IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

CLQ-CS: An Innovative Subspace Clustering Technique for Dimensionality Reduction

Behera Gayathri[1]

Assistant Professor, CSE Department, GITAM Institute of Technology, GITAM University, Visakhapatnam, India [1]

Abstract: High dimensional data clustering is an emerging research field as; it is becoming a major challenge to cluster high dimensional data due to the high scattering of data points. Most of the traditional clustering methods are not that appropriate to handle high dimensional data. In this paper, a new algorithm "CLQ-CS" is proposed based on the subspace clustering algorithm called CLIQUE and the Cuckoo Search strategy. The proposed "CLQ-CS" algorithm consists of two phases. The data pre-processing is the first phase where CLIQUE is used for subspace relevance analysis to find the dense subspaces. In the second phase a global search strategy called cuckoo search is introduced to cluster the subspaces detected in the first phase. The problem of losing some of the regions that are actually densely populated due to the high scattering of data points in high-dimensional space can be overcome. The experiments performed on large and high dimensional synthetic and real world datasets demonstrate that CLQ-CS performs with a higher efficiency and better resulting cluster accuracy. Moreover, the proposed algorithm not only yields accurate results when the number of dimensions increases but also outperforms the individual algorithms when the size of the dataset increases.

Keywords: CLIQUE, Cuckoo Search, High Dimensional Data, Subspace Clustering, CLQ-CS.

I. INTRODUCTION

Data mining refers to extracting or "mining" knowledge from large amounts of data. The main goal of the data mining process is to extract information from a dataset and transform it into an understandable structure for further use. Data mining is an essential step in the process of knowledge discovery. Knowledge discovery consists of an iterative sequence of the steps data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation [1].

Cluster analysis is a very important technique in data mining. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. Clustering is the process of dividing different sets of data points into different clusters in such a way that the points within a cluster are more similar to each other than those in different clusters. Clustering is considered as a part of unsupervised learning and there are different types of clustering techniques available like hierarchical based clustering, partitioned based clustering, density based clustering and grid based clustering.

Subspace clustering is the task of detecting groups of clusters within different subspaces of the same dataset. There are several problems which are encountered in clustering of high-dimensional data [2]. Traditional clustering algorithms can usually obtain more accurate results in general low-dimensional data clustering but does not often get the desired clustering results in high-dimensional data clustering due to the impact of the 'curse of dimensionality'. When the number of dimensions increases, the distance between any two points in a given dataset converges. The discrimination of the nearest and farthest point becomes difficult and hence the concept of distance becomes meaningless. In high dimensional data, most of the dimensions could be irrelevant and can mask existing clusters in noisy data. Hence subspace clustering algorithms localize the search for relevant dimensions allowing them to find clusters that exist in multiple subspaces of the dataset.

II. RELATED WORK

A. Cuckoo Breeding Behaviour

The Cuckoo breeding behaviour is interesting because of their aggressive reproduction strategy. A number of species can be seen laying their eggs in the nests of other host birds (often other species). There are three basic types of brood parasitism: intra-specific brood parasitism, cooperative breeding, and nest takeover. Some host birds can engage direct conflict with the intruding cuckoos. If a host bird discovers that the eggs are not their own, they will either throw these alien eggs away or simply abandon its nest and build a new nest somewhere else. Some of the female cuckoos of the New World brood-parasitic Tapera species are specialized in the mimicry of the egg pattern and colour of a few chosen host species [3]. Hence the probability of their eggs being abandoned reduces and their reproductivity increases.

The timing of egg-laying of some species is also interesting. In general, the cuckoo eggs hatch slightly earlier than their host eggs; this is the reason why, a nest where the host bird just laid its own eggs is often chosen

DOI 10.17148/IJARCCE.2016.51158

by parasitic cuckoos. Once the first cuckoo chick is hatched, the host eggs are evicted by blindly propelling them out of the nest. This increases the cuckoo chick's share of food provided by its host bird. Studies also show that a cuckoo chick can also mimic the call of host chicks to gain access to more feeding opportunity.

B. L´evy Flights

Various studies have shown that flight behaviour of many animals and insects has demonstrated the typical characteristics of L´evy flights [4, 5]. Some of the recent studies show that fruit flies or Drosophila melanogaster, take a series of straight flight paths with a sudden 90^o turn to explore their landscape leading to an L´evy-flight-style intermittent scale free search pattern. Even light can be related to L´evy flights. Such behaviour has been applied to optimization and optimal search, and preliminary results show its promising capability. [6,7]

C. Cuckoo Search

The Cuckoo Search, uses the following three idealized rules:

1) Each cuckoo lays one egg at a time and dumps its egg in a randomly chosen nest;
2) The best nests with high quality of eggs will carry over to the next generations;
3) The number of available host nests is fixed and the egg laid by a cuckoo is discovered by the host bird with a probability p_a [0, 1].

