Packet Identification By Using Data Mining Techniques

Safaa O. Al-Mamory
College of Information Technology, Babylon University, Iraq
safaa_vb@yahoo.com
Ali Hussein Ali
Directorate-General for Education in Qadisiyah District
ali.hussein8783@yahoo.com

Abstract

Accurate internet traffic identification and classification are fundamental to numerous network activities, including network management and security monitoring, traffic modeling and network planning, accounting and Quality of Service provision. With the development of network, P2P as new generation of network technology is widely used. Starting from the first popular one (Napster), a number of new P2P based multimedia file sharing systems have been developed (FastTrack, eDonkey, Gnutella, Direct Connect, etc.). A fundamental types of networks architectures in today's world are Client/ server and Peer to Peer. A promising approach that has recently received some attention is traffic classification using *machine learning* techniques. The term data mining is used for methods and algorithms that allow analyzing data in order to find rules and patterns describing the characteristic properties of the data. The aim of this research is to classify traffic accuracy which can be achieved by using *machine learning* techniques such as K-Means and Birch algorithms. This system depends on the extracted attributes and then use it in the proposed system to distinguish all types of packets. The goal of system of packet identification is to detect the types of packets and identification of application usage and trends , also identification of emerging applications diagnosing anomalies is critical for both network operators and end user in term of data security and service availability.

Keywords: Data Mining, Machine Learning , Clustering Algorithms (K-Means, Birch)

الخلاصة

التصنيف والتعريف الدقيق لحركة المرور على الانترنيت هو أمر أساسي للعديد من أنشطة الشبكة التي تتضمن إدارة الشبكات ,تخطيط الشبكات, المراقبة الأمنية، ونوعية الخدمة المقدمة . مع تطور الشبكات نشأ لدينا جيل جديد وهو تقنية نظير إلى نظير واستخدمت على نطاق واسع . بدأ من خلال تطوير (Napster)والذي يعتبر الأكثر شيوعاً من بين الأنظمة التي تستخدم أنظمة مشاركة الملفات مثل (FastTrack, eDonkey, Gnutella, Direct Connect, etc .)

الأنواع الأساسية لمعماريات الشبكات في الوقت الحاضر هي العميل / الخادم و النظير لنظير . الطريقة الواعدة التي نالت بعض الاهتمام هي تصنيف الحزم باستخدام تقنيات التعلم الآلي . أن مفهوم تنقيب البيانات يعبرعن الطرق والخوارزميات التي تسمح بتحليل البيانات لغرض إيجاد القواعد والأنماط لوصف الخصائص المميزة للبيانات .

الهدف من نظام تعريف الحزم هو لتحديد أنواع الحزم ,تحديد استعمالها واتجاهات التطبيق . كما يمكن التعرف على التطبيقات الناشئة لان تشخيص التشوهات أمر بالغ الأهمية لكل من المشغل والمستخدم للشبكة من ناحية أمن البيانات وتوفر الخدمة .

المفردات الرئيسية: تنقيب البيانات , التعلم الآلي , خوارزميات التجميع (Birch , K–Means) .

1. Introduction

Identifying application is essential for effective network planning and monitoring the trends of applications. The network traffic classification becomes more challenging because modern applications complicated their network behaviors. The objective of traffic classification is to understand the type of traffic carried on the Internet to protect the network resources [Jenefa *et al*, 2013].

With the development of network, P2P as new generation of network technology is widely used. Starting from the first popular one (Napster), a number of new P2P based

multimedia file sharing systems have been developed (FastTrack, eDonkey, Gnutella, Direct Connect, etc.). A fundamental types of networks architectures in today's world are Client/ server and Peer to Peer.

The term Peer-to-Peer (P2P) is used to refer to distributed systems without any central control, where all the nodes (called peers) are equivalent in functionality. In a P2P system, peers can collaborate and communicate with each other without utilizing expensive and difficult to maintain central infrastructure [Milojicic, 2002].

Data mining is the extraction of hidden predictive information from large databases and also a powerful new technology with great potential to analyze important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Every user need to collect and use the tremendous amounts of information is growing in a very large manner. Initially, with the advent of computers and means for mass digital storage, users has started collecting and storing all sorts of data, counting on the power of computers to help sort through this combination of information. Unfortunately, these massive collections of data stored on disparate structures very rapidly became overwhelming. This initial confusion has led to the creation of structured databases and database management systems. [Vijayarani *et al.*, 2011].

Machine learning techniques generally consists of two parts: model building and then classification. A model is first built using training data. This model is then inputted into a classifier that then classifies a data set. Machine learning techniques can be divided into the categories of supervised and unsupervised[McGregor *et al.*, 2004].

A number of methods have been proposed to identify and to classify the traffic into applications. In this paper, we use the Machine Learning Techniques (k-Means , Birch) to perform Packets Identification Process.

Section 2 describes the related works. In section3 proposed system is explained . The experimental results are discussed in section 4. Conclusions are given in section 5.

2. Related Works

This section surveys some of the recent and most related works to the proposed system.

Andrew *et al.* [Andrew *et al.*, 2005], proposes approach which uses supervised Machine-Learning to classify network traffic. Uniquely, we use data that have been hand-classified (based upon flow content) to one of the number of categories. Sets of data consisting of the (hand- assigned) category combined with description of the classified flows (e.g., flow length, port numbers, time between consecutive flows) are used to train the classifier. We show that in its most basic form a Naïve Bayes classifier is able to provide 65% accuracy for data in the same period and can achieve over 95% accuracy when combined with a number of simple refinements.

