International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 5 Issue 6, June-2016, Impact Factor: 1.544

A Bee Colony Optimization Based Improved to Hierarhical Clustering

Preeti[1], Manju[2]

Student, M.Tech (CSE) PDM College of Engineering, Bahadurgarh, Haryana
Asst. Prof, M.Tech (CSE), PDM College of Engineering, Bahadurgarh, Haryana

ABSTRACT

Clustering is having the significance to generate effective data segments based on feature and dimension evaluation. In this work, an ABC improved Hierarchical clustering method is defined. The method has divided the clustering activities for different types of bees with their relative role assignment. The scout bee defined as the controller and the onlooker bees are defined as the cluster agents. Based on these agents, the cluster segments are evaluated under feature specific parametric evaluation. The work is applied on two benchmark datasets. The proposed work model has improved the accuracy and reliability clustering process.

Keywords: ABC, Clustering, Segmentation, Dimension, parameter Specific

I. INTRODUCTION

Data clustering is the self featured and processed method without specification of any known class or the pre-trained knowledge or characterization. Data processing and streaming comes under ubiquitous computing. This processing is based on different measures, features and domains. The clustering is the data processing activity defines to categorize, prune and filter the data. It not able to divide the data in different segments but also able to identify the existence of any outlier in dataset. Clustering can be defined with specification of data patterns and specification of data characterization. Clustering is a process that can be defined as a pre-processing stage or can be defined as an individual process method that can be used to generate the data segments. Data discovery with specification of associated rules and relative feature pattern is an essential and requirement measure. The clustering not only able to improve the efficiency but also able to improve the accuracy of any data processing activity. As the clustering method provides the ability to process on selected data segment. This reduced size data processing improves the efficiency of any algorithmic model. The selective feature based analysis, working on most relevant data and outlier pruning improves the reliability of any mining model or algorithm. But the features and ability of clustering method also raise some challenges for clustering methods.

The size and dimension of available data pool increases the criticality of clustering process. The processing such larger dataset size increases the computational complexity and processing speed. Larger memory and high speed processors are the basic requirement of clustering method and environment. Clustering process cannot be accomplished by performing the one time mapping, because of this it is considered as the continuous or the increment process model that can be used to generate effective clusters. Another feature of clustering method is the dynamic form of available data. Some of the datasets are defined with relative timestamp as well as updated regularly. Such kind of data updation need more clear observation respective to the changing behavior of dataset. The clustering model is required to keep the multiple views on dataset under different perspectives. These views can be respective to attributes, dimension, or based on analytical features. Clustering is also able to identify the outlier from the dataset. This outlier is defined as the abnormal information that can be generated over the dataset. The parameter specific evaluation and feature driven computation can be applied to improve the clustering quality and results.

Clustering methods can be applied variety of areas and domains including the text based and numerical processing. The extensive domains of clustering include image, semi-structured data clustering, document clustering, gene expression processing etc. The clustering is having the significance to the web data and the data processing in effective data forms. An attribute specific estimation and evaluation is required based on the inclusive data forms. These data forms are also defined relative to the associated patterns and mining activity. The featured evaluation along with similarity measures are required to observe. The behavior observation and the problem navigation are requited with each of the processing iteration so that the cluster tracking will be done. The number of elements, regions and the pattern condition is also considered while forming these clusters. The position specific analysis with data points specification is required to generate the clusters effectively.

In this paper, an ABC improved clustering method is defined to generate the data segments. The proposed dynamic method used the agent specific approach for generating the effective clusters. In section I, the basic requirement and characterization of clustering is presented. In section II, the work defined by earlier researchers for cluster optimization is presented. In section III, the proposed research methodology is presented. In section IV, the results obtained from the work are presented. In section V, the conclusion of the work is presented.

II. RELATED WORK

Data Clustering is the essential data processing activity used to filter the available data pool and to generate the required data patterns. Many researchers already submitted clustering work on organized, semi-organized, unorganized and incomplete datasets. In this section, some of the work provided by earlier researchers is presented and discussed. Author[1] has defined self organized analysis method to provide clustering on incomplete data. Author examined the data under different parameters to handle the features of incomplete data and provided the dimension specific clustering. The method is defined to generate the effective clustering solution under practical approach. This defined dataset is processed under iterative manner and to provide the cluster formation to improve the cluster quality. Author[2] defined a study specific work on clustering for different datasets and for different applications. The streamed data analysis and task specific mining method is provided for evolution of data elements. The mining method is here applied under feature analysis to process the clustering data and to provide the effective cluster formation.

