Network Traffic Classification based on Unsupervised Approach

Pallavi Singhal
L.M.College of Science and Technology
A-Sector,Shastri Nagar, Jodhpur

Rajeev Mathur, Ph.D
L.M. College of Science and Technology
A-Sector,Shastri Nagar, Jodhpur

Himani Vyas
L.M. College of Science and Technology
A-Sector,Shastri Nagar, Jodhpur

ABSTRACT

The IP network engineering, management and control are highly benefited by Network traffic classification and application identification. There are many popular methods available namely port-based and payload-based but they have shown some disadvantages, and the machine learning based method is a potential one. Unsupervised learning deals with a class of problems in which one seeks to determine how the data are organized. The difference between it and supervised learning in that the learner is given only unlabeled examples. Unsupervised learning is a way to form 'natural groupings' or clusters of patterns. Unsupervised learning is useful to the problem of density estimation in statistics. One form of unsupervised learning is clustering.

General Terms

Network traffic classification, clustering, machine learning, unsupervised

Keywords

Classification, clustering, machine learning, semi supervised, unsupervised , supervised

INTRODUCTION

Network traffic is composed of many applications, including Internet-like applications (Web, mail, online games, P2P...) and other proprietary ones. Faced by this diversity, Internet Service Providers (ISPs) are more and more interested in identifying the different applications sources of the traffic. The traditional “Port-based Classification method” [2] relies on linking a well-known port number with a specific application, so as to identify different Internet traffic. This traffic classification technique use well-known port numbers to identify Internet traffic. Port-based classification has been shown to be ineffective because many recently developed applications do not communicate on standardized ports. Other method is “Deep Packet Inspection" [2]. In this approach, the packet payloads are analyzed to see whether or not they contain characteristic signatures of known applications. This technique can be extremely accurate when the payload is not encrypted. Another promising approach to traffic classification is the use of machine learning. Machine learning techniques [2] can be divided into the categories of unsupervised and supervised. The Supervised approach requires the training data to be labelled before the model is built. The goal of those methods is how to improve accuracy of classification. Unsupervised techniques don't need hand labelled traces, they are just based on the inner similarity among all flows within a training set to group several clusters. A semi-supervised methodology that classifies (or equivalently, identifies) network flows by using only flow statistics is analyzed and implemented.

Network traffic is composed of many applications, including Internet-like applications (Web, mail, online games, P2P...) and other proprietary ones. Faced by this diversity, Internet Service Providers (ISPs) are more and more interested in identifying the different applications sources of the traffic. The traditional “Port-based Classification method” [2] relies on linking a well-known port number with a specific application, so as to identify different Internet traffic. This traffic classification technique using well-known port numbers to identify Internet traffic. Port-based classification has been shown to be ineffective because many recently developed applications do not communicate on standardized ports. Other method is “Deep Packet Inspection" [2]. In this approach, the packet payloads are analyzed to see whether or not they contain characteristic signatures of known applications. This technique can be extremely accurate when the payload is not encrypted. Another promising approach to traffic classification is the use of machine learning.

This methodology is based on machine learning principles, consists of two components: a learner and a classifier. The goal of the learner is to discern a mapping between flows and traffic class from a training data set. Subsequently, this learned mapping is used to obtain a classifier. Traditionally, learning is accomplished using a fully labelled training data set, as has been previously considered in the traffic classification context. Obtaining a large, representative, training data set that is fully labelled is difficult, time consuming, and expensive.

On the contrary, obtaining unlabeled training flows is inexpensive. They develop and evaluate a technique that enables us to build a traffic classifier using flow statistics from both labelled and unlabeled flows. Specifically, we build the learner using both labelled and unlabeled flows and show how unlabeled flows can be leveraged to make the traffic classification problem manageable. Labelled data means the input set for which the class to which it belong is known. Unlabeled dataset is one for which class to which it belongs is unknown and is to be properly classified. There are three main advantages of semi supervised approach. First, we can obtain fast and accurate classifiers by training with a small number of labelled flows mixed with a large number of unlabeled flows. Second, our approach is robust and can handle previously unseen flows. Furthermore, our approach allows iterative development of the classifier by allowing flexibility of adding unlabeled flows to enhance the classifier's performance.

