Diagnosis of Diabetes by Applying Data Mining Classification Techniques

Comparison of Three Data Mining Algorithms

Tahani Daghistani, Riyad Alshammari
Health Informatics Department, College of Public Health and Health Informatics
King Saud Bin Abdulaziz University for Health Sciences (KSAU_HS)
King Abdullah International Medical Research Center (KAIMRC)
Minisrty of National Guard Health Affairs
Riyadh, KSA

***Abstract*—Health care data are often huge, complex and heterogeneous because it contains different variable types and missing values as well. Nowadays, knowledge from such data is a necessity. Data mining can be utilized to extract knowledge by constructing models from health care data such as diabetic patient data sets. In this research, three data mining algorithms, namely Self-Organizing Map (SOM), C4.5 and RandomForest, are applied on adult population data from Ministry of National Guard Health Affairs (MNGHA), Saudi Arabia to predict diabetic patients using 18 risk factors. RandomForest achieved the best performance compared to other data mining classifiers.**

Keyword—Diabetes; Data mining; Self-Organizing Map; Decision tree; Classification

I. Introduction

Saudi Arabia is facing financial challenge due to the prevalence of diabetes. The Ministry of Health (MOH) in Saudi Arabia and Institute for Health Metrics and Evaluation (IHME) implemented, as collaboration, the assessment of burden based on the direct cost of diabetes from integrated health information system in 2014 [1]. Based on the established system, the current estimated cost of diabetes is 17 billion Riyals (US $4.5 billion) with expectation to increase the cost to 27 billion Riyals (US $7.2 billion) in the case that undiagnosed people are documented. Moreover, if pre-diabetes people become diabetes the cost will increase to 43 billion Riyals (USD 11.43 billion). The cost includes medications, visits, and lab tests, which also varies based on the patient's stage. The high cost of treating diabetes plus the expected growth of diabetes will put Saudi Arabia face to face with financial and health challenges in near future. Prevention, monitoring and controlling are the most effective actions to face such a health care challenge.

Data mining techniques assist health care researchers to extract knowledge from large and complex health data. With the evolution of information technology, data mining provides a valuable asset in diabetes research, which leads to improve health care delivery, increase support to decision-making and enhance disease management [2]. Data mining techniques include pattern recognitions, clustering, classification and association.

Diabetes is one of the main topics for medical research due to the longevity of the diabetes and the huge cost on the health care providers. Early detecting of diabetes ultimately reduces cost on health care providers for treating diabetic patients [3-8], but it is a challenging task. For early detecting of diabetes, researchers can take advantage of the patient's health care data to convert raw data into meaningful information and extract hidden knowledge by applying data mining such as decision tree or SOM to construct an intelligent predictive model.

SOM or Kohonen maps is a machine-learning tool that is used to analyze heterogeneous data and provides supervised or unsupervised learning model [9-11]. Hence, SOM maps high dimensional data to be more meaningful by identifying similarities. In this research article, decision trees, namely C4.5 and RandomForest, are compared with SOM to build a classification model to predict diabetic patients using retrospective data collected from hospital database systems. The data sets are extracted from the hospital information management system from the Ministry of National Guard Health Affairs (MNGHA), Saudi Arabia. The National Guard Health hospitals provide optimum health care to their employees, dependents, other eligible patients and private patients. The data sets are collected from four hospitals in the three largest regions in Saudi Arabia in terms of populations. The hospitals are: i) King Abdulaziz Medical City (SANG) in Riyadh, Central Region; ii) King Abdulaziz Medical City in Jeddah, Western Region; iii) Imam Abdulrahman Al Faisal Hospital in Dammam, Eastern Region; and v) King Abdulaziz Hospital in Alahsa, Eastern Region. The contribution of this study is utilizing the data mining techniques to construct intelligent predictive model using real healthcare data that are extracted from hospital information systems using 18 risk factors.

The rest of the paper is organized as follows. The literature review is given in Section II. Methodology is presented in Section III. Results and discussion are given in Section IV. Finally, conclusions and future work are presented in Section V.

