Contents lists available at www.innovativejournal.in

INNOVATIVE JOURNAL

Asian Journal of Computer Science And Information Technology

Journal Homepage: http://innovativejournal.in/ajcsit/index.php/ajcsit

AJCSIT

ISSN 2249 - 5126

AN IMPROVED CLASSIFICATION USING BAYESIAN TECHNIQUE FOR MEDICAL DOMAIN

B. Murugeshwari *

Professor, Department of Information Technology,Velammal Institute of Technology

Contents lists available at www.innovativejournal.in

INNOVATIVE JOURNAL

Asian Journal of Computer Science And Information Technology

Journal Homepage: http://innovativejournal.in/ajcsit/index.php/ajcsit

AJCSIT

ISSN 2249 - 5126

AN IMPROVED CLASSIFICATION USING BAYESIAN TECHNIQUE FOR MEDICAL DOMAIN

B. Murugeshwari *

Professor, Department of Information Technology,Velammal Institute of Technology

CrossMark

ARTICLE INFO

Corresponding Author:
Dr.B.Murugeshwari
Professor, Department of Information Technology,Velammal Institute of Technology
niyansree@gmail.com

Keywords: Naive Bayesian classification, medical classification, Data mining

crossref MEMBER
CROSSREF.ORG
THE CITATION LINKING BACKBONE

DOI:http://dx.doi.org/10.15520/ajcsit.v6i7.49

ABSTRACT

Data This paper presents a novel classification scheme to improve classification performance when few training data are available. In the proposed scheme, data are described using the discretized statistical features. Solving the classification in a classifier combination framework and theoretically analyze the performance benefit. Furthermore classification method is proposed to aggregate the naive Bayes (NB) predictions of the correlated data. It investigates an analysis on prediction error sensitivity of the aggregation strategies. Finally, a large number of experiments are carried out on two large-scale real-world data sets to evaluate the proposed scheme. The experimental results show that the proposed scheme can achieve much better classification performance than existing classification methods.

I.INTRODUCTION

To classify datasets in medical domain and to sort patients accordingly using Naive bayes algorithm implemented using weka tool. Flow statistical feature based traffic classification can be enhanced by feature discretization. Particularly, feature discretization is able to dramatically affect the performance of Naive Bayes(NB).NB is oneof the earliest classification methods applied in Internet traffic classification which is a simple and effective probabilistic classifier employing the Bayes theorem with naive feature independence assumptions . Since independent features are assumed, an advantage of the NB classifier is that it only requires a small amount of training data to estimate the parameters of a classification model. However, the performance degradation of NB traffic classifier is reported in the existing works. The main reason for the underperformance of a number of traditional classifiers including NB is the lack of the feature discretization process. For example, feature discretization can effectively improve the accuracies of the support vector machine (SVM) and K-Nearest neighborhood(KNN) algorithms at the price of lower classification speed. More interestingly, NB with feature discretization demonstrates not only significantly higher accuracy but also much faster classification speed.

The coding for implementing Naive Bayes algorithm was done by java. The target of the sample code is to present an example which trains a simple Naive Bayes Classifier in order to detect whether a person is affected by diabetes. It returns either true positive or false negative. Declaration of attributes is done. For example in diabetes there are totally 8 attributes and 1 class which can be denoted by 0 or 1. The attributes used are Number of times pregnant, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), Body mass index (weight in kg/(height in m)^2), Diabetes pedigree function, Age (years) and Class variable (0 or 1). class value 1 is interpreted as "tested positive for% diabetes".

