BVG Trust

J. Sci. Trans. Environ. Technov. 2016, 9(3) : 161-165

Scientific Transactions in Environment and Technovation

FSS-DKM: A Hybrid Big Data clustering approach using feature subset selection and distance based k-means algorithm

D.Anitha* **and M.V.Srinath**

Department Of Computer Science, Department of Master of Computer Applications, Sengamala Thayaar Educational Trust Womens College Mannargudi-614001, Thiruvarur District, Tamil Nadu, India

Abstract

Partitioning a dataset into identical clusters is a vital operation applied in data mining. Currently, there is an enormous effort is necessary to partition a big data into clusters. Clustering algorithms have developed as an alternative powerful meta-learning tool to precisely analyze the huge volume of data created by modern applications. In particular, their main objective is to classify data into clusters such that objects are gathered in the same cluster when they are similar permitting to specific metrics. This paper proposes a novel hybrid algorithm for big data clustering. The proposed method is designed with the help of Feature Subset Selection (FSS) algorithm and Distance based K-Means (DKM) algorithm. FFS helps to reduce the dimension of the dataset, it removes the irrelevant data from the dataset. A distance based k-means algorithm is best suited for implementing the clustering operation over big data. The experimental results for the proposed method is taken and compared with the existing method. The proposed method results better results in terms of two evaluation criteria such as error rate and adjusted rand index (ARI). The proposed method results better performance than the existing method and also provides lesser execution time.

Keywords: Big data, Clustering, Distance based K-Means, Feature Subset Selection (FSS), Error Rate, and Adjusted Rand Index (ARI).

INTRODUCTION

In current digital era, rendering to huge progress and development of the internet and online world skills such as big and powerful data servers, we aspect a huge volume of information and data day-by-day from many various resources and services which were not accessible to humankind just a few decades ago (Chen, *et al.* 2012). Huge quantities of data are created by and about people, interactions and their things. Various groups argue about the possible benefits and costs of examining information from Google, Twitter, Wikipedia, Verizon, Facebook, and every space where huge groups of people leave digital suggestions and deposit data (Fan and Bifet 2013, McAfee, *et al.* 2012). This data arises from available various online services and resources that have been recognized to serve their customers. Services and resources such as Cloud Storages, Cloud Storages, Social Networks and etc., yield big volume of data and it also requisite to manage and recycle that data or some other analytical aspects of the data. Though, this huge volume of data can be actually useful for corporations and people, it can be difficult as well. Hence, a big data or big volume of data has its own insufficiencies as well. They require big storages and this volume of data makes operations such as process operations, analytical operations, retrieval operations, very problematic and enormously time consuming (Marz and Warren, 2015).

Big Data can be acquired, stored, handled, and examined in different ways (Wu, *et al.* 2014). The sources yield diverse characteristics, including Volume, Velocity, Variety, Veracity and Value of analyzed data. These characteristics are called the 5Vs and defined as:

1. The volume of data kept today is increasing. For instance in 2013, every day Facebook and Twitter generate respectively, 7 terabytes (10^{15}) and 10 terabytes.
2. Velocity denotes both the frequency at which data are created, captured and collective.
3. Variety refers that huge data are raw, semi-structured or unstructured (however unstructured data must, for use, be structured). Moreover 80% of the planet's created data is unstructured (text, images, video, voice, etc).
4. Veracity defines the uncertain or imprecise data. This characteristic defines to the trustworthiness or dirtiness of the data. Due to the diverse forms of Big Data, quality and accuracy are less manageable.
5. Value is defined as the form of data that we examine. In fact, created data should be turned into value treatable and accessible.

One way to overcome these hard problems is to have big data clustered in a compressed format which is still a revealing version of the complete data. Such clustering algorithms are aimed to yield a good quality of clusters or summaries (Aggarwal and Reddy, 2013). Clustering

*Corresponding Author : email: *anithamca.raj@gmail.com* or clustering analysis, one of the main approaches in data mining and it is the process of placing similar data in one cluster or group and dissimilar data in other cluster. When distributing with big data, a data clustering problem is one of the most significant issues. Frequently data sets, particularly big data sets, comprise of some clusters (groups) and it is essential to discover the clusters. Clustering approaches have been applied to various significant problems, for instance, to determine healthcare trends in patient records, to remove duplicate records in address lists, to recognize new classes of stars in planetary data, to split data into groups which are useful, meaningful, to cluster millions of documents or web pages. To overcome these applications and many others a diversity of clustering algorithms has been established (Shirkhorshidi, *et al.* 2014). There exist some restrictions in the existing clustering approaches, most algorithms involve scanning the data set for several times, hence they are inappropriate for big data clustering. Moreover, there is a lot of applications in which tremendously large or big data sets requisite to be explored, but which are much too large to be handled by traditional clustering methods.