NEED FOR NETWORK TRAFFIC CLASSIFICATION

Network traffic classification is extensively required mainly for many network management tasks such as flow prioritization, traffic shaping/policing, and diagnostic monitoring. Similar to network management tasks, many network engineering problems such as workload characterization and modelling, capacity planning, and route provisioning also benefit from accurate identification of network traffic. Many network operators are interested in tools to manage traffic, such that traffic critical to business or traffic with real time constraints is given higher priority service on their network. Critical for the success of any such tool is its ability to accurately, and in real time, identify and categorize network flow by the application responsible for the flow. This task of mapping flows to the network applications that generate the traffic is called traffic classification.

At present, the development of the TCP/IP (Transmission Control Protocol/Internet Protocol) technology means based on Internet towards to a depth direction, such as the deployment of new generation infrastructure, the development of new technology and the emergence of new application patterns and demands.

Compared with the rapid development of the Internet, there is little research of network behaviours. The number of application layer protocols and end-users is increasing rapidly. Because of this Deployment, the efficient management of network resources is a complicated task. [1] With traditional network management methods, it is difficult to obtain a comprehensive view from the state of the network and simultaneously discover important details from the network traffics. Traffic classification mechanisms are useful tools that help the allocation, control and management of resources in TCP/IP networks. Internet not only has the volatile, heterogeneity, dynamic, but also the strong society. The growing success of Internet gives great facilities regarding acquisition of information. The huge quantity of data available on the network corresponds to a large variety of themes and forms (Text, images, video). This information can be created, stored, obtained or modified by all kinds of people worldwide this user behaviour has an important effect on the Internet. So it is an interesting direction to understand such a system's statistical and dynamic property, and Internet users' behavioural character. In addition, research of Internet and users' behaviour is an important step of many network management tasks.

Accurate identification and categorization of network traffic according to application type is an important element of many network management tasks such as flow prioritization, traffic shaping/ policing, and diagnostic monitoring. For example, a net-work operator may want to identify and throttle (or block) traffic from peer-to-peer (P2P) file sharing applications to manage its bandwidth budget and to ensure good performance of business critical applications. Similar to network management tasks, many network engineering problems such as workload characterization and modelling, capacity planning, and route provisioning also benefit from accurate identification of network traffic.
The classical approach to traffic classification relies on mapping applications to well-known port numbers and has been very successful in the past. To avoid detection by this method, P2P applications began using dynamic port numbers, and also started disguising themselves by using port numbers for commonly used protocols such as HTTP and FTP. Many recent studies confirm that port-based identification of network traffic is ineffective.

Network traffic analysis is done in order to resolve the above problems. Almost all activities related to the network are linked to traffic. Network traffic is an important carrier to record and reflect the Internet and user activities; it is also an important composition of network behaviour. Through the analysis of network traffic statistics, we can master the network statistical behaviour indirectly. With the variety of applications emerging, besides the traditional applications (e.g. http, email, web, and ftp), new applications such as P2P (Peer to Peer) have gained strong momentum. So it will be an interesting work to classify traffic and identify applications. A number of areas, such as trend analysis means refer to the concept of collecting information and attempting to spot a pattern, or trend, in the information and dynamic access control is the process that evaluates resource access, can benefit from it. At the same time, accurate classification of Internet traffic define it as the density of data present in the Internet is an important basis of network security and traffic engineering is the process of managing the allocation of network resources to carry traffic subject to constraints. Traffic statistics of different application types, reflecting user behaviour while using the network, so it can be useful to help network administrators to control traffic such that traffic critical to business is given higher priority service on their network.

LITERATURE SURVEY

At present, the main types of network application include HTTP, P2P, SMTP, POP3, Telnet, DNS, and FTP, etc. discusses the level of traffic analysis, and demonstrates which levels we are concerned about. Meanwhile, several techniques presented in the literature are surveyed.

The following different techniques used for network traffic classification.
Port Number Mapping.
2. Payload-based Analysis.
3. Machine Learning.

