

Domácí úkol č. 1 do předmětu PV056

Prerevizity:

Nainstalovaný program WEKA 3, lze stáhnout na adrese
<http://www.cs.waikato.ac.nz/ml/weka/>

Datasey:

Datasey ke stáhnutí na <http://archive.ics.uci.edu/ml/datasets.html>. Datasey označené hvězdičkou jsou v archivu metal-data.tar ve studijních materiálech

<https://is.muni.cz/auth/el/1433/jaro2011/PV056/?fakulta=1433;obdobi=5105;predmet=585450>

Každý student má jinou sadu datasetů viz tabulka.

UČO	DATASEY
255496	Abalone, ann*
374580	Post-Operative Patient, adult*
356530	Balance Scale, allbp*
325073	Teaching Assistant Evaluation, australian*
269281	Car Evaluation, diabetes*
255821	Chess (King-Rook vs. King-Pawn), fluid*
374386	Chess (King-Rook vs. King), german_cont*
208013	SpamBase, german_f*
256726	Contraceptive Method Choice, german_n*
359312	Cylinder Bands, Statlog Project (Heart)
255756	Dermatology, segment*
255658	Echocardiogram (vymazat sloupec name), Wine (1. atribut class label)
324899	Glass Identification, letter*
172564	Zoo, yeast*
374160	Congressional Voting Records(class label je 1. atribut), optical*
348646	Ionosphere, Connectionist Bench (Sonar, Mines vs. Rocks, all data)
173390	Soybean (Small), page*
374370	Mammographic Mass, pendigits*
324545	Poker Hand (pracovat pouze s trénovacími daty), vaw40*
143357	Lung Cancer (atribut 1 je class label), vehicle*
256368	MAGIC Gamma Telescope, quisclas*
325001	Nursery, Blood Transfusion Service Center, vowel*
389924	Libras Movement (movement libras.data), satimages*
359500	Acute Inflammations, vaw21*

Zadání:

1. Stáhněte si přidělené datasey. Mají obvykle dva soubory *.names a *.data. Soubory *.data zkontrolujte, zda jsou ve formátu hodnot oddělených čárkou a třídu mají jako

poslední hodnotu, popř. je do něj převedte.

2. Vytvořte odpovídající soubory *.names, aby odpovídaly požadovanému formátu C4.5 popsanému dále. Ve staženém souboru *.names může být popis jednotlivých parametrů, v opačném případě je potřeba soubor projít a parametry popsat. Je možné také importovat soubor *.data jako *.csv do WEKY, kde na první řádek napíšete jména atributů oddělená čárkou, poslední jméno bude třída. WEKA provede konverzi. Zkontrolujte, zda datové typy jsou správně rozpoznány. I v tomto případě se očekává odevzdání souboru *.names.
3. U každého datasetu proveďte klasifikaci s defaultním nastavením parametrů, použijte cross-validation (10).
4. Po doběhnutí algoritmu uložte celý výstup do souboru dataset_algoritmus.log, kde algoritmus $\in \{NB, IB1, IB3, JRip, DS, J48, SMO, PART, MLP\}$ Povšimněte si celkové správnosti (accuracy, počet správně klasifikovaných příkladů) a času potřebného pro sestavení modelu. Vytvořte soubor results.txt, tam zaznamenejte poznámky o netypickém průběhu, např. pokud algoritmus nedoběhne a o případných dodatečných úpravách dat, jako vymazání sloupce id...
5. U algoritmu s nejvyšší celkovou správností (accuracy), při více stejných vyberte ten s nejnižším časem (je-li i čas stejný, vyberte si) a zkuste nastavit jiné vstupní parametry. Sledujte tendence vývoje celkové správnosti. Nalezněte nejlepší nastavení parametrů. Zaznamenejte postup a kombinace parametrů do souboru results. Uložte celý výstup WEKY do souboru cislo_dataset_algoritmus.log pro první tři nejlepší výsledky, kde cislo označuje pořadí.
6. Vypracovaný domácí úkol (logy, results.txt) uložte do odevzdávací skříně včetně souborů s koncovkou .names (těch, které jste sami vytvořili) do adresáře se svým jménem do **7.4.2011**.
7. Informace o splnění se objeví v poznámkovém bloku.
8. Konkrétní dotazy řešte stručným mailem na 208230@mail.muni.cz, do předmětu napište minimálně kód PV056. Případné nejasnosti obecného rázu přes diskusní fórum předmětu.

Formát C4.5

Soubor.data – co řádek to záznam, hodnoty atributů oddělené čárkou, poslední hodnota na řádku je třída. Záznam není ukončen tečkou. Každý atom (nenumernická hodnota atributu) musí být v seznamu hodnot (popisu) korespondujícího atributu. Atomy nesmí být v uvozovkách, obsahovat mezery ani jiné bílé znaky. Chybějící hodnotu vyjadřuje otazník. Záznamy s chybějící třídou nejsou povoleny.

Např.

a2,39,a4,c2

a4,30,a1,c2

a1,9,a2,c1

Soubor.names – popis atributů a jejich hodnot. První řádek obsahuje seznam možných hodnot třídy oddělených čárkou a ukončených tečkou. Tyto musí korespondovat s posledními

hodnotami na řádcích v Soubor.data. Všechny řádky obsahují popis atributů v pořadí, v jakém se nacházejí v Soubor.data. Chybějící hodnota (?) se do výčtu nezahrnuje. Popis atributu je následující:

jméno_atributu: [continuous | CSV seznam hodnot u ordinárního atributu].

Např.

c1,c2.

at1: a1,a2,a3,a4.

at2: continuous.

at3: a0,a1,a2,a3,a4.

Žádné komentáře ani prázdné řádky nejsou povoleny. Poznámky lze umístit do samostatného souboru Soubor.info.

Soubor.data, Soubor.names, popř. Soubor.info musí být ve stejném adresáři.

Poznámka 1: Při použití automatické konverze WEKY by vypadal příklad takto (uloženo jako Soubor.csv):

at1,at2,at3,class

a2,39,a4,c2

a4,30,a1,c2

a1,9,a2,c1

Poznámka 2: Obsahuje-li váš dataset sloupec unikátních hodnot (id), odstraňte ho.

algoritmy

Pouze klasifikační algoritmy, ve WECE záložka classify.

- Naive Bayes
- IB1
- IBk (pro $k = 3$, nastavit jako parametr KNN)
- JRip
- DecisionStump
- J48
- PART
- SMO
- Multilayer Perceptron