International Journal of Engineering & Technology, 7 (3.4) (2018) 47-51

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Comparative Analysis of Clustering Techniques in Cloud For Effective Load Balancing

Akankshya Aparajita[1], Shrabanee Swagatika[2], Debabrata Singh[3]

1,2Department of Computer Science & Engineering, ITER, SOA , Bhubaneswar, India
3Department of CSIT, ITER, SOA , Bhubaneswar, India
**Corresponding author E-mail: 1akankshyaaparajita29@gmail.com,*

Abstract:

Clustering is used as an important procedure in the process of data mining, where information of large datasets are transformed into meaningful and concise data. It performs activities like pattern representation, using of clustering algorithms and their validation, data abstraction and finally result generated. Clustering has many categories of algorithms such as partition-based, hierarchical-based, density-based, grid-based etc. Partition-based is the centroid-based clustering. Hierarchical-based clustering is link-based. Density-based is clustering is focused on area of higher density in the dataset. Grid-based clustering relies on size of the grid. In this paper, we discussed different clustering techniques as well as, a detailed review on the partition-based and hierarchical-based algorithms. Finally we compare clustering algorithms on the basis of attributes like time complexity, capacity of handling large datasets, scalability, sensitivity to outliers and noise, and also discussed result after solving a particular dataset implemented in cloud computing environment.

Keywords: *Clustering; Data mining; Partition-based; Hierarchical-based; Density-based; Grid-based; Cloud computing.*

1. Introduction

The world is full of raw data. Each day we come to explore some new information which need to be preserved for further analysis and representation. The most important and crucial part of data management is to classify the data under some defined categories or clusters[1], [2]. To have an about a new object, we need to know its characteristics and features and accordingly classify them based on some similarity or dissimilarity and other definite standards [3], [4], [5]. We can define "Classification is a data analysis class where a model classifier is constructed to predict categorical levels such as safe or risky for loan application data, buy or sell for share market data, treatment of A,B or C for medical data etc.". Classification is a two-step process. In the first step a classifier is build describing the predefined sets of data classes or concepts this step is also known as learning step or training phase or supervised learning. In the second step the model is used for classification [3], [4], [6].In unsupervised learning, there are no availability of defined categorical data [7], [8]. So, clustering in unsupervised learning is possible by isolating a limited unlabeled data set into limited and discrete management of common concealed information structures [9], [4]. This helps in vector quantization [10], probability density function estimation[3] and entropy maximization [11].

Clustering plays an important role in grouping, representation of patterns analysis and choice making and machine learning (which involves data mining, classifying the given patterns, retrieval of document etc). Sometimes there are problems already provide with the statistical models where decision can be made through a few assumptions. Under these confinements, clustering method is perfectly suited for proper identification of the interrelationships among the data sets to get an evaluated structure [12].

The objective of this paper is to overview the center ideas and strategies in clustering large subsets with its foundation in insights and decision making where key ideas and clustering techniques usedin machine learning and other fields are referred and then make a survey on the clustering techniques and compare them on various basis of benefits to solve a problem.

Next section we introduced the clustering procedure and in Section 3 we studies some literature. Section 4 categories the clustering algorithms and Section 5 we implementing the clustering techniques in cloud computing environment with clustering algorithm comparisons. Finally in Section 6 we conclude our paper with some references.

2. Clustering Procedure

The steps involved in clustering activity are depicted in Fig.1. The activities mainly involve four sections i.e.

a) Clustering or grouping and clustering algorithm validation
b) Data abstraction and output generation
c) Knowledge or meaningful information
d) Pattern Representation and proximity measures.

