Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 82 (2016) 115 – 121

Procedia Computer Science

Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia

Performance Analysis of Data Mining Classification Techniques to Predict Diabetes

Sajida Perveen[a]*, Muhammad Shahbaz[a], Aziz Guergachi[b], Karim Keshavjee[c]

[a]*Department of Computer Science & Engineering, University of Engineering & Technology, Lahore, Pakistan*
[b]*Ted Rogers School of Information Technology Management, Ryerson University, Toronto, Ontario, Canada*
[c]*University of Victoria, School of Health Informatics, Victoria, British Columbia, Canada*

Abstract

Diabetes Mellitus is one of the major health challenges all over the world. The prevalence of diabetes is increasing at a fast pace, deteriorating human, economic and social fabric. Prevention and prediction of diabetes mellitus is increasingly gaining interest in healthcare community. Although several clinical decision support systems have been proposed that incorporate several data mining techniques for diabetes prediction and course of progression. These conventional systems are typically based either just on a single classifier or a plain combination thereof. Recently extensive endeavors are being made for improving the accuracy of such systems using ensemble classifiers. This study follows the adaboost and bagging ensemble techniques using J48 (c4.5) decision tree as a base learner along with standalone data mining technique J48 to classify patients with diabetes mellitus using diabetes risk factors. This classification is done across three different ordinal adults groups in Canadian Primary Care Sentinel Surveillance network. Experimental result shows that, overall performance of adaboost ensemble method is better than bagging as well as standalone J48 decision tree.

Peer-review under responsibility of the Organizing Committee of SDMA2016
Keywords: Diabetes Mellitus; Ensemble method; Base Learner; Bagging; Adaboost and Decision tree

* Corresponding authors. Tel.: +92-556601721.
E-mail address: sajida.uaar@gmail.com

Peer-review under responsibility of the Organizing Committee of SDMA2016
doi:10.1016/j.procs.2016.04.016

1. Introduction

Diabetes mellitus (DM), commonly known as diabetes, is a chronic and one of the dramatically increasing metabolic diseases in the world[6, 11]. It is associated with an abnormal increase in the level of glucose (hyperglycemia) in blood, ensued either owing to the inadequate production of insulin by pancreas (Type 1 diabetes) or the cells failure in effective response to insulin produced by pancreas (Type 2 diabetes)[13]. The downside of all this variability in plasma glucose (hyperglycaemia, hypoglycemia) is that it leads to severe damage to many of the body's vital systems especially blood vessels and the nervous system[10]. While its causes are not yet entirely understood, scientists believe that both genetic factors and environmental triggers are involved therein[8]. However, diabetes used to be most prevalent in adults and once called "adult-onset" diabetes. It is now widely believed that diabetes mellitus is closely related with the aging process.

According to Canadian Diabetes Association (CDA), between 2010 and 2020, the number of people diagnose with diabetes in Canada is expected to escalate from 2.5 million to about 3.7 million[7]. Unfortunately, worldwide the picture is no different from this. According to the International Diabetes Federation, number of individuals with diabetes mellitus has reached 382 million in 2013[14] that bring 6.6% of the world's total adult population with diabetes. Health care expenditures for diabetes are anticipated to be $490 billion for 2030, accounting for 11.6% of the total health care expenditures in the world[2]. Furthermore, diabetes is a potentially independent contributing risk factor to microvascular complications. Its patients are likely to be more vulnerable to an elevated risk of microvascular damage thereby exposing them to cardio vascular disease two to fourfold more as compared to no diabetic individuals. This micro vascular damage and consequent cardio vascular disease ultimately lead to retinopathy, nephropathy and neuropathy[8]. Studies revealed that the life expectancy for people with diabetes might get curtailed by as much as 15 years[17].

Given the above narrated consequences, early stage detection and diagnosis of diabetes is the need of the day. In this context, Electronic Medical Records (EMRs) play a crucial role by keeping track of repeated clinical measurements related to particular patient's condition over time. To provide a rapid and minutely detailed analysis of medical data, diabetes risk scoring models as well as their various algorithms has been widely investigated. Schwarz et al. [16] provided a comprehensive survey of these models with their specificity and sensitivity. However, as these risk scoring models involve human intervention though to some extent in deciding criteria and risk score, it may expose the results to the human error.

Data mining is a prominent tool set in medical databases. This promising approach improves sensitivity and/or specificity of disease detection and diagnosis by opening a window of comparatively better resources. It also substantially reduces accompanied cost by bypassing unwanted and expensive medical tests[9]. Extensive studies regarding diabetes prediction has been undergone for several years. Recently, some reports have compared different learning techniques. Such comparisons are generally a few and conducted on Pima Indian diabetic database with a limited number of data sets.