Current research of network traffic analysis mainly focuses on the bit-level, packet level, flow-level and stream-level. Bit-Level's research mostly concern network traffic's quantitative characteristics, such as network cable transmission rate and throughput's changes. Packet-Level cares the arrival procedure of the IP packet, delay and packet loss rate [3]. Packet network engineering team to protect networks against network element failure and support the rapid growth of traffic volume. So far, this approach has been successful in maintaining simple, scalable, highly available, and robust networks. It is important to realize that in packet networks which do not perform call admission control, there is often no way to control the amount or types of traffic entering the network.

The provisioning problem therefore lies in figuring out how much excess capacity is required to provide robustness (e.g. resilience to multiple simultaneous link failures) and scalability. The current tools for network management, such as SNMP (Simple Network Management Protocol), are limited in their capabilities, since they only provide highly.

Aggregated statistics about the traffic (e.g. average traffic load over five minute intervals) and do not give insight into traffic dynamics on time scales appropriate for events such as packet drops. The basis of flow partition is the address and protocol. For example, in [4] defines flow as series of packet exchanges

between two hosts, identified by a 5-tuple (source IP address, source port, destination IP address, destination port, application protocol).

These four layers mainly think of arrival procedure of flow, inter-arrival time etc local characteristics. [5] Stream level data is certainly not as precise as passive measurements of flow or packet level data; we demonstrate that it is sufficient for exposing some different types of unusual traffic across a whole network.

It also has the benefit of generating much smaller data sets than other two level measurements, which tends to a significant issue in large, heavily used networks. The stream only records the host pair and application traffic from source to destination. Any implementation in a router would not be needed by this methodology and removing the complex processing makes low requirement to passive monitoring mechanism. Stream level as a define 3-tuple (source IP address, destination address, application protocol). The goal is to focus on statistical characteristics of the long-term flow about backbone network. The granularity of traffic within the above four layers increases from small to large, and the time scale of concern increases gradually. At different time scale, network traffic performs different behaviour regularly. In this project the level of concern is flow. Then goal is to classify different flows and specify their application types.

Machine Learning-Based Approaches

Machine Learning is an important research direction of modern artificial intelligence. The ability of continually gaining new knowledge or skills, re-organizing knowledge structure to improve their performance, has let it become a widely used method in network traffic classification. The machine learning procedure can be divided into two steps: classification model building and then classification.

Machine learning techniques [2] can be divided into the categories of supervised, unsupervised and Semi-supervised. The supervised approach requires the training data to be labelled before the model is built. The goal of those methods is how to improve accuracy of classification. Unsupervised techniques don't need hand-labelled traces, they are just based on the inner similarity among all flows within a training set to group several clusters. A semi-supervised methodology that classifies (or equivalently, identifies) network flows by using only flow statistics is analyzed and implemented.
This methodology is based on machine learning principles, consists of two components: a learner and a classifier. The goal of the learner is to discern a mapping between flows and traffic class from a training data set. Subsequently, this learned mapping is used to obtain a classifier.

Machine learning is a scientific field that is concerned with the design and development of algorithms that allow computers to evolve behaviours based on empirical data that means term data refers to groups of information that represent the qualitative or quantitative attributes of a variable or set of variables., such as from sensor data or databases. Machine learning is the field of research devoted to the formal study of learning systems. This is a highly interdisciplinary field which borrows and builds upon ideas from statistics, computer science, engineering, cognitive science, optimization theory and many other disciplines of science and mathematics.
A learner can take use of examples to capture characteristics of interest of their unknown underlying probability distribution. Data can be seen examples illustrating relations between observed variables.
A important concentration of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data; the difficulty lies in the fact that the set of all possible behaviours given all possible inputs is too large to be covered by the set of observed examples (training data).

Hence the learner must generalize from the given examples and produce a useful output in new cases. Artificial intelligence, probability theory and statistics, data mining, pattern recognition, adaptive control, computational neuroscience and theoretical computer science are closely related fields..