Sebastian *et al.*[Sebastian *et al.*, 2005], proposes a novel method for ML-based flow classification and application identification based on statistical flow properties. We use a feature selection technique for finding the optimal set of flow attributes and evaluated the effectiveness of our approach. Quantify the influence of different attributes on the learning. The results show that some separation of the applications can be achieved depending on the particular application. The average accuracy across all traces is 86.5%.

While some applications seem to have more characteristic attributes and can be well separated others intermingle and are harder to identify.

Jeffrey *et al.* [Jeffrey *et al.,* 2006], proposes approach to evaluate three different clustering algorithms, namely K-Means, DBSCAN, and AutoClass, for the network traffic classification problem. Our analysis is based on each algorithm's ability to produce clusters that have a high predictive power of a single traffic class, and each algorithm's ability to generate a minimal number of clusters that contain the majority of the connections. This is very useful because these clusters have a high predictive power of a single category of traffic. K-Means algorithm is more suitable for this problem due to its much faster model building time.

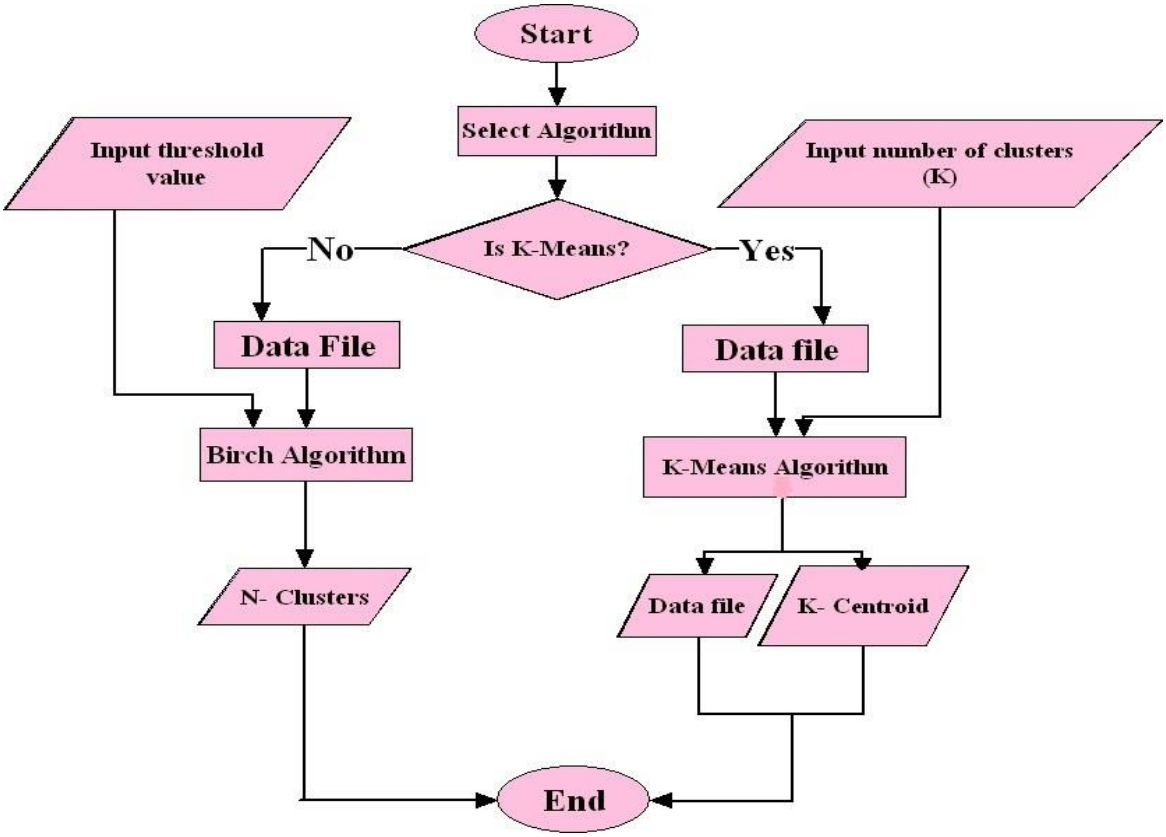
Bernaille. *et al.* [Bernaille. *et al.,* 2006], proposes approach which uses a simple K-Means clustering algorithm to perform classification using only first five packets of the flow , aiming at applying on the real time classification. Conduct the experiments to test the effectiveness of using the statistics of the first 5 packets to identify p2p traffic by decision tree algorithms. The results show that it is sufficient to use first 5 packets of flow to identify p2p application and could achieve high accuracy with low cost in distinguishing p2p traffic at the beginning of the TCP flow establishment. Another key advantage is the ability to accommodate both known and unknown p2p flows during development of the classifier and identify encrypted p2p traffic as well.

Gerhard *et al.* [Gerhard *et al.,* 2007], proposes a novel Network Data Mining approach that applies the K-Means clustering algorithm to feature datasets extracted from flow record. Training data are divided into clusters of time intervals of normal and anomalous traffic. While the data mining process is relatively complex, the resulting cluster centroids can be used to detect anomalies in new on-line monitoring data with a small number of distance calculations. This allow deploying the detection method for scalable real-time detection, e.g. as part of intrusion detection system. Applying the clustering algorithm separately for different services (identified by their transport protocol and port number) improve the detection quality.

