The comparison of Self-organized map and k-means algorithms used for 4G network performance evaluation

Shaoxuan Wang (shaoxuan867865371@gmail.com)
Universitat Politècnica de Catalunya (UPC)

Research Article

Keywords: mobile networks, 4G network, Long Term Evolution (LTE) networks

Posted Date: April 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-435517/v1

Type of the Paper (Article, Review, Communication, etc.)

The comparison of Self-organized map and k-means algorithms used for 4G network performance evaluation

Shaoxuan Wang [1]

[1] *Universitat Politècnica de Catalunya (UPC),* Barcelona, Spain;
* Correspondence: Shaoxuan.wang@upc.edu;

Abstract— With the increasing complexity of mobile networks, it has become more and more difficult to perform effective management of mobile networks, which has led to more data to be evaluated and optimized. This article focuses on the performance evaluation of Long Term Evolution (LTE) networks by using two unsupervised learning techniques. Besides, this paper aims to identify the pros and cons of these two clustering algorithms as well. To achieve the above goals, different dimensional datasets for learning a process based on two classic unsupervised clustering methods are introduced to this work. A Self-organized map (SOM) neural network and k-means are as a comparison algorithm and the sample data with three different degree correlation coefficients features with 63 LTE cells, which is from a major European city. The purpose behind using these two methods is to see how different dimensions of the datasets can be used for testing clustering effectiveness and we propose a method based on the features extracted from key performance indicators (KPIs) and Euclidean distance is used as the evaluation standard for the distance between different clusters and samples within clusters. The comparing results show that k-means has a better cluster performance in low dimension data set, whereas the SOM's performance unsatisfactory. On the other hand, the SOM's clustering performance is better than k-means in high dimension and big data set and it could visualize results. It was verified that there is a significant difference in the obtained results using different clustering algorithms.

1. Introduction

In recent years, the large-scale construction of LTE networks has ensured the coverage advantages of 4G networks, but the huge network scale has further increased the difficulty of network optimization. At the same time, Mobile Access Networks produce a large number of Operations, Administration, and Maintenance (OAM) data used by operators for network operational assurance. These data include multiple and diverse performance measurements and indicators that characterize the behavior of the radio cells. The task of how to organize and manage limited wireless resources has become more and more important and urgent, which the telecom operators need to face. For this goal, the emergence of intelligent network planning tools will provide a reference for solving this problem and it could adjust the allocation of wireless resources in real-time by evaluating the cells and user performance.

To measure the above problems, mobile network operators (MNOs) increasingly focus on creating tools and processes, which is not only designed to help engineers maintain and optimize mobile radio network but also aims to make the network more autonomy. Machine learning (ML) has recently become a hot technology, aiming to balance the computational complexity of the problem and accuracy issues, and aroused widespread concern in the mathematical optimization community [1]. This trend has inspired researchers to use limited resources to solve the problem of wireless network optimization methods based on machine learning [2].

Unsupervised learning is a type of automatic learning method, which is a type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision [3]. The difference from supervised learning is that unsupervised learning usually makes use of human-labeled data. But, unsupervised learning allows for the modeling of probability densities over inputs. It does not think over samples' labels and the datasets are treated as random variables, the goal of it is to look for the probability density distribution that accommodates them [4]. Usually, the methods of unsupervised learning are to sort out the datasets with similar characteristics; this dealing process is called clustering, which consists of separating each group with different feature characters in sets called clusters. The output of these methods represents familiarity or similarity of the information presented at the input. These methods have a memory capacity, learning ability, and pattern recognition capacity [5].

The main purpose of this work is to detect mobile cells showing similar performance based on the KPIs collected from the network, thereby determining cell groups with different performances. Since there is no basic fact about the performance of the cells deployed in the network, the system should apply unsupervised learning techniques to find groups of cells that exhibit similar performance. Besides, the system should classify these groups according to the performance of their cells, depending on the performance level specified by the mobile network characteristics. This work also aims to analyze the pros and cons of different traditional ML algorithms (i.e., k-means, spectral clustering, and Gaussian Mixture Models) and neural networks for cell clustering. Therefore, we make a deep analysis of the Self-Organizing Map (SOM) and k-means in data analyzing and processing. We compared the SOM model and k-means algorithm for the exploration of large datasets with different features consisting of temporally referenced values of numeric attributes with the use of clustering by using Matlab and represented different parameters of RAN performance by suitable feature vectors and compared the clustering outcomes of the SOM and k-means.

The rest of the paper is organized as follows. Section 2 provides a background and overview of the related work. Section 3 presents the OAM data collected from the LTE network. Section 4 describes the clustering scenario with the SOM model and k-means algorithm being explained. Section 5 provides simulation results and discussion. Finally, Section 6 draws the conclusions.

2. Related Work

The latest developments in the field of data processing capabilities have paved the way for the use of machine learning (ML) technology to provide the impetus for the exploration of solutions for automatically evaluating mobile network performance [6]. At the same time, data for assessing the network performance based on the key performance indicators (KPI) are general used in the wireless mobile network. For instance, as described in [5] and [7], a cellular network fault diagnose system related to applying support vectors machine (SVM) along with a Continuous Time Markov Chain (CTMC) analytical model and an automatic diagnosis system based on SOM technology to KPIs collected from mobile networks has been developed respectively, and their methods get a better network fault evaluation when compared to a traditional Proactive Self-Healing method. At the same time, authors in [8]present a mobile cells/UEs capacity architecture, which is enabling data-driven intelligence for performance monitoring in mobile networks and they are comparing four supervised learning (SL) algorithms (i.e., random forest, shallow multi-layer perceptron, support vector regression and k-nearest neighbors) with deep learning in predicting and estimating cell and user throughput in the downlink in busy hours from radio measurements collected on a cell basis in the Operation Support System (OSS), the results show that these four non-SL methods have a better performance in evaluating since they provide higher accuracy with reduced datasets. Network performance also could be evaluated by the anomaly detection in the wireless area, for instance, authors in [9] investigate a network intrusion detection based on comparing SOM, k-means, Ad-

versarial Learned Anomaly Detection (ALAD), and Deep Auto-encoding Gaussian Mixture Model (DAGMM), and the results show that k-means works the better than the other three. [10] set up a detection model, which is consisting of a support vector machine (SVM), spectral cluster, and deep learning methods to evaluate the wireless performance by detecting anomaly UEs, and results show that traditional ML methods, such as spectral cluster have a better performance in anomaly detection. The data using in this work has clear positive and negative labels and it is easy to calculate the accuracy of each algorithm. Authors in [11] [12] mainly focus their work on UEs clustering by analyzing the UEs activity and page views on the social network based on the SOM model and k-means. In this study, the results show that the SOM has a better performance in clustering analysis. At the same time, authors in [13] compared SOM, k-means, and hierarchical clustering algorithm for classified telecommunication UEs dataset in the 3G network, and the results show that SOM and k-means have a better performance than hierarchical clustering algorithm on classified data clustering. Different from previous works, our task is using the SOM and k-means algorithm to analyze the performance of the LTE radio access network and evaluate the quality of cluster results based on the SOM and K-means, respectively. At the same time, similar research to our work was conducted in [14]and this article analyzes the call detailed records by comparing the detected anomalies with ground truth information to verify whether the two algorithms correctly assess user mobility and traffic usage and it only uses a feature corresponding to mobile user activity, namely counting the number of incoming and outgoing calls and text messages. Finally, a study closer to ours was conducted in [15], which compared three clustering algorithms using performance indicators. The research revealed that there were no significant differences in the obtained results with both Expectation-Maximization (EM) using Gaussian Mixture Models (GMM) and spectral clustering in LTE cells clustering based on different KPIs when compared to the ones obtained with K-means. However, although closer to our work than the aforementioned investigation, this article focuses on network optimization and cell subscription capacity evaluation based on comparing three traditional ML methods. Contrary to our research, the main contributions of this paper can be summarized as follows:

- The achievement of a proposed unsupervised method for the number of clusters of LTE cell performance. For this goal, different groups of KPIs are tested to evaluate the cluster results.
- Compared to [13] using classified UEs datasets under the 3G network, we investigate the performance of all unsupervised learning algorithms on clustering based on different correlated groups dataset to analyze the results of LTE cell clusters in the low dimensional field.
- Compared to [15] using traditional ML algorithms (i.e., EM-GMM and spectral clustering), this is the first time that k-means has been compared with SOM neural network (NN) for LTE high/low-dimensional datasets clustering in RAN.

3. Collected Data

The analyzed data samples come from the performance measurement results of different cells extracted from 4G networks deployed in major European cities. From September 12 to September 26, 2017, a 15-day measurement and monitoring were carried out, which is including 11 macrocell base stations (BS) and 6 microcell BSs distributed within a total of 63 cells and a range of approximately 3.2 km2 LTE cell data. The geographic distribution of BS is shown in Figure 1. The red dot represents the macrocell BS, and the blue cross represents the microcell BS. The macrocell BS contains 63 LTE cells, and the microcell BS is configured with 7 UMTS cells, they are microcells BS 2, 11, 13, 14, 15, 16, and 17. for example, microcell 2, 11, 13, 16, 17 are working and covering 2 cells with 42 Mbps bandwidth respectively, microcell 15 is working and covering 4 cells with 42Mbps. For the macrocells under different base stations, BS1, 6, 7, and 8 respectively cover 6 cells

with a bandwidth of 20 MHz, and also cover 3 cells with a bandwidth of 10 MHz. Base stations 3 and 10 have the same number of cells, and they include 3 cells with a cell bandwidth of 20 MHz and 3 cells with a bandwidth of 10 MHz. Meanwhile, the base stations 4 and 5 have the same number of cells, which are comprised of three 20MHz bandwidth of the cell respectively. Finally, base station 9 is comprising two cells with a cell bandwidth of 20MHz and 10MHz bandwidth respectively.

A total of 29 characteristic performance metrics are collected for each cell. These characteristics are illustrated in Table 1, which is grouped into 9 categories. Except that the average/maximum of each eNB UE, CQI, and SINR PUCCH / PUSCH mean measurement are sampling per one hour, the other characteristics of the measured values are sampled for 15 minutes. In this way, each 4G cell contains 1404 samples (14 days with 96 samples per day, and 60 samples/day on the last day). There are a total of 85,977 input samples (i.e., we have 63 cells, and there are 1404 samples in each cell, except for the U, V, and W cells of BS 10 to retain one week of data (each cell has only 579 samples), and each input sample contains 29 different characteristics. Table 1 provides the global statistics (maximum, average, and minimum) of the collected performance measurements to illustrate their range of variation.

