Comparative analysis of classification algorithms

R. Muhamedyev, K. Yakunin, S. Iskakov, S. Sainova, A.Abdilmanova
Institute of Information and Computational Technologies, Ministry of Education and Science of the Republic of Kazakhstan, 125 Pushkina st., Almaty 050010, Kazakhstan 31
saida@gmail.com

Y. Kuchin
"GeoTechnoServiss"
Almaty, Kazakhstan
ykuchin@mail.ru

***Abstract*—machine learning algorithms are widely used in classification problems. Certainly, recognition quality of algorithms is important indicator, but the ability of the algorithm to learn is more significant. In this work the learning curves experiment was performed in order to identify which of the three learning rates occur when training the machine learning algorithms: overfitting, perfect case and underfitting. Neural Network, k-Nearest Neighbors and Naïve Bayes were chosen for this experiment, since their results in previous experiments were reasonable for the log data. Also this paper contains a comparative analysis of those recognition algorithms applied to the log data of Inkai uranium deposits in Kazakhstan.**

Index Terms —machine learning; artificial neural network; k-NN; Naïve Bayes; learning curves; quality indicators; accuracy; precision; recall.

I. INTRODUCTION

To organize a process of extraction of uranium by in-situ leaching, it is important to know the composition and thickness of lithologic layers. Evaluation of the lithologic structure is performed experts by using data of geophysical investigation (logging data). However, there is a certain ability to automate the process and reduce errors. One way is to use machine learning methods.

Machine Learning (ML) is an extensive subsection of artificial intelligence, studying the methods of constructing algorithms with ability to learn. There is a wide variety of tasks and successful applications of machine learning. The methods of machine learning refer to a broad class of algorithms: from decision trees, metric algorithms, such as k-NN, Support Vector Machines (SVM) to artificial neural networks (ANN). ANN is widely used in problems of classification and pattern recognition [1]. There are some works [1,2,3] describe the results of applying ANN for the interpretation of geophysical investigation data for borehole during uranium production.

Since 1970s ANN have been used in the problems of petrography, as a means of analyzing logging data, in lithology, in estimation of mineral sources of raw materials, etc. Researchers [1, 2, 3] propose and describe some of the results of applying neural networks for the interpretation of geophysical investigation data for borehole during uranium production.

This paper contains a comparative analysis of different recognition algorithms applied to the log data of Inkai uranium deposits in Kazakhstan. Also there are results and analysis of conducted experiment to detect the ability of the algorithm to learn. The experiment is conducted using the data mining environment RapidMiner v5.3. It was carried out and learning curves were plotted for comparative analysis of ML methods.

In the work we attempted to explore which of three learning rates implemented algorithm belongs to: overfitting, perfect case or underfitting. Our suggested decision is to construct the learning curves of each machine learning (ML) algorithm, using such parameters as accuracy, weighted mean precision and weighted mean recall. With help of learning curves experiment, it is possible to define the state of the ML algorithms. Such operation is needed to manipulate algorithm's properties to achieve better results in recognition of lithological types.

The purpose of this work is to find out whether that classification algorithms trainable and what to fix if they are not.

In ANN the number of hidden layers and the number of training examples significantly affect its accuracy, in linear regression – its order, in k-NN – the radius of circumference of the closest neighbors and so on. Actually, it is important to take into account the ability of the algorithm to learn, overfit or underfit. The right balance between underfit and overfit means finding such an algorithm and its parameters that would be able to show acceptable results on the test set (or cross validation). Underfitting algorithm will show the same poor results both in the training and in test sets, while overfitting algorithm will demonstrate good results in the training set and poor results in the test set.

II. DESCRIPTION OF INVESTIGATED ALGORITHMS

During this experiment the following classification algorithms have been investigated:

1) Feedforward artificial Neural Network (ANN);

2) k-Nearest-Neighbors (k-NN);

3) Naïve Bayes.

