Journal of Engineering and Applied Sciences 13 (Special Issue 13): 10802-10806, 2018
ISSN: 1816-949X

Comparative Analysis of Machine Learning Techniques, Clustering Algorithms Impact in Internet of Things

[1]Ahmed Burhan Mohammed, [2]Ahmad Abdullah Mohammed AL-Mafrji
[1]College of Arts, University of Kirkuk, Kirkuk, Iraq
[2]College of Education for Pure Sciences, University of Kirkuk, Kirkuk, Iraq

Abstract: The sensors scattered around the world and the precision devices that work on the internet have formed the so-called internet of things. As these devices expose large amounts of information to the central data storage for decision. Taking the right decision in real-time requires analyzing these data for the purpose of getting the right thing done. In order to make the right decisions on people and things using data mining techniques and machine learning algorithms helps make decisions. Internet of things that inject large amounts of data needs to be studied, analyzed 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 study, presents two clustering algorithm simple k-means and Self Organizing Map (SOM) in industrial data used IoT devices. Next, comparing the clustering models of 2 algorithms output in IoT dataset that improved the SOM is better than k-means but it is slower in creating the model.

Key words: Internet of Things (IoT), big data, machine learning, k-means, SOM, self organizing map

INTRODUCTION

In the recent decades, Big Data (BD) and Internet of Things (IoT) has been one of the major interesting research subject's due implementations started getting more attention (Moon *et al*., 2017). Academics motivated on unindustrialized big data analytics solutions consuming machine learning models. A lot off developments in the filed of Machine Learning (ML) due to its ability to extract hidden features and patterns even in highly complicated datasets (Banerjee *et al*., 2017).

The world's attention has been directed to the research and development of Internet of Things (IoT) and widely by the research institutes and scientific research workers in smart cities. The term IoT consists of two main words, the internet and the things (Ray, 2016). As 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 on the internet mainly.

With the world moving towards Industry 4.0, IoT has expanded a noticeable location in all fields. Essentially, IoT tolerates the connection between people and things in anytime anywhere through devices which can transmit data with anything over any network. The most two important technical issues in energy are efficiency and scalability that must be completely addressed in order to construct high-performance IoT systems (Ray, 2016).

As the volume of data collected by various IoT stations increases, a huge challenge for IoT application is big data management and analytics. Although, big data can potentially benefit from data compression techniques, the chances are that compression will reduce an insignificant amount of data such that it would not worth the effort (Xiao *et al*., 2018).

Big data: Michael Cox and David Ellsworth in 1997 was named the term first "Big Data". Big data is the data that collects from internet communication mobile devices, social networking, video sharing, sensors and smart devices from IoT and etc. Big data consists of extensive data sets collection primarily in the characteristics of volume, variety, velocity or variability for the analysis, manipulation and efficient storage these are scalable architecture requirement (Chen *et al*., 2015).

The 4 Vs physiognomies of big data, research in 2001 introduced the 3 Vs data management perception. Volume, Velocity and Variety are known as 3 Vs. Next, IBM added one more V-Veracity, to the 3 Vs (Chen *et al*., 2015). Essential characteristics of 4 Vs (Variety, Velocity, Volume and Veracity) are pronounced in the following.

Literature review: Marjani *et al*. (2017) mention the significant relationship between big data and IoT. Big IoT

Corresponding Author: Ahmed Burhan Mohammed, College of Arts, University of Kirkuk, Kirkuk, Iraq

data analytics enables data minor and scientists to analyze huge amounts of unstructured data that can be harnessed using traditional tools. Used data mining techniques that help in making predication identifying, recent trends, finding hidden information and making decision (Marjani *et al*., 2017).

Yerpude and Singhal (2017) represent the relationship between big data and business analytic. The smart environment evolved which consist of transmitting the data onto the smart network of Internet of Thing (IoT). For getting the right decision, decision making model used on the data gathered from (IoT) devices by the business analytic. Conclude that the data analytic in a business field gives the right decision at the right time. Moreover, it is the successful key in business (Yerpude and Singhal, 2017).

Borthakur *et al*. (2017) present a comprehensive review of employing k-means clustering algorithm on clinical speech data. Analyzing the data collected by the smart devices, k-means quantitative clustering algorithm used which is based on the foggy architecture of the smart devices. That proved the ability of large data to research in the analysis of data smart devices (Borthakur *et al*., 2017).

