International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 09 | Dec-2015 www.irjet.net p-ISSN: 2395-0072

An Efficient Analysis for High Dimensional Dataset Using K-Means Hybridization with Ant Colony Optimization Algorithm

Prabha S .[1], Arun Prabha K.[2]

[1] *Research Scholar, Department of Computer Science, Vellalar College for Women Tamilnadu, India*
[2] *Head and Assistant Professor, Department of Computer Technology (IT & CT) Tamilnadu, India*

Abstract - ***Data mining is the process of discovering meaningful, new correlation patterns and trends by the large amount of data are stored. Clustering is the useful technique for the discovery of data distribution and patterns in the underlying data. The purpose of clustering is grouping similar data. K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume K clusters) fixed a priori. Proposed the well-known Ant Colony Optimization algorithm can be applied to K-Means clustering problems. The Ant Colony algorithm is based on the behavior of ants in searching of food. The Ant converge is used to find a shortest path, a near-optimum solution for the target problem. A new method of K-Mean clustering in which is calculate initial centroid instead of random selection, due to the number of iterations is reduced. Ant Colony Optimization algorithm is to evaluate the efficiency with respect to accuracy in improving the fitness values among the ants. Finally concluded the proposed scenario yields superior performance than the existing scenario through Extended Particle Swarm Optimization Algorithm.***

Key Words: *K-Means, Clustering, Partical Swarm Optimization, Ant Colony Optimization.*

1.DATA MINING

The term "data mining" refers to the finding of relevant and useful information from databases. Data mining and Knowledge discovery in the databases is a new interdisciplinary field, merging ideas from statistics, machine learning, databases and parallel computing. Data mining should have been more appropriately named" knowledge mining from data". Knowledge mining a shorter term may not reflect the emphasis on mining from large amounts of data. Data mining tools predict future trends and behaviour, allowing businesses to make proactive, knowledge-driven decisions. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems. Data mining techniques can be broadly classified as Predictive and Description.

Data mining is the process of discovering meaningful patterns and relationships that lie hidden within very large databases. Data mining is a part of a process called knowledge discovery in databases (KDD). This process consists basically of steps that are performed before carrying out data mining, such as data selection, data cleaning, pre-processing, and data transformation.[6]

There are many other terms carrying a similar or slightly different meaning to data mining such as knowledge mining from databases, knowledge extraction, Data/pattern analysis, Data archaeology and Data dredging. A standard definition for data mining is the non-trivial extraction of implicit, previously unknown, and potentially useful knowledge from data.

1.1 Ant Colony Optimization (ACO)

An Ant Colony Optimization algorithm (ACO) is essentially a system based on agents which simulate the natural behaviour of ants, including mechanisms of cooperation and adaptation. In the use of this kind of system as a new metaheuristics was proposed in order to solve combinatorial optimization problem. This new metaheuristics has been shown to be both robust and versatile – in the sense that it has been successfully applied to a range of different combinatorial optimization problems.

ACO algorithms are based on the following ideas: Each path followed by an ant is associated with a candidate solution for a given problem. When an ant follows a path, the amount of pheromone deposited on that path is proportional to the quality of the corresponding candidate solution for the target problem. When an ant has to choose between two or more paths, the path(s) with a larger amount of pheromone have a greater probability of being chosen by the ant.

As a result, the ants eventually converge to a short path, hopefully the optimum or a near-optimum solution for the

target problem, as explained before for the case of natural ants. In essence, the design of an ACO algorithm involves the specification of an appropriate representation of the problem, which allows the ants to incrementally construct/modify solutions through the use of a probabilistic transition rule, based on the amount of pheromone in the trail and on a local, problem-dependent heuristic. A method to enforce the construction of valid solutions, that is, solutions that is legal in the real-world situation corresponding to the problem definition.

