An Improved Algorithm for Bayes Classifier to Handle Correlated Attributes

Sharmishtha Panagare
Mahakal Institute of Technology, Ujjain

Kshitij Pathak
Mahakal Institute of Technology, Ujjain

ABSTRACT

Classification can be defined as a target function which maps attribute value of objects to predefined class. One objective is to divide the objects into proper group and other objective is to predict the class of unknown records. Bayesian classifier classifies and predicts the class of objects on the basis of posterior probability based on some prior probability. Earlier work does not handle the effect of correlated attributes on the performance and the accuracy of classifier. In this paper a novel approach using association rules is defined to predict the class of unknown records even if the attributes are correlated. In medical and health care systems it is necessary to generate an outcome which can work well in all conditions and give the beneficial results.

KEYWORDS

Data mining, Bayes classification, Decision Tree, Association rule.

1. INTRODUCTION

Data mining concepts are used to find out the hidden and useful knowledge from a large database. The knowledge can be gathered in multiple fields. The discovered knowledge can be use to make some important decisions in critical and lengthy problems. Classification techniques employed for gathering the knowledge in a proper form that can be use by the administrators to make some decisions. Medical diagnosis is an important field in data mining for solving the major issues in health care system. The purpose of using data mining in medical field is to get better knowledge with reduced cost and timing because the health care environment is always information reach to extract the useful information. In medical and health care systems the quality of service should be high. Any wrong decisions in medical field can create a blunder if it fails to work correctly. With increased accuracy and less efforts the decision support system should solve the critical problems in health related systems. The classifiers can be made on the basis of some knowledgebase and previous results and learning technology. Previously used classifiers worked on decision tree and Support Vector machines and Genetic algorithms, some other data mining techniques used according to their capability to make decisions correct and their level of accuracy. The classifier used here is to predict the chances of heart attack on the basis of some knowledge gathered. Heart diseases are difficult to find out within the time and cause maximum no. of deaths every year.

1.1 Classifiers

Classifiers made on the basis of classification technique used. Classification is a supervised learning method used to predict the diseases in health care systems. There is some classification techniques used in data mining to make the decisions correctly. The techniques include Decision Tree, Rule set classifier, neural networks, Neuro fuzzy, Naïve Bayes, Support vector machine.

Number of health care organizations is still struggle for the utilization of data that is collected through an organization OLTP system and that is not integrated for decision making and pattern analysis. Sometimes OLAP solution provides a multidimensional view of data that is found in relational database. Here the stored data in two dimensional formats make it effective to analyze large amount of data with very fast response time.

Classification Techniques

Depend on domain application various classification techniques available. Here some of the data mining techniques are described with illustration of their applications in health care and medical diagnosis.

Decision Tree

Decision Tree is simple to implement and a popular classification technique. It can handle high dimensional data and requires no domain knowledge. But it can suffer from repetition also. With the use of suitable attributes the performance of Decision Tree can be increase.

Naïve Baye Classifiers

For predicting the chances of heart attacks we are using a Naïve bayes classifier. It is a statistical classifier which assumes no dependency between attributes. It tries to maximize the posterior probability in determining the class. The classifier works with great accuracy. Naive bayes classifier works consistently with the increase and decrease of attributes. Naïve Bayes classifier works well with real world situations.

K-Nearest Neighbor

It is the method of classifying objects based on closest training data. It is a type of instance bases learning. It is the simplest algorithm. But with less accuracy it can degrade the performance sometimes.

1.2 Association Analysis

One of the important terms included here is association analysis. It can be very useful for discovering interesting relationship that is there in large datasets and this uncovered relationship can be shown by association rules or set of frequent items present there. In medical diagnosis, Web mining, Bioinformatics and scientific data analysis association analysis can be applied. To form an accurate classifier association rule mining is used sometimes. In general association analysis is use to discover so many relationships that are hidden in large datasets. That uncovered relationships represented in the form of association rules. One example is that we can have a rule from a large dataset that

{Pencil} – > {Paper}

Here the rule shows relationship between Pencil and Paper that if a person buys a pencil then he can buy a paper also. The sailors use these relations for their products sale. The data can also be represented in binary format sometimes that the rows show the transaction and columns show an item. This means that we can take an item in binary form that if the item is there in transaction then we can take it one and if not then zero.