Alam *et al*. (2016) studied effects and ability of eight data mining algorithms for IoT data. Archive that C4.5 and C5.0 has better accuracy but with high memory efficient and high process. Finally, ANN and DLANN show highest accuracy by modelling high-level data abstraction but are computationally expensive (Alam *et al*., 2016).

Meidan *et al*. (2017) investigate that as differential impact on the machine learning algorithms applying for IoT device data to detect the unauthorized devices. Random forecast applies to extract feature of network traffic data set. White list is identified to aim the accurately IoT devices. Multi-class classifier examined for each type. Perfect classification of white list archive as the best accuracy result (Meidan *et al*., 2017).

Thangaraju *et al*. (2017) discuss the challenges and strategies of comparative analysis of two clustering algorithms have been applied. Besides, 2 data sets are used in this study taken from UCL data set repositories. Experiment results archive the k-means algorithm have better accuracy among all cases (Thangaraju *et al*., 2017).

MATERIALS AND METHODS

Filtered cluster: New significant attribute added by filter that characterizes the clusters specified to each instance by quantified clustering algorithm (Wei *et al*., 2016). Either the clustering algorithm is constructed with the initially batch of data or references are sequential clustered model file to use instead.

Mathematically, 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 nonempty intersection with each neighbourhood of x. A filter base F that has x as a cluster point may not converge to x. The limit inferior of F is the infimum of all cluster nodes of F. The limit superior of F is the supremum of all cluster nodes of F. The filter F be convergent when its limit inferior is low and its limit superior is high (Hahn, 2018).

Simple k-means: k-means algorithm is a type of clustering that is used for investigative data analysis of unidentified data. k-means is a method of vector quantization and is quite significantly used in data mining (Khoda, 2017). The aim of this algorithm is to find groups in the data, the number of collections represented by the variable K. The algorithm works to allocate each data point to one of K sets based on the provided features. This algorithm aims to minimize the squared error function J (Neureiter *et al*., 2016).

Self organize map: The SOM algorithm is one of the algorithms that depend on the neural connections between nodes in a 2-dimensional scheme. This contract is related to its neighbors according to certain topologies that help in the process of interdependence and there are two types of rectangular and hexagonal ideologies (Onuki *et al*., 2018).

RESULTS AND DISCUSSION

In this part, discuses the steps of the study and the experimental to get the result.

Steps of work: Downloads data set from the link https://www.kaggle.com/pitasr/industrialiot (Anonymous, 2016), this data was collected by applying industrial Demand/Response (DR) by internet of things. Data is for facility energy management systems. Which can be used for academic purpose. It contains 16382 instances which splits in tow parts train 11467 and test 4914 and includes 7 attributes itemized below: Demand_Response {Numeric} area {Numeric}, season {Numeric}, energy {Numeric}, cost {Numeric}, pair no {Numeric} and distance {Numeric} (Fig. 1).

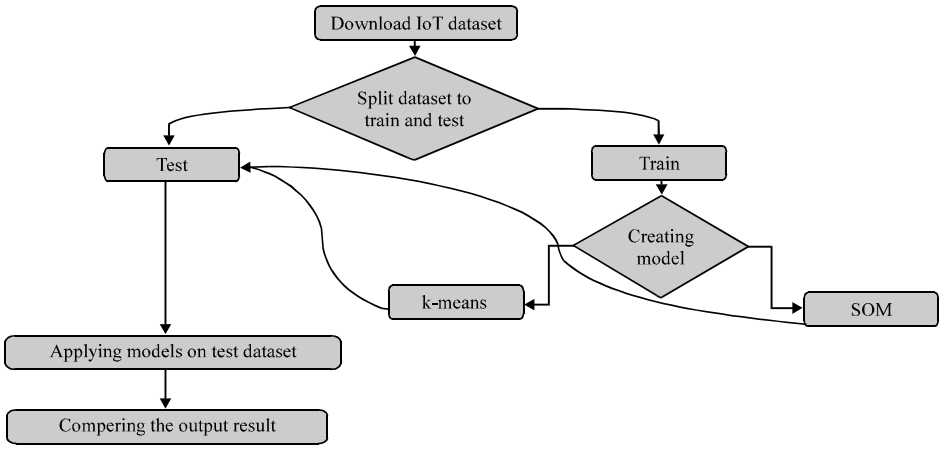

Fig. 1: Steps of work

Table 1: Clustering result for creating model

	Clustering output creating model by train data set			
	k-means		SOM	
Cluster ID	No. of instances clustered	Percentage of clustering (%)	No. of instances clustered	Percentage of clustering (%)
1	3796	33	3325	29
2	3307	29	3766	33
3	1123	10	3267	28
4	3241	28	1109	10

