Diagnostics of diabetic polyneuropathy in children and adolescents using data mining methods

O S Krotova¹, I V Moskalev¹, O M Nazarkina² and L A Khvorova¹
¹Altai State University, 61 Lenin prospect, Barnaul, 656049, Russia
²Maternity and Child Healthcare Centre, 179 Gushchina str., Barnaul, 656019, Russia

E-mail: kr.olga0910@gmail.com, khvorovala@gmail.com

Abstract. The article considers the application of data mining methods to develop a diagnostic model for one of the most common and dangerous diabetes mellitus complications – diabetic polyneuropathy characterized by damage to peripheral nerve fibers. The article explores the possibility of diagnosing diabetic polyneuropathy with the use of machine learning methods. The data base of the study includes 3204 anonymized medical records of children and adolescents with type 1 diabetes residing in the territory of Altai krai. 1100 records hold data on diabetic polyneuropathy. Medical records contain different information: patient complaints, medical case history, test results. The attribute space is represented by 40 different indicators. In the course of the study, we considered the differences between the attribute values in two groups, built the attribute space ensuring the best classification quality. The article presents the results of applying various methods to transform the source data, evaluates the quality of the resulting model. The implementation of all the study phases was carried out with the use of the Python high-level programming language.

1. Introduction
Diabetes mellitus is a dangerous chronic disease, often accompanied by serious complications. One of the most common complications of type 1 diabetes in children and adolescents is diabetic polyneuropathy. Diabetic polyneuropathy is characterized by a whole complex of clinical and subclinical syndromes caused by the damage to peripheral nerve fibers, leading to deterioration in the quality of life. Moreover, it often becomes a cause of disability.

The incidence rate of diabetic polyneuropathy in children and adolescents varies widely from 5 to 50%. The data discrepancy can be explained by the lack of standardized diagnostic criteria and the use of diagnostic methods with varying degree of specificity [1]. According to clinical guidelines for standardization and optimization of the healthcare delivery to diabetes mellitus patients, in the Russian Federation, it is recommended that type 1 diabetes patients of any age be screened for diabetic polyneuropathy annually after the 5-year-period from the onset of the disease [2].

Mandatory diagnostic methods for diabetic polyneuropathy include complaints assessment and examination of the patient. However, children and adolescents often do not have any complaints throughout the development of polyneuropathy [1]. To obtain the most objective information, electro-neuromyography (ENMG) is performed to identify the earliest preclinical signs of damage to the lower and upper limbs nerve fibers [3].
The results of the study [4] show that the actual prevalence of diabetes mellitus complications is 20-50% higher than the one recorded. This is due to the fact that a large number of patients are in adverse conditions in terms of the healthcare availability. Children and adolescents living in the countryside have an opportunity to be screened for diabetes complications only during their hospital stay.

The purpose of this study is to develop a model for diagnosing diabetic polyneuropathy with the use of modern data mining methods based on various patient data. Such models will allow patients to diagnose complications without the use of neurophysiological diagnostic techniques, and can also be used as a support system for medical decision making in diagnostically unclear cases.

Software implementation of all the data analysis and modeling phases was carried out using the Python high-level programming language. Python is a popular programming language for solving problems of data analysis and machine learning. A large number of modules and libraries that provide turnkey solutions for working with data and extensive documentation make the Python programming language a convenient and demand tool for solving both applied and fundamental problems in the field of data processing and data mining.

The data base of the study comprises the anonymized data from 3204 medical records of children and adolescents from Altai krai suffering from type 1 diabetes. The original data sample used in the study included 40 indicators: age, height and weight of the patient, the disease duration, the patient’s complaints and the level of glycated hemoglobin (HbA1c), hormonal tests, complete blood count and biochemical blood test. The task of diagnosing diabetic polyneuropathy is a binary classification task. The output parameter can take two values: 1 – there is a complication, -1 – there is no complication. Diabetic polyneuropathy was documented in 1100 medical records, while in 2104 records, no complication was detected following the results of the patient examination and ENMG.

