INTRUSION DETECTION SYSTEM BY AGGREGATING CORRELATED NAIVE BAYES PREDICTIONS

R.Gowtham [1] C.Rohinth Sundar [1] Balaji Srinath [1],K.Seetharaman[2]
U.G Students[1] Assistant professor[2],
Department Of Computer Engineering, Velammal College Of Engineering And Technology, Madurai, India

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 aredescribed using the discredited 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 trafficdatasets to evaluate the proposed scheme. The experimentalresults show that the proposed scheme can achieve much betterclassification performance than existing state-of-the-art trafficclassification methods. It is useful to tackle a number of network security problems including lawful interception and intrusion detection. In addition, traffic classification also plays an important role in modern network management, such as quality of service (QoS) control. 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. Recently, substantial attention has been paid on the application of machine learning techniques to statistical features based traffic classification.

I. PROBLEM EXISTING SYSTEM

Traffic classification techniques such as dynamic port numbers and user privacy protection. 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.
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.
Itfound that the main reason for the underperformance of number of traditional classifiers including NB is the lack of the feature discretization process.
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.
wehave to only manually label very few samples as supervised training data since traffic labeling is time-consuming.

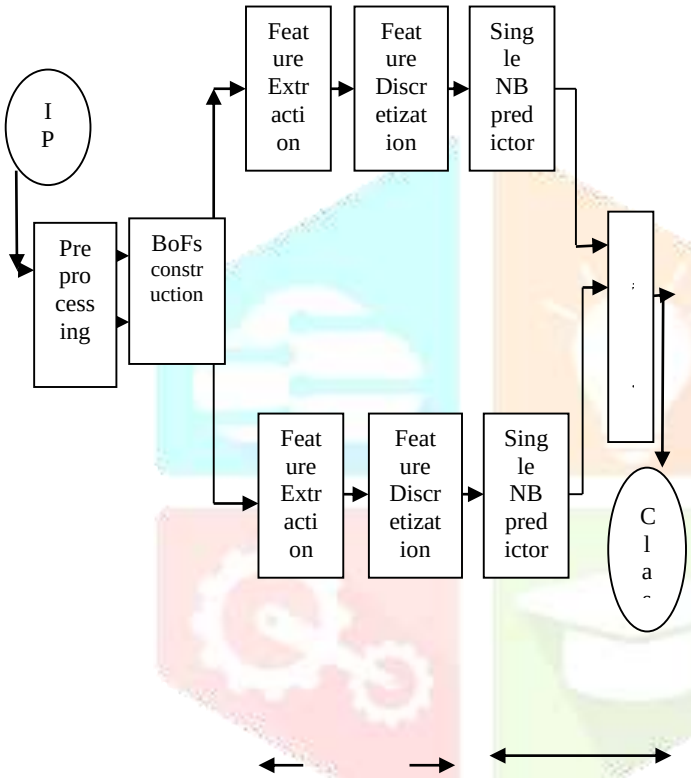
II. SYSTEM IMPLEMENTATION

NB is one of the earliestclassification methods applied in Internet traffic classification which is a simple and effective probabilistic classifier employingthe Bayes' theorem with naive feature independence assumptions.
It assumes independent features.
NB classifier is that it only requires a smallamount of training data to estimate the parameters of a classificationmodel.

ADVANTAGES:

NBwith feature discretization demonstrates not only significantlyhigher accuracy but also much faster classification speed.
NB effectively improves the accuracies of the support vector machine (SVM) and -NN algorithms at the price of lower classification speed.
NB-based traffic classifier improves classification with a small set of training samples.

III. SYSTEM ARCHITECTURE



-Aggregation of c

Analyzing the Data set:

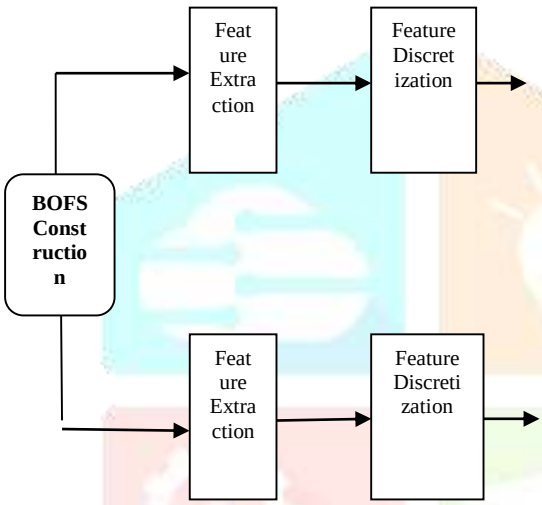
A data set (or dataset) is a collection of data, usually presented in tabular form. Each column represents a particular variable. Each row corresponds to a given member of the data set in question. It lists values for each of the variables, such as height and weight of an object or values of random numbers. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows.The values may be numbers, such as real numbers or integers, for example representing a person's height in centimeters, but may also be nominal data (i.e., not consisting of numerical values), for example representing a person's ethnicity. More generally, values may be of any of the kinds described as a level of measurement. For each variable, the values will normally all be of the same kind. However, there may also be "missing values", which need to be indicated in some way.

Classification Process:

It is based on a flow-level traffic classification. The system captures IP packets crossing a target network and constructs traffic flows by checking the headers of IP packets It isflow-level traffic classification. A flow consists of successive IP packets with the same 5-tuple: source IP, source port, destination IP, destination port, and transport layer protocol.It uses heuristic way to determine the correlated flows and model them. If the flows observed in a certain period of time share the same destination IP, destination port, and transport layer protocol, they are determined as correlated flows and form a BoF. For the classification purpose, a set of flow statistical features are extracted and discretized to represent traffic flows.

A BoF-Based Classification Framework:

In this a set of correlated flows are generated by the same application, which is modeled using a bag of flows BoF.A novel approach is proposed for traffic classification, namely aggregation of correlated NB predictions, which consists of two steps. In the first step, the single NB predictor produces the posteriori class-conditional probabilities for each flow. In the second step, the aggregated



predictor aggregates the flow predictions (posteriori probabilities) todetermine the final class for BoFs.

Aggregation of Correlated NB Predictions

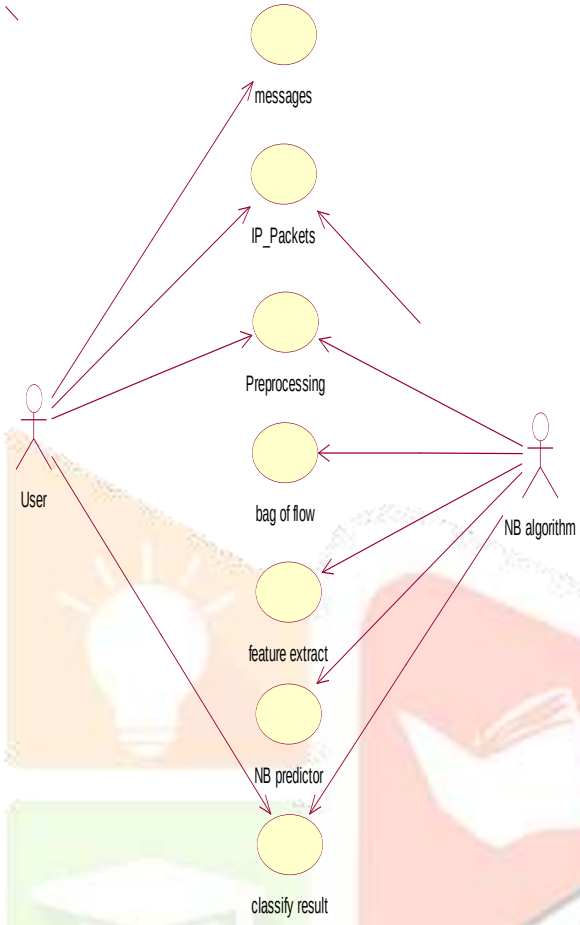
Single NB Predictor: NB algorithm to producea set of posterior probabilities as predictions for each testingflow. It is different to the conventional NB classifier which directlyassigns a testing flow to a class with the maximum posteriorprobability. Considering correlated flows, the predictionsof multiple flows will be aggregated to make a final prediction.

Aggregated Predictor: Under Kittler's theoretical framework, a number of combination methods can be derivedfrom the Bayesian decision theory which can be used for aggregatedpredictor.

