Domáca úloha č.2 k predmetu PV056

Prerekvizity:

Nainštalovaný program Weka 3, ktorý si môžete stiahnuť na adrese: `http://www.cs.waikato.ac.nz/ml/weka/`

Datasety:

Datasety si môžete stiahnuť na adrese: `http://archive.ics.uci.edu/ml/datasets.html`.

Každý z vás má svoju vlastnú sadu datasetov. Pridelené datasety nájdete v tabuľke:

UČO	Dataset
72665	Parkinsons Data Set
106451	Ionosphere Data Set
113869	Car Evaluation Data Set
207622	Soybean (Large) Data Set
255651	Yeast Data Set
255838	Chess (King-Rook vs. King-Pawn) Data Set
256365	Tic-Tac-Toe Endgame Data Set
256699	Breast Cancer Wisconsin (Original) Data Set
324709	Pima Indians Diabetes Data Set
324751	Wine Quality Data Set
325154	Hepatitis Data Set
333617	Lymphography Data Set
356530	Flags Data Set
359185	Heart Disease Data Set
359226	Wall-Following Robot Navigation Data
359305	Echo-cardiogram Data Set
359331	SPECT Heart Data Set
359747	Internet Advertisements Data Set
359789	Mammographic Mass Data Set
359860	Wine Data Set
374368	Dermatology Data Set
374595	Balance Scale Data Set
388433	Mushroom Data Set
395607	Blood Transfusion Service Center Data Set
395881	Libras Movement Data Set
395986	Hill-Valley Data Set
396273	Japanese Credit Screening Data Set
396477	Lung Cancer Data Set
409930	Chess (King-Rook vs. King) Data Set
410345	Census Income Data Set
417148	Horse Colic Data Set

Zadanie:

- Na analýzu použite dátové sady, ktoré ste si vytvorili v prvej úlohe, prípadne si stiahnite dataset a predspracujte ho tak, ako bolo popísané v prvej úlohe.
- V tejto úlohe by ste si mali precvičiť analýzu za pomoci Zmiešaných metód strojového učenia.
- Konkrétne sa jedná o metódy: *Bagging* a *Vote* ktoré nájdete na záložke classify, medzi *meta* klasifikátormi.
- Vašou úlohou bude za pomoci týchto klasifikátorov dosiahnuť lepšie výsledky klasifikácie ako pri prvej úlohe, alebo minimálne porovnateľné.
- Pri oboch algoritmoch nastavujete ďalšie algoritmy, ktoré vykonajú samotnú analýzu. Použite tieto algoritmy: J48, RandomForest, NaiveBayes, SMO a ďalšie ľubovolné 3.
- Pri *Baggingu* nastavujete len jeden klasifikátor, preto vykonajte analýzu na každom z vašich 7 algoritmov, ale pokúste sa nastaviť parametre baggingu (a vybraného algoritmu) tak, aby ste dosiahli ešte lepších výsledkov.
- Pri *Vote* môžete nastaviť viacero klasifikátorov. Môžete ich ľubovolne miešať, takže pri riešení je prípustná akákoľvek podmnožina vašich 7 klasifikátorov. Dôležité je, aby ste si všimli parameter "combinationRule" a pohrali sa sním tak, aby ste dostali čo najlepší výsledok. Vašou úlohou bude dosiahnuť v tomto prípade lepší výsledok (alebo aspoň porovnateľný) ako bol Váš najlepší dosiahnutý výsledok na tejto dátovej sade z minulej úlohy spomedzi všetkých použitých algoritmov. V tomto prípade mi odovzdáte prvých 7 najlepších výsledkov. Vo väčšej polovici prípadov vyžadujem, aby ste použili kombináciu aspoň 3 algoritmov.
- Dáta predspracujte klasicky, tak ako v prvej úlohe. Samozrejme, môžete sa s nimi pohrať aj viac ak to uznáte za vhodné.
- V prípade *Baggingu* mi výsledky zapíšte do tabuľky v nasledovnom formáte:

Alg.	**Acc. 1.úl.**	**Acc. 2.úl.**	**Param. baggingu**	**Param. algoritmu**	**Zlepšenie o**
J48	**xx.x**	**xx.x**	**...**	**...**	**+/-xx.x**

- V prípade *Vote* mi výsledky zapíšte do tabuľky v nasledovnom formáte (zoradené od najlepšieho po najhorší):

Algoritmy	**Alg. param.**	**Accuracy**	**Vote params**	**Best 1.úl.**	**Zlepšenie o**
Alg A Alg B Alg C ...	params A params B params C ...	**xx.x**	**...**	**...**	**+/-xx.x**
Alg B Alg E Alg A ...	params B params E params A ...	**xx.x**	**...**	**...**	**+/-xx.x**

- Vytvorte súbory unexpected-bagging.txt a unexpected-vote.txt a zaznamenajte do nich poznámky o netypickom priebehu, ak napríklad algoritmus nedobehne, alebo o prípadných dodatočných úpravách dát.
- Ak niekto chce, môže si vyskúšať túto úlohu aj naprogramovať. Weka má relatívne dobrú dokumentáciu na webe aj s praktickými ukážkami a programovanie je veľmi intuitívne a jednoduché. Osobne si myslím, že naprogramovanie tejto úlohy bude pre vás rýchlejšie ako keby ste to mali vyklikávať v GUI.

- Vypracovanú úlohu (2x tabuľka s výsledkami (bagging.pdf. vote.pdf), arff súbor s vašou dátovou sadou, 2x unexpected.txt (unexpected-bagging.txt, unexpected-vote.txt)) odovzdajte do Odevzdávárny zazipované v jednom súbore do **12.05.2012 12:00**.
- Ak sa to rozhodnete úlohu naprogramovať, zašlite mi aj zdrojové kódy vášho riešenia.
- Súbory s riešením prosím nevkladajte do žiadneho podadresára! (Povolené sú len súbory s naprogramovaným riešením aby boli v zvlášť adresári)
- Informácie o splnení úlohy vám zadám do poznámkového bloku.
- V prípade nesplnenia úlohy vám budem nútený zadať mínusové body, ktoré sa vám odpočítajú od bodov získaných v záverečnej skúške.
- Ak by ste mali nejaké nejasnosti, alebo by ste si nevedeli rady, napíšte mi stručný e-mail na 173001@mail.muni.cz a do predmetu mailu zadajte aspoň kód predmetu. Všeobecné otázky prosím riešte cez diskusné fórum.