IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

A Hybrid Clustering Technique Combining ACO Algorithm with K-Means

Aarti Pardeshi[1], Neetesh Gupta[2]

M. Tech Student, Computer Science and Engg, Technocrats Institute of Technology and Science, Bhopal, India[1]

Assistant Professor, Computer Science and Engg, Technocrats Institute of Technology and Science, Bhopal, India[2]

Abstract: Data mining is an extraction of knowledge from large data set. It is an evolving technology which is a direct result of the increasing use of computer databases in order to store and retrieve information effectively. Optimization is essential for a huge amount of data processing. So that optimization is a challenging issue in data mining. Ant colony optimization (ACO) is an evolutionary computation technique. Proposed methodology is the iterative process of ACO algorithm, inertia weight adjustment is usually expected to make particles have stronger global searching capability in early stage to prevent premature convergence and have stronger local search capability in latter stage to accelerate convergent speed. In other words, the inertia weight should vary nonlinearly along with the process of decreasing slowly, then rapid and then slowly again so as to attain fast convergence speed in prophase and have local search capability to a certain degree at the later stage, too. In this dissertation ACO based k-means clustering is applied to generate clusters. And provide multimodal and higher dimensional complicated optimization problems, and can accelerate convergence speed, improve optimization quality effectively in comparison to the algorithms of ACO K-means.

Keywords: data Mining, Clustering, Ant colony optimization, K-means.

I. INTRODUCTION

Data mining is the task of extracting interesting patterns from large amounts of data. The data can be stored in databases, data warehouses, or other information repositories. It is also known as knowledge discovery in databases (KDD).The architecture of a typical data mining system may have the major components as database, data warehouse, or other information repository; their server which is responsible for fetching the relevant data based on the user's data mining request; knowledge base which is used to guide the search, or evaluate the interestingness of resulting patterns; data mining engine which consists of a set of functional modules for tasks; pattern evaluation module which interacts with the data mining modules so as to focus the search towards interesting patterns; and graphical user interface which communicates between users and the data mining system, allowing the user interaction with system.

Optimization become a huge problem among researchers in years of 60's, it take the researchers attention in the developing the powerful algorithm which can remove this problem. In the years of 60's-70's of 19th century, the three basic approaches were proposed as an evolutionary algorithm in order to get rid of the problem of optimization. These approaches are, evolutionary programming (EP), evolutionary strategies and genetic algorithm.

Ant Colony Optimization (ACO) is a population-based approach inspired by the observation of behavior of real ants colony. In ACO, solutions of the problem are constructed within a stochastic iterative process, by adding solution components to partial solutions [5]. Each individual ant constructs a part of the solution using an artificial pheromone, which reflects its experience accumulated while solving the problem, and heuristic information dependent on the problem [6] [7].Ant colony algorithm(ACO) [12] was first put forward in the 1990s by the Italian scholar D. Dorigo et al. It is a novel simulated evolutionary algorithm that is proposed for solving combinatorial optimization problems according to the ant colony behavior, which is also called Ant System. It turns out to be an effective method to solve the QAP allocation problem, TSP traveling salesman problem and JSP scheduling problem. Therefore, the ant colony algorithm(ACO) has certain advantages in solving complex optimization problems, especially the computer network routing optimization problem.

II. RELATED WORK

One of the challenges for clustering [11] resides in dealing with data distributed in separated repositories, because most clustering techniques require the data to be centralized. One of them, k-means, has been elected as one of the most influential data mining algorithms for being simple, scalable and easily modifiable to a variety of contexts and application domains. Although distributed versions of k-means have been proposed, the algorithm is still sensitive to the selection of the initial cluster prototypes and requires the number of clusters to be specified in advance. Traditional

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

approach [12] to clustering is to fit a model (partition or prototypes) for the given data. We propose a completely opposite approach by fitting the data into a given clustering model that is optimal for similar pathological data of equal size and dimensions. Cluster ensemble [13] approaches make use of a set of clustering solutions which are derived from different data sources to gain a more comprehensive and significant clustering result over conventional single clustering approaches. Unfortunately, not all the clustering solutions in the ensemble contribute to the final result. Cluster structure ensemble [14] focuses on integrating multiple cluster structures extracted from different datasets into a unified cluster structure, instead of aligning the individual labels from the clustering solutions derived from multiple homogenous datasets in the cluster ensemble framework. The genetic algorithms (GAs) [15] generally determine the number of clusters automatically. However, they typically choose the genes and the number of genes randomly. If we can identify the right genes in the initial population then GAs have better possibility to produce a high quality clustering result than the case when we randomly choose the genes.

