International Journal of Mathematical, Engineering and Management Sciences
Vol. 4, No. 5, 1239–1250, 2019
https://dx.doi.org/10.33889/IJMEMS.2019.4.5-098

ACOCA: Ant Colony Optimization Based Clustering Algorithm for Big Data Preprocessing

Neelam Singh
Department of Computer Science and Engineering
Graphic Era Deemed to be University, Dehradun, Uttarakhand, India
Corresponding author: neelamjain.jain@gmail.com

Devesh Pratap Singh
Department of Computer Science and Engineering
Graphic Era Deemed to be University, Dehradun, Uttarakhand, India
E-mail: devesh.geu@gmail.com

Bhasker Pant
Department of Computer Science and Engineering
Graphic Era Deemed to be University, Dehradun, Uttarakhand, India
E-mail: pantbhaskar2@gmail.com

(Received May 31, 2018; Accepted June 28, 2019)

Abstract
Big Data is rapidly gaining impetus and is attracting a community of researchers and organization from varying sectors due to its tremendous potential. Big Data is considered as a prospective raw material to acquire domain specific knowledge to gain insights related to management, planning, forecasting and security etc. Due to its inherent characteristics like capacity, swiftness, genuineness and diversity Big Data hampers the efficiency and effectiveness of search and leads to optimization problems. In this paper we explore the complexity imposed by big search spaces leading to optimization issues. In order to overcome the above mentioned issues we propose a hybrid algorithm for Big Data preprocessing ACO-clustering algorithm approach. The proposed algorithm can help to increase search speed by optimizing the process. As the proposed method using ant colony optimization with clustering algorithm it will also contribute to reducing pre-processing time and increasing analytical accuracy and efficiency.
.

Keywords- Big Data, ACO, Clustering, Optimization, Preprocessing.

1. Introduction

Corporations, over the last decade, are becoming accustomed with a data-driven strategy to achieve objective services by reducing risks and perk up performance. Dedicated data analytics platforms and services are being employed to gather, accumulate, administer and investigate large data sets, called Big Data. Knowledge detection from Big Data has now become a demanding predicament. Big Data sets are distinguished because of their colossal sample size and high dimensionality. The conventional tactics used in classical statistical methods are not relevant for analyzing such gigantic data set. Modeling the intrinsic heterogeneity of Big Data requires enhanced machine learning tools, architecture and algorithm. These algorithms are required to be distributive in nature and made communication-efficient. Displaying the inborn heterogeneity of Big Data requires improved machine learning devices, techniques and algorithms.

Data mining includes machine learning which focuses on the presumption of models from a-priori known data by automatic means. Machine learning deals with the interpretation of the information and assumption of the discovered blueprints for use on future obscured information (Singh et al., 2014). The enhanced the representation of data the better the performance of machine learners. Amid the previous decade, scientific savvy machine-learning frameworks have been generally received in various monstrous and complex information concentrated fields, for example, space science, science, climatology, solution, fund and economy. Data mining involves an assortment of techniques like clustering, classification, regression, association, decision support system etc. to gain meaningful insights from the outsized amount of data (Han et al., 2011).

Clustering is one of the sought after technique applied to find hidden structure in a dataset. Based on resemblance i.e. similarity data is grouped into clusters. Clustering can be categorized as an optimization problem as the most favorable (optimal) position of unknown cluster center needs to be predicted. Big data clustering is a foremost issue as often big data sets consists of assemblage (clusters) which needs to be discovered. Clustering is playing a vital role in areas related to wireless sensor network's based application (Mahmood et al., 2013), data mining applications, bioinformatics, geographical information system, search engines etc. many clustering algorithms have been devised to handle these problems. Existing clustering methods suffer limitations like scrutinizing data multiple times which is not possible for big data clustering. Performance improvement of clustering algorithms is another sought after research area when it comes to handle and analyze big data. One of the main benchmark to be achieved for performance enhancement is computation time. Big data comes with huge volume and variety and traditional clustering algorithm fails to achieve time efficiency with these data sets.

