MEDICAL DIAGONOSIS USING DATA-MINING

Zain Momin, Asad Siddiqui, Huzaifa Vakil, Sohel Tharani
Dr. Anupam Choudhary
Department of Computer Engineering,
Rizvi College Of Engineering, Bandra, Mumbai.

Abstract:

The amount of data that is collected from the healthcare industry is enormous and hence to go through that data for the improvement of the diagnosis in the future is next to impossible .In summary, the amount of data that is collected is not mined properly. Even though it will be very advantageous for us and the doctors that are giving the diagnosis to plough through the data for simple and effective diagnosis. Our research, hence, focuses on this niche concept of exploiting the collection of data for diseases like diabetes, hepatitis and heart diseases and to devise a decision tree that will make intelligent decisions medically to help the physicians. Various such decision tree algorithms are C4.5 algorithm, ID3 algorithm and CART algorithm. These types of algorithms are used to compare the effectiveness and correction rate among them.

1. Introduction:

There are at least millions of people that are diagnosed and treated in the healthcare industry on a monthly basis. To go through the diagnosis and the treatment of every one of them is inhumane. But in this world of technology, we don't need to do all the work if a set system using specific algorithms can do our tedious work. The amount of information that can be churned into important diagnosis and treatment is unimaginable. The process becomes easier for both the physicians and the doctors concerned to treat the disease and give the proper diagnosis. A support system solely based on the decisions that were made and which can analyze the different factors such as relationships between the past patient and the current patient, all the diseases in the population, history of the family, and test results, would prove very useful and would be revolutionary in the field of medicine. This concept, called the Decision Support System (DSS) is very broad because of many approaches and the domain which has no end as the patients in the healthcare industry are never-ending based on which decisions are made. It can be summarized as the computerized system that helps make decisions. A DSS application consists of many subsystems .However, the development of a system like this one is a very daunting and tedious task. Many factors are attributed in making such a system but inadequate information has been identified as a major challenge. Hence it, has become our responsibility in such a technical age to take care of such problems like inadequate information and make powerful medical decision support system (MDDS) to support the exploding demand and the increasingly difficult and excruciating diagnosis decision process.

The medical diagnosis is a very difficult process as there are various factors that come into the picture. Every patient has their own story to tell and their own background. To suffice such information and provide diagnosis using specific algorithms is very difficult and that is why we use the concepts of soft computing methods such as decision tree classifiers have shown great potential to be applied in the development of MDSS of various diseases, majorly aimed at diseases like heart diseases, diabetes and many other. The aim is identification of the most important risk factors based on the classification rules that need to be extracted.

2. Overview of related work:

According to the World Health Organization (WHO) fact sheet on diabetes, an estimated 3.4 million deaths are caused due to high blood sugar. Also heart disease is a leading cause of death in world, WHO claims 3.8 million and 3.4 million deaths in males and females, respectively.

Up to now, many studies have reported that they have focused on medical diagnosis. Data Mining Techniques Used in Diagnosis System Classification technique is the most frequently used data mining tasks with a majority of the implementation of Bayesian classifiers, neural networks, and Association Rule. The data mining techniques that have been applied to medical data include Apriori and FPGrowth, decision tree algorithms like ID3, C4.5, C5, and CART, and, Naïve Bayesian, combination of K-means, Self Organizing Map (SOM) and Naïve Bayes. Different approaches have been applied on these studies to achieve high accuracies i.e. 77% or higher using different datasets. Some examples are given below:

D. Senthil Kumar, G. Sathyadeviand S. Sivanesh's results concluded the idea of medical diagnosis and decision tree algorithm for effective classification and also found 83.184% accuracy with cart algorithm which is greater than ID3.

Robert Detrano's experimental results showed correct classification accuracy of approximately 77.00% with a logistic-regression-derived discriminant function.

Vector Machines gets support from Ischemic Heart Disease (IHD) which have high accuracy and serve as excellent classifiers and predictors. Nonlinear proximal Support Vector Machines (PSVM) uses Classifiers when it is tree based.

Polat and Gunes designed an expert system to diagnose the diabetes disease based on principal component analysis. To diagnose the diabetes, Polat et al. developed a cascade learning system.

The Centers for Disease Control and prevention shows gestational diabetes which was presented by National Center for Chronic Disease Prevention and Health Promotion.

