UDC 616.37-002
Doi: 10.31772/2587-6066-2019-20-2-153-159

For citation: Kononova N. V., Mangalova E. S., Stroev A. V., Cherdantsev D. V., Chubarova O. V. [Applied classification problems using ridge regression]. *Siberian Journal of Science and Technology.* 2019, Vol. 20, No. 2, P. 153–159. Doi: 10.31772/2587-6066-2019-20-2-153-159

Для цитирования: Кононова Н. В., Мангалова Е. С., Строев А. В., Черданцев Д. В., Чубарова О. В. Прикладные вопросы классификации с использованием гребневой регрессии // Сибирский журнал науки и технологий. 2019. Т. 20, № 2. С. 153–159. Doi: 10.31772/2587-6066-2019-20-2-153-159

APPLIED CLASSIFICATION PROBLEMS USING RIDGE REGRESSION

N. V. Kononova[1*], E. S. Mangalova[2], A. V. Stroev[3], D. V. Cherdantsev[3], O. V. Chubarova[4]

[1]Siberian Federal University
79, Svobodny Av., 660041, Krasnoyarsk, Russian Federation
[2]OOO “RD Science”
19, Kirova St., Krasnoyarsk, 660017, Russian Federation
[3]Krasnoyarsk State Medical University named after Prof. V. F. Voino-Yasenetsky
1, Partizana Zheleznyaka St., Krasnoyarsk, 660022, Russian Federation
[4]Reshetnev Siberian State University of Science and Technology
31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660037, Russian Federation
[*] E-mail: koplyarovanv@mail.ru

The rapid development of technical devices and technology allows monitoring the properties of different physical nature objects with very small discreteness of the data. As a result, one can accumulate large amounts of data that can be used with advantage to manage an object, a multiply connected system, and a technological enterprise. However, regardless of the field of activity, the tasks associated with small amounts of data remains. In this case the dynamics of data accumulation depends on the objective limitations of the external world and the environment. The conducted research concerns high-dimensional data with small sample sizes. In this connection, the task of selecting informative features arises, which will allow both to improve the quality of problem solving by eliminating “junk” features, and to increase the speed of decision making, since algorithms are usually dependent on the dimension of the feature space, and simplify the data collection procedure (do not collect uninformative data). As the number of features can be large, it is impossible to use a complete search of all features spaces. Instead of it, for the selection of informative features, we propose a two-step random search algorithm based on the genetic algorithm uses: at the first stage, the search with limiting the number of features in the subset to reduce the feature space by eliminating “junk” features, at the second stage - without limitation, but on a reduced set features. The original problem formulation is the task of supervised classification when the object class is determined by an expert. The object attributes values vary depending on its state, which makes it belong to one or another class, that is, statistics has an offset in class. Without breaking the generality, for carrying out simulation modeling, a two-alternative formulation of the supervised classification task was used. Data from the field of medical diagnostics of the disease severity were used to generate training samples.

Keywords: small samples, supervised classification, ridge-regression, quantile transformation, meta-classifier, significance of features, genetic algorithm.

ПРИКЛАДНЫЕ ВОПРОСЫ КЛАССИФИКАЦИИ С ИСПОЛЬЗОВАНИЕМ ГРЕБНЕВОЙ РЕГРЕССИИ

Н. В. Кононова[1*], Е. С. Мангалова[2], А. В. Строев[3], Д. В. Черданцев[3], О. В. Чубарова[4]

[1]Сибирский федеральный университет
Российская Федерация, 660041, г. Красноярск, просп. Свободный, 79
[2]ООО «АрДиСайнс»
Российская Федерация, 660017, г. Красноярск, ул. Кирова, 19
[3]Красноярский государственный медицинский университет имени профессора В. Ф. Войно-Ясенецкого
Российская Федерация, 660022, г. Красноярск, ул. Партизана Железняка, 1
[4]Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева
Российская Федерация, 660037, г. Красноярск, просп. им. газ. «Красноярский рабочий», 31
[*]E-mail: koplyarovanv@mail.ru

Бурное развитие технологий и техники обеспечивают возможность мониторинга свойств объектов различной физической природы с очень малой дискретностью. В результате накапливаются большие объемы данных, которые можно использовать с пользой для управления объектом, многосвязной системой, техноло-

гическим предприятием. Однако, вне зависимости от сферы деятельности, остаются задачи, связанные с небольшими объемами данных, динамика их накопления зависит от объективных ограничений внешнего мира и окружающей среды.

