INTERNATIONAL JOURNAL OF RESEARCH AND DEVELOPMENT IN ENGINEERING SCIENCES
e-ISSN: 2582 - 4201 Volume: 2 | Issue: 6 | Year: 2020

Medical Assistance through Data Analysis using Big Data

Dr. K. SUJATHA
Professor, Department of Computer Science and Engineering,
Dadi Institute of Engineering and Technology (DIET), Anakapalle
Jawaharlal Nehru Technological University, Kakinada

Y.DIVYA
M.Tech Student, Department of Computer Science and Engineering,
Dadi Institute of Engineering and Technology (DIET), Anakapalle
Jawaharlal Nehru Technological University, Kakinada

Abstract *- In this paper, we are proposing a modified algorithm for classification. This algorithm is based on the concept of the decision trees. The proposed algorithm is better than the previous algorithms. It provides more accurate results. We have tested the proposed method on the example of patient data set. Our proposed methodology uses greedy approach to select the best attribute. To do so the information gain is used. The attribute with highest information gain is selected. If information gain is not good then again divide attributes values into groups. These steps are done until we get good classification/misclassification ratio. The proposed algorithms classify the data sets more accurately and efficiently.*

Keywords: *Big Data, Data mining, Naive Bayes, Classification*

1. INTRODUCTION

Decision Trees: The well-known machine learning techniques. A decision tree is composed of three basic elements:

a) A decision node specifying a test attributes.
b) An edge or a branch corresponding to the one of the possible attribute values which means one of the test attribute outcomes.
c) A leaf which is also named an answer node contains the class to which the object belongs. In decision trees, two major phases should be ensured:

Building the tree: Based on a given training set, a decision tree is built. It consists of selecting for each decision node the Appropriate 'test attributes and also to define the class labeling each leaf.

Classification: In order to classify a new instance, we start by the root of the decision tree, then we test the attribute specified by this node. The result of this test allows moving down the tree branch relative to the attribute value of the given instance.

This process will be repeated until a leaf is encountered. The instance is then being classified in the same class as the one characterizing the reached leaf. Decision trees have also been used for intrusion detection [3]. The decision trees select the best features for each decision node during the construction of the tree based on some well-defined criteria. One such criterion is to use the information gain ratio.

2. RELATED WORK

Naïve bays classifier is also a very good and accurate method for the data classification. Naive Bayes classifier [17] is a probabilistic classifier based on the Bayes theorem, considering strong (Naive) independence assumption. Thus, a Naive Bayes classifier believes that all attributes (features) independently contribute to the probability of a certain decision. Considering the characteristics of the underlying probability model, the Naive Bayes classifier can be trained very efficiently in a supervised learning setting. This could yield much better results in many complex real-world situations, especially in the field of computer-aided diagnosis [16] [17].

Here it is assumed that all variables are independent. Hence only the variances of the variables for each class need to determine and not the entire covariance matrix. The RnD tree is a modern method for data classification. Accurate results provided by this method are also attracting so many researchers to this method. The RnD tree [18] algorithm can be applied to both classification and regression problems. Random trees are a collection or assembly of tree predictors that is called forest [18]. The classification works as follows:

the random trees classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of "votes". In the case of regression the classifier response is the average of the responses over all the trees in the forest.

A recursive Bayesian classifier is introduced in [7]. Lots of improvement is already done on decision tree induction method for 100 % accuracy and many of them achieved the goal also but main problem on these improved methods is that they required lots of time and complex extracted rules. The main idea is to split the data recursively into partitions where the conditional independence assumption holds. A decision tree is a mapping from observations about an item to conclusions about its target value [9, 10, 11, 12 and 13].

