Impact of Classification Algorithms in Diabetes Data: A Survey

Velmurugan T
D.G.Vaishnav College

Impact of Classification Algorithms in Diabetes Data: A Survey

K.Saravananathan[1], T.Velmurugan[2]
[1]Assistant Professor, SRM Arts and ScienceCollege, Kattankulathur, KanchipuramDt, Tamil Nadu, India.
[2]Associate Professor, PG and Research Department of Computer Science, D. G. Vaishnav College, Chennai, India
E-Mail: [1]greatsaro@yahoo.co.in, [2]velmurugan_dgvc@yahoo.co.in

***Abstarct:*Data Mining is one of the most motivating areas of research in medical field that is become increasingly popular in health organizations. Data Mining plays an important role for uncovering new trends in healthcare organization which in turn helpful for all the parties associated with this field. Diabetes referred to by doctors as diabetes mellitus, describes a group of metabolic diseases in which the person has high blood glucose (blood sugar), either because insulin production is inadequate, or because the body's cells do not respond properly to insulin, or both. Patients with high blood sugar will typically experience polyuria (frequent urination), they will become increasingly thirsty (polydipsia) and hungry (polyphagia). This survey explores the utility of various Data Mining techniques such as classification algorithms and in particular, the use of classification algorithms C4.5, C5.0 and kNN. Classification techniques have been widely used in the medical field for accurate classification of decease. Finally, this research work suggests among the considered algorithms, a comparative study is carried out and the best algorithm is identified from the different researcher's viewpoints based on its performance.**

Key words: Diabetes Data, C4.5 Algorithm, k-Nearest Neighbor Algorithm, Classification Algorithms.

I. INTRODUCTION

Data mining refers to extracting or "mining" knowledge from store house of data. Data Mining is the process of extracting unknown knowledge from large volumes of raw data. Data Mining (DM) is the process of analyzing large quantities of data and summarizing it into useful information. Data mining has been defined as "the nontrivial extraction of previously unknown, implicit and potentially useful information from data.There are several major data mining techniques have commenced and used in data mining. Data mining techniques are used in health management for, Diagnosis and Treatment, Healthcare Resource Management, Customer Relationship Management and Deceit Anomaly Detection. Data mining techniques can help Physicians identify adequate treatments and best practices, and Patients receive better and more affordable harmonize services. Data mining applications in healthcare can have immense potential and usefulness. However, the success of healthcare data mining pivots on the availability of clean healthcare data. In this respect, it is critical that the healthcare industry look into how data can be better apprehended, stored, prepared and mined.

Data mining is the process of examining data from different perspectives and capitulatingit into useful information. The main goal of data mining is to discover new patterns for the users and to discern the data patterns to provide meaningful and useful information for the users. Data mining is applied to find useful patterns to help in the essential tasks of medical scrutiny and treatment. This project aims for mining the relationship in Diabetes data for efficient allocation. The data mining methods and techniques will be inquired to identify the suitable methods and techniques for efficient classification of Diabetes dataset and in mining useful patterns.

The organization of this paper is analytic as follows. Section II discuss about the use of diabetes data using data mining techniques. Section III states the role of classification algorithms for diabetes data. Section IV deals with C4.5, kNN and ID3 algorithms for diabetes data set. Finally conclusion is given in section V.

II. DIABETES DATA ANALYSIS USING DATA MINING TECHNIQUES

Diabetes is the most banal endocrine disease in all populations and all age groups. According to the World Health Organization, it impinges around 194 million people worldwide and that number is expected to increase to at least 300 million by 2025. Diabetes has become the fourth enhancing cause of death in developed countries and there is abundant evidence that it is reaching wide ranging proportions in many developing and newly industrialized nations with evidence pointing to avoidable factors such as sedentary lifestyle and poor diet.

Diabetes describes a metabolic incoherence characterized by chronic hyper glycaemia with disturbances of carbohydrate, fat and protein metabolism resulting from defects in insulin secretion, insulin action, or both. The future complications include progressive development of the specific complications of retinopathy with potential blindness, nephropathy that may lead to renal failure and/or neuropathy with risk of foot ulcers, amputation, Charcot joints and features of autonomic dysfunction, including sexual dysfunction, known as micro vascular complications. People with diabetes are also at a greatly increased risk of cardiovascular, peripheral vascular and cerebrovascular disease, known as macro vascular aggravations. In many research the Pima diabetes datasets are taken for analyzing performance of classification algorithms. The Pima are the groups of Indian peoples live in Southern Arizona. The positive and negative record of Pima data set complicates the classification task.Abdulla et al [1] worked on predictive analysis of diabetic treatment using a relapsed based data mining technique. The support vector machine algorithm was used for experimental analysis. The data sets Datasets of Non Communicable Diseases (NCD) was analyzed for finding out the effectiveness of different treatment types for different age groups.

