

Existuje teda mnoho rôznych faktorov, ktoré môžu ovplyvniť konečný výsledok credit scoringu. Záleží teda len na individuálnych preferenciách banky, ktorým z nich venuje viac pozornosti a úsilia a ktorým menej.

Kapitola 3

Základné pojmy a teória

Definícia 3.1.

Neorientovaný graf G (ďalej len *graf*) je usporiadaná dvojica (K,E), kde K je *množina vrcholov* a E je podmnožina množiny všetkých dvojprvkových podmnožín K. Prvky množiny E sa nazývajú *hrany*.

Definícia 3.2.

Vrcholy k_i a k_j sú *spojené* v grafe $G = (K,E)$, pokiaľ je medzi nimi hrana. V texte budeme túto vlastnosť označovať $\{k_i,k_j\} \in E$.

Definícia 3.3.

Majme grafy $G_1 = (K_1,E_1)$ a $G_2 = (K_2,E_2)$. Hovoríme, že graf G_1 je *podgrafom* grafu G_2, pokiaľ súčasne platí $K_1 \subseteq K_2$ a $E_1 \subseteq E_2$.

Ak naviac platí, že $K_1 = K_2$, hovoríme, že graf G_1 je *faktorom* grafu G_2.

Definícia 3.4.

Hovoríme, že graf alebo podgraf je *úplný*, pokiaľ je každý jeho vrchol spojený so všetkými ostatnými.

Definícia 3.5.

Podgraf G_a grafu G *indukovaný množinou* $a \subseteq K$ je graf, ktorý vznikol z G vylúčením všetkých vrcholov, ktoré nepatria do a spolu so všetkými hranami, ktoré spájajú iné vrcholy, než vrcholy z a.

Definícia 3.6.

Nech $a \subseteq K$. Hovoríme, že množina a je *kľuka*, pokiaľ indukuje úplný podgraf a pokiaľ po pridaní ľubovoľného vrcholu indukuje podgraf, ktorý už nie je úplný.

Poznámka 3.7.

Hovoríme, že kľuka je *maximálny úplný podgraf*.

V ďalšej časti budeme predpokladať znalosť základných pojmov zo štatistiky, akými sú napr. *náhodný vektor*, *náhodný výber*, *nezávislosť* alebo *podmienená hustota*. Ich definície sú uvedené napríklad v [1] v kapitole 2 a 3.

Definícia 3.8.

Náhodné vektory **Y**, **Z** sú *podmienene nezávislé* pri danom vektore **X**, pokiaľ podmienená hustota $f_{YZ|X}(y,z;x)$ spĺňa vzťah

$$f_{YZ|X}(y,z;x) = f_{Y|X}(y;x) \cdot f_{Z|X}(z;x)$$

pre všetky hodnoty y, z a pre všetky x také, že $f_X(x) > 0$.

V texte budeme túto skutočnosť označovať $\mathbf{Y} \perp \mathbf{Z} \mid \mathbf{X}$.

Definícia 3.9.

Graf podmienenej nezávislosti náhodného vektora **X** je graf $G = (K,E)$, kde $K = \{1,2,...,k\}$ a dvojica (i,j) nie je v množine hrán E práve vtedy, keď $\mathbf{X}_i \perp \mathbf{X}_j \mid \mathbf{X}_{K\setminus\{i,j\}}$.

Definícia 3.10.

Majme náhodný vektor $\mathbf{X} = (X_1,X_2,...,X_k)$, kde i-tá zložka môže nadobúdať hodnôt $0,1,...,r_i-1$, kde $i = 1,2,...,k$.

Pravdepodobnosti $p(\mathbf{x}) = P(X_1=x_1, X_2=x_2,..., X_k=x_k)$ je možné zoradiť do tabuľky.

Hovoríme, že k-rozmerný náhodný vektor má *rozdelenie dané tabuľkou pravdepodobností* $p(\mathbf{x})$, pokiaľ jeho hustota f je daná nenulovou tabuľkou pravdepodobností p tak, že

$$f(\mathbf{x}) = p(\mathbf{x}),$$

$$p(\mathbf{x}) > 0 \text{ pre každé } \mathbf{x},$$

$$\sum_{\mathbf{x}} p(\mathbf{x}) = 1.$$

Definícia 3.11.