Проводимые исследования касаются данных небольших объемов выборок и размерности признаков объектов, которая может считаться высокой относительно количества изучаемых объектов. В связи с этим возникает задача отбора информативных признаков, что позволит как улучшить качество решения задачи за счет исключения «мусорных» признаков, так и повысить скорость принятие решения, поскольку алгоритмы обычно зависимы от размерности признакового пространства, и упростить процедуру сбора данных (не собирать неинформативные данные). Поскольку количество признаков может быть велико, полный перебор всех пространств признаков оказывается невозможным. Вместо этого для отбора информативных признаков предложен двуступенчатый алгоритм случайного поиска, основанный на применении генетического алгоритма: на первом этапе с ограничением количества признаков в подмножестве для сокращения признакового пространства за счет исключения «мусорных» признаков, на втором этапе – без ограничения, но по сокращенному набору признаков.

Исходная формулировка проблемы представляет собой задачу классификации объектов с учителем, когда класс объекта определен экспертом. Значения признаков объектов меняются в зависимости от его состояния, что обусловливает принадлежность тому или иному классу, то есть статистики обладают смещенностью в классе.

Без нарушения общности для проведения имитационного моделирования использовалась двухальтернативная постановка задачи классификации с учителем, для генерации обучающих выборок были использованы данные из области медицинской диагностики степени тяжести заболевания.

Ключевые слова: малые выборки, классификация с учителем, ридж-регрессия, квантильное преобразование, мета-классификатор, значимость признаков, генетический алгоритм.

Introduction. By solving the problem of classification the essential components are objects selection, feature reduction and distance criteria (norming).

For feature reduction it is necessary to pay attention to the following aspects:

– the possible accuracy of classification algorithm which can be valued with cross-validation algorithms for any feature set. If the set is insufficient for model building the accuracy of an examined classification algorithm will be limited by the lack of information;

– time to build a classifier: the size of feature space implicitly defines learning time. Amount of irrelevant features can unnecessary increase classifier building time;

– the number of objects required for learning a sufficiently accurate classifier: other conditions being equal the greater number of features is used in the model the greater number of objects must be which are necessary to achieve the required classification accuracy. With a large number of features and a small number of objects the risk of retraining the model is high;

– the cost of classifying a new object using the trained classifier: in many practical applications, for example, in medical diagnostics features are the observed symptoms as well as the results of diagnostic tests. Different diagnostic test may have various costs and associated risks. For example, an invasive exploratory operation can be much more expensive and riskier than a blood test. This presents us with the problem of choosing a subset of features when training the classifier.

The problem of choosing a feature space is related to the identification task and it is to choose a subset of features from the larger set of often mutually redundant, possibly irrelevant features with different measurement costs and / or risks. An example of such a task of considerable practical interest is the problem of forming the feature space for solving the problem of classifying the disease severity.

The literature suggests various approaches to select a subset of features. Some of them include finding the optimal subset based on a specific quality criterion [1], uses an exhaustive wide search [2], to find the minimum combination of features sufficient to build an accurate model from observations. Since the complete enumeration of all feature combinations is impossible because of the large number of features and combinations, most approaches to the choice of the feature subset imply the monotony of a certain measure – classification accuracy. If it is assumed that adding features does not impair accuracy, then the branch and bound method can be used for searching [3; 4]. However, in many practical applications the monotony assumption is not satisfied.

Some authors consider the use of heuristic search (often in combination with the branch and bound method) [5–11], as well as randomize algorithms [12; 13] and genetic algorithms [14–17], to select a subset of features and its further use with the decision tree or the method of nearest neighbours.

General formulation of the problem. Suppose there are many objects $\left\{O_i, i=\overline{1,n}\right\}$, where n is a sample size which is described by a known features set $\left\{p_i, i=\overline{1,m}\right\}$, measured in absolute ($m_1$) and rank ($m_2$) scales: $m_1+m_2=m$. For each object there is an indication of the teacher to which class it belongs: $O_i \in Z_l$, $l=\overline{1,L}$, L is a number of classes. Denote feature measurements for each object by a set of values $\left\{\left(z_i, x_i^j\right), i=\overline{1,s}, j=\overline{1,m}\right\}$, where x_i^j is a value of p_j feature of O_i object, z_i is a number of the class, n_l is a number of class objects Z_l, $\sum_{l=1}^{L} n_l = n$.