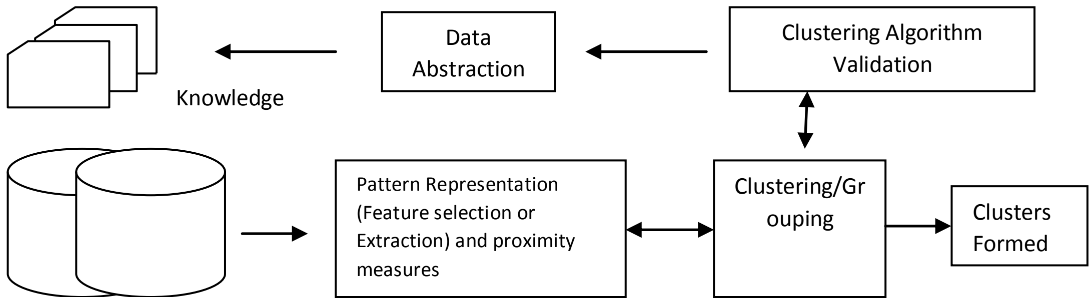

Fig. 1: Activities involved in clustering techniques

Pattern representation means number of patterns available and, their type and size for the clustering. Feature extraction means to choose or identify the distinguished features and find the most effective feature to use in clustering. Feature selection means to transform or add some features to the original ones to get more efficient features [12-14]. Proximity measures in the process of clustering is done in terms of distance (such as Euclidean distance) which differs based on similarity between the patterns [2], [8], [15]. It can be in terms of simple distance measures or to find the similarity between the patterns [16].
Clustering is an optimization problem to partition the datasets. Clustering can be done in number of ways i.e. hard or fuzzy. In hard clustering, data is partitioned into clusters whereas, in fuzzy clustering, each pattern has different membership degree for each clusters. Partitional clustering algorithms recognizes the partition to optimize the clustering criteria. Hierarchical clustering algorithms have a series of partitions based on clustering criteria to form clusters based on similarity. Other clustering method are probabilistic [17] and graph-theoretic [18].
Data abstraction helps in getting a basic and brief portrayal of dataset. The simplicity can be provided by machine learning or it is human-oriented. It gives a brief illustration of each cluster of data objects in terms of cluster prototypes or representatives such as centroids. The result needs to be accessed whether it is a good one or poor one based on certain conditions prevailing. Finally our aim is to get meaningful information from the raw data by performing clustering so that we can solve the problem efficiently.
Clustering has been connected in a wide assortment of fields running from designing (machine learning, electrical designing, mechanical building), computer sciences (data mining, examining spatial database, record gathering), medical and therapeutic sciences(genetics, microbiology, pathology), social sciences and economics [19], [20]. Clustering methods have been encouraged in the field of pattern recognition [2], image processing [21].Many terms such as unsupervised learning [8], vector quantization [22], observation learning[16], numerical taxonomy.

3. Literature Survey

Clustering came to picture since lineage dating by Aristotle.A review on data clustering was done on data clustering by Jain, Murty and Flynn [1] to present a taxonomy of clustering techniques and some applications of clustering algorithm. Hansel and Jaumard at the same time came out with the clustering problems under calculation logics. Kolatch examined the applications of algorithms used in clustering of spatial database systems and record gathering. Then, Berkhin introduced the whole idea of data mining [23]. Furthermore, Murtagh surveyed on hierarchical clustering algorithms [24] and Baraldi investigated on fuzzy and neural network algorithms [25]. Some more overview papers can likewise be found in [26] and [14].
Not only the survey papers many researches on clustering have been performed likewise in huge numbers. Rauber, Pampalk and Paralic displayed observational outcomes for five clustering techniques [27]. Wei, Lee and Hsu set the emphasis on examining quick calculations for expansive databases [28]. Scheunders thought about a few clustering methods for picture quantization, with accentuation on computational time, and also the likelihood of obtaining global optima [29]. Assessments and applications on clustering techniques to examine quality information from DNA tests are also portrayed in [30], Exploratory evaluations on clustering techniques based on hierarchical algorithms were outlined by Chih-Ping Wei, Yen-Hsien Lee, and Che-Ming Hsu [28].