Another issue which is often faced by most of the clustering algorithm is that the algorithms work successfully on pure numeric or categorical data but fails to perform efficiently on mixed categorical and numerical data types (Fahad et al., 2014). Big data set contains a large number of features which often leads to 'curse of dimensionality'. It becomes futile to measure the distance between pair of points as the data becomes sparse as the amount of dimensions amplify.

In order to overcome the above mentioned issues we propose a hybrid algorithm for Big Data preprocessing ACO based clustering algorithm approach. The proposed algorithm can help to increase search speed by optimizing the process.

To decipher discrete optimization problems Dorigo and Di Caro, (1999) proposed Ant colony optimization a nature –inspired metaheuristic algorithm, which impersonates the behavior of ants to search food (Handl et al., 2003). Ants correspond with each other by means of pheromone tracks and attempt to uncover the shortest trail from food resource to nest. Ant colony optimization can be applied to hard combinatorial optimization tribulations to unearth solutions in the rational quantity of computation time. Application of ACO algorithm is to discover fairly accurate solutions to complicated optimization tribulations ranging from travelling salesman problem, quadratic assignment problem, job scheduling problem etc. (Liu and Fu, 2010).

As the proposed method using ant colony optimization with clustering algorithm it will also contribute to reducing pre-processing time and increasing analytical accuracy and efficiency.

The paper has been ordered as follows. Section 2 establishes the background work done in the field. Section 3 focuses on the Ant colony optimization method and k-means clustering algorithms. Section 4 provides an overview of the experimental setup and results. In Section 5, we conclude the work by providing further research objectives and future scope.

2. Background and Overview of K-Means and ACO Algorithm's

Big data is exemplified by large volumes, varieties, rapidity and authenticity posing a multifaceted challenge to data mining tasks. Traditional methods often lose their potential to handle such colossal dataset even if powerful computer clusters are employed. Many new robust and scalable machine learning and data mining algorithms, procedures and technologies are building up to handle big data leading to new analytical methods, data types and storage architectures. Classical algorithms are being extended with new methodologies based on swarm intelligence as they improve the quality of result extracted making them more competent. ACO a.k.a. ant colony optimization is always been the most thriving procedure(Handl et al., 2003) which deploy the natural foraging conduct of ants to find the shortest conduit for a food source. These algorithms are used in optimization areas together with machine learning (Menéndez et al., 2016).

Data clustering holds a critical role in big data analysis as the dataset consists of groups and it requires to be clustered accordingly. Areas like document clustering for grouping web pages, healthcare analysis, astronomical analysis etc. deploy clustering methods (Kurasova et al., 2014).

Considering a dataset X of data items X_1, X_2,…….X_i with a feature set x_1, x_2……….x_j, such that *'i'* is the quantity of data items and *'j'* is the quantity of features. Taken into account big data both *'m'* and *'n'* values are comparatively high enough to be handled by conventional clustering algorithms.

Existing clustering methods have several limitations pertaining to big data like:

- Category of the dataset: Conventional algorithm usually works on numeric or categorical data which is unlikely for big data as it can have a mix of both.
- Data Scan: Running data scan for multiple times increasing time complexity of the process. The high colossal volume of big data often obstructs the efficiency of clustering task (Singh et al., 2017).
- Choice of initial centers: Selection of varying initial centers for a clustering algorithm such as k-means may provide dissimilar results.
- It becomes a complicated task to find termination criteria for the hierarchical clustering algorithm.
- Input parameters: Traditional clustering algorithms suffers 'curse of dimensionality' as they work efficiently and effectively when fewer input parameters are available which is not in the case of the high dimensional big dataset.

Data clustering is an NP-complete problem associated with the discovery of groups based on some similarity or dissimilarity measure. Mathematically it can be viewed as an optimization problem such that:

Let in a clustering data space S there exist a set of '*n*' object {X_1, X_2……..X_n} with '*K*' clusters {C_1,C_2…..C_k}. Constrictions for clustering partitions can be defined as:

$$min(f)$$
$$s.t.\begin{cases} C_i \neq \phi, \forall_i \in \{1,2,........k\} \\ C_i \cap C_j = \phi, i \neq j \ and \ i, j \in \{1,2,\ldots,k\} \\ \bigcup_{i=1}^{k} C_i = S \end{cases} \tag{1}$$

'*f*' is the function that validates the cluster. There are various legitimacy functions used like cosine similarity, Euclidean, Manhattan etc. Clustering aims to group data based on the theory of minimum intra-group divergence and maximum inter-group divergence (Jiang et al., 2011; Xing et al., 2016).

