Global Journal of
Information Technology:
Emerging Technologies

Volume 10, Issue 2, (2020) 106-112

www.gjit.eu

Comparative analysis of clustering techniques in the internet of things

Ahmed Burhan Mohammed*, Department of Energy, University of Kirkuk, Kirkuk 36001, Iraq

Suggested Citation:
Mohammed, A. B. (2020). Comparative analysis of clustering techniques in the Internet of Things. *Global Journal of Information Technology: Emerging Technologies. 10*(2), 106–112. https://doi.org/10.18844/gjit.v10i2.4746

Received June 13, 2020; revised August 08, 2020; accepted October 17, 2020.
Selection and peer review under responsibility of Assoc. Dr. Devkan Kaleci, Inonu University, Turkey.

Abstract

One of the most important topics in the last decade is the Big Data (BD) and how to link it and benefit from its consumption in different fields, included as the introduction in this research analysis of the BD belonging to devices of the Internet of Things. The concept of managing objects and exploring devices is connected to the Internet and sensors deployed in the world, all these devices are pumping a lot of data through the Internet of Things (IoT) into the world. In order to make the right decisions for people and things, BD using data mining techniques and machine language algorithms help make decisions. The Internet of Things that insert large amounts of data need to be studied, analysed and disseminated in order to access valuable, useful and bug-free information for the purpose of making the right decision and avoiding problems. In this paper, two clustering algorithms simple K-means and self-organising map (SOM) in IoT are presented. Next, comparing the clustering models' output in the IoT data set that improved the SOM is better than K-means, but it is slower in creating the model.

Keywords: Internet of things (IoT), big data, machine learning, filtered cluster, K-means, SOM.

* ADDRESS FOR CORRESPONDENCE: **Ahmed Burhan Mohammed**, Department of Energy, University of Kirkuk, Kirkuk 36001, Iraq.
E-mail address: ahmedlogic79@yahoo.com

1. Introduction

In the recent decades, Big Data (BD) and Internet of Things (IoT) have been one of the major interesting research subjects due to the implementations getting more attention [1]. Academics are motivated by the unindustrialised BD analytics solutions consuming machine learning (ML) models [2]. A lot of the developments in the field of ML are due to its ability to extract hidden features and patterns even in highly complicated data sets [3].

The world's attention has been directed to the work and development of Internet of Things (IoT), and widely towards the research institutes and scientific research workers in smart cities. The term IoT consists of two main words, the Internet and the Things [4]. The Internet includes a set of connected devices on the Internet, which operate within the Internet, but things include all types of computers and smart devices and devices used in our daily lives, which rely mainly on the Internet [5]. With the world moving towards Industry 4.0 [3], IoT has expanded to notable locations in all fields. Essentially, IoT tolerates the connection between people and things anytime and anywhere through devices that can transmit data with anything over any network. The two most important technical issues in energy are efficiency and scalability that must be completely addressed in order to construct high-performance IoT systems [4].

As the volume of data collected by various IoT stations increases, a huge challenge for IoT application is BD management and analytics [6]. Although BD can potentially benefit from data compression techniques, the chances are that compression will reduce an insignificant amount of data such that it would not be worth the effort [5].

1.1. Big data (BD)

Michael Cox and David Ellsworth in 1997 were the first to use the term 'BD'. BD is data that are collected from Internet communication, mobile devices, social networking, video sharing, sensors and smart devices from IoT, etc. [7]. BD consists of extensive data sets primarily in the characteristics of volume, variety, velocity and/or variability for the analysis, manipulation and efficient storage that are scalable architecture requirements [8].

Four V's Physiognomies of BD research in 2001 introduced the three V's data management perception.

Volume, velocity and variety are known as three V's. Next, IBM added one more V, veracity, to the three V's.

Essential characteristics of the four V's (variety, velocity, volume and veracity) are defined below.

2. Related work

Marjani et al. [9] mentioned the significant relationship between BD and IoT. Big IoT data analytics enable minor data and scientists to analyse huge amounts of unstructured data that can be harnessed using traditional tools and data mining techniques that help in making predictions, identifying recent trends, finding hidden information and making decisions.

Yerpude and Singhal [10] showed the relationship between BD and business analytics. The smart environment was evolved by transmitting the data onto the smart network of Internet of Thing (IoT). For getting the right decision, the decision-making model was used on the data gathered from (IoT) devices by the business analytic. It can be concluded that the data analytic in a business field gives the right decision at the right time. Moreover, it is the successful key in business.

Borthakur et al. [11] presented a comprehensive review of employing K-means clustering algorithm on the clinical speech data. For analysing the data collected by smart devices, K-means quantitative clustering algorithm was used, which is based on the foggy architecture of smart devices. However, it proved the ability of large data to work in the analysis of data from smart devices.