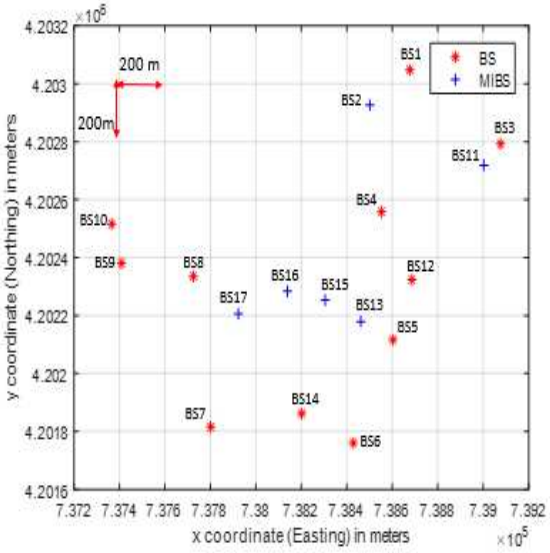

Figure 1. Base stations distribution of LTE.

Table 1. Characteristic performance measurements collected per cell

Category	**Features**	**Values (Max/Mean/Min)**
UEs	Average #*UEs* in UL	48.2/1.3/0
	Average #*UEs* in DL	100.6/1.5/0
	Max #*UEs* in UL	79.6 /4.98/0
	Max #*UEs* in DL	213/5.9/0
	Total Max #*UEs* in eNB	1388/712/3
	Total Average #*UEs* in eNB	1186/606/0
	Carrier Aggregation capable UEs (%)	45/7.45/ 0
Data volume	Data Traffic (MB)– Hourly data traffic in UL	2307/ 19/0
	Data Traffic (MB)– Hourly data traffic in DL	4431/208.3/0

Throughput	Max Cell Throughput (Mbps) in UL		48.2/2.87/0
	Max Cell Throughput (Mbps) in DL		232.8/ 31.2 /0
	Mean Cell Throughput (Mbps) in UL		87/12.7/ 0
	Mean Cell Throughput (Mbps) in DL		26/ 0.7/ 0
Physical Resource Block (PRB) Utilization	Average PRB Usage per Time Transmission Interval (TTI) (%) in UL		95.8/4.6/ 0
	Average PRB Usage per TTI (%) in DL		97.9/7.6/ 0
Handover (HO) Failure Indicators	Intra eNB HO Failure Rate (%)		66.7 /0.36 / 0
	Inter eNB HO over X2 Failure Rate (%)		25/ 0.17/ 0
	Inter eNB HO over S1 Failure Rate (%)		2.8/0.01/ 0
Radio Resource Control / Data Radio Bearer (RRC(DRB) Failure Indicators	RRC Drop Ratio (%)		15.9/ 0.6/0
	DRB Setup Failure Rate (%)		44.4/ 0.72/ 0
	RRC Setup Failure Rate (%)		8.97/0.41/ 0
Channel Quality	Average CQI		15/9.83/ 0
	Average Physical Uplink Share Channel (PUSCH) SINR		34/ 13 /-9
	Average Physical Uplink Control Channel (PUCCH) SINR		28/ 7 /-10
	MCS Distribution (%)	Low (MCS0-9)	15.5/7.84/0
		Medium (MCS10-19)	32.1/16.7/0
		High (MCS20-28)	52.4/31.6/0
Circuit Switched fallback (CSFB) attempts	CSFB attempts in idle mode		1421/ 24.7/ 0
	CSFB attempts in connected mode		898/10.7/ 0
Latency	Intra eNB Latency in DL (ms)		526/0.87/ 0
	Intra eNB Latency in UL (ms)		1381/22.9/ 0

4. Clustering Methodology Using SOM and k-means

4.1. General

For both SOM and k-means, the same simulation and experiment were executed, exploring architectures of 4 clusters/neurons and selecting the solution with the ratio of intra-cluster and inter-cluster distance among the achieved results. After finding these two methods, intra-cluster and inter-cluster distances were calculated for the entire dataset, and with the latter, comparing similarities and differences of the cluster centroid at SOM and k-means algorithms, which are based on three degrees' Pearson correlation coefficient features. Moreover, CPU occupation and simulation time were determined to make a statistical comparison of these two algorithms under high dimension with huge input samples. Before training the SOM model and k-means, input data have to be normalized. Otherwise, directly using raw data as input data, might cause huge deviations because of the range of numerical values taken but each of the features differs greatly. Therefore, standardization is applied using the following formula[16]:

$$O_{ij} = \frac{x_{ij} - min(x_j)}{max(x_j) - min(x_j)} \quad (1)$$

Where $\boldsymbol{x_{ij}}$ is the raw value of feature *j* in data item *i*, $\mathbf{min}(\boldsymbol{x_j})$ **and** $\mathbf{max}$ $(\boldsymbol{x_j})$ are the minimum and maximum data values of feature *j* in whole data items respectively.$\boldsymbol{O_{ij}}$ is the normalized value and the result of normalization is in the range of [0,1].

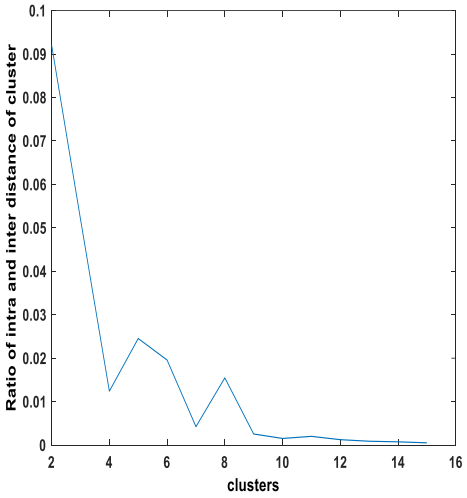

Figure.2 Optimal number of clusters fitting values of clusters

All the simulation work was programmed by using Matlab and three groups of data were tested. In each test, the best results for the structure from 2 to 15 clusters were achieved. Two clusters were used as a minimum cluster and fifteen as a maximum cluster. In every test, 3000 iterations were carried out, reorganizing randomly the database in each one of them, because the algorithm is a heuristic process.

The "elbow" method help data scientists select the optimal number of clusters by fitting a range of values for k. If the line chart resembles an arm, then the "elbow" (the point of inflection on the curve) is a good indication that the underlying model fits best at that point. In the visualizer, "elbow" will be annotated with a dashed line. In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. Figure 2 shows an elbow graph to determine the number of clusters, the X-axis represents the number of clusters, Y-axis represents the ratio of intra-centroid and inter-centroids' distance of clusters [17]. Intra-centroid cluster distance is also called within-cluster distance, which is representing the distance of samples in the cluster to the cluster centroid. On the contrary, the inter-cluster distance is the distance between two different clusters' centroids. These are two very important unsupervised learning clustering indicators. We can find that when k <4, the curve drops rapidly; From this work, we can find that when k > 4, the curve tends to have fluctuated and more and more stable in the end. We consider inflection point 4 to be the best value of k through the elbow method. Therefore, we set the initial number of clusters to 4, at the same time, the neurons of the SOM model are the same as well.

4.2. Self-organizing maps algorithm

SOM is a kind of unsupervised neural network, the main part of its algorithm. The idea is to project the n-dimensional input data onto certain representations. By reducing the data dimension, visual clustering can be used to obtain intuitive representations[18]. Its structure includes a vector input layer and a competitive output layer. It allows

visualizing the output via a competing layer[7]. SOM could change its structure according to external stimulation to make suitable clusters.

In the network, SOM is a single neural network, whose N nodes are distributing in a grid mode. Most distribution modes are hexagonal and rectangular. These could make a high dimension data project a low dimension space[19]. In the SOM model, an input node is widely connected with other nodes, and each other stimulates each other, and the strength of its interaction is determined by the connection weight. The connection weights include the weights between the input layer and the competing layer neurons, and the competing output layer nodes. The former represents the response of neurons to external input, and the latter represents the interaction between neurons. Figure.3 shows the topology of SOM and the training steps of the SOM network are as follows:

- Initialization: Assign and normalize the weight vector $W_j = [w_{j1}, w_{j2}, \dots, w_{jn}]^T$, where j=1,…, p, of each neuron node in the competition layer.
- Select the winning neuron: Randomly select an input sample X_i in the training set and normalize it. The weight vector W_j of each neuron is compared with the input vector for similarity. The Euclidean distance is selected so that the winning neuron w_q satisfies:

$$min_j\{\|w_j - X\|\} = \|w_q - X\| \quad (2)$$

- Adjust the weight: The winning neurons and their topological neighbors are moved closer to the input vector. Adjust learning rate $\rho(t)$ and neighborhood function $h_{ci}(t)$ The update rule for the prototype vector of neurons is:

$w_j(t+1) = w_j(t)+\rho(t)*h_{ci}(t)[X- w_j(t)]$, $j\epsilon N_j(t)$ (3)Where t is the time, $\rho(t)$ is the learning rate, which range is from 0 to 1, $hci(t)$ is the neighborhood neuron and $h_{ci}(t)$ is usually a Gaussian function, which is centered around the winning neurons.

- The iterative operation, set t=t+1, repeat work in 2 and 3 until the network is convergence.

The SOM network is defined in the measurement vector space. Since only the samples in the cluster have the greatest similarity, they can be mapped to adjacent neurons in the competition layer. Therefore, in the topological neighborhood of the competitive layer, updating neurons with weights close to the input sample can make the neurons of the competitive layer sensitive to the corresponding samples[20].

