Decision Tree Classifiers in Bioinformatics

Inese Polaka, *Riga Technical University*, Igor Tom, *United Institute of Informatics Problems of the National Academy of Sciences of Belarus*, Arkady Borisov, *Riga Technical University*

***Abstract* – This paper presents a literature review of articles related to the use of decision tree classifiers in gene microarray data analysis published in the last ten years. The main focus is on researches solving the cancer classification problem using single decision tree classifiers (algorithms C4.5 and CART) and decision tree forests (e.g. random forests) showing strengths and weaknesses of the proposed methodologies when compared to other popular classification methods. The article also touches the use of decision tree classifiers in gene selection.**

***Keywords* – bioinformatics, cancer classification, decision tree classifiers, gene expression, gene microarray data analysis**

I. Introduction

Gene expression microarrays allow monitoring and studying gene expression profiles. Microarrays contain data of up to several hundred patients and tens of thousands of genes simultaneously. These microarrays can be used for diagnostics and monitoring of illnesses as well as patients' response to medication. For all these tasks it is important to identify the profiles of similar gene expressions that can point to groups of sick/healthy people, different types of cancer, etc. These tests have only recently become widely available and need research to find the best fitting analysis methods. For this purpose many scientists are looking at the machine learning methods to find those that perform well on data with this specific character.

The choice of the classification method is not definite and different classification algorithms fit different problems – there is no one dominant method. Lee et al. [1] propose support vector machine (SVM) as the method that is most likely to provide the best classification results while working with high-dimensional data and/or missing data [2] - [4]. However, the right data pre-processing can show a significant improvement on other methods. Decision trees are proven to be as effective as other classifiers and exceed the efficiency of other classifiers for particular problems. Researchers also give preference to decision tree classifiers because of their ability of relevant gene selection and scalability, as well as model accuracy and easy interpretation.

This article presents reviews of papers published since 2000 that use decision tree methods for cancer classification in gene expression data.

The paper is organized as follows. An overview of gene expression microarray technology and related classification problems is presented in Section 2. Decision tree classification methods (single decision tree classifiers and decision forests) are introduced in Section 3. A review of relevant articles is presented in Section 4. Finally, some concluding remarks are made.

II. Classification in Bioinformatics

Microarray technology enables scientists to explore gene expressions of thousands of genes simultaneously. The patterns that are hidden in this amount of data are crucial for diagnosis and monitoring of diseases like cancer and can only need a fraction of the whole gene set. The methods that were initially used to analyze the data were mostly statistical but the introduction of machine learning tools to bioinformatics problems has shown to pay off, mostly in classification tasks that in this particular field are diagnostics – sick patients/control group and types of illness, and drug response monitoring via short time series of gene expressions. The peculiarity of these tasks is not only the high-dimensional data but also the evaluation of classifiers and results. It takes into account not only the accuracy of classification models but also their biological relevance [5]. These models can reveal underlying processes, gene interaction and marker genes. For example, decision trees provide information about gene interaction by their stepwise splitting of the data set – each split reveals one gene and the hierarchical structure shows the nature of interaction.

III. Decision Tree Classifiers

Decision tree classifiers recursively partition the instance space using hyperplanes that are orthogonal to axes. The model is built from a root node which represents an attribute and the instance space split is based on function of attribute values (split values are chosen differently for different algorithms), most frequently using its values. Then each new sub-space of the data is split into new sub-spaces iteratively until an end criterion is met and the terminal nodes (leaf nodes) are each assigned a class label that represents the classification outcome (the class of all or majority of the instances contained in the sub-space). Setting the right end criterion is very important because trees that are too large can be overfitted and small trees can be underfitted and suffer a loss in the accuracy in both cases. Most of the algorithms have a mechanism built in that deals with overfitting; it is called pruning.

Each new instance is classified by navigating them from the root of the tree down lo a leaf, according to the outcome of the tests along the path [6].

Although decision trees produce efficient models, they are unstable – if the training data sets differ only slightly, the resulting models can be completely different for those two sets. Due to that, decision trees are often used in classifier ensembles.