2. Text mining
The information contained in medical records has both a structured and an unstructured form. Part of the data, for example, test results, is kept in a table form, another part, for example, patient complaints and the medical case history is kept in the form of records in natural language.

An important indicator for this study is the presence of the symptoms characteristic of polyneuropathy in the patient. The main cause for diabetic polyneuropathy is considered to be chronic hyperglycemia. The main manifestations of diabetic polyneuropathy are the pain symptom, paresthesia and reduced tendon reflexes. However, often in children and adolescents, there are no symptoms throughout the whole course of the disease [1].

The NLTK (Natural Language Toolkit) library of the Python programming language is designed for symbolic and statistical processing of the natural language contains detailed documentation, graphic representations, supports working with many languages, including Russian. Using the NLTK, we analyzed the presence of hyperglycemia, pain and other symptoms characteristic of polyneuropathy in two groups (group 1 – records where polyneuropathy was documented, group 2 – records where polyneuropathy was not diagnosed). The results of the analysis are presented in table 1.

| Symptom                                      | Group 1 | Group 2 |
|----------------------------------------------|---------|---------|
| Intermittent / persistent hyperglycemia      | 653 (59.4%)  | 736 (35%)  |
| Pain syndrome (limb pains, headaches)       | 621 (56.4%)  | 569 (27%)  |
| Other symptom (muscle cramps, limbs weakness, numbness, tingling, etc.) | 606 (56.9%)  | 642 (30.5%)  |

The considered symptoms are much more common in patients with diabetic polyneuropathy than in patients having no complications. However, each symptom is associated with polyneuropathy only in half of the cases.
3. Attribute selection

The quality of the models is greatly influenced by the attributes that are used for their learning that is why the most important task is the quality origination and transformation of the source data. The first phase of data origination is the exclusion of non-informative parameters. Classical methods for attribute selection are statistical data analysis and the application of sequential attribute selection algorithms.

The database contains information from various medical records. Some patients were admitted to the hospital several times in a row, while others were admitted only once. There is also data on patients who were previously treated in another hospital. This fact makes it difficult to use statistical criteria. During the statistical data analysis, the differences in the descriptive statistics of two groups were studied. The greatest differences are observed in the patients’ age and the disease duration, the content of lymphocytes, segmented neutrophils, alkaline phosphatase, potassium and creatinine in blood.

Table 2. Descriptive statistics of features with the greatest differences.

|        | mean | std  | min  | max  | 25%  | 50%  | 75%  |
|--------|------|------|------|------|------|------|------|
| **Group 1** |      |      |      |      |      |      |      |
| age    | 12.30| 3.57 | 0    | 17   | 10   | 13   | 15   |
| disease duration | 66.03 | 43.65 | 0     | 208  | 32   | 60   | 91   |
| lymphocytes | 40.79 | 11.75 | 3    | 78   | 34   | 41   | 49   |
| segmented neutrophils | 50.64 | 12.22 | 5.3  | 89   | 43   | 50   | 58   |
| alkaline phosphatase | 397.34 | 288.28 | 0.82 | 1556.7 | 196.52 | 308.75 | 507.7 |
| potassium | 4.56  | 0.5  | 0.84 | 5.91 | 4.26 | 4.59 | 4.88 |
| creatinine | 63.76 | 63.76 | 4.24 | 180.5 | 55.17 | 62.24 | 70.75 |
| **Group 2** |      |      |      |      |      |      |      |
| age    | 9.64 | 4.26 | 0    | 17   | 6    | 10   | 13   |
| disease duration | 32.49 | 34.32 | 0     | 189  | 7    | 22   | 48   |
| lymphocytes | 44.01 | 12.41 | 2    | 82   | 32   | 44   | 52   |
| segmented neutrophils | 47.51 | 12.46 | 4.7  | 89.8 | 39   | 47.9 | 55   |
| alkaline phosphatase | 460.54 | 293.24 | 22.61 | 2010.2 | 244.7 | 369.7 | 637.57 |
| potassium | 4.46  | 0.49 | 1.24 | 6.2  | 4.2  | 4.5  | 4.7  |
| creatinine | 60.06 | 14.59 | 4.63 | 212.78 | 50.82 | 58.63 | 67.4 |