It is necessary to build a classification algorithm, develop procedures for setting the parameters of algorithms.

The above task is complicated by the fact that the number of features is comparable to the sample size. In fact, this means data that are highly sparse in multidimensional space and have no pronounced interclass boundaries. That is, it is hardly possible to build a satisfactory quality classifier.

In this case the question arises: what is the ratio of the number of features to the sample size considered as a threshold of significant sparseness and what to do when the set of features is large and the data volume is limited? The first question is the subject of further research and is closely related to the stability characteristic of the classifier. The second question is considered in the further sections of the article.

Due to the small amount of input data leading to the high sparseness in the feature space it is necessary to order the features according to the degree of their influence on the quality of classification, in other words, to reduce the number of features discarding the ones of little significance.

An insignificant feature is proposed to consider the one which, being excepted, does not worsen the classification.

Selection of significant features and classification. Feature reduction experiments were performed using a standard genetic algorithm [18; 19]. The results are based on a sliding exam for the problem of classifying an object with the following parameters of the genetic algorithm:

1. Population size: 100;
2. Number of generations: 200;
3. Selection: ranking method;
4. Crossover probability: 0.6;
5. Mutation probability: 0.001;
6. Probability of choosing an individual with the highest rank: 0.6.

Each individual in the population is a variant of the feature subset to solve the problem of classification. Based on the total number of features *m* a classification can be made.

To solve the classification problem it is proposed to use its regression formulation. This is possible when objects form different classes according to their states. This means that the class is a group of similar objects in a certain state. The state of the object is classified according to the values of a specific set of features which are the measurements of technological parameters and the results of diagnostic tests. Due to the fact that the transition of an object from one state to another under the influence of various loads, disruptions in work and environmental influences can be interpreted as a sequential transition from one class to another. The classification task is presented as a regression task where the classes are ordered according to the state of the object. Thus, objects with a light form of deviation of technological parameters and a slight wear of resource form class 1; class 2 is made up of objects with a higher (medium) level of inconsistency; objects with significant deviations (severe stage) form class 3.

As a result of building the regression dependence each new object will be assigned a value from 1 to 3 instead of a class number (1 – light, 2 – medium, 3 – severe). For example if forecasts 1.1 and 1.4 are obtained for two objects, the probability that the first object has a slight degree of deviation is higher than for the second one although both of them will be attributed to the state of slight deviation.

If there are *m* features there are 2^{m} possible subsets of features. For large values of *m* the complete search of features takes considerable time which may not be consisted with the restriction of waiting for the result of the algorithm.

Each individual is a binary vector of dimension *m*. If the bit is 1, it means that the corresponding attribute is selected to build the classifier. The value 0 indicates that the corresponding attribute is excluded from the classifier.

The average square of residual is chosen as a fitness function. Since the classification problem is solved as a regression problem, MSE additionally penalizes (for example, in comparison with MAE) large errors of classification when the forecast deviation from the real value of the class is more than 1.5, i. e. the error is more than one class (instead of the light stage the classifier predicts the severe condition and vice versa).

For numerical modeling and building the regression dependence a ridge regression was used (linear regression with a regularization parameter) [20]. The ridge regression is used if it occurs:

– data redundancy;

– correlated independent variables (multicollinearity);

– strong differences in the eigenvalues of the characteristic equation or the proximity to zero of several of them.

All of the above properties of features quite often take place in practice when the removal of technological parameters is of a distributed nature.

We use a linear model: $y = f(x, \beta)$, where f is a linear operator (linear functional dependence), β is model parameters.

We assume that the vector of coefficients of the linear regression model β is found by the least squares method:

$$\sum_{i=1}^{n}\left(f\left(x_i,\beta\right)-y_i\right)^2 \to \min_{\beta_i}. \tag{1}$$

Analytical solution of this problem: $\beta = (X^TX)^{-1}X^TY$, however, when the matrix X^TX is degenerate, the solution is not unique, but if it is poorly conditioned it is not stable. Therefore, regularization of the parameter β is introduced, for example according to the following rule:

$$Q(\beta) = Y - X\beta^2 + \lambda\beta^2 \to \min_{\beta}, \tag{2}$$

where $\lambda > 0$ regularization parameter.

