A Compendium on the Contrasting Features of Data Mining Algorithms for the Diagnosis of Diabetes

Jinali Samir Gandhi
D.J.Sanghvi College of Engineering,
Mumbai.

Khushali Deulkar
D.J.Sanghvi College of Engineering,
Mumbai.

ABSTRACT

Data mining technology offers a user-oriented technique to hidden and novel patterns in the data. Because of rising diseases in the world, we are unable to evaluate all types of diseases and how to consume correct medicine for various diseases. Techniques of data mining are quite useful in finding the medicinal decision for the suitable diseases. Diabetes monitoring system is beneficial to diabetes patients. The diabetes system is useful not just for diabetes patients but also for those suspecting they are diabetic. The primary goal of this paper is to conduct a comparative analysis of decision tree algorithms namely ID3 and CART method focusing on diagnosis of diabetes.

General Terms

Data-mining,diabetes-mellitus, diabetes-diagnosis, comparison, cart-algorithm, id3-algorithm.

Keywords

Data mining, diagnosis, diabetes, cart algorithm, id3 algorithm, comparison.

1. INTRODUCTION

Data mining for diagnosing and predicting disease in medical field plays a crucial role. Various data mining algorithms are there which are accessible for comprehensive and deeper understanding of medical data offering solutions for complicated problems. [6]Data mining, in healthcare, may be used for providing evaluation of medical centers in order to supply better resources, detecting and preventing diseases at an early stage; and cost savings from expensive and unwanted medical tests. Many data mining methods are used by various researchers for treatment and diagnosis of various diseases like diabetes, cancer, stroke etc.

Today, diabetes is increasing rapidly because of lack of exercise and obesity. In human body, insulin is most essential hormone and in case it is not correctly produced then huge amount of sugar is thrown out from body and leads to all types of diabetes. [1]In the domain of data mining, many collaborative and other classification approaches are suggested for diabetes diagnosis. Ensemble classifiers are taken into account as definite in prediction and performance precisions when compared to single classifiers. They offer more flexible structure and select amidst various alternatives to give best solution, greater precision and high predictive performance.

2. LITERATURE REVIEW

Many Researchers have worked on this concept and used various techniques for the diagnosis of diabetes, since many years. Different algorithms, methods and platforms were used by different researchers.

Table-1 shows the brief summary of the work developed till now, in this domain, by different researchers.

Table 1. Literature survey on the different methods of diabetes diagnosis.

Previous Work Year and Author Name	Development and Algorithm
(Murthy.et.al, 1994)	Utilized ANNs and confirmed the requirement for replacing and preprocessing missing values in datasets.
(Friedl.et.al 1997)	Used decision trees in a form of flowchart for classifying and predicting different instances with representation using internodes and nodes.
(Turney, 2000)	The framework utilized an ensemble classifier that is grounded on support vector machine, neural networks and fuzzy systems to learn membership functions, which are then to be used in genetic algorithms.
(Rokach.et.al, 2005)	Recommended a framework for diagnosing diabetes for Pima Indian diabetes dataset. SVM classifier was used for predicting diabetes patients. Using F-score feature selection was done and to obtain optimal set of features, k-mean clustering techniques were used.
(Bashir.et.al, 2011)	A hybrid model was recommended for improving diabetes diagnosis and classification precision.

3. ID3

Decision Tree based on Information Gain (ID3) works on the criterion greedy search that uses top-down manner. Entropy is a measure which is used for dividing the examples into subsets. [2] The homogeneity for a given dataset is calculated by it. In case of complete homogeneity, entropy will be zero. Then it is divided equally for having entropy value one. Decrease in entropy forms the basis for information gain. An attribute having extreme entropy would return maximum information gain value. Hence, an attribute of highest information gain will be chosen as the splitting attribute. Then based on discovery of the most homogenous branch, a decision tree will be constructed.

3.1 Run time analysis

The research is carried out on considered group of 640 diabetes datasets that are obtainable easily from online repositories. The evaluation of subjects denotes that both datasets consist of diverse characteristics hence diverse results are gained. Sensitivity, f-measure, specificity and accuracy are used for evaluating performance of such ensemble methods. Shown below are the formulas of such performance measures.

Where FN, TN, FP TP are false negatives, true negatives, false positives and true positives respectively. Fig 1 shows the flow of a decision tree for ID3 algorithm.

[11]For model building, testing and learning RapidMiner5 is used. For handling class variance problem, stratified sampling is used for fewer sick individuals and more healthy persons are there. Furthermore, 10-fold cross endorsement is applied that uses ten percent testing set and ninety percent data as training set. The data is distributed into 10 mutually special sets and every time 9sets are utilized for training and another one set is made use of for testing. Ten times this is repeated so that every time the test and training sets are different. Theoutcomes are then being around over the ten iterations. The comparison of ensemble techniques is presented.

4. CART

A robust data-analysis and DM tool, CART searches automatically for significant relationships and patterns and swiftly uncovers concealed structure even in extremely complex data. Statistical methods were used in the former, whereas in the latter, CART. To detect diabetes, we present CART in this article. [4]The CART technique iteratively separates the data set, in accordance with a principle that maximizes the splitting of classes, generating a tree like decision structure. The CART technique is known technically as binary recursive partitioning. The procedure is binary due to splitting of parent nodes precisely into recursive and two child nodes as the process may be repeated by treating every child node as a parent

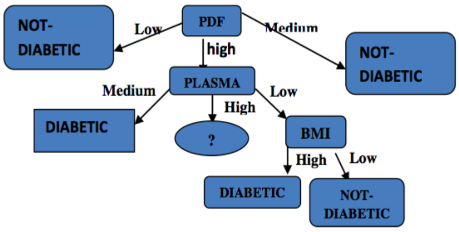

Fig 1: ID3 –Decision Tree

Decision Tree based on Gini index (CART) is useful in measuring the level of impurity of a certain data and a binary tree is constructed wherein every internal node outputs precisely two classes for certain attribute. [8]For each attribute Gini index is calculated and then attribute having lowest Gini index is chosen as the splitting attribute.