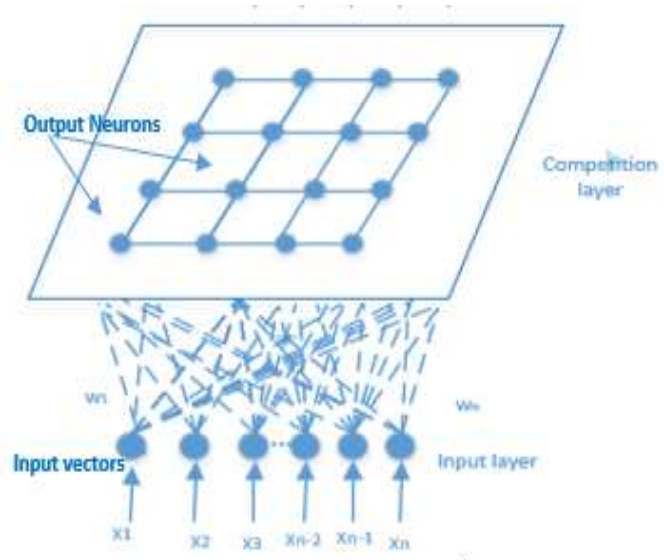

Figure.3 Network topology of SOM

4.3. k-means algorithm

The basic idea of the k-means algorithm is to take the mean value of the data samples in each cluster subset as the representative point of the cluster and through iteration. The process divides the data set into different categories so that the evaluation of clustering. The criterion function of energy can reach the optimal so that each cluster is generated. Compact and independent between classes [22]. In the iterative process, the objects in the clustering set are constantly moved until the ideal cluster set is obtained, and each class uses this class. The average value of the objects in the middle is expressed. Using the cluster obtained by k-means, the similarity of the objects in the cluster is very high, and the degree of dissimilarity between different cluster objects is also very high [23]. The basic principle of k-means for samples clustering can be depicted as follows:

- Initialization: Suppose the number of cluster centers in a given sample space is K, and the K initial centers are $C = \{C_1, C_2, C_3, \dots, C_k\}$,X=$\{X^{(1)}, X^{(2)}, \dots, X^n\}$ represents all input sample vectors. $S_i = \{X | X \in S_i\}$, which is representing all sample sets belonging to the i-th cluster center. Setting the threshold of iteration is ξ.
- Sample division:$X^p \in S_i$, if $\|X^p - C_i\| \leq \|X^p - C_j\|$, j=1, 2,…, K, and $i \neq j$.
- Calculate new cluster centers: $C_i^* = \frac{1}{N_i} \sum_{X^p \in S_i} X^p$, N_i is the number of samples in the cluster S_i.
- Check convergence: If $\|C_i^* - C_i\| < \xi$, iteration stopped, otherwise repeat 2 and 3.

The advantage of the k-means clustering algorithm is: it can cluster dynamically and has a certain adaptive. The convergence of the algorithm depends on the characteristics of the sample and the number of different regions that it can form. After the k-means algorithm is divided, the objects in the same cluster have the greatest similarity or correlation, while the objects in different clusters the similarity between them are as small as possible, and the k-means clustering method finds that spherical clusters are very suitable in small and medium-sized databases.

5. Performance evaluation

5.1. Low dimension clusters with different correlation features

In this simulation, firstly, we use 2-dimension data as an input, whose characters are low, medium, and high degree features. For instance, CQI (Channel Quality Indicator) is the downlink spectral efficiency indicator, which is indicated the quality of the network measured from terminal UEs, i.e. how much downlink throughput a UE, under certain radio conditions (i.e., interference conditions) [24]. Therefore, CQI could reflect the specific radio conditions accurately [25]. The downlink throughput or downlink network throughput is the rate of successful message delivery from the communication channel to the UEs. Throughput is usually measured in Mega-bits per second (Mbps), and sometimes in data packets per second (p/s) or data packets per time slot. This simulation is to explore the low correlation feature on data set clustering. Due to the correlation coefficient has different degrees of strength, therefore, we can calculate correlation-ship between different features. The correlation coefficient could be divided into three degrees, which are high, medium, and low. For instance, the high degree range is 0.6 to 1, and the medium's range is from 0.4 to 0.6. Finally, the low degree is from 0 to 0.4. These two features have a low correlation; whose correlation coefficient is 0.225. Drop rate is an important indicator in mobile communication, also known as the call interruption rate, which refers to the

probability of unexpected communication interruption during mobile communication. The call drop rate reflects the quality of mobile network communication to a certain extent. The latency is the time takes to get a packet from a specific point. The latency time is generally the sum of response delay and transmission delay and it is also measured in ms. These two features have a high correlation and its correlation coefficient is 0.512. Downlink data traffic is the rate of message delivery from the base station to different UEs and it is usually measured in Mega-bits. The downlink maximum throughput or downlink network throughput is the highest rate of successful message delivery from the communication channel to the UEs. Throughput is usually measured in Mega-bits per second (Mbit/s or Mbps), and sometimes in data packets per second (p/s or pps) or data packets per time slot. This simulation is to explore the low correlation feature on data set clustering. These two features have a high correlation and its correlation coefficient is 0.873.

From Table.2 we can find that the SOM intra cluster's distance is a little bit larger than k-means when they all have 4 clusters. For instance, clusters #1, #2, #3 and #4 in SOM are all larger than the same cluster of k-means. The distance gap among these three clusters is 0.12,0.14,0.02 and 0.17 respectively. The inter-cluster distance of them is very close when compared to the intro clusters. According to the shorter intra distance of clusters, the better cluster performance [26], we can conclude that the intra-cluster distance of k-means and SOM are closer. At the same time, Table 2. shows the average distance of inter-cluster centroids as well. For instance, cluster #1 and #4's distance is almost the same, the #2 and #3 are also similar, that means inter-distance of SOM and k-means in low dimension have little difference, the larger inter centroids' distance, the better cluster performance, therefore, we can only set the intra-cluster distance as evaluation criteria. Figure.4 illustrates the scatter distribution of SOM and k-means cluster centroid in 63 sample cells within throughput and CQI features. We can find the coordinate of the centroid in cluster#3 at the SOM model is the same as cluster#4 at the k-means algorithm.

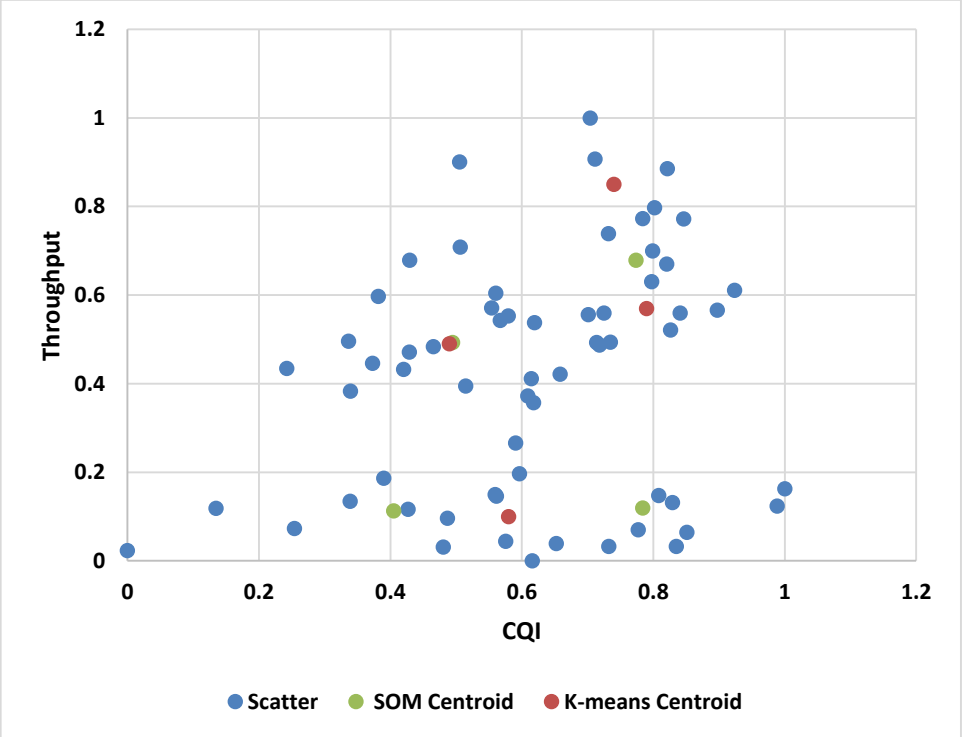

Figure.4 Scatter of Throughput and CQI by using SOM and k-means cluster

Comparing Table.3 and 4 we can find, the k-means cluster #3 and #4's characteristic is similar to SOM cluster #3 and #2. The only difference between these two kinds of the cluster is #1and #4, for instance, the character of k-means clauter#2 is the lowest downlink throughput and the CQI span of cluster #2 is very large, it from 0.013 to 1, which include

the low to high CQI within similar downlink throughput. But the character of cluster #4 at SOM has the lowest CQI and downlink throughput. So we can conclude that cluster #4 in SOM is similar to cluster#2 at k-means, however, samples in #2 of k-means have covered the samples of #4 at SOM. Finally, cluster#1at k-means and cluster#1of SOM are also different. For cluster#1of SOM, it has a medium CQI and high downlink throughput. At the same time, cluster#1 of k-means has high CQI and medium downlink throughput, therefore, these two clusters' character is very close. All in all, we can find in the low dimension features, there are certain differences, but their clustering characteristics are generally similar in a cluster of SOM and k-means, and the clustering result of k is more intuitive.

