ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 12, December 2017

Application of Classification Algorithms for Disease Diagnosis Using Big Data Analytics

[1] Shobana.V, [2] Dr.K.Nandhini
[1] Ph.D Research Scholar, [2] Assistant Professor
Department of Computer Science, Tirupur, Tamilnadu, India. Chikkanna Government Arts College,

Abstract: **- A numerous amount of data is generated in the healthcare sector and mining of those data yields good results. It will also be helpful in diagnosing a particular disorder or a disease and also helps in predicting the future of the disease occurrence. Various classification algorithms in data mining have been used in research to predict a particular disease. Classification is used to find out in which group each data instance is related within a given datas et. It classifies the data into different classes based on some conditions There are many classification algorithms which includes C4.5, ID3, k-nearest neighbor, Naive Bayes, SVM, and ANN etc., which are used for classification. There are three forms of classification approaches, namely Statistics, Machine Learning and Neural Network for classification. The main objective of this study is to provide a compact source of reference for the researchers who want to use decision tree which is an important tool of data mining technology in their area of work. With this aim in mind, we compared widely used decision tree algorithms to classify types of disease and compared their performances according to six performance metrics (ACC(%), MAE, PRE, REC, FME, and Kappa Statistic). We hope that this study can provide a useful overview of the current work in this field and highlight how to apply decision tree algorithms as a tool of data mining technology. While considering these approaches this paper provides an inclusive survey of different classification algorithms of decision trees and their features and limitations.**

Keywords: **Big data, decision trees, classification, disease, performance metrics, distributed decision trees.**

I. INTRODUCTION

Classification techniques are capable of processing a large amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels. It can be used for classifying the newly available data. Classification procedure is a recognized method for repeatedly making such decisions in new situations. Here if we assume that problem is a concern with the construction of a procedure that will be applied to a continuing sequence of cases in which each new case must be assigned to one of a set of pre defined classes on the basis of observed features of data. Creation of a classification procedure from a set of data for which the exact classes are known in advance is termed as pattern recognition or supervised learning. Contexts in which a classification task is fundamental include, for example, the initial diagnosis of a patient's disease in order to select immediate treatment while awaiting perfect test results. Three main historical stands of research can be identified: statistical, machine learning and neural network. All groups have some objectives in common.

(i) Statistical Procedure Based Approach

Two main phases of work on classification can be identified within the statistical community. The first "classical" phase concentrated on extension of Fisher's early work on linear discrimination. The second, "modern" phase concentrated on more flexible classes of models many of which attempt to provide an estimate of the joint distribution of the features within each class which can in turn provide a classification rule [1]. Statistical procedures are generally characterized by having an precise fundamental probability model which provides a probability of being in each class instead of just a classification. Also it is usually assumed that the techniques will be used by statisticians and hence some human involvement is assumed with regard to variable selection and transformation and overall structuring of the problem.

(ii) Machine Learning Based Approach

Machine Learning is generally covers automatic computing procedures based on logical or binary operations that learn a task from a series of examples. Here we are just concentrating on classification and so attention has focus on decision-tree approaches in which classification results from a sequence of logical steps. These classification results are capable of representing the most complex problem given sufficient data. Other techniques such as genetic algorithms and inductive logic procedures (ILP) are currently under active improvement and its principle would allow us to deal

All Rights Reserved © 2017 IJERCSE

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 12, December 2017

with more general types of data including cases where the number and type of attributes may vary. Machine Learning approach aims to generate classifying expressions simple enough to be understood easily by the human and must mimic human reasoning sufficiently to provide insight into the decision process [1].Like statistical approaches background knowledge may be used in development but operation is assumed without human interference.

(iii) Neural Network

The field of Neural Networks has arisen from diverse sources ranging from understanding and emulating the human brain to broader issues of copying human abilities such as speech and can be use in various fields such as banking, legal, medical, news, in classification program to categorize data as intrusive or normal. Generally neural networks consist of layers of interconnected nodes where each node producing a non-linear function of its input and input to a node may come from other nodes or directly from the input data. Also, some nodes are identified with the output of the network.