The new framework known as duo-mining tool was presented by Jaya Rama Krishniah et al., which is used for diagnosing diabetes. Many classification algorithms like KNN, SVM, and decision tree was applied by Jaya Rama Krishniah for type-2 diabetes. SVM algorithm has highest accuracy among all the algorithm with value of 96.39%.

Aljarullah et al., proposed J48 algorithm to diagnose type-2 diabetes which is used for constructing a decision tree. The accuracy of the model is 78.68%.

Adidela et al., presented the type of diabetes by using Fuzzy ID3 method. The author uses the system for predicting the disease from data set as it initially clusters the data and applies the classification algorithms on clustered data. The author presented a combination of classification method where they developed EM algorithm for clustering and fuzzy ID3 algorithm to attain decision tree for each cluster.

G. Parthiban et al, applied Naïve Bayes method to diagnose heart related problems which are occurring in diabetic patients.

3. About the datasets

The aim of this study is evaluation and development of a Clinical Decision Support System for the treatment of patients with Diabetes and Heart Disease. According to survey and World Health Organization (WHO), every year, the leading of deaths is Heart Disease.

Elsevier, in the journal of American College of Cardiology, the death rate due to cardiovascular diseases (CVD) declined by a significant 41 % in US between 1990 and 2016, whereas in Indi it rose by 34 % from 155.7 to 209.1 deaths per one lakh population in the same period. In India, the leading cause of mortality is Cardiovascular Diseases (CVDs). In India, the heart ailments caused more than 2.1 million deaths in all ages that is more than a quarter of all deaths.

Diabetes is also called as diabetes mellitus. It is a disease that result in too much sugar in blood (high blood glucose). Diabetes mellitus is a metabolic disorders in which there are high blood sugar levels. The number of people with diabetes has risen from 108 million in 1980 to 422 million in 2014. The global prevalence of diabetes among adults over 18 years of age has risen from 4.7% in 1980 to 8.5% in 2014. There are currently 246 million diabetic people all over the world. According to the International Diabetes Federation, the number of diabetic patients would rise to 380 million by 2025. Due to the rapid change in lifestyles, the deaths due to diabetes had increased by 50% between 2005 and 2015. Now in India, the seventh most common cause of death is diabetes.

All these datasets used in this study are taken from UCI KDD Archive. Also this study used Cleveland Clinic Foundation dataset known as "Cleveland Clinic Foundation Heart Disease Dataset"

3.1. Experimental Data

We have used three medical datasets namely, heart disease and diabetes datasets. All these datasets are obtained from www.kaggle.com. After classifying the diseases we are going compare the attribute selection measure algorithms such as ID3, C4.5 and CART. The heart disease dataset of 268 patients is used in this experiment and has 76 attributes, 14 of which have a linear valued and are relevant. The diabetic dataset of 874 patients with includes 9 attributes.

NO	NAME	DESCRIPTION
1	Age	Age in years
2	Sex	Male=1; female=0
3	Chest Pain	CP type(typical angina=1; atypical angina=2; non-anginal pain=3; asymptomatic=4)
4	Rest blood pressure	Resting blood pressure(in mm Hg on admit to hospital)
5	Cholesterol	Serum cholesterol in mg/dl
6	Fasting BS	(Fbs>120mg/dl)(true=1; false=2)
7	Thalach	Maximum heart rate
8	Exang	Exercised angina(yes=1; no=0)
9	Oldpeak	ST depression exercise relative to rest
10	Slope	The slope of the peak exercise segment(upsloping = 1; flat=2; down sloping=3)
11	Ca	Number of major vessels(0-3) flourosopy colored
12	Thal	(normal=3; fixed defect=6; reversible defect=7)
13	Rest ECG	restingECG results (0 = normal; 1 = having ST-T wave abnormality; 2 = showing probable or define left ventricular hypertrophy by Estes' criteria)
14	Num	Classes of diagnosis(healthy= 0; possible heart disease=1)

