The information-mathematical system of the borrower's solvency prediction

V.A. Alekseeva[1], Yu.E. Kuvayskova[1]

[1]*Ulyanovsk State Technical University, Severny Venec, 32, 432027, Ulyanovsk, Russia*

Abstract

The paper is about research of the algorithms, methods of the classification and prediction objects' groups, depiction of the information-mathematical system, which is created on these algorithms' basis. They use variety methods of the machine learning and their compilations – aggregative classifier, all of these is for the solution of the classification's problem, particularly borrower's solvency prediction. This helps to make previous preparation of the source data, which also contents discretisation, missed data's recovery and detection of the important factors for statistics, how to use these methods of the classification and create cogeneration models, how to analyze quality of these models using statistical measures, to predict objects' groups.

Keywords: machine learning; aggregative classifier; statistical data analysis; classification; solvency; prediction

1. Introduction

Consider the problem of objects' binary classification [1], in which every object $K_i (i = 1,..., N)$ is characterized *m*-measured vector of the features $(X_1 ... X_m)$, which can be numeral or nonnumeric value and can create sample for the further researching. Using value of these features we need to predict value of the binary characteristics of objects y. Example for this type of matter is matters of technical diagnostics and classification of the object's condition [7,8], detection the fact of emission or absence the only signal, normal or abnormal element's condition etc.

There is a solution of the problem of binary classification by the example of credit score, which is in borrower creditworthiness assessment [10].

The increase the amount of debt on loans, increase the risk of loan default, also rivalry on the credit market – all of these need improvement of known techniques of the assessment and prediction of the borrower's solvency with the aim to more accurate assessment of the credit risk and making the right decision in the case of issuance of credit. Known methods cannot help finding more accurate models for solvation this problem.

Information-mathematical system was made in the aim of decrease the amount of borrowers' debts and to provide return of the credits, this system allow assessing borrower's creditworthiness at the stage of making decision of issuance of credit. For the assessing creditworthiness was used methods of machine learning [2] with aggregation different classifiers on the basis of decision trees, neural net, discriminant analysis, Bayesian classifier, SVM, logit regression etc.

2. Aggregative classifiers

Nowadays there are a lot of models and methods for the solution the problem of prediction the class of objects. Next methods was used for the analysis of credit risks' assessment: decision trees [14], neural nets [13], discriminant analysis [2], Bayesian classifier [5], SVM, logit regression [6], bagging decision trees, fuzzy inference models [10], created function method. Every method has advantages and disadvantages. For example, there is no possibility to use created function method for data's prediction, which set of characteristic values disagrees at least with one set from learning sample. For using Bayesian attitude it needs to bring given data to interval scale to variables were discrete, otherwise important information will be lost. There is no general model, which one can help assess belonging of the object to one of classes with high accuracy.

Depends on concrete case every method of machine learning can be the best one from the side of prediction's accuracy, so it offers joint using of different classifiers, which are made on variety parts of learning sample [3]. If use nine methods listed above, so it is possible to get $2^9 - 9 - 1 = 502$ all kinds of combinations of different models using method of full enumeration.

To decide belonging borrower to one of the classes (creditworthy or not) on the basis of results of parallel application to the original sample of certain methods of the classification, aggregation results is possible on three grounds:

- by average value (the possibility of object belongs to class $y = 1$ ("creditworthy client") shall be considered as arithmetical average of belonging probabilities of object to class $y = 1$, which were found out using all nine methods of classification);
- by median (first of all, expansion is ranging, which contains results base methods of classification in the combination, probability is counted through calculation result of average classifier in the case of their odd number or in the case of even number probability is counted through half-sum of results of average classifier);
- by voting (result of the aggregative classifier in this case is average result of classification's basic methods, which gave fact of the belonging object to $y = 1$ class with $\geq 0{,}1$ probability).

There is a solution algorithm for the assessment of clients' creditworthiness on the basis of aggregative classifies, it contains next stages:

1) Formation and processing of original sample. This stage consists in dividing sample into learning one (for making classification models) and test one (for checking accuracy of the made models), recover missed data [9], discretisation of some characteristics and searching aspects, which influence on the output characteristic y;
2) Parallel creation of nine classification models on the learning sample;
3) Creation aggregative classifier;
4) Prediction on the basis of test sample of new clients' creditworthiness using all constructed models;
5) Achievement of the prediction result of creditworthiness of every client. It evaluates average probability value of all constructed models on this stage;
6) Choice of the best model, which means model with the highest accuracy of the prediction. The accuracy is found out using certain measures [12].

3. Information-mathematical system of credit score

Information-mathematical system of credit score was made in the basis of listed above algorithm. It allows predict the class of the object (for example, borrowers' creditworthiness) using learning sample. Software package was devised in the programming support environment Matlab R2014a, which contains all of methods initial data computing and amount of algorithms machine learning, which are needed for solvation the classification problem. Initial data is information about clients, which is personal details and relevant class of the creditworthiness "old" clients; personal details of "new" clients; personal details, credit history and credit transaction terms and conditions of borrowers, who repay a loan.