Liu *et al.* [Liu *et al.,* 2007], proposes approach which describes the use of supervised and unsupervised Machine-Learning to classify network traffic by application. Compare overall accuracy before and after log transformation of data, and it proved that the accuracy can be improved at least 10% after log transformation. We show that the K-Means perform well at traffic classification, with an accuracy of 90%. Thus, new applications can be identified by grouping a separate cluster.

Bin Liu [Bin Liu., 2011], This paper proposes and evaluates a semi-supervised clustering for classifying P2P traffic using three P2P traffic metrics are *IP Address Discreteness, Success Rate of Connections, and Bidirectional Connections Rate*. Comparing with supervised machine learning methods, the proposed semi-supervised approach has two main advantages: first, high precision can be obtained by training with a small number of labeled samples mixed with a large number of unlabeled samples. Adding unlabeled samples can enhance the classifier's performance, second, this method can handle both seen and unseen applications. Using P2P traffic metrics, this approach identifies P2P traffic at host level now.

Ilya *et al.*. [Ilya *et al.,* 2012], proposes a clustering technique based on the Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm and Latent Semantic Analysis LSA-methods for clustering these large, high dimensional datasets in Russian and English languages. First step in dimension reduction is removing "noisy"



Figure(1):Training Phase.

parts of speech from the document model. The next step is selecting the most informative terms in the model. The last step is Latent Semantic Analysis (LSA) of the term matrix. LSA assumes that words that are close in meaning will be close to each other in text and use singular value decomposition (SVD) to reduce the number of terms while preserving the similarity of structure among documents.

Ghousia *et al.* [Ghousia *et al.*, 2013], proposes a method for effective clustering by selecting initial centroids. Firstly, this algorithm evaluates the distance between data points according to criteria; then tries to find out nearest data points which are similar; then finally selects actual centroids and formulates better clusters. According to the results of new solution, the improved k-means clustering algorithm provides more accuracy and effectiveness rather than previous one.

3. Problem Formulation And Methodology

The proposed system is designed to identify packets by using clustering algorithms to achieve the system functionality. The purpose of this stage is to obtain groups or clusters which should exhibit high intra-cluster similarity and high inter-cluster dissimilarity. The structure of the proposed system consists two phases. They are training and testing. Figure (1) illustrates training phase.

In the first phase, this model can be derived or learned from training data using clustering algorithms. Dataset is divided into two parts; training phase and testing phase. At the beginning of training phase, selecting algorithm K-Means or Birch. K-Means algorithm, begins with the assigned value K, and partitions the dataset into K disjoint clusters. The output of training phase are two files; first one contains input data and new column which represent predicted output of the algorithm. The second file contains selected centroid based on the input K value which can be used in the testing phase. A flow chart of K-Means algorithm is summarized in Figure (2).