In this case, the host bird can either throw the egg away or abandon the nest, and build a completely new nest. Each egg in a nest represents a solution and a cuckoo egg represents a new solution. The aim is to use the new and potentially better solutions (cuckoos) to replace a not-so-good solution in the nests. For this present work each nest is considered to have only a single egg.

D. Subspace clustering algorithms

The subspace clustering algorithms are of two types based on the searching technique as top-down search and bottom-up search methods [8]. CLIQUE [9] algorithm is a combination of density and grid based clustering and it uses an Apriori approach to find clusterable subspaces. Once the dense subspaces are found they are sorted by coverage. The coverage is defined as the fraction of the dataset covered by the dense units in the subspace. The subspaces with the greatest coverage are kept and the rest are pruned. ENCLUS [10] is a subspace clustering method which is based on the CLIQUE algorithm. ENCLUS introduces the concept of 'Subspace Entropy'. The algorithm is based on the observation that, a subspace with clusters typically has lower entropy than a subspace without clusters. The lower is the entropy value the more is the probability of having clusters within a subspace. MAFIA [11] is another extension of CLIQUE that uses an adaptive grid based on the distribution of data to improve efficiency and cluster quality; it also introduces parallelism to improve scalability. Cell-based Clustering (CBF) [12] addresses scalability issues associated with many bottom-up algorithms.CBF uses a cell creation algorithm that creates optimal partitions by repeatedly examining minimum and maximum values on a given dimension which results in the generation of fewer bins (cells). CLTree [13] uses a modified decision tree algorithm to select the best cutting planes for a given dataset. It uses a decision tree algorithm to partition each dimension into bins, separating areas of high density from areas of low density.

Density-based Optimal projective Clustering (DOC) is a type of hybrid method of the grid based approach used by the bottom up approaches and the iterative improvement method from the top-down approaches [14]. PROCLUS [15] was the first top down subspace clustering algorithm. Similar to CLARANS [16], PROCLUS samples the data, then selects a set of k medoids and iteratively improves the clustering. The algorithm uses a three phase approach consisting of initialization, iteration, and cluster refinement. ORCLUS [17] is an extended version of the algorithm PROCLUS which looks for non-axis parallel subspaces. This algorithm arises from the observation that many datasets contain inter-attribute correlations. The algorithm is divided into three steps: assign clusters, subspace determination, and merge. A Fast and Intelligent Subspace Clustering Algorithm uses Dimension Voting, FINDIT [18] is similar in structure to PROCLUS and the other top-down methods, but uses a unique distance measure called the Dimension Oriented Distance (DOD). The algorithm mainly consists of three phases, namely sampling phase, cluster forming phase, and data assignment phase. δ-clusters algorithm uses a distance measure that attempts to capture the coherence exhibited by subset of instances on subset of attributes [19]. The algorithm takes the number of clusters and the individual cluster size as parameters. Clustering On Subsets of Attributes (COSA) [20] is an iterative algorithm that assigns weights to each dimension for each instance, not each cluster. Starting with equally weighted dimensions, the algorithm examines the k nearest neighbours (knn) of each instance. These neighbourhoods are used to calculate the respective dimension weights for each instance.

III.CLQ-CS ALGORITHM

Phase 1: Pre-processing using CLIQUE

The first algorithm which was proposed for dimension-growth subspace clustering in high dimensional space was the CLIQUE (CLustering In QUEst) [21] algorithm. The process of clustering starts from single-dimensional subspaces and grows upward to higher-dimensional ones when the dimension-growth subspace clustering is considered. CLIQUE can be considered as an integration of density-based and grid-based clustering methods. In the proposed clustering technique CLIQUE is used in the pre-processing phase for subspace relevance analysis to find the dense subspaces. This process involves the following steps:

DOI 10.17148/IJARCCE.2016.51158

Step 1: In the first step, CLIQUE partitions the d-dimensional data space into non-overlapping rectangular units.
Step 2: Identifying the subspaces that are dense.
Step 2.1: Identifying the dense units
1. The zero density valued dimensions are discarded.
2. The threshold value is calculated, which is the immediate density value which is greater than the smallest density value of all dimensions.
3. The dimensions whose density values does not satisfy the threshold value are been removed since, a unit is considered as dense only if the fraction of total data points contained in it exceeds the threshold value.
Step 2.2: Identifying the Subspaces of High Coverage
The dimensions which are having dense units can be considered as n-dimensional subspaces of high coverage, where n is the number of dimensions which are dense.

