HandsOn Assignments 2: Data Mining

A) Dataset

- Pick a dataset from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml) Requirements: min. 1000 instances, min 15 attributes, must have class labels for classification tasks
- Register the dataset you picked with your group number in the TUWEL Wiki to avoid that everyone is taking the same dataset (each dataset should be used only once, so first-come, first serve)
- Download the dataset, analyze its characteristics (size, attribute types, value ranges, sparsity, min/max values, ...), and describe this in the report.

B) Classification: Analysis of Train/Test Set Splits, Classifier Performance and Parameters

(1) Using WEKA, pick two classification algorithms that are sufficiently different (i.e. no two tree-learners, no two slight variations of the same classifier such as SVM with two different kernels). Describe why you chose the respective classifiers.

(2) Preprocess the data to have it in a form suitable for training these classifiers, and describe any preprocessing required (transcoding, scaling,...), the reasons for it, and how you performed it.

(3) If the dataset is far too large for the classifier to process, you may choose a subsampling strategy, selecting a reasonable amount of data items (describe the reasoning applied and how you did it. Consider repeatability of experiments if possible).

(4) Train these two classifiers in several versions and evaluate results using a reasonable classification quality measure and interpreting the results using both graphs summarizing the performance as well as, potentially, confusion matrices. Document the experiments, parameters, and describe/summarize the findings in the report (note: describe and analyze your findings - do not simply copy the WEKA output into the report without further intellectual analysis). For the individual experiments you should vary:

   a. Parameters: If the classifier has specific parameters, explore their effect with different settings using 10-fold cross-validation and document the results.
   b. Scaling: where possible, try different scaling approaches (min/max, zero mean/unit variance, length) using the best parameters identified above and observe the difference in classification performance using 10-fold cross-validation. Analyze the reasons for the effects observed, test useful and also non-useful scalings (!) and summarize your findings as well as analyze reasons why specific scalings make sense in a given setting.
   c. Training/test set splits: Using the most promising parameter settings and scaling identified above, evaluate the effect of different training and test set splits. Starting initially with a very small training set, increase the training set size in reasonable increments (10 different training/test set splits, e.g. from 5%/95% in 10%-increments to 95%/5% train/test) and observe changes in the performance of the classifiers depending on training set size.

(5) Provide a summary of this part, analyze the results, addressing questions such as

   - What trends do you observe in each set of experiments?
   - How easy is it to interpret the classifier and its performance?
   - Which classes are most frequently mixed-up? (and why?)
   - How does the performance increase/decrease with different parameter settings?
   - Do both classifiers show the same behavior in performance degradation / robustness against
     - different correct/incorrect/missing scalings?
     - smaller and larger variations in parameter settings?
   - Can you observe characteristics such as over-learning?
   - How does the performance change with different amounts of training data being available? What are the best scalings (per attribute/per vector) and why?

HandsOn Assignments 2: Data Mining

C) Missing Values:

(1) Write a little program/script that allows you to create new versions of the dataset, replacing x% of selected attributes by missing values (e.g. a "?" symbol). Missing values should be distributable
   a. randomly across attributes
   b. with specific percentages of missing values per attribute. (Note: make sure to keep the initialization parameter of the random number generator as a parameter to the tool to ensure repeatability of results). Briefly describe the script, configuration parameters, settings etc. in the report.

(2) Generate a small number of different datasets and describe them in your report, varying
   i. a small, medium, large fraction of missing values in 2 or 3 attributes
   ii. a small, medium, large fraction of missing values randomly distributed across all attributes

(3) Implement different strategies to deal with these missing values and describe their implementation, by
   a. ignoring the respective attributes completely in the dataset
   b. replacing the missing attribute values by the mean / median value of that attribute in the entire dataset
   c. replace the missing attribute by the mean/median value of that attribute in the respective class
   d. (for the curious ones: try to learn association rules or use regression to "predict" the missing value from other attributes)

(4) Pick one of the two classifiers from exercise B above and train it with the different strategies to deal with missing values (using 10-fold cross-validation, best parameter setting from the experiments above). Document experiment settings and summarize the results in the report. Analyze the effect of increasing percentages of missing values in single attributes and across several attributes. For the same amount of missing values, does the classifier performance degrade identically, irrespective of which attribute these missing values occur in? Analyze the behavior according to questions such as
   a. Do missing values in some attributes cause more damage than in others? If so, why? Why not?
   b. How do the different replacement strategies work? Which ones have the most positive effect on classifier performance?
   c. How to the strategies degrade with increasing fractions of values of a specific attribute missing?

D) Clustering:

1. Pick two conceptually different clustering algorithms (either two from WEKA or one from WEKA and the SOM using the SOMToolbox) and explain them briefly, as well as the reason for choosing them.
2. Run these clustering algorithms on the data analyzing and compare the results of different parameter setting. Specifically, specify different numbers of clusters and see how the cluster cardinality and semantics that may be discovered, changes.
   a. How even/uneven is the distribution of items onto clusters?
   b. How easy is it to interpret these clusters?
   c. How well do the clusters correspond to the classes provided? (i.e. use class labels to measure the purity of clusters
3. Test different scalings of the data and evaluate the effect on the clustering.
   a. What is the effect of different scalings on the clustering algorithms?
   b. How even/uneven is the distribution of items onto clusters?
   c. How well do the clusters correspond to the classes provided? (i.e. use class labels to measure the purity of clusters
   d. Is the effect of different (correct and incorrect) scalings comparable to the performance differences observed in the classification task?

HandsOn Assignments 2: Data Mining

E) Summarize your report

1. Summarize your overall findings and lessons learned
2. Provide feedback on the exercise in general: which parts were useful / less useful; which other kind of experiment would have been interesting, … (this section is, obviously, optional and will not be considered for grading. You may also decide to provide that kind of feedback via the feedback mechanism in TISS – in any case I'd appreciate learning about it to adjust the exercises for next year.)

General advice:

- Formatting guidelines: Follow the formatting guidelines provided by the ACM for their proceedings series. You can use the templates provided by the ACM (LaTeX strongly recommended, but Word/OpenOffice also ok)
  ACM Proceedings Style File: LaTeX2e - Strict Adherence to SIGS style
  http://www.acm.org/sigs/pubs/proceed/template.html
- Provide the names and student ID numbers of each group member as part of the author information.
- There is no page limit, but try to use common sense and the structure/questions provided in the task description to determine the space needed. Use graphs to visualize findings.
- Upload in TUWEL at zip/tgz/rar file (filename should include the student ID numbers of all group members involved, plus the name of the dataset analyzed), containing
  - the report (PDF)
  - the scripts/programs your wrote and
  - subsidiary information that may be needed to repeat your experiments
- Make sure the report has some proper structure following as far as possible the structuring provided by the questions posed in the task description.
- Follow general guidelines for scientific papers, i.e. enumerate and label all figures, equations, and refer to them in the text of the report.
- Collaboration between groups is welcome, but make sure each group uses a different dataset.
- Try to perform at least part of the experiments within a group together, and specifically discuss the results amongst each other (rather than subdividing tasks such as one does classification, the other clustering; or each taking one classifier and then working independently - this is not a recommended strategy).
- Try to understand what the results tell you, note down any peculiar observations you make, try to provoke "wrong" behavior of the algorithms (over-learning, strange parameters, test wrong encodings, absurd scalings, 99% missing values in an attribute,...) and report these findings as well
- Explore WEKA beyond the activities required in this assignment - it's an exciting tool (with some limitations)

Submission Deadline: January 19 2014 (but try to do most of the assignment before that)