Building Ensembles Using Decision Tree Metrics Based Meta-Trees

Peter Kokol, University of Maribor, Slovenia; E-mail: kokol@uni-mb.si
Gregor Stiglic, University of Maribor, Slovenia; E-mail: gregor.stiglic@uni-mb.si

ABSTRACT

Ensembles of classifiers have become one of the most popular techniques in machine learning research. The problem of selecting the most appropriate classifier for classification that is known from machine learning is also an important issue in the ensemble building methods. This paper tries to find a method that would select the best performing ensemble building method based on dataset characteristics instead of the most appropriate classifier. The proposed approach captures the characterization of the dataset from the metrics of the tree induced from the dataset. On 15 benchmark datasets, the proposed meta-tree based method discovered some strong and simple rules that could be used in future research in the field of basic ensemble building method selection.

1. INTRODUCTION

Combining multiple classifiers in ensembles is one of the standard and most important techniques for improving classification accuracy in machine learning. This paper compares performance of four different ensemble methods – Bagging [1], Boosting [2], Random Subspacing [3] and Random Forests [4]. There are numerous methods of combining different classifiers into ensembles, but there is no universal method of how to achieve the best accuracy on all datasets using only one ensemble creation method. The results of four above mentioned ensemble building methods are used to build meta-tree which would be capable of selecting the most appropriate ensemble building method based on the characteristics of the given machine learning problem. A variety of data characterization techniques have been developed, however their quality still needs to be improved. In our approach the idea of capturing the characterization from the metrics of the tree induced from the dataset is used [5].

The paper begins with the description of the ensemble building techniques in chapter 2 which is followed by chapter 3 on meta-tree building details. Next section describes our experiments and results of meta-tree learning. Conclusions and proposal of future work are discussed in the final chapter.

2. ENSEMBLE CREATION TECHNIQUES

To create an ensemble of classifiers two components are needed: a set of diversely trained classifiers and a mechanism that composes the single predictions into an overall outcome. This paper compares four most popular methods of combining the classifiers into ensembles which are described in more detail below.

2.1. Bagging

To compose ensemble from base classifiers using bagging, each classifier is trained on a set of *n* training examples, drawn randomly with replacement from the original training set of size *m*. Such subset of examples is also called a bootstrap replicate of the original set. Each such subset contains, on average, 63.2% of the original training set [6]. A set of classifiers is then used to classify the example using the majority vote of the ensemble.

The vital element of the bagging technique is the instability of the classifiers. If perturbations in the learning set can cause significant difference in the classifier construction, than bagging can improve accuracy of ensemble.

2.2. Boosting

Boosting is represented by the AdaBoost.M1 algorithm described in [2], which is the most commonly used algorithm for boosting ensembles. To use boosting it is assumed that the base classifier can handle weighted examples. In case where this is not possible we use sampling of the training set examples according to a weight distribution.

In AdaBoost algorithm classifiers are trained sequentially. Each classifier is trained on the dataset based on the misclassification of the previously generated classifier. Weights of the examples are updated according to the classification accuracy of the previous classifier by lowering weights of correctly classified examples and increasing weights of misclassified examples. After the training process is finished the predictions are made using weighted vote of the individual classifiers.

Boosting was tested by many researchers who proved that it can be declared as one of the best ensemble methods [7, 8, 9]. It was also applied to decision trees based ensembles and it can be considered as one of the best classification methods [10].

As each other classification method boosting also contains some drawbacks. One of the most important is overfitting although early literature mentions that boosting would not overfit even when running for a large number of iterations [11]. Recent research clearly shows overfitting effects when boosting is used on datasets with higher noise content [7, 12].

2.3. Random Subspacing

The ensemble method also called Random Subspacing was proposed by Ho in [3] and is based on multiple decision trees constructed systematically by pseudo-randomly chosen features from the training dataset. Each tree is constructed using randomly chosen features which cause higher diversity of ensemble members. Therefore Random Subspacing method can achieve nearly monotonic increase in generalization accuracy while preserving high accuracy on training data, provided that the features are sufficient to distinguish all samples belonging to different classes, or that there is no intrinsic ambiguity in the datasets [13].

