Scientific Journal of Riga Technical University
Computer Science. Information Technology and Management Science
2011
Volume 49

Mining Online Store Client Assessment Classification Rules with Genetic Algorithms

Anna Galinina[1], Serge Parshutin[2], [1-2]*Riga Technical University*

***Abstract* – The paper presents the results of the research into algorithms that are not meant to mine classification rules, yet they contain all the necessary functions which allow us to use them for mining classification rules such as Genetic algorithm (GA). The main task of the research is associated with the application of GA to classification rule mining. A classic GA was modified to match the chosen classification task and was compared with other popular classification algorithms – *JRip*, *J48* and *Naive Bayes* classifier. The paper describes the algorithm proposed and the application task as well as provides a comparative analysis of the obtained results with other algorithms.**

***Keywords* – genetic algorithms, classification rule mining, data mining, classification**

I. Introduction

The paper describes a task associated with the application of genetic algorithms to mining classification rules from data. The goal of the research is to provide a solution to the classic problem of data mining, which can be solved by a number of methods that are referred to as classification rule mining algorithms.

Topicality of this research is that there are other algorithms that are not meant to mine classification rules, yet they contain all the necessary functionality to apply them to mining rules. Genetic algorithm is one of these algorithms. The paper studies and analyzes the genetic algorithm options for mining classification rules as well as other data mining methods that can be applied to forecasting whether the e-shop client will re-purchase within 90 days or not.

The theoretical background of this paper focuses on the genetic algorithms and classification concepts and principles. In the paper, a genetic algorithm, which is different from the classical one, is proposed. This particular genetic algorithm is adapted to mining the classification rules from statistical data. As other possible solutions to this problem, the algorithms *JRip* [8], *J48* [4][8] and the *Naive Bayes* classifier were tested [11]. Genetic algorithm is an evolutionary computational technique that belongs to the set of optimization algorithms, which can be described as a natural heuristic algorithm. The algorithm is based on an iterative improvement of an existing solution [1]. A set of solutions is used instead of a single solution, and it is called a population, where each individual is a solution to a specific problem. Solution set is iteratively improved by following the natural evolution and applying genetic algorithm operators: crossover, mutation and selection.

Genetic algorithm represents the optimization techniques [19], [21] that simulate a natural evolutionary process leading to survival and producing the largest number of offspring of individuals, who are better fitted to environmental conditions. The fitness depends on genes of the individual, inherited from parents [1], [5].

In 1859 Charles Darwin published his famous work "The emergence of individuals", presenting the basic principles of evolutionary theory:

1. Heredity – offspring inherits signs (genes) from parents;
2. Variability – the descendants almost always are not identical;
3. Natural selection – survival of the fittest [10].

By improving the value of fitness function, the genetic algorithm searches and finds a suitable solution to the task. The solution may be optimal or close to it [2], [7], [10].

II. Mining Classification Rules with GA

Mining classification rules with genetic algorithm is a complex and laborious process which requires the GA to be modified to solve a data mining task [16], [20]. The proposed genetic algorithm has two phases. During the first phase the best individuals in each population are added to the archive. Then in the second phase the archive of the best found individuals is used to mine the classification rules. Figure 1 shows the first phase and contains the next eight steps:

1. Generating the first population. Individuals are created randomly. The binary encoding is used. If the generated random number is less than or equal to 0.5, the gene will be assigned "0", otherwise the value "1" is assigned to the specific gene. The length of chromosome depends on the number of descriptive attributes and their value sets, because the number of genes that is used to encode the attribute is equal to the number of values the attribute can take [15]. With respect to the dataset used, it was determined that the length of each chromosome length would be equal to 35 genes.
2. Evaluation of individuals. The fitness of each individual depends on how good it is, how bad it is and on support of the rule that is encoded by the individual. The support is calculated using equation (1). How good the individual is or "confidence" is calculated using equation (2), and the equation (3) is used to evaluate how bad the individual is. When all calculations are made, the fitness of an individual is calculated using equation (4).

$$S = \frac{|A+B|}{B}, \quad (1)$$

$$C = \frac{|A+B|}{|A|}, \tag{2}$$

$$\bar{C} = 1 - C, \tag{3}$$

$$\begin{cases} C > \bar{C}, & F = (C - \bar{C}) * S = (2 * C - 1) * S, \\ C \le \bar{C}, & F = 0, \end{cases} \tag{4}$$

where: C – "benefit" of the individual;
$\bar{C}$ – "badness" of the individual;
S – support of a given class B;
$|A|$ – number of records that meet the condition A;
$|B|$ – number of records that meet the condition B;
$|A+B|$ - number of records, which comply with the rule: „If A, then B";
F – fitness of an individual.

