MULTILEVEL CLASSIFIER USE IN A PREDICTION TASK

Arnis Kirshners and Arkady Borisov

Riga Technical University
Institute of Information Technology
1 Kalku str., Riga, LV-1658
Latvia
arnis.kirsners@rtu.lv, arkadijs.borisovs@cs.rtu.lv

Abstract: *This study proposes a multi-level classifier to predict heart necrosis or risk based only on the descriptive parameters of a new laboratory animal when solving a pharmacology task. The operation of the multi-level classifier is based on data mining tasks and algorithms. The construction of the multi-level classifier uses intelligent data analysis techniques like classification, clustering and prediction. The classification process includes splitting of the data set, feature selection based on their information, growing of a decision tree using the C4.5 algorithm and induction of a conditional rule set. The clustering is based on finding groups of similar objects in heart contraction power data. The prediction is carried out using only descriptive parameters that are projected onto obtained decision tree determining a connection between descriptive parameters of an animal and the class obtained in clustering. Based on the acquired results a rule set is selected from database that is the basis for finding occurrence frequency statistics used in the calculation of the potential risk.*

Keywords: *Multilevel classifier, Machine learning, Classification trees, Classification rules, Clusterization, Prediction task*

1 Introduction

Nowadays the demand is rising for data mining methods and algorithms in the areas that traditionally use mathematical statistics [1-3] for result evaluation. For example, pattern recognition methods are used to determine the shape of an object [4]. One of such areas that use mathematical statistics and pattern recognition is pharmacology. The data acquired in pharmacological research have to be assessed and processed. The obtained results are often processed using mathematical statistics methods, e.g. assessing statistical significance with analysis of variance [2], statistical significance with normally distributed data [3] or using correlation to evaluate parameters. Of course, there are attempts to use data mining methods to accomplish these tasks, e.g. brainstem syndrome determination [4], using inductive decision trees, but there are tasks in pharmacology that are difficult to formalize and impossible to solve using simple methods. There are data that describe features of laboratory animals (rats) [5], their heart necrosis risk estimation and membership in a group of similar objects (class) that are acquired as a result of pharmacological experiments. The solving of this task requires a set of methods or a multi-level processing chain that is able to solve tasks that are difficult to formalize. Similar problems dealing with descriptive parameters and object groups are solved within data mining classification and forecasting tasks [6, 7] that search for the relationships between these variables. The study proposes an approach that performs cardiac necrosis risk prognosis based only on descriptive parameters of a laboratory animal. The implementation of the proposed method in pharmacology would decrease the time spent in experiments and the number of laboratory animals used.

2 Problem Formulation

The study requires data analysis based on pharmacological experiments to determine the percentage of necrotic heart tissue (cardiac necrosis) risk in 'Isolated heart' [5] experiments. A classifier has to be made based on the analysis of the acquired results that would carry out prediction task determining the cardiac necrosis risk based only on the descriptive parameters of a laboratory animal (specially bred rats), e.g. weight, food supplements and parameters describing plasma of an animal. The creation of such classifier will accelerate the pace of pharmacological experiments, decrease their costs and the number of animals used in the experiments.

The construction of the classifier is based on the use of data mining methods and algorithms while solving pharmacology tasks predicting the potential cardiac necrosis risk of a laboratory animal. The classifier has to solve several problems related to the specifics of the data set and attributes, for example:

- Find the intercorrelation of parameters describing laboratory animals due to the chance of present attributes that correlate or have no correlation at all;
- Find a classifier that suits the task best and gives results that are easy to interpret for pharmacology experts;
- The classifier has to link descriptive parameters of a laboratory animal with heart contraction power data or class while covering the whole range of the classes;

- Create a rule base based on the rules induced by the classifier that describes the range of necrosis risk values in each class;
- Calculate the frequency of risk occurrences.

The prognosis in its turn has to provide an answer to the question: what the potential risk will be given the descriptive parameters of a new laboratory animal and accomplish the following tasks:

- Perform classification of a new laboratory animal;
- Select a rule set from the created rule base that will be the basis of the potential heart necrosis or risk range calculations for the new laboratory animal.