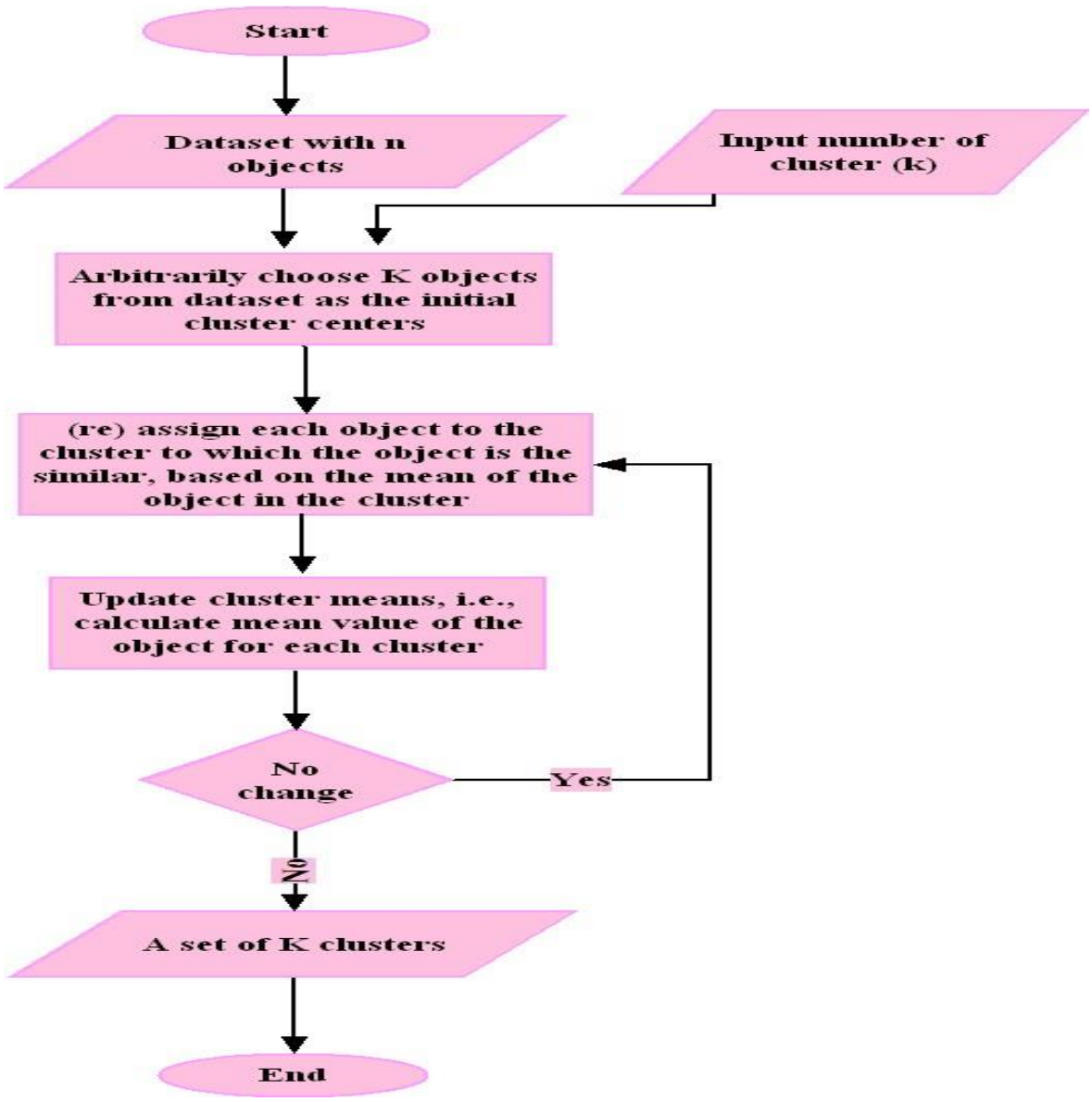


Figure (2): Main Steps of K-Means Algorithm.

Training phase with birch algorithm also start with number of parameters [Tian *et al.,* 1996]*:*

- Memory (*M*): 5% of data set
- Disk space (*R*): 20% of *M*
- Distance equation: *D*2
- Quality equation: weighted average diameter (*D*)
- Initial threshold (*T*): 0.0
- Page size (*P*): 1024 bytes

Given N_1 d-dimensional data points in a cluster: $\{\vec{x}_i\}$ where i = 1,2,…N_1, and N_2 data points in another cluster: $\{\vec{x}_j\}$ where j = N_1+1,N_1+2,…,N_1+N_2, the average inter-cluster distance D_2, average intra-cluster distance D_3 and variance increase distance D_4 of the two clusters are defined as [Tian *et al.*,1996]:

$$D2 = \left(\frac{\sum_{i=1}^{N_1}\sum_{j=N_1+1}^{N_1+N_2}(\vec{X}_i - \vec{X}_j)^2}{N_1 N_2}\right)^{\frac{1}{2}} \quad \text{.....(1)}$$

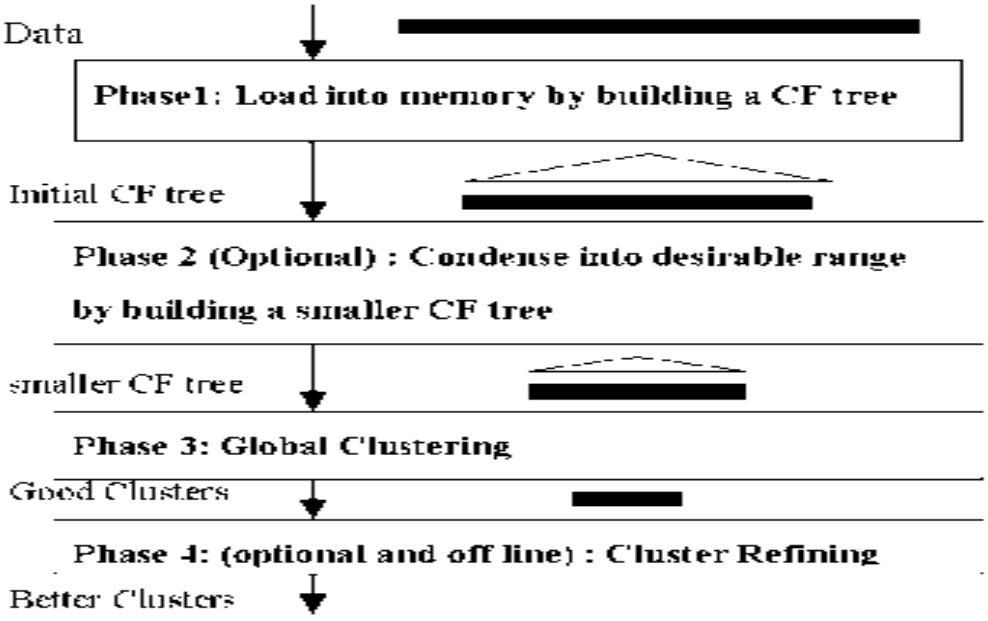
$$D3 = \left(\frac{\sum_{i=1}^{N_1+N_2}\sum_{j=1}^{N_1+N_2}(\vec{X}_i - \vec{X}_j)^2}{(N_1+N_2)(N_1+N_2-1)}\right)^{\frac{1}{2}} \quad \text{.....(2)}$$

$$D4 = \sum_{k=1}^{N_1+N_2}\left(\vec{X}_k - \frac{\sum_{l=1}^{N_1+N_2}\vec{X}_l}{N_1+N_2}\right)^2 - \sum_{i=1}^{N_1}\left(\vec{X}_i - \frac{\sum_{l=1}^{N_1}\vec{X}_l}{N_1}\right)^2 - \sum_{j=N_1+1}^{N_1+N_2}\left(X_j - \frac{\sum_{l=N_1+1}^{N_1+N_2}\vec{X}_l}{N_2}\right)^2 \quad \text{.....(3)}$$

The main steps in Birch algorithms can be summarized as follow : [Tian *et al.*,2009].

