UDC 336.717.061

EVALUATION OF PERSONAL CREDITABILITY ON THE BASIS OF ARTIFICIAL INTELLIGENCE METHODS

M. Smirnova, Hu Xiaohui
Anning West Road, 88, Lanzhou, Gansu Province, 730070, China. E-mail: malaya.km@mail.ru
A. Judina
Kremenchuk Mykhailo Ostrohradskyi National University
vul. Pershotravneva, 20, Kremenchuk, 39600, Ukraine. E-mail: iyusa@ukr.net

Purpose. To investigate the possibilities of improving existing methods for determining the assessment of the personal creditworthiness of potential borrowers based on classification algorithms for further usage in the field of personal credit banking systems. **Methodology.** When solving the problem, general methods of data mining have been used, such as the decision tree, to identify potential borrowers, as well as existing approaches to machine learning to assess personal credit based on the probabilistic Naïve Bayes classifier. **Findings.** In this paper, the issue of determining personal creditworthiness has been considered, the existing approaches and in the issues of determining personal creditworthiness have been analyzed. The analysis of the personal data of potential borrowers has been carried out to reveal the possibilities of reducing the dimension of many factors included in the credit model. The developed model has been investigated by the Weka software application, which made it possible to visualize the results of identifying potential borrowers. **Originality.** Today the definition of personal credit evaluation in the field of personal banking credit systems does not have one best solving method. The article presents the results of our own research in this field, which can be used in the development of general methods of estimating personal creditworthiness. **Practical value.** This paper shows the practical possibility of reducing the dimension of the classification model without reducing its accuracy, which makes it possible to reduce the cost of constructing a classification tree and a probabilistic table in the process of machine learning.

Key words: personal creditworthiness, decision tree, machine learning, Weka, Naïve Bayes algorithm.

ОЦІНЮВАННЯ ПЕРСОНАЛЬНОЇ КРЕДИТОСПРОМОЖНОСТІ НА ОСНОВІ МЕТОДІВ ШТУЧНОГО ІНТЕЛЕКТУ

М. Смірнова, Ху Сяохуей
Ланьчжоуський Транспортний Університет
вул. Аньнін Вест, 88, м. Ланьчжоу, провінція Ганьсу, 730070, Китай. E-mail: malaya.km@mail.ru
А. Юдіна
Кременчуцький Національний Університет ім. Михайла Остроградського
вул. Першотравнева, 20, м. Кременчук, 39600, Україна. E-mail: iyusa@ukr.net

Досліджено можливості поліпшення існуючих методів визначення оцінки персональної кредитоспроможності потенційних позичальників на основі алгоритмів класифікації з метою подальшого використання в області банківських систем персонального кредитування. Розглянуто питання визначення персональної кредитоспроможності, проаналізовано існуючі підходи, методики і методи можливості машинного навчання в питаннях визначення персональної кредитоспроможності. Проведено аналіз персональних даних потенційних позичальників для виявлення можливостей скорочення розмірності множини факторів, що входять в модель кредитування. Розроблено модель досліджена засобами програмного додатка Weka, що дало можливість візуалізувати результати виявлення потенційних позичальників. Представлено результати власних досліджень по даному напрямку, які можуть бути використані при розробці загальної методики оцінювання персональної кредитоспроможності. Як результат, в даній роботі показано практичну можливість зменшення розмірності класифікаційної моделі без зменшення її точності, що дозволяє скоротити витрати на побудову класифікаційного дерева і ймовірнісної таблиці в процес машинного навчання.

Ключові слова: персональна кредитоспроможність, дерево прийняття рішень, машинне навчання, інтелектуальний аналіз даних, Weka, Naive Bayes алгоритм.

PROBLEM STATEMENT. Together with the rapid development of the economy, consumer credit is constantly growing. It allows users to use various types of loans. In this case banks often face with a credit risk factor, when a person for various reasons is not able to repay his loan on time.