Author[3] has improved the purity of clustering method under entropy specific data categorization. Author provided the distance specific analysis on data points and distance specific under different parameters. Author used the Shannnon concept with clustering method to provide integration to cluster formation. Author also identified the outlier to improve the efficiency of clustering algorithm. Author[4] has used the hybrid clustering method using KMeans clustering and Neural network approach. Author generated the data clustering based on the automated mask generation method to discover the hidden patterns and to provide the boundaries specific membership analysis. This visualization specific method provided the integration in an emergent method. The data mining method and relative feature analysis was provided to discover the hidden similarities so that the membership characterization will be achieved. Author[5] has used the shift specific clustering method for educational data and provided the aspect driven mining method integration for improving the clustering results. Author processed the educational data to generate the pattern prototype for resolving the associated difficulties and to provide the cluster quality enhancement using clustering approach. Author defined a feature trained method for improving the quality of clustering methods.

A work on microarray[6] data processing and decomposition using clustering method was provided by Wang. The work combined the FCM method with empirical decomposition method to reduce the noise effect and to generate the effective clustering structure. Author processed the structural information of dataset under fuzzy operation to generate more reasonable results. Another work on temporal[7] clustering was provided by Yang by using the concept of weight assignment. A feature cut specific information organization and reduction was provided by the author along with time series specification. Author processed the benchmark datasets to generate effective clusters underweight processing method. The proposed ensemble algorithm used the partitioned method to improve the quality of formed clusters. An optimization to clustering method was provided using PSO[8] approach for Web usage data. Author processed the heterogeneous data by combining the hierarchical clustering and PSO approach.

The similarity measure based clustering effect was verified by the author to improve the degree of applicability. A work on Affinity Propagation based clustering method was provided on large scale datasets. The complexity driven analysis

along with cluster formation in global environment was provided by the author. The data point specific similarity analysis along with adaptive hybrid algorithm was provided by the author. A work on stock[10] data analysis using clustering approach based on featured SVD method. The proposed hybrid method used the singular decomposition method to generate the features using Canopy and KMeans algorithm and implement them in Hadoop environment. The proposed massive method is defined for time series data and able to provide the effective conclusion for data clustering. Another work on fuzzy[11] adaptive clustering method was provided for relational data. Author used the Euclidian distance based similarity analysis to identify the degree of membership and to generate the effective clustering results. The outlier identification and interpretation under noise class estimation was provided by the author.

A comparative study on distance and similarity measures for mix attributes based clustering method was provided by Prasetyo et. al.[12] Author provided the prototype and feature driven analysis on multiple datasets. The ratio specific mismatch analysis was provided to generate the similarity specific clustering results. Reddy[13] used the labeling specific categorical data processing under entropy method for generating the clusters. A future specific similarity analysis was provided to generate the clusters and to identify the outlier. The label specific method has improved the quality of generated clusters. A work on data clustering on incomplete data using fill method and tolerance set specific dissimilarity analysis was provided by Hua ai[14]. The method applied the probability hypothesis for generating the clusters and verified it under constraint tolerance method. This data object based constraint processing has generated the effective clusters. Tsai et. al.[15] has used the grid based clustering method for intuitive neighbor generation under relationship analysis.

A neighbor checkpoint based analysis and evolution under noise and correctness parameter was provided to improve the quality of clustering method. A feature driven measure relative to cluster framing was provided to improve the quality of cluster formation. Wang[16] also used the fuzzy clustering method to process multiple medoid for processing the large dataset. Author applied the incremental clustering under complex pattern formation so that the pattern specific clusters will be formed and effective segmented data will be achieved. Another work on customized KMeans method for uncertain features and measurement form was provided by the author and relatively constraint specific functionality observation was provided. A probability distribution based uncertain data processing with real means and standard division processing was provided by the author. Author used the realistic measure for cluster generation and distance computation in multi-dimension data environment.

III. RESEARCH METHODOLOGY

Clustering is one of the traditional and global task that can be applied on different data forms, dimension and application domains. The requirement and significance of this method exist as process stage of many feature generation and prediction methods. Because of this, there is always a requirement to improve the effectiveness of clustering method. In this present work, an improvement to the hierarchical clustering method is provided by integrating the Artificial Bee Colony Optimization Algorithm. The proposed method is here considered as an agent driven method in which the process stages are defined respective to the role of each bee. The method is defined on dimension independent datasets. The data is here presented as the pool with specification of number of clusters. After setting up these required clusters, the random centers are generated. This whole work of center setup and environment initialization is here defined by the controller bee called scout bee.

