Gene Expression Analysis of Leukemia Samples Using Visual Interpretation of Small Ensembles: A Case Study

Gregor Stiglic[1], Nawaz Khan[2], Mateja Verlic[1], and Peter Kokol[1]

[1] University of Maribor, FERI, Smetanova 17, 2000 Maribor, Slovenia
[2] School of Computing Science, Middlesex University, The Burrough, Hendon,
London NW4 4BT, UK
{gregor.stiglic,kokol}@uni-mb.si,
N.X.Khan@mdx.ac.uk

Abstract. Many advanced machine learning and statistical methods have recently been employed in classification of gene expression measurements. Although many of these methods can achieve high accuracy, they generally lack comprehensibility of the classification process. In this paper a new method for interpretation of small ensembles of classifiers is used on gene expression data from real-world dataset. It was shown that interactive interpretation systems that were developed for classical machine learning problems also give a great range of possibilities for the scientists in the bioinformatics field. Therefore we chose a gene expression dataset discriminating three types of Leukemia as a testbed for the proposed Visual Interpretation of Small Ensembles (VISE) tool. Our results show that using the accuracy of ensembles and adding comprehensibility gains not only accurate but also results that can possibly represent new knowledge on specific gene functions.

Keywords: gene expression analysis, machine learning, decision trees.

1 Introduction

Gene expression analysis is a novel technique that in contrast to measurement of a single gene transcription enables measurement of all genes in an organism at once. Finding combinations of genes whose expression levels distinguish different groups of diseases is a complex task that is usually solved by different machine learning or statistical algorithms. While most of the algorithms gain very accurate results in classification of gene expression samples, there is still very limited number of algorithms that can offer a good interpretation of the results that were gained using advanced machine learning techniques.

Methods like bagging, boosting and random forests, which combine decisions of multiple hypotheses, also called ensemble methods, are some of the strongest existing machine learning methods. Ensemble methods are learning algorithms that build a set of classifiers which are used to classify new instances by combining their predictions. It was shown that ensembles clearly outperform the single classifiers in terms of classification accuracy [1-5].

One of the main drawbacks of ensemble classifiers is weak comprehensibility of the produced classification models. Many times it is possible to convert all single

J.C. Rajapakse, B. Schmidt, and G. Volkert (Eds.): PRIB 2007, LNBI 4774, pp. 189–197, 2007.

models from an ensemble to a set of rules, but such rule sets quickly become too complex to be comprehensible. 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. Sometimes this process can be even simpler – e.g. when working with decision tree (DT), rules can be extracted directly from the branches of a tree.

Another option when improving the comprehensibility of classification process is introduction of classification visualization. One of the first papers where visualization of high-dimensional classifiers is presented was written by Melnik [6], where visual interpretation of neural networks is described. An extensive work in visualization of multiple and single DTs that also includes their interpretation was done by Urbanek in [7]. He presents a tool for interactive visual interpretation of DT forests. Another paper by Frank and Witten [8] presents a technique that uses a two-dimensional visualization based on class probability 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.

Above mentioned examples demonstrate use of visual interpretation in classical machine learning problems, while it should also be mentioned that there were some experiments that combine visualization and microarray classification process. A study that uses Support Vector Machines and tries to interpret the results using visualization was presented by Caragea et al. [9]. A similar study in terms of visualization of microarray data to interpret results of classification was conducted by Lee et al. [10]. Their tool called GeneGobi is mostly based on statistical instead of machine learning methods. Another tool was developed by Curk et al. [11] where visualization is used for setting the experiments and interpretation of results, which represents a major simplification of experimental process in microarray analysis.

The following sections of this paper present a case study where a novel Visual Interpretation of Small Ensembles (VISE) method [12] is used on a microarray dataset discriminating three types of Leukemia that was initially presented by Armstrong et al. [13]. In contrast to experiments described in [12] another version of VISE tool was used where DTs are generated based on bagging instead of boosting DTs. Section 2 contains a presentation of virtual interpretation of small ensembles. It is followed by a section describing the experimental settings and results. Section 4 presents a validation study by providing an interpretation of the results in the context of rule sets and then by comparing the proposed adaptations with the combined and simple DTs for leukemia grouping. In the last section, the main contribution of this paper is summarized and several issues for future works are indicated.