K-Nearest-Neighbors (k-NN) algorithm

The algorithm is based on counting the number of objects in sphere (hypersphere) of each class centered in a recognizable object. The object belongs to the class of objects that represent majority in this sphere. This method assumes that the unit weight is chosen for all objects. If the weights are not the same, it can be summed up their weights instead of counting the number of objects. Thus, if in the sphere around the recognizable object 10 standard objects of class A and 15

erroneous/borderline objects of class B, then the point will be assigned to Class A [3].

Also, the weights of the objects in the sphere can be described as an inverse proportionally dependence from the distance to the recognizable object. Thus, the closer the object is, the more important it is for the recognizable object.

As a result, the metric classifier could be written in this form:

$$a(u;\ X^l) = \arg\max_{y \in Y} \sum_{i=1}^{l} \left[y_u^{(i)} = y\right] w(i,u) \qquad (1)$$

where, w(i,u) – weight of i^{th} neighbor of recognizable object u,
a(u;X^l) – class of object u, recognized by sample X^l.

The radius of the hypersphere can be both fixed and dynamic. And in the case of the dynamic radius, the radius of each point is selected so that the number of objects in each sphere is fixed. Thus, when recognizing in areas with varying sample density, the number of "neighboring" objects (which we need for recognition) would be equal, i.e. in areas with low density there will be no situations where the data is insufficient to recognize [3].

Neural Network architecture

Neural network architecture is a number of neurons in every layer and the number of those layers. Moreover, the optimal number of neurons in the hidden layers depends on the number of input and output neurons. The recognition results may impair because of too many or too small number of weights. Unfortunately, there is no a specific formula, theory or method to determine the optimal number of neurons, so the only accurate method is to reconsider and testing the various options [3].

However, it is impossible to get an exact solution, due to the combinatorial explosion in finding solutions. Nevertheless, for search the optimal architecture was reconsidered various options and found the suitable parameters with the help of data mining environment RapidMiner v5.3.

In neural network architecture different parameters can be changed such as learning rate, training cycles, number of layers and number of hidden layers in each layer. We will achieve different results by applying different values for these parameters.

Naïve Bayes

A Naive Bayes classifier is a simple probabilistic classifier with strong (naive) independence assumptions which based on applying Bayes' theorem. In simple terms, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature [4].

The advantage of the Naive Bayes classifier is that it only requires a small amount of training data to estimate the means and variances of the variables necessary for classification. Because independent variables are assumed, only the variances of the variables for each label need to be determined and not the entire covariance matrix [4].

III. Quality indicators

In this experiment three main error indicators to estimate the quality of the classification: accuracy, weighted mean recall, weighted mean precision have been used. These values more clearly show the recognition quality since they are independent (explicitly) from the number of objects in the test set: precision and recall are often used in data mining problems.

Accuracy is the relative number of correctly classified examples or in other words the percentage of correct predictions:

$$accuracy = \frac{TP+FP}{N} \qquad (2)$$

They are calculated on the basis of characteristics of TP and FP:

$$Precision = \frac{TP}{TP+FP} * 100\% \qquad (3)$$

Precision characterizes how many of the received from the classifier positive answers are correct. The greater the precision is, the fewer the number of false hits. It measures a share of true detections among all detected objects.

Weighted mean precision is the weighted mean of all per class precision measurements. It is calculated through class precisions for individual classes.

The metric of precision does not give us a full notion about whether all results of classifier are correct. For this purpose there is one more metric called recall:

$$Recall = \frac{TP}{TP+FN} * 100\% \qquad (4)$$

The metric of recall characterizes the ability of the classifier to "guess" the greatest possible number of positive responses from the expected number. It measures a share of true recognition from all the objects we are interested in. Note that the false-positive responses have no effect on this metric.

Weighted mean recall is the weighted mean of all per class recall measurements. It is calculated through class recalls for individual classes.

