Open access Journal International Journal of Emerging Trends in Science and Technology

A Survey of Clustering Algorithms for Traffic Classification

Author
J.Nithisha
P.G Scholar, Jeppiaar Engineering College, Chennai
Email: *nithisha.j@gmail.com*

ABSTRACT: *Network traffic in the world wide is predicted to increase every year double the times. So, traffic occurs in the network. To manage a network traffic classification is necessary because of increasing number of users and Qos. Classification algorithm provides a major role in traffic classification (i.e. flow or packet classification .In this paper different classification algorithms used are discussed. Traffic classification algorithm divided into supervised and unsupervised algorithm. Unsupervised algorithm uses unlabelled data to process batches of flows. So it can identifies a new classes of traffic application. Supervised algorithm works well for known dataset (flow) .Here, different classification algorithms like k-means, Model based clustering, identity based clustering, and k medians are presented.*

INTRODUCTION

Clustering is the process of partitioning the data into groups of similar objects ((i.e. homogeneous group).Each partitioned group is called a cluster which consists of objects that are similar between themselves. Clustering analysis is very important step in network traffic classification. Network traffic classification is the essential task in networking. It is the process of analysing traffic flows and group them into different categories of network applications .Existing measurement algorithm like Deep Packet Inspection (DPI), port based method were popular until last decades. These approaches for network traffic classification had diminished nowadays. Since, DPI method was time consuming and insecure because it needs to look through the content and then passes the traffic flows into corresponding destination. In port based method it uses the port number to filter the traffic classes. Sometimes Port number may carry valuable information.so these method also not suitable because of security purpose. Network traffic classification is the process of grouping similar flows (application wise) according to some predefined criterion. Unsupervised machine learning algorithm used for network traffic classification.

The rest of this paper is organized as follows. Section II describes basic idea about flows and the classification process. In the Section III, different clustering algorithm have been discussed. Some of the work is reviewed in Section IV. Finally in the section V we present the conclusion of this work

RELATED WORK

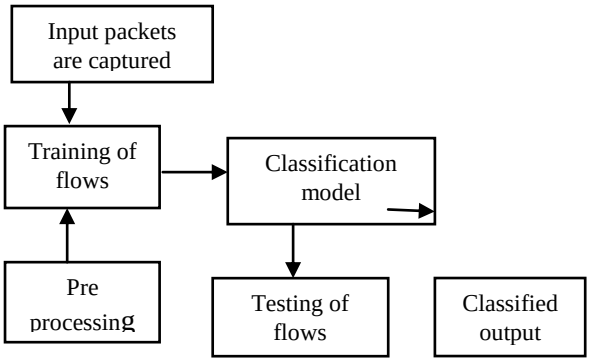
Flow Classification Model

Flow is a packet moved from one place to another within the period of time. Flows are represented by five tuple information.

Tuple→

src ip	dst ip	src port	dst port	tos

Flows are classified as unidirectional or bidirectional. In Unidirectional flows series of packets share the five tuple information. Bidirectional flows is a pair of flows going in opposite direction between same source and destination ip address. Flows have properties like packet length, packet inter arrival time, number of bytes transferred. The following system model illustrates how the flows are formed.



Classification Model

Classification Metrics

We use two metrics to evaluate the performance of classification algorithm. In order to evaluate the performance confusion matrix is used. It represented by rows and columns and row represents predicted class and column represents actual classes. It uses positive and negative values.

True Positive: Total percentage of flows classified as class A.

False Positive: Total percentage of flows not belonging to class A.

False Negative: Total percentage of flows incorrectly classified and not belongs to class A

True Negative: Total percentage of flows correctly classified as not belonging to class A.100-FP

The Classification metrics like overall accuracy, precision and recall are evaluated.

CLASSIFICATION ALGORITHMS

A) K-MEANS

In [1], discussed K-Means clustering is one of the well-known partition methods. K-means groups flows into subsets or clusters which are generated by same traffic application. The procedure follows a simple way to classify the traffic application based on k-cluster centres. (K centre-one for each cluster).For example, consider two flows that generated the application which have same packet length and inter arrival time. The centres are defined by different packet length or application. The closest application or which has nearby inter arrival time are considered as cluster.

