IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 9, Issue 5 (Mar. - Apr. 2013), PP 65-68
www.iosrjournals.org

Automatic Clustering Using Improved Harmony Search

Ajay Kumar Beesetti[1] , Dr.Rajyalakshmi Valluri[2], K.Subrahmanyam[3], D.R.S. Bindu[4]

[1]*(Department of Electronics & Communications Engineering, ANITS, Visakhapatnam, India)*
[2]*(Department of Electronics & Communications Engineering, ANITS, Visakhapatnam, India)*
[3]*(Systems Engineer, Infosys Limited, India)*
[4]*(Systems Engineer, Infosys Limited, India)*

Abstract: *The paper presents automatic clustering using Harmony Search based clustering algorithm. In this algorithm, the capability of Improved Harmony search is used to automatically evolve the appropriate number of clusters as well as the locations of cluster centers. By incorporating the concept of variable length in each harmony vector, our strategy is able to encode variable number of candidate cluster centers at each iteration. The CH cluster validity index is used as an objective function to validate the clustering result obtained from each harmony memory vector. The proposed approach has been applied onto well-known datasets and experimental results show that the approach is able to find the appropriate number of clusters and locations of cluster centers.*

Keywords *– Automatic Clustering, Harmony Search, Harmony Memory Vector, Cluster Centers*

I. Introduction

Clustering is the task of partitioning the given unlabeled dataset into a number of groups such that objects in the same group are similar to each other and dissimilar to the objects in the other groups. The dissimilarities are assessed based on the attribute Values representing the objects. Although, classification seems to be a good tool for distinguishing or grouping of data yet it requires a large collection of labeled data basing on which the classifier model is developed. Additional advantage of clustering techniques is that it is adaptable to changes and helps in finding useful feature that distinguishes different groups. Cluster analysis has been widely applied in various research areas such as market research, pattern recognition, data analysis and image processing, data mining, statistics, machine learning, spatial database technology, biology. It may serve as a preprocessing step for other algorithms such as attribute subset selection and classification, which would operate on the resulting clustered data and the selected attributes. Clustering methods have been classified as partitioning methods, Hierarchical methods and Density-based methods grid-based methods. A partitioning algorithm organizes the objects into k partitions (K<=n, the no of objects in the dataset), where each partition represents a cluster. The clusters are formed to optimize an objective criterion, so that the objects within a cluster are similar to each other and dissimilar to objects in the other clusters. A hierarchical clustering method works by grouping data objects into a tree of clusters. The problem of partitioned clustering has been developed from a number of fields like the statics[1], graph theory[2], expectation maximization[3],artificial neural networks[4], Evolutionary Computing [5], swarm intelligence[6] and so on. Though there is plethora of papers on the methods to partition the data into clusters there exists a problem of finding the number of clusters in the dataset. Finding the correct number of clusters is too difficult because there are no class labels as is the case in classification task.

II. Harmony Search

Calculus has been used in solving many engineering and scientific problems. These calculus based problem solving techniques can be used only if the objective functions used can be differentiable. These techniques are futile when the objective function is step-wise, discontinuous, multi-modal or when the decision variables are discrete rather than continuous. This is one of the reasons which lead the research community towards the metaheuristic algorithms which have been inspired by biological evolution, animal behavior, or metallic annealing. Harmony search is a music-inspired metaheuristic algorithm. Interestingly, there exists an analogy between music and optimization: each musical instrument corresponds to each decision variable; musical note corresponds to variable value; and harmony corresponds to solution vector. Just like musicians in Jazz improvisation play notes randomly or based on experiences in order to find fantastic harmony, variables in the harmony search algorithm have rand34om values or previously-memorized good values in order to find optimal solution.

Harmony Search (HS) is a heuristic, based on the improvisation process of the jazz musicians [7],[8].Starting with some initially generated harmonies , the musicians will try to improve the harmonies in each iteration. This analogy is used in HS to optimize the given objective function .Like how jazz musicians improvise the new harmonies by improvisation, HS algorithm creates new solutions based on the past solutions and on random modifications. The basic HS algorithm described in the literature is as follows