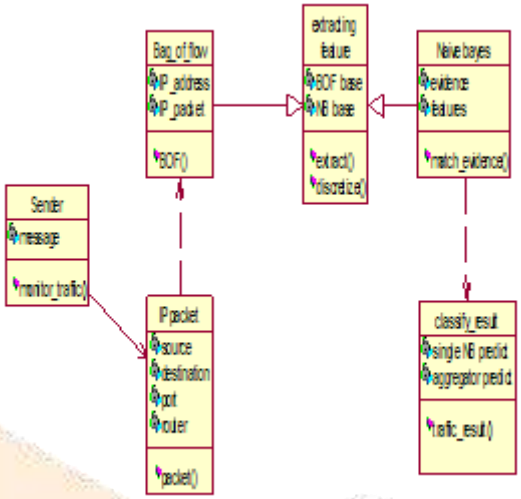
Multi boosting

The effect of combining different classifiers can be explained with the theory of bias-variance decomposition. Bias refers to an error due to a learning algorithm while variance refers to an error due to the learned model. The total expected error of a classifier is the sum of the bias and the variance. In order to reduce bias and variation, some ensemble approaches have been introduced: Adaptive Boosting (AdaBoost) ,Bootstrap Aggregating (Bagging),Wagging and Multiboosting. This is why the idea emerged of combining both in order to profit from the advantages of both algorithms and obtain an overall error reduction.

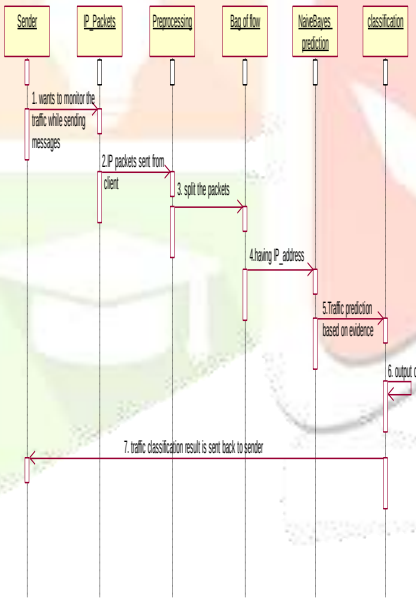
USECASE DIAGRAM:-



CLASS DIAGRAM:-



SEQUENCE DIAGRAM:-



ALGORITHM DESCRIPTION

Naive Bayes PredictionsDefinition:

A **Naive Bayes classifier** is a simple probabilistic classifier based on applying **Bayes' theorem** with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". Naive Bayes belongs to a group of statistical techniques that are called 'supervised classification' as opposed to 'unsupervised classification.' In 'supervised classification' the algorithms are told about two or more classes to which texts have previously been assigned by some human(s) on whatever basis.

Explanation

In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods.

In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of naive Bayes classifiers.[1] Still, a comprehensive comparison with other classification methods in 2006 showed that Bayes classification is outperformed by more current approaches, such as boosted trees or random forests.[2]

An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

Examples

Sex classification

Problem: classify whether a given person is a male or a female based on the measured features. The features include height, weight, and foot size.

Training

Example training set below.

sex	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

The classifier created from the training set using a Gaussian distribution assumption would be:

sex	mean (height)	variance (height)	mean (weight)	variance (weight)	mean(foot size)	variance(foot size)
male	5.855	3.5033e-02	176.25	1.2292e+02	11.25	9.1667e-01
female	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667e+00

Let's say we have equiprobable classes so P(male)= P(female) = 0.5. There was no identified reason for making this assumption so it may have been a bad idea. If we determine P(C) based on frequency in the training set, we happen to get the same answer.

Testing

Below is a sample to be classified as a male or female.

sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

We wish to determine which posterior is greater, male or female. For the classification as male the posterior is given by

$$posterior(male) = \frac{P(male)\,P(height|male)\,P(weight|male)\,P(footsize|male)}{evidence}$$

For the classification as female the posterior is given by

$$posterior(female) = \frac{P(female)\,P(height|female)\,P(weight|female)\,P(footsize|female)}{evidence}$$

The evidence (also termed normalizing constant) may be calculated since the sum of the posteriors equals one.

The evidence may be ignored since it is a positive constant. (Normal distributions are always positive.) We now determine the sex of the sample.

