Clustering Techniques for Traffic Classification: A Comprehensive Review

Kate Takyi[1], Amandeep Bagga[2], Pooja Goopta[3]

[1,2,3]Department of Computer Applications, Lovely Professional University
Jalandhar-Delhi G.T Road, 144411 Phagwara, Punjab, India
[1]katetakyi@yahoo.com; [2]amandeep1@lpu.co.in; [3]pooja.19580@lpu.co.in

Abstract: **The threat of malicious content on a network requires network administrators and users to accurately detect desirable traffic flow into their respective networks. To this effect, several studies have found it imperative to classify traffic flow, and to use traffic classification in various applications such as intrusion detection, monitoring systems, as well as pattern detection in various networks. Research into machine learning techniques of clustering emerged due to the inefficiencies and drawbacks of the traditional port-based and payload-based schemes. The classic K-means technique of clustering, in combination with other methods and parameters, can be used to build newer unsupervised and semi-supervised approaches to meliorate the quality of service in networks. In this paper, we review twelve of the existing clustering techniques. The review covers their contribution to clustering methods, the existing challenges, as well as recommendations for further research in clustering traffic flows.**

Keywords: **Machine Learning, clustering techniques, Traffic Classification, QoS, K-means algorithm.**

I. INTRODUCTION

The concept of Real Time traffic classification using clustering techniques or algorithms involves the process of identifying and grouping similar packet traces of traffic under unique headings. In this paper, the term 'traffic' typifies IP traffic, internet traffic, private network traffic, as well as packets or data flows within a network. Clustering techniques have several applications in a network system. For instance, they are the precursors to intrusion detection systems, which use anomaly detection [1], [2], [4]. Furthermore, traffic classification is also applied in network security applications, network management, traffic analysis [34] and quality of service enhancement in networks [3], [29]. The traditional method of classifying traffic includes the Port based approach and the Payload method. The Port-based approach makes use of known ports in the list of registered ports ascribed from the Internet Assigned Numbers Authority (IANA) [5]. The use of dynamic ports and encryption of the IP layer made it difficult to find the genuine port number [6]. The use of dynamic ports and encryption of the IP layer, however, made it difficult to find the genuine port number [6]. The approach could also not classify encrypted packets in the traffic flow. In the Payload method of classification, the packets in the flow payload are examined carefully to find known signatures [7]. Classification is done based on the similarities of these signatures with the knowledge-based signatures, which have already been trained with the classifier [8]. This was challenged by the difficulty in updating the database of application signature, to enable effective comparisons.

Machine Learning (ML) approaches are advanced in terms of accuracy, performance, and complexity compared to traditional approaches. ML enables a system to train itself with existing database, from which the system later infers appropriate decisions regarding the traffic classification [31]. The approach of learning falls under three groups specifically supervised, unsupervised and semi-supervised. The supervised approach involves using an algorithm to learn from labelled training data set. The training dataset then serves as a guide to infer new examples that will also be fed to the system [33]. On the other hand, the input dataset in unsupervised learning is unlabeled. Hence, prior knowledge of the output is not known. The semi-supervised learning combines both supervised and unsupervised learning approaches. It makes use of some amount of labelled data coupled with unlabeled data, and thus makes use some form of supervision in the classification data [9]. The unlabeled data, however, forms the majority of the dataset fed as input to the system. Clustering techniques fall under unsupervised and semi supervised learning and mostly has to deal with association of some characteristic features [31]. There are different methods of clustering, namely classic K-means, hierarchical, density-based, grid-based, probabilistic model-based, and hybrid models [14], [17], [18], [32]. The classic K-Means method divides the dataset into a disjointed set of clusters and exemplifies each cluster with its centroid, whereas the Hierarchical clustering method focuses on generating a clustering hierarchy [18]. The Probabilistic model-based clustering presumes that data is formed by an assortment of the inherent probability distributions among various populations intended to be described by its characteristics. With the density-based model, clusters are defined to comprise groups of dense connected points and can develop in any direction forming arbitrary shapes. Grid-based partitions give data spaces to a multi-resolution structure of grids with a finite number of cells [17].