2 Interpretation Tool

Usually as the number of classifiers in ensemble increases it means an increase of complexity and decrease of comprehensibility, assuming that single models combined in an ensemble are comprehensible models (e.g. DTs or a set of rules). This paper demonstrates a novel tool for visual interactive interpretation of ensembles consisting of three DTs. It is based on idea that a small ensemble can increase the accuracy and still keep the complexity of the ensemble as low as possible. To ensure the diversity of induced DTs is high enough we use a simple variant of bagging [14] technique for building DTs. Training set is split into three equal parts, where the first DT 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 in training set. Default pruning settings are used to achieve lower complexity levels of generated DTs. All DTs used are standard C4.5 trees as implemented in Weka environment [15]. The same environment was used as a core for the developed small ensembles interpretation tool.

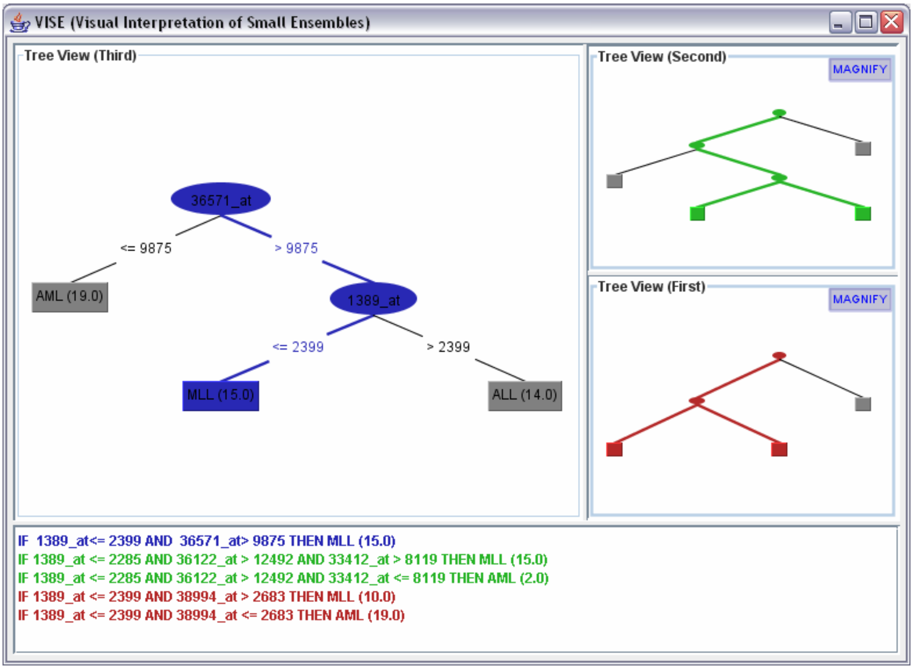
Main screen of the VISE (Visual Interpretation of Small Ensembles) tool is presented in Fig. 1. Primary DT window can be seen on the left hand side of the screen, while on the opposite side the other two DTs are displayed in smaller windows. Each of the trees on the right side can be magnified and transferred to the main window by switching the main and one of the two side windows containing simplified visualization of the tree. Bottom of the screen contains a set of rules that are extracted from the above trees in an interactive way. Interaction is an integral part of the tool; therefore user is allowed to select branches of trees that he is interested in, either by decision at the terminal node of the branch or by features (i.e. nodes) that are included in the branch. The first interactive step is selection of a significant branch (according to expert's opinion) 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.

First step is followed by automatic extraction of rules that can be done in two ways:

1. Using the training set examples, a single or a group of branches is selected (and rules are extracted from them) which contain the examples that were used when the selected branch was built.
2. In case 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 DTs)

This way user is able to observe which rules (i.e. DT branches) could possibly vote against decision of the main DT. Using this knowledge we are able to understand how and why an ensemble would vote differently in case of using a single DT for specific samples that fit in the selected branch of the tree.

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



```
IF  1389_at<= 2399 AND  36571_at> 9875 THEN MLL (15.0)
IF 1389_at <= 2285 AND 36122_at > 12492 AND 33412_at > 8119 THEN MLL (15.0)
IF 1389_at <= 2285 AND 36122_at > 12492 AND 33412_at <= 8119 THEN AML (2.0)
IF 1389_at <= 2399 AND 38994_at > 2683 THEN MLL (10.0)
IF 1389_at <= 2399 AND 38994_at <= 2683 THEN AML (19.0)
```

Fig. 1. Main screen of VISE tool

The informative value of resulting rules is marked by their color that represents their origin and by their decision class. The following section demonstrates usage of the tool on a gene expression dataset discriminating three types of Leukemia.

