An Ensemble Model of Data Mining Approach for Enhancing the Performance of Diabetes Mellitus Diagnosis

[1] S.Hemalatha, [2] T.Kavitha, [3] T.M.Saravanan, [4] K.Chitra

Article Info
Volume 82
Page Number: 6251 - 6257
Publication Issue:
January-February 2020

Article History
***Article Received**: 18 May 2019*
***Revised**: 14 July 2019*
***Accepted**: 22 December 2019*
***Publication**: 30 January 2020*

Abstract:
Data mining is a significant creditworthy technique in fields such as banking, communication, education, advertising and healthcare. Since data mining is a remarkable resource in the domain of medical databases, our research focuses on employing ensemble data mining techniques on medical databases.Of many chronic diseases, diabetes mellitus is slowly but surely becoming a major threat to all ages across the globe and a central public health issue. In order to prevent and control the diabetes, several researches are ongoing. The main intend of this paper is to introduce a new ensemble approach, which plays a key role in the domain of medical data mining, more than ever, in detecting diabetes mellitus. Ensemble methodology is one of the growing tactics for strengthening classifier accuracy. Typically, ensemble is an effective technology that combines multiple foundation classifier predictions. Support vector machine, Random Forest and Adaboost are very efficient algorithm and play a vital role in delivering the highest level of precision rate. When in ensemble of these algorithms offer more accuracy than when applied in single. Here an ensemble technique has been applied to improve the accuracy rate of diagnosing diabetes mellitus.To boost the accuracy rate of diagnosis of diabetes mellitus, an integrated strategy was adopted here in this paper.

Keywords: *EnsembleData mining Techniques, Diabetes mellitus, SVM, Adaboost and Random Forest.*

1. INTRODUCTION

Worldwide, Diabetes mellitus is now becoming a serious health issue and influencing countries with low and middle incomes mostly. It is projected that 285 million people worldwide are suffering with diabetes and are predicted to experience up to 436 million by 2030. In order to better cope with this health issue, it is prudent to follow a realistic and feasible approach that takes into account the role of people with diabetes on the economic and health hierarchy of the world.

Clinics and other health organizations are equipped with enormous quantities of data that give the medical records of people a challenge in analysing. It really is crucial to discern between type 1 as well as type 2 diabetes. Type 2 Diabetes among the wider sections of people is more common. It exists all but utterly among adults, but now it also crop up on children. Perhaps there is a sudden rise in type 2 diabetes worldwide owing to the mechanical life trends and obesity of today. Further, to address this rising health challenge, world leaders have perpetrated to reduce the burden of diabetes, which is considered as one of four priority diseases.

Data mining empowers data to be excavated from a large bulk of data. It is the way that data mines facts.Data mining techniques typically involve domain market and data awareness, learning information, simulation, development, and implementation.Besides this, certain other data mining techniques are also available such as outer detection, regression, clustering, classification, prediction, association rules, and sequential patterns analysis. R-language and Oracle Data mining are Well-known tools for data mining. Data mining is often used in various sectors or fields namely banking, supermarkets, communications, insurance, education, bioinformatics supermarkets, retail, e-commerce service providers. Ultimately, it's all about studying the past and predicting the possibilities. Due to the massive and rapid raise in the magnitude of medical information, data mining strategies in this field are incredibly effective.

An Effective machine-based information and decision support technologies can make it a lot easier for medical practitioners to do the same effectively in their work. For the effective and efficient implementation of an automated system, greater provision of systematic review of various techniques is desired. A survey was conducted here on various research papers and an outline review was addressed on the Data Mining techniques for the diagnosis and prognosis of Diabetes Mellitus disorders and even some key issues are highlighted.

The rest of the content of this paper is presented as defies: the overview of diabetes mellitus and data mining is given in Sections 2-3. Section 4 articulates the findings for most used DMT algorithms for the diagnosis and prognosis of diabetes mellitus. Section 6 covers the suggestions for the future development of Data Mining Technology methodologies and their impact in medical field and Section 7 affords a narrow conclusion.

2. About Diabetes Mellitus

2.1 What is Diabetes Mellitus?

Diabetes mellitus is a disorder where blood sugar levels are exceptionally high due to a shortage of insulin naturally produced to reach its requires. It is the result of insufficient pancreatic insulin generation or inadequate processing of blood glucose in the body.

