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

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

An Efficient Accumulative Constraint Based Leader Ant Clustering

S. Sridevikarumari

Assistant Professor, Pioneer College of Arts and Science, Jothipuram, Coimbatore, India

Abstract: The thesis entitled "An Efficient Accumulative Constraint Based Leader Ant Clustering" is based on the Ant colony optimization clustering algorithm. Ant-based clustering can be divided into two groups. The first group of methods directly mimics the clustering behavior observed in real ant colonies. The second group is less directly inspired by nature. Clustering task can be considered as the most important unsupervised learning problems, which deals with finding a structure in a collection of unlabeled data. To this end, it conducts a process of organizing objects into groups. These algorithms have recently been shown to produce good results in a wide variety of real-world applications. In recent years, research on and with the ant-based clustering algorithms has reached a very promising state. Clustering with constraints is a developing area of machine learning, improve the efficiency of analysis and express the intractability results. It is an interactive process where a user can run the constraints number of times to refine previous clustering results. In this research, An efficient and fast constraint based Leader Ant Clustering provide three new variants algorithm (MCALA, MEALA and CEALA) are proposed that implements the following constraints: the must-link, cannot-link constraints, ε –constraints and accumulative constraints. The main aim of this research is to improve the accuracy of the clustering techniques. The real data sets from the Machine Learning repository namely Glass, Iris, Wine, Thyroid and Soybean are used in this experiment. These accumulative constraints algorithms have been compared to other constraint based clustering algorithms such as K-Means clustering with constraints and the original Leader Ant clustering algorithm. The average accuracy of the proposed MCALA, MEALA and CEALA are found to be higher than the COP-K-Means, MCLA, MELA and CELA.

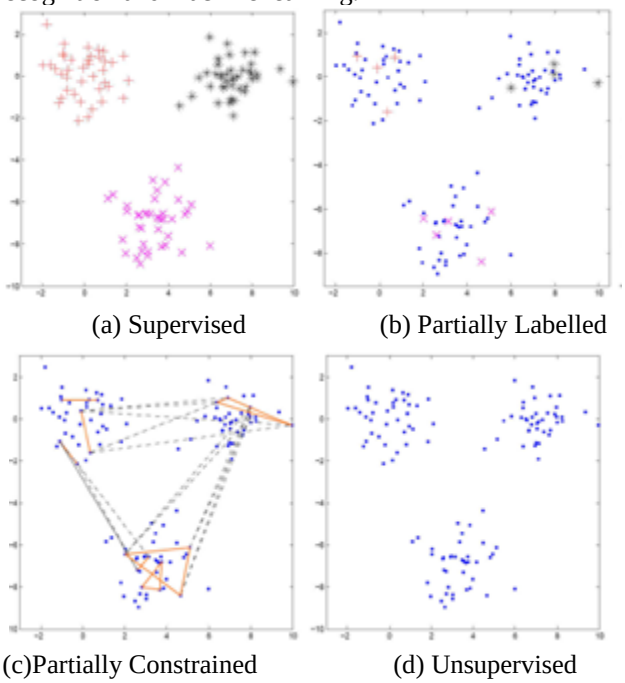
Keywords- clustering, constraint, artificial ants

I. INTRODUCTION

One of the most fundamental modes of understanding and learning is organizing data into sensible grouping. Cluster analysis is defined as the formal study of methods and algorithms for grouping, or clustering, objects according to measured or perceived intrinsic characteristics or similarity. The data clustering (unsupervised learning) is distinguished from classification or discriminant analysis (supervised learning) due to the absence of category information.

The aim of clustering is to obtain structure in data and is therefore exploratory in nature.The growth in the amount of data, on the other hand, the variety of available data like text, image, and video has also increased drastically. The popularity of RFID tags or transponders due to their low cost and small size has led to the deployment of millions of sensors that transmit data regularly. Eg: E-mails, blogs, Transaction data. Many of these data streams are unstructured, adding to the difficulty in analyzing them.

Due to increase in the volume and the variety of data, there should be advancements in methodology to automatically understand, process, and summarize the data. Data analysis techniques can be widely categorized into the following types. Confirmatory or Inferential and Exploratory or Descriptive. In pattern recognition, data analysis is mainly concerned with predictive modeling and the task is also referred to as learning. A clear distinction is made between learning problems that are Supervised (classification) and Unsupervised (clustering).The Supervised (classification) involving only labeled data while the Unsupervised (clustering) involving only unlabeled data. Fig. 1 illustrates this spectrum of different types of learning problems of interest in pattern recognition and machine learning.



