Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (3): 416-422, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.416.422

Application of Classification based Data Mining Technique in Diabetes Care

Abdullah A. Aljumah, Mohammad Khubeb Siddiqui and Mohammad Gulam Ahamad
College of Computer, Engineering and Sciences, Salman bin Abdulaziz University,
Alkharj-11942, P. Box 151, Kingdom of Saudi Arabia

Abstract: The present research work relates data mining to medical informatics. The proposed work shows various models for each type of diabetic intervention and analysis is carried out using classification based data mining technique. The Area Under Curve (AUC) of ROC (Receiving Operating Characteristics) plots are calculated, the confusion matrix is formed, through which accuracy and cost of interventions have been evaluated. The AUC of ROC for all six modes of diabetic interventions are obtained and have been distinguished which mode of intervention is more appropriate. The accuracy and AUC of the model depend on the cost of model which is always inversely proportional to the cost of the model. Present analysis predicts that smoking cessation is the best intervention followed by exercise, diet, weight and drug for the diabetic control. Therefore, the results are quite impressive in predicting the diabetic intervention control resulting in high AUC of ROC and high accuracy with lowest cost.

Key word: Data mining, classification, ROC, AUC, confusion matrix, diabetes

INTRODUCTION

The latest survey reveals that the diabetic patients have been dramatically increased in Saudi Arabia and reports show that it is in the epidemic form in the industrialized and developed countries. The diabetes is an endocrine disease and classified into two categories: type1 (insulin dependent) and 'type 2' (non-insulin dependent) (Gan, 2003).

The important aspect of controlling diabetes is that the heightened blood glucose levels are to be brought to normal levels with following further leading to hyperglycaemia. As mentioned by WHO (2005) NCD report of Ministry of Health, Saudi Arabia, (WHO, NCD risk factor, standard report of Ministry of Health, Saudi Arabia, following are the six types of diabetic interventions discussed below:

- Drug
- Diet
- Weight reduction
- Smoke cessation
- Exercise
- Insulin

The literature survey reveals many results on diabetes, the diabetic data warehouse were formed a large integrated health care system in the New Orleans area of USA with 30,383 diabetic patients. They used the classification and regression tree approach to analyze the data sets (Breault *et al*., 2002). The diabetes in Saudi Arabia had been investigated and found that the overall prevalence of diabetes adults in KSA is 23.7%. They further recommend a longitudinal study to demonstrate the importance of modifying risk factors for the development of diabetes and reducing its prevalence in KSA (Al-Nozha *et al*, 2004). The factors contribute to improvement in glycemic control, published in 2001-2005. Data mining indicated that intensity of education did not predict changes in HbA1c levels (Sigurdardottir *et al*., 2007). The insulin therapy and episodes of severe hypoglycaemia in preschool children were investigated (Yokotaa *et al*., 2005).The therapeutic management and control of diabetes with cardiovascular modification in case of type diabetes in France had been studied. The study was proposed to 575 diabetologists across France. However, both awareness and application of published recommendations need to be reinforced (Charpentier *et al*., 2003). The diabetic control study in the Western Pacific Region to identify factors associated with glycemic control and hypoglycemia. A cross-sectional clinic based study on 2312 children and adolescents (aged b18 years; 45% males) from 96 paediatric diabetes centres in Australia, China, Hong Kong, Indonesia, Japan, Malaysia, Philippines, Singapore, South Korea, Taiwan and Thailand was conducted. Clinical and management

