A Survey of Classification Algorithms for Network Traffic

R.Deebalakshmi [1]Dr.V.L.Jyothi[2]
[1]Research scholar, Sathyabama University, Chennai,deepar11@gmail.com
[2]Professor/Head of Dept, Jeppiaar Engineering College, Chennai

***Abstract*- Network traffic in the world wide is calculated to rise every year twice the times. To keep pace and profit from this increased amount of flows efficiently. And offer new services. Some efficient techniques needed. Day by day new applications are invented and they have heterogeneous nature in network environment and communication between these new devices also a critical part. improving the network performance, establish proper service policies in router, handling network security risks, management of network operations and provide Qos services to users in internet. To solve these issues classification techniques are used. In this survey different classification algorithms are discussed. K-means algorithm, classification using clustering algorithm, Classification based on Fuzzy Kernel K-means Clustering, Support vector machine algorithm, and self-learning classifier Bayesian classification, C5.0 and traffic classification using correlation information and robust network traffic algorithms are presented.**

I. INTRODUCTION

Traffic classification has customary ever-increasing attention in the most recent years. Traffic Classification is only way to categorize different applications and protocols that be present in a network. To improving the network performance, a different action takes place from the categorized traffic like monitoring, detection, control and optimization. After the classification packets are categorized as application or protocol. They are called as flags. These flags facilitate the router establish proper service policies to be applied for those flows.

Classification is essential for network security solutions due to large increase in internet usage with much application that lead to a huge mixture of traffic flowing over networks. And efficient classification methods are useful for internet service providers to ensure quality of services to their customers.

It is the fundamental block that is necessary to permit any traffic management operations, from differentiating traffic pricing and treatment like policing, shaping. And security operations like firewalling, filtering, anomaly detection.

Network traffics are 3 major types.

- Sensitive -packets are delivered on time examples are online gaming, video conferencing and web browsing
- Best-effort -packet time is non-detrimental examples are email, peer-to-peer.
- Undesired -delivery of spam and traffic created by worms, botnets and other malicious attacks.

The two traditional methods for classifying traffic:

- Categorize the packets based on standard port numbers; it is not efficient scheme because some application use dynamic port numbers and other applications port number. The port-based categorization is useless for classifying peer to peer applications, 30–70 % of the Internet traffic is classified as "unknown" by this scheme.
- Categorize the packet based on the payload. Payload information taken from layer 4. In this scheme signature in payload packet checked up to date. Because of signature concept dynamic port number use problem in port based method is solved. it is slow and takes more processing power.
- Classification based on a statistical method, in this method traffic activity based on packet arrival time and session time.

Flow Description and Classification

Flow is defined as packets moved from one place to other place over a period of time. And flow is classified based on following tuples. Flows are taken from forward and backward direction for classification. Flows categorized based on length of packet, packet arrival time and bytes transferred.

Flow information table as shown below.

The following scheme show how the flows are formed. Traffic traced at the gateway of the network and categorized for more study.

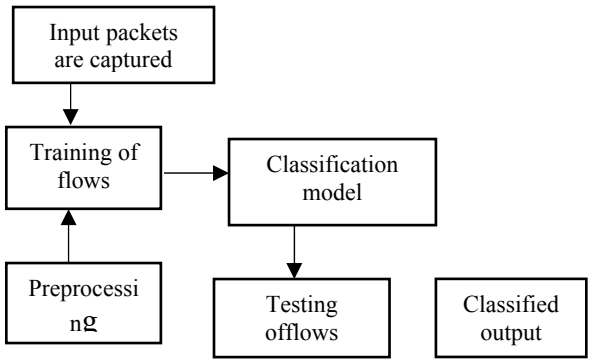

Figure 1.Classification Model

Classification performance measures

Generally three metrics are used to evaluate the performance of classification algorithm accuracy, precision and recall.

$$\text{Accuracy} = \frac{\sum_{i-1}^{n} TPi}{\sum_{i-1}^{n} TPi + \sum_{i-1}^{n} FPi} * 100$$

$$\text{Precision} = \frac{TP}{TP + FP} * 100$$

$$\text{Recall} = \frac{TP}{FN + TP} * 100$$

True Positive: percentage of flows matched to given class.
False Positive: percentage of flows not matched to given class.
False Negative: percentage of flows incorrectly matched and not belongs to given class.
True Negative: percentage of flows correctly matched as not belonging to given class.
The paper is arranged as follows. Section II describes basic idea about flows and the classification process. In the Section IV, different clustering algorithm has been discussed. Finally, the section III summarizes the conclusion of this work.