Clustering is an iterative progression involving i) cluster head selection based on the data set associated with cluster ii) updating the data within the cluster based on cluster centroid in the space. This is the theme used by one of the unsurpassed K-means algorithm (Menéndez et al., 2016).

2.1 Customary K-Means Algorithm

When speed is a priority k-means proves to be advantageous for a large dataset in comparison to other algorithms as it exhibits time complexity O(tkn) where '*t'* is the quantity of iterations required for convergence, '*k*' is cluster value and '*n'* is the count of objects. While algorithms like agglomerative hierarchical clustering have time complexity as O(n3).

2.1.1 Steps for K-Means Algorithm

Start

Input: Number of preferred clusters, '*k*' and a dataset X={x_1, x_2, x_3 ...x_n} comprising of *'n'* data items.

Output: k clusters

Method:

Let X ={x_1, x_2, x_3…x_n} be the set of data points

(i) Decide the amount of clusters '*k*' to be produced.

(ii) The cluster centroids are selected arbitrarily.

(iii) Using similarity criteria find the distance amid individual data point and cluster core.

(iv) Data item with the lowest amount of inter-assembly distance and highest intra-assembly distance is allocated to a cluster.

(v) Again the cluster head is recalculated.

(vi) Also the distance among individual data item and new derived cluster cores is reevaluated.

(vii) Repeat from step third till further reassignment else stop.

Although K-means algorithm exhibit low-computational cost with higher efficiency when large datasets are clustered so it is considered to be one of the prominent solutions for big data analysis still it faces certain limitations like prior selection of cluster number, well suited for single-view data clustering problems and also the algorithm converges to local minima rather than giving a global optimum outcome.

Several extensions of standard extensions have been building up like kernel k-means, spherical k-means, Minkowski metric weighted k-means, fuzzy c-means etc. (Kurasova et al., 2014). These extensions are formulated to amplify the rapidity and competence of the clustering task. K-means algorithm has been optimized using Bayesian information criterion so that the appropriate number of clusters can be determined in X-means method (Soheily-Khah, 2016).

Applying a brute-force approach to solve combinatorial clustering problem to achieve a global optimal solution is not feasible as the increase in dataset volume shoot up the cluster quantity making it computationally exhaustive. To analyze this random probability distribution to achieve global optima heuristic approaches are considered (Fong et al., 2014).

Meta-heuristics approaches aka nature-inspired optimization methods, imitating the swarm behavior of living creatures have come into subsistence, exploiting search agents to represent a scrupulous amalgamation of centroid arrangement.

Handl et al. (2003) proposed ant based algorithm to be a multi-agent process to combinatorial optimization tribulations like Traveling Salesman Problem (TSP). Application and extension of these ant- based schemes for discrete optimization problems is an enduring research commotion owing to its robustness, exceptional distributed computation mechanism and strong affinity and compatibility with other methods.

Ant colony inspired clustering methods were first explored by (Deneubourg et al., 1991) by instituting an elementary model to cluster data based on the arbitrary movement and pick and drop of ant colony according to the similarity between data object in space. This model was further extended for numerical data analysis by Lumer and Faieta (1994) by proposing the LF algorithm.

2.2 The Ant Colony Optimization Approach

Ant algorithm follows the given steps:

(i). Initialization phase

a. Data elements are indiscriminately speckled on the toroidal grid.
b. Every agent (ant) which is also positioned at a subjective location on the grid arbitrarily picks a data item.

(ii). Sorting Phase (Iterative phase)

a. Random selection of an agent
b. Based on step size agent moves through the grid and probabilistically choose whether to plunge an item or not.
c. If it is a drop-decision then agent at the current lattice position or in the neighborhood drops it and start searching for a new data element to be picked up.

This continues until the agent successfully picks a data item.

(iii) This process is repeated for the rest of the agents.