The selected twelve techniques for our review, to the best of our knowledge, have generated the most proficient results in the literature. The results obtained from the works are also discussed in terms of their accuracy in classification, performance, complexity in computations and run time where applicable. Due to limited space, evaluation results of the

978-1-5386-4692-2/18/$31.00 ©2018 IEEE

works discussed, achievements, limitations and gaps will be summarized in a table. The remaining sections of the paper will be discussed as follows; Literature Review, Discussion of Results, Challenges in Clustering Classification, Considerations for Future Research, and Conclusion.

II. LITERATURE REVIEW

McQueen [11] proposed a non-hierarchical method of partitioning captioned as the K-means algorithm. Lloyd [12] adopted this method to partition datasets into clusters based on a predefined number of initially selected centroids (k). By this method, the centroid of the K number of clusters (C_k) is iteratively computed using the Euclidean distance, until a convergence measure is reached. The aim of this algorithm is to utilize the Euclidean Distance to diminish the errors that occur in computing the mean squares from the objective function. The algorithm proved to be efficient with *O(jkn)* computational complexity where k represents the number of clusters, j depicts the number of iterations, and n equals to the total number of objects. The algorithm, which terminates at a local optimal, produces closely related clusters. It is also computationally faster, as compared to the hierarchical method, which is characterized by high complexity of $O(n^2)$. A weak or strong initialization produces poor and good clusters, respectively.

Incorporating the K-means clustering, Hirvonen and Laulajainen [22] proposed a two-phased classification of traffic for better Quality of Service (QoS) management. The aim of their work is to provide an efficient classifier that is able to make out target applications and detect unknown flows (noise) in the network, which could not be trained during the process. Their classification approach is based on flow behavior and the process comes in phases, namely assignment phase and labelling phase. The assignment phase assigns the flows to a cluster. The product of density measure and the phase threshold value determine the coverage of a particular cluster. The labelling phase uses the proposed algorithm to assign the appropriate label to the flows. From these phases, a decision is made and fed into the system to update the classifier for future reference.

Zhang et al. came up with the BIRCH method which incorporated scalability into the clustering model [13]. They used clustering feature tree (CF tree) with an in-memory structure and multilevel clustering to process large datasets in two main steps. Each step has an additional optional phase. In the first step, large datasets or data objects are compressed into a compact in-memory CF tree with the underlying clustering structure intact. By digesting into more suitable ranges, an optional smaller CF tree can be built. In the second step, they used an agglomerative algorithm with other flexible clustering methods to produce initial clusters, which were then refined based on their centroids.

Guha et al. used a hierarchical clustering algorithm, termed as Clustering using Representatives (CURE), to cluster larger dataset [15]. They hypothesized that CURE can withstand distortions caused by outliers and that this approach was best suited for arbitrary-shaped and non-spherical shaped clusters with wide variances. To differentiate the procedure of CURE from BIRCH, initial data sampling is randomly conducted and partitioned before clustering as compared to BIRCH which clusters all the data points from the start. The partitions are then clustered partially for the removal of outliers to be conducted. With no outliers present, the partial clusters are clustered again to produce finer clusters to be labelled to disk.

Ester et al. aimed at bringing out clusters shaped arbitrarily, called DBSCAN [14]. The Density based approach factored the quality of clusters that will be produced by considering the algorithms capacity to identify noise. With the origination of a density-based opinion of a cluster, parameters *Eps* and *MinPts* were defined. *Eps* reflects the density reachability possessed by clusters, and it characterizes the highest radius value of a point (P) neighbourhood. *MinPts* on the other hand, refers to the density connectivity, which is the lowest value of points in number with an *Eps* neighbourhood. Commencing from an arbitrary point a, the clusters are formed by finding if for any a, the distance to the P is less than or equal to *Eps*. The process is performed iteratively to include new points. DBSCAN possess a sensitive characteristic to its setting of parameters which are not easy to compute.