The data set used for classifier training and testing is acquired in pharmacological experiments with specially bred laboratory animals. The animals were fed specific food supplements for a certain period of time that improve heart action. The experiments gave results about heart contraction power in a given period of time for each heart used in the experiment. The obtained heart contraction power data were analyzed using data mining methods and algorithms determining relationships between similar object groups; as a result a class or membership to one of the groups was assigned to each animal.

3 Problem Solution

The basis of the multilevel classification approach was taken from methods that were developed to forecast demand for a product based on historical demand data and the descriptive parameters of a new product [6, 7].

The problem to be solved is related to classification and prediction [8]. The classification task searches for correlations between the descriptive parameters of a laboratory animal and heart contraction power data. Clustering provides aggregation of objects into groups or clusters based on their similarity or distance between objects. Prediction task in its turn is based on the descriptive parameters of a new laboratory animal and the obtained classification rules when predicting the potential risk. The prediction is made based on a trained classifier and given only the descriptive parameters of the new laboratory animal, which shows that the term 'prediction' is used as part of data mining and its solution cannot be obtained using classical prediction methods.

There are several multiple classifier topologies, e.g. parallel where several classifiers work in parallel and the results are compiled at one point, and serial where classifiers work one after another decreasing the number of possible classes. As the solution to the task defined in the study a multi-level classifier is proposed. Whereas the proposed solution uses only one classifier it is considered a classifier with multiple levels of data processing. The term 'multi-level classification' is used with intent because the implementation of the processes is carried out in multiple levels and other intelligent data analysis processes can be embedded between these levels, e.g. clustering, predicting etc.

The proposed multi-level classifier (see Fig.1.) initially splits the data set into two subsets in the first level: a set of descriptive parameters and a set of short time series. Clustering of short time series (see Fig.1. painted in grey) is not described in detail in this article giving only a brief overview of the process. The data set describing laboratory animals is split based on the *Group* attribute values into a number of subsets that matches the number of unique discrete values that this attribute holds. The significance and influence of this attribute point to the heart necrosis risk because the given attribute shows the type of food supplement that was given to the laboratory animals. As later research showed this attribute did not correlate with other attributes but the credibility and adequacy of the obtained results depends on it, therefore it was chosen for the initial data set split. Therefore the animals that had value 1 for this attribute were put into a new data subset (DATA SET 1) and the rest of the data set was split accordingly creating DATA SETs 2 till 5. When the *Group* attribute is removed from the rest of the data set other attributes can be split based on their significance or information measures [8]. Such a split of attributes is necessary for the creation of new data sets and construction of an inductive decision tree [8]. The split of the attributes is then used in the second level of the multi-level classifier when creating the inductive decision tree. This concludes the first level of the multi-level classifier and the middle step between levels is implemented merging the data. The most informative attributes selected previously are appended by a class acquired in clustering forming a new data set for each group of clusters.

Short time series clustering was performed using the *k-means* algorithm [9] with cluster range from two to maximum number [8]. The use of predefined cluster range increases the speed of algorithm execution. The maximum cluster number is calculated using Equation (1) and rounding the result.

$$C_{max} = \sqrt{n} \tag{1}$$

The distance metric used in the algorithm was *Euclidean* distance [9]. The obtained clustering results that describe the distribution of laboratory animals in cluster according to their closeness against each other are appended to the data set as a new attribute named '*Class*'. The resulting clustering algorithm errors are assessed together with classification errors that will be described later.

