RIGA TECHNICAL UNIVERSITY
Department of Computer Science and Information Technology
Institute of Information Technology

Inese POLAKA
Student of doctoral study program „Information Technology"

EVOLUTIONARY INDUCTION OF DECISION TREE CLASSIFIER ENSEMBLES USING CLASS DENSITY STRUCTURE

Summary of Doctoral Thesis

Scientific supervisor
Dr.habil.sc.comp., Professor
A. Borisovs

Riga 2014

UDK 004.85.021(043.2)
Po 185 e

Poļaka I. Evolutionary induction of Decision tree classifier ensembles using class density structure. Summary of Doctoral Thesis. - R.:RTU Press, 2014. – 37 pages.

Printed according to the decision of the RTU Institute of Information Technology Board meeting, February 7, 2014, Protocol No. 14-2

This work has been supported by the European Social Fund within the project „Support for the implementation of doctoral studies at Riga Technical University”.

ISBN

DOCTORAL THESIS
IS SUBMITTED FOR THE DOCTOR'S DEGREE IN
ENGINEERING SCIENCE AT RIGA TECHNICAL UNIVERSITY

The defence of the thesis submitted for the doctoral degree in engineering science (Information Technology) will take place at an open session at the Faculty of Computer Science and Information Technology of Riga Technical University, in 1/3 Meza Street, auditorium 202, at 14^{30}, on May 19, 2014.

OFFICIAL REVIEWERS

Professor, Dr.habil.sc.ing. Zigurds Markovičs
Riga Technical University, Latvia

Associated professor, Dr.dat. Jānis Zuters
University of Latvia, Latvia

Professor, Dr.sc.ing. Alexander Bozhenyuk
Southern Federal University, Taganrog Technological Institute, Russia

DECLARATION

I hereby confirm that I have developed this thesis submitted for the doctoral degree at Riga Technical University. This thesis has not been submitted for the doctoral degree at any other university.

Inese Poļaka …….
signature

Datums …….

The doctoral thesis is written in Latvian and includes introduction, 5 sections, result analysis and conclusions, 48 tables, 37 figures, overall it consists of 141 pages. The bibliography contains 76 references.

CONTENTS

GENERAL DESCRIPTION OF THE THESIS

Background

The world medicine is moving from symptomatic diagnostics to systems biology approach where biological tests about the person's genetic or immunological profile are used in diagnostics, treatment prognostics and monitoring etc. This approach has to deal with large data amounts dimension-wise but with very few records because the tests are expensive and have not yet been introduced to everyday praxis of health care. None the less this data has to be analyzed to find biological markers that point to disease processes. Therefore the data analysis part of the task involves mathematical and statistical analysis methods (called biostatistics) and other intellectual data analysis approaches including data mining and machine learning methods (bioinformatics).

Actuality

The Human Genome Project was started in 1990 and announced the completed human genome sequence in 2003. Since then researchers in biomedicine could use genes as explanative markers of diseases by identifying the genes and their expression levels in diseased and healthy tissue. This led to gene microarray technologies that allowed scanning thousands of genes in one test. This technology also allowed studying immune responses of individuals by using protein microarrays. All this led to a new systems biology perspective in diagnostics, monitoring and prognostics of diseases and treatments. But there is still a lot of unknown information about genes and proteins, their functions and relations. Therefore the last decade made bioinformatics very popular but the emerging techniques, approaches and methods still are not accurate enough and there is still a lot of knowledge in the biomedical data that is yet to acquire.

Problem statement

It is not possible to analyze the huge amounts of data for humans alone therefore there is a need for computational techniques and methods. Most of the methods used in similar studies have shown good results with single datasets but there is no dominating method that fits equally well to all data sets. Support Vector Machines (SVM) and Naive Bayes method (NB) have shown good accuracies in some of the studies but not equally good in all and the interpretation of the results is close to impossible. There is a small subset of genes or proteins that explain the knowledge about the disease held in the data but these methods do not provide such smaller feature subsets and meaningful description of the obtained results is at this point impossible. Another method popular in bioinformatics is Random Forests and other

tree-based classifiers. Although they show slightly worse classification accuracies, the obtained classification models are easily interpretable, hold a small panel (subset) of informative features (biomarkers) and show relationships among these biomarkers. This knowledge obtained during background and literature study led to a decision to implement decision tree based classifiers. But these should be adapted to work well in the specific data and use the inner structure of the data in the process of building a classifier. The proposed methodology application area is shown in Figure 1.

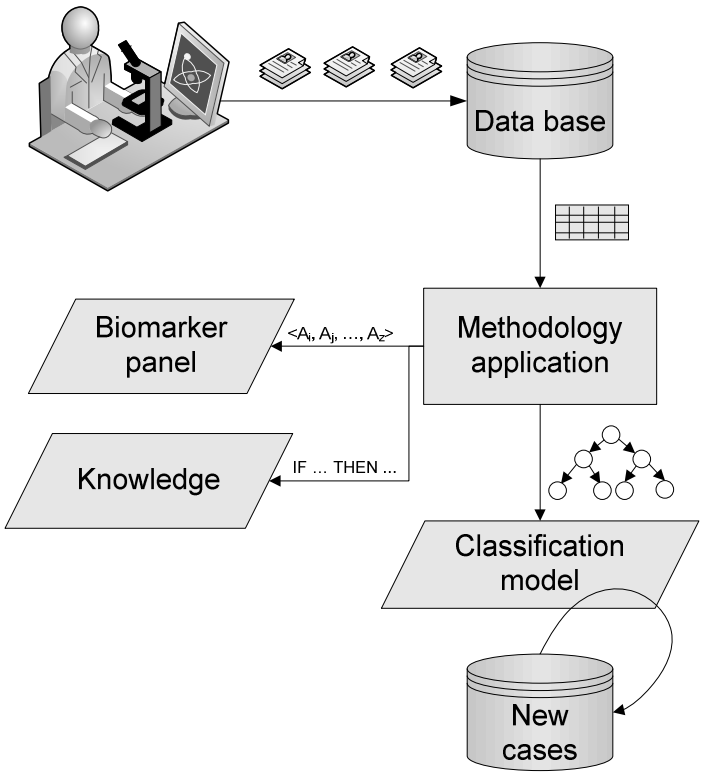

Figure 1. The process of methodology application

Motivation

The motivation to solve the problem using data mining and machine learning methods is based on their specifics:

- They work with non-parametric data without specific demands toward the data;
- They work well with highly dimensional data and some methods have built in feature selection (dimensionality reduction) approach;
- Although small number of records in such high dimensionality is a setback that influences the accuracy of the methods, they work well with the data sets with few records.

The specific data are acquired by analyzing biologic material of patients using microarrays that hold several thousand genes or antibodies (the scanned microarray looks as depicted in Figure 2.

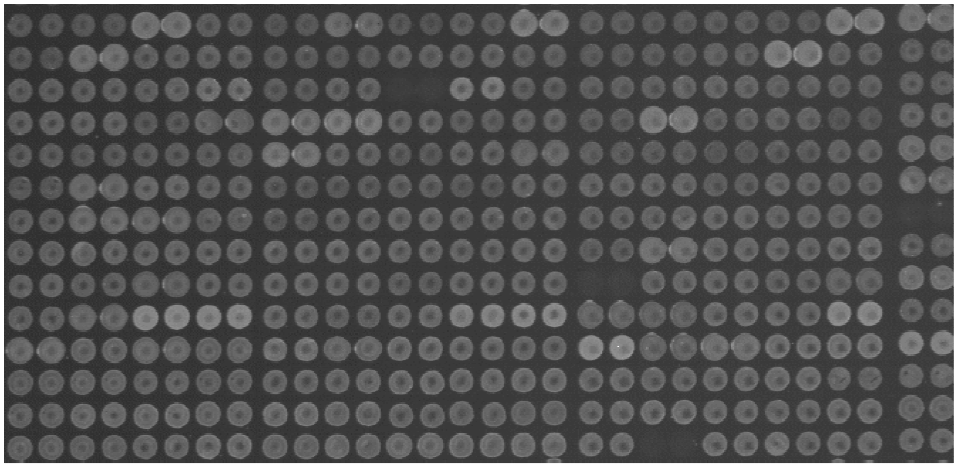

Figure 2. A part of a scanned microarray

The spots of the scanned microarray are transformed to continuous numbers based on the intensity of the red and the green dye added to the genes or proteins before scanning. The data set is then transformed to a matrix where lines represent patients and the results of their tests (x_{ij}) are the values of the continuous number attributes that represent genes or antibodies (see Table 1).

Table 1

Data set with transformed scanned microarray data

ID	$Gene_1$	$Gene_2$	$Gene_3$	...	$Gene_j$
$Patient_1$	x_{11}	x_{12}	x_{13}		x_{1j}
$Patient_2$	x_{21}	x_{22}	x_{23}		x_{2j}
$Patient_3$	x_{31}	x_{32}	x_{33}		x_{3j}
...	...	...	...	...	...
$Patient_n$	X_{n1}	x_{n2}	x_{n3}		x_{nj}

To solve the diagnostic task with high accuracy and quality, a methodology has to be developed that searches for the most accurate classification model and uses the inner structure of the data even if it costs extra calculations and extra time.

Research goal and tasks

The **goal** of the research is to develop bioinformatics classification methodology that uses inner structure of the classes and genetic algorithms to find classification models.

The **tasks** of the thesis set forth to reach the goal are the following:

- Analyse other methods and approaches used in similar studies in this field and described in available scientific literature.
- Develop an approach that allows representing the inner structure of the classes and using it in classifier construction.
- Develop a hybrid method that enables finding a quasi-optimal decision tree ensemble classifier using genetic algorithms.

- Develop a unified methodology that would construct diagnostic models using the developed methods.
- Evaluate the efficacy of the developed methodology, methods and approaches comparing it to the results of the available alternative methods.

Research object and subject

The **object** of the thesis is biomedical diagnostics. The **subject** of the thesis is data mining and machine learning methods.

Research hypotheses

The IT-based technical side of the research is based on the fact stated by biomedical experts: the disease with the same symptomatic manifestation can have different biomedical profiles that slightly vary and can be described by a comparatively small biomarker panel.

In the development of biomedical diagnostic model proposed in this thesis the following hypotheses were defined:

1. There is a significantly smaller attribute set that holds the majority of knowledge about the disease (its diagnostic properties) represented by the data.
2. The use of information about the data inner structure (finding the differing subtypes) improves the performance of classification algorithms that can successfully build complex models that not only hold the information about profiles of different classes but also incorporate the additional information about the structure of the data.
3. In a finite search space the genetic algorithms can find quasi-optimal classifiers that are accurate and easily interpretable for further use in diagnostics.
4. Ensemble classifiers are more accurate in the specific bioinformatics data if new atomic classifiers incorporated in the ensemble add significant information.