III. PROPOSED METHODOLOGY

Proposed methodology will be use Ant Colony optimization (ACO) technique behalf of GA [6] for optimized clustering. Ant colony optimization is a technique for optimization that was introduced Marco Dorigo in 1991.The inspiring source of ant colony optimization is the foraging behavior of real ant colonies. This behavior is used in artificial ant colonies for the search of approximate solutions to discrete optimization problems, to continuous optimization problem.

ACO is Multi-agent approach for solving complex combinatorial optimization problems. However, unlike GA, ACO has no evolution operators such as crossover and mutation. In ACO, the potential solutions, called particles, fly through the problem space by following the current optimum particles.

In this section, the object function is created to combine intra cluster compactness and inter cluster separation. Based on this function, we derive objective function which only contains data samples and cluster assignments, while the data samples are given, a new model is established for hard cluster assignments of k-means clustering. ACO algorithm is introduced in this part; the operations of ACO are presented finally as show in figure1.

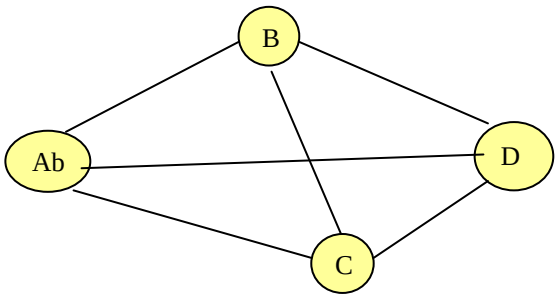

Figure 1: Ant System

IV. ANT COLONY ALGORITHMS

Ant Colony Algorithms are typically used to solve minimum cost problems. Problem may usually have N nodes and A undirected arcs .There are two working modes for the ants: either forwards or backwards. Pheromones are only deposited in backward mode. The ants memory allows them to retrace the path it has followed while searching for the destination node as show in figure 2 .Before moving backward on their memorized path, they eliminate any loops from it. While moving backwards, the ants leave pheromones on the arcs they traversed.

The ants evaluate the cost of the paths they have traversed. The shorter paths will receive a greater deposit of pheromones. An evaporation rule will be tied with the pheromones, which will reduce the chance for poor quality solutions. At the beginning of the search process, a constant amount of pheromone is assigned to all arcs. When located at a node i an ant k uses the pheromone trail to compute the probability of choosing j as the next node:

$$P_{ij}^{k} = \begin{cases} \dfrac{\tau_{ij}^{\alpha}}{\sum_{l \in N_i^k} \tau_{il}^{\alpha}} & \text{if } j \in N_i^k \\ 0 & \text{if } \notin N_i^k \end{cases}$$

Where N_i^k is the neighborhood of ant k when in node i.

When the arc (i,j) is traversed , the pheromone value changes as follows:

$$\tau_{ij} \leftarrow \tau_{ij} + \Delta\tau^{k}$$

DOI10.17148/IJARCCE.2017.6713

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

V. PROPOSED ARCHITECTURE

Proposed methodology is the iterative process of ACO algorithm, inertia weight adjustment is usually expected to make particles have stronger global searching capability in early stage to prevent premature convergence and have stronger local search capability in latter stage to accelerate convergent speed.