Diabetes data using various data mining techniques which involved, Naive Bayes, J48, Neural networks, Decision trees, kNN, Fuzzy logic and Genetic Algorithms based on accuracy and time was analyzed by VelidePhani Kumar et al [2]. They found that out of various data mining techniques which were employed to analyze the diabetes data. J48 took least time. A decision support system which combined the strengths of both OLAP and data mining was developed by RupaBagdi et al [3]. This system would predict the future state and generate useful information for effective decision making. They also compared the result of the ID3 and C4.5 decision tree algorithms. The system could discover hidden patterns in the data and it also enhanced real-time indicators and discovered bottlenecks and it improved information visualization.

A research work carried out by K. Rajesh et al. [4] to classify Diabetes Clinical data and predict the likelihood of a patient being affected with Diabetes. The training dataset used for data mining classification was the Pima Indians Diabetes Database they applied Different classification techniques and found out that C4.5 classification algorithm was the best algorithm to classify the data set. Jayalakshmi and Santhakumaran [5] use the ANN method for diagnosing diabetes, using the Pima Indian diabetes dataset without missing data and produce 68.56% classification accuracy. In study (Pradhan and Sahul, 2011) suggested an ANN based classification model for classifying diabetic patients. It shows the average accuracy of 72.2%. The three classification algorithms naïve Bayes, IB1 and the C4.5 to predict the diabetes control used and analyzed by Huang et al. [6], they gives some best idea about the algorithms. The c4.5 has got the highest accuracy of 95% and proved that the c4.5 is the stable classifier. Christobel and Sivaprakasam [7] applied kNN method to the Pima Indian Diabetes dataset. With 10-fold cross validation it gives the average accuracy of 71.94%.The algorithm is improvised and it gains the accuracy for the same dataset as much as 73.38%.

III. ROLE OF CLASSIFICATION ALGORITHMS FOR DIABETES DATA

In present era various public and private healthcare institutes are producing enormous amounts of data which are difficult to handle. So, there is a need of powerful automated Data Mining tools for analysis and interpreting the useful information from this data. This information is very valuable for healthcare specialist to understand the cause of diseases and for providing better and cost effective treatment to patients. The combination of four supervised machine learning algorithms, Classification and Regression Tree (CART), Adaboost algorithm, Logiboost algorithm, Grading algorithm. The experimental result shows the performance analysis of different meta-learning algorithms and also compared on the basis of misclassification and correct classification rate, the error rate focuses True Positive, True Negative, False Positive and False Negative and Accuracy [8]. The researchers used different classification techniques like RIPPER classifier, Decision Tree, Artificial neural networks (ANNs), and Support Vector Machine (SVM) are analyzed on cardiovascular disease dataset [9].

In different classification algorithms the researchers used K-means clustering algorithm and prepare the ANN technique for the heart diseases. Also here the MAFIA algorithm used to extract data appropriate to heart attack from warehouse. The ANN is trained with the selected patterns for the effective prediction of heart attack [10].Hu et al. used different classification method suchas LibSVMs, C4.5, BaggingC4.5,AdaBoostingC4.5, and Random Forest on seven Microarray cancer data sets. Numerous Microarray data classification algorithms have been proposed in recent years. Most of them have been adapted from current data mining and machine learning algorithms [11]. This research work performed comparative analysis of above mentioned classification method using 10-fold cross validation approach on the data set obtained from Kent Ridge Bio Medical Dataset repository.

Real-life data mining applications are interesting because they often present a different set of problems for data miners. One such real-life application that we have done is on the diabetic patients databases. Valuable lessons are learnt from this application. In particular, we discover that the often neglected preprocessing and post-processing steps in knowledge discovery are the most critical elements in determining the success of a real-life data mining application [12].

IV. ID3 AND k-NN ALGORITHMS FOR DIABETES DATA.

Thek-nearest neighbor algorithms look suitable for the most similar instances; the whole dataset should be searched. Many of researchers in them research find the classification accuracy on six public datasets is comparable with C5.0, ID3 and kNN which has a few representatives from training dataset with some extra information to represent the whole training dataset the selection of each representative they used the optimal. The kNN Model significantly reduces the number of the data tuples in the final model for classification with a 90.41% reduction rate on average.

