COMPARATIVE ANALYSIS OF NAIVE BAYS CLASSIFIER AND DECISION TREE C4.5 ON CREDIT PAYMENT DATA SET

S.B.Siledar[1], S.R.Chaudhary[2]

[1]*Dept. of CSE, MIT, Aurangabad(M.S)*
[2]*Dept. of CSE, MIT, Aurangabad(M.S)*

Abstract

A whole world is revolving around data. Current techno-world is generating huge amount of data. For research and analysis purpose data mining is essential. Data is collected from various resources, it needs preprocessing. Z-Normalization technique is one of the transformation technique which is used in preprocessing. Accuracy of Naïve Bays and Decision tree algorithm is increased after applying Z-score normalization on UCI - default of credit card clients data set.

Keywords: *Data mining, Z-score normalization, Naïve Bays classification, Decision Tree C4.5, Z-score normalization, Naïve Bays classification, Decision Tree C4.5*

General Terms: *Data Mining, Classification Techniques*

----------***----------

1. INTRODUCTION

Naïve Bays classification and decision tree 4.5 classification are based on supervised learning. In supervised learning, Known data objects are used as a training data set, then records without classes are classify based on training data model. Accuracy of classification may be increased if normalized data is used

2. CLASSIFICATION TECHNIQUES, NORMALIZATION

2.1 Decision Tree Classifier

To build decision tree doesn't require domain knowledge expert. It handles multidimensional data. Most of the decision tree algorithms like ID3,C4.5 and CART are based on top down approach. Training set consists of tuples with associated class labels. Training set is recursively divided into smaller partitions as the tree is being built.

In decision tree attribute selection method is used to decide partition. Main purpose of this method is to reduce randomness while partitioning data such that one can easily classify data. There are various attribute selection approaches such as Information Gain, Gain Ratio and Gini Index. ID3 is based on Information Gain. C4.5 uses Gain Ratio as a attribute selection method.

Gain ratio calculates the information with respect to classification of based in the partition. C4.5 uses pessimistic pruning[1].

2.1 Naïve Bays Classifier

It is a statistical classifier. It is based on probability that a given data belongs to a class. In Bays classification :
X – Given data to be classify
H – Hypothesis such that the data tuple ***X*** belongs to a specified class *C*.
$P(H|X)$ - is the **posterior probability**
$P(H)$ is the **prior probability**
$P(X|H)$ is the posterior probability of ***X*** conditioned on *H*
$P(\boldsymbol{X})$is the prior probability of ***X***.

Bayes' theorem is

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Naïve bays classifier has used Bayes theorem to calculate conditional probability[1].

2.2 Z-score Normalization

In ***z*-score normalization** (or *zero-mean normalization*), the values for an attribute, *S*, are normalized based on $\bar{s}$ −the mean (i.e., average) and σ_s − standard deviation of *S*[1].

A value, *bi* , of *S* is normalized to $b'_i$by computing

$$b'_i = \frac{(b_i - \bar{s})}{\sigma_s}$$

3. LITERATURE SURVEY

Jiangtao Ren[2] discusses Bayesian classification on uncertain data. They use the class conditional probability estimation with probability distribution function. Kabir Md

Faisal [3] combines k-means clustering method with Naïve Bayesian classification algorithm to improve the performance. Bhavesh Patankar Dr. Vijay Chavda[4] compares Naïve Bayesian classification with Decision Tree and KNN and concludes decision tree is better that the other two algorithms. S. Karthika and N. Sairam[5] proposes a Naïve Bayes classification for the classification of person in education field.

Badr Hssina, Abdelkarim[6] focuses on compareision of ID3, C4.5 ,CART and C5.0 algorithms. C4.5 is better than ID3 and CART. C5.0 is faster than C4.5.

Shiju Sathyadevan and Remya R. Nair[7] highlights theID3, C4.5 and Random Forest algorithms. They experiment it on Credit Approval dataset. Random Forest performs better on this type of data.

4. PERFORMANCE ANALYSIS

Experiment is done on UCI - default of credit card clients data set.. It contains 24 attributes and 30000 records. First accuracy of naïve bays classifier and C4.5 classifier is calculated without applying normalization. After applying Z-score normalization on Credit amount attribute of data set accuracy of both classifier is calculated.

Table 1: Classification Before Normalization

Classification Techniques	**Correctly Classified Instances**	**Incorrectly Classified Instances**
Naïve Bays Classifier	19574 (65.24%)	10426 (34.74%)
Decision Tree C4.5	24588 (81.96%)	5412 (18.04 %)

Table 2: Summary Before Normalization

Summary	**Naïve Bays Classifier**	**Decision Tree C4.5**
Kappa statistic	0.2015	0.353
Mean absolute error	0.3624	0.283
Root mean squared error	0.523	0.3763

Table 3: Classification after Z-score Normalization

Classification Techniques	**Correctly Classified Instances**	**Incorrectly Classified Instances**
Naïve Bays Classifier	20230 (67.43%)	9770 (32.56%)
Decision Tree C4.5	24588 (81.96%)	5412 (18.04 %)

Table 4: Summary after Z-score Normalization

Summary	**Naïve Bays Classifier**	**Decision Tree C4.5**
Kappa statistic	0.2165	0.353
Mean absolute error	0.3426	0.283
Root mean squared error	0.507	0.3763

5. CONCLUSION

Experimental result before Z-score normalization is shown in Table I and II. Naïve Bays classifier accuracy is improved after Z-score normalization as shown in Table III whereas Decision Tree C4.5 performance remains same before and after normalization of credit payment data set. Naïve Bays classifier is faster than Decision Tree C4.5

REFERENCES

[1]. Jiawei Han, Micheline Kamber, Jian Pei, “Data Mining Concepts and Techniques”, Third Edition, Elsevier Inc.

[2]. Jiangtao Ren, Sau Dan Lee, Xianlu Chen_, Ben Kao, Reynold Cheng and David Cheung , “Naïve Bayes Classification of Uncertain Data”, 2009

[3]. Faisal KM, Mofizur RC, Alamgir H, Kesav D. ,“Enhanced Classification Accuracy on Naive Bayes Data Mining Mod¬els”, International Journal of Computer Applications, 2011

[4]. Bhavesh Patankar Dr. Vijay Chavda, “A Comparative Study of Decision Tree, Naive Bayesian and k-nn Classifiers in Data Mining”, International Journal of Advanced Research in Computer Science and Software Engineering, 2014

[5]. S. Karthika and N. Sairam, “A Naïve Bayesian Classifier for Educational Qualification”, Indian Journal of Science and Technology, July 2015

[6]. Badr Hssina, Abdelkarim Merbouha,Hanane Ezzikouri,Mohammed Erritali, “A comparative study of decision tree ID3 and C4.5”, International Journal of Advanced Computer Science and Applications, 2013

[7]. Shiju Sathyadevan and Remya R. Nair, “Comparative Analysis of Decision Tree Algorithms: ID3, C4.5 and Random Forest”, Computational Intelligence in Data Mining, Springer 2015