Analýza dat Meningoencephalitis

Štěpán Sem

4IZ450 – Dobývání znalostí z databází

LS 2009/2010

1 Úvod

Cílem této práce je řešení vybrané úlohy z oblasti dobývání znalostí (predikce, klasifikace, deskripce) z databází nad zadanými daty (Tsumoto, 2000). Data jsem se rozhodl analyzovat v systému *Weka*[1] (jak lze vytušit z úvodního listu[2]), zadání úlohy jsem definoval na základě jednoho z typů doporučených úloh uvedených přímo v popisu analyzovaných dat.

Zpracování probíhalo volně podle kroků popsaných metodikou *CRISP-DM*.

2 Porozumění problematice

Meningitida (též zánět mozkových blan) se řadí mezi neurologické infekční choroby. Může být *bakteriálního* nebo *virového* původu – bakterie nebo viry proniknou do oblasti pod podlebicí,[3] kde způsobují bolestivý zánět. Pokud toto nastane, pacient je diagnostikován jako „meningoencephalitis“.[4] Jestliže u pacienta vznikne absces (hnisavý zánět vzniklý zničením okolní tkáně), diagnostikujeme jej jako „brain abscess“. Tento typ zdravotních problémů vyvolává převážně meningitida bakteriálního původu.

2.1 Definice úlohy

Zadání úlohy formuluji na základě nástinu *diferenční diagnózy*. Diferenční diagnózu popisují autoři takto: zjistíme počet buněk v mozkomíšním moku. Pokud převládají *vícejaderné* buňky, diagnóza zní bakteriální meningitida. Jestliže převládají buňky s *jedním jádrem*, diagnóza zní meningitida virového původu. Pro potvrzení přítomnosti mozkového abscesu se použije vyšetření pomocí počítačové tomografie (*CT* – Computed tomography).

Ve své úloze jsem stanovil minimální práh relevance rozdílu typů buněk na 5 %. Získal jsem tedy pět přípustných tříd: `BM` (meningitida bakteriálního původu), `BM-AB` (s potvrzeným mozkovým abscesem), `VM` (meningitida virového původu), `VM-AB`, `UN` (původ meningitidy nejistý, nebyl splněn požadavek minimální diference 5 %)

3 Porozumění datům

Datový soubor obsahuje 140 záznamů s 38 možnými atributy. Jednotlivé atributy popisují jednak samotné pacienty (věk, pohlaví), jednak výsledky vyšetření, dosavadní průběh onemocnění a způsob léčby.

[1] Konkrétně verze 3.7.0.

[2] Obrázek ptáka Weka převzat z `http://www.flickr.com/photos/hadevereux/2376258665/`

[3] Též dura mater – lat. tvrdá plena mozková. Zevní obal centrálního nervového systému. V lebce je pevně přimknut k lebeční kosti, v páteři je však umístěn volně a vytváří durální vak, v němž je uložena mícha. V lebce v ní probíhají žilní splavy sinus durae matris. Zdroj: `http://lekarske.slovniky.cz/pojem/dura-mater`

[4] Bohužel jsem nenalezl dostatečně výstižné české adjektivum.

3.1 Popis datových atributů

Tabulky 1-6 popisují význam atributů v datovém souboru; u některých se mi nepodařilo zjistit přesný význam.

Atribut	Interpretace
AGE	věk
SEX	pohlaví

Tabulka 1: Osobní informace

Atribut	Interpretace
DIAG	diagnóza
Diag2	odvozený z DIAG

Tabulka 2: Diagnóza

4 Příprava dat

V datovém souboru jsou data rozdělena do dvou bloků (druhý je nadepsán „New Samples“); nabízela se možnost použít druhý blok jako testová data, nicméně vzhledem k relativně nízkému počtu instancí ve druhém bloku (19) jsem oba bloky sloučil a dále s daty pracoval jako s jediným blokem.

Co se týče obsahových změn, bylo v první řadě bylo nutné sjednotit označení chybějících hodnot. V datovém souboru se totiž v jednotlivých atributech označují různě – v některém znak „-“ znamená chybějící hodnotu, v některém označení třídy („+“ a „-“). Další způsoby zahrnují mezeru, dvě po sobě ná-

Atribut	Interpretace
COLD	zimnice
HEADACHE	bolest hlavy
FEVER	horečka
NAUSEA	nevolnost
LOC	ztráta vědomí
SEIZURE	epileptický záchvat
ONSET	stav pacienta na počátku

Tabulka 3: Nedávné příznaky

Atribut	Interpretace
BT	teplota
STIFF	ztuhlost krku
KERNIG	
LASEGUE	
GCS	Glasgowská škála kómatu
LOC_DATA	odvozený – ztráta vědomí
FOCAL	