IV. Analysis of learning curves

Data from 30 boreholes of Inkai uranium deposits have been selected: 15 boreholes for training and 15 boreholes for testing. First, one borehole for training it was applied and all the main quality indicators' values were recorded. Then on trained module one borehole was applied for testing, which is different from previous borehole. Its values were recorded too. Next time one more borehole was applied and so on until fifteenth borehole.

The graphs of dependency between the indicators were plotted for ANN, kNN and Naïve Bayes algorithms:

1) accuracyTest and accuracyTrain on training and test sets (see Fig. 1, 2 and 3);

2) accuracyTest, weighted mean recallTest and weighted mean precisionTest on test set (see Fig. 4, 5 and 6);

3) accuracyTrain, weighted mean recallTrain and weighted mean precisionTrain on training set (see Fig. 7, 8 and 9).

As we can see from Fig. 1, 2 and 3 below, all algorithms behaved as expected according to the theory, i.e. "in normal situation when the number of training samples increases, the

error in training set slightly increases (accuracyTrain), but the error in test set reduces (accuracyTest)". A small difference in the end of graphs is noticeable. In the graph of kNN algorithm greater distance between lines and the tendency of divergence of the lines can be noticed. This indicates that the training of kNN algorithm is deteriorated. In the graph of Naïve Bayes algorithm is noticeable changes till eighth borehole and then stabilizes, but its values around 57%, while other algorithms are around 65%. First half of graph indicates that Naïve Bayes tried to be trained and in the second half of graph has been trained as good as it can, but it is less than other algorithm's values.

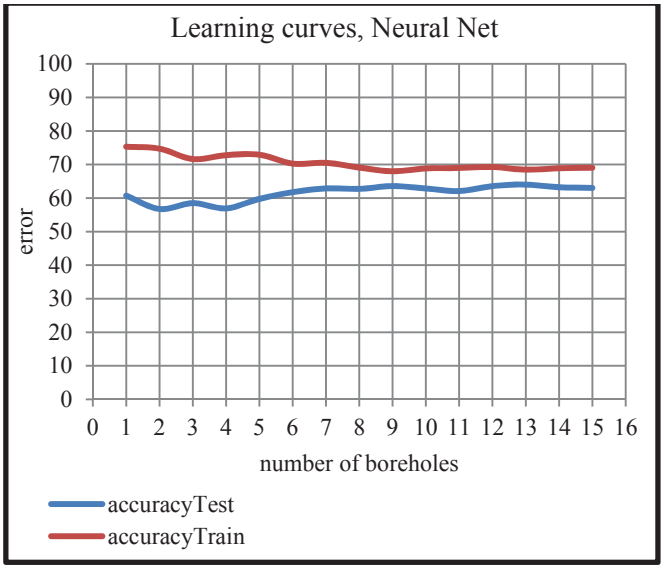

Fig. 1. Dependency graph between accuracyTest and accuracyTrain indicators applying to NN algorithm

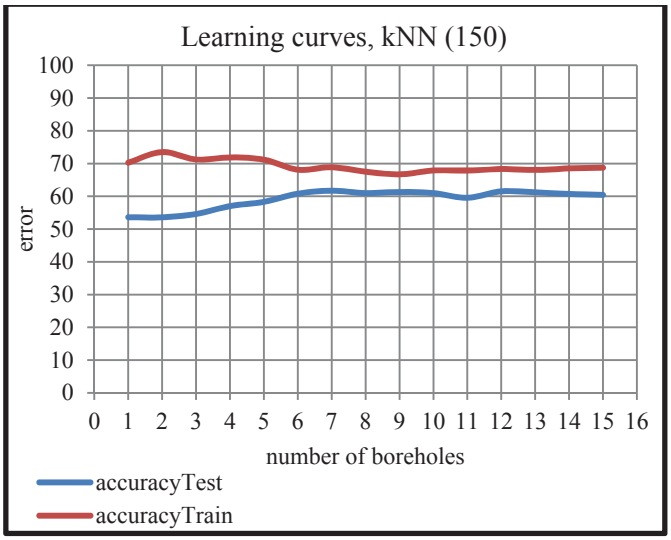

Fig. 2. Dependency graph between accuracyTest and accuracyTrain indicators applying to kNN algorithm

