MACHINE LEARNING MODEL FOR EVALUATING THE EFFECTIVENESS OF TREATMENT FOR ERYSIPELAS

D.A. Zhurman, R.R. Kotyubeev,
Scientific supervisor: S.V. Aksyonov
Tomsk Polytechnic University
E-mail: daz18@tpu.ru

Introduction

At the moment, erysipelas is the 4th most common in the world among infectious diseases and, in most cases, can be cured. The main symptoms of this disease are fever, pressure, palpitations, headaches, loss of sleep and appetite [1]. Therefore, it is proposed to develop a machine learning model that, based on these symptoms, is able to evaluate the effectiveness of treatment and, if necessary, adjust or change it.

Thus, the aim of this work is to develop a machine learning model for evaluating the effectiveness of the treatment of erysipelas.

Since the goal of the work is very complex, at this stage only the following tasks will be solved:

1. Collect data for model training, that is, digitize patient records
2. Highlight the features that will be key in assessing the effectiveness of treatment.
3. Classify patients according to how quickly and easily treatment was received.

This study is significant because with the wrong treatment the symptoms of the disease can go away, but relapse can occur in the future. The developed algorithms and knowledge can serve as the basis for further studies to assess the effectiveness of the treatment of other diseases. Also, the developed model may be useful for medical insurance companies, as it will allow you to correctly determine the condition of the patient, and thereby adjust the amount of insurance payments for the treatment of the patient.

Data preparation

The data is represented as a collection of the digitalized patients' history. All patients in the data were treated in a hospital with one diagnosis – Erysipelas. Dataset has 58 files.

Then, to reduce the number of word forms, we applied lemmatization using the pymorphy2 library. Lemmatization is a method of morphological analysis, which boils down to reducing the word form to its original vocabulary form (lemma).

After that we downloaded the list of stop words from NLTK library and completed it with other meaningless words from inspections. Examples of stop words from inspections: 'loc', 'localis', 'st', 'локальный','статус'.

Then we applied Tf-Idf vectorization using n-grams from 1 to 3 inclusive. TF is a number of times the term a occurred in the text divided by number of all words in the text. IDF is logarithm of dividing a total number of documents by number of documents in which the term a occurs. Tf-Idf is multiplication of TF and IDF [2].

The result of the Tf-Idf vectorization is shown in figure 1 as a Pandas DataFrame.

	абсцесс	абсцесс учитывать	абсцесс учитывать практически
48	0.013847	0.013847	0.013847
32	0.000000	0.000000	0.000000
2	0.000000	0.000000	0.000000

Fig. 1. The result of the Tf-Idf vectorization

After that we added the number of days which patient was in the hospital and applied normalization of the table.

Thus, the data on all patients is a table consisting of 58 rows and 9083 columns. Each line corresponds to the medical history of one patient.

Dimension reduction

In order to visualize and classify data, it was necessary to reduce the dimension. That is, reduce the number of columns from 9083 to 2.

The graph in figure 2 shows that the principal component method is not suitable for reducing the dimension of a given dataset. This is due to the fact that 50 components describe 90% of the variance, that is, in this dataset there are 50 main components.

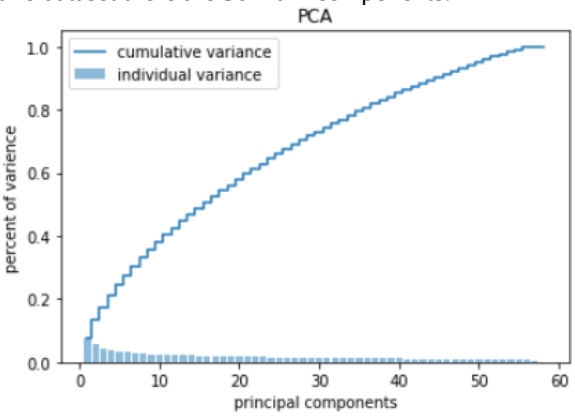

Fig. 2. The result of PCA

Thus, to reduce the dimension, it is necessary to use other methods of reducing the dimension, such as UMAP, Isomap, MDS, and TNSE. Previously, it was necessary to divide patients into 3 groups.

The average treatment time for erysipelas is 10 days. Therefore, it is necessary to classify patients into three groups:

1. in the hospital less than 10 days - a quick recovery;

2. in the hospital from 10 to 12 days - the average recovery;

3. The hospital has more than 12 days - a long recovery [3].

As can be seen from the graphs in figure 3, the UMAP and Isomap methods reduce the dimension in such a way that it is impossible to separate one group of patients from another. Using the MDS and TNSE methods, the following relationship was found: the longer the patient received treatment, the closer to the center he is on the chart.