The most common three types of diabetes are as follows:

Type 1, as well referred as juvenile diabetes, is usually found in infancy, but often not.

Type 2, affects up to 95 per cent of cases. This happens more often in adolescents due to overweight and less physical childhood activity.

Gestational, comes into play during most of the pregnancy period of women.

The great news is that a great deal of research is carrying on to make perfect sense and reliably to avoid the diabetes. Diabetes is a life-long disease that must be managed to stay alive.

2.2 Health issues related to Diabetes

Over the period of time, the people suffering with diabetes may go with some health issues such as heart diseases, nerve damage, dental, eye sight, stroke etc.,

2.3 The prevalence of Diabetes

There were millions of people with diabetes according to the 2017 National Diabetes Statistics Survey. Diabetes affects the highest percentage of people over 65 years of age. Roughly around 90 – 95 percent of prevalence is for Type 2 diabetes among adults.

2.4 Symptoms of diabetes

General signs of the diabetes are as follows Loss of weight, increased hunger, excessive tiredness, augmented thirst, repeated urination, blurry vision, unhealing sores, etc.,

3. Datamining

Data mining is the practice of automatically or semi-automatically exploring, recapitulates and interpreting enormous volume of data to find appropriate functional patterns and principles fromdifferent perceptions.It can also be called data mining knowledge discovery in data.

Data mining is often used and attractive technology in various sectors like medical, business, bioinformatics, finance, marketing, etc.,. This falls all over the study by means of established methods and algorithms to predict future trends with real and descriptive statistics. Research's core undertaking is to seek the shielded truth, and information technology provides proactive resources to facilitate, precise, and consistent the research process. Data Mining is one such basic tool bag of the investigator which supports them in the field of science.

3.1Future and progress of data mining

Healthcare's success may well rely on using data mining to minimize healthcare costs, classify treatment regimens and guidelines, assess efficiency, track false policy and medical claims, and eventually increase patient care quality.

4. Review of Literature on DataMining techniques

From the review of literature, the below mentioned DataMining techniques has been applied on various diabetes dataset for the diagnosis of diabetes mellitus. This review gives the detailed knowledge of all the techniques used under the condition of medical problems and the tools which are employed over them. When applied to various data with different parameters, each algorithm has its own pros and cons. Accuracy rate of diagnosing diabetes has been detected by implementing various algorithms has been given in the below table no1

Table no 1. Comparison of Various algorithms and its measured accuracy rate

S.No	Techniques used	Accuracy achieved in %
1.	K-Means and Logistic regression	83.2
2.	Naïve Bayes	69.11
3.	Ada boost and bagging	85.99
4.	K-nearest	83.49
5.	Support Vector Machine	86.3
6.	Back Propagation	83.11
7.	C4.5	80
8.	Random Forest	86.45
9.	Gaussian Process	78.54
10.	Re-Rx with j48graft	80.98
11.	Rough set and Bat optimization	86.76
12.	Spider monkey optimization	85
13.	J48 and Naïve bayes	86
14.	CART	75

5. Proposed Ensemble Approach

Here, a new prototype has been designed to overcome the limitations in other algorithms. As a replacement for sole algorithm, an ensemble technique will support for betterment of accuracy level in diagnosis of diabetes mellitus. When applied singly, Random forest, support vector machine are giving better accuracy rate than other algorithms. Even, the combined algorithms are standing next to these two algorithms. Also adaboost is a healthier algorithm for providing accuracy in prediction.

Almost any ensemble development algorithm needs to generate and aggregate individual models from a strong-level context. It is reasonable to assume that, rather than the combination, most existing algorithms concentrate solely on the implementation of the base models. Ensemble methodology is one of the growing tactics for strengthening classifier accuracy. So by ensemble algorithms like Random Forest with AdaBoost and Support Vector machine, a novel and enhanced approach have been found for precise diagnosis of diabetes mellitus. It has been depicted in the diagrammatical representation in Fig no.1.