Alam et al. [12] studied the effects and ability of eight data mining algorithms for IoT data. Archives C4.5 and C5.0 have better accuracy, but only with high memory-efficient processes. Finally, ANN and DLANN show the highest accuracy by modelling high-level data abstraction, but they are computationally expensive.

Meidan et al. [13] investigated the differential impacts on the ML algorithms applied for IoT device data to detect unauthorised devices. Random Forecast was applied to extract the feature of network traffic data set. White List was identified to aim for accurate IoT devices. Multi-class classifier examined each type. The perfect classification of White List archive is the best accuracy result.

Yerpude and Singhal [14] provided in-depth analysis of the work of IoT data statistical analysis that impact the IoT data on demand forecasting. The data were collected from smart tools and devices. Various forecasting models were examined. Furthermore, the accuracy of the forecasting model was verified by the result that experimented to get the error value and relevance.

Thangaraju et al. [15] discussed the challenges and strategies of comparative analysis of two clustering algorithms applied. Besides, two data sets are used in this work taken from UCL data set repositories. Experiment results achieved by the K-means algorithm have a better accuracy among all cases.

3. Methodology

3.1. Filtered cluster

New significant attributes were added by the filter, which characterised the clusters specified to each instance by a quantified clustering algorithm [16]. Either the clustering algorithm is constructed with the initial batch of data or references are sequential clustered as a model file to use, instead.

Mathematically, the filter is exceptional subset of a partially well-arranged set. When X is a topological area and x a node of X. F is a filter applied on X called cluster if and only if every quantity of F has a non-empty intersection with each neighbourhood of x [17]. A filter base F that has x as a cluster point may not converge into x. The limit inferior of F is the infimum of all the cluster nodes of F. The limit superior to F is the supremum of all cluster nodes of F. The filter F is convergent when its limit is inferior or low and its limit is superior or high. [18]

3.2. Simple K-means

K-means algorithm is a type of clustering that is used for investigative data analysis of unidentified data [19]. K-means is a method of vector quantisation and is quite significantly used in data mining. The aim of this algorithm is to find groups in the data and the number of collections is represented by the variable K. The algorithm works to allocate each data point to one of the K sets based on the provided features. This algorithm aims to minimise the squared error function J [20].

3.3. Self-organised map (SOM)

The SOM algorithm is one of the algorithms that depends on the neural connections between nodes in a two-dimensional scheme [21]. This contract is related to its neighbours according to certain topologies that help in the process of interdependence, and there are two types of rectangular and hexagonal ideologies [22].

4. Experiment and results

This section discusses the steps of the work and the experiment to get the result.

4.1. Steps of the work

Downloads the data set from the link https://www.kaggle.com/pitasr/industrialiot [23]. The data were collected by applying Industrial Demand/Response by Internet of Things. Data are for facility energy management systems, which can be used for academic purpose. It contains 16,382 instances, (which is categorised in two parts: train 11,467 and test 4,914) and includes seven attributes as follows: DEMAND_RESPONSE {Numeric}, Area {Numeric}, Season {Numeric}, Energy {Numeric}, Cost {Numeric}, pair no {Numeric} and Distance {Numeric}.

4.2. Experimental results

4.2.1. Creating and applying clustering model

The clustering results for creating models from training instances over simple K-means and SOM algorithms are specified in Table 1.

Table 1. Clustering result for creating model

	Clustering output creating model on train data set			
	K-means		SOM	
Cluster ID.	No. of instances clustered	Percentage of clustering	No. of instances clustered	Percentage of clustering
1	3,796	33%	3,325	29%
2	3,307	29%	3,766	33%
3	1,123	10%	3,267	28%
4	3,241	28%	1,109	10%

It appears from the aforementioned investigations in Table 1 that the creating model highlights the highest number of clustering instances in cluster 1 with K-means, which has the highest percentage (32%). But the highest number of clustering with SOM is cluster 2, when presenting the results from Table 1. On the other hand, the lowest number of clustering instances is in cluster 3 with the lowest percentages. But the lowest clustering number of clustering is cluster 4, depending on the output result of SOM.

The incoming result of applying the cluster model on the test data set instances over simple K-means and SOM algorithms is quantified in Table 2.