- Phase 1: Scan dataset once, build a CF tree in memory.
- Phase 2: (Optional) Condense the CF tree to a smaller CF tree.
- Phase 3: Global Clustering.
- Phase 4: (Optional) Clustering Refining (require scan of dataset).

Figure(3) shows the phases of the birch algorithm as below :



Figure(3): Shows The Phases Of The BIRCH Algorithm [Tian *et al.*, 2009**].**

In the second phase, based on this obtained knowledge, testing phase for both algorithms (K-Means, Birch) is starting. The results of birch algorithm in training phase are N-clusters. In order to be performed by the testing operation, the results must first

calculate the value of centroid for each N-clusters, and then use this value to complete the testing phase. The value of centroid in this case can be found by calculate the mean value for each clusters. Figure (4) summarised the testing operation for proposed system.

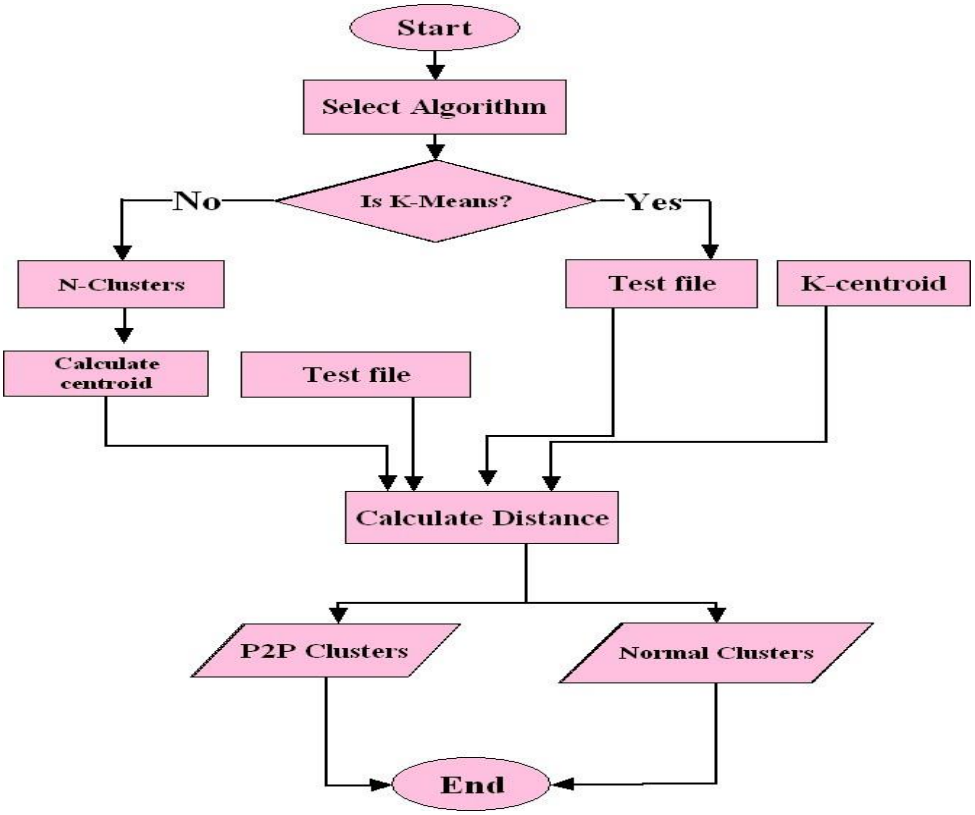


Figure (4):Testing Phase.

Testing phase for both algorithms is similar after calculating centroid value for birch algorithm. We apply Euclidian distance equation (4) for each packet in test file with center and then compare between two distances and select the smaller from them.

$$d(x,y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \qquad \text{.....(4)}$$

where $x = (x_1, \ldots, x_m)$ and $y = (y_1, \ldots, y_m)$ are two input vectors with m quantitative features. In the Euclidean distance equation, all features contribute equally to the function value.

4. Experiments And Results

To evaluate the different results, there are standard metrics that have been developed for evaluating packets identification process. Accuracy, specificity, sensitivity, precision are the four most famous metrics that have already been used [J.Han.,2006]. Accuracy metric can be calculated by the ratio between the true detection (positive , negative) to the total detection (true, false) [J.Han.,2006].

$$\text{Accuracy} = \text{True}_{(pos+neg)} / (\text{True}_{(pos+neg)} + \text{False}_{(pos+neg)}) \quad \ldots.(5)$$

The remaining three metrics sensitivity, specificity and precision can be described as follow, sensitivity is also referred to as the *true positive (recognition) rate* (that is, the proportion of positive tuples that are correctly identified) [J.Han.,2006].

$$\text{Sensitivity} = \text{True}_{(pos)} / \text{True}_{(pos+neg)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(6)$$

while specificity is the *true negative rate* (that is, the proportion of negative tuples that are correctly identified) [J.Han.,2006].

$$\text{Specificity} = \text{True}_{(neg)} / (\text{True}_{(neg)} + \text{False}_{(pos)}) \quad \ldots\ldots\ldots..\ldots.(7)$$

The precision metric can be used to assess the percentage of tuples labeled as *"p2p"* that actually are *"p2p"* tuples. This measure is defined as [J.Han.,2006] :

$$\text{Precision} = \text{True}_{(pos)} / (\text{True}_{(pos)} + \text{False}_{(pos)}) \quad \ldots\ldots\ldots\ldots..\ldots.(8)$$

where $\text{True}_{(pos)}$ is the number of true positives (*"p2p"* tuples that are correctly classified as such). $\text{True}_{(neg)}$ is the number of true negatives (*"normal"* tuples that are correctly classified as such), $\text{False}_{(pos)}$ is the number of false positives (*"normal"* tuples that were incorrectly labeled as *"p2p"*), and $\text{False}_{(neg)}$ is the number of false negative (*"p2p"* tuples that are incorrectly labeled as *"normal"*) [J.Han.,2006].

