ISSN No. 0976-5697

Volume 2, No. 4, July-August 2011

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Optimizing Fuzzy Clustering using Swarm Intelligence in Data Mining

Poonam Chalotra*
Computer Science Deptt.
SBBSIET,
Jalandhar India
poonamchalotra@gmail.com

Harpreet Kaur
Computer Science Deptt.
SBBSIET,
Jalandhar India
er.harpreetarora@gmail.com

Abstract: Data mining is a powerful new technology, which aims at the extraction of hidden predictive information from large databases. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The process of knowledge discovery from databases requires fast and automatic clustering of very large datasets. It deals with large databases that impose severe computational requirements on clustering analysis. A family of nature inspired algorithms, known as Swarm Intelligence (SI), has recently emerged as tool to meet such requirements on number of real world clustering problems. Algorithms based on Swarm Intelligence are inspired from the collective intelligence emerging from the behavior of a group of social insects like bees, termites and wasps. In this paper we have discussed the use of Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO) for clustering in Data Mining. The performance of Fuzzy C-means is enhanced when used with PSO optimization and ACO optimization.

Keywords: Swarm Intelligence (SI), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Fuzzy C-means Clustering (FCM).

I. INTRODUCTION

Data mining is a powerful new technology, which aims at the extraction of hidden predictive information from large databases. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The process of knowledge discovery from databases necessitates fast and automatic clustering of very large datasets with several attributes of different types [3]. This poses a severe challenge before the classical clustering techniques. Cluster analysis is a technique for breaking data down into related components in such a way that patterns and order becomes visible. It aims at sifting through large volumes of data in order to reveal useful information in the form of new relationships, patterns, or clusters, for decision-making by a user. Clusters are natural groupings of data items based on similarity metrics or probability density models. Cluster analysis has the virtue of strengthening the exposure of patterns and behavior as more and more data becomes available [6]. A cluster has a center of gravity which is basically the weighted average of the cluster. Membership of a data item in a cluster can be determined by measuring the distance from each cluster center to the data point [5]. Recently a family of nature inspired algorithms, known as Swarm Intelligence (SI), has attracted several researchers from the field of pattern recognition and clustering. Clustering techniques based on the SI tools have reportedly outperformed many classical methods of partitioning a complex real world dataset. Swarm Intelligence is a relatively new interdisciplinary field of research, which has gained huge popularity in these days. Algorithms belonging to the domain, draw inspiration from the collective intelligence emerging from the behavior of a group of social insects (like bees, termites and wasps). In a community, these insects even with very limited individual capability can cooperatively perform many complex tasks necessary for their survival. Problems like finding and storing foods, selecting and picking up materials for future usage require a detailed planning, and are solve by insect colonies without any kind of supervisor or controller.

II. DATA MINING AND CLUSTERING

Data Mining or Knowledge discovery refers to a variety of techniques that have developed in the fields of databases, machine learning and pattern recognition [3]. The process of finding useful patterns and information from raw data is often known as Knowledge discovery in databases or KDD. Data mining is a particular step in this process involving the application of specific algorithms for extracting patterns (models) from data [4]. Data mining is a powerful new technology, which aims at the extraction of hidden predictive information from large databases. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The process of knowledge discovery from databases necessitates fast and automatic clustering of very large datasets with several attributes of different types [3].

Clustering is a widely used technique in data mining application for discovering patterns in underlying data. It is an important tool for a variety of applications in data mining, statistical data analysis, data compression and vector quantization, aims gathering data into clusters (or groups) such that the data in each cluster shares a high degree of similarity while being very dissimilar to data from other clusters [2]. Clustering has many applications, including part family formation for group technology, image segmentation, information retrieval, web pages grouping, market segmentation, and scientific and engineering analysis [1]. The goal of clustering is to group data into clusters such that the similarities among data members within the same cluster are maximal while similarities among data members from different clusters are minimal. Representing data by fewer clusters necessarily loses certain fine details, but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, the clustering is a typical unsupervised learning technique for grouping similar data points. It assigns a large number of data points to a smaller number of groups such that data points in the same group share the same properties while, in different groups, they are dissimilar. Data mining deals with large databases that impose on clustering analysis additional severe computational requirements. Most traditional clustering algorithms are limited in handling datasets that contain categorical attributes. However, datasets with categorical types of attributes are common in real life data mining problem. For these data sets, no inherent distance measure, like the Euclidean distance, would work to compute the distance between two categorical objects.