2. ANALYSIS OF RELATED WORK

Kahkashan Kouser, Sunita, [2013], proposed the K-Means clustering algorithm is applied on flower data set is studied, here various distance function such as Euclidean distance, Manhattan distance and Chebyshev distance function is used for analyzing the result of number of iterations. The overall accuracy of Chebyshev is greater as compared to Euclidean distance function and Manhattan distance. Different approaches were used to measure the distance among various data objects which is the most significant step of creating cluster. So special consideration should be given to choose distance function and it should be chosen according to dataset and number of cluster (13)

Gnanapriya.S and shivaranjani. P [2013] has proposed an Clustering is a machine learning technique that places data elements into related groups. Clustering can be defined as a process of organizing objects into groups whose members are similar in some way. The primary goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. K-Means (KM) is one of the widely used algorithms in clustering techniques. KM is the simplest unsupervised learning algorithms that can solve the well-known clustering problem. Ant Colony Optimization (ACO) is one of the most popular evolutionary algorithms inspired from nature and utilized in the field of clustering. Thus ACO can be defined as the adaptive heuristic search algorithm premised on the evolutionary ideas of natural behavior of ants. ACO is used to initialize the KM clustering algorithm. This will help in better clustering result with lesser error rates.

Kapil Agrawal, Renu Bagoria,[2014] has proposed an artificial ant colony capable of finding the shortest path. Ants of the artificial colony are able to generate successively shorter feasible path by using information accumulated in the form of a pheromone trail deposited on the edges of the graph. We describe the problem of Dijkstra's algorithm that solves the single-source shortest-path problem when all edges have non-negative weight. To solve that problem using Ant colony optimization. Ant colony optimization is already used in too many areas from graph related problems to the medical problem and study of Genomics.

Overview of Existing system:

Extended Particle Swarm Optimization (ECPSO) is a computational method that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. PSO optimizes a problem by having a population of candidate solutions, here dubbed particles, and moving these particles around in the search-space according to simple mathematical formulae over the particle's position and velocity. Each particle's movement is influenced by its local best known position and is also guided toward the best known positions in the search-space, which are updated as better positions are found by other particles. This is expected to move the swarm towards the best solutions. Each particle keeps track of its coordinates in the solution space which are associated with the best solution (fitness) that has achieved so far by the particle. This value is called personal best, (pbest). Another best value that is tracked by the PSO is the best value obtained so far by any particle in the neighbourhood of that particle. This value is called gbest. Hence to overcome all these issues The Ant Colony Optimization algorithm (ACO) is utilized in the current work.

Challenges In The Existing Scenario

- The method easily suffers from the partial optimism, which causes the less exact at the regulation of its speed and the direction.
- The method cannot work out the problems of scattering and optimization.
- The method cannot work out the problems of non-coordinate system, such as the solution to the energy field and the moving rules of the particles in the energy field.

3. PROPOSED SCENARIO

The ACO meta-heuristic is to assign each pixel to a cluster during the classification process. This assignment is done through a probability, which is inversely dependent to the distance (similarity) between the pixel and cluster centers and a variable, τ, representing the pheromone level. Pheromone is defined to be dependent on the minimum distance between each pair of cluster centers and inversely dependent on the distances between each pixel and its cluster center. So the pheromone gets larger when cluster centers are far apart and clusters tend to be more compact (our criterion for best solution), making the probability of assigning a pixel to that cluster high. Pheromone evaporation is considered to weaken the influence of the previously chosen solutions. The

algorithm starts by assigning a pheromone level τ and a heuristic information η to each pixel. Then each ant will assign each pixel to a cluster with the probability P is obtained.

Clustering is a distribution of data into groups of similar objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. In data mining, K-Means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. Though the K-Means is one of the best clustering algorithms, the quality is based on the starting condition and it may converge to local minima. The Ant Colony Optimization algorithm (ACO) is one of the most widely used probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs. It develops an iterative solution for any problem at hand. The intermediate solutions can be used to arrive at the final solution.
The proposed algorithm is tested using data from different datasets and the results shows that K-Means based on ACO can lead to improved solutions in terms of entropy and accuracy of clusters

Features of Proposed Scenario

- Inherent parallelism.
- Positive Feedback accounts for rapid discovery of good solutions.
- Can be used in dynamic applications (adapts to changes such as new distances, etc).