II. Literature Review

In the literature, SOM has been applied in health care data. Mäkinen et al. [12] used SOM algorithm to detect association

between certain risk factors and complications. They used SOM as an unsupervised method to cluster biochemical profiles. A 7 x 10 grid of hexagonal map units with Gaussian neighborhood function were used to present similarities and differences between variables. Tirunagari et al. [13] applied SOM to cluster heterogeneous diabetes data. They were able to reduce the dimensionality of the data and demonstrate the similarities between patients by placing them in groups using the U-matrix. As a result, the profiles of patients who need self care management were grouped clearly and easily were identified.

In another study, Tirunagari [14] used the SOM to recognize the behavior of self care based on survey data collected from type I diabetic patients. The visualization result improved understanding pattern of various behaviors as well as detecting patients who need to adjust their lifestyle. Zarkogianni et al. [15] proposed personalized hybrid model by combining Compartmental Models **(**CMs) and Self-Organizing Map. The model helped patients with Type I Diabetes Mellitus to predict the metabolic behavior. Luboschik et al. [5] used SOM as part of an early detecting system to predict Neuropathy complications in diabetic patients. By using the computational and visual methods of SOM, they were able to identify characteristics of diabetic Neuropathy patients.

Other data mining algorithms had been applied to classify diabetic patients. Farran et al. [16] used non-laboratory attributed to classify the diabetes by applying 4 data mining models that were logistic regression, k-nearest neighbors (k-NN), multifactor dimensionality reduction and Support Vector Machines (SVM). They achieved an accuracy of 85% for diabetic patients. Barakat et al. [17] applied SVM on data collected from a national survey in the Sultanate of Oman that investigated the prevalence of diabetes mellitus. They achieved a sensitivity of 93% and 94% for accuracy and specificity.

Moreover, Ganji et al. [18] used (FCS-ANTMINER) on public diabetes data set (Pima Indians Diabetes data set [19]). They obtained an accuracy of 84%. Huang et al. [20] employed three data mining algorithms that were Naive Bayes, IB1 and C4.5 to predict diabetes on data gathered from Ulster Community and Hospitals Trust (UCHT) between 2000 and 2004. They were able to achieve an accuracy of 98%. Furthermore, Al Jarullah A. [21] employed C4.5 data mining algorithm on Pima Indians Diabetes data set [19]. He achieved an overall accuracy of 78%.

From the literature review, data mining algorithms have been used to predict diabetes using public data or private data. However, the data sets are either small in size (less than 10,000 records) or collected from one region (mostly one hospital). In this research study, the data sets are collected from 4 large hospitals in Saudi Arabia. The model extracted from the data could assist in improving healthcare plans that are delivered for diabetic patients.

III. METHODOLOGY

To achieve the study objectives, study method consists of several phases, which are collection of data and attribute selection, data mining algorithms and evaluation criteria.

A. Data sets and Attributes Selection

In this work, the data sets are collected from Ministry of National Guard Health Affairs (NGHA) databases from the highest three populated regions in Saudi Arabia, where the databases have all patients visit information such as laboratory and medications, etc. These regions are: central region (Riyadh city), western region (Jeddah city) and eastern region (Alahsa and Dammam cities). The latest Saudi census showed that more than 66% of the country total population lives in these three regions and the largest city on these three regions are Riyadh city (The capital and the largest city in the Central region); Jeddah (the largest city in the Western region; iii) Dammam; and Alahsa (the largest two cities in the Eastern region) [22].

The data sets consist of 66,325 diabetic and non-diabetic instances. The study used data from the hospital Information System in MNGHA from the 2013 to 2015. Hospital databases are extremely exposed to inconsistent values, noisy and missing input values from the data because the data are coming from heterogeneous sources. There are several considerations that are followed and assured throughout the data extraction process by the information systems in MNGHA to insure the accuracy of the data. In addition, the data sets are gone through manual inspections to ensure the data are consistent and accurate.

All adult patients who have diabetes are included while pediatric diabetic patients are excluded. The data used for the study did not include identification information in order to not violate the patient privacy.

Detailed information about demographic variables is summarized in Table1. Furthermore, the data set divided into training and test data sets as follows:

- Data from 2013 to 2014 represents a training set that is used to construct and train the model.
- Data from 2015 represents a test set that is used to test the model and estimate the accuracy rate.