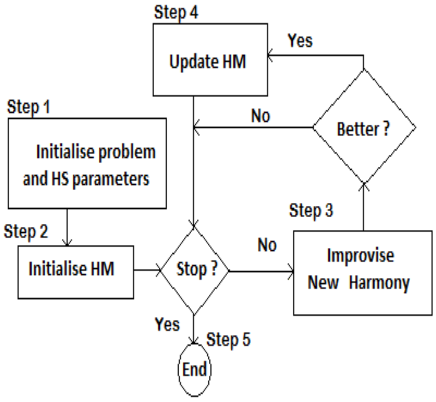

Fig 1: Harmony Search Flowchart

The HS algorithm initializes the Harmonic Memory HM with randomly generated solutions. The number of randomly generated solutions that are stored in the HM is determined by the Harmonic Memory Size (HMS).Once the Harmonic memory is filled with random solutions new solutions are created iteratively as follows. A New Harmony vector is created based on three rules 1) memory consideration 2) pitch adjustment 3) random selection. Generating a new harmony is called improvisation .The parameters that are used in the generation process are Harmony Memory Consideration Rate (HMCR) and Pitch Adjusting Rate (PAR). Each decision variable is set to the value of the corresponding variable of one of the solutions in the HM with a probability of HMCR, and an additional modification of this value is performed with a probability of PAR. Otherwise (with a probability of 1 HMCR), the decision variable is set to a random value. After a new solution has been created, it is evaluated and compared to the worst solution in the HM. If its objective value is better than that of the worst solution, it replaces the worst solution in the HM. This process is repeated, until a termination criterion is met. More detailed description of the algorithm can be found in [7][9].Madhavi et al proposed an improved harmony search (IHS) algorithm that uses variable PAR and bandwidth in improvisation step. In their method PAR and bw change dynamically with generation number as expressed below:

$$\mathrm{PAR(gn)} = \mathrm{PAR}_{\min} + \frac{(\mathrm{PAR}_{\max} - \mathrm{PAR}_{\min})}{\mathrm{NI}} \times \mathrm{gn} \ \ldots.. (1)$$

Where PAR (gn) is the pitch adjusting rate for each generation, $\mathrm{PAR}_{\min}$ is the minimum pitch adjusting rate, $\mathrm{PAR}_{\max}$ is the maximum pitch adjusting rate and gn is the generation number.

$$\mathrm{bw\ (gn)} = \mathrm{bw}_{\max} \exp\ (\mathrm{c.gn}) \ \ldots.(2)$$

$$\mathrm{c} = \frac{\ln\frac{(\mathrm{bw}_{\min})}{(\mathrm{bw}_{\max})}}{\mathrm{NI}} \ldots\ldots\ldots (3)$$

Where bw (gn) is the bandwidth for each generation, $\mathrm{bw}_{\min}$ is the minimum bandwidth and $\mathrm{bw}_{\max}$ is the maximum bandwidth. Recently, other variants of harmony search are also having been proposed.

III. Automatic Clustering Using Harmony Search

Cluster analysis identifies and classifies objects individuals or variables on the basis of the similarity of the characteristics they possess. It seeks to minimize within-group variance and maximize between-group variance. The result of cluster analysis is a number of heterogeneous groups with homogeneous contents: There are substantial differences between the groups, but the individuals within a single group are similar. Data may be thought of as points in a space where the axes correspond to the variables. Cluster analysis divides the space into regions characteristic of groups that it finds in the data. We can show this with a simple graphical example:

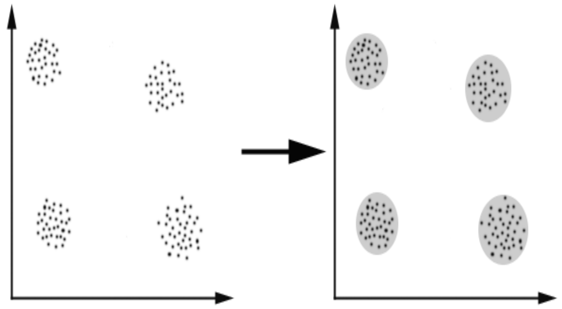

Fig 2: Clustering Example

A. Selecting the objective functions:

Cluster analysis partitions the set of observations into mutually exclusive groupings in order to best represent distinct sets of observations within the sample. Cluster analysis is not able to confirm the validity of these groupings as there are no predefined classes. Several clustering validity measures have been developed. The result of a clustering algorithm can be very different from each other on the same data set as the other input parameters of an algorithm can extremely modify the behavior and execution of the algorithm. The aim of the cluster validity is to find the partitioning that best fits the underlying data. The process of evaluating the quality of a clustering is called the clustering assessment. Two measurement criteria have been proposed for evaluating and selecting an optimal clustering scheme in the literature

- Compactness: The members of each cluster should be as close to each other as possible. A common measure of compactness is the variance.
- Separation: The clusters themselves should be widely separated.

There are three common approaches measuring the distance between two different clusters: distance between the closest member of the clusters, distance between the most distant members and distance between the centers of the clusters. Numerical measures that are applied to judge various aspects of cluster validity are classified into the following three types [10], [11].

1. External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
2. Internal Index: Used to measure the goodness of a clustering structure without respect to external information.
3. Relative Index: Used to compare two different clustering of clusters.

Sometimes these are called as criteria instead of indices. The performance of the automatic clustering algorithm depends upon the clustering objective function that is being optimized. In this work we choose the CH index [12] as the objective function. The idea behind the CH measure is to compute the sum of squared errors(distances) between the k^{th}cluster and the other k - 1 clusters, and compare that to the internal sum of squared errors for the k clusters (taking their individual squared error terms and summing them, a pooled value so to speak). In effect, this is a measure of inter-cluster (dis)similarity over intra-cluster (dis)similarity.

