A comparative study on the pre-processing and mining of Pima Indian Diabetes Dataset

Amatul Zehra[1], Tuty Asmawaty[1], M.A M. Aznan[2]

[1]Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang, Kuantan, Pahang 26300, Malaysia
tuty@ump.edu.my
[2]Kulliyah of Medicine, International Islamic University Malaysia, P.O Box 141, Kuantan, Pahang 25710, Malaysia
aznan@iium.edu.my

Abstract. Data mining in medical data has successfully converted raw data into useful information. This information helps the medical experts in improving the diagnosis and treatment of diseases. In this paper, we review studied data mining applications applied exclusively on an open source diabetes dataset. Type II Diabetes Mellitus is one of the silent killer diseases worldwide. According to the World Health Organization, 346 million people are suffering from diabetes worldwide. Diagnosis or prediction of diabetes is done through various data mining techniques such as association, classification, clustering and pattern recognition. The study led to the related open issues of identifying the need of a relation between the major factors that lead to the development of diabetes. This is possible by mining patterns found between the independent and dependant variables in the dataset. This paper compares the classification accuracies of non-processed and pre-processed data. The results clearly show that the pre-processed data gives better classification accuracy.

Keywords: Diabetes prediction; Type II Diabetes Mellitus; Data Mining; Data pre-processing

1 Introduction

Diabetes Mellitus has become a common health problem nowadays, which would affect people and lead to various disablements like cardio vascular disease, visual impairments, leg amputation and renal failure if diagnosis is not done in the right time [1]. Diabetes can affect people due to the lack of insulin in the blood. Insulin is a natural hormone secreted by the pancreas, which acts as a key to unlock the body cells so that sugar, starch and food molecules can be absorbed and hence be utilized by the cells to generate energy required for daily life. Insulin deficiency is due to either of the two conditions. First is when the pancreas does not produce insulin at all. This leads to type I diabetes mellitus (T1DM) which is usually found by birth. Second state is when the body does not respond correctly to the insulin produced by the pancreas and hence the glucose that is consumed by the person is locked inside the

blood instead of entering into the cells of the body. This ineffective insulin leads to type II diabetes mellitus (T2DM). Among these, type I diabetes is usually diagnosed in children and type II is the most common form which affects adults [2].

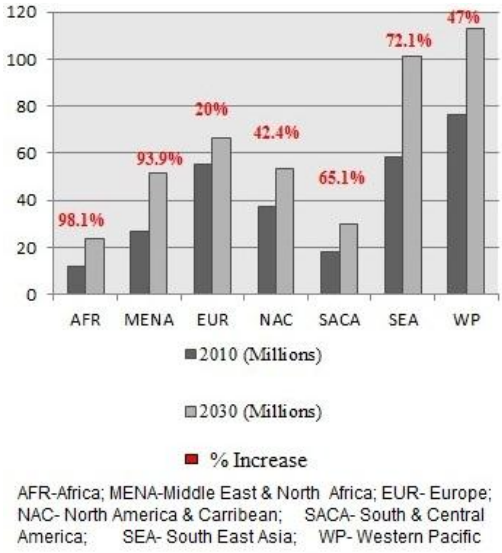

Fig. 1. Region-wise estimated rise in diabetics by 2030 (Diabetes Atlas 4th Edition, International Diabetes Federation)

1.1 Diabetes-A Global Threat

The International Diabetes Federation has estimated an alarming rise in the number of diabetics by the year 2030, Fig. 1 [3]. A sharp rise in diabetics has been observed in Asian region with 138 million Asians including 14.9% Malaysians [4]. From 1996 to 2006, the number of diabetics in Malaysia had increased by almost 80% and reached to 1.4 million adults above the age of 30. Among those, almost 36% were undiagnosed, resulting in complications that required more intensive medical care, putting great strain on the existing overstretched health services [5].

