Machine learning in the analysis of medical data

O D Kazakov¹, A V Averchenkov², N Yu Kulagina¹

¹Bryansk State University of Engineering and Technology, 3, Stanke Dimitrov Av., Bryansk, 241037, Russia
²Bryansk State Technical University”, 7, 50 years of October Av., Bryansk, 241035, Russia

Abstract. The article is devoted to the using of machine learning algorithms for the diagnosis of human diseases on the basis of General blood analysis. The value of this research is in the strategy of the development of the information society in the Russian Federation during 2017 - 2030. The author has developed a model that will allow us to diagnose certain chronic obstructive pulmonary diseases or allergic rhinitis with a sufficient degree of probability. This model will also allow to make a conclusion about the absence of diseases in humans. To implement the methods of machine learning to solve problems in the context of the research topic, the authors obtained impersonal data of the results of a General blood test of more than a thousand people in the city of Bryansk. The results of the study can be the basis for the development of a module which help us to support medical decision-making medical information system.

1. Introduction

The research is carried out in order to develop the Strategy of the information society development in the Russian Federation during 2017-2030 in accordance with the decree of the President of the Russian Federation dated 09.05.2017 № 203

In modern conditions of the development of the social and economic systems in the field of health care there is a process of serious scientific, technical and technological changes. Modern medicine has accumulated a huge amount of information about patients and the development of diseases. There are many methods of diagnosis and research — from blood tests to genetic tests. Thus, the health system has a lot of information about its patients: text data, tabular data, time series, display and video (three-dimensional image, genomic images). There is no doubt that they need to be analyzed as for the predicting of the diagnosis, before the diagnosis is made, and to select the target therapy, depending on the patient's medical history. The effectiveness of such analysis depends on the support of information technology. Moreover, machine learning could be used to support medical decision-making by medical professionals who has no sufficient experience in the professional field. However, it should be noted that there are a number of practical problems associated with the implementation of modern information technologies, including artificial intelligence technologies

The purpose of our study is to create and to learn the models of diagnosis of human diseases based on the data of the General blood test.

To achieve this goal, we solved the following tasks:

• We obtain depersonalized data from the results of a General blood test.
• We analyzed all data.
• We built and learnt a model for the diagnosis of chronic obstructive pulmonary disease and allergic rhinitis based on classification methods.
• We made cross-validation and optimized the model.
• We developed a model for supporting medical decision-making based on machine learning algorithms.

The model of diagnosis of chronic obstructive pulmonary diseases and allergic rhinitis is based on classification methods.

2. Data collection and preliminary analysis
For implementing the methods of machine learning and to solve problems in the context of the research topic, we took impersonal data of the results of a General blood test of more than a thousand people in the city of Bryansk. This data you can see in tabular form (Table 1.)

| №  | Diagnosis                                      | Hemo   | Leukocytes | Lymphocytes | Monoocytes | Erythrocytes | Eosiphyles |
|----|-----------------------------------------------|--------|------------|-------------|------------|--------------|------------|
| 0  | Chronic obstructive pulmonary disease         | 120    | 5.2        | 69          | 7          | 1.90         | 1          |
| 1  | Allergic rhinitis                             | 118    | 6.8        | 44          | 10         | 1.89         | 2          |
| 2  | Chronic obstructive pulmonary disease         | 112    | 8.8        | 46          | 7          | 1.98         | 2          |
| 3  | Chronic obstructive pulmonary disease         | 117    | 8.7        | 62          | 13         | 1.80         | 4          |
| 4  | Chronic obstructive pulmonary disease         | 119    | 6.8        | 75          | 9          | 1.80         | 1          |
| 5  | Chronic obstructive pulmonary disease         | 120    | 10.7       | 13          | 7          | 1.90         | 7          |
| 6  | Chronic obstructive pulmonary disease         | 124    | 7.5        | 18          | 5          | 1.80         | 11         |
| 7  | Chronic obstructive pulmonary disease         | 147    | 6.1        | 55          | 5          | 4.50         | 1          |
|    |                                              |        |            |             |            |              |            |
| 1099 | Chronic obstructive pulmonary disease        | 123    | 6.9        | 52          | 4          | 4.15         | 8          |
| 1100 | Chronic obstructive pulmonary disease        | 104    | 6.8        | 43          | 7          | 1.60         | 1          |
| 1101 | Chronic obstructive pulmonary disease        | 127    | 5.0        | 55          | 6          | 4.05         | 4          |
| 1102 | Chronic obstructive pulmonary disease        | 136    | 6.8        | 68          | 1          | 4.26         | 4          |