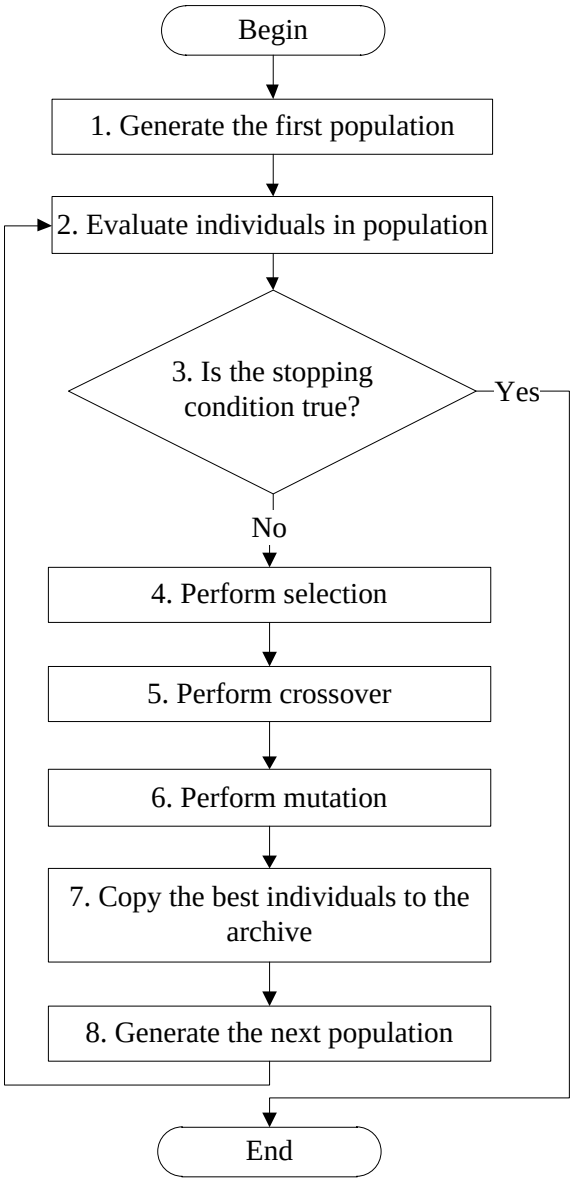

Fig. 1. Processes in the first phase of the proposed algorithm

3. Checking the stopping condition. The number of iterations was used as the stopping condition. If the defined number of iterations is successfully reached, and the desired number of populations is generated, then the algorithm steps into the second phase, otherwise the algorithm remains in the first phase [17], [18].
4. Selection. The roulette wheel selection was applied.
5. Crossover. The *N*-point crossover strategy was used.
6. Mutation. As the binary encoding was chosen to represent the individuals, the flipping concept was implemented for the mutation – 0 changes to 1 and vice versa.
7. Copying the best individuals to the archive. At the beginning of the process it is defined how many individuals will be copied to the archive. An individual can be copied to the archive only once, no duplicates are allowed.
8. Generating the next population. The strategy when parents compete with children for the right to be included in the next generation is used. Also the elitism was applied to overcome the loss of the fittest individuals [6], [12], [13].

In the second phase of the algorithm, which is shown in Figure 2, the classification rule mining process begins [9]. The second phase contains the next seven steps:

1. Sorting the individuals. Individuals in the archive are sorted by their fitness in descending order. The best individual will remain in the first position.
2. Moving the best individual *X* to the rule set *R*. The rule, encoded in the individual *X*, is included in the rule set *R*. The individual is excluded from the archive.
3. Excluding covered records. The records in the training set, covered by the rule that is encoded in the individual *X*, are excluded from the training set.
4. Checking the stopping condition. The algorithm stops in two cases: if the training set is empty or if the archive is empty – all individuals are excluded from the archive. Otherwise the cycle continues.
5. Re-evaluating each individual in the archive. The fitness of each individual in the archive is re-calculated using the records that remain in the training set.
6. Classifying test records. The rule set *R* is the classifier that is tested during this step. The rules are activated in the order they were included in the rule set. The first rule that matches the values of descriptive attributes of a record in the test set is activated. The target attribute value from an active rule is assigned to the test record. If no rule is activated by a test record, it will be charged as an error. The second option is to manually create a rule that does not depend on the values of the descriptive attributes, and assigns the most frequent target attribute value to the record.