Corresponding Author: Mohammad Khubeb Siddiqui College of Computer, Engineering and Sciences, Salman bin Abdulaziz University, Alkharj-11942, P. Box 151, Kingdom of Saudi Arabia, Tel: +966547006246

details were recorded and finger-pricked blood samples were obtained for central glycated hemoglobin (HbA1c) (Craig *et al*., 2006). The prevalence of diabetes mellitus and islet auto antibodies in an adult population from Southern Spain and the prevalence of Type 2 diabetes and LADA are high in the south of Spain (Soriguer-Escofet *et al*., 2002). Diabetes due to specific mechanisms and diseases is divided into two subgroups; diabetes in which genetic susceptibility is clarified at the DNA level and diabetes associated with other diseases or conditions (Richards *et al*., 2001). The data of Smoking habits were reported to the Swedish National Diabetes Register (NDR) the trend in the proportion of smoking in diabetes and to study associations between smoking, glycaemic control and micro albuminuria and their studies concluded that Smoking in patients with diabetes was widespread, especially in young female type 1 and in middle-aged type 1 and 2 diabetes patients and should be the target for smoking cessation campaigns. Smoking was associated with both poor glycaemic control and microalbuminuria, independently of other study characteristics (Nilsson *et al*., 2004). Investigated on weight loss goal among participants enrolled in an adapted Diabetes Prevention Program (DPP), findings highlight the importance of supporting participants in lifestyle interventions to initiate and maintain dietary self-monitoring and increased levels of physical activity (Harwell *et al*., 2011).

Application of data mining in health care is an attracting field of research. Proposed research work provides an overview of this emerging field, clarifying how data mining techniques is applicable on healthcare analysis to predict the mode of diabetic intervention control. In the last few years, the 'data mining' technique has been increasingly used in the medical and clinical diagnostics. In general, data mining is the analysis of observation data sets to find unsuspected relationship and we will summarize the data in novel ways that are understandable and useful to the common man and medical fraternity. In this paper, we give a methodological review of data mining, focusing on diabetic data analysis process using classification technique and highlighting ROC plots and formation of confusion matrix to give the performance measures of related to best mode of diabetic intervention control.

MATERIALS AND METHODS

WHO's 2005 data collection: The dataset was collected from World Health Organisation' s (WHO), NCD risk factor, a standard report of Ministry of Health, Saudi Arabia, 2005 (WHO, 2005). This is publicly available dataset on the web portal of WHO (http://www.emro.who.int/ncd/pdf/stepwise_saa_05.pdf)

Process of model building: The approach of process of model building is carried out in data mining technique the steps of generation of model are described in the process model as given in the Fig. 1 data mining process of model building.

Database design: The database was designed in oracle 10g database. The data of all six tables 'drug', 'diet', 'weight', 'smoke_cessation', 'exercise' and 'insulin' interventions had been designed and merged into a single table named 'diab_treatment' in oracle 10 g database.

Connecting server to client: In present case the server is Oracle database 10 g and connection is established with Oracle data miner 10 g. To establish connection it requires certain important privileges from server side (SYSDBA) to designed schema.

Building model: The oracle data miner applies the mechanism of building, testing and applying a model in order to find the mode of diabetic intervention control.

Data mining technique-classification: Classification function predicts class membership for categorical target. The algorithm used in classification technique is 'Support Vector Machine'.

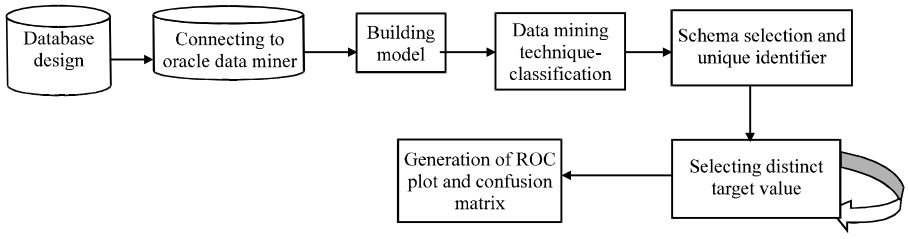

Fig. 1: Data mining process of model building

Schema selection and unique identifier: After selecting data mining classification technique and the algorithm 'support vector machine', the tool requires to select the schema, table and unique identifier of that Table.

