Global Scientific JOURNALS

GSJ: Volume 6, Issue 9, September 2018, Online: ISSN 2320-9186
www.globalscientificjournal.com

LION OPTIMIZATION ALGORITHM USING DATA MINING CLASSIFICATION AND CLUSTERING MODELS

S.Sowmiyasree [#1]

M.Phil Full-Time / Research Scholar

Dr.P.Sumitra [#2]

Assistant Professor

PG and Research Department of Computer Science and Applications

Vivekanandha College of Arts And Sciences for Women (Autonomous)

Elayampalayam

ABSTRACT

Spectral clustering has been a popular data clustering algorithm. This category of approaches often resort to other clustering methods, such as K-Means, to get the final cluster. The potential flaw of such common practice is that the obtained relaxed continuous spectral solution could severely deviate from the true discrete solution. In this paper we propose to impose an additional orthonormal constraint to better approximate the optimal continuous solution to the graph cut objective functions. Such a method, called spectral rotation .In literature we optimizes the spectral clustering objective functions better than K-Means, and improves the clustering accuracy. In many applications, data objects are described by both numeric and categorical features. The K-Mean++ algorithm is one of the most important algorithms for clustering. However, this method performs hard partition, which may lead to misclassification for the data objects in the boundaries of regions, and the dissimilarity measure only uses the user-given parameter for adjusting the significance of attribute. In this paper, first, we combine mean and K-Mean++ centroid to represent the prototype of a cluster, and employ a new analysis based on co-occurrence of values to survey the dissimilarity between data objects and prototypes of clusters. This survey also takes into account the significance of different attributes towards the clustering process. Then we present our LOA algorithm for clustering mixed data. Finally, the performance of the different method is analysis by a series of real world datasets in comparison with that of traditional clustering algorithms.

Keywords – Classification Model, LOA, KNN, K-Prototype, Clustering Process

I. INTRODUCTION

Data mining is process of analyzing of bulk amount of data to automatically discover the interesting regularities or associations which in turn lead to improved understanding of the original processes. To find the useful classes or patterns using decision making. There are two categories of data mining are:

- Descriptive Data mining
- Predictive Data mining

Descriptive data mining, it generalizes or summarizes the general properties of the data in the database. Predictive data mining is searched to inference on the present data to make predictions .Classification maps data into predefined groups, it is often referred to as supervised learning as the classes are determined prior to examining the data. Classification algorithms usually require that the classes be defined based on the data attribute values. Classification is the technologies used for classify the data and predict the accuracy for the future work with the use of behind and present data. Clustering is an automatic learning technique aimed at grouping a set of objects into subsets or clusters. The goal is to create clusters that are coherent internally, but substantially different from each other. In plain words, objects in the same cluster should be as similar as possible, whereas objects in one cluster should be as dissimilar as possible from objects in the other clusters. Automatic document clustering has played an important role in many fields like information retrieval, data mining, etc. The aim of this thesis is to improve the efficiency and accuracy of document clustering. In this proposed system two clustering algorithms and the fields where these perform better than the known standard clustering algorithms.

In clustering, it is the distribution and the nature of data that will determine cluster membership, in opposition to the classification where the classifier learns the association between objects and classes from a so called training set, i.e. a set of data correctly labeled by hand, and then replicates the learnt behavior on unlabeled data.

In this paper, we survey a different type clustering algorithm to cluster these types of data. In proposed method, we implement the different centroid and mean to represent the prototype of a cluster, and use a new measure to evaluate the dissimilarity between data objects and the prototype of a cluster. In comparison with other algorithm, our algorithm has two main contributions:

- Firstly, by using the different clustering centroid survey analysis can preserve the uncertainty inherence in data sets for longer time before decisions are made, and are therefore less prone to falling into local optima in comparison with other clustering algorithms.
- Secondly, survey analysis takes account of the significance of different attributes towards clustering by using the new measure to evaluate the dissimilarity between the data objects and the cluster's prototype

II. LITERATURE REVIEW

MAZIAR YAZDANI et al [1] describe a new optimization algorithm that is called Lion Optimization Algorithm (LOA), is introduced. LOA is constructed based on simulation of the solitary and cooperative behaviors of lions such as prey capturing, mating, territorial marking, defense and the other behaviors. In order to evaluate performance of the introduced algorithm, we have tested it on a set of various standard benchmark functions. The results obtained by LOA in most cases provide superior results in fast convergence and global optima achievement and in all cases are comparable with other meta heuristics.