With this type of card it is possible to regulate and stimulate the development of commercial banks. Modern intelligent systems, based on machine learning, are able to provide banks with an opportunity for a better and more comprehensive assessment of the personal credit risk of customers [1].

The goal of this research is to determine the possibility of improving existing methods for determining the assessment of the personal creditworthiness of potential borrowers based on classification algorithms.

MATERIAL AND RESULTS. For registration of personal loan customers need to undergo a series of procedures to issue appropriate forms and provide the necessary documents for the registration of the application. In turn, the bank takes to the application of a potential borrower, conducts a credit assessment of the personal data of the applicant, and conducts a number of necessary procedures, for a subsequent decision to grant a loan. To make a decision, the bank needs a comprehensive and scientific understanding basic information about the customer and his credit history, as well as assessing the situation with loans, to control the risk, avoid possible debt and reduce the resulting losses.

Currently, commercial banks use a credit score card to assess the creditworthiness of potential borrowers, which includes a number of general as and special issues, on the basis of which the report is generated on the client's creditworthiness.

The study on the evaluation of personal credit based on a credit score cards (scoring cards), indicators used are often not the same, and some are not the most important criteria cause a great impact on the credit rating, which in turn can lead to different results.

Based on the formula:

$$ER = \frac{b+c}{a+b+c+d}, \qquad (1)$$

where ER – a loan;

a – «good» customers;

b – «bad» clients;

c – «bad» clients who declare themselves as «good»;

d – «good» customers who declare themselves as «bad».

Get:

$$ER_1 = \frac{c}{a+c}, \qquad (2)$$

$$ER_2 = \frac{b}{b+d}, \qquad (3)$$

where ER_1 – preliminary risk% of «bad» customers classifying themselves as «good»;

ER_2 – preliminary risk % of «good» clients classifying themselves as «bad» [2].

There is an unobservable random variable Q_i, called the creditworthiness of the borrower *i,* which includes all the information necessary to assess credit risk. And since there is no opportunity to observe Q_i , the bank uses the available information about the borrower, such as a successful return of previous loans, financial condition of the borrower, as well as other factors that contribute to the assessment Q_i [2].

The likelihood that the client will receive a loan is an increasing function and determined by next formula:

$$P_{rc} = f(Q_i). \qquad (4)$$

An example of this assessment of personal creditworthiness based on scoring for potential customers is Fig. 1.

Schedule a personal credit ratings based scoring assessment, it shows a direct dependence of the amount marked «right» answers (factors) from the number of questions (pre-determined factors) with overcoming the critical threshold of 15 points to determine the possibility of issuing a loan.

Figure 1 – Graph of personal credit ratings based scoring

For a detailed consideration of the possibility of lending by commercial banks on the basis of scoring cards, the entire evaluation system can be divided into several criteria, one of them is a time criterion. According to that all loans can be divided depending on the loan term into the following categories:

- Short-term loans - up to 1 year.
- Medium-term loans - from 2 to 5 years.
- Long-term loans - from 5, 10 years and more [3].

Based on a set of conditions for issuing a loan, as one of the options, all potential borrowers, authenticated by other criteria, can be strictly eliminated by the time criterion. A fragment of the scheme for determining the issuance of a loan for the sample of the credit amount, credit rate and, by the time criterion is shown in Fig. 2.

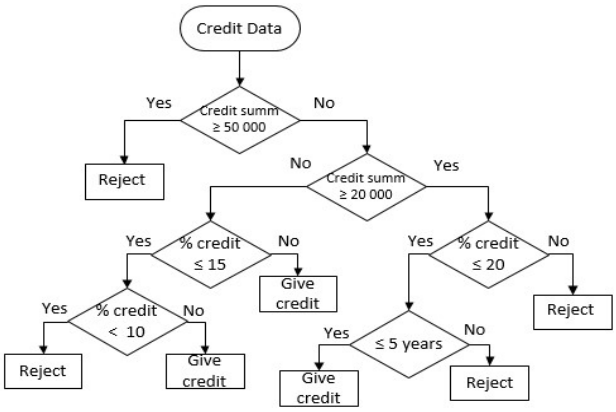

Figure 2 – A fragment of the scheme for determining the issuance of a loan

As software for the implementation of the task, the authors selected the software Weka (Waikato Environment for Knowledge Analysis), which is a set of visualization tools and algorithms for the intellectual analysis of data and the solution of prediction problems, together with a graphical user interface for accessing them.

