Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 156 (2019) 114–122

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Young Scientist Conference on Computational Science

Analysis course of the disease of type 2 diabetes patients using Markov chains and clustering methods.

Ilya V. Derevitskii [a,*], Sergey V. Kovalchuk[a]

[a] *ITMO University, Saint Petersburg, Russia*

Abstract

The main idea of this research is the creation of a new approach to the prediction of chronic diabetes course. This approach is based on dividing patients into several clusters. We used Machine Learning methods for it. Next, we created a diagram of the course of disease in form Markov chains for each cluster of patients. Diagram in form directed graph. Nodes and edges are conditions of patients and transactions between these conditions. Edges values are the probability of transactions. The method based on 7000 histories of diabetes course. A medical specialist can use this method for creating a prediction of diabetes course for a patient.

Keywords: Markov chain, oriented state graph, dynamic model of the type 2 diabetes course, the trajectory of the disease course, non-insulin-dependent diabetes, event clustering, the probability of the disease course complicating, machine learning.

1. Introduction

Type 2 diabetes is one of the most common chronic diseases in the world,[1],[2],[3]. For health care, this type of diabetes is one of the highest priority problems. This disease is associated with many comorbid diseases. It leads to new disability and increases cardiovascular risk. The importance of researching this topic was explained. Next, we will create predictive models for type 2 diabetes course.

There are several approaches to solve the problem of predicting the complications of type 2 diabetes, which we can see in the different literature. The first approach is based on machine learning methods. Scientists create a prediction model using retrospective data. These methods estimate the probability of presence/absence of complications. The final model solves the problem of binary classification [4] and [5]. In approach number 2, scientist considers the time interval between the time of the start of observation and time appearance of diabetic complications, [6], [7]. In the third approach, the dynamics of medical parameters which affect on the trajectory of diabetes is examined. These parameters are selected using prior medical information — the example of this approach [8]. The fourth approach comprises solving the problem using clustering patients followed by the creation of a

1877-0509
10.1016/j.procs.2019.08.186

classification model [9]. Often these approaches provide high-quality of solving the problems. However, these solutions do not create a complete dynamic trajectory of the disease.

In this study, we solve the problem of creating the probabilistic scheme for the course of the disease. This scheme is based on information from electronic medical records of a large number of patients. A probabilistic scheme of transitions of patients between different diabetes states is constructed, and this scheme centers around personal medical history and the dynamics of all important parameters.

We study the patient's medical information space. It is important for assessing the probability of a patient moving from one disease state to another. This space is multidimensional, complex, and multicomponent. Expert analysis is not enough to analyze and describe it. It is necessary to create a specific method of processing medical information and creating diabetes trajectories.

The method should include a module for visualizing the dynamics of the course of the disease.

In this study, a method for predicting a diabetes course using dividing patients into clusters is described along with the creation of a scheme of transitions between diabetic states, which is typical for a given cluster of patients.

The study objectives include: creating a probabilistic model of the course of diabetes; clustering patients with diabetes; analysis of the course of the disease within each cluster; detailed interpretation of the course of the disease within the clusters. Also, the task of the study is identifying the typical patterns of the course of type 2 diabetes and the analysis of their causes for different groups of patients.

2. Case Study: Type 2 DM

We conducted this study based on 6864 medical records. Medical Almazov center of the Ministry of Health of Russia provided it. 6455 patients suffer from diagnosed type 2 diabetes; 99 patients have type 1 diabetes, 310 patients have two types of diabetes. There are 3868 men and 2996 women. (This study is a continuation of the research Graph-Based Predictive Modelling of Chronic Disease Development: Type 2 DM Case Study).

3. The First Stage of the Study - Clustering Patients

We calculate the probability of transition between patients conditions as well as the probability of particular diabetic complications. Probability of transition depends on the personal history of diabetes course. Therefore scientist should consider the individual characteristics of the patients. These characteristics include recent blood and urine tests, the last known measurement of physiological parameters (imt, weight, waist circumference, etc.), the presence of associated diseases associated with diabetes, drug use, and other indicators.

In this study, we propose a method of considering individual medical parameters, which is important for the course of the disease. This method is based on dividing patients into groups. In a group, patients have a similarity of selected parameters.

Our first step was choosing cluster for particular patients based on the similarity of individual characteristics for a particular group.

Clustering was performed using the Python programming language tools, using the sklearn package. The Parameters included gender, age, IHD, weight, body mass index, smoking, maximum glucose level over the observed period.

K-means method was used for clustering. The number of clusters was calculated on the basis of maximizing the metric of the ratio of intercluster distance to intracluster. Also, maximizing the silhouette score metric was applied. We chose the number of clusters from the interval, which include 2 to 25 clusters. The best metrics are when patients are divided into 17 groups.