The table contains data that reflect the results of the General blood test according to the diagnosis. You can see the diagnoses we took into account in the General sample: chronic obstructive pulmonary diseases, allergic rhinitis and we took the absence of diseases. For the purpose of preliminary analysis we have analyzed the data distributions (Figure 1)
As a result of this visual analysis, it can be concluded that the selected indicators have a normal distribution and can participate in the process of forming the machine learning algorithm. For the final determination of the composition of the indicators that we can take into account, we will calculate the corresponding correlation coefficients (Table 2).

Table 2. Pearson correlation coefficient

Figure 1. Data analysis
Because of the analysis, we can say that there is no any ties between the values of the General blood test. Therefore, to develop a machine learning algorithm for the diagnosis of human diseases, we take the following indicators:

- Hemoglobin.
- Leukocytes.
- Lymphocytes.
- Monocytes.
- Erythrocytes.
- Eosiphyles.

Let us look onto the machine learning process

### 3. Machine learning models

The main method of research is a computational experiment. The cycle of construction of our model is based on classification methods. Therefore, during our computational experiment, we will use the classification learning algorithm - Stochastic Gradient Descent.

Here, the gradient approach will be considered as a way to select the vector of synaptic weights \( \omega \) in the linear classifier.

So \( y^*: X \to Y \) – target dependence is known only on the objects of the training sample: \( X^i = (x_i, y_i)^i_{i=1}, y_i = y^*(x_i) \)

We will find the algorithm \( \alpha(x, \omega) \), approximating dependence \( y^* \). In the case of a linear classifier, the desired algorithm has the form:

\[
\alpha(x, \omega) = \varphi\left( \sum_{j=1}^{n} \omega_j x^j - \omega_b \right),
\]

where \( \varphi(z) \) plays the role of activation function (in the simplest case, we can use \( \varphi(z) = \text{sign}(z) \)).

According to the principle of minimization of empirical risk it is enough to solve the optimization problem

\[
Q(\omega) = \sum_{i=1}^{n} L(\alpha(x, \omega), y_i) \to \min,
\]

where \( L(\alpha, y) \) – loss function.

For the minimization we apply the method of gradient descent. This is a step-by-step algorithm, and at each iteration of this algorithm the vector \( \omega \) changes in the direction of the greatest decrease of the functional \( Q \) (in the direction of the anti-gradient):

\[
\omega := \omega - \eta \nabla Q(\omega),
\]
where $\eta$ – a positive parameter, called learning rate.

Algorithm Stochastic Gradient for the gradient descent can be represented as follows:

**Initial data:**

1. $X^i$ – training sample
2. $\eta$ – the pace of learning
3. $\lambda$ – the smoothing parameter of the functional $Q$

**The final data:**
1. The vector of weights $\omega$

**The basis:**
1. To initialize weights $\omega_j$, $j = 0,...,n$;
2. To initialize the functionality: $Q(\omega) = \sum_{i=1}^l L(\omega(x,\omega),y_i)$
3. To repeat:
   3.1 Chose the object $x_i$ from $X^i$ (for example, at random);
   3.2 Calculate the output value of the algorithm $\alpha(x,\omega)$ and a mistake: $\xi_i = L(\alpha(x,\omega),y_i)$;
   3.3 Take a step of gradient descent: $\omega = \omega - \eta L^i(\alpha(x,\omega),y_i)\alpha'(\omega,x_i)x_i$
   3.4 Evaluate the function: $Q = (1-\lambda)Q + \lambda \xi_i$
4. Until the value $Q$ will stabilize and values $\omega$ won't stop changing.