Transition rule equation defines the probability to select the next node. It represents the possibility for ant '*k*' to go from one grid position '*i*' to next grid position '*j*' on the *t*th tour

$$\mathrm{p}_{\mathrm{ij}}^{\mathrm{k}}(t)=\frac{\left[\tau_{ij}(\tau)\right]^{\alpha}.\left[\eta_{ij}\right]^{\beta}}{\sum_{l\in j_i^k}\left[\tau_{il}(t)\right]^{\alpha}.\left[\eta_{il}\right]^{\beta}} \tag{2}$$

In the given equation:

τij: pheromone trail
ηij: visibility between the two nodes,
α, β: trail intensity and visibility controlling amendable parameters

2.3 Ant Colony Optimization in Clustering

Both clusterings as well as classification had been implemented along with ACO in order to yield accurate results as these algorithms are likely to get trapped in local minima whereas ACO performs a global search. Yang and Kamel (2006) illustrated a multi-ant colonies procedure for clustering based on parallel and independent ant colonies. Menéndez et al. (2014) discussed the evolutionary algorithmic approach used by various famous bio-inspired methods to deal with clustering predicament. ATTA an adaptive clustering algorithm to cluster data on the basis of corpse's behavior of ants was discussed (Handl et al., 2003). Niknam and Amiri (2010) implemented a hybrid approach using ACO for clustering. Implementation of clustering based on ACO and K-medoids optimization methods to handle uneven clustering problem (Peng et al., 2014). Utilization of ACO procedure for finding long paths through the data for graph based clustering of spectral imagery has been employed (Ashok and Messinger, 2012).

A review on nature stimulated metaheuristic algorithms for partitional clustering has been presented by Nanda and Panda (2014). Yang and Li (2013) implemented map-reduce based big data clustering using ants for parallel clustering. Adaptive Data clustering gained by pheromone interaction among ants based on agents is realized using Ant Colony algorithm (Menéndez et al., 2016). A study on various clustering approaches in relation to the evolutionary algorithm was presented (Jafar and Sivakumar, 2010).

3. Proposed Method

The ACOCA algorithm means ACO with K-means algorithm. The proposed approach presents a hybrid algorithm that integrates the K-means process amid ACO method. The ACO technique is a well-recognized optimization technique and K-means is known for its competent clustering. In order to unite the efficacy of both the algorithms this approach has been proposed to preprocess big data. The projected algorithm is designed to incorporate the features of the above mentioned algorithms.

In this work, the ACO based filtering step is considered with clusters taken as input. The idea for this refinement is taken into consideration as in any clustering algorithm it is impossible to obtain 100% quality from the acquired clusters. It is prone to erroneousness known as mis-clustering.

The approach follows the stochastic and investigative ideologies of ant colony based method with the deterministic and probing philosophies of the K-means algorithm.

3.1 ACO Based K-Means - ACOCA

Input: Dataset of the highest magnitude

Step 1: Input number of clusters '*K'* such that the number of ants is the same as the amount of Clusters.

```
Step 2: Initialize Cluster centroids 'K' using ACO Optimization.
        Initialize Number of Ants 'A'
        For i=1 to m (iterations)
```

a) Based on the given likelihood for each ant, let every article *'x'* assimilate in one cluster

$$\mathrm{p}_{\mathrm{ij}}^{\mathrm{k}}(t)=\frac{\left[\tau_{ij}(\tau)\right]^{\alpha}.\left[\eta_{ij}\right]^{\beta}}{\sum_{l \in j_i^k}\left[\tau_{il}(t)\right]^{\alpha}.\left[\eta_{il}\right]^{\beta}} \tag{3}$$

b) Reckon the pickup probability, drop probability.
c) For each ant, ensure if all occurrences have been visited.
d) Allot individual data articles to their neighboring centroid and estimate the objective function value for every ant *a*
e) From M solutions evaluated save the best elucidation using the given steps:
   i. grade the ants elucidation;
   ii. Eliminate an ant from ranking with centroids less than *k.*
   iii. pick the finest ant *a∗* (iteration- paramount result);
f) Based on the finest solution update the pheromone level for all items.

$$\tau_{ij}(t)\xleftarrow{test}\delta.\tau_{ij}(t)+(1-\delta).\tau_0 \tag{4}$$

Stop Iteration.

