Using Visual Interpretation of Small Ensembles in Microarray Analysis

Gregor Stiglic, Matej Mertik, Vili Podgorelec, Peter Kokol
University of Maribor, Slovenia
{gregor.stiglic, matej.mertik, kokol}@uni-mb.si

Abstract

Many different classification models and techniques have been employed on gene expression data. These computational methods are in rapid and continuous evolution and there is no clear consensus on which methods are best to cope with the complex microarray data analysis. Currently ensembles of classifiers are regarded as one of the best classification techniques as they can achieve excellent classification accuracy in comparison to single classifiers methods. One of their main drawbacks is their incomprehensibility. This paper addresses the important issue of the tradeoff between accuracy and comprehensibility when building ensembles and proposes a novel visual technique for interactive interpretation of the knowledge from the small ensembles consisting of only a few decision trees. This way we can achieve better accuracy compared to single classifier, but still maintain a certain level of comprehensibility in small ensembles. The results show that our small ensembles outperform the single classifiers and still retain comprehensibility. Our study also points out that in order to take advantage of our proposed method we need more effective small ensemble building techniques.

1. Introduction

Ensemble methods are learning algorithms that build a set of classifiers which are used to classify new instances by combining their predictions. Empirical studies show that ensembles are often much more accurate than the individual classifiers [1, 2, 3, 4]. One of the main drawbacks of the ensemble classifiers is the incomprehensibility of the produced classification models. Usually it is possible to convert all single models from an ensemble to a set of rules, but such rule sets quickly become too complex to be comprehensible. Since the purpose of most data mining systems is to support decision making the need for knowledge interpretation in these systems is apparent.

Many methods have been developed to improve the comprehensibility of incomprehensible classification algorithms like neural networks or ensembles of classifiers. The main scheme for such methods is rule extraction, that is, symbolic rules are extracted from the 'black-box' model. Most usual method is simple rule extraction from all components of a classification model that is followed by aggregation of the extracted rules. One of first such systems was presented by Setiono in [5], where the neural network is pruned and the outputs of hidden units are discretized. The rule extraction algorithm is executed iteratively for each sub-network constructed from hidden units with many outputs.

Another neural-network based extraction algorithm was presented by Craven and Shavlik [6], which aims to extract rules that map inputs directly to outputs. The whole problem is solved by a set of queries to the oracle from which an ID2-of-3 decision tree is constructed. Similar approach was used by Domingos [7] where he generalized the concept of oracle queries. His idea was that any 'black-box' model can be captured in a comprehensible classification model by using additional 'artificial' examples that are labeled according to the original model. Similar approach was used by Zhou et al. in [8], where they build ensemble of artificial neural networks, which is used to generate instances and then extracts symbolic rules.

It can be seen that most of the previous work on interpretation of 'black-box' classification models relies on rule extraction techniques. Another possibility to improve the comprehensibility of classification process is introduction of classification visualization. There were only a few papers proposing solutions in this field. One of the first is paper by Melnik [9], where he concentrates on visualization of high-dimensional classifiers. An extensive work in visualization of multiple and single decision trees that also includes their interpretation was done by Urbanek in [10]. He presents a tool for interactive visual interpretation of decision tree forests. Another paper by Frank and Witten [11] presents a technique that uses a two-dimensional visualization based on class probability

Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)
0-7695-2517-1/06 $20.00 © 2006 IEEE

estimates. All above mentioned papers suggest that visual interpretation of classification models is worth further research to help both experts and non-experts understand the most accurate classification techniques.

The rest of this paper is organized as follows. In Section 2, our method for virtual interpretation of small ensembles is presented. Next section presents our proposed simple small ensemble building technique, which is followed by Section 4 where we present results and comparison to other classification methods. In the last section, the main contribution of this paper is summarized and several issues for future works are indicated.