III. CLUSTERING USING SWARM INTELLIGENCE

Clustering techniques based on the Swarm Intelligence tools have reportedly outperformed many classical methods of partitioning a complex real world dataset. Swarm Intelligence is a relatively new interdisciplinary field of research, which has gained huge popularity in these days. Data mining and particle swarm optimization do not have common properties. However, they can be integrated to form a method. Cluster analysis has become an important technique in exploratory data analysis, pattern recognition, machine learning, neural computing, and other engineering. The clustering aims at identifying and extracting significant groups in underlying data. The four types of clustering: partitioned clustering, hierarchical clustering, density based clustering and grid-based clustering. Document clustering is a fundamental operation used in unsupervised document organization, automatic topic extraction, and information retrieval. Fast and high-quality document clustering algorithms play an important role in effectively navigating, summarizing, and organizing information. In the field of clustering, K-means algorithm is the most popularly used algorithm to find a partition that minimizes mean square error (MSE) measure. Although K-means is an extensively useful clustering algorithm, it suffers from several drawbacks. The objective function of the K-means is not convex and hence it may contain local minima. Consequently, while minimizing the objective function, there is possibility of getting stuck at local minima [7]. The performance of the K-means algorithm depends on the initial choice of the cluster centers. Besides, the Euclidean norm is sensitive to noise or outliers. Hence K-means algorithm should be affected by noise and outliers [10], [8]. In addition to the K-means algorithm, several algorithms, such as Genetic Algorithm (GA) [8], [11] and Self-Organizing Maps (SOM) [9], have been used for document clustering. A hybrid document clustering algorithm based on PSO is proposed in [12]. The PSO clustering algorithm performs a globalized search in the entire solution space. The results of the experiments showed the hybrid PSO algorithm can generate more compact clustering results than the K-means algorithm. An evolutionary PSO learning-based method to optimally cluster *N* data points into *K* clusters is introduced in [14]. The hybrid PSO and K-means, with a novel alternative metric algorithm is called Alternative KPSO-clustering (AKPSO) method. This is developed to automatically detect the cluster centers of geometrical structure data sets. In AKPSO algorithm, the special alternative metric is considered to improve the traditional K-means clustering algorithm to deal with various structure data sets.

The general idea for data clustering is that isolated items should be picked up and dropped at some other location where more items of that type are present. Ramos et al. [16] proposed *ACLUSTER* algorithm to follow real ant-like behaviors as much as possible. In that sense, bio-inspired spatial transition probabilities are incorporated into the system, avoiding randomly moving agents, which encourage the distributed algorithm to explore regions manifestly without interest. The strategy allows guiding ants to find clusters of objects in an adaptive way. In order to model the behavior of ants associated with different tasks (dropping and picking up objects), the use of combinations of different response thresholds was proposed. There are two major factors that should influence any local action taken by the ant-like agent: the number of objects in its neighborhood, and their similarity. Lumer and Faieta [15] used an average similarity, mixing distances between objects with their number, incorporating it simultaneously into a response threshold function like the algorithm proposed by Deneubourg et al. [14]. Kuo et al. [19], [18] proposed ant K-means (AK) clustering method. AK algorithm modifies the K-means as locating the objects in a cluster with the probability, which is updated by the pheromone, while the rule of updating pheromone is according to total within cluster variance (TWCV). Tsai et al. [17] proposed a novel clustering method called ant colony optimization with different favor algorithm which performed better than the fast self-organizing map (SOM) K-means approach and genetic K-means algorithm.

Web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the Web. Web usage mining has become very critical for effective Web site management, creating adaptive Web sites, business and support services, personalization, and network traffic flow analysis and so on. Abraham and Ramos [20] proposed an ant clustering algorithm to discover Web usage patterns (data clusters).

IV. FUZZY C-MEANS CLUSTERING

The central idea in fuzzy clustering is the non-unique partitioning of the data in a collection of clusters. Clustering is the process of recognizing natural groupings or clusters in multidimensional data based on some similarity measures [21]. Distance measurement is generally used for evaluating similarities between patterns. The conventional clustering algorithms in data mining like k-means algorithm have difficulties in handling the challenges posed by the collection of natural data which is often vague and uncertain. The modeling of imprecise and qualitative knowledge, as well as handling of uncertainty at various stages is possible through the use of fuzzy sets. Fuzzy logic is capable of supporting to a reasonable extent, human type reasoning in natural form by allowing partial membership for data items in fuzzy subsets. Integration of fuzzy logic in data mining has become a powerful tool in handling natural data. In contrast to clustering data objects in a unique cluster, fuzzy clustering algorithms result in membership values between 0 and 1 that indicate the degree of membership for each object to each of the clusters. Fuzzy c-means clustering involves two processes: the calculation of cluster centers and the assignment of points to these centers using a form of Euclidian distance. This process is repeated until the cluster centers stabilize. The algorithm is similar to k-means clustering in many ways but it assigns a membership value to the data items for the clusters within a range of 0 to 1. So it incorporates fuzzy set's concepts of partial membership and forms overlapping clusters to support it. The objective function of the fuzzy clustering is to minimize the equation:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \left\| x_i - c_j \right\|^2 \qquad (3)$$

Where m is any real number greater than 1, it is set to 2 by Bezdek, u_{ij} is the degree of membership of x_i in the cluster j and $||xi - zj||^2$ is the Euclidean distance from sample points x_i to cluster center z_j.

