Vinod Mahajan, Bhupendra Verma / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 2,Mar-Apr 2012, pp.1249-1252

Implementation of Distance Based Semi Supervised Clustering and Probabilistic Assignment Technique for Network Traffic Classification

Vinod Mahajan*, Bhupendra Verma**
*Department of Computer Science &Engineering, T.I.T., Bhopal
** Department of Computer Science & Engineering,T.I.T.,Bhopal

ABSTRACT
Network Traffic Classification is an important process in various network management activities like network planning, designing, workload characterization etc. Network traffic classification using traditional techniques such as well known port number based and payload analysis based techniques are no more effective because various applications uses port hopping and encryption technique to avoid detection. Recently machine learning techniques such as supervised, unsupervised and semi supervised techniques are used to overcome the problems of traditional techniques. In this work we use semi supervised machine learning approach and proposed distance based semi supervised clustering and probabilistic assignment technique for network traffic classification. This technique used only flow statistics to classify network traffic. It permits to build the classifier using both labeled and unlabeled instances in training dataset.

Keywords – **Clustering, Classification, Supervised, Semi Supervised, Unsupervised.**

1. Introduction
In recent years, Internet has obtained rapid development in number of new technologies, applications and number of users. It results in great challenges to classify network traffic and associate them with different applications or protocols. At the same time accuracy of traffic classifier is a main basis of network security and traffic engineering [1].

Network Traffic Classification is a process of analyzing network traffic and classifies them into different types of applications and protocols presents in a networks [2]. In past classical techniques such as port number based and payload analysis techniques are more popular to classify the network traffic. Port number based technique [3,4] requires to access the header of the packet to inspect the port number and associate them to application on the basis of IANA's list of well known port number and registered port number [5]. This technique fails to classify the traffic accurately because many application uses dynamic port negotiation and because of ambiguity in port number assignment to application by IANA. Payload Analysis [6, 7] technique was introduced to solve the problems of port number based technique. It needs to access the payload of the packet to find the specific pattern in the payload to classify the traffic. This technique fails because various applications use encryption techniques to avoid detection and legality and privacy law does not allow scanning users payload.

Machine learning is now promising approach to classify network traffic. It uses only flow statistics such as duration, protocol_types, services, flags etc to classify the traffic and does not need to access the header and payload of the packet. Machine learning approach is classified as [8, 9] unsupervised, 6upervised and semi supervised approach. Supervised approach needs the labeled instances to train the classifier. Decision tree, Support Vector Machine, Naïve Bayes etc. are supervised algorithms. Supervised approach has following limitations; first, labeled instances are rare and difficult to obtain. Second, it forces mapping of instances to one of the known class without detecting new ones. Unsupervised approach is a class of machine learning in which unlabeled instances are used and based on the inner similarity between instances clusters (groups) are formed. K-Means, DBSCAN, CLARANS etc are unsupervised algorithms. It has limitation in assigning label to cluster after clustering so that new instances properly mapped to applications. Semi supervised approach is a combination of supervised and unsupervised approach [10].

To overcome the limitations of supervised and unsupervised approach a distance based semi supervised clustering and probabilistic assignment technique for network traffic classification is proposed. The proposed technique permits both labeled and unlabeled instances to build the traffic classifier.

Remainder of this paper is organized as follows. Section II presents related work in semi supervised approach. Section III describes proposed work. In section IV Dataset and analysis tool used in experiment is described whereas in section V includes performance evaluations. Finally last section concludes this work.

2. Related work
Much work has been done in the field of network traffic classification. This section explains works related to semi supervised approach only.

In 2007 Jeffery Erman et al.[11,12] proposed a semi supervised traffic classification technique consists of two steps clustering and classification. For experimental purposes traces are collected from the internet link of a large university in which 29 application are identified. The authors categorized traces as 1hour campus, 10 hour residential and 1 hour wireless LAN. They performed various experiment on this work. In first experiment 64000

unlabeled flows are provided for clustering after these flows is clustered, the fix numbers of random flows in each cluster are labeled. The results show that 94% accuracy is achieved by two labeled flows per cluster and K=400. The second set of experiments 80,800 and 8000 labeled flows are mixed with random number of unlabeled flows to generate the training dataset. The accuracy will increase when five or more flows are labeled per cluster.

In 2008 Chuanliang chan et al.[13] proposed two graph based semi supervised methods(i.e. spectral graph transducer, Gaussian fields approach) and one semi supervised clustering (MPCK Means) method to perform intrusion detection .KDD CUP 99 dataset is used for experimental purpose and PPrecision and PRecall and PF Measure is used to evaluate the clustering results..The authors compared two semi supervised classification with other traditional supervised algorithm and finds that performance of their approach are much better than other .Also show that the performance of MPCK means is better than KMeans.