The first hypothesis is directly based on the medical fact. The hypothesis will be tested by carrying out sets of experiments with smaller attribute sets (up to 200 features) and evaluating the knowledge left in the smaller data sets by building classification models and assessing their accuracy in the smaller sets as well as in full data sets. If classification accuracy is equal or very close the hypothesis will be considered to be proven.

The second hypothesis is based on the fact that there are inner class structures that allow better discrimination between classes. The drawback here is that the data set is rather small record-wise and these records have to provide information about both –

the original class specifics as well as the inner class structure description, which means that there has to be additional knowledge mined from the data sets. This increases the complexity of classification models that tend to become overfitted or display unnecessary record-based information that does not propagate in the classes. The second hypothesis will be tested by comparing classification accuracies that are obtained in the initial data sets with the accuracies obtained in data sets whose inner structure was used in classifier induction. If the classification results are better in the set of experiments that use inner class structures (decomposed classes), the hypothesis will be considered to be proven.

The third hypothesis is based in the idea that there are optimal classification models that can be obtained using classical algorithms. Therefore by searching the classifier set using genetic algorithms an optimal or quasi-optimal classifier can be found in this set. This hypothesis will be tested by classifying data sets with the classical methods and with genetic algorithms that search for decision tree classifiers and their ensembles. If the found decision tree classifiers and their ensembles would be more accurate or equally accurate while being more easily interpretable (transparent for medical field experts), this hypothesis will be considered proven.

The fourth hypothesis is based on the data set complexity – more complex data sets ask for classifiers that can explain more knowledge and create more complex models. Single tree classifiers get more complex with more depth but this way the classifiers tend to be overfitted to the data due to the small record set. Decision tree ensembles can keep separate classifiers simple while explaining complex structures using several classifiers. This hypothesis will be tested by comparing the classification accuracies of separate trees with that of decision tree classifiers. If the decision tree ensembles will be more accurate, this hypothesis will be considered proven.

Research methods

The research is based on mathematical and statistical analysis, data mining, machine learning, genetic algorithm and experimental study methods. It also uses literature analysis to gain knowledge about the previous researches and current state of art.

Scientific novelty

The scientific novelty of the thesis is based on the methodology that is developed in the research work of the thesis. The methodology proposes two new methods that can be commonly used in similar bioinformatics studies and experiments. The methods are the following:

- The developed approach to use data structure in classification is implemented in data decomposition method that allows presenting the inner structure of the

class in a way that classical data mining and machine learning methods can use it in classifier design.

- The genetic algorithm was modified and adapted to work with decision tree classifiers and their ensembles. The method developed in the thesis can be used for both tasks – searching for single as well as ensemble classifiers.

Practical value

The methodology and separate methods can be used in bioinformatics tasks with similar features – finding relationships and knowledge (diagnostic, prognostic etc.) in the data that would discriminate among different groups. The developed methodology and methods work well with highly dimensional data with few records like gene expression, protein expression and other data. The methods developed in this thesis not only increase the classification accuracy but the resulting classification models are transparent and easily interpretable, which widens the possible application field to tasks that not only need an accurate classification but also explain the knowledge behind its reasoning. The methods are not field-specific but novel in different areas (after a vast literature and similar study research the author has not found similar methods or algorithms that implement the functions of the methods developed in the thesis). The class decomposition method might also become useful in tasks that have complex data and there is a suspicion that the class structure might be more complex – the classes do not have radial forms but rather form different high density areas that are easily described by class decomposition and used in classifier design. This same approach as in class decomposition can be adapted to work with an expert using the same class structure description.

Approbation

The results of the research were presented at the following international conferences:

1. *Riga Technical University 54th International Scientific Conference*, Riga, Latvia, October 14-16, 2013.
2. *European Conference on Data Analysis 2013*, Luxembourg, Luxembourg, July 10-12, 2013.
3. *Applied Information and Communication Technology 2013*, Jelgava, Latvia, April 25-26, 2013.
4. *Riga Technical University 53rd International Scientific Conference*, Riga, Latvia, October 10-12, 2012.
5. *Workshop on Data Mining in Life Sciences*, Berlin, Germany, July 20, 2012.
6. *Applied Information and Communication Technology 2013*, Jelgava, Latvia, April 26-27, 2012.

7. *21st European Meeting on Cybernetics and Systems Research*, Vienna, Austria, April 10-13, 2012.
8. *Riga Technical University 52nd International Scientific Conference*, Riga, Latvia, October 12-25, 2011.
9. *8th International and Practical Conference 'Environment. Technology. Resources'*, Rezekne, Latvia, June 20-22, 2011.
10. *17th International Conference on Soft Computing MENDEL*, Brno, Czech Republic, June 15-17, 2011.
11. *Riga Technical University 51st International Scientific Conference*, Riga, Latvia, October 11-15, 2010.

Publications

The research results of this thesis have been published in 14 scientific articles:

1. Poļaka, I., Borisovs, A. The Application of Class Structure to Classification Tasks. *Information Technology and Management Science*. Nr.16, 2013, 114.-120.lpp. Cited in: VINITI, EBSCO, CSA/ProQuest.
2. Poļaka, I., Borisovs, A. Genethic Algorithm and Tree Based Classification in Bioinformatics// In: European Conference on Data Analysis 2013: Book of Abstracts, conference in Luxembourg, Luxembourg, July 10-12, 2013. – Luxembourg: GFKL, 2013. – p. 107.
3. Poļaka, I. Clustering Algorithm Specifics in Class Decomposition.// In: Applied Information and Communication Technology 2013 (AICT2013): Proceedings of the 6th International Scientific Conference, Latvia, Jelgava, April 25-26, 2013. – Jelgava, Latvia: LLU. – pp. 29-36. Cited in: Thomson Reuters ISI Web of Science.
4. Poļaka, I., Borisovs, A. The Impact of Cluster Stability on Class Decomposition in Antibody Display Data// Information Technology and Management Science. – 2012. – Vol. 15. – pp. 70-75. Cited in: VINITI, EBSCO, CSA/ProQuest.
5. Poļaka, I., Borisovs, A. Class Decomposition in Bioinformatics Analyzing Omics Data. No: Proceedings of Workshop on Data Mining in Life Sciences (DMLS'2012): Workshop on Data Mining in Life Sciences (DMLS'2012), Germany, Berlin, July 20, 2012. – Berlin: Springer-Verlag Berlin Heidelberg, 2012. – pp. 158-167.
6. Poļaka I., Borisovs A. Robust Dimensionality Reduction in Bioinformatics Data // 21st European Meeting on Cybernetics and Systems Research (EMCSR 2012): Book of Abstracts, Austria, Vienna, April 10-13, 2012. – Vienna, Austria: BCSSS. – pp. 286-289.
7. Poļaka I. Genetic Algorithm for Random Tree Generation in Bioinformatics Data // Proceedings of the 5th International Scientific Conference on Applied

Information and Communication Technologies (AICT2012), Latvia, Jelgava, April 26-27, 2012. – Jelgava, Latvia: LLU. – pp. 335-340. Cited in: Thomson Reuters ISI Web of Science.

8. Poļaka I., Borisovs A. Impact of Antibody Panel Size on Classification Accuracy // Scientific Journal of RTU – 2011. – 5th series, Computer science. Vol. 49 – 2010. – pp. 85-90. Cited in: VINITI, EBSCO, CSA/ProQuest.
9. Grabusts P., Poļaka I. Estimation of the Efficiency of Knowledge Acquisition Techniques Using Clustering // Proceedings of the Ninth International Scientific School MA SR, Russia, Saint Petersburg, June 28-July 2, 2011. – Saint Petersburg, Russia: IPME RAS, 2011. – pp. 131-137.
10. Poļaka I., Borisovs A. Impact of Feature Selection on Classifier Testing Validity // Proceedings of the 17th International Conference on Soft Computing MENDEL, Czech Republic, Brno, June 15-17, 2011. – Brno, Czech Republic: MENDEL SCC, 2011. – pp. 411-418. Cited in: Thomson Reuters ISI Web of Science.
11. Poļaka I. Feature Selection Approaches in Antibody Display Data Analysis // Proceedings of the 8th International and Practical Conference, June 20-22, 2011, Volume II, Latvija, Rēzekne, 20.-22. jūnijs, 2011. - 16.-23. lpp.
12. Poļaka I., Borisovs A. Using Data Structure Properties in Decision Tree Classifier Design // Scientific Journal of RTU – 2010. – 5th series, Computer science. Vol. 44 – 2010. – pp. 111-117. Cited in: VINITI, EBSCO, CSA/ProQuest.
13. Poļaka I., Tom I., Borisovs A. Decision Tree Classifiers in Bioinformatics // Scientific Journal of RTU – 2010. – 5th series, Computer science. Vol 44. – pp. 118-123. Cited in: VINITI, EBSCO, CSA/ProQuest.
14. Poļaka, I., Borisovs, A. Clustering-Based Decision Tree Classifier Construction. Technological and Economic Development of Economy, 2010, Vol.16, Iss.4 – pp. 765-781. Cited in: Taylor&Francis.

Main results

The main results are the following:

- Analysis of other methods and approaches used in similar studies in this field was carried out.
- An approach that allows describing the inner structure of the classes and using it in classifier construction was developed.
- A hybrid method that searches for optimal decision tree ensemble classifier using genetic algorithms was developed.

- A unified methodology that would construct diagnostic models using the developed methods (class decomposition and genetic algorithm based decision tree classifier ensemble construction) was developed.
- The efficacy of the developed methodology, methods and approaches was evaluated and compared to the results of the available alternative methods drawing conclusions about the possibilities of the developed methods and methodology.

Structure and contents of the thesis

The **First section** describes the definitions of the tasks being solved in the study as well as field background and working specifics working with bioinformatics tasks.

The **Second section** includes analysis of similar studies featured in scientific articles that are available in various databases – their tasks and proposed solutions, methods and algorithms. It also gives the information about the most popular and accurate methods that are used in studies with similar field, specifics and tasks.

The **Third section** presents the overview and detailed description of the popular methods found in the Second section as well as methods and approaches that are used in this thesis to build the proposed methodology.

The **Fourth section** describes the developed methodology, explains the used approaches and gives detailed information about the methodology and its implementation for the design of diagnostic models.