Nech $\mathbf{X} = (X_1,X_2,...,X_k)$ je náhodný vektor a G graf podmienenej nezávislosti náhodného vektora $\mathbf{X}$. *Grafický model pre vektor* $\mathbf{X}$ je rodina pravdepodobnostných rozdelení vektora $\mathbf{X}$, ktoré spĺňajú podmienené nezávislosti dané grafom G.

Definícia 3.12.

Grafický model nazývame *saturovaný*, pokiaľ je jeho graf úplný.

Kapitola 4

Algoritmy použité pri analýze

4.1. Deviancia

Testová štatistika s názvom "deviancia" sa používa pri kontrole zhody napozorovaných údajov so zvoleným matematickým modelom. To znamená, že po vytvorení modelu budeme schopní určiť, či vierohodne odpovedá skutočnosti alebo musíme zvoliť iný, ktorý by ju lepšie reprezentoval. V tejto podkapitole si najskôr vysvetlíme niektoré základné pojmy, neskôr zadefinujeme samotnú devianciu.

DÔLEŽITÉ POJMY

Poznámka 4.1.1.

Označme $K = \{1,2,...,k\}$ indexovú množinu a $a,b,c \subseteq K$ jej podmnožiny. V ďalšom texte budeme používať toto označenie.

Definícia 4.1.2.

Majme náhodný výber tvorený N pozorovaniami k-rozmerného náhodného vektora $\mathbf{X} = (X_1,X_2,...,X_k)$ daného tabuľkou pravdepodobností $p(\mathbf{x})$. *Vierohodnostnú funkciu* definujeme ako súčin cez všetky realizácie náhodného vektora $\mathbf{X}$

$$f_K(\mathbf{x}) = p_K(\mathbf{x}) = \prod_i p_K(i)^{\delta(i,x)},$$

kde i prebieha $r_1 \times r_2 \times ... \times r_k$ možných hodnôt vektora $\mathbf{X}$ a δ je charakteristická funkcia definovaná vzťahom

$$\delta(i,x) = 1 \quad \text{pre } i = x,$$
$$\delta(i,x) = 0 \quad \text{inak}$$

Poznámka 4.1.3.

Uvažujme N nezávislých realizácií $x^1, x^2,...x^N$, náhodného vektora $\mathbf{X}$.

Nech $n_K(\mathbf{x})$ označuje četnosť danej realizácie x, teda

$$n_K(\mathbf{x}) = \sum_{l=1}^{N} \delta(x, x_l).$$

Marginálna tabuľka četností potom bude

$$n_a(\mathbf{x_a}) = \sum_{\mathbf{x_b}} n_K(\mathbf{x}),$$

kde $\mathbf{x_b}$ je zložka rozdeleného vektora $\mathbf{x} = (x_a, x_b)$, teda n_a je funkciou $\mathbf{x_a}$.

V ďalšom texte budeme používať nasledujúce značenie :

$p_K(\mathbf{x}) = p(\mathbf{x}) = p$, $p_a(\mathbf{x_a}) = p_a$,

$n_K(\mathbf{x}) = n(\mathbf{x}) = n$, $n_a(\mathbf{x_a}) = n_a$.

Definícia 4.1.4.

Informačnou divergenciou nazveme funkciu dvoch hustôt :

$$I(f;g) = E_f \log \frac{f(x)}{g(x)}.$$

Informačná divergencia meria “vzdialenosť” hustôt f,g .

Tvrdenie 4.1.5.

Logaritmická vierohodnosť vyjadrená pomocou informačnej divergencie je

$$l(p,n) = l(\frac{n}{N};n) - NI(\frac{n}{N};p).$$

Dôkaz :

Nájdeme v publikácii [3], str. 214 – 215.

Poznámka 4.1.6.