The paper is organized as follows: Section 2 presents the some of the existing big data clustering algorithms. Section 3 discusses about the proposed clustering algorithm. Section 4 discusses the comparative results and discussion. Finally, the paper is concluded in section 5.

RELATED WORKS

Cai et al (Cai, *et al.* 2013) proposed a multi view k-means clustering algorithm on big data. A robust large-scale multi-view clustering method to participate heterogeneous representations of largescale data. Hatamlou (Hatamlou, 2013) introduced a heuristic optimization approach for data clustering. Related to other population-based algorithms, the black hole algorithm (BH) starts with an initial population of candidate solutions to an optimization problematic and an objective function that is designed for them. At every iteration of the black hole algorithm, the finest candidate is nominated to be the black hole, which then starts dragging other candidates around it, called stars. If a star gets too near to the black hole, it will be accepted by the black hole and is gone forever. In such a case, a new star (candidate solution) is arbitrarily created and placed in the search space and starts a new search.

Amiri and Mahmoudi (Amiri and Mahmoudi, 2016) proposed a fuzzy cuckoo optimization algorithm for data clustering. This algorithm clusters a large dataset to previous estimated clusters numbers using meta-heuristic algorithm and ideal the results by fuzzy logic. Initially, the algorithm produces a random solutions equal to cuckoo population and with length dataset objects and with a cost function analyses the cost of each solution. Lastly, fuzzy logic tries for the optimal solution. Gutierrez-Rodríguez et al (Gutierrez-Rodríguez, *et al.* 2015) introduced a new pattern-based clustering algorithm for numerical datasets, which does not requisite an a priori discretization on numerical features. The new algorithm extracts, from a collection of trees created through a new induction procedure, a small subset of patterns valuable for clustering.

Hong et al (Hong, *et al.* 2015) proposed a genetic algorithm (GA) to find appropriate attribute clusters. Though, in their approaches, numerous chromosomes signify the same attribute clustering result (feasible solution) due to the combinatorial property, and hence the search space is greater than necessary. This training develops the performance of the GA-based attribute clustering process depends on the grouping genetic algorithm (GGA). In the technique, the general GGA representation and operators are utilized to lessen redundancy in the chromosome representation for attribute clustering. Hassanzadeh and Meybodi (Hassanzadeh and Meybodi, 2012) presented a hybrid approach based on firefly algorithm and k-means. It is revealed how firefly algorithm can be used to discovery the centroid of the user specified number of clusters. Then, the algorithm extended to use k-means clustering to sophisticated centroids and clusters.

MATERIALS AND METHODS

This section proposes a hybrid big data clustering algorithm. Usually, irrelevant and redundant features are severely damage the accuracy and efficiency of data mining and analysis. Moreover, the raw dataset comprises the missing values that may degrade or worsen the cluster accuracy. The proposed clustering approach is based on feature subset selection algorithm and distance based k-means algorithm. Features subset selection algorithm helps to remove the redundant

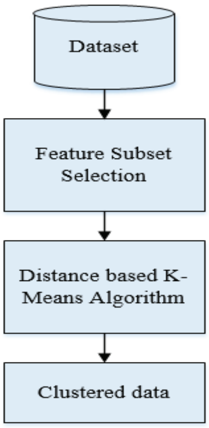

Figure 1. Hybrid big data clustering Approach

features from the raw dataset. Distance based k-means clustering algorithm clusters the dataset into distinct clusters. The flow of the proposed method is shown in figure 1.

Feature Subset Selection Algorithm

Consider the given dataset D into two subsets A and B such as $D = A \cup B \ and \ A \cap B = \emptyset$, where A = ($a_1$, a_2,…,a_m), B = (b_1, b_2,…,b_n). The symmetric uncertainty measure is used to cluster the features into groups. It is defined as follows:

$$SU\ (X,Y) = \frac{2\ Gain(X|Y)}{H(X) + H(Y)} \qquad (1)$$

Here, H(X) and H(Y) are the entropy of the discrete random variable X and Y. Suppose p(x) is the prior probabilities for all the values of X, then H(X) is defined as follows:

$$H(X) = -\sum\nolimits_{x \in X} p(x) log_2 p(x) \qquad (2)$$

Gain(X|Y) is the amount by which the entropy of Y decreases,

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum\nolimits_{x \in X} p(x|y) log_2 p(x|y) \qquad (4)$$

Here, H(X|Y) is the conditional entropy, which is defined by the following:

$$Gain(X\,|\,Y) = H\ (X) - H\ (X\,|\,Y) = H\ (Y) - H\ (Y\,|\,X) \qquad (3)$$

where, p(x|y) is the posterior probabilities of X given the values of Y.