This paper focuses to investigate the possible solutions for the group of people who are at a risk of developing type II diabetes in future. We aimed to study type II diabetes because this type can be prevented by adopting proactive measures. We propose to design classifiers and develop a prediction model based on existing data. For this purpose, we intend to use Pima Indian diabetes dataset. Eventually, the model would be able to answer the need for significant and urgent requirement to: (i) stop sharp rise in diabetes, (ii) grow public health awareness, and (iii) prevent the onset of this disease.

The paper is organized as follows: section 2 gives a brief overview on type II

diabetes followed by the review of the prediction and diagnostic models related to diabetes. Section 3 is the proposed study of this review paper and in the end is the conclusion.

2 Type II diabetes

Type II diabetes is sometimes called non-insulin dependent diabetes or adult-onset diabetes [3]. At least 90% of all cases of diabetes are victims of this type. It strikes a person due to insulin resistance and relative insulin deficiency, either of which may be present at the time that diabetes becomes clinically evident. The diagnosis of type II diabetes usually occurs after the age of 40 but can occur earlier, especially in populations with high diabetes prevalence. Type II diabetes can remain undetected for many years and the diagnosis is often made from associated complications or incidentally through an abnormal blood or urine glucose test. It is often, but not always, associated with obesity, which itself can cause insulin resistance and lead to elevated blood glucose levels.

The normal range of fasting blood glucose level is between 4.0-5.6 mmol/L. After consuming a meal, the blood glucose level rises in the blood and can reach up to 7.8mmol/L. Any value higher than these ranges indicates the prevalence of diabetes. After two hours of having a meal, the blood glucose level drops again, Figure 1(b) [6]. There is also a condition called pre-diabetes. It is that state, where the blood glucose level is higher than the normal range but not high enough to be stated as diabetes.

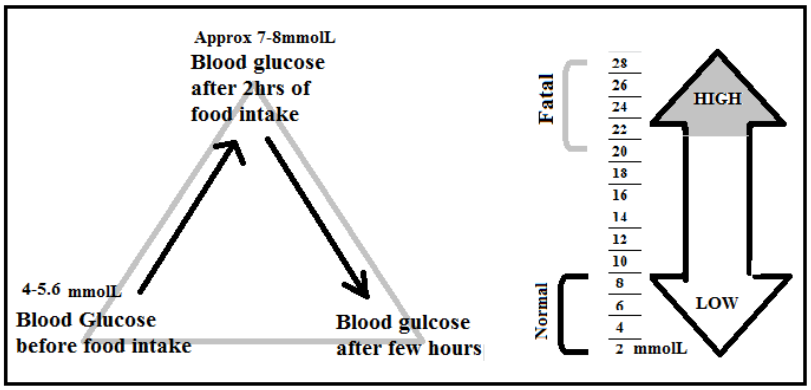

Fig. 2. Ranges of normal and high blood glucose levels (www.diabetes.co.uk)

Individuals can be categorized into three groups namely, ‘healthy, ‘pre-diabetics’ and ‘diabetics’. Blood glucose levels for all three categories vary accordingly. Table 1 shows the normal and post-meal ranges of blood glucose levels for these groups.

Table 1. Fasting and post meal normal ranges of blood glucose. (www.diabetes.co.uk)

Individuals	Fasting	2 hours after a meal
Healthy	4.0 - 5.6 mmol/L	<7.8 mmol/L
Pre-diabetics	5.6 - 7.0 mmol/L	<7.8 mmol/L
Diabetics	> 7 mmol/L	≥ 7.8 mmol/L

The need for avoidance and better management of type II diabetes has been an important issue since ages. Medical practitioners and researchers have investigated and continue to find solutions to overcome this disease. Various researches and studies are done on predicting the blood glucose levels for type II diabetes patients for a short term. Most of the predictions helped to decide the diet control and physical activities in order to maintain a healthy life [7].