On the basis of this example there are different applications for neural networks that involve recognizing patterns and making simple decisions about them. In airplanes we can use a neural network as a basic autopilot where input units reads signals from the various cockpit instruments and output units modifying the plane's controls appropriately to keep it safely on course. Inside a factory we can use a neural network for quality control.

II DECISION TREE ALGORITHMS IN DATA MINING FOR CLASSIFYING DISEASES

Decision trees are trees that classify instances by sorting them based on feature values Given a set S of cases, C4.5 first grows an initial tree using the divide-and-conquer algorithm as follows: ಠ If all the cases in S belong to the same class or S is small, the tree is a leaf labeled with the most frequent class in S. Otherwise, choose a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition S into corresponding to subsets S1, S2… according to the outcome for each case, and apply the same procedure to each subset[7]. Decision Trees are the most popular architectures widely used in data mining [8]. These architectures use a divide-and-conquer strategy in order to partition the instance space into decision regions. At first, a root node is designated by using a test. Then, the value of related test attribute splits the data set and, the process is repeated until the determined stopping criterion is provided. At the end of the tree, each node is named as leaf node. Each leaf node denotes the class. Also, each branch indicates a path defined as a decision rule. The classification is handled by using each decision rule for a new sample. As a summary, the decision tree architectures consist of a root node, branches, internal nodes, and leaf nodes. There are three main steps for classification by using decision trees: The first step is the learning process. The model is constructed on the training data. Hence, this model is presented by classification rules. In the second step, a test is selected in order to calculate the model accuracy. The model is accepted according to the value of this test. If this value is considerably accepted, the model could be used for the classification of a new datum. At last, the third step includes the usage of the model for a classification or prediction of a new data [9].

2.1 Decision tree algorithms

(a) J 48

J 48 is modified version of C4.5 [10]. The principle of this algorithm is to use divide-and-conquer strategy. Also, it uses pruning for the construction of the tree in order to avoid over-fitting problem. Maximum gain information is used as a splitting criterion. It calculates overall entropy of the training data and entropy for each attribute according to classes. Then it takes the differences between overall entropy and entropy achieved for each attribute. This value is called as the gain information. Then, the attribute which has the highest gain information is selected for splitting.

$$Entropy\ (S) = -\sum_{j=1}^{n} \frac{freq\ (C_j, S)}{|S|} \cdot \log_2 \frac{freq\ (C_j, S)}{|S|}$$

$$Gain\ (S, A) = Entropy\ (S) - \sum_{i \in A}^{k} \frac{|S_i|}{|S|} Entropy\ (S_i)$$

where S is a training set explained in terms of k attributes, and C ={C1 ,C2 ,...,C n } defines n classes.

(b) CART

Classification and Regression Trees (CART) is a kind of statistical technique. It is a tree structure technique that produces binary trees. Each internal nodes has two outgoing edges according to the selected test attribute. There are various types of impurity criteria by using CART such as Gini Index, Symmetric Gini Index, Twoing, Ordered Twoing, Class Probability for Classification Tree, Least Squares, Least Absolute Deviation for Regression Trees, Multi Variable Splitting Criterion and The Linear Combinations Method. For example, Gini Index is an impurity-based criterion. The pruning of the constructed tree is done by cost–complexity pruning. The divergences among the probability distributions of the target attribute's value are measured by cost-complexity pruning [11].

$$Gini\ (y, S) = 1 - \sum_{c \in dom(y)} \left(\frac{|\sigma_{y=c_j} S|}{|S|} \right)^2$$

All Rights Reserved © 2017 IJERCSE

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 12, December 2017

(c). NBTree
NBTree combines Naive Bayes Classification and Decision Trees [12]. The decision tree constructed by NBTree algorithm uses Naïve Bayes Classifiers. The tree contains univariate splits. This algorithm uses Bayes rule in order to find the probability of each class given the instance. This algorithm assumes that the attributes are conditionally independent given the label. NBTree classifier generally has higher accuracy rate than a naïve Bayes classifier.