Machine learning algorithms are categorized in to different type.
1. Supervised learning
2. Un-Supervised learning
3. Semi-Supervised learning

Un-Supervised learning

Unsupervised learning is a class of problems in which one seeks to determine the organization of data. The difference between it and supervised learning in that the learner is given only unlabeled examples. Unsupervised learning is a way to form 'natural groupings' or clusters of patterns. Unsupervised learning is useful to the problem of density estimation in statistics. One form of unsupervised learning is clustering.

Techniques used for un-supervised learning

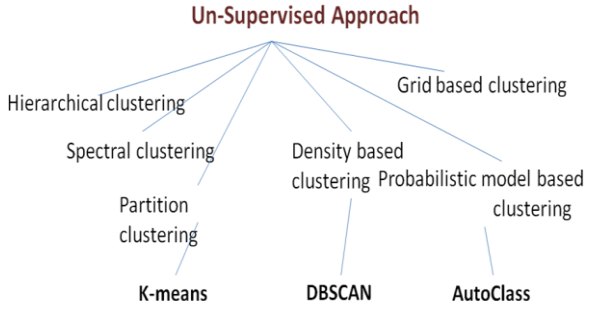

Fig 1: Techniques for Un-Supervised Approach

Fig. shows so many techniques used for unsupervised learning, few of them is presented here in Table 1.

Table 1 : Comparison of Un-Supervised Approach

K-means	DBSCAN	AutoClass
Partition based algo.	Density based algo.	Probabilistic model based algo.
Fastest at clustering	Much faster at clustering	Slow at clustering
Took least time to build classification model	Took largest time to build classification model	Took largest time to build classification model i.e. time consuming
No. of clusters need to specify	Automatically not determine no. of clusters	Automatically determine no. of clusters
Every object may be assigned to cluster i.e few may be misclassified	Every object that is not part of cluster is categorized as noise.	Every object is assigned to cluster
Produces clusters that are spherical in shape.	Produces clusters that are non-spherical in shape.	Clusters are of any shape.
Overall accuracy is good.	Overall accuracy is lower.	Accuracy is good.

Advantages of using k-mean clustering

The advantages of using k-mean clustering is that with a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small). For globular clusters K-Means may produce tighter clusters than hierarchical clustering.

Implementation of k – mean clustering

We have implemented the above for 3 cases taking first balance_training.xls dataset consisting of 492 records, 4 features and 6 clusters. The second case had glass_training.xls dataset consisting of 158 records , 9 features and 4 clusters. The final case had iris data_train30.xls consisting of 120 records, 4 features and 4 clusters.

CASE 1 :