On the other hand, this study follows the adaboost and bagging Data Mining ensemble techniques using J48 (c4.5) decision tree as a base learner along with standalone data mining technique J48 (c4.5). More specifically, the dataset used in this study for disease diagnosis and decision making is obtained from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) database. That is Canada's first multi-disease EMR-based surveillance system. Firstly; The objective of this study is to evaluate the performance of aforementioned techniques of data mining to accurately classify patients with diabetes mellitus using diabetes risk factors across three different ordinal adults groups in CPCSSN, namely (i) young adults (ii) middle aged adults (iii) adults older than 55. Secondly; to identify the best ensemble framework for J48 decision tree that would help identify the diabetes patients efficiently and most importantly, with high accuracy. The rest of paper is organized as follows: Section 2 presents material and method. Section 3 describes results, evaluation and discussion. Conclusion is given in section 4.

2. Material and method

The dataset used in this study is obtained from the CPCSSN database (http://cpcssn.ca/). CPCSSN database contains 667907 records for a period ranging from 2003 through 2013. Each record contains several features including important risk factors such as vital signs, diagnosis and demographics that will be used for diabetes prediction. This data have previously been used in[7] to validate the performance of Framingham diabetes risk model in Canadian population which investigated the 8 year risk for developing diabetes. An abstract detail of those relevant risk factors selected in this study is provided in Table 1 that includes age, sex, systolic blood, diastolic blood pressure, high density lipoprotein (HDL) triglycerides (TRG), body mass index (BMI), and fasting blood glucose (FBG). Out of 667,907 patients, 40,042 patients were diagnosed as diabetic, which constitutes about 6% of the total patients. The ascertainment of diagnosis for diabetes for each patient is based on the most recent laboratory results.

Table 1. Characteristics of the population in the Canadian primary care sentinel surveillance network database

Predictors	Findings
Demographic (Gender, Age)	
Male, sample size, %	287964, 43.27
Female, sample size, %	379561, 57.04
Male age mean (SD),Years	47.27±25.10
Female age mean(SD),Years	49.53±24.84
Vital Signs/ clinical measures	
Systolic blood pressure, mean (SD), mm Hg	121.94± 16.95
Diastolic blood pressure mean (SD), mm Hg	73.3 ± 12.4
Unknown disease frequency, %	393344, 59
COPD frequency, %	15926, 2.38
Dementia frequency, %	12007, 1.79
Depression frequency, %	62682, 10
Diabetes Mellitus frequency, %	40317, 6
Epilepsy frequency, %	5553, 0.83
Hypertension frequency, %	88615, 13
Osteoarthritis frequency, %	47606, 7
Parkinson's Disease frequency, %	1825, 0.2
Lab Values	
FG, mean (SD), mmol/L	5.54 ± 1.91
Triglycerides, mean (SD), mmol/L	1.43± 1.21
HDL, sample size, mean (SD), mmol/L	1.38 ± 0.41
BMI, mean (SD), kg/m^2	26.54± 7.37

The data on clinical measurements are partial at this stage; approximately 660,745 patients do not have information for all the risk factors that are considered relevant in this study for the prediction of diabetes. Hence, upon performing sanity checks, the final data set resulted in a total of 4,678 participants of which 4,301 are non-diabetic and 377 diabetic. Since the study goal is to compare the performance of aforementioned data mining algorithms across three different age groups therefore CPCSSN datasets is divided into three research cohorts namely D18-35, D36-55 and D > 55 with the cutoff age group of 18-35, 36-55 and more than 55 years respectively. For instance, D18-35 contains only data of those patients those ages are between18 to 35 year and diagnosed as diabetic positive/negative based on most recent laboratory test results.

2.1. Experimental methodology

This study systematically involves three representative data mining techniques for predictive data mining task. That includes standalone J48 decision tree, ensemble techniques bagging and adaboost using J48 as a base learner. These methods are combined for generating knowledge to make it useful for decision making. Each method will produce different results to classify patients with diabetes mellitus comprising the available variables in each dataset created from CPCSSN dataset that are then compared and evaluated using AUROC (Area under receiver operating characteristic curve). The experimentation is performed using WEKA.