2. Visual Interpretation of Ensembles

There was a lot of research done in the field of reducing the number of models that are combined in an ensemble. Usually as the number of models increases this means an increase in the comprehensibility of the ensemble, assuming that single models combined in an ensemble are comprehensible models (e.g. decision trees or a set of rules). This paper proposes a novel tool for visual interactive interpretation of ensembles consisting of three decision trees. The tool was developed with the idea of possible extension to the number of decision trees, but still keeping the complexity of the ensemble as low as possible. Usually the data mining tools are not used only by data mining experts, but are used in numerous different fields. With this in mind we wanted to keep our ensemble interpretation tool as simple as possible, so that it can be also used by non-experts. Fig. 1 presents the main screen of the VISE (Visual Interpretation of Small Ensembles) tool. The main decision tree window can be seen on the left hand side of the screen, while on the opposite side the other two decision trees are displayed. Each of the trees on the right side can be magnified in the main window by switching the main and one of the two side windows containing reduced visualization of the tree. At the bottom of the screen we can observe a set of rules that are extracted from the above trees in an interactive way. Interaction is allowed, because as we mentioned above, usually we get too many rules when extracting them from ensembles of decision trees. Therefore we allow user to select the branches of the trees that he is interested in, either by decision at the terminal node of the branch or by features that are included in the branch. The first interactive step is selection of a significant branch in a tree, which is followed by automatic extraction of the rule from this branch and all the rules that could possibly contribute to the decision from the remaining two trees.

The automatic extraction of the rules can be done in two ways:

- using the training set examples, we mark the branches (and extract rules from them) which contain the examples that were used in building of the selected branch
- if there are too few examples in the selected branch, we artificially create the examples whose attribute values correspond to the selected branch and label them using a robust and accurate ensemble (in our case we use Random Forests ensemble consisting of 100 decision trees)

For each small ensemble we can also get the quick accuracy estimation using 10-fold cross-validation.

3. Experimental Settings

Five microarray analysis datasets from Kent Ridge Bio-medical Data Set Repository [12] were used in our experiments, where we evaluated our proposed ensemble building method. In this initial stage of small ensembles interpretation research, we used very simple method of ensemble building which we call “Triple Trees”. We compare the accuracy of our proposed method to J48 decision trees [13], three-step boosting, boosting using 100 iterations and Random Forests using 100 decision trees. All tests were done using WEKA tool [13], which was also used for development of our VISE tool.

Triple Trees

In our VISE tool Triple Trees implementation we use J48 decision trees as a base model. To ensure we build diverse trees we split the training set into three equal parts. The first decision tree is generated from the first two thirds, the second from the last two thirds and the last tree from first and last third of the examples from the training set. Default pruning settings are used to achieve the low complexity of the generated decision trees. To evaluate the possibility of extending the number of trees to five, we also tested the 5-trees method, which is the same as Triple Trees, but uses five instead of three decision trees.

Boosting

To avoid misinterpretation of the term boosting the AdaBoost.M1 algorithm described in [14] was used, which is the most commonly used algorithm for boosting ensembles. To use boosting it is assumed that