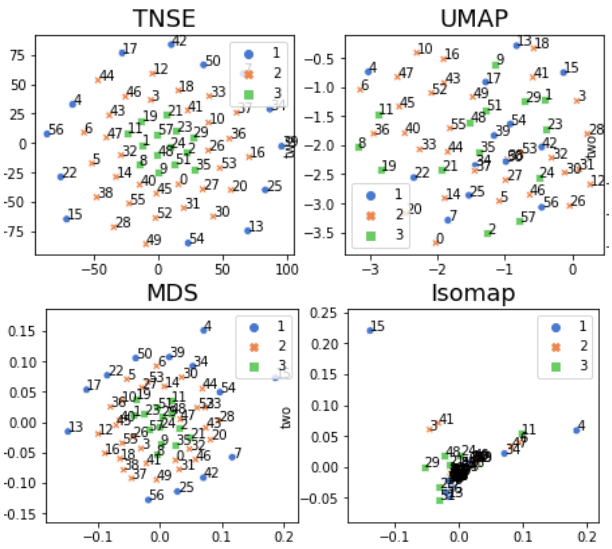

Fig. 3. The result of various methods of dimension reduction

We applied these methods with different parameters values. However, it was possible to separate the various groups of patients on the graph only using the MDS and TNSE methods. From these two methods, TNSE with standard parameters showed the best separation, so it will be used in future work.

Classification

After the dimensionality reduction was made, it was necessary to apply a classifier in order to separate the groups of patients from each other.

First of all, the Nearest Neighbors, Linear SVM, RBF SVM, Gaussian Process, Decision Tree, Random Forest, Neural Net, AdaBoost, Naive Bayes, QDA classifiers were used on the entire data set.

The best accuracy with the preservation of the distribution form and without overfitting was shown by the QDA classifier, shown in figure 4.

After that, the data set was divided into test and training samples. Then reclassification was performed. The best accuracy with the preservation of the distribution form and without overfitting was again shown by the QDA classifier, shown in figure 5.

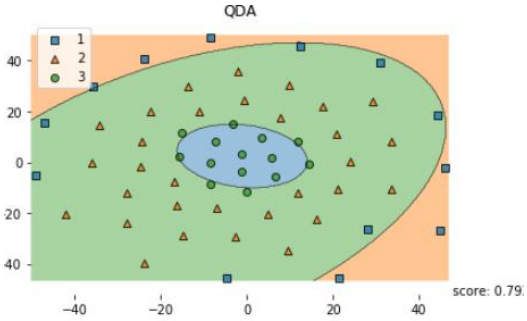

Fig. 4. The result of QDA classification

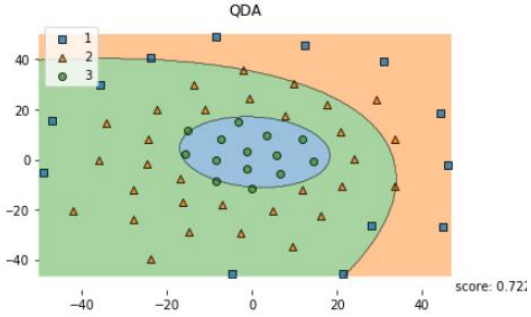

Fig. 5. The result of QDA classification with splitting

The decrease in accuracy and overfitting is connected with a small amount of data, since the sample size is only 58 patients.

Conclusion

In conclusion, patients were classified according to the length of hospital stay. In the problem of reducing dimensionality, the TSNE method showed the best result, and in the classification problem, the QDA method. The dependence was found: the longer the patient was in the hospital, the closer he was to the center on the chart. In the future, when such indicators as pressure, temperature and complaints, ellipses that divide the distributions into 3 groups will be used to evaluate the effectiveness of treatment, they will be divided into different sectors depending on medical indicators.

List of references

1. New policlinica / Erysipelas // URL: http://www.newpoliclinica.ru/zab/rozhistoe-vospalenie/. – (accessed 14.12.2019).
2. MonkeyLearn / What is TF-IDF // URL: https://monkeylearn.com/blog/what-is-tf-idf/. – (accessed 14.12.2019).
3. Journal of Clinic Trials / Randomized Controlled Trial of Short Course Intravenous Therapy for Cellulitis and Erysipelas of the Lower // URL: https://www.omicsonline.org/open-access/randomized-controlled-trial-of-short-course-intravenous-therapy-for-cellulitis-and-erysipelas-of-the-lower-limb-switch-2167-0870.1000200.php?aid=35758. – (accessed 14.12.19)