Maximálne vierohodným odhadom $\hat{p}(\mathbf{x})$ nazveme funkciu, ktorá maximalizuje logaritmickú vierohodnostnú funkciu $l(p,n)$.

Definícia 4.1.7.

Nech G_0 je úplný graf a G jeho faktor, v ktorom chýba f hrán. Potom *devianciu* grafického modelu M definujeme ako

$$dev^{(f)} = 2[l(\hat{p}^S;n) - l(\hat{p}^M;n)],$$

kde $\hat{p}^S$ je maximálne vierohodný odhad pravdepodobnosti $p(\mathbf{x})$ v saturovanom modeli S, určenom grafom G_0 a $\hat{p}^M$ je maximálne vierohodný odhad $p(\mathbf{x})$ v modeli M, ktorý je určený grafom G.

Poznámka 4.1.8.

Deviancia je testová štatistika pre test modelu M s grafom G proti saturovanému modelu S grafu G_0.

Definícia 4.1.9.

Nech G_0 je úplný graf, G_1 jeho faktor, v ktorom chýba f_1 hrán a ktorý má devianciu $dev^{(f_1)}$, a G_2 faktor, v ktorom chýba f_2 hrán, pričom $f_2 > f_1$, a ktorý má devianciu $dev^{(f_2)}$. Symbolicky to môžeme zapisovať $G_0 \supset G_1 \supset G_2$.

Potom definujeme *diferenciu deviancií* modelov s grafmi $G_1 \supset G_2$ predpisom

$$dev^* = -\,[dev^{(f_1)} - dev^{(f_2)}] = dev^{(f_2)} - dev^{(f_1)}$$

Poznámka 4.1.10.

Diferencia deviancií je testová štatistika pre test modelu M_2 s grafom G_2 proti modelu M_1 s grafom $G_1 \supset G_2$.

Deviancia $dev^{(f)}$ (prípadne diferencia deviancií dev^*) má za platnosti modelu M (prípadne M_2) asymptoticky χ^2_s rozdelenie, kde s je počet stupňov voľnosti, o ktorých bude pojednávané neskôr.

Veta 4.1.11.

Nech x^1, x^2,...x^N je náhodný výber z k-rozmerného rozdelenia daného tabuľkou pravdepodobností $p(\mathbf{x})$.

Potom diferenciu deviancií môžeme vyjadriť v tvare

$$dev^* = 2\sum_x n(x)\log\frac{\hat{p}_1(x)}{\hat{p}_2(x)}$$

kde $p_1(x)$ je maximálne vierohodný odhad pravdepodobností v grafickom modeli s grafom G_1 a $p_2(x)$ je maximálne vierohodný odhad pravdepodobností v grafickom modeli s grafom $G_1 \supset G_2$.

Dôkaz :

Nájdeme v [4], str. 21.

Definícia 4.1.12.

Logaritmicko-lineárny rozvoj hustoty f_K pre rozdelenie dané k-rozmernou tabuľkou pravdepodobností je

$$\log f_K(x) = \sum_{a \subseteq K} u_a(x_a),$$

kde sčítame cez všetky možné podmnožiny a množiny K a kde u-členy u_a sú súradnicovou projekciou funkcií tak, že $u_a(x) = u_a(x_a)$ a spĺňajú obmedzenie $u_a(x) = 0$, pokiaľ $x_i = 0$ pre $i \in a$.

Definícia 4.1.13.

Náhodná veličina X má *Bernoulliho rozdelenie*, pokiaľ nadobúda len hodnotu 1 a to s pravdepodobnosťou $1–p$ a hodnotu 0 s pravdepodobnosťou p.

Hustota Bernoulliho rozdelenia je definovaná vzťahom

$$f_x(x) = p^x(1–p)^{1–x}, \text{ kde } x = 0,1 \text{ a } 0 < p < 1.$$

Pre hustotu *dvojrozmerného Bernoulliho rozdelenia* potom platí

$$f_{12}(x_1,x_2) = p_{12}(x_1,x_2), \text{ pre } x_1 \in \{0,1\} \text{ a } x_2 \in \{0,1\}.$$

Príklad 4.1.14.