In carrying out investigations on the basis of available databases [3] authors have analyzed the factors, affect the evaluation of personal credit in lending by financial institutions.

The test database with a set of data for Weka consists of 1000 customer records represented in the form of rules - sets of 21 attributes.

The attributes are:

- checking_status
- duration
- credit_history
- purpose
- credit_amount
- savings_status
- employment
- installment_commitment
- personal_status
- other_parties
- residence_since
- property_magnitude
- age
- other_payment_plans
- housing
- existing_credits
- job
- num_dependents
- own_telephone
- foreign_worker
- class

To carry out training and testing procedures the database was broken into training DB (800 entries) and test DB (200 entries), the records in the database are represented by a set of attributes, for each of which certain boundaries of the partition are distinguished.

At initial loading of a DB the system divided all clients into age categories in the range from 19 to 75, selecting the average value of customer category in 47 years. Fig. 3 shows a graph of the age index of the client base at the initial stage of creditworthiness assessment. Where the x-axis represents the age category of the clients, and the y-axis represents the number of people of the given age category.

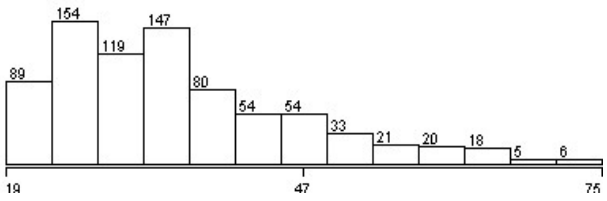

Figure 3 – Graph of the age index

Based on a study of existing models of evaluation of personal credit, the authors are building their own method of determining the assessment of the personal creditworthiness of potential borrowers based on the integrated decision tree for further use in the field of personal banking credit systems.

Decision tree (classification tree or regression tree) - decision support tools, used in statistics and analysis of data for forecast models.

In intellectual data analysis, decision trees can be used as mathematical and computational methods to help describe, classify and summarize a set of data that can be written as follows:

$$(x, Y) = (x_1, x_2, x_3, \ldots, x_n, Y). \quad (5)$$

The dependent variable Y is the target variable, which must be analyzed, classified and generalized. The vector x consists of input variables: $x_1, x_2, \ldots x_n$ etc., which are used to perform this task.

Using the classification tree a primary analysis of the data potential creditworthy customers. That gives a strict division by the consumer criterion of «intention». Fig. 4 shows a part of the classification tree is responsible for the selection criteria for a consumer «purpose».

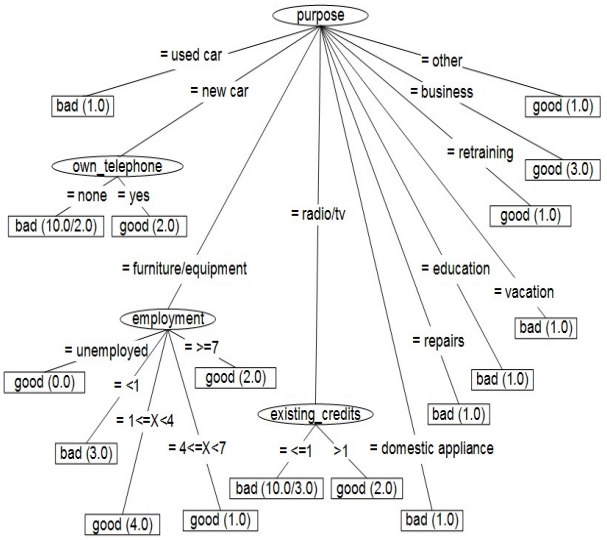

Figure 4 – Fragment of a classification tree on consumer criteria «purpose»

The task of classification is, refer to a previously unknown entity to a particular class. In this case, the definition of creditworthy customers under existing loans. Based on the data provided, we had two classes: positive and negative. Then the output of the classifier can be observed four different situations:

TP true positive: number of examples predicted positive that are actually positive;

1) FP false positive: number of examples predicted positive that are actually negative;
2) TN true negative: number of examples predicted negative that are actually negative;
3) FN false negative: number of examples predicted negative that are actually positive.