In 2009 Levi Lesis and Jorg Sander [14] proposed semi supervised algorithm called as SSDBSCAN. This algorithm requires only one input parameter, does not need user intervention and automatically finds noise objects. The authors used both artificial and real world datasets for experimental purpose and compare SSDBSCAN with HISSCLU and finds that their approach is better to find the cluster in datasets.

The Liu bin and Tu Hao in 2010 [15] proposed semi supervised clustering methods based on particle swarm optimization (PSO) algorithm and two host feature name IP address discreteness and success rate of connections. They collected experimental dataset from the router of their university which contained 7 classes. To evaluate their approach they used precision. Result showed that 85% accuracy achieved when 100 or more labeled samples are used in training dataset.

In 2010 Amita Shrivastav et al. [16] proposed a semi supervised approach based on clustering algorithm. This approach has two steps clustering and classification. KDD CUP-99 dataset is used for experimentation. They compare their approach with SVM based classifier. The experimental result showed that accuracy of proposed classifier lies between 70% and 96% for various datasets.

3. Proposed Technique

In this paper a semi supervised approach is used for network traffic classification. It has two phases training phase and testing phase. The proposed technique, a distance based semi supervised clustering and probabilistic assignment is used in training phase to build (train) the classifier. It permits both labeled and unlabeled instances used in training dataset.

Following are the steps in proposed technique

1. Data Preprocessing

Normalization is used for data preprocessing, where the attribute values are scaled so as to fall within a small specified range such as 0.0 to 1.0. In this work for normalization the attribute values are divided by the largest value for that attribute present in the dataset.

2. Distance based semi supervised clustering using K-Means

The K-Means algorithm is used to partition the dataset into number of clusters. The K-Means algorithm uses Euclidean distance measure to find the similarity between instances.

3. Probabilistic assignment

For assigning label to clusters formed in second step the probabilistic assignment technique as in [11] is used. To complete the mapping, within each cluster an assessment is performed to find out to which class has maximum probability of data instances belongs. On the basis of this probability the class label is assigned to clusters. This mapping forms the basis for classification model.

The Classifier build (train) in training phase by proposed technique is used to classify the traffic in testing phase.

4. K-Means Clustering

The K-Means clustering algorithm [10, 20] is a simple and popular analysis method. The following are the steps in K-Means clustering as in [20]-

1. Define K= number of clusters.
2. Partition the training dataset into K clusters and assign the training instances as the following:
   2.1 Take the first K training instances as a single element cluster.
   2.2 Assign each of the remaining (N-K) training instances to the cluster with the nearest centroid. After each assignment, centroid of the gaining cluster is recomputed.
3. Take each instance in sequence and compute its distance from the centroid of each of the clusters. If a instance is not currently in the cluster with the closest centroid, switch this instance to that cluster and update the centroid of the cluster gaining the new instance and the cluster losing the instance.
4. Iterate step 3 until convergence is achieved.

In this paper we use the Euclidean distance as the similarity measure as required in step 2 and step 3 in K-Means clustering algorithm. Euclidean distance [10] is defined as:

$$\text{Dist}(X, Y) = \left(\sum_{i=1}^{n} (Xi - Yi)^2\right)^{1/2} \quad (1)$$

Where X=(X1, X2… Xn) and Y = (Y1, Y2... Yn) are two n dimensional data instances [34].

5. Experimental Setup

This section describes the dataset and analysis tool used in the experiment.

5.1. Dataset Description

This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 [17]. This dataset contains 41 features such as duration, Protocol_type, service, flag etc. The raw training data was about four gigabytes of compressed binary TCP dump data from seven weeks of network traffic. This was processed into about five million connection records. Each connection is labeled as either normal, or as an attack, with exactly one specific attack type. Each connection record consists of about 100 bytes. Attacks fall into four main categories Probe, DoS, U2R and R2L.

The dataset used for experiments contains 8,000 records. We take only 8000 instances form KDD CUP dataset because to test the effectiveness of proposed approach by using whole dataset requires more computational time and resources. This data set is divided into training dataset which contains 6000 records and test dataset which contains 2000 records. Training dataset contains 2400 labeled instances and 3600 unlabeled instances. Both training and test dataset contains all 41 features.

5.2 Analysis Tool

The experiments were conducted using MATLAB 7.3. The name MATLAB stands for matrix laboratory. MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation [18].

6 Experimental Results

6.1 Performance Evaluation Parameters

To perform the evaluation the following evaluation parameters are used.