3. METHODOLOGY

[18]Classification is a form of data analysis that extracts models describing important data classes. These models also called as classifiers are used to predict categorical (discrete, unordered) class labels. This analysis can help us for better understanding of large data. Classification has numerous applications, including fraud detection, target marketing, performance prediction, manufacturing, credit risk and medical diagnosis. Data Classification is a two-step process. They are: Learning Step and Classification Step

A. *Learning Step:*

In this step classification model is constructed. A classifier is built describing a predetermined set of data classes or concepts. In learning step or training phase, where classification algorithm builds the classifier by analyzing or "learning from" a training set made up of database tuples and their associated class labels.

This step is also known as supervised learning as the class label of each training tuple is provided. This learning of the classifier is "supervised" by telling to which class each training tuple belongs. In unsupervised learning or clustering, the class label of each training tuple is not known, and the number or set of classes to be learned may not be known in advance.

B. *Classification Step:*

In this step, the model is used to predict class labels for given data and it is used for classification. First, the predictive accuracy of the classifier is estimated. To measure the classifiers accuracy, if we use the training set it would be optimistic, because the classifier tends to over fit the data i.e., during learning it may incorporate some particular anomalies of the training data that are not present from which the classifier cannot be constructed. The accuracy of a classifier on a given test set is the percentage of test tuples that are correctly classified by the classifier. The associated class label of each test tuple is compared with the learned classifier's class prediction for the tuple. If the accuracy of the model or classifier is considered acceptable, the model can be used to classify future data tuples or objects for which the class label is not known.

C. *Decision Tree Induction:*

A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. The topmost node in a tree is the root node. Given a tuple K, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class predicate for that tuple. Decision trees are easily converted to classification rules. The construction of decision does not require any domain knowledge or parameter setting. It can handle high dimension data. The learning and classification steps are simple and fast. It has good accuracy. Decision tree Induction algorithm can be used in many applications like medicine, manufacturing and production etc.

4. PROPOSED METHOD

```
DTC (in T: table; C: classification
attribute)
return decision
tree
{
  if (T is empty) then return(null);
 /* Base case 0 */
 N: = a new node; if (there are no
 predictive attributes in T)
 /* Base case 1 */
}
```

```
      Then label N with most common
value of C in T (deterministic tree) or
with frequencies of C in T
(probabilistic tree)
else if (all instances in T have the
same value V of C) /* Base case 2 */
then label N, "X.C=V with probability 1"
else
{
for each attribute A in T compute AVG
ENTROPY(A,C,T);
AS := the attribute for which
AVGENTROPY(AS,C,T) is minimal; if (AVG
ENTROPY(AS,C,T) is not substantially
smaller than ENTROPY(C,T))
 /* Base case 3 */

Then label N with most common value of
C in T (deterministic tree) or
with frequencies of C in T
(probabilistic tree).

Else {label N with AS; for each value V
of AS do
{
  }
}
return N;
N1:= DTC (SUBTABLE (T,A,V),C)

/* Recursive call */
if (N1 != null) then make an arc from
N to N1 labeled V;
}
SUBTABLE (in T : table; A : predictive
attribute; V : value)
return table;
{
      T1 := the set of instance X in T
such that X.A = V;
}
      T1 := delete column A from
T1; return T1

/* Note: in the textbook this is called I(p(v 1 ).....p(v)) */

ENTROPY (in C : classification attribute;

T :
table)
return real number;
{
for each value V of C, let p(V) :=
FREQUENCY(C,V,T);
Return Vp(V )
log2k(p(V ))
/* By convention, we consider 0 log (0) to be 0. */

}
/* Note; In the textbook this is called "Remainder (A)" */

AVG ENTROPY (in A: predictive
attribute; C: classification
attribute; T: table)
Return real number; {return V
   FREQUENCY (A, V, T) ·····
   ENTROPY(C, SUBTABLE (T,A, V))
}
FREQUENCY (in B: attribute; V:
value; T: table)
return real number;
{
return #{ X in T| X.B=V} / size (T);
}
```