(a) Supervised (b) Partially Labelled

(c)Partially Constrained (d) Unsupervised

Fig. 1 Learning problems

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

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

In Figure 1: (a), the Supervised (classification) involving only labeled data In (b), dots correspond to points without any labels. Points with labels are denoted by plus signs, asterisks, and crosses. In (c), the must-link and cannot-link constraints are denoted by solid and dashed lines, respectively. In (d), the unsupervised involving only unlabeled data.

The communication between the agents during the search process is not direct, instead they communicate indirectly by modifying the environment faced by each other. There is no single 'Ant Model', rather there exists a family of models, each inspired by a different aspect of ant behavior.These models include those inspired by: i. Ant-foraging behavior ii. Brood-sorting behavior iii. Cemetery formation behavior and iv. Cooperative transport.Ant colonies show high degrees of parallelism, self-organization and fault tolerance. These characteristics are essential for the computer systems.The nature inspired methods like ant-based clustering techniques have found success in solving clustering problems.

Ant-based clustering algorithms have been used in a large variety of applications. They are applied for web usage mining. Web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the web.The ant colony inspired methods can also be applied in many stages of the Electrocardiogram Interpretation Process.

II. Literature Survey

Many researches have been conducted on learning methods with constraints: partitioning algorithm [4,11], hierarchical algorithm [5], EM clustering with pairwise constraints [1,16,17], density-based algorithm [10], incremental constrained clustering [3], Support Vector Machine clustering with constraints [15] and co-clustering with constraints [8,9]. Constraints have been used in the classification problem to improve results such as in classification with pairwise constraints using SVM [7] and in discriminative learning framework with pairwise constraints for video object classification [13].

In [11], Wagstaff et al. proposed a K-Means algorithm with must-link and cannot-link constraints named COPKMeans .The COP-KMeans attempts to find a partition that minimizes the vector quantization error and also satisfies all the constraints. Wagstaff's empirical results show that using constraints can improve the performance of clustering [12].

In [5], Davidson et al. proposed the Hierarchical Clustering with constraints algorithm. The key idea is to use the constraints as guidelines in the clustering process. Their experimental results indicate that a small amount of constraints can improve the dendogram quality with respect to cluster purity and "tightness".

Ant colonies provide a means to formulate some powerful nature-inspired heuristics for solving the clustering problems. Several clustering methods based on ant behavior have been proposed in the literature. In several species of ants, workers have been reported to form piles of corpses – cemeteries – to clean the nests.

Cheretien and others [19,18] have performed experiments with the ant Lasius Niger to study the organization of cemeteries. Other experiments on the ant Phaidole Pallidula are also reported in Deneubourg et al. [20]. Brood sorting is observed by Franks and Sendova-Franks [21] in the ant Leptothorax Unifasciatus. Workers of this species gather the larvae according to their size. Franks and Sendova-Franks [21] have intensively analyzed the distribution of brood within the brood cluster.

Ant-based clustering sorting was first introduced by Deneubourg et al. [20] to explain the above mentioned phenomena of corpse clustering and larval sorting in ants.

It is an instance of the broad category of ant algorithms [22]. Ant-based clustering algorithms are based upon the brood sorting behavior of ants. Larval sorting and corpse cleaning by ant was first modeled by Deneubourg et al. for accomplishing certain tasks in robotics. Their work was actually focused on clustering objects by using group of real world robots. Their model is known as basic model (BM).

V. Ramos, F. Muge and P. Pina [23] noticed that the SACA would generate a large quantity of small clusters. They modified the Ant-based clustering by changing the movement paradigm. While the previous works all relied on random moving ants, their ants would move according to a trail of pheromones left on clustering formations. This would reduce the exploration of empty areas, where the pheromone would eventually evaporate. This algorithm was applied to the classification of stone images. They studied the performance of the algorithm on continuous clustering [24] and showed that this improves clustering performance for the ant-based clustering system.

III. METHODOLOGY

Clustering with constraints has become a topic of significant interest for many researchers because it allows to take into account the knowledge from the domain, expressed as a set of constraints, and thus to improve the efficiency of the analysis.

A. Leader Ant Clustering with Constraints

The three new methods proposed in this research rely on the Leader Ant Clustering algorithm, the main principles of this approach are considered. The new variants and the introduction of the three constraints - must-link, cannot-link, ε-constraints and accumulative constraints in the clustering process.

B.The Leader Ant Clustering Algorithm

The Leader Ant clustering algorithm (LA) is inspired from the chemical recognition system of ants.