Tabulka 4: Fyzické vyšetření při přijetí

Atribut	Interpretace
WBC	množství bílých krvinek
CRP	C-Reactive protein
ESR	sedimentace
CT_FIND	odvozený – výsledky CT
EEG_WAVE	
EEG_FOCUS	
CSF_CELL	množství buněk v mozkomíšním moku
Cell_Poly	množství vícejaderných buněk
Cell_Mono	množství buněk s jedním jádrem
CSF_PRO	množství proteinu v mozkomíšním moku
CSF_GLU	množství glukózyv mozkomíšním moku
CULT_FIND	odvozený – zda jsou virus či bakterie známy
CULTURE	název viru či bakterie

Tabulka 5: Laboratorní vyšetření při přijetí

Atribut	Interpretace
THERAPY2	způspob léčby
CSF_CELL3	množství buněk v mozkomíšním moku tři dny po ošetření
CSF_CELL7	množství buněk v mozkomíšním moku sedm dní po ošetření
C_COURSE	klinické příznaky při propouštení
COURSE	odvozený z C_COURSE
RISK	rizikové faktory
RISK(Grouped)	odvozený z RISK

Tabulka 6: Terapie a průběh

sledující mezery, prázdný znak (v CSV[5] dva oddělovače polí bezprostředně za sebou). Rovněž stojí za zmínku, že některé tabulkové kalkulátory (např. *Calc*) nežádoucím způsobem „přeformátují“ otevřený soubor (znak „-“ nahradí znakem „0“; některá čísla s desetinnou tečkou interpretují jako datum...). Pro tuto činnost jsem shledal ideálním tabulkový kalkulátor *Gnumeric* (neprovádí žádné přeformátování).

V dalším kroku jsem vytvořil odvozený atribut `DIAG`, který slouží pro klasifikaci do jedné z pěti tříd popsaných výše. Dále jsem ze souboru odstranil atributy, které sloužily jako podklady pro vytvoření atributu `DIAG` nebo bylo jejich další použití z jiného důvodu nežádoucí (například atributy vzniklé seskupením z „podkladových“) – jedná se o `DIAG` (z původního datového souboru), `DIAG2`, `CT-FIND`, `CSF-CELL`, `CELL-POLY`, `CELL-MONO`, `CULT-FIND`, `CULTURE`, `CSF-CELL3`, `CSF-CELL7`, `C-COURSE`.

Takto upravený CSV soubor lze již snadno převést do formátu ARFF a dále s ním pracovat v systému Weka.

5 Modelování

5.1 Selekce atributů

K selekci atributů vhodných pro klasifikaci jsem použil *metodu filtru*, konkrétně kritérium χ^2 (*ChiSquaredAttributeEval*) s metodou *Ranker* (ohodnotí kritériem každý z atributů). Volím 5 % hladinu významnosti ($\alpha = 0,05$) a protože řeším úlohu klasifikace do 5 tříd, potřebuji znát hodnotu kvantilu (o pěti stupních volnosti).[6]

$$\chi^2_{1-\alpha}(5) = \chi^2_{0,95}(5) = 11,1$$

V dalším zpracování použiji pouze atributy, pro které platí

$$\chi^2 \geqq 11,1\ ,$$

což splňují `THERAPY2`, `RISK`, `LOC-DAT`, `CRP`, `RISK(Grouped)`, `FOCAL`, `ONSET`, `STIFF` a `SEX` (uvádím je sestupně dle hodnoty kritéria). Úlohu jsem tedy redukoval na klasifikaci do jedné z pěti tříd na základě devíti vysvětlujících atributů.

5.2 Vlastní řešení

Pro řešení klasifikační jsem se rozhodl použít *bagging* (*Bagging*), *boosting* (*AdaBoostM1*) a *kombinování modelů* (*ensemble*, *EnsembleSelection*) a porovnat účinnost jednotlivých metod. Pokud neuvedu jinak, ponechávám standardní nastavení systému. Jako dílčí klasifikátor jsem u baggingu i boostingu použil *J48*. Při kombinování modelů jsem zahrnul *naivní bayesovský klasifikátor*, *logistickou*

[5] Comma-separated values.

[6] Zdroj: STATISTIKA - TABULKY `http://statistika.vse.cz/download/materialy/tabulky.pdf`

regresi, *vícevrstvý perceptron* a *J48*. Pro vyhodnocování modelů jsem použil desetinásobnou *křížovou validaci*, srovnání úspěšnosti jednotlivých metod ilustruje následující tabulka.