5. RESULTS

Fig.1: Main Screen

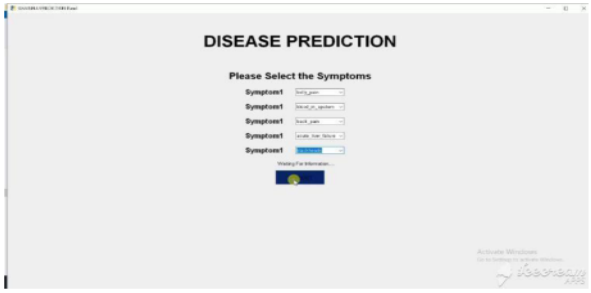

Fig.2: Starting Prediction

Fig.3: Result Box with Prediction

Fig.4: Symptoms Selection

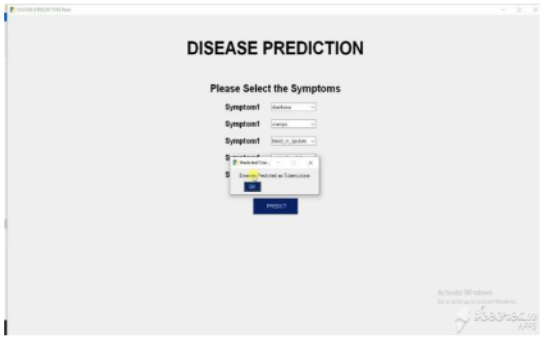

Fig.5: Showing Result

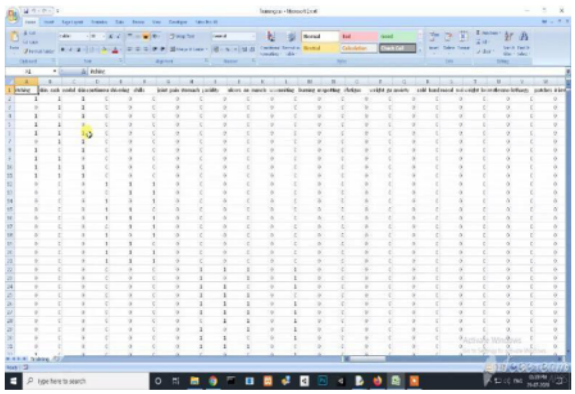

Fig.6: Dataset Overview

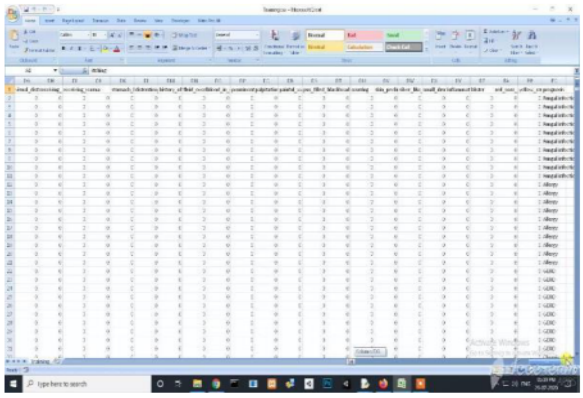

Fig.7: Dataset Classification

6. CONCLUSION

In this paper, we have proposed a modified algorithm for classification. This algorithm is based on the concept of the decision trees. The proposed algorithm is better than the previous algorithms. It provides more accurate results. We have tested the proposed method on the example of patient data set. The framework would definitely lessen the human exertion, lessen the cost and time imperative in terms of HR and mastery, and increment the symptomatic exactness. The forecast of illnesses utilizing data Mining applications and some unsafe undertaking as the information found are unessential and monstrous as well. In this situation, learning of the medicinal information is possible through information mining devices which proven to be useful and it is very fascinating.

The scope of this paper could be commercial use of the application or further research purposes such as to detect the location of users and estimate which disease is more prevalent in a particular region and also to get results month wise that the frequency of a particular disease is diagnose d the most and spread awareness according to it in that region. This paper gave diagram of utilization of information mining secures in regulatory, clinical, inquire about, the more, instructive parts of Clinical Predictions is paper set up that while the current down to earthlization of information mining in wellbeing related use is constrained, there exists an extraordinary entail for information mining systems to enhance ferret parts of Clinical Predictions.