Table 2. Clustering result for applying model

	Clustering output applying model on the test data set			
	K-Means		SOM	
Cluster ID	No. of instances clustered	Percentage of clustering	No. of instances clustered	Percentage ofclustering
1	3,796	33%	3,325	29%
2	3,307	29%	3,766	33%
3	1,123	10%	3,267	28%
4	3,241	28%	1,109	10%

Table 2 shows that the output clustering result for SOM algorithm has the highest percentage (33%) in cluster 2, but in cluster 1 the highest percentage (33%) was for the K-means algorithm. Additionally, the instances were clustered in the lowest percentage (10%) in cluster 3 by K-means algorithm, but cluster 3 returned the lowest personage (10%) by SOM. After conducting the clustering and testing of the samples on the data and obtaining the clustered nodes according to the clustering algorithms, the speed of implementation of the model and the error rate were found, as shown in Table 3.

Table 3. The incorrectly clustering instances number

	Incorrectly clustering instance with time		
	Incorrectly clustered instances		Time taken to a build model
Algorithm	No. of instances	Percentage	Time in Seconds
K-Means	100.0	0.8721	0.09s
SOM	73.0	0.6366	41.64s

5. Conclusion

The objective of the present work was to investigate the comparative analysis between two clustering algorithms in IoT. The improvement of the SOM algorithm gives better clustering result from K-means. Moreover, the survey and experimental application of algorithms were carried out.

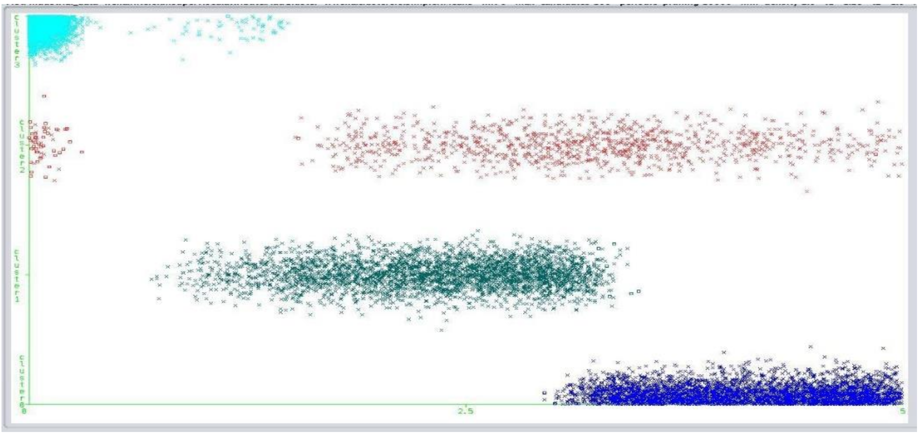

Figure 2. Clustering output for K-means

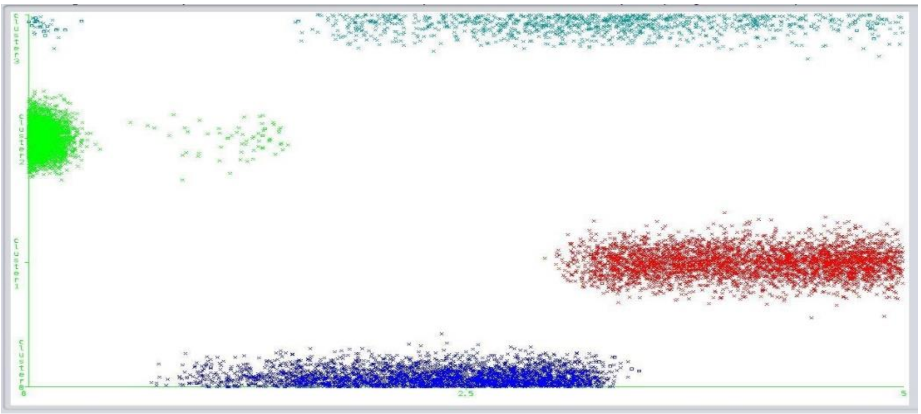

Figure 3. Clustering output for SOM

According to Figure 2, the appearance of some of the instances being shifted to other clusters, when applying the model, can be seen. Based on BD and data analysis using algorithms, BD was found to have good results in analysing and clustering data of IoT. The results propose that K-Means has 100 incorrect clustering instances, but SOM has a minimum number of incorrectly clustering instances. Furthermore, SOM is better than K-Means in account of incorrectly, but SOM needs more time to create the model. So, the K-means is faster than SOM. Finally, SOM achieved a better clustering algorithm.

References

[1] A. Moon, J. Kim, J. Zhang, H. Liu, and S. Woo Son, “Understanding the impact of lossy compressions on IoT smart farm analytics,” in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 4602–4611.

[2] M. Banerjee, J. Lee, and K.-K. R. Choo, “A blockchain future for internet of things security: A position paper,” *Digit. Commun. Netw.*, vol. 4, no. 3, pp. 149-160, Aug. 2018.