7. Classifying error calculation. The classification error is calculated by using the equation (5):

$$CE = \frac{NP}{N}, \quad (5)$$

where NP – the number of incorrectly classified records;
N – the number of records in the test set;
CE – classification error.

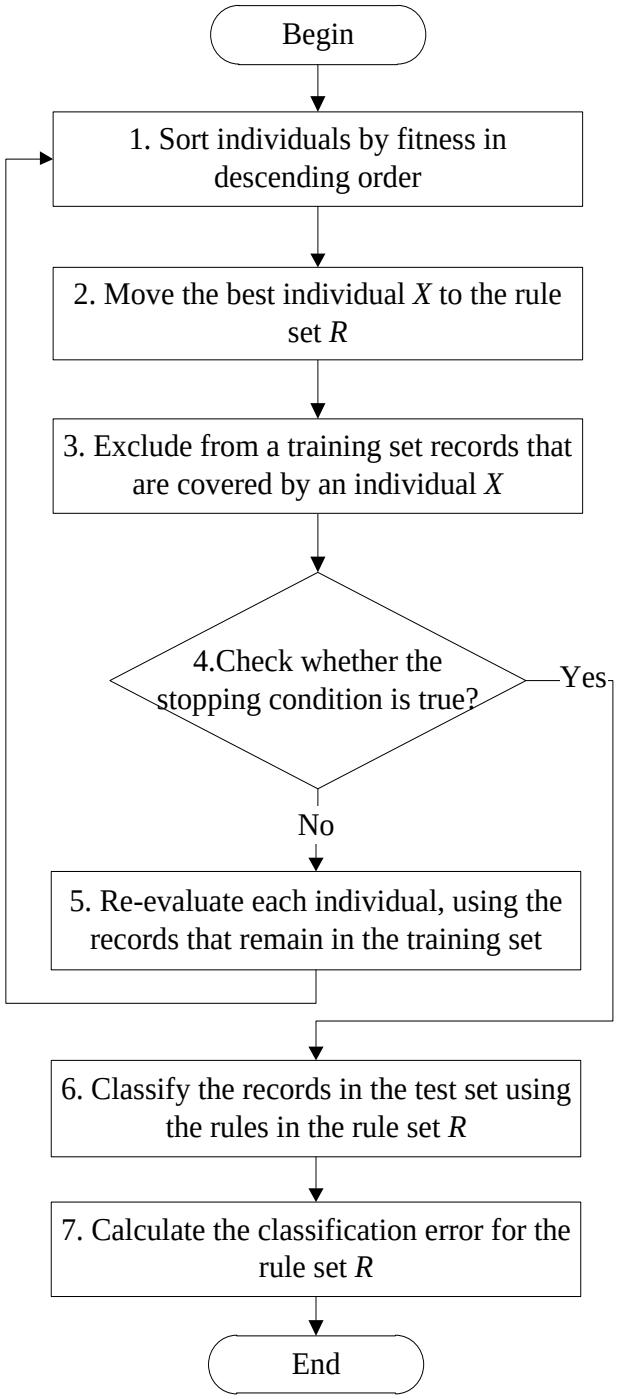

Fig. 2. Processes in the second phase of the proposed algorithm

III. The Developed Software

To evaluate the proposed algorithm, a specialised application was developed in *Microsoft Visual Studio 2010* environment, using the programming language *Visual Basic 2010*. The software implements the proposed genetic algorithm and mines classification rules. The program loads the information about the attributes from a **.csv* file and loads the datasets from the SQL Server 2008 database.

A number of specific tasks are implemented in the software environment:

1. Applying the modified genetic algorithm to classification rule mining;
2. Loading the information on attributes from a simple file in the **.csv* format;
3. Loading the datasets from the SQL Server 2008 database which stores the training and test sets;
4. Saving mined rules in a file;
5. Saving the classifier testing results in a file.

