Traffic classification using K-means algorithm

Title: Machine Learning Project April 2015

Institute: BITS PILANI – Pilani campus

Area: Wide Area Networks

**Abstract** - Work in this paper involves dealing with traffic classification which is important in order to optimize performance of different kind of traffic and application. Here basically k-means algorithm of machine learning is used on the statistical features of a flow an then we see whether same kind of samples are clustered together or not.

*Keywords*- Traffic Classification, K-means algorithm, Features.

.

## Table of Contents

1. Introduction 4
2. Background 4
3. Problem 4
4. Solution 5
5. K-means Clustering 6
6. Data Set and Feature Extraction 7
7. Algorithm Implementation 8
8. Summary 9
9. Future Work 9
10. References 10

# Introduction

### What is Traffic flow?

In networks, traffic flow or network flow is a sequence of packets (data) from a source to a destination. RFC 2722 defines traffic flow as "an artificial logical equivalent to a call or connection.

### What is Traffic Classification?

Traffic Classification is a process with classifies network traffic based on various features of the flow of the data packets. Based on classification different identified classes can be treated differently in a network to improve Quality of service or reduce congestion in the network.

# Background

Traffic is generally classified based on the source of application, payload content and some flow statistical features. The traditional well known port based methods are ineffective in current scenario therefore the statistical features based methods have started replacing traditional methods. One of the case where traditional classification techniques are increasingly losing their effectiveness are in cases where explicit “anti-classifier” obfuscation techniques are employed by the user applications. Also packet inspection for traffic classification which is port based traffic classification, and payload based traffic classification have their own limitations.

# Problem

Earlier traffic classification was basically based on transport layer port numbers, registered with IANA representing a particular application but now there is a shift from particular port number assignment to dynamic port assignment because of increasing support for peer to peer applications that hide identity by assigning dynamic ports. Thus we need some other methods in order to classify data.

# Solution

**How to approach a solution?**

To solve this problem we will need to know different statistical and observable features about our data flow and also what features are relevant and to what extent. Since already mentioned that traditional approaches are becoming increasingly ineffective and there is shift to flow based approaches, we will use machine learning in order to see whether we have some success with it.

**Design**

One of the solution to this problem by doing a clustering on the unsupervised set in order to identify how many different kind of classes exits and what are their general trends on a given networks.

The Unsupervised learning approaches provide a learning technique which doesn’t require any pre-determined or defined knowledge instead they discover natural groups (clusters) in the data using the difference between the features of different samples. It focuses on finding patterns in the input data. It clusters instances with similar properties into groups.

Here we will be using K-mean clustering in order to identify clusters and infer their properties. Number of means will be based on general heuristics of a details of commonly used classes and test it. After this labelling the data set based on the features of the class and at the end testing the algorithm using the real class of each sample.

**K-means Clustering**

**What is K-Means Clustering?**

Simply speaking it is an algorithm to classify or to group your objects based on attributes (features) into K number of group. K has to be a positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Thus, the purpose of K-mean clustering is to classify the unsupervised data set based on the similarities between features.

**How the K-Mean Clustering algorithm works?**

Each centroid will have a cluster number. For each data, we calculate the distance to all centroid and get the minimum distance. This data is said belong to the cluster that has minimum distance from this data. Since we are not sure about the location of the centroid, we need to adjust the centroid location based on the current updated data. Then we assign all the data to this new centroid. This process is repeated until no data is moving to another cluster anymore or we can ourselves set the no of iteration. Mathematically this loop can be proved convergent.

**What are the applications of K-mean clustering?**

Applications of the K-mean clustering ranges from unsupervised learning of neural network, Pattern recognitions, Classification analysis, Artificial intelligent, image processing, machine vision, etc. It is used whenever you have several objects and each object have several attributes and you want to classify the objects based on the attributes, then you can apply this algorithm.

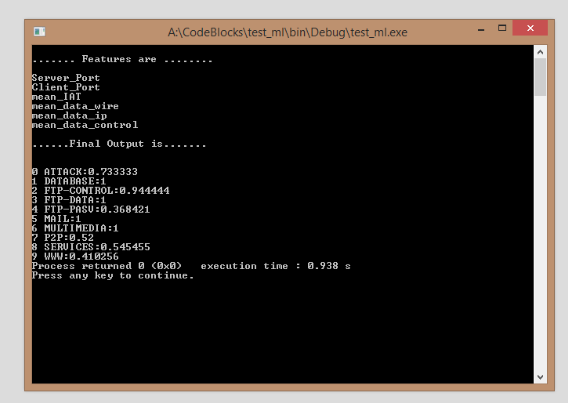
**What are the weaknesses of K-Mean Clustering?**

Similar to other algorithm, K-mean clustering has many weaknesses example, when the numbers of data are not so many, initial grouping will determine the cluster significantly. Number of cluster, K, must be determined beforehand. We may converge to local minimum. We never know which attribute contributes more to the grouping process since we assume that each attribute has the same weight. Thus it makes sense to use this algorithm when we have large data set.

