An Analysis of Particle Swarm Optimization with Data Clustering-Technique for Optimization in Data Mining.

Amreen Khan[1], Prof. Dr. N.G.Bawane [2], Prof. Sonali Bodkhe[3]

1 Master of Technology [CSE], Computer Science & Engineering Department,
GHRCE, Nagpur.

2 Professor, Faculty of P.G.Depatrment of Computer science & Engineering,

3 Professors, Dept of CSE, GHRCE, Nagpur.

***Abstract*—Data clustering is a popular approach for automatically finding classes, concepts, or groups of patterns. Clustering aims at representing large datasets by a fewer number of prototypes or clusters. It brings simplicity in modeling data and thus plays a central role in the process of knowledge discovery and data mining. Data mining tasks require fast and accurate partitioning of huge datasets, which may come with a variety of attributes or features. This imposes severe computational requirements on the relevant clustering techniques. A family of bio-inspired algorithms, well-known as Swarm Intelligence (SI) has recently emerged that meets these requirements and has successfully been applied to a number of real world clustering problems. This paper looks into the use of Particle Swarm Optimization for cluster analysis. The effectiveness of Fuzzy C-means clustering provides enhanced performance and maintains more diversity in the swarm and also allows the particles to be robust to trace the changing environment.**

Keywords- Particle Swarm Optimization (PSO), Fuzzy C-Means Clustering (FCM), Data Mining, Data Clustering .

I. INTRODUCTION

Particle Swarm Optimization (PSO) was originally designed and introduced by Eberhart and Kennedy [1]. The PSO is a population based search algorithm based on the simulation of the social behavior of birds, bees or a school of fishes. PSO originally intends to graphically simulate the graceful and unpredictable choreography of a bird folk. Each individual within the swarm is represented by a vector in multidimensional search space. This vector has also one assigned vector which determines the next movement of the particle and is called the velocity vector. The PSO also determines how to update the velocity of a particle. Each particle updates its velocity based on current velocity and the best position it has explored so far; and also based on the global best position explored by swarm [2].

The PSO process then is iterated a fixed number of times or until a minimum error based on desired performance index is achieved. It has been shown that this simple model can deal with difficult optimization problems efficiently. The PSO was originally developed for real valued spaces but many problems are, however, defined for discrete valued spaces where the domain of the variables is finite. Classical examples of such problems are: integer programming, scheduling and routing [4]. In 1997, Kennedy and Eberhart introduced a discrete binary version of PSO for discrete optimization problems [5]. In binary PSO, each particle represents its position in binary values which are 0 or 1. Each particle's value can then be changed (or better say mutate) from one to zero or vice versa. In binary PSO the velocity of a particle defined as the probability that a particle might change its state to one.

Upon introduction of PSO algorithm, it was used in number of engineering applications. Using binary PSO, Wang and Xiang [6] proposed a high quality splitting criterion for codebooks of tree-structured vector quantizers (TSVQ). Using binary PSO, they reduced the computation time too. Binary PSO is used to train the structure of a Bayesian network [7]. Although PSO is successfully used in number of engineering applications, but this algorithm still has some shortcomings. In novel binary PSO, the velocity of a particle is its probability to change its state from its previous state to its complement value, rather than the probability of change to 1. In this new definition the velocity of particle and also its parameters has the same role as in real valued version of the PSO. There are also other versions of binary PSO. In [8] authors add birth and mortality to the ordinary PSO. Also fuzzy system can be used to improve the capability of the binary PSO.

The remainder of the paper is organized as follows: Section II surveys Data Mining and Data Clustering. The Fuzzy C-Means Clustering is presented in Section III. Finally, Section IV concludes the paper.

II. DATA MINING AND DATA CLUSTERING

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 . This poses a severe challenge before the classical clustering techniques. 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; this 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). When acting as a community, these insects even with very limited individual capability can jointly (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 solved by insect colonies without any kind of supervisor or controller. Particle Swarm Optimization (PSO)) is another very popular SI algorithm for global optimization over continuous search spaces. PSO has attracted the attention of several researchers all over the world resulting into a huge number of variants of the basic algorithm as well as many parameter automation strategies.

Data mining has been called exploratory data analysis, among other things. Masses of data generated from cash registers, from scanning, from topic specific databases throughout the company, are explored, analyzed, reduced, and reused. Searches are performed across different models proposed for predicting sales, marketing response, and profit. Classical statistical approaches are fundamental to data mining. Automated AI methods are also used. However, systematic exploration through classical statistical methods is still the basis of data mining. Some of the tools developed by the field of statistical analysis are harnessed through automatic control (with some key human guidance) in dealing with data. A variety of analytic computer models have been used in data mining. The standard model types in data mining include regression (normal regression for prediction, logistic regression for classification), neural networks, and decision trees. These techniques are well known.