[3] P. P. Ray, “A survey on Internet of Things architectures,” *J. King Saud Univ.—Comput. Inf. Sci.*, vol. 30, no. 3, pp. 291–319, 2016.

[4] J. Huang and K. Hua, “Managing the Internet of Things,” Inst. Eng. Technol., London, U.K., Tech. Rep., 2017.

[5] H. Geng, *Internet of Things and Data Analytics Handbook Edited*. Hoboken, NJ, USA: Wiley, 2017.

[6] L. Xiao, X. Wan, X. Lu, Y. Zhang, and D. Wu, “IoT security techniques based on machine learning,” 2018, *arXiv:1801.06275*. [Online]. Available: http://arxiv.org/abs/1801.06275

[7] S. Bin, L. Yuan, and W. Xiaoyi, “Research on data mining models for the Internet of Things,” in *Proc. Int. Conf. Image Anal. Signal Process. (IASP)*, Apr. 2010, pp. 127–132.

[8] F. Chen, P. Deng, J. Wan, D. Zhang, A. V. Vasilakos, and X. Rong, “Data mining for the Internet of Things: Literature review and challenges,” *Int. J. Distrib. Sensor Netw.*, vol. 11, no. 8, 2015, Art. no. 431047.

[9] M. Marjani *et al.*, “Big IoT data analytics: Architecture, opportunities, and open research challenges,” *IEEE Access*, vol. 5, pp. 5247–5261, Mar. 2017.

[10] S. Yerpude and T. K. Singha, “Internet of Things and its impact on business analytics,” *Indian J. Sci. Technol.*, vol. 10, no. 5, pp. 1–6, Feb. 2017.

[11] D. Borthakur, H. Dubey, N. Constant, L. Mahler, and K. Mankodiya, “Smart fog: Fog computing framework for unsupervised clustering analytics in wearable Internet of things,” 2017, *arXiv:1712.09347*. [Online]. Available: http://arxiv.org/abs/1712.09347

[12] F. Alam, R. Mehmoodb, I. Katiba, and A. Albeshria, “Analysis of eight data mining algorithms for smarter Internet of Things (IoT),” in *Proc. Int. Workshop Data Mining IoT Syst. (DaMIS)*, 2016, pp. 437–442.

[13] Y. Meidan *et al.*, “Detection of unauthorized IoT devices using machine learning techniques,” 2017, *arXiv:1709.04647*. [Online]. Available: http://arxiv.org/abs/1709.04647

[14] S. Yerpude and T. K. Singha, “Internet of Things and its impact on business analytics,” *Indian J. Sci. Technol.*, vol. 10, no. 5, pp. 1–6, Feb. 2017.

[15] G. Thangaraju, J. Umarani, and V. Poongodi, “Comparative study of clustering algorithms: Filtered clustering and K-means cluttering algorithm using WEKA,” *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 5, no. 9, Sep. 2017.

[16] M. Wei, S. H. Hong, and M. Alam, “An IoT-based energy-management platform for industrial facilities,” *Appl. Energy*, vol. 164, pp. 607–619, Feb. 2016.

[17] J. Hahn, “Information & Environment: IoT-powered recommender systems,” Simons Found., Cornell Univ. Library (CUL), Ithaca, NY, USA, Tech. Rep., 2018.

[18] Khoda, “A survey on various techniques in Internet of Things (IoT) implementation: A comparative study,” *Int. J. Future Revolution Comput. Sci. Commun. Eng.*, vol. 3, no. 11, Nov. 2017.

[19] F. Chen, “Data mining for the Internet of Things: Literature review and challenges,” *Distrib. Sensor Netw.*, vol. 11, no. 8, Aug. 2015, Art. no. 431047.

[20] C. Neureiter, M. Uslar, D. Engel, and G. Lastro, “A standards-based approach for domain specific modelling of smart grid system architectures,” in *Proc. 11th Int. Conf. Syst. Eng. (SoSE)*, Kongsberg, Norway, Jun. 2016, pp. 1–6.

[21] V. Chaudhary, R. S. Bhatia, and A. K. Ahlawat, “A novel self-organizing map (SOM) learning algorithm with nearest and farthest neurons,” *Alexandria Eng. J.*, vol. 53, no. 4, pp. 827–831, Dec. 2014.

[22] Y. Onuki *et al.*, “A comparative study of disintegration actions of various disintegrants using Kohonen's self-organizing maps,” *J. Drug Del. Sci. Technol.*, vol. 43, pp. 141–148, Feb. 2018.

[23] *Industrial Internet of Things Data Demand/Response (DR) data for IoT Analytics*. Accessed: 2016. [Online]. Available: https://www.kaggle.com/pitasr/industrialiot