Selecting distinct target value: Selecting target column of table, after selection of target column, it shows the distinct values of that column. It's a manual iterative process that is to be applied for each distinct value of the target column.

Generation of ROC plots and confusion matrix: The final stage states the ROC graph, confusion matrix, cost, accuracy, area under curve for each and every distinct target value 'intervention'.

Classification: The data mining process has various techniques. Classification technique is one of the prominent techniques to predict the character type target attribute in dataset. Classification of a collection consists of dividing the items that make up the collection into categories or classes. In the context of data mining, classification is done using a model that is built on historical data. The goal of predictive classification is to accurately predict the target class for each record in new data.

A classification task begins with build data (also known as training data) for which the target values (or class assignments) are known. Different classification algorithms uses different techniques for finding relations between the predictor attribute values and the target attribute values in the build data, which were summarized in model. A classification model can also be applied to data that was held aside from the training data to compare the predictions to the known target values; such data is also known as test data or evaluation data. The comparison technique is called testing a model, which measures the model's predictive accuracy. The application of a classification model to new data is called applying the model and the data is called apply data or scoring data.

While applying the classification mining techniques on ODM needs the target attribute the target should be discrete value e.g. 0 or 1, male or female etc.

In present case the target attribute is 'treatment' attribute that includes the value of mode of interventions that acts as discrete value, means the distinct value of this column shows the all six mode of diabetic interventions.

Algorithm for classification technique: ODM provides the several algorithms for classification and here Support Vector Machine (SVM) had been adopted.

ROC: ROC 'Receiver Operating Characteristics' is the powerful statistical technique used in signal detection in radio frequency band. Nowadays this has been a marvellous technique in medical data mining under classification analysis to classify the medical data. The ROC has long been used in signal detection theory to depict the trade off between hit rate and false alarm rate of classifiers (Egan, 1975). It allows us to visualize the trade-off between the rate at which the model can accurately recognize 'yes' cases versus the rate at which it mistakenly identifies 'no' cases as 'yes' for different portions of the test set. Any increase in the true positive rate occurs at the cost of an increase in the false positive rate. The area under curve of ROC graph is measurement of accuracy of the model.

The ROC curve allows us to explore the relationship between the sensitivity and specificity of a clinical test for a variety of different thresholds points, thus allowing the determination of an optimal value. It is often a test is to be carried out, which provides a result on a continuous measure. The vertical and horizontal axis of ROC curve represents the true positive rate and false positive rate, respectively.

Area under curve (AUC): The area under curve measures the discriminating ability of a distinct classification model. The prediction can be based on AUC value. If the AUC value is larger the probability of positivity of the case is more compare to negativity of the case. The AUC is a portion of the area of the unit square its value always lies between 0 and 1.0.

The classification accuracy of ROC is indicated below in grade point system:

- 1.0: perfect prediction
- 0.9: excellent prediction
- 0.8: good prediction
- 0.7: mediocre prediction
- 0.6: poor prediction
- 0.5: random prediction
- <0.5: wrong prediction

EXPERIMENTAL ANALYSIS: CLASSIFIER PERFORMANCE

In our experiment we had used the oracle data miner version 10.2.0.3.0.1; build 2007 for the mining activity and analysis of data. The ODM acts as a client and oracle 10 g database release 10.2.0.3.0 acts as a server (http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/downloads/index.htm).

The ODM version 10.1 is data mining software embedded in the Oracle 10 g Database Enterprise Edition (EE) that enables us to discover new insights hidden in data. The main advantage of using ODM is that all data mining processing occurs within the Oracle database. Since ODM and Oracle database are compatible to each other, as they are the product of oracle corp. Other mining tools force us to extract the data from the database before the actual mining process. This may cause number of flaws in these mining tools. Another important aspect of ODM is that, we get secured and stable data management which enhance the productivity.

