

# EVALUATION OF DATA MINING CLASSIFICATION AND CLUSTERING TECHNIQUES FOR DIABETES

Tuba PALA<sup>1</sup> and Ali Yilmaz Camurcu<sup>2</sup>

<sup>1</sup>Duzce University, Golyaka Vocational School,

<sup>2</sup>Fatih Sultan Mehmet Vakif University, Department of Computer Engineering

## Abstract

*Analysis of the diagnosis and treatment records by computer programs in the field of medicine constitutes a treatment-supporting factor. Like in many fields, in the field of medicine, too, the use of the methods of data mining has been increasing. It has been aimed to develop the system which will give doctors help for effective treatment and early diagnosis with the result to be obtained in the medical data sets by realizing medical decision support system design in which the methods of data mining are used. After the pre-processing stage in the data mining process, in the data classification stage, Support Vector Machines (SVM), Naive Bayes, Decision Trees, Artificial Neural Networks (ANN), Multilayer Perceptron (MLP), Logistic Regression (LR) algorithms have been used. The success evaluation of data mining classification algorithms have been realized through the data mining programs Weka and RapidMiner. Multilayer Perceptron algorithm has been the best algorithm with the highest success percentage in both of the programs; Decision Trees has been the algorithm which has the lowest success percentage again in both of the programs. This study has indicated that data mining can be a useful tool in the medical field. Doctors can be provided convenience in the progress of the disease and the treatment for patients for whom prediction of disease is made by making morbidity or non-morbidity prediction of the diabetes which is seen on many people in the world and in our country beforehand.*

**Keywords:** *Medical Decision Support System, Data Mining, Classification, K-Means Clustering, Diabetes Data Set*

## 1. Introduction

The Diabetes is a metabolic disease which reveals itself through hyperglycemia. It appears as a result of insulin deficiency or inefficacy, and chronically progresses. Being a hormone secreted by the pancreas, insulin provides the glucose in blood to be used by tissues. Without insulin, tissues cannot effectually use the nutrients taken in and the glucose in blood increases. This metabolic malfunction leads to modifications in many organs and diabetic individuals have an important place in the health system. The diabetes will probably continue to be the most important disease and death reason with its worldwide increasing number in the future, as well (Harrison İç Hastalıkları, 2004). If the insulin hormone is totally missing, this diabetes is called “Type 1 Diabetes (the diabetes dependent on insulin)” (The Society of Endocrinology and Metabolism of Turkey, 2012). Generally, it is seen in children or patients at young ages.

If there is the insulin hormone in the body, but its amount is low or there is resistance to insulin in the tissues, this diabetes is called “Type 2 Diabetes”. Type 2 Diabetes is the most prevalent metabolism disease in the adult society. In developed countries, 5-10% of the society is diabetic (Lakso, 2003). Being generally accepted as a middle-age disease, type 2 diabetes has recently started to be seen also at young ages (International Diabetes Federation, 2003). Due to its chronic complications, economic burden of the diabetes on the national health budgets is quite much (Williams, 1997). Type 2 diabetes is a serious global health problem in most countries. Type 2 Diabetes has developed with aging population, increasing urbanization, diet changes, decreasing physical activities, unhealthy life style, behavior patterns and fast

cultural and social changes (Pickup, 2003). In this study, many classification algorithms have been implemented on PIMA Diabetes data set by UCI and the performance of this algorithm has been analyzed by the data mining tool WEKA. According to the analysis results, C4.5 Decision Tree algorithm has been the algorithm which has the best classification accuracy with a classification accuracy of 91% (Vijayarani and Sudha, 2013). In this study, 10 different data mining algorithms have been used in the diabetes prediction of the patients. While evaluating the algorithms, their accuracy, sensitivity, specificity values have been examined (Kartikeyani *et al.*, 2012).