The efficiency of the algorithm and the necessary calculation time depends on the values of parameters defined at the beginning of the rule mining. Thus, the values of parameters are task specific. A user-friendly interface was developed (see Figure 3) to ensure the ability to define values for parameters and to follow the main process.

Fig. 3. User interface

IV. The Main Dataset and Data Pre-processing

The described algorithm was applied for solving the task, issued in the international competition "Data Mining Cup" in 2010. The dataset contains information on individual transactions of an online shop; the data include customer descriptive data and purchased items. Initially, the dataset contained 32428 records: 6051 record of class "1" and 26377 records of class "0". Each dataset entry was described with 33 attributes – 32 descriptive and one target attribute. The target attribute was a discrete attribute, showing whether the customer made another purchase within the next 90 days from the moment of the present transaction. Descriptive attributes contained both discrete and continuous attributes describing different transaction parameters – whether the customer was a new client or a registered one; sale and delivery dates; payment method, type and weight of the product, etc.

Several continuous attributes were excluded from the dataset, others were discretized and remained in the attribute set together with other qualitative attributes. For example, the weight of the purchased items cannot be used as numbers (kilograms), as there will be problems in the data analysis process, using the described algorithm. Therefore, the attribute "weight" was divided into five categories from 1 – very lightweight, till 5 – very heavy. One of the further researches will include the modification of the described algorithm to process not only attributes with a finite set of values, but also continuous attributes. During the data pre-processing some attributes were merged, and new attributes were created. An example of attribute combinations: using the transaction date and the user registration date, a new attribute was created, showing whether the transaction was made by a new user or by a user that had registered earlier. If the dates match, then the attribute has value "1" – a new user, if not, then it will be equal to "0" – a user that has registered earlier. The values of the selected attributes were digitized and a vocabulary of the values was formed. This step was included in data pre-processing to simplify the algorithm evaluation process. A set of attributes was chosen, where 13 attributes hold descriptive data, and the 14th was the target attribute.

The number of the class "0" records in the default dataset was more than four times greater than the number of the class "1" records. To lessen the impact of this inequality on the classification result, the classes were aligned. All the 6,051 records of class "1" were copied, and the same number of records of class "0" was randomly copied to the new dataset. As a result, the new dataset with 12,102 records was formed, having all records randomly sorted. The dataset was divided into two subsets – 70% (8471 records) for a training set and 30% (3631 records) for a test set.

V. The Main Results

The efficiency of the proposed genetic algorithm was evaluated using the developed software. In addition the data mining tool *Weka 3.6.4* [14] was used to evaluate other classification algorithms – the rule mining algorithm *JRip*, the decision tree classifier *J48* and the *naive Bayes* classifier; and to perform a comparative analysis with the genetic algorithm.

The obtained results are summarized in Table I, where the classification error and the number of generated rules are given for each of the evaluated algorithms.

TABLE I
Classification Results

	Algorithm	**Classification Error**	**Number of extracted rules**
1	*J48* (Decision Trees)	43.73%	79
2	*Genetic Algorithm*	44%	2
3	*JRip* (Rule extraction)	45.25%	2
4	*Naïve Bayes*	45.83%	-

The genetic algorithm has mined two rules. It was noticed that for some attributes rules contain all values, in other words – these attributes do not have any impact on the final result. Thus, these attributes were removed from the rules. The post-processed rules, mined with the genetic algorithm, are shown in Figure 4. If the values of descriptive data of a client match any of the rules found, then it is assigned class "0", which means that the client will not re-purchase within the next 90 days.