The diabetic interventions database is designed in Oracle10g. To establish the connection between oracle10g database and ODM tool it requires certain important privileges from SYSDBA (administrator) to the designed schema. The granted privileges successfully connect ODM tool to oracle10g database. The table designed in schema shows the diabetic interventions i.e. drug, diet, weight, smoke cession, exercise and insulin.

We had applied the classification based data mining technique. In our case ODM takes 'treatment' attribute from the table 'diab_treat' from oracle database as the target attribute. This is categorical attribute for prediction of the results and their performance of diabetes interventions with the help of ROC graphs and confusion matrix. The AUC of ROC analysis is obtained for the prediction of diabetic intervention. The confusion matrix consists of four cells which are defined as below:

- **TP** = **True positive:** If the instance is positive and it is classified as positive
- **TN** = **True negative:** If the instance is negative and it is classified as negative
- **FP** = **False positive:** If the instance is negative and it is classified as positive
- **FN** = **False negative:** If the instance is positive and it is classified as negative

For calculating the predictive accuracy of the model the formula as:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

Confusion matrix: The AUC works as efficiency measure of the unsupervised data produced by classification technique. During the testing we get correct and incorrect classification from each class. This result is formed as confusion matrix. A confusion matrix (Kohavi and Provost, 1998) contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix.

Table 1: Classified result of diabetes intervention

Intervention	Area under curve	Accuracy	Cost
Drug	0.50	0.83	5
Diet	0.84	0.83	5
Weight	0.83	0.86	4
Smoke cessation	0.91	0.90	3
Exercise	0.62	0.83	5
Insulin	0.77	0.86	4

Confusion matrix is a visualization tool typically used in supervised learning in unsupervised learning it is typically called a matching matrix. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. One benefit of a confusion matrix is that it is easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another). When a data set is unbalanced (when the number of samples in different classes vary greatly) the error rate of a classifier is not representative of the true performance of the classifier.

Cost: The Cost is an indication of the damage done by an incorrect prediction and is useful in comparing one model to another. Lower cost means better model (http://www.oracle.com/technetwork/indexes/downloads/index.html#database).

Below are the Confusion matrices for all six types of diabetic intervention. The ROC plots have been generated after classifying data using Support Vector Machine (SVM) algorithm.

Table 1 shows three columns named as interventions, AUC and cost. While calculating the accuracy of model cost plays the vital role. Lower cost means better model (http://www.oracle.com/technetwork /indexes/downloads/index.html#database). For each intervention we had designed different models and selecting among them the best model that satisfies the accuracy rate, where the cost is the key factor.

RESULTS AND DISCUSSION

The present research work relates data mining to medical informatics; it's one of the exploring applications of data mining. Various models have designed for each type of diabetic intervention using classification technique.

The ROC's of diabetic intervention which depicts excellent, good and worthless treatment tests on the different graph for six types of interventions. The accuracy of the treatment depends how well the treatments nature which classify the various modes of

treatment. Accuracy is measured by AUC of ROC curves. This is a statistical analysis for classifying the accuracy of the diabetic intervention which has been exhibited by point system described in the section 2.3.2. In the present investigation, ROC curves had been generated from clinical prediction rules using ODM. The ROC curves in the Fig. 2 to 7 are generated from the study of clinical findings, which predict the various diabetic interventions. The study compared the different diabetic interventions and found that smoking cessation is best (AUC = 0.9) compared to other interventions.

Confusion matrix for drug intervention

	Predicted class		
Actual class		Other	Drug
	Other	25	0
	Drug	5	0

Accuracy = TP+TN/(TP+FP+TN+FN)-(1)
Overall predictive accuracy = 25+0/(25+0+5+0) = 0.83