Phase2: Cuckoo Search via L´evy Flights

```
begin
Objective function $f(x)$, $x = (x_1, ..., x_d)^T$
Generate initial population of “n” host nests $x_i$ (i = 1,2,. .n)
while (t < MaxGeneration) or (stop criterion)
        Get a cuckoo randomly by L´evy flights
        evaluate its quality/fitness $F_i$
        Choose a nest among n (say, j) randomly
      if ($F_i > F_j$),
         replace j by the new solution;
      end
   A fraction ($p_a$) of worse nests
         are abandoned and new ones are built;
   Keep the best solutions (or nests with quality solutions);
   Rank the solutions and find the current best
end while
Postprocess results and visualization
end
```

The phase 2 implements the cuckoo search algorithm via L’evy flights .The algorithm begins with the objective function f(x). An initial population x_i (i = 1, 2, ..., n) is generated where “n” represents the number of host nests. If the condition t<MaxGeneration is true then a cuckoo is randomly chosen by L’evy flights. The fitness of the randomly chosen cuckoo is calculated and considered as F_i. Then a nest is randomly chosen among “n” nests which is considered as “j”. If $F_i > F_j$, then we can replace j with the new solution. The egg laid by a cuckoo is discovered by the host bird with a probability $p_a \in [0,1]$. A fraction p_a of the n nests are replaced by the new nests. The best solutions are been retained. Each time, the solutions are ranked and the current best solution is found. While generating new solutions $x^{(t+1)}$ for, say, a cuckoo ‘i’, a L´evy flight is performed

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \oplus \text{L'evy}(\lambda)$$

where α > 0 is the step size which should be related to the scales of the problem of interests. In most cases, we can use α=1. The product ⊕ means entrywise multiplications. The L´evy flight essentially provides a random walk while the random step length is drawn from a L´evy distribution

$$\text{L´evy} \sim u = t^{-\lambda} \qquad (1 < \lambda \leq 3)$$

which has an infinite variance with an infinite mean.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

CLQ-CS has been implemented on Zoo [22] dataset which is a real dataset from the UCI machine learning repository and electoral poll dataset which is a synthetic dataset prepared to test this algorithm. Validating the results of clustering is very important. Cluster Accuracy [23] is used to evaluate the obtained clustering results.
The evaluation metric used in the present clustering technique is given below

$$\text{Cluster Accuracy} = \frac{1}{N}\sum_{i=1}^{T} X_i$$

where ,
N = Number of data points in the dataset
T = Number of resultant clusters
X_i = Number of data points of majority class in cluster i

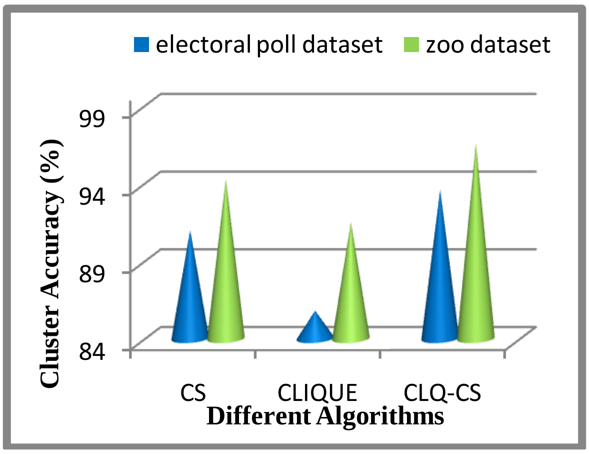


Fig.1: Comparison of Cluster Accuracy for different algorithms

The graph in fig.1 shows that the cluster accuracy for CS (Cuckoo Search) is more compared to CLIQUE, but CLQ-CS algorithm exceeds both CS and CLIQUE algorithms.