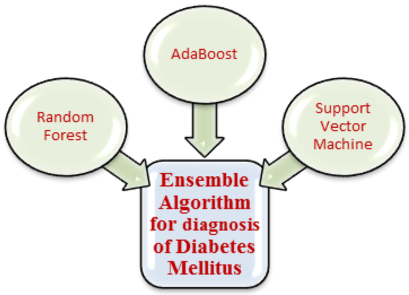

Fig no.1. Ensemble Model for Diagnosis of Diabetes mellitus

5.1 System Flow of Ensemble Model

Main processes involved in this work are

1. Surveillance
2. Data Pre-processing using feature selection and normalization
3. Developing the ensemble model with Random Forest, Support Vector Machine, Adaboosting
4. Training and testing the Model on data

5.1.1 Surviellance:

The dataset used here is uci machine learning repository. In particular, pima indians diabetes database consists of female patient records with at least 21years of age.

The datasets are made up of an eight clinical predictor variables and one dependent- target variable called outcome with value 0 or 1 for the indication of positive or negative diabetes. Clinical predictor variables are also called as independent variables such as

1. Number of times pregnant
2. plasma Glucose
3. Diastolic blood pressure
4. SkinThickness
5. Insulin
6. BMI
7. Diabetes Pedigree Function
8. Age

5.1.2 Data pre-processing

Feature Selection

Feature selection plays the vital role in classification, Since it can have a significant impact on the classifier's accuracy. This limits the number of data sets dimensions, so that, the data becomes more coherent and simple to employ on it. Filter approaches should be used to achieve results for large datasets in less time. Euclidean Distance is used for feature selection and the formula used for this selection is as follows

$$\mathbf{d(A,B) = \{\Sigma i\ (Ai - Bi)2\ \}½}$$

Where,

Ai and Bi = features

Σ= Summation

d(A,B) =Euclidean distance between features Ai and Bi

Normalization

Normalization is a standard technique regularly implemented as an element of data training for machine learning. The purpose of Normalization is to adjust the values of numeric columns in the database to use a standard scale without biasing variations in value ranges or losing information. In this dataset, two attributes plasma glucose and diabetes pedigree function has variation in their value ranges from 44 to 199 and 0.078 to 2.42 respectively. Hence, Normalization is applied to fix this problem.
The Z-score normalization is used here for normalizing the data. Thus the formula for z-score normalization is as:

$$\mathbf{N=(ai-\mu)/\ \sigma}$$

Where,
N=normalized value of an attribute
ai=original value of an attribute
μ = mean
σ = standard deviation

5.1.3 Ensemble Model

Ensemble is a Machine Learning technique whose methods are meta-algorithms that combine several machine learning techniques into one optimal predictive model in order to reduce variance, bias or improve predictions. This approach enables improved predictive performance when compared to that of a single model. There are various methods of ensembling such as bagging, boosting, ada-boosting, stacking, voting, averaging etc. We have applied voting based ensembling method on PIMA Indian diabetes dataset.

Thus the Fig no 2 illustrates the pictorial representation of work flow of the proposed model.

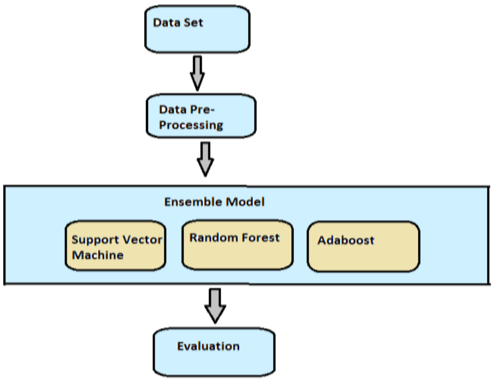

Fig no 2. An Ensemble Model System flow

To handle the class imbalance issue, the AdaBoost approach was espoused. Besides, the Random forest with AdaBoost and SVM are used for classifier model constrcution. SVM is used as the weak classifier, and Random forest with AdaBoost is used as the strong classifier.