In this study, the hybrid model that has been developed to classify the PIMA diabetes data set is mentioned. The model is composed of 3 stages. At the first stage, the misclassified records have been deleted from the data set by K-Means. At the second stage, the process of feature extraction has been realized by using genetic algorithm (GA) and CFS. At the third stage, the data set has been classified with the classification algorithm K-NN. In this study, it has showed to have got better results than the previous studies with a classification accuracy of 96.68% for K=5. Weka data mining tool has been used (Karegowda *et al.*, 2012). In this study in which the diabetes prediction is done by using PIMA diabetes data set, Decision Trees algorithm has been used. It includes two stages (AlJarullah, 2011). The purpose of this study is to form a model for an accurate prediction by benefiting from the data of diabetic patients. For this reason, application of data mining techniques on the diabetes data and making rules can reduce the probability of catching this disease in the future, and it will help the patients diagnosed as diabetics to have early treatment, and their disease to be brought under control.

## 2. Data Mining

Substantial and fast development of technology has enabled data mining systems to develop, too. Developments in data mining systems have enabled data collection, data base formation, data management and easier data transfer into the electronic environment, and reliable and cheap data storage. Data mining, as can be understood from its name, is a technique that tries to extract mine-like valuable and variant information from data stacks. The purpose in data mining is the data collected in data bases to be analyzed by examining via mathematical and statistical methods, and the available rules, structures or some different unforeseen information to be revealed.

### A. Data Mining Application Programs

RapidMiner has been developed using Java code by Yale University. For categorization and regression, there are many characteristics such as a variety of algorithms, decision trees, Bayesian, logical clusters, association rules and clustering algorithms, data preprocessing, normalizing, filtering properties, genetic algorithm, artificial neural networks, data analysis in 3D. RapidMiner can import data from the databases Oracle, Microsoft SQL Server, PostgreSQL, MySQL (Dener *et al.*, 2009). WEKA (Waikato Environment for Knowledge Analysis) is an open-source-code Data Mining application development program which has been developed on the Java platform which is used by several people in the world today and developed in Waikato University, New Zealand.

### B. Data Mining Models

#### a. Data Mining Categorization Models

Human mentality tends to categorize and classify the objects, events, situations around. Thus, he can understand and talk about the objects and events better. Categorization process in data mining is the process of categorizing the available data according to the determined features and of predicting this data category when some new data are added.

**Decision Tree:** Decision Tree is one of the most frequently used categorization models to analyze data. A decision tree is composed of root, trunk and leaf joints. Considering that the structure of a tree develops from root to leaves, it's formed from top to bottom. The most outer joint is the root joint. Each inner joint of the tree is separated to make the best decision with the help of algorithms (Quinlan, 1993). Tree leaves form the category tags, namely, categorical characteristics, making the data in the data set groups.

**Artificial Neural Networks:** Artificial Neural Networks (ANN) normally enable people to learn from similar or different events they experience, face, observe, to get new knowledge, to generalize those events, relating them with each other, to learn from their mistakes if they make any, and to make decision using all of these in an event they come across. ANN is inspired from human's problem solving with the abilities of thinking, observing, learning from mistakes, trial-error, that is, in a more general speaking, learning. An artificial neural network is an information processing system based on human cognition simulation. It is composed of plenty of calculating neural units attached together. These units are called neurons. The neurons in the nervous system form the network getting attached. ANN is a model developed on layers.

The neurons in the network are arranged all along the layers. A neuron in a layer is connected to all of the neurons in the next layer. Every neuron in the layers takes informative signals from all of the neurons in the former layer, and multiplies every informative signal with their massive values. Process is done with the activation function to get the output, collecting weighted inputs. The output of this function is transmitted to all of the neurons in the next layer. This process is completed after done by the neuron in the output layer, too (Beh Boon *et al.*, 2011).

**Multilayer Perceptron:** MLP (Multilayer Perceptron) is an artificial neural network model which is mostly used and learns best (Delen *et al.*, 2005). It is known that Multilayer Perceptron has a very strong function in classifying prediction problems (Delen *et al.*, 2005). The purpose of this model is to minimize the difference between the target result (output) of the network and the attained result. This model is expressed as propagation algorithm since it makes the mistake spreading it over the network, or as back-propagation since it uses back-propagation learning algorithm in the process of that the multilayer perceptron is getting trained.

**Naive Bayes:** Bayes classifiers are statistical classifiers. Bayes predicts the membership probabilities of the data, that is, their probability about belonging to a specific category. Bayes classifier is based on the Bayes theorem explained below:

A sample in a data set is composed of the input values  $X = \{x_1, x_2, x_m\}$ . If it is pretended that the total number of the categories is  $m$ , the probability calculations are done with Equation 1 for the sample whose category is to be determined (Jiangtao *et al.*, 2009).

