Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200451

Risk Factors for Chronic Diabetes Patients

Oleg METZKER [a], Kirill MAGOEV [a], Stanislav YANISHEVSKIY [b], Alexey YAKOVLEV [b] and Georgy KOPANITSA [a]
[a] *ITMO University, Saint Petersburg, Russia*
[b] *Almazov National Medical Research Centre, Saint Petersburg, Russia*

Abstract. Specific predictive models for diabetes polyneuropathy based on screening methods, for example Nerve conduction studies (NCS), can reach up to AUC 65.8 – 84.7 % for the conditional diagnosis of DPN in primary care. Prediction methods that utilize data from personal health records deal with large non-specific datasets with different prediction methods. It was demonstrated that the machine learning methods allow to achieve up to 0.7982 precision, 0.8152 recall, 0.8064 f1-score, 0.8261 accuracy, and 0.8988 AUC using the neural network classifier.

1. Introduction

Every third patient with diabetes suffers from diabetic polyneuropathy (DPN). This complication is a severe problem for a chronic patient because it can be accompanied by neuropathic pain and lead to a decrease in the quality of life. Neuropathic pain significantly disrupts sleep, negatively affects daily activity and life satisfaction. The earlier the diagnosis is found, the easier the disease can be managed. Several studies on the prediction of diabetes mellitus that utilize Linear regression (LR) developed predictive models based on records with up to 967 independent variables available from large electronic health record archives resulted in up to 0.80% Area Under The Curve Receiver Operating Characteristics (AUC-ROC) and enabling to formulate the most likely trajectory of comorbidities. Application of the ensemble models approach based on the naïve Bayes, Linear regression (LR), Instance-Based Learner, support vector machine (SVM), artificial neural networks (ANN), decision trees, and random forest (RF) help to increase the accuracy of diabetes prediction with the accuracy of up to 74%. The goal of this study is the implementation of machine learning methods for early risk identification of diabetes polyneuropathy based on structured electronic medical records.

2. Methods

The dataset contains a total number of 238590 laboratory records. Each record contains patient and episode identifiers, a timestamp, and a varying number of measured parameters, the total number of which is 31 (27 laboratory tests, retinopathy, nephropathy, age, and gender). The records include information 5846 diabetic patients. Diagnosis served as the source of information about the target class values. To identify subclasses within the class of patients with diabetes, the classification problem was solved. Each experiment ran in the setting of stratified 5-fold cross-validation (i.e., random 80% of patients were used for training and 20% for testing, target class ratios in

the folds were preserved). The AUC was calculated based on an average of 5 curves (one curve per fold in the setting of 5-fold cross-validation).

3. Results

There are two large clusters (#0 and #3) that are opposite to the percentage of patients with polyneuropathy, namely the smallest and the largest rate of polyneuropathy. Cluster # 0 has an increased rate of patients with polyneuropathy: 59 %. In cluster #3, on the contrary, the reduced number of patients with polyneuropathy is 22 %. However, these are both female clusters. In male clusters #1 and #2, the percentage is close to the average. Cluster #4 is slightly below average. In clusters 0, 1, and 2, patients are about the same age. The difference in cluster 1 is that there is an outlier on the left, and it is male in contrast to cluster 0. Cluster 3 patients are young men, younger than patients cluster 0, 1, and 2, and also male. Cluster 4 are young male patients, which explains their low percentage of polyneuropathy. Cluster 5 patients of different ages with a small portion of polyneuropathy are mostly men. It is evident that the female gender and the age of more than 50 years is a risk factor for the development of polyneuropathy in patients with diabetes mellitus. However, there is a particular group of young men (cluster 4) with the probability of polyneuropathy. As a result of the cluster analysis, several groups of patients were identified: younger men and women (clusters 4 and 3, the share of polyneuropathy is significantly lower), older men and women (clusters 1 and 2: men, 0: women, the share of polyneuropathy is significantly higher). The most significant features were those of neutrophils and glucose level in the blood and the urine. The correlation matrix shows the high correlation of MPW, RBC, HGB, and polyneuropathy. If we discuss the main features (indicators) affecting the development of polyneuropathy in diabetes, the presence of a decrease in the relative number of segmented neutrophils, which supports the postulate of the main contribution of monocyte cells in the process of inflammation development [13], as well as signs of unsatisfactory glycemic control manifested in glucoseuria and hyperglycemia, become clinically important events.

4. Discussion and conclusion

Clustering shows differences within patients with diabetes mellitus. Decision trees demonstrate pathophysiological mechanisms. It was demonstrated that inclusion of just two features, namely “nephropathy” and “retinopathy”, allows to increase the performance even further, achieving up to 0.7982 precision, 0.8152 recall, 0.8064 f1-score, 0.8261 accuracy, and 0.8988 AUC using the neural network classifier.

Acknowledgements

This work financially supported by the government of the Russian Federation through the ITMO fellowship and professorship program. This worked was supported by a Russian Fund for Basic research 18-37-20002.