The **Fifth section** describes the empirical study – experiment sets, the reasoning behind experiment design and the results of the experiments that were carried out to test the hypotheses of this research and evaluate the developed methods, approaches and methodology comparing them to the other popular methods in bioinformatics. It also gives a more detailed analysis of classification accuracy dependence on feature panel size, improvements gained using class decomposition, comparison between efficacy of similar methods and the developed genetic algorithm based classification method and the methodology. This section also gives reasoning behind choices of parameters, approaches and selection of methods.

THE SUMMARY OF THESIS CHAPTERS

1. BIOMEDICAL DIAGNOSTICS USING BIOINFORMATICS

This section describes the process of obtaining biomedical data and definition of the tasks of the thesis or its parts. The methods chosen for the thesis task solution are decision tree based classifiers (due to their interpretability and built-in feature subset selection) that are induced using genetic algorithms to find optimal or quasi-optimal classifiers and the description of the data structure. The inner structure of the data is analyzed to find high density areas that would describe different subtypes of the same disease and that can be used in decision tree based classifier design. To describe the inner structure of the data, the classes are decomposed by solving a cluster analysis task.

Thesis task definition

The task being solved in this thesis using computer science methods is the systems biology (biomedical) diagnostics task. Given are gene or protein expression data as well as diagnosis for each record (gene/protein expression value vector). The solution is a model describing the gene/protein groups (biomarker panels) that points to the specific diagnosis of each gene/protein vector that has been defined by a 'golden standard' method (medical method different from the genetic/proteomic method).

So the aim of the diagnostics task is to find knowledge in the data about biomarkers that are underlying different diseases and diagnoses. This knowledge is discovered in this thesis using data mining and machine learning approach because of their low requirements towards data and adaptation possibilities to the bioinformatics task specifics.

Each data mining task is defined by primitives [22] that describe the task. The task of this research is defined by the following primitives:

- Task data: thousand and more data attributes with continuous values (gene or protein levels), diagnosis as the target attribute (disease or healthy donor label) as well as records that correspond the patients and their tests (one test of one patient is a vector of all attribute values and a label of the target attribute).
- Knowledge to be discovered: classification model that associates attribute values and their relationships with target class.
- Related knowledge: the data is normalized using background and noise levels to equalize signal strengths across different tests.
- Evaluation of the discovered knowledge and profiles: the discovered profiles and classification (diagnostic) models are evaluated using classification accuracy; the results should also be interpretable, forming a biomarker panel.

- Visualization of the discovered models: the implemented classification approach is based on decision tree classifiers due to their accuracy and interpretability, therefore the visualization is a tree graph where nodes are attributes (biomarkers – genes or proteins), arches are split values and leaves are target classes.

Formal definition of the classification task

Whereas the data has gene or protein expression levels and target class labels, the diagnostic model induction is performed using supervised learning. This means that the task is a classification task.

In classification task there is given a data set with records $x_i = x_i^{A1}, x_i^{A2} \ldots, x_i^{Ai} \ldots, x_i^{An}, x_i^C$, which are value vectors of the attributes *A1, A2, …, An* (gene or protein expression levels) and the target attribute *C*. The solution is a classification model that maps the target attribute value using the vector of other attributes: $x_i^C = f(x_i^{A1}, x_i^{A2} \ldots, x_i^{Ai} \ldots, x_i^{An})$. The aim of the task is assigning the target attribute value to previously unseen attribute value vectors.

During learning or classifier construction phase the algorithm uses data set *X*, where the values of attributes *A1 … An*, and target attribute *C* are given. Then the algorithm searches for relations that map the vector $x_i = x_i^{A1}, x_i^{A2} \ldots, x_i^{Ai} \ldots, x_i^{An}$ to the target attribute value set *C*. This mapping or function is the resulting classification model also called classifier. When all attribute value vectors (records) of the learning data set have been used to find functions or rules that map the vectors to the class set *C*, the induced classifier is evaluated using an unseen test record set that holds the vectors $x_j = x_j^{A1}, x_j^{A2} \ldots, x_j^{Ai} \ldots, x_j^{An}$, which have to be mapped to the result class set *C* using the previously induced classifier.

The geometric interpretation of a classification class in a two-dimensional space is given in Figure 3. Different data points (vectors or records) are depicted in two-dimensional space according to two attributes (*x* and *y*). The dashed lines show classifier hyperplanes that separate different classes (depicted as black and grey dots). The right pane shows a picture of a new record (white dot) being classified according to the same classifier (hyperplanes). Whereas the unclassified dot belongs to the area assigned to the dark class, the new record is classified as belonging to the dark class.

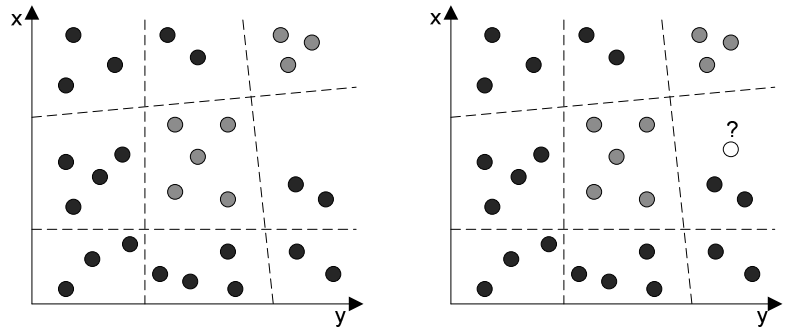

Figure 3. Classifier visualization and class assignment to a new record

2. SOLUTION OF BIOMEDICAL DIAGNOSTICS TASK USING BIOINFORMATICS METHODS

This section describes other studies in the field of bioinformatics solving biomedical diagnostics task. Initially all biomedical data were analyzed using statistics methods but since Golub et al. study in 1999 [16] the data mining and machine learning approaches have been becoming ever more popular in the analysis of the data to discover new knowledge and diagnostic and prognostic models [62].

Classification

Golub et al. [16] used classification and cluster analysis approaches in the analysis of gene expression data to discover relations and knowledge about leukaemia. This was the breaking point in biomedical data analysis and has since been cited more than 800 times in various biomedical and bioinformatics articles in IEEE, ACM and other journals.

Since then a lot of different machine learning methods have been applied to gene and protein expression data. Many of the researchers prefer decision tree based methods including ensembles due to their accuracy and interpretability [13, 29, 30, 31, 42, 43]. But the best accuracy is most often achieved using support vector machines (SVM) and Naive Bayes classifier based algorithms but they do not give useful information about biomarkers and relations in the data [42, 43, 74].

Many of the studies are also dedicated to feature selection task, which improves the classification accuracy but also loses information and transparency [13, 41, 42, 74].

Cluster analysis

Golub et al. [16] also pointed out that there are morphologically similar diseases with different pathogeneses that also inspired the class decomposition approach developed in this thesis. The different subtypes of the diseases present similar symptoms but have different responses towards treatment. This once again emphasizes the need for class structure analysis.

Cluster analysis is often used in bioinformatics in the task of class discovery, which is similar to classification task but approaching it from a different side, without using the known classes. Most often this approach is used for different treatment outcome data – the researchers are looking for subgroups of patients that would explain different treatment outcomes and the most popular method is hierarchical clustering [2, 47, 66, 73], which is also often used for attribute cluster analysis.

Genetic algorithm and decision tree based classifier hybrid methods

There are surprisingly few studies about decision tree classifier and genetic algorithm hybrids. The most popular algorithm and tool GATree that was developed until 2010 [34, 48] uses highly adapted genetic algorithm with different representation of the trees and therefore changed operators. They also use search over the full classifier space that is next to impossible in bioinformatics data with its high dimensionality. Also other proposed approaches involved complex tree coding and full classifier space search that is too time-consuming for bioinformatics [1, 3, 15, 19].

3. MACHINE LEARNING METHODS USED IN BIOINFORMATICS

The third section gives description of classical data mining and machine learning methods used in the thesis – both in development of the proposed methods and for comparative analysis. The first subsection describes the pre-processing step, the second gives detailed description of overcoming 'the curse of dimensionality' [55, 57, 59, 60]. The third subsection describes the most popular classification methods in bioinformatics while the fourth subsection describes the clustering methods. The fifth section gives detailed information about genetic algorithms.

The proposed class decomposition method implements cluster analysis [58, 61, 63]; therefore the most popular clustering methods (see literature review in the second section) are described in this section:

- k-means clustering [70] that constructs clusters based on their centres and the distance between an object and the centre (it is added to the closest cluster);
- hierarchical clustering [11] that implements stepwise merging (or splitting) of the objects and clusters based on the closest (or furthest) distance.

The distance used in the comparative analysis of the methods and in cluster analysis that is a part of class decomposition is the Euclidean distance [17]. The metric for the distance between clusters is Ward's distance [72].

The proposed classification method uses decision tree induction algorithm that is based on the C4.5 algorithm [64] using Information Gain metric in determining of the splitting attribute and value but limiting the tree structure to a binary tree.

The proposed classification method is also based on genetic algorithm [26] that is depicted in Figure 4.

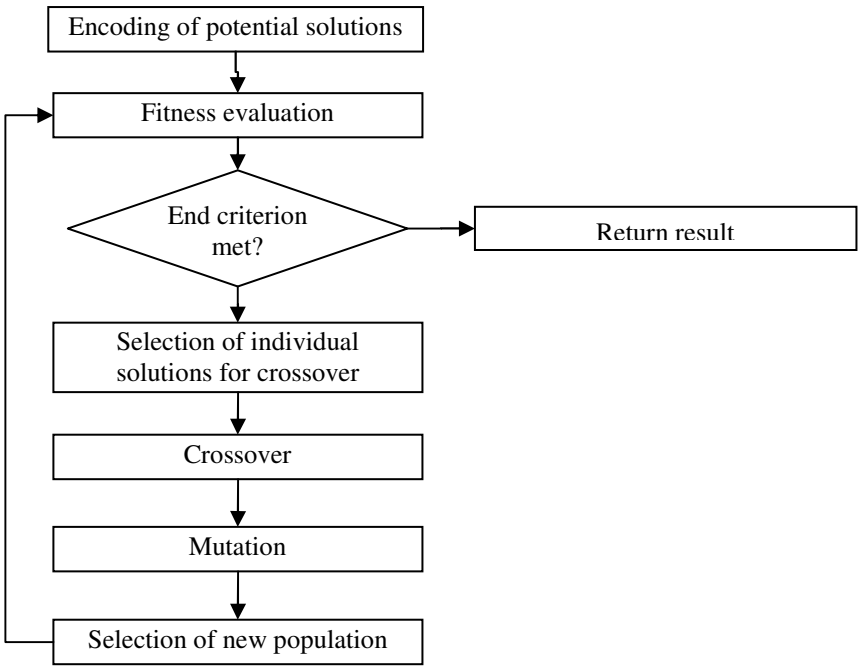

Figure 4. The diagram of genetic algorithms

The comparative analysis of the developed methods and methodology is carried out using four classification methods that are popular and accurate in bioinformatics (see literature review in Section 2):

- Naive Bayes (NB) method [32] that uses probabilistic approach that uses all attributes without any dimensionality reduction;
- Support Vector Machines (SVM) [6] is the most popular and accurate method in most of the studies even though the models induced by SVMs are very complex and close to non-interpretable to medical staff with such high dimensionality;
- C4.5 [64] is the most popular decision tree classifier in the analyzed bioinformatics studies; it has a built-in feature selection and the induced models are transparent and easy to understand making them also useful for medical staff;
- Random Forest [7] is an ensemble of decision tree classifiers that also uses the Random Subspace method (like the proposed method) and shows very good accuracies in different studies.