Step 3: Using local best result, measure Global best Ant such that the number of ants is equal to global best solutions.

Step 4: Euclidean Distance metrics is used to assign data points to the closest center.

Step 5: New cluster center is evaluated by calculating the mean of the data points of the clusters.
a) For the identified data articles verify the preliminary and concluding F-measure.
b) Reallocate the data article to another cluster when final F-measure is larger than the initial.
c) Estimate the entropy

Step 6: Till convergence occur (i.e. no reassignment) repeat steps 4 & 5.

3.2 Evaluation Measures

To analyze the effectiveness of the proposed scheme various evaluation measures have been considered.

Ant agents will move on a grid where data items are scattered. Pickup and drop probability is the measures for the picking and dropping of a data item of the dataset. These operations rely on the likeness and compactness of the data items.

Pickup Probability:

$$P_p=\left(\frac{I_1}{I_1+f}\right)^2 \tag{5}$$

Drop Probability:

$$P_d = \left(\frac{f}{I_2 + f} \right)^2 \tag{6}$$

where '*f*' means entropy of the data item prior to it is selection and moving into a different cluster. Pick up and dropping probabilities are represented by 'I_1' and 'I_2' as threshold constant. Entropy and F-measure are measured for quality evaluation.

Entropy:

Entropy can be formulated as the summation of the probabilities with which an affiliate of cluster '*j*' fit into class '*i*'. The probability is given by '$p_{i,j}$'. The given formula is used to find entropy.

$$E_j = -\sum_i p_{i,j} \log\left(p_{i,j}\right) \tag{7}$$

For a group of clusters taken as a whole entropy is evaluated as the addition of the entropies of all cluster weighed by means of the volume of each cluster.

$$E_{cs} = -\sum_{j=1}^{m} \frac{n_j * E_j}{n} \tag{8}$$

F-Measure:

It is measured based on the recall and precision value such that

$$Recall(\, i, j\,) = m_{ij}/m_i \tag{9}$$

$$Precision(\, i, j\,) = m_{ij}/m_j \tag{10}$$

where 'm_{ij}' is the frequency of elements of class '*i*' in cluster '*j*', 'm_j' is the frequency of elements of cluster '*j*' and 'm_i' is the frequency of elements of class '*i*'.

The F measure for cluster '*j*' and class '*i*' can be computed using the following equation:

$$F(i,j) =(2 * Recall(\, i, j\,) * Precision(\, i, j\,)) / ((Precision(\, i, j\,) + Recall(\, i, j\,)) \tag{11}$$

The total assessment of F-measure is calculated as the weighted average of all values for the F-measure according to equation 12:

$$\sum_i \frac{n_i}{n} \max\{F(i, j)\} \tag{12}$$

4. Performance and Results

The proposed algorithm's performance is tested by taking movie dataset from IMDB repository. The dataset comprises of 28 attributes and the Pearson correlation plot for the attributes has been shown (Figure 1). ACOCA based clustering is applied on the dataset and the result is plotted in the form of clusters (Figure 2).

The proposed algorithm is then evaluated and judge against the standard k-means algorithm by using various evaluation measures including Entropy, F-measure and run time. The cluster analysis based on the given parameters is presented (Table 1, Figure 3).