Table.2. Location of centroids and intra-cluster distance with SOM and k-means (low correlation features)

	SOM Centroid		***k*-means Centroid**		**Intra-cluster distance**		**Inter-cluster distance**	
	X(SOM)	***Y(SOM)***	***X(k-means)***	***Y(k-means)***	***SOM***	***k-means***	***SOM***	***k-means***
Cluster#1	0.40	0.11	0.79	0.57	0.29	0.17	0.41	0.37
Cluster#2	0.78	0.11	0.58	0.10	0.36	0.21	0.35	0.42
Cluster#3	0.49	0.49	0.74	0.85	0.34	0.32	0.30	0.37
Cluster#4	0.77	0.67	0.49	0.49	0.31	0.14	0.39	0.40

Table.3. Description of Obtained Cell Patterns of k-means Cluster

Patterns	**Characteristics**	**Comments**
Cell pattern#1 (Cluster#1)	High CQI and medium downlink throughput	The inter-cell interference among these cells in #1 is relatively small and traffic flows usage is huge
Cell pattern #2 (Cluster#2)	Lowest downlink throughput	The CQI span of cluster #2 is very large, from 0.013 to 1, but the throughput is the lowest, which means #2's inter-cell interference is high
Cell pattern#3 (Cluster#3)	Medium CQI and medium downlink throughput	The inter-cell interference among these cells in #3 is relatively small and traffic flows usage is huge
Cell pattern#4 (Cluster#4)	Highest CQI and downlink throughput	Little inter-cell interference among these cells in #3 and best quality of UEs

Table.4. Description of Obtained Cell Patterns of SOM Cluster

Patterns	**Characteristics**	**Comments**
Cell pattern#1 (Cluster#1)	Medium CQI and high downlink throughput	The inter-cell interference among these cells in #1 is relatively moderate and traffic flows usage is huge
Cell pattern #2	Highest CQI and highest downlink throughput	Little inter-cell interference among these cells in #2 and the best quality of UEs

(Cluster#2)		
Cell pattern#3 (Cluster#3)	Medium CQI and medium downlink throughput	The inter-cell interference among these cells in #3 is relatively small and traffic flows usage is huge
Cell pattern#4 (Cluster#4)	Lowest CQI and downlink throughput	The cells in #4 have far from the base station and the interference of these cells are high

From Table.5 we can find that SOM intra-cluster distance is smaller than k-means when they are in the first three clusters. But in cluster #4, the intra-cluster distance of SOM is quite larger than k-means. At the same time, Table.4 also shows the average distance between inter-cluster centroids, we can find that the inter-cluster distance of k-means and SOM are closer. For instance, only cluster #4's distance has a big difference between SOM and k-means, but SOM performance in #2 and #3 are better than K-means, but the inter-distance of #1 in K-means is better than SOM, and we can comprehensively consider the performance of SOM and k intra-cluster distance and inter-cluster distance. Therefore, according to the shorter distance of intra clusters and the larger inter-cluster distance the better cluster performance. Comparing the clustering performance in Table.2, no matter the intro-cluster distance of SOM or k-means in Table.5 are all smaller than Table. 2. Therefore, we conclude that k-means cluster performance is worse than SOM in medium correlation features at this simulation and the scatter of k-means. Figure.5 shows the distribution of SOM and k-means cluster centroids within drop rate and latency.

Comparing Table.6 and 7 we can find, the k-means cluster #1, #3 and #4's characteristic is similar to SOM cluster#1 #2, and #4. At the same time, the similar two kinds of the cluster are #1 in k-means and #1 in SOM, whose performance is all higher drop rate and latency. And the #3 in k-means and #2 in SOM have similar performance, which is moderate latency and drop rate. Finally, the performance of #4 at K-means and SOM are having the same characters, which all have the lowest latency and drop rate. So the only subtle difference is #2 in k-means and #3 in SOM, they all have high latency and medium drop rate. We can conclude that the SOM and k-means are having a similar performance under medium correlation features in low dimensions.

From Table.8 we can find that the intra-cluster distance of SOM is close to k-means but smaller than k-means. For instance, in these four clusters, the gap distance between SOM and k-means are different, which are 0.03,0.1,0.04 and 0.3 respectively. Table.8 also shows the inter-cluster centroids, we can find that the inter-cluster distance of k-means and SOM are closer. For instance, cluster #1 distance is almost the same, the SOM performance in #2 and #3 are better than k-means, but the inter-distance of #4 in k-means is better than SOM, and we can comprehensively consider the performance of SOM and k intra-cluster distance and inter-cluster distance. According to the shorter the distance of inter-clusters and the larger inter-cluster distance the better cluster performance. Therefore, we can conclude that k-means cluster performance is worse than SOM in high correlation features at this simulation. Figure.6 illustrates the scatter distribution of SOM and k-means cluster centroid in 63 sample cells within throughput and data traffic features. Comparing to Table.1, 4, and 7, we can find that the higher value of the correlation coefficient, the better cluster performance we will get no matter in the SOM model or k-means.

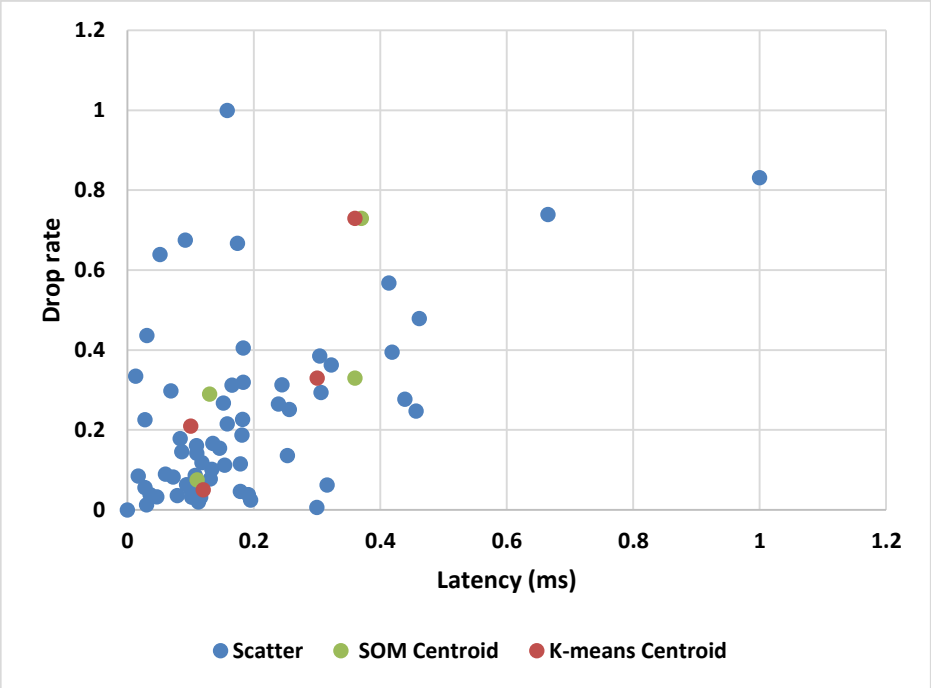

Figure.5 Scatter of Latency and Drop rate by using SOM and K-means cluster

Table.5. Location of centroids and intra-cluster distance with SOM and K-means (medium correlation features)

	SOM Centroid		K-means Centroid		Intra-cluster distance		Inter-cluster distance	
	X(SOM)	*Y(SOM)*	*X(k-means)*	*Y(k-means)*	*SOM*	*k-means*	*SOM*	*k-means*
Cluster#1	0.36	0.33	0.36	0.73	0.14	0.36	0.33	0.56
Cluster#2	0.11	0.075	0.30	0.33	0.09	0.1	0.43	0.32
Cluster#3	0.37	0.73	0.1	0.21	0.06	0.08	0.54	0.32
Cluster#4	0.13	0.29	0.12	0.05	0.32	0.11	0.32	0.39

Table.6. Description of Obtained Cell Patterns of k-means Cluster

Patterns	Characteristics	Comments
Cell pattern#1 (Cluster#1)	Highest drop rate and latency	The inter-cell interference among these cells in #1 is big and traffic flows usage is huge
Cell pattern #2 (Cluster#2)	Moderate drop rate and latency	The inter-cell interference among these cells in #2 is not big and channel quality is moderate.
Cell pattern#3 (Cluster#3)	Moderate drop rate and latency	The inter-cell interference among these cells in #3 is not big and channel quality is moderate
Cell pattern#4 (Cluster#4)	lowest drop rate and latency	Little inter-cell interference among these cells and best channel quality

Table.7. Description of Obtained Cell Patterns of SOM Cluster

Patterns	Characteristics	Comments

Cell pattern#1 (Cluster#1)	Highest latency and drop rate	The inter-cell interference among these cells in #1 is big and traffic flows usage is huge
Cell pattern #2 (Cluster#2)	Medium latency and the drop rate	The inter-cell interference among these cells in #2 is not big and channel quality is moderate.
Cell pattern#3 (Cluster#3)	High latency and the low drop rate	The inter-cell interference among these cells in #3 is not big and channel quality is moderate
Cell pattern#4 (Cluster#4)	Lowest drop rate and latency	Little inter-cell interference among these cells and best channel quality

Comparing Table.9 and 10 we can find, the k-means cluster #1and #3's characteristic is similar to SOM cluster #2 and #3. These two group clusters are all medium data traffic. At the same time, the similar two kinds of the cluster are #1 in k-means and #4 in SOM, whose performance is all higher data traffic and downlink throughput. Finally, the performance of #2 at k-means and #4 at SOM are having the same characters, which all have the lowest data traffic and throughput. Therefore, we can conclude that the SOM and k-means are having a similar performance under the high correlation features in the low dimension. And the specific number of each cluster in k-means.

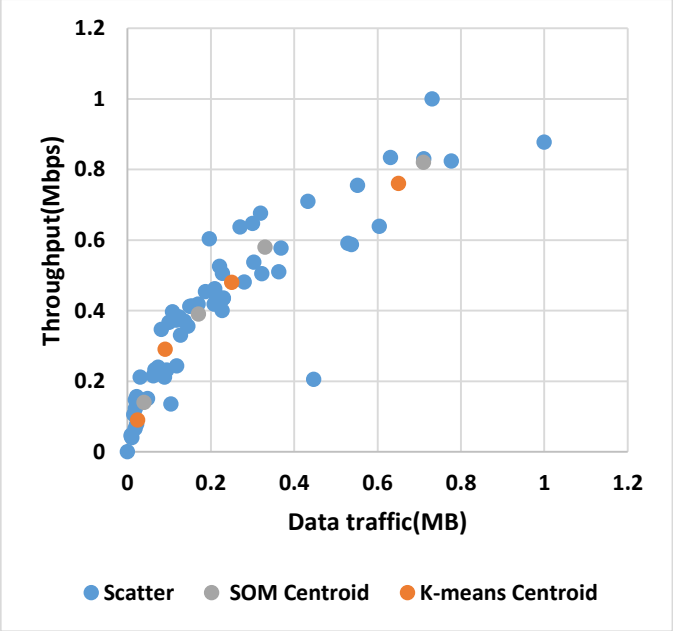

Figure.6. Scatter of data traffic and throughput by using SOM and k-means cluster

Table.8. Location of centroids and intra-cluster distance with SOM and k-means (high correlation features)

	SOM Centroid		*k*-means Centroid		Intra-cluster distance		Inter-cluster distance	
	X(SOM)	*Y(SOM)*	*X(k-means)*	*Y(k-means)*	*SOM*	*k-means*	*SOM*	*k-means*
Cluster#1	0.17	0.39	0.25	0.48	0.08	0.11	0.31	0.30
Cluster#2	0.71	0.82	0.024	0.09	0.14	0.24	0.52	0.40
Cluster#3	0.04	0.14	0.09	0.29	0.09	0.13	0.44	0.30