Besides, t h e escapable ascent of clinical information will build the entail for information mining systems that enhances quality and reduces cost of social insurance. This system has large scope as it has the following features which are:

a) Automation of Disease Diagnosis.
b) Paper free work helping the environment.

References

[1]. A. Nayak, M. Pai and R. Pai, "Prediction Models for Indian Stock Market", *Procedia Computer Science*, vol. 89, pp. 441-449,2016.

[2]. W. Fenghua, X. Jihong, H. Zhifang and G.Xu, "Stock Price Prediction Based on SSA and SVM", *Procedia Computer Science*, vol. 31, pp. 625-631, 2014.

[3]. Bini, T. Mathew, "Clustering And Regression Techniques For Stock Prediction", *Procedia Technology*, vol. 24, pp. 1248-1255, 2016.

[4]. F. Elberzagher and K. Holl, "Towards Automated Capturing and Processing of User Feedback for Optimizing Mobile Apps", *Procedia Computer Science*, vol. 110C, pp. 215-221, 2017.

[5]. A. Izzah and R. Widyastuti, "Prediksi Harga Saham Menggunakan Improved Multiple Linear Regression Untuk Pencegahan Data Outlier", *KINETIK*, vol. 2, no. 3, pp. 141-150, 2017.

[6]. A. Khandeparkar, R. Gupta and B.Sindhya, "An Introduction to Hybrid Platform Mobile Application Development", *International Journal of Computer Applications*, vol. 118, no. 15, 2015.

[7]. W.M. Lee, Beginning Android Application Development 1st Edition Kindle Edition, John Willey and Sons Inc., 2012.

[8]. A. Singh, S. Sharma and S. Singh, "Android Application Development using Android Studio and PHP Framework", *International Journal of Computer Applications*, vol. xx, no. xx.

[9]. Robert K. Lai, Chin-Yuan Fan, Wei-Hsiu Huang and Pei-Chann Chang, "Evolving and clustering fuzzy decisiontree for financial Time series data forecasting", An International Journal of Expert Systems with Applications,Vol.36, No.2, pp. 3761-3773, March 2009.

[10]. Shyi-Ming Chen and Yu-Chuan Chang, "Multi- Variable Fuzzy Forecasting Based On Fuzzy Clustering andFuzzy Rule Interpolation Techniques", Information Sciences, Vol.180, No.24, pp. 4772-4783, 2010.

[11]. S AbdulsalamSulaimanOlaniyi, Adewole, Kayode S., Jimoh, R. G," Stock Trend Prediction Using RegressionAnalysis – A Data Mining Approach", ARPN Journal of Systems and Software Volume 1 No. 4, JULY 2011,Brisbane, Australia.

[12]. D. Bhuriya, G. Kaushal, A. Sharma and U.Singh, "Stock Market Predication Using A Linear Regression," in International Conference on Electronics, Communication and Aerospace Technology, 2017.

[13]. P. Guo, M. Waqar, H. Dawood , M. B.Shahnawaz and M. A. Ghazanfar, "Prediction of Stock Market by Principle Component Analysis," in 13th International Conference on Computational Intelligence and Security,2017. Authorized licensed use limited to: UNIVERSITY OF BIRMINGHAM.

[14]. "M. Chen, S. Mao and Y. Liu. Big data: Asurvey".

[15]. "P. B. Jensen, L. J. Jensen and S. Brunak.Mining electronic health records: Towards better research applications and clinical care".

[16]. "Yulei wang1, Jun yang2, Viming.Big Health Application System based on Health Internet of Things and Big Data".

[17]. 17. "S.-M. Chu,W.-T. Shih,Y.-H. Yang, P.-C.Chen and Y.-H. Chu. Use of traditional Chinese medicine in patients with hyperlipidemia: A population-based study in Taiwan".

[18]. 18. "S. Zhai, Chang, R. Zhang and Z. M. Zhang.Deepintent: Learning attentions for online advertising with recurrent neural networks"