First a training dataset is taken and their mean and standard deviation is found out. After that it needs to be analyzed whether the classified part is "tested positive" or "tested negative". After finding out the result of the training set, using that rest of the datasets can be classified

II.RELATED WORK

JunZhang et al in Internet Traffic classification by aggregating correlated naïve bayes predictions presented a traffic classification scheme to improve classification performance when few training data are available. In this scheme, traffic flows are described using the discretized statistical features and flow correlation information is modeled by bag-of-flow (BoF). It was solved by the BoF-based traffic classification in a classifier combination framework and theoretically analyze the performance benefit. Furthermore, a new BoF-based traffic classification method was presented to aggregate the naive Bayes (NB) predictions of the correlated flows. It also presents an analysis on prediction error sensitivity of the aggregation strategies. Finally,a large number of experiments are carried out on two large-scale real-world traffic datasets to evaluate the scheme presented. The experimental results show that this scheme can achieve much better classification performance than existing traffic classification methods.

Anupama Kumar et al in Efficiency of design trees in predicting students academic performance said Classification methods like decision trees, rule mining, Bayesian network etc can be applied on the educational data for predicting the students behavior, performance in examination etc. This prediction could help the tutors to identify the weak students and help them to score better marks. The C4.5 decision tree algorithm was applied on student's internal assessment data to predict their performance in the final exam. The outcome of the decision tree predicted the number of students who are likely to fail or pass. The result was given to the tutor and steps were taken to improve the performance of the students who were predicted to fail. After the declaration of the results in the final examination the marks obtained by the students are fed into the system and the results were analyzed. The comparative analysis of the results states that the prediction has helped the weaker students to improve and brought out betterment in the result. To analyze the accuracy of the algorithm, it is compared with ID3 algorithm and found to be more efficient in terms of the accurately predicting the outcome of the student and time taken to derive the tree.

Hao Zhang in Deiscriminative Nearest Neighbour Classification for Visual category recognition considered visual category recognition in the framework of measuring similarities, or equivalently perceptual distances, to prototype examples of categories. This approach permits recognition based on color, texture, and particularly shape, in a homogeneous framework. While nearest neighbour classifiers are natural in this setting, they suffer from the problem of high variance (in bias-variance decomposition) in the case of limited sampling. Alternatively, one could use support vector machines but they involve time consuming optimization and computation of pairwise distances. Here a hybrid of these two methods were used which deals naturally with the multiclass setting, has reasonable computational complexity both in training and at run time, and yields excellent results in practice. The basic idea is to find close neighbors to a query example and train a local support vector machine that preserves the distance function on the collection of neighbors. This method can be applied to large, multiclass data sets for which it outperforms nearest neighbor and support vector machines, and remains efficient when the problem becomes intractable for support vector machines. A wide variety of distance functions can be used and our experiments show state-of-the-art performance on a number of benchmark data sets for shape and texture classification (MNIST, USPS, CUReT) and object recognition (Caltech- 101). On Caltech-101 we achieved a correct classification rate of 59.05%(±0.56%) at 15 training images per class, and 66.23%(±0.48%) at 30 training images.

III. Materials and Methods

A. *Naive bayesie classification*

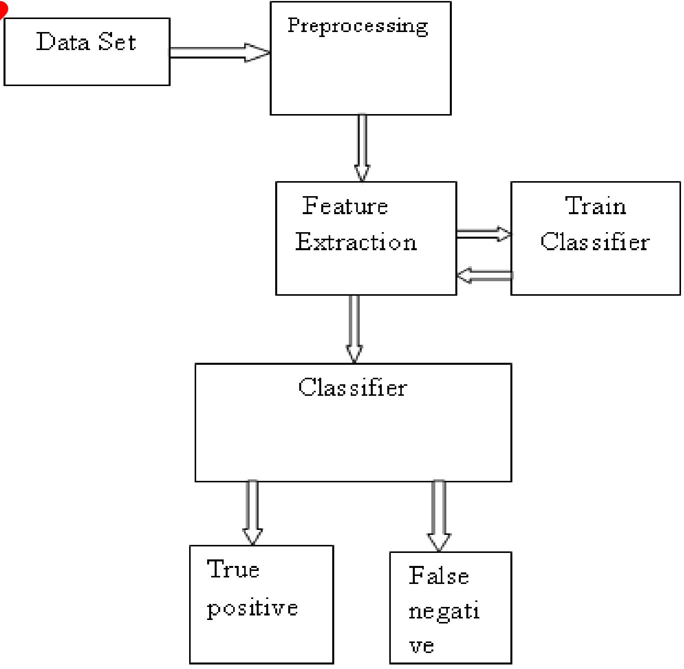
The classification can be done by finding out the rate of sensitivity and specificity. Sensitivity (also called the *true positive rate*, or the recall rate in some fields) measures the proportion of actual positives which are correctly identified as such the condition). Specificity (sometimes called the *true negative rate*) measures the proportion of negatives which are correctly identified as such. These two measures are closely related to the concepts of type I and type II errors. A type I error (or error of the first kind) is the incorrect rejection of a true null hypothesis. It represents false positive. Usually a type I error leads one to conclude that a supposed effect or relationship exists when in fact it doesn't. Examples of type I errors include a test that shows a patient to have a disease when in fact the patient does not have the disease. A type II error is the failure to reject a false null hypothesis. With respect to the non-null hypothesis, it represents a false negative. Examples of type II errors is blood test failing to detect the disease it was designed to detect, in a patient who really has the disease.