The new data set with the appended class obtained in clustering is used to find correlations between the descriptive parameters of laboratory animals and the class obtained in the clustering using machine learning algorithm *C4.5* [10] that is based on inductive decision trees. Algorithm *C4.5* is based on the classical *ID3* algorithm [8, 10] but works with continuous data, which is crucial for this case. The algorithm finds the error level δ that is based on training set data and

statistics basics. Algorithm *C4.5* considers each node to be a leaf. In its execution it performs temporary cuts on internal nodes and calculates the number of misclassified records considering the number of all records belonging to the leaf of

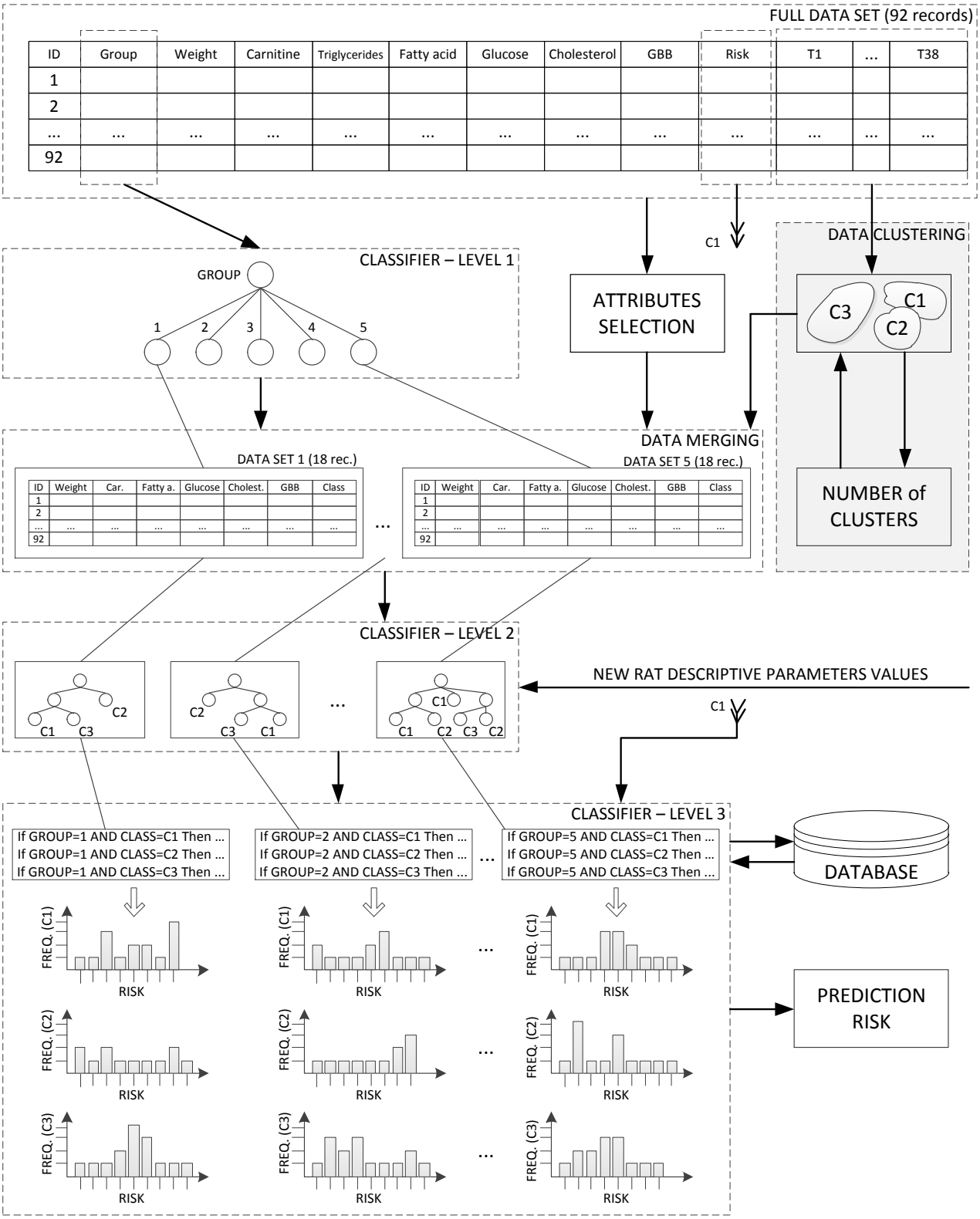

Fig.1. Multilevel classifier

this class. It is usually assumed that the confidence level of algorithm *C4.5* is 25%, which is used to calculate confidence limit given by the Equation (2).

$$25\% = P\left(\frac{k/n - q}{\sqrt{[q(1-q)]/n}} > z\right) \tag{2}$$

where P – probability that a randomly selected record from the data set will belong to one of the classes;
k – number of misclassified records;
n – number of records in the data set;
q – the exact level of error;
z – confidence limit.