The regularized least squares solution is as follows:

$$\beta = \left(X^TX + \lambda I\right)^{-1} X^TY. \tag{3}$$

The increase in the parameter λ leads to the decrease in the norm of the parameter vector and the increase in the efficiency of the feature space dimension.

The error is estimated using a sliding exam since the size of the training sample is small and the construction of the classifier takes little machine time. This paper proposes the use of two-step feature selection algorithm:

1. At the first step the primary selection of features is carried out. It means the exclusion of the most "junk" ones. We restrict the individuals so that the number of features in subsets is less than 0.2*n, where n is the number of observations in the training set, further additions of features in theory will lead to the increase in the risk of retraining. For example in the task of predicting the severity of the disease 25 features (of 94 in the original sample) are revealed, i.e. no evaluable feature subset can contain more than 25 features. If, as a result of crossing or mutation we get an individual who is not suitable for this condition then such an individual is thrown out and replaced by the following one suitable for this rule.

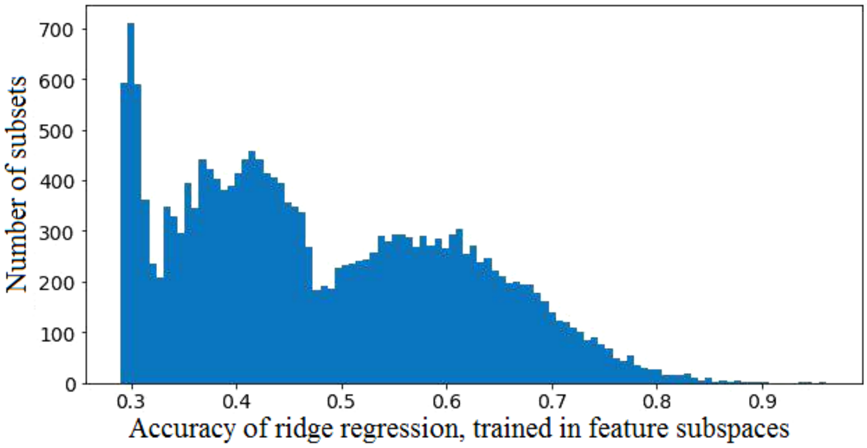

Fig.1. The distribution of the accuracy of classifiers trained on various attribute subspaces that were individuals in the course of optimization by the genetic algorithm for any generation

Рис. 1. Распределение точности классификаторов, обученных по различным признаковым подпространствам, которые являются индивидами в ходе оптимизации генетическим алгоритмом на любом поколении

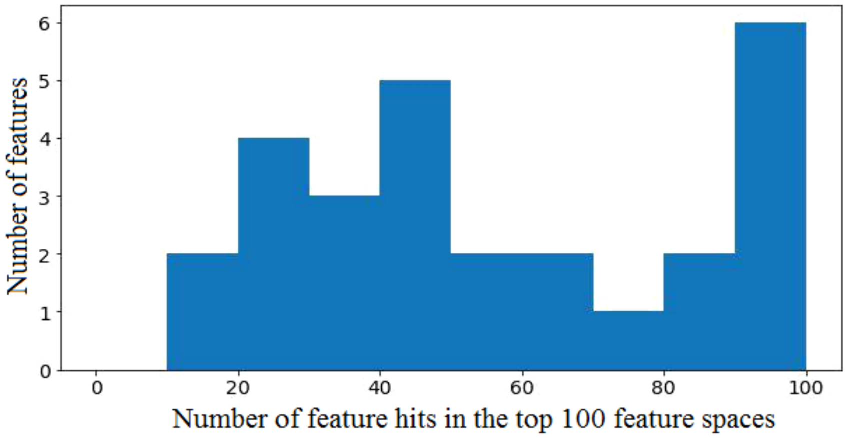

Fig. 2. Distribution of features by the number of hits in the subsets from the pool of the best subsets

Рис. 2. Распределение признаков по количеству попаданий в подмножества из пула лучших подмножеств

Using the genetic algorithm we form sets of the best solutions that could be obtained at any iteration. Fig. 1 shows the accuracy distribution of forecasts for MSE. It is to be noted that the accuracy of 0.699 is provided by a simple average for this problem. If a subset of features gives the accuracy worse than 0.699, it means that there is a retraining effect and the subset contains no important features.

