An Improved Classification of Network Traffic Using Adaptive Nearest Cluster Based Classifier

D.Thuthi Sarabai., M.Sc.,M.Phil[1] ,R. Krissna Priya MCA, (M.Phil)[2]

[1] *HOD, Department of Computer Applications, KG College of Arts and Science,Coimbatore. Tamil nadu, India.*

[2] *Assistant Professor, Department of Computer Science, KG College of Arts and Science,Coimbatore. Tamil nadu, India.*

Abstract— **In modern network security and management architecture, Classification of Traffic plays a major role in recent years. In particular, the process of intrusion detection and QOS control is considered as a essential components in traffic classification. Recent method uses statistical feature related classification approach with machine learning techniques for Traffic classification. Earlier method used several machine learning classifiers for classification purpose. Due to the lack of classifier performance in each aspect, the overall classification of traffic affected while least size of training data are used. To deal with this process, this paper proposes an efficient classifier called adaptive nearest cluster based classifier (ANCC-classifier). The proposed classifier classifies the traffic by collecting the statistical feature based correlated information. Such information is obtained by analysing the normal and abnormal flow of network. The present system is analysed in theoretical and experiential perspectives. Experimental result provides improved performance when compared with the other state of art methods.**

Keywords— **Traffic classification, adaptive nearest cluster based classifier, Nearest Neighbour classifier, network security**

I. INTRODUCTION

Traffic classification [1] is an automated practice which characterizes computer network traffic in relation to a mixture of parameters into a several traffic classes. All resultant traffic class can be considered in a different way with the intention of distinguish the service inferred for the user. As the Internet turn into the mainly service providers, critical communications infrastructure try to retrofit functionality, incorporating privacy, reliability, Multiple service qualities and security into a "best effort" construction initially planned to hold up a research environment. With the aim of protect, prioritize or prevent definite traffic, providers require to employ technology for traffic classification: connecting traffic flows with the applications or its types and that generated them. Once the centre of attention is on detecting particular applications, the concept traffic identification is sometimes used. In spite of the rising dependence on the Internet, there is fundamentally no methodically reproducible corpse of research on worldwide Internet traffic characteristics as a result of the sensitivity and representative restrictions on contribution traffic data. In spite of these limits, economic realities and security concerns have stimulated recent advances in traffic classification. Situational responsiveness of traffic is necessary to mitigation, prevention and response to new arrival of malware, which can abruptly and quickly intimidate legitimate service on network links. Perhaps as significant, the elevated cost of organizing and functioning Internet infrastructure forces providers to repeatedly look for conduct to optimize their network engineering or else increase revisit on capital investments, as well as content-sensitive pricing and application based service differentiation. As a result, the state of the art in traffic classification has practised a major improvement in the past few years, calculated in the various publications and research groups decisive on the topic. Different interests contain led to a fragmented, heterogeneous and rather incompatible landscape.

At present, network environment in conventional methods suffer from a various realistic issues namely encrypted applications and dynamic ports. Recent research has been aimed on the application of machine learning methodologies for traffic classification derived from flow statistical features [2]. Machine learning can automatically look for and illustrate practical structural patterns in a complete traffic data set that is supportive to perform traffic classification [3], [4]. Still, the difficulty of accurate classification of present network traffic is in research related to statistical features which has not been solved.

II. RELATED WORK

In this section, considerable research works were surveyed based on different traffic classification machine learning methods. The literature has been classified into two categories namely unsupervised and supervised methods. The supervised traffic classification methods suggest the supervised training data and construct a conditional function which can forecast the output class for several testing flow. This method often suffers from payload-based traffic classification problems namely user's privacy and encrypted applications. In [5] Moore et.al presented supervised naive Bayes techniques for classifying the network traffic based on statistical flow features. Then in [6] Williams et.al presented supervised methods namely Bayesian network, naive Bayes tree, naive Bayes with kernel density estimation and C4.5 decision tree. After that in [7] Erman et al. presented a unidirectional statistical features for traffic classification in the network centre and presented a method with the potential of calculating the missing features. Finamore et al. in [8] presented application signatures employing statistical categorization of payload and applied supervised algorithms, namely SVM, to carry out traffic classification. Close to the supervised methods anchored in flow statistical features, these payload-based methods necessitate adequate supervised training data. In [9] McGregor et al. presented an