Ankerst et al. [16] attempted using OPTICS to overcome inherent DBSCAN drawbacks. The proposed algorithm emerged as less sensitive to the parameter settings. In accordance with the structure of density based methods, OPTICS generates a clustering order that stores information equating to a wide array of the parameter setting. The algorithm scales well with varying values of Eps (ε) within a range of 10, 000 to 100, 000. This gives OPTICS the advantage of being linear and running very fast with the number of data points.

Subramani et al. [27] adopted a hybrid of OPTICS and DBSCAN to tackle the issue of selecting an appropriate density threshold in social network community detection. The selection of a suitable density threshold contributes to the production of substantive clusters. With density defined by a distance function, OPTICS usage enabled the authors to select a good Eps parameter distance value in DBSCAN and also to realize the outcomes of using alternating density threshold values. The issue of whether a true definition for a community in social networks is feasible is an open ended query that is derived from the analysis of the authors.

Research into IP and Traffic classification using unique flow characteristics also proved to be very efficient. Zander et al. [19] suggested an automated method of classification, an unsupervised method based on the statistical flow characteristics of NetMate [10]. Using the Expectation Maximization algorithm [20] and AutoClass Algorithm [21], the packets are first partitioned into bi-directional flows for the computation of flow characteristics. Together with the flow

model's attributes, the classes can be learned for further classification of new flows. Results can be extracted for evaluation and other QoS purposes.

Semi-supervised techniques have also led to a new dimension of research in clustering approaches. Erman et al. [23] proposed one of the earliest semi-supervised works in clustering using supervised and unsupervised methods. Packet Milestones were used as a design consideration. The authors researched into classifying traffic using flow-based characteristics in applications and proposed a semi-supervised method for classifying traffic from known and unknown applications. The classifier is trained by comparingtraffic flows tomostly unlabeled flows, whereas labelled flows are minimally incorporated.

To make the accuracy of the clustering method of classification better, Wang et al. [24] suggested a semi-supervised strategy called set based constrained K-means. The statistical features of a flow are extracted along with some background information of the TCP/IP flows. Using Gaussian mixture density, the observed data and derived constraints are modelled. The authors established that introducing discrete features in flow clustering can increase the level of clustering accuracy. Based on only how the flow features are similar or dissimilar, they are grouped according the 5 tuple labels which are source and destination IP, source and destination port, as well as protocol used by the port. Flows that bear similarities from different applications are likely to be grouped into a particular cluster.

Within the framework of software defined networks (SDN), Wang et al. [25] classified clusters by combining Quality of Service requirements with the implementation of Deep Packet Inspection. .. They detected incoming flows possessing long lives with an SDN switch. With values of Hurst packet, port and average packet inter-arrival time as inputs into a mapping function, traffic flows were classified into their respective QoS classes. Statistical features were gathered and class queues were formed from the flows. The flows were then classified into their respective QoS classes.

For the purpose of Quality of Service, using a generative model (Hidden Markov's model, HMM) for semi-supervised sequence learning, Dianotti et al. [28] proposed a novel packet level method of traffic classification. The usage of this HMM sequence qualifies this approach to be in line with semi-supervision. Using the characteristics of packet size and inter packet time, the authors' classification was based on the aggregated characteristics using real network traffic and estimation. This made their approach usable on encrypted traffic as well.