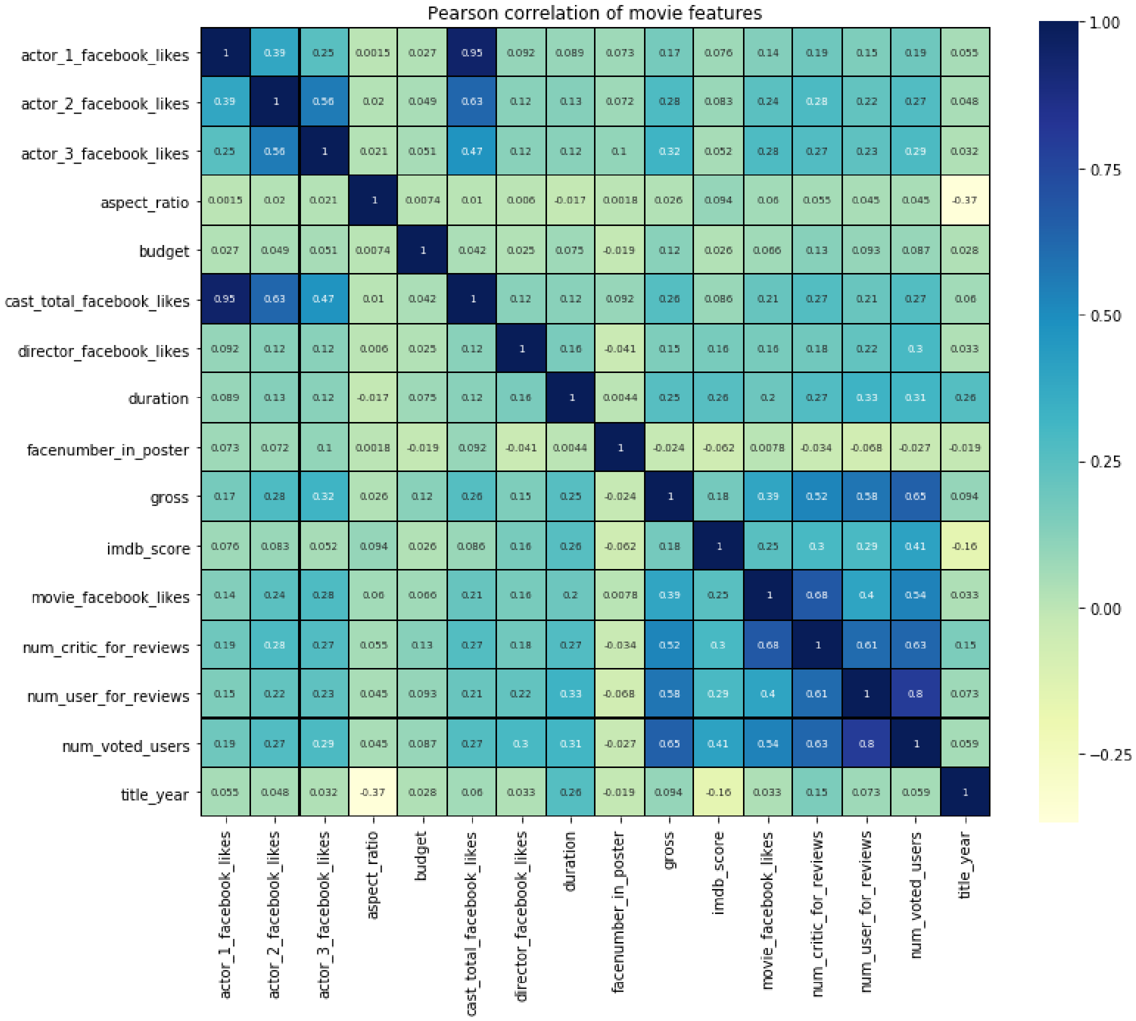

Figure 1. Pearson correlation plot for movie attributes

The clusters generated according to the proposed algorithm have been shown (Figure 2) with three distinguishable clusters as per our experimental setup.

The results of the algorithms are evaluated in terms of the entropy, F-measure and run time (in seconds). A comparison between the conventional k-means algorithm with ACOCA is presented in Table 1 which shows an enhancement in F-measure and run time. Also there is a decrease in entropy in comparison to the traditional methods which tends to alter as the size of the dataset increases due to raising in the amount of iterations.