Table 3.1 Heart Disease dataset

Table 3.1.2. Diabetes Dataset

3.2. Measures for Attribute Selection

Machine learning and data mining uses many different metrics to build and evaluate models. We have implemented the ID3, C4.5 CART algorithm and tested them on our experimental datasets. Using confusion matrix produced by them the accuracy of these algorithms can be examined. We employed four performance measures: precision, recall, F-measure and ROC space. In order to calculate the four measures a distinguished confusion matrix (sometimes called contingency table) is obtained. The classification results are represented by confusion matrix. It contains information about predicted and actual classifications done by a classification system. The cell which denotes the number of samples classifies as false while they were false(i.e., TN), and the cell that denotes the number of samples classified as true while they were actually true (i.e., TP). The number of samples misclassified are denoted by other two cells. Specifically, the cell denoting the number of samples classified as true while they actually were false (i.e., FP), and the cell denoting the number of samples classified as false while they actually were true (i.e., FN). The precision, recall, F-measure are easily calculated once the confusion matrixes are constructed using these formulae:

Recall= TP/ (TP+FN)
Precision = TP/ (TP+FP)
F_measure = (2*TP)/ (2*TP+FP+FN)

No	Attribute Name	Description
1	Number of pregnancies	Numerical values
2	concentration of plasma glucose	glucose concentration in a 2 hours in an oral glucose tolerance test
3	Diastolic blood pressure	In mm Hg
4	Triceps skin fold thickness	Skin thickness in mm
5	2-Hour serum insulin	Insulin (mu U/ml)
6	BMI	(weight in kg/(height in m)^2)
7	Diabetes pedigree function	A function – to analyze the presence of diabetes
8	Age	Age in years
9	Class	1 is interpreted as “tested positive for diabetes and 0 as negative

Less formally, the percentage of the actual patients (i.e. true positive) among the patients that got disease is measured by precision; while, the percentage of the actual patients that were discovered is measured by recall; the balance between precision and recall is given by F-measure. A ROC (receiver operating characteristic) space is defined by true positive rate (TPR) and false positive rate (FPR) as y and x axes respectively, which depicts relative tradeoffs between false positive and true positive.

TPR= TP/ (TP+FN)
FPR= FP/ (FP+TN)

3.3. ID3 Algorithm:

In decision tree learning, ID3 (Iterative Dichotomizer 3) is an algorithm invented by Ross Quinlan used to generate a decision tree from a dataset. ID3 is used in Natural Language Processing (NLP) and Machine Learning domains and is precursor to C4.5 algorithm.

The ID3 algorithm begins with the original set S and consider S as the root node. On every iteration of the algorithm, the algorithm passes and iterates through every unused attribute of the set S and calculates the entropy H(S) of that attribute. The attribute with smallest entropy (or largest information gain) value is selected. The subsets of data is produced by splitting or partitioning the set S. Considering only attributes that are never selected before, the algorithm continues to recur on each subset.

ID3 does not guarantee an optimal solution. The ID3 algorithm's optimality can be improved by using backtracking during the search for the optimal decision tree at the cost of possibly taking longer.

ID3 can overfit the training data. The ID3 algorithm usually produces small trees, but it does not always produce the smallest possible decision tree.

ID3 is harder to use on continuous data. If the values of any given attribute is continuous, then more data is split on this attribute, and searching for the best value to split by can be time consuming.

3.3.1. ID3 Metrics:

3.3.1.1. Entropy:

Entropy H(S) is a measure uncertainty amount in the set S.

$$\mathrm{H}(S) = \sum_{x \in X} -p(x) \log_2 p(x)$$

Where,
S – Calculation of entropy of current dataset.
X – The set of classes in S
p(x) – The proportion of the number of elements of class x to the number of elements in set S.

3.3.1.2. Information Gain:

Information gain is the amount of information that is gained by know value of the attribute, which is the entropy of the distribution before the split minus the entropy of the distribution after it.

$$IG(S, A) = \mathrm{H}(S) - \sum_{t \in T} p(t)\mathrm{H}(t) = \mathrm{H}(S) - \mathrm{H}(S|A).$$

Where,

Position	Schema Name	Table Type Name	Parameter Type
1	*<schema name>*	*<input table type>*	In
2	*<schema name>*	<Parameter table type>	In
3	*<schema name>*	*<Result output table type>*	Out
4	*<schema name>*	*<PMML output table type>*	Out

H(S) – Entropy of set S
T – The subsets created from splitting set S by attribute A such that $S = U_t\ t$
H(t) – Entropy of subset t

3.4. C4.5 algorithm:

An appropriate action or a decision that is made among a given set of actions for a particular condition or case, is by definition, a decision tree. Decision tree is a classifier for the determination of an appropriate action or decision among a set of actions for the given case. Historical association with the different outcomes is analyzed and to effectively identify the factors for different outcomes of the decision. Tree like structures and their possible combinations that are made is by definition a decision tree.The nodes can be a decision node or a leaf node.Leaf node is the node that mentions the value of the dependent (target) variable.Decision node is the node that contains one condition that specifies some test on an attribute value. The division of branches happen in categories such as subtrees or nodes, these divisions are the result of condition.