3 Review of the type II diabetes prediction and diagnosis models

Due to rising cost of health care, it is useful to assist patients to control diabetes by themselves. In many instances, early information related to diabetes might help in avoidance, curing and appropriate treatment of the disease. Many computer programs or systems were developed and are being developed by emulating human intelligence that could be used to assist the users or patients in managing diabetes [8].We assessed different systems such as artificial intelligence systems, mobile phone applications and specially designed devices for the prediction and diagnosis of diabetes. The focus of this paper is to investigate for a model to predict and diagnose diabetes in the long run. Most of the models have been developed to diagnose diabetes and predict the blood sugar level for a short term. However, according to the authors' knowledge, there are rarely any systems developed to predict the onset of diabetes in the long run. In the next section, a brief review on all related systems is done.

3.1 Data mining applications and the Pima Indian Diabetes Dataset (PIDD)

There are several studies found in the literature that have used various techniques on the Pima Indian Diabetes dataset to train and test data. This paper focuses only on the data mining techniques used for classification on the same dataset. The National Institute of Diabetes and Digestive and Kidney Diseases of the NIH originally owned the Pima Indian Diabetes Database (PIDD) [9].

The database has n=768 patients each with 9 numeric variables. The data refers to females of ages from 21 to 81. Out of the nine condition attributes, six attributes describe the result of physical examination, rest of the attributes are of chemical examinations. The independent or target variable is the class variable (diabetes = 1 (yes), diabetes =0 (no)), represented by the 9^{th} variable. The attributes are:

1. number of times pregnant.

2. 2-hour OGTT plasma glucose.
3. diastolic blood pressure
4. triceps skin fold thickness
5. 2-hour serum insulin
6. BMI
7. diabetes pedigree function
8. age
9. class variable(0,1)

The aim is to use the first 8 variables to predict the value of the 9th variable (diabetes=yes (1) diabetes = no (0)).

Although the owners claim that this dataset does not have any missing values, it is found that there are many missing values. For that, we need to pre-process the data before using it. The data processing techniques, when applied prior to mining, can considerably improve the overall quality of the patterns mined and/or the time required for the actual mining. Data preprocessing is a significant step in the knowledge discovery process, since quality decisions must be based on quality data. In the 768 cases of the Pima Indian Diabetes Dataset (PIDD), 5 patients had a glucose of 0, 11 patients had a body mass index of 0, 28 others had a diastolic blood pressure of 0, 192 others had skin fold thickness readings of 0, and 140 others had serum insulin levels of 0 which is physically impossible. After deleting these cases there were 392 cases with no missing values (130 tested positive cases and 262 tested negative).

A brief review on the various data mining techniques used previously on the PIDD is shown in Table 2.

Ilango et al. proposed a Hybrid Prediction Model with F-score feature selection approach to identify the optimal feature subset of the Pima Indians Diabetes dataset [10]. The features of diabetes dataset are ranked using F-score and the feature subset that gave the minimal clustering error was the optimal feature subset of the dataset. The correctly classified instances determined the pattern for diagnosis and were used for further classification process. The improved performance of the Support Vector Machine classifier measured in terms of Accuracy of the classifier, Sensitivity, Specificity and Area Under Curve (AUC) proved that the proposed feature approach improved the performance of classification. The proposed prediction model achieved a predictive accuracy of 98.9427.

Bushra M. Hussan preprocessed the PIDD successfully by supplying missing values using the KNN mutation, then clustered using K-means with k value equal to 2 [11]. The first result of algorithm execution on the original data showed accuracy of 81%. Later the data was further improved by preprocessing process and then applied the algorithm again which gave an accuracy of 94%. Furthermore they applied the algorithm on new instances (almost 700 records), they got the accuracy of 97%.

Table 2. List of data mining techniques used on PIDD and their accuracy levels.