A. Single Decision Tree Classifiers

The most popular algorithms that build single decision trees for classification are C4.5 and CART (Classification And Regression Trees). Decision trees were first proposed by J. Ross Quinlan in [7] describing algorithm ID3 that was used as a basis for other decision tree classifiers that were created changing evaluation functions and construction parameters. Algorithm C4.5 was proposed in [8] and CART algorithm was presented in [9] by Breiman et al. Both algorithms divide attribute space in a similar manner but they differ in tree structure, split criteria and pruning method.

Algorithm C4.5 usually uses Information gain or Gain ratio as the criteria to choose the attribute for each split. Information gain is the change in entropy of information if the state of information is changed. Let *C* be the class attribute with values $\{c_1, c_2, \dots, c_n\}$ and *A* attribute with values $\{a_1, a_2, \dots, a_k\}$, *H(C)* be the entropy of the class attribute, and *H(C|A)* conditional entropy that shows entropy of *C* if state of attribute *A* is known, Information gain is:

$$I(C,A) = H(C) - H(C|A) \tag{1}$$

The entropy of attribute *C* is:

$$H(C) = -\sum_{i=1}^{n} P(C = c_i) log_2(P(C = c_i)) \tag{2}$$

where $P(C = c_n)$ is the relative frequency of class value c_n. And the conditional entropy is:

$$H(C|A) = -\sum_{j=1}^{k} P(A = a_j) H(C|A = a_j). \tag{3}$$

Information gain favors attribute with higher number of values. To avoid that, gain ratio can be used. This criterion penalizes a large number of attribute values by dividing Information gain with entropy of the attribute itself:

$$IG(C,A) = \frac{I(C,A)}{H(A)} \tag{4}$$

where the entropy of attribute A is calculated as follows:

$$H(A) = -\sum_{j=1}^{k} P(A = a_j) log_2(P(A = a_j)). \tag{5}$$

CART algorithm in its turn usually uses Gini index as splitting criteria. Gini index is calculated as:

$$G(C) = 1 - \sum_{i=1}^{n} P(C = c_i). \tag{6}$$

CART and C4.5 have also other differences like pruning method, missing values handling and others [10]. Pruning examines and substitutes subtrees of the whole tree with a leaf or a branch of the subtree where necessary. C4.5 uses reduced error pruning that analyzes if a subtree replacement with a leaf leads to less error. This technique requires a separate data set for pruning, which can be a drawback but it examines every subtree once and is much faster than other techniques [11]. CART uses minimal cost complexity pruning technique which assigns costs to subtrees based on the error from pruning and the size of the subtree [10]. This technique does not require a separate data set for pruning.

B. Decision Forests

Decision Forests is an ensemble methodology, which builds a predictive model by integrating multiple models (decision trees); it can be used for improving prediction performance as well as stability of classifiers [6]. The most popular methods are bagging, boosting and Random forests.

Bagging was first introduced by Breiman [12] in 1996. In bagging for each trial t=1,2,…,T a training set of size N is sampled with replacement from the original instances (the training set is the same size as the original set but some instances may not appear in it while some instances appear more than once). Then a classifier is built for each generated set and the final classifier is formed by aggregating the T classifiers. To classify a new instance, a vote for class k is recorded by every classifier, and the final assigned class is the class with the most votes [13].

Boosting was first introduced by Freund and Schapire [14] when they proposed AdaBoost algorithm. Boosting maintains a weight for each instance – the higher the weight, the more the instance influences the classifier. At each trial, the vector of weights is adjusted to reflect the performance of the classifier, with the result that the weight of misclassified instances is increased. The final classifier also aggregates the learned classifiers by voting, but each classifiers vote is a function of its accuracy [13].

Random forests use a large number of unpruned decision trees, which are created by randomizing the split at each node of the decision tree. The number of attributes used to determine the decision at a node of the tree is predefined and is less than the original number of attributes. The attributes are chosen randomly and the best split among those attributes is chosen. The classification of a new sample is performed using majority vote [6].

IV. Literature Review

In recent years researchers have been using machine learning tools to classify cancer (discriminating healthy individuals from cancer patients and discriminating among various types of cancer) for diagnostic purposes in microarray data. Both simple decision tree classifiers (e.g. C4.5 and CART) and their ensembles are used for various classification tasks. Although decision tree classifiers can be used for multi-class tasks, most of the problems discussed in the papers are associated with data with binary classes.