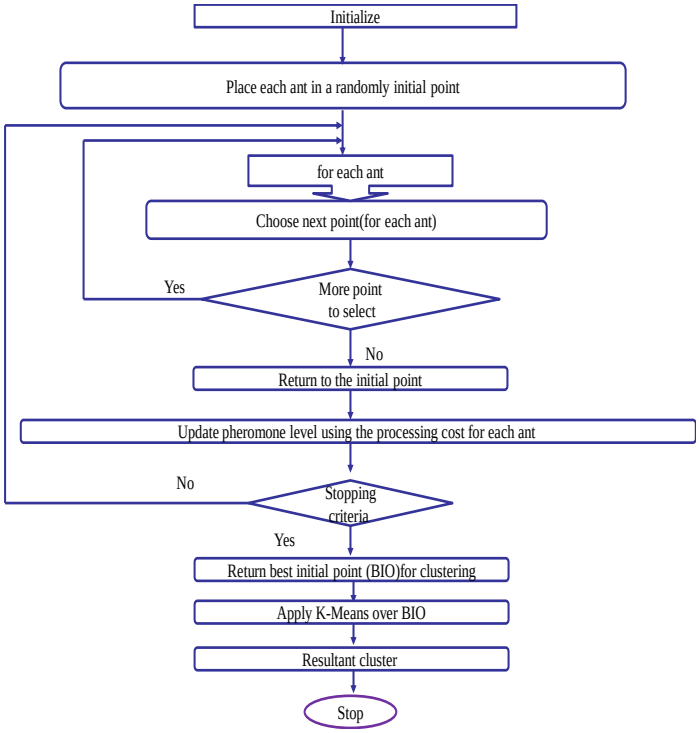

Figure 2: Proposed Architecture

In other words, the inertia weight should vary nonlinearly along with the process of decreasing slowly, then rapid and then slowly again so as to attain fast convergence speed in prophase and have local search capability to a certain degree at the later stage, too. The process is consistent with the decrease piece ($[0, \pi]$) of cosine function. Hence, this paper chooses the cosine function to simulate the inertia weight nonlinear changes.

VI. RESULT ANALYSIS

Comparing of Antclust and Genclust [1] algorithm's clustering performance are shown in Table 1, and the results of GA-means s clustering performance are shown in second column Table 1. Comparing to graph in figure 3,4, we can find our algorithm outperforms than GA-means. And we can see our algorithm overcomes some problems existing in GA based k-means.

Table 1: Comparing To ANT-K-Means Algorithm And GA-K-Means

			F-Measure		Computation Time	
S. No	**CP/ Alpha**	**MP/ Pick Up gamma**	**GenClust**	**AntClust**	**GenClust**	**AntClust**
1	0.88	0.79	83.7024	90.9452	4.75803	3.54122
2	0.834	0.814	83.2857	90.9365	3.63482	3.68162
3	0.931	0.916	83.7745	92.2051	4.90923	4.73924
4	0.496	0.485	83.8292	90.6719	5.61604	5.13562
5	0.708	0.81	84.0481	91.6781	5.47564	5.21043
6	0.849	0.842	83.3032	91.7467	3.18242	3.04724
7	0.533	0.523	83.0443	91.8914	2.90162	2.11524
8	0.945	0.944	83.3839	91.032	3.46322	3.28162
9	0.836	0.823	83.6979	90.9408	3.79082	3.61922

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

There are number of parameaters has been used to compare the proposed approch with exixting technique. with the help of graph this work is showing the diffrences.

Precision represent random probability that shown the random probability of selected item-set is relevant as shown in equation

$$\text{Precision} = \frac{T_P}{T_P + F_P} \dots \dots \dots \dots 2$$

Recall represent random probability that shown the random probability of selected item-set is search.

$$Recall = \frac{T_P}{T_P + F_P} = \text{Sensitivity} \dots\dots 3$$

F-measure combines score of precision and recall. It's the harmonic mean of precision and recall, the traditional F-measure.

$$F - \text{measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \dots\dots\dots 4$$

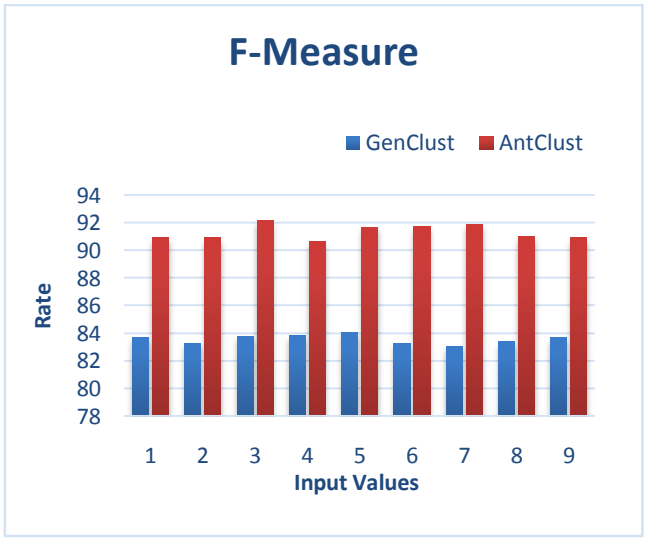

Figure 3: F_Measure

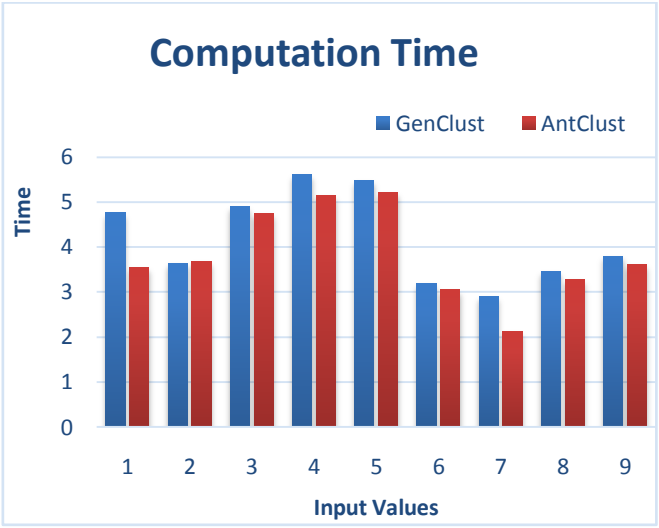

Figure 4: Average Computation Time

For result analysis in this paper four different parameter has been taken ie Average computation time and F-Measure.