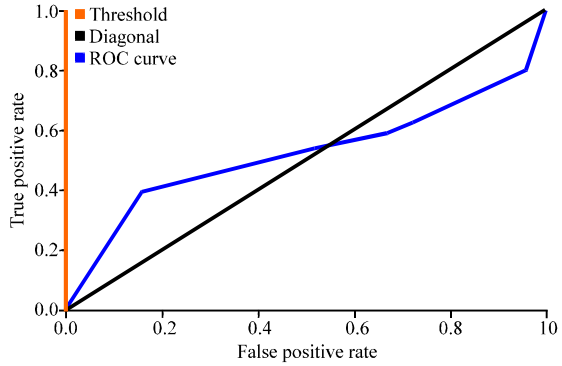

Fig. 2: ROC plot for drug intervention

Confusion matrix for diet intervention

	Predicted class		
Actual class		Other	Dite
	Other	24	1
	Diet	4	1

Overall predictive accuracy = 24+1/(24+1+4+1) = 0.83

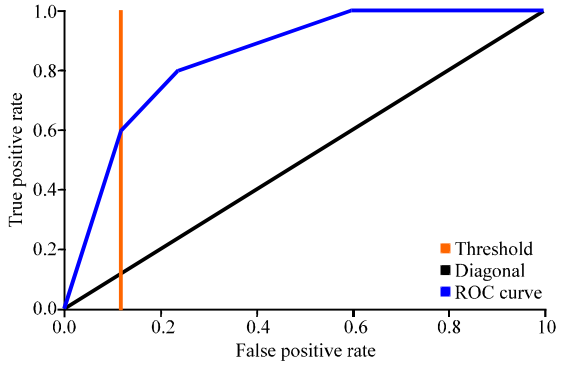

Fig. 3: ROC plot for diet intervention

From the experimental analysis, the cost of the model is found inversely proportional to the accuracy of the model. As we are aware that lesser is the cost of the model the better is the accuracy of the model. The cost of the model predicts the accuracy of the intervention. The Cost analysis of each intervention reveals different models. The best model is selected here on the basis of lowest cost of the model. Initially we would like to correlate AUC of each model, with the properties of AUC values. The AUC value for smoking cessation intervention is 0.9 indicates the excellent prediction. The AUC = 0.8 for

Confusion matrix for weight intervention

	Predicted class		
Actual class		Other	Weight
	Other	25	0
	Weight	4	1

Overall predictive accuracy = 25+1/ (25+0+4+1) = 0.86

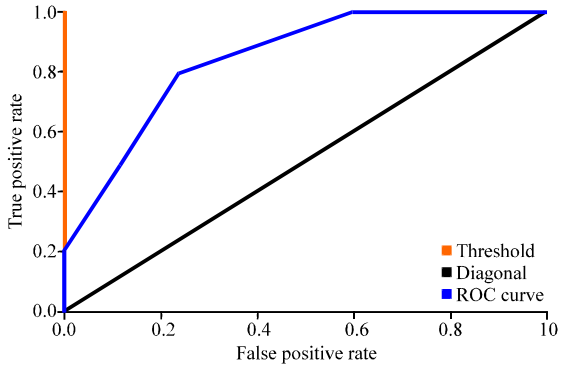

Fig. 4: ROC plot for weight intervention

Confusion matrix for smoke cessation intervention

	Predicted class		
Actual class		Other	Smoke cessation
	Other	24	1
	Smoke cessation	2	3