The error level δ in its turn is calculated according to Equation (3) that uses the acquired value to calculate subtree cutting:

$$\delta = \frac{f + \frac{z^2}{2n} + z\sqrt{\frac{f}{n} - \frac{f^2}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \tag{3}$$

where f – observed level of error that is characterized by the ratio of misclassified records to the number of records in the data set;
z – standard deviation when the confidence level is 25% being 0,69.

An important process is decision tree pruning that decreases the number of tree nodes. The use of this condition emphasizes that decrease in node number also decreases the number of classes covered by the decision tree, which asks for a balance that is regulated using pruning coefficient.

In the third level the proposed classifier makes *„If ... Then"* rules based on the values of attributes *Group*, *Class* and *Risk*. The left side of the rule holds attributes *Group* and *Class* and the right side holds the values of attribute *Risk*. The advantage of these rules is their ability to process numeric values on both sides of the rules, e.g. one of the rules could be IF GROUP=1 AND CLASS=C1 THEN RISK=43. A rule is made for each record in the data set and saved in the database.

3.1 Calculation of classification accuracy

The calculation of classification accuracy is based on confusion matrix (see Table 1) [8]. The matrix holds results according to classification and the classifier total accuracy *CA* is calculated according to these assessments and

Table 1. Confusion matrix

		Predicted class	
		C1	C2
Actual class	C1	True positive (TP)	False negative (FN)
	C2	False positive (FN)	True negative (TN)

Equation (4). The assessment of classification accuracy describes the adequacy of the constructed model, e.g. one can estimate the accuracy of the model when recognizing one or another class. There are also other parameters that describe

$$CA = \frac{TP + TN}{TP + FN + FP + TN} \tag{4}$$

accuracy, e.g., sensitivity shows the classifier accuracy when assigning the positive class (C1) and specificity shows the accuracy when assigning the second class (C2).

3.2 Calculation of risk occurrence frequency

From the rules that characterize the class of laboratory animals (C_n) and attributes *Group* un *Risk* induced in the third level of classifier and stored in the database value ($x_n \in C$) occurrence frequency statistic set C is made from elements (x_n, y) see Fig.2 where y is occurrence frequency of x_n in the *n-th* cluster. The obtained occurrence frequency statistics are also stored in the database and used when predicting the potential risk of a new laboratory animal. Risk distribution function is made from the induced rules and is reflected in risk occurrence frequency table where parameter *FREQ.(C_n-cluster number)* shows how many times attribute *GROUP(n-food supplement type)* has occurred with the corresponding risk assessment in the data set.

3.3 Risk prediction

The potential risk of a new laboratory animal is predicted based on the risk occurrence statistics. The potential risk is predicted in three ways – one is an expert opinion based on the distribution function, the second uses mathematical expectation to determine the most possible risk based on the risk assessment range in the corresponding group cluster and the third assesses the value of possible risk according to the occurrence frequency distance evaluation and calculation differences between the current and minimum value in the corresponding group cluster. The expert opinion assessment does not ask for further explanation. Mathematical expectation while assessing the potential risk is calculated as follows: take the obtained risk occurrence frequency statistic (see Fig.2.), assign a probability to each assessment considering the occurrence frequency statistics (see Table 2) and calculate mathematical expectation [11],