Třída	Ensemble		Bagging		Boosting	
	TP	FP	TP	FP	TP	FP
BM-AB	0,667	0,074	0,667	0,057	0,5	0,066
BM	0,45	0,033	0,4	0,042	0,3	0,067
UN	0,2	0,015	0	0,007	0,2	0,015
VM-AB	0,588	0,024	0,471	0,033	0,588	0,041
VM	0,938	0,25	0,925	0,35	0,938	0,267
vážený $\bar{x}$	0,764	0,161	0,729	0,218	0,721	0,176

Tabulka 7: Srovnání klasifikačních metod

Je zajímavé, že přes rozdílný přístup jednotlivých metod vyšly relativně velmi podobné výsledky. TP a FP označují *TP Rate* a *FP Rate* z výstupu systému. Uvedl jsem tato kriteria hodnocení metod, neboť zastoupení jednotlivých tříd v datech není vyvážené.

Třída	BM-AB	BM	UN	VM-AB	VM
Zastoupení	18	20	5	17	80
Zastoupení [%]	0,13	0,14	0,04	0,12	0,57

Tabulka 8: Zastoupení tříd

Příklad stromu J48 vytvořeného během *baggingu*:

```
J48 pruned tree
------------------

THERAPY2 = multiple
|   FOCAL = -: BM (7.0/1.0)
|   FOCAL = +: BM_AB (2.0)
THERAPY2 = ABPC+CZX: BM_AB (12.0/3.0)
THERAPY2 = FMOX+AMK: BM (2.0)
THERAPY2 = ABPC: VM (3.0)
THERAPY2 = ope: BM_AB (3.0/1.0)
THERAPY2 = Dara_P: BM_AB (1.0)
THERAPY2 = ABPC+FMOX: BM (2.0)
THERAPY2 = LMOX: BM (1.0)
THERAPY2 = PCG: BM (2.0)
THERAPY2 = ABPC+LMOX: BM (3.0)
```

```
THERAPY2 = PIPC+CTX: VM (0.0)
THERAPY2 = no_therapy
|   LOC_DAT = -
|   |   CRP <= 4.5: VM (47.0/1.0)
|   |   CRP > 4.5: BM_AB (2.0)
|   LOC_DAT = +
|   |   SEX = M: UN (4.0/1.0)
|   |   SEX = F
|   |   |   STIFF <= 1: VM_AB (2.0)
|   |   |   STIFF > 1: VM (5.0)
THERAPY2 = ABPC+CTX: BM_AB (1.0)
THERAPY2 = INH+RFP: VM_AB (2.0)
THERAPY2 = ABPC+CEX: UN (2.0)
THERAPY2 = Zobirax
|   FOCAL = -
|   |   LOC_DAT = -: VM (13.0/1.0)
|   |   LOC_DAT = +: VM_AB (4.0/1.0)
|   FOCAL = +: VM_AB (5.0)
THERAPY2 = ARA_A: VM (12.0)
THERAPY2 = INH: VM (0.0)
THERAPY2 = globulin: VM (3.0)
```

Vzhledem k tomu, že vytvořené modely mají stejnou váhu hlasu a mohou jich být vytvořeny desítky, jejich interpretace může být poněkud obtížná (díky nepřehlednosti). V každém případě je k jejich správné interpretaci třeba názorů experta.