4. DEVELOPMENT OF A MACHINE LEARNING METHOD BASED METHODOLOGY FOR BIOMEDICAL DIAGNOSTIC MODEL INDUCTION

The methodology will induce diagnostic models using machine learning methods. Therefore the necessary steps in methodology are the following:

- Data preparation and pre-processing;
- Class decomposition [54];
- Classifier induction (finding the most accurate decision tree based classifiers using genetic algorithms) [51];
- Classifier testing and accuracy evaluation [50];
- Result interpretation.

The process of the methodology and its steps is depicted in Figure 5. The developed methods are indicated by double line around the box.

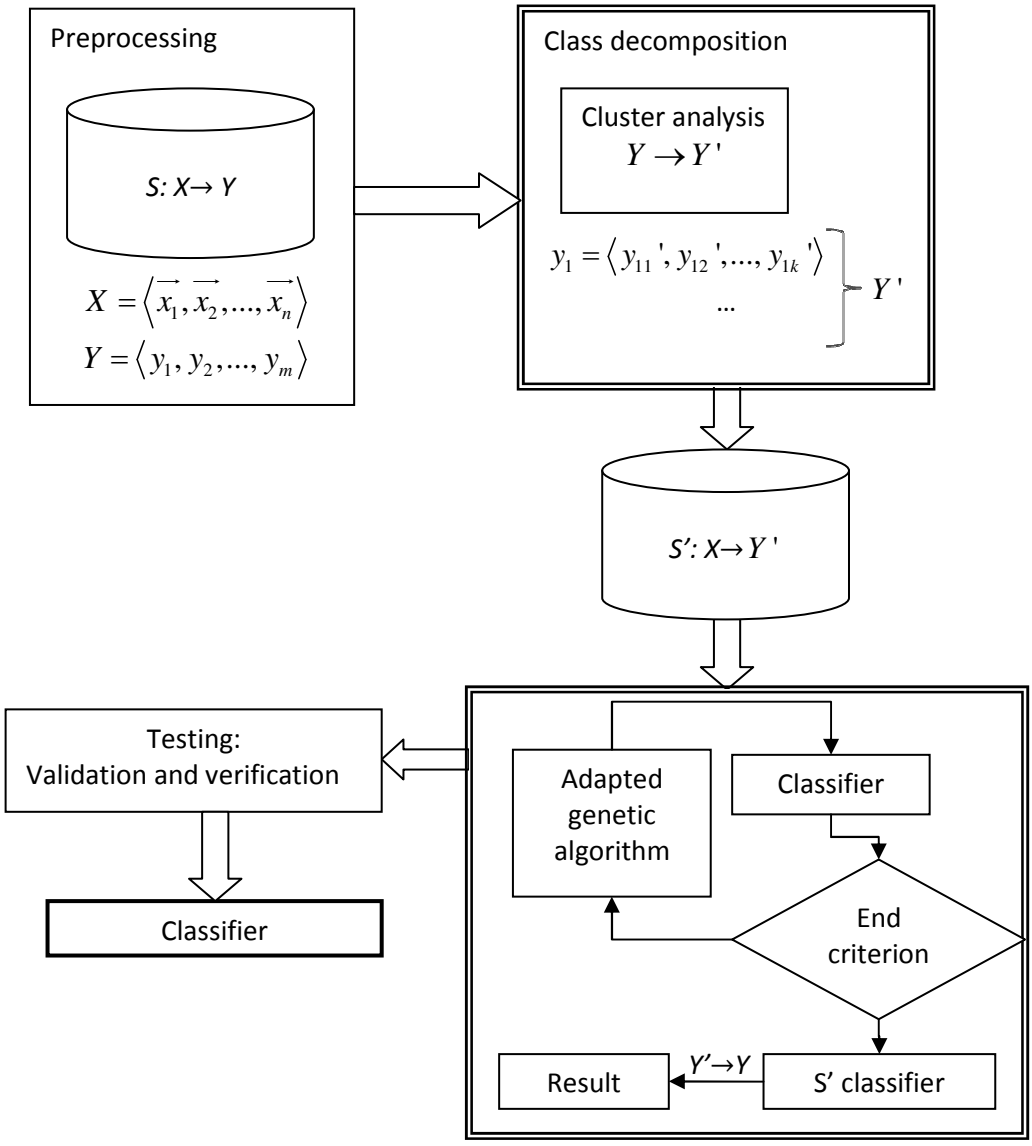

Figure 5. The proposed methodology

Data preparation and pre-processing

The data obtained using gene expression microarrays or phage displays are first scanned in as different intensity points and are first normalized using the software that comes with the scanner to equalize the intensity of points in microarrays and over different tests as well as normalized against background intensity. The scans are transformed to data tables with continuous values for each of the points (representing genes, proteins etc.).

The obtained data then are transposed according to gene or protein mapping to tables that hold data of different patients. These data sets are then merged with those of comparative patient groups according to a unified gene/protein mapping. The names of proteins/antibodies in the data sets donated by Latvian Biomedical Study and Research centre are pre-processed to replace antibody names with identifiers due to the sensitive nature of this information.

The missing data in the data sets are replaced with the values that are calculated using the two closest neighbours according to the first level Minkowski distance. The missing value of attribute A_n in the value vector x_i, whose closest neighbours are x_a and x_b is calculated as follows:

$$A_n^{x_i} = \frac{A_n^{x_a} + A_n^{x_b}}{2} \quad (1)$$

Class decomposition

Taking into account that there are different subtypes of diseases proven in recent studies, the inner structure of the positive classes is analyzed to find high density areas (using cluster analysis) that are used as subclasses of the initial classes. The task of the thesis is not finding real verified disease subtypes but rather making the classification task simpler and therefore increasing the accuracy of classifiers using natural object subgroups. The process of finding disease subtypes using cluster analysis in this thesis is called class decomposition.

Figure 6 shows a hypothetical situation with positive class (triangles), decomposed into two subclasses for easier classification, and negative class (quadrangles).

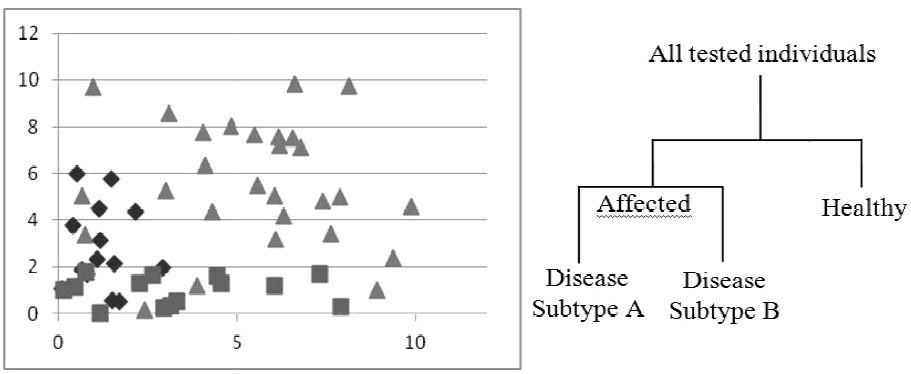

Figure 6. Clustering result in the attribute space and dendrogram

The difference between the most simple decision tree based classifier for one positive class is shown on the left-hand side of Figure 7 and the most simple decision tree based classifier for subclasses is shown on the right-hand side. The second graph has less false positive cases and therefore increased sensitivity and overall accuracy.

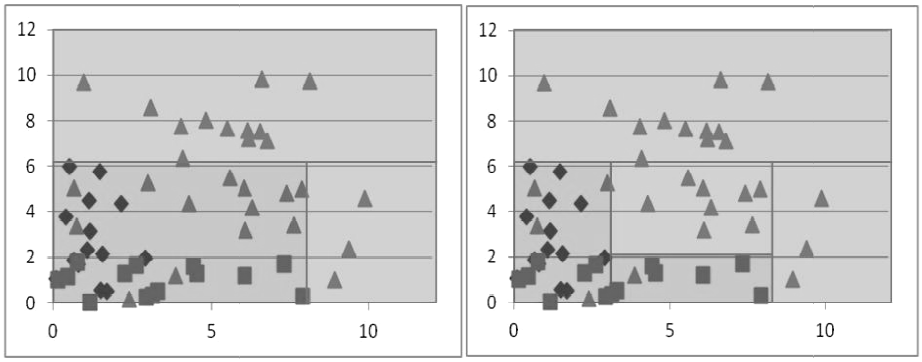

Figure 7. Classification hyperplanes that show the classes:
classical classifiation on the left; classification for decomposed classes on the right

Whereas the class decomposition process uses natural high density areas of the positive class objects, they are found using cluster analysis. The steps of class decomposition are the following:

1. Data set preparation;
2. Splitting the data set according to classes;
3. Clustering of the positive record subsets, finding high density areas that represent biomedical disease subtypes that differ in one class;
4. Assigning labels to the high density areas that can be used in classification task;
5. Merging the data set;
6. Using different subtypes in the further analysis (classification, monitoring, prognostics etc.) of the data set and interpreting them at the end of the analysis process.

The cluster analysis is performed using hierarchical agglomerative clustering. The analysis of other studies showed k-means and hierarchical clustering approaches to be the most popular in bioinformatics but the method analysis process carried out in the thesis showed that k-means algorithm found outliers and made clusters of few records that were outliers [52]. The results of hierarchical cluster analysis proved to find larger object groups that can be interpreted as subclasses.