From the entire set of features a pool of the best solutions (subsets) is selected from the first (left) peak of the distribution density of the accuracy of classifiers trained in various attribute subspaces. Further, according to this pool, we calculate the number of inclusions of the feature in the feature spaces. The feature distribution by the number of hits is shown in fig. 2

There is a sharp drop in the number of feature hits in the best feature subspaces. All features that fall into a substantially small number of subsets are excluded, provided that the excluded feature does not fall into the top 20 feature spaces.

Thus, 67 features are cut off at this stage. It is also to be noted that one feature gets into the best subsets a lot more times and can be initially included in the best subset.

2. At the second step we launch a new optimization process with no limitation on the number of features. Fig. 3 shows distribution of the feature number in the best in reproducible classification accuracy of feature subspaces.

Fig. 4 shows the accuracy distribution of forecasts for MSE. It should be noted that the proportion of individuals close to the best selected increased, as did the average accuracy of the classifiers and the best accuracy compared to the solutions found at the first step.

Fig. 5 shows the distribution of features by the number of individuals in the top 100 and table contains the number of specific features in the most suitable individuals top 5, top 50 and top 100.

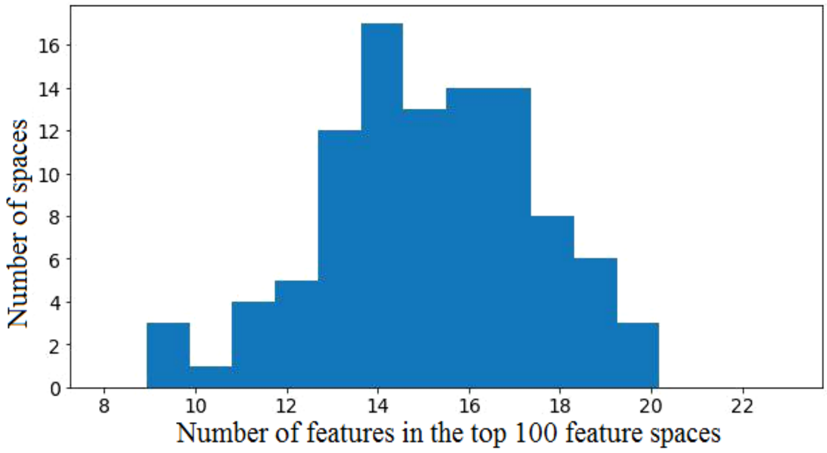

Fig. 3. Distribution of the features number in the best in reproducible classification accuracy of feature subspaces

Рис. 3. Распределение количества признаков в лучших по воспроизводимой точности классификации признаковых подпространствах

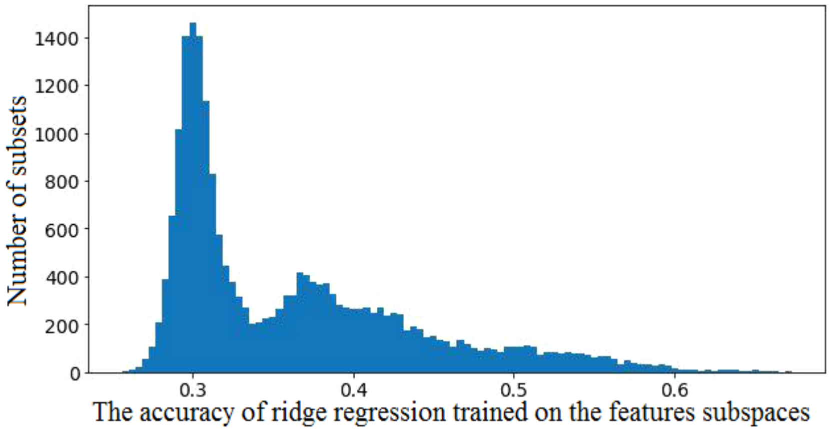

Fig. 4. Accuracy distribution of MSE classifiers after optimization without limitation on the number of features