2.4. Random Forests

Breiman upgraded the idea of bagging by combining it with the random feature selection for decision trees. This way he created Random Forests, where each member of the ensemble is trained on a bootstrap replicate as in bagging. Decision trees are than grown by selecting the feature to split on at each node from randomly selected number of nodes. Number of chosen features is set to $log_2(k+1)$ as in [4], where k is the total number of features.

Random Forests are the ensemble method that works well even with noisy content in the training dataset and are considered as one of the most competitive methods that can be compared to boosting [14].

3. BUILDING META-TREES

There was a lot of research done in the field of meta-learning methods based on data characterization in the domain of machine learning [5, 15, 16]. Meta-learning is based on set of meta-attributes that usually characterizes the dataset, and search for the optimal correlation between these attributes and the performance of learning algorithms. This paper focuses on research of the correlation between basic decision tree metrics as meta-attributes and performance of different ensemble building methods as learning algorithms.

To build meta-trees one should follow the three basic steps:

- describe the characterization of the dataset by definition of meta-attributes

- learn meta-tree using as much as possible different datasets to estimate the value of meta-attributes and measure the accuracy of different ensemble building techniques
- evaluate the results of built meta-tree

Many techniques extracting the characterization of dataset have been developed, such as data characterisation techniques (DCT) [17] including simple measures (e.g. number of attributes, classes et al.), statistical measures, and information theory-based measures. Inspired by work of Bensusan et al. [16] and Peng et al. [5] we use simple decision tree metrics to define meta-attributes, but two different decision tree algorithms are used – C4.5 [18] and Logistic Model Trees (LMT) [19].

Eight metrics were defined that consist of C4.5 pruned and unpruned number of leaves and number of nodes, LMT number of leaves and nodes and C4.5 leaves and node pruning ratio. Pruning ratio is computed as number of leaves divided by number of nodes in pruned tree by the same value in unpruned tree. Another two attributes that represent number of instances and features were added.

4. EXPERIMENTAL RESULTS

This section describes the experiments that were performed to build and evaluate meta-learning decision tree for ensemble method selection. Most of the experiments were done using WEKA toolkit for machine learning [20]. To train the meta-tree 10 UCI [21] datasets were selected and tested the built tree on 5 UCI datasets which are presented in Table 1.

Instance and feature sampling when learning the meta-tree were used to make learning datasets more diverse. This way each dataset consisted of 100 to 200 sampled instances and *n/2* to *n* features for each experiment, where *n* represents the number of features of the original dataset. Ensemble building methods were run five times with five different settings of forest dimension for 20 times on 10 training set datasets. This way 1000 different datasets were used in the phase of learning the meta-tree.

A part of our experiment was also a classic comparison of four different ensemble methods using different number of decision trees. To observe the accuracy at different number of trees in ensembles five groups of experiments were selected. The first and smallest set of ensembles consisted of 6 trees and each next set consisted of the number in previous experiment multiplied by 2. Following this formula the size of ensembles equals $6 \cdot 2^i$, where *i=[1..5]*. The average accuracy of 10-fold cross-validation on 10 training datasets using different number of decision trees in ensembles is shown in Figure 1.

Table 1. List of datasets and their details

Training Set				
Dataset	Attributes	Continuous	Instances	Classes
cmc	9	2	1473	3
tic-tac-toe	9	0	958	2
segment	19	19	1500	7
balance-scale	4	4	625	3
ecoli	7	7	336	8
vowel	14	10	990	11
vehicle	18	18	846	4
dermatology	34	1	366	6
heart-statlog	13	13	270	2
liver	7	7	345	2
Test Set				
diabetes	8	8	768	2
glass	9	9	214	7
wine	13	13	178	3
hepatitis	19	6	155	2
sonar	60	60	208	2

It can be seen in Fig.1 that Bagging gains less on accuracy comparing to other methods when the number of classifiers increases. While on the other hand Random Forest method still increases significantly from when increasing the number of trees from 24 to 48. The later can be explained with the fact that Random Forest method produces the most diverse sets of classifiers as it is using Bagging and even some kind of Random subspacing when generating classifiers. At around 100 classifiers the accuracy of all methods becomes stable so the ensemble with 96 trees is used in most of experiments.