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (1)$$

$p(x|C_i)$  : the probability for a sample from Category  $i$  to be  $x$

$P(C_i)$  : the first probability of Category  $i$

$p(x)$  : the probability for any sample to be  $x$

$P(C_i|x)$  : the probability for sample  $x$  to be from Category  $i$  ( last probability)

**Support Vector Machines:** Support vector machines are a popular method for binary categorization. SVMs can be seen as a perceptron which tries to find a hyper plane that separates the data. The perceptron simply tries to find a hyper plane. It does this process in a way that it separates the data of the hyper plane best. However, it is preferred for the hyper plane to separate the categories as much as possible because it is sought that the hyper plane generalizes the invisible data best with this way of separation. The structure that technically best calculates the separation of data is geometric boundaries. The distance of the hyper plane means the largest boundary which will separate the closest points in the data set, saying in general, means the largest boundary which will categorize the data best (Menon Krishna, 2009). It is not enough for the data in the data set to be in the right side, in the right category of the hyper plane in SVM. Also, they should be in some distance to the hyper plane for a good generalization. Determination of the most appropriate hyper plane directly affects the success of the categorization.

**Logistic Regression:** The statistical analysis method which is used to express the relationship between a dependent variable and one or more than one independent variables numerically is called Regression Analysis. The purpose in Regression Analysis is to calculate to what extent the independent variables affect the dependent variable, that is, to predict the value of the dependent variable starting out of predicted variables. Logistic Regression is primarily used to predict binary or multi-category dependent variable categories; because response variable is separate, and cannot be directly modeled with linear regression. Therefore, instead of prediction of the point, prediction model of the event occurring is built. In case of the ratio to be more than 50% in the binary categorization problem, category '1' is meant to be assigned to the category determined as category '0' for other cases (Beh Boon *et al.*, 2011). The purpose of Logistic Regression Analysis is to calculate to what extent independent variables affect dependent variables.

#### b. Data Mining Clustering Model

K-Means is the best and mostly used clustering algorithm. It is used to divide into k number of categories which have been determined according to quality and characteristics of the records in the data set beforehand. Categorization is done by locating the records in the data set in the closest cluster centers. The success of the clustering process is determined by intra-cluster similarity to be maximum, inter-cluster to be at minimum. In other words, objects in the cluster are to be located in the closest way, clusters in the most distant way.

### 3. Application

In this part, an application reported in an article called “Hybrid Prediction Model for Type 2 Diabetic Patients” (Patil *et al.*, 2010) has been realized on the diabetes disease data with two data mining programs, and the prediction performance rates ensuing as a result of the application have been compared with the programs and the article from which the sample is taken. The application of the designed model has been done using the data mining programs RapidMiner and WEKA. The diabetes data set has been taken from the web site of UCI (UC-Irvine archive of machine learning datasets (UCI Machine Learning Repository, 2012)). The data comprise statistically important information about diabetic patients. The diabetes data set consists of 768 patients and 9 attributes. There are no missing data in the data set. Detailed portray of the data set is seen in Table 1.

Table 1. The Diabetes Data Set

| No | Name of the attribute                     | Definition of the attribute   |
|----|---|---|
| 1  | Number of pregnancy                       | Numerical values  |
| 2  | Plasma glucose concentration              | Glucose concentration in the 2 <sup>nd</sup> hour in oral glucose tolerance test                  |
| 3  | Diastolic blood pressure                  | mm Hg   |
| 4  | Triceps derma thickness (mm)              | Triceps derma thickness   |
| 5  | Serum insulin in the 2 <sup>nd</sup> hour | Insulin (mu U/ml)   |
| 6  | Body mass index                           | ( Kg weight/(m height)^2)   |
| 7  | Diabetes family history                   | Whether there is the diabetes or not in the family  |
| 8  | Age                                       | Age   |
| 9  | Class                                     | 1 – test result for the diabetes positive (yes)<br>0 - test result for the diabetes negative (no) |

#### A. Data Set Pre-Processing Procedure

The data get prepared for the analysis before the data mining operation. The pre-processing for the diabetes data has been reported.