The cancer classification problem is defined as follows. Given a training set $T = \{(t_1, c_1), (t_2, c_2), \dots, (t_n, c_n)\}$, where t_i is an m-dimensional vector of gene expression values, m is the total number of genes, $t_i = (x_i^1, x_i^2, \dots, x_i^m)$, $m \gg n$ and $c_i \in C$ is the class label of the i-th vector where C is the set of classes; a test set $S = \{s_1, s_2, \dots, s_l\}$, where each s_1 is an m-dimensional gene expression data vector; find a classification function f that assigns class value c to each S with maximal accuracy.

A. Algorithms C4.5 and CART

Dudoit et al. [15] studied the performance of different discrimination methods for the tumor classification based on gene expression data. For this purpose they used nearest-neighbor classifiers, linear discriminant analysis and classification trees (CART algorithm). To estimate the accuracy of the classifiers the authors used 10-fold cross validation. They also used classifier aggregation for CART classifiers to avoid instability – bagging and boosting methods were used to aggregate maximum ‘exploratory’ classifiers by weighted voting. These methods were applied to three cancer gene expression data sets: Lymphoma, Leukemia and NCI 60. The data sets were pre-processed by imputing missing data using k nearest-neighbor algorithm, normalizing the data and selecting the most relevant genes based on the ratio of their between-group to within-group sums of squares. To evaluate classifiers the authors observed test set error rates (30% of the data were left out of the whole set to test the built classifiers for each run), observation-wise error rates (the proportion of times an observation was classified incorrectly). The performance of CART classifiers was intermediate and aggregated tree predictors were generally more accurate. The test set errors for Lymphoma data set were in the range between 0 and 20%. CART algorithm had about 10% error rate and the best of the tree classifiers – boosting had misclassified ~ 5% of the samples. In the Leukemia data set (two classes) the test set error rates were in the range between 0 and 20%. CART boosting outperformed other decision tree classifiers and had an error rate of ~5%. For three class problem in the Leukemia data set the test set error range was between 4 and 8%, and CART boosting had misclassified ~5% of the test samples. The accuracy of the classifiers in NCI 60 data set was much lower – the error was between 40 and 60% and CART boosting showed ~48% error. The authors concluded that although other classifiers had higher accuracy (linear discriminant and nearest neighbor methods showed 100% accuracy), they ignore the relations among different genes, whereas decision trees are capable to exploit and reveal interactions among genes.

Lu and Han [5] discuss the solution of the cancer classification problem using machine learning tools. The authors used Fisher’s linear discriminant analysis, weighted voting of informative genes – GS method, Naïve Bayes method, neural networks, decision trees, Nearest neighbor analysis, CAST, max-margin classifiers, SVM and aggregated classifiers (boosting). They applied these methods to publicly available cancer cDNA microarray data sets – Colon-cancer, Ovarian-cancer, Leukemia, Lymphoma, NCI 60 and another NCI data set. They observed that all of the classification methods performed well and none of the methods is superior to the rest. The difference between classical classification tasks and cancer classification makes the performance of classification methods worse. Although other methods achieved 100% accuracy (like Naïve Bayes classifier) decision trees allowed the authors to explore the gene interactions and assess interactions between genes. One of the main reasons is that these methods do not give much biological information; besides, they do not use available information about gene interaction and significance of known genes.

Lee et al. [1] compared performance of 21 methods that were applied to seven cancer data sets. The methods that were used for experiments included SVMs, neural networks, discriminant analysis methods, CART and aggregating classifiers. They also tested three gene-selection approaches and tested the efficiency using all of the classification methods. All the methods showed similar results. The performance of CART algorithm was average when compared with other methods with the same pre-processing procedures. SVM showed accuracy higher than 90% on most data sets outperforming other classifiers including CART that showed results between 44% and 90% accuracy. Aggregating tree classifiers mostly increased the performance and outperformed other classical methods (accuracy between 68% and 99% for various data sets) but none of the algorithms was dominant for all data-sets. The authors also concluded that Random forests was the best method among the tree methods when the number of classes is moderate.