Koeficienty u_0, u_1, u_2, u_{12} v *logaritmicko-lineárnom rozvoji dvojrozmerného Bernoulliho rozdelenia*, ktorý je charakterizovaný logaritmom hustoty

$$\log f_{12}(x_1,x_2) = u_0 + x_1u_1 + x_2u_2 + x_1x_2u_{12},$$

sa nazývajú u-členy.

URČENIE STUPŇOV VOĽNOSTI

Počet stupňov voľnosti u diskrétnych dát je odvodený od u-členov v logaritmicko-lineárnom rozvoji hustoty Bernoulliho rozdelenia. Člen $u_a(x_a)$ je charakterizovaný množinou indexov a, a zložkami vektora x_a. Tento u-člen môžeme považovať za nulový, pokiaľ má vektor x_a niektorú zložku nulovú (podľa definície 4.1.12.) a taktiež pokiaľ obsahuje vo svojej indexovej množine a chýbajúcu hranu v grafe. A práve táto vlastnosť je pre nás dôležitá, pretože podľa [2], str. 86, je počet stupňov voľnosti rovný počtu u-členov, ktorých vektor x_a neobsahuje nuly, no napriek tomu sú tieto členy nulové a to z dôvodu, že vo svojej indexovej množine obsahujú hrany, ktoré v príslušnom grafe chýbajú.

Tvrdenie 4.2.1.

Deviancia v grafickom modeli M, ktorý odpovedá prípadu $\mathbf{X}_b \perp \mathbf{X}_c \mid \mathbf{X}_a$, je

$$dev(\mathbf{X}_b \perp \mathbf{X}_c \mid \mathbf{X}_a) = 2 \sum n_{abc} \log \frac{n_{abc} n_a}{n_{ab} n_{ac}},$$

kde sčítame cez všetky prvky tabuľky četností.

Pokiaľ označíme počet prvkov v marginálnej tabuľke danej $\mathbf{X}_a,\mathbf{X}_b,\mathbf{X}_c$ ako r_a,r_b,r_c, potom počet stupňov voľnosti pre devianciu grafického modelu M je

$$r_a\,(r_b - 1)\,(r_c - 1).$$

Dôkaz :

Nájdeme v [4], str. 24.

Poznámka 4.2.2.

Označme p,q,r počty prvkov množín $a,b,c \subseteq K$. Pokiaľ sú všetky premenné binárne, je počet stupňov voľností $2^p(2^q-1)(2^r-1)$ a odpovedá počtu u-členov, ktorých vektor x_a neobsahuje nuly a ktoré v indexovej množine obsahujú aspoň jeden prvok z množiny b a jeden z množiny c.

4.2. IPF algoritmus

Pre samotný výpočet deviancie potrebujeme odhadnúť hodnoty neznámych parametrov, čiže hodnoty tabuľky pravdepodobností. Marginálne pravdepodobnosti, ktoré odpovedajú kľukám, vieme odhadnúť pomocou relatívnych četností. Pre odhad združených pravdepodobností použijeme tzv. *IPF algoritmus*. Jeho názvom je anglická skratka výrazu Iterative Proportional Fitting. Je to iteračný algoritmus a jeho detailnejší popis nájdeme v [3] na stranách 112 – 117.

STRUČNÝ POPIS ALGORITMU

Označme f rozdelenie náhodného vektora $\mathbf{X} = (X_1,X_2,...,X_k)$ so známymi marginálnymi rozdeleniami na množinách $a_1,a_2,...,a_m \subseteq K = \{1,2,...,k\}$. Tieto množiny nemusia byť disjunktné, no žiadna nesmie byť časťou niektorej inej. Naviac musia spĺňať podmienku

$$\bigcup_{i=1}^{m} a_i = K.$$

Tieto vlastnosti spĺňajú kľuky daného grafu, preto ich aj budeme v ďalšom postupe voliť za tieto podmnožiny.