**Data Cleaning:** Since the plasma glucose concentration cannot be 0 in a living person, these data are deleted. Then, the 0 inputs in all of the observations are deleted. As a result, when the data out of 768 diabetic patient data, as stated above, are deleted, 625 diabetic patient data remain.

**Data Transformation:** Attribute values are put into certain intervals with the mean and the standard deviation values by doing data transformation with Z-score normalization.

**The Model Applied on the Data Set:** The model created to predict the survivability of the diabetic patients has been designed as in Figure 1. The prediction performance rates of the model and the programs have been evaluated by applying the programs RapidMiner and WEKA on the data set of the designed model.

As the test method, “10 times cross-check” has been used while operating the algorithms. When the created models are being compared, accuracy value has been used. In the evaluation of the model performance, the parameter of accuracy ratio has been used. Accuracy ratio is the ratio of the number of the accurately classified samples to the total number of the samples (2) [20].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (2)$$

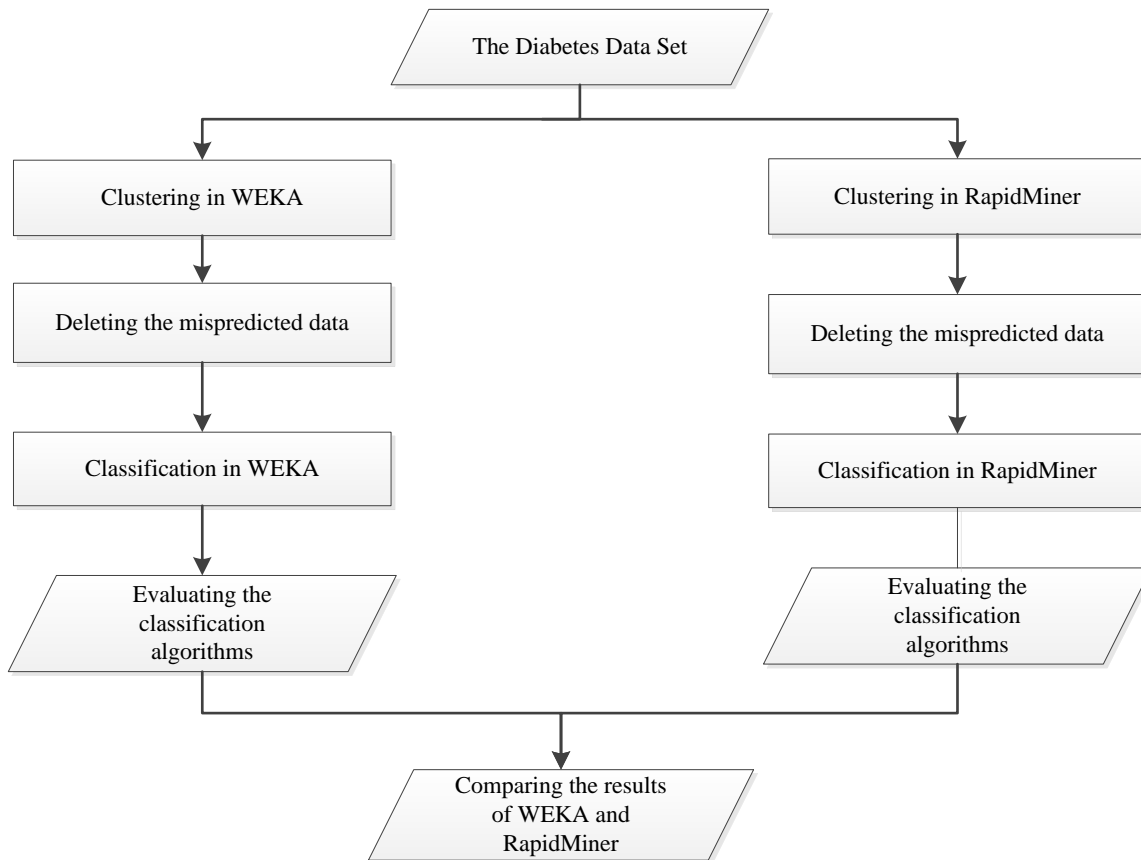


Figure 1. Block Diagram of the Model Applied on the Diabetes Data Set

### B. Application in Weka

In all of the operations conducted in WEKA, the data set has been transformed into the Arff file format which Weka program can read. The diabetes data set has been clustered with K-Means clustering algorithm, and the result of the clustering is that the new data set is to be created deleting the data (patient records) whose classes are mispredicted.