The data set prepared in the pre-processing phase contains $n = n_a + n_b$ results of antibody/gene tests where n_a are tests of cancer patients and n_b are tests of healthy donors. One test equals a vector $\vec{x} = \{x_1, x_2, \dots, x_m\}$ that holds expression of m genes/antibodies. The data set n_a is split into object groups that correspond to the density areas determined by the cluster analysis process. As a result of clustering each

vector $\vec{x}$ of n_a belongs to exactly one density area (subclass) C_i so that the records of the same group are more similar to other objects in the same subclass than to objects in another subclass. The measure of the similarity in this case is Euclidean distance. And the difference between two subclasses is measured according to Ward's method [56].

To determine the number of clusters, the distances between objects and clusters will be used in the following manner for cluster set S with clusters C_{m} that hold records x_{n}:

$$\max_{C_i, C_j \in S} (d(C_i, C_j)) \tag{2}$$

$$min_{x_a, x_b \in C_i} (d(x_a, x_b)) \tag{3}$$

The quality of the obtained clusters will be tested using Gap statistic [55] and cluster stability (robustness) over 20 iterations [53].

Classification method

Whereas the data used in the research have very high dimensionality, the classification method used to analyze the data has to be scalable. It also has to be transparent so that the results can be interpreted and used in biomedicine.

Decision tree based classifiers are scalable and have shown good accuracy (see previous chapters); they also build data models that are easy to visualize and interpret, finding panels of meaningful biomarkers and displaying relationships among them. Decision tree based classifiers have also been proven to be robust to noise and attribute value scales and other descriptive parameters.

The developed methodology uses an additional method called Random Subspace method to deal with highly dimensional data (also used in Random Forests with different implementation). This method is modified for the task and forms a gene pool for every gene in a chromosome in the genetic algorithm based classification providing a finite and rather small search space for the genetic algorithm [56]. Using different gene pools for different decision tree classifiers in an ensemble allows using the most informative attributes without overfitting. Another mechanism that was built in into the methodology developed in this thesis is setting limits on separate decision tree classifier sizes (depth levels). This mechanism is based on Occam's razor principle that the less complex/smallest classifier is most likely to be the true representation of data.

The algorithm schema of the developed classification method based on genetic algorithms and decision tree ensembles is shown in Figure 8.

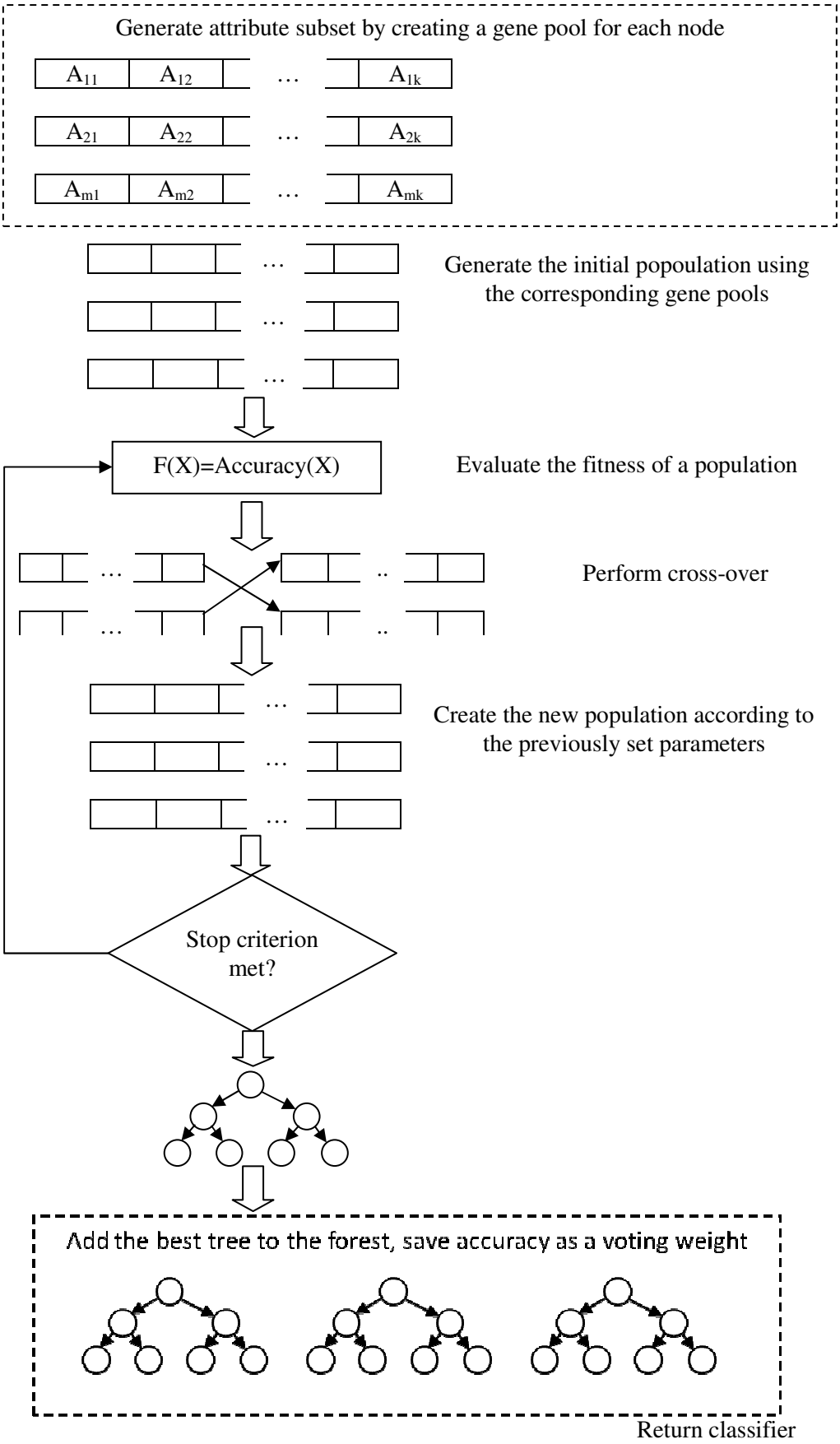

Figure 8. The poposed classification algorithm

The genetic algorithm is used in the developed methodology to find the best (most accurate) decision tree classifiers. Each single classifier is coded in the chromosome using the attributes in each node and then finding the split values based on entropy measures (see Figure 9). Decision tree depth is a parameter that is set a priori by the user. Also the number of the trees in the classification ensemble and genetic algorithm parameters like the size of population, mutation and crossover probabilities are parameters that are set before classification.

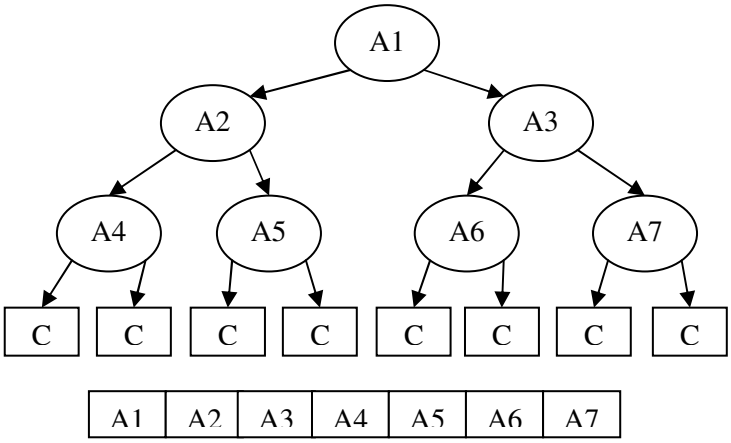

Figure 9. Encoding of a decision tree classifier

Decision tree ensembles are used because of their ability to represent complex data but coding a whole ensemble and work with decision tree ensembles is difficult and resource-consuming. Therefore ensembles are constructed from individual accurate trees built from different random attribute subsets. When determining a class for a new record, all of the trees vote and their votes are weighted by their classification accuracies on test sets.

The positive trait of genetic algorithms is that they use random changes in the classifier, which is also their negative trait because results are unstable across different runs. To evaluate the methodology and the developed algorithm 100 runs are used and the result is approximated from these runs, finding the best properties.

The result interpretation is also different for this method. Whereas all different subtypes of the disease have to be classified as positive, the classical evaluation method – the confusion matrix, has to be adapted for these specifics and the case for three positive subtypes (+1, +2, +3) and a negative class value is shown in Table 2. In this case the total classifier accuracy is calculated as follows:

$$\frac{TP|T^{+1} + TP|F^{+1} + TP|T^{+2} + TP|F^{+2} + TP|T^{+3} + TP|F^{+3} + TN}{TP|T^{+1} + TP|F^{+1} + TP|T^{+2} + TP|F^{+2} + TP|T^{+3} + TP|F^{+3} + TN + FP + FN}$$

Classifier sensitivity or the rate of true positive recognition is calculated as follows:

$$\frac{TP|T^{+1} + TP|F^{+1} + TP|T^{+2} + TP|F^{+2} + TP|T^{+3} + TP|F^{+3}}{TP|T^{+1} + TP|F^{+1} + TP|T^{+2} + TP|F^{+2} + TP|T^{+3} + TP|F^{+3} + FN}$$

Classifier specificity or the rate of true negative recognition is calculated as follows:

$$\frac{TN}{TN + FP}$$

Table 2

Confusion matrix for the decomposed classes

		Assigned value			
		+1	+2	+3	-
Real value	+1	TP|T+1*	TP|F+2**	TP|F+3	FN***
	+2	TP|F+1	TP|T+2	TP|F+3	FN
	+3	TP|F+1	TP|F+2	TP|T+3	FN
	-	FP#	FP	FP	TN##

*True positive, true +1; **True positive, false +2;
***False negative; #False positive; ##True negative

5. EXPERIMENTAL ANALYSIS

To test the hypothesis number one that small attribute subsets can be used in diagnostics more efficiently than the full data sets, the accuracy of the most popular classification methods will be tested on the full and the reduced data sets for comparative analysis. The dimensionality will be reduced to 10, 20, 50, 100 and 200 attributes, using subset selection and ranking feature selection approaches. The difference between the full data set error and the best reduced data set error (using the most informative attributes) is presented in Table 3. The classification method/algorithm names are abbreviated as follows: NB – Naive Bayes classifier, SVM – Support Vector Machines (Sequential Minimal Optimization algorithm), C4.5 – C4.5 decision trees based on J48 algorithm, RF – Random Forest; the row that ends with avg is the average classification accuracy increase for this data set; the data set Gastro-intestinal inflammatory disease is abbreviated as GIS.

