A Review of Data Clustering Techniques and Enhancement of Data Clustering using Hybrid Clustering Model of K-Means and PSO Clustering

Vinod Sharma, Nitish Salwan, Sandeep Singh, Navneet Singh Babra & Prabhsimran Singh

E-mail : vinod.daviet@gmail.com, salvishu5050@gmail.com, sandeep.sandhu50@yahoo.com, nsbabra@hotmail.com, prabh_singh32@yahoo.com

***Abstract* – Clustering is the classification of patterns (observations, data items, or features) into groups (clusters). Cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity The patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster. Clustering is useful in several exploratory pattern-analysis, grouping, decision-making, and machine-learning situations, including data mining, document retrieval, image segmentation, and pattern classification. In This paper we will presents an overview of pattern clustering methods, with a goal of providing useful information and references to fundamental concepts accessible to the broad community of clustering researchers.**

Keywords – Data clustering, Hierarchical clustering, K-Means clustering algorithm, Hybrid PSO algorithm.

I. INTRODUCTION

Data clustering is the process which divides a dataset into some groups or classes. It lets the data objects of the same group have high similarity, and the data objects of different groups have large differences. The similarity is often using the distance between the objects. The data clustering usually has two classes, namely the supervised clustering and the unsupervised clustering. Under the supervised clustering, learning algorithm has an external guidance signal, which offers the class marks for its data vectors. For the unsupervised clustering, there is not an external directive signal, and the algorithm groups the data vectors based on distance from each other.

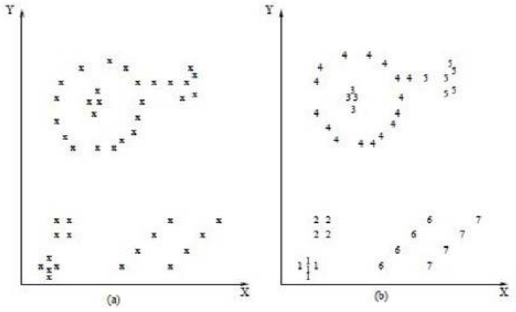

Fig. 1 : Data Clustering

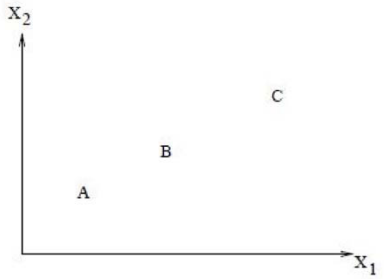

Fig. 2: A and B are More Similar than A and C

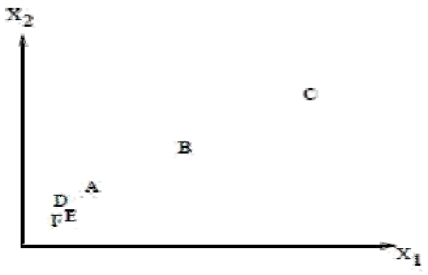

Fig. 3: After a Change B and C ae more similar than A and B

II. COMPONENTS OF A DATA CLUSTERING PROCESS

Data clustering activity involves the following steps.

(1) Pattern representation (optionally including feature extraction and/or selection),

(2) Definition of a pattern proximity measure appropriate to the data domain,

(3) Clustering or grouping,

(4) Data abstraction (if needed),

(5) Assessment of output (if needed).

Pattern representation refers to the number of classes, the number of available patterns, and the number, type, and scale of the features available to the clustering algorithm.

Feature selection is the process of identifying the most effective subset of the original features to use in clustering.

Feature extraction is the use of one or more transformations of the input features to produce new salient features.

Pattern proximity is usually measured by a distance function defined on pairs of patterns.

III. DATA CLUSTERING TECHNIQUES

Different Approaches are being used for the purpose of identifying data sets belonging to their respective clusters.

A. Hierarchical Clustering

A hierarchical data Clustering algorithm yields a *dendrogram* representing the nested grouping of patterns and similarity levels at which groupings change. The operation of a hierarchical clustering algorithm is illustrated using the two-dimensional data set in Figure 4 .This figure depicts seven patterns labeled A, B, C, D, E, F, and G in three clusters.