The algorithm defined by following steps.

STEPS

X={x1, x2…xn} be a set of inter arrival time.Y= {y1, y2….yn} be the set of centers.

1. Initialize, k cluster centre (inter arrival time).
2. Calculate the Inter arrival time for each each flow and cluster centre.
3. Assign each flows into cluster whose inter arrival time is similar or minimum.
4. Recalculate 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. Recalculate the time between each flow and new obtained cluster center.
6. Stop when no of flows reassigned otherwise repeat from step 3

It is easy to understand and Gives best result when dataset are different. However, it is not suitable for different dataset.

B) MODEL BASED CLUSTERING

K-mean is inefficient because of estimation of number of clusters. In [4], they discussed model based clustering in their paper, it provided better accuracy than k-mean.Model based clustering with association rule mining techniques provide better accuracy than k-mean. Moreover, the rules are produced automatically and make the algorithm to work independent of dataset.

This algorithm works well for smaller applications but not suitable for larger application.

C) IDENTIFIER BASED CLUSTERING ALGORITHM

Each node in the network has unique ID. Nodes know the id of neighbour node. Cluster head is chosen by the following rules.

LOWER ID CLUSTER ALGORITHM

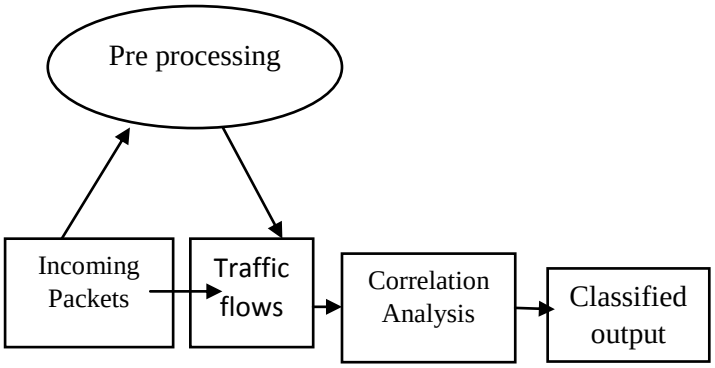
In [6] the algorithm suggests, the node with minimum ID is chosen as a cluster head. The id of the nearby node has higher node id than of cluster head. A node which lies within the range of two or more cluster head is called gateway node. Gateway is mainly used for routing between clusters. Each node is assigned with different ID.

Characteristics of node

1. A node which has higher value nearby node than itself is called cluster head.
2. A node which is near by two or more cluster head is called gateway.
3. Remaining nodes are ordinary node.

D) Traffic classification using Correlation

K-nearest neighbour is a non-parametric classifier and does not require training set. However, it doesn't provide better classifier performance when the



Training data set is small. In [3], the author proposed a non parametric approach that uses traffic classification using correlation information. It improves the classification performance for very few training data set.

In this, the incoming packets are collected and using this trace, traffic flows are formed. The correlation information is analysed and it is forwarded to the classifier. Finally the classifier classifies the flows based on application wise and produces output .The NN classifier produces only 60%overall accuracy when considering small flows. In [5] they used bag of flows to model correlation information. Further evaluated three classification methods like AVG_NN, MIN-NN, and MVT_NN for improving performance.

E) CLASSIFICATION USING BAYESIAN ANALYSIS TECHNIQUE

In [9] Moore and Zuev proposed 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, webservices,p2p, multimedia and games. The authors used accuracy as a classification metric to evaluate performance of classifier. This results depicted that naïve bayes techniques have 65 % accuracy in classification. To improve the classification accuracy, two refinements were performed using naïve bayes kernel estimation and fast correlation based filter method. It gives 95% as the overall accuracy. This work is extended by [10] using application of Bayesian approach in neural network. This results produces 99% accuracy.

F) AUTOCLASS

Autoclass algorithm is one of the unsupervised machine learning algorithm which used to produce best cluster from the given data set. This algorithm automatically selects the number of clusters and provides soft clustering on the data. The auto class algorithm pre-configured with the clusters so it is easy to find and determine the clusters of flow from the given traffic statistics. The algorithm uses EM (expectation maximization) algorithm [14].