The data sets consist of a total of 18 attributes. The attributes include gender, age and region as demographic variables; patient's measurements such as BMI and blood pressure in addition to 11 various lab tests. The Data sets contain 36,811 male (55.50%) and 29,514 females (44.50%), all of them at least 14 years old and older. More than half of the total patients (64.47%) have diabetes; male diabetic patients represent 36.34% of the total diabetic patients, while female diabetic patients represent 28.13% of the total diabetic patients as shown in Table 1.

TABLE I. DISTRIBUTION OF DEMOGRAPHIC VARIABLES

Variables	*Diabetes*	*Non-diabetes*	*Total*
Region			
Central	34039 (62.87%)	20102 (37.13%)	54141 (81.63%)
Eastern	8012 (72.28%)	3073 (27.72%)	11085 (16.71%)
Western	708 (64.42%)	391 (35.58%)	1099 (1.66%)
Gender			
Male	24104 (36.34%)	12707 (19.16%)	36811 (55.50%)
Female	18655 (28.13%)	10859 (16.37%)	29514 (44.50%)
Age			
13-19	118 (0.28%)	460 (1.95%)	578 (0.87%)
20-34	1120 (2.62%)	2947 (12.51%)	4067 (6.13%)
35-44	2070 (4.84%)	2416 (10.25%)	4486 (6.76%)
45-64	16226 (37.95%)	7723 (32.77%)	23949 (36.11%)
65-84	20602 (48.18%)	8447 (35.84%)	29049 (43.80%)
>85	2623 (6.13%)	1573 (6.67%)	4196 (6.33%)
Overall total	42759 (64.47%)	(35.53%)	66325

Lab test data are described statistically and summarized in Table 2 in order to provide more understanding of lab tests data which are considered as attributes in the study.

TABLE II. STATISTIC DESCRIPTION OF LAB TEST DATA

	N	*Minimum*	*Maximum*	*Mean*	*Std. Deviation*
eGFR	66325	2.00	220.00	78.3311	40.83988
MCV	66325	.00	129.80	86.9547	7.58909
MCH	66325	12.20	59.80	28.0319	2.91036
MCHC	66325	27.10	373.00	317.5521	38.98827
RDW	66325	10.00	99.00	15.2312	2.42988
Plt	66325	3.00	999.00	273.7028	125.00519
MPV	66325	.00	90.00	8.5508	1.38118
WBC	66325	.00	319.60	9.3527	5.81372
RBC	66325	1.18	8.71	4.1693	.83756
Hgb	66325	5.10	232.00	114.5618	26.71571
Hct	66325	.04	54.70	.9110	4.44215
Valid N	66325				

B. Data Mining Algorithms:

R software [23] is used to employ SOM algorithm in order to predict diabetes patients. Kohonen package in R implements SOM as unsupervised algorithm as well as supervise algorithm. The *bdk* and *xyf* are supervised functions of SOM in R. The returned output obtained from calling both functions is used for prediction in this study.

Since SOM has a number of parameters, selecting the appropriate parameters, such as type of SOM, network size and training algorithm, is important. Parameters have direct impact on the classification performance as well as computational time [9]. The values for parameters are summarized in Table 3. On the other hand, Weka [24] data mining tool is used to run C4.5 and RandomForest decision trees using the default parameters.

TABLE III. SOM PARAMETERS

Parameters	***Meaning***	***Value***
Grid	To determine the size of map	20 x 20, hexagonal
Rlen	To determine number of iterations	1000
Alpha	To determine the learning rate for start and stop	[0.05, 0.01]
Radius	To determine the initial neighborhood. The value is decreased during training linearly	[90%, 85%]

C. Evaluation Criteria:

To select the best performance data mining algorithms in predicting diabetic patients, two standard matrices have been applied, which are Recall and Precision. Recall, Eq. 1, will reflect the number of diabetic instances who are correctly classified, which we need in such system. It is calculated using:

$$\text{Recall} = TP/(TP + FN) \quad (1)$$

While Precision, Eq. 2, represents the relevant instances that are correctly classified. It is calculated using:

$$\text{Precision} = TP/(TP+FP) \quad (2)$$

True Positive (TP) implies that diabetic patients who are classified as diabetic patients, whereas False Negative (FN) implies that diabetic patients who are classified as non-diabetic patients. On the other hand, False Positive (FP) implies that non-diabetic patients who are classified as diabetic patients. Commonly, the best learning algorithm is going to be selected based upon the performance of the classifiers in terms of high Recall, and Precision.

