Dosage planning for diabetes patients using data mining methods

Emirhan Gülçin Yıldırım, Adem Karahoca, Tamer Uçar

* Bahcesehir University, Software Eng. Dept., Istanbul 34353, Turkey

Abstract

The condition of high blood sugar (glucose) level is called as diabetes mellitus. Cause of this disease can be either insufficient insulin production or improper response of body cell to insulin. Diabetes patients should use different drugs in order to keep their blood sugar level within normal range values. The purpose of this study is to develop a data mining model for which will predict a suitable dosage planning for diabetes patients.

Medical records of 89 different patient records were used in this study. 318 diabetes assays were extracted using these patient records. ANFIS and Rough Set methods were used for dosage planning objective.

According to the results of ANFIS and Rough Set methods, ANFIS is a more successful and reliable method for diabetes drug planning objective when compared to Rough Set method.

Keywords: Diabetes, Dosage Planning, ANFIS, Rough Set Theory

1. Introduction

Diabetes mellitus which is often simply referred to as diabetes is a disorder that is caused by decreased production of insulin or by decreased ability to use insulin, for this reason glucose levels in the blood increases [19]. Diabetes increases the risks of developing heart disease, kidney disease, blindness, nerve damage and blood vessel damage [1]. Diabetes mellitus is a widespread disease in the world and researchers attach importance for the diagnosing and predicting of this disease. Diabetes mellitus is divided into two types: type 1 and type 2 diabetes. Type 1 diabetes (insulin dependent diabetes, IDDM) occurs when the pancreas no longer produces any or very little insulin as a result of autoimmune destruction of beta cells. It usually develops in childhood or adolescence and affects 10% of people with diabetes. Different from type 1, type 2 diabetes (non insulin dependent diabetes, NIDDM) occurs when the pancreas does not produce enough insulin to meet the body’s needs or the insulin is not metabolized effectively. Type 2 diabetes usually occurs later in life and affects 90% of people with diabetes [2].

Human body needs energy for activation. The carbohydrates are broken down to glucose. Glucose is the primarily importance energy source for the human body cells. Insulin is needed for glucose transport to body cells [3]. The human body needs steady steady blood glucose. These supply with insulin and glucagon hormones which produced by pancreas. Insulin hormones produced by the beta cells of the islets of langerhans and glucagon...
hormones are produced by the alpha cells of the islets of langerhans in the pancreas. When the blood glucose increases, beta cells are stimulated and insulin given to the blood. Insulin enables blood glucose to get in to the cells and this glucose used for energy. So blood glucose kept in a narrow range [4].

For separation of type 1 and type 2, blood analysis is performed on fast blood sugar, insulin, C-peptide, triglyceride and acetone levels. At type 1 diabetes; fast blood sugar level is over 180 mg/dl, insulin level is under 2 mcg/dl, blood does not include C-peptide and urine includes acetone. At type 2 diabetes; fast blood sugar level is over 125 mg/dl, insulin level is over 20 mcg/dl, blood includes C-peptide and urine does not include acetone. Type 2 diabetes patients have high triglyceride level. This is not valid for type 1 diabetes patients because, type 1 diabetes patients usually have normal triglyceride level.

Nowadays, insulin injection is the only treatment method for type 1 diabetes. Daily insulin requirement varies depending on the patients’ body mass index, age, food consumption and activity level. Treatment method has various types which depend on patients’ blood glucose level, age, body mass index, genetic, food consumption and activity level.

Oral medicine treatment applied for type 2 diabetes. Beside this treatment, some of different treatments are proved. A suitable diet and making some exercises is important. Drug treatment starts with metformin (Glucophage) or gliclazide (Diamicron) included medicines. If patients are not overweight or obese, they will begin to use gliclazide, otherwise they begin to use metformin. However, the blood glucose level cannot decrease metformin and gliclazide used together. After using drugs, if blood glucose level cannot decrease, then short acting insulin injection should be used.

Insulin dosage adjustments are based on blood glucose levels. Nowadays, doctors generally use two types of insulin injection; regular insulin and NPH (Neutral Protamine Hagedorn) insulin. Regular insulin starts being absorbed in 30-60 minutes, and in most individuals, their peak action is within 2 to 3 hours after injection. Regular insulin is often used before eating to control the large rise of blood glucose that often occurs after a meal.

NPH insulin is slow or long acting insulin. NPH insulin starts being absorbed 3-4 hours after being injected and their peak action is after 7-9 hours. The best act is injection at bedtime to control the morning glucose of the next day [4].