4.1 Run Time Analysis

In CART method a very high accuracy rates for non-diabetic patients are there. Of the 240 non-diabetic patients in the dataset, 220 are accurately classified as non-diabetic by the decision tree. Moreover, for the 300 persons classified as non-diabetic by the decision tree, 230 are non-diabetic.

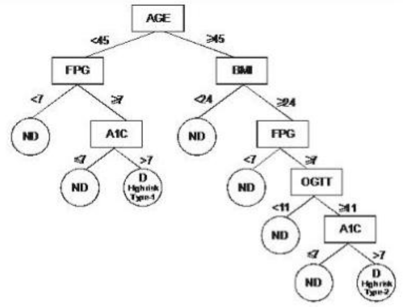

Fig 2: Decision Tree via Cart method

Fig 2, represents the flow of decision tree for Cart algorithm. For predicting diabetic cases, the accuracy rates are lower however still may be deemed to be adequate for the purpose of application of data mining. Particularly for 262 diabetes cases of the dataset, the decision tree correctly classifies 162 as diabetic patients, 69.47 percent accuracy rate. Moreover, the decision tree as truly diabetic classifies 182, or 75.21 percent out of 242 individual

5. COMPARATIVE ANALYSIS

Id3 and Cart algorithms are both used for diabetes detection, diagnosis but they have certain different characteristics and thus certain advantages and disadvantages. Table 2, presents the contrast between the behaviors of the two algorithms.

Table 2. Comparison of the two algorithms on different factors.

Comparative Factor	ID3	CART
Respective Formula	$Entropy = \sum_{i=1}^{m} -p_i log_2 p_i$	$Gain_{ratio} = \frac{\text{Information n gain}}{\text{Splitting info}}$ $Accuarcy = \frac{\text{TP + TN}}{\text{TP + FP + TN + FN}}$
Possibility of Errors	• It uses, the shortest decision tree from learning data,[11] which might not always be the best classification.	• With decision tree’s performance deemed sufficient, it can be again interpreted as follows. [11]The results reveal that the most significant factor connected with the beginning of diabetes, with persons older than age forty showing remarkable higher risk of diabetes when compared to their BMI is second significant factor related to the onset of diabetes.

Advantages	• It is an easy and simple algorithm compared to the other form of algorithm. • Its run time depends on the problem.	• The method of predictions that is used by this process is very accurate. • [11]This proficiency is better than any other process that any analyst will use. • In most of the cases the rendition of most of the problems is quite simple and easy for the users to understand. This simplicity is not only utile for the quick assortment of the new observations, but is also necessary for the explanation of why the particular inputted data was arranged in the particular manner.
Disadvantages	• The major drawback of this algorithm is that the new data that is given as input cannot be modified. • [10]In order to modify the data a new tree has to be made which is very tedious and complex. • These algorithms are much classified in terms of priorities. The priority of any particular data is visible clearly and it is easy for users to differentiate between the more prior and the less prior data types.	• Cart may have many precarious decision trees. It also includes modifications of the data. • There can be changes in the observations that could result in the increase or the decrease in the tree size.

6. CONCLUSION

As we know, diabetes is a threat to the population of the entire world, methods to detect it and cure it at early stages have to be developed. Data mining as we know, uses different algorithms for the early detection. Decision trees are basically used, to check and confirm if all symptoms and conditions point to the problem of diabetes mellitus. Hence in this paper we have compared the two famous decision tree algorithms and come to a conclusion that, CART algorithm has more advantages over ID3 in certain conditions. All of their, performance, data sets, flowcharts etc have been successfully compared for a better understanding of the entire process of diabetes diagnosis. The recommended research concentrates on the improvement of accuracy for diabetes datasets and disease diagnosis performance using decision trees. This will have a vast future scope in efficient diagnosis of diabetes through the use of data mining. It would help the medical professional to predict the disease in a better and efficient manner.

7. REFERENCES

[1] Murthy.et.al, S. (1994). A system for induction of oblique decision trees. Journal of artificial intelligence research, 11-43.

[2] Fayyad.et.al, U. (1996). From data mining to knowledge discovery in databases. AI magazine , 33-54.

[3] Friedl.et.al, M. (1997). Decision tree classification of land cover from remotely sensed data. Remote sensing of environment , 32-77.

[4] Denison.et.al, D. (1998). A bayesian CART algorithm. AFM Smith - Biometrika , 10-43.

[5] Turney, P. (2000). Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. Journal of artificial intelligence research , 44-78.

[6] Mellitus, D. (2005). Diagnosis and classification of diabetes mellitus. Diabetes care , 10-21.

[7] Rokach.et.al, L. (2005). Top-down induction of decision trees classifiers-a survey. Systems, Man, and Cybernetics , 99-189.

[8] Bashir.et.al, S. (2011). An Efficient Rule-based Classification of Diabetes Using ID3, C4.5 & CART Ensembles . College of Electrical and Mechanical Engineering , 33-54.

[9] Kumar.et.al, L. (2012). ID3 Algorithm Performance of Diagnosis For Common Disease . International Journal of Advanced Research in Computer Science and Software Engineering , 55-70.

[10] Jalernrat, S. (2013). Data Mining Using Decision Tree Algorithms. University of the Thai Chamber of Commerce Journal , 11-43.

[11] More.et.al, A. (2013). ID 3 Algorithm. 11-67.