Advances in Information Mining, ISSN: 0975–3265, Volume 2, Issue 1, 2010, pp-13-17

Ant based rule mining with parallel fuzzy cluster

Sankar K.[1] and Krishnamoorthy K.[2]
[1]Department of Master of Computer Applications, KSR College of Engineering, Tiruchengode, san_kri_78@yahoo.com
[2]Department of Computer Science and Engineering, SONA College of Technology, Salem, kkr_510@yahoo.co.in

Abstract- Ant-based techniques, in the computer sciences, are designed those who take biological inspirations on the behavior of these social insects. Data clustering techniques are classification algorithms that have a wide range of applications, from Biology to Image processing and Data presentation. Since real life ants do perform clustering and sorting of objects among their many activities, we expect that an study of ant colonies can provide new insights for clustering techniques. The aim of clustering is to separate a set of data points into self-similar groups such that the points that belong to the same group are more similar than the points belonging to different groups. Each group is called a cluster. Data may be clustered using an iterative version of the Fuzzy C means (FCM) algorithm, but the draw back of FCM algorithm is that it is very sensitive to cluster center initialization because the search is based on the hill climbing heuristic. The ant based algorithm provides a relevant partition of data without any knowledge of the initial cluster centers. In the past researchers have used ant based algorithms which are based on stochastic principles coupled with the k-means algorithm. The proposed system in this work use the Fuzzy C means algorithm as the deterministic algorithm for ant optimization. The proposed model is used after reformulation and the partitions obtained from the ant based algorithm were better optimized than those from randomly initialized hard C Means. The proposed technique executes the ant fuzzy in parallel for multiple clusters. This would enhance the speed and accuracy of cluster formation for the required system problem.

1. INTRODUCTION

Research in using the social insect metaphor for solving problems is still in its infancy. The systems developed using swarm intelligence principles emphasize distributiveness, direct or indirect interactions among relatively simple agents, flexibility and robustness [4]. Successful applications have been developed in the communication networks, robotics and combinatorial optimization fields.

1.1 ANT COLONY OPTIMIZATION

Many species of ants cluster dead bodies to form cemeteries, and sort the larvae into several piles [4]. This behavior can be simulated using a simple model in which the agents move randomly in space and pick up and deposit items on the basis of local information. The clustering and sorting behavior of ants can be used as a metaphor for designing new algorithms for data analysis and graph partitioning. The objects can be considered as items to be sorted. Objects placed next to each other have similar attributes. This sorting takes place in two-dimensional space, offering a low-dimensional representation of the objects. Most swarm clustering work has followed the above model. In the work, there is implicit communication among the ants making up a partition. The ants also have memory. However, they do not pick up and put down objects but rather place summary objects in locations and remember the locations that are evaluated as having good objective function values. The objects represent single dimensions of multidimensional cluster centroids which make up a data partition.

1.2 CLUSTERING

The aim of cluster analysis is to find groupings or structures within unlabeled data [5]. The partitions found should result in similar data being assigned to the same cluster and dissimilar data assigned to different clusters. In most cases the data is in the form of real-valued vectors. The Euclidean distance is one measure of similarity for these data sets. Clustering techniques can be broadly classified into a number of categories [6]. Hard C Means (HCM) is one of the simplest unsupervised clustering algorithms for a fixed number of clusters. The basic idea of the algorithm is to initially guess the centroids of the clusters and then refine them. Cluster initialization is very crucial because the algorithm is very sensitive to this initialization. A good choice for the initial cluster centers is to place them as far away from each other as possible. The nearest neighbor algorithm is then used to assign each example to a cluster. Using the clusters obtained, new cluster centroids are calculated. The above steps are repeated until there is no significant change in the centroids. Hard clustering algorithms assign each example to one and only one cluster. This model is inappropriate for real data sets in which the boundaries between the clusters may not be well defined. Fuzzy algorithms can partially assign data to multiple clusters. The strength of membership in the cluster depends on the closeness of the example to the cluster center. The Fuzzy C Means algorithm (FCM), allows an example to be a partial member of more than one cluster. The FCM algorithm is based on