6.1.1 Evaluation Matrix

In a multi class prediction, the result on a test data set is denoted as a two dimensional Evaluation matrix with row and column for each class. Where row represents actual class and column represents predicted class for a matrix element. Correct classified instances are measures from the diagonal of confusion matrix. It is a useful tool for analyzing how well your classifier can recognize instances of different classes [8, 10].

Table 1. Evaluation Matrix

Actual Class	Predicted As	
	CS	IT
CS	TP	FN
IT	FP	TN

Where True Positive (TP) is The number of correctly classified member of Class CS (interested) as a class CS. True Negative (TN) is The number of tuples that are correctly classified as not a member of class CS. False Positive (FP) is The number of tuples that are incorrectly classified as belonging to class CS. False Negative (FN)is The number of tuples that are incorrectly classified as not belonging to class CS.

From the Evaluation matrix following parameters will be computed.

Overall Accuracy: It is the ratio of sum of TP of all classes to the number of instances present in the dataset.

Precision: It is the ratio of number of instances correctly classified to the number of instances that are correctly and incorrectly identified [8,19].

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (2)$$

Recall: It is the ratio between the numbers of instances of class correctly classified and total number of instances of that class. It is equivalent to the true positive rate (TPR) [8, 19].

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (3)$$

6.2 Performance Evaluation

It is necessary to evaluate the performance of the technique being implemented. We evaluating the performance of classifier at number of clusters equal to 50.To do so Evaluation matrix is computed for test dataset and it is shown in Table 2. Amongst all the testing instances it is determined how many instances are incorrectly classified and correctly classified.

Table 2. Evaluation Metrics for Test Dataset

Actual Class	Predicted as				
	Normal	Probe	Dos	U2R	R2L
Normal	388	0	0	10	02
Probe	3	382	6	4	5
Dos	10	0	390	0	0
U2R	23	0	0	351	26
R2L	7	0	0	8	385

From Table 2 we calculate the overall accuracy of the classifier and it is 94.8% at number of cluster =50.

Table 3shows the precision and recall of each class calculated from the evaluation matrix. Fig. 1 and fig. 2 are plotted from the Table 3. Several observations can be made from the table 4, fig. 1 and fig.2. First, more than 92% precision achieved for all classes. Second, probe class achieved 100% precisions i.e. the instances belong to other class are not classified as belongs to probe class. Third, the normal class achieved lowest precision indicates that the other instances are misclassified as belongs to this class as compared to others. Forth, more than 95% recall achieved for all classes. Fifth, DoS class achieve 97.5% recall i.e. the instances belong to this class are more correctly classified. Sixth, U2R class achieved lowest recall values i.e. the large number of instances belongs to this class are misclassified as compared to the other classes.

Table 3. Precision and Recall

Class	Precision (%)	Recall (%)
Normal	90.02	97
Probe	100	95.5
DoS	98.48	97.5
U2R	94.63	87.75
R2L	92.10	96.25

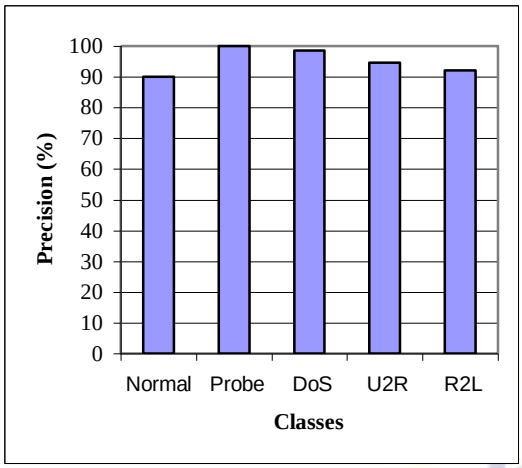

Figure 1. Precision of classes

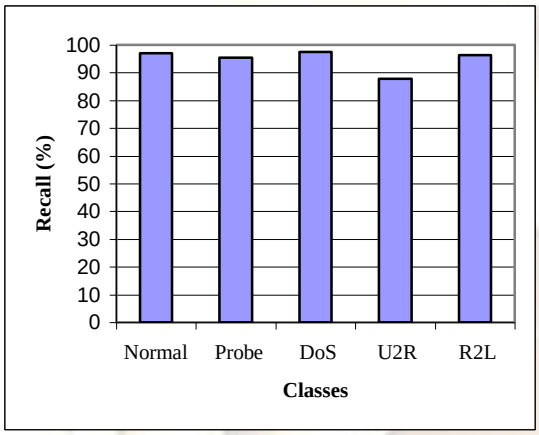

Figure 2. Recall of classes

7. Conclusion

In this paper we presented a distance based semi supervised clustering and probabilistic assignment technique to build the classifier and it has been achieved successfully. It permits both labeled and unlabeled instances to be used in training the classifier. The classifier achieves 94.8% accuracy at K=50. It is observed that the accuracy of the classifier depends on the number of clusters and initial centroids selected in K-Means algorithm. This technique will be used in real time traffic classification in future.