Table 2: Clustering result for applying model

	Clustering output applying model on test data set			
	k-means		SOM	
Cluster ID	No. of instances clustered	Percentage of clustering (%)	No. of instances clustered	Percentage of clustering (%)
1	3796	33	3325	29
2	3307	29	3766	33
3	1123	10	3267	28
4	3241	28	1109	10

Creating and applying clustering model: The returning determination of clustering creating model form training instances over simple k-means and self organizing map algorithms is specified by Table 1.

It appears from the aforementioned investigations that in Table 1. Creating model highlights the highest number of clustering instances are in cluster 1 with k-means and has the highest percentages 32%. But the highest number of clustering with SOM is cluster 2. when presenting to the results from Table 1. On the other hand, the lowest number of clustering instances are in cluster 3 with the lowest percentages. But, the lowest clustering number of clustering is cluster 4 depending the output result of SOM.

The incoming result of applying cluster model on test data set instances over simple k-means and self organizing map algorithms is quantified by Table 2 and Fig. 2.

Table 3: The incorrectly clustering instances number

	Incorrectly clustering instance with time		
	Incorrectly clustered instances	Time taken to build model	
Agorithm	No. of istances	Percentage	Time (sec)
k-means	100.0	0.8721	0.09
SOM	73.0	0.6366	41.64

According to Table 2 shows that the output clustering result for self-organize map algorithm has the highest percentage 33% in cluster 2 but the cluster 1 is returned the highest percentage 33% with k-means algorithm. Additionally, the instances were clustered in the lowest percentage 10% in cluster 3 by k-means algorithm but the 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 it was found that the speed of implementation of the model and the error rate is as shown in Table 3 and Fig. 3.

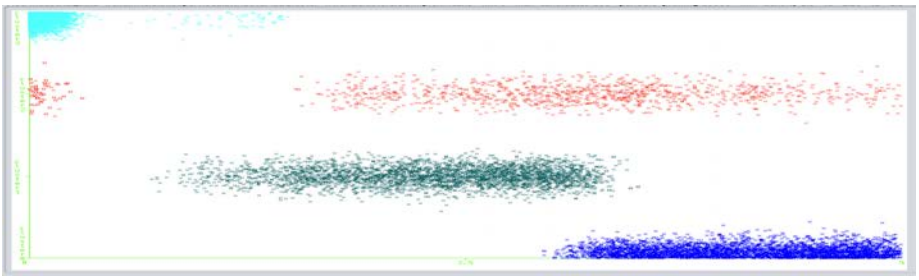

Fig. 2: Clustering output for k-means

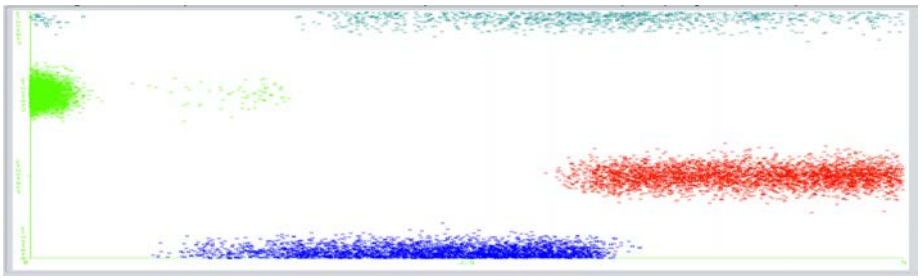

Fig. 3: Clustering output for SOM

CONCLUSION

The objective of the present study is to investigate the comparative analysis between two clustering algorithms in IoT. Improve that the SOM algorithm has better clustering result from k-means. Moreover, the survey experimental applying algorithms done.