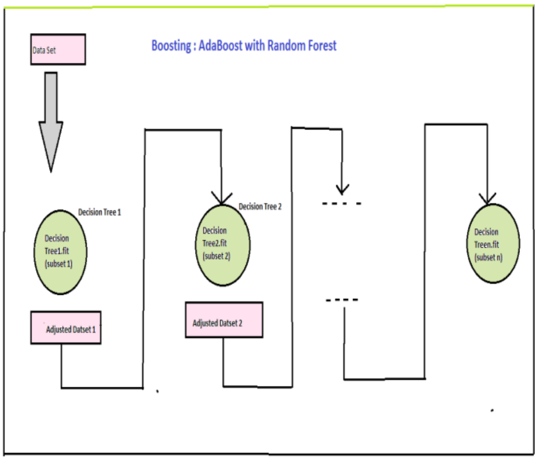
Data Set
Boosting : AdaBoost with Random Forest
Decision Tree 1
Decision Tree1.fit (subset 1)
Adjusted Datset 1
Decision Tree 2
Decision Tree2.fit (subset 2)
Adjusted Datset 2
Decision Treen.fit (subset n)

Fig no.3. Adaptive Boosting Algorithm with Random Forest

Functioning of Random Forest Algorithm with Adaboost

Random Forest is one of the so-called classification methods because a learner's committee is created and each casts a vote for a given instance's predicted label.

The Working of Random Forest algorithm is as follows

Step 1 − First, selection of random samples from a given dataset should be done.

Step 2 – Construction of a decision tree for a given sample.

Step 3 – Weighted error rate of decision tree1 to be calculated.

Step 4 – Weight of this decision tree1's weight in the ensemble to be calculated.

Step 5 – Increase the weight of wrongly classified points, so that same dataset with updated weight will be the output.

Step 6 - Repeat the steps 2 -5 until reaching the targeted number of trees to train

Step 7 – Get the final predicted result from the decision tree n.

Functioning of Support Vector Machine

The support vector machine algorithm's purpose is to discover a hyperplane in an N-dimensional space which classifies the data points distinctly. SVM is a supervised algorithm for machine learning that can be used for problems of classification or regression.

The Working of SVM is as follows

Step 1 - Create a hyperplane that separates the dataset into classes.

Step 2 - Find the points, called support vectors that are closest to both the classes.

Step 3 - Next, find the nearness (Margin) between dividing plane and the support vectors.

Step 4 - The aim of an SVM algorithm is to maximize this very margin. When the margin reaches its maximum, the hyperplane becomes the optimal one.

Step 5 - The SVM model tries to enlarge the distance between the two classes by creating a well-defined decision boundary.

Initially, an investigation is performed using support vector unit, then experiment is performed using random forest ensemble with Adaboost. The simulation is undertaken using information gathered from the repository of the UCI machine. At last, there is a comparison of findings and a conclusion statement is made. And accuracy of these algorithms is measured. The concluded statement declares that ensemble technique is performing better than other classifiers in terms of measuring accuracy level.

6. Suggestions and supplementary enrichments

- **Diverse domains**

However, the strategies can also extended to different sectors such as accounting, ecommerce, banking, marketing, molecular biology, etc.

- **Diverse repositories on Healthcare domain**

Similar methods can be extended to different clinical data sets such as autism, diabetes, heart disease, psychological disorders, etc.

- **Ensemble various approaches**

For higher performance, several other individual models like linear system, regression, etc. can also be used in bagging or boosting units.

- **Future of datamining with cloud**

Additionally, these machine learning techniques can be provided as a cloud service

7. Conclusions

The severity of Diabetes Mellitus is expected to rise by 55% by 2035. There are several ways of diagnosing the disease; data mining algorithms are one of the best. Due to the rapid production of clinical data, data mining techniques are taking an active role in supporting decision-making and prediction systems in the healthcare field, primarily diabetes mellitus. In this sense, in order to detect diabetes, an experiment is sought out on several reports and research by other authors. After reviewing different research papers, the general conclusion is that rather than using single methodology, DMT ensemble paradigms can be used to function better in terms of accuracy, margin of error, computational time and reliability. This new solution has been implemented with the random forest technique with Adaboost and SVM and offers highest precision rate than other algorithms.