GarouDan (talk)Probably wrong, I did this calculus (twice) and didn't return the same results. Please fix it with the correct ones. The problem is that here we have variance and we need standard deviation

P(male) = 0.5

P(height | male) = 1.5789 (A probability distribution greater than 1 is OK. It is the area under the bell curve that is equal to 1. The formula for calculating probability distribution is P(height | male) = (sample height - mean male height) / standard deviation of male height)

P(weight | male) = 5.9881e-06

P(foot size | male) = 1.3112e-3

posterior numerator (male) = their product = 6.1984e-09

P(female) = 0.5

P(height | female) = 2.2346e-1

P(weight | female) = 1.6789e-2

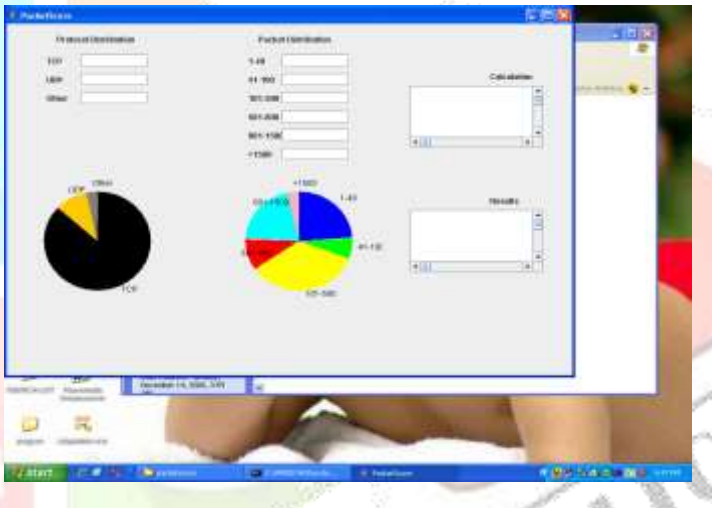
P(foot size | female) = 2.8669e-1

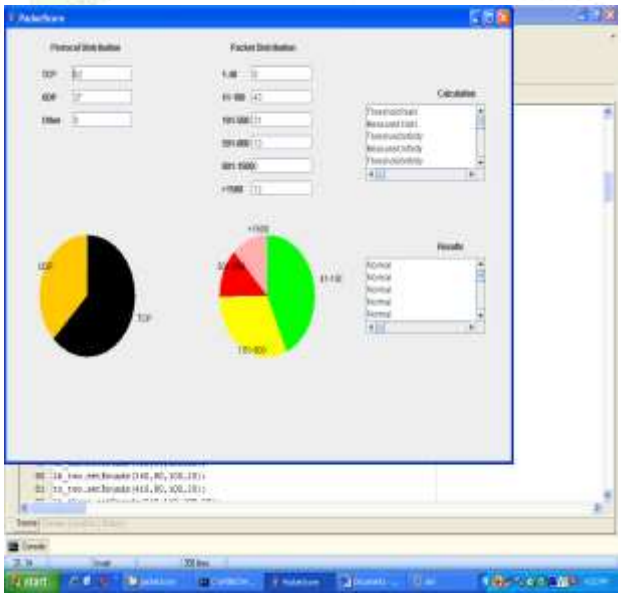
posterior numerator (female) = their product = 5.3778e-04

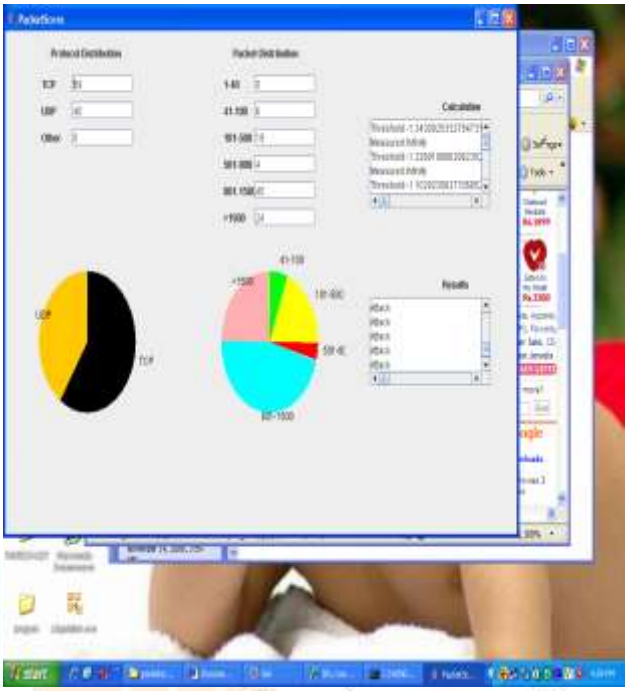
$$evidence = P(male)P(height|male)P(weight|male)P(footsize|male) + P(female)P(height|female)P(weight|female)P(footsize|female)$$