We chose the sequential backward selection method for the implementation of the sequential attribute selection. The algorithm, using the specified criterion, sequentially removes the attributes from the attribute set until the specified number of attributes remains. The difference between the values of the classification quality metric after and before the elimination of an individual attribute is used as the criterion. The task of diagnosing diabetic polyneuropathy, like many other tasks of medical diagnosis, is non-linear. The application of linear transformations methods to the solution of the task under consideration is not effective. Therefore, the sequential backward selection was carried out with the use of the Support Vector Machines with a kernel from the radial basis function (RBF) from the Python scikit-learn library and the proportion of valid answers as a selection criterion. Currently, the libraries of the Python programming language do not feature a ready-made solution for the method of sequential backward attribute selection. However, this algorithm is quite simple to implement using standard libraries.

Figure 1 shows the dependence of the classification quality on the number of attributes. The maximum proportion of correct answers – 81% – is achieved when the model is trained on 9 or 16 attributes. Thus, using statistical data analysis and sequential attribute selection, we determined the attributes to build and train the classification model: the patient age, the disease duration, the presence
of intermittent and persistent hyperglycemia, pain symptom, other symptoms that can be caused by diabetic polyneuropathy, glycated hemoglobin level, the content of total protein, cholesterol, alkaline phosphatase, potassium, segmented neutrophils and red blood cells. Some of these parameters do not have significant statistical differences in two groups of medical records, but their combination can affect the quality of classification significantly.

![Figure 1. Dependence of the classification quality on the number of attributes used to train the model.](image)

The studied data contain a large number of missing values. Before proceeding to the next phase of the data analysis, it is necessary to delete the cases containing the missing values in the selected attributes. For further transformations and model training, a data sample containing complete data on 12 attributes was generated. A new data sample includes 1987 medical records.

4. Visualization

Methods of reducing the dimensionality of data provide the ability to compress data, visualize, get rid of noise, and also find a representation of the data that will give more information in the course of further analysis. The most popular method of data dimensionality reduction is the principal component analysis (PCA). The given task is nonlinear; therefore, the kernel principal component analysis was chosen as the method of data dimensionality reduction. The kernel PCA is applied to perform nonlinear mapping, which converts the data into a higher dimensional space. Then, in this higher dimensional space, the standard PCA is used to map the data back into the lower dimensional space. The kernel PCA algorithm is implemented in the KernelPCA class of the sklearn.decomposition module [5]. The use of KernelPCA is similar to the class for standard PCA, and we can determine the kernel using the kernel parameter.

The difficulty of using kernel PCA with RBF as a kernel to reduce dimensionality is in the fact that it is necessary to a priori determine the hyperparameter gamma. Gamma is the cutoff parameter, an increase in the gamma value leads to a softer solution boundary. To find the optimal value for this parameter we used a grid search.

Figure 2 illustrates the result of the original data set dimensionality reduction from 12 to 3 attributes.
Figure 2. Data set visualization with the application of the dimensionality reduction to 3 entries.

Figure 2 shows that the data set under study is comprised of two intersecting clusters that cannot be separated with the help of linear models.

5. Modeling
The method of Support Vector Machines (SVM) with the kernel from the radial basis function (RBF) was chosen as a data classification model. The choice of the classifier can be explained by the fact that the considered data are not linearly separable and the use of linear classifiers will not provide desirable classification accuracy.