FSS utilizes bottom-up hierarchical clustering to cluster features into groups. It first splits the two greatest similar features into one cluster. The similarity among two features is calculated using eqn (1). Then, the algorithm recursively combines the two most similar clusters at each stage. Assume that there are *m* features in the cluster a_i, and *n* features in the cluster a_j. The similarity between a_i and a_j defined as the eqn (5).

$$S_{ij} = \frac{1}{mn} \sum\nolimits_{x \in a_i} \sum\nolimits_{y \in a_j} |S \cup (x,y)| \qquad (5)$$

Next, FSS chooses one illustrative feature from every cluster to practice final feature subset. FSS recognizes the representative feature of each cluster by the measure of F-Completeness, which is definite in this paper. Assume that there are n objects in the raw dataset. For the feature cj, if there are m objects with lost the value of aj, F-Completeness of feature cj is defined as eqn (6).

$$FC = 1 - \frac{m}{n} \qquad (6)$$

FSS chooses the feature with the extreme F-Completeness as the representative feature of the cluster.

The specifics of the FSS algorithm are outlined in Algorithm 1.

Algorithm 1: *Featur Subset Selection*

Input : D-the raw dataset - the clustering threshold.

Output: S-selected feature subset

```
1.  Begin θ
2.  Compute similarity between every pair of features
3.  Merge the two most similar features into one cluster;
4.  Repeat
5.  Calculate similarity between every two clusters;
6.  Merge the two most similar into one cluster;
7.  Until
8.  The maximum similarity between every two clusters
    is less than
9.  Calculate F-Completeness of every feature;
10. S= θ
11. Add the feature with the extreme F-Completeness of
    each cluster into S;
12. Return Φ
13. End
```

The algorithm stops if the extreme similarity between any two clusters is lesser than the predefined threshold.

Distance based K means Clustering

A novel distance based K-means algorithm is proposed to cluster the data. DKM consists of the following two major steps:

1. Compute the distance among each object and every cluster to allocate the objects to the nearest cluster.
2. Compute the mean for each cluster to update the cluster centers.

In order to measure the distance among the incomplete object $o_{k,}$ and the cluster center v_i, the proposed method defines a distance measure as follows.

$$Distance = \frac{m}{I_k} \sqrt{\sum_{j=1}^{m} (o_{kj} - v_{ij})^2 I_{kj}} \qquad (7)$$

$$I_{kj} = \begin{cases} 0, if\, o_{kj} = * \\ 1, if\ o_{kj} \neq * \end{cases} for\ 1 \leq j \leq m, 1 \leq k \leq n \qquad (8)$$

Suppose that v_{ij} is the j-th feature of the center v_i of the

$$I_k = \sum_{j=1}^{m} I_{ki} \qquad (9)$$

i-th cluster c_i which is updated by eqn (8).

$$v_{ij} = \frac{\sum_{o_k} I_k * o_{kj}}{\sum_{o_k} I_k} \qquad (10)$$

$$I_k = \begin{cases} 0, if\, o_{kj} = * \\ 1, if\ o_{kj} \neq * \end{cases}$$
$$for\ 1 \leq j \leq m, 1 \leq k \leq n \qquad (11)$$

Assume that the objects with the feature subset nominated by FSS form the data set D′. To progress the cluster efficiency, the proposed approach only clusters the data subset D′ and utilizes the cluster result as the result of the raw data set D. Thus, the proposed FSS-DKM algorithm is outlined in Algorithm 2.

Algorithm 2: Distance based K-Means

Input: D′-the dataset, k-the number of clusters

Output: V- dataset consisting of cluster clusters

```
1. K-dataset containing the labels of all the objects
2. Select randomly k objects as the initial cluster centers
3. Repeat
4.    Calculate the distance between every object and each cluster center using eqn (7) and (8);
5.    Distribute every object into its nearest cluster;
6.    Update all the feature values of each cluster center using eqn (9) and (10);
7. Until
8.    All the cluster centers do not change;
9. Return V and K.
```

RESULTS AND DISCUSSION

This section presents the experimental analysis of the proposed FSS-DKM method and the comparative analysis for the proposed and the existing PDPCM approach. As the proposed cluster performance is based on the amount of missing values. Six types of missing ratios are produced such as 1%, 3%, 5%, 10%, 15% and 20% substances with missing values. In order to obtain the effectiveness of the proposed approach, two well-known evaluation criteria, E* and Adjusted Rand Index (ARI), are used in the experiment. E* is utilized to assess the error between ideal cluster centers v^i ideal and cluster centers v_i^* produced by a specific algorithm according to eqn (12).

$$E_* = \sqrt{\sum_{i=1}^{c} \| v_{ideal}^{i} - v_*^{i} \|^2} \qquad (12)$$

A lower value E* of specifies that the algorithm yields more accurate cluster centers. ARI (U,U′) is applied to extent the agreement between two partitions of a set of objects, where U signifies the ground truth labels for the objects in the data set. Here U′ means a partition produced by a specific algorithm. A greater value of ARI(U,U′) specifies that the algorithm produces a better cluster result.