In the next experiment the first meta-learning tree was built using all 1000 runs of ensemble building on different datasets. Fig. 2 represents the number of wins for each method and number of experiments where two or more methods achieved the best results. Due to 10-fold cross-validation only 33 experiments ended with a tie of two or more methods that produced the same best accuracy. Those 33 experiments were excluded from the meta-learning training set which consisted of 967 examples, each containing 8 decision tree metrics and 2 statistical measures (number of instances and features). After this step the first C45 based meta-learning tree was built using all four ensemble building methods as decisions. The accuracy of this tree was 74.77%.

From Fig. 2 it can be observed that Boosting won in the majority of cases. These results can also be the consequence of the fact that Random Forest, Bagging and Random Subspacing methods share a lot of characteristics in the way they build ensembles of classifiers. Therefore an additional experiment was performed where Boosting method was compared to Non-Boosting methods (Random Forest was used as a representative of this group). The accuracy on training set was 84.4%.

Next experiment compares small (6 trees) and large (96 trees) ensembles and shows how ensemble methods are performing when the number of classifiers increases (Fig. 3).

Figure 1. Average accuracy for different ensemble methods using different number of decision trees

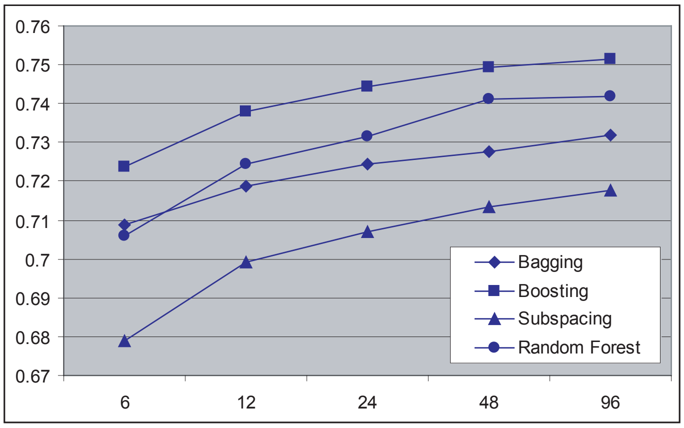


Figure 2. Number of wins and tied situations of four ensemble building techniques

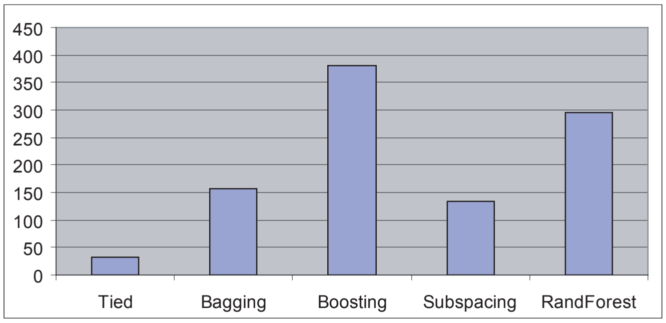
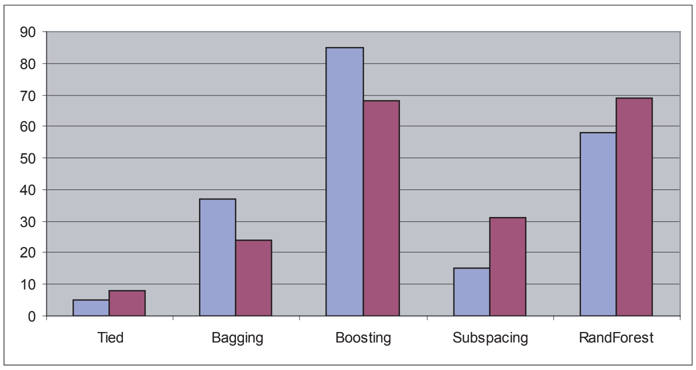


Figure 3. Comparison of four ensemble methods on small and large ensembles



It can be seen that Boosting is not dominant anymore when large number of classifiers is used. It is also interesting to observe a drop in accuracy of Bagging performance and an increase of accuracy at Random Subspacing method. This could again be due to the fact that Random Forest shares a lot of characteristics with Bagging and Random Subspacing. Therefore another observation was done where "Bagging – Random Subspacing" and "Boosting - Random Forest" comparison is done independently (Fig. 4).