Program allows making previous preparation of initial data: recover missed information; characteristics' discretisation; coding nonnumeric data; selection statistically worthy characteristics. All of listed above classification methods instantiates in the program, which includes aggregative classifier with the possibility of selecting criteria of aggregation (by average value, by median, by voting).

Method of *L*-fold cross-check is used for making classifiers for getting unbiased estimator of quality parameter. The essence of this method is in division original sample to *L* non-crossing parts, which are approximately equals to each other by the extent. It is possible to choose *L*'s value, it varies from 3 to 10. In turn every part serves as test sample, rest ones aggregates to learning sample. Summative assessment of the classifier's quality is defined by averaging mistakes in all *L* test sample. It allows exclude possibility of fudge to the best prediction.

In conclusion constitutes values of quality of created models for three cutoff thresholds (cutoff threshold – value, which if target is higher than the target become the one from class $y = 1$): cutoff threshold 0,5; definitive cutoff threshold; custom cutoff threshold. Definitive classification threshold is the least deviation between mistakes of I-class and II-class.

Quality control of created classification models and aggregative classifiers makes with helping of next characteristics [12]: mistakes of I-class and II-class, ROC curves, area under ROC curve, MSPE and percent of right predictions creditworthy clients and right prediction percent of non-creditworthy clients.

Customer can estimate which method or method combination gives the best result for objects and make prediction for original set of characteristic values using specified criteria. Working process of aggregative classifier is making by program, so optimal method combination is formed automatically using special criteria, after this customer can differ compare results of aggregative classifier and basic classification methods.

4. Case study of developed system of credit score

As the first example there are results of program working on realization aggregative classifier for sample of German bank's clients, which includes 900 borrowers, who have 20 characteristics (status of current checking account, credit history, loan purpose, credit length, loan proceeds, average balance on the savings account, work experience in the last place, income in %, family status, guarantors, permanent residence in the last place, data on property, age, available loans, type of housing, number of previous loans in this bank, type of activity, number of dependents, phone availability, citizenship), and one dependent binary variable (borrower is creditworthy and non-creditworthy). This program provides previous data processing, including characteristics' discretisation and coding nonnumeric data, such as citizenship of client, education, family status etc., with numbers. Nine different classification methods and aggregative classifier are analyzed. Aggregation was made by average value. It is possible to make aggregation using all of three characteristics. A 10-fold cross-check was used in this classification.

For target sample is got optimal aggregative classifier with 0,5 cutoff threshold, which contains next methods: neural nets, logit regression, bagging decision trees, created function method, fuzzy inference models. There is results of program in tab.1. The best classification result is got with helping of aggregative classifier, because of mean-root error of aggregative classifier is less than other methods; the highest percent of right prediction of creditworthy clients is in two methods: aggregative classifier and bagging decision trees, but I-class error of aggregative classifier is lower; aggregative classifier gives average value for prediction for non-creditworthy clients, but with minimal II-class error.

Table 1. Results of German bank's borrowers' classification .

Classifier	MSE	Creditworthy ($y=1$)		Non- creditworthy ($y=0$)	
		Right prediction, %	I-class error, %	Right prediction, %	II-class error, %
Neural net	0,1743	84,1	56,8	44,2	15,8
Discriminant analysis	0,1862	84,5	48,0	57,0	16,7
Bayesian classifier	0,2012	76,2	32,8	62,4	28,5
SVM	0,1653	88,5	47,7	63,2	15,1
Decision trees	0,2395	79,1	46,1	57,6	26,8
Logit regression	0,1852	88,7	50,2	50,4	13,3
Bagging decision trees	0,1532	88,1	49,3	51,1	11,9
Created function method	0,4576	35,7	4,8	95,3	70,2
Fuzzy inference models	0,1845	79,3	39,1	68,2	23,5
Aggregative classifier	0,1552	88,5	36,5	61,9	11,0

There are just three levels of quality of classifiers in this table. Also, program allows form diagrams, which show areas under ROC curves (AUC). Fig.1 shows such diagram for target sample. ROC curve [12], also known as curve of errors, shows correlation between deal of right positive classifications from whole number of negative classifications with variation threshold of decision rule. AUC level allows assay diagram of ROC curve. The more AUC level is higher, the more classifier is accurate. Diagram show that aggregative classifier and bagging decision trees gives the most accurate classification result, but AUC of aggregative classifier has higher value.

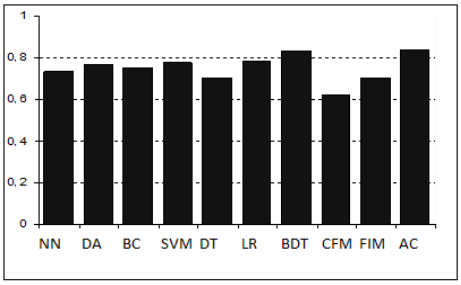

Fig.1. Areas under ROC curves for target sample.

The paper [4] analyses sample of borrowers of German banks but with bigger extent (1000 examinations). Decrease number of examination inconspicuous changes results of classification.

Researching of data about creditworthiness of Australian borrowers was made similar. Names of variables and their values were coded for Privacy Policy. Data includes one of dependent binary variable, which means creditworthiness (takes value 0 in case of non-creditworthy client or 1 in case of creditworthy client) and 14 independent characteristics. There are 690 examinations.