B (k) - between cluster sums of squares

W (k) - within cluster sum of squares

$$CH (k) = [B(k)/(k-1)]/ [W(k)/(N-k)] \quad ....(4)$$

Now, if B(k) and W(k) are both measures of difference/dissimilarity, then a larger CH value indicates a better clustering, since the between cluster difference should be high, and the within cluster difference should below.

B. Search Variable Representation

In this work we have been using a centroid based representation for the search variable. Given a dataset consisting of K instances each having D dimensions the search variable is constructed as follows containing variable number of cluster centers.

$\overline{Z}_{i,G}$ =

$T_{i,1}$	$T_{i,2}$	..	$T_{i,Kmax}$	$\overline{m}_{i,1}$	$\overline{m}_{I,2}$	.	$\overline{m_{i,Kmax}}$

Table 1: Dataset consisting of K instances, showing variable number of cluster centers where, Zi, G is the ith search variable in Gth iteration

$m_{i,j}$ is the jth cluster center

$T_{i,j}$ is a Threshold value
If $T_{i,j}$<0.5 then the corresponding cluster center mi,j is not active in the search variable
If $T_{i,j}$>0.5 then the corresponding cluster center mi,j is active in the search variable K_{max} is the user specified number representing the maximum number of clusters.

Dataset	No.of Attributes	Expected No.of Clusters	No. of clusters obtained
Iris	4	3	3
Wine	13	3	3
Bupa	7	2	2
Balance	5	3	3

Table 2: Simulation results for four data sets.

IV. Results

The algorithm is implemented using MATLAB 7.0. AMD Athlonprocessor (2.3GHz) and 1GB RAM. The simulation results showing the effectiveness of the Harmony Search based clustering has been provided for four real life datasets: Iris, Wine, Bupa and balance data

The parameters for the algorithm are set as follows HMCR=0.9, PAR_{max}=0.8 and PAR_{min}=0.4 BW_{min}=0.01 and BW_{max}=0.1. The average (corrected to decimal) of the output of 15 individual runs of the algorithm have been reported as the number of clusters obtained.

V. Conclusion

In this paper, the problem of automatic clustering has been investigated and presented a Harmony Search based strategy for crisp clustering of real world datasets. An important feature of this clustering algorithm is that it is able to find the optimal number of clusters automatically, that is the number of clusters does not have to be known in advance. Our experimental results show that Harmony Search can be a candidate optimization technique for clustering. Future research may extend the single objective automatic clustering using Harmony Search to handle multi-objective-optimization.

References

[1] E.W. Forgy, *Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of classification, Biometrics, 21, (1965) 768-9.*

[2] C. T. Zahn, *Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on Computers C-20 (1971), 68–86.*

[3] M.H. Law, M.A.T. Figueiredo, and A.K. Jain, Simultaneous Feature Selection and Clustering Using Mixture Models, *IEEE Transactions Pattern Analysis and Machine Intelligence, vol. 26, no. 9, pp. 1154-1166, Sept.2004.*

[4] N. R. Pal, J. C. Bezdek, E. C.-K. Tsao, *Generalized clustering networks and Kohonen's self-organizing scheme, IEEE Trans. Neural Networks, vol 4, (1993) 549–557.*

[5] S. Das, A. Abraham, and A. Konar, Metaheuristic *Clustering, Studies in Computational Intelligence, SpringerVerlag, Germany, 2009.*

[6] A. Abraham, S. Das, and S. Roy, *Swarm Intelligence Algorithms for Data Clustering, Soft Computing for Knowledge Discovery and Data Mining, O. Maimon and L. Rokach (*Eds.), *Springer Verlag, Germany, pp. 279-313,2007.*

[7] Z.W. Geem, J.H. Kim, and G.V. Loganathan, *"A new heuristic optimization algorithm: harmony search", Simulation, vol. 76, 2001,pp. 60-68.*

[8] Z.W. Geem, M. Fesanghary, J. Choi, M.P.Saka, J.C. Williams, M.T. Ayvaz, L. Li, S.Ryu, and A. Vasebi, *"Recent advances in harmony search", in Advance in Evolutionary Algorithms, WitoldKosiński, Ed. Vienna : ITeachEducation and Publishing, 2008, pp.127-142*

[9] M. Mahdavi, M. Fesanghary, and E. Damangir, *"An improved harmony search algorithm for solving optimization problems", Applied Mathematics and Computation. vol. 188, 2008, pp. 1567-1579.*

[10] M. Halkidi, Y. Batistakis and M. Vazirgiannis, *Cluster validity methods:part I, SIGMOD Rec., Vol. 31, No. 2, pp. 40-45, 2002*

[11] M. Halkidi, Y. Batistakis and M. Vazirgiannis, *Cluster validity methods:part II, SIGMOD Rec., Vol. 31, No. 3, pp. 19-27, 2002*

[12] Calinski, T., Harabasz, *J: A dendrite method for cluster analysis. Communications in Statistic 3 (1974) pp127*