minimizing the objective function. The drawback of clustering algorithms like FCM and HCM, which are based on the hill climbing heuristic, is, prior knowledge of the number of clusters in the data is required and they have significant sensitivity to cluster center initialization. The proposal of this work moves in the direction of constructing C fuzzy means clustering with ant colony optimization (parallel ant agents) in evolving efficient rule mining techniques. In this thesis, the proposal introduces the problem of combining multiple partitionings of a set of objects without accessing the original features. The system first identify several application scenarios for the resultant `knowledge reuse' framework that the system call cluster ensembles. The cluster ensemble problem is then formalized as a combinatorial optimization problem in terms of shared mutual information in building rule mining techniques. In addition to a direct maximization approach, the system proposes three effective and efficient techniques for obtaining high-quality combiners.

2. RELATED WORKS

Andrea Baraldi and Palma Blonda,[1] describe, equivalence between the concepts of fuzzy clustering and soft competitive learning in clustering algorithms was proposed on the basis of the existing literature. Moreover, a set of functional attributes is selected for use as dictionary entries in the comparison of clustering algorithms. Alfred Ultsch systems for clustering with collectives of autonomous agents follow either the ant approach of picking up and dropping objects or the DataBot approach of identifying the data points with artificial life creatures. In DataBot systems the clustering behaviour is controlled by movement programs. Julia Handl and Bernd Meyer Sorting and clustering methods inspired by the behavior of real ants are among the earliest methods in ant-based meta-heuristics. The system revisits these methods in the context of a concrete application and introduces some modifications that yield significant improvements in terms of both quality and efficiency. Firstly, re-examine their capability to simultaneously perform a combination of clustering and multi-dimensional scaling. In J.Handl, J.Knowles and M.Dorigo Ant-based clustering and sorting is a nature-inspired heuristic for general clustering tasks. It has been applied variously, from problems arising in commerce, to circuit design, to text-mining, all with some promise. However, although early results were broadly encouraging, there has been very limited analytical evaluation of the algorithm. Alexander Strehl, Joydeep Ghosh introduces the problem of combining multiple partitioning of a set of objects into a single consolidated clustering *without* accessing the features or algorithms that determined these partitioning. The system first identify several application scenarios for the resultant `knowledge reuse' framework that we call *cluster ensembles*. The cluster ensemble problem is then formalized as a combinatorial optimization problem in terms of shared mutual information. In addition to a direct maximization approach, the system proposes three effective and efficient techniques for obtaining high-quality combiners (consensus functions). The first combiner induces a similarity measure from the partitioning and then reclusters the objects. The second combiner is based on hypergraph partitioning. The third one collapses groups of clusters into meta-clusters which then compete for each object to determine the combined clustering. Due to the low computational costs of the techniques, it is quite feasible to use a supra-consensus function that evaluates all three approaches against the objective function and picks the best solution for a given situation. The system evaluates the effectiveness of cluster ensembles in three qualitatively different application scenarios: (i) where the original clusters were formed based on non-identical sets of features, (ii) where the original clustering algorithms worked on non-identical sets of objects, and (iii) where a common data-set is used and the main purpose of combining multiple clusterings is to improve the quality and robustness of the solution. Promising results are obtained in all three situations for synthetic as well as real data-sets. Nicolas Labroche, Nicolas Monmarch´e and Gilles Venturini introduces a method to solve the unsupervised clustering problem, based on a modeling of the chemical recognition system of ants. This system allow ants to discriminate between estimates and intruders, and thus to create homogeneous groups of individuals sharing a similar odor by continuously exchanging chemical cues. This phenomenon, known as ”colonial closure”, inspired us into developing a new clustering algorithm and then comparing it to a well-known method such as K-MEANS method. The previous literature work on fuzzy cluster depicted above insists on the following parameters. The first one handles the functional attribute with the theoretical analysis. The second and third one deal with the cluster object movement issues on synthetic data sets. The fourth and fifth one deals with heuristic ant optimization model with trial repetition. Sixth and seventh authors utilized unsupervised cluster with class tree structuring. The final one uses c-fuzzy mean cluster in the sequential way. This motivates us to precede our proposal on ACO with c-fuzzy means. Based on the C-Fuzzy sequential clustering of ACO Problem, we derived a parallel fuzzy ant clustering model to improve the attribute accuracy rate and faster execution on the proposed problem domain.