Cluster#4	0.33	0.58	0.65	0.76	0.11	0.41	0.31	0.53

Table.9. Description of Obtained Cell Patterns of k-means Cluster

Patterns	Characteristics	Comments
Cell pattern#1 (Cluster#1)	Medium downlink throughput and data traffic	The inter-cell interference among these cells in #1 is not big and traffic flows usage is moderate
Cell pattern #2 (Cluster#2)	Lowest downlink throughput and data traffic	The CQI span of cluster #2 is very large, from 0.013 to 1, but the throughput is the lowest, which means #2's inter-cell interference is high
Cell pattern#3 (Cluster#3)	Medium downlink throughput and data traffic	The inter-cell interference among these cells in #3 is not big and traffic flows usage is moderate
Cell pattern#4 (Cluster#4)	Highest data traffic and downlink throughput	Little inter-cell interference among these cells and best quality of UEs

Table.10. Description of Obtained Cell Patterns of SOM Cluster

Patterns	Characteristics	Comments
Cell pattern#1 (Cluster#1)	Highest downlink throughput and data traffic	The inter-cell interference among these cells in #1 is relatively small and traffic flows usage is huge
Cell pattern #2 (Cluster#2)	Medium downlink throughput and data traffic	Little inter-cell interference among these cells in #2 and the best quality of UEs
Cell pattern#3 (Cluster#3)	Medium downlink throughput and data traffic	The inter-cell interference among these cells in #3 is moderate and traffic flows usage are medium
Cell pattern#4 (Cluster#4)	Lowest downlink throughput and data traffic	The cells in #4 have far from the base station and the interference of these cells are high

5.2. High dimension clusters with different samples

Different from section 5.1 using the low dimension and different correlation features, we will use the high dimension features (i.e. 29 features) to test SOM and k-means performance. In the first stage, from Table.11, we can find that the overall difference is not obvious between the average distance of cluster and samples in k-means and SOM. For instance, cluster #2 and #3 in k-means and SOM is similar and closer. However, the only huge difference is cluster #4, which is 0.93 and 0.53 respectively. The average distance of different cluster centroid in k-means is smaller than SOM. The average distance of cluster centroid and samples and an average distance of different cluster centroid in k-means are similar, for example, these two values in cluster #3 are 0.56 and 0.64 respectively, which is much closer. From this, we can conclude that, firstly, k-means can cluster high-dimensional data, but the intra-cluster distance and the inter-cluster distance of the cluster are very close. On the other hand, SOM has a good performance in the distance of cluster centroid. Therefore, for high dimension datasets, researchers prefer to using the other algorithms except for the k-means. Table.12. shows the cluster hits of SOM and k-means

and Figure.7 shows the sample's distribution at SOM topology, we can find that the number of hits in cluster#1 and #2 are similar between these two methods, which are 12 and 23 in k-means and 11 and 21 in SOM. The only difference between the hits' number between SOM and k-means are cluater#3 and #4.

In the second stage, we try to obtain cell behavior based on the short-term performance of the cell observed during the entire measurement time domain. It is different from the long-term behavior capturing the cell behavior within 24 hours (1 day), In the short-term community behavior analysis, we hope to perform a similar cluster analysis, but now we will capture the performance indicators of the behavior of the cells with a sampling period of 1 hour as the input sample, and the entire 15 days will be divided into 350 hours (there are only 14 hours of data in the last day), and the U, V, and W cells of BS 10 also have only one week of cell behavior data. Therefore, based on the preprocessing of 63 groups of cell data, we obtained a total of 21341 hours of data in this work. Finally, according to the average of the entire measurement period of all cells, 29-dimensional features are selected as the column vector, and 21341 sets of time samples are used as the input data array of the row vector (one for each feature). Therefore, the input data sample section is expressed as a 21341x29 matrix.

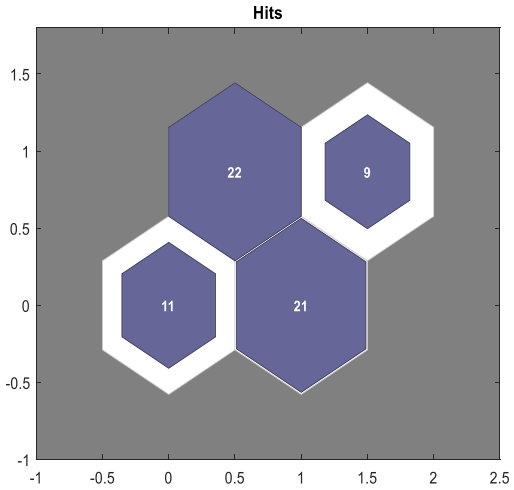

Figure.7. Hits of SOM for 63 samples with 29 features

From Table.13, we can find the distribution of the hits of SOM topology is located in neurons #1, also named cluster#1. On the contrary, most samples of k-means are distributed in cluster#2, which value is 20159. Figure 8 also shows the distribution of the hits of samples in the SOM topology. Such as, different from k-means, SOM's most samples are located in cluster#1 and its value 18101, and SOM's distribution is more evenly distributed when compared to k-means.

Table.11. distance comparison between k-means and SOM

	The average distance of cluster centroid and samples (k-means)	**The average distance of cluster centroid and samples(SOM)**	**The average distance of different cluster centroid (k-means)**	**The average distance of different cluster centroid(SOM)**
Cluster#1	0.21	0.35	0.67	1.46
Cluster#2	0.38	0.43	1.25	1.93
Cluster#3	0.56	0.64	0.97	1.8
Cluster#4	0.93	0.53	1.13	1.27

Table.12. Hits of K-means and SOM for 63 samples with 29-dimension features

	Cluster#1	**Cluster#2**	**Cluster#3**	**Cluster#4**
***k*-means**	12	23	8	20
SOM	11	21	22	9

Table.14 illustrates that the overall difference is not big between the average distance of cluster and samples in k-means and SOM except for cluster #4. For instance, cluster#1 and #2 in k-means and SOM is closer and the only huge difference is cluster#4, which are 1.1 and 0.023 respectively. The average distance of different cluster centroid in k-means is higher than SOM. However, the average distance of cluster centroid with samples and an average distance of different cluster centroid in k-means are different, and the degree of it is smaller than the SOM algorithm. For example, these two values in cluster#3 are 0.01 and 1.02, respectively, which is almost a hundred times the gap. And the other distance's gap is also huge, which is two or three times gap as well. From Table.15, we also could find that, although SOM is better than the k-means in high-dimensional large datasets clustering, SOM neural network algorithm has high complexity and slow learning speed. For instance, the CPU occupation of SOM is much higher than the k-means, the simulation time is almost a 5 times gap when compared to k-means. Therefore, although the k-means algorithm is more concise and fast than SOM with a large dataset in high dimension, the cluster performance and accuracy are worse than SOM. Therefore, researchers prefer to using the other algorithms except for the k-means to cluster complicated and high dimensional datasets. We can conclude that for large and complex datasets clustering, the SOM's performance is better than the k-means and SOM has a good cluster effective on complex data.

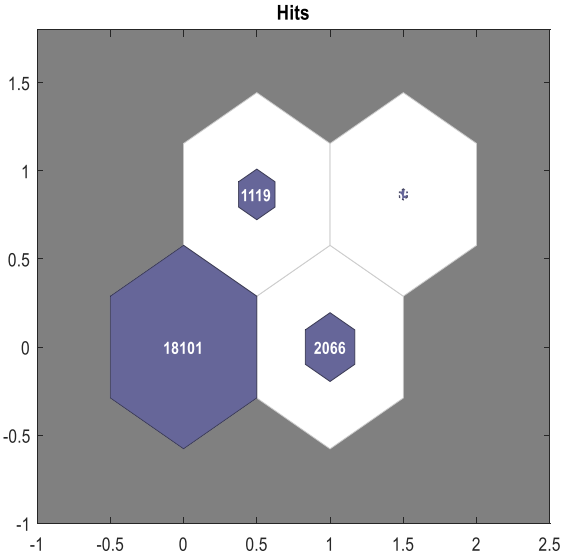

Figure.8. Hits distribution of SOM for 21341 samples with 29 features

Table.13. Hits of k-means and SOM for 21341 samples with 29-dimension features

	Cluster#1	**Cluster#2**	**Cluster#3**	**Cluster#4**
***k*-means**	1145	20159	35	2
SOM	18101	2066	1119	55

Table.14. Distance comparison between k-means and SOM for 21341 samples with 29-dimension features

	The average distance of cluster centroid with samples (k-means)	The average distance of cluster centroid with samples (SOM)	The average distance of different cluster centroid (k-means)	The average distance of different cluster centroid (SOM)
Clsuter#1	0.028	0.023	0.51	1.01
Clsuter#2	0.01	0.015	0.42	1
Clsuter#3	0	0.049	0.01	1.02
Clsuter#4	1.1	0.023	0.82	2.78

Table.15. Performance comparison between k-means and SOM for 21341 samples with 29-dimension features

	CPU Occupation	Simulation time (s)
***k*-means**	34%	201
SOM	78%	1618

6. Conclusions

This article discusses the performance evaluation of real LTE networks using unsupervised learning techniques. Since the data set to be applied to the clustering algorithm depends on the set of defined targets, it can be easily pointed out that the definition of these targets is the key aspect of the proposed method. Therefore, the target of each KPI should be specified by the mobile network operator according to the performance level required by the network.

Regarding the clustering results using SOM and K-means, the optimal number of clusters given by the elbow method is mainly four, one of the clusters mainly contains the best performing cells, and the other mainly consists of the worst-performing cells, the last two are mainly composed of cells with moderate performance. Besides, different dimensions of data are used in the simulation of this article. In low-dimensional data clustering, compared with the results obtained using K-means, the results of clustering using SOM are not significantly different. Therefore, considering the ease of adjusting the input parameters of the K-means algorithm, this method is considered the best method of the two methods. In comparison to other high dimensional data set, LTE cell data has many KPIs and they are also acting as vectors in the high dimensional feature space, the results of clustering using SOM are significantly different from K-means and it gets better performance in cells clustering, but it cost much more time and CPU occupation when working at the simulation. All in all, using big data analytics and mining in network optimization also means that network performance tuning can now be done on a highly scalable method and our method successfully leverages big data to identify regions of interest in almost real-time, which could provide business value in terms of reducing operational expenditure for a cellular network operator.