After classification tree learning methods software Weka following characteristics were obtained:

- J48 pruned tree
- Number of Leaves: 86
- Size of the tree: 116
- Correctly Classified Instances 672 84 %
- Incorrectly Classified Instances 128 16 %
- Kappa statistic 0,5885
- Mean absolute error 0,2447
- Root mean squared error 0,3498
- Relative absolute error 58,3615 %
- Root relative squared error 76,4131 %
- Total Number of Instances 800

Where:

Correctly classified instances – Shows the number of correctly classified clients.

Incorrectly classified instances of instances - Shows the number of incorrectly classified clients [3].

Kappa statistic – measures the agreement of prediction with the true class – 1,0 signifies complete agreement.

The Mean absolute error – is the average of all absolute errors [4].

Root mean squared error (RMSE) – is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed. The RMSD represents the sample standard deviation of the differences between predicted values and observed values.

Relative absolute error – the relative absolute error takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor [5].

Root relative squared error – is the root relative squared error is relative to what it would have been if a simple predictor had been used. More specifically, this simple predictor is just the average of the actual values. Thus, the relative squared error takes the total squared error and normalizes it by dividing by the total squared error of the simple predictor. By taking the square root of the relative squared error one reduces the error to the same dimensions as the quantity being predicted [6].

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,939	0,193	0,849	0,939	0,892	0,600	0,850	0,902	good
0,607	0,061	0,810	0,607	0,694	0,600	0,850	0,759	bad
0,84	0,294	0,837	0,84	0,833	0,600	0,850	0,859	Weighted Avg.

TP Rate – true positive rate, called the sensitivity of the classification algorithm.

FP Rate – false positive rate, the specificity of the classification algorithm [5].

Precision – share of objects, called classifiers are positive and are actually positive.

Recall – share index object of a positive class of all the objects of positive class, who found the algorithm.

F-Measure – the way to combine precision and recall in an aggregated quality criterion. What might be called the average harmonic between precision and recall.

ROC Area – a graph that allows to estimate the quality of a binary classification, and displays the relationship between the proportion of the total number of objects trait carriers correctly classified as bearing a sign and a share of objects from the total number of objects not bearing characteristic, misclassified as a carrier indication by varying the threshold decision rule [6].

=== Confusion Matrix ===

a	b	classified as
527	34	a=good
94	145	b=bad

According to the received data, the following studies were carried out to improve the characteristics of Number of Leaves, Size of the tree, Correctly Classified Instances, Recall and ROC Area.

According to the records of the original database, the size of the originally constructed tree has the size 116, 86 leaves, the accuracy is 84%, the percentage of objects in the positive class is 0,840, and the ROC Area is 0,850.

According to the confusion matrix we have 527 clients as TP (true positive), 145 - FN (false negative), 34 - FP (false positive), and 145 - FN (false negative), which in total amount to 179 customers that are not ranked either one particular class and, accordingly, are listed in the «risk zone».

To verify the accuracy of the data analysis model in addition to the classification tree, test database was also analyzed using the Naïve Bayes classifier.

Naïve Bayes classifier is a method of controlled learning (teaching with a teacher), and a probabilistic method of classification. Naïve Bayes classifier allows you to lock the uncertainty about the models, in principle, by the sense of the definition of probability of results.

Abstractly, Naïve Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector $x = (x_1, x_2, ...x_n)$ representing some n features (independent variables), it assigns to this instance probabilities $p(C_i \mid x_1, ...x_n)$ for each of 'i' possible outcomes or classes C_i. [7].