3. FUZZY ANT CLUSTERING

Ant-based clustering algorithms are usually inspired by the way ants cluster dead nest mates into piles, without negotiating about where to gather the corpses. These algorithms are characterized by the lack of centralized control or a priori information, which makes them very appropriate candidates for the task at hand. Since the Fuzzy ants algorithm does not need initial partitioning of the data or a predefined number of clusters, it is very well suited for the Web People Search task, where the system do not know in advance how many clusters (or individuals) correspond to a particular document set (or person name in the case). A detailed description of the algorithm is given by Schockaert et al. It involves a pass in which ants can only pick up one item as well as a pass during which ants can only pick up an entire heap. A fuzzy ant-based clustering algorithm was introduced where the ants are endowed with a level of intelligence in the form of IF / THEN rules that allow them to do approximate reasoning. As a result, at any time the ants can decide for themselves whether to pick up a single item or an entire heap, which makes a separation of the clustering in different passes superuous. The system has experiment with a different number of ant's runs and fixed the number of runs to 800000 for the experiments. In addition, the system has also evaluated different values for the parameters that determine the probability that a document or heap of documents is picked up or dropped by the ants and kept following values for the experiments:

Table 1: Parameter settings for fuzzy clustering

n1	probality of dropping one item	1
m1	probality of picking up one item	1
n2	probality of dropping an entire heap	5
m2	probality of picking up a heap	5

3.1 Hierarchical Clustering

The second clustering algorithm the system applies is an agglomerative hierarchical approach. This clustering algorithm builds a hierarchy of clustering's that can be represented as a tree (called a dendrogram) which has singleton clusters (individual documents) as leaves and a single cluster containing all documents as root. An agglomerative clustering algorithm builds this tree from the leaves to the top, in each step merging the two clusters with the largest similarity. Cutting the tree at a given height gives a clustering at a selected number of clusters. The system have opted to cut the tree at different similarity thresholds between the document pairs, with intervals of 0.1 (e.g. for threshold 0.2 all document pairs with similarities above 0.2 are clustered together). For the experiments, the system has used an implementation of Agnes (Agglomerative Nesting) that is fully described.

3.2 Fuzzy Ant Parallel System

Clustering approaches are typically quite sensitive to initialization. In this thesis, the system examine a swarm inspired approach to building clusters which allows for a more global search for the best partition than iterative optimization approaches. The approach is described with cooperating ants as its basis. The ants participate in placing cluster centroids in feature space. They produce a partition which can be utilized as is or further optimized. The further optimization can be done via a focused iterative optimization algorithm. Experiments were done with both deterministic algorithms which assign each example to one and only one cluster and fuzzy algorithms which partially assign examples to multiple clusters. The algorithms are from the C-means family. These algorithms were integrated with swarm intelligence concepts to result in clustering approaches that were less sensitive to initialization.

4. EXPERIMENTAL SIMULATION ON ANT BASED PARALLEL CLUSTER

The system implementation of fuzzy ant based parallel clustering algorithm for rule mining used three real data sets obtained from UCI repository. The data sets were Iris Human Data Set, Wine Recognition Data Set, and Glass Identification Data Set. The simulation conducted in matlab normalizes the feature values between 0 and 1. The normalization is linear. The minimum value of a dataset specific feature is mapped to 0 and the maximum value of the feature is mapped to 1. Initialize the ants with random initial values and with random direction. There are two directions, positive and negative. The positive direction means the ant is moving in the feature space from 0 to 1. The negative direction means the ant is moving in the feature space from 1 to 0. Clear the initial memory. The ants are initially assigned to a particular feature within a particular cluster of a particular partition. The ants never change the feature, cluster or the partition assigned to them.