REFERENCES

1. Han Wu, Shengqi Yang , Zhangqin Huang, Jian He, Xiaoyi Wang, “Type 2 diabetes mellitus prediction model based on data mining”, Informatics in Medicine Unlocked vol.10, pp. 100–107, 2018.
2. Ramalingaswamy Cheruku, Damodar Reddy Edla, Venkatanareshbabu Kuppili, Ramesh Dharavath ,”RST-BatMiner: A fuzzy rule miner integrating rough set featureselection and Bat optimization for detection of diabetes disease “,Applied Soft Computing, vol. 67, pp. 764–780, 2018.
3. Fikirte Girma Woldemichael, Sumitra Menaria ,”Prediction of Diabetes using Data Mining Techniques”, Proceedings of the 2nd International Conference on Trends in Electronics and Informatics ,IEEE Conference Record: # 42666; IEEE Xplore ISBN:978-1-5386-3570-4, DOI: 10.1109/ICOEI.2018.8553959, 2018. (IEEE Transactions)
4. Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, and Hua Tang,”Predicting Diabetes Mellitus With Machine Learning Techniques”, Frontier Genetics, vol. 9, pp. 515, 2018
5. Beatriz López, Ferran Torrent-Fontbona, Ramón Vinas,José Manuel Fernández-Real,”Single Nucleotide Polymorphism relevance learning with RandomForests for Type 2 diabetes risk prediction”, Artificial Intelligence in Medicine, vol. 85,pp. 43–49, 2018

6. Himansu Das, Bighnaraj Naik, H. S. Behera, Himansu Das,"Classification of Diabetes Mellitus Disease (DMD): A Data Mining (DM) Approach", Progress in Computing, Analytics and Networking, vol. 710, pp 539-549, 2018.
7. Ioannis Kavakiotis , Olga Tsave , Athanasios Salifoglou , Nicos Maglaveras, Ioannis Vlahavas , Ioanna Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal, vol.15, pp. 104–116, 2017.
8. Md. Maniruzzaman , Nishith Kumar , Md. Menhazul Abedin , Md. Shaykhul Islam , Harman S. Suri , Ayman S. El-Baz , Jasjit S. Suri, "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm",Computer Methods and Programs in Biomedicine, vol. 152, pp. 23–34, 2017.
9. Ramalingaswamy Cheruku, Damodar Reddy Edla, Venkatanareshbabu Kuppili, "SM-RuleMiner: Spider monkey based rule miner using novel fitnessfunction for diabetes classification", Computers in Biology and Medicine, vol.81, pp. 79–92, 2017.
10. Santiago Esteban, Manuel, Rodríguez Tablado, Francisco E. Peper, Yamila S. Mahumud, Ricardo I. Ricci, Karin S. Kopitowski , Sergio A. Terrasa, "Development and validation of various phenotyping algorithms for Diabetes Mellitus using data from electronic health records", Computer Methods and Programs in Biomedicine, vol. 152, pp. 53–70, 2017.
11. Sajida Perveen, Muhammad Shahbaz, Aziz Guergachi, Karim Keshavjee, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes, Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia, Procedia Computer Science vol. 82, pp. 115 – 121, 2016.
12. YoichiHayashi n, ShonosukeYukita ,"Rule extraction using Recursive-Rule extraction algorithm with J48 graft combined with sampling selection techniques for the diagnosis of type2 diabetes mellitus in the Pima Indian dataset ", Informatics in Medicine Unlocked, vol. 2, pp. 92–104,2016.
13. Tahani Daghistani, Riyad Alshammari, " Diagnosis of Diabetes by Applying Data Mining Classification Techniques:Comparison of Three Data Mining Algorithms", (IJACSA) international Journal of Advanced Computer Science and Applications, Vol. 7, No. 7,2016.
14. Tarig Mohamed Ahmed," Using Data Mining to develop model for Classifying Diabetic Patient control level Based on Historical Medical Records", Journal of Theoretical And Applied Information Technology, vol.87. no.2, 2016.
15. K.Vembandasamy, T.Karthikeyan, " Novel Outlier Detection In Diabetics Classification Using Data Mining Techniques", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 2 , pp 1400-1403, 2016.
16. Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, "Diagnosis Of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.5, No.1, 2015.
17. Thirumal P. C. And Nagarajan N., "Utilization Of Data Mining Techniques For Diagnosis Of Diabetes Mellitus - A Case Study",, ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 1, 2015.
18. Jia Zhu, Qing Xie , Kai Zheng ,"An improved early detection method of type-2 diabetes mellitus using multiple classifier system", Information Sciences, vol. 292, pp. 1–14, 2015.
19. J. Pradeep Kandhasamy, S. Balamurali, "Performance Analysis of Classifier Models to Predict Diabetes Mellitus", Procedia Computer Science, vol. 47,pp. 45 – 51, 2015.