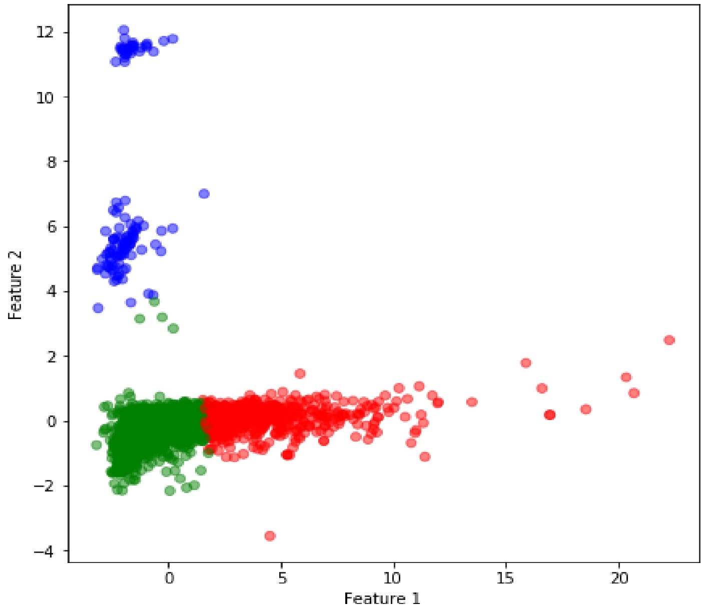

Figure 2. ACOCA cluster formation

Table 1. Cluster analysis

Parameters/ Evaluation Measure	**k-means**	**ACOCA**
Number of Clusters	3	3
Entropy	0.3315	0.2745
F-measure	0.8576	0.9332
Run time (second)	0.62	0.11

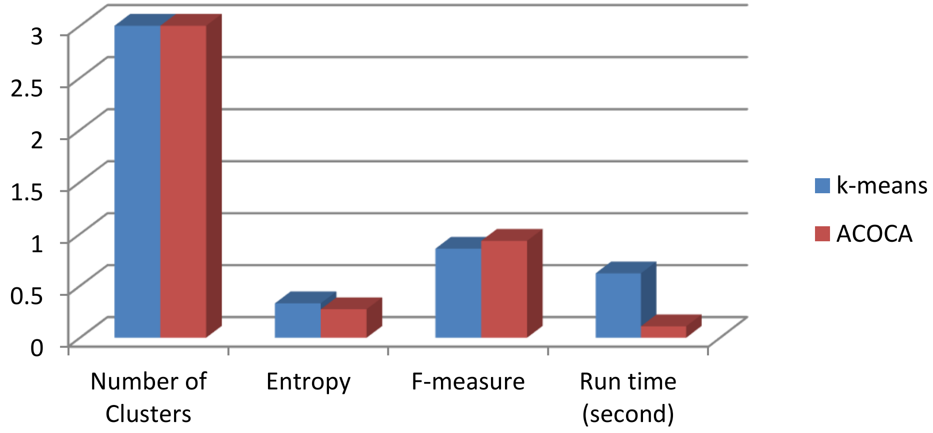

Figure 3. Cluster analysis

The complexity of the proposed algorithm increases as it involves pre-calculation of pick and drops probabilities and pheromone update step has also been increased as compared to the k-means algorithm but based on the evaluation parameters it shows significant improvement in the results.

5. Conclusion

In this paper we analyzed the cluster performance by integrating ant colony optimization with k-means to preprocess big data sets. Using this hybrid approach the accuracy of cluster formation has been approved by improving the run time performance and lessening the amount of iterations required for the algorithm to congregate. The algorithm executes well and gives improved cluster quality. The algorithm also shows better performance as compared to the conventional k-means algorithm with big data sets. The hybridization of optimization techniques shows that the drawbacks of conventional algorithms can be overcome and can be used for large data sets. The algorithm can further be extended on a parallel or distributed environment to achieve greater accuracy and runtime performance.

Conflict of Interest

The authors confirm that this article contents have no conflict of interest.

Acknowledgement

The authors acknowledge and express the gratitude for the support of Graphic Era Deemed to be University, in Dehradun, Uttarakhand, India.

References

Ashok, L., & Messinger, D.W. (2012, May). A spectral image clustering algorithm based on ant colony optimization. In *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVIII,* (Vol. 8390, p. 83901P). International Society for Optics and Photonics.

Deneubourg, J.L., Goss, S., Franks, N., Sendova-Franks, A., Detrain, C., & Chrétien, L. (1991, February). The dynamics of collective sorting robot-like ants and ant-like robots. *In Proceedings of the First International Conference on Simulation of Adaptive Behavior on From Animals to Animats, (pp. 356-363).*

Dorigo, M., & Di Caro, G. (1999). Ant colony optimization: a new meta-heuristic. In *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406),* (Vol. 2, pp. 1470-1477). IEEE.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A.Y., & Bouras, A. (2014). A survey of clustering algorithms for big data: taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, *2*(3), 267-279.