III. DISCUSSION OF RESULTS

From Table I, it can be deduced that all proposed works achieved some level of accuracy ranging from 80% to above 90% indicating that clustering techniques are better for network traffic classification. Also, supervised and semi-supervised methods that incorporated the K-means, either as an aggregation or adopting its advantages, achieved higher percentages of accuracy compared to the others. From Table I, Hirvonen and Laulajainen [22] used the classic K-means in an unsupervised technique and resulted in classifying 97.8% of target applications. Similarly, from Table II, Erman et al. [23] using the Classic K-means in a semi-supervised technique achieved an over 90% accuracy in classifying flows.In addition, Wang et al.'s [24] semi-supervised SBCK, which has the Classic K-means and Gaussian mixture model (hybrid approach), resulted in 96% to 99% accuracy with feature discretization. They also obtained accuracy of 94% to 97% without feature discretization. SBCK also had better run time of 5 seconds compared to K-means of 13 seconds. The above methods with Classic K-means yielded better results compared to Zander et al.'s [19] probabilistic clustering approach, which obtained accuracy between 85% to 95%. Although Lloyd's [12] approach is one of the earliest to produce closely related clusters, its high sensitivity to noise remains a challenge. The Hierarchical agglomerative method used by Zhang et al. [13] overcame this drawback. The Hierarchical and density based methods adopted by some authors considered the run time of the proposed algorithms. In terms of run time, Ester et al.'s [14] DBCAN performs better than an existing density based algorithm CLARANS by a factor range of 250 to 1900. In spite of the similarity of Ankrest et al.'s [16] OPTICS to DBSCAN in run time, it could achieve a lower complexity of $O(n)$ using grid objects. The hybrid approach of the above methods in Subramani et al. [27], from Table I, is able to define and give a clearer understanding of the clustering structure. However, its runtime complexity is not discussed by the authors. The most interesting derivation is that, methods that aimed to improve quality of service also achieved better results with accuracy above 90%. In Table II, that the approaches used by Dianotti et al. [28] and Wang et al. [25] achieved accuracy greater than 90%, which makes their methods more effective than the K-means approach adopted by Erman et al [23]. Thus, incorporating quality of service features into the K-means method is more likely to produce higher percentages of accuracy.

IV. CHALLENGES IN CLUSTERING CLASSIFICATION

Even though the clustering technique of network traffic classification has yielded higher results interms of accuracy and performance, some challenges still persist.

The method of clustering itself has a challenge of how to produce good and non-overlapping clusters. The definition of a good cluster depends on the purpose for which the clustering is to be used or what it seeks to achieve. Another challenge is how to reduce the error rate. Roughan et al. [26] investigated the origin of this problem using statistical signatures of the flows, utilizing algorithms from machine learning and Nearest Neighbours for the purpose of Quality of Service. Their evaluation resulted that flows consisting of different

applications are more prone to errors. As the number of mixed applications increases, the error rate also increases. The challenge of a better clustering technique with low computational complexity is another challenge in Network traffic classification. To the best of our knowledge, there is no proposed work that has achieved a lower computational complexity than K- means and overcoming the drawbacks of K-means at the same time.

V. CONSIDERATIONS FOR FURTHER RESEARCH

Existing algorithms and methods which havehad the greatest impact on clustering traffic flowshave been discussed in the paper.These algorithms and most of the existing work are focused on features like packet size, inter-arrival time, including some QoS features as well [30]. However, their over-concentrationon particular applications that are traversing through the network limits their capacity to classify service classes efficaciously for better QoS.In [29], the authors aimed at overcoming this limitation, but only focused on internet video traffic without adequate attention on other types of traffic that can traverse through the same network. We therefore recommend further research into Quality of Service approaches to clustering. QoS levels provided by networks form an important aspect to many networks and service providers, therefore developing a more effective algorithm that uses some QoS parameters like throughput, packet loss, packet fragmentation, and delay will be of great value.

VI. CONCLUSION

Researchers have keen interest in developing more accurate methods of classifying and identifying real-time traffic patterns in network security and other network solutions. A lot of models have been formulated based on the existing unsupervised and semi-supervised methods of clustering. These models comprise techniques, which demonstrate the algorithm's capability to handle noise and its performance and ability to classify a large dataset of real time network traffic. Although classic K-means approach has served as a relevant model for the development of several semi-supervised clustering approaches, related computational complexity impedes its ability to work with limited computational resources. However, to our utmost knowledge, there is limited research on how the algorithms will perform under certain QoS parameters are incorporated, which is our aim to investigate in the future.