Confusion Matrix (CM) is the measurement of performances of packets identification having these elements, Table (1) shows the elements of CM [Kusum K. B., 2010]:

1. **True Positive (TP):** Number of packets that are correctly classified as p2p.
2. **True Negative (TN):** Number of packets that are correctly classified as normal.
3. **False Positive (FP):** Number of normal packets that are classified as p2p.
4. **False Negative (FN):** Number of p2p packets that are classified as normal.

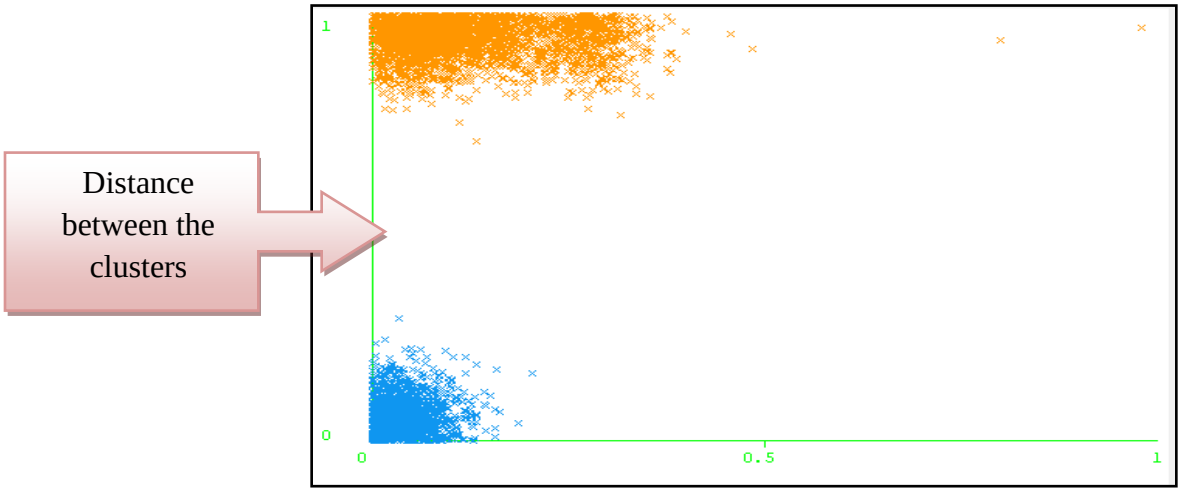
Table (1) : A confusion matrix for positive and negative tuples.

Confusion Matrix (Standard Metrics)		Predicted Class	
		P2P	Normal
Actual Class	P2P	TP	FN
	Normal	FP	TN

Confusion Matrix (CM) is a useful tool for analyzing how well your classifier can recognize tuples of different classes.CM is a square matrix in which each column corresponds to the predicted class, while rows correspond to the actual classes [Fatin, 2011].

4.1. Evaluate the performance of K-Means algorithm

Applying K-Means algorithm on the training data sets generates numbers of centers can be used in testing phase. Figure(5) explains the testing process with number of clusters (K=2). Presence of two areas can be observed; upper area with brown color represents the normal class and lower area with blue color represents a p2p class.



Figure(5):The testing process with number of clusters (K=2).

Experiment 1 : As mentioned in equation (5) the accuracy value represents the ratio between the true detection (true positive , true negative) to the total data set. After terminating the test phase with two clusters (K=2),true positive, true negative , false positive and false negative are calculated. Table (2) presents the CM related to the standard metric for proposed system, the number of successful prediction for normal packets is (2,939) from the total (2,986) , while failed in (47) packets. But the number of successful predication for p2p packets is (1,535) from the total (1,535). The elapsed time for testing phase with two clusters (K=2) is (78 sec).

Table (2): CM for K-Means Algorithm with K=2 .

Confusion Matrix (Standard Metrics)		Predicted Class	
		P2P	Normal
Actual Class	P2P	TP(1,535)	FN(0)
	Normal	FP(47)	TN(2,939)

Table (3) describes the remaining experiments with four performance metrics. Accuracy, sensitivity , specificity, and precision are four performance measures defined in equations(5), (6),(7) ,and (8) respectively .

Table (3): The performance of K-Means algorithm .

Experiment	No. of cluster	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	precision	Time in sec
1	2	1,535	2,939	47	0	98.9	100	98	97	78
2	3	1,535	2,706	280	0	93.8	100	90	84	89
3	4	1,535	2,939	47	0	98.9	100	98	97	90
4	5	1,099	2,939	47	436	89.3	71	98	95	106
5	6	1,099	2,939	47	436	89.3	71	98	95	113
6	7	1,535	2,900	86	0	98	100	97	94	123
7	8	1,535	2,904	82	0	98	100	97	95	131
8	9	1,535	2,904	82	0	98	100	97	95	236
9	10	1,535	2,900	86	0	98	100	97	95	152

First and third experiments with yellow color have the same value of accuracy but with different test time, First experiment with two clusters represents the best one because it has same accuracy(98.9) with the fewer time(78 sec).