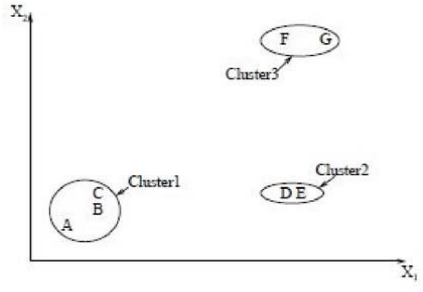

Fig. 4 : Points falling in three clusters

B. Hard Data Clustering vs. *fuzzy Data Clustering*

A hard clustering algorithm allocates each pattern to a single cluster during its operation and in its output. A fuzzy clustering method assigns degrees of membership in several clusters to each input pattern. A fuzzy clustering can be converted to a hard clustering by assigning each pattern to the cluster with the largest measure of membership.

B. (a) Hard Data Clustering vs. *fuzzy Data Clustering*

(1) Select an initial fuzzy partition of the *N* objects into *K* clusters by selecting the *N* * *K* membership matrix U. An element Uij of this matrix represents the grade of membership of object **xi** in cluster **cj**. Typically, uijE (0,1).

(2) Using U, find the value of a fuzzy criterion function, e.g., a weighted squared error criterion function, associated with the corresponding partition. Reassign patterns to clusters to reduce this criterion function value and recompute U.

(3) Repeat step 2 until entries in U do not change significantly.

C. K-Means Clustering Algorithm

(1) Choose *k cluster* centers to coincide with *k* randomly-chosen patterns or *k* randomly defined points inside the hyper volume containing the pattern set.

(2) Assign each pattern to the closest cluster center.

(3) Recompute the cluster centers using the current cluster memberships.

(4) If a convergence criterion is not met, go to step 2. Typical convergence criteria are: no (or minimal) reassignment of patterns to new cluster centers, orminimal decrease in squared error.

Several Versions of K – Means are available and some of them try to find good initial partition so that it is able to find minimum global value.

D. K-Means Clustering Algorithm

PSO algorithm originated from the study of social behaviors of bird flock. Researchers found that birds always changed direction, dispersed and clustered all of a sudden. Their behaviors were unpredictable, but even then the whole always kept the consistency and each of them preserved the optimum distance as well from each other. Through studying behaviors of the similar biotic population, found that there is a social information sharing mechanism in the biotic-population. A Particle Swarm Algorithm maintains apopulation of a certain number of particles, and every particle stands for a

potential solution of the problems. The particles are flying in an n-dimensional space and their position adjustment depending on the experience of themselves and their neighbors.

The clustering process terminates when one of the following conditions is satisfied:

1) The number of iterations exceeds a predefined maximum.
2) When change in the cluster centroids is negligible.
3) When there is no cluster membership change.

Algorithm

1. Initialize each particle to contain Nc randomly selected cluster centroids.
2. For t = 1 to t_{max}

   do

   i. For each particle i do

   ii. For each data vector Z_p

   a. Calculate the Euclidean distance $d(Z_p,M_{ij})$ to all cluster centroids C_{ij}

   b. Assign Z_p to Cluster C_{ij} .

   c. Calculate the fitness using equation (3)

   iii. Update the global best and loc.

E. Hybrid Model for K-Means and PSO Clustering

The convergence rate of K-Means algorithm is faster than the PSO algorithm, but the former usually is not accurate clustering. In order to improve the capability of the PSO algorithm, using the result of the K-means algorithm as an initialized particle, so that it can improve the convergence rate of the PSO algorithm. The process of this mixed clustering

Hybrid PSO algorithm (KPSO) can be described as following:

1) Execute K-means algorithm, and assign the *K* cluster centric vectors from the K-means algorithm to a particle of the particle swarm, then initialize the other particles of the particle swarm randomly according to the norm of data vectors;
2) Execute the PSO algorithm as presented above.

IV. CONCLUSION

In this paper, a method based on combination of theparticle swarm optimization (PSO) and the k-mean algorithm. The combined method has the advantage of both PSO and k-means methods while does not inherent their drawbacks. As the PSO algorithm successfully searches all space during the initial stages of a global search we used PSO algorithm at earlier stage of PSO-KM. As long as the particles in swarm being close to the global optimum, the algorithm switches to k-means as it can converge faster than PSO algorithm. It was detected the proper stage for switching from PSO to k-means using the fitness function. The result of experiment on five datasets including real and synthetic data showed the hybrid algorithm outperforms K-means and PSO clustering.