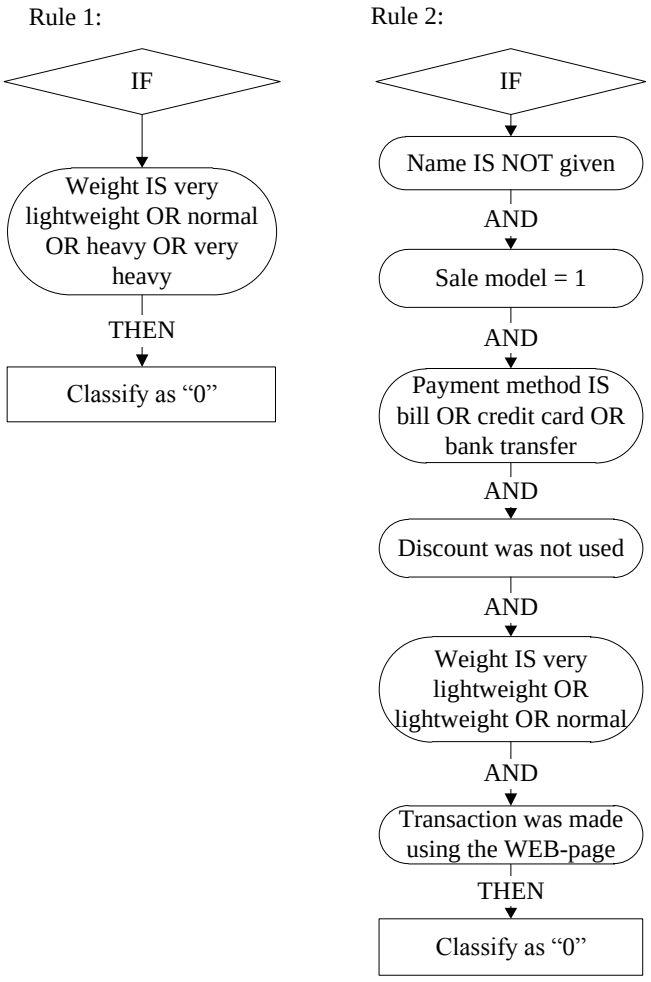

Fig. 4. Classification rules mined with the genetic algorithm

As can be seen, both rules classify clients that will not repurchase within the defined period of time. To classify clients that will re-purchase, a negative selection has been applied – if no rules are activated, then the client is assigned class "1" – the client will re-purchase within the defined period of time.

The obtained results show that the number of rules mined slightly affects the precision of the classifier. The decision tree classifier *J48* created 79 rules, which is 39.5 times greater than the number of rules, created by the genetic algorithm or by *JRip* algorithm. At the same time it gave the increase of precision by 0.27% and by 1.52%, respectively, to GA and *JRip* (see Table I). The larger number of rules may give better results, on the other hand; it will make the classifier less general and more specific to the dataset used for building the classifier. Balance between precision and number of rules should be found each time when building the classifier.

The best result for the genetic algorithm was obtained with these parameters: population size – 100 individuals; mutation probability – 0.3; the number of mutating genes – 5; application of two point crossover; use of the elitism – one best individual each time was copied to the next population; in each iteration 5 best individuals were copied to the archive (without duplicates). From Table I it can be seen that the genetic algorithm has a high error rate – 44%. At the same time, the efficiency of the GA is at one level with other algorithms evaluated on the same datasets, showing that the genetic algorithm is efficient enough to be applied in classification rule mining task.

VI. Conclusions

The objective of this research was to study the classification rule mining features of the genetic algorithm. Being an evolutionary optimisation algorithm, the classification rule mining is not the classical task to be solved with GA.

In the scope of the research, the idea of the GA was studied, and it was concluded that the genetic algorithm has all the necessary features to mine rules from statistical data. To practically evaluate the GA in classification rule mining, the classical GA was modified, a unique individual fitness evaluation method was proposed, as well as an algorithm for extracting the final rules set from the one created during the evolution process was developed. To perform the practical experiments, the specialised software was written.

The efficiency of the proposed GA-based rule mining method was evaluated on the pre-processed Data Mining Cup 2010 dataset. Also, the efficiency of the other classification algorithms – *J48*, *JRip* and *Naïve Bayes*, was evaluated on the same dataset. The results have shown that precision of the genetic algorithm is at the same level with other algorithms (see Table I). Comparing by precision and the number of rules mined, the efficiency of GA is better than efficiency of other algorithms. Having obtained these results, it can be concluded that the genetic algorithm, at least, is not worse than the others algorithms used and can be applied to mine classification rules from statistical data.

As a shortcoming of the proposed method, the capability of processing only discrete attributes can be mentioned. The main objective of future research is supposed to be the modification of the presented algorithm in order to gain the option of processing not only discrete data, but also continuous data.