Naïve Assumption of "class conditional independence":

$$P\left(\frac{X}{C_i}\right) = \prod_{k=1}^{n} P\left(\frac{X_k}{C_i}\right) \quad (6)$$

$$P\left(\frac{X}{C_i}\right) = P\left(\frac{X_1}{C_i}\right) \cdot P\left(\frac{X_2}{C_i}\right) \cdot \ldots \cdot P\left(\frac{X_n}{C_i}\right) \quad (7)$$

The results of checking the accuracy of data mining models for the test database using the Naïve Bayes classifier.

=== Summary ===

Correctly Classified Instances	613	76,625 %
Incorrectly Classified Instances	187	23,375 %
Kappa statistic	0,4132	
Mean absolute error	0,281	
Root mean squared error	0,4048	
Relative absolute error	67,0416 %	
Root relative squared error	88,4404 %	
Total Number of Instances	800	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,870	0,477	0,811	0,870	0,839	0,417	0,813	0,909	good
0,523	0,130	0,631	0,523	0,752	0,417	0,813	0,609	bad
0,766	0,373	0,757	0,766	0,759	0,417	0,813	0,820	Weighted Avg.

=== Confusion Matrix ===

A	b	Classified as
488	73	a=good
114	125	b=bad

The constructed probabilistic classification model has an accuracy of 76,625%, with an indicator of the proportion of objects of the positive class at 0,766, and ROC Area 0,813.

As the data taken from the confusion matrix, we have 488 clients as TP (true positive), 125 - FN (false negative), and 73 - FP (false positive), and 114 - FN (false negative), which add up to 187 clients, who are included in the «risk zone».

Carrying out the analysis of the testing data of the training and verification databases of two different estimation systems gave us the initial notions of the correctness of the projected model. The verification database consists of 200 clients, whose data we need to process.

Table 1 provides comparative data on the learning and testing data of the regressive tree (J48 pruned tree) and the data of the Naive Bayes classifier.

Table 1 – Learning and testing data J48 pruned tree and Naive Bayes

	J48 pruned tree		Naive Bayes	
	Training set	Test set	Training set	Test set
Number of Leaves:	86	86		
Size of the tree:	116	116		
Correctly Classified Instances	84%	75,5 %	76,625 %	79%
Incorrectly Classified Instances	16%	24,5 %	23,375 %	21%
Kappa statistic	0,5885	0,4095	0,4132	0,4775
Mean absolute error	0,2447	0,3049	0,281	0,2696
Root mean squared error	0,3498	0,424	0,4048	0,4022
Relative absolute error	58,3615%	71,6285%	67,041 %	63,3378%
Root relative squared error	76,4131%	91,2328%	88,441 %	86,5455%
Total Number of Instances	800	200	800	200

Table 1 shows, that J48 was obtained during training Correctly Classified Instances of 84%, and when testing 75,5%. At the same time for Naive Bayes, when training was received Correctly Classified Instances was 76,625%, and when tested 79%. So, we can conclude that the method of Naive Bayes gave a better solution than C4.5.

The next stage of the study calculation factors became insignificant and their impact on our model using different ranking methodologies which will reduce the gap for Correctly Classified Instances when computing the above two methods.

After applying the ranking various methods on the training database, it was obtained by the data set, which for more convenience of consideration was divided into two parts with 10 attributes each. The first part represented attributes of a more significant nature, and the second less significant.

In Table 2 a list of attributes defined as «most significant» in the classification model using various ranking methods is presented.

Table 3 shows the attributes defined as «least significant» by ranking methods.