(d). BFTree
Best-First decision tree (BFTree) learning process uses the procedure defined for standard decision trees [13]. It handles categorical and numerical variables. While standard decision tree induction process expands in dept-first order, the best-first decision tree induction process expands the "best" node first.

(e). LADTree
LADTree is a classification method. In this approach, decision trees are combined with the predictive accuracy of boosting into a set of classification rules [14]. Then, they are adapted to the multiclass LogitBoost and AdaBoost. LogitBoost algorithm is fused with AdaBoost induction. Multi-class alternating decision trees are built by using this structure. A single variable is selected for the splitter node. The algorithm aims to minimize the least squares between the working return and the mean value of the examples.

(f). REPTree
REPTree is decision tree learner which builds a decision and regression tree [15]. The gain information is used in order to split the data set. It prunes the tree in order to reduce error pruning. The main purpose is not only to decrease the effectiveness of the noisy instances but also to decrease the complication in the classification progress. REPTree is a fast algorithm.

(g). RANDOMTree
RANDOMTree is a multiple random tree algorithm [16]. A non-tested attribute is selected randomly from the whole data set without using a training set. A limit is predefined. The tree has been constructed until the depth of the tree exceeds this predefined limit. If the depth of the tree exceeds this limit, it stops. The training set is used in order to update the statistics of each node. The class attribute contains different classes. Each node sets down the number of records classified as different classes. The same process is applied for the classification by using a decision tree. Each data record is read in order to update multiple random trees. It is necessary to complete one scan of the data. The classification of a datum x is performed by averaging the probability outputs from multiple random trees.

(h). Random forest
Random Forest is a group of tree predictors [17]. A random vector is used. It is sampled independently by using the same distribution θkis handled from the old vectors θ1, θ2, … , θk−1. X is defined as an input vector. The construction of the tree is handled on the training set by using the random vector θk. The resulting is defined with h (X, θk). If a large numbers of trees are generated, they are voted in order to find the most popular class. The procedure is called as random forests. It is a classifier. Each tree has a cost as a vote for the class selected the most popular at input X.

(i). LMT
Logistic Model Trees (LMT) associates tree induction and logistic regression technique [18]. It constructs a single tree. It contains binary splits on numeric attributes, multi-way splits on nominal ones, and logistic regression models at the leaves. The logistic regression states at each parent node and a leaf node gathers all parent models for estimating a probability for each classes. Also, a trimming process is used to generalize the model.

(j). FT
Functional Trees (FT) is a kind of multivariate tree [19]. They are able to explore multiple representation way by using decision tests. This algorithm works as a multivariate tree learning algorithm. Decision nodes with multivariate tests, leaf nodes made predictions using linear functions are used for generating the functional trees. This algorithm can handle missing values. It can be applied into binary, numerical or categorical data.

(k). Decision stump
Decision Stump (DS) is a single level decision tree [20]. It is a kind of machine learning model. The decision tree constructed by DS consists of one root node and this node connects to the leaf nodes. Just a single input attribute is used in order to make a prediction. It is regularly employed with a bagging or boosting algorithm.

2.2 Performance Metrics in Decision tree Algorithms

Classification Accuracy (ACC) is a widely used measure to evaluate a classifier. It is just defined as the degree of right predictions of a classifier. It can take values range from 0 to 100(%).

$$\mathrm{ACC} = \mathrm{Num(Test\ Samples\ Rightly\ Classified)} / \mathrm{Num(Total\ Samples)}$$

Precision (PRE) is a kind of measure. It assures a specific class which has been forecasted. It can be thought as percentage of times that the classifier is correct in its classification of positive samples.

$$\mathrm{PRE} = \mathrm{True\ Positive}/(\mathrm{TruePositive} + \mathrm{FalsePositive})$$