References

[1] M. Mitchell, *An Introduction to Genetic Algorithms*. Cambridge, Massachusetts: A Bradford Book The MIT Press, 1998.
[2] D. E. Goldberg, Genetic *Algorithms in Search, Optimization & Machine Learning*. Reading MA: Addison-Wesley, 1989.
[3] J. R. Quinlan, *C4.5: Programs For Machine Learning.* San Mateo: Morgan Kaufmann Publishers, 1993.
[4] M. Mitchell, H. McGraw, *Machine Learning*. USA: R.R. Donnelley & Sons Comapny, 1997.
[5] S.N. Sivanandam, S.N. Deepa, *Introduction to Genetic Algorithms.* Berlin: Springer, 2008, pp. 39-129.
[6] R. Sarker, K-H. Liang, C. Newton, *A new multiobjective evolutionary algorithm*. Eur J Oper Res 2002, pp. 12–23.
[7] J. Horn, N. Nafpliotis and E. Goldberg, *A niched Pareto genetic algorithm for multiobjective optimization*. Orlando, USA: IEEE; 1994, pp. 82-87.
[8] R. Kohavi, R. Quinlan, *Decision Tree Descovery*: *Handbook of Data Mining and Knowledge Discovery*. USA: University Press, 1999, pp. 267-276.
[9] W.W. Cohen, *Fast Effective Rule Induction*, „Machine Learning: Proceedings of the Twelfth Conference" (ML95). California, 1995, pp. 115-123.
[10] S. Dehuri, A. Ghosh and R. Mall, *Genetic Algorithms for Multi-Criterion Classification and Clustering in Data Mining*. International Journal of Computing & Information Siences – Vol. 4, No. 3 (2006), pp. 143-154.
[11] A. Jain, M. N. Murty and P. J. FLynn, *Data Clustering: A Review*. ACM Computing Surveys, Vol. 31, No. 3, September 1999, pp. 364-423.
[12] A. Konak, D. Coit, A. Smith, *Multi-objective optimization using genetic algorithms: A tutorial*, Elsevier Ltd, 2005. [Online]. Available: http://www.rci.rutgers.edu/~coit/RESS_2006_MOGA.pdf [Accessed: 2006].
[13] W. Lu, I. Traore, *Detecting New Forms of Network Intrusion Using Genetic Programming*, Department of Electrical and Computer Engineering, University of Victoria, Canada, 2004. [Online]. Available: http://www.ece.uvic.ca/~wlu/COI04.pdf [Accessed: 2004].
[14] Weka, The University of Waikato, *Software Waikato*, The University of Waikato. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/
[15] D. Whitley, *A Genetic Algorithm Tutorial*, Computer Science Department, Colorado State University Fort Collins, 1993. [Online]. Available: http://www.cs.colostate.edu/~genitor/Pubs.html [Accessed 1995].
[16] A. Sukovs, L. Aleksejeva, K. Makejeva, *Datu ieguve: Pamati*. Rīga: RTU, SIA "Drukātava", 2007.
[17] Г. Вороновский, К. Махотило, *Генетические алгоритмы, искусственные нейронные сети и проблема виртуальной реальности*. Харьков: Основа, 1997.
[18] М.Т. Джонс, *Программирование искусственного интеллекта в приложениях* (перевод с английского Осипова, А.). Москва: ДМК Пресс, 2004, стр. 112-139.
[19] А. Еремеев *Генетические алгоритмы и оптимизация.* Омск: Издательство Омского государственного университета, 2008.
[20] И. Чубукова *Data Mining*. Киев: Бином, 2006.
[21] А. Шалыто. Генетические алгоритмы – Санкт-Петербург: СПбГУ, 2009.

Anna Galinina, MSc student, Riga Technical University, Institute of Information Technology, 1 Kalku Street, Riga, LV - 1658, Latvia
E-mail: gestija18@inbox.lv.
Anna Galinina received her BSc degree in computer management and computer science from Riga Technical University in 2011. Currently she is an MSc student at the Faculty of Computer Science and Information Technology at Riga Technical University. Her research interests include genetic algorithms, data mining and knowledge discovery.

Serge Parshutin, Lecturer, Riga Technical University, Institute of Information Technology, 1 Kalku Street, Riga, LV - 1658, Latvia,
E-mail: serge.parshutin@rtu.lv.
Serge Parshutin received his Ph.D. degree in information technology from Riga Technical University in 2011. Now he is a research fellow at the Faculty of Computer Science and Information Technology and a Lecturer at the Department of Modelling and Simulation, Riga Technical University. His research interests include data mining and knowledge discovery, intelligent information systems, intelligent agent technology, evolutionary computing and decision support.