4. Clustering Algorithms Categories

There are various number of clustering algorithms which can be categorized into several groups[32] i.e.
1. Partitioning Based (which contains K-means, K-medoids, PAM, CLARA, and CLARANS algorithms)
2. Hierarchical Based (which contains BIRCH, ROCK and CURE algorithms)

4.1. Partitioning-Based Clustering:

Here clusters are formed quickly. Initial clusters are determined and redistributed towards an association. Clustering in this case, occurs by partitioning datasets into number of subsets of data where each subset represent a cluster. To form a cluster certain compulsions must be followed i.e. each cluster must contain minimum one object and each object must be a part of a cluster. The basic idea here is the center of the cluster which can also be named as centroid or medoid. K-means and K-medoids [31] are two popular partitioning-based clustering techniques. In K-means, center of cluster are modified iteratively to form new clusters until a convergence is met, whereas, in K-medoids, it is just an advancement of K-means clustering to work with different data. There are many other clustering algorithms in this category such as PAM, CLARA, CLARANS [30]. Here relatively there is low time complexity and high efficiency. This clustering does not work with non-convex data which are sensitive towards number of clusters.
K-means: It is the most popular squared error-based clustering algorithm [28]. Here k is the clusters to be formed. To get a final solution, there needs to be a stopping criteria i.e. there is a definite number of relocations. There is a cluster head to represent each cluster known as centroid and with every iteration the centroid may change to form new clusters.

Initially the given dataset is randomly divided into k number of clusters based on some prior information

$$D=\sqrt{((i1-j1)^2)+((i2-j2)^2)+\cdots+((in-jn)^2)}$$

Where i= (i1, i2, i3…..in) and j= (j1, j2,j3……jn) are the two data points in the cluster.

The centroid is calculated each time a new member is added to the cluster.

To verify whether each datapoint is assigned to the appropriate cluster, we need to compare the distance to its cluster and also to the other cluster.

The relocations takes place until and unless there is no possibility of changing of locations.

This algorithm can work for both compact and hyper-spherical clusters. The time complexity of K-means is O(Nkd) and its space complexity is O(N+K) where K and d is less than N. It works on numerical data set.

But still it has some disadvantages. It is not able to tackle high dimensional data and is not always efficient as it is not an universal method for identifying initial partitions and number of clusters k. Moreover, this iterative procedure of optimization cannot guarantee to achieve global optima convergence. It cannot handle noise.

K-medoids:K-medoid is an extension to the K-means clustering algorithm. The objective of this clustering algorithm is to find less sum of error from the calculation of medoids. Here there is a compulsion to select the lowest cost of configuration. It is more robust to noise than K-means algorithm. Medoid is selected on the basis of minimal average dissimilarity of that object from other objects of the cluster. This algorithm has two steps: Build step and Swap step. In Build step, k centrally located objects are sequentially selected as medoids and in Swap step, the objective function is reduced by interchanging the selected object and non-selected object until there is no possibility of decrease in the objective function.

From the given data points of the datasets k random points are initially selected as medoids.

Then each data point is associated to the closest medoid by using a common distance metrics i.e. Manhattan distance which can be defined as:

$$D=\sum(|i1-j1|+|i2-j2|+\cdots+|in-jn|)$$

Where, i= (i1, i2, i3....in) and j= (j1, j2, j3....jn) are the two data points in the cluster.

For each non-selected object and selected object pair, total swapping cost is calculated. If the swapping cost is less than zero, then the selected object is replaced by the non-selected object. It can handle high dimension data.

Step 2and 3 are iteratively followed until any change in the medoid is possible.

It can handle high dimension data as well as can handle large data set.

PAM (Partitioning Around Medoids): PAM is same as K-medoid .PAM helps to minimize the dissimilarity between the representatives of each cluster i.e. the medoid of each cluster and its members.

Randomly select some data items as medoids.

Assign rest of the data items to the closest medoids.

For each medoid, swap the medoid and non-medoid item and then the total cost of the configuration is calculated.