Data clustering is the process of identifying natural groupings or clusters, within multidimensional data, based on some similarity measure (e.g. Euclidean distance) [9, 10]. The term “clustering” is used in several research communities to describe methods for grouping of unlabeled data. These communities have different terminologies and assumptions for the components of the clustering process and the context in which clustering is used.

Clustering algorithms are used in many applications, such as data mining [11], compression [12], image segmentation [13-15], machine learning [16], etc. A cluster is usually identified by a cluster center (or *centroid*). Data clustering is a difficult problem as the clusters in data may have different shapes and sizes .Furthermore; it is usually not known how many clusters should be formed [17]. Most clustering algorithms are based on two popular techniques known as hierarchical and partitional clustering [18, 19]. In hierarchical clustering, the output is "a tree showing a sequence of clustering with each clustering being a partition of the data set" [19]. Such algorithms have the following advantages [18] the number of clusters need not be specified *a priori*, and they are independent of the initial conditions.

However, hierarchical clustering techniques suffer from the following drawbacks:

- They are static, i.e. data points assigned to a cluster cannot move to another cluster.
- They may fail to separate overlapping clusters due to a lack of information about the global shape or size of the clusters.

On the other hand, partitional clustering algorithms partition the data set into a specified number of clusters. These algorithms try to minimize certain criteria (e.g. a square error function) and can therefore be treated as optimization problems. The advantages of hierarchical algorithms are the disadvantages of the partitional algorithms and *vice versa*. Partitioned clustering techniques are more popular than hierarchical techniques in pattern recognition [9].

Partitioned clustering aims to optimize cluster centers, as well as the number of clusters. Most clustering algorithms require the number of clusters to be specified in advance. Finding the "optimum" number of clusters in a data set is usually a challenge since it requires *a priori* knowledge, and/or ground truth about the data, which is not always available. The problem of finding the optimum number of clusters in a data set has been the subject of several research efforts [20, 21].This paper uses a new approach called Fuzzy C means Clustering (FCM) using a Particle Swarm Optimization algorithm the approach uses some of the ideas presented by Kuncheva and Bezdek [23].

III. FCM CLUSTERING

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method was developed by Dunn in 1973 and improved by Bezdek in 1981 and it is frequently used in pattern recognition.

FCM is a method of clustering which allows a data point to belong to two or more clusters. The detailed description of FCM method can be found in [24, 25]. Several methods have been used for estimating the optimal number of clusters. Fuzzy clustering allows each feature vector to belong to more than one cluster with different membership degrees (between 0 and 1) and vague or fuzzy boundaries between clusters. Growth in both the theory and applications of this clustering methodology has been steady since its inception

FCM is a pretty standard least squared errors model that generalizes an earlier and very popular non-fuzzy (or hard, which in this context simply means not "soft") c-means model that produces hard (or crisp or non-soft) clusters in the data. And, in turn, FCM itself can be generalized in many, many ways. For example, including but not limited to: the memberships have been generalized to include possibilities; the prototypes have evolved from points to linear varieties to hyper quadrics to shells to regression functions, etc.; the distance used has been generalized to include non-inner

product induced and hybrid distances; there are many relatives of FCM for the dual problem called relational fuzzy c-means which is useful when the data are not object vectors, but instead, relational values (such as similarities) between pairs of objects, as for example, often happens in data mining; there are many acceleration techniques for FCM; there are very large data versions of FCM that utilize both progressive sampling and distributed clustering; there are many techniques that use FCM clustering to build fuzzy rule bases for fuzzy systems design; and there are numerous applications of FCM in virtually every major application area of clustering.

It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|X_i - c_j\right\|^2 \qquad 1 \le m < \infty$$

Where *m* is any real number greater than 1,

uij is the degree of membership of *xi* in the cluster *j*,

xi is the *i*th of d-dimensional measured data,

cj is the d-dimension center of the cluster,

and ||*|| is any norm expressing the similarity between any measured data and the center.

By iteratively updating the cluster centers and the membership grades for each data point, FCM iteratively moves the cluster centers to the "right" location within a data set. Performance depends on initial centroids. For a robust approach there are two ways which is described below.

1) Using an algorithm to determine all of the centroids. (For example: arithmetic means of all data points)

2) Run FCM several times each starting with different initial centroids.

With fuzzy *c*-means, the centroid of a cluster is computed as being the mean of all points, weighted by their degree of belonging to the cluster. The degree of being in a certain cluster is related to the inverse of the distance to the cluster.

