GRD Journal for Engineering | Volume 2 | Issue 5 | April 2017
ISSN: 2455-5703

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

Anusha Chaudhary
Student
Department of Computer Science & Engineering
IMS Engineering College, Ghaziabad

Abstract

Data clustering is an approach for automatically finding classes, concepts, or groups of patterns. It also aims at representing large datasets by a few number of prototypes or clusters. It brings simplicity in modelling data and plays an important 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 computational requirements on the clustering techniques. Swarm Intelligence (SI) has 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 allows the particles to be robust to trace the changing environment. Data structure identifying from the large scale data has become a very important in the data mining problems. Cluster analysis identifies groups of similar data items in large datasets which is one of its recent beneficiaries. The increasing complexity and large amounts of data in the data sets that have seen data clustering emerge as a popular focus for the application of optimization based techniques. Different optimization techniques have been applied to investigate the optimal solution for clustering problems. This paper also proposes two new approaches using PSO to cluster data. It is shown how PSO can be used to find the centroids of a user specified number of clusters.

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

I. INTRODUCTION

Particle Swarm Optimization (PSO) was designed and introduced by Russell Eberhart and James Kennedy in 1995 [1].The PSO is a population based search algorithm based on the simulation of the social behavior of birds flock, bees or a school of fishes. PSO originally intends to 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 assigned vector which determines the next movement of the particle and is called the velocity vector. PSO determines how to update the velocity of a particle. Each particle updates its velocity and the best position it has explored so far based on the global best position explored by swarm particle [2]. The PSO process iterated a fixed number of times until a minimum error based on desired performance index is achieved. It has been shown that this simple model deal with difficult optimization problems. The PSO was originally developed for real valued spaces but many problems are defined for discrete valued spaces. 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 i.e 0 or 1. Each particle value can be changed from 1 to 0 or vice versa. In binary PSO, the velocity of a particle defined as the probability that a particle might change its state to 1. PSO algorithm used in the number of engineering applications. Using binary PSO, Wang and Xiang [6] proposed a splitting criterion for code books of tree structured vector quantizes. Using binary PSO, they reduced the computation time. Binary PSO is used to train the structure of a Bayesian network [7]. In 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 definition the velocity of particle and its parameters has the same role as in real version of the PSO, there are also other versions of binary PSO. In [8] authors add birth and mortality to the ordinary PSO. Fuzzy system can also be used to improve the capability of the binary PSO.

II. PARTICLE SWARM OPTIMIZATION (PSO)

The concept of Particle Swarms introduced for simulating social behaviors of human, which has become very popular as an efficient search and optimization technique. PSO does not require any gradient information of the function to be optimized, it uses only primitive mathematical operators. The algorithm maintains a population of particles where each particle represents a potential solution. The aim of PSO is to find the particle position that results in the best evaluation of a given fitness function. Each particle represents a position in Nd dimensional space, and is “flown" through this multi-dimensional search space, adjusting its position toward both the particle best position found thus far and the best position in the neighborhood. Particles velocity is accelerated toward its previous best position and towards a neighborhood for best solution. Equation (1) and (2) are velocity and position equations:

$$V_i(t+1)=W*V_i(t)+c_1r_1(pbest_i(t)-X_i(t))+c_2r_2\,(gbest_i(t)-X_i(t)) \quad (1)$$

$$X_i(t+1)=X_i(t)+V_i(t) \quad (2)$$

Where t is iteration count, $V_i(t)$ is velocity of particle i at time t, $X_i(t)$ is position of particle *i* at time *t*, *W* is Inertia weight, $pbest_i(t)$ is the best position found by particle itself, $gbest_i(t)$ is the position found by swarm, random values r_1, r_2 in the range of (0,1) particles that explore wide search space and c_1, c_2 are positive acceleration constant and control the weight balance of $pbest_i(t)$ and $gbest_i(t)$.

Equation (1) is used for recording its current position X_i, and velocity V_i indicates speed along dimensions in a problem space. The best fitness values are updated at each generation, according to the equation (3),

$$P_i(t) \begin{cases} P_i(t) & f(X_i(t+1)) \leqslant f(X_i(t)) \\ X_i(t+1) & f(X_i(t+1)) > f(X_i(t)) \end{cases} \quad (3)$$

Where, f indicates the fitness function, $P_i(t)$ indicates best fitness values and *t* indicates iteration count. As clustering problem is an optimization problem that locates optimal centroids of centers. This gives an opportunity to apply particle swarm optimization (PSO) algorithm for clustering problems. PSO clustering algorithm performs a globalized search in the solution space, equation (4) - (7) represent the process, from [2],[6], for data clustering process, single particle represents the N_c cluster centroid vectors. So, each particle x_i as follows:

$$X_i = \{m_{i1}, m_{ij}\ , \ldots, m_{iNc}\} \quad (4)$$

Where m_{ij} indicates j^{th} cluster centroid vector of i^{th} particle in cluster C_{ij}. Fitness value of particle is measured by using quantization error expression:

$$J_e = \frac{\sum_{j=1}^{N_c} \left[\sum_{\forall z_p \in C_{ij}} d\left(z_p, m_j\right) / \left|C_{ij}\right| \right]}{N_c} \quad (5)$$