This paper presents a novel classification scheme to improve classification performance when few training data are available. In the proposed scheme, data are described using the discretized statistical features. We solve the classification in a classifier combination framework and theoretically analyze the performance benefit. Furthermore classification method is proposed to aggregate the naive Bayes (NB) predictions of the correlated data. We also present an analysis on prediction error sensitivity of the aggregation strategies. Finally, a large number of experiments are carried out on two large-scale real-world datasets to evaluate the proposed scheme. The experimental results show that the proposed scheme can achieve much better classification performance than existing classification methods.

B. *Architecture of Proposed system*

The fig describes the classification of test datasets based on training sets and determination of true positive and false negative values for each class enlisted in the dataset.

Figure I Architecture of Proposed System



1) Data Analysis and preprocessing: A data set (or dataset) is a collection of data. Here the data is collected from various sources. The data will be in the .txt format(text format).Data creation can be done by combining all the data receive the dataset and preprocess it by removing unwanted comment statements manually. Standardize all features or normalize all training samples. The first row contains the attribute names followed by each data row with attribute values listed in the same order. By doing preprocessing data can be taken separately. The attributes used for classifying diabetes patients are (i)Number of times pregnant, (ii) Plasma glucose concentration a 2 hours in an oral glucose tolerance test, (iii) Diastolic blood pressure (mm Hg), (iv) Triceps skin fold thickness (mm) , (v) 2-Hour serum insulin (mu U/ml), (vi) Body mass index (weight in kg/(height in m)^2) , (vii) Diabetes pedigree function , (viii) Age (years) , (ix)Class variable (0 or 1). Class value 1 is interpreted as "tested positive".

2) Using WEKA tools as Interface: Weka(Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java The Weka (pronounced Weh-Kuh) workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. It consists of collection of data mining techniques. It is used for classifying the dataset using various classifying techniques. Weka tool is implemented using GUI interface. Clustering of various classification algorithms are inbuilt.

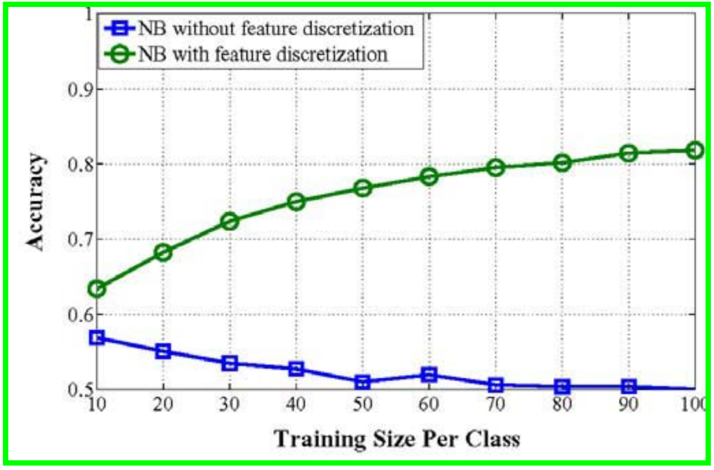
3)Classification Using Naive Bayes Algorithm: A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem with strong (naive) independence assumptions. The Bayesian recipe is simple, optimal, and in principle, straight forward to apply.