1) The Chemical Recognition System of Ants

In the biological system, each ant possesses its own odor called label that is spread over its cuticle (its "skin"). This

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

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

label acts as an identity card and is partially determined by the genome of the ant and by the substances extracted from its environment. During their youth, ants learn to distinguish the labels of the colony members and learn a neuronal template of what should be the label of a nestmate. This template is continually updated and is used at each meeting between two ants, to decide if they should accept each other and exchange chemical cues. The continuous chemical exchanges between the nestmates lead to the establishment of a colonial odor that is shared and recognized by every nestmates, according to the "Gestalt theory" [6].

2) The Leader Ant Model

The underlying model of the Leader Ant algorithm (LA), although inspired by real ants system, has been adapted to match more specifically the objectives of the clustering problem and for performance purposes. In LA, an artificial ant is described by three parameters.

• The genome is associated with an unique object of the data set;

• The template is the same for all artificial ants and is either fixed or computed experimentally as the mean value of the distance values $d(i,j)$ estimated between Nb_{learn} couples of ants i and j randomly selected.

$$template = \frac{\sum_{Nb_{learn}} d(i,j)}{Nb_{learn}} \quad (1)$$

with $d(i,j)$, the distance value between the object associated to the ant i and the object of the ant j.

• The label reflects the nest membership of each artificial ant. At the beginning, this value is set to zero as no hypothesis is made concerning the initial membership of ants.

LA is a one-pass agglomerative algorithm that iteratively selects at random a new ant a (that has not been already assigned to a nest), and determines its label or nest membership by simulating $Nb_{meetings}$meetings with randomly selected ants from each existing nest k in $[0, NbMaxNests]$.

During these meetings, the ant a estimates the similarity of its genome with those of ants from the evaluated nest k. At the end, the distance $D(a,k)$ between the anta and the nest k is computed as the mean distance over the $Nb_{meetings}$ meetings.

$$D(a,k) = \frac{1}{Nb_{meetings}} \sum_{j=1}^{Nb_{meetings}} d\left(a, ant_j^k\right) \quad (2)$$

Where ant_j^k is the jth, $j \in [1, Nb_{meetings}]$, randomly selected ant from nest k. If no nest exists or if the mean distance value is under the template value, the ant creates its own new nest.

$$NbMaxNests = NbMaxNests + 1 (create\ a\ new\ nest)$$

$$Label_a = NbMaxNests \quad (3)$$

In the opposite case, the ant joins the nest with the lowest mean distance value by setting its label as follows:

$$Label_a = \arg min_{k \in [1, NbMaxNests]} D(a,k) \quad (4)$$

Finally, when all ants are assigned to a nest, the smallest nests whose size is under a fixed percentage of the total number of objects n can optionally be deleted and their ants reassigned to the other clusters.

C. Leader Ant Clustering with Must-Link and Cannot-Link Constraints (MCLA)

The MCLA algorithm (fig.4) which integrates the must-link and cannot-link constraints to LA algorithm is presented. Must-link constraints are transitive; must-link constraints (si, sj) and (sj, sk) imply that there exists a must-link constraint (si, sk). Thus, the two constraints can be combined into a single must-link constraint, namely (si, sj, sk). So, a given collection Con= of must-link constraints can be transformed into an equivalent collection $M = \{M1, M2, \ldots, Mr\}$ of constraints, by computing the transitive closure of $Con = [4]$.

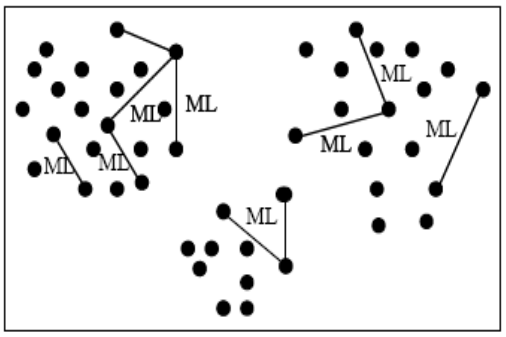


Fig. 2 Construction the transitive closure of the Must-link (ML) for theMCLA Algorithm

D. Leader Ant Clustering with Must-Link and E-Constraints (MELA)

In this section MELA algorithm is proposed (figure 4) that implements the must-link constraints and ε- constraint in LA algorithm. Similarly to MCLA algorithm, the transitive closure is constructed for the must-link constraints.