Lee et al. [16] studied the impact of different dimension reduction methods on algorithms C4.5 and SVM. Six dimension reduction methods – three linear (PCA, Linear discriminant analysis and linear MDS) and three non-linear (Graph embedding, Isomap and LLE) methods were applied to 10 different cancer data sets with binary classes and then the classifiers were tested on the reduced data. Without the use of dimension reduction the average accuracy was very similar for both classifiers. The use of linear dimension reduction methods did not result in the expected improvement of efficiency – the accuracy dropped whereas it rose significantly when the non-linear dimension reduction methods were applied in the same situation. When classification algorithms were applied to reduced data, C4.5 outperformed SVM in the most of the data sets. The accuracy of algorithm C4.5 improved reaching 100% using non-linear dimensionality reduction methods and Lung cancer and Prostate cancer data sets, the accuracy reached 96,9% for DLBCL data set, 95% in Leukemia and Lymphoma data sets and 63,3% for Ovarian cancer data sets.

Another research on dimension reduction was conveyed by Horng et al. [17]. They introduced a new method of gene selection based on C4.5 algorithm. The first step in the proposed algorithm (called Resampling) is to increase the number of virtual samples and avoid the curse-of-dimensionality problem. Samples are randomly chosen and a decision tree model is built for each new set of samples. Then all internal nodes of the generated trees are gathered and the genes that appear most frequently are chosen for classification (the authors suggest taking 6-10 genes). The authors used different approaches for classification – Naïve Bayes method, Decision trees, SVM etc. These methods were applied to 13 public tumor microarray sets. When the authors tested the new gene selection approach combined with Decision tree classifiers, their accuracy was higher using most data sets than the performance of the same classification methods using different gene selection methods (GEMS and HykGene). The authors also tested other classification methods with the reduced set and the results varied for different methods applied to different data sets. The accuracy of all classifiers was lower using the 9 tumors and the 14 tumors data sets that can be explained by sample to class ratio – the data sets included many classes and few samples for each class making it harder for classification algorithms to build accurate models and reveal the patterns that were in the data. C4.5 algorithm had 65% to 70% accuracy using these data sets. The accuracy of C4.5 using Leukemia data set was about 97%, the performance using Colon tumor, SRBCT, DLBCL and Lung cancer data was around 95%. The performance using Prostate tumor and Brain tumor data sets was around 90% and slightly worse (but above 80%) for the rest of the data sets. The use of the proposed gene selection method improved the performance of C4.5 algorithm that was up to 10% higher than the accuracy shown in other researches using the same data sets.

B. Decision forests

Huang et al. [18] introduce Improved decision forest (IDF), which, unlike the classical Decision forests, can use the same feature several times so that the most informative genes can contribute more to the class assignment. This method was applied to Hepatotoxicity data as well as cancer data to classify types of cancer (colon cancer, leukemia and lymphoma), using full data sets (all genes) as well as reduced numbers of genes (200, 100 and 50). The ‘signal-to-noise’ gene selection method proposed by Golub et al. [19] was used to filter the genes and reduce the dimensionality . The results were compared to six other classification methods including Random forests, bagging and boosting using C4.5 classifier. The Improved decision forest and Random forest showed better results (2-3 % more accurate) than SVM and kNN classifiers particularly on full data sets (thousands of genes). The accuracy of IDF using the Hapatotoxicity data set was 90% using all genes for training and rose to slightly above 91% when the gene selection method was applied. The performance of all classifiers evened out when the gene set was reduced to 50 genes (~91% accuracy). The accuracy of IDF using Colon data set was 82% using all genes (this method outperformed others) and rose above 83% when the gene selection method was applied. For both of these data sets bagging performed very well, showing accuracy that was at most 2% below the accuracy of IDF. The accuracy of IDF using Leukemia data set was highest (97%) when the gene selection was not applied. The performance using the reduced data set dropped slightly (1-2%). The performance of IDF using Lymphoma data set was average (accuracy ~95% with or without gene selection) but Bagging outperformed other methods and showed stable performance for different dimensionalities (96.6% accuracy). Authors conclude that bagging suffered less from the curse of dimensionality showing stable results using data sets with different numbers of genes. This shows the scalability benefit of tree methods that is of high importance in tasks like microarray data analysis.