As the main machine learning environment of the medical diagnostics model, we will focus on the distribution of programming languages Python – Anaconda. For the realization of the algorithm we will use not the entire sample, but a random data. This will increase the speed of learning. (Figure 2)

```python
train_data, test_data, train_labels, test_labels =
cross_validation.train_test_split(iris_frame[['Hemoglobin', 'Leukocytes', 'Lymphocytes', 'Monocytes', 'Erythrocytes', 'Eosinophils']],
                           iris_frame[['target']],
                           test_size = 0.3, random_state = 0)

# print train_data[:20]
# print test_data[:20]
# print train_labels[:20]
# print test_labels[:20]

# Создаем модель Stochastic Gradient Descent:
model = linear_model.SGDClassifier(alpha=0.001, n_iter=100, random_state = 0)
# Обучаем модель на
model.fit(train_data, train_labels)
model_predictions = model.predict(test_data)

print ("accuracy: ", metrics.accuracy_score(test_labels, model_predictions))
print (metrics.classification_report(test_labels, model_predictions))
```

**Figure 2.** Realization of the algorithm Stochastic Gradient Descent

For the evaluating the algorithm, we calculate the basic quality metrics of the constructed model (Figure 3)
Figure 3. Values of quality metrics of the constructed model

The data you can see [0 1 2] are used diagnoses: chronic obstructive pulmonary diseases, allergic rhinitis and the absence of diseases.

The main metric is the indicator of an accuracy. It reflects the proportion of correct answers in the constructed model. The indicator is calculated as the ratio of the number of correct answers to the number of all answers. So, according to the parameter «accuracy» we can say, that we have good result of model training.

The parameter «precision» calculated by the formula (Guido S., Mueller A. C., 2016):

$$precision = \frac{true\_positives}{true\_positives + false\_positive}, \quad (4)$$

where:
- true_positives – number of the responses, which seems to be right, and they really were right;
- false_positives – number of the responses, which seems to be right, and they really were not right.

If we compare the value of the indicator «precision» with the standard indicator, we can state the average result of the training model.

The indicator «recall» calculated by the formula (Guido S., Mueller A. C., 2016):

$$recall = \frac{true\_positives}{all\_positive}, \quad (5)$$

where:
- false_positives – the number of all correct answers in the sample.

In our case, we got a good result.

4. Cross-validation

Next, we will test the outcome of the learning. That is, we try to make sure, that this result was not the result of obtained due to a randomly good division of data into training and test samples. To verify this, we apply the so-called «cross-validation». Cross-validation is a method of empirical evaluation of the real ability of algorithms, learned by some precedents.

Ten times we will split the data into training and test samples. Than the result of the algorithm will be averaged (Figure 4).

Final value «accuracy» deteriorated to 0.65
Cross-validation revealed the shortcomings of our model. If we leave the model at this stage of training, we will receive unreliable results of diagnosis of diseases. Therefore, we will try to improve the model by optimizing the coefficients «alpha» and «n_iter» (Figure 5).

![SGDClassifier(alpha=0.0008999999999999998, average=False, class_weight=None, epsilon=0.1, eta0=0.0, fit_intercept=True, l1_ratio=0.15, learning_rate='optimal', loss='hinge', n_iter=75, n_jobs=1, penalty='l2', power_t=0.5, random_state=0, shuffle=True, verbose=0, warm_start=False)](image)

**Figure 5.** The search for optimal values of coefficients «alpha» and «n_iter»

So we found optimal value alpha=0.0009, n_iter=75

Then we retrained our model of diagnosis of diseases and got a very good result of accuracy, which is 0.87.

It can be stated, that we have optimized the classification algorithm and got very good results on the test data.