In Fig. 4 there is a sharp decline of two indicators: wm precision and wm recall, which means the model started training and it classifies not good enough yet. But two indicators are stabilized on applying the fifth borehole, while the accuracy1 increased a little bit. On applying the twelfth borehole there was a sharp rise and then it was stable. It means the model has improved and recognize stably well.

In Fig. 5 there is a sharp decline of two indicators: wm precision and wm recall, which means the model started training and it classifies not well enough yet. The wm precision started to grow on applying the sixth borehole and decreased then, while wm recall is stable. Both indicators are stable on applying twelfth borehole. It means the model was trained up to its maximum and further improvements are not expected.

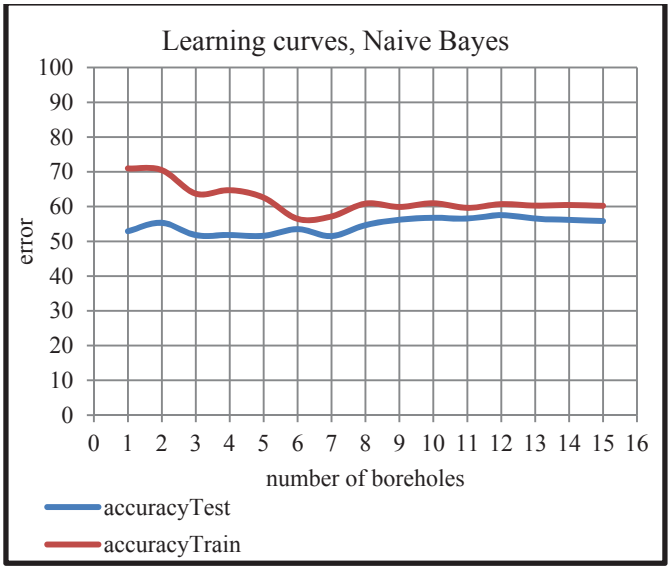

Fig. 3. Dependency graph between accuracyTest and accuracyTrain indicators applying to Naïve Bayes algorithm

In Fig. 6 there is a sharp decline of two indicators in the beginning of graph: wm precision and wm recall, which means the model started training and it classifies not well enough yet. The tendency of sharp changes in growing and decreasing of wm precision and wm recall indicators are observed till applying fifth borehole, then they stabilized. It means the model was trained up to its maximum and further improvements are not expected.

ANN algorithm indicators are higher to 5-15% by the best values of wm precision and wm recall.

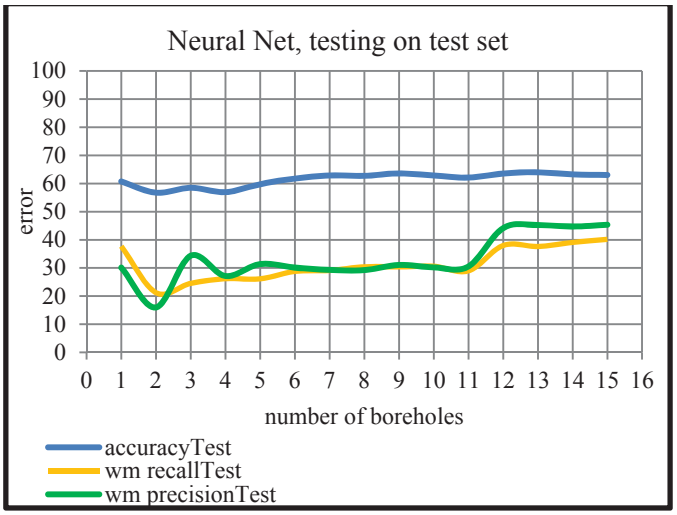

Fig. 4. Dependency graph between accuracyTest, weighted mean recallTest and weighted mean precisionTest on test set applying to NN algorithm