Using the concept of information entropy, C4.5 algorithm builds decision tree from a set of training data. The set of samples that are classified are by definition, a training set. C4.5 chooses one attribute of the data at each node of the tree that most definitively splits it into subsets in one class or the other. Its criteria for the end result is the normalized information gain i.e. the information gain that results from choosing an attribute for splitting the data from the training set. The attribute that is given the responsibility of making the decision is the one with the highest normalized information. Highest normalized information gain attribute is selected as the winner. The C4.5 algorithm then proceeds recursively until meeting some stopping criteria such as the minimum number of cases in a leaf node.C4.5 algorithm works through its training set recursively until meeting the number of cases that are there in the leaf node. It works through its cases recursively before meeting the cases that are there in the leaf node.

Discrete and continuous values are supported in the C4.5 decision tree which in turn is supported by PAL. The decision tree functions implemented in support of Pal has support for discrete as well as continuous values. The Pal implementation has the reduced error pruning (REP) algorithm as a method of pruning

3.4.1. Prerequisites

- The column order and column number of the predicted data are the same as the order and number used in tree model building.
- The last column of the training data is used as a predicted field and is of discrete type. The predicted data set has an ID column.
- The table that will store the tree model will be mostly a column table.
- The target column of training data must not have null values, and other columns should have at least one valid value (not null)

Table 3.4.1 Created With C4.5:

This function will create a decision tree from the training data:-

3.4.2. Procedure Generation:

The table of signature should contain the following record:-

3.4.3. Procedure Calling:

The procedure name will be as exact as written in the procedure generation.
The parameters, input and output tables must be of the specified types in the signature table.

Table	Column	Column Data Type	Description	Constraint
Training / Historical Data	Columns	Varchar, nvarchar, integer, or double	Table used to build the predictive tree model	Discrete value: integer, varchar or nvarchar Continous value: integer or double
	Last column	Varchar, nvarchar, or integer	Target variable (class label)	

Table 3.4.2 Signature Table:

Name	Data Type	Default Value	Description	Dependency
Percentage	Double	1	Specifies the percentage of the input data to be used to build the tree model.	
MIN RECORDS OF PARENT	Integer	1	Specifies the stop condition: if the number of records is less than the parameter value, the algorithm will stop splitting	
MIN RECORDS OF LEAF	Integer	0	Promises the minimum number of records in each leaf.	
MAX DEPTH	Integer	Number of columns in the input table which contains the training data	Specifies the stop condition: if the depth of the tree model is greater than the parameter value, the algorithm will stop splitting.	
THREAD NUMBER	Integer	1	Number of threads.	
CONTINUOUS COL	Integer	Detection from input data.	Indicates which columns attributes are continuous. The default behavior is : String or integer Double: continuous	
SPLIT THRESHOLD	Double	1e-5	Specifies the stop condition: if the information gain ratio is less than this value, the algorithm will stop splitting.	
<name of target value>	Double	Detected from input data	Specifies the probability of every class label.	
SELECTED FEATURES	Varchar	Detected from input data	A string to specify the features that will be processed. The pattern is “X_1, Xn”, where X_i is the corresponding column name in the data table. If this parameter is not specified, all the features will be processed.	

DEPENDENT VARIABLE	Varchar	Detected from input data	Column name in the data table used as dependent variable. If this parameter is not specified, the last column of the training data will be used as dependent variable.	
IS OUTPUT RULES	Integer	0	If the parameter turns out to be 1, the decision rules are extracted and saved from the tree model to the result table which in turn is used to save the PMML model.	
PMML EXPORT	Integer	0	0 – PMML does not export tree model 1,2 – PMML exports the tree model	Only valid when *IS OUTPUT RULES* is 0.