5. Conclusions
The decision of the tasks, delivered as part of our research, let us speak about the acceleration of the process of serious scientific, technical and technological changes in the modern development of socio-economic systems in the field of health. The preparation of data for the implementation of machine learning methods took most of the time, reserved for the research. Based on our capabilities, we were able to obtain depersonalized data from the results of a General blood test of more than a thousand people in the city of Bryansk. The following diagnoses were included in the General sample: chronic obstructive pulmonary disease, allergic rhinitis and no disease.

The main method of research is a computational experiment. As the main algorithm of machine learning model of diagnosis of human diseases based on the data of the General blood analysis, we have chosen Stochastic Gradient Descent. The distribution of programming languages was chosen as the main environment for machine learning of the medical diagnostics model Python – Anaconda. The ratio of the number of correct answers to all responses were 0.7. Cross-validation revealed the shortcomings of our model. Summary value «accuracy» deteriorated to 0.65. Therefore, we improved the model by optimizing the coefficients alpha and n_iter. We retrained our model of diagnosis of diseases and received accuracy which is equal to 0.87.

It can be stated, that we have optimized the classification algorithm and got very good results on the test data.

The practical significance of the results of this study is to support medical decision-making by specialists of medical institutions in small towns and villages. The results of machine learning models of diagnosis of human diseases based on the data of General blood analysis can be the basis for the development of analytical medical information system. The proposed model of the process of supporting medical decision-making based on machine learning algorithms will improve the quality of medical services

6. References
[1] Guido S, Mueller A C. (2016) Introduction to Machine Learning with Python: A Guide for Data Scientists. O Reilly Media, Inc, USA, United States
[2] Huang K 2008 Machine learning: modeling data locally and globally (Berlin: Springer)
[3] Li C, Pei Z, Li B, & Zhang Z 2009 A New Fuzzy K-Nearest Neighbors Algorithm Intelligent Decision Making Systems
[4] Xue W, Liu P & Liu D 2013 Improved Chameleon algorithm using weighted nearest neighbors graph
[5] Oladipupo T 2010 Types of Machine Learning Algorithms *New Advances in Machine Learning* (InTech)
[6] Bjornsdotter M 2010 *Machine Learning for Functional Brain Mapping Application of Machine Learning* (InTech)
[7] Lu Z 2010 Alarming Large Scale of Flight Delays: an Application of Machine Learning
[8] Motrenko A, Strijov V and Weber G-W 2014 Sample size determination for logistic regression *Journal of Computational and Applied Mathematics* 255 pp 743–752
[9] Zaitsev A A, Burnaev E V and Spokoiny V G 2013 Properties of the posterior distribution of a regression model based on gaussian random fields *Automation and Remote Control* 74(10) pp 1645–1655.
[10] Turkov P, Krasotkina O and Mottl V 2012 The bayesian logistic regression in pattern recognition problems under concept drift *Pattern Recognition (ICPR-2012)* 21st International Conference pp 2976–2979
[11] Genrikhov, I. E., Djukova, E. V., and Zhuravlev, V. I. (2017). On full regression decision trees. Pattern Recognition and Image Analysis, 27(1):1–7.
[12] Baskin, I. I., Marcou, G., Horvath, D., and Varnek, A. (2017). Chapter 12. regression models. In Tutorials in Chemoinformatics. First Edition. Edited by Alexandre Varnek, pages 193–208. John Wiley & Sons Ltd United Kingdom.
[13] Sivogolovko, E. and Novikov, B. (2012). Validating cluster structures in data mining tasks. In Proceedings of the 2012 Joint EDBT/ICDT Workshops EDBT-ICDT ’12, pages 245–250. Association for Computing Machinery (ACM)
[14] Mashechkin, I. V., Petrovskiy, M. I., Popov, D. S., and Tsarev, D. V. (2015). Applying text mining methods for data loss prevention. Programming and Computer Software, 41(1):23–30. and support vector machine. Bulletin of Experimental Biology and Medicine, 156(5):706–709.