BAYESIAN METHODS FOR DATA MINING

Martin Mézl
Master Degree Programme (2), FEEC BUT
E-mail: xmezlm00@stud.feec.vutbr.cz

Supervised by: Jiří Sekora
E-mail: sekora@feec.vutbr.cz

ABSTRACT

This paper deals with Bayesian methods used in the data mining process. Part of the work consists of a theoretical overview of two selected methods i.e. maximum likehood classifier and naive Bayesian classifier both of which are also being used for classification. Finally, these methods have been programmed in Matlab and will be used for analysis of a particular database.

1 ÚVOD

V současné době je data mining jedním z nejmocnějších nástrojů pro analýzu dat. Důvodem je existence rozsáhlých databází, které shromažďují mnoho různých údajů. Data mining je soubor metod určených k analýze velkých datových souborů a právě kvůli velikosti zpracovávaných dat se data mining postupem času vyčlenil z vědního oboru statistiky. Cíle data miningu jsou různorodé, ale můžeme definovat dvě základní úlohy: klasifikaci a predikci. Klasifikace se snaží najít určité rysy chování dat a vyvodit obecné závěry. Predikce se užívá pro odhad budoucích hodnot ze znalosti hodnot předešlých a rysů systému. Existuje celá řada data miningových technik, např. rozhodovací stromy, využití umělých neuronových sítí, učení založené na instancích, Bayesovské metody, aj. Tento článek se zaměřuje na Bayesovské metody používané v data miningu a jejich aplikaci na konkrétní databázi.

2 ROZBOR

Bayesovské metody používané v data miningu vycházejí z Bayesovy věty o podmíněné pravděpodobnosti. Přestože se jedná o pravděpodobnostní metody, jsou díky svým velice dobrým výsledkům a jednoduché algoritmizaci zkoumány a používány ve strojovém učení, především pro klasifikaci, ale v některých aplikacích i pro predikci. V současné době jsou prezentovány studie, kdy se Bayesovské metody kombinují s rozhodovacími stromy nebo genetickými algoritmy pro zlepšení výsledků daných metod. Bayesův vztah (1) slouží pro výpočet aposteriorní pravděpodonosti P(H|E), tedy podmíněné pravděpodobnosti, že platí hypotéza H při pozorování evidence E. Vychází z apriorní pravděpodobnosti hypotézy P(H), pravděpodobnosti výskytu evidence P(E) a podmíněné pravděpodobnosti P(E|H), která popisuje pozorování evidence E

v případě, že platí hypotéza H.[1, 3]

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{1}$$

2.1 METODA NEJVĚTŠÍ VĚROHODNOSTI

V reálných případech klasifikace se dostáváme do situace, kdy máme více hypotéz v prostoru hypotéz T a rozhodujeme, která je pro danou evidenci nejpravděpodobnější. V tomto případě přechází jmenovatel Bayesova vztahu do tvaru $\sum_t P(E|H_t)P(H_t)$, který vyjadřuje úplnou pravděpodobnost evidence E pro všechny hypotézy H_t. Pro všechny hypotézy dostáváme hodnoty aposteriorní pravděpodobnosti a vybíráme hodnotu maximální. Zpravidla nás nezajímá konkrétní hodnota pravděpodobnosti, proto můžeme vztah upravit zanedbáním jmenovatele, který je pro všechny hypotézy stejný. Další úpravou je předpoklad, že všechny hypotézy jsou stejně pravděpodobné a tedy, že nezáleží na jejich pravděpodobnosti $P(H_t)$. Takto jsme obdrželi vztah (2), podle kterého určujeme hypotézu s největší věrohodností.[1, 2]

$$H_{ML} = \arg\max P(E|H_t), t \in T \tag{2}$$

2.2 NAIVNÍ BAYESOVSKÝ KLASIFIKÁTOR

Nedostatkem první uvedené metody je, že uvažuje pouze vliv jedné evidence pro posuzování pravděpodobnosti jednotlivých hypotéz. Pokud chceme sledovat vliv více evidencí, a v praxi to tak velmi často bývá, používáme rozšíření metody na Naivní bayesovský klasifikátor (NBK). Vycházíme z předpokladu, že jednotlivé evidence jsou při platnosti dané hypotézy podmíněně nezávislé, Bayesův vztah potom přechází do podoby (3).

$$P(H|E_1,...,E_K) = \frac{P(H)}{P(E_1,...,E_K)} \prod_{k=1}^{K} P(E_K|H) \tag{3}$$

Pro klasifikaci pomocí této metody budeme opět uvažovat hypotézy z prostoru hypotéz T a vybíráme hypotézu s největší aposteriorní pravděpodobností H_{MAP} podle vztahu (4).