Overall predictive accuracy = 24+3/ (24+1+2+3) = 0.9

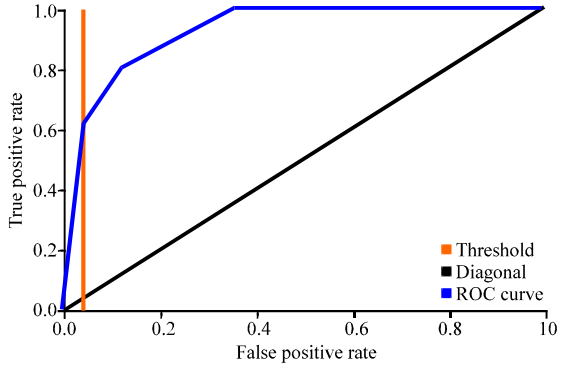

Fig. 5: ROC plot for smoke cessation intervention

Confusion matrix for exercise intervention

	Predicted class		
Actual class		Other	Exercise
	Other	25	0
	Exercise	5	0

Overall predictive accuracy = 25+0/(25+0+5+0) = 0.8

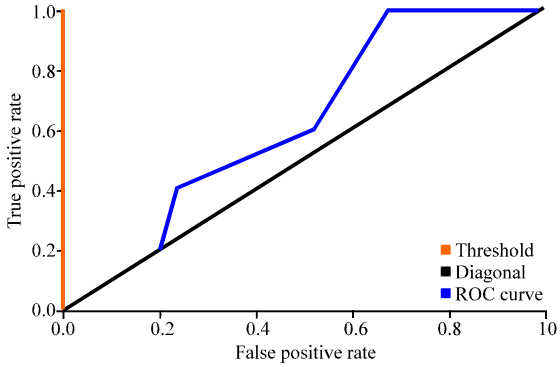

Fig. 6: ROC plot for exercise intervention

Confusion matrix for insulin intervention

	Predicted class		
Actual class		Other	Insulin
	Other	23	2
	Insulin	2	3

Overall predictive accuracy = 23+3/ (23+2+2+3) = 0.86

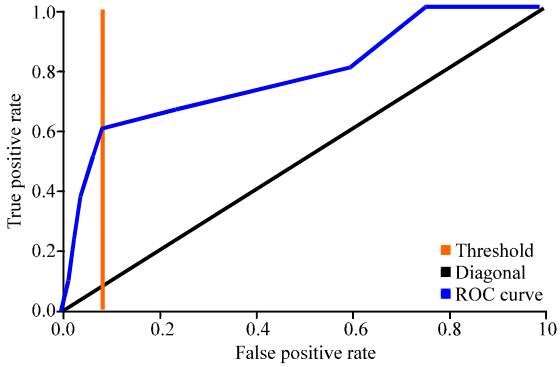

Fig. 7: ROC plot for insulin intervention

drug, diet control, exercise, weight and insulin interventions indicates the good prediction.

It is quite evident from the Table 1 that smoking cessation plays a vital role in controlling diabetes. Therefore, the results are quite impressive in predicting the diabetic interventions resulting high AUC of ROC, high accuracy with lowest cost. The smoking causes the increased rate of diabetes in both men and women and quieting smoking reduces the rate of diabetes (Will *et al*., 2001). Smokers experience four times the risk of death from all cardiovascular disease and three times the risk of death from coronary artery disease (Winslow *et al*., 1996). Present results are in excellent agreement of this fact. Therefore, quitting smoking would reduce the diabetic rate. In the context we would like to stress upon that smoking is one of the causes for elevated diabetic rate and quitting smoking obliviously is predicted as best way to control the rate of diabetes. For non smoking diabetic patients as per our study we have got high accuracy for weight control, diet control and physical exercise interventions. The diabetic control can be done through drug intervention also. The study reveals that the diabetic control over these factors would bring down the diabetic rate. Besides, this as shown in Table 1. Classified result of diabetes intervention. The first line treatment of type2 diabetes are diet control, weight reduction and physical exercises, despite this, if the blood glucose levels are elevated then the treatment through drug is to be followed and insulin injections are advised even after taking oral tablets. The results with regard to the diet, drug, weight, exercise and insulin the AUC of ROC, accuracy and cost are well matching.

CONCLUSION

In conclusion our experimental study states a close relationship of application of data mining technique on health care. In the light of data mining we had analysed the diabetes intervention control. The prevalence of diabetes is increasing among Saudi patients. In this study different mode of diabetic intervention controls are predicted. The experimental result states that the accuracy of smoking cessation model is very high which is calculated as 0.9, which directly indicates that quitting smoking reduces the rate of diabetes. Smoking is one of the causes for elevated diabetic rate and quitting smoking obliviously is predicted as best way to control the rate of diabetes. For non smoking diabetic patients as per our study we have got high accuracy of 0.8 for models of weight control, diet control and physical exercise.