The algorithm needs a fuzzification parameter m in the range [1,n] which determines the degree of fuzziness in the clusters. When m reaches the value of 1 the algorithm works like a crisp partitioning algorithm and for larger values of m the overlapping of clusters is tend to be more. The algorithm calculates the membership value μ with the formula,

$$\mu_j(x_i) = \frac{\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{p} \left(\frac{1}{d_{ki}}\right)^{\frac{1}{m-1}}} \qquad (4)$$

Where,
$\mu_j(x_i)$: is the membership of x_i in the j^{th} cluster
d_{ji} : is the distance of x_i in cluster C_j
m: is the fuzzification parameter
p: is the number of specified clusters
d_{ki}: is the distance of x_i in cluster C_k

We modify the degree of fuzziness in x_i's current membership and multiply this by x_i. The product obtained is divided by the sum of the fuzzified membership. In this way new centroids are calculation with these membership values using equation (4) for clusters.

$$C_j = \frac{\sum_i \left[\mu_j(x_i)\right]^m x_i}{\sum_i \left[\mu_j(x_i)\right]^m} \qquad (5)$$

Where
C_j: is the center of the j^{th} cluster
x_i: is the i^{th} data point
μ_j: the function which returns the membership
m: is the fuzzification parameter

V. CONCLUSION

Clustering techniques based on the SI tools have reportedly outperformed many classical methods of partitioning a complex real world dataset. In this paper, we have reviewed many algorithms for data clustering. We concluded that the clustering in Data mining can be improved with the use of Swarm Intelligence techniques like PSO and ACO for optimization.

VI. REFERENCES

[1] Pham, D.T. and Afify, A.A., "Clustering techniques and their applications in engineering", Journal of Mechanical Engineering Science, 2006.

[2] Jain, A.K. and Dubes, "R.C. Algorithms for Clustering Data", Prentice Hall.

[3] Mitra S, Pal S.K. and Mitra P, "Data mining in soft computing framework: A survey", IEEE Transactions on Neural Networks, 2002.

[4] Inmon, W. H., "The data warehouse and data mining", Communication, ACM, 1996.

[5] Fayyad U., Uthurusamy R., "Data mining and knowledge discovery in databases", Communication, ACM, 1996.

[6] Berkhin P., "A Survey of Clustering Data Mining Techniques", http://citeseer.ist.psu.edu/berkhin02survey.

[7] Selim S.Z, Ismail M.A., "K-means Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality", IEEE Transaction on Pattern Analysis and Machine Intelligence.

[8] Jones G., Robertson A., Santimetvirul C., Willett P., "Non-hierarchic document clustering using a genetic algorithm", Information Research, 1(1).

[9] Merkl D., "Text mining with self-organizing maps", Handbook of data mining and knowledge, Oxford University Press, 2002.

[10] Wu K.L., Yang M.S., "Alternative C-means Clustering Algorithms", Pattern Recognition, 2002, 35, 2267-2278.

[11] Raghavan V.V, Birchand K., "A clustering strategy based on a formalism of the reproductive process in a natural system", Proceedings of the Second International Conference on Information Storage and Retrieval, 1979, 10-22.

[12] Cui X, Potok TE., "Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm", Journal of Computer Sciences (Special Issue), ISSN 1549-3636, pp. 27-33,2005.

[13] Fun Y, Chen C.Y., "Alternative KPSO-Clustering Algorithm", Tamkang Journal of Science and Engineering, 8(2), 2005, 165-174.

[14] Deneubourg JL, Goss S, Franks N, Franks AS, Detrain C, Chretien L., "The dynamics of collective sorting: Robot-like ants and ant-like robots", Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animats, Cambridge, MA: MIT Press, 1, 356-365,1991.

[15] Lumer E.D, Faieta B., "Diversity and Adaptation in Populations of Clustering Ants", Clio D, Husbands P, Meyer J and Wilson S (Eds.), Proceedings of the Third International Conference on Simulation of Adaptive

Behaviour: From Animals to Animats 3, Cambridge, MA: MIT Press, 501-508, 1994.

[16] Ramos V, Muge, F, Pina, P., “Self-organized data and image retrieval as a consequence of inter-dynamic synergistic relationships in artificial ant colonies”, Soft Computing Systems - Design, Management and Applications, Proceedings of the 2nd International Conference on Hybrid Intelligent Systems, IOS Press, 500-509, 2002.

[17] Tsai C.F, Tsai C.W, Wu H.C, Yang T., “ACODF: a novel data clustering approach for data mining in large databases”, Journal of Systems and Software, Volume 73, Issue 1, 133-145, 2004.

[18] Kuo R.J, Wang H.S, Hu T.L, Chou S.H., “Application of ant K-means on clustering analysis, Computers & Mathematics with Applications”, Volume 50, Issues 10-12, 1709-1724,2005.

[19] Admane L, Benatchba K, KoudilM, Siad L,Maziz S (2006), “AntPart: an algorithm for the unsupervised classification problem using ants”, Applied Mathematics and Computation (http://dx.doi.org/10.1016/j.amc.2005.11.130).

[20] Abraham A, Ramos V., “Web Usage Mining Using Artificial Ant Colony Clustering and Genetic Programming”, 2003 IEEE Congress on Evolutionary Computation (CEC2003), Australia, IEEE Press, ISBN 0780378040, 1384-1391.

[21] Jain, A.K., Murty, M.N. and Flynn, P.J., “Data clustering: A review. ACM Computing Survey”, 1999, 31 (3), 264-323.