Design an optimal classifier on the following basis : P(ωi) (priors)(x | ωi) (class-conditional densities). Have some knowledge on training data {(xi,ωi)}. Use the samples to estimate the unknown probability distributions. Given training data set D. Need to estimate probabilistic parameters, no need for complicated training process as in neural networks .Estimate Maximum Likelihood for every attribute under each class and build confusion matrix. Determine true positive and false positive values using NB predictor. True positive value gives the number of data correctly classified in the given class. False positive value gives the number of data wrongly classified in the given class .

4)Multi-Boosting for Estimating Accuracy: Multi-boosting is implemented to compare the efficiency of various classification techniques .In this module comparison of J48, Random forest techniques are made with Naive Bayes classifier .This is to enhance the classification of datasets by analysing purity value of each classifier .

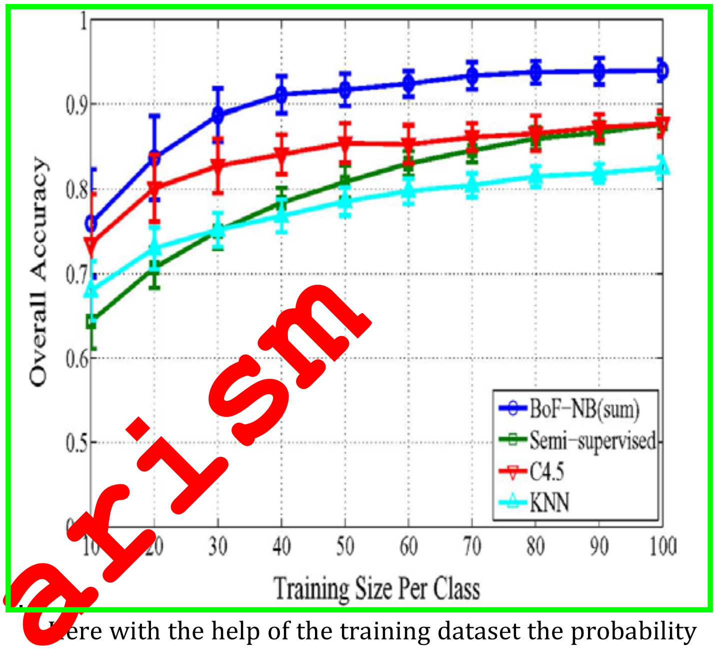
IV EXPERIMENTAL RESULTS

Fig II shows accuracy of classification with NB discretized and not discretized data.



Here x-axis is taken as Training size per class i.e the amount of training data that is taken for classifying and y-axis is the accuracy in which the datasets are classified. Discretization is the process of portioning continous attributes so that accuracy can be improved. As the training datasets are increased accuracy increases. Hence Naive Bayes with feature discretization produces 81% accuracy. Without feature discretization, all the attributes may collide, as a result of that classifying becomes tougher and thereby reduces the accuracy as the training datasets increases.The lowest accuracy that can reach is 0.5 which means 50% accuracy only

Fig III shows accuracy of various classification algorithms.



Here with the help of the training dataset the probability of test datasets can be found out. To improve the accuracy, discretization needs to be used. Here classification using Naive Bayes produces an accuracy of about 95%. While comparing with other algorithms, Naive Bayes classifies efficiently. C4.5 can produce an accuracy of about 88%. KNN can produce about 82% accuracy. Semi-Supervise can produce about 87%.