This bee also controls the cluster data switching from one iteration to other. The process of error specific analysis and the validation of clustering method is also decided by the scout bee. Once the centers are identified, the next work defined by this clustering model is to assign an onlooker bee for each center. These onlooker bees will work as the agent for clustering environment and perform the evaluation for each individual cluster. Now for each cluster, the onlooker bee assign the employee bee defined with each data instance. The feature, distance and dimension specific analysis is defined by the employee bee. This estimation is defined respective to the cluster center. Once the distance information is collected, this information is submitted to the cluster centers. Here, the onlooker bee perform the estimation of the distance information respective to other cluster and identify the actual qualifier to the data member. After forming one level of clustering, the scout bee controls the clustering method and identifies the cluster formation. This process is repeated till the clusters are not formed effectively. The clustering model defined and suggested in this work is shown here in figure 1.

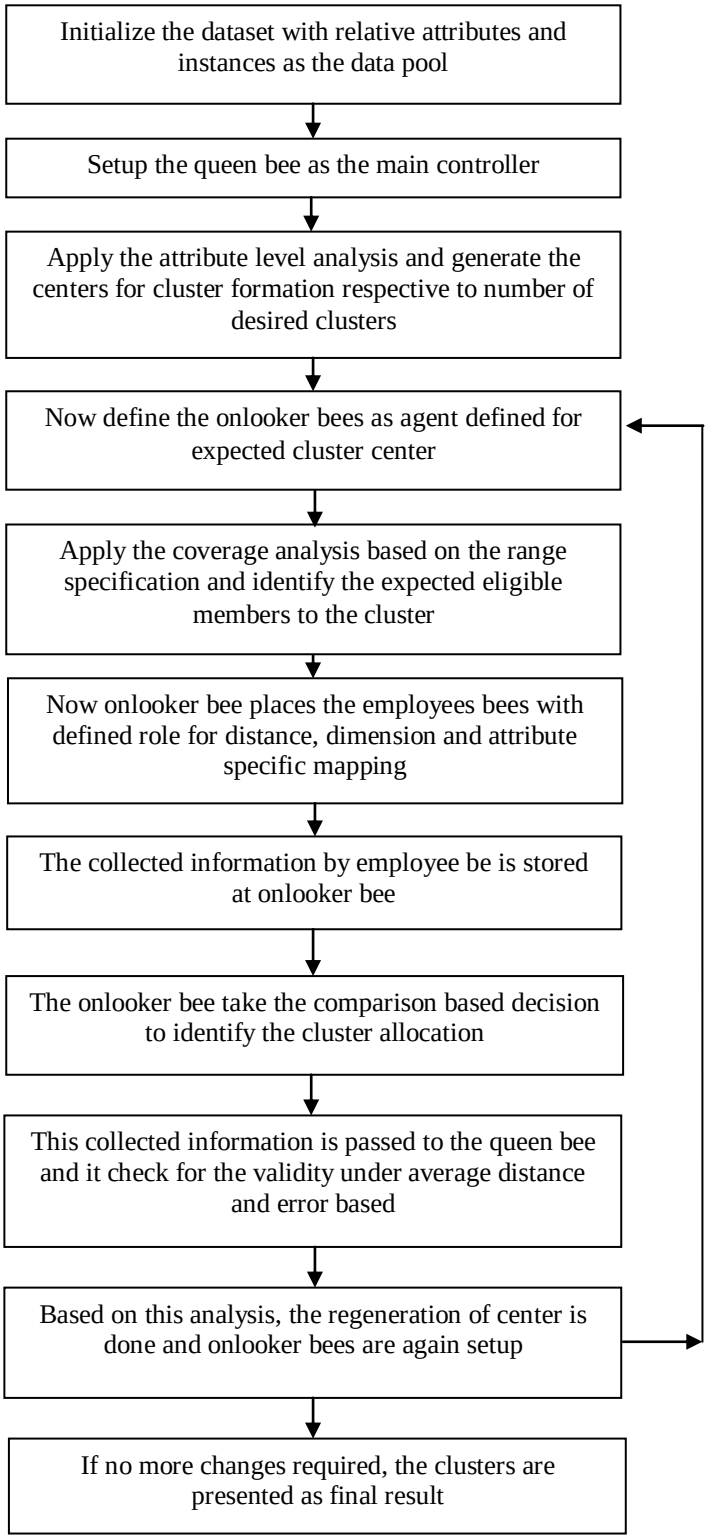

Figure 1 : Proposed Cluster Optimization Method