Table 2. KMeans Algorithm Clustering Classes

| Attribute Clusters | The Number of the Samples | The number of the data whose classes are mispredicted |
|--------------------|---------------------------|---|
| Cluster 1 → yes    | 625                       | 194   |
| Cluster 0 → no     |                           |   |

When the classifying algorithms whose success percentages are the highest are applied on the new data set, the results are as in Table 3.

Table 3 Accuracy Values of WEKA Classifying Algorithms

| <b>Classifying Algorithms</b> | <b>Accuracy(%)</b> |
|-------------------------------|--------------------|
| <b>Decision Trees</b>         | 93.50              |
| <b>ANN</b>                    | 95.12              |
| <b>NaiveBayes</b>             | 96.28              |
| <b>SVM</b>                    | 97.21              |
| <b>LR</b>                     | 98.60              |
| <b>MLP</b>                    | 98.83              |

The classifying accuracy values of this application have given very good results. The algorithm which has the lowest accuracy value is Decision Trees with 93.50%. ANN has an accuracy percentage of 95.12%, Bayes of 96.28%. In this application, MLP, 98.83%, Logistic Regression, 98.60%, have given the highest and the closest results. SVM algorithm is the third best algorithm again with a high success percentage of 97.21%.

### C. Application in Rapidminer

After the diabetes data imported to RapidMiner from Excel have been clustered with K-Means clustering algorithm, the clustering classes are as in Table 4.

**Table 4** Clustering Classes of the KMeans Clustering Algorithm

| <b>Attribute Clusters</b>             | <b>The Number of the Samples</b> | <b>The number of the data whose classes are mispredicted</b> |
|---------------------------------------|----------------------------------|--|
| Cluster 1 → alive<br>Cluster 0 → dead | 625                              | 176  |

The misclassified data has been deleted, and the classifying algorithms have been applied on the remaining data.

**Table 5** Accuracy Values of RapidMiner Classifying Algorithms

| <b>Classifying Algorithms</b> | <b>Accuracy (%)</b> |
|-------------------------------|---------------------|
| <b>Decision Trees</b>         | 91.90               |
| <b>NaiveBayes</b>             | 95.30               |
| <b>ANN</b>                    | 97.76               |
| <b>SVM</b>                    | 97.98               |
| <b>LR</b>                     | 98.65               |
| <b>MLP</b>                    | 99.10               |

When Table 5 is examined, the algorithm which has the lowest accuracy value is the Decision Trees algorithm with a success percentage of 91.90%. Bayes, 95,30%, ANN, 97,76%, have shown higher

success than the decision trees algorithm. While the SVM and LR algorithms follow the ranking, the MLP algorithm has been the best algorithm with a success percentage of 99,10%.

#### 4. Results and Discussion

In this study, the model created with the data mining programs WEKA and RapidMiner using the diabetes data that we have pre-processed has been applied. In the model applied, firstly, the data have been separated into two different clusters (the diabetic and the non-diabetic) with K-Means clustering algorithm. After this clustering process, clustering operation has been realized on the remaining data by deleting the data whose classes are mispredicted, in other words, mis-clustered. The Support Vector Machines, Naive Bayes, Decision Trees, Artificial Neural Networks, Multilayer Perceptron and Logistic Regression algorithms out of the classifying algorithms in the RapidMiner and WEKA programs have been used in the process of classifying. And the success rates of these algorithms have been compared.

When the results are evaluated in general, the success percentages for RapidMiner and WEKA are between 91-100%. This has demonstrated that the success of the applied model is very good. Both of the programs possess very high success rates for all of the algorithms. When the comparison is done in terms of the algorithms, the Multilayer Perceptron algorithm has been the best algorithm with the highest success percentage in both of the programs; the Decision Trees algorithm has been the algorithm whose success percentage is the lowest in both. When we compare the results in terms of data mining, except the decision trees algorithm, close but better results have been obtained in the RapidMiner program.