Select the cluster formed having low cost.

If there is no change in medoidthen again repeat step 2 to step 4.

PAM works efficiently for small data sets. It can also handle high dimensionality data and noise like K-medoids. It can work on

In sequence, the individual data are allocated to their closest clusters as per the Euclidean distance of the centroid. The Euclidean distance is calculated as follows

categorical dataset. The time complexity is O(K(N-K)2). It deals with numerical dataset.

CLARA (Clustering Large Applications): This algorithm focusses on sampling. For better approximation on calculation of medoids, CLARA draws multiple samples and results to the best clustering [37]. Here the quality of clustering is based on the average inequality of all objects in the entire data set. It gives more satisfactory result than PAM. It works on numerical data. It has time complexity of O(Ks^2+K(N-K)). It does not handle high dimensional data and does not have the ability to handle noise.

Fix a number of sample

Draw the sample of objects randomly and apply PAM algorithm to find the medoids.

For each object, determine which medoid is most similar to the object.

Calculate average dissimilarity obtained. If the value is less than the minimum value then the medoid found by PAM is the best set of medoids.

CLARANS (Clustering Large Applications based on Random Search): In CLARANS [30], a sample of neighbor is drawn in each step of search whereas, in CLARA , a sample of nodes is drawn at the beginning of search.

Input the maximum number of neighbors examined and number of local minima.

Initialize a counter variable assigning it to 1 and a minimum cost as a large number.

Randomly choose a node and check for a random neighbor of the node based on the condition thatif random neighbor has lower cost then only consider it and go to step 3.

Calculate the cost differential of the two nodes.

If condition in step 3 is not satisfied then another neighbor is selected and the cost of current cost to be less than minimum cost so that it will be the best node.

It can deal with large data set but cannot handle high dimensional data neither can handle noise. It has high time complexity of O(N^2).

4.2. Hierarchical-Based Clustering:

It has a proximity matrix to assemble the data into a hierarchical structure [28]. The output generated from the hierarchical clustering algorithms is always a binary tree, where the root node represents the whole data set and the leaf node represents the subsets of the data. The intermediate nodes represent how much the subsets are proximal to each other and the height of the tree represent the distance between each pair of data points of the cluster. The clustering technique results in dividing the binary tree at several different levels. This clustering category can be further classified as agglomerative algorithms and divisive algorithms.

Agglomerative clustering algorithms contain N clusters where, each cluster has only one object [28]. Many addition operation are applied to the given clusters resulting in all the object belonging to the same cluster. Divisive clustering [8] follows the reverse procedure and is not cost-effective to use. So agglomerative clustering algorithms are mostly used. Lack of robustness is the disadvantages of agglomerative algorithms. They cannot handle noise and outliers. They form spherical shapes of clusters. Hierarchical relations among clusters are easily detected and they are relatively scalable.

BIRCH (Balanced iterative reducing and clustering using hierarchies):It is an unsupervised algorithm to perform

hierarchical clustering over large data sets [1], [28]. It can incrementally and dynamically form clusters from the incoming multi-dimensional data objects to produce best quality of clustering for a given set of resources. It only requires a single scan of database. It has a compact representation. A CF (clustering feature) tree IS constructed incrementally as a hierarchical data structure.
In the first phase, according to the scanned database, an in-memory CF tree is initially made.
An arbitrary algorithm technique is then used to classify the leaf nodes of the CF tree into clusters.
Data points are taken as inputs in the form of real-valued vectors and number of clusters is also taken.
Then CF is defined as (N, LS, SS)
Where N is the total number of data points, LS is the linear sum and SS is the square sum of data points.
Here the tree concept is used with 2 parameters branching factor B and threshold T. Each leaf node has child node and clustering feature which represent associated cluster.
It has two pointers i.e. previous node and next node with a definite tree size.
A page of size P contains the node and helps in finding the branching factor and also the tree size.
The algorithm scans all leaf entries to rebuild smaller tree and removes outliers and crowded sub-clusters.
This agglomerative hierarchical algorithm is used to smaller sets of cluster to specify desired number of clusters. If inaccuracies were found then again data points are redistributed to form a new set of clusters. It cannot tackle high dimensional data. The time complexity of BIRCH clustering algorithm is O(N).