In Raikwal, J. and Saxena, K. [13] did a research over a medical data set they made a comparison between kNN and SVM them result was after implementing the two algorithm showed that kNN is a quit good classifier but when applying kNN algorithm over small data set and it is accuracy decrease when it applies over large data set it performs poor results. Karegowda, A. G., Jayaram, M. and Manjunath, A. [14] made a paper using cascading k-means clustering and kNN classifier over diabetic patient them result was quite good.kNN approach has been used in different data analysis applications such as pattern recognition, data mining, databases and machine learning due to its simplicity and high accuracy. It has been recognized as one of the top 10 algorithms in data mining [15]. Pima Indian diabetes dataset is complex due to its missing values. A class wise k-nearest algorithm have been designed and tested against Pima Indian diabetes dataset. Here testing data is classified in to class label corresponding to the lowest distance. Accuracy achieved for C-kNN is 78.16% [16].

A class of decision tree learning methods to perform supervised, batch (non-incremental) inductive learning (e.g., concept learning) and classification tasks. ID3 and ASSISTANT are two instances of this class. The learned decision tree should capture relevant relationships between attributes' values and class information [17]. Common Disease diagnosis system using data mining techniques namely ID3, Neural Network is implemented in [18] using .NET platform .its Web-based, user-friendly, scalable, reliable and expandable system. It can also answer complex “what if” queries which traditional decision support systems cannot. C. Apte and S. Weiss [19] describe the use of decision tree and rule induction in data-mining applications. In this paper also deals some major state-of-the-art tree and rule mining methodologies, as well as some recent advances. [20] A new decision tree induction technique in which uncertainty measure is used for best attribute selection. This is based on the study of priority based packages of SDFs (Sequence Derived Features). The present research work results the creation of better decision tree in terms of depth than the existing C4.5 technique. The tree with greater depth ensures more number of tests before functional class assignment and thus results in more accurate predictions than the existing prediction technique.

The two entropy-based methods i.e. C4.5 and ID3 for which the pre-processed data and the original data are induced from the decision trees. The methods that are explained [21] may tell us that the collections of attributes that describe objects are extended by the new attributes and secondly, the original attributes are replaced by the new attributes. A new hybrid learning algorithm named ELM-Tree is proposed to deal with the over-partitioning problem in DT induction. It adopts the uncertainty reduction heuristics, and embeds ELMs as its leaf nodes when the information gain ratios of all the cut-points are smaller than a given uncertainty coefficient. Besides, a parallel ELM-Tree model is proposed for big data classification, which is proved to be effective in reducing the computational time.

Previous works on parameter optimization as well as results from those studies confirm that learning performances vary widely if the parameter settings changes even on the same dataset. For instance, in [22] the authors discuss the effect that parameters have on the performance of the Evolutionary Algorithms like the population size, the selection method, the crossover, and mutation operators. [23]. The researchers clarify the methods like revise the terminology, which is unclear and confusing, thereby providing a classification of such control mechanisms, and survey various forms of control which have been studied by the evolutionary computation community in recent years. A broad range of ID3 parameter set values were applied to the Grid Search algorithm to explore as large an area of the parameter space as possible whilst keeping processing cost down by using relatively large step sizes. In the same part of the study the same parameter range/step values were applied to the SGA to examine the ability of the SGA to explore the same ID3 parameter set space at a lower processing cost. The second part of the study was an attempt to see how the process of searching for the optimimal ID3 parameter set using SGA can vary by modifying one of the SGA’s parameters, the Crossover rate that may be one of the main parameters affecting the exploration power of the SGA and performance [24].

Tang, Ping-Hung, and Ming-Hseng Tseng [25] are used different classification algorithms crisp k-NN, fuzzy k-NN, and weighting fuzzy k-NN are compared. For weighting of features, two types of coding including binary-coded genetic algorithms (BGA) and real-coded genetic algorithms (RGA) are evaluated. Experiments are conducted on the Wisconsin diagnosis breast cancer (WDBC) dataset and the Pima (PIMA) Indians diabetes dataset, and the classification accuracy, false negative, and computation time. This research work analyses various research papers which are exactly utilized the classification algorithms for the diabetes data classification. A comparison of different approaches given by many researchers is summarized in table 1. Also, table 1 contains the techniques used, their usage and the results of the experiments.