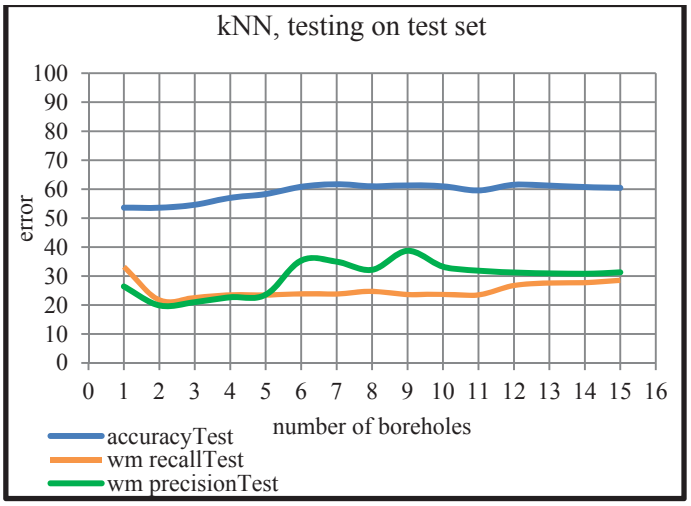

Fig. 5. Dependency graph between accuracyTest, weighted mean recallTest and weighted mean precisionTest on test set applying to kNN algorithm

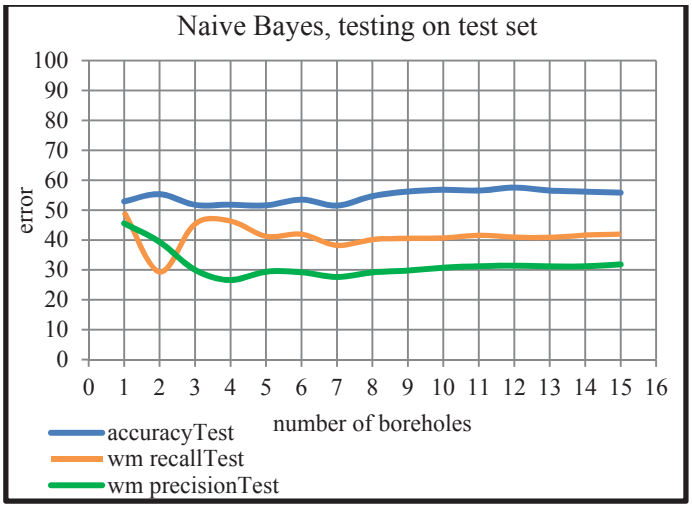

Fig. 6. Dependency graph between accuracyTest, weighted mean recallTest and weighted mean precisionTest on test set applying to Naïve Bayes algorithm

We can see, relying on the wm recall indicator for comparison of the Fig. 7, 8 and 9, that Naïve Bayes algorithm shows 5-20% better result than two other algorithms. KNN keeps up with 30%, ANN is on 40-45%, while Naïve Bayes is 50%. Also it is observed only on Naïve Bayes' graphs that wm recall indicator located higher than wm precision in both cases of test and training sets, which means this algorithm's share of true recognition is better than others.

In Fig. 7 the changes of quality indicators on applying to 5 and 12 boreholes are observed. The model is improved on fifth borehole; we can see it from increasing of the wm precision. Then the model is stable, i.e. it stably recognizes. It increases again on twelfth borehole and becomes stable then, which means that the model has improved and stabilized.

In Fig. 8 is observed rise of precision indicator between 5th and 12th boreholes. The model was tried to improve, but after 12th borehole decreased and stabilized.