Recall (REC) measure the capability of a prediction model for selecting the samples from the same class.

$$\mathrm{REC} = \mathrm{TruePositive}/(\mathrm{Truepositive} + \mathrm{FalseNegative})$$

Whereas a true positive can be defined as a positive sample identified with the same label correctly, a false positive can

All Rights Reserved © 2017 IJERCSE

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 12, December 2017

be defined as a negative sample identified incorrectly with the positive label. Also, a true negative means that a negative sample identified with the same label correctly. On the other hand, a false negative is a positive sample identified with the negative label incorrectly.

F-Measure (FME) is the harmonic mean of precision and recall.

$$\mathrm{FME} = (2 * \mathrm{PRE} * \mathrm{REC})/(\mathrm{PRE} + \mathrm{REC})$$

Mean Absolute Error (MAE) can be explained as the average of the absolute values of the prediction errors. It demonstrates the deviations from the true probability by calculating the absolute value of differences.

$$\mathrm{MAE} = \frac{\sum_{j=1}^{c}\sum_{i=1}^{m} \left| o(i,j) - p(i,j) \right|}{mxc}$$

Kappa Statistic is a measure of agreement between two categorical variables, X and Y. It is used to make comparison for the ability of different raters. The Kappa Statistic takes values between 0 and 1. The observed agreement between X and Y is shown by p0 and the expected agreement by chance is shown by pe and the kappa statistic is defined as follows:

$$K = \frac{p_0 - p_e}{1 - p_e}$$

Table 1: Comparing the various decision tree algorithms in disease classification

$$\text{Confidence} = \mathrm{P(B/A)} = \frac{\mathrm{P(A \cap B)}}{\mathrm{P(A)}}$$

The above table shows the results of various decision tree algorithms against various performance metrics.

Algorithm	Tree Size	Leaf	ACC(%)	MAE	PRE	REC	FME	Kappa Statistic
NBTree	5	3	75.00	0.1548	0.773	0.75	0.749	0.684
J48	21	11	66.25	0.1673	0.581	0.638	0.603	0.568
LADTree	21	14	66.25	0.1312	0.668	0.663	0.663	0.568
BFTree	17	9	65.00	0.1378	0.666	0.65	0.651	0.556
LMT	1	1	65.00	0.1797	0.622	0.65	0.626	0.545
Random Forest	10 Trees		65.00	0.1264	0.665	0.65	0.646	0.551
FT	5	3	63.75	0.1673	0.581	0.638	0.603	0.529
Random Tree	51		62.50	0.125	0.627	0.625	0.625	0.523
REPTree	13		62.50	0.1702	0.634	0.625	0.625	0.523
CART	21	11	58.75	0.1715	0.577	0.588	0.573	0.470
DS	Single Level		41.00	0.2288	0.185	0.413	0.256	0.239

III. DECISION TREES IN BIG DATA

Jingxiang Chen et al. proposed A Distributed Decision Tree Algorithm and Its Implementation on Big Data Platforms. Their work presents KS-Tree, a distributed decision tree algorithm that scales well in large data sets. The algorithm is implemented using MPI and is now running on top of a wide range of systems, including Hadoop file system and distributed database systems (such as Teradata, Greenplum and Aster). They compare its implementation with the state-of-the-art decision tree techniques implemented in R and Spark with some public data sets of both large and small sizes. The examples show that the new algorithm enjoys competitive prediction accuracy results. They also demonstrates that their algorithm can also be as a variable selection tool. Thus decision trees can be well implemented with larger datasets.

IV. CONCLUSION

This study implies the fact that the decision trees are well suited for classification of various diseases. It is more flexible than any other classification algorithms which classifies the data more accurately based on the class labels. The various performance metrics of decision tree classification helps in finding out the exact state of decision trees. Thus this is a very accurate classifier comparing to all other classifiers.