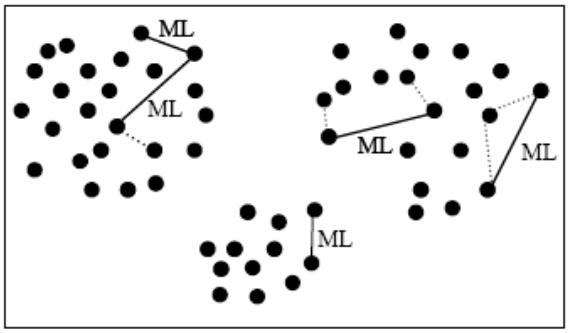


Fig. 3 Construction the transitive closure of the Must-link (ML) for the MELA Algorithm

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

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

It is important to notice that the transitive closure has to satisfy the ε-constraint: for any transitive closure M_k and for any point $s_p \in M_k$, there must be another point $s_q \in M_k$ such that $d(s_p, s_q) \leq \varepsilon$. If not, some points of the data set have to be found to complement the transitive closures. Figure 3 show that there are some points which relate to the transitive closure (by dashed line) and allow the ε-constraint to be satisfied. If there exists a transitive closure which cannot find the points to satisfy the ε-constraint, the algorithm will stop and the output is "No solution".

E. An Efficient Accmulative Constrained Clustering Algorithm

The MCALA algorithm (figure 5) which integrates the must-link, cannot-link and accumulative constraints to LA algorithm is presented. Given a set of points, a set of constraints and a partition of the point set into k subsets. Whether there is a partition of S into k subsets such that all the constraints in C ∪ {ML(s_i, s_j)} are satisfied. If so, output one such partition $\prod$ .The algorithm takes an input a single constraint at a time and depending on the properties of the constraint will attempt to greedily optimize the objective function f. If the constraint does not improve f then the constraint is passed over and the user chooses another. Furthermore, since finding a clustering to satisfy a must-link constraint between two points that are already constrained by cannot-link constraints cannot be done efficiently, it is also informed that the user of this situation pass over the constraints.

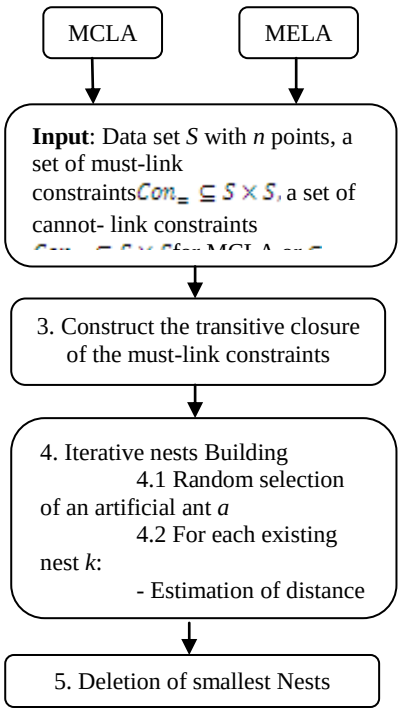


Fig. 4 MCLA and MELA algorithm

The transitive closure computation in the algorithm can be carried out as follows. Construct an undirected graph G with n node, one node for each point in the data set, and an edge between two nodes if the corresponding points appear together in the must-link constraint. Then, the connected components of G give the sets of objects in the transitive closure. Therefore, the connected components can be found in $O(n+m)$ time [2], where m is the number of must-link constraints (the edges of G). The set of transitive closure is used to reduce the number of points which has to be considered during the clustering.

IV. EXPERIMENTAL RESULTS

This section deals with the experimental evaluation of the proposed approach. The performance evaluation of the proposed approaches is evaluated based on the accuracy.

a) Data Sets

The real data sets from the Machine Learning repository named Glass, Iris, Wine, Thyroid and Soybean are used in this experiment. Tables 4.1 shows the number of points n, the number of attributes m and the number of clusters k in the partition of each data set.

TABLE 1 MAIN CHARACTERISTIC OF THE DATASETS FOR MCLA, LA and MCALA

Files	n	m	K
Glass	214	9	6
Iris	150	4	3
Wine	178	13	3

The experiments are reported to evaluate the efficiency of the proposed algorithm. The results of MCLA and LA algorithm are compared with the proposed Must-link and Cannot-link accumulative constraint (MCALA).

b) Evaluation Method

The data set used for the evaluation includes a "correct answer" or label for each data point. The labels are used in a post-processing step for evaluating the performance of our approaches.