3 Experimental Settings and Results

This section highlights the details of our study and key findings that were obtained by applying the VISE tool to Leukemia microarray dataset. In the original research by Armstrong et al. [13] clustering algorithms revealed that lymphoblastic leukemias with MLL translocations can clearly be separated from conventional acute lymphoblastic and acute myelogenous leukemias. The same dataset consisting of 72 tissue samples, each of them containing 12582 gene expression measurements was used in our experiment. In the original study a dataset was split in a training set containing 57 samples and testing set with another 15 samples. In our study all 72 samples (24 ALL, 20 MLL, 28 AML) were used in a single dataset, while 10-fold cross validation was used for accuracy estimations. Basic DT that was used to extract rules from a small ensemble of three DTs is presented in Figure 2 where number in parentheses indicates that all examples from training set were correctly classified.

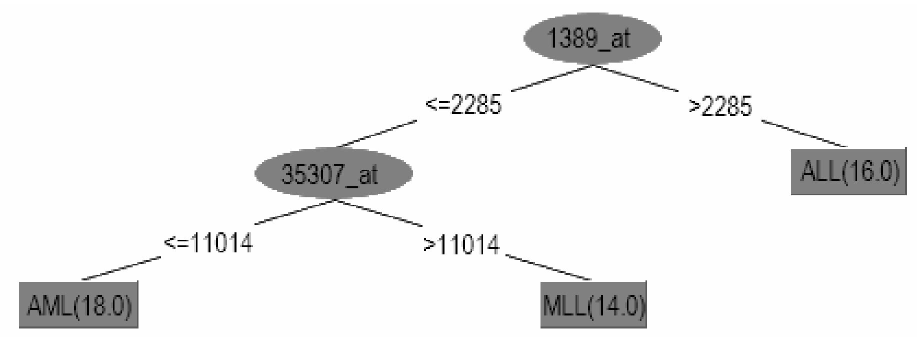


Fig. 2. Primary decision tree induced by VISE

Rules that were directly extracted from small ensemble are presented in Table 1. All rules extracted from primary DT are displayed in bold and are followed by rules that are fired in other two DTs using the corresponding samples from the selected primary tree branch. When evaluating the accuracy of decision trees that were built using Leukemia dataset [13] it was indicated that three decision trees together reached an average 10-fold cross validation accuracy rate of 90.5% compared to 84.2% that was achieved by single decision trees.

For easier understanding and rule interpretation gene id to gene description mappings are presented in Table 2. When interpreting the results from VISE tool it should be noticed that among the rules fired in secondary DTs it is possible to find rules that are voting against the rule extracted from primary DT. Those rules could also be called opposing rules and should be taken into consideration when interpreting results. In our case there are two genes that are included in such rules – i.e. genes with identification numbers 41503_at and 38046_at.

Table 1. Rules fired for each branch in the primary DT

AML Branch
IF 35307_at NOT EXPRESSED AND 1389_at NOT EXPRESSED THEN AML
IF 1389_at NOT EXPRESSED AND 38046_at NOT EXPRESSED THEN AML
IF 41503_at EXPRESSED AND 1389_at NOT EXPRESSED THEN MLL
IF 41503_at NOT EXPRESSED THEN AML
MLL Branch
IF 35307_at EXPRESSED AND 1389_at NOT EXPRESSED THEN MLL
IF 1389_at NOT EXPRESSED AND 38046_at EXPRESSED THEN MLL
IF 1389_at NOT EXPRESSED AND 38046_at NOT EXPRESSED THEN AML
IF 41503_at EXPRESSED AND 1389_at NOT EXPRESSED THEN MLL
ALL Branch
IF 1389_at EXPRESSED THEN ALL
IF 1389_at EXPRESSED THEN ALL
IF 41503_at NOT EXPRESSED THEN AML
IF 41503_at EXPRESSED AND 1389_at EXPRESSED THEN ALL

Table 2. Gene descriptions for easier interpretation of results in Table 1

Gene ID	Description
35307_at	Homo sapiens mRNA for GDP dissociation inhibitor beta
1389_at	Human common acute lymphoblastic leukemia antigen (CALLA) mRNA, complete cds
38046_at	Homo sapiens mRNA for Prer protein
41503_at	Homo sapiens mRNA for KIAA0854 protein, complete cds

4 Interpretation of Results

This section provides an expert evaluation of results and shows the differences between traditional gene expression analysis techniques and VISE tool in terms of results interpretation. Evaluation is based on rules that were extracted from DTs and are presented in Table 1.