In Fig. 9 the changes of quality indicators in the beginning of training till 5th borehole are observed. Between 5 and 12 boreholes there a little changes. The model is improved as well as it could till 5th borehole; we can see it from increasing of the wm precision and wm recall in the start of curves. After 12th borehole the model is stable, i.e. it stably recognizes. Now it is stable and could not be improved more.

It is obvious that comparing Fig. 7, 8 and 9 on the whole the Neural Network shows better results by accuracy, precision and recall indicators. As well as in the test set, Neural Network shows that it is more trainable than other algorithms and it could be improved, while other two algorithms keep stability.

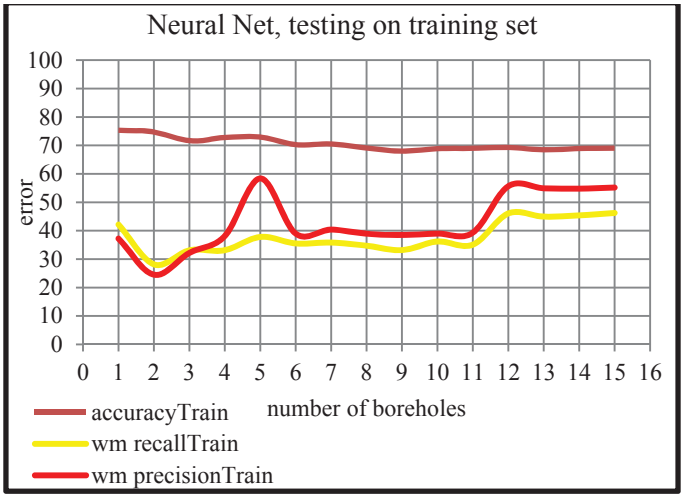

Fig. 7. Dependency graph between accuracyTrain, weighted mean recallTrain and weighted mean precisionTrain on training set applying to ANN algorithm

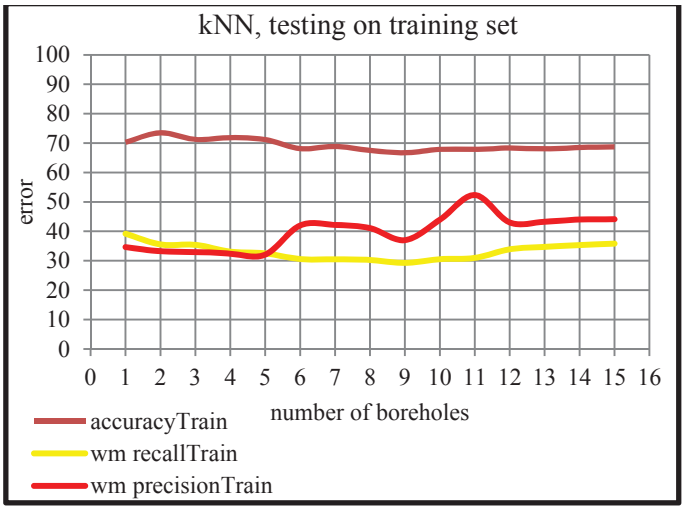

Fig. 8. Dependency graph between accuracyTrain, weighted mean recallTrain and weighted mean precisionTrain on training set applying to kNN algorithm

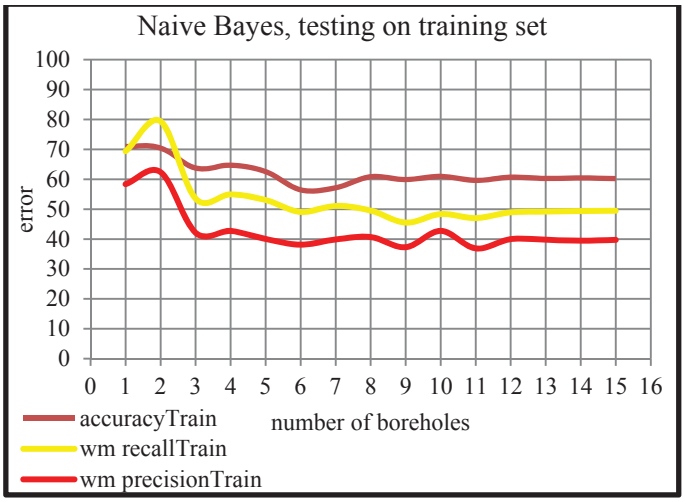

Fig. 9. Dependency graph between accuracyTrain, weighted mean recallTrain and weighted mean precisionTrain on training set applying to Naïve Bayes algorithm