Fourth and fifth experiments with blue color have the same value of accuracy but with different test time; Fourth experiment with five clusters represents the best one because it has accuracy (89.3) with the fewer time (106sec).

Seventh, eighth, and ninth experiments with green color have the same value of accuracy but with different test time. Seventh experiment with eight clusters represent the best one because it has accuracy (98) with the fewer time (123 sec).

Figure (6) shows relationships between the number of clusters and the accuracy value results for the different experiments .

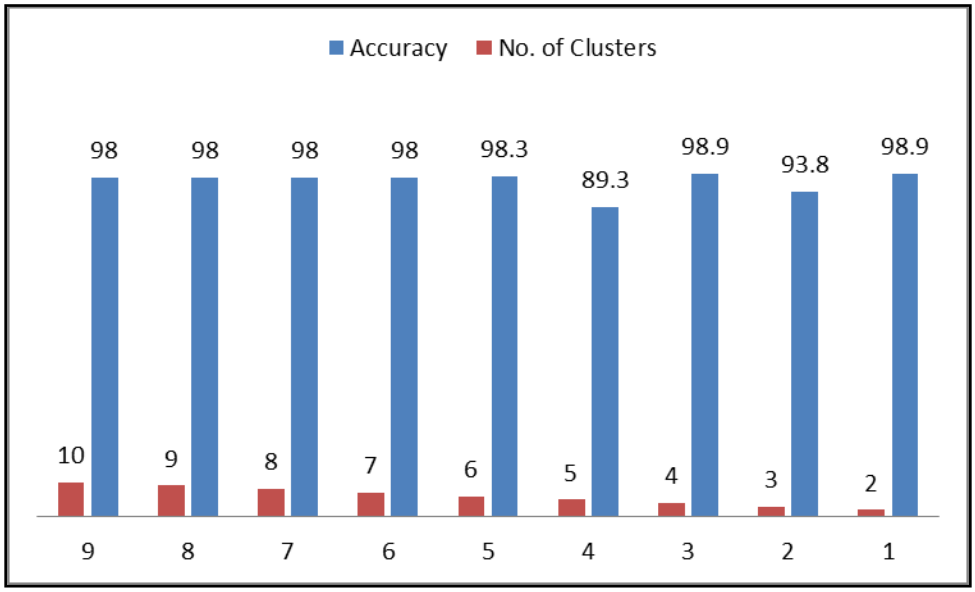


Figure (6):The relationships between the number of clusters and the accuracy values.

4.2. Evaluating the performance of Birch algorithm

Applying Birch algorithm on the training datasets generates numbers of clusters depending on the threshold value. In the testing phase, We calculate the mean value for each clusters and use it as a center. Running Birch algorithm with a range value of threshold between (0.1,0.2,0.3,…,1) produces a number of clusters depending on this value. Same performance metrics which are used with K-Means algorithm can be used to evaluate the results of Birch algorithm. Accuracy, sensitivity, specificity, and precision are four performance metrics used to evaluate the results of Birch algorithm.

Experiment 2 : Applying Birch algorithm with (0.2) threshold value generates (6) clusters which can be used to calculate the accuracy value in the testing phase. Table (4) displays the results of this experiment .

Table (4): CM for Birch Algorithm with threshold value(0.2) and K=6 .

Confusion Matrix (Standard Metrics)		Predicted Class	
		P2P	Normal
Actual Class	P2P	TP(1,441)	FN(94)
	Normal	FP(107)	TN(2,879)

Table (5) describes the remaining experiments with four performance metrics. Accuracy, sensitivity , specificity, and precision are four performance measures defined in equations(5), (6),(7) ,and (8) respectively .

Fifth ,sixth, seventh, and eighth experiments with yellow color have the same value of accuracy but with different test time, fifth and seventh experiments with three clusters represent the best accuracy because they have accuracy (98) with the fewer time (56 sec). Ninth experiment with blue color represents best accuracy value for birch algorithm with two clusters .

Table (5): The performance of Birch algorithm .

Experiment	Threshold	No. of cluster	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	precision	Time in sec
1	0.1	16	1420	2729	257	115	91.7	92	91	84	59
2	0.2	6	1441	2879	107	94	95	93	96	93	56
3	0.3	5	1459	2900	86	76	96	95	97	94	57
4	0.4	4	1469	2926	60	66	97	95	97	96	58
5	0.5	3	1493	2942	44	42	98	97	98	97	56
6	0.6	3	1499	2953	33	36	98	97	98	97	59
7	0.7	3	1499	2953	33	36	98	97	98	97	56
8	0.8	3	1510	2964	22	25	98	98	99	98	57
9	0.9	2	1524	2979	7	11	99	99	99	99	57

Figure (7) shows the relationship between the number of clusters and the accuracy value resulted from the different experiments .