IV. RESULT AND DISCUSION

In Table 4, two different measurements were calculated for each algorithm for assessing how well each model and to be used to evaluate algorithm's performance compared to each other. C4.5 and RandomForest achieved Recall and Precision over 90% on the training data set while SOM (bdk and xyf) was able to achieve Recall and Precision over 79% on the training data set.

To choose the best algorithms in terms of high performance, according to the evaluation criteria, all algorithms are evaluated on an unseen data set (test data set). The algorithm/model who achieved the highest performance in terms of high Recall and Precision is considered to be the best one. It can be seen that RandomForest achieves the highest Recall and Precision on the test data set as indicated in Table 4.

TABLE IV. RESULT OF THE CLASSIFIERS

Models	***Recall***	***Precision***
SOM- bdk – Training data	0.79	0.83
SOM- bdk – Test data	0.69	0.48
SOM- xyf – Training data	0.79	0.84
SOM- xyf – Test data	0.79	0.46
C4.5 – Training data	0.965	0.92
C4.5 – Test data	0.776	0.67
RandomForest – Training data	1.0	1.0
RandomForest – Test data	0.904	0.68

The reason behind that SOM could not perform higher than decision trees due to the fact that the SOM constructs the its model from only the first SOM grid layer. The multi-layer classification capability of SOM could improve the performance. However, the multilayer capability is not available in R software [23].

In this study, SOM and decision tree techniques are applied to predict diabetic patients using 18 risk factors (attributes). The most common risk factors among the model constructs from the algorithms are as the following: i) gender; ii) age; iii) blood pressure; iv) Body Mass Index (BMI); v)

region; and vi) several lab tests such as Hematocrit (Hct), hemoglobin (Hgb), Platelet count (Plt) and Mean Platelet Volume (MPV).

The extracted knowledge from the research conducted among the samples (patient records) from MNGHA can be generalized to the wider diabetic population in Saudi Arabia since the data sets (samples) are collected from the largest populated region in Saudi Arabia where more than 66% of the total country population lives.

V. CONCLUSION

Model constructed from the data mining algorithms could help to support decision making in different fields including health care field. In this research, real health care data sets have been collected from MNGHA databases that contain 18 attributes. Furthermore, three data mining algorithms have been evaluated, namely SOM (bdk and xyf), C4.5 and RandomForest to construct data mining models to predict diabetic patients using real health care data sets.

The results show that the constructed data mining model could assist health care providers to make better clinical decisions in identifying diabetic patients. Additionally, the model could be further developed for patient protection. In the future, the results can be utilized to create a control plan for diabetes because diabetic patients are normally not identified till a later stage of the disease or the development of complications.

ACKNOWLEDGMENT

This study was funded by the King Abdullah International Medical Research Center (KAIMRC), National Guard, Health Affairs, Riyadh, Saudi Arabia with research grant No. SP15/064.

REFERENCES

[1] Mokdad AH, Tuffaha M, Hanlon M, El Bcheraoui C, Daoud F, et al. (2015) Cost of Diabetes in the Kingdom of Saudi Arabia, 2014. J Diabetes Metab 6: 575

[2] Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., & Hua, L. (2012). Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems, 36(4), 2431-2448.

[3] Li, R., Zhang, P., Barker, L. E., Chowdhury, F. M., & Zhang, X. (2010). Cost-effectiveness of interventions to prevent and control diabetes mellitus: a systematic review. Diabetes care, 33(8), 1872-1894.

[4] Lin, J. H., & Haug, P. J. (2006). Data preparation framework for preprocessing clinical data in data mining. In AMIA Annual Symposium Proceedings (Vol. 2006, p. 489). American Medical Informatics Association.

[5] Luboschik, M., Röhlig, M., Kundt, G., Stachs, O., Peschel, S., Zhivov, A., ... & Schumann, H. (2014). Supporting an Early Detection of Diabetic Neuropathy by Visual Analytics.