Fong, S., Deb, S., Yang, X.S., & Zhuang, Y. (2014). Towards enhancement of performance of K-Means clustering using nature-inspired optimization algorithms. *The Scientific World Journal*, Article ID 564829.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Handl, J., Knowles, J., & Dorigo, M. (2003, July). Ant-based clustering: a comparative study of its relative performance with respect to k-means, average link and 1d-som. In *Proceedings of the Third International Conference on Hybrid Intelligent Systems, IOS Press*.

Jafar, O.M., & Sivakumar, R. (2010). Ant-based clustering algorithms: a brief survey. *International Journal of Computer Theory and Engineering*, *2*(5), 787-796.

Jiang, L., Ding, L., Peng, Y., & Zhao, C. (2011, July). An efficient clustering approach using ant colony algorithm in mutidimensional search space. In *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, (Vol. 2, pp. 1085-1089). IEEE.

Kurasova, O., Marcinkevicius, V., Medvedev, V., Rapecka, A., & Stefanovic, P. (2014, November). Strategies for big data clustering. In *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, (pp. 740-747). IEEE.

Liu, X., & Fu, H. (2010). An effective clustering algorithm with ant colony. *Journal of Computers*, *5*(4), 598-605.

Lumer, E.D., & Faieta, B. (1994, July). Diversity and adaptation in populations of clustering ants. In *Proceedings of the Third International Conference on Simulation of Adaptive Behavior: From Animals to Animats 3: From Animals to Animats, 3* (pp. 501-508). MIT Press.

Mahmood, A., Shi, K., Khatoon, S., & Xiao, M. (2013). Data mining techniques for wireless sensor networks: a survey. *International Journal of Distributed Sensor Networks*, *9*(7), 406316.

Menéndez, H.D., Otero, F.E., & Camacho, D. (2014, September). MACOC: a medoid-based ACO clustering algorithm. In *International Conference on Swarm Intelligence*, (pp. 122-133). Springer, Cham.

Menéndez, H.D., Otero, F.E., & Camacho, D. (2016). Medoid-based clustering using ant colony optimization. *Swarm Intelligence*, *10*(2), 123-145.

Nanda, S.J., & Panda, G. (2014). A survey on nature inspired metaheuristic algorithms for partitional clustering. *Swarm and Evolutionary Computation*, *16*, 1-18.

Niknam, T., & Amiri, B. (2010). An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis. *Applied Soft Computing*, *10*(1), 183-197.

Peng, L., Dong, G.Y., Dai, F.F., & Liu, G.P. (2014). A new clustering algorithm based on ACO and K-medoids optimization methods. *IFAC Proceedings Volumes*, *47*(3), 9727-9731.

Singh, N., Garg, N., & Pant, J. (2014). Document clustering using feature selection based on multiviewpoint and link similarity measure. *International Journal Computer Technology & Applications*, *5*(3), 1151-1155.

Singh, N., Singh, D.P., & Pant, B. (2017, December). A comprehensive study of big data machine learning approaches and challenges. In *2017 International Conference on Next Generation Computing and Information Systems (ICNGCIS)*, (pp. 80-85). IEEE.

Soheily-Khah, S. (2016). *Generalized k-means-based clustering for temporal data under time warp*, (Doctoral dissertation). Université Grenoble Alpes, English. NNT: 2016GREAM064. tel-01680370v2.

Xing, E.P., Ho, Q., Xie, P., & Wei, D. (2016). Strategies and principles of distributed machine learning on big data. *Engineering*, *2*(2), 179-195.

Yang, J., & Li, X. (2013, October). Mapreduce based method for big data semantic clustering. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, (pp. 2814-2819). IEEE.

Yang, Y., & Kamel, M.S. (2006). An aggregated clustering approach using multi-ant colonies algorithms. *Pattern Recognition*, *39*(7), 1278-1289.