GDP dissociation inhibitor (GDI) is a protein that controls the GDP-GTP exchange reactions. GTP-binding proteins involve in trafficking of molecules between cellular organelles. GDIs slow the rate of dissociation of GDP and release GDP from membrane-bound Rabs [16]. The GDI beta gene is vulnerable to inversion/deletion mutation and may cause leukemia. The association of GDI and its expression involving cellular transport have been reported by many researchers, for example [17] and [18]. It is evident from many researches that GDI expression is responsible for chronic myelogenous leukemia.

Common acute lymphocytic leukemia antigen (metallo endopeptidase; neutral endopeptidase) is an important cell surface marker in the diagnosis of human acute lymphocytic leukemia (ALL) [19]. It is present on leukemic cells of pre-B phenotype, which represent 85% of cases of ALL. Yagi et al. [20] and Fasching et al. [21] have suggested that the specific antigen receptor may be present at birth in some patients with ALL, suggesting a prenatal origin for the leukemic clone. They also have showed that some patients with ALL characterized by specific translocations have been demonstrated to have cells showing the translocation at the time of birth. This is because Lymphoblasts antigen receptors are unique to a particular patient. Sheikh et al. [22] has reported of peripheral blood lymphocytosis caused by CD23, CD25 in addition to CD5 and CD10. The expression of antigens for ALL have been reported my many researchers. For example, Ogawa et al. [23], Cutrona et al. [24] and Shipp [25] have reported the close correlation between expression of CD10/neutral endopeptidase and tumor development.

Red protein (RER protein; IK factor; cytokine IK) involves in the negative regulatory pathway of constitutive MHC Class II antigens expression. It expressed at similar levels in fetal and adult tissues in developmental stage. A lower expression of mRna for the protein may lead to fetal brain placenta COT 25-normalized squamous cell carcinoma, B cell metastatic chondrosarcoma and colon tumor.

Transcription factor ZHX2 involves in transcription factor activities and regulates the transcription [26, 27]. The irregular expression of mRNA may lead to lymphoma, B-cell lymphatic leukemia and lung and spleen lymphoma.

The rules above, although, show the direct association of GDI and lymphoblastic leukemia antigen to the ALL and AML, some of the features of leukemia exhibit a mixed type of leukemia, for example, MLL. The morphological features and immunophenotypic profile of the leukemia is not readily classifiable and may be influenced by some other expressions, for example, expression of Prer proteins and Transcription factors. The importance of these genes that influence the classification of leukemia cannot be ignored.

5 Conclusions and Future Work

From the previous section it is evident that results obtained from VISE tool can reveal potential new knowledge and make interpretation of results a simple task for bioinformatics experts. It was shown that in most cases it is enough to select a few crucial genes that are sufficient for improvement of classification accuracy. But a step further enables extraction of additional rules and significant genes that can be decisive for comprehensibility of classification results.

Another important aspect of VISE tool is the interactiveness of the classification process. It enables interaction with the expert in a way where it can be specified which rules (i.e. DT branches) are important for him and does not rely only on automatic feature selection like most of other methods.

As usual in the gene expression research we should emphasize that all the results are obtained from datasets containing a low number of samples. The increase of datasets that will provide us with more samples in the future also brings some new challenges. We can expect more complex classifiers which will also be more accurate. Therefore one of the main aims for the future is reduction of produced classifiers when working with many of them at once as it is the case in ensembles of classifiers.

References

1. Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, boosting and variants. Machine Learning 36(1/2), 525–536 (1999)
2. Dietterich, T.G.: An experimental comparison of three methods for constructing ensembles of decision tress: Bagging, boosting and randomization. Machine Learning 40(2), 139–158 (2000)
3. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Proceedings of the 13th International Conference on Machine Learning, pp. 148–156. Morgan Kauffman, San Francisco (1996)
4. Kuncheva, L., Whitaker, C.: Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. Machine Learning 51, 181–207 (2003)
5. Hall, L.O., Bowyer, K.W., Banfield, R.E., Bhadoria, D., Kegelmeyer, W.P., Eschrich, S.: Comparing Pure Parallel Ensemble Creation Techniques Against Bagging. In: The Third IEEE International Conference on Data Mining, Melbourne, Florida, pp. 533–536 (November 2003)