Author Contributions: All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the program of China Scholarships Council (No. 201808390034).

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Bengio, Y.; Lodi, A.; Prouvost, A.; Machine learning for combinatorial optimization: A methodological tour d'horizon. arXiv:1811.06128. Available Online: https://arxiv.org/abs/1811.06128 (accessed on 12 April 2020)

[2] Zappone, A.; Di Renzo, M.; Debbah, M.; Lam, T. T.; Qian, X.; Modelaided wireless artificial intelligence: Embedding expert knowledge in deep neural networks towards wireless systems optimization, arXiv:1808.01672. Available Online: https://arxiv.org/abs/1808.01672 (accessed on 19 February 2020)
[3] Rodríguez León C.; García Lorenzo MM.; Adecuación a metodología de minería de datos para aplicar a problemas no supervisados tipo atributo-valor [online], Universidad y Sociedad, 2016,8, pp. 43–53.
[4] Nicolas Andres M.R.; Bayron Alexis C. E.; Lilia Edith A. P.; Comparison between K-means and Self-Organizing Maps algorithms used for diagnosis spinal column patients. Info. in Medi. Unlocked, 2019,16, pp. 529-551.
[5] Kumar, Y.; Farooq, H.; Imran, A.; Fault prediction and reliability analysis in a real cellular network. In Proceedings of the Wireless Communications and Mobile Computing Conference (IWCMC), Valencia, Spain, 26-30 June 2017
[6] Jessica, M.; Furqan A.; Mario G.L.; Jarno N.; Big data-driven automated anomaly detection and performance forecasting in mobile networks.in Proceedings of IEEE Globecomm 2020 workshop on AI-Enabled 5G/6G Networks: Automation, Openness, and Radio Access, Taipei, Taiwan. 7-11 December. 2020
[7] Gomez-Andrades, A.; Munoz, P.; Serrano, I.; Barco, R.; Automatic root cause analysis for LTE networks based on unsupervised techniques, IEEE Trans. Vehi. Tech., 2016, 65, pp. 2369–2386,
[8] Carolina, G.; Matías, T.; Salvador L.R.; Juan L. B.L.; María Luisa M.A.; Estimating pole capacity from radio network performance statistics by supervised learning," IEEE Trans netw. serv. manag., 2020, 17, 2090–2101.
[9] Ahsanul Kabir, Md.; Xiao L.; Unsupervised learning for network flow based anomaly detection in the era of deep learning, In Proceedings of 2020 IEEE 6th International Conference on Big Data Computing Service and Applications (Big Data Service). Oxford, UK, 3-6 August 2020
[10] Ma, J.; Lin, S.; Big data enabled anomaly user detection in mobile wireless networks, In Proceedings of 2019 IEEE 5th International Conference on Computer and Communications (ICCC), Chengdu, China, 6-9 December 2019.
[11] Y Chen et all.; The comparison of SOM and K-means for text clustering, Compu. Infor., 2010,3,1-7.
[12] Singh, G.; Kaur, A.; Comparative analysis of k-means and Kohonen-SOM data mining algorithms based on behaviors in sharing information on facebook, Intern. Jour Engi. Comp. Sci., 2017, 6, 20990-20993.
[13] Lavneet S., Savleen S.; Parminder K.D.; Applications of clustering algorithms and Self Organizing Maps as data mining and business intelligence tools on real world data sets," In Proceedings of2010 International Conference on Methods and Models in Computer Science (ICM2CS-2010), New Delhi, India, December. 13-14, 2010.
[14] Parwez M. S.; Rawat D. B.; Garuba, M.; Big data analytics for user-activity analysis and User-Anomaly Detection in Mobile Wireless Network," IEEE Trans. Ind. Info., 2017,13, 2058–2065.
[15] Santos, R.; Sousa, M.; Vieira, P.; Queluz M. P., Rodrigues, A.; An unsupervised learning approach for performance and configuration optimization of 4G networks. In Proceedings of 2019 IEEE Wireless Communications and Networking Conference (WCNC), Marrakesh, Morocco, 15-18 April 2019.
[16] Wang Y.K.; Cheng R.H.; Zheng, P..; Che Y.; Spatial differentiation of water quality in river networks in Shanghai and its response to land use in riparian zones, Jour. Ecol. Rur. Envi., 2019, 35,925-932.
[17] Slobodan P.; A comparison between the Silhouette Index and the Davies-Bouldin Index in labeling IDS clusters, Compu. Scie., 2006,14, 1-12.
[18] Cai, R.R.; Zhang, H.W.; Bu, H.L.; Zhang, Y., Research on variation of runoff and sediment load based on the combination patterns in the middle and lower yellow river (Chinese), Shuili Xuebao, 2019,5, 732-742.
[19] Leonardo E. B. S.; Donald C. W.II.; An information-theoretic-cluster visualization for Self-Organizing Maps, IEEE Trans. Neur. Netw. learning syst., 2018,29, 161–169.
[20] Zhou H., Li G.M.; Zhang G.Y.; SOM + K-means two-phase clustering algorithm and its application, Mode. Electr. Techno, 2010,33, 113-116.
[21] Orkphol, K.; Yang, W.; Sentiment analysis on microblogging with K-means clustering and artificial bee colony Intern. Jour. Compu. Intelli. Applic., 2019,18,1950017.
[22] Shi, Z.Z.; Knowledge discovery (Chinese), 2nd ed. Tsinghua University: Beijing, China, 2011; vol.3 pp.111-122.
[23] Dr. G.S.; A. K.; Comparative analysis of K-means and Kohonen-SOM data mining algorithms based on student behaviors in sharing information on facebook, Intern. Jour. Engi. Compu. Scie., 2017, 6, Apr.,20990-20993.
[24] Chiu, P.; Reunanen, J.; Luostari, R.; Holma H.; Big Data Analytics for 4.9G and 5G Mobile Network Optimization, In Proceedings of the IEEE 85nd Vehicular Technology Conference (VTC Spring), Sydney, NSW, 4-7, June 2017.
[25] 3GPP TS 36.101, Evolved Universal Terrestrial Radio Access (EUTRA)User Equipment (UE) radio transmission and reception, V10.9.0, 2013-02.
[26] Yanchi L.; Zhongmou L.; Hui X.; Xuedong G.; Junjie W.; Understanding of internal clustering validation measures, In Proceedings of at the 2010 IEEE International Conference on Data Mining, Sydney, NSW, Australia, 13-17 December 2010

Figures

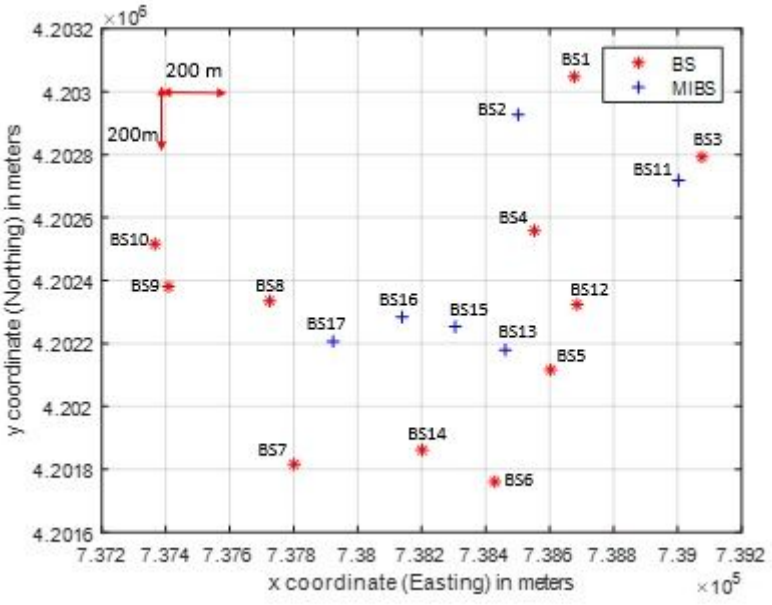

Figure 1

Base stations distribution of LTE.