Here figure 1 has shown the proposed clustering method using bee colony optimization. The figure has shown each of the process stage and its evaluation under error specific analysis. The method is later on applied on two different sample sets and the analytical results are obtained under MSE and MAE parameters. The results obtained from the work are described in next section.

IV. RESSULTS

In this paper, an ABC improved clustering method is proposed to generate the clusters. The method is implemented in matlab environment and for two different datasets. The datasets are here defined of numerical data with specification of

different number of attributes and different dimensions. The evolution of the clustering method is here done under MSE(Mean Square Error) and MAE (Mean Absolute Error). The comparative analysis is here generated against the CMeans Clustering method. The comparative evaluated results are defined in this section.

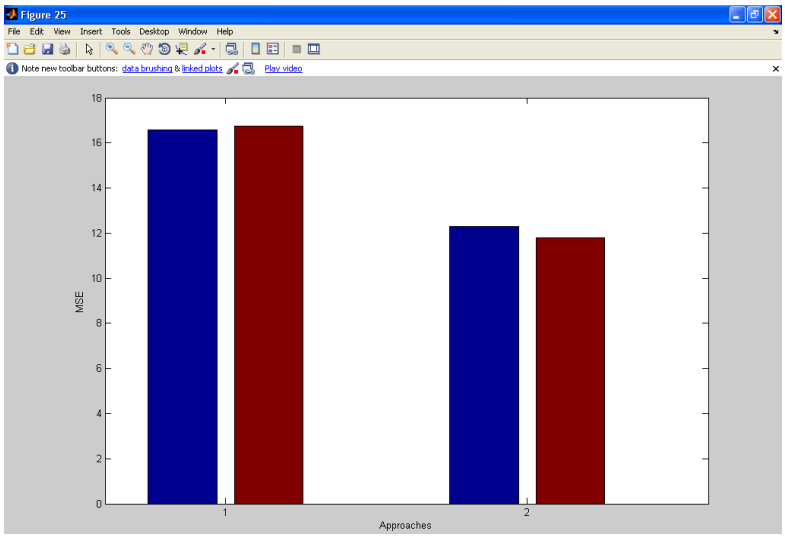

Figure 2 : MSE Analysis (Existing Vs. Proposed)

Here figure 2 has showed the MSE analysis of this presented method for generated two clusters. Here x axis shows the cluster index and y axis shows the MSE estimation. The results shows that the proposed method has reduced the error in cluster formation.

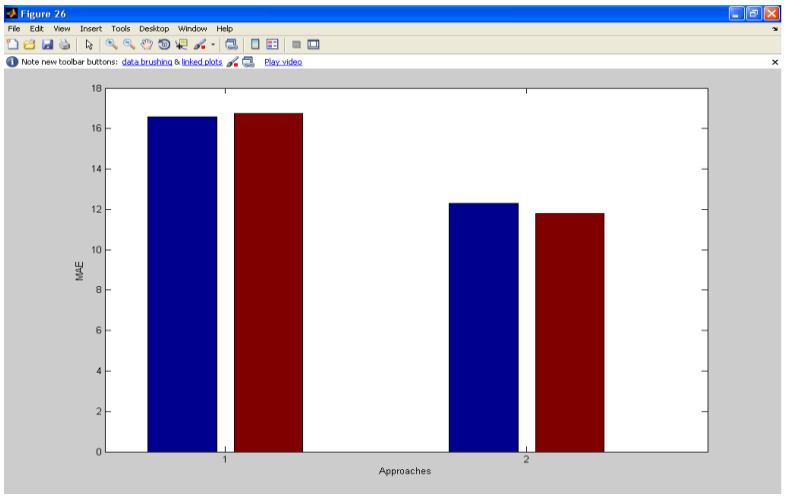

Figure 3: MAE Estimation (Existing Vs. Proposed)

Here figure 3 shows the MAE based estimation generated for existing and proposed work. The figure shows that the proposed work model has improved the clustering reliability by reducing the MAE. The method has formed the clusters in more effective way.