The aim of this study is to create a data mining model which will be used for drug dosage planning for diabetes patients. We developed both an ANFIS and a Rough Set model for this dosage planning objective.

2. Background

Data mining techniques are being used for existed comprehensive diabetic records for decades. Data mining, (knowledge discovery in databases) is the process of analyzing data and summarizing the useful information which can be used to make predictions for future experiments. In this section, some of the remarkable studies will be briefly explained [17].

There have been a lot of studies about using data mining techniques with diabetes mellitus data. One of them is Data mining for the Diagnosis of Type 2 Diabetes from three Dimensional Body Surface Anthropometrical Scanning Data. This study is about investigating what the risk factors for anthropometrical data of type 2 diabetes. Neural network, logistic regression, decision tree and rough set data mining approaches are used to predict diabetes. This study reported that BMI (body mass index) and WHR (waist hip ratio) are massive for diabetes disease [2].

Another study is an expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system for diagnosis of diabetes disease. It aims improving the diagnostic accuracy of diabetes disease combining PCA and ANFIS. There were two classes as healthy and patient. The obtained sensitivity, specificity and MSE values by PCA-ANFIS for diabetes are 85.71%, 92.0% and 0.262 [1].

Another one is a comparative study on diabetes disease diagnosis using neural networks. In this study a comparative Pima Indian diabetes disease diagnosis was realized. Pima Indian diabetes dataset is used. There were 768 samples. Multilayer neural network structure and a probabilistic neural network structure were used. As a result of the study, classification accuracy was 82.37% for multilayer neural network and 66.78% for ANFIS [5].

A further study outcome of educational interventions in type 2 diabetes: WEKA data mining analysis. The aim is to enhance diabetes related self-care that contributes to good metabolic control which minimizes the occurrence of
both acute and chronic complications. Data were analyzed with WEKA. Data mining educational content and intensity of education did not predict changes in HbA1c levels [6].

An altered study is Data Mining Diabetic Databases: Are Rough Sets a Useful Addition. This study is about analyzing a diabetic dataset with rough sets. Rough set is applied to Pima Indian Diabetic Database. Rosetta software is used. 392 data is selected randomly. It divided into 300 training records and 92 testing records. In this study discretization method is used best with Johnson algorithm. The obtained accuracy rate is 82.6% [7].

Besides all these, another study is developing an expert-system for diabetics by supporting with ANFIS. The aim was arranging an expert system to help dosage planning. It used data mining with a knowledge based on diabetic patients. 390 patients’ records are used. 300 records used for training and 90 records used for checking [8].

The last study is prediction of cyclosporine A blood levels: an application of the adaptive-network-based fuzzy inference system (ANFIS) in assisting drug therapy. In this study therapeutic drug monitoring is observed. The aim of the study is predicting the concentration of cyclosporine A level in blood. ANFIS is used for predicting cyclosporine A level in blood. 654 TDM assays are collected from 138 patients. 473 records are used for training and 181 records are used for checking [9].

3. Material

For this research, diabetics’ data is collected from hospitals in Turkey. After collecting data, information extracting is performed by using data segmentation process. The collected data are arranged for use in data mining techniques.

3.1. Preparing Diabetes Data Set

Diabetes mellitus assays carried out in 2008 and 2009 on diabetes patients who made laboratory tests in State Hospitals which were included in this study. In this research, collected blood and urine samples were analyzed in State Hospital Laboratories.

Data on the total of 318 diabetes assays were collected from 89 patients (44 type1 DM, 45 type 2 DM). The data collected for each assays were gender, age of patient (years), body mass index of patient, genetic (parents of patients were diabetic or not), blood insulin (mcg/dl), C-peptide, fast blood sugar (FBS)(mg/dl), urine in blood (mg/dl), creatinine (mg/dl), total cholesterol (mg/dl), triglyceride (mg/dl), high density lipoprotein (HDL)(mg/dl), low density lipoprotein (LDL)(mg/dl), uric acid (mg/dl), acetone in urine (mmol/l) and glucose in urine (mmol/l). Diabetic patients use drugs and injections to reduce blood insulin level in a narrow range.