Označme g^0 počiatočné rozdelenie náhodného vektora **X.** Cieľom tohto algoritmu je potom nájsť rozdelenie g^{∞}, ktoré bude mať rovnakú štruktúru ako g^0 a marginálne rozdelenia ako *f*. Táto podmienka by potom znamenala, že informačná divergencia $I(f;g^{\infty})$ je minimálna, to znamená, že logaritmická vierohodnostná funkcia $l(g^{\infty})$ je maximálna a preto pravdepodobnosti $p^{\infty}(\mathbf{x})$ sú maximálne vierohodné odhady pravdepodobností $p(\mathbf{x})$.

4.3. Metóda generovania grafov

Na záver ešte potrebujeme nájsť spôsob, akým by sme vybrali vhodný graf, ktorý by verne reprezentoval rozdelenie napozorovaných údajov. Jednou z možností by bolo generovať všetky možné grafy a postupne u nich testovať zhodu s dátami testovou štatistikou, napríklad vyššie spomínanou devianciou. Avšak už pri 7 vrcholoch by sme dostali 2 097 152 grafov ! Okrem tohto pomerne vysokého počtu by celý výpočet trval hodiny a namiesto jedného vhodného grafu, ktorý by vyhovoval našim údajom, by sme dostali celú radu.

Pre naše potreby zvolíme spôsob, ktorý sa nazýva *jednoduchá dvojkroková selekčná metóda*. Jej detailný popis nájdeme v [3]. Vo výpočtoch vychádzame na začiatku z úplného grafu G^*. V prvom kroku spočítame pre každú hranu *devianciu vynechanej hrany*. Pod týmto pojmom tu rozumieme diferenciu deviancií dvoch grafických modelov G_1 a G_2 (ktoré sú faktormi úplného grafu G^*). Spočítame ju predpisom

$$dev^* = dev^{(f_2)} - dev^{(f_1)},$$

pričom $f_2 = f_1 + 1$ a teda $G_2 = G_1 \backslash \{i,j\}$. Pokiaľ je táto hodnota menšia než kritická hodnota χ^2 s príslušným stupňom voľnosti na zvolenej hladine, túto hranu vynecháme a nový graf označíme G'. V ďalšom kroku spočítame pre každú vyradenú hranu *devianciu pridanej hrany* podľa vzorca

$$dev^* = dev^{(f_1)} - dev^{(f_2)},$$

kde $f_2 = f_1 - 1$ sú počty chýbajúcich hrán a $G_2 = G_1 \cup \{i,j\}$. Pokiaľ bude táto hodnota vyššia než kritická hodnota χ^2, hranu vrátime späť do grafu. Týmto postupom získavame výsledný graf *G*.

Kapitola 5

Aplikácia na napozorované údaje

Táto kapitola je venovaná praktickému využitiu vyššie uvedených postupov a algoritmov pri metóde credit scoringu. K analýze sme využili program *Selekce.nb* naprogramovaný v matematickom softwari Mathematica 4.1. Jeho detailný popis nájdeme v [4]. Naša upravená verzia sa nachádza v súbore *Selekcia.nb*.

DATABÁZA ÚDAJOV

Pre vytvorenie modelu sme použili dáta z archívu Univerzity Mníchov [5]. Databáza obsahuje údaje o 1000 klientoch jednej nemeckej banky. O každom z nich máme k dispozícii celkovo 21 údajov, budeme však používať len tieto :

PREMENNÁ	POPIS	HODNOTA	OZNAČENIE
1.uver	bonita klienta	úver bol poskytnutý úver nebol poskytnutý	*0* *1*
2.ucet	zostatok na bankovom účte klienta	...≤ 0DM 0DM ≤...≤ 200DM 200DM ≤...	*0* *1* *2*
4.doveryhodnost	predchádzajúce splatenie úveru	s problémami žiadne minulé úvery všetko splatené načas	*0* *1* *2*
6.vyska_uveru	výška klientom žiadaného úveru	0DM ≤...≤ 2500DM 2500DM ≤...≤ 5000DM 5000DM ≤...	*0* *1* *2*
10.pohlavie	pohlavie klienta	muž žena	*0* *1*
12.byvanie	doba bývania na súčasnom mieste	...≤ 1 1 ≤...≤ 4 4 ≤...	*0* *1* *2*
20.telefon	telefón	nie áno	*0* *1*

Títo klienti v minulosti žiadali o spotrebiteľský úver a bankou boli pri tom požiadaní o vyplnenie krátkeho dotazníka. Otázky v ňom sa týkali ich finančnej, sociálnej i osobnej stránky. Údaje sme upravili tak, aby každá premenná mohla dosahovať maximálne 3 hodnôt. Dosiahli sme toho zlúčením niektorých kategórií v pôvodných dátach.