II. CLASSIFICATION ALGORITHMS

A. ***K-Means***

In [1], K-Means clustering algorithm is famous partition scheme. Packets traced in computer networks. From traced packets flow is defined, and flows are grouped into clusters. Flows classified based on k-cluster centres. The cluster centres are defined by packet length and packets arrival time of application. Applications which are having same packet length or nearby packet length are compared with existing cluster, if it is matched combined to existing cluster otherwise new cluster group if formed. The algorithms continues still all traced packets are made as a cluster groups
The algorithm steps are given below.
STEPS
A= {a1, a2…an} collection of packet inter arrival time. B= {b1, b2….bn} collection of cluster centres
1. Initialize, k cluster centre.
2. Calculate the packet Inter arrival time for all flow
3. Allot each flows into cluster whose packet inter arrival time is equal or minimum.
4. Redefine the cluster centre by

Yi= (1/Ki) $1 \sum_{i=1}^{n} X\, i$

Ki- represents the number of flows in the i th cluster.
5. Redefine the time and cluster centre for all flows.
6. Stop when all flows assigned otherwise repeat from step 3.
K-mean algorithm is simple to comprehend, and gives better result for varying data set. And it is faster compared with hierarchical clustering if k value is small, even though it is not suitable for diverse cluster size.
.

B. Classification using Clustering and Association rule mining

In [2] the goal of classification is to get healthy and unfailing clustering by applying the data mining techniques. This method classifies traffic also produce the behavior pattern report of flows. For Classification traditional clustering steps and for association rule apriori algorithm is used.
The classification metrics reaches very high values use of Model-based clustering and rule base categorization (overall accuracy 94% and perfect HTTP accuracy). apriori algorithm is used to produce Behavior pattern report for all network applications. Using rule evaluation parameters flow attributes interaction is analyzed. After the clustering association rules applied this process is independent of data set, so algorithm identifies flow details, feasible nameless applications.
STEPS

1. Get the traffic traces.
2. Clustering algorithm is used to implement classification
3. Association rules are derived
4. Association between flow parameter is defined
5. Final result is fed back to the classification model.

This method gives better accuracy using Model based clustering with association rule mining, rules are framed automatically for all datasets it is suitable for next generation applications.

C. Traffic Classification using Clustering Algorithm

In [3] this methods present a different approach to classify traffic by make use of the unique character of applications while they communicate on a network. It uses two unsupervised clustering algorithms, namely K-Means and DBSCAN, that have beforehand not been used for network traffic classification. The K-Means algorithm produces clusters that are spherical in shape whereas the DBSCAN algorithm has the ability to produce clusters that are non-spherical. The AutoClass algorithm uses a Bayesian approach and can automatically determine the number of clusters.
AutoClass**:** AutoClass uses the Expectation Maximization (EM) algorithm. The EM algorithm has two steps: an expectation step and a maximization step. The initial expectation step guesses what the parameters are using pseudo-random numbers. In the maximization step, the mean and variance are used to re-estimate the parameters continually until they converge to a local maximum. These local maxima are recorded and the EM process is repeated.
K-Means**:** The K-Means algorithm partitions objects in a data set into a fixed number of K disjoint subsets. For each cluster, the partitioning algorithm maximizes the homogeneity within the cluster by minimizing the square-error.
DBSCAN**:** The DBSCAN (Density Based Spatial Clustering of Applications with Noise) algorithm was chosen as a representative of density-based algorithms. The DBSCAN algorithm is based on the concepts of density reach ability and density-connectivity. Although DBSCAN has lower accuracy

compared to K-Means and AutoClass, DBSCAN produces better clusters.

The three algorithm performance compared by ability of making cluster classes with more connections. Autoclass gives more accuracy than other two methods, DBSCAN gives more connection among clusters and k-mean accuracy is lower than other two but it is efficient in building faster cluster classes.

D. Classification based on Fuzzy kernel K-means Algorithm

In [4] This system discuss Internet traffic classification based on fuzzy kernel K-Means clustering to solve the drawback of the fuzzy K-Means clustering algorithm to meet the requirements of the Internet network classification.. This method overcomes the dependence of clustering algorithm on sample distribution form. Fuzzy kernel K-means clustering consists of three main phases, namely preprocessing phase, clustering phase and mapping phase.

