International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol.3, Issue.1, Jan-Feb. 2013 pp-504-508 ISSN: 2249-6645

A PSO Optimized Layered Approach for Parametric Clustering on Weather Dataset

Shikha Verma, [1] Kiran Jyoti[2]
[1]*Student, Guru Nanak Dev Engineering College Ludhiana, Punjab*
[2]*Asstt. Prof, Guru Nanak Dev Engineering College Ludhiana, Punjab*

Abstract: *Clustering is the process to present the data in an effective and organized way. There are number of existing clustering approaches but most of them suffer with problem of data distribution. If the distribution is non linear it gives impurities in clustering process. The propose work is about to improve the accuracy of such clustering algorithm. Here we are presenting the work on time series oriented database to present the work. Here we are presenting the three layer architecture, where first layer perform the partitioning based on time series and second layer will perform the basic clustering. The PSO is finally implemented to remove the impurities from the dataset.*

Keywords: *Clustering, CMeans, KMeans, Dataset, Feature*

I. INTRODUCTION

Clustering is a fundamental problem in machine learning and has been approached in many ways. Clustering can be considered the most important unsupervised learning technique; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data. There are different clustering algorithms .The k-centers clustering algorithm is one of the popular algorithm. Clustering is used because following some reasons:

- Simplifications
- Pattern detection
- Useful in data concept construction
- Unsupervised learning process

A new clustering approach called Affinity Propagation was introduced by Brenden J. Frey and Delbert Dueck in 2007. Affinity Propagation has many advantages over other clustering algorithms like simplicity, general applicability, lower errors and good performance. Affinity Propagation takes as input a collection of real valued similarities between data points. It is distance based clustering algorithm and similarity measure are calculated in the form of Euclidean distance. Real valued messages are exchanged between data points until a high quality set of exemplars and corresponding clusters gradually emerges.

Each data point i sends a responsibility r (i, k) to candidate exemplar indicating how much it favors the candidate exemplar over other candidates .Availability a(i,k) is sent from candidate exemplar to data points indicating how well the candidate exemplar can act as a cluster center for the data points. At the end of each iteration, availability and responsibility are combined to identify the exemplar.

$$r(i,k)=s(i,k)-\max\{\sum a(i,k')+s(i.k')\}$$

where s(i,k) is the similarity between data point i and candidate exemplar k ,a(i,k') is availability of other candidate exemplar k' to data point i , s(i,k') is the similarity between data point i and other candidate exemplar k.

$$a(i,k)= \min\{0,r(k,k)+\sum\max\{0,r(i',k)\}\}$$

where a(i,k) is the availability of candidate exemplar k to data point i ,r(k,k) is self responsibility of candidate exemplar k , r(i',k) is the responsibility from other data point I' to the sane candidate exemplar k. and other candidate exemplar k

$$a(k,k) =\sum\max\{0.r(i',k)\}$$

Where a (k, k) is self availability of candidate exemplar k and r(i',k) is the responsibility from other data point I' to the sane candidate exemplar k.

Clustering is "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters

A. Clustering as unsupervised classification

As for Classification, we have a multivariate dataset; a set of cases, each of which has multiple attributes/variables. Each case is represented as a record, (or row), and to each attribute variable there is an associated column. The whole comprises the caseset. The attributes may be of any type; nominal, ordinal, continuous, etc…, but the in Cluster Analysis none of these attributes is taken to be a classification variable. The objective is to obtain a rule which puts all the cases into groups, or clusters. it may then define a nominal classification variable which assigns a distinct label to each cluster. We do not know the final classification variable before we do the clustering; the objective is to define such a variable.

Hence, clustering is said to be an unsupervised classification method; unsupervised, because we do not know the classes of any of the cases before we do the clustering. We do not have class values to guide (or "supervise") the training of our algorithm to obtain classification rules.

We should note that one or more of the attributes in our case set may in fact be nominal classification variables. However, as far as our Clustering of the case set is concerned they are just nominal variables and are not treated as classification variables for the clustering analysis.

B. Distance Matrices

If we have two cases, C1 and C2 say, which have continuous variables x and y, taking values (x_1, y_1) and (x_2, y_2) respectively, then we can plot the two cases in x-y space as in Figure 1.

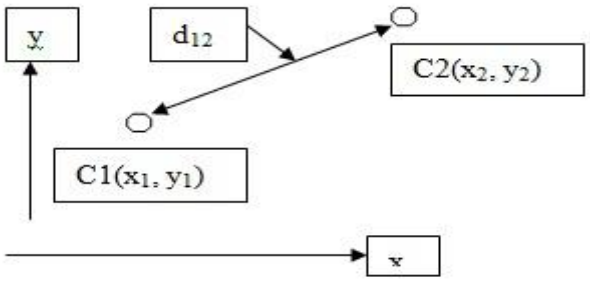

Figure 1: Euclidean distance

Using Pythagorus's Theorem, we may write

$$d_{12} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

This is the Euclidean distance between the two cases, "in the x-y state-space". We often say that we have defined a "distance metric" when we have defined the formula for the distance between two cases.

If the two cases are characterized by p continuous variables ($x_1, x_2, \ldots x_i, \ldots x_p$ say) rather than two (i.e. x, y), [Note: $x \rightarrow x_1$ so x(case 2) $\rightarrow x_1$(case 2) $\equiv x_1(2)$; similarly $y_2 \rightarrow x_2(2)$], then we may generalize the Euclidean distance to:

$$d_{12} = \sqrt{\sum_{i=1}^{p} (x_i(2) - x_i(1))^2}$$

This can be generalized further to the Minkowski metric:

$$d_{