V. REFERENCES

[1] Hartigan, J.A.; Wong, M.A.; "A K-Means Clustering Algorithm" Journal of the Royal Statistical Society. Series C (AppliedStatistics), vol. 28, no. 1, pp 100-108, 1979

[2] Nagy, G.; "Feature Extraction on Binary Patterns" IEEE Trans. Systems Science and Cybernetics, vol. 5, issue 4, pp 273 – 278, 1969

[3] Coleman, G.B.; Andrews, H.C.; "Image segmentation by clustering" Proceedings of the IEEE,vol. 67, issue: 5, pp 773 –785, 1979

[4] Bezdek, J.C.; Dunn, J.C.; "Optimal Fuzzy Partitions: A Heuristic for Estimating the Parameters in a Mixture of Normal Distributions" IEEE Trans. Computers , vol. C-24, issue 8, pp 835 – 838, 1975

[5] Bezdek, James C.; "A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms" IEEE Trans. Pattern Analysis and Machine Intelligence, vol. PAMI-2, issue 1, pp 1 – 8, 1980

[6] Eberhart, R.; Kennedy, J.; "A new optimizer using particle swarm theory", Sixth Int. Symp. Micro Machine and Human Science, pp 39 – 44, 1995

[7] Backer, Eric; Jain, Anil K.; **"**A Clustering Performance Measure Based on Fuzzy Set Decomposition", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. PAMI-3, issue 1, pp 66 – 75, 1981

[8] Selim, Shokri Z.; Ismail, M. A.; "K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality" IEEE Trans. Pattern Analysis and Machine Intelligence, vol. PAMI-6, issue 1, pp 81 – 87, 1984

[9] Gath, I.; Geva, A.B.; **"**Unsupervised optimal fuzzy clustering**"** IEEE Trans. Pattern Analysis

and Machine Intelligence, vol. 11, issue 7, pp 773 – 780, 1989

[10] Xie, X.L.; Beni, G.; **"**A validity measure for fuzzy clustering**"** IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 13, issue: 8, pp 841 – 847, 1991

[11] Krishnapuram, R.; Keller, J.M.; "A possibilistic approach to clustering", IEEE Trans. Fuzzy Systems, vol. 1, issue 2, pp 98 – 110, 1993

[12] Hathaway, R.J.; Bezdek, J.C.; "Optimization of clustering criteria by reformulation", IEEE Trans. Fuzzy Systems, vol. 3, issue 2, pp 241 – 245, 1995

[13] Pal, N.R.; Bezdek, J.C.; "On cluster validity for the fuzzy c-means model" IEEE Trans. Fuzzy Systems, vol. 3, issue 3, pp 370 – 379

[14] Pedrycz, W.; Waletzky, J.; " Fuzzy clustering with partial supervision" IEEE Trans. Systems, Man, and Cybernetics, vol. 27, issue 5, pp 787 – 795, 1997

[15] Kwon, S.H.; "Cluster validity index for fuzzy clustering" Electronics Letters ,vol. 34, issue 22, pp 2176 – 2177, 1998

[16] Mu-Chun Su; Chien-Hsing Chou; "A modified version of the K-means algorithm with a distance based on cluster symmetry" IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, issue 6, pp 674 – 680, 2001

[17] van der Merwe, D.W.; Engelbrecht, A.P.; "Data clustering using particle swarm optimization" Conference on Evolutionary Computation , vol. 1, pp 215 – 220, 2003

[18] Ahmadi, A.; Karray, F.; Kamel, M.; "Multiple Cooperating Swarms for Data Clustering" IEEE Swarm Intelligence Symposium, pp 206 – 212, 2007

[19] Ahmadyfard, A.; Modares, H.; "Combining PSO and k-means to enhance data clustering" Int. Symp. Telecom., pp 688 – 691, 2008

[20] Minghui Zhou; Junnian Wang; "A Clustering Based Niching Particle Swarm Optimization for Locating Multiple Optimal Solutions" Second Int. Conf. Intelligent Computation Technology and Automation, vol. 1, pp 211 – 214, 2009

[21] Li Shi-Wei; Qian Xiao-Dong; "Date clustering using Principal Component Analysis and Particle Swarm Optimization" 5th Int. Conf. Comp. Sc. and Education, pp 493 – 497, 2010