Table 3.4.3 optional parameters

Table	Column	Column Data Type	Description	Constraint
Tree model of JSON format	1st column	Integer	ID	
	2nd column	CLOB, varchar, or nvarchar	Tree model saved as a JSON string	Table should be a column. Every unit's minimum length of row is 5000.
Tree model of PMML format	1st column	Integer	ID	
	2nd column	CLOB, varchar, or nvarchar	Tree model in PMML format	

Table 3.4.3. Output Tables of C4.5 Algorithm:

3.5. CART Algorithm:

Classification and regression trees (CART) is a non-parametric technique that will process the dependent variable which can be categorical or numerical and will produce output which will be either classification or regression trees respectively. Based on values of variables in the data set, trees are created by putting rules on this data. Depending on how well a rule splits the values of variables so as to differentiate observations, a rule gets selected. After selection of a rule, the rule is applied recursively to each child node. The rule stops when CART finds no further improvement or if some conditions are met to stop further processing. The whole notion of expanding the tree is to make sure the child nodes being created are the most optimal.

The innovations of CART are:

1. Deciding how big the tree should grow.
2. Using two-way splitting of node into child nodes.
3. Tree validation and automatic testing.
4. Missing values can be found.

3.6. Comparison of CART, ID3 and C4.5 algorithms:

	Type of attribute	Missing values	Strategy for pruning	Criteria for splitting	Outlier detection
ID3	Categorical values processed	Can't handle missing values	No pruning	Gain of information	Vulnerable to outliers
CART	Categorical and Numeric	Can handle missing values	Cost-complexity	Towing criteria	Outliers can be handled

	values processed		pruning		
C4.5	Categorical and Numeric values processed	Can handle missing values	Error based pruning	Gain ratio	Vulnerable to outliers

Table 3.5.1 comparison of all algorithms

4. Conclusion

Decision tree algorithms are some of the most efficient and powerful methods of classification. The data sets we have used will test the efficiency and correction rate of the algorithms. As expected through our intuition and prior knowledge of the algorithms, CART algorithm showed the best performance of all the algorithms. The outcomes we have reached will give great advances to doctors making decisions on a patient with heart disease or diabetes. We have surveyed the data through decision tree algorithms ID3, CART and C4.5 and we have concluded that CART algorithm has the best rate of success with regards to performance of rules and accuracy. The results of all this processed data is stored in a database so doctors can access it with ease and to allow them to look at the data and form similarities with the diagnoses. To further improve this decision support, the medications that patients are consuming should also be added to the data set to make the results even more accurate.

References:

[1] Umair Abdullah (2008). “Analysis of Effectiveness of Apriori Algorithm in Medical Billing of Data Mining1” IEEE. pp. 1-5.

[2] Jenn-Lung Su, Guo-Zhen Wu, I-Pin Chao (2001). “THE APPROACH OF DATA MINING METHODS FOR THE MEDICAL DATABASE” IEEE. pp. 1-3

[3] Safwan Mahmud Khan Md. Rafiqul Islam Morshed U. (n.d). “Medical Image Classification Using of an Efficient Data Mining Technique” IEEE, pp. 1-6.

[4] Yanwei Xing, Jie Wang and Zhihong Zhao (2007). “Combination of data mining methods with new medical data to predicting outcome of Coronary Heart Disease” IEEE. pp. 1-5.

[5] Ranjit Abraham, Jay B.Simha, Iyengar (n.d). A comparative analysis of discretization methods for Medical Data mining with Naïve Bayesian classifier. IEEE. pp. 1-2.

[6] Syed Zahid Hassan and BrijeshVerma,(n.d). A Hybrid Data Mining Approach for a Knowledge Extraction and Classification in Medical Databases. IEEE. pp. 1-6.

[7] National Center for Chronic Disease Prevention and a Health Promotion. Gestational Diabetes. Centers for Disease Control and Prevention. U.S. Department of Health and Human Services; 2011. Available from: http://www.cdc.gov/

[8] Jaya Rama Krishnaiah VV, Chandra Shekar DV, Satya Prasad R, Rao KRH. An study about type-2 diabetes suing duo mining approach. International Journal of Computational Engineering Research. 2012; 2(6):33–42.

[9] AljarullahAA. Decision tree discovery for the diagnosis of a type II diabetes. International Conference on Innovative in Information Technology; 2011. pp. 303–7.

[10] Adidela DR, Lavanya DG, Jaya SG, Allam AR. Application of fuzzy ID3 to predict diabetes. Int J AdvComput Math Sci. 2012; 3(4):541–5.

[11] http://www.ics.uci.edu/~mlearn/MLRepository.html