The average classification results across all classification methods show accuracy improvements in all data sets except melanoma antibody data set. But even in this data set the reduction of the dimensionality from 1229 attributes to 200 attributes costs in average 0,47% drop in the classification accuracy. Although the error increases in some of the other cases (few method-data set combinations), the error rate drops in the most of the cases, which proves hypothesis number one.

The main developed method for classification efficacy improvement is class decomposition. This method can also be used with other classification methods than the one proposed in this thesis, therefore it is tested with the classification methods that are used for comparative analysis in this thesis. The average results are presented

in Figure 10 (‘(a/b)’ is short for antibody data set; ‘(g)’ is short for gene expression data set).

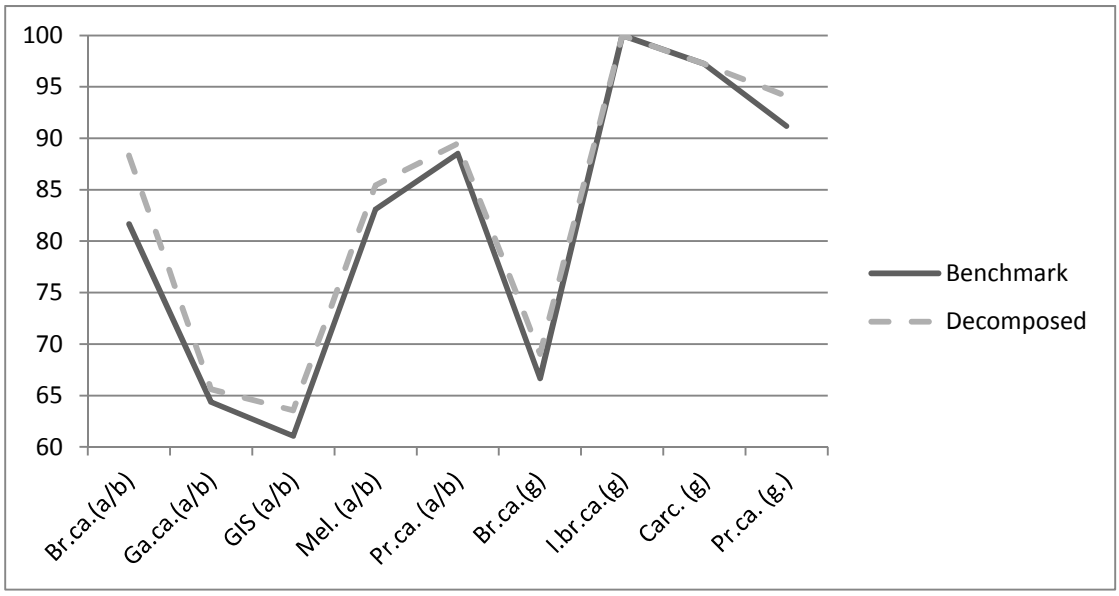

Figure 10. The best classification accuracy in each data set with and without class decomposition

The overall results show improvement in the average accuracy across all classification methods in all data sets except inflammatory breast cancer gene expression data set where classification accuracy is 100% for both cases and the small carcinoma data set where results are equal, which proves hypothesis number two.

Table 3

The growth of classification accuracy after attribute set reduction

Data set	No. of attributes in the reduced subset				
	10	20	50	100	200
Breast cancer NB	**1,79**	**2,5**	**2,38**	**2,26**	**2,62**
Breast cancer SVM	-0,71	0,48	0,36	**1,43**	0,83
Breast cancer C4.5	**1,67**	**1,19**	0,83	-0,71	-1,67
Breast cancer RF	-0,24	0,24	0	0,48	0,12
Breast cancer avg	0,63	1,1	0,89	0,86	0,48
Gastric cancer NB	-4,7	-3,66	-3,66	0,98	0,12
Gastric cancer SVM	-11,65	-6,65	-6,77	-6,22	-3,48
Gastric cancer C4.5	-4,76	-0,3	-0,61	-2,13	-0,3
Gastric cancer RF	**3,35**	**4,88**	**5,06**	**4,27**	**6,65**
Gastric cancer avg	-4,44	-1,43	-1,49	-0,78	0,75
GIS NB	0,5	**2,06**	**5,2**	**2,35**	**3,77**
GIS SVM	-3,84	0	**2,49**	**2,92**	**2,35**
GIS C4.5	0,43	**1,71**	0,64	0,14	**1,21**
GIS RF	0,28	**2,56**	**4,56**	0,85	**6,55**
GIS avg	-0,66	1,58	3,22	1,57	3,47

Data set	No. of attributes in the reduced subset				
	10	20	50	100	200
Melanoma NB	-7,11	-7,99	-7,52	-4,49	-1,17
Melanoma SVM	-16,03	-13,18	-6,65	-2,68	-2,62
Melanoma C4.5	**1,98**	**3,21**	**1,69**	**2,86**	0
Melanoma RF	-6,65	-4,55	-1,17	**1,87**	**1,92**
Melanoma avg	-6,95	-5,63	-3,41	-0,61	-0,47
Prostate cancer NB	-7,54	-2,9	-1,64	-2,42	0
Prostate cancer SVM	-11,4	-8,7	-5,99	-5,41	-1,84
Prostate cancer C4.5	**4,93**	**2,9**	**4,73**	**5,12**	**1,35**
Prostate cancer RF	-3,77	-0,58	**5,12**	**3,48**	**2,8**
Prostate cancer avg	-4,44	-2,32	0,56	0,19	0,58
Average	-3,17	-1,34	-0,05	0,25	0,96

The efficacy of class decomposition is also linearly dependent on subclass stability (the robustness of clusters). The subclass instability and the corresponding maximum accuracy are shown in Table 4 ('a/b' is short for antibody data set, 'g' is short for gene expression data set).

The Pearson correlation coefficient for class instability and maximum accuracy increase is -0,76 at $p<0,05$. This leads to a conclusion that more stable subclasses lead to higher classification accuracy.To test the developed genetic algorithm based decision tree induction method, the initial gene expression and antibody data sets were initially split into training and test sets for 10 folds of cross-validation to eliminate the influence of randomly chosen training and test sets.

Table 4

The maximum increase in classification accuracy
and the corresponding cluster instability

Data set	Max accuracy increase	Average displaced objects
Breast cancer (a/b)	15,39	0,00
Gastric cancer (a/b)	2,50	0,04
Gastro-intestinal disease (a/b)	13,93	0,01
Melanoma (a/b)	2,62	0,04
Prostate cancer (a/b)	1,00	0,33
Breast cancer (g)	4,77	0,01
Inflammatory breast cancer (g)	3,12	0,02
Carcinoma (g)	8,33	0,00
Prostate cancer (g)	-	0,03

The results of classification (average accuracy across all 10 folds) are presented in Table 5; the abbreviations of classification methods: GACT – the developed Genetic Algorithm generated Classification Tree, GARF – the developed Genetic Algorithm generated Random Forest, NB – Naive Bayes method, SVM – Support Vector Machines, RF – Random Forest.

Table 5

The classification accuracies of the developed and the most popular classification methods

Data set	GACT	GARF	NB	SVM	C4.5	RF
Br.c.(g)	**78,57%**	**80,95%**	**78,57%**	69,05%	66,67%	64,29%
I.br.c.(g)	98,89%	**100,00%**	84,44%	54,44%	72,22%	**100,00%**
Br.c.(a/b)	**92%**	**93%**	88%	88%	**92%**	83%
Carc.(g)	80,56%	**94,44%**	86,11%	**100,00%**	86,11%	83,33%
Ga.c.(a/b)	59,33%	63,00%	**65,33%**	**66,00%**	59,00%	55,00%
GIS (a/b)	60,00%	**67,14%**	55,36%	59,64%	55,36%	**85,00%**
Mel.(a/b)	78,82%	**81,18%**	73,24%	79,41%	**81,18%**	**97,35%**
Pr.c.(g)	82%	**93%**	66%	**93%**	83%	82%
Pr.c.(a/b)	79%	84%	83%	**87%**	78%	**94%**

The cells with the bold text (top results) represent one of the two best classification accuracies for each data set (if the second and the third best results are equal, both are emphasized using bold text).

Genetic Algorithm generated Classification Trees (GACT) or one tree ensembles show top results in only two data sets, because most of the data sets are too complex to be represented by a single decision tree with a limited depth. Genetic Algorithm generated Random Forests (GARF) achieve top results in seven out of nine cases (the best result in four cases). This shows that the results of the developed classification method without class decomposition are comparable to the accuracy of the most popular and accurate methods, while still being easily interpretable, which proves hypotheses number three and four.

The results of the developed methodology (class decomposition with Genetic Algorithm generated Random Forests and the specific case of one tree ensembles) are shown in Table 6. The experimental setup is the same as previously comparing the results of the whole methodology instead of just classification methods.

Table 6

The classification accuracies of the developed methodology
and the most popular classification methods

Data set	GACT	GARF	NB	SVM	C4.5	RF
Br.c.(g)	**85,71%**	**85,71%**	78,57%	69,05%	66,67%	64,29%
I.br.c.(g)	98,89%	**100,00%**	84,44%	54,44%	72,22%	**100,00%**
Br.c.(a/b)	**92%**	**97%**	88%	88%	**92%**	83%
Carc.(g)	97,22%	**100,00%**	86,11%	**100,00%**	86,11%	83,33%
Ga.c.(a/b)	60,67%	**66,00%**	65,33%	**66,00%**	59,00%	55,00%
GIS (a/b)	61,43%	**69,29%**	55,36%	59,64%	55,36%	**85,00%**
Mel.(a/b)	81,47%	**82,65%**	73,24%	79,41%	81,18%	**97,35%**
Pr.c.(g)	87%	**94%**	66%	**93%**	83%	82%
Pr.c.(a/b)	83,00%	**90,50%**	82,50%	87,00%	78,00%	**94,00%**

The table shows improvement in the first two columns, which is due to the implementation of the class decomposition in the methodology. The GARF based methodology using class decomposition is one of the top results in all data sets (the best result in six out of nine data sets), which proves that the methodology and the methods developed in the thesis are not only more accurate but also create easily interpretable classification models that can be used in further research in the field of application (biomedicine in this case).