Table 2 – List of significant attributes

Methods	Number of attribute									
Correlation AttributeEval	1	2	5	15	6	9	13	14	8	20
GainRatio AttributeEval	1	20	2	5	3	6	10	15	4	14
InfoGain AttributeEval	1	3	2	4	6	5	12	7	9	15
OneR AttributeEval	3	7	20	8	6	11	4	19	10	12
ReliefF AttributeEval	1	3	7	4	9	8	19	12	6	17
SymmetricalUncert AttributeEval	1	2	3	5	6	4	20	15	9	12

Table 3 – List of insignificant attributes

Methods	Number of attribute									
Correlation AttributeEval	3	4	7	12	19	17	16	18	10	11
GainRatio AttributeEval	9	12	7	17	19	13	16	11	8	18
InfoGain AttributeEval	20	14	10	17	19	18	11	13	8	16
OneR AttributeEval	15	18	17	1	14	16	2	9	13	5
ReliefF AttributeEval	2	10	18	14	13	5	11	16	15	20
SymmetricalUncert AttributeEval	7	14	10	17	19	8	18	16	11	13

After analyzing the ranked data, we can say that such attributes as checking_status, 3 (credit_history), 4 (purpose), and 6 (savings_status) have the highest influence on the accuracy of the model. And such attributes as residence_since, 13 (age), 16 (existing_credits), 17 (job) and 18 (num_dependents) practically do not influence the accuracy of the model. To confirm this conclusion, attributes that are the least significant were removed from the training database in turn. The result of decreasing the dimension of the set of attributes is presented below in the summary Table 4.

Based on the results of the application of the ranking, for the decision tree the main improvement was reduction the estimation model tree size by 29 points (from 116 to 87) with the result Correctly Classified Instances not decreased. Also for the Bass Classifier can be noted minor changes in the Correctly Classified Instances, which practically did not affect the initial ability to accurately assess creditworthiness.

Table 4 – Result of decreasing the dimension of attributes

	J48 pruned tree		Naive Bayes	
	Training set	Test set	Training Set	Test set
all 21 attributes				
Number of Leaves:	86	86		
Size of the tree:	116	116		
Correctly Classified Instances	84%	75,5 %	76,625 %	79%
Incorrectly Classified Instances	16%	24,5 %	23,375 %	21%
elimination 18 attribute				
Number of Leaves:	60	60		
Size of the tree:	87	87		
Correctly Classified Instances	83,75%	72,5 %	77%	78,5 %
Incorrectly Classified Instances	16,25%	27,5 %	23%	21,5 %
elimination 16 attribute				
Number of Leaves:	60	60		
Size of the tree:	87	87		
Correctly Classified Instances	83,75%	72,5 %	76,5 %	78,5 %
Incorrectly Classified Instances	16,25%	27,5 %	23,5 %	21,5 %
elimination 13 attribute				
Number of Leaves:	60	60		
Size of the tree:	87	87		
Correctly Classified Instances	83,75%	72,5 %	76%	78%
Incorrectly Classified Instances	16,25%	27,5 %	24%	22%

At the same time, the total confusion matrix also underwent small changes. The summary data is presented in Table 5.

Table 5 shows the training and test data at the initial and final stages of the study. Due to the deteriorating Correctly Classified Instances in both cases unclassified clients was, in general, 3 more people. It was noted that some of the clients «unclassified» by the J48 method were successfully classified by the NB method, and vice versa - those clients that NB could not recognize were successfully classified by the J48 method.

Table 5 – Summary data

J48 pruned tree				Naive Bayes			
on the start of research							
train	527	34		488	73		
	94	145		114	125		
			128				187
testing	117	20		124	13		
	29	34		29	34		
			49				42
in finish							
train	509	52		485	74		
	78	161		118	121		
			130				192
testing	115	24		123	14		
	31	34		30	33		
			51				44

Based on the results of the studies, the following conclusions can be drawn:

1) The use of existing methods of ranking significant attributes of classification models can significantly reduce the dimensionality of the vector of input parameters of the model with virtually no loss of accuracy of the model, which reduces the costs of constructing a classification tree and a probability table in the process of machine learning.

2) Sharing Classifiers based on decision tree and probabilistic classifiers allows you to build more accurate classification model, allowing to reduce a share of those clients, on which each of the models separately can't give an unambiguous decision on the possibility of lending.