G. JAGATHEESHKUMAR et al [2] describe a new clustering algorithm for textual documents, which is a combination of K - means and LOA. The proposed clustering algorithm is applied over the pre - processed text documents, from which the text clusters are formed. Initially, the LOA is utilized for selecting the better cluster centre points and the k - means perform the clustering operation by refining the already formed clusters. F-measure, purity and entropy. The experimental results attained by kLOA are compared with the existing techniques and the proposed kLOA outperforms the existing algorithms. In future, the quality of clusters is planned to be increased further by feeding semantic knowledge to the algorithm.

NAVNEET et al [3] develops a novel big data classification algorithm based on a nature inspired meta-heuristic algorithm (lion optimization algorithm). Lion optimization algorithm is an optimization algorithm base d on the hunting and social behavior of the lion. The developed algorithm uses the K-mean clustering to generate the pride and nomad. The performance of the proposed algorithm CALOA is also compared and has been found better than other algorithms of literature i.e. J48, ACO, PSO and DE. In future the algorithm can be extended for regression purpose. Moreover, the proposed CALOA algorithm can be analyzed for various applications domains to handle different types of big datasets.

MING-YI SHIH et al [4] describe a new two-step clustering method is presented to find clusters on this kind of data. In this approach the items in categorical attributes are processed to construct the similarity or relationships among them based on the ideas of co-occurrence; then all categorical attributes can be converted into numeric attributes based on these constructed relationships. Finally, since all categorical data are converted into numeric, the existing clustering algorithms can be applied to the dataset without pain. Furthermore, the number of subset i is set to one-third of number of objects in this paper. Although experimental results show that it is feasible, how to set this parameter precisely is worth more study in the future.

IZHAR AHMAD et al [5] presented a system design approach for the K - Mean and K - Prototype Algorithms Performance Analysis. The system architecture in this research is presents a detail discussion of the k - means and k - prototype to recommend efficient algorithm for outlier detection and other issues relating to the database clustering. The tracking and monitoring application will have local database which will store all the information of the user and their groups including the privileges. After the

analyses are performed, the information is transferred to the result outcome.

This proposed solution method ensures that the initial centroids and k values are generated depend on the distribution of the data received from the data grid. These proposed system architecture presents the better accuracy compared to the original k - means and k - prototype algorithm.

JINCHAO JI et al [6] describe a fuzzy c-mean type clustering algorithm to cluster these type of data. In our method, we integrate the fuzzy centroid and mean to represent the prototype of a cluster, and use a new measure to evaluate the dissimilarity between data objects and the prototype of a cluster. In comparison with other algorithm, our algorithm has two main contributions: Firstly, by using the fuzzy centroid our algorithm can preserve the uncertainty inherence in data sets for longer time before decisions are made, and is therefore less prone to falling into local optima in comparison with other clustering algorithms. Secondly, our algorithm takes account of the significance of different attributes towards clustering by using the new measure to evaluate the dissimilarity between the data objects and the cluster's prototype.

III. RESEARCH PROBLEM

Organizing data into sensible groupings arises naturally in many scientific fields. It is, therefore, not surprising to see the continued popularity of data clustering. It is important to remember that cluster analysis is an exploratory tool; the output of clustering algorithms only suggest hypotheses. While numerous clustering algorithms have been published and new ones continue to appear, there is no single clustering algorithm that has been shown to dominate other algorithms across all application domains.

A clustering method that satisfies the requirements for one group of users may not satisfy the requirements of another. As mentioned earlier, ''clustering is in the eye of the beholder" – so indeed data clustering must involve the user or application needs. Clustering has numerous success stories in data analysis. In spite of this, machine learning and pattern recognition communities need to address a number of issues to improve our understanding of data clustering.

Below is a list of problems and research directions that are worth focusing in this regard.

- There needs to be a suite of benchmark data (with ground truth) available for the research community to test and evaluate clustering methods.
- We need to achieve a tighter integration between clustering algorithms and the application needs.
- A fundamental issue related to clustering is its stability or consistency. A good clustering principle should result in a data partitioning that is stable with respect to perturbations in the data.
- Choose clustering principles according to their satisfiability of the stated axioms. Despite Kleinberg's impossibility theorem, several studies have shown that it can be overcome by relaxing some of the axioms.
- Given the inherent difficulty of clustering, to decide both (i) data representation and (ii) appropriate objective function for data clustering.