TABLE I: Summary of Unsupervised Clustering Methods

Author	Objectives	Clustering Method	Clustering Parameters	Limitations	Results
Lloyd [12] K -Means	To diminish the errors that occur in computing the mean squares in cluster formation	Classic K-means	Distance function as a parameter setting	* Sensitive to noisy data * poor clustering resulting from poor initialization of centroids	Produces closely related clusters compared to the traditional hierarchical methods
Zhang *et al*. [13] BIRCH	To use a limited amount of resources to process large datasets	Hierarchical (agglomerative algorithm)	* Clustering feature tree (CF tree) * Multilevel approach of clustering	Sensitivity to insertion of data points	* Handle outliers (noise), * Higher workload base performance * Time : clusters large datasets in less than 15 seconds (within 10-14 seconds) to K means (minimum within 12 – 44 seconds range), CLARANS (Minimum of 816 seconds)
Ester *et al*. [14] DBSCAN	To better the quality of clusters using the algorithm's capability to identify noise	Density Based Clustering	Density reachability *(Eps)*, Maximum radius of neighborhood *(MinPts)*	* Sensitivity to parameter settings *(Eps* and *MinPts)*. * Difficulty in computing parameters	* Accuracy: Able to identify and detect noise points while CLARANS assigns to nearest cluster * Run time: with increasing database size, DBSCAN performs better than CLARANS by a factor range of 250 to 1900. * Complexity of time which is fair enough
Guha *et al*. [15] CURE	To Identify non-spherical shaped clusters, arbitrary shaped clusters and withstand outliers in large datasets	Hierarchical	* Representative points for clusters * Shrinking factor	High computational complexity (cost) with higher dimensional space of input size (from large datasets)	* Produces high quality clusters. * Time: 50% lower execution time compared to BIRCH with increasing number of points.
Ankerst *et al*.	To overcome	Density Based	* Density	Challenge of	* Reachability plot is insensitive to input

[16]OPTICS	the limitations of DBSCAN's sensitivity to its parameters		reachability *(Eps)*, Maximum radius of Neighbourhood (*MinPts)* * Augmented Clustering ordering / structure	managing the clustering order with increasing updates of the database taking place	parameters when compared to DBSCAN and other clustering methods * Run Time: Fairly same as DBSCAN with its parameter setting, but lower other parameter settings such as tree based special index *(O (n log n))*, or using grid objects (*O(n))*
Subramani *et al*. [27]	To select an appropriate density threshold in social network community detection	Hybrid Approach (OPTICS & DBSCAN)	Density threshold parameter	Computational complexity of the hybrid approach not discussed	* Community definition is liable to lead to sudden change and relies on the application assumptions used. * Hybrid approach gives clear understanding into clustering structure * Ease of density threshold selection using the proposed method.
Zander *et al*. [19]	* To improve the overall intra class homogeneity * To overcome traditional methods of classification limitations.	Probabilistic Clustering Approach (Expectation Maximization and mixture models (AutoClass)	* Statistical flow characteristics * Intra class Homogeneity as a metric.	Performance on increasing datasets and runtime complexity not considered	Achieves an average 85% accuracy of clustering the flows with some applications achieving as high as close to 95%
Hirvonen and Laulajainen [22]	To provide an efficient classifier that is able to identify target applications and classify network flows in applications that are untrained as unknown.	Classic K – Means	* Flow behaviours * density measure * phase threshold value	* Calculation and determination of threshold values not discussed. * The evaluation compared its efficiency to only pure port based classification and not to other renowned existing works * Computational heaviness of the proposed work is not discussed	* Classifies 97.8 % of target applications * precision: detection of untrained flows from applications

TABLE II: Summary of Semi-supervised Clustering Methods

Author	Objectives	Clustering Method	Clustering Parameters	Limitations	Results
Erman et al. [23]	To build a fast and accurate classifier that can adapt to known and unknown applications.	Classic K - means	Distance function, Flow characteristics, packet Milestones	Do not compare results and performance with other works or classifiers	A high flow and byte accuracy is achieved with over 90% accuracy.
Wang et al. [24] SBCK	To improve upon the accuracy of clustering method of classification	Hybrid: Probabilistic Hierarchical (K - means with Gaussian Mixture Model)	Flow Statistical Features, Feature Discretization, Log Likelihood,	K means outperforms SBCK for small datasets in terms of run time SBCK – 0.4 seconds, K-means – 0.2 seconds.	* Accuracy: SBCK – 94 to 97 percent, K-means – 73 to 81 percent, EM – 90 to 93 percent (at higher levels of K =500) * Feature Discretization: SBCK – 96 to 99 percent accuracy * Run time: SBCK – 5 seconds, K-Means – 13 seconds for large datasets