2.1.1. J48 decision tree

J48 decision tree is an open source java implementation of commonly known C4.5 supervised classification algorithm in WEKA. It is an evolution and extension of ID3 algorithm developed by Quinlan. It is a fraction between information gain and its splitting information. Quinlan[4] presented a comprehensive detail related to J48 decision tree.

$$Gain\_Ratio(D,A) = \frac{Entropy(D)\sum_{j=1}^{l}(\mathrm{p_j} \times Entropy(\mathrm{p_j}))}{Spliting\_Info}$$

2.1.2. Bagging

Bagging (Breiman,1996), derived for bootstrap aggregating is one of the simple but powerful independent ensemble methods[3] to improve the accuracy of unstable learning algorithms i.e. decision tree, rule learning algorithms[12]. In bagging dataset is distributed into various bootstrap replicates. Each replicate is drawn independently from the original dataset with replacement; on average each replicate contains 63.2% of the original data[12]. The process is carried out by repeatedly running the weak learner on various bootstraps. The classifier learned from weak learner at each iteration is combined into strong composite classifier in order to gain high accuracy than any single component classifier could do individually.

$$sign(\sum_{t=1}^{L} f_t(\mathrm{x})$$

2.1.3. Adaboost

Adaboost an acronym for Adaptive Boosting is one of the well-known ensemble methods proposed by Freund and Schapire[15]. It is an iterative process that produces strong classifier which consists of a sequence of weighted classifiers that complement one another. These base learners trained on different subsets are drawn deterministically from original dataset. The main idea behind this method is that at each following iteration more emphasis is given on examples that were are misclassified in previous iteration. The amount of emphasis is quantified by a weight that is assigned to every instance in the training replicate at each step.

$$H(\mathrm{x}_i) = sign\sum_{t=1}^{T} \alpha_t h_t(\mathrm{x}_i)$$

3. Results, evaluation and discussion

As mentioned earlier, the data used in this study is obtained from the CPCSSN database. The 9 potentially relevant risk factors associated with the prediction of diabetes mellitus, as proposed in literature [1, 5] are selected in this study as tabulated in Table 1.

Table 2. Study sample distribution among different age group with Chi-square test

	Diabetic		Non diabetic	
Age group	N	%	N	%
18-35	4	1.06	183	5.25
36-55	51	13.57	194	5.56
Older than 55	322	85.41	3107	89.07
Total	377	100.0	3484	100.0

Chi-square

Chi-square value	Df	p-value
46.85	2	0.000

We conducted chi-square test in order to identify statistical significance of age groups and diabetes, particularly to explore the association across different ordinal age groups and diabetes prevalence. Table 3 shows the results considering a significant level of 0.05. The result demonstrated a highly significant difference among age groups and diabetes prevalence. This means that those with older age had higher likelihood to develop diabetes than those with younger age. That means, age is a significant influencing factor for diabetes.

Since the objective of this study is to evaluate the performance of standalone J48 decision tree, two ensemble techniques bagging and adaboost using J48 as a base classifier across three different age groups therefore CPCSSN dataset is divided into three research cohorts namely D18-35, D36-55 and D > 55 with the cutoff age group of 18-35, 36-55 and more than 55 years respectively. In summary, there are two types of ensemble methods and one standalone J48 decision tree and three types of datasets. Table 2 shows the diabetes prevalence ratio across different age group in whole CPCSSN dataset. The final data set resulted in a total of 3,861 participants of which 3,484 are non- diabetic and 377 diabetic as shown in Table 2.

Each record in the above mentioned cohorts is augmented with a diabetes status positive/negative based on most recent laboratory test results. This status is considered as class label for each instance in the data. In the present study holdout method is used to evaluate the performance of classifiers. Therefore, we split the datasets into training and testing sets. The 60 % portion of data reserved for model induction and rest of the 40% is used to test the accuracy of the trained model. All experimentation is carried out in 10 independent runs in order to obtain sustained and reliable results. The mean is calculated for 10 runs, as each run renders a distinct result given the randomness of ensemble learning. To assess the overall performance or the discriminative capability of binary classifiers in Canadian primary care patients Area under Receiver Operating Characteristic (AROC) curves is used as a tool.

The AROC curve basically represents the combination of sensitivity and specificity[3, 11]. Theoretically, the AROC can assume values between 0 and 1, where an ideal classifier will take the value of 1. However, the practical lower bound for random classification is 0.5 that means the classifier with no discriminative capability whereas classifiers with an AROC significantly higher than 0.5 have at least some ability to discriminate. Fig.1 depicts the experimental results of the study using adaboost, bagging and J48 respectively. In the results, the area under AROC for bagging ensemble method with large dataset is 0.98%, showing a high reliability of discriminative capability among all the methods. It can also be derived, the larger the sample size, the greater is the performance of bagging. Overall, adaboost ensemble method outperformed across three different age groups and also demonstrates its unique characteristic of dealing with small sample size. J48 decision tree also yielded better performance with relatively larger sample size.