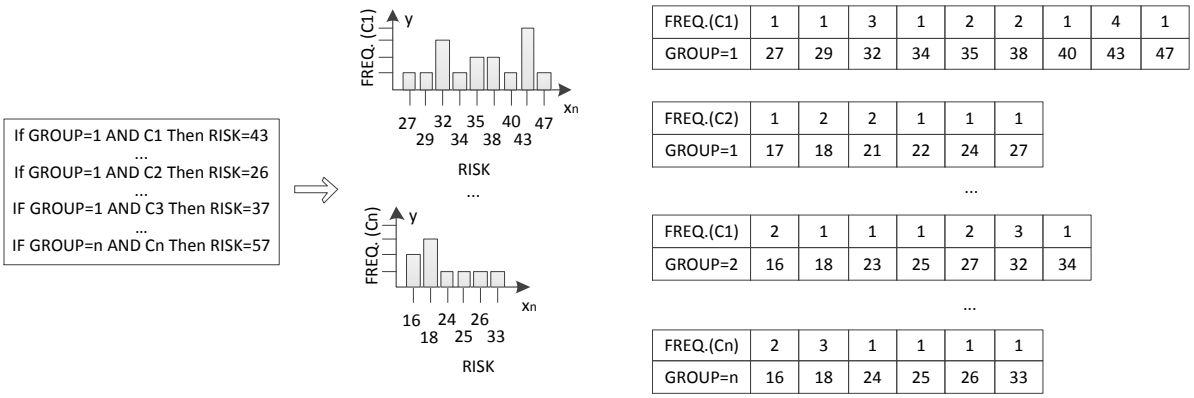

Fig.2. Acquiring risk occurrence frequency statistics

that determines the potential risk value 37 based on the descriptive parameters of the new laboratory animal and the rules stored in the data base and induced in the classifier training. The potential values are determined by first

Table 2. Calculation of mathematical expectation

										Sum
FREQ.(C1)	1	1	3	1	2	2	1	4	1	
GROUP=1 (RISK VALUE)	27	29	32	34	35	38	40	43	47	
Probability	0.06	0.06	0.19	0.06	0.13	0.13	0.06	0.25	0.06	1
Mathematical expectation	1.62	1.74	6.08	2.04	4.55	4.94	2.40	10.8	2.82	**37**

calculating the distance between the occurrence frequency values (see Table 3) and then calculating the difference between the minimum value 1 and the number of occurrence frequency (FREQ.). Then the field of distance and

Table 3. Calculation of the potential risk

FREQ.(C1)	1	1	3	1	2	2	1	4	1
GROUP=1 (RISK VALUE)	27	29	32	34	35	38	40	43	47
Distance between occurrence frequencies		2	3	2	1	3	2	3	4
Difference between given and min value of FREQ.			-2		-1	-1		-3	
Sum: Distance + Difference		2	1	2	**0**	2	2	**0**	4
Assessment:					35			43	

difference are added and the fields that have the minimum sum value (0 in the example that is highlighted in bold) are chosen. The obtained result (the potential risk value) for this new laboratory animal is 35 or 43.

4 Experimental results

The study included a set of experiments to find the most suitable classifier for solving the pharmacology task. The accuracy of various classifiers was studied working with the proposed data set and also the influence of laboratory animal descriptive parameters on research results was explored. The study analyzes the classifier errors and assesses their performance when different numbers of classes were used. Data mining tools like *Weka* and *Orange Canvas* were used in the experiments to facilitate the acquisition of the results and *MS Excel 2010* was used for calculations.

4.1 Used data

The used data set consisted of 92 records. Each record was characterized by descriptive parameters of laboratory animals – *Weight* (rat weight); *Group* (food supplement type improving heart action); *Risk* (the evaluation of necrosis in pharmacological experiments); *Carnitine*; *Triglycerides*; *Fatty acids*; *Glucose*; *GBB*; *Cholesterol* (these six parameters were acquired form the blood count of the laboratory animals) and normalized time series consisting of 38 periods (*T1 … T38*) that was acquired from heart contraction power data analysis. All parameters held continuous data and the food supplement type *Group* had discrete values.