V. Conclusion

The graphs of learning curves experiment clearly show learning rate of ML algorithms. All algorithms are trainable enough. They are not overfitting or underfitting. And additional graphs with wm recall and wm precision shows more details on every stage of training the algorithms, when they have been improved or when stabilized.

These graphs of experiment prove that despite the fact that two (kNN and ANN) of three algorithms are properly trainable, the Neural Network recognizes better than other algorithms. Despite that Naïve Bayes algorithm's values of wm recall and wm precision are good enough, its accuracy is deteriorated. Almost in all cases ANN algorithm values is higher to 5-15% from others.

According to the results of experiment the ANN algorithm have shown the best results, so this algorithm with set parameters will be used in further investigations.

Acknowledgment

This work has been performed with the support grant No. 2318/GF3 of the Ministry of Education and Science of the Republic of Kazakhstan.

References

[1] E. Amirgaliev, S. Iskakov, Y. Kuchin, R. Muhamediyev, I. Ualiyeva, E. Muhamedyeva, "Recognition of rocks at uranium deposits using machine learning methods," Joint issue: Journal of the East Kazakhstan State University/Computational Technologies. Institute of Computational Technologies, Siberian Branch of the Russian Academy of Sciences. Volume "Information and communication technologies".- Ust-Kamenogorsk, EKSTU named after D. Serikbayev, ICT, 2013. ISSN1561-4212, 1560-7534. C.232-240. (in russian)

[2] R. Muhamediyev, E. Amirgaliev, S. Iskakov, Y. Kuchin, E. Muhamedyeva, "Integration of Results of Recognition Algorithms at the Uranium Deposits," JACIII, 2014, Vol.18 No.3. pp. 347-352

[3] E. Amirgaliev, Z. Isabaev, S. Iskakov, Y. Kuchin, R. Muhamediyev, E. Muhamedyeva, K. Yakunin, "Recognition of rocks at uranium deposits by using a few methods of machine learning," Soft Computing in Machine Learning. Advances in Intelligent Systems and Computing, Volume 273, 2014, pp. 33-40. ISBN: 978-3-319-05532-9 (Print) 978-3-319-05533-6 (Online) (indexed by ISI Proceedings, EI-Compendex, DBLP, SCOPUS, Google Scholar and Springerlink) http://link.springer.com/chapter/10.1007/978-3-319-05533-6_4

[4] RapidMiner documentation [electronic resource]: http://docs.rapidminer.com

[5] R. Muhamedyev, S. Iskakov, P.Gricenko, K. Yakunin Y. Kuchin, "Integration of results from Recognition Algorithms and its realization at the uranium production process", Proceedings of 8th IEEE International Conference on Application of Information and Communication Technologies - AICT2014, Kazakhstan, Astana, 15-17 October 2014, p.188-191, ISBN 987-1-4799-4120-92, IEEE Catalog Number CFP1456H-PRT. (Engineering Index (EI) and EI Compendex and IEEE Xplore TM IEEE Catalog Number CFP1456H-ART, ISBN 978-1-4799-4119-3)

[6] E. Amirgaliev, S. Iskakov, Y. Kuchin, R. Muhamediyev, E. Muhamedyeva (2013) "Estimation of quality of lithological members at the uranium deposits", Proceedings of the 3rd International science-practical conference «ICT: Science,

education, innovations» Almaty, 20 May 2013. pp. 485-493. ISBN 9965-476-59-4