Na priloženej diskete sa nachádzajú súbory :

- ***Selekcia.nb*** – upravená verzia programu *Selekce.nb*, ktorá obsahuje samotný program na analýzu údajov, vykreslenie výsledného grafu a výsledky ďalej zadaných 3 príkladov
- ***Data.xls*** – databáza v programe Microsoft Excel 2003. Súbor začína názvami premenných. Tie, ktoré sme použili, sú zvýraznené červenou farbou. V ostatných riadkoch sú samotné údaje.
- ***Data.txt*** – databáza vo formáte "txt". Tieto dáta používa program pri výpočtoch. V prvom riadku je počet premenných, v druhom ich názvy. Ďalej nasledujú zaznamenané údaje oddelené medzerami, kde každý riadok reprezentuje údaje iného klienta.

VÝSTUP PROGRAMU

Po skončení výpočtu sa detailne vypíšu všetky medzikroky, ktoré sa počas behu programu uskutočnili. Na začiatku je uvedený zoznam použitých premenných a samotný výpis je rozdelený podľa fáz jednoduchej dvojkrokovej selekčnej metódy pre generovanie grafov. Tú sme bližšie popisovali v podkapitole 4.3. V prvej fáze nájdeme informácie o vynechaní konkrétnej hrany z úplného grafu na základe jej deviancie. Vo fáze druhej už vychádzame z týchto vynechaných hrán, pričom tu máme k dispozícii deviancie pridaných hrán, príslušný počet stupňov voľnosti a kritickú hodnotu χ^2 rozdelenie, s ktorou je deviancia porovnávaná. Na základe tohto výsledku rozhodneme, či hranu vrátime alebo nevrátime späť do grafu. Na konci program vykreslí výsledný graf a uvedie jeho celkovú devianciu a stupne voľnosti. Na základe porovnania s kritickou hodnotou sa vypíše tvrdenie o tom, či graf dobre popisuje štruktúru podmienených závislostí v databáze.

Pri voľbe jednotlivých premenných zaradených do modelu sme ich počet postupne zväčšovali. Analýzu sme urobili pre tri rôzne kombinácie :

Príklad č.1 – 4 premenné :
1.uver, 4.doveryhodnost, 6.vyska_uveru, 12.byvanie

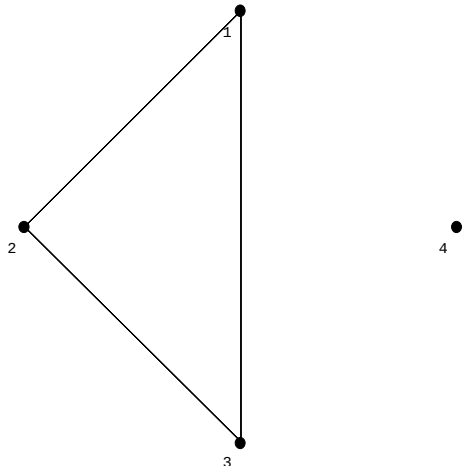
Príklad č.2 – 5 premenných :
1.uver, 2.ucet, 4.doveryhodnost, 6.vyska_uveru, 20.telefon

Príklad č.3 – 6 premenných :
1.uver, 2.ucet, 4.doveryhodnost, 10.pohlavie, 12.byvanie, 20.telefon