This algorithm used to separate the traffic flows into either bidirectional or unidirectional flows. Then identify the classes based on application.

G) SELF LEARNING CLASSIFIER

Earlier methods like deep packet inspection and port based approaches were unfamiliar because of availability of training set. Unsupervised algorithm used as a viable alternative to classify the flows which are not trained. In [7] author proposed a new method called SeLeCT, a Self-Learning Classifier for Internet Traffic. It uses unsupervised algorithm to automatically groups the flow into homogeneous group based on application. It doesn't require prior knowledge of training set to identify the flows. 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.

H) C5.0 CLASSIFIER

In [13] author discussed, C5.0 is the decision tree based algorithm and use the concept of machine learning algorithm. It is easier to use and memory efficient. It generates the decision tree based on set of training set .The trees are used to classify the set of test cases. The c5.0 classifier uses command line interface to generate the rules for decision tree and test the classifier. The experiment was executed many times using different set of training and test cases. It produced 98% accuracy when accurate testing and training data used.

SUMMARY OF REVIWEDPAPER

Author	Classification Algorithm	Features	Data traces	Considered traffic
Jeffrey Erman [2]	k-mean	Inter arrival time	Auckland iv,Calgary	Irc,pop3,http ,Limewire,NNTP,Socks
Jun zhang [3]	NN	Packet size,bytes	Wide ,isp	Dns,http,imap,ftp,p2p
Umang chaudry [4]	Model based clustering	Packets,bytes	WITS,CRAWDAD	http,smtp,dns,mail
Ratih Agarwal [6]	Lower ID clustering	Transmission range,packets	Gps	Smtp,http
Moore and Zuev [9]	Bayesian Technique	Packet size, Inter arrival time, Flow duration	Proprietary Hand based traces	Mail,p2p,www
Ngugen and Armitage [11]	Supervised Naïve bayes	Inter arrival time, packet length	Traces collected from game server	Http,dns,smtp
Luigi Grimaudo [7]	Self learning classifier	Flow size,subset, Inter arrival time	Traffic traces from isp	http,gmail,Rtsp,BitTorrent, POP3,Telnet, eMule

CONCLUSION

In this study different clustering algorithm for traffic classification is reviewed. The basic concept of clustering and different clustering techniques are discussed. Clustering is a significant task in traffic classification. Clustering is the process of grouping objects (flows) into classes of similar objects. There are different types of clustering algorithm such as k-means, model based clustering identifier based algorithm are presented. In this paper, unsupervised algorithms for traffic classification are discussed, it uses unlabelled data to partition the flows.

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. Jun Zhang, Yang Xiang, Wanlei Zhou, Yu Wang, Unsupervised traffic classification using flow statistical properties and IP packet payload, Journal of Computer and System Sciences 79 (2013) 573–585.
3. 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.
4. Uman K Chaudhary , Ioannis Papapanagiotou ,Flow classification using clustering and association rule mining
5. Priti K.Doad and Mahip M.Bartere ,Survey on Clustering Algorithm & Diagnosing Unsupervised Anomalies for Network Security , International Journal of Current Engineering and Technology ISSN 2277 – 410.
6. Ratish Agarwal,Survey of clustering algorithms for MANET, International Journal on Computer Science and Engineering Vol.1(2), 2009, 98-104
7. 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, NO. 2, JUNE 2014
8. Arthur Callado, Carlos Kamienski," A Survey on Internet Traffic Identification and Classification" in IEEE 2011.
9. A. Moore and D. Zuev, "Internet traffic classification using Bayesian analysis techniques," in ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS) 2005, Banff,
10. 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.
11. 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.
12. Jun Zhang, Yang Xiang, Wanlei Zhou, Yong Xiang, and Yong Guan "An Effective Network Traffic Classification Method with Unknown Flow Detection "in IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, VOL. 10, NO. 2, JUNE 2013
13. 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).
14. A.P.Dempster,N.M paird, and D.B.Rubin.Maximum likelihood from icomplete data via the EM algorithm.