Preprocessing Phase: In the stage of preprocessing phase, network monitors can record statistics such as source-destination IP pairs and connection characteristics for each flow. A flow is defined to be as a series of packet exchanges between two hosts, identifiable by the 5-tuple (source address, source port, destination address, destination port, transport protocol). Prior to the ML modeling, feature selection can be executed off-line.

Clustering Phase: In the stage of clustering phase, it employs a machine learning approach called fuzzy kernel K-means clustering to partition a training data set that consists of scarce labeled flows combined with abundant unlabeled flows. Clustering partitions the training data set into disjoint clusters such that flows within a group are similar to each other whereas flows in different groups are as different as possible.

Mapping Phase: In the stage of mapping phase, it uses the available labeled flows to obtain a mapping from the clusters to the different known classes. Eventually, the identification output would be applied to network activities such as network surveillance, QoS.

This algorithm solves dependency on the shape of sample space in peer to peer application, and gives more cluster accuracy.

E. SVM

In [5] machine learning SVM it uses supervised learning algorithm and analyze data used for classification and regression. in supervised learning data sets are already trained and labeled in database, input data's are compared with trained data sets and labeled as known flows. If no trained set that method is called unsupervised algorithm, here input flows are grouped as clusters.

There are two types of SVM.

1. Linear SVM and
2. Non- linear SVM.

Linear SVM:

In linear SVM, it has linear n-dimensional vector and if there is a possibility of separating n-1 hyper plane vector it is called a linear classifier. This hyper plane is used to classify data. And optimal hyper plane will give largest separation or margin , Bad selection of hyper plane will give noise data classification. Distance of data compared with hyper plane so maximum value stetted for it.

Non-Linear SVM:

In non-linear SVM, the classifier is produced by applying the essence deception to maximum-margin hyper planes. They are efficient in high dimensional spaces. They are still efficient in cases where number of dimensions is greater than the number of samples. It uses a subset of training points in the decision function and so it is memory efficient.

Disadvantages:

SVM has some disadvantages. Particularly, if the number of features is much greater than the number of samples, the method is likely to give poor performance. They do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

One-Class SVM:

Another alternative of SVM is one-class SVM. It is an unsupervised algorithm that learns a decision function for originality detection that it classifies new data as similar or different to the training set. Basically, it also separates all the data points from the origin and maximizes the distance from this hyper plane to the origin. The optimal hyper plane is the one that represents the largest separation, or margin, between the two classes.

Disadvantages:

If the number of samples is more than the creation of optimal boundary is difficult and it performance is also affected. It uses only unsupervised algorithm. Hence, it cannot identify new traffic.

F. Classification using Correlation algorithm

In[6] K-nearest neighbor is not uses training set. Though, it do not provide good performance. No over fitting of datasets, it handles huge amount of classes and performances is degraded by small training data sets. It follows the non parametric approach. In this algorithm discovers the correlated information from input flows this information is integrated with trained data set.

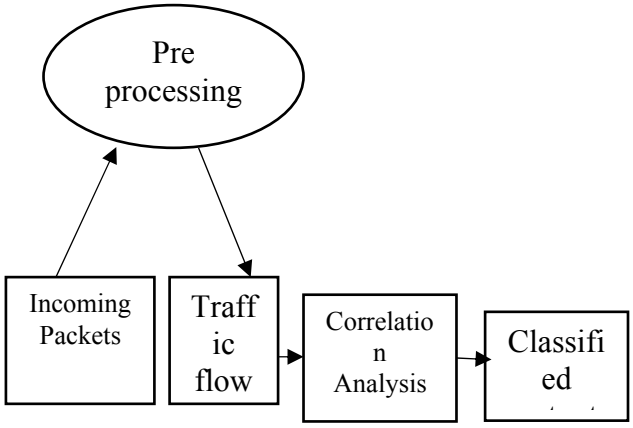

Figure 2: classification using correlation

In pre-processing traffic traced from networks and IP flows are formed. From IP flow correlation information is discovered and it is forwarded to the classifier. Classifier produce the classified output flows based on application wise.

G. Classification using Bayesian analysis algorithm

In [7] machine learning Naïve Bayes technique to categorize internet traffic based on application. The traffic in the internet applications were classified into different categories, e.g. mail, web services, p2p,multimedia and games. The authors used accuracy as a classification metric to evaluate performance of classifier. This result depicted that naïve bayes techniques have 65 % accuracy in classification. Two refinements were performed for improving classification accuracy using naïve bayes kernel estimation and fast correlation based filter method. It gives 95% as the overall accuracy. This work is extended by [8] Using application of Bayesian approach in neural network. This result produces 99% accuracy.