REFERENCES

[1] Jian- Minwang, Cheng-Lu Qian, Chun- Hui Che and Hai-Tao He, "Study on process of Network Traffic Classification using Machine Learning ", *The Fifth annual China Grid Conference*, 2010, 262-266

[2] Alberto Dainotti, Walter de Donato, Antonio Pescape and Pierluigi Salvo Rossi, "Classification of Network Traffic via Packet Level Hidden Markov Models", *In IEEE GLOBECOM* , New Orleans, LO, Dec.2008, 1-5.

[3] Patrick Schneider, *TCP/IP traffic Classification Based on port numbers*, Diploma Thesis, Division of Applied Science, Harvard University, 29 Oxford Street, Cambridge, MA 02138, USA, 1-6.

[4] Alok Madhukar and Carey Willamson,"A Longitudinal study of P2P Traffic Classification", *in proceeding of the 2nd IEEE International Symposiun on MASCOTS'06*, Los Alamitos, California, 2006, 179-188.

[5] IANA, "Internent Assigned Numbers Authority", http://www.iana.org/assignments/port-numbers

[6] Subhabrata Sen, Oliver Spatscheck, and Dongmesi Wang, "Accurate, Scalable In- network Identificatio of P2P Traffic using application Signature", *WWW 2004*, New York, USA ACM, May 17-22, 2004, 512-521.

[7] Geza Szabo, Istyan Szabo, and Daniel Orincsay,"Accurate Traffic Classifciation", *in IEEE international symposium on a World of Wireless Mobile and Multimedia Networks (WoWMoM)*, Espoo, Finland, June 2007, 1-8.

[8] I.Witten and E.Frank, *Data mining: practical machine learning tools and techniques* (Second Edition, Morgan Kaufmann Publishers, 2005).

[9] Thuy T.T. Nguyen, Grenville Armitage, " A Survey of Techniques for Internet Traffic Classification using Machine Learning", *In IEEE communication surveys and tutorials*, 2008, 1-21.

[10] Jiawei Han and Micheline Kamber, *Data mining – concepts and technique* (Second Edition Morgan Kaufmann publishers an imprint of Elsevier, 2005).

[11] Jeffrey Erman, Anirban Mahanti, Martin Arlitt, Ira Cohen and Carey William Son, "Semi-Supervised Network Traffic Classlification", *in proc. of ACM SIGMETRICS'07*, San Diego, California, USA, *vol.35*, June 2007, 369-370.

[12] Jeffrey Erman, Anirban Mahanti, Martin Arlitt, Ira Cohen, and Carey Williamson, *Offline/Realtime Traffic Classification Using Semi-Supervised Learning*, Technical Report, Department of Computer Science, University of Calgary, February 2007. (Manuscript 22 pages)

[13] Chuanliang Chen, Yunchao Gong and Yingjie Tian, " Semi-Supervised Learning Methods for Network Intrusion Detection", *in proc. of IEEE International Conference on Systems Man and Cybernetics (SMC)*, 2008 , 2603-2608.

[14] Levi Lelis and Jorg Sander,"Semi-Supervised Density-Based Clustering", *in proc. of Nineth IEEE International Conference on Data Mining* , Miami, FL , Dec 2009, 842-847.

[15] Liu Bin and Tu Hao, "An Application Traffic Classification Method Based on Semi-Supervised Clustering", *A 2nd International Symposium on Information Engineering and Electronics Commerce (IEEC)*, 2010,1-4.

[16] Amita Shrivastav and Aruna Tiwari,"Network Traffic Classification using Semi-Supervised Approach", *Second International Conference on Machine Learning and computing (ICMLC)*, 2010, 345-349.

[17] KDD CUP 1999 dataset available at ,http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[18] Rudra Pratap, *Getting started with MATLAB 7* (OXFORD University press, Indian Edition).

[19] Nigel Williams Sebastian Zander, Grenville Armitage, *Evaluating Machine Learning Algorihtms for Automated Network Application Identification*, CAIA Technical Report 060410B, March 2006, 1-14.

[20] Teknomo and Kardi, K-Means Clustering Tutorials, http:\\people.revoledu.com\kardi\ tutorial\kMean\.