Učení z textů

Systémy s umělou inteligencí – úloha 3

LS 2012

1 Odevzdání a hodnocení

1. Úlohu vypracovává každý student samostatně.
2. Odevzdání má dvě části:
   (a) demonstrace experimentů, předvedení výsledků
      - na cvičení 14.5. dle rozvrhu,

   (b) odevzdání zprávy o řešení
      - mailem na klema@labe.felk.cvut.cz do 14.5. 23:59,

3. Za úlohu lze získat 15 bodů
   (a) 5 bodů za demonstraci experimentů
   (viz bod 2a, lze získat pouze na uvedeném cvičení),
   (b) 10 bodů za zprávu o řešení
   (viz bod 2b),
   (c) za každý započatý týden zpoždění odevzdání zprávy ztrácíte 3 body.

2 Zadání

Učení z textových dat, obecné zadání:
Vytvořte a testujte klasifikátory, které na základě textu diskusního příspěvku přidělí tento příspěvek do jedné z dvaceti různých diskusních skupin. Cílem je vytvořit srozumitelný klasifikátor s maximální klasifikační přesností. Důležitou součástí řešení je vhodné předzpracování vstupních textů. Kromě faktického cíle je důležité komentovat jednotlivé kroky řešení a ilustrovat tím pochopení role jednotlivých kroků učení z příkladů.

Postup řešení:

1. Seznamte se s 20_Newsgroups daty na stránce:
   http://people.csail.mit.edu/jrennie/20Newsgroups/.

2. Seznamte se s nástrojem strojového učení WEKA:
   viz http://www.cs.waikato.ac.nz/ml/weka/ a WEKA_guide.pdf.

3. K dispozici jsou 2 varianty vstupních dat v ARFF formátu, lze využít i původní textová data:

   (a) **Zjednodušená ARFF, soubor news_1000.arff.** Počet sledovaných slov je omezen na 1000, výběr je dán pouze pořadím slov. Tento soubor je zvladatelný pro většinu klasifikačních algoritmů, nelze ale dosáhnout optimální klasifikační přesnosti, řada významných slov chybí.

   (b) **Plná ARFF, soubor news_all_sparse.arff.** Každý z dokumentů je charakterizován vektorem četností téměř 54000 slov. Z důvodu toho, že většina slov se nevyskytuje v každém konkrétním dokumentu, je použita takzvaná zahuštěná reprezentace řídké matice četností. Tento soubor může být pro klasifikační algoritmy nezvladatelný, s využitím metod selekce příznaků lze ale dosáhnout rozumné klasifikační přesnosti v rozumném čase. Soubor je vytvořen z veřejně dostupného souboru http://people.csail.mit.edu/jrennie/20Newsgroups/20news-bydate-matlab.tgz.

   (c) **Původní textová, viz bod 1.** Adresář původních textů je dostupný na: http://people.csail.mit.edu/jrennie/20Newsgroups/20news-bydate.tar.gz. Lze navrhnout vlastní způsob předzpracování těchto dat s použitím stemmerů implementovaných přímo v prostředí WEKA (weka.filters.unsupervised.attribute.StringToWordVector), popřípadě uvažovat zachycení sekvenční povahy textu (nezáleží pouze na četnosti výskytu slov, ale také jejich interakci). Případně použít specializované nástroje typu Bow (http://www.cs.cmu.edu/~mccallum/bow/).

4. Pokud je počet atributů nebo počet příkladů příliš velký, použijte filtry.

(a) Pro předvýběr relevantních slov lze použít například míru weka.attributeSelection.GainRatioAttributeEval.

(b) Jaká slova se jeví jako nejdůležitější? Odpovídá to očekáváním?

(c) Instance lze filtrovat například filtrem weka.filters.unsupervised.instance.RemovePercentage (náhodné filtrování) nebo weka.filters.unsupervised.instance.RemoveWithValues (omezení na podmnožinu kategorií).

5. Na základě dat sestavte symbolický klasifikátor (rozhodovací strom – weka.classifiers.trees.J48, množina rozhodovacích pravidel – weka.classifiers.rules.DecisionTable, weka.classifiers.rules.JRip).

   (a) vytvořený model interpretujte (popište slovně), je srozumitelný?, dává smysl?, které diskusní skupiny se daří nebo naopak nedaří oddělit (vysvětlete matici záměn, popřípadě použijte F-míru)?

   (b) ověřte přesnost modelu křížovou validací,

   (c) křížovou validaci použijte i k oprimalizaci parametrů klasifikátoru nebo volbu jeho variant,,

   (d) sestavte křivku učení (zkušenost může být parametrizována rostoucícm počtem trénovacích příkladů nebo relevantních slov), lze manuálně opakovaným použitím filtru weka.filters.unsupervised.instance.RemovePercentage, návod na automatické vytvoření křivky učení je na http://weka.wikispaces.com/Learning+curves.

6. Na základě dat sestavte neuronovou síť (vícevrstvý perceptron – weka.classifiers.functions.MultilayerPerceptron).

   (a) ověřte přesnost modelu křížovou validací,

   (b) křížovou validaci použijte i k oprimalizaci parametrů klasifikátoru nebo volbu jeho variant,

   (c) srovnejte vlastnosti modelu s modelem symbolickým vytvořeným v předchozím kroku,

   (d) sestavte křivku učení (zkušenost může být parametrizována rostoucícm počtem trénovacích příkladů nebo relevantních slov).

7. Diskutujte silné a slabé stránky výše uvedených klasifikačních algoritmů.

   (a) algoritmy hodnoťte z hlediska přesnosti, srozumitelnosti, doby učení, robustnosti a požadavků na množství trénovacích dat.

8. O řešení vypracujte písemnou zprávu

   (a) zpráva bude obsahovat řešení pro body 3-7,

   (b) do zprávy zařaďte i jakákoli zajímavá pozorování nad rámec bodů 3-7, napište krátké shrnutí,

   (c) název odevzdaného ZIP archivu = vaše uživatelské jméno.