	Technique applied	Accuracy %
1	F-score Feature Selection, k-means Clustering and SVM	98.94
2	K-means algorithm	97
3	Cascading K-means Clustering and K-Nearest Neighbor Classifier	96.67
4	b-Colouring Technique in Clustering Analysis	93.67
5	Feature Weighted Support Vector Machines and Modified Cuckoo Search	93.58
6	Cascaded K-Means and Decision Tree C4.5	93.33
7	Rough sets	82.6
8	Prediction Model Discovery Using RapidMiner	80
9	Ensemble model (SVM, Discriminant analysis and Bayesian Network)	76.03
10	Neural Network and Fuzzy k-Nearest Neighbor Algorithm	74.32

Karegowda et al. proposed a model that consisted of three stages [12]. In the first stage, K-means clustering was used to identify and eliminate incorrectly classified instances. In the second stage Genetic algorithm (GA) and Correlation based feature selection (CFS) was used in a cascaded fashion for relevant feature extraction, where GA rendered global search of attributes with fitness evaluation effected by CFS. Finally in the third stage a fine tuned classification was done using K-nearest neighbor (KNN) by taking the correctly clustered instance of first stage and with feature subset identified in the second stage as inputs for the KNN. Experimental results showed the cascaded K-means clustering and KNN along with feature subset identified GA_CFS enhanced classification accuracy of KNN. The proposed model obtained the classification accuracy of 96.68% for the PIDD.

Vijayalakshmi et al. developed a clustering algorithm used for predicting diabetes based on graph b-colouring technique [13]. They implemented, performed experiments, and compared their approach with KNN Classification and K-means clustering. The results showed that the clustering based on graph colouring outperforms the other clustering approaches in terms of accuracy and purity. The proposed technique presented a real representation of clusters by dominant objects that assures the inter cluster disparity in a partitioning and used to evaluate the quality of clusters.

Giveki et al. proposed a model that consisted of three stages [14]. Firstly, Principal Component Analysis (PCA) is applied to select an optimal subset of features out of set of all the features. Secondly, Mutual Information is employed to construct the Feature Weight Support Vector Machine by weighing different features based on their degree of importance. Finally, classification accuracy of SVMs, MCS is applied to select the best parameter values. The proposed MI-MCS-FWSVM method obtains 93.58% accuracy on the PIDD.

Jayaram et al. presented the development of a hybrid model for classifying Pima Indian diabetic database (PIDD) [15]. The model consisted of two stages. In the first stage, the K-means clustering was used to identify and eliminate incorrectly classified

instances. The continuous data was converted to categorical form by approximate width of the desired intervals, based on the opinion of medical expert. In the second stage a fine tuned classification was done using Decision tree C4.5 by taking the correctly clustered instance of first stage. Experimental results signify that cascaded K-means clustering and Decision tree C4.5 has enhanced classification accuracy of C4.5. Further rules generated using cascaded C4.5 tree with categorical data are less in numbers and easy to interpret compared to rules generated with C4.5 alone with continuous data. The proposed cascaded model with categorical data obtained the classification accuracy of 93.33 % when compared to accuracy of 73.62 % using C4.5 alone for PIMA Indian diabetic dataset.

Breault proposed the idea of using rough sets on the PIDD for the first time [16]. He first pre-processed the data and discretized it by making intervals of data. He used the equal frequency binning criteria for the same purpose. Then he created reducts by using Johnson reducer algorithm and classified using the batch classifier with the standard/tuned voting method (RSES). The rules were constructed for each of the 10 randomizations of the PIDD training sets from above. The test sets were classified according to defaults of the naïve Bayes classifier, and the 10 accuracies ranged from 69.6% to 85.5% with a mean of 73.8% and a 95% CI of (71.3%, 76.3%).

Han et al. used data mining technique through RapidMiner for diabetes data analysis and diabetes prediction model [17]. A decision tree was used for prediction with 72 % of accuracy. ID3 Algorithm was also used for this purpose which gave 80 % accurate results.

An ensemble model of three classifiers is used on the PIDD by Pujari et al [18]. Classification performance of SVM (support vector machine), discriminant analysis and Bayesian network was investigated individually with the help of gain chart and response chart for both training and testing set. Results indicated that the ensemble model achieved an accuracy of 76.03% on test data set.