3.1 MEASURES

Rule Pruning

Rule pruning is a common place technique in data mining. The main goal of rule pruning is to remove irrelevant terms that might have been unduly included in the rule. Rule pruning potentially 12 increases the predictive power of the rule, helping to avoid its over fitting to the training data. Another motivation for rule pruning is that it improves the simplicity of the rule, since a shorter rule is usually easier to be understood by the user than a longer one. The rule pruning procedure is called, when the ant completes the construction of its rule. The strategy for the rule pruning procedure is similar to that suggested by, but the rule quality criteria used in the two procedures are very different.

The basic idea is to iteratively remove one-term-at-a-time from the rule while this process improves the quality of the rule. More precisely, in the first iteration one starts with the full rule. Then it is tentatively tried to remove each of the terms of the rule each one in turn and the quality of the resulting rule is computed using a given rule-quality function. It should be noted that this step might involve replacing the class in the rule consequent, since the majority class in the cases covered by the pruned rule can be different from the majority class in the cases covered by the original rule. The term whose removal most improves the quality of the rule is effectively removed from it, completing the first iteration. In the next iteration it is removed again the term whose removal most improves the quality of the rule, and so on. This process is repeated until the rule has just one term or until there is no term whose removal will improve the quality of the rule.

Pheromone Updating

Recall that each $term_{ij}$ corresponds to a segment in some path that can be followed by an ant. At each iteration of the WHILE loop of Algorithm I all $term_{ij}$ are initialized with the same amount of pheromone, so that when the first ant starts its search, all paths have the same amount of pheromone. The initial amount of pheromone deposited at each path position is inversely proportional to the number of values of all attributes, and is defined by Equation

$$r_{ij}(t=0) = \frac{1}{\sum_{i=1}^{a} b_i} \qquad (1)$$

Where a is the total number of attributes, and b_i is the number of possible values that can be taken on by attribute A_i.

The value returned by this equation is normalized to facilitate is use, which combines this value and the value of the heuristic function. Whenever an ant constructs its rule and that rule is pruned the amount of pheromone in all segments of all paths must be updated. This pheromone updating is supported by basic ideas, namely:

- The amount of pheromone associated with each $term_{ij}$ occurring in the rule found by the ant (after pruning) is increased in proportion to the quality of that rule.

ACO PROCESS

1. Let (t) be the total pheromone deposited on path ij at time t, and (t) be the heuristic value of path ij at time t according to the measure of the objective function.

$$P_{ij}(t) = \frac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}(t)]^{\beta}}{\sum[\tau_{ij}(t)]^{\alpha}[\eta_{ij}(t)]^{\beta}} \qquad (2)$$

Where α and β = parameters that control the relative importance of the pheromone trail versus a heuristic value.

2.Let q be a random variable uniformly distributed over [0, 1], and ∈ [0, 1] be a tunable parameter. The next

node j that Ant k chooses to go:

$$j = \begin{cases} \underset{J}{arg\,max} \left\{ [\tau_{il}(t)]^{\propto} [\eta_{ij}(t)]^{\beta} \right\} & \text{if } q \leq q_0 \end{cases} \quad (3)$$

Where J = a random variable selected according to the probability distribution of P_{ij}(t). The pheromone trail is changed both locally and globally.

3. Local updating is intended to avoid a very strong path being chosen by all the Ants. Every time a path is chosen by an Ant, the amount of pheromone will change by applying the local trail updating formula:

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

where τ_o= initial value of pheromone; δ = tuning parameter (0 ≤ δ ≤ 1); and the symbol $\xleftarrow{step}$is used to show the next step. Upon completion of a tour by all ants in the colony, the global trail updating is done as follows:

$$\tau_{ij}(t) \xleftarrow{iteration} \rho.\tau_{ij}(t) + (1 - \rho).\Delta \tau_{ij} \quad (5)$$

where 0 ≤ ρ ≤ 1; (1 - ρ) = evaporation (i.e., loss) rate; and the symbol $\xleftarrow{iteration}$is used to show the next iteration.