4.2 Attribute selection

Data classification is based on information carried by attributes; if there is correlation between attributes then these attributes are considered to be interrelated [8]. But if the attributes do not correlate, it means that the connection between the attributes is weak. Therefore the informativity of the attributes should be experimentally evaluated or divided according to importance of the attribute. There are many approaches to implementation of this procedure. *Weka* offers many attribute selection resources [8], this study implements *CfsSubsetEval* attribute evaluation method and search method *BestFirst*. The CFS subset evaluation method evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred [8]. It searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility by setting the number of consecutive non-improving nodes allowed. Best first may start with the empty set of attributes and search forward, or start with the full set of attributes and search backward, or start at any point and search in both directions (by considering all possible single attribute additions and deletions at a given point) [8]. The results were as follows: *Fatty acids*, *Cholesterol* and *Carnitine* showed 100%, *Glucose* 90%, *Weight* 80%, and *GBB* showed 20% intercorrelation. As a result, these six laboratory animal descriptive attributes were used in construction of the multi-level classifier.

4.3 Finding the most adequate classifier

To choose the classifier that would be the most suitable for this task, a set of experiments were carried out using the full data set with all attributes. A model was built in *Orange Canvas* environment that allowed changing the number of clusters during clustering process. The most suitable classifier was determined during classification process using various classifiers (*Naive Bayes*, *k-Nearest Neighbor* and *C4.5*) and 10 fold cross-validation [12] based on mean classification accuracy measure in the whole range of clusters. According to classification accuracy evaluation (see Table 4) the best results using different cluster numbers were achieved using inductive decision tree algorithm *C4.5*, which was chosen for further experiments and became the basis of the multi-level classifier construction.

Table 4. Classification accuracy with full data set

	Cluster								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Mean**
Naive Bayes	0.48	0.33	0.30	0.24	0.26	0.25	0.25	0.11	0.23
kNN	0.50	0.40	0.35	0.34	0.30	0.29	0.34	0.20	0.32
C4.5	0.50	0.37	0.41	0.29	0.34	0.20	0.23	0.15	**0.34**

It also should be noted that inductive decision trees are easily interpretable to visualize the results and ease the process of laboratory animal descriptive parameter classification.

4.4 Construction of inductive decision tree

The proposed classifier constructs an inductive decision tree for each of the obtained subsets in the second level using C4.5 algorithm, e.g., for the first subset DATA SET 1 (see Fig.3.) and for the second subset DATA SET 2 (see Fig.4.);

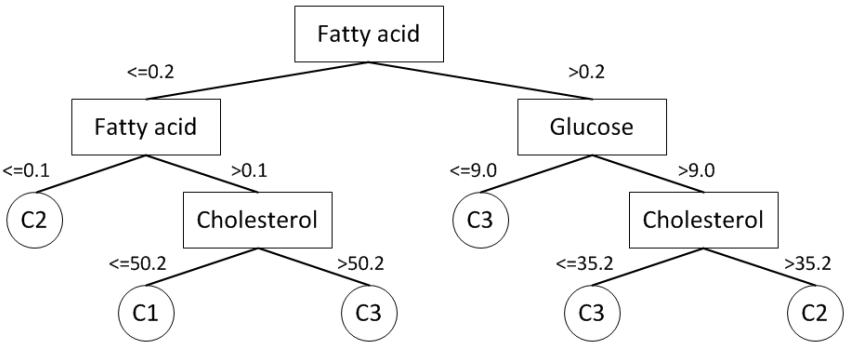

Fig.3. The inductive decision tree for the DATA SET 1 subset

decision trees for the rest of the subsets are constructed similarly. The decision tree enables determining relationships between descriptive parameters of a laboratory animal and the class obtained in clustering that points to changes in heart contraction power in the 'isolated heart' experiments. The nodes of a decision tree characterize descriptive parameters of a laboratory animal and the leaves of the tree show changes in heart contraction power. The arcs between the nodes and the leaves show splitting values of the decision tree.