Data set = balance_training.xls, No.of records = 492, Features = 4, Clusters = 6

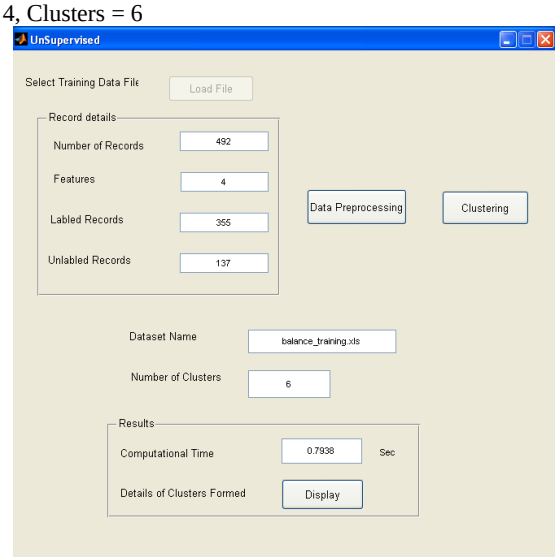

Fig. 2 : Window showing inputs and computational time for dataset 1

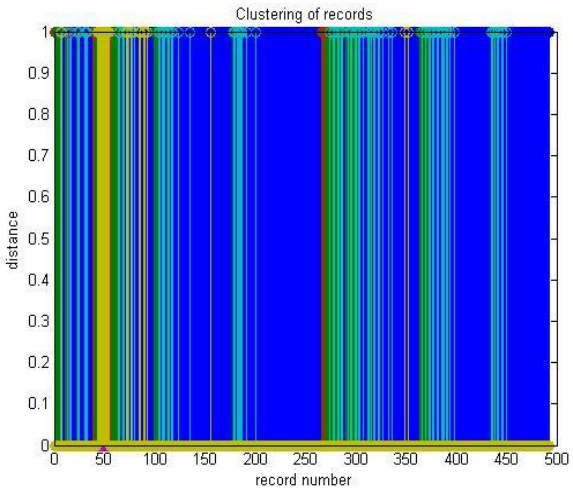

Fig. 3 : Figure showing clustering of records for dataset 1

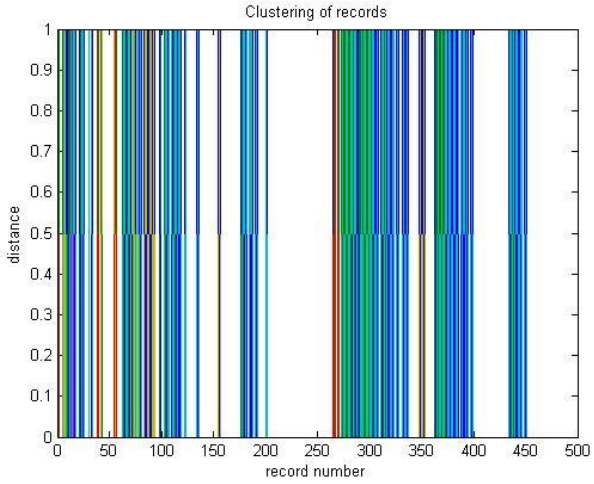

Fig. 4 : Figure showing clustering of records for dataset 1

CASE 2 :

Data set = glass_training.xls
No.of records = 158, Features = 9, Clusters = 4

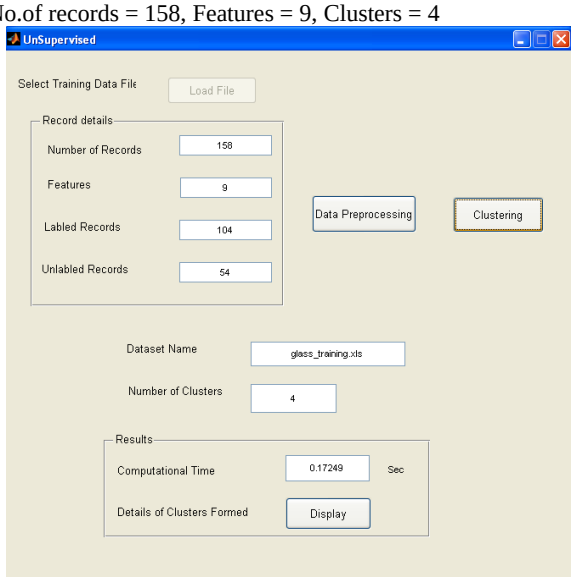

Fig. 5 : Window showing inputs and computational time for dataset 2

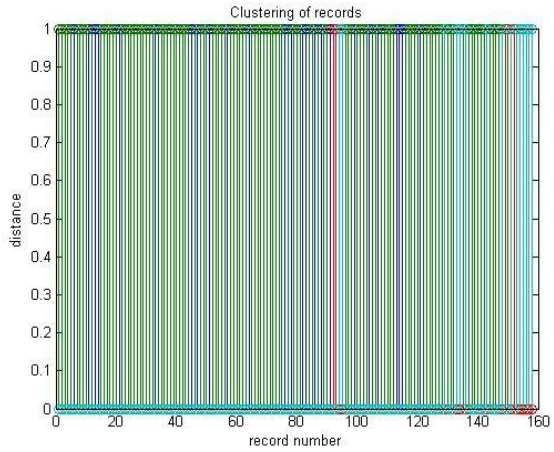

Fig. 6 : Figure showing clustering of records for dataset 2

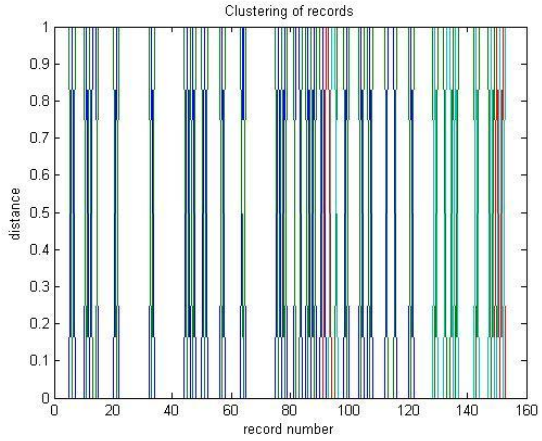

Fig. 7 : Figure showing clustering of records for dataset 2

CASE 3 :