Author	Objectives	Clustering Method	Clustering Parameters	Limitations	Results
Dianotti et al. [28]	To develop a multiclassifier for higher accuracy to achieve a better Quality of Service	Hybrid (Hidden Markov's Model with Packet features)	Packet size, inter packet time,	Do not compare its performance with other classifiers	classifies more than 90% applications correctly
Wang et al. [25]	To realize an accurate traffic classification for Improved Quality of service	Hybrid (Machine learning & Deep Packet Inspection)	QoS requirements, average packet inter arrival time, Hurst parameter, packet length	Issue of packet loss not addressed.	Test accuracy exceeds 90% which performs better than the existing K-means method in [23]

REFERENCES

[1] Z. S. Hosseini, S. J. S. M. Chabok and S. Kamel. "DOS intrusion attack detection by using of improved SVR." In *Technology, Communication and Knowledge (ICTCK)*, 2015 International Congress on, pp. 159-164, IEEE, 2015.

[2] A. Garg and P. Maheshwari. "Identifying anomalies in network traffic using hybrid Intrusion Detection System." *In Advanced Computing and Communication Systems (ICACCS), 2016 3rd International Conference on*, vol. 1, pp. 1-6. IEEE, 2016.

[3] C. M. Tseng, G. T. Huang, and T. J. Liu. "P2P traffic classification using clustering technology." *In System Integration (SII), 2016 IEEE/SICE International Symposium on*, pp. 174-179, IEEE, 2016.

[4] M. Ahmed, A. N. Mahmood, and J. Hu. "A survey of network anomaly detection techniques." *Journal of Network and Computer Applications*, vol. 60, pp. 19-31, 2016.

[5] Internet Assigned Numbers Authority (IANA), http://www.iana.org/assignments/port-numbers, as of May 25, 2017.

[6] T. T. Nguyen and G. Armitage. "A survey of techniques for internet traffic classification using machine learning." *IEEE Communications Surveys & Tutorials 10,* no. 4 (2008), pp. 56-76, 2008.

[7] A. W. Moore. and K.Papagiannaki. "Toward the Accurate Identification of Network Applications." In *PAM*, vol. 5, pp. 41-54, 2005.

[8] F. Dehghani, N. Movahhedinia, M. R. Khayyambashi, and S. Kianian. "Real-time traffic classification based on statistical and payload content features." In *Intelligent Systems and Applications (ISA), 2010 2nd International Workshop on*, pp. 1-4. IEEE, 2010.

[9] O. Chapelle, B. Scholkopf and A. Zien. "*Semi-supervised learning* (chapelle, o. et al., eds.; 2006)[book reviews]." *IEEE Transactions on Neural Networks,* vol.20, no. 3, pp. 542-542, 2009

[10] NetMate, http://sourceforge.net/projects/netmate-meter (as of August 2005)

[11] J. MacQueen. "Some methods for classification and analysis of multivariate observations." In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14, pp. 281-297, 1967.

[12] S. Lloyd. "Least squares quantization in PCM." *IEEE transactions on information theory*, vol.28, no. 2, pp. 129-137, 1982.

[13] T. Zhang, R. Ramakrishnan, and M. Livny. "BIRCH: A new data clustering algorithm and its applications." *Data Mining and Knowledge Discovery*, vol.1, no. 2, pp. 141-182, 1997.

[14] M. Ester, H.P. Kriegel, J. Sander, and X. Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." In *Kdd*, vol. 96, no. 34, pp. 226-231, 1996.

[15] S. Guha, R. Rastogi, and K. Shim. "CURE: an efficient clustering algorithm for large databases." In *ACM Sigmod Record*, vol. 27, no. 2, pp. 73-84, ACM, 1998.