Repeat

For one epoch /* One epoch is n iterations of random ant movement */

For all ants

With a probability Prest the ant rests for this epoch

If the ant is not resting then with a probability Pcontinue the ant continues in the same direction else it changes direction

With a value between Dmin and Dmax the ant moves in the selected direction

The new Rm value is calculated using the new cluster centers calculated by recording the

position of the ants known to move the features of clusters for a given partition.
If the partition is better than any of the old partitions in memory then the worst partition is removed from the memory and this new partition is copied to the memories of the ants making up the partition.
If the partition is not better than any of the old partitions in memory Then
With a probability P Continue Current the ant continues with the current partition
Else
With a probability 0.6 the ant moves to the best known partition, with a probability 0.2 the ant moves to the second best known partition, with a probability 0.1 the ant goes to the third best known partition, with a probability 0.075 the ant goes to the fourth best known partition and with a probability 0.025 the ant goes to the worst known partition Until Stopping criteria The stopping criterion is the number of epochs.

Table 2- Parameter Values

Parameter	Value
Number of ants	30 * c * # features
Memory per ant	5
Iterations per epoch	50
Epochs	1000
P_{rest}	0.01
$P_{continue}$	0.75
$P_{ContinueCurrent}$	0.20
D_{min}	0.001
D_{max}	0.01

Note the multiplier 30 for the number of ants allows for 30 partitions.
Three data sets Glass Data Set, Wine Data Set, Iris Data Set were evaluated from a mixture of five Gaussians. The probability distribution across all the data sets is the same but the means and standard deviations of the Gaussians are different. Of the three data sets, two data sets had 500 instances each and the remaining one data set had 1000 instances each. Each instance had two attributes. To visualize the Iris data set, the Principal Component Analysis (PCA) algorithm was used to project the data points into a 2D and 3D space.

5. RESULTS AND DISCUSSIONS

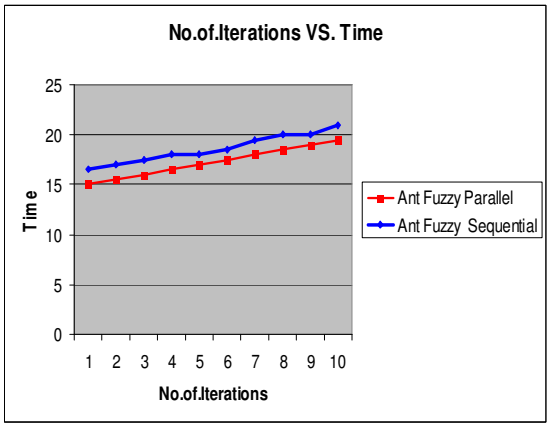
The ants move the cluster centers in feature space to find a good partition for the data. There are less controlling parameters than the previous ant based clustering algorithms. The previous ant clustering algorithms typically group the objects on a two-dimensional grid. Results from 18 data sets show the superiority of the algorithm over the randomly initialized FCM and HCM algorithms. For comparison purposes, Table 2 shows the frequency of occurrence of different extrema for the ant initialized FCM and HCM algorithms and the randomly initialized FCM and HCM algorithms.