Suppose we have a set of items S = {s1, s2, s3… s_n} and a set of transaction R = {r1, r2 … r_n}.Every transaction R_i contains a set of items that is taken from S. An item set can have zero or more items in an item set. If an item set have k items in it then we call it a k- item set. For example the set of items like {Bread, Butter, Milk, Sugar} is an item set of 4 items that is 4-itemset.

An Association Rule can be expressed in the form as A -- > Y where A and B are disjoint item sets. That is A ∩ B = Ø. An association rule's strength is measured with Support and Confidence. Support specifies how frequently a rule is applicable in a dataset. Confidence specifies how frequently items in B appears in transaction that also contain A.

Support and Confidence can be defined as

$$\text{Support s (A - > B)} = \frac{\sigma\text{ (A U B)}}{N}$$

$$\text{Confidence, c (A - >B)} = \frac{\sigma\text{ (A U B)}}{\sigma\text{ (A)}}$$

It can be also say that if a rule has support count very less than it just happens by chance and it may not be very much profitable for the sailors. That's why it can also be use to eliminate the rules that are of no interest. Confidence shows the reliability for a rule and also gives conditional probability of B given A. Sometimes Brute force approach is used for mining association rule's confidence and support but may be costly because so many rules can be formed from a large dataset. The number of rules that we can have from a data set of d items is calculated by the formula below

$$S = 3^d - 2^{d+1} + 1$$

The above formula is used to calculate the rules from a data set that can be used to predict the future behavior of that dataset.

2. LITERATURE SURVEY

In [1] the major element in any medical research is to extract some meaningful information from large data sets. In this study a novel algorithm that is base on Bayes classifier is used gives capability to extract useful features and effective features. The Paper gives a computational advantage over the previously-employs KNN classifier where the training data are summarized, rather than stored.

The classification of data from the University of California, Irvine (UCI) machine learning data set repository was performed to evaluate the effectiveness of hybrid classifier on real world data, and it also helps to compare it with other classifiers.

In [2] the study it is examined the potential use of classification based data mining techniques for example the rule based, Naive Bayes and Artificial Neural Network, Decision Tree. Here for data preprocessing and effective decision making ODANB that is One Dependency Augmented Naïve Bayes classifier and NCC2 Naïve Credal Classifier 2 are used. It is an extension of Naive Bayes to improve the probabilities that aims at delivering robust classification when we have some small or incomplete data sets.

In [3] the paper describes an attempt to predict and efficiently diagnosis the patient for heart diseases with reduced number of factors or attributes using classifications. In a previous study Asha et al developed an intelligent heart disease prediction system is made using three classifiers Decision Tree, Naïve Bayes and Neural Networks. Naive Bayes performed with very good prediction probability of 96.6%.with use of 13 attributes and this work is different in the use of less number of attributes with the same performance. In place of 13 attributes only 6 attributes are used using genetic search and three classifiers Naïve Bayes, Decision Tree, and Classification by clustering are used to predict the chances of heart diseases with the same performance. Naïve Bayes did not affect the same or less attributes but Classification by clustering performed poor in comparison to the others. And intensity of the disease is unpredictable here in the study.

In [4] the work focused on predicting accurately the patients with ill class rather than the healthy class. A novel learning tool that is biased minimax probability machine (BMPM). It directly controls the worst case accuracies to add in a bias towards ill class. BMPM provides a more accurate way to handle medical diagnosis. In the study the performance of the model is evaluated and compared with three classifiers that is k-nearest neighbor, Naïve Bayes and C4.5 with to data sets that is heart disease data set and a breast cancer dataset.

In [5] the study of different algorithms are used and analyzed for the detection of heart diseases by KNN, Decision Tree and Naïve Bayes. With 15 Attributes the classifiers evaluated the accuracies in which Decision Tree outperforms and Bayesian classifier also have accuracy equivalent to it sometimes.

In [6] it gives a Decision Support in Heart Disease Prediction System (DSHDPS) by the use of Naïve bayes a data mining technique. some basic attributes used such as age, sex, blood sugar, blood pressure to predict the heart diseases. The Naïve bayes is use to make predictive capable models.

In [7] here the performance of clustering and classification algorithms analyzed with the heart dataset. The performance evaluation is done through different classifiers like Naïve bayes, Naïve bayes updateable. Function (SMO), Decision Tree, and clustering algorithms K-means. Performance of classifier is calculated by cross validation and cluster is evaluated by mode of classes. Mainly on the basis of 14 attributes the performance of the classifier calculated. Here result of the evaluation found that NB tree shows the highest prediction capability in comparison to cluster algorithms.