CONCLUSION

Data Clustering is having the significance as the independent process method as well as an intermediate pre-processing stage. In this present work, an improved clustering method using Artificial Bee colony optimization is provided. The generated results shows that the method has reduce the error rate and provided more effective cluster formation.

REFERENCES

[1]. V. T. N. Chau, "A Robust Self-Organizing Approach to Effectively Clustering Incomplete Data," Knowledge and Systems Engineering (KSE), 2015 Seventh International Conference on, Ho Chi Minh City, 2015, pp. 150-155.

[2]. Yogita and D. Toshniwal, "Clustering techniques for streaming data-a survey," Advance Computing Conference (IACC), 2013 IEEE 3rd International, Ghaziabad, 2013, pp. 951-956.

[3]. H. Venkateswara Reddy, P. Agrawal and S. Viswanadha Raju, "Data labeling method based on cluster purity using relative rough entropy for categorical data clustering," Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on, Mysore, 2013, pp. 500-506.
[4]. S. S. R. Abidi and J. Ong, "A data mining strategy for inductive data clustering: a synergy between self-organising neural networks and K-means clustering techniques," TENCON 2000. Proceedings, Kuala Lumpur, 2000, pp. 568-573 vol.2.
[5]. V. T. N. Chau, P. H. Loc and V. T. N. Tran, "A Robust Mean Shift-Based Approach to Effectively Clustering Incomplete Educational Data," 2015 International Conference on Advanced Computing and Applications (ACOMP), Ho Chi Minh City, 2015, pp. 12-19.
[6]. Y. F. Wang, Z. G. Yu and V. Anh, "Fuzzy C-means method with empirical mode decomposition for clustering microarray data," Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on, Hong Kong, 2010, pp. 192-197.
[7]. Y. Yang and K. Chen, "Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations," in IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 2, pp. 307-320, Feb. 2011.
[8]. S. Alam, G. Dobbie, Y. S. Koh and P. Riddle, "Clustering heterogeneous web usage data using Hierarchical Particle Swarm Optimization," Swarm Intelligence (SIS), 2013 IEEE Symposium on, Singapore, 2013, pp. 147-154.
[9]. X. Liu, M. Yin, J. Luo and W. Chen, "An improved affinity propagation clustering algorithm for large-scale data sets," 2013 Ninth International Conference on Natural Computation (ICNC), Shenyang, 2013, pp. 894-899.
[10]. Y. Xie, A. Wulamu, Y. Wang and Z. Liu, "Implementation of time series data clustering based on SVD for stock data analysis on hadoop platform," 2014 9th IEEE Conference on Industrial Electronics and Applications, Hangzhou, 2014, pp. 2007-2010.
[11]. R. N. Dave and S. Sen, "Robust fuzzy clustering of relational data," in IEEE Transactions on Fuzzy Systems, vol. 10, no. 6, pp. 713-727, Dec 2002.
[12]. H. Prasetyo and A. Purwarianti, "Comparison of distance and dissimilarity measures for clustering data with mix attribute types," Information Technology, Computer and Electrical Engineering (ICITACEE), 2014 1st International Conference on, Semarang, 2014, pp. 276-280.
[13]. H. V. Reddy, B. S. Kumar and S. Viswanadharaju, "A Data Labeling Method for Categorical Data Clustering Using Cluster Entropies in Rough Sets," Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on, Bhopal, 2014, pp. 444-449.
[14]. K. Hua-Ai, "Method of Data Clustering Incomplete Fill Based on Constraint Tolerance Set Dissimilarity," Intelligent Systems Design and Engineering Applications (ISDEA), 2014 Fifth International Conference on, Hunan, 2014, pp. 615-620.
[15]. C. F. Tsai and S. C. Huang, "An effective and efficient grid-based data clustering algorithm using intuitive neighbor relationship for data mining," Machine Learning and Cybernetics (ICMLC), 2015 International Conference on, Guangzhou, 2015, pp. 478-483.
[16]. Y. Wang, L. Chen and J. P. Mei, "Incremental Fuzzy Clustering With Multiple Medoids for Large Data," in IEEE Transactions on Fuzzy Systems, vol. 22, no. 6, pp. 1557-1568, Dec. 2014.
[17]. Y. Peng, Q. Luo and X. Peng, "UCK-means :A customized K-means for clustering uncertain measurement data," Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on, Shanghai, 2011, pp. 1196-1200.