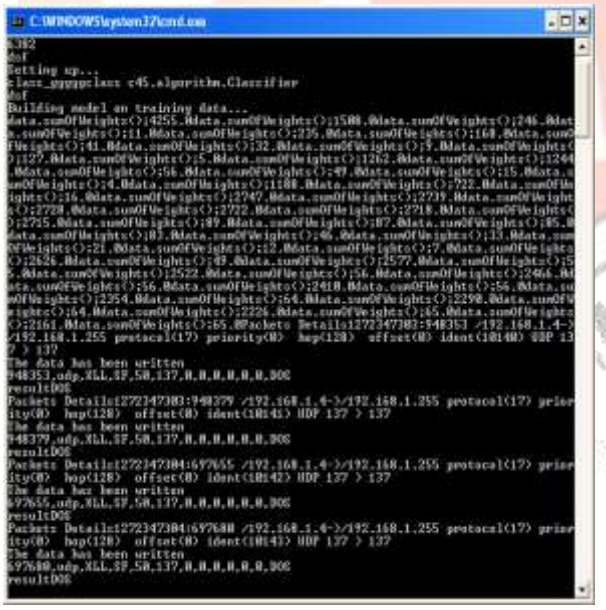
Since posterior numerator is greater in the female case, we predict the sample is female.

IV. RESULT ANALYSIS









V. CONCLUSION

In this paper, we proposed a new traffic classification scheme which can effectively improve the classification performance in the situation that only few training data are available. The proposed scheme is able to incorporate flow correlation information into the classification process. We presented a theoretical analysis on why and how the proposed scheme does work. A new BoF-NB method was also proposed to effectively aggregate the correlation naive Bayes (NB) predictions. The experiments performed on two real-world network traffic datasets demonstrated the effectiveness of the proposed scheme. The experimental results showed that BoF-NB with the sum rule outperforms existing state-of-the-art methods by large margins. This

study provides a solution to achieve high-performance traffic classification without time-consuming training samples labelling.

VI. REFERENCES

[1] T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *Commun. Surveys Tuts.*, vol. 10, no. 4, pp. 56–76, 4th Quarter 2008.

[2] Y. Xiang, W. Zhou, and M. Guo, "Flexible deterministic packet marking: An ip traceback system to find the real source of attacks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 20, no. 4, pp. 567–580, Apr. 2009.

[3] Snort 2011 [Online]. Available: http://www.snort.org/

[4] Bro 2011 [Online]. Available: http://bro-ids.org/index.html

[5] H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee, "Internet traffic classification demystified:Myths, caveats, and the best practices," in *Proc. ACM CoNEXT Conf.*, New York, 2008, pp. 1–12.

[6] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "BLINC: Multilevel traffic classification in the dark," in *Proc. SIGCOMM Comput. Commun. Rev.*, Aug. 2005, vol. 35, pp. 229–240.

[7] A. W. Moore and D. Zuev, "Internet traffic classification using bayesian analysis techniques," in *SIGMETRICS Perform. Eval. Rev.*, Jun. 2005, vol. 33, pp. 50–60.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York: Wiley, 2001.

[9] N.Williams, S. Zander, and G. Armitage, "A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification," in *Proc. SIGCOMM Comput. Commun. Rev.*, Oct. 2006, vol. 36, pp. 5–16.

[10] Y.-S. Lim, H.-C. Kim, J. Jeong, C.-K. Kim, T. T. Kwon, and Y. Choi, "Internet traffic classification demystified: On the sources of the discriminative power," in *Proc. 6th Int. Conf., Ser. Co-NEXT'10*, New York, 2010, pp. 9:1–9:12, ACM.