CURE (Clustering Using Representative): It is an proficient algorithm for large databases those are robust and has size variances[37]. But in this clustering algorithm, some enhancements are required:
Random sampling
Partitioning
Labelling data on disks
Handling outliers
It is less expensive as compared to BIRCH. CURE clustering algorithm has a good execution time.
First set a target sample number S. Then select S sample points scattered all around in the cluster.
Then shrunk them towards the mean of the clusters (centroid).
These are the representative which will be further used for cluster merging approach.
After merging, S sample points are selected from previous cluster to form a new set of cluster.
The cluster merging procedure stops when target K cluster is filled.
CURE develops more sophisticated cluster shapes than BIRCH [30]. As it uses a well scattered points to represent the group, so there is no chaining effect or any limitation of centroid sizes. Here the random sampling and partitioning helps in reduction of computational complexities. It has low time complexity of O(S^2*S) [28].

ROCK (Robust Clustering using links): It is classified under the category of agglomerative hierarchical clustering algorithm [28]. One must have to install a CBA package to use this algorithm. ROCK improves the quality of clustering by using links. It has high time complexity of O(N^2*log N).
A random sample from database is drawn.
Clusters formed from the sampled points are used to associate the remaining data points on disk to the appropriate clusters.
The samples are deployed with hierarchical clustering algorithm is applied.

5. Implementing Clustering in Cloud Computing Environment and Comparing the Clustering Algorithms

Traditional techniques fail to handle the heterogeneous data available in the cloud. Clustered storage used to increase efficiency in performance, reliability and capacity. Clustering helps in balancing the workload among all the servers and provides each individual server the authority to access all the files regardless of the server location.
K-means is one of the most common clustering algorithms which is used in the cloud environment. K-means algorithm is applied to the Google Cloud using Google App Engine with Cloud SQL. In the deploying procedure, first we need to download the Google App Engine Plug-in from Google's official site. Here Cloud SQL is used to create database and table. Cloud computing can effectively improve the processing speed of the clustering algorithms. As we know, SaaS (Software as a Service) is an important part of cloud computing architecture model. With increase in the number of users, demand of cloud services and the number of clouds have tend to increase. Each user expects to be provided with efficient software services and this makes the task difficult. So these services need to be evaluated according to certain attributes to enable service discovery in Enterprise Cloud Bus architecture. The clustering process involved the following steps:
Obtaining SaaS data
Transformation of the SaaS data
Changing it to RDBMS/MDBMS format
Implementation of the clustering algorithm
Final cluster generated
Here in cloud computing, implementation of hard clustering algorithm like K-means leads to form a single cluster where redundancy and overlapping is not possible so it is better to apply soft clustering technique like Fuzzy C-Means clustering where individual data objects can be there in multiple clusters.

5.1 Clustering Algorithms Comparisons

Table 1: Comparisons of different Clustering Algorithms

Clustering algorithms	Clustering category	Time complexity	Scalability	For large data sets	Capability of handling high dimension-al data	Shape of suitable data set	Sensitive to the sequencing of input data	Sensitive to outliers and noise
K-means	Partition-based	O(Nkd) (Low)	Average	Yes	No	Convex	Highly	Highly
K-medoid	Partition-based	O(k(N-k)^2) (High)	Low	No	No	Convex	Moderately	Little
PAM	Partition-based	O(k^3*n^2) (High)	Low	No	No	Convex	Moderately	Little
CLARA	Partition-based	O(ks^2+k(N-k)) (Middle)	High	Yes	No	Convex	Moderately	Little