Table 1: Results Comparison

REF.	TECHNIQUES	UTILITY	Results
[1]	Vector Machine Algorithm	Regression based Technique	The elderly diabetes patients should be given an assessment and a treatment plan that is suited to their needs and lifestyles.
[2]	J48, Decision trees, kNN	-	J48 is best suitable
[3]	ID3 and C4.5 algorithms	Decision support system	Discovers hidden patterns in the data and can, it enhances real-time indicators and discovers bottlenecks and it improves information visualization.
[4]	C4.5 classification	Classify the data	A classification rate of 91% was obtained for C4.5 algorithm.
[5]	ANN	Classification accuracy	The various missing value techniques to improve the classification accuracy.
[6]	naïve Bayes, IB1, C4.5	Classification accuracy	The use of data mining as an exploratory tool, particularly as the domain is suffering from a data explosion due to enhanced monitoring and the (potential) storage of this data in the electronic health record.
[7]	kNN method	Average accuracy	Improved kNN’s accuracy with imputation and scaling method is 3.2% higher than the standard Weka's Implementation of kNN.
[8]	CART, Adaboost, Logitboost, Grading algorithms	Gini measure of impurity	Regression technique algorithm is the best as compared to adaboost, logitboost, Grading algorithm.
[9]	ANN and SVM	Comparative study	Cross validation is used to measure the unbiased estimate of these prediction models.
[10]	K-means and ANN	-	The data items are selected perfectly with the help of MAFIA algorithm. K-means clustering algorithm also used for selection of heart attack data set.
[11]	Decision tree, SVM	Comparative study	In different methods are significantly more accurate than C4.5.
[12]	Data mining	knowledge discovery	A step-by-step approach to help the health doctors explore their data and to understand the discovered rules better.
[13]	SVM and kNN	Accuracy	K-NN performs poor results as the size of data set increases it is best fit for small data set. SVM is complex classifier and here we implement leaner kernel.
[14]	k-Means Algorithm and kNN	-	k-means clustering and KNN classifier over diabetic patient them result was quite good.
[15]	Knn algorithm	Simplicity and high accuracy	Each and every algorithm gave best result for different datasets used.
[16]	K-means, kNN, and CkNN	Classification accuracy	kNN is best.
[17]	ID3 algorithm	Categoricalattributes	Decision tree learning, a supervised inductive learning method that can be extended to handle noisy and/or incomplete data.
[18]	ID3 and Neural Network	Common Disease diagnosis system	Relationships between medical factors related to heart disease, to be established.
[19]	Decision tree	Knowledge discovery	Using decision tree algorithm to predict step by step solutions.
[20]	C4.5	Decision tree induction methodology	The uses of new prediction technique by the drug discoverer are clearly demonstrated.
[21]	C4.5 and ID3	ELM-Tree and Preprocessing	A new hybrid learning algorithm named ELM-Tree is proposed to deal with the over-partitioning problem in DT induction.
[22]	ID3	Evolutionary Algorithms	-
[23]	ID3	Meta optimization	-
[24]	ID3	Genetic algorithm	Veryimportant result in choosing the crossover operator

			when using genetic algorithms, particularly in job shop.
[25]	kNN	Classification of diseases	The classification accuracy, false negative, and computation time are reported.

V. CONCLUSION

The main goal of this compendium is to get best classification algorithms that describes given data in multiple aspects. These algorithms are necessary for automatic classification tools. Particularly, this paper analyses several classification algorithms such as Naïve Bayes, Decision trees, k Nearest Neighbor and SVM for their use in medical data such as diabetes disease. Also, this paper has provided the abstract of data mining techniques used for medical data mining besides the diseases they classified. It also analysis the importance of locally frequent patterns and the mining techniques used for the purpose. Most of the medical opinion use different classification algorithms. Only few algorithms give better performance. Comparatively the C4.5 algorithm gives the better performance than the other classification algorithms. Still the extemporization of C4.5 algorithm is required to amplify accuracy, handle large amount of data, reduce the space requirement for large amount of datasets, support new data types and reduce the error rate. Also, this survey finds that the C5.0 algorithm is the potentially suitable algorithm for any kind of medical diagnoses with some constrained extraction. The classification algorithms helpto medical officers in order to decide and conclude the depth of diabetes disease in its prediction. So, this survey work identify that the role of classification algorithms in analyzing diabetes data are decisive and play a dynamic role.

REFERENCES

[1] Aljumah, Abdullah A., Mohammed GulamAhamad, and Mohammad KhubebSiddiqui. "Application of data mining: Diabetes health care in young and old patients." Journal of King Saud University-Computer and Information Sciences, Vol. 25, 2013, pp. 127-136.

[2] VelidePhani Kumar, Lakshmi Velide, “A data mining approach for prediction and treatment of diabetes disease” in international journal of science inventions today Vol. 3, 2014, pp. 626-629.

[3] RupaBagdi, Prof. PramodPatil, “Diagnosis of Diabetes Using OLAP and Data Mining Integration” in International Journal of Computer Science & Communication Networks, Vol. 2, 2012, pp. 314 – 322.