Computation Time requirement for any clustering algorithm is need to be minimized. As shown in figure 4 proposed AntClust algorithm required less amount of time as compare to existing Genclust algorithm.

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

VII. CONCLUSION

ACO k-means clustering, a new model is established by transforming the clustering problem to 0-1 integer programming problem and we introduce ACO algorithm skillfully, both intra cluster compactness and inter-cluster separation are considered in the objective function. double parameters nonlinear adjustment, which considering the mutual influences of the inertia weight and learning factors on the particle's velocity updates., we can acquire the final clustering results, this algorithm overcomes the local convergent of the traditional algorithms and good results have been obtained. The performances of our clustering results indicate that our method has the potential performance improvement. But if the data samples are large, we should process huge amount of data and take a lot of time. In future work, it is necessary to simplify the chromosomes lengths and come up to a faster algorithm to tackle data sets. Optimizing the fitness function and changing function by studying the features of data sets can improve the accuracy of the results.

REFERENCES

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth "From Data Mining to KDD in Databases" pp. 0738-4602 1996.
[2] Xun Zhu1, Hongtao Deng2, Zheng Chen3 "A Brief Review On Frequent Pattern Mining" PP-4-11 2011 IEEE.
[3] Thair Nu Phyu "Survey of Classification Techniques in Data Mining" Vol I Imecs 2009, March 18 - 20, 2009, Hong Kong.
[4] Rupalihaldulakar, prof. Jitendraagrawal" Optimization of Association Rule Mining through Genetic Algorithm" (IJCSE) Vol. 3 No. 3 Mar 2011.
[5] Xiaojun Chen, Xiaofei Xu, Joshua Zhexue Huang, and Yunming Ye "TW-k-Means: Automated Two-Level Variable Weighting Clustering Algorithm for Multi-view Data" in IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 4, April 2013
[6] EnmeiTu , Longbing Cao , Jie Yang , Nicola Kasabov "A novel graph-based k-means for nonlinear manifold clustering and representative selection" in Elsevier transaction of Neuro computing 143 (2014) 109–122
[7] T. Velmurugan "Performance based analysis between k-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data" Elsevier transaction, pp 134–146, 2014
[8] GrigoriosTzortzis "The Min Max k-Means clustering algorithm" in Elsevier transaction of Pattern Recognition 47 (2014) 2505–2516
[9] Cheng-Huang Hung , Hua-Min Chiou b, Wei-Ning Yang "Candidate groups search for K-harmonic means data clustering" in Elsevier transaction of Applied Mathematical Modelling 37 (2013) 10123–10128
[10] Huang, J.Z.; Ng, M.K.; HongqiangRong; Zichen Li, "Automated variable weighting in k-means type clustering," in Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.27, no.5, pp.657-668, May 2005
[11] M.C. Naldi, R.J.G.B. Campello, Evolutionary k-means for distributed data sets, Neurocomputing, Volume 127, 15 March 2014, Pages 30-42, ISSN 0925-2312,
[12] Mikko I. Malinen, RaduMariescu-Istodor, PasiFränti, K-means□ : Clustering by gradual data transformation, Pattern Recognition, Volume 47, Issue 10, October 2014, Pages 3376-3386, ISSN 0031-3203
[13] Zhiwen Yu, Le Li, Yunjun Gao, Jane You, Jiming Liu, Hau-San Wong, Guoqiang Han, Hybrid clustering solution selection strategy, Pattern Recognition, Volume 47, Issue 10, October 2014, Pages 3362-3375.
[14] Zhiwen Yu, Le Li, Hau-San Wong, Jane You, Guoqiang Han, Yunjun Gao, Guoxian Yu, "Probabilistic cluster structure ensemble, Information Sciences", Volume 267, 20 May 2014, Pages 16-34.
[15] MdAnisur Rahman, MdZahidul Islam, A hybrid clustering technique combining a novel genetic algorithm with K-Means, Knowledge-Based Systems, Volume 71, November 2014, Pages 345-365, ISSN 0950-7051