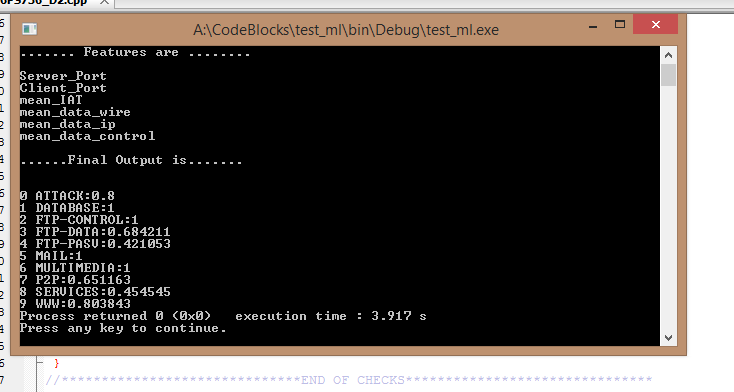
**Data Set and Feature Extraction**

Data Set chosen is benchmark data publically distributed for study and research purposes. A very small subset of features were chosen from total of 249 features which are easily observable in any flow. Server port, client port , mean inter-arrival time, mean of bytes in Ethernet packet , mean of total bytes in IP packet, mean of data control packets. We may do data pre-processing or

Use different features in order to optimize performance. The main aim of the experiment was to implement clustering algorithm and observer how it is clustering the data and using the features of flow ii cn order to optimize our classification.



Above output is for data set of 217 samples and the one below is for 1510 sample set.



**Algorithm Implementation**

Start with initial k centres randomly initialized

1. INITIALIZE CLUSTER FOR EACH sample to which they belong

By finding centre which is at minimum distance

1. SET CLUSTER POPULATION for each cluster
2. SET THE CENTER OR CLUSTERS formed using above steps.
3. SET THE entropy FOR EACH sample which is sum of square of difference between attributes of sample and its corresponding cluster centre.
4. BEGIN LOOP for a given no of iterations (maximum)
   1. FOR each sample find the new nearest cluster centre after the adjustment above in step 3.
   2. If there is a difference between old and new cluster centre then change the cluster number corresponding to the sample and adjust cluster population and find new centre properties corresponding to both old and new cluster centre
5. END of LOOP

We will have cluster no corresponding to each sample in the end.

**Summary**

As we have discussed above k-means is a very simple yet powerful algorithm, it has its own power and weakness. In order to fully utilize this algorithm a big data set is essential, feature extraction and selection plays a major role in this algorithm and we need to play with data and its features in order to explore its limits.

**Future Work**

There is a lot of scope in this area since there is a high demand for correct classification algorithm in network security and Quality of service thus we need to use the wide category of features and extract most relevant features for classification. Another line of thought can be to use most basic information available about a packet in a flow and only use information in a packet in order to classify instead of the knowledge of whole flow.

# References

M. Tamilkili et al; A Survey on Recent Traffic Classification Techniques Using Machine Learning Methods ; International Journal of Advanced Research in Computer Science and Software Engineering; India; Volume 3, Issue 12, December 2013;Pages:368 to 373

Indra Bhan Arya et al; Internet Traffic Classification: An Enhancement in Performance using Classifiers Combination; (IJCSIT) International Journal of Computer Science and Information Technologies; Bhopal ,India ; Vol. 2 (2) ; 2011; Pages:663 to 667

Andrew W. Moore et all; Discriminators for use in flow-based classification; Queen Mary, University of London, Department of Computer Science; August 2005; Pages:16

Jamuna .A et all; Efficient Flow based Network Traffic Classification using Machine Learning; International Journal of Engineering Research and Applications (IJERA); Vol. 3, Issue 2; March -April 2013; Pages: 1324 to 1328

Christopher m. Bishop; Pattern Recognition and Machine Learning; Springer;

Tom M. Mitchell; Machine Learning; Mc Graw Hill

Web sources:

http://en.wikipedia.org/