RESULTS AND CONCLUSIONS

The goal of the thesis was to develop bioinformatics methodology that uses data structure description and genetic algorithms to construct classification models. The goal was effectively achieved; and in the process the following steps were completed giving the following results:

- An analytical study of similar researches and studies revealed the most popular and accurate approaches and methods in this field:
    - Naive Bayes classifier, Support Vector Machines, C4.5 and Random Forests in classification,
    - k-means and hierarchical methods in clustering.
- An approach that describes the inner structure of a class and that can be used in classification process was developed for use in bioinformatics classification (diagnostic) task and experimentally tested evaluating its parameters and effect on classification;
- A hybrid classification method based on decision tree classifier ensembles and genetic algorithms was developed and experimentally tested on bioinformatics data with very high dimensionality;
- A unified methodology that implements the developed methods and approaches was proposed and tested on bioinformatics data with very high dimensionality;
- A comparative analysis was carried out and conclusions were drawn about the performance of the developed approaches, methods and methodology.

All developed methods and the methodology were experimentally analyzed to test the hypotheses defined for this study:

- The first hypothesis was proven by reducing (by selection) the feature set and testing the preserved information by constructing classification models; the classifiers built on the reduced data sets were as accurate (no information loss) or more accurate (reduced noise and redundancy), which leads to a conclusion that only a part of the full attribute set (gene or antibody panel) is necessary to describe the significant patterns in the data;
- The second hypothesis was proven by constructing classifiers in the initial data set and in the data set where class inner structure was described; the results showed class accuracy increase in 8 data sets out of 9 (in one case the initial accuracy was 100%, which could not have been improved but which was

repeated); this leads to a conclusion that the application of the developed class decomposition method improves classification accuracy;

- The third hypothesis was proven by comparing the classification results of tree-based classifiers generated by genetic algorithms to those of classical methods; the results show that in 7 cases out of 9, the developed decision tree ensemble based algorithm was one of the top two results for the corresponding data set, which leads to a conclusion that in most cases the developed algorithm has similar or better accuracy than the classical methods;
- The fourth hypothesis was tested in the same experiment set and the results showed that in all cases genetic algorithm generated classification trees (specific case of ensembles, where the ensemble holds only one tree) were outperformed by genetic algorithm generated random forests (where ensembles held 10 trees).

The efficacy of the developed methodology was tested by comparative analysis of the results obtained by the application of the methodology and the results achieved by classical classification methods. The results showed that the accuracy of the methodology outperforms other methods in six data sets out of nine and shows the second best result in others. This result is the most significant because no one of the other methods showed the same reliability; other best results were achieved by different methods which performed poorly in other data sets. The overall conclusions of the research are the following:

- The classification method accuracies improved in seven cases out of nine when class decomposition was applied, which means that the application of class decomposition (using description of the class inner structure) increases classification accuracy;
- Cluster analysis during the process of detection of high density areas showed that the most suitable method for class decomposition is hierarchical agglomerative clustering;
- Class decomposition has to be stable to bring the most increase in classification (the Pearson correlation coefficient between subclass stability and increase in classification accuracy is 0,76);
- The developed classification in method showed comparable results to classical classification methods while keeping the classification models simple and transparent, which means that the developed method is a better choice for biomedical tasks;
- The developed classification method works better with classifier ensembles (starting from 10 trees), because special case of the method when the classifier

ensemble consisted of one tree showed worse accuracies than ensembles of 10 and more trees in all of the nine analyzed cases;

- Whereas the size of the classifiers induced by the developed method is limited, it also selects a smaller attribute subset (biomarker panel) in the process;
- The developed methodology, which implements the developed methods, showed the best classification accuracies in six out of nine studied cases and is the second best in other cases, which means that the developed methodology builds more accurate classifiers than classical classification methods;
- The decision tree ensembles constructed by the developed methodology consist of up to ten binary trees with up to six levels, which means that the constructed classifiers are transparent and easy to understand;
- The developed methodology showed one of the two best accuracies in all of the studied data sets that shows that the developed method is also stable across different data sets, which is a unique result – other methods show good accuracies in only a few data sets.

REFERENCES

1. Aitkenhead M. J. A co-evolving decision tree classification method// Expert System Applications. – 2008. – Vol. 34, No. 1. – pp. 18-25.
2. Alizadeh A. A., Eisen M. B., Davis R. E., Ma C., Lossos I. S., Rosenwald A., Boldrick J. C., Sabet H., Tran T., Yu X., Powell J. I., Yang L., Marti G. E., Moore T., Hudson J. J., Lu L., Lewis D. B., Tibshirani R., Sherlock G., Chan W. C., Greiner T. C., Weisenburger D. D., Armitage J. O., Warnke R., Staudt L. M. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling// Nature. – 2000. – Vol. 403, Issue 6769. – pp. 503-511.
3. Barros R. C., Basgalupp M. P., de Carvalho A. C. P. L. F., Freitas A. A. A Survey of Evolutionary Algorithms for Decision Tree Induction// IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews. – 2012. – Vol. 42, Issue 3. – pp. 291-312.
4. Basgalupp M., de Carvalho A., Barros R. C., Ruiz D., Freitas A. Lexicographic multi-objective evolutionary induction of decision trees// International Journal of Bio-Inspired Computation. – 2009. – Vol. 1, No. 1/2. – pp. 105-117.
5. Bellman R. E. Adaptive Control Processes: A Guided Tour. – Prnceton, NJ: Princeton University Press, 1961. – 255 p.
6. Boser B. E., Guyon I. M., Vapnik V. N. A training algorithm for optimal margin classifiers// In Proceedings of the 5th Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, July 27-29, 1992. – New York, NY: ACM Press, 1992. – pp. 144-152.
7. Breiman L. Bagging Predictors// Machine Learning. – 1996. – Vol. 24, No. 2. – pp. 123-140.
8. Breiman L., Friedman J., Olshen R., Stone C. Classification and Regression Trees. – Belmont: Wadsworth Int. Group, 1984. – 368 p.
9. Brown G. Encyclopedia of Machine Learning// Sammut C., Webb G.I., Eds. – Berlin, Heidelberg, Springer-Verlag, 2010. – pp. 312-320.
10. Büssow K., Konthur Z., Lueking A., Lehrach H., Walter G. Protein Array Technology: Potential Use in Medical Diagnostics// American Journal of PharmacoGenomics. – 2001. – Vol. 1, Issue 1. – pp. 37-43.
11. Carugo, O., Eisenhaber F., Eds. Data Mining Techniques for the Life Sciences (Methods in Molecular Biology). – Totowa, NJ: Humana Press, 2010. – 420 p.
12. Dietterich T. G. Ensemble Methods in Machine Learning// In Proceedings of the 1st International Workshop on Multiple Classifier Systems, Cagliari, Italy, June 21-23, 2000. – Lecture Notes in Computer Science. – Vol. 1857. – New York: Springer Verlag, 2000. – pp. 1.-15.
13. Dudoit S., Fridlyand J., Speed T. P. Comparison of discrimination methods for the classification of tumors using gene expression data// Journal of the American Statistical Association. – 2002. – Vol. 97, Issue 457. – pp. 77–87.
14. Freund Y., Schapire R. E.. Experiments with a new boosting algorithm// In Proceedings of the 13th International Conference on Machine Learning, Bari, Italy, July 3-6, 1996. - San Francisco: Morgan Kaufmann Pub., 1996. – pp. 148-156.
15. Fu Z., Golden B. L., Lele S., Raghavan S., Wasil E. Diversification for better classification trees// Comput. Oper. Res. – 2006. – Vol. 33, No. 11. – pp. 3185-3202.
16. Golub T. R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield C. D., Lander E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring// Science. – 1999. – Vol. 386. – pp. 531-537.

17. Gan G., Ma C., Wu J. Data clustering - theory, algorithms, and applications. – Philadelphia, USA: Society for Industrial and Applied Mathematics, 2007. – 489 p.

18. Guyon I., Weston J., Barnhill S., Vapnik V. Gene selection for cancer classification using support vector machines// Machine Learning. – 2002. – Vol. 46. – pp. 389-422.

19. Haizhou D., Chong M. Study on constructing generalized decision tree by using dna coding genetic algorithm// In Proceedings of the International Conference on Web Information Systems and Mining, Shanghai, China, November 7-8, 2009. – Washington, DC, USA: IEEE Computer Society, 2009. – pp. 163–167.

20. Hall D. A., Ptacek J., Snyder M. Protein Microarray Technology// Mech Ageing Dev. – 2007. – Vol. 128, Issue 1. – pp. 161–167.

21. Hall M. A. Correlation based Feature Subset Selection for Machine Learning. – 1998. – Disertācija Vaikato universitātē (Hamilton, Jaunzēlande) – 198 p.

22. Han J., Kamber M. Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. – Gray J., Ed. - San Mateo: Morgan Kaufmann Pub., 2000. – 550 p.

23. Harrington C. A., Rosenow C., Retief J. Monitoring gene expression using DNA microarrays// Current Opinion in Microbiology. – 2000. – Vol. 3, Issue 3. – pp. 285-291.

24. Hinneburg E., Keim D. A. Clustering Techniques for Large Data Sets From the Past to the Future// In Proceedings of the International Conference on Knowledge Discovery and Data Mining, Sandiego, California, USA, August 15-18, 1999. – New York, NY: ACM Press, 1999. – pp. 141-181.

25. Ho T. K. The Random subspace method for constructing decision forests// IEEE Trans Pattern Analysis and Machine Intelligence. – 1998. – Vol. 20, Issue 8. – pp. 832-844.

26. Holland J. Adaptation in Natural and Artificial Systems. Ann Arbor, MI: University of Michigan Press, 1975. – 183 p.

27. Holte R. C. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets// Machine Learning. – 1993. – Vol. 11. – pp. 63-91.

28. Horng J., Wu L., Liu B., Kuo J., Kuo W., Zhang J. An expert system to classify microarray gene expression data using gene selection by decision tree// Expert Systems Applications. – 2009. – Vol. 36, Issue 5. – pp. 9072-9081.

29. Hu H. Mining patterns in disease classification forests// Journal of Biomedical Informatics – 2010. – Vol. 43, Issue 5. – pp. 820-827.

30. Huang J., Fang H., Fan X. Decision forest for classification of gene expression data// Computers in Biology and Medicine. – 2010. – Vol. 40, Issue 8. – pp. 698-704.

31. Huber W., von Heydebreck A., Vingron M. Analysis of Microarray Gene Expression Data// In Handbook of Statistical Genetics. – Hoboken, NJ: John Wiley & Sons, 2004. – pp. 203-230.

32. John G. H., Langley P. Estimating Continuous Distributions in Bayesian Classifiers// In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, August 18-20, 1995. – San Mateo: Morgan Kaufmann Pub., 1995. – pp. 338-345.