Hu [20] proposed a new method for discovering relevant gene interactions called Recursive random forests (RRF) that is based on Random forests – in the first step a robust random forest is generated to classify gene expression data by recursively applying Random forest algorithm. Then the generated trees are analyzed to find the most frequently used co-occurring genes (interaction patterns), which could mean that these interactions are disease-relevant and can be used for disease classification. He applied this method to four cancer datasets – Breast cancer, NCI 60, Thrombocythemia and Michigan group lung dataset. First the pathways of the data were ranked and the top 10% were used for building Random forests – he removed one pathway at a time and tested the other *n-1* pathways with random forests. This was repeated recursively until there was only one pathway left. The group of pathways with the smallest error was then used to explain the observed sample types. Frequent itemset mining was then applied to this group to find co-occurring genes from different pathways. The author determined the most relevant genes that contribute most to the classification process and compared the gene subsets found to those discovered by Random forest using its Mean Decrease in Accuracy feature evaluation and 85% of the found relevant genes overlapped the most informative genes found by Random forest. The accuracy of RRF using Breast cancer data was 90.9% whereas the accuracy of the Random forest was 88.8%. The accuracy using Thrombocythemia data set was 90% for RRF and 82.5% for Random forests. The performance of RRF using NCI60 data set was 88%, Random forests showed 84% accuracy. The accuracy using Lung cancer data set was 81.2% for RRF and 75.3% for Random forests. This research shows that the proposed method for pathway analysis performs better in phenotype classification than the standard Random forests method. This approach also helps to discover potential interactions between genes.

Zintzaras and Kowald [21] used Forest classification tree and Forest SVMs to classify four types of tumor in prostate gene expression data. At threshold split value of 0.001 and using 100 markers, the classification tree consisted of 29 terminal nodes and achieved perfect classification. Forest SVM performed worse and its performance improved when the set of genes used for classification increased to 200 and more genes. The authors note that Decision tree classifiers allow exploring the data structure and relevant genes and they provide easy to understand decision rules.

Diaz-Uriarte and Alvarez de Andrez [22] introduce a new approach to gene selection for classification based on Random forests. They also use Random forests for cancer classification in gene expression data comparing its performance with kNN, SVM, Diagonal linear discriminant analysis (DLDA) and Shrunken centroids. For gene selection, the authors use measures of variable importance of Random forest – the decrease on classification accuracy. They iteratively find the least important variables (genes) and discard the worst 20% of the variables without recomputing the importance of the variables at each step. Then the solution with the smallest number of genes and an acceptable out-of-bag error rate (previously set based on preferable range within u standard errors of the minimum error rate of all forests) is chosen for classification. The authors used Leukemia, Breast cancer, NCI60, Adenocarcenoma, Brain cancer, Colon cancer, Lymphoma, Prostate cancer and SRBCT data sets. The efficiency of the classifiers using that proposed gene selection method is in most cases comparable to the efficiency of Random forests and also comparable to the efficiency of other classification methods. The accuracy of Random forests using Leukemia data set was 94.9%. This result was average and the method was outperformed by 4 other methods (SVM showed the best result – 98.6% accuracy). The performance of all methods using Breast cancer data set with two classes was very similar – Random forests using the proposed gene selection method had 66,8% accuracy and the best result was 67,4% (Shrunken centroids). Using the Breast cancer data set with three classes the Random forests (using the proposed gene selection method) had 65.4% accuracy and outperformed other methods. The accuracy of Random Forests using NCI60 data set was 74.8% (the best result was 75.4% using Shrunken centroids). The RRF method also outperformed other methods using Adenocarcenoma data set with 87.5% accuracy. The results using Brain cancer data set were very good using RRF (84.6%) that was outperformed only by SVM (86.2% accuracy). The accuracy of RRF using Colon cancer data set was 87.3% that was outperformed only by Shrunken centroids (87.8% accuracy). The performance of RRF using Lymphoma data set was very good (99.1% accuracy) and was outperformed only by *k* nearest neighbor method (00.2% accuracy). The RRF using the proposed gene selection method and Prostate cancer data set was the best and had 92.3% accuracy. The performance of RRF using SRBCT data set was average and had 97.9% accuracy (the best result was 98.9% for DLDA method). This research showed that there is no one best method for all data sets and Random forests perform as good or in most cases even better than other classification methods.