3.There are several definitions for Δ (Dorigo et al.

1996; Dorigo and Gambardella 1997). They use three algorithms:

i. Ant Colony algorithm

$$\Delta \tau_{ij}(t) = \sum_{k=1}^{M} \tau m_{ij}^{k}(t) \quad (6)$$

$$\tau m_{ij}^{k}(t) = \int_{O}^{1/G^k} (m) \quad \text{if } (i,j) \in T^k(m) \quad (7)$$

$$\text{if } (i,j) \notin T^k(m)$$

Where G^k (m) = value of the objective function for the tour T^k(m) taken by the K-th ant at iteration m.

ii. Ant Colony Optimization–Iteration Best

$$\Delta \tau_{ij}(t) = \int_{0}^{1\backslash G^{K_{ib}}} \text{if } (i,j) \in \text{tour done by ant } K_{ib} \quad (8)$$

Where $G^{k_{ib}}$ (m) = value of the objective function for the ant taken the best tour at iteration m.

iii. Ant Colony Optimization –Global Best

$$\Delta \tau_{ij}(t) = \int_{0}^{1\backslash G^{K_{gb}}} \text{if } (i,j) \in \text{tour done by ant } K_{gb} \quad (9)$$

Where$G^{k_{gb}}$ = value of the objective function for the ant with the best performance within the past total iteration.

3.2 ALGORITHM STEPS

ACO Based K-Means

Input Dataset of maximum Size

Step 1: Input number of clusters K where number of ants is equal to Number of Clusters.

Step 2: Initialize Cluster Centroids 'K' using ACO Optimization.

Initialize Number of Ants 'M'

For i=1 to m (iterations)

a) For each ant, let each pixel x belong to one cluster with the probability

$$P_{ij}(t) = \{ \frac{[\tau_{ij}(t)]^{\alpha} [\eta_{ij}(t)]^{\beta}}{\sum [\tau_{ij}(t)]^{\alpha} [\eta_{ij}(t)]^{\beta}} \}$$

b) Save the best solution among the M solutions found.

c) Update the pheromone level on all pixels according to the best solution using

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

End For loop

Step 3: Calculate Global best Ant from local best solution, such that number of ants (no. of cluster K) is equal to global best solutions.

Step 4: Assign data points to the closer cluster using Euclidian Distance metrics.

$$\arg\, min_c \sum_{i=1}^{k} \sum_{x \in c_i} d(x, \mu_i) = \arg\, min_c \sum_{i=1}^{k} \sum_{x \in c_i} \left| X - \mu_I \right|_2^2$$

Where C_i is cluster centre of i^{th} cluster.

Step 5: Calculate the new cluster centre of each cluster by calculating mean of the data points belonging to that clusters.

Step 6: Repeat step 4 & 5 until the convergence mets (no change in cluster centroids).

4. Results and discussion

Experimental analysis is indented to be of use to researchers from all fields who want to study algorithms experimentally. The dataset is obtained from the UCI Machine Learning Repository to test the performance of proposed algorithm against other algorithms. At the same time, its properties are also empirically studied. The experimental result are summarized and discussed in the following section.

Purity Measures

Purity is an external evaluation criterion of cluster quality. It is the percent of the total number of objects (data points) that were classified correctly, in the unit range. The overall purity of the clustering solution is obtained as a weighted sum of the individual cluster purities and is given as

$$\text{Purity} = \sum_{j=1}^{m} \frac{n_j}{n} P_j \qquad (1)$$

Where n_j is the size of cluster j, m is the number of clusters, and n is the total number of data points.