33. Jones B. R., Crossley W. A., Lyrintzis A. S. Aerodynamic and Aeroacoustic Optimization of Airfoils via a Parallel Genetic Algorithm// In Proceedings of the 7th Symposium on Multidisciplinary Analysis and Optimization, St. Louis, MO, USA, September 2-4, 1998. – Reston: AIAA, 1998. – pp. 1088-1096.

34. Kalles D, Papagelis A. Lossless fitness inheritance in genetic algorithms for decision trees// Soft Computing. – 2010. – Vol. 14. – pp. 973-993.

35. Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms, Second Edition. – Hoboken, NJ: John Wiley & Sons, Inc., 2011. – 552 p.

36. Kohavi R., John G. H. Wrappers for feature subset selection// Artificial Intelligence. – 1997. – pp. 1-2.

37. Kohavi R., Quinlan J. R. Decision-tree discovery// Handbook of Data Mining and Knowledge Discovery. – Klosgen W., Zytkow J. M., Eds. – Oxford: Oxford University Press, 2002. – pp. 267-276.

38. Kohavi R., Provost F. Glossary of terms// Applications of Machine Learning and the Knowledge Discovery Process. – 1998. – Vol. 30, No. 2/3. – pp. 2-3.

39. Kononenko, I. Estimating attributes: analysis and extensions of RELIEF// In Proceedings of the European Conference on Machine Learning on Machine Learning, Catania, Italy, April 6-8, 1994. – New York: Springer-Verlag, 1994. – pp. 171-182.

40. Kretowski M., Grzes M. Evolutionary induction of cost-sensitive decision trees// In Proceedings of the 16th international conference on Foundations of Intelligent Systems, Bari, Italy, September 27-29, 2006. – Berlin, Heidelberg: Springe-Verlag, 2006. – pp. 121-126.

41. Lee G., Rodriguez C., Madabhushi A. An Empirical Comparison of Dimensionality Reduction Methods for Classifying Gene and Protein Expression Datasets// In Proceedings of the Bioinformatics Research and Applications: Third International Symposium, Atlanta, GA, USA, May 7-10, 2007. – Berlin, Heidelberg: Springer-Verlag, 2007. – pp. 170-181.

42. Lee J. W., Lee J. B., Park M.. Song S. H. An extensive comparison of recent classification tools applied to microarray data// Computational Statistics & Data Analysis. – 2005. – Vol. 48, Issue 4. – pp. 869-885.

43. Lu Y., Han J. Cancer classification using gene expression data// Information Systems. – 2003. – Vol. 28, Issue 4. – pp. 243-268.

44. MacQueen J. Some methods for classification and analysis of multivariate observations// In Proceedings of the Fifth Berkeley Symp. on Math. Statist. and Prob., Berkeley, CA, USA, June 21-July 18, 1965, Vol. 1. – Berkeley, CA: University of California Press, 1967. – pp. 281-297.

45. Mingers J. An Empirical Comparison of Pruning Methods for Decision Tree Induction// Machine learning. – 1989. – Vol. 2, Issue 4. – pp. 227-243.

46. Mishra D., Shaw K., Mishra S., Rath A. K., Acharya M. Hash based biclustering for class discovery from gene expression data: A pattern similarity approach// In Proceedings of the 3rd International Conference on Electronics Computer Technology (ICECT), Kanyakumari, India, April 8-10, 2011, Vol. 2. –- Washington, DC: IEEE Computer Society, 2011. – pp. 137-141.

47. Monti S., Tamayo P., Mesirov J., Golub T. Consensus clustering - A resampling-based method for class discovery and visualization of gene expression microarray data// Machine Learning. – 2003. – Vol. 52, Issue 1-2. – pp. 91-118.

48. Papagelis A., Kalles D. GA Tree: genetically evolved decision trees// In Proceedings of the 12th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '00), Vancouver, Canada, November 13-15, 2000. – Washington, DC: IEEE Computer Society, 2000. – pp. 203-207.

49. Poļaka, I., Borisovs, A. Genethic Algorithm and Tree Based Classification in Bioinformatics// Studies in Classification, Data Analysis, and Knowledge Organization. – Heidelberg: Springer, 2014, (accepted).

50. Poļaka, I., Borisovs, A. The Application of Class Structure to Classification Tasks// Scientific Journal of RTU – 2013. – Information Technology and Management Science (accepted).

51. Poļaka, I., Borisovs, A. Genethic Algorithm and Tree Based Classification in Bioinformatics// In: European Conference on Data Analysis 2013: Book of Abstracts, conference in Luxembourg, Luxembourg, July 10-12, 2013. – Luxembourg: GFKL, 2013. – p. 107.

52. Poļaka, I. Clustering Algorithm Specifics in Class Decomposition.// In: Applied Information and Communication Technology 2013 (AICT2013): Proceedings of the 6th International Scientific Conference, Latvia, Jelgava, April 25-26, 2013. – Jelgava, Latvia: LLU. – pp. 29-36.

53. Poļaka, I., Borisovs, A. The Impact of Cluster Stability on Class Decomposition in Antibody Display Data// Information Technology and Management Science. – 2012. – Vol. 15. – pp. 70-75.

54. Poļaka, I., Borisovs, A. Class Decomposition in Bioinformatics Analyzing Omics Data. No: Proceedings of Workshop on Data Mining in Life Sciences (DMLS'2012): Workshop on Data Mining in Life Sciences (DMLS'2012), Germany, Berlin, July 20, 2012. – Berlin: Springer-Verlag Berlin Heidelberg, 2012. – pp. 158-167.

55. Poļaka I., Borisovs A. Robust Dimensionality Reduction in Bioinformatics Data // 21st European Meeting on Cybernetics and Systems Research (EMCSR 2012): Book of Abstracts, Austria, Vienna, April 10-13, 2012. – Vienna, Austria: BCSSS. – pp. 286-289.

56. Poļaka I. Genetic Algorithm for Random Tree Generation in Bioinformatics Data // Proceedings of the 5th International Scientific Conference on Applied Information and Communication Technologies (AICT2012), Latvia, Jelgava, April 26-27, 2012. – Jelgava, Latvia: LLU. – pp. 335-340.

57. Poļaka I., Borisovs A. Impact of Antibody Panel Size on Classification Accuracy // Scientific Journal of RTU – 2011. – 5th series, Computer science. Vol. 49 – 2010. – pp. 85-90.

58. Grabusts P., Poļaka I. Estimation of the Efficiency of Knowledge Acquisition Techniques Using Clustering // Proceedings of the Ninth International Scientific School MA SR, Russia, Saint Petersburg, June 28-July 2, 2011. – Saint Petersburg, Russia: IPME RAS, 2011. – pp. 131-137.

59. Poļaka I., Borisovs A. Impact of Feature Selection on Classifier Testing Validity // Proceedings of the 17th International Conference on Soft Computing MENDEL, Czech Republic, Brno, June 15-17, 2011. – Brno, Czech Republic: MENDEL SCC, 2011. – pp. 411-418.

60. Poļaka I. Feature Selection Approaches in Antibody Display Data Analysis // Proceedings of the 8th International and Practical Conference, June 20-22, 2011, Volume II, Latvija, Rēzekne, 20.-22. jūnijs, 2011. - 16.-23. lpp.

61. Poļaka I., Borisovs A. Using Data Structure Properties in Decision Tree Classifier Design // Scientific Journal of RTU – 2010. – 5th series, Computer science. Vol. 44 – 2010. – pp. 111-117.

62. Poļaka I., Tom I., Borisovs A. Decision Tree Classifiers in Bioinformatics // Scientific Journal of RTU – 2010. – 5th series, Computer science. Vol 44. – pp. 118-123.

63. Poļaka, I., Borisovs, A. Clustering-Based Decision Tree Classifier Construction. Technological and Economic Development of Economy, 2010, Vol.16, Iss.4 – pp. 765-781.

64. Quinlan J. R. C4.5: Programs for Machine Learning. – San Mateo: Morgan Kaufmann Pub., 1993. – 302 p.

65. Quinlan J. R. Simplifying decision trees// International Journal of Man-Machine Studies. – 1987. - Vol. 27. – pp. 221-248.

66. Slonim D. K., Tamayo P., Mesirov J. P., Golub T. R., Lander E. S. Class prediction and discovery using gene expression data// In Proceedings of Fourth Annual International Conference on Computational Molecular Biology, Tokyo, Japan, April 8-11, 2000. – New York, NY: ACM, 2000. – pp. 263-272.

67. Sreekumar A., Nyati M. K., Varambally S., Barrette T. R., Ghosh D., Lawrence T. S., Chinnaiyan A. M. Profiling of cancer cells using protein microarrays: discovery of novel radiation-regulated proteīns// Cancer Res. – 2001. – Vol. 61. – pp. 7585-7593.

68. Steinfeld I., Navon R., Ardigò D., Zavaroni I., Yakhini Z. Clinically driven semi-supervised class discovery in gene expression data// *Bioinformatics. – 2008. – Vol.* 24, Issue 16. – pp. i90-i97.

69. Tan C. P., Lim K. S., Lai W. K. Multi Dimensional Features Reduction of Consistency Subset Evaluator on Unsupervised Expectation Maximization Classifier for Imaging Surveillance Application// International Journal of Image Processing. – 2008. – Vol. 2, Issue 1. – pp. 18-26.

70. Tan P. N., Steinbach M., Kumar V. Introduction to Data Mining. – Boston: Pearson Addison-Wesley, 2006. – 769 p.

71. Tibshirani R., Walther G., Hastie T. Estimating the Number of Clusters in a Dataset via the Gap Statistic// Journal of the Royal Statistical Society, Series B. – 2000. – Vol. 63. – pp. 411-423.

72. Ward J. H., Jr. Hierarchical Grouping to Optimize an Objective Function// Journal of the American Statistical Association. – 1963. – Vol. 58. – pp. 236-244.

73. Witten I. H., Frank E. Data mining: Practical machine learning tools and techniques (Second edition). – San Francisco, CA: Morgan Kaufmann, 2005. – 560 p.

74. Yu Z., Wong H. S. Class discovery from gene expression data based on perturbation and cluster ensemble// IEEE Trans Nanobioscience. – 2009. – Vol. 8, Issue 2. – pp. 147-160.

75. Zhao H. A multi-objective genetic programming approach to developing pareto optimal decision trees// Decis. Support Syst. – 2007. – Vol. 43, No. 3. – pp. 809-826.

76. Zintzaras E., Kowald A. Forest classification trees and forest support vector machines algorithms: Demonstration using microarray data// Comput. Biol. Med. – 2010. – Vol. 40, Issue 5. – pp. 519-524.