Data set = iris data_train30.xls
No.of records = 120
Features = 4
Clusters = 4

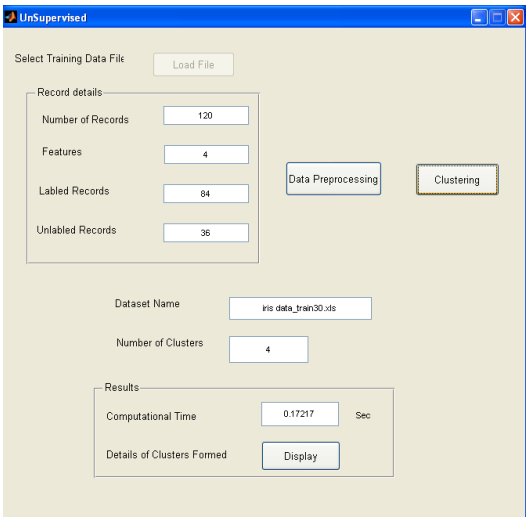

Fig. 8 : Window showing inputs and computational time for dataset 3

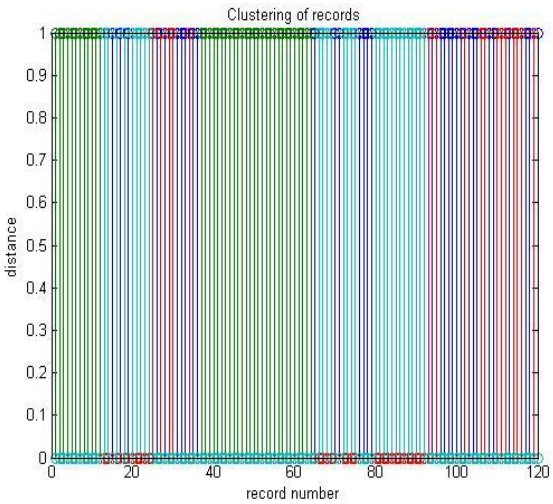

Fig. 9 : Figure showing clustering of records for dataset 3

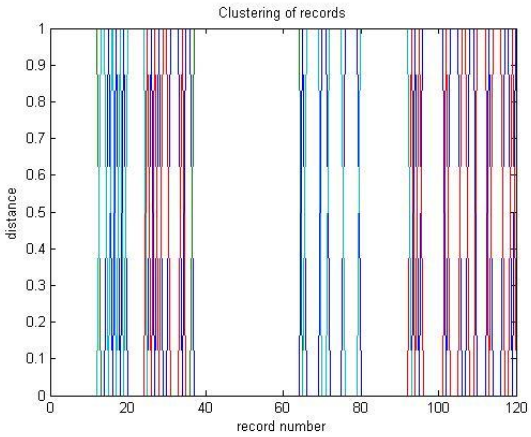

Fig. 10: Figure showing clustering of records for dataset 3

TABLE 2 : Table showing Cluster wise distribution for individual record

Record No	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	0	1	0	0
2	0	2	0	0
3	0	3	0	0
4	0	4	0	0
5	0	5	0	0
6	0	6	0	0
7	0	7	0	0
8	0	8	0	0
9	0	9	0	0
10	0	10	0	0
11	0	11	0	0
12	0	0	0	13
13	0	0	0	14
14	0	0	0	0
15	15	0	0	16
16	0	0	0	0
17	17	0	0	18
18	0	0	0	0
19	19	0	0	20
20	0	0	0	21
21	0	0	0	22
22	0	0	0	23
23	0	0	0	24
24	0	0	0	0
25	0	0	25	0
26	26	0	0	0
27	0	0	27	0
28	28	0	0	0
29	0	0	29	0
30	0	0	30	0
31	31	0	0	0
32	32	0	0	0
33	33	0	0	0
34	0	0	34	0
35	35	0	0	0
36	36	0	0	0
37	0	37	0	0
38	0	38	0	0
39	0	39	0	0
40	0	40	0	0
41	0	41	0	0
42	0	42	0	0
43	0	43	0	0
44	0	44	0	0
45	0	45	0	0
46	0	46	0	0
47	0	47	0	0
48	0	48	0	0
49	0	49	0	0
50	0	50	0	0
51	0	51	0	0
52	0	52	0	0
53	0	53	0	0
54	0	54	0	0
55	0	55	0	0
56	0	56	0	0
57	0	57	0	0
58	0	58	0	0
59	0	59	0	0
60	0	60	0	0
61	0	61	0	0
62	0	62	0	0
63	0	63	0	0
64	0	64	0	0
65	65	0	0	66
66	0	0	0	67
67	0	0	0	68
68	0	0	0	69
69	0	0	0	0
70	70	0	0	0
71	71	0	0	0
72	0	0	0	72
73	0	0	0	73
74	0	0	0	74
75	0	0	0	75
76	76	0	0	0
77	77	0	0	0
78	78	0	0	0
79	79	0	0	80
80	0	0	0	81
81	0	0	0	82
82	0	0	0	83
83	0	0	0	84
84	0	0	0	85
85	0	0	0	86
86	0	0	0	87
87	0	0	0	88
88	0	0	0	89
89	0	0	0	90
90	0	0	0	91
91	0	0	0	92
92	0	0	0	93
93	0	0	93	0
94	94	0	0	0
95	0	0	95	0
96	96	0	0	0
97	97	0	0	0
98	98	0	0	0
99	99	0	0	0
100	100	0	0	0
101	101	0	0	0
102	0	0	102	0
103	103	0	0	0
104	104	0	0	0
105	105	0	0	0
106	0	0	106	0
107	0	0	107	0
108	108	0	0	0