CLARANS	Partition-based	O(N^2) (High)	Average	Yes	No	Convex	Highly	Little
BIRCH	Hierarchical-based	O(N) (Low)	High	Yes	No	Convex	Moderately	Little
CURE	Hierarchical-based	O(S^2*log S) (Low)	High	Yes	Yes	Arbitrary	Moderately	Little
ROCK	Hierarchical-based	O(N^2*log N)(High)	Average	No	Yes	Arbitrary	Moderately	Little

6. Conclusion

Clustering technique is an essential process used to summarize data. It especially helps in the field of data analysis and data mining. We need to be very accurate while choosing the right clustering algorithm to produce the best quality of clusters. Every clustering algorithm has some strength and weakness. The clustering solution of the dataset after applying four different algorithms has been compared. Furthermore, soft clustering in cloud computing environment is more preferable than hard clustering technique. More work needs to be done in future to promote load balancing and efficient service providence in cloud environment.

References

[1] Rui Xu and Donald Wunsch., "Survey of clustering algorithms", IEEE Transactions on neural networks, Vol-16, Issue- 3, pp. 645-678, 2005.

[2] Banerjee, A., Pati, B. and Panigrahi, C.R, "SC2: A Selection-Based Consensus Clustering Approach", Progress in Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2016, 2, p.177, 2017.

[3] Yang, Haibo, and Mary Tate, "A descriptive literature review and classification of cloud computing research", CAIS 31, 2, 2012.

[4] Ambroise, Christophe, Geniève Sèze, Fouad Badran, and Sylvie Thiria, "Hierarchical clustering of self-organizing maps for cloud classification", Neurocomputing 30, no. 1-4, pp. 47-52, 2000.

[5] Weinhardt, Christof, Arun Anandasivam, Benjamin Blau, Nikolay Borissov, Thomas Meinl, Wibke Michalk, and Jochen Stößer, "Cloud computing–a classification, business models, and research directions", Business & Information Systems Engineering, Vol-1, no. 5, pp.391-399, 2009.

[6] Swagatika Shrabanee, Rath Amiya Kumar, Pattnaik Prasant Kumar, Markov Chain Model And PSO Technique for Dynamic Heuristic Resource Scheduling for System Level Optimization of Cloud Resources, ARPN Journal of Engineering and Applied Sciences. 13(7): 375-393, 2018.

[7] Fritzke, Bernd. "Growing cell structures—a self-organizing network for unsupervised and supervised learning", Neural networks, Vol-7, no. 9, pp.1441-1460, 1994.

[8] Sindhwani, Vikas, Partha Niyogi, and Mikhail Belkin, "Beyond the point cloud: from transductive to semi-supervised learning", In Proceedings of the 22nd international conference on Machine learning, pp. 824-831. ACM, 2005.

[9] Andrea Baraldi and Ethem Alpaydin, "Constructive feed forward ART clustering networks", IEEE transactions on neural networks 13, no. 3, pp. 645-661, 2002.

[10] Panda, Prasanta Kumar, and Swagatika Shrabanee. "Energy consumption in cloud computing and power management", International Journal of Advanced Research in Computer Science, Vol- 3, no. 2, 2012.

[11] Bernd Fritzke, "Some competitive learning methods", Artificial Intelligence Institute, Dresden University of Technology , 1997.

[12] Zhang, Qingchen, Laurence T. Yang, Zhikui Chen, and Peng Li, "PPHOPCM: privacy-preserving high-order possibilistic c-means algorithm for big data clustering with cloud computing", IEEE Transactions on Big Data, 2017.

[13] Chris Fraley, and Adrian E. Raftery, "MCLUST: Software for model-based cluster analysis", Journal of classification 16, no. 2, pp.297-306, 1999.

[14] Anil K. Jain, Robert P. W. Duin, and Jianchang Mao, "Statistical pattern recognition: A review", IEEE Transactions on pattern analysis and machine intelligence, Vol-22, no. 1, pp.4-37, 2000.