Pradhan et al. proposed a model based on Neural Network and Fuzzy k-Nearest Neighbor Algorithm [19]. They first pre-processed the data by eliminating the records containing the missing values from the Pima Indian Diabetes Dataset. The Fuzzy k-Nearest Neighbor algorithm is used to train the Neural Networks. Finally, the entire training set is used as test set to calculate the classification accuracy.

4 Comparison of non-processed data and pre-processed data

As seen in the review above, many researches are done on the prediction and diagnosis of diabetes [20,21]. After the literature survey, we have found that most of the data mining techniques applied on the PIDD were pre-processed. The PIDD has 8 attributes out of which a few attributes contain values that are out of the normal range. Also there are many missing values i.e. the value is 0 instead of an actual value. Therefore, the pre-processing of data is necessary for efficient data mining of patterns in the PIDD.

In this section we compare the accuracy of classification on the PIDD when the data is not pre-processed and when it is pre-processed. For this purpose, we have used the WEKA tool. The full form of WEKA is Waikato Environment for Knowledge

Learning. Weka is an open source software that was developed by the students of the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains [21]. Data preprocessing, classification, clustering, association, regression and feature selection types of standard data mining tasks are supported by Weka.

4.1 Data Preprocessing

To improve the quality of the results obtained after mining and the effectiveness of the complete mining process, data preprocessing is done [10]. Researchers and practitioners realize that in order to use data mining tools on the database effectively, data preprocessing is essential for successful data mining. After observing the Pima Indians Diabetes dataset, we found the need to pre-process the data in two steps.
Firstly, it is seen that the dataset has the value zero for missing data. We removed all the instances which had the value zero for a particular field where having a zero as a value was impossible. Therefore, the instances which have missing values were eliminated.
Next, we did the process of data discretization. Data discretization is defined as a process of converting continuous data attribute values into a finite set of intervals and associating with each interval some specific data value [22]. There are no restrictions on discrete values associated with a given data interval except that these values must induce some ordering on the discretized attribute domain. Discretization significantly improves the quality of discovered knowledge and also reduces the running time of various data mining tasks such as association rule discovery, classification, and prediction.

4.2 Classification accuracy on non-processed data

We have used the same dataset for the comparison. We have chosen five classification techniques to compare. The following Table 3 shows the accuracy results on non-processed data for five techniques.

Table 3. Classification accuracies of five classifiers applied on non-processed data

Technique	Correctly classified	Incorrectly classified
Naïve Bayes	76.3%	23.69%
Multilayer Perceptron	75.39%	24.6%
Decision Table	71.22%	28.77%
J48	73.82%	26.17%
Simple Cart	75.13%	24.86%

Furthermore, we used the same set of classifiers on the pre-processed data. The following Table 4 shows the accuracy results

Table 4. Classification accuracies of five classifiers applied on pre-processed data

Technique	Correctly classified	Incorrectly classified
Naïve Bayes	80.3%	19.69%
Multilayer Perceptron	81%	18%
Decision Table	85.2%	14.79%
J48	80%	19%
Simple Cart	79.6%	20.39%

5 Results

This paper focused on the importance of data pre-processing for data mining. We used the PIMA Indian Diabetes Dataset for the study. The data was first classified without pre-processing it and the results were noted. Then the same set of data was pre-processed that is the removal of missing values and data discretization. Classification was done after the two step process of data pre-processing. After the comparison between the accuracies of classification on non-processed and pre-processed data, it showed that the classification accuracy increases when the data is pre-processed. Hence the data mining accuracy depends a lot on the pre-processing of data.