V.Conclusion

There is no one universal method that fits all of the tasks but with right data pre-processing decision trees and their ensembles can be very efficient and outperform other methods. Better results can be achieved using aggregation of decision tree classifiers like bagging, boosting and Random forests. Decision trees are very attractive for researchers because they are interpretable for experts that don’t have any knowledge about machine learning methods. Taking into account the specific character of the gene expression data, decision trees have another advantage – they are scalable and can work well with data with high dimensionality (they outperformed other methods on full data sets as well as reduced data sets with 200 genes). Decision tree models also allow exploring data structure and provide decision rules. They can provide important information about gene interactions that can be studied further to explain the effect of marker genes. Also the construction process of decision tree models is relatively fast and they are featured in various data mining and analysis tools.

Acknowledgements

This work has been partly supported by Latvia-Belarus Co-operation programme in Science and Engineering within the project ‘Development of a complex of intelligent methods and medical and biological data processing algorithms for oncology disease diagnostics improvement’.

References

[1] **J. W. Lee**, **J. B. Lee**, **M. Park**, **S. H. Song**. “An extensive comparison of recent classification tools applied to microarray data.” Computational Statistics & Data Analysis, Vol. 48, Issue 4, pp. 869-885, Apr. 2005.

[2] **A. Statnikov**, **C. F. Aliferis**, **I. Tsamardinos**, **D. Hardin**, **S. Levy**. “A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis.” Bioinformatics, Vol. 21, Issue 5, pp. 631-643, Mar. 2005.

[3] **B. Twala**, **M. Phorah**. “Predicting incomplete gene microarray data with the use of supervised learning algorithms.” Pattern Recognition Letters, In Press, Corrected Proof, Available online 11 May 2010.

[4] **M. Z. Man**, **G. Dyson**, **K. Johnson**, **B. Liao**. “Evaluating Methods for Classifying Expression Data.” Journal of Biopharmaceutical Statistics, Vol. 14, Issue 4, pp. 1065-1086, Nov. 2004.

[5] **Y. Lu**, **J. Han**. “Cancer classification using gene expression data.” Information Systems Vol. 28, Issue 4, pp. 243-268, Jun. 2003.

[6] **L. Rokach**, **O. Z. Maimon**. *Data mining with decision trees: theroy and applications.* Singapore: World Scientific, 2008.

[7] **J. R. Quinlan**. “Induction of Decision Trees.” Machine learning, Vol. 1, Issue 1, pp. 81-106, 1986.

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

[9] **L. Breiman**, **J. Friedman**, **R. Olshen**, **C. Stone**. *Classification and Regression Trees*. Belmont, CA: Wadsworth Int. Group, 1984.

[10] **R. Kohavi**, **J. R. Quinlan**. “Decision-tree discovery,” in *Handbook of Data Mining and Knowledge Discovery*, W. Klosgen and J. M. Zytkow, Eds. Oxford: Oxford University Press, 2002, pp. 267-276.

[11] **J. R. Quinlan**. “Simplifying decision trees.” International Journal of Man-Machine Studies, Vol. 27, Issue 3, pp. 221-248, Sep. 1987.

[12] **L. Breiman**. “Bagging predictors.” Machine Learning, Vol. 24, Issue 2, 123-140, Aug. 1996.

[13] American Association for Artificial Intelligence. *The Thirteenth National Conference on Artificial Intelligence*, August 4-8, 1996, Portland, Oregon, USA. Menlo Park, CA: AAAI Press; Cambridge, MA: MIT Press, 1996.

[14] **P. Vitanyi**, Ed. *Second European Conference on Computational Learning theory (Lecture Notes in Computer Science Vol. 904)*, March 13 - 15, 1995, Barcelona, Spain. London: Springer-Verlag, 1995.