unsupervised methods which groups the traffic flows into a less number of clusters employing the expectation maximization (EM) algorithm and physically label each cluster to an application. Then in [10] Zander et al. utilized AutoClass to cluster traffic flows and presented a measure called intraclass homogeneity for group estimation. Then in [11] Bernaille et al. presented the k-means algorithm for traffic clustering by using a payload analysis tool. Erman et al. [12] evaluated the DBSCAN, AutoClass and K-means algorithms for traffic clustering on two experiential data traces. The experimental research proved that traffic clustering be able to create high-purity clusters as the number of clusters is put as a great deal larger than the amount of real applications.

III. PROPOSED WORK

In In this section, the proposed classifier of adaptive nearest cluster based classifier (ANCC-classifier) is used for traffic classification in networks. The proposed system initially works by generating the unknown flows from the undefined applications. During training, large number of unlabelled flows of traffic and less number of labelled flows of traffic is combined and forms a traffic cluster. In this situation, labelled flow sets are extended by flow label propagation which automatically finds correlated flows in the unlabelled set of traffic. Then traffic clusters are mapped into the corresponding application related traffic classes with the help of labelled flows. From the categorized flows of traffic, the proposed classifier called as of adaptive nearest cluster based classifier (ANCC-classifier) is trained for compound classification on the correlated flows of traffic adaptively with the replacement of individual classification of traffic flows.

A. LABELLED FLOW SET EXTENSION

In this section flow labelled propagation is used for extending labelled flow of traffic in unknown applications. Flow label propagation classifies the traffic flows by related to the statistical features of flow levels. In general, a flow contains five tuples of successive IP Packets. Those five tuples are {Source_IP, Source port, destination_IP, destination_port, transport protocol}.

Consider the pre-labelled flows of traffic to generate the supervised data for mapping the cluster applications. Assume the labelled flow of traffic set as A={X1, X2….} with the labels L= {y1, y2……}, where each flow is a real vector and the dimension of the vector is found by the number of flow statistical properties. The unlabelled flows in collected randomly and it can be denoted as B= {z1, z2…….}. By merging the unlabelled and labelled flow sets, the training set *T* for traffic clustering is obtained as follows(1):

$$T = A \cup B \qquad \rightarrow(1)$$

B. NEAREST NEIGHBOUR CLASSIFICATION TECHNIQUE

In this section, the NN classifier is used for traffic flows of networks. It can be inferred as that a BoF can be used and classified by combining the prediction values of flows generated by the Nearest Neighbour classifier. The prediction value of a flow x generated by the NN classifier is initialized by employing its minimum distance to the training samples of class ω.

$$D(x) = \min_{x' \in \omega} ||x - x'||^2 \qquad \rightarrow(2)$$

The distance value of Query to class is obtained by connecting the distance of flow with the average function as follows:

$$d_Q^{avg} = \frac{1}{||Q||} \sum_{x \in Q} d(x) \qquad \rightarrow(3)$$

At last the flows in Query are classified into the minimum distance of class Query.

Algorithm

1. Estimate statistical features of entire traffic flows.
2. Built BOFs of traffic
3. For a given Query BOF ={x1,……x(n)}
4. Compute prediction values for all labelled and unlabeled flows
5. Combine prediction values to predict the training class samples
6. Assign entire flows in Query to training sample classes

C. ADAPTIVE NEAREST CLUSTER BASED CLASSIFIER (ANCC-CLASSIFIER)

In this section, the proposed ANCC classifier is used for traffic classification in networks. In this, k-means clustering is employed to split the traffic flows into clusters as C = {C1, C2… Ck}, in order to reduce the within-cluster sum of squares:

$$\arg\min_C \sum_{i=1}^{k} \sum_{x(j) \in C(i)} ||x(j) - m(i)|| \qquad \rightarrow(4)$$

where **m**(i) denotes the centroid of C(i) and it is the mean of flows in C(i). The conventional k-means algorithm employs an iterative modification method. For a given set of an initial randomly selected k centroids {$\mathbf{m}_{01}$, $\mathbf{m}_{02}$. . . $\mathbf{m}_{0k}$}, the algorithm performs by changing between the assignment and the update stages due to the adaptive nature of flow characteristics.