IV. METHODOLOGY

A. K Means Clustering

K-means clustering is a data mining the machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The K-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields. The cluster's mean is then recomputed and the process begins again. Here's how the algorithm works:

- The algorithm arbitrarily selects k points as the initial cluster centers ("means").
- Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
- Each cluster center is recomputed as the average of the points in that cluster.
- Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

B. K-Mean++

K-means++ is the most important flat clustering algorithm. The objective function of K- means is to minimize the average squared distance of objects from their cluster centers, where a cluster center is defined as the mean or centroid μ of the objects in a cluster C:

The ideal cluster in K-means is a sphere with the centroid as its center of gravity. Ideally, the clusters should not overlap. A measure of how well the centroids represent the members of their clusters is the Residual Sum of Squares (RSS), the squared distance of each vector from its centroid summed over all vectors

- Reassigning objects to the cluster with closest centroid
- Recomputing each centroid based on the current members of its cluster.

The following termination conditions as stopping criterion for using termination process

1. A fixed number of iterations I has been completed.
2. Centroids μ i do not change between iterations.
3. Terminate when RSS falls below a pre-established threshold.

Algorithms

Step 1. procedure KMEANS(X,K)
Step 2. {s1, s2, · · · , sk}
SelectRandomSeeds(K,X)
Step 3. for i ← 1,K do
Step 4. μ(Ci) ← si
Step 5. end for
Step 6. repeat
Step 7 min k~x n −~μ(C k)k C k = C k [{~x n }
Step 8. for all C k do
Step 9. μ(C k) = 1
Step 10. end for
Step 11. until stopping criterion is met
Step 12. end procedure

The proposed algorithm fall within a subcategory of the flat clustering algorithms, called Model-based clustering. The model-based clustering assumes that data were generated by a model and then tries to recover the original model from the data. This model then defines clusters and the cluster membership of data. The proposed algorithm is a generalization of K-Means algorithm in which the set of K-centroids as the model that generate the data. It alternates between an expectation step, corresponding to reassignment, and a maximization step, corresponding to re computation of the parameters of the model.

C. LION OPTIMIZATION

The LOA is a meta heuristic algorithm that imitates the behavior of lions. The initial population of lions is created and the percentage of resident and nomadic lions is specified. The lioness from each pride is sent for hunting its prey. The fitness of the lion is computed and the probability of success in hunting the prey is computed. Similarly, for ach nomadic lion the prey is generated and the fitness is computed. The lions with the least fitness value are removed.

The initial population of lions is created and the percentage of resident and nomadic lions is specified. The lioness from each pride is sent for hunting its prey. The fitness of the lion is computed and the probability of success in hunting the prey is computed. Similarly, for ach nomadic lion the prey is generated and the fitness is computed. The fitness value is computed by means of the following equation.

$$f_i=\sum_{i=1}\sum_{dm \in Ccem} |d_m-C_{cem}||^2 \quad \text{---- 1}$$

$$S_p=f_i/\sum_{i=1}\sum f_n \quad \text{----2}$$

The success probability is computed by taking the fitness value of the i^{th} data and n which is the total count of cluster centroids. In equation (1), d_m is the document, C_{cen} is the centroid of the cluster and C is the total number of clusters. This fitness value ranges from 0 to 1 and it indicates the similarity between the textual data and the centroid of the cluster. The pride of lions is arranged in ascending order with respect to the fitness value and the lions with minimal similarity are discarded. K-means algorithm is applied to enhance the so formed cluster.

The distance between the textual data and the cluster centre is found out and the data with minimal distance must be considered as a part of the cluster. As the maximum candidates of a cluster are fixed, the relevant data are grouped together. By this way, the initial cluster points are selected by the LOA and the formed clusters are enhanced by the k-means algorithm. Hence, all the textual data in the dataset are allotted to the most relevant data cluster.

Algorithm-
Input : Ndata
Output : Nclusters

Step 1 : Begin
For all
Preprocess the data;
Step 2 : Initialize the population of lions and other parameters;

Step 3 : Distribute the prey;
Step 4 : For each pride of lions
Step 5: Randomly select a lioness for hunting;

Step 6: Compute the fitness of the lions by eqn.1;
Step 7: Compute the probability of success by eqn.2;
Step 8: Arrange the fitness values in ascending order;
Step 9: Disregard the lions with greater value;
Step10: Apply k-means and greedy selection;
End for;
Step 11: For each nomadic lion
Perform the same steps as in pride of lion;
Step 12: Compute fitness value;
Step 13: Compare the fitness value of nomadic lion and the pride of lion;
Step 14: If (fitness(nomadic lion) < fitness(lion in pride) Exchange the lion to pride;
Step 15: Discard the nomadic lions with greater value;
Step 16: Apply k-means and greedy selection;
Step 17: Compare and swap the lions if better value is
found;
Step 18 : Save the best solution;
Step 19: End if;
Step 20: End for;
Step 21: End;
Step 22: While(termination condition not met)