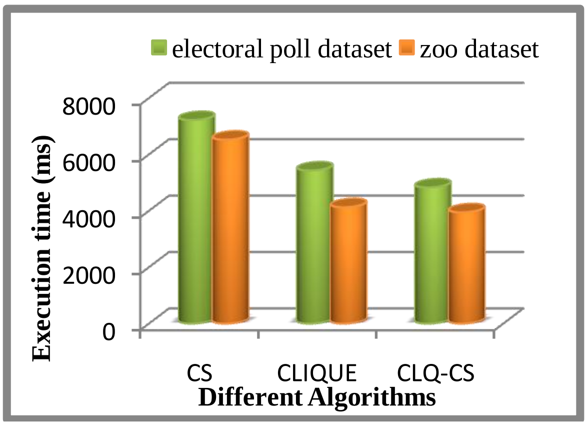


Fig. 2: Comparison of Execution Time for Different Algorithms

The graph in fig.2 shows that the execution time for CLIQUE is less compared to CS, but the execution time for CLQ-CS algorithm is even much lesser than the CLIQUE algorithm.

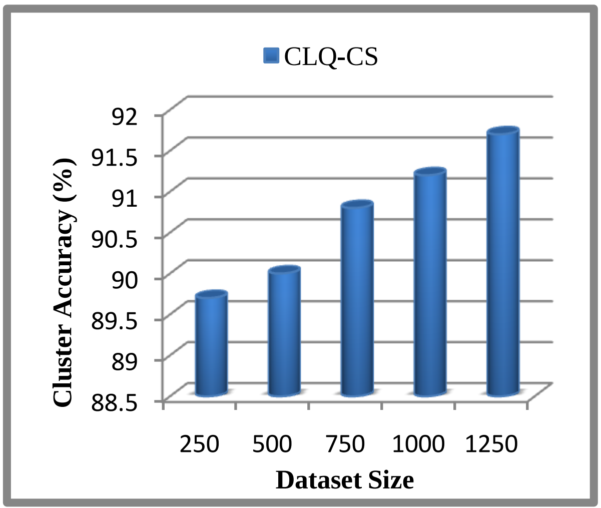
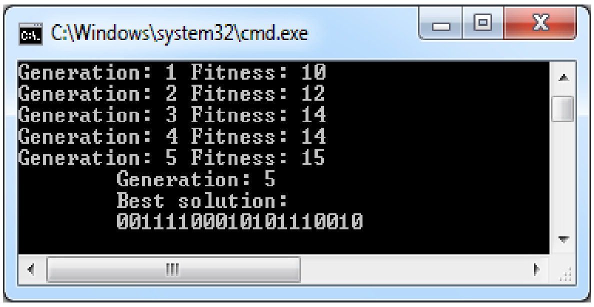


Fig. 3: Scalability with Dataset Size

The graph in fig.3 shows that the cluster accuracy for CLQ-CS algorithm increases when the dataset size increases from 250 to 500. The cluster accuracy increases even when the dataset size is increased from 500, 750, 1000 up to 1250. This shows that CLQ-CS algorithm is highly scalable with increasing sizes of data.



```
Generation: 1 Fitness: 10
Generation: 2 Fitness: 12
Generation: 3 Fitness: 14
Generation: 4 Fitness: 14
Generation: 5 Fitness: 15
        Generation: 5
        Best solution:
        0011110001010111001 0
```

Fig. 4: Fitness of each generation and the best solution

The fig. 4 shows the fitness value of each generation. The fittest generation is the generation 5 with fitness value 15. The best solution is also been displayed. The solution which is the best is kept and others are been discarded or abandoned.

V. CONCLUSION

In this paper a CLQ-CS clustering technique is proposed based on CLIQUE subspace clustering algorithm and cuckoo search strategy. The proposed “CLQ-CS” technique consists of two phases in which the first phase performs pre-processing of data using CLIQUE to find the dense subspaces. In the second phase a global search strategy called cuckoo search via L´evy flights is implemented to cluster the subspaces detected in the first phase. The experiments performed on large and high dimensional synthetic and real world data sets show that the current CLQ-CS technique performs with a higher efficiency and better resulting cluster accuracy. Moreover, the algorithm not only yields accurate results when the number of dimensions increases but also outperforms the individual algorithms when the size of the dataset increases. The present clustering technique CLQ-CS can be extended to a more complicated case where each nest has multiple eggs representing a set of solutions. An additional direction to explore is that, the pre-processing phase can be more rigorously done so as to increase the cluster accuracy further, at the same time taking care not to decrease the cluster accuracy and increase the execution time.

REFERENCES

[1] Data Mining: Concepts and Techniques, Second Edition, Jiawei Han and Micheline Kamber.

[2] [Online] Available: https://en.wikipedia.org/wiki/Clustering_high-dimensional_data

[3] Payne R. B., Sorenson M. D., and Klitz K., The Cuckoos, Oxford University Press, (2005).