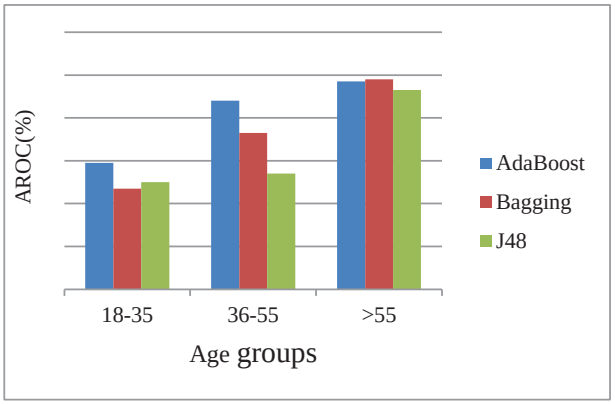

Fig. 1. Comparison of ensembles and J48 decision tree across three different age groups in CPCSSN dataset

4. Conclusion

Decision tree is one of the most powerful and widely applied techniques for classification and prediction. Our study constructed reasonably good models with higher performance to classify diabetic patients, across three age groups in the Canadian population, using bagging adaboost as well as J48 decision tree. The dataset used in this study is obtained from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) database. Evaluation of results indicates that adaboost ensemble method outperforms than bagging as well as standalone J48 decision tree. In future, similar ensemble approaches can be applied on other disease datasets such as hypertension, coronary heart disease and dementia. Furthermore, diverse individual techniques like Naïve Bayes, SVM and neural networks etc. can be incorporated as base learners in ensemble framework.

References

1. C. A. R., G. A. and K., N. 2011. Validating the CANRISK prognostic model for assessing diabetes risk in Canada's multi-ethnic population. *Chronic diseases and injuries in Canada*. 32, 1(Dec. 2011).
2. Carlo, B G., Valeria, M. and Jesús, D. C. 2011. The impact of diabetes mellitus on healthcare costs in Italy. *Expert review of pharmacoeconomics & outcomes research*. 11, (Dec. 2011),709-19.
3. Brown, G., Wyatt, J. L. and Tiňo, P. 2005. Managing diversity in regression ensembles. *The Journal of Machine Learning Research*, *6*, 1621-1650.
4. J., R. Q. C4. 5: programs for machine learning.2014. *Elsevier*. 28(June. 2014).
5. Jian-jun, D., Neng-jun, L., Jia-jun, Z., Zhong-wen, Z., Lu-lu, Q., Ying, Z. and Lin, L. Evaluation of a risk factor scoring model in screening for undiagnosed diabetes in China population. *Journal of Zhejiang University Science B.* 12, 1 (Oct. 2011), 846-852.
6. Kandhasamy, J. P., and S. B. Performance Analysis of Classifier Models to Predict Diabetes Mellitus. *Procedia Computer Science.* 47, (2015), 45-51.
7. Morteza, M., Franklyn, P., Bharat, S., Linying, D., Karim, K. and Aziz G. 2015. Evaluating the Performance of the Framingham Diabetes Risk Scoring Model in Canadian Electronic Medical Records. *Canadian journal of diabetes* 39, 30(April. 2015), 152-156.
8. Nahla B., Andrew, P. B. and M., N. B. 2010. Intelligible support vector machines for diagnosis of diabetes mellitus. *Information Technology in Biomedicine, IEEE Transactions.* 14, (July. 2010), 1114-20.
9. R., D. C. 2009. Data mining in healthcare: Current applications and issues. School of Information Systems & Management, Carnegie Mellon University, Australia. 5(Aug. 2009).

10. Rian, B. L. and E, I. 2015. The Early Detection of Diabetes Mellitus (DM) Using Fuzzy Hierarchical Model. Procedia Computer Science. 59, 31(Dec. 2015), 12-9.
11. Seokho, K., Pilsung, K., Taehoon, K., Sungzoon, C., Su-jin, R., and Kyung-Sang, Y. 2015. An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction. *Expert Systems with Applications*. 42, 1 (jun. 2015), 4265-4273.
12. Thomas G. D. 2000. Ensemble methods in machine learning. In Multiple classifier systems. Springer Berlin Heidelberg. 21(June. 2000), 1-15.
13. Vijiyarani, S. and Sudha, S. 2013. Disease prediction in data mining technique–a survey. 2, (2013), 17-21.
14. V., A. K. and R., C. 2013. Classification of Diabetes Disease Using Support Vector Machine. *International Journal of Engineering Research and Applications*. *3, (April. 2013)*, 1797-1801.
15. Yoav, F. and Robert, E. S. Experiments with a new boosting algorithm. *InICML*. 96, 3(July. 1996), 148-156.
16. Schwarz, P. E., J., L., J. L., and J., T. 2009. Tools for predicting the risk of type 2 diabetes in daily practice. *Hormone and metabolic research= Hormon-und Stoffwechselforschung= Hormones et métabolisme* .41, (Feb. 2009), 86-97.
17. Choi, S. B., Kim, W. J., Yoo, T. K., Park, J. S., Chung, J. W., Lee, Y. H., ... & Kim, D. W. 2014. Screening for prediabetes using machine learning models. *Computational and mathematical methods in medicine*, *2014*.