In[8] using some data mining techniques like Naïve bayes, Decision tree and Neural networks, a prototype is developed that is Intelligent Heart Disease Prediction System (IHDPS). Each technique has its unique strength for predicting the chances of disease. It is implemented on .Net Platform and a user friendly and web based system.

Here in [9] a novel type of Naïve Bayes classifier that is also known as noisy-threshold classifiers is applied here for predicting the carcinoid heart disease (CHD). 11 Attributes are used here for this. Area under the receiver operating characteristics (ROC) curve of noisy-threshold is compared with these classifiers.

In [10] the system uses the incremental learning algorithm. Here an incremental learning algorithm called modified dynamic weight majority voting (DWMV) Learn++.

3. RESEARCH GAP

Sometimes correlated attributes changes the performance of bayes classifier because for such attributes the conditional probability no longer holds.

Suppose we have following probabilities

$$P(X = 0|Z = 0) = 0.4 \qquad P(X = 1|Z = 0) = 0.6,$$

$$P(X = 0|Z = 1) = 0.6, \qquad P(X = 1|Z = 1) = 0.4,$$

Where X is a binary attribute and Z is a binary class variable. Now suppose there is another binary attribute C that is perfectly correlated with X when Z = 0, but it is independent of X when Z = 1.

For simplicity we can also assume that the class conditional probabilities for C are the same as for X. We have given a record with attributes X = 0, C = 0, We can compute its posterior probabilities as follows:

$$P (Z = 0|X = 0, C = 0) = \frac{P(X = 0|Z = 0)P(C = 0|Z = 0)P(Z = 0)}{P(X = 0, C = 0)}$$

$$= \frac{0.16 \ {*}P(Z = 0)}{P(X = 0,C = 0)}$$

$$P (Z = 1|X = 0, C = 0)=\frac{P(X = 0|Z = 1)P(C=0|Z = 1)P(Z=1)}{P(X = 0, C = 0)}$$

$$= \frac{0.36{*}P(Z = 1)}{P(X = 0, C = 0)}$$

If P(Z = 0) = P(Z = 1), then this naïve Bayes classifier would assign the record to class 1. However, the true is,

$$P(X = 0, C = 0)|Z= 0) = P(X = 0|Z = 0) = 0.4,$$

Because X and C are perfectly correlated when Z = 0.As a result the posterior probability for Z = 0 is

$$P(Z = 0|X = 0, C = 0) = \frac{P(X = 0,C = 0|Z = 0)P(Z = 0)}{P(X = 0, C = 0)}$$

$$= \frac{0.4 \ {*}P(Z = 0)}{P(X = 0,C = 0)}$$

Which is larger than for Z = 1. The record should have been classified as class 0.

4. PROBLEM STATEMENT

To design and develop an approach using association rules to classify the records using Bayesian classifier having correlated attributes.

5. PROPOSED SOLUTION

The proposed solution for handling the correlated attributes in a dataset presented an algorithm that finds whether the attributes are correlated if the given attributes have some correlation between them then one of them is selected and other is discarded. The step by step procedure finds the attributes and relation between them and the prior and posterior and prior probabilities for the given dataset. The algorithm is described below:

```
Input: Attribute A1, A2 ...$A_n$ with values v1, v2.. $v_n$

Output: Class

1. Split all the attribute value in parts.
2. length = count the splits
           Or
   length = length (parts)
3. Generate association rules between the attributes by
   taking min confidence threshold as 100
4. For all the rules generated, left hand side and right
   hand side of the rules are identified as correlated
   attributes
5. Let there be two class 'No' & 'Yes'
   Classno = count of the tuple having 'No' value in
   class attribute
   Classyes = count of the tuple having 'Yes' value in
   class attribute
6. priorprobabilityno = 1
7. priorprobabilityyes = 1
8. If correlation exist between parts
   {Let correlation exist between A1 & A3 }
   Then for calculating priorprobabilityno & prior
   probabilityyes either part A1or part A3 will be
   deleted.
9. For i = 1 to length
   {
     x = parse( text = parts[i])
      m = eval(k)
   priorprobabilityno = priorprobabilityno *
                 $\frac{\text{(length (which(m \& class no)))}}{\text{length (which (class no))}}$
   priorprobabilityyes = priorprobabilityyes *
                 $\frac{\text{(length (which(m \& class yes)))}}{\text{length (which (class yes))}}$
     }
10. posteriorprobabilityno = priorprobabilityno
                  *(length (which (class no))/10)

    posteriorprobabilityyes = priorprobabilityyes
                  *(length (which (class yes))/10)
11. If(posteriorprobabilityno> posteriorprobabilityyes )
      Then output of the class is no
      Else
     Output of the class is yes
```