Рис. 4. Распределение точности классификаторов по MSE после процесса оптимизации без ограничения на количество признаков

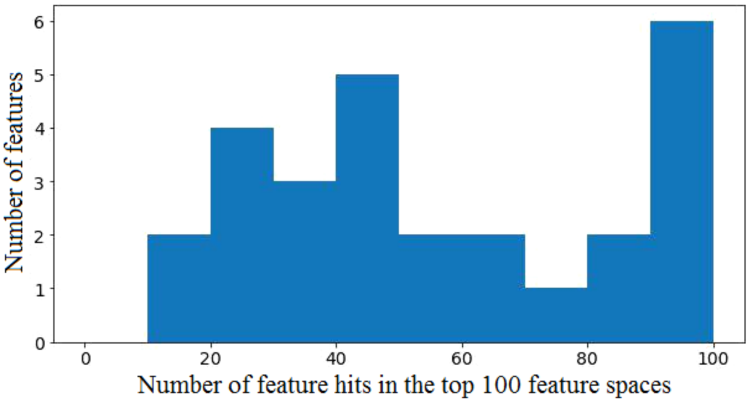

Fig. 5. Distribution of features by the number of hits in the top 100 feature spaces

Рис. 5. Распределение признаков по количеству попаданий в топ-100 признаковых пространств

The number of hits in the top best individuals

Sequence number of the feature	Number of hits in the top 5 best individuals	Number of hits in the top 50 best individuals	Number of hits in the top 100 best individuals
1	5	44	78
2	2	16	35
3	0	7	28
4	0	21	48
5	0	11	30
6	0	12	27
7	5	49	90
8	5	50	100
9	0	18	43
10	1	19	43
11	5	45	91
12	0	13	37
13	5	35	64
14	0	20	47
15	0	5	16
16	5	50	90
17	0	18	41
18	2	30	61
19	2	37	82
20	0	9	18
21	5	45	93
22	1	15	25
23	5	50	100
24	5	45	86
25	3	27	54
26	1	12	26
27	5	30	57

Conclusion. The paper considers the procedure of feature selection for small volumes of the original training set and a significant number of features describing the state of objects. To solve the problem the genetic algorithm for the feature subspaces formation was used. The approach to the classifier construction differs from the classical one and is a construction of the regression dependence of the class value output on the input feature values. Such implementation allows estimating the probability of belonging to a particular class through the regression value. Application of this approach to the classification problem is possible in cases when the class of an object depends on its state and the classes can be ordered (ranked). Data on the state of human health were taken as a basis of simulation samples generation for numerical experiments but without violation of generality the approach can be extended to other industries and fields of activity including the diagnosis of changes in the state of equipment during its operation.

Acknowledgment. This work was carried out with the support of the Krasnoyarsk Territory Foundation of Science in the framework of the project: "Development and implementation of a decision support system for the acute pancreatitis diagnosis and treatment in the Krasnoyarsk Territory".

Благодарности. Исследование выполнено при поддержке Красноярского краевого фонда науки в рамках реализации проекта: «Разработка и внедрение системы поддержки в принятии решений при диагностике и лечении острого панкреатита на территории Красноярского края».

References

1. Vafaie H., De Jong K. Robust Feature Selection Algorithms. *Proceedings of the IEEE International Conference on Tools with Artificial Intelligence*. 1993, P. 356–363.

2. Cormen T. H., Leiserson C. E., Rivest R. L., Stein C. Introduction to Algorithms. 3rd edition. The MIT Press. 2009, 1320 p.

3. Narendra P., Fukunaga K. A Branch and Bound Algorithm for Feature Subset Selection. *IEEE Transactions on Computers*. 1977, Vol. 26, P. 917–922.

4. Foroutan I., Sklansky J. Feature Selection for Automatic Classification of non- Gaussian Data. *IEEE Transactions on Systems, Man and Cybernetics*. 1987, Vol. 17, P. 187–198.

5. Kira K., Rendell L. A Practical Approach to Feature Selection. *Proceedings of the Ninth International Conference on Machine Learning (Morgan Kaufmann)*. 1992, P. 249–256.

6. Modrzejewski M. Feature Selection Using Rough Sets Theory. *Proceedings of the European Conference on Machine Learning (Springer)*. 1993, P. 213–226.