In this article study in which “Hybrid Prediction Model for Type 2 diabetic patients” [14] makes the prediction of morbidity or non-morbidity in different situations analyzing the data of the diabetic patients, the WEKA data mining program has been used; and the data clustered with K-Means have been classified with the decision trees algorithm after mis-clustered 192 data are deleted; and the classifying success percentage has been found to be 92.38%. In the study we have conducted, the decision trees algorithm have been found to be higher than the result obtained in this article with an accuracy percentage of 93.50% in the WEKA program.

Doctors can be provided convenience in the progress of the disease and the treatment for patients for whom prediction of disease is made by making morbidity or non-morbidity prediction of the diabetes which is seen on many people in the world and in our country beforehand. Or it can be procured for the patients to live more comfortably with this disease or to be completely cured by trying some other treatment methods and taking some other precautions.

#### References

- AlJarullah, AA., “Decision Tree Discovery for the Diagnosis of Type II Diabetes”, International Conference on Innovations in Information Technology 2011
- Beh Boon, C.; Jafri, M. Z. M.; Lim Hwee, S.: "Mangrove Mapping in Penang Island by Using Artificial Neural Network Technique", Open Systems (ICOS), 2011 IEEE Conference on, (2011) 245-249.
- Coşkun, C., Baykal, A.,”Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması”, Akademik Bilişim 11, İnönü Üniversitesi, <http://ab.org.tr/ab11/bildiri/67.pdf>, (accessed: 31.05.2012)
- Delen, D., Walker, G., Kadam, A., “Predicting breast cancer survivability: a comparison of three data mining methods”, Artificial Intelligence in Medicine, vol 34, June 2005, pp113-127



Dener, M., Dörterler, M., Orman, A., "Açık Kaynak Kodlu Veri Madenciliği Programları: WEKA'da Örnek Uygulama", Akademik Bilişim'09 - XI. Akademik Bilişim Konferansı Bildirileri, 11-13 Şubat 2009 Harran Üniversitesi, Şanlıurfa.

"Harrison İç Hastalıkları Cilt 2", Bölüm 333 Diabetes Mellitus Syf: 2109-2137, (2004)

<http://archive.ics.uci.edu/ml/datasets> (accessed: 10.11.2012)

<http://www.temd.org.tr/> (accessed: 09.11.2012)

International Diabetes Federation, World Diabetes Foundation. Diabetes Atlas. 2nd Ed. Brussels: International Diabetes Federation Publ.; 2003

Jiangtao, R.; Sau Dan, L.; Xianlu, C.; Ben, K.; Cheng, R.; Cheung, D.: "Naive Bayes Classification of Uncertain Data", Data Mining, 2009. ICDM '09. Ninth IEEE International Conference on, (2009) 944-949.

Karegowda, AG., Jayaram, MA., Manjunath, AS., "Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-3, February 2012

Karthikeyani, V., Begum, IP., Tajudin, K., Begam, IS., "Comparative of Data Mining Classification Algorithm (CDMCA) in Diabetes. Disease Prediction", International Journal of Computer Applications (0975 – 8887) Volume 60– No.12, December 2012 Lakso M.: Epidemiology And Diagnosis Of Type 2 Diabetes. In: Goldstein BJ Müller Wieland D, Eds. Testbook Of Type 2 Diabetes. New York: Martin Dunitz Taylor&Francs Group; 2003. P. 1-12.

Menon Krishna, A., "Large-Scale Support Vector Machines: Algorithms and Theory", Research Exam, University of California, San Diego, 2009

Patil, M.B.; Joshi, R.C.; Toshniwal, D.: "Hybrid Prediction Model For Type-2 Diabetic Patients", Expert Systems with Applications 37 (2010), 8102–8108,

Pickup, J. C., & Williams, G. (2003). Textbook of diabetes (3rd ed.). Oxford: Blackwell Science.

Quinlan, J. R. (1993), C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, California.

Vijayarani, S., Sudha, S., "Disease Prediction in Data Mining Technique – A Survey", International Journal of Computer Applications & Information Technology Vol. II, Issue I, January 2013 (ISSN: 2278-7720)

Williams R.: Public Health Problems Of Diabetes Mellitus And Its Cost To The Community. In: Williams G, Pickup J, Eds. Textbook Of Diabetes. 2nd Ed. Vol. I. Cambridge: Blackwell Sci; 1997. Ch. 4.1-4.4.