The adaptive nature of clustering performs the ellipsoid clustering which covers the each class region with ellipsoids. Each ellipsoid represents a range of points in which Ellipsoids can be produced by an incremental learning process. The ellipsoids are fashioned, constricted, or enlarged regularly at the each training sample. The gradient vector of the point on the boundary finds a direction next to which nearby data points are well removed. This gradient vector is utilized to compute local feature relevance and weighting statistical features consequently. The feature relevance *R*(*x*) can be given as

$$\mathbf{R(x)} = |\mathbf{N_d} \cdot \mathbf{u_i}| = |\mathbf{N_{di}}| \qquad \rightarrow(5)$$

Where d is the nearest point on the boundary to the query, N_d the gradient vector at point d(i) the vector unit. After r is transformed from R(x) by a scheme it could be applied in the weighted distance computation during the NN classification.

Algorithm: ANCC
Input: large flow set T; Label set
Output: Classification of traffic flows

1.Create k clusters C={C(1),C(2)….C(k)} and obtain the centroids {m(1),…..m(k)} by performing k-means on flow set T

1.1. Find ellipsoid E(n) which is nearest to query q

1.2. Find the point d which is the nearest point to q on the E(n).

1.3. Compute the gradient vector N (d)

1.4. Compute $R(x)$ and transform it to r by the scheme

1.5 Use r in weighted distance computation and apply K-NN rule

1.6 Generate the adaptive cluster

2. Predict the flow using NCC classifier based on generated clusters

3. Assign all flows into class according to the majority vote of clusters

4. End.

IV. EXPERIMENTAL RESULTS

In this section, a large number of experiments are performed to systematically estimate the proposed method. Then, the proposed method OF Adaptive nearest Cluster Based Classifier (ANCC-Classifier) is compared with two methods such as when unknown applications are measured in the traffic classification experiments and a traffic classification using nearest neighbour (NN).

Two common metrics are used to calculate the classification performance with un known application situation framework.

Overall accuracy is defined as the ratio of the amount of all correctly classified flows to the sum of all testing flows

$$\text{Accuracy} = \frac{number\ of\ correctly\ classified\ flows}{number\ of\ testing\ flows}$$

F-measure is calculated by

$$\text{F-measure} = \frac{2 \times precision \times recall}{precision + recall}$$

Where precision is defined as the ratio of correctly classified flows over all predicted flows in a class and recall is defined as the ratio of correctly classified flows over all ground truth flows in a specified class.

The comparative graph for the classification of unknown flows with unknown class is illustrated below:

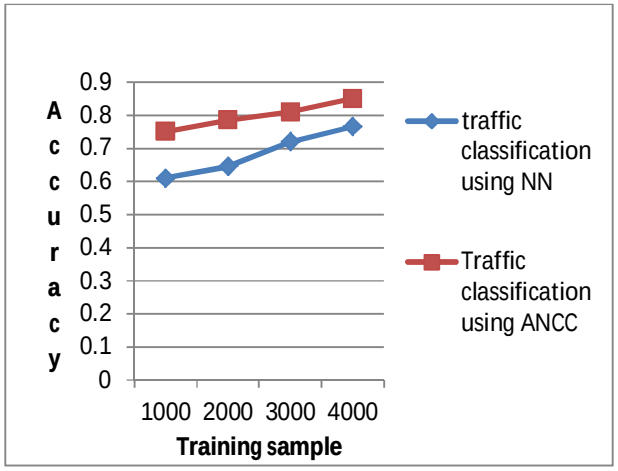

Figure 1.Accuracy comparison graph

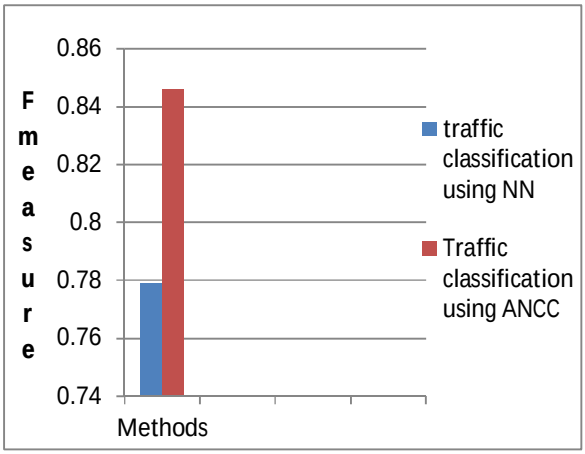

Figure 2: F-measure comparison graph

Thus the above graph in figure 1 and 2 shows that proposed system of traffic classification using Adaptive nearest Cluster Based Classifier (ANCC-Classifier) provides higher accuracy and F-measure when compared with existing method of nearest neighbour (NN) traffic classification with unknown application.

V. CONCLUSION

The present work proposes Adaptive nearest Cluster Based Classifier (ANCC-Classifier) for traffic classification in network. A comprehensive analysis is performed on the system architecture and performance gain from both empirical and theoretical view, which powerfully holds the proposed approach. The system performs by detecting the unknown flows by flow propagation method. The proposed system adaptively clusters the traffic flows from unknown applications and classifies the categorized flows based on clusters. Experimental system provides better accuracy and F-

Measure result when compare with the existing system of work.

REFERENCES

[1] Alberto Dainotti and Antonio Pescapé, University of Napoli Federico II Kimberly C. Claffy, University of California San Diego," Issues and Future Directions in Traffic Classification" IEEE Network • January/February 2012

[2] T.T. Nguyen and G. Armitage, "A Survey Of Techniques for Internet Traffic Classification Using Machine Learning," IEEE Comm. Surveys Tutorials, vol. 10, no. 4, pp. 56-76, Oct.-Dec. 2008.

[3] P. Haffner, S. Sen, O. Spatscheck, and D. Wang, "ACAS: Automated Construction of Application Signatures," Proc. ACM SIGCOMM, pp. 197-202, 2005.

[4] A.W. Moore and D. Zuev, "Internet Traffic Classification Using Bayesian Analysis Techniques," ACM SIGMETRICS Performance Evaluation Review (SIGMETRICS), vol. 33, pp. 50-60, June 2005.

[5] A.W. Moore and D. Zuev, "Internet Traffic Classification Using Bayesian Analysis Techniques," ACM SIGMETRICS Performance Evaluation Review (SIGMETRICS), vol. 33, pp. 50-60, June 2005.

[6] N. Williams, S. Zander, and G. Armitage, "A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification," Proc ACM SIGCOMM, vol. 36, pp. 5-16, Oct. 2006.

[7] J. Erman, A. Mahanti, M. Arlitt, and C. Williamson, "Identifying and Discriminating between Web and Peer-to-Peer Traffic in the Network Core," Proc. 16th Int'l Conf. World Wide Web, pp. 883-892, 2007.

[8] Finamore, M. Mellia, M. Meo, and D. Rossi, "KISS: Stochastic Packet Inspection Classifier for UDP Traffic," IEEE/ACM Trans. Networking, vol. 18, no. 5, pp. 1505-1515, Oct. 2010.

[9] McGregor, M. Hall, P. Lorier, and J. Brunskill, "Flow Clustering Using Machine Learning Techniques," Proc. Passive and Active Measurement Workshop, pp. 205-214, Apr. 2004.

[10] S. Zander, T. Nguyen, and G. Armitage, "Automated Traffic Classification and Application Identification Using Machine Learning," Proc. IEEE Ann. Conf. Local Computer Networks, pp. 250-257, 2005.

[11] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian, "Traffic Classification on the Fly," Proc ACM SIGCOMM, vol. 36, pp. 23-26, Apr. 2006.

[12] J. Erman, M. Arlitt, and A. Mahanti, "Traffic Classification Using Clustering Algorithms," Proc ACM SIGCOMM, pp. 281-286, 2006.