Při použití *boostingu* vypadá situace o něco lépe, protože modely („instance“ stromu *J48*) klasifikující obtížnější příklady získají hlas s vyšší vahou – za pomoci této dodatečné informace se lze v modelech o něco lépe orientovat. Následuje příklad výpisu stromu s vahou 2,23:

```
J48 pruned tree
------------------

THERAPY2 = multiple
|   SEX = M
|   |   CRP <= 0.5: VM_AB (2.85/0.33)
|   |   CRP > 0.5: BM_AB (5.58/0.98)
|   SEX = F: BM (2.46)
THERAPY2 = ABPC+CZX
|   STIFF <= 1
|   |   RISK = n: BM_AB (2.14)
|   |   RISK = LC: UN (0.0)
|   |   RISK = bechet: UN (0.0)
|   |   RISK = sinusitis: UN (0.0)
|   |   RISK = broncho: UN (2.52)
```

```
|   |   RISK = myeloma: UN (0.0)
|   |   RISK = LC_DM: UN (0.0)
|   |   RISK = DM: UN (0.0)
|   |   RISK = hepatits: UN (0.0)
|   |   RISK = TB: UN (0.0)
|   STIFF > 1
|   |   ONSET = SUBACUTE: BM_AB (2.14)
|   |   ONSET = ACUTE
|   |   |   CRP <= 4.9
|   |   |   |   CRP <= 2.4
|   |   |   |   |   FOCAL = -: BM_AB (2.79)
|   |   |   |   |   FOCAL = +
|   |   |   |   |   |   LOC_DAT = -: BM_AB (2.14)
|   |   |   |   |   |   LOC_DAT = +: BM (2.52)
|   |   |   |   CRP > 2.4: BM (2.52)
|   |   |   CRP > 4.9: BM_AB (4.6)
|   |   ONSET = CHRONIC: BM_AB (0.0)
|   |   ONSET = RECURR: BM_AB (0.0)
THERAPY2 = FMOX+AMK: BM (0.33)
THERAPY2 = ABPC
|   SEX = M: BM (2.52)
|   SEX = F: VM (4.27)
THERAPY2 = ope
|   SEX = M: BM_AB (2.14)
|   SEX = F: UN (2.52)
THERAPY2 = Dara_P: BM_AB (0.33)
THERAPY2 = ABPC+FMOX
|   STIFF <= 3: BM_AB (2.52)
|   STIFF > 3: BM (6.41)
THERAPY2 = LMOX: BM (0.33)
THERAPY2 = PCG: BM (0.33)
THERAPY2 = ABPC+LMOX: BM (0.65)
THERAPY2 = PIPC+CTX: BM (0.33)
THERAPY2 = no_therapy
|   LOC_DAT = -
|   |   ONSET = SUBACUTE: BM_AB (2.46/0.33)
|   |   ONSET = ACUTE
|   |   |   STIFF <= 0: BM (21.12/4.55)
|   |   |   STIFF > 0
|   |   |   |   CRP <= 4: VM (10.41)
|   |   |   |   CRP > 4: BM (2.52)
|   |   ONSET = CHRONIC: BM (0.0)
|   |   ONSET = RECURR: BM (0.0)
|   LOC_DAT = +
|   |   SEX = M
|   |   |   FOCAL = -: VM_AB (2.52)
```

```
|   |   |   FOCAL = +: UN (4.27)
|   |   SEX = F: VM_AB (3.82/1.3)
THERAPY2 = ABPC+CTX: BM (2.85/0.33)
THERAPY2 = INH+RFP: VM_AB (0.33)
THERAPY2 = ABPC+CEX: UN (0.33)
THERAPY2 = Zobirax
|   LOC_DAT = -: VM (12.83/3.17)
|   LOC_DAT = +
|   |   CRP <= 2.4: VM_AB (3.76)
|   |   CRP > 2.4: VM (3.17/0.65)
THERAPY2 = ARA_A
|   CRP <= 0.3: VM_AB (6.99/1.95)
|   CRP > 0.3: VM (6.41)
THERAPY2 = INH: VM (0.33)
THERAPY2 = globulin: VM (0.98)

Number of Leaves  :  52

Size of the tree :  74


Weight: 2.23
```

Ještě lépe by na tom z hlediska interpretovatelnosti mělo být použití výstupu *skládání modelů*. V systému Weka sice není vypsán konkrétní vnitřní stav (kterému z použitých modelů bychom měli „více důvěřovat“), nicméně by bylo možné systém upravit tak, aby tyto informace byly vypisovány (případně je jistě poskytují některé jiné systémy).

6 Zhodnocení (využitelnosti) výsledků

Vzhledem k relativně vysoké úspěšnosti klasifikace může tento postup (užití metod dobývání znalostí z databází) znamenat určité vodítko pro lékaře, po posouzení expertem by měla existovat nezanedbatelná šance na nalezení určitých zobecněných závislostí mezi vysvětlujícími atributy a stanovenou diagnózou (jedné z pěti tříd). Mohly by vyvstat určité pochybnosti ohledně zařazení vícevrstvého perceptronu mezi kombinaci modelů (kvůli problematické interpretaci vah uvnitř sítě), nicméně lze použít přiblížení – závislost typu $A \rightarrow C$ s vahou w (viz (Berka, 2003)), kde A značí vysvětlující atribut, C příslušnou třídu a w příspěvek (váhu) „pravidla“. Jinak lze samozřejmě použít jednu ze dvou zbylých metod, které pracují pouze se modely stromu *J48*.

Otázkou zůstává relevantnost zjištěných faktů vzhledem k nízkému počtu záznamů v datovém souboru.

Použité zdroje

CRoss Industry Standard Process for Data Mining – Process Model. Dostupné z: `http://www.crisp-dm.org/Process/index.htm`. Citováno 29.5.2010.

BERKA, P. *Dobývání znalostí z databází*. Academia, 2003. ISBN 80-200-1062-9.

TSUMOTO, S. Guide to the meningoencephalitis Diagnosis Data Set, 2000.