Pre lepšiu predstavu, uvádzame kompletný výstup programu pre príklad č. 1 :

```
Zvolene PREMENNE : {1.uver,4.doveryhodnost,6.vyska_uveru,12.byvanie}

----------------------------
1. krok selekcneho algoritmu
----------------------------

Deviancia odobranej hrany {1,2} je 67.8270
Kritická hodnota chi-kvadrat rozdelenia s 18 stupnami volnosti je 28.8693
67.8270 > 28.8693 => hranu {1,2} z grafu NEVYLUCIME

Deviancia odobranej hrany {1,3} je 31.6645
Kritická hodnota chi-kvadrat rozdelenia s 18 stupnami volnosti je 28.8693
31.6645 > 28.8693 => hranu {1,3} z grafu NEVYLUCIME

Deviancia odobranej hrany {1,4} je 15.8844
Kritická hodnota chi-kvadrat rozdelenia s 18 stupnami volnosti je 28.8693
15.8844 < 28.8693 => hranu {1,4} z grafu VYLUCIME

Deviancia odobranej hrany {2,3} je 35.3679
Kritická hodnota chi-kvadrat rozdelenia s 24 stupnami volnosti je 36.4150
35.3679 < 36.4150 => hranu {2,3} z grafu VYLUCIME

Deviancia odobranej hrany {2,4} je 24.3793
Kritická hodnota chi-kvadrat rozdelenia s 24 stupnami volnosti je 36.4150
24.3793 < 36.4150 => hranu {2,4} z grafu VYLUCIME

Deviancia odobranej hrany {3,4} je 24.3533
Kritická hodnota chi-kvadrat rozdelenia s 24 stupnami volnosti je 36.4150
24.3533 < 36.4150 => hranu {3,4} z grafu VYLUCIME

Po 1. kroku sme z grafu vylucili tieto hrany : {{1,4},{2,3},{2,4},{3,4}}

----------------------------
2. krok selekcneho algoritmu
----------------------------

Deviancia pridanej hrany {1,4} je 0.1209
Kritická hodnota chi-kvadrat rozdelenia s 2 stupnami volnosti je 5.9914
 0.1209 <  5.9914 => hranu {1,4} do grafu NEVRATIME

Deviancia pridanej hrany {2,3} je 19.5191
Kritická hodnota chi-kvadrat rozdelenia s 8 stupnami volnosti je 15.5073
19.5191 > 15.5073 => hranu {2,3} do grafu VRATIME

Deviancia pridanej hrany {2,4} je 6.6075
Kritická hodnota chi-kvadrat rozdelenia s 4 stupnami volnosti je 9.4877
 6.6075 <  9.4877 => hranu {2,4} do grafu NEVRATIME
```

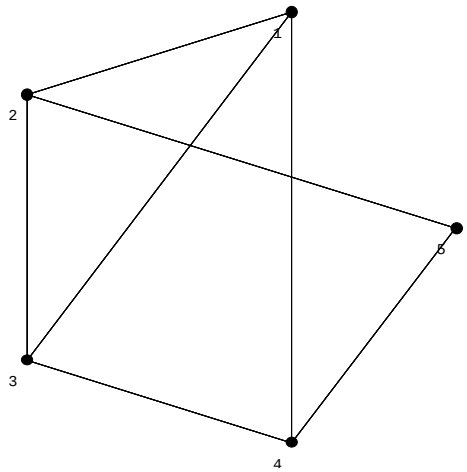
```
Deviancia pridanej hrany {3,4} je 4.8713
Kritická hodnota chi-kvadrat rozdelenia s 4 stupnami volnosti je 9.4877
 4.8713 <  9.4877 => hranu {3,4} do grafu NEVRATIME

Graf zodpovedajuci nasim udajom :
================================
```