[15] **S. Dudoit**, **J. Fridlyand**, **T. P. Speed**. “Comparison of discrimination methods for the classification of tumors using gene expression data.” Journal of the American Statistical Association, Vol. 97, Issue 457, pp. 77–87, Mar. 2002.

[16] **I. Mandoiu**, **A. Zelikovsky**, Eds. *Third international Conference on Bioinformatics Research and Applications*, May 7 – 10, 2007, Atlanta, GA, USA. Berlin, Heidelberg: Springer-Verlag, 2007.

[17] **J. Horng**, **L. Wu**, **B. Liu**, **J. Kuo**, **W. Kuo**, **J. Zhang**. “An expert system to classify microarray gene expression data using gene selection by decision tree.” Expert Systems Applications, Vol. 36, Issue 5, pp. 9072-9081, Jul. 2009.

[18] **J. Huang**, **Hong Fang**, **Xiaohui Fan**. “Decision forest for classification of gene expression data.” Computers in Biology and Medicine, Vol. 40, Issue 8, pp. 698-704, Aug. 2010.

[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**, **E. S. Lander**. “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.” Science, Vol. 286, pp. 531-537, Oct. 1999.

[20] **H. Hu**. "Mining patterns in disease classification forests." Journal of Biomedical Informatics, In Press, Corrected Proof, Available online 23 June 2010.

[21] **E. Zintzaras**, **A. Kowald**. "Forest classification trees and forest support vector machines algorithms: Demonstration using microarray data." Comput. Biol. Med., Vol. 40, Issue 5, pp. 519-524, May 2010.

[22] **R. Díaz-Uriarte**, **S. Alvarez de Andrés**. "Gene selection and classification of microarray data using random forest." BMC bioinformatics, Vol. 7, Issue 1, Jan. 2006.

Inese Polaka is first year graduate student at Riga Technical University. She finished her Master studies at Riga Technical University majoring in Information Technology in 2010 taking a Mg. sc. ing. degree.
Her research interests include machine learning methods and classification tasks in bioinformatics, decision tree classifiers, classifier efficiency improvement methods, use of ontology in machine learning, ontology based classifier design, descriptive statistics, and exploratory data analysis.

Igor Tom received the University diploma in 1977 from the Belarusian State University of Informatics and Radioelectronics. Since 1977, he worked as a researcher and was the postgraduate, in 1984-1986, at the Institute of Engineering Cybernetics (IEC) of National Academy of Sciences of Belarus (NASB). In 1986 he received the PhD degree in Computer Science from IEC NASB. From 1977 to 1998 he was an Engineer/Junior Researcher/Senior Researcher/Leading Researcher of Man-Machine Systems Modeling at IEC NASB. Since 1999 he is Chief of the Department Bioinformatics at the United Institute of Informatics Problems of NASB. His current research interests are in the fields of the development of intelligent methods of data analysis and creation of information technologies for medical and industrial applications. He is the author and co-author more than 150 scientific publications, including 50 papers. His last publications are devoted to hybrid intelligent data analysis methods for solving classification tasks, data clustering and revealing of decision rules.
Awards and memberships: Belarus State Award for Works in Military Area (1984), National Academy of Sciences of Belarus Award for Advances in Informatics Area (1998), the member of Belarusian Association of Artificial Intelligence (since 1999), the member of Belarusian Society of Operations Research (since 1998).

Arkady Borisov holds a Doctor of Technical Sciences degree in Control in Technical Systems and the Dr.habil.sci.comp. degree.
He is Professor of Computer Science in the Faculty of Computer Science and Information Technology at Riga Technical University (Latvia). His research interests include fuzzy sets, fuzzy logic and computational intelligence. He has 205 publications in the area.
He has supervised a number of national research grants and participated in the European research project ECLIPS.