[4] Brown C., Liebovitch L. S., Glendon R., L´evy flights in Dobe Ju/’hoansi foraging patterns, Human Ecol., 35, 129-138 (2007).

[5] Reynolds A. M. and Frye M. A., Free-flight odor tracking in Drosophila is consistent with an optimal intermittent scale-free search, PLoS One, 2, e354 (2007)

[6] Pavlyukevich I., L´evy flights, non-local search and simulated annealing, J. Computational Physics, 226, 1830-1844 (2007).

[7] Shlesinger M. F., Zaslavsky G. M. and Frisch U. (Eds), L´evy Flights and Related Topics in Phyics, Springer, (1995).

[8] Lance Parsons, Ehtesham Haque, Huan Liu,“Subspace Clustering for High Dimensional Data: A Review”, Sigkdd Explorations, Vol. 6, Issue 1.

[9] R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In Proceedings of the 1998 ACM SIGMOD international conference on Management of data, pp. 94-105. ACM Press, 1998.

[10] C.-H. Cheng, A. W. Fu, Y. Zhang,"Entropy based subspace clustering for mining numerical data. In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 84-93. ACM Press, 1999.

[11] S. Goil, H. Nagesh, A. Choudhary,"Mafia: Effcient and scalable subspace clustering for very large data sets", Technical Report CPDC-TR-9906-010, Northwestern University, 2145 Sheridan Road, Evanston IL 60208, June 1999.

[12] J.-W. Chang, D.-S. Jin,"A new cell-based clustering method for large, high-dimensional data in data mining applications", In Proceedings of the 2002 ACM symposium on Applied computing, pp. 503-507. ACM Press, 2002.

[13] B. Liu, Y. Xia, P. S. Yu.,"Clustering through decision tree construction", In Proceedings of the ninth international conference on Information and knowledge management, pages 20{29. CM Press, 2000.

[14] C. M. Procopiuc, M. Jones, P. K. Agarwal, T. M.Murali, "A monte carlo algorithm for fast projective clustering", In Proceedings of the 2002 ACM SIGMOD international conference on Management of data, pp. 418-427. ACM Press, 2002.

[15] C. C. Aggarwal, J. L. Wolf, P. S. Yu, C. Procopiuc, J. S. Park. Fast algorithms for projected clustering", In Proceedings of the 1999 ACM SIGMOD international conference on Management of data, pp. 61-72. ACM Press, 1999.

[16] R. Ng, J. Han,"Effcient and effective clustering methods for spatial data mining", In Proceedings of the 20th VLDB Conference, pp. 144-155, 1994.

[17] C. C. Aggarwal, P. S. Yu.,"Finding generalized projected clusters in high dimensional spaces", In Proceedings of the 2000 ACM

DOI 10.17148/IJARCCE.2016.51158

SIGMOD international conference on Management of data, pp. 70-81. ACM Press, 2000.

[18] K.-G. Woo, J.-H. Lee.,"FINDIT: A Fast and Intelligent Subspace Clustering Algorithm using Dimension Voting", Ph.D thesis, Korea Advanced Institute of Science and Technology, Taejon, Korea, 2002.

[19] J. Yang, W. Wang, H. Wang, P. Yu δ-clusters: Capturing subspace correlation in a large data set. In Data Engineering, 2002. Proceedings. 18th International Conference on, pp. 517-528, 2002.

[20] J. H. Friedman, J. J. Meulman,"Clustering objects on subsets of attributes", [Online] Available: http://citeseer.nj.nec.com/friedman02clustering.html, 2002.

[21] Jyoti Yadav, Dharmender Kumar,"Subspace Clustering using CLIQUE:An Exploratory Study", IJARCET International Journal of Advanced Research in Computer Engineering & Technology Vol. 3, Issue 2, February 2014.

[22] Zoodataset: [Online] Available: https://archive.ics.uci.edu/ml/datasets/Zoo

[23] Behera Gayathri, A.Mary Sowjanya "Dimensionality Reduction using Clique and Genetic Algorithms" IJCST International Journal of Computer Science and Technology Vol. 6, Issue 3, September 2015.

BIOGRAPHY

Behera Gayathri is working as an Assistant Professor in the department of Computer Science and Engineering at GITAM Institute of Technology, GITAM University, Andhra Pradesh, India. She has received her M. Tech degree in Computer Science and Technology from Andhra University College of Engineering (A), Andhra Pradesh, India. Her research interests include Data Mining, Subspace Clustering and Big data analysis.

DOI 10.17148/IJARCCE.2016.51158