

Domácí úkol č. 2 do předmětu PV056

Prerekvizity:

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. Domácí úkol se týká předzpracování dat, viz. přednáška na toto téma.

2. Datasets upravte stejně jako v předchozím úkolu (včetně odstranění atributu id, je-li přítomen).
3. Použijte vždy metodu předzpracování (pokud je možná u daného datasetu) a spusťte klasifikaci. Metody jsou uvedeny dále plus jednu navíc si vyberte dle svého uvážení. Pokud u vašeho datasetu nelze použít žádnou z navrhovaných možností, použijte tři jiné dle vlastního výběru.
4. U každého datasetu proveďte klasifikaci s defaultním nastavením parametrů, použijte cross-validation (10).
5. Po doběhnutí algoritmu uložte celý výstup do souboru dataset_metoda_algoritmus.log. Povšimněte si celkové správnosti (accuracy, počet správně klasifikovaných příkladů) a času potřebného pro sestavení modelu. Porovnejte s výsledkem na neupravených datech. Vytvořte soubor results.txt, tam zaznamenejte celý postup a pozorování, která jste učinili. Přidejte poznámky o netypickém průběhu, např. pokud algoritmus nedoběhne...
6. Pro každý dataset a algoritmus najděte nejlepší kombinaci metod předzpracování, která zvýší accuracy.
7. Vypracovaný domácí úkol (logy, results.txt) uložte do odevzdávacího zip do **15.5.2011**.
8. O splnění budete informováni mailem.
9. Konkrétní dotazy řešte stručným mailem na 208230@mail.muni.cz, do předmětu napište minimálně kód PV056. Případné nejasnosti obecného rázu přes diskusní fórum předmětu.

algoritmy

Pouze klasifikační algoritmy, ve WECE záložka classify.

- Naive Bayes
- IBk (pro $k = 3$, nastavit jako parametr KNN)
- J48

metody předzpracování

1. diskretizace reálných hodnot (defaultní nastavení, unsupervised, supervised)
2. výběr atributů (supervised weka.attributeSelection.CfsSubsetEval BestFirst, potom zkuste sami najít nejlepší možnou podmnožinu atributů)
3. odstranění atributu s velkým počtem různých hodnot
4. odstranění atributu s hodnotou, která je stejná pro všechny řádky (nebo téměř pro všechny)
5. přidání odvozeného atributu
6. vlastní nápad :)