In this research, drug dosage planning is made by using data mining methods. These drugs are Glifix, Diamicron, Glucophage, Humilin-M, Novorapid and Insulin Lantus. 6 different data sets are formed for each of these drugs. ANFIS algorithm is used for planning the degree of opponent medicine for both types of diabetes patients. Some population statistics for the training and checking sets of used medicines are shown in the following table:

<table>
<thead>
<tr>
<th>Drug</th>
<th>Checking Data (n)</th>
<th>Training Data (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glifix</td>
<td>12</td>
<td>25</td>
</tr>
<tr>
<td>Diamicron</td>
<td>28</td>
<td>58</td>
</tr>
<tr>
<td>Glucophage</td>
<td>39</td>
<td>78</td>
</tr>
<tr>
<td>Humilin-M</td>
<td>12</td>
<td>25</td>
</tr>
<tr>
<td>Novorapid</td>
<td>59</td>
<td>120</td>
</tr>
<tr>
<td>Insulin Lantus</td>
<td>62</td>
<td>125</td>
</tr>
</tbody>
</table>

The following table lists the whole parameters that are used for building data mining models:
In this research 17 different parameters are used. We can explain the meaning of these parameters as follows:

Gender can take two values male or female. For male patients gender takes 0 value, for female patients it takes 1 value. Age is segmented in 8 partitions which are “0-7”, “8-14”, “15-25”, “26-35”, “36-45”, “46-60”, “61-70” and “70+”. BMI is divided into five parts, between 10-18.5 under weight, 18.51-25 normal, 25.01-30 over weight, 30.01-40 obese and 40.01-70 over obese. Type 1 diabetic patients’ body mass indexes are usually normal and underweight but type 2 patients’ are overweight and obese in general. In general, type 2 diabetes disease is inherited. If the patient’s parents already suffered from diabetes mellitus, then genetic is 1, otherwise genetic is 0. Type 1, diabetes patients’ blood insulin level is under 2 mcg/dl and type 2 diabetes mellitus patients’ blood insulin level is over 20 mcg/dl. Between 2-20 mcg/dl insulin level is normal. Healthy people and type 2 diabetic patients’ blood has C-peptide in their blood. Only type 1 patients cannot have C-peptide in their blood. C-peptide is a hormone secreted with insulin in the blood. At type 1 diabetes, insulin produced any or very little therefore type 1 diabetes patients’ blood have not got C-peptide in their blood. Fast blood sugar (FBS) is divided into 9 parts. Under 100 mg/dl is normal values. Between 100-125 mg/dl can be a prediabetes, more than 125 mg/dl is diabetic patient. At type 1 diabetes, value of FBS is more than 180 mg/dl. The urine level is normal among 10-50 mg/dl. For overage people the acceptable range is 51-75 mg/dl, over 75 mg/dl is an attention for kidney disease. Also high level creatinine causes kidney disease. Over 2 mg/dl for creatinine is acceptable risk of kidney disease. Diabetes triggers high cholesterol and heart disease consequently cholesterol level have to be kept in a narrow range. Over 240 mg/dl is high cholesterol and a risk of heart disease. Until 200 mg/dl is normal level for cholesterol. The patient must be careful for cholesterol level between 200-240 mg/dl because this level is near high level. If the cells could not use glucose to produce energy, so burn fat cells. Then fat cells mobilize and passing through the blood. In this case, triglyceride level increases. As burning fat cells, the body needs just bit insulin. Therefore, type 2 diabetes patients have high triglyceride level. This condition is not valid for type 1 diabetes patients because at type 1 diabetes pancreas cannot produce insulin. Normal degree of triglyceride level is under 150 mg/dl normal, between 150-199 mg/dl is near high level and over 200 mg/dl is considered high level. Over 60 mg/dl for HDL level is high and optimal condition. It considered protective against heart disease. Between HDL and LDL levels have inverse ratio. High LDL level is increased risk of heart disease. The optimal LDL level is under 100 mg/dl. The near optimal LDL level is 100-129 mg/dl, between 130-159 mg/dl is borderline high LDL level, 160-190 mg/dl is high LDL level and over 190 mg/dl is very high LDL level. High uric acid level is associated with type 2 diabetes. Over 8 mg/dl is accepted high level for uric acid. If the insulin level is approximate to zero, acetone will appears in urine. At type 1, in the urine of patients has included acetone. High level fast blood sugar (FSB) is caused glucose in urine.

4. Methods

In this study, we implemented an ANFIS and a Rough Set classification method for dosage planning purpose. In this section, both of these methods will be briefly described.

4.1. Adaptive Neuro Fuzzy Inference System (ANFIS)

Fuzzy inference system is mapping a given input to an output using fuzzy logic. The fuzzy inference systems are used in fields such as automatic control, data classification, decision analysis, expert systems and computer vision. The Fuzzy logic which is based on the linguistic expression is an artificial intelligence technique. The fuzzy logic approach is published by [10] to define the complicated systems. Adaptive Neuro Fuzzy Inference System (ANFIS)
is the combination of ANN and the fuzzy logic. ANFIS is a multilayer feed forward network which uses ANN learning algorithms and fuzzy reasoning to characterize an input space to an output space [11]. Takagi and Sugeno proposed the first systematically fuzzy modeling. The fuzzy inference system’s process consists of several components which are membership functions, fuzzy logic operators, and if-then rules [12].