6. IMPLEMENTATIONS

The language used here is R [11]. R is a language and environment which is used for statistical computing and graphics. It is a GNU project and it is similar to the S which was developed at Bell Laboratories by John Chambers and colleagues. R can be a different implementation of S. The code is shown for finding the correlated attributes and the class Yes on No from the given dataset. The input and output calculated and also the time for computing is shown below.

INPUT:-

Table :- Bayc

tid	howner	mstatus	faulty
1	Y	S	NO
2	N	M	NO
3	N	S	NO
4	N	M	NO
5	N	D	YES
6	N	M	NO
7	Y	D	NO
8	N	S	YES
9	N	M	NO
10	N	S	YES

TUPLE TO CLASSIFY: - bayc$howner=='N' & bayc$mstatus == 'M'

OUTPUT:-

```
priorprobabilityno = 0.3265306
priorprobabilityyes =  0
posteriorprobabilityno =  0.2285714
posteriorprobabilityyes =  0
Predicated class is no"
```

TIME TAKEN:-

User	System	Elapsed
0.020	0.000	0.023

7. CONCLUSION

In this paper the study and implementation of a classifier shown for predicting the class of objects accurately. The classifier studied here is Bayesian classifier which sometimes does not provide optimal results due to the presence of correlated attributes. Proposed approach uses the concept of association rules to overcome this problem. Proposed algorithm classifies the records in feasible time.

Future work is to compare the classifier with the other classifiers that whether they can work with correlated attributes and with minimum time and maximum accuracy. The classifier can also be implemented to predict for the improvement of medical and health care support systems.

8. REFERENCES

[1] Michael L. Raymer, *Member*, Travis E. Doom, Leslie A. Kuhn, William F. Punch , Knowledge Discovery in Medical and Biological Datasets Using a Hybrid Bayes Classifier /Evolutionary Algorithm, IEEE Transactions on Systems, Man, and Cybernetics—Part b: Cybernetics, vol. 33, no. 5, october 2003

[2] K.Srinivas, B.Kavihta Rani, Dr. A.Govrdhan, K.Srinivas et al. / (IJCSE) Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks.

[3] M. Anbarasi, E. Anupriya, N.CH.S.N.Iyengar, Enhenance Prediction of with feature selection using Genetic Algorithm, IJEST,VOL 2, 2010

[4] Kaizhu Huang, Haiqin Yang, Irwin King, Michael R. Lyu, Maximizing sensitivity in Medical Diagnosis Usin Bayes Minimax Probability Machine, IEEE Transactions on Biomedical engineering, vol. 53, no. 5, may 2006

[5] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni Predictive Data Mining for Medical:An Overview of Heart Disease Prediction, International Journal of Computer Applications, March 2011.

[6] Mrs.G.Subbalakshmi ,Mr. K. Ramesh M.Tech, Asst. Professor, Mr. M. Chinna Rao M.Tech,(Ph.D.) Asst. Professor, Decision Support in Heart Disease Prediction System using Naive Bayes, G.Subbalakshmi et al. / Indian Journal of Computer Science and Engineering (IJCSE)

[7] P. Santhi, V. Murali Bhaskaran,Improving the Performance of Data Mining Algorithm in Health Care Data, IJCST Vol. 2, Issue 3, September 2011

[8] Sellappan Palaniappan , Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques. IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008

[9] Marcel A.J. van Gerven a, Rasa Jurgelenaite a, Babs G. Taal b,Tom Heskes a, Peter J.F. Lucas, Predicting carcinoid heart disease with the Noisy-threshold classifier, Artificial Intelligence in Medicine (2007).

[10] Shima Aghtar, Mcmaster University A New Incremental Classification Approach Monitoring the Risk of Heart Disease. Open Access Dissertations and Theses.(2012)

[11] www.r.projrct.org