ACKNOWLEDGMENT

This research was funded by a grant from the Deanship of Scientific Research in Salman bin Abdulaziz University, Ministry of Higher Education, Kingdom of Saudi Arabia.

REFERENCES

Al-Nozha, M.M., M.A. Al-Maatouq, Y.Y. Al-Mazrou, S.S. Al-Harthi, M.R. Arafah and M.Z. Khalil *et al*., 2004. Diabetes mellitus in Saudi Arabia. Saudi Med. J., 25: 1603-1610.

Breault, J.L., C.R. Goodall and P.J. Fose, 2002. Data mining a diabetic data warehouse. Artif. Intell. Med., 26: 37-54.

Charpentier, G., N. Genes, L. Vaur, J. Amar, P. Clerson, J.P. Cambou and P. Gueret, 2003. Control of diabetes and cardiovascular risk factors in patients with type 2 diabetes: A nationwide French survey. Diabetes Metab., 29: 152-158.

Craig, M.E., T.W. Jones, M. Silink and Y.J. Ping, 2006. Diabetes care, glycemic control and complications in children with type 1 diabetes from Asia and the Western Pacific Region. J. Diab. Complications, 21: 280-287.

Egan, J.P., 1975. Signal Detection Theory and ROC-Analysis. Academic Press, New York, USA., ISBN-13: 9780122328503, Pages: 277.

Gan, D., 2003. Diabetes Atlas. 2nd Edn., International Diabetes Federation, Brussels, Belgium.

Harwell, T.S., K.K. Vanderwood, T.O. Hall, M.K. Butcher and S.D. Helgerson, 2011. Factors associated with achieving a weight loss goal among participants in an adapted Diabetes Prevention Program. Primary Care Diabetes, 5: 125-129.

Kohavi, R. and F. Provost, 1998. Glossary of terms. Mach. Learn., 30: 271-274.

Nilsson, P.M., S. Gudbjornsdottir, B. Eliasson and J. Cederholm, 2004. Smoking is associated with increased HbA_{1c} values and microalbuminuria in patients with diabetes-data from the National Diabetes Register in Sweden. Diabetes Metab., 30: 261-268.

Richards, G., V.J. Rayward-Smith, P.H. Sonksen, S. Carey and C. Weng, 2001. Data mining for indicators of early mortality in a database of clinical records. Artificial Intell. Med., 22: 215-231.

Sigurdardottir, A.K., H. Jonsdottir and R. Benediktsson, 2007. Outcomes of educational interventions in type 2 diabetes: WEKA data-mining analysis. Patient Educ. Counseling, 67: 21-31.

Soriguer-Escofet, F., I. Esteva, G. Rojo-Martinez, S.R. de Adana and M. Catala *et al*., 2002. Prevalence of Latent Autoimmune Diabetes of Adults (LADA) in Southern Spain. Diabetes Res. Clin. Pract., 56: 213-220.

WHO, 2005. NCD risk factor standard report of Ministry of Health. Saudi Arabia.

Will, J.C., D.A. Galuska, E.S. Ford, A. Mokdad and E.E. Calle, 2001. Cigarette smoking and diabetes mellitus: Evidence of a positive association from a large prospective cohort study. Int. J. Epidemiol., 30: 540-546.

Winslow, E., N. Bohannon, S.A. Brunton and H.E. Mayhew, 1996. Lifestyle modification: Weight control, exercise and smoking cessation. Am. J. Med., 101: 25S-31S.

Yokotaa, I., S. Amemiya, K. Kida, N. Sasaki and N. Matsuura, 2005. Past 10-year status of insulin therapy for preschool-age Japanese children with type 1 diabetes. Diabetes Res. Clin. Pract., 67: 227-233.