6. Melnik, O., Pollack, J.B.: Theory and scope of exact representation extraction from feed-forward networks. Cognitive Systems Research 3(2) (2002)
7. Urbanek, S.: Exploring Statistical Forests. In: Proc. of the 2002 Joint Statistical Meeting, Mira DP (2002)
8. Frank, E., Hall, M.: Visualizing Class Probability Estimators. In: Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat, Croatia (2003)
9. Caragea, D., Cook, D., Honavar, V.: Visual Methods for Examining Support Vector Machine Results, ISU Technical Report (December 2005)
10. Lee, E.K., Cook, D., Wurtele, E., Kim, D., Kim, J., An, H.: GENEGOBI: Visual Data Analysis Aid Tools for Microarray Data. In: Computational Statistics 2004 Symposium (COMPSTAT 04) (2004)
11. Curk, T., Demsar, J., Xu, Q., Leban, G., Petrovic, U., Bratko, I., Shaulsky, G., Zupan, B.: Microarray data mining with visual programming. Bioinformatics 21(3), 396–398 (2005)
12. Stiglic, G., Mertik, M., Podgorelec, V., Kokol, P.: Using Visual Interpretation of Small Ensembles in Microarray Analysis. In: Proceedings of Computer Based Medical Systems, Salt Lake City, UT, USA (2006)
13. Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R., Korsmeyer, S.J.: MLL translocations specify a distinct gene expression profile that distinguishes a unique leukaemia. Nat. Genet. 30(1), 41–47 (2002)
14. Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996)
15. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco (2005)
16. Bachner, D., Sedlacek, Z., Korn, B., Hameister, H., Poustka, A.: Expression patterns of two human genes coding for different rab GDP-dissociation inhibitors (GDIs), extremely conserved proteins involved in cellular transport. Hum. Mol. Genet. 4(4), 701–708 (1995)
17. Cutrona, G., Tasso, P., et al.: CD10 is a marker for cycling cells with propensity to apoptosis in childhood ALL. Br. J. Cancer 86(11), 1776–1785 (2002)
18. Fasching, K., Panzer, S., Haas, O.A., et al.: Presence of clone-specific antigen receptor gene rearrangements at birth indicates an in utero origin of diverse types of early childhood acute lymphoblastic leukemia. Blood 95(8), 2722–2724 (2000)
19. Kawata, H., Yamada, K., Shou, Z., Mizutani, T., Yazawa, T., Yoshino, M., Sekiguchi, T., Kajitani, T., Miyamoto, K.: Zinc-fingers and homeoboxes (ZHX) 2, a novel member of the ZHX family, functions as a transcriptional repressor. Biochem. J. 373(Pt 3), 747–757 (2003)
20. Ogawa, H., Iwaya, K., Izumi, M., Kuroda, M., Serizawa, H., Koyanagi, Y., Mukai, K.: Expression of CD10 by stromal cells during colorectal tumor development. Hum. Pathol. 33(8), 806–811 (2002)
21. Sheikh, S.S., Kallakury, B.V., Al-Kuraya, K.A., Meck, J., Hartmann, D.P., Bagg, A.: CD5-negative, CD10-negative small B-cell leukemia: variant of chronic lymphocytic leukemia or a distinct entity? Am. J. Hematol. 71(4), 306–310 (2002)
22. Shipp, M.A., Tarr, G.E., Chen, C.Y., Switzer, S.N., Hersh, L.B., Stein, H., Sunday, M.E., Reinherz, E.L.: CD10/neutral endopeptidase 24.11 hydrolyzes bombesin-like peptides and regulates the growth of small cell carcinomas of the lung. Proc. Natl. Acad. Sci. USA 88(23), 10662–10666 (1991)
23. Shisheva, A., Sudhof, T.C., Czech, M.P.: Cloning, characterization, and expression of a novel GDP dissociation inhibitor isoform from skeletal muscle. Mol. Cell Biol. 14(5), 3459–3468 (1994)

24. Strausberg, R.L., Feingold, E.A., et al.: Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc. Natl. Acad. Sci. USA 99(26), 16899–16903 (2002)
25. Toyoda, M., Nakamura, M., Makino, T., Kagoura, M., Morohashi, M.: Sebaceous glands in acne patients express high levels of neutral endopeptidase. Exp. Dermatol. 11(3), 241–247 (2002)
26. Weitzdoerfer, R., Stolzlechner, D., Dierssen, M., Ferreres, J., Fountoulakis, M., Lubec, G.: Reduction of nucleoside diphosphate kinase B, Rab GDP-dissociation inhibitor beta and histidine triad nucleotide-binding protein in fetal Down syndrome brain. J. Neural Transm. Suppl. 61, 347–359 (2001)
27. Yagi, T., Hibi, S., Tabata, Y., et al.: Detection of clonotypic IGH and TCR rearrangements in the neonatal blood spots of infants and children with B-cell precursor acute lymphoblastic leukemia. Blood 96(1), 264–268 (2000)