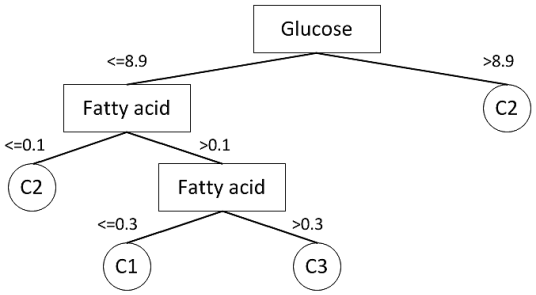

Fig.4. The inductive decision tree for the DATA SET 2 subset

4.5 Prediction of the potential risk for a new laboratory animal

To predict the potential risk of cardiac necrosis for a new laboratory animal let the descriptive parameters of the animal be as follows: Group=2; Glucose=7.4; Fatty acid=0.1 etc.; they are projected onto the corresponding constructed decision tree (see Fig.5.), assigned by the value of *Group* attribute, being 2 in the example. Therefore the projection is

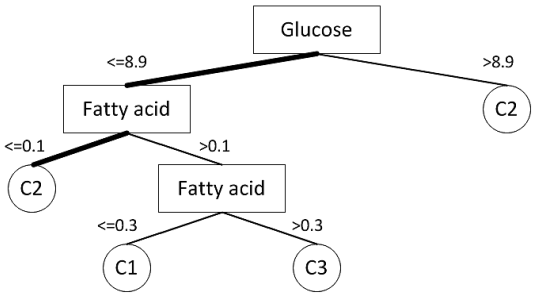

Fig.5. Decision tree with the parameters of the new animal projected onto it

made onto subset DATA SET 2. As a result, one of the decision tree leaves is reached that points to the class, which is used to acquire corresponding rules from the database. These rules serve as a basis to obtain the frequency of risk occurrence based on the distribution function that is used to predict the potential risk. The full prediction process is shown in Fig. 6.

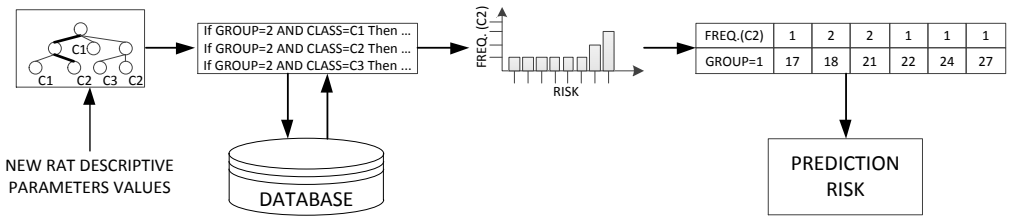

Fig.6. Prediction of the potential cardiac necrosis risk for a new laboratory animal

4.6 Evaluation of classification performance

The performance of the created multi-level classifier is assessed using each of the obtained subsets using 10-fold cross-validation and different number of clusters. The use of the 10-fold cross-validation helps avoiding the adaptation of the

Table 5. Classification accuracy with data subsets

	Cluster				
	2	**3**	**4**	**5**	**6**
DATA SET 1	0.55	0.70	0.60	0.50	0.25
DATA SET 2	0.80	0.85	0.45	0.50	0.40
DATA SET 3	0.90	0.80	0.65	0.65	0.60
DATA SET 4	0.85	0.75	0.55	0.45	0.20
DATA SET 5	0.80	0.90	0.55	0.30	0.20
Mean	0.78	**0.8**	0.56	0.48	0.33

classifier to the training set data. The results (see Table 5) show that the proposed multi-level classifier reaches the highest accuracy when three clusters are used to cluster the data subsets using the average values. The evaluation was

not carried out using seven or more clusters because the table shows a clear trend for the accuracy to decrease when the number of clusters is increased. Based on this evaluation the proposed multi-level classifier determines the necessary number of clusters to cluster the data subsets.

5 Conclusion

The obtained results and the experiments made in the study show that the classifier is effective and it provides implementation of the pharmacology task determining the risk of cardiac necrosis based only on the descriptive parameters of a laboratory animal.

The parameters that correlate and that can be used as a basis for building a multi-level classifier with high classification accuracy have been determined among descriptive parameters of an animal. Also the most appropriate classifier has been determined that can carry out the pharmacology task and the decision trees used in the classifier and their results are easy to interpret for users. The prediction of the potential cardiac necrosis risk has also been carried out. It is experimentally proven that exactly the use of a multi-level classifier approach is more than 2.3 times more effective (see Table 5) than the use of a simple classifier (see Table 4) for this task.