Cluster Results in terms of E* and ARI

Table 1 presents the comparison of cluster accuracy in terms of E* and Table 2 presents the comparison of cluster accuracy in terms of ARI.

Table 1. Cluster result in terms of E*

Missing Ratio	FSS-DKM (proposed)	PDPCM (existing)
1%	12.24	29.21
3%	18.16	32.70
5%	17.12	39.87
10%	21.16	46.07
15%	27.24	51.61
20%	33.62	59.08

Table 2. Cluster results in terms of ARI

Missing Ratio	FSS-DKM (proposed)	PDPCM (existing)
1%	0.99	0.95
3%	0.97	0.94
5%	0.96	0.91
10%	0.93	0.87
15%	0.90	0.84
20%	0.87	0.76

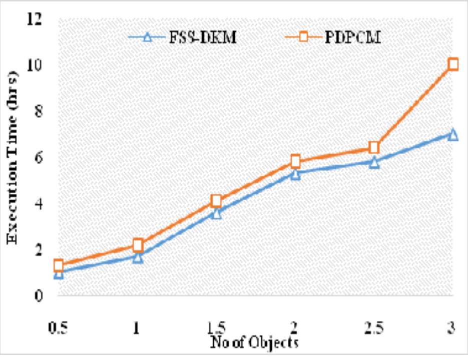

Figure 2. Execution time analysis

Table 1 and 2 proves that the proposed FSS-DKM results better performance than the existing PDPCM approach in terms of E* and ARI.

Execution Time Analysis

Figure 2 shows the comparative results for execution analysis for the proposed and existing method. The proposed method results lesser execution time than the existing method. It provides better and efficient clustering of data in big data environment.

CONCLUSION

This paper proposed a hybrid big data clustering algorithm which is based on the feature subset selection algorithm and distance based k-means clustering algorithm. Feature subset selection supports to reduce the dimension of the raw dataset. Hence, it improves the cluster accuracy of the big data. The proposed algorithm designs a new distance measure to calculate the distance among one incomplete object and one complete object, which is used to cluster big data and regulate the weight of each object. The implementation results of the proposed FSS-DKM provides better cluster accuracy on both evaluation criteria namely E* and ARI. The comparison results proves a clear superiority of FSS-DKM over the PDPCM algorithm. Moreover, it takes significantly less time for clustering high-dimensional big data.

REFERENCES

Aggarwal, C. C. and Reddy, C. K., 2013. *Data clustering: algorithms and applications*: CRC Press.

Amiri, E. and Mahmoudi, S., 2016. *Efficient protocol for data clustering by fuzzy Cuckoo Optimization Algorithm*, Applied Soft Computing, P. 15-21.

Cai, X., Nie, F. and Huang, H. 2013. *Multi-View K-Means Clustering on Big Data*, IJCAI.

Chen, H., Chiang, R.H. and Storey, V.C. 2012. Business Intelligence and Analytics: From Big Data to Big Impact', *MIS quarterly,* 36 : 1165-1188.

Fan, W. and Bifet, A., 2013. Mining big data: current status, and forecast to the future, *ACM sIGKDD Explorations Newsletter,* 14:1-5.

Gutierrez-Rodríguez, A.E., Martínez-Trinidad, J.F., García-Borroto, M. and Carrasco-Ochoa, J.A., 2015. Mining patterns for clustering on numerical datasets using unsupervised decision trees, *Knowledge-Based Systems,* 82: 70-79.

Hassanzadeh, T. and Meybodi, M.R. 2012. A new hybrid approach for data clustering using firefly algorithm and K-means, 2012 16th *CSI International Symposium on Artificial Intelligence and Signal Processing* (AISP), P. 007-011.

Hatamlou, A., 2013. 'Black hole: A new heuristic optimization approach for data clustering', *Information sciences,* 222: 175-184.

Hong, T.P., Chen, C.H. and Lin, F.S., 2015. Using group genetic algorithm to improve performance of attribute clustering, *Applied Soft Computing,* 29 : 371-378.

Marz, N. and Warren, J. 2015. *Big Data: Principles and best practices of scalable realtime data systems*; Manning Publications Co., 2015.

McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D. and Barton, D., 2012. Big data, The management revolution. *Harvard Bus Rev,* 90 (10): 61-67.

Shirkhorshidi, A.S., Aghabozorgi, S., Wah, T.Y. and Herawan, T. 2014. Big data clustering: a review. In: *Computational Science and Its Applications–ICCSA 2014,* Springer, P. 707-720.

Wu, X., Zhu, X., Wu, G.Q. and Ding, W. 2014. Data mining with big data, *IEEE Transactions on Knowledge and Data Engineering,* 26 (1) : 97-107.