Anna Galiņina, Sergejs Paršutins. Klasifikācijas likumu ieguve ar ģenētiskajiem algoritmiem e-veikala klientu novērtēšanai
Darba mērķis bija izpētīt ģenētiskā algoritma (ĢA) iespējas, lai piemērotu klasifikācijas likumu ieguvei no statistikas datiem. Ģenētiskais algoritms ir evolucionāras optimizācijas algoritms un var būt pielietots vairākās sfērās, taču klasifikācijas likumu ieguve nav klasiskais uzdevums ģenētiskā algoritma pielietošanai. Darbā izpētīts ĢA darbības princips un noteikts, ka ģenētiskajam algoritmam piemīt visas nepieciešamās īpašības klasifikācijas likumu ieguvei no datiem. Ģenētiskā algoritma efektivitātes praktiskajai novērtēšanai klasifikācijas likumu ieguves uzdevumā, tika modificēts ĢA pamata algoritms un piedāvāta metode indivīdu piemērotības novērtēšanai. Izstrādāta metode klasifikācijas likumu kopas veidošanai, izmantojot likumu kopu, kas atlasīta ar ģenētisko algoritmu.
Piedāvātā metode aprobēta ar iepriekš sagatavotu datu kopu, kuru izsniedza starptautiskajā konkursā Data Mining Cup 2010. Pielietojot izvēlēto datu kopu, tika novērtēta arī efektivitāte šādiem klasifikācijas algoritmiem: J48, JRip un Naïve Baiyes. Rezultāti parādīja, ka ģenētiskā algoritma precizitāte atrodas vienā līmenī ar citu izmantoto klasifikācijas algoritmu precizitāti. Salīdzinot algoritmu efektivitāti arī pēc iegūto likumu skaita, var secināt, ka piedāvātais ģenētiskais algoritms ir efektīvāks pār citiem algoritmiem. Pēc iegūtajiem rezultātiem var secināt, ka piedāvātais modificētais ģenētiskais algoritms nav sliktāks par citiem izmantotajiem klasifikācijas algoritmiem un to var pielietot klasifikācijas likumu ieguvei no statistikas datiem. Par piedāvātā algoritma nepilnību var minēt to, ka tas spēj apstrādāt tikai kategoriskus atribūtus. Turpmāko pētījumu mērķis būs nepārtraukto atribūtu apstrādes iespējas pievienošana piedāvātajam ģenētiskajam algoritmam.

Анна Галынина, Сергей Паршутин. Извлечение правил классификации с помощью генетических алгоритмов для оценки клиентов интернет магазина
Целью работы являлось изучение возможностей генетического алгоритма (ГА) применительно к извлечению правил классификации из имеющихся статистических данных. ГА является эволюционным алгоритмом оптимизации различных решений, тем не менее, извлечение правил из данных не является классической задачей для применения генетического алгоритма. В рамках проведённых исследований был изучен принцип действия ГА и установлено, что генетический алгоритм обладает всеми свойствами, необходимыми для извлечения правил из данных. С целью практической оценки ГА в рамках решения упомянутой задачи был модифицирован основной алгоритм действия ГА и предложен метод оценки пригодности индивидов. Также разработан алгоритм формирования конечного множества правил на основе отобранных генетическим алгоритмом.
Предложенный метод был опробован на предварительно подготовленном множестве данных, использованном на международном конкурсе Data Mining Cup 2010. Также на выбранном множестве данных была оценена эффективность таких алгоритмов классификации, как J48, JRip и Naïve Bayes. Результаты показали, что по точности классификации ГА находится на одном уровне с остальными классификаторами. Сравнивая эффективность по числу сформированных правил, ГА также оказался более эффективен, нежели остальные алгоритмы. Имея данные результаты, можно заключить, что предложенный модифицированный генетический алгоритм не хуже остальных использованных алгоритмов и может быть использован для извлечения правил классификации из данных. Недостатком предложенного алгоритма является то, что он способен обрабатывать только категорийные атрибуты. Целью дальнейших исследований станет устранение данного недостатка.