H. Self learning Classifier

In [9] author discussed a new method called SeLeCT, a Self-Learning Classifier for Internet Traffic. It uses unsupervised algorithm to automatically group the flow into homogeneous group. It doesn't require prior knowledge of training set to identify the traffic flows. Automatically this algorithm groups the flows into homogeneous clusters using statistical features. It also simplifies the label assignment by assigning labels to the cluster based on application. Furthermore, it uses self seeding approach to process next batches of flows before assigning labels to previous cluster. The author evaluates the performance of SeLeCT using different traffic traces collected from ISP located in the different continents. The experiments showed that it achieves overall accuracy .The accuracy is achieved is nearly 98% and it discover new protocols and application in the traffic traces.

I. C5.0 Classifier

In [10] C5.0 is the decision tree based algorithm and use the concept of ML algorithm. It is simple and memory efficient. the decision trees are generated based on training set .decision trees are used to classify the flow classes. The c5.0 classifier uses command line interface to produce the rules for decision tree and test the classifier. The test was executed many times using different set of training flows and attributes like packet size, packet length, number of flows.

J. Robust Traffic Classification Algorithm

In[11] The Robust Traffic Classification (RTC) works by combining both supervised and unsupervised machine learning techniques to meet this challenge. It has the capability of identifying the traffic of unknown applications as well as accurately discriminating predefined application classes. In addition, it uses a new method for automating the RTC scheme parameters optimization process.

Existing traffic classification methods suffer the problem of zero-day traffic due to a lack of new zero day traffic samples in the classifier training stage. How to obtain sufficient new traffic samples becomes a key question for fundamentally solving this problem. The system intend to build a robust classifier by extracting zero-day samples and incorporating them into the training stage.

The traffic classification algorithm works in three stages: unknown discovery, "bag of flows" (BoF)-based traffic classification, and system update. The module of unknown discovery aims to automatically find new samples of zero-day traffic in a set of unlabeled traffic randomly collected from the target network. The module of BoF-based traffic classification takes pre labeled training samples and zero-day traffic samples as input to build a classifier for robust traffic classification. To achieve fine-grained classification, the module of a system update can intelligently analyze the zero-day traffic and construct new classes to complement the system's knowledge. RTC uses the algorithms of random forest and K-means are employed to perform supervised classification and unsupervised learning (clustering).

3. Conclusion

Classification is needed for overwhelming packets with heterogeneous nature in network environment and offer new service, and communication between new devices. In this study different clustering algorithm for traffic classification is reviewed. The challenges faced in the each classification algorithm and suggested ways to improve the performance of classification accuracy is discussed. This survey also describes about the general characteristics and methodologies used in clustering algorithm. Accuracy is selected as a performance metrics for clustering algorithms. This survey will lead to easily identify the clustering algorithm for traffic classification and helps to choose the clustering algorithm effectively when different traffic traces collected in different application.

IV SUMMARY OF CLUSTERING ALGORITHMS

Author	Classification Algorithm	Features	Data traces	Considered traffic	Overall Accuracy
Jeffrey Erman [3]	k-mean	Inter arrival time	Auckland iv,Calgary	Irc,pop3,http ,Limewire,NNTP,Socks	92%
Jun zhang [6]	NN	Packet size,bytes	Wide ,isp	Dns,http,imap,ftp,p2p	60-80%
Umang chaudry [2]	Model based clustering	Packets,bytes	WITS,CRAWDAD	http,smtp,dns,mail	91-95%
Ratih Agarwal [12]	Lower ID clustering	Transmission range,packets	Gps	Smtp,http	Based on chosen node
Moore and Zuev [7]	Bayesian Technique	Packet size, Inter arrival time, Flow duration	Proprietary Hand based traces	Mail,p2p,www	95-98%
Ngugen and Armitage [13]	Supervised Naïve bayes	Inter arrival time, packet length	Traces collected from game server	Http,dns,smtp	96-99%
Luigi Grimaudo [9]	Self learning classifier	Flow size,subset, Inter arrival time	Traffic traces from isp	http,gmail,Rtsp,BitTorrent, POP3,Telnet, eMule	97-98 %
Williams et al [14]	C4.5 decision tree	Inter arrival time, Packet length	NLANR	Smtp,Dns	93.5-99%
Jeffrey Erman [4]	AutoClass	Mean packet length ,flow duration	Auckland IV,Calgary	Smtp,socks, dns http	87-93.5%
Alice Este et al [15]	SVM(support Vector machine)	Packet size,bytes	CAIDA	http,ftp,pop3,bittorent	

REFERENCES

[1] T.T. Nguyen, G. Armitage, A survey of techniques for Internet traffic classification using machine learning, IEEE Commun. Surveys Tutor. 10 (4) (2008) 56–76.