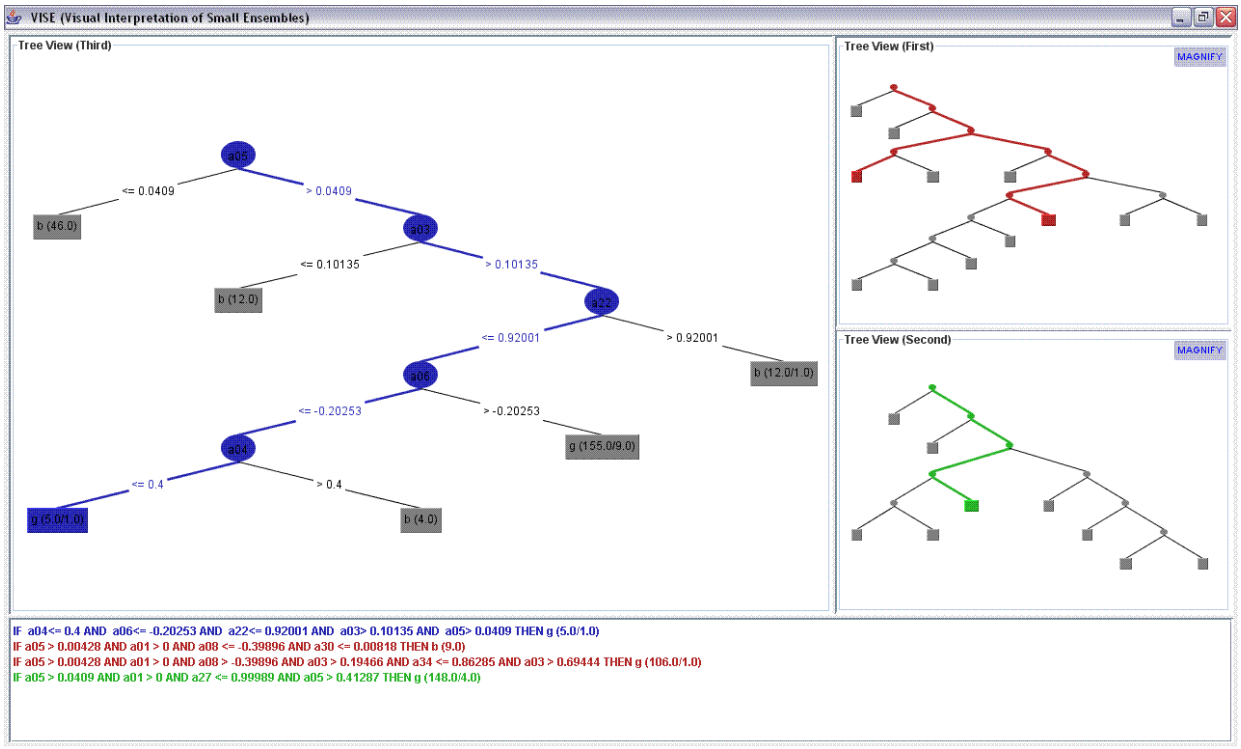

Figure 1. Main user interface of VISE tool

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 most accurate ensemble methods [1, 2, 15], that was also applied to decision tree based ensembles [16].

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 [17]. Recent research clearly shows overfitting effects when boosting is used on datasets with higher noise content [15, 18].

In our experiment boosting is used with two different settings. In the first case we use only five iterations and use it as a small ensemble building technique in VISE tool. In the second version we use 100 iterations, which is the most usual setting.

4. Datasets

Five widely used publicly available gene expression datasets that were used in our experimental evaluation of the proposed method are presented in this chapter.

Leukemia1 dataset (amlall)

The original data comes from the research on acute leukemia by Golub et al. [19]. Dataset consists of 38 bone marrow samples from which 27 belong to acute lymphoblastic leukemia (ALL) and 11 to acute myeloid leukemia (AML). Each sample consists of probes for 6817 human genes. Golub used this dataset for training. Another 34 samples of testing data were used consisting of 20 ALL and 14 AML samples. Because we used leave-one-out cross-validation, we were able to make tests on all samples together (72).

Breast cancer dataset (breast)

This dataset was published in [20] and consists of extremely large number of scanned gene expressions. It includes data on 24481 genes for 78 patients, 34 of which are from patients who had developed distance metastases within 5 years, the rest 44 samples are from

Table 2. Comparison of accuracy for J48 decision tree and 4 ensemble algorithms

Dataset	J48	3-Trees	5-Trees	5-Boost	Boosting
amlall	83.58	92.16	92.88	89.46	91.20
breast	64.81	70.92	72.46	72.01	86.97
lung	97.06	96.88	97.09	97.44	97.51
mll	85.11	90.70	91.67	89.02	89.55
prostate	86.83	88.20	88.93	89.99	94.25
Average	**83.47**	**87.77**	**88.60**	**87.58**	**91.89**

patients who remained healthy from the disease after their initial diagnosis for interval of at least 5 years.

Lung cancer dataset (lung)

Lung cancer dataset includes the largest number of samples in our experiment. It includes 12533 gene expression measurements for each of 181 tissue samples. The initial research was done by Gordon et al. [21] where they try to classify malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung.

Leukemia2 dataset (mll)

This Leukemia dataset tries to discern between 3 types of leukemia (ALL, MLL, AML). Dataset contains 72 patient samples, each of them containing 12582 gene expression measurements. Data was collected by Armstrong et al. and results published in [22].

Prostate Tumor dataset (prostate)

Prostate Tumor dataset contains 52 prostate tumor samples and 50 non-tumor (labeled as "Normal") prostate samples with around 12600 genes. The original study was conducted by Singh et al. in [23].

5. Results

Each classification method was tested for accuracy using 10-fold cross-validation that was repeated 20 times to achieve higher accuracy. Table 1 shows that Triple Trees can effectively improve the accuracy of classification comparing to simple J48 decision trees. Using some advanced techniques that are specialized for building compact ensembles it would be possible to even further improve the results.