[6] Nuwangi, S. M., Oruthotaarachchi, C. R., Tilakaratna, J. M. P. P., & Caldera, H. A. (2010, December). Utilization of Data Mining Techniques in Knowledge Extraction for Diminution of Diabetes. In Information Technology for Real World Problems (VCON), IEEE 2010 Second Vaagdevi International Conference on (pp. 3-8).

[7] Wang, K. J., Adrian, A. M., Chen, K. H., & Wang, K. M. (2015). An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus. Journal of biomedical informatics, 54, 220-229.

[8] Shivakumar, B. L., & Alby, S. (2014, March). A Survey on Data-Mining Technologies for Prediction and Diagnosis of Diabetes. In Intelligent Computing Applications (ICICA), IEEE 2014 International Conference on (pp. 167-173).

[9] Ballabio, D., Vasighi, M., & Filzmoser, P. (2013). Effects of supervised Self Organising Maps parameters on classification performance. Analytica chimica acta, 765, 45-53.

[10] Wehrens, R., & Buydens, L. M. (2007). Self-and super-organizing maps in R: the Kohonen package. Journal of Statistical Software, 21(5), 1-19.

[11] Wijayasekara, D., & Manic, M. (2012, June). Visual, linguistic data mining using Self-Organizing Maps. In Neural Networks (IJCNN), The 2012 International Joint Conference on (pp. 1-8). IEEE.

[12] Mäkinen, V. P., Forsblom, C., Thorn, L. M., Wadén, J., Gordin, D., Heikkilä, O., ... & Groop, P. H. (2008). Metabolic phenotypes, vascular complications, and premature deaths in a population of 4,197 patients with type 1 diabetes. Diabetes, 57(9), 2480-2487.

[13] Tirunagari, S., Poh, N., Aliabadi, K., Windridge, D., & Cooke, D. (2014, December). Patient level analytics using self-organising maps: A case study on Type-1 Diabetes self-care survey responses. In Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on (pp. 304-309). IEEE.

[14] Tirunagari, S., Poh, N., Hu, G., & Windridge, D. (2015). Identifying Similar Patients Using Self-Organising Maps: A Case Study on Type-1 Diabetes Self-care Survey Responses. arXiv preprint arXiv:1503.06316

[15] Zarkogianni, K., Litsa, E., Vazeou, A., & Nikita, K. S. (2013, November). Personalized glucose-insulin metabolism model based on self-organizing maps for patients with Type 1 Diabetes Mellitus. In Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on (pp. 1-4). IEEE.

[16] B. Farran, A. M. Channanath, K. Behbehani, T. A. Thanaraj, Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from kuwait– a cohort study., BMJ Open 3 (5). 2012

[17] N. Barakat, A. Bradley, M. Barakat, Intelligible support vector machines for diagnosis of diabetes mellitus, Information Technology in Biomedicine, IEEE Transactions on 14 (4) (2010) 1114–1120.

[18] M. F. Ganji, M. S. Abadeh, A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis, Expert Systems with Applications 38 (12) (2011) 14650 – 14659

[19] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, R. S. Johannes, Using the adap learning algorithm to forecast the onset of diabetes mellitus, Johns Hopkins APL Technical Digest 10 (1988) 262–266.

[20] Y. Huang, P. McCullagh, N. Black, R. Harper, Feature selection and classification model construction on type 2 diabetic patients'data, Artificial Intelligence in Medicine 41 (3) (2015) 251–262.

[21] A. Al Jarullah, Decision tree discovery for the diagnosis of type ii diabetes, in: Innovations in Information Technology (IIT), 2011 International Conference on, 2011, pp. 303–307.

[22] Central Department of Statistics & Information (CDSI), Statistical yearbook, http://www.cdsi.gov.sa/ar/1805/, June, 2016

[23] R, (n.d.), r-project.org, Retrieved 15 November 2015, from https://cran.r-project.org/bin/windows/base/

[24] Machine Learning Group at the University of Waikato. (2015). *Weka 3: Data Mining Software in Java.* Retrieved December 23, 2015 from http://www.cs.waikato.ac.nz/ml/weka/