CONCLUSIONS. This article presents the results of a study on the topic: «Evaluation of personal creditability on the basis of artificial intelligence methods» during which it was discussed existing approaches, methods and techniques in the field of machine learning of credit systems. As part of the research work the simple model was proposed decision tree, on the basis of which the authors built further studies. For determine the credit worthiness of borrowers the authors conducted studies using a decision tree, and the naive Bases classifier. The study attributes depending different ranking methods were investigated, due to this it was possible to reduce the dimensionality of the model and reduce the proportion of customers for which the system could not provide the correct solution.

ACKNOWLEDGEMENTS. We would like to thank Alla Chudakova and Olia Petrushenka for their moral support and wonderful contribution towards completion of this research.

REFERENCES.

1. Shepeleva, M.V. (2006), Modely kreditnogo scoringa [Models of credit scoring] Donetsk, Available at: http://masters.donntu.org/ 2006/kita/shepeleva/ library/metod%20scoring.pdf

2. Liu, Y. (2002,) The evaluation of classification models for credit scoring. Arbeitsbericht 02/2002, Institut fur Wirtschaftsinformatik.

3. Remco, R., Bouckaert, Peter, Reutemann, Alex, Seewald, David, Scuse (2013), WEKA Manual for Version 3-7-8, University of Waikato, Hamilton, New Zealand.

4. Witten, I. H. and Frank, E. (2005), Data Mining: Practical machine learning tools and techniques. 2nd edition Morgan Kaufmann, San Francisco

5. Howard, James (2012), Predictive Analysis available at: https://www.ibm.com/developerworks /ru/library/bd-javaweka/

6. Analyzing GeneXproTools Models Statistically Section2 (2015), available at: https://www.gepsoft.com /gxpt4kb/Chapter10/Section2/SS15.htm

7. Analyzing GeneXproTools Models Statistically Section2 (2007), available at: https://www.gepsoft.com/ gxpt4kb/Chapter10/Section1/SS07.htm

8. Webb, G. I., Boughton, J. and Wang, Z. (2005), "Not So Naive Bayes: Aggregating One-Dependence Estimators". Machine Learning, 58(1), 5–24. doi: 10.1007/s10994-005-4258-6

9. Macine Learning Repository (2014), University of California, Irvine, available at: http://www.ics.uci.edu/~mlearn/MLSummary.html

ОЦЕНКА ПЕРСОНАЛЬНОЙ КРЕДИТОСПОСОБНОСТИ НА ОСНОВЕ МЕТОДОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

М. Смирнова, Ху Сяохуэй

Ланьчжоуский Транспортный Университет

ул. Аньнин Вест, 88, г. Ланьчжоу, провинция Ганьсу, 730070, Китай. E-mail: malaya.km@mail.ru

А. Юдина

Кременчугский Национальный университет им. Михаила Остроградского

ул. Первомайская, 20, г. Кременчуг, 39600, Украина. E-mail: iyusa@ukr.net 2

Исследованы возможности улучшения существующих методов определения оценки персональной кредитоспособности потенциальных заёмщиков на основе алгоритмов классификации с целью дальнейшего использования в области банковских систем персонального кредитования. Рассмотрен вопрос определения персональной кредитоспособности, проанализированы существующие подходы, методики и методы возможности машинного обучения в вопросах определения персональной кредитоспособности. Проведен анализ персональных данных потенциальных заёмщиков для выявления возможностей сокращения размерности множества факторов, входящих в модель кредитования. Разработанная модель исследована средствами программного приложения Weka, что дало возможность визуализированного представления результатов выявления потенциальных заёмщиков. Представлены результаты собственных исследований по данному направлению, которые могут быть использованы при разработке общей методики оценивания персональной кредитоспособности. Как результат, в данной работе показана практическая возможность уменьшения размерности классификационной модели без уменьшения ее точности, что позволяет сократить затраты на построение классификационного дерева и вероятностной таблицы в процесс машинного обучения.

Ключевые слова: персональная кредитоспособность, дерево принятия решений, машинное обучение, Weka, Naive Bayes алгоритм.

Стаття надійшла 15.12.2017.