It is also interesting that more complex datasets (i.e. using larger number of features) cause bigger difference in accuracy when comparing single classifier accuracy with the ensemble or even small ensembles. But we have to be careful, because more complex datasets cause even more complex decision trees or rules, which means large ensembles are totally incomprehensible in such cases.

6. Discussion

In our paper we present a novel method of interpreting small ensembles consisting of three or five decision trees. Our method is interactive, so that even non-experts are able to identify rules which could be interesting to them. Also, a new small ensemble building technique is presented, which was mainly developed only to test the interpretation tool and to generate diverse decision trees. This technique can still be further developed or completely replaced by a specialized algorithm for building small ensembles. During the development phase of our VISE tool, another interesting paper was published by Zhou and Li [24] which proposes a novel Tri-Training technique where three decision trees are used and could also be used as a base small ensemble building algorithm in our tool.

Further improvement of the tool is also needed in the filed of interpretation of decision on instance by instance basis, which would enable user to get more background knowledge of the decision for each instance that can be chosen by the user. Another important issues that could be explored in the future is rule ranking, which would even further simplify the work for non-experts.

7. References

[1] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting and variants," Machine Learning, 36(1/2):525–536, 1999.

[2] T.G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision tress: Bagging, boosting and randomization," Machine Learning, 40(2):139–158, 2000.

[3] Y. Freund and R.E. Schapire, "Experiments with a new boosting algorithm," In Proceedings of the 13th International Conference on Machine Learning, pages 148–156. Morgan Kauffman, 1996.

[4] L. Kuncheva and C. Whitaker, "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy," Machine Learning, Vol. 51, pp. 181-207, 2003.

[5] R. Setiono, "Extracting rules from neural networks by pruning and hidden-unit splitting," Neural Computation, Vol. 9, No. 1, January 1997, pages 205-225.

[6] M.W. Craven and J.W. Shavlik, "Extracting treestructured representations of trained networks," Advances in Neural Information Processing Systems, 8, pp.24--30, 1996.

[7] P. Domingos, "Knowledge acquisition from examples via multiple models," In Proc. of the 14th International Conference on Machine Learning, pp. 98 – 106, Morgan Kauffman, 1997.

[8] Z.H. Zhou, and M.L. Zhang, "Ensembles of multi-instance learners," In: Proceedings of the 14th European Conference on Machine Learning (ECML'03), Cavtat-Dubrovnik, Croatia, LNAI 2837, 2003, pp.492-502.

[9] O. Melnik, and J.B. Pollack, "Theory and scope of exact representation extraction from feed-forward networks," Cognitive Systems Research 3(2), 2002.

[10] Urbanek, S, "Exploring Statistical Forests," in Proc. of the 2002 Joint Statistical Meeting, Mira DP, 2002.

[11] E. Frank, and M. Hall, "Visualizing Class Probability Estimators," Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat, Croatia, 2003.

[12] J. Li, and H. Liu, "Ensembles of cascading trees," in Proc. IEEE International Conference on Data Mining (ICDM 2003), IEEE Computer Society, Melbourne, Florida, pp. 585.

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

[14] Y. Freund, and 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.

[15] G. Rätsch, T. Onoda, and K.R. Müller, "Soft margins for AdaBoost," Machine learning, Vol. 42, No. 3, pp. 287-320, 2001.

[16] 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.

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

[18] 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.

[19] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," Science, vol. 286, no. 5439, pp. 531-537, 1999.

[20] L. J. van 't Veer, H. Dai, M. J. van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. Der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer", Nature, no. 415 pp. 530-536, 2002.

[21] G. J. Gordon, R. V. Jensen, L.-L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswami, W. G. Richards, D. J. Sugarbaker, and R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," Cancer Research, no. 62, pp. 4963-4967, 2002.

[22] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukaemia," Nat Genet., vol. 30 no. 1, pp. 41-47, 2002.

[23] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, W.R. Sellers, "Gene expression correlates of clinical prostate cancer behavior", Cancer Cell, Vol. 1, 2002, pp. 203-9

[24] Z.-H. Zhou and M. Li, "Tri-training: exploiting unlabeled data using three classifiers," IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1529-1541.