The Support Vector Machines with the RBF kernel can be dealt as a nonlinear generalization of the linear SVM classifier, which allows building models with the use of dividing surfaces of various shapes. When solving a nonlinear problem using SVM, the training data is transformed into a feature space of higher dimension using the display function and the linear SVM model is trained to classify data in this new feature space. The problem is that the construction of new features in the computational plan is very expensive, especially if there is high-dimensional data. Therefore, they use the replacement of the scalar product by the kernel function (kernel trick). One of the most widely used nuclear functions is represented by the RBF core, or Gaussian core.

Training and fine-tuning of the classifier are performed using the classes and methods of the scikit-learn library. To select the optimal values of the model’s hyperparameters, we used a grid search implemented in the GridSearchCV () class.

The data sample is not balanced; it contains more cases related to -1 class. The classifier was fine-tuned taking this fact into account and a model that best defines both classes was selected. The classification quality was assessed by means of such metrics as precision, recall, $F$-score, and accuracy. The test data sample encompassed 596 medical records.

Table 3 presents the values of quality metrics of the resulting classification model. The proportion of correct answers in the test sample equaled 80%.
Table 3. Values of evaluation metrics to assess the classification quality for the built model.

| class | precision | recall | F-score |
|-------|-----------|--------|---------|
| -1    | 0.81      | 0.86   | 0.83    |
| 1     | 0.78      | 0.71   | 0.74    |

Accuracy = 0.80

6. Conclusion
Clinically, the resulting model has a low classification quality and can not be used in medical practice. However, the achieved accuracy allows stating that the methods and approaches of data mining can be used to search for hidden patterns in the course of the disease and the development of complications.

The possibility of using data mining methods to diagnose one of the most common and dangerous complications of diabetes mellitus - diabetic polyneuropathy is considered. The possibility of diagnosing diabetic polyneuropathy using machine learning methods has been studied. The information base of the study included 3204 anonymized medical extracts from children and adolescents of the Altai krai with type 1 diabetes mellitus, in 1100 of which the presence of diabetic polyneuropathy was recorded. Medical extracts contain different information: patient complaints, medical history, test results. The feature space is represented by 40 different indicators. In the course of the study, the differences between the values of the attributes in the two groups were studied, the space of signs was constructed, which ensures the best classification quality. The results of applying various methods to transform the source data are analyzed, the quality of the resulting model is evaluated.

Carrying out such research and the introduction of automated models for the diagnosis of various diseases will enable a transition to a qualitatively new level of using modern information technologies in healthcare management and the provision of medical care.

References
[1] Dedov I I, Kuraeva T L, Peterkova V A and Shcherbacheva L N 2002 *Diabetes Mellitus in Children and Adolescents* (Moscow: Universum Publisher) pp 268–70
[2] Dedov I I et al. 2019 *Diabetes Mellitus* 22(S1) 104–106
[3] Alimova I L 2016 *Russian Bulletin of Perinatology and Pediatrics* 61(3) 114–23
[4] Suntsov Yu I, Maslova O V and Dedov I I 2010 *Problems of Endocrinology* 56(1) 3–8
[5] Rashka S *Machine Learning* (Birmingham: Packt Publishing) p 154
[6] Bjerg L, Hulman A, Charles M, Jorgensen M E and Witte D R 2018 *J. of Diabetes and its Complications* 32(4) 393–99
[7] Fitri A, Sjahriar H, Bachtiar A, Ichwan M, Fitri F I and Rambe A S 2019 *Open Access Macedonian Journal of Med. Sc.* 7(16) 2626–2629
[8] Jelinek H F, Cornforth D J and Kelarev A V 2016 *J. of Diabetic Complications & Medicine* 1(2) 1000108
[9] Turkyilmaz H, Guzel O, Edizer S and Unalp A 2017 *Turkish J. of Med. Sc.* 47(3) 942–46
[10] Qin Let al. 2019 *Chinese J. of Epidemiology* 40(12) 1578–84