Inese Poļaka, Igors Toms, Arkādijs Borisovs. Lēmumu koku klasifikatori bioinformātikā

Rakstā piedāvāts literatūras apskats, analizējot zinātniskos rakstus, kas apskata klasifikācijas koku un to ansambļu metožu izmantošanu klasifikācijas uzdevuma risināšanai bioinformātikā. Apskatīts vēža klasifikācijas uzdevums, kurā nosaka vēža tipu vai pacienta diagnozi (slims vai vesels) pēc gēnu ekspresijas datiem (mikrorežģa formāta dati).
Apskatīti vairāki raksti, kas analizē dažādu klasifikācijas metožu pielietošanas iespējas šādu bioinformātikas uzdevumu risināšanā un salīdzina to veiktspēju, izmantojot dažādas datu kopas un pirmapstrādes pieejas. Klasifikatoru salīdzināšanā ņemts vērā arī īpatnējais datu raksturs – dati satur vairākus tūkstošus atribūtu (gēnu) un salīdzinoši maz ierakstu (daži desmiti vai simti), kas apgrūtina klasisko datu ieguves metožu darbību. Apskatītajos rakstos aprakstītās lēmumu koku metodes šajā rakstā tiek salīdzinātas pēc to efektivitātes (klasifikācijas kļūda/precizitāte), kas uzrādīta vairākās populārās gēnu mikrorežģa datu kopās (leikēmijas, limfomas u.c. datu kopas).
Rakstā arī apskatītas uz lēmumu koku izmantošanu balstītas metodes, kas izmantotas gēnu atlasei. Šādas metodes ir, piemēram, gēnu lietderības noteikšana pēc lēmumu koku klasifikatoru konstruēšanā izmantotās atribūtu informatīvuma novērtēšanas pieejas (Information Gain u.c.) un gadījuma lēmumu koku mežu ģenerēšana, nosakot visbiežāk izmantotos gēnus, kas tiek atlasīti tālākajam darbam.
Kopumā lēmumu koku klasifikatoru veiktspēja ir līdzvērtīga vai pārspēj citas klasiskās metodes, veicot pareizu datu pirmapstrādi. Lēmumu koku klasifikatoru ansambļu veiktspēja lielākoties pārspēj vienkāršu lēmumu koku klasifikatoru veiktspēju, ņemot vērā šādu klasifikatoru nestabilitāti. Lēmumu koku priekšrocība ir arī to vieglā interpretējamība un to spēja atklāt sakarības datos, kas var palīdzēt atklāt gēnu lomu slimības diagnostikā un ārstēšanā.

Инесе Поляка, Игорь Том, Аркадий Борисов. Деревья решений в биоинформатике

В статье предложен обзор литературы, анализ научных статей, которые рассматривают применение методов деревьев решений и их ансамблей для решения задач классификации в биоинформатике. Рассматривается задача классификации рака, которая определяет тип рака или диагноз пациента (больной или здоровый) по данным экспрессии генов (данные формата микрочипов).
Рассматриваются статьи, в которых анализируются возможности применения различных методов классификации в области биоинформатики при решении подобных задач и сравнивается их производительность с помощью различных наборов данных и подходов предобработки. При сравнении классификаторов также принимается во внимание особый характер данных - данные содержат несколько тысяч признаков (генов) и относительно небольшое число записей (несколько десятков или сотен), что осложняет работу классических методов добычи данных. Методы деревьев решений, рассматриваемые в статьях, сравниваются в данной статье по их эффективности (ошибка/точность классификации), показанной в экспериментах с популярными наборами данных генных микрочипов (наборами данных о лейкемии, лимфоме и другими).
В статье также обсуждается использование методов на основе деревьев решений для отбора генов. Такие методы включают в себя, например, использование подходов к оценке информативности атрибутов (Information Gain и т.д.), которые используются при построении классификаторов деревьев решений, и генерацию случайных лесов деревьев решений для определения наиболее часто используемых генов, которые отбираются для дальнейшей работы.
В целом, классификаторы деревьев решений по производительности равны или превосходят другие традиционные методы, производя правильную предварительную обработку данных. Ансамбли классификаторов деревьев решений в значительной степени превосходят простые классификаторы деревьев решений по производительности с учетом нестабильности классификаторов. Преимущество методов деревьев решений заключается в том, что их легко интерпретировать, и они способны обнаруживать взаимосвязи в данных, которые могут помочь определить роль гена в диагностике и лечении заболеваний.