The method proposed in this work is developed especially for the task of laboratory animal cardiac necrosis risk prognostics. The study does not include the comparative analysis of the obtained results with similar methods because their descriptions were not found in the analysed literature. Therefore the accuracy of the proposed method was compared to the results acquired in the pharmacological experiments using 10-fold cross-validation.

The prediction methods of the potential risk re yet to be discussed with experts of pharmacology to choose the most appropriate solution that would satisfy both parties involved in the experiment and for the obtained results to suit pharmacology standards controlling the predicted risk.

The research carried out shows that data mining methods and algorithms can be used to solve pharmacological tasks.

Further research could approbate the proposed multi-level classifier using other data sets and analyze the obtained results to facilitate estimation of the extensive nature of the proposed multi-level classifier. Another possible way is to create a risk occurrence frequency evaluator based on other functions. It is possible that normalization of the descriptive attributes of laboratory animals would increase the accuracy of classification but the contrary is also possible; it can only be proven experimentally.

Acknowledgements: The authors thank the lead researcher of the pharmaceutical pharmacology laboratory Dr. phrm. Edgars Liepins and researcher Janis Kuks of Latvian Institute of Organic Synthesis for the expressed interest and the provided research data on heart contraction power that were used for experiments in this study.

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

References:

[1] Ballester I., Gonzalez R., Nieto A., Zarzuelo A., Sanchez de Medina F., Monochloramine induces acute and protracted colitis in the rat: Response to pharmacological treatment, *Elsevier, Life Science 76*, 2005, p.2695-2980.

[2] Di Angelantonio S., Nistri A., Moretti M., Clementi F., Gotti C., Antagonism of nicotinic receptors of rat chromaffin cells by N,N,N-trimethyl-1-(4-trans-stilbenoxy)-2-propylammonium iodide: a patch clamp and ligand binding study, *British Journal of Pharmacology 129*, 2000, p. 1771-1779.

[3] Zhang W.-F., Liu C.-C., Yan H., Clustering of temporal gene expression data by regularized spline regression and an energy based similarity measure, *Elsevier, Pattern Recognition 43*, 2010, p. 3969-3976.

[4] Pilih, I.A., Mladenic, D., Prevec, T.S., Lavrac, N., Data analysis of patients with severe head injury, Book chapter in *Intelligent Data Analysis in Medicine and Pharmacology*, (eds. Lavrac, Keravnou, Zupan), Kluwer 1997, p. 131-148.

[5] Liepinsh E., Vilskersts R., Zvejniece L., Svalbe B., Skapare E., Kuka J. et al., Protective effects of mildronate in an experimental model of type 2 diabetes in Goto-Kakizaki rats, *British Journal of Pharmacology*, 2009, **157**: 1549–1556.

[6] Kirshners A., Parshutin S., Borisov A., Combining clustering and a decision tree classifier in a forecasting task, *Automatic Control and Computer Sciences*, Vol.44, N3, 2010, p. 124-132.

[7] Thomassey S., Fiordaliso A., A hybrid sales forecasting system based on clustering and decision trees, *Decision Support Systems*, Volume 42, Issue 1, 2006, p. 408-421.

[8] Written I.H., Frank E., *Data mining: practical machine learning tools and techniques - 2nd edition*, Amsterdam etc.: Morgan Kaufmann, 2005.

[9] Han J., Kamber M., *Data Mining: Concepts and Techniques - 2nd edition*, Morgan Kaufmann publishers, UK, 1993

[10] Quinlan J.R., *C4.5: Programs for Machine Learning*, Amsterdam etc.: Morgan Kaufmann publishers, USA, 2006.

[11] L.Gringlaz, E.Kopitov, *The Higher Mathematics for economists* part 1 (examples of solving problems with computer), Riga, 2003.

[12] Kohavi R., A Study of Cross_Validation and Bootstrap for Accuracy Estimation and Model Selection, *Proc.14th Int. Conf. on Artificial Intelligence*, San Mateo, CA: Morgan Kauffman, 1995, pp. 1137–1143.