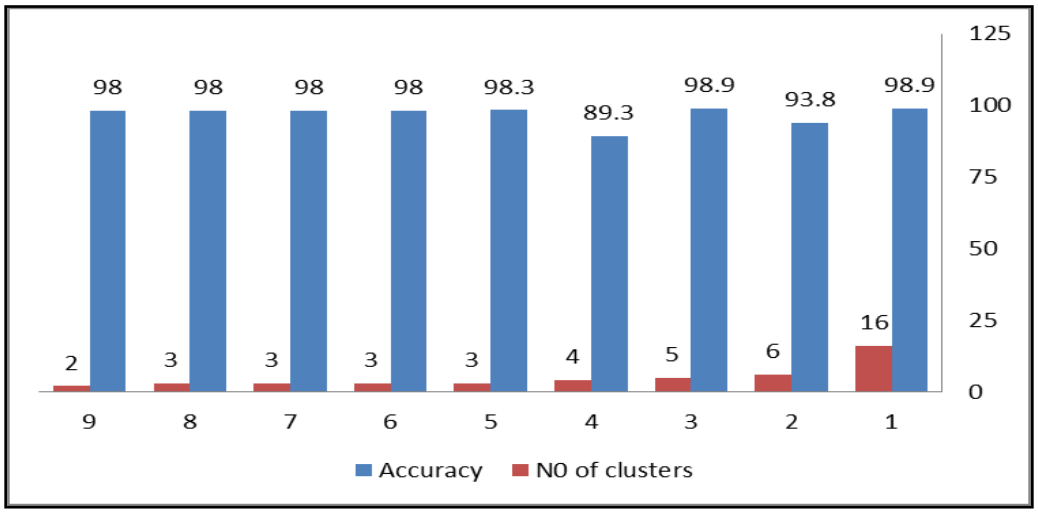


Figure (7): The relationships between the number of clusters and the accuracy values.

5.Conclusions

The proposed system has two clustering algorithms (K-Means, Birch) , after applying, the performance measure can detect the high and best accuracy value. The best and high accuracy value with K-Means algorithm in this work is (98.9%) with smaller elapsed time 78 sec. Birch algorithm has the best accuracy value (99%) with smaller elapsed time 57 sec and 0.9 threshold value .Speed training (building) and testing model, using a K-Means algorithm do not exceed (3) minutes compared with some systems that exceeded hours. Experimental results show that the K-Means and Birch algorithms are efficient for clustering in all data sets .Birch algorithm provides the highest accuracy with the highest threshold value and smallest number of clusters.

6. References

Andrew W., Denis Z., (2005)," Internet Traffic Classification Using Bayesian Analysis Techniques", Canada.

Bin Liu,(2011)," A Semi-Supervised Clustering Method For P2P Traffic Classification", JOURNAL OF NETWORKS, VOL. 6, NO. 3 .

Fatin N. M., Norita Md. N. and Kamaruzzaman S., (2011), "Identifying False Alarm Rates for Intrusion Detection System with Data Mining", International Journal of Computer Science and Network Security (IJCSNS), Vol. 11, No. 4, pp. 22-28.

Gerhard M., Sa Li, Georg C., (2007),"Traffic Anomaly Detection Using K- Means Clustering", Germany.

Ghousia U., Usman Ahmad, Mudassar A., (2013), "Improved K-Means Clustering Algorithm by Getting Initial Cenroids", World Applied Sciences Journal 27 (4): 543-551, ISSN 1818-4952.

Ilya K., Alexandr G.,(2012), "Application of BIRCH to text clustering", Proceedings of the 14th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" RCDL-2012,PereslavlZalesskii, Russia.

J.Han, M.Kamber, (2006),"Data Mining: Concepts And Techniques " ,second edition, Morgan Kaufmann, New Yourk,pp.1-745.

Jeffrey E., Martin A., Anirban M.,(2006)," Traffic Classification Using Clustering Algorithms", Italy .

Jenefa, A.; S.E Vinodh Ewards ,(2013)," Application Identification using Supervised Clustering Method ", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 3, Issue 2, pp.1334-1339.

Kusum K. Bharti, S. Shukla and S. Jain, (2010), "Intrusion Detection Using Clustering", International Journal of Computer Applications, Vol. 1, Issue 2, pp. 125-138.

L.Bernaille, R.Teuxeira , I. Akodkenous, A. Soule, and K. Slamatian, (2006), " Traffic Classification on the fly", ACM SIGCOMM Computer Communication Review, vol. Vol.36,No.2 .

Liu Y., Li Wei , Li Yunchun, (2007), "Network Traffic Classification Using K-Means Clustering", Second International Multi symposium on Computer and Computational Sciences , China.

McGregor, A.; M. Hall, P. Lorier, and J. Brunskill, (2004), "Flow Clustering Using Machine Learning Techniques," in *PAM 2004*, Antibes Juan-les-Pins, France.

Milojicic, D. S. e. a. (2002), "Peer-to-peer computing," HP Labs, Tech. Rep.

Sebastian Z., Thuy N., Grenville A., (2005),"Automated Traffic Classification and Application Identification usingMachine Learning", Proceedings of the IEEE Conference on Local Computer Networks.

Tian Zhang, Raghu R., Miron L ,(1996): BIRCH: An EfficientData Clustering Method for Very Large Databases. SIGMODConf: 103-114 .

Tian Zhang, Raghu Ramakrishnan, MironLivny, (2009),"BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies", Presented by Zhao Li.

Vijayarani, S.; Dr. A.Tamilarasi, N.Murugesh.(2011),"A NEW TECHNIQUE FOR PROTECTING SENSITIVE DATA AND EVALUATING CLUSTERING PERFORMANCE", International Journal of Information Technology Convergence and Services (IJITCS) Vol.1, No.2.