IV. CONCLUSION

The PSO is a is an efficient global optimizer for continuous variable problems .The advantages of the PSO are very few parameters to deal with and the large number of processing elements, so called dimensions, which enable to fly around the solution space effectively. Algorithm modifications improve PSO local search ability

Many algorithms have been devised for clustering. They are divided into two categories: the parametric approach and the nonparametric approach. The clustering method described in this paper is a parametric approach. It starts with an estimate of the local distribution, which efficiently avoids pre-assuming the cluster number. Then the seed clusters that come from a similar distribution are merged by this clustering program was applied to both artificial and benchmark data classification and its performance is proven better than the well-known k-means algorithm.

REFERENCES

[1] R. Eberhart, and J. Kennedy, (1995) A New Optimizer Using Particles Swarm Theory, Proc. Sixth International Symposium on Micro Machine and Human Science (Nagoya, Japan), IEEE Service Center, Piscataway, NJ, pp. 39-43.

[2] J. Kennedy, and R Eberhart, (1995), Particle Swarm Optimization, IEEE Conference on Neural Networks, pp. 1942-1948, (Perth, Australia), Piscataway, NJ, IV, 1995.

[3] J. Kennedy and R. Eberhart. Swarm Intelligence. Morgan Kaufmann Publishers, Inc., San Francisco, CA, 2001.

[4] A. P. Engelbrecht. (2005), Fundamentals of Computational Swarm Intelligence. Wiley, 2005.

[5] Kennedy, J.; Eberhart, R.C. (1997), A discrete binary version of the particle swarm algorithm, IEEE Conference on Systems, Man, and Cybernetics, 1997.

[6] M. Fatih Tasgetiren. & Yun-Chia Liang, (2007), A Binary Particle Swarm Optimization Algorithm for Lot Sizing Problem Journal of Economic and Social Research vol 5. Elsevier pp. 1-20.

[7] Wen-liang Zhong, Jun Zhang, Wei-neng Chen, (2007), A novel discrete particle swarm optimization to solve traveling salesman problem, Evolutionary Computation, 2007. CEC 2007. IEEE Congress on, Singapore, Sept. 2007, pp. 3283-3287.

[8] J. Sadri, and Ching Y. Suen, (2006), A Genetic Binary Particle Swarm Optimization Model, *IEEE Congress on Evolutionary Computation*, Vancouver, BC, Canada, 2006.

[9] A.K. Jain, M.N. Murty, P.J. Flynn, Data Clustering: A Review, ACM Computing Surveys, vol. 31(3), 264-323,1999.

[10] A.K. Jain, R. Duin, J. Mao, Statistical Pattern Recognition: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22 (1), 4-37, 2000.

[11] D. Judd, P. Mckinley, A.K. Jain, Large-scale Parallel Data Clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20 (8), 871-876, 1998.

[12] H.M. Abbas, M.M. Fahmy, Neural Networks for Maximum Likelihood Clustering, Signal Processing, vol. 36(1), 111-126, 1994.

[13] G.B. Coleman, H.C. Andrews, Image Segmentation by Clustering, Proc. IEEE, vol. 67, 773-785, 1979.

[14] S. Ray, R.H. Turi, Determination of Number of Clusters in K-Means Clustering and Application in Color Image Segmentation, Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99), Calcutta, India, 137-143, 1999.

[15] C. Carpineto, G. Romano, A Lattice Conceptual Clustering System and Its Application to Browsing Retrieval, Machine Learning, vol. 24(2), 95-122, 1996.

[16] C.-Y. Lee, E.K. Antonsson, Dynamic Partitional Clustering Using Evolution Strategies, In The Third Asia-Pacific Conference on Simulated Evolution and Learning, 2000.

[17] G. Hamerly, C. Elkan, Learning the K in K-means, 7th Annual Conference on Neural Information Processing Systems, 2003.

[18] H. Frigui and R. Krishnapuram, A Robust Competitive Clustering Algorithm with Applications in Computer Vision, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21(5), 450-465, 1999.

[19] Y. Leung, J. Zhang, Z. Xu, Clustering by Space-Space Filtering, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22(12), 1396-1410, 2000.-12

[20] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On Clustering Validation Techniques, Intelligent Information Systems Journal, Kluwer Pulishers, vol. 17(2-3), 107-145, 2001.-13

[21] S. Theodoridis and K. Koutroubas, Pattern Recognition, Academic Press, 1999.-14

[22] C. Rosenberger and K. Chehdi, Unsupervised Clustering Method with Optimal Estimation of the Number of Clusters: Application to Image

Segmentation, International Conference on Pattern Recognition (ICPR'00), vol. 1, 1656-1659 (2000).

[23] L. Kuncheva and J. Bezdek, Nearest Prototype Classification: Clustering, Genetic Algorithms, or Random Search?, IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, vol. 28(1), 160-164, 1998.

[24] Xie XL, Beni G: A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1991.

[25] Tibshirani R, Walther G, Hastie T**:** Estimating the number of clusters in a datasets via the Gap statistic. *Royal Statistical Society: Series B (Statistical Methodology)* 2001.