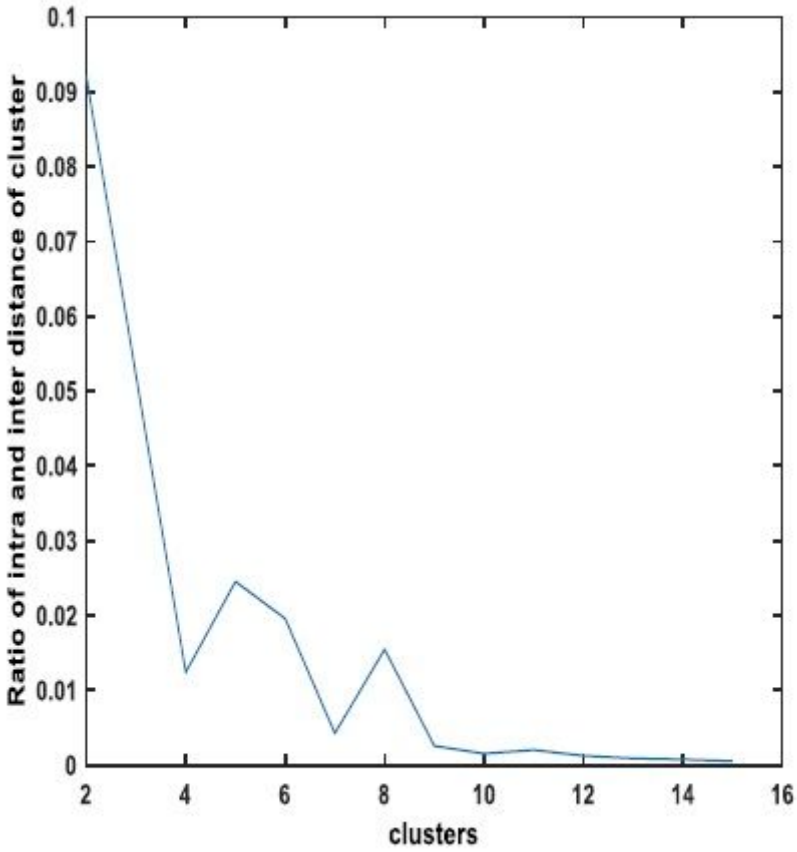

Figure 2

Optimal number of clusters fitting values of clusters

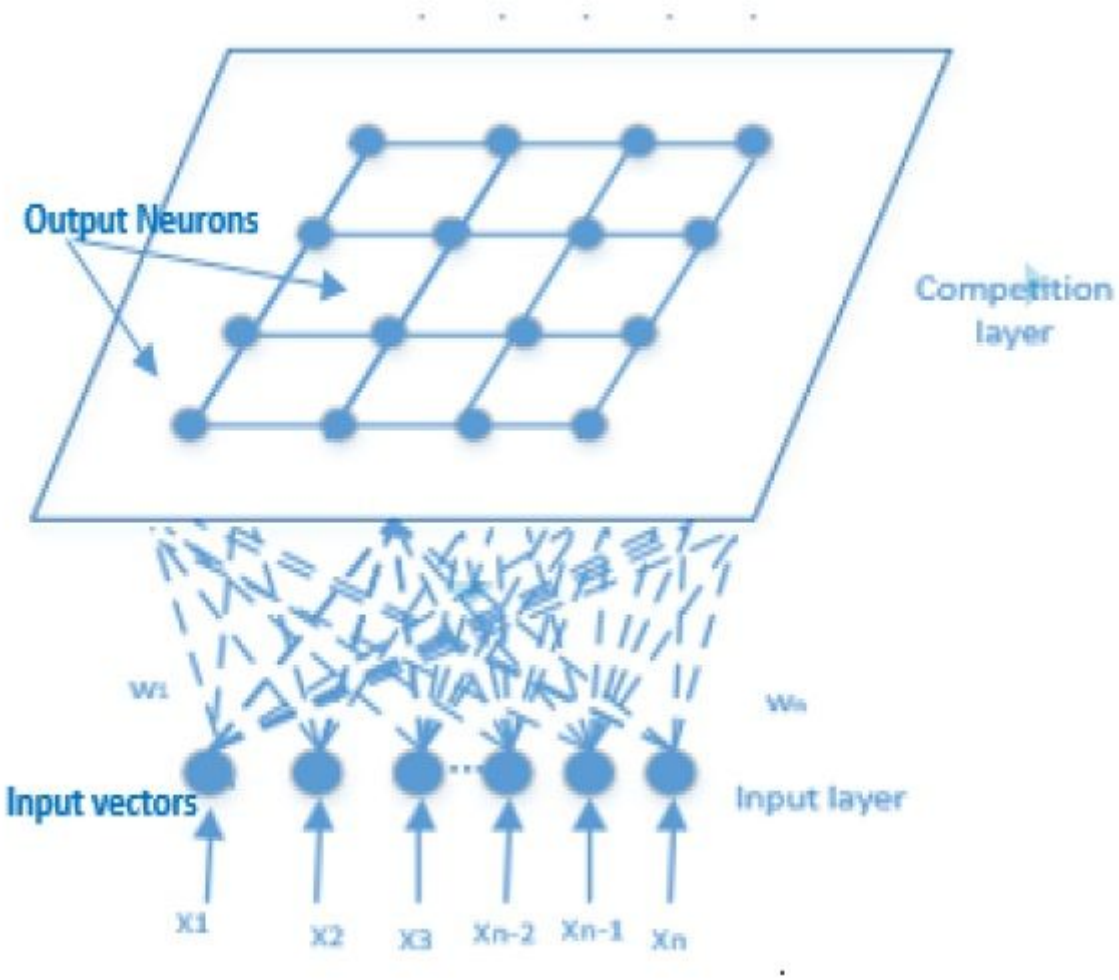

Figure 3

Network topology of SOM

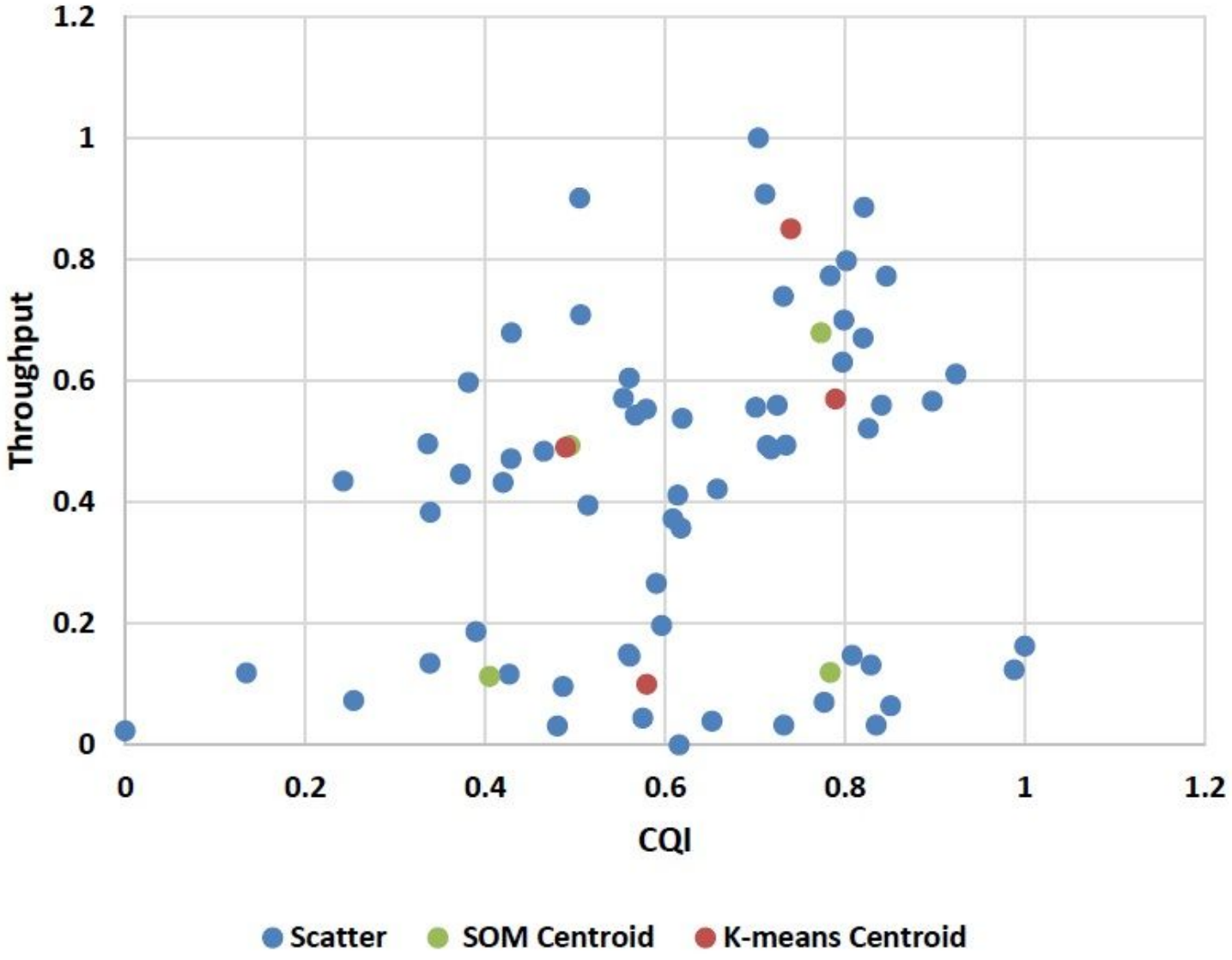

Figure 4

Scatter of Throughput and CQI by using SOM and k-means cluster

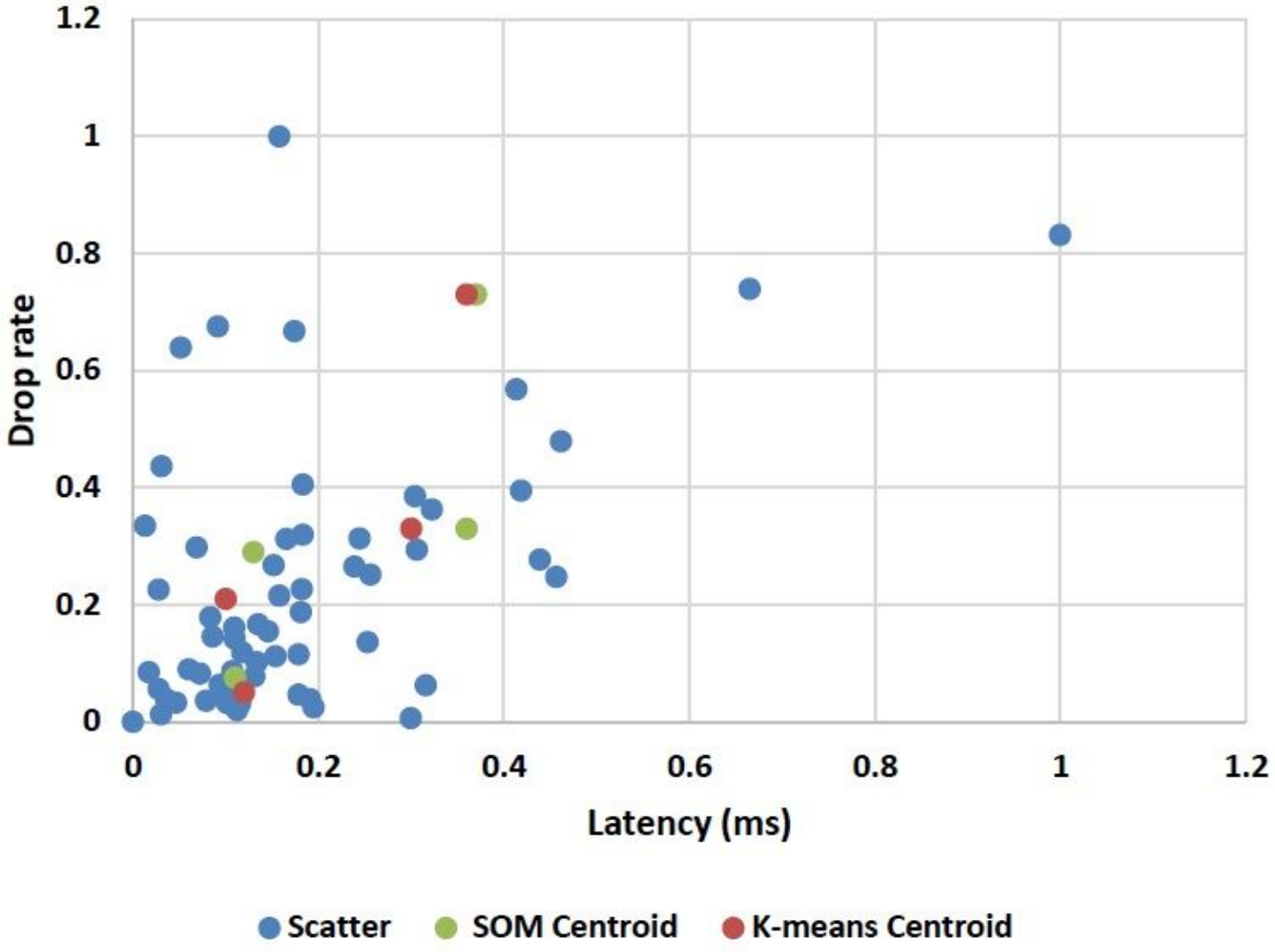

Figure 5