To evaluate the similarity between the expected partition and the partition produced by our method, the Rand Index [14] is used. This measure is based on (n*(n-1))/2 pairwise comparisons between the n points of a data set D. For each pair of points $x_i$and $x_j$in D, a partition assigns them either to the same cluster or to different clusters.

Let us consider two partitions $P_1$and P_2, and let a be the number of decisions where the point $x_i$is in the same cluster as $x_j$in $P_1$and P_2. Let b be the number of decisions where the two points are placed in different clusters in both partitions. A total agreement can then be calculated using:

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

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

$$Rand(P_1, P_2) = \frac{a+b}{n*(n-1)/2}$$

This measure is used to evaluate the performance of proposed approach in all the experiments.

c) Comparison between LA,MCLA and MCALA

In this section, the evaluation of MCALA is compared with the MCLA and LA algorithm as it implements the same constraints.

The constraints were generated as follows: for each constraint, two points are randomly picked from the data set and checked their labels. If they have the same label, a must-link constraint is generated. Otherwise, generate a cannot-link constraint is generated. The Must-link Accumulative and cannot-link Accumulative constraints are also generated. 50 constraints of must-link and cannot-link is used as in [19], then 100 runs are conducted for each data set and the results are averaged.

The proposed Must-link and Cannot-link Accumulative Constraint in LA (MCALA) is compared with MCLA and LA on three aspects: average minimum accuracy, average mean accuracy and average maximum accuracy over 100 runs. Figure 6 shows the result for the glass data set. The graphical representation clearly shows that, the proposed MCALA approach has higher accuracy than the other. In the data sets, MCALA performs better than MCLA and LA.

TABLE 2 CLUSTERING ACCURACY FOR MCLA, LA AND MCALA WITH GLASS DATASET

	LA	MCLA	MCALA
Average Minimum Accuracy	68	70	72
Average Mean Accuracy	80	85	88
Average Max Accuracy	83	89	92

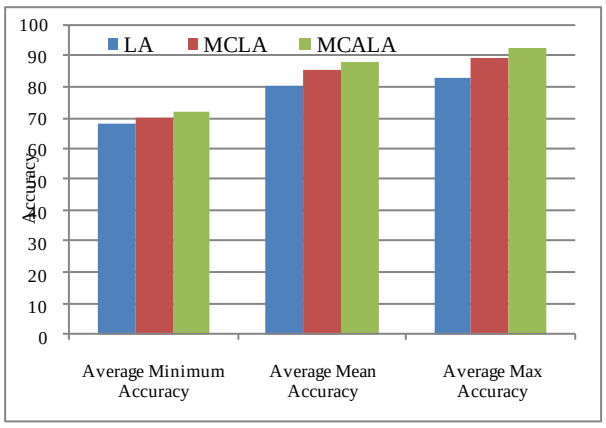


Fig. 6 Accuracy Comparison with Glass data set

V. CONCLUSION AND SCOPE FOR FUTURE WORK

Ant-based clustering algorithms are an appropriate alternative to traditional clustering algorithms. The algorithm has a number of features that make it an interesting study of cluster analysis. The nature of the algorithm makes it fairly robust to the effects of outliers within the data.An efficient algorithm is developed for these sufficient conditions that incrementally allows feedback and tested it using simulated feedback from small amounts of labeled data. In this paper, new constraint based clustering algorithms named MCALA, MEALA and CEALA that rely on the Leader Ant clustering algorithm. The must-link, cannot-link constraints, ε- constraints and accumulative constraints are integrated to improve the quality of clustering. the accumulative constraints provide significant accuracy.

REFERENCES

[1] S. Basu, I. Davidson and K. Wagstaff, Constrained Clustering: Advances in Algorithms, Theory, and Applications, Publisher Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, First Edition, August 18, 2008.

[2] T. Corment, C. Leiserson, R. Rivest and C. Stein, Introduction to algorithms, Second Edition. MIT Press and McGraw-Hill, Cambridge, 2001.

[3] I. Davidson, M. Ester and S.S. Ravi, "Efficient incremental constrained clustering". In Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007, August 12-15, San Jose, California, USA.

[4] I. Davidson, M. Ester and S.S. Ravi, "Clustering with constraints: Feasibility issues and the K-means algorithm", in proc. SIAM SDM 2005, Newport Beach, USA .

[5] I. Davidson,M. Ester and S.S. Ravi, "Agglomerative hierarchical clustering with constraints: Theoretical and empirical results", in Proc. of Principles of Knowledge Discovery from Databases, PKDD 2005.