[15] Gu, Guofei, Roberto Perdisci, Junjie Zhang, and Wenke Lee, "BotMiner: Clustering Analysis of Network Traffic for Protocol-and Structure-Independent Botnet Detection", In USENIX security symposium, vol. 5, no. 2, pp. 139-154. 2008.

[16] Ryszard S. Michalski, and Robert E. Stepp, "Automated construction of classifications: Conceptual clustering versus numerical taxonomy",IEEE Transactions on pattern analysis and machine intelligence,pp.396-410,1983.

[17] Victor L. Brailovsky, "A probabilistic approach to clustering", Pattern Recognition Letters 12, no. 4, pp.193-198, 1991.

[18] D.Singh,B. Pattanayak,C. Panda, "Analysis of an Improved Energy Balanced Routing Protocol for Wireless Sensor Network", IEEE International conference on Communication and Signal processing(ICCSP-2016),pp-1503-1507, 2016.

[19] John H. Holland, "Adaptation in natural and artificial systems. An introductory analysis with application to biology, control, and artificial intelligence", Ann Arbor, MI: University of Michigan Press, pp. 439-444, 1975.

[20] Jianchang Mao, and Anil K. Jain, "Texture classification and segmentation using multi resolution simultaneous autoregressive models", Pattern recognition 25, no. 2, pp.173-188, 1992.

[21] Mahendiran, A., N. Saravanan, N. Venkata Subramanian, and N. Sairam, "Implementation of K-means clustering in cloud computing environment", Research Journal of Applied Sciences, Engineering and Technology, Vol- 4, no. 10, pp.1391-1394, 2012.

[22] Karen L. Oehler, and Robert M. Gray, "Combining image compression and classification using vector quantization", IEEE transactions on pattern analysis and machine intelligence 17, no. 5, pp. 461-473,1995.

[23] Berkhin, Pavel, "A survey of clustering data mining technique", In grouping multidimensional data, pp. 25-71. Springer, Berlin, Heidelberg, 2006.

[24] Fionn Murtagh, "A survey of recent advances in hierarchical clustering algorithms", The Computer Journal 26, no. 4, pp. 354-359, 1983.

[25] Andrea Baraldi, and Palma Blonda, "A survey of fuzzy clustering algorithms for pattern recognition", IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 29, no. 6, pp.778-785, 1999.

[26] Swagatika, Shrabanee, and Pattnaik P. K., "Design Criteria of SOA for Cloud Based Infrastructure Resource Management as a Service", International Journal of Instrumentation, Control & Automation (IJICA),Vol- 1, no-1 ,2011.

[27] A.Rauber, J. Paralic, and E. Pampalk, "Empirical evaluation of clustering algorithms," J. Inf. Org. Sci., vol. 24, no. 2, pp. 195-209, 2000.

[28] Chih-Ping Wei, Yen-Hsien Lee, and Che-Ming Hsu, "Empirical comparison of fast clustering algorithms for large data sets", In System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on, pp. 10-pp. IEEE, 2000.

[29] Scheunders, Paul, "A comparison of clustering algorithms applied to color image quantization", Pattern Recognition Letters, Vol-18, no. 11-13, pp.1379-1384, 1997.

[30] Boomija, M. D., and M. Phil, "Comparison of partition based clustering algorithms", Journal of Computer Applications, Vol-1, no. 4, pp.18-21, 2008.

[31] Ahmad, Amir, and Lipika Dey, "A k-mean clustering algorithm for mixed numeric and categorical data", Data & Knowledge Engineering, Vol-63, no. 2, pp.503-527, 2007.

[32] D. Singh, B.K. Pattanayak, "Analytical Study of an Improved Cluster based Routing Protocol in Wireless Sensor Network", Indian Journal of Science and Technology, Vol 9(37), DOI: 10.17485/ijst/2016/v9i37/97947, October 2016.