7. Liu H., Setiono R. Chi2: Feature Selection and Discretization of Numeric Attributes. *Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence*. 1995.

8. John G., Kohavi R., Peger K. Irrelevant Features and the Subset Selection Problem. *Machine Learning: Proceedings of the Eleventh International Conference (Morgan Kaufmann)*. 1994, P. 121–129.

9. Kohavi R., Frasca B. Useful Feature Subsets and Rough Set Reducts. Third International Workshop on Rough Sets and Soft Computing. 1994.

10. Kohavi R. Feature Subset Selection as Search with Probabilistic Estimates. AAAI Fall Symposium on Relevance. 1994 .

11. Koller D., Sahami M. Toward Optimal Feature Selection. *Machine Learning: Proceedings of the Thirteenth International Conference (Morgan Kaufmann)*. 1996.

12. Liu H., Setiono R. A Probabilistic Approach to Feature Selection – A Filter Solution. *Proceedings of the Thirteenth International Conference on Machine Learning (Morgan Kaufmann).* 1996.

13. Liu H., Setiono R. Feature Selection and Classification – A Probabilistic Wrapper Approach. *Proceedings of the Ninth International Conference on Industrial and Engineering Applications of AI and ES*. 1996.

14. Siedlecki W., Sklansky J. A Note on Genetic Algorithms for Large-scale Feature Selection. *IEEE Transactions on Computers*. 1989, Vol. 10, P. 335–347.

15. Punch W., Goodman E., Pei M., Chia-Shun L., Hovland P., Enbody R. Further Research on Feature Selection and Classification Using Genetic Algorithms. *Proceedings of the International Conference on Genetic Algorithms (Springer)*. 1993, P. 557–564.

16. Brill F., Brown D., Martin W. Fast Genetic Selection of Features for Neural Network Classiers. *IEEE Transactions on Neural Networks*. 1992, Vol. 3(2), P. 324–328.

17. Richeldi M., & Lanzi P. Performing Effective Feature Selection by Investigating the Deep Structure of the Data. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (AAAI Press).* 1996, P. 379–383.

18. Goldberg D. Genetic Algorithms in Search, Optimization, and Machine Learning. New York, Addison-Wesley, 1989.

19. Mitchell M. An Introduction to Genetic algorithms. Cambridge, MA: MIT Press. 1996.

20. Dreiper N., Smit G. Applied regression analysis. 1986, 351 p.

Kononova Nadezhda Vladimirovna – Cand. Sc., associate professor; Informational systems department, Siberian Federal University. E-mail: koplyarovanv@mail.ru.

Mangalova Ekaterina Sergeevna – software developer; “RD Scienc” (Research. Development. Science.). E-mail: e.s.mangalova@hotmail.com.

Stroev Anton Vladimirovich – Postgraduate Student; Krasnoyarsk State Medical University named after Prof. V. F. Voino-Yasenetsky. E-mail: antoxa134@mail.ru.

Cherdantsev Dmitry Vladimirovich – Dr. Sc., Professor; Krasnoyarsk State Medical University named after Prof. V. F. Voino-Yasenetsky. E-mail: gs7@mail.ru**.**

Chubarova Olesya Victorovna – Cand. Sc., associate professor; System analysis and operations research department, Reshetnev Siberian State University of Science and Technology. E-mail: kuznetcova_o@mail.ru.

Кононова Надежда Владимировна – кандидат технических наук, доцент; кафедра информационных систем, Сибирский федеральный университет. E-mail: koplyarovanv@mail.ru.

Мангалова Екатерина Сергеевна – программист-разработчик; ООО «АрДиСайнс» (Research. Development. Science.). E-mail: e.s.mangalova@hotmail.com.

Строев Антон Владимирович – аспирант; Красноярский государственный медицинский университет имени профессора В. Ф. Войно-Ясенецкого. E-mail: antoxa134@mail.ru.

Черданцев Дмитрий Владимирович – доктор медицинских наук, профессор; заведующий кафедрой хирургических болезней, Красноярский государственный медицинский университет имени профессора В. Ф. Войно-Ясенецкого. E-mail: gs7@mail.ru.

Чубарова Олеся Викторовна – кандидат технических наук, доцент, доцент; кафедра системного анализа и исследования операций, Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева. E-mail: kuznetcova_o@mail.ru.