Figure 4. Direct comparison of ensemble building methods (large ensembles)

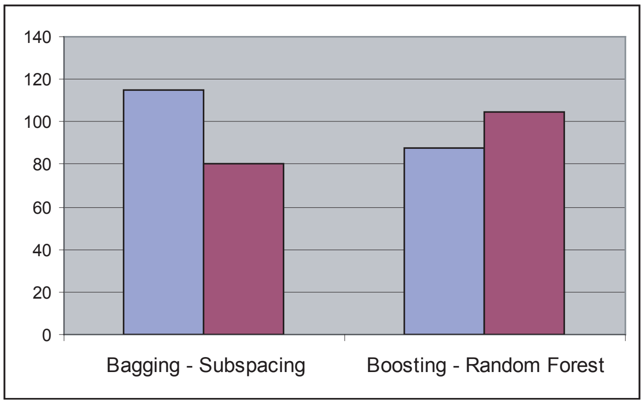


Figure 5. Simplified meta-learning tree (1-Boosting, 4-NonBoosting)

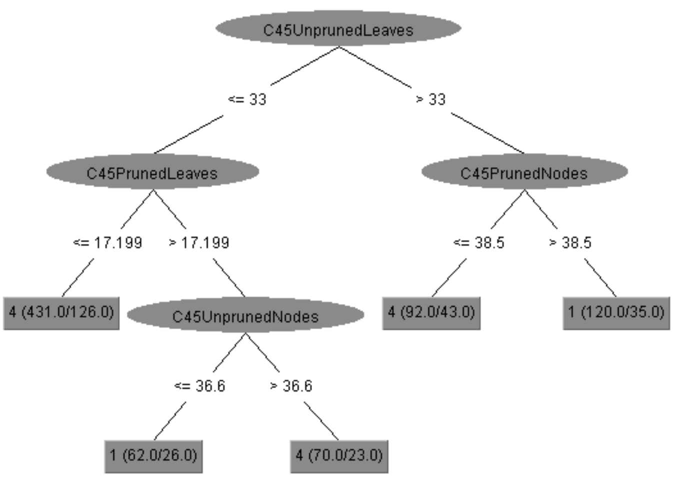


Fig. 4 shows that Random Forests in direct comparison even more evidently outperform Boosting in large ensembles. On the other hand Bagging outperformed Random Subspacing, but this comparison depends on the number of instances and the dimensionality of selected datasets very much. On average our training datasets had 14.2 features which could be to few for the Boosting method to perform better.

To simplify the observation of the large generated C45 meta-tree the number of instances per leaf is set to 60 and generated fairly simple meta-tree with the accuracy of 68%. The simplified tree (Fig. 5) shows two strong rules that characterize the main difference between boosting and non-boosting methods. Left and right most branch of the tree, which include more then 70% of training samples, can be transformed to the following rules:

- IF C45UnprunedLeaves <= 33 AND C45PrunedLeaves <= 17.2 THEN UseNonBoosting
- IF C45UnprunedLeaves > 33 AND C45PrunedNodes > 38.5 THEN UseBoosting

The above rules show that trees with lower complexity dictate use of non-boosting methods while trees with the higher measured complexity suggest using the boosting method of ensemble building.

Based on the observations from Fig. 5 it can be said that C45 metrics contribute the most useful information for the ensemble building method selection.

5. DISCUSSION AND FUTURE WORK

This paper presents the meta-tree approach to selection of the most appropriate ensemble building method for different datasets in machine learning domain. The accuracy of four most popular methods was measured for combination of classifiers into ensembles and were used for building C45 meta-tree based on C45 and LMT based decision tree metrics. This method of selecting the ensemble building technique enables assessment of the dataset complexity through decision tree metrics.

Our experiments show that some very simple rules can be extracted from the generated meta-trees which can help us understand which meta-features are the most important in the selection of ensemble building method problem. From the obtained results it is also obvious that there is no assurance that a single best ensemble building method can be found for all datasets.