[16] M. Ankerst, M. M. Breunig, H. P. Kriegel and J. Sander. "OPTICS: ordering points to identify the clustering structure." *In ACM Sigmod record*, vol. 28, no. 2, pp. 49-60. ACM, 1999.

[17] R. Agrawal, J. Gehrke, D. Gunopulos and P. Raghavan. "*Automatic subspace clustering of high dimensional data for data mining applications*". Vol. 27, no. 2. ACM, 1998.

[18] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. "*Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann", Second Edition, Morgan Kaufman Publishers, 2016.

[19] S. Zander, T. Nguyen, and G. Armitage. "Automated traffic classification and application identification using machine learning." In *Local Computer Networks*, 2005, *30th Anniversary, The IEEE Conference on*, pp. 250-257, IEEE, 2005.

[20] A. McGregor, M. Hall, P. Lorier and J. Brunskill. "Flow clustering using machine learning techniques." *In International Workshop on Passive and Active Network Measurement*, pp. 205-214, 2004.

[21] P. Cheeseman and J. Stutz. "Bayesian classification (autoclass): Theory and results in advances in knowledge discovery and data mining eds." *Articles FALL*, pp. 51, 1996.

[22] M. Hirvonen and J. P. Laulajainen. "Two-phased network traffic classification method for quality of service management." In *Consumer Electronics, 2009. ISCE'09. IEEE 13th International Symposium on*, pp. 962-966. IEEE, 2009.

[23] J. Erman, A. Mahanti, M. Arlitt, I. Cohen and C. Williamson. "Offline/realtime traffic classification using semi-supervised learning." *Performance Evaluation 64*, no. 9, pp. 1194-1213, 2007.

[24] Y. Wang, Y. Xiang, J. Zhang, W. Zhou, G. Wei and L. T. Yang. "Internet traffic classification using constrained clustering." *IEEE Transactions on Parallel and Distributed Systems25*, no. 11, pp. 2932-2943, 2014.

[25] P. Wang, S. C. Lin and M. Luo. "A framework for QoS-aware traffic classification using semi-supervised machine learning in SDNs." In *Services Computing (SCC), 2016 IEEE International Conference on*, pp. 760-765. IEEE, 2016.

[26] M. Roughan, S. Sen, O. Spatscheck and N. Duffield. "Class-of-service mapping for QoS: a statistical signature-based approach

to IP traffic classification." In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pp. 135-148. ACM, 2004.

[27] K. Subramanian, A. Velkov, I. Ntoutsi, P. Kroger and H. P.Kriegel. "Density-based community detection in social networks." In *Internet Multimedia Systems Architecture and Application (IMSAA), 2011 IEEE 5th International Conference on*, pp. 1-8. IEEE, 2011.

[28] A. Dainotti, W.De Donato, A.Pescape and P.S.Rossi. "Classification of network traffic via packet-level hidden markov models. "In *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE*, pp. 1-5. IEEE, 2008.

[29] H. Jiang, A. W. Moore, Z. Ge, S. Jin and J. Wang. "Lightweight application classification for network management." In *Proceedings of the 2007 SIGCOMM workshop on Internet network management*, pp. 299-304. ACM, 2007.

[30] S. H. Yoon, J. S. Park, M. S. Kim, C. Lim and J. Cho. "Behavior signature for big data traffic identification." *In Big Data and Smart Computing (BIGCOMP)*, 2014 International Conference on, pp. 261-266. IEEE, 2014.

[31] U. R. Hodeghatta and U. Nayak. "Unsupervised Machine Learning." In *Business Analytics Using R-A Practical Approach*, pp. 161-186. Apress, 2017.

[32] P. Berkhin. "A survey of clustering data mining techniques". In *Grouping multidimensional data*, vol 25, pp. 71, 2006.

[33] B. Hu & Y. Shen. "Machine learning based network traffic classification: a survey." *Journal of Information & Computational Science*, vol. 9, no.11, pp. 3161-3170. 2012.

[34] D. Achunala, M Sathiyanarayanan & B. Abubakar. "Traffic classification analysis using omnet++." *In Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, pp. 417-422. Springer, Singapore 2018.