[4] K. Rajesh, V. Sangeetha, “Application of Data Mining Methods and Techniques for Diabetes Diagnosis”, International Journal of Engineering and Innovative Technology, Vol. 2, 2012, pp. 224-229.

[5] Jayalakshmi, T. and A. Santhakumaran, “A novel classification method for diagnosis of diabetes mellitus using artificial neural networks”. International Conference on Data Storage and Data Engineering, Vol. 9, 2010, pp: 159-163.

[6] Huang, Y., P. McCullagh, N. Black and R. Harper, “Feature selection and classification model construction on type 2 diabetic patients data”. Artificial Intelligence in Medicine, Elsevier, Vol. 41, 2007, pp. 251-262.

[7] Christobel, Y.A. and P. Sivaprakasam, “Improving the performance of k-nearest neighbor algorithm for the classification of diabetes dataset with missing values”. International Journal of Computer Engineering and Technology, Vol. 3, 2012, pp. 155-167.

[8] Sanjay Kumar Sen and Dr. Sujata Dash, “Application of Meta Learning for the Prediction of Diabetes Disease”, International Journal of Advanced Research in Computer Science and Management Studies, Vol. 2, 2014, pp. 396-401.

[9] M. Kumari and S. Godara, “Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction”, Vol. 2, 2011, pp. 304-308

[10] S. B. Patil and Y. S. Kumaraswamy, “Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network”, European Journal of Scientific Research, Vol. 31, 2009, pp. 642-656.

[11] Hong. H, Jiuyong Li, Ashley Plank, Hua Wang, and Grant Daggard, "A comparative study of classification methods for microarray data analysis", Proc. of the fifth Australasian conference on Data mining and analytics, Vol. 61, 2006, pp. 33-37.

[12] Hsu, Wynne, Mong Li Lee, Bing Liu, and Tok Wang Ling. "Exploration mining in diabetic patients databases: findings and conclusions." International conference on Knowledge discovery and data mining, 2000, pp. 430-436.

[13] Raikwal, J. &Saxena, K. “Performance Evaluation of SVM and K-Nearest Neighbor Algorithm over Medical Data set”. International Journal of Computer Applications, Vol. 50, 2012, pp. 35-39.

[14] Karegowda, A. G., Jayaram, M. and Manjunath, A. “Cascading K-means Clustering and KNearest Neighbor Classifier for Categorization of Diabetic Patients”. International Journal of Engineering and Advanced Technology**,** Vol. 1, 2012, pp. 147-151.

[15] Wu Xindong, *et al*. Top 10 algorithms in data mining. Knowledge and Information Systems, Vol. 14, 2008, pp. 1-37.

[16] Christobel, Y. Angeline, and P. Sivaprakasam. "A New Class wise k Nearest Neighbor (CKNN) Method for the Classification of Diabetes Dataset”. International Journal of Engineering and Advanced Technology Vol. 2, 2013, pp. 396-400.

[17] Quinlan J.R., “Introduction of decision trees”, Machine learning, Vol. 1, 1986, pp. 81-106.

[18]SellappanPalaniappanRafiahAwang, “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, International Journal of Computer Science and Network Security, Vol. 8, 2008, pp. 108-115.

[19] C. Apte and S. Weiss, “Data Mining with Decision Trees and Decision Rules”, Future Generation Computer Systems, Vol. 13, 1997, pp. 197-210.

[20] M. Singh, P. K. Wadhwa and P. S. Sandhu, “Human Protein Function Prediction using Decision Tree Induction”, International Journal of Computer Science and Network Security, Vol. 7, 2007, pp. 92-98.

[21] Wang, Ran, et al. "Learning ELM-Tree from big data based on uncertainty reduction". Fuzzy Sets and Systems, Vol. 258, 2015, 79-100.

[22] A. E. Eiben and S. K. Smit, “Parameter tuning for configuring and analyzing evolutionary algorithms”. Swarm and Evolutionary Computation, Vol. 1, 2011, pp. 19–31.

[23] A. E. Eiben, R. Hinterding, and Z. Michalewicz, “Parameter control in evolutionary algorithms”, Transactions on Evolutionary Computation , Vol. 3, 1999, pp. 124–141.

[24] Magalhães-Mendes, Jorge. "A comparative study of crossover operators for genetic algorithms to solve the job shop scheduling problem." WSEAS transactions on computers, Vol. 12, 2013, pp. 164-173.

[25] Tang, Ping-Hung, and Ming-Hseng Tseng. "Medical data mining using BGA and RGA for weighting of features in fuzzy k-NN classification." In Machine Learning and Cybernetics, International Conference, Vol. 5, 2009, pp. 3070-3075.