V. RESULTS AND DISCUSSION

The following **Fig 5.1** describes experimental result for existing K-Mean++ algorithm analysis. The table contains weight of text document, weight of clustering Text document and average details

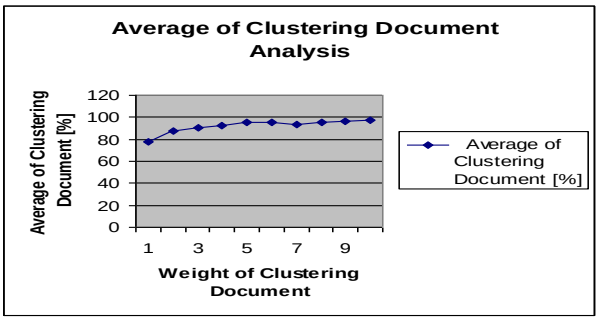

Fig 5.1 Existing system-Average Clustering Documents

The following **Fig 5.2** describes experimental result for K-Mean++ and K-LOA system analysis. The figure contains weight of text document, cluster document with R, G, and B document clustering details are shown

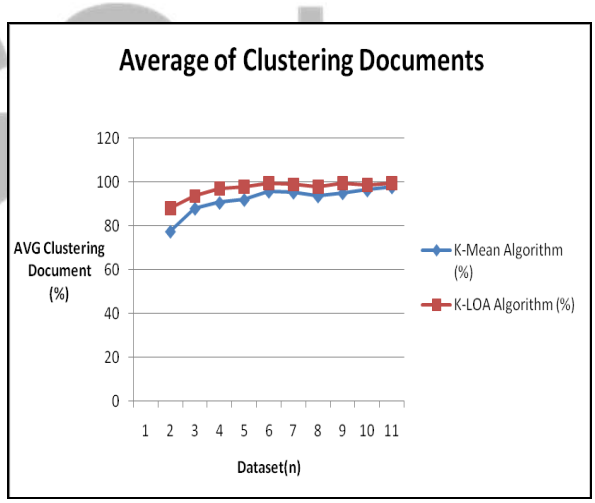

Fig 5.2 Proposed systems - Average of Clustering Documents

V. CONCLUSION

This paper has survey worked on data clustering by using lion optimization algorithm The analysis algorithm has used k-mean clustering with Schwarz criteria for the initiation purpose. The prides formed by the algorithm is classified by using the LOA algorithm. The comparison of the algorithm with other cluster based classification algorithm by using the parameter like accuracy, recall etc. shows the significant improvement over other algorithms.

REFERENCES

1. Yazdani, M. and Jolai, F."Lion Optimization Algorithm (LOA): A nature-inspired metaheuristic algorithm", Journal of Computational Design and Engineering, 3(1), pp.24-36, 2016.
2. Xiaofen Lu and Ke Tang, Xin Yao,Classification-Assisted Differential Evolution for Computationally Expensive Problems, IEEE Congress on Evolutionary Computation (CEC), 2011./
3. Navneet, Nasib Singh Gill, "A Novel Algorithm For Big Data Classification Based On Lion Optimization, IEEE Congress on Evolutionary", IEEE, Computation (CEC), 2017.
4. Maziar Yazdani and Fariborz Jolai, "Lion Optimization Algorithm Based K- Means For Textual Data Clustering", International Journal of Pure and Applied Mathematics, Volume 117 No. 22 2017, 167-171.
5. .Jagatheeshkumar and Dr. S. Selva brunda, "Lion Optimization Algorithm (LOA): A nature-inspired meta heuristic algorithm ,"Journal of Computational Design and Engineering 3 (2016) 24–36.
6. Ming-Yi Shih*, Jar-Wen Jheng and Lien-Fu Lai, "A Two-Step Method For Clustering Mixed Categroical And Numeric Data",Tamkang Journal of Science and Engineering, Vol. 13, No. 1, pp. 11-19 (2014).
7. Izhar Ahmad, "K- Mean And K-Prototype Algorithms Performance Analysis ",American Review of Mathematics and Statistics, March 2014, Vol. 2, No. 1,pp. 95-109.
8. Jinchao Ji and Wei Pang, "A Fuzzy K-Prototype Clustering Algorithm For Mixed Numeric And Categorical Data", Elsevier, Knowledge-Based Systems 30 (2012) 129–135.