Optimal aggregative classifier for target sample was found 0,5 with cutoff threshold, which contains next methods: neural nets, logit regression, Bayesian classifier and fuzzy inference models. Results of working are in the Table 2. The best result was made with aggregative classifier.

Table 2. Results of classification of Australian bank's borrowers.

Classifier	MSE	Creditworthy ($y=1$)		Non-creditworthy($y=0$)	
		Right prediction, %	I-class error, %	Right prediction, %	II-class error, %
Neural net	0,1258	88,6	56,2	67,3	13,6
Discriminant analysis	0,2511	75,8	42,1	54,2	25,2
Bayesian classifier	0,1253	84,6	58,3	62,8	21,8
SVM	0,1648	87,2	42,2	55,1	20,1
Decision trees	0,3519	69,8	51,0	56,8	18,4
Logit regression	0,2157	78,9	42,9	61,5	12,6
Bagging decision trees	0,1642	76,8	45,2	54,5	20,3
Created function method	0,5862	58,1	31,2	66,8	25,8
Fuzzy inference models	0,1683	87,6	48,6	57,0	16,1
Aggregative classifier	0,1146	89,1	31,8	63,1	12,1

These examples show possibilities of this information-mathematical system of credit score. Method is selected from all of possible methods and it allows predict creditworthiness or non-creditworthiness of clients at the same time, minimizing mean-root error and I-class, II-class errors and maximizing AUC level. Using current method it is possible to find in which class is target using specified set of values. Also, this program allows renovate models if there is new data.

5. Conclusion

It considered using nine known methods of machine learning and their combinations for solvation the problem of binary classification of objects. It is not possible to explain effectiveness just one of the methods, because for different samples, even for different parts of one sample, is possible to get variety results. These methods and algorithm of making aggregative classifiers are realized in terms of information-mathematical system of credit score.

This program allows find the best model or optimal aggregative classifier, Classifier was the best one for targets samples. In the case with German borrowers the most accurate prediction was received by using combination of next methods: neural nets, logit regression, bagging decision trees, created function method, fuzzy inference models; classifier includes neural nets, logit regression, Bayesian classifier and fuzzy inference models in case with Australian data. Aggregative classifier helps to get the purpose – increase of prediction accuracy of creditworthiness clients of the bank.

This system of the credit score can be used for any problem of binary classification, for prediction of technical condition of objects [7,8] in particular and for prediction of signal presence or absence.

References

[1] Ayvazyan SA, Buchstaber VM, Enyukov IS, Meshalkin LD. Applied Statistics: Classification and Dimension Reduction. Moscow: Finance and Statistics, 1989; 607 p.

[2] Alekseeva VA. Using of mining techniques in problems of binary classification. Izvestiya of the Samara Scientific Center of the Russian Academy of Sciences 2014; 16(6-2): 354–356.

[3] Alekseeva VA. Construction of an aggregative binary classifier. Modern problems of design, production and operation of radio engineering systems 2015; 1-2(9): 211–214.

[4] Alekseeva VA. The use of machine learning methods for binary classification. Automation of Control Processes 2015; 3(41): 58–63.

[5] Bidyuk PI, Terent'ev AN. Construction and methods of learning Bayesian networks. Informatics and Cybernetics 2004; 2: 140–154.

[6] Vasiliev NP. Experience in calculating the parameters of logistic regression by the Newton-Raphson method for estimating winter hardiness of plants. Mathematical Biology and Bioinformatics 2011; 6(2): 190–199.

[7] Klyachkin VN, Karpunina IN, Kuvayskova YuE, Khoreva AS. The Machine learning methods application for technical diagnostics. Scientific Bulletin of the UVAU GA (I) 2016; 8: 158–161.

[8] Kuvayskova YuE, Barth AD, Fedorova KA. Application of methods of fuzzy logic and machine learning in solving the problem of technical diagnostics. Informatics and Computer Science: a collection of scientific papers of the VIII All-Russian Scientific and Technical Conference of Postgraduates, Students and Young Scientists, 2016;160–166.

[9] Little RJA, Rubyn DB. Statistical analysis of data with omissions. Moscow: Finance and Statistics, 1990; 336 p.

[10] Shtovba SD. Identification of nonlinear dependencies using fuzzy logic in the Matlab. Scientific and practical journal Exponenta Pro: mathematics in applications 2003; 2(2): 9–15.

[11] Shunina YuS, Alekseeva VA, Klyachkin VN. Forecasting the customers" creditworthiness through machine learning methods. Finance and Credit 2015; 27(651): 2–12.

[12] Shunina YuS, Alekseeva VA, Klyachkin VN. Criteria of quality of qualifiers work. Bulletin of Ulyanovsk State Technical University 2015; 2(70): 67–70.

[13] Yasnitsky LN. Introduction to Artificial Intelligence. Moscow: Publishing Center "Academy", 2005; 176 p.

[14] Yakupov AI. Application of decision trees for modeling the creditworthiness of commercial bank clients. Artificial Intelligence 2008; 4: 208–-213.