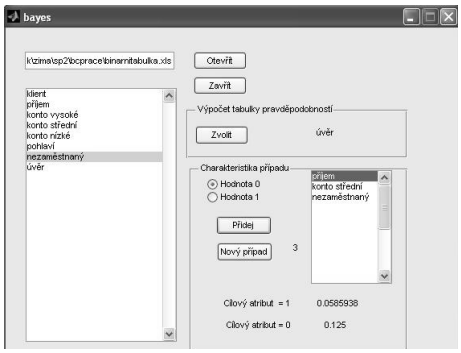
$$P(H|E_1,...,E_K) = \arg\max P(H_t) \prod_{k=1}^{K} P(E_K|H), t \in T \tag{4}$$

Všechny veličiny, které pro výpočet potřebujeme, získáme přímo z datového souboru jako pravděpodobnosti určené na základě četností výskytů jednotlivých hodnot. Na rozdíl od dalších data miningových metod, jako jsou například rozhodovací stromy nebo využití asociačních pravidel, při užití této metody neprohledáváme veškeré možné kombinace evidencí a hypotéz. Tento fakt dává menší výpočetní náročnost, která při zpracování velkých souborů hraje velkou roli.[1, 3]

Tyto dvě metody jsou základní Bayesovské metody používané v data miningových oblastech a můžeme říct, že se staly standardem. V dnešní době je publikováno mnoho metod, které jsou rozšířením těchto metod (např. semi-naivní bayesovský klasifikátor, iterativní klasifikátor, aj.). Popis těchto metod je nad rámec této práce.

2.3 ALGORITMIZACE METOD

Obě výše uvedené metody byly naprogramovány v prostředí Matlab pomocí funkcí. Pro zjednodušení jsme předpokládali binární rozdělení hodnot atributu, možné hodnoty jsou tedy z dvouprvkové množiny {0,1}. To znamená, že pokud má daná veličina spojité rozdělení, musí být hodnoty převedeny do dvou množin podle vhodného kritéria. Nejběžnějším kritériem je rozdělení pomocí hodnoty entropie[1]. Metody jsou realizovány jako funkce, které po zadání vstupních atributů vrací hodnoty aposteriorní pravděpodobnosti. Vstupními proměnnými jsou zvolené atributy (evidence) a jeden cílový atribut, který odpovídá dané hypotéze. Z matematického hlediska se určují pouze relativní četnosti a pravděpodobnosti jednotlivých evidencí. Pro metodu největší věrohodnosti obdržíme aposteriorní pravděpodobnosti pro všechny testované hypotézy a pro naivní bayesovský klasifikátor obdržíme pravděpodobnosti, které odpovídají hodnotám, zda hypotéza nastane, či nenastane. Obě funkce byly otestovány na tabulce převzaté z [1]. Výsledky odpovídají numerickému výpočtu. Tyto funkce jsou naprogramovány tak, aby je bylo možné snadno implementovat pro budoucí práci do uživatelského rozhraní GUI Matlabu. Příklad jednoduché implementace naivního bayesovského klasifikátoru je na obr. 1.



Obrázek 1: Příklad jednoduché aplikace NBK

3 ZÁVĚR

V rámci této práce byly popsány a naprogramovány dvě základní bayesovské metody používané v data miningu. Tyto metody našly velké uplatnění v dané oblasti hlavně kvůli výpočetní nenáročnosti a velmi dobrým výsledkům, které jsou v některých aplikacích srovnatelné s umělými neuronovými sítěmi. Tato práce je součástí mé diplomové práce, kde budou uvedené postupy spolu s dalšími metodami implementovány v systému, který bude sloužit pro analýzu rozsáhlých datových souborů. Tento systém bude použit pro analýzu databáze pacientů kardiologického oddělení Fakultní nemocnice Brno Bohunice se zaměřením na hledání závislostí mezi jednotlivými atributy.

REFERENCE

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

[2] DUMOUCHEL, William. *Bayesian Data Mining in Large Frequency Tables*. The American Statistician, Aug 1999(53).

[3] LAVRAČ, Nada. *Selected Techniques for Data Mining in Medicine.* Artificial Intelligence in Medicine 16(1999). Elsevier Science B.V., 1999.