REFERENCES

[1] D. Michie, D.J. Spiegelhalter, C.C. Taylor "Machine Learning, Neural and Statistical Classification", February 17, (20044).

[2] https://selecthub.com/business-intelligence/bi-vs-big-data-vs-data-mining/

[3] Shaik Razia1 and M. R. Narasinga Rao "Machine Learning Techniques for Thyroid Disease Diagnosis - A Review" in Indian Journal of Science and Technology, Vol 9(28), DOI: 10.17485/ijst/2016/v9i28/93705, July 2016

[4] Zhang GP, Berardi VL. An investigation of neural networks in thyroid function diagnosis. Health Care Management Science. 1998; 1:29–37.

[5] Anupam S, Ritu T, Prabhdeep K, Janghel RR. Diagnosis of thyroid disorders using artificial neural networks. paper presented at the Advance Computing Conference, 2009.

All Rights Reserved © 2017 IJERCSE

[6] Md. Dendi Maysanjaya, Hanung Adi Nugroho, Noor Akhmad Setiawan "A Comparison of Classification Methods on Diagnosis of Thyroid Diseases" 978-1-4799-7711-6/15/$31.00 © 2015 IEEE.

[7] Dr. Rajesh Verma, Rajkumar "Classification Algorithms for Data Mining: A Survey" in International Journal of Innovations in Engineering and Technology (IJIET), Vol. 1 Issue 2 August 2012 ISSN: 2319 – 1058.

[8] D. T. Larose, Discovering knowledge in data: An introduction to data mining, John Wiley & Sons, (2005) 385 http://dx.doi.org/10.1002/0471687545.

[9] Ebru Turanoglu-Bekar, Gozde Ulutagay , Suzan Kantarcı-Sava " Classification of Thyroid Disease by Using Data Mining Models: A Comparison of Decision Tree Algorithms" Oxford Journal of Intelligent Decision and Data Science 2016 No. 2 (2016) 13-28

[10] J. R. Quinlan, C4.5: Programs for machine learning, San Mateo, California: Morgan Kaufmann, (1993).

[11] L. Rokach, O. Maimon, Classification Trees, Data Mining and Knowledge Discovery, Springer New York Dordrecht Heidelberg London, (2010) 149-174.

[12] R. Kohavi, Scaling up the accuracy of naïve bayes classifiers:a decision tree hybrid, In: Proc. of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, (2006) 202-207.

[13] H. Shi, Best-first decision tree learning, (MSc Thesis, University Waikato, 2007, Hamilton NZ).

[14] G. Holmes, B. Pfahringer, R. Kirkby, et al; Multiclass Alternating Decision Trees, Proceedings of the 13th European Conference on Machine Learning (ECML), (2002) 161-172. http://dx.doi.org/10.1007/3-540-36755-1_14.

[15] I. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Mateo, CA, (2000).

[16] W. Fan, H. Wangi, P. S. Yu, et al; Is a random model better? On its accuracy and efficiency, In: Proceedings of Third IEEE International Conference on Data Mining (ICDM), (2003) 51.http :// dx.doi. org /10.1109 /ICDM.2003. 1250902

[17] L. Breiman, Random forests, Mach. Learning, 45 (1) (2001) 5-32. http : // dx.doi.org / 10.1023 / A:1010933404324

[18] N. Landwehr, M. Hall, E. Frank, Logistic model trees, Mach. Learning, 59 (2005) 592-98. http :// dx.doi.org /10.1007 /s10994- 005- 0466- 3

[19] J. Gama, Functional trees, Mach. Leaning, 55 (3) (2004) 219-50. http :// dx.doi.org/10.1023 /B: MACH. 0000027782. 67192. 13.

[20] J. J. Oliver, D. Hand, Averaging Over Decision Stumps, in Machine Learning, In Proceedings of European Conference on Machine Learning (ECML), Catania, Italy, (1994). http://dx.doi.org/10.1023/10.1007/3-540-57868-4_61

All Rights Reserved © 2017 IJERCSE