Using Exploration Trees in the Loan Applications Domain

P. Berka[1,2]

Abstract: Induction of decision trees belongs to the most popular algorithms used in machine learning and data mining. When building a decision tree, we recursively partition the attribute space in a top-down way at each branching node looking for best attribute to make a split. The quality of the attribute is evaluated on the basis of its ability to separate examples of different classes. This process will result in a single tree that can be use both for classification of new examples and for description the partitioning of the training set. But due to the used greedy search strategy, this tree need not to split the training data in the best way with respect to the classes.

In the paper we propose an alternative approach that is related to the idea of finding all interesting relations (usually association rules, but in our case all interesting trees) in given data. When building the so called exploration trees, we consider not a single best attribute for branching but more good attributes for each split. This modification of the tree learning algorithm will result in more trees, each partitioning the data in a little different way giving thus alternative knowledge for segmentation of the data with respect to the classes. The exploration trees can of course be used for classification tasks as well; in this case we can use either single tree or work with a whole ensemble of created trees.

The proposed method will be compared with the standard C4.5 algorithm on several data sets from the loan application domain. One of the data sets comes from the ECML/PKDD Discovery Challenge workshop, the other sets are taken from the UCI Machine Learning Repository. The results show, that among the exploration trees created for different data sets, there was always a tree that better splits the training data (had higher classification accuracy on training data) than the tree created (by greedy search) using C4.5. We are aware of the fact, that trees in C4.5 are tuned to perform well on the testing data (to avoid over-fitting) but if the task is to find and describe segments of training data related to the class attribute, our method gives better results.

[1] Department of Information and Knowledge Engineering
University of Economics
W. Churchill Sq. 4, 130 67 Prague, Czech Republic
berka@vse.cz

[2] Department of Computer Science and Mathematics
University of Finance and Administration
Estonska 500, 10100 Prague, Czech Republic