Purity Measures:

Datasets	No. of Attributes	No. of Data Points	ECPSO Purity	ACO Purity
Glass	10	214	0.89	1
Iris	6	150	0.50	0.68
Seeds	7	210	0.18	0.34

Table.1 Existing and Proposed Purity Values

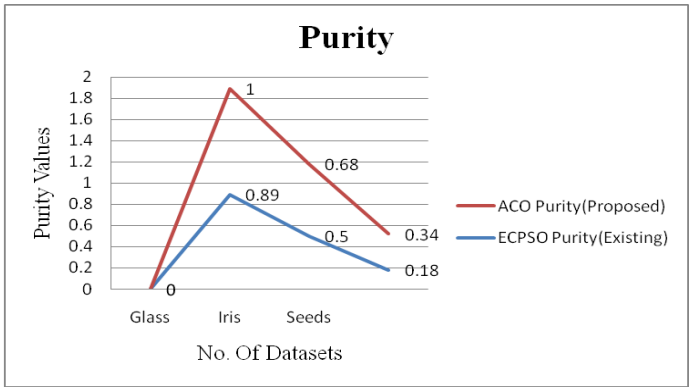

Figure.1 Existing and Proposed Purity Values

From the above figure (1), the comparison of existing and proposed system in terms of Purity is observed. In the x axis is to plot the Number of datasets and in y axis is to plot the Purity values. In existing scenario, the Purity values are lower by using ECPSO algorithm. The Purity values of existing scenario by using Glass dataset is 0.89. In proposed system, the Purity value is higher by using the ACO algorithm is 1.The Purity value of existing scenario by using Iris dataset is 0.5. In proposed system, the Purity value is higher by using the ACO algorithm is 0.68. The Purity value of existing scenario by using Seeds dataset is 0.18. In proposed system, the Purity value is higher by using the ACO algorithm is 0.34. From the result, concluded that proposed system is superior in performance.

5. Conclusion and Future Work

Ant Colony Optimization (ACO) is proposed to improve K-Means clustering. Though the K-Means is one of the best clustering algorithms and it may converge to local minima. K-Means clustering is a common approach, which is based on initial centroid selected randomly. Ant colony optimization exploits a similar mechanism for solving optimization problem. An Ant Colony Optimization algorithm (ACO) is essentially a system based on agents which simulate the natural behavior of ants, including mechanisms of cooperation and adaptation. The result showed that, concerning predictive accuracy and improves the purity of fitness value obtained by Ant Colony Optimization algorithm which is better than the Extended Partical Swarm Optimization algorithm. On the other hand, ACO has consistently found much simpler (smaller) rule lists than ECPSO. Therefore, ACO seems particularly advantageous when it is important to minimize the number of discovered rules and rule terms (conditions), in order to improve comprehensibility of the discovered knowledge.

FUTURE WORK

Research study is the complete in-depth analysis on a specific area. The research will have impact on the future and is an on-going activity that never ends. This research work can be enhanced with the following features:

- To investigate the performance of other kinds of heuristic function and pheromone updating strategy.
- And also the metaheuristics algorithm to minimize the iteration process for classifying high dimensional data with maximum accuracy.

REFERENCES

[1] **Mrs. Nidhi Singh, Dr.Divakar Singh**, "The Improved K-Means with Particle Swarm Optimization" [2013].
[2] **Kahkashan Kouser, Sunita.A**, "Comparative study of K Means Algorithm by Different Distance Measures" [2013].
[3] **Gnanapriya. S and shivaranjani. P**, "Initialization K-Means using ant colony optimization" ISSN 2319-5991 www.ijerst.com, vol. 2, no. 2, may [2013].
[4] **Kapil Agrawal, Renu Bagoria**, "Ant Colony Optimization: efficient way to find shortest path International Journal of Advanced Technology & Engineering Research (IJATER)", [2014].
[5] **Millie Pant, Radha Thangaraj, and Ajith Abraham**, "A New PSO Algorithm with Crossover Operator for Global Optimization Problems" [2007].
[6] **Fayyad. U., Uthursamy.R.,** "Data mining and Knowledge discovery in Databases", Communication of the ACM, Pages 24-27, [1996].