[6] B. Hölldobler and E. Wilson (1990), The Ants, Chapter colony odor and kin recognition. p. 197-208. Spinger Verlag, Berlin, Germany.

[7] N. Nguyen and R. Caruana, "Improving classification with pairwise constraints: A margin-based approach", in proc. of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD'08).

[8] R.G. Pensa, C. Robardet and J.-F. Boulicaut (2006), "Co-Classification sous contraintes", in proc CAp'06, Trégastel, France, p.155-170, Presses Universitaires de Grenoble.

[9] R.G. Pensa and J.-F. Boulicaut (2008), "Co-Classification sous contraintes par la somme des résidus qudratiques", in proc. des 8ème Journées Francophones Extraction et Gestion de Connaissances EGC'08, p.655-666.

[10] C. Ruiz, M. Spiliopoulou and E. Menasalvas, "C-DBSCAN: Densitybased clustering with constraints", In RSFDGrC 2007, LNAI 4482, pp. 216-223, Springer-Verlag Berlin Heidelberg.

[11] K. Wagstaff, C. Cardie, S. Rogers and S. Schroedl, "Constrained Kmeans clustering with background knowledge", in: Proc. Of 18th Int. Conf. on Machine Learning ICML'01, p. 577 - 584.

[12] K. Wagstaff, Intelligent clustering with instance-level constraints, PhD Thesis of Computer Science, 2002, Cornell University, USA.

[13] R. Yan, J. Zhang, J. Yang and A. Hauptmann, "A discriminative learning framework with pairwise constraints for Video Object Classification", in IEEE Conference on Computer Vision and pattern recognition (CVPR), 2004, Washington, DC.

[14] K. Wagstaff, "When is Constrained Clustering Beneficial, and Why ?", in proc. of the Twenty-first National Conference on Artificial Intelligence (AAAI), July 2006.

[15] Y. Hu, J. Wang, N. Yu and X.-S. Hua, "Maximum Margin Clustering with Pairwise Constraints", in proc. of the Eighth IEEE International Conference on Data Mining(ICDM) , 253-262, 2008

[16] Q. Zhao and D.J. Miller, "Mixture modeling with pairwise, instancelevel class constraints", Neural Computation, 17(11): 2482-2507, November 2005.

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

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

[17] M.H. Law, Clustering, Dimentioanality Reduction, and Side Information, PhD Thesis of Computer Science , Michigan State University, USA, 2006.
[18] L. Chretien, "Organization Spatiale du Materiel Provenant de L'excavation du nid chez Messor Barbarus et des Cadavres d'ouvrieres chez Lasius niger {Hymennopterae: Formicidae)", Ph.d. thesis, Universite Libre dr Bruxelles, 1996.
[19] E. Bonabeau, M. Dorigo and G. Theraulaz, "Swarm Intelligence: from natural to artificial systems", Oxford University Press, Inc., New York, NY, 1999.
[20] J.-L. Deneubourg, S. Gross, N. Franks, A. Sendova-Franks, C. Detrain and L. Chretien, " The dynamics of collective sorting: Robot-like ants and ant-like robots", In Proceedings of the First International Conference on Simulation of Adaptive Behavior: From Animals to Animats, Cambridge, MA, MIT Press, 1991, pp. 356-363.
[21] N.R. Franks and A.B. Sendova-Franks, "Brood sorting by ants: distributing the workload over the work surface", Behav. Ecol. Sociobiol., 1992, 30:109-123.
[22] M. Dorigo , E. Bonabeau and G. Theraulaz, "Ant algorithms and stigmergy", Future Generation Computer Systems, 16(8), 2000, pp. 851-871.
[23] V. Ramos, F. Muge and P.Pina, "Self-Organized Data and Image Retrieval as a Consequence of Inter-DynamicSynergistic Relationships in Artificial Ant Colonies", In J. Ruiz-del-Solar, A. Abrahan and M. Koppen Eds., Soft-Computing Systems – Design, Management and Applications, Frontiers in Artificial Intelligence and Applications: IOS Press, Amsterdam, v. 87, 2002, pp. 500-509.
[24] V. Ramos and Ajith Abraham, "Swarms on continuous data", In CEC'03 Congress on Evolutionary Computation, IEEE Press, 2003, pp. 1370-1375.

BIOGRAPHY

Mrs. S. Sridevikarumari completed MCA., M.Phil., in Computer Science and currently working as an Assistant Professor, Dept. of Computer Applications in Pioneer College of Arts and Science. Ten years of experience in teaching and presented papers in various National and International conferences. Area of research is Data mining.