Clearly, appear some of instances was shifted to other cluster when applying the model. The results as shown in study, propose that k-means have 100 incorrect clustering instances but the SOM have minimum number of incorrectly clustering instances. Furthermore, the SOM is better than k-means in account of incorrectly but the SOM need more time to create the model. So, the k-means is faster than SOM. Finally, SOM achieved as better clustering algorithm.

REFERENCES

Alam, F., R. Mehmood, I. Katib and A. Albeshri, 2016. Analysis of eight data mining algorithms for smarter Internet of Things (IoT). Procedia Comput. Sci., 98: 437-442.

Anonymous, 2016. Industrial internet of things data Demand/Response (DR) data for IoT analytics. Kaggle, San Francisco, California, USA.

Banerjee, M., J. Lee and K.K.R. Choo, 2017. A blockchain future for internet of things security: A position paper. Digital Commun. Networks, 4: 149-160.

Borthakur, D., H. Dubey, N. Constant, L. Mahler and K. Mankodiya, 2017. Smart fog: Fog computing framework for unsupervised clustering analytics in wearable internet of things. Proceedings of the 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), November 14-16, 2017, IEEE, Montreal, Canada, ISBN:978-1-5090-5991-1, pp: 472-476.

Chen, F., P. Deng, J. Wan, D. Zhang and A.V. Vasilakos *et al*., 2015. Data mining for the internet of things: Literature review and challenges. Intl. J. Distrib. Sens. Networks, 2015: 1-14.

Hahn, J., 2018. The bibliotelemetry of Information and environment: An E valuation of IoT-powered recommender systems. Digital Libraries, 1: 1-10.

Khoda, A., 2017. A survey on various techniques in internet of things (IoT) implementation: A comparative study. Intl. J. Future Revolution Comput. Sci. Commun. Eng., 3: 259-264.

Marjani, M., F. Nasaruddin, A. Gani, A. Karim and I.A.T. Hashem *et al*., 2017. Big IoT data analytics: Architecture, opportunities and open research challenges. IEEE. Access, 5: 5247-5261.

Meidan, Y., M. Bohadana, A. Shabtai, M. Ochoa and N.O. Tippenhauer *et al*., 2017. Detection of unauthorized iot devices using machine learning techniques. Cryptography Secur., 1: 1-13.

Moon, A., J. Kim, J. Zhang, H. Liu and S.W. Son, 2017. Understanding the impact of lossy compressions on IoT smart farm analytics. Proceedings of the 2017 IEEE International Conference on Big Data (Big Data), December, 11-14, 2017, IEEE, Boston, Massachusetts, ISBN:978-1-5386-2716-7, pp: 4602-4611.

Neureiter, C., M. Uslar, D. Engel and G. Lastro, 2016. A standards-based approach for domain specific modelling of smart grid system architectures. Proceedings of the 2016 11th Conference on System of Systems Engineering (SoSE), June 12-16, 2016, IEEE, Kongsberg, Norway, ISBN:978-1-4673-8728-6, pp: 1-6.

Onuki, Y., A. Kosugi, M. Hamaguchi, Y. Marumo and S. Kumada *et al*., 2018. A comparative study of disintegration actions of various disintegrants using Kohonen's self-organizing maps. J. Drug Delivery Sci. Technol., 43: 141-148.

Ray, P.P., 2016. A survey on internet of things architectures. J. King Saud Univ. Comput. Inf. Sci., 30: 219-319.

Thangaraju, G., J. Umarani and V. Poongodi, 2017. Comparative study of clustering algorithms: Filtered clustering and K-means cluttering algorithm using WEKA. Intl. J. Innovative Res. Comput. Commun. Eng., 5: 15115-15124.

Wei, M., S.H. Hong and M. Alam, 2016. An IoT-based energy-management platform for industrial facilities. Appl. Energy, 164: 607-619.

Xiao, L., X. Wan, X. Lu, Y. Zhang and D. Wu, 2018. IoT security techniques based on machine learning. Cryptography Secur., 1: 1-20.

Yerpude, S. and T.K. Singhal, 2017. Internet of things and its impact on business analytics. Indian J. Sci. Tech., 10: 1-6.