V CONCLUSION

The main purpose of this paper is to classify the dataset in medical domain. Datasets include details of patient and their medical history. Naive Bayes classifier is used to discretize the dataset using few training sets and extract their features and thereby classifies the test datasets. This helps in easy classification of various patients and their diseases. If one wants to search for any details immediately then this can be handy. Even new patients can be classified using the feature extracted from the previous datasets.

REFERENCES

1. Y.Wang, Y. Xiang, and S.-Z. Yu, "An automatic application signature construction system for unknown traffic," *Concurrency Computat.:Pract. Exper.*, vol. 22, pp. 1927–1944, 2010.
2. Y.S.Lim,H.C.Kim,J.Jeong,C.K.Kim,T.T.Kwon,andY.Choi, "Internet traffic classification demystified: On the sources of the dis- criminative power," in Proc. 6th Int. Conf., Ser. Co-NEXT'10, New York, 2010, pp. 9:1–9:12, ACM.
3. Y. Xiang, W. Zhou, and M. Guo"Flexibldeterministic packet marking: An iptraceback system to find the real source of attacks," IEEE Trans. Parallel Distrib. Syst., vol. 20, no. 4, pp. 567–580, Apr. 2009.

4. T. T. Nguyen and G. Armitage, “A survey of techniques for internet trafficclassificationusingmachinelearning,”Commun.SurveysTuts., vol. 10, no. 4, pp. 56–76, 4th Quarter 2008.
5. H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee,“Internettrafficclassificationdemystified:Myths,caveats,and the best practices,” in Proc. ACM CoNEXT Conf., New York, 2008, pp. 1–12.
6. .N.Williams,S.Zander,andG.Armitage,“Apreliminaryperformancecomparisonoffivemachinelearning algorithms for practica lip traffic flowclassification,”inProc.SIGCOMMComput.Commun. Rev.,Oct. 2006, vol. 36, pp. 5–16.
7. .J. Ma, K. Levchenko, C. Kreibich, S. Savage, and G. M. Voelker, “Unexpectedmeans of protocol inference,” in *Proc. 6th ACM SIGCOMMConf. Internet Measurement*, New York, 2006, pp. 313–326.
8. .L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian,“Traffic classification on the fly,” in *Proc. SIGCOMM Comput.Commun. Rev.*, Apr. 2006, vol. 36, pp. 23–26.
9. J. Erman, M. Arlitt, and A. Mahanti, “Traffic classification using clustering algorithms,” in *Proc. SIGCOMM Workshop on Mining Network Data*, New York, 2006, pp. 281–286.
10. W. Moore and D. Zuev, “Internet traffic classification using bayesian analysis techniques,” in SIGMETRICS Perform. Eval. Rev., Jun. 2005, vol. 33, pp. 50–60.
11. S. Zander, T. Nguyen, and G. Armitage, “Automated traffic classification and application identification using machine learning,” in *Proc.Ann. IEEE Conf. Lcal Computer Networks*, Los Alamitos, CA, 2005,pp. 250–257.
12. T. Karagiannis, K. Papagiannaki, and M. Faloutsos, “BLINC: Multi- level traffic classification in the dark,” in Proc. SIGCOMM Comput. Commun. Rev., Aug. 2005, vol. 35, pp. 229–240
13. .R.O.Duda,P.E.Hart,andD.G.Stork,PatternClassification. New York: Wiley, 2001.
14. Snort 2011 [Online]. http://www.snort.org/.
15. Bro 2011 [Online]. Available: http://bro-ids.org/index.html

Internet Traffic Classification by Aggregating Correlated Naive Bayes Predictions

Jun Zhang, *Member, IEEE*, Chao Chen, Yang Xiang, *Senior Member, IEEE*, Wanlei Zhou, *Senior Member, IEEE*, and Yong Xiang, *Senior Member, IEEE*