6 Conclusion

In this paper, various investigations on prediction and diagnosis of type II diabetes mellitus using data mining techniques are present. Various classification techniques are used after pre-processing of the data in PIDD. In this paper we have done a comparison of the accuracy of classification done on non-processed and pre-processed data. We have come to a conclusion that the pre-processed data gives us better accuracy results rather than non-processed data. This shows the importance of pre-processing in the data mining techniques

References

1. World Health Organization, http://www.who.int/topics/diabetes_mellitus/en/, (Last access date: 30th September 2012)
2. Pobi, S., Hall, LO.: Predicting juvenile diabetes from clinical test results. In: 2006 International Joint Conference on Neural Networks (IJCNN), pp. 2159 – 65 (2006)
3. International Diabetes Federation, http://www.idf.org/diabetesatlas/5e/regional-overviews, (Last access date: 30th September 2012)

4. 'Sharp rise of diabetics in Asia', The Malay Mail, 3rd December 2010, (Last access date: 13th October 2011).
5. 'Alarming rise in number of diabetics in Malaysia', The Star, 11th January 2010, (Last access date: 12th October 2011).
6. Normal and diabetic blood sugar level ranges. URL: www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html, (Last access date: 14th October 2011
7. Lauritzen, J.N., Arsand, E., Vuurden, K.V., Bellika, J.G., Hejlesen, O.K., Hartvigsen, G.: Towards a mobile solution for predicting illness in type 1 diabetes mellitus: Development of a prediction model for detecting risk of illness in type 1 diabetes prior to symptom onset. In: 2011 2nd International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronics Systems Technology (Wireless VITAE), pp 1-5 (2011)
8. Barakat, N.H., Bradley, A.P., Barakat, M.N.H.: Intelligible support vector machines for diagnosis of diabetes mellitus. In: IEEE transaction on information technology in Biomedicine. 14:4 (2010)
9. Pima Indian Diabetes Database, Url: www.ics.uci.edu/~mlearn/MLRepository.html, Access on 24th June 2013.
10. Ilango, B.S., Ramaraj, N.: Hybrid Prediction Model with F-score Feature Selection for Type II Diabetes Databases. A2CWiC 2010, September 16-17, India (2010)
11. Hussan, B.M.. Data Mining based Prediction of Medical data Using K-means algorithm. Basrah Journal of Science (A). Vol.30(1),pp. 46-56 (2012)
12. Karegowda, A.G., Jayaram, M.A., Manjunath, A.S.: Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients. In: International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-3, February (2012)
13. Vijayalakshmi, D., Thilagavathi, K.: An Approach for Prediction of Diabetic Disease by Using b-Colouring Technique in Clustering Analysis. In: International Journal of Applied Mathematical Research, 1 (4) pp. 520-530 Science Publishing Corporation www.sciencepubco.com/index.php/IJAMR (2012)
14. Giveki, D., Salimi, H., Bahmanyar, G.R., Khademian, Y.: Automatic Detection of Diabetes Diagnosis using Feature Weighted Support Vector Machines based on Mutual Information and Modified Cuckoo Search
15. Karegowda, A.G., Punya, V., Jayaram, M.A., Manjunath, A.S.: Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4.5. International Journal of Computer Applications. 45–12, (2012)
16. Breault, J.L.: Data Mining Diabetic Databases: Are Rough Sets a Useful Addition?
17. Han, J., Rodriguze, J.C., Beheshti, M.: Diabetes data analysis and prediction model discovery using RapidMiner. Second International Conference on Future Generation Communication and Networking.96-9 (2008)
18. Pujari, P., Vishwavidyalaya, G.G.: Ensemble Data Mining Model for Classification of Pima Indian Diabetes Data set.
19. Pradhan, M., Sahu, R.K.: Predict the onset of diabetes disease using Artificial Neural Network. Intl J Comp Sci & Emerging Tech. 2:303-11 (2011)
20. Gani, A., Gribok, A.V., Lu, Y., Ward, W.K., Vigersky, R.A., Reifman, J.: Universal glucose models for predicting subcutaneous glucose concentration in Humans. Proceedings of the IEEE Transactions on Information Technology in Biomedicine. 14: 157-65 (2010)
21. Sparacino, G., Zanderigo, F., Corazza, S., Maran, A., Facchinetti, A., Cobelli, C.: Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series. IEEE transactions on biomedical engineering. 54:931-37 (2007)
22. Devi, R., Khemchandani, V.: Application of Data Mining Techniques For Diabetic DataSet. In: Computing For Nation Development (2010)