In this research, Takagi and Sugeno type fuzzy if-then rules are used such that the output of each rule is a linear combination of input variables plus a constant term. The final output is the weighted average of each rule’s output. ANFIS is a fuzzy rule based classifier in which the rules are learnt from examples that use a standard back propagation algorithm [13].

4.2. Rough Set Theory

The rough set theory is developed by Pawlak. It is interested in classificatory analysis of data sets. Rough Set analysis aim is to synthesize approach of concepts from the acquired data. Its principle is every object of the universe is associated some information. Because, the same information with some elements interdependent by the result [14].

From this origin, two different elements can be indiscernible in view of the available information. Information related with objects of the universe generates a relation on its elements [18].

Any subset of the universe is characterized by two ordinary sets. These are lower and upper approximations. The lower approximation of the target set consists of only those objects. That can positively identified as members of the set. The upper approximation includes all objects but some objects of upper approximation may not be the members of the target set. The lower approximation’s objects has the probability=1 while the upper approximation’s objects has the non-zero probability [15].

The rough set occur combining of the lower and upper approximations. The Rough Set methods are applied as a component of hybrid solutions in data mining. The Rough Set data information model is collected in a table. The main purposes of analyzing data with Rough Set are data reduction, missing value handling, feature selection and feature extraction [16].

5. Findings

The results of this study will be mentioned in this part. Both ANFIS and Rough Set methods are applied to the data sets. As mentioned in Section 3, there are six data sets which are Glifix, Diamicron, Glucophage, Humilin-M, Novorapid and Insulin Lantus. The sensitivity results of ANFIS and Rough Set are shown in Table 3. The comparison of ANFIS and Rough Set RMSE rates are shown in Table 4.
### Table 3. Sensitivity results for ANFIS and Rough Set methods

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sensitivity (ANFIS)</th>
<th>Sensitivity (Rough Set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glifix</td>
<td>82%</td>
<td>80%</td>
</tr>
<tr>
<td>Diamicron</td>
<td>78%</td>
<td>70%</td>
</tr>
<tr>
<td>Glucophage</td>
<td>75%</td>
<td>67%</td>
</tr>
<tr>
<td>Humilin-M</td>
<td>72%</td>
<td>65%</td>
</tr>
<tr>
<td>Novorapid</td>
<td>73%</td>
<td>47%</td>
</tr>
<tr>
<td>Insulin Lantus</td>
<td>74%</td>
<td>60%</td>
</tr>
</tbody>
</table>

### Table 4. RMSE results for ANFIS and Rough Set methods

<table>
<thead>
<tr>
<th>Dataset</th>
<th>RMSE (ANFIS)</th>
<th>RMSE (Rough Set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glifix</td>
<td>18%</td>
<td>50%</td>
</tr>
<tr>
<td>Diamicron</td>
<td>11%</td>
<td>10%</td>
</tr>
<tr>
<td>Glucophage</td>
<td>21%</td>
<td>61%</td>
</tr>
<tr>
<td>Humilin-M</td>
<td>19%</td>
<td>25%</td>
</tr>
<tr>
<td>Novorapid</td>
<td>18%</td>
<td>32%</td>
</tr>
<tr>
<td>Insulin Lantus</td>
<td>18%</td>
<td>50%</td>
</tr>
</tbody>
</table>

If we compare ANFIS and Rough Set methods, according to the sensitivity and RMSE values, ANFIS generated better results than Rough Set method for all data sets. Considering the better RMSE values of ANFIS method, we can also state that ANFIS obtained more reliable results than Rough Set.

### 6. Conclusion

In this study, we tried to determine the degree of drug amount dosage for diabetes patients. 318 diabetic assays were used in order to generate a suitable data mining solution. As a result of these assays, more than one drug was used. These data were classified into 6 classes, according to the types of used drugs. As a result of this classification, six data sets are obtained which are Glifix data set, Diamicron data set, Glucophage data set, Humilin-M data set, Novorapid data set and Insulin Lantus data set. To predict the dosage levels for these drugs, 2 different data mining methods are applied on the data sets. Considering the results which are stated in Section 5, ANFIS is a more reliable method than Rough Set algorithm for dosage planning of diabetes patients.

### References