[2] Uman K Chaudhary, Ioannis Papapanagiotou, Flow classification using clustering and association rule mining.

[3] Jeffrey Erman, Martin Arlitt, Anirban Mahanti, "Traffic Classification Using Clustering Algorithms", University of Calgary, 2500 University Drive NW, Calgary, AB, Canada..

[4] Chengjie GU, Shunyi ZHANG, Xiaozhen XUE, "Internet Traffic Classification based on Fuzzy Kernel K-means Clustering", International Journal of Advancements in Computing Technology, Volume 3, Number 3, April 2011.

[5] Www. wikipedia.org

[6] J. Zhang, Y. Xiang, Y. Wang, W. Zhou, Y. Xiang, Y. Guan, Network traffic classification using correlation information, IEEE Trans. Parallel Distrib. Syst. (2012)1–15.

[7] A. Moore and D. Zuev, "Internet traffic classification using Bayesian analysis techniques," in ACM International Conference on Measurementand Modeling of Computer Systems (SIGMETRICS) 2005, Banff.

[8] Alberta, Canada, June 2005. T. Auld, A. W. Moore, and S. F. Gull, "Bayesian neural networks for Internet traffic classification," IEEE Trans. Neural Networks, no. 1, pp. 223–239, January 2007.

[9] Luigi Grimaudo, Marco Mellia, Elena Baralis and Ram Keralapura, SeLeCT: Self-Learning Classifier for Internet Traffic, IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, VOL. 11, and NO. 2, JUNE 2014.

[10] Bujlow tombaz, Tahir, Jenns Peddersen, A method for classification of network traffic based on C5.0 Machine Learning Algorithm, to appear in International Conference on Networking and Communications (ICNC 2012).

[11] J. Zhang ; School of Information Technology, Deakin University, Melbourne, Australia ; X. Chen ; Y. Xiang ; W. Zhou, "Robust Network Traffic Classification ," IEEE/ACM Transactions on Networking (Volume:23 , Issue: 4), pp. 1257–1270, August 2015.

[12] Ratish Agarwal,Survey of clustering algorithms for MANET, International Journal on Computer Science and Engineering Vol.1(2), 2009, 98-104.

[13] T. Nguyen and G. Armitage, "Training on multiple sub-flows to optimise the use of Machine Learning classifiers in real-world IP networks," in Proc. IEEE 31st Conference on Local Computer Networks, Tampa, Florida, USA, November 2006.

[14] N. Williams, S. Zander, and G. Armitage, "A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification," Special Interest Group on Data Communication (SIGCOMM) Computer Communication Review, vol. 36, no. 5, pp. 5–16, 2006

[15] Alice Este, F. Gringoli, L. Salgarelli, Support vector machines for tcp traffic classification, Computer Networks 53 (14) (2009) 2476–2490.

[16] Aceto. G, Dainotti. A, Donato. W, and Pescap`e. Portload. A, "Taking the best of two worlds in traffic classification", In INFOCOM IEEE Conference on Computer Communications Workshops, 2010, pages 1–5, 15 2010.

[17] Alberto Dainotti, Antonio Pescap´e, and K.C. Claffy, "Issues and future directions in traffic classification", Network, IEEE, 26(1), 35 –40, January-February 2012.

[18] Chengjie GU, Shunyi ZHANG, Xiaozhen XUE, "Internet Traffic Classification based on Fuzzy Kernel K-means Clustering", International Journal of Advancements in Computing Technology, Volume 3, Number 3, April 2011.

[19] Dashevskiy, M., Luo, Z. "Reliable probabilistic classification of internet traffic", Int. J. Inf. Acquis. 6(2), 133–146, 2009.

.

[20] Kim. H, "Internet traffic classification demystified: myths, caveats, and the best practices," in Proc. ACM CoNEXT Conf., 2008, pp. 1–12.

[21] Mikhail Dashevskiy, Zhiyuan Luo "Two methods for reliable classification of network traffic", Springer-Verlag 2012.