There are still several open issues that have not been discussed in this paper. In particular more metrics could be included, especially the metrics of complexity in meta-datasets. In the future we will incorporate some ideas of fractal dimensions that can be used for the measurement of complexity and try to find some common points with the already used metrics. Another important aspect of the research would be the comparison of our proposed method with the similar methods that are in principle used for single classifier selection.

REFERENCES

[1] L. Breiman, "Bagging predictors," *Machine learning*, Vol. 24, No. 2, pp. 123-140, 1996.
[2] Y. Freund, R.E. Schapire, "Experiments with a New Boosting Algorithm," in *Proc. of the Thirteenth International Conference on Machine Learning*, Morgan Kaufmann, pp. 148-156, 1996.
[3] T.K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 8, pp. 832-844, 1998.
[4] L. Breiman, "Random forests," *Machine learning*, Vol. 45, No. 1, pp. 5-31, 2001.
[5] Y. Peng, P. Flach, P. Brazdil, and C. Soares, "Decision tree-based data characterization for meta-learning," *in Proc. of the ECML/PKDD-02 Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning (IDDM-2002)*, Helsinki University Printing House, pp. 111-122, 2002.
[6] P. Melville, R.J. Mooney, "Creating diversity in ensembles using artificial data," *Information fusion*, Vol. 6, pp. 99-111, 2005.
[7] G. Rätsch, T. Onoda, and K.R. Müller, "Soft margins for AdaBoost," *Machine learning*, Vol. 42, No. 3, pp. 287-320, 2001.

[8] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Machine learning*, Vol. 36, No. 1-2, pp. 105-139, 1999.

[9] T.G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine Learning*, Vol. 40, No. 2, pp. 139-157, 2000.

[10] H. Drucker and C. Cortes, "Boosting decision trees," in *Advances in Neural Information Processing Systems 8: Proceedings of NIPS'95*, Vol. 8, pp. 479-485, 1996.

[11] R. Meir, G. Rätsch, "An Introduction to Boosting and Leveraging," *Machine Learning Summer School 2002*, pp. 118-183, 2002.

[12] A. Grove and D. Schuurmans, "Boosting in the limit: Maximizing the margin of learned ensembles," in *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pp. 692-699, 1998.

[13] T.K. Ho, "Random Decision Forests," in *Proceedings of the 3rd International conference on Document Analysis and Recognition*, Montreal, Canada, August 14-18, pp. 278-282, 1995.

[14] T.G. Dietterich, "Ensemble Learning," in *The Handbook of Brain Theory and Neural Networks*, 2nd ed., M.A. Arbib, Ed. Cambridge, MA: The MIT Press, pp. 405-408, 2002.

[15] P. Brazdil, C. Soares, and J. P. da Costa, "Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results," *Machine Learning*, Vol. 50, No. 3, pp. 251-277, 2003.

[16] H. Bensusan, C. Giraud-Carrier, and C. Kennedy, "A higher-order approach to meta-learning," in *Proceedings of the ECML'2000 workshop on Meta-Learning: Building Automatic Advice Strategies for Model Selection and Method Combination*, 2000, pp. 109-117.

[17] C. Linder, and R. Studer, "AST: Support for Algorithm Selection with a CBR Approach," in *Proceedings of the 16th International Conference on Machine Learning*, Workshop on Recent Advances in Meta-Learning and Future Work. 1999.

[18] J.R. Quinlan, "*C4.5: Programs for Machine Learning*," Morgan Kaufmann, San Mateo, CA, 1993.

[19] N. Landwehr, M. Hall, and E. Frank, "Logistic Model Trees," in *Proc. of 14th European Conference on Machine Learning*, pp. 241-252, 2003.

[20] I.H. Witten, and E. Frank, "*Data Mining: Practical machine learning tools with Java implementations*," Morgan Kaufmann, San Francisco, CA, 1999.

[21] C.L. Blake, and C.J. Merz, "UCI repository of machine learning databases," 1998 [Online]. Available: www.ics.uci.edu/~mlearn/MLRepository.html

Related Content