Scatter of Latency and Drop rate by using SOM and K-means cluster

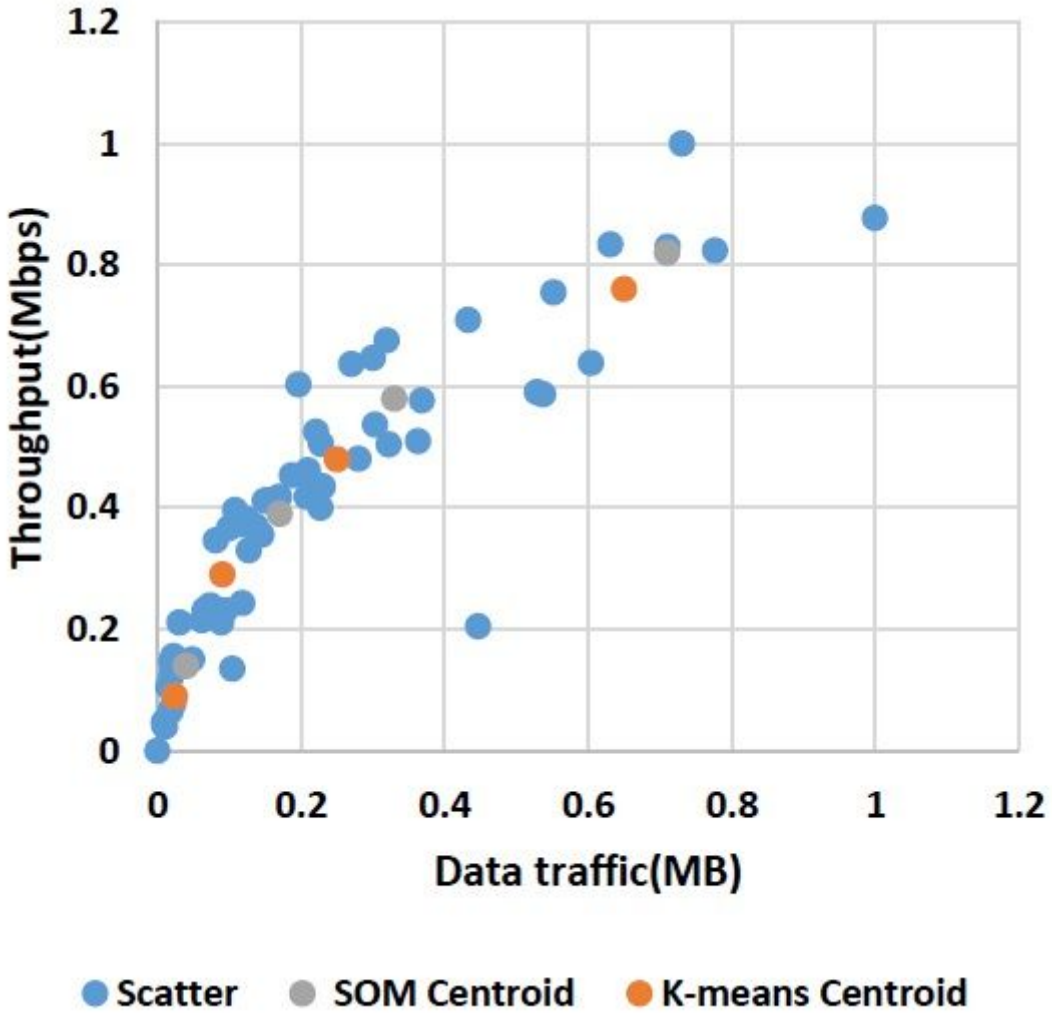

Figure 6

Scatter of data traffic and throughput by using SOM and k-means cluster

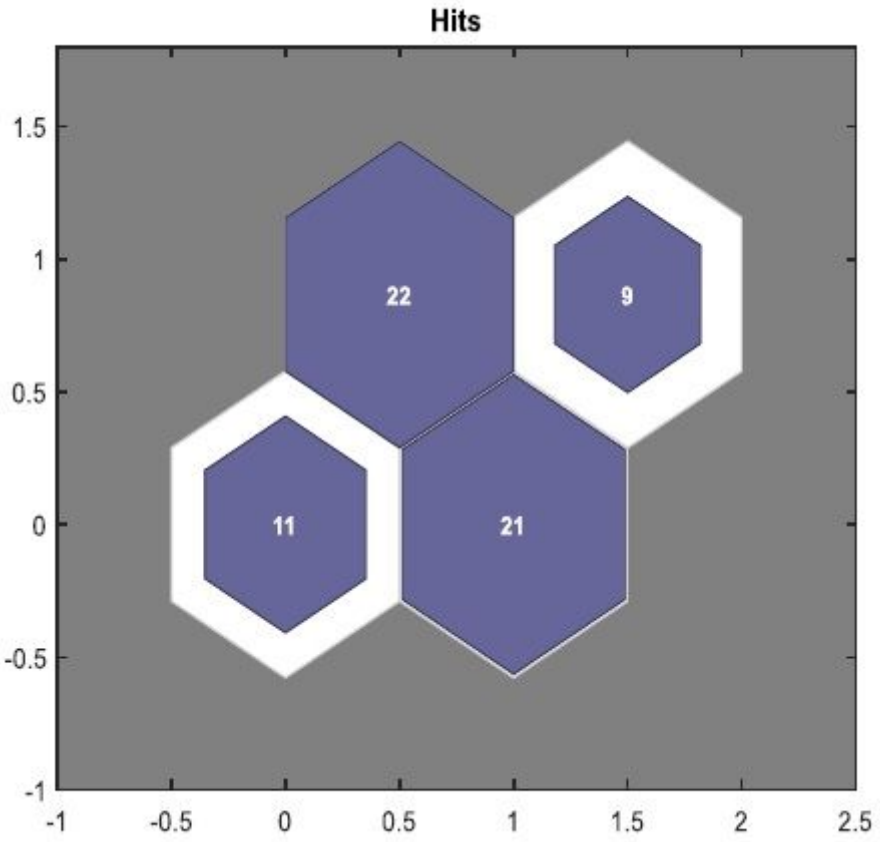

Figure 7

Hits of SOM for 63 samples with 29 features

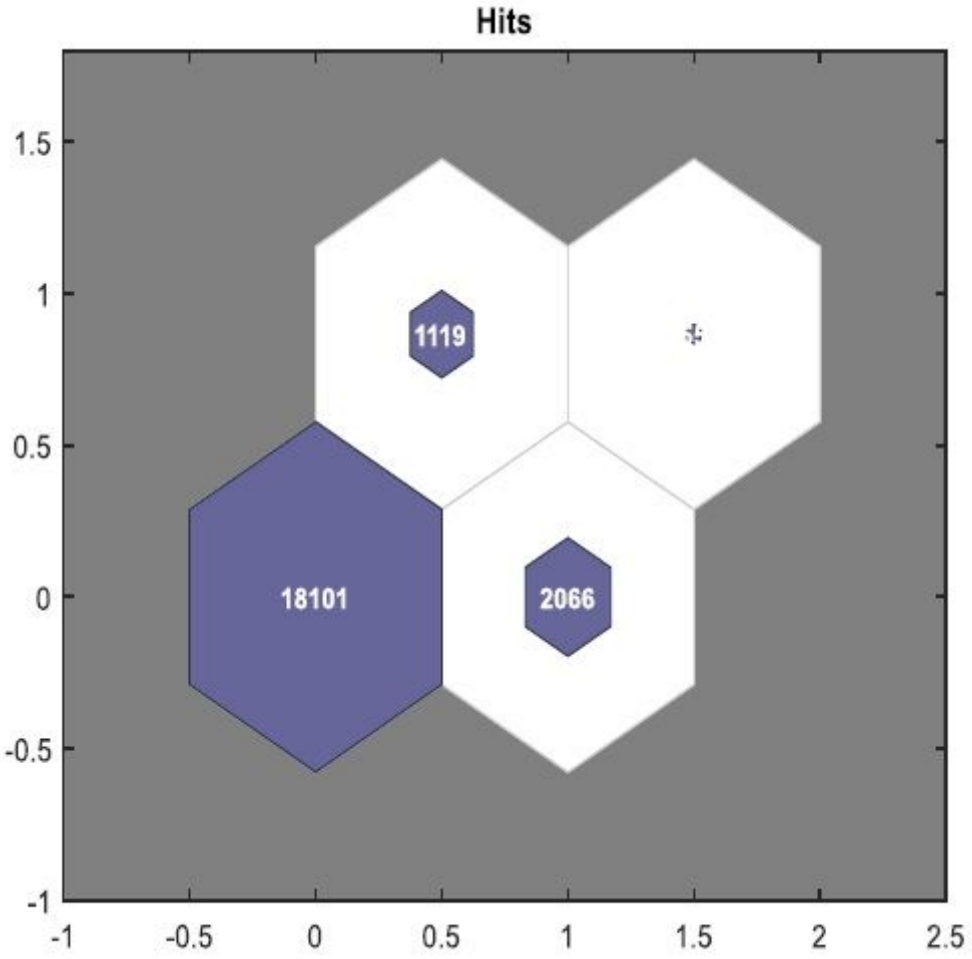

Figure 8

Hits distribution of SOM for 21341 samples with 29 features