Where, $|C_{ij}|$ indicates number of data vectors that belongs to cluster C_{ij}, d indicates Euclidean distance between each data vector to the centroid. This Euclidean distance can be calculate as:

$$d(z_p, m_j) = \sqrt{\sum_{k=1}^{N_d} (z_{pk} - m_{jk})^2} \quad (6)$$

Where, k indicates the dimension, N_d indicates number of parameter of each data vector, N_0 indicates number of cluster centroid to be formed, z_p indicates p^{th} data vector and m_j indicates centroid vector of cluster j. The cluster centroid vectors are recalculated by using following expression:

$$m_j = \frac{1}{n_j} \sum_{\forall z_p \in C_j} z_p \quad (7)$$

Where, n_j indicates number of data vectors in cluster j and C_j indicates subset of data vectors from cluster j.

III. DATA MINING AND DATA CLUSTERING

Data mining is a technology which aims at the extraction of hidden information from large databases. Data mining tools predict future trends and behavior to make businesses proactive. The process of knowledge discovery from databases fast and automatic clustering of very large datasets. A family of nature inspired algorithms, known as *Swarm Intelligence* (SI), has several researchers from the field of pattern recognition and clustering. Clustering techniques based on the SI tools that have reportedly outperformed many classical methods of partitioning a complex real world dataset.

Swarm Intelligence is a relatively new field of research which gained huge popularity. Algorithms belonging to the domain draw inspiration from the collective intelligence emerging from the behavior of a group of insects (like termites, bees and wasps). These insects with very limited individual capability can jointly perform many complex tasks for their survival. Problems like finding and storing foods, selecting and picking up materials for future usage need a planning, and are solved by insect colonies. PSO has also attracted the attention of several researchers all over the world resulting into a huge number of variants of the basic algorithm.

Data mining has been called exploratory data analysis. In which, data generated from cash registers, scanning, topic specific databases throughout the company, are explored, analyzed, reduced, and reused. 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 many research communities to describe methods for grouping of unlabeled data. These communities have different terminologies and assumptions for the components of clustering process.

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 centroid (or cluster centre). 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 data set" [19]. Such algorithms have the following advantages [18] the number of clusters are independent of the initial conditions.

Hierarchical clustering techniques suffer from the following drawbacks:

1) They are static as data points assigned to a cluster cannot move to another cluster.
2) They may also fail to separate overlapping clusters due to a lack of information about the size of the clusters.

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 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 or the number of clusters. Most clustering algorithms require the number of clusters to be specified. Finding the "optimum" number of clusters in a data set is usually a challenge it requires a priori knowledge and 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 [20, 21].This paper also uses a new approach called Fuzzy C means Clustering (FCM) using a Particle Swarm Optimization algorithm.

IV. 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, 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. Fuzzy clustering allows each feature vector to belong to more than one cluster with different membership degrees between 0 and 1.

FCM is a pretty standard least squared errors model that generalizes an earlier and very popular non-fuzzy (or hard, which means not "soft") c-means model that produces hard clusters in the data. FCM itself can be generalized in 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, etc. this distance 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 instead of relational values 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.
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,
$u_{ij}$is the degree of membership of $x_i$in the cluster *j*,
$x_i$is the *i*th of d-dimensional measured data,
$c_j$is the d-dimension center of the cluster,
and ||*|| expressing the similarity between any measured data and the center.
By 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.
There are two ways which is described below:

1) Using an algorithm to determine all of the centroids.
2) Run FCM several times each starting with different centroids.

Fuzzy *c*-means the centroid of a cluster is computed as the mean of all points, weighted by their degree to the cluster.

V. CONCLUSIONS

The PSO is an efficient global optimizer for continuous variable problems. The advantages of the PSO are very few parameters to deal with the large number of processing elements. 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 estimation of the local distribution, which efficiently avoids pre-assuming the cluster number. Then the clusters that come from a similar distribution are merged by this clustering program which was applied to both artificial and benchmark data classification and its performance is proved better than the k-means algorithm. The proposed approach is designed to overcome the major problems associated with the existing technique i.e. hybrid subtractive and PSO clustering algorithm and the other methodologies.

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. Engel Brecht. (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] Mahamed G.H. Omran, Andries P Engelbrecht, and Ayed Salman Dynamic Clustering using Particle Swarm Optimization with Application in Unsupervised Image Classification PWASET Volume 9 November 2005 ISSN 1307-6884.