109	109	0	0	0
110	0	0	110	0
111	0	0	111	0
112	0	0	112	0
113	113	0	0	0
114	0	0	114	0
115	0	0	115	0
116	0	0	116	0
117	117	0	0	0
118	118	0	0	0

119	0	0	119	0
120	120	0	0	0

Conclusion

So with the above tables and figures we can make out that the entire set of records are distributed in the clusters. We have provided a complete table for the iris data_train30.xls dataset with 120 records. The records are divided into 4 clusters which can be seen as the 4 colours in the figure and the table provides us with classification of each record in one of the four clusters.
Similarly for set 1 and set2 we have provided the classification of the records in various clusters.

References

[1] The Monitoring System Based on Traffic Classification Ali Asghar Yarifard and Mohammad Hossein Yaghmaee Islamic Azad University, Qaenat Branch, Iran Department of Computer Engineering, and Ferdowsi University.

[2] L. Yingqiu, Li Wei, L. Yunchun, "Network Tra_c Classi_cation Using K- means Clustering , School of Computer Science and Engineering, Beihang University,Beijing 100083, China, 2007 IEEE.

[3] [C. Barakat, P. Thiran, G. Iannaccone, C. Diot, P. Owezarski, "Modeling Internet backbone traffic at the flow level", IEEE Trans. on Signal Processing Special Issue on Networking, 2003.

[4] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, et al. "Packet-level traffic measurements form the sprint IP backbone", IEEE Trans. On Networks, 2003, 17(6): 6-16.

[5] T. He, h. Zhang, z. Li, "A methodology for analyzing backbone network traffic at stream-level" , Communication Technology Proceedings, 1, 2003, PP. 98-102

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

[7] P. Ha_ner, S. Sen, O. Spatscheck, D. Wang, "ACAS: Automated Construction of Application Signatures, "SIGCOMM'05 MineNet Workshop, New York: ACM Press, 2005, PP. 197-202.