Table 3- Frequency of different extrema from parallel fuzzy based ant clustering, for Glass (2 class) Iris and Wine data set

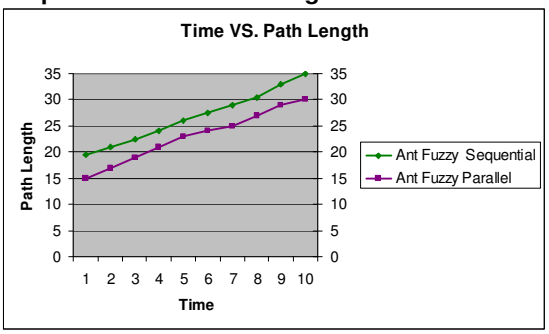
Data Set	Extrema	Frequency HCM, and Initialization	Frequency HCM, random Initialization	Sequential C-Fuzzy ACO (Existing)	Parallel C-Fuzzy ACO (Proposed)
Glass (2 class)	34.1320	19	3	31	27.8
	34.1343	11	19	32.12	28.5
	34.1372	19	15	32.36	29.1
	34.1658	1	5	32.89	29.82
Iris	6.9981	50	23	5.3938	4.23
	7.1386	0	14	5.8389	4.3658
	10.9083	0	5	8.3746	5.3256
	12.1437	0	8	10.6434	8.2356
Wine	9.3645	20	2	5.2369	3.2567
	11.3748	15	20	8.2356	5.236
	13.8483	12	18	10.2356	8.3656

The ant initialized parallel ant fuzzy algorithm always finds better extrema for the Iris data set and for the Wine data set the ant initialized algorithm finds the better extrema 49 out of 50 times. The ant initialized HCM algorithm always finds better extrema for the Iris data set and for the Glass (2 class) data set a majority of the time. For the different Iris, the ant initialized parallel algorithm finds a better extrema most of the time. The ACO approach was used to optimize the clustering criteria, the ant approach for parallel C Means, found better extrema 64% of the time for the Iris data set. The ant initialized parallel C fuzzy ACO finds better extrema all the time. The number of ants is an important parameter of the algorithm. This number only increases when more partitions are searched for at the same time; as ants are (currently) added in increments (Graph 1 and Graph 2). The quality of the final partition improves with an increase of ants, but the improvement comes at the expense of increased execution time.

Graph 1: Number of Iterations Vs Time



Graph 2: Time Vs Path Length



7. CONCLUSION

The system discussed a swarm inspired optimization algorithm to partition or creates clusters of data. The system described it using the ant paradigm. The approach is to have a coordinated group of ant's position cluster centroids in feature space. The algorithm was evaluated with a soft clustering formulation utilizing the fuzzy c-means objective function and a hard clustering formulation utilizing the hard c-means objective function. The presented clustering approach seems clearly advantageous for the data sets where it is expected there will be lots of local extrema. The cluster discovery aspect of the algorithm provides the advantage of obtaining a partition at the same time it indicates the number of clusters. That partition can be further optimized or accepted as is. This is in contrast to some other schemes which require partitions to be created with different numbers of clusters and then evaluated. The results are generally a superior optimized partition (objective function) than obtained with FCM/HCM. One needs a large number of random initializations to be competitive in terms of skipping some of the poor local extrema which was done with the ant-based algorithm. It has provided enhanced final partitions on average than a previously introduced evolutionary computation clustering approach for several data sets. Random initializations have been shown to be the best approach for the c-means family and the ant clustering algorithm results in generally better partitions than a single random initialization. The parallel version of the ants algorithm could operate much faster than the current sequential implementation, thereby making it a clear choice for minimizing the chance of finding a poor extrema when doing c-means clustering. This algorithm should scale better for large numbers of examples than grid-based ant clustering algorithms.

REFERENCES

[1] Baraldi A. and Blonda P. (1999a) *IEEE Transactions on Systems, Man, and Cybernetics,* 29(6), 778-785.

[2] Kanade P.M. and Hall L.O. (2003) *IEEE Transactions on Fuzzy Systems* , 11(2), 227-232.

[3] Handl J. and Meyer B. (2002) *Springer-Verlag*, 2439, 913-923.

[4] Handl J., Knowles J. and Dorigo M. (2003) *IOS Press, Amsterdam, the Netherlands*, 204-213.

[5] Strehl A. and Ghosh J. (2002) *Journal of Machine Learning Research* 3, 583-617.

[6] Labroche N., Monmarche N. and Venturini G. (2002) *France: IOS Presss*, 345-349.