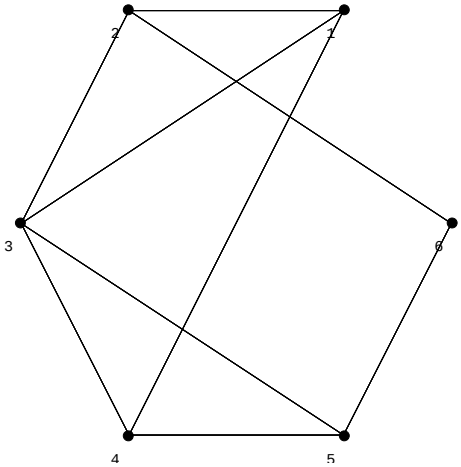

```
Deviancia vysledneho grafu G je 33.0047
Kriticka hodnota chi-kvadrat rozdelenia s 34 stupnami volnosti je 48.6024
=> NEZAMIETAME zhodu vysledneho grafu s udajmi.
```

Z tohto výsledného grafu vyplýva, že *1.uver* (vrchol č.1) závisí na premenných *4.doveryhodnost* (vrchol č.2) a *6.vyska_uveru* (vrchol č.3). Od vrcholu č.4, ktorý reprezentuje znak *12.byvanie* nevedie žiadna hrana, to znamená, že dĺžka pobytu klienta na jednom mieste nemá vplyv na jeho kredibilitu. Ukázala sa tu tiež závislosť medzi predchádzajúcim splatením úveru a výšky žiadaného úveru.

Po analýze premenných definovaných pre príklad č.2 a 3 program zvolil nasledujúce grafické modely :



Graf z príkladu č.2



Graf z príkladu č.3

Príklad č.2 – program v tomto prípade odhalil skutočnosť, že bonitu klienta ovplyvňujú premenné *2.ucet* (vrchol č.2), *4.doveryhodnost* (vrchol č.3) a *6.vyska_uveru* (vrchol č.4). Fakt, či žiadateľ vlastní alebo nevlastní telefón – premenná *20.telefon* (vrchol č.5) nemá nič spoločné s jeho schopnosťou splácať poskytnutý bankový úver.

Príklad č.3 – v našom poslednom príklade sa potvrdili výsledky z tých predošlých. Hlavne to, že aj tu je prvý znak – *1.uver* (vrchol č.1), ovplyvňovaný zostatkom na bankovom účte klienta a skúsenosťami z pridelenia úveru v minulosti. K nášmu prekvapeniu nám tu vyšla závislosť na pohlaví klienta. To môžeme napríklad prisúdiť skutočnosti, že ak o úver žiada žena s nižším príjmom, jej predpoklad splatenia tohto úveru je horší. Absencia prepojenia s informáciou o bývaní a vlastníctve telefónu sa potvrdila.

Záver

Táto práca si kládla za cieľ prezentovať aplikáciu grafických modelov na konkrétnu situáciu z finančnej praxe. Z využitím poznatkov z teórie grafov a matematického softwaru Mathematica 4.1 sa podarilo zistiť závislosti medzi viacerými dostupnými informáciami o klientoch – žiadateľoch o bankový úver.

Samotný program by bolo teoreticky možné použiť na viac znakov naraz, avšak s pribúdajúcim počtom vrcholov grafu by sa zvyšovala časová a programová náročnosť výpočtu. Naša metóda obecne zistila závislosť bonity klienta na jeho aktuálnej finančnej situácii – výške zostatku na jeho bankovom účte, veľkosti úveru či skúseností so splácaním z minulosti. Naopak ako nevýznamné znaky sa ukázali napríklad jeho aktuálna dĺžka bývania na rovnakom mieste, či vlastníctvo telefónu.

Na základe týchto pozorovaní môžeme konštatovať, že banka nemusí uchovávať všetky údaje o klientoch a takisto by mohla zjednodušiť dotazník, ktorý by bol takto príjemnejší pre jej klientov.

Literatúra

[1] Anděl, J. : *Základy matematické statistiky*.
Matfyzpress, Praha, 2005

[2] Andersen, E. B. : *Introduction to the Statistical Analysis of Categorical Data*.
Springer, Berlin, 1997

[3] Whittaker, J. : *Graphical Models in Applied Multivariate Statistics*.
Wiley, New York, 1990

[4] Svobodová, B. : *Analýza kategoriálních finančních dat.*
Diplomová práce, MFF UK, 2003

[5] Archív Univerzity Mníchov : http://www.statistik.lmu.de/service/datenarchiv