***Abstract*—This paper presents a novel traffic classification scheme to improve classification performance when few training data are available. In the proposed scheme, traffic flows are described using the discretized statistical features and flow correlation information is modeled by bag-of-flow (BoF). We solve the BoF-based traffic classification in a classifier combination framework and theoretically analyze the performance benefit. Furthermore, a new BoF-based traffic classification method is proposed to aggregate the naive Bayes (NB) predictions of the correlated flows. We also present an analysis on prediction error sensitivity of the aggregation strategies. Finally, a large number of experiments are carried out on two large-scale real-world traffic datasets to evaluate the proposed scheme. The experimental results show that the proposed scheme can achieve much better classification performance than existing state-of-the-art traffic classification methods.**

***Index Terms*—Traffic classification, network security, naive Bayes.**

I. Introduction

APPLICATION oriented traffic classification is a fundamental technology for modern network security. It is useful to tackle a number of network security problems including lawful interception and intrusion detection [1]. For example, traffic classification can be used to detect patterns indicative of denial of service attacks, worm propagation, intrusions [2], and spam spread. In addition, traffic classification also plays an important role in modern network management, such as quality of service (QoS) control. Many open source and commercial tools [3], [4] with traffic classification function have been deployed and there is an increasing demand on the development of modern traffic classification techniques [1], [5].

While traditional traffic classification techniques may rely on the port numbers specified by different applications or the signature strings in the payload of IP packets, modern techniques normally utilize host/network behavior analysis or flow level statistical features by taking emerging and encrypted applications into account [6], [7]. Recently, substantial attention has been paid on the application of machine learning techniques to statistical features based traffic classification [1]. In the state-of-the-art traffic classification methods, Internet traffic is characterized by a set of flow statistical properties and machine learning techniques are applied to automatically search for structural patterns. These methods can address the problems suffered from by the traditional methods, such as dynamic port numbers and user privacy protection.

Recent research shows that flow statistical feature based traffic classification can be enhanced by feature discretization. Particularly, feature discretization is able to dramatically affect the performance of naive Bayes (NB). NB is one of the earliest classification methods applied in Internet traffic classification [7], which is a simple and effective probabilistic classifier employing the Bayes' theorem with naive feature independence assumptions [8]. Since independent features are assumed, an advantage of the NB classifier is that it only requires a small amount of training data to estimate the parameters of a classification model. However, the performance degradation of NB traffic classifier is reported in the existing works [5], [9]. Lim *et al.* found that the main reason for the underperformance of a number of traditional classifiers including NB is the lack of the feature discretization process [10]. For example, feature discretization can effectively improve the accuracies of the support vector machine (SVM) and k-NN algorithms at the price of lower classification speed. More interestingly, NB with feature discretization demonstrates not only significantly higher accuracy but also much faster classification speed.

Considering complex network situation, a difficult question is that how to obtain a high-performance statistical feature based traffic classifier using a small set of training data. The solutions to this question are essential to address a number of difficult problems in the field of network security and management. For instance, in practice, we may only manually label very few samples as supervised training data since traffic labelling is time-consuming, especially for new applications and encrypted applications. Moreover, a big challenge for current network management is to handle a large number of emerging applications, where it is almost impossible to collect sufficient training samples in a limited time. These observations motivate our work.

In this paper, we provide a solution to effectively improve NB-based traffic classifier with a small set of training samples.

Manuscript received December 06, 2011; revised April 29, 2012; accepted October 03, 2012. Date of publication October 09, 2012; date of current version December 26, 2012. This work was supported in part by ARC Discovery Project DP1095498, in part by ARC Linkage Project LP100100208, and in part by ARC Linkage Project LP120200266. The associate editor coordinating the review of this manuscript and approving it for publication was C.-C. Jay Kuo.

The authors are with the School of Information Technology, Deakin University, Melbourne, 3125, Australia (e-mail: jun.zhang@deakin.edu.au; chao.chen@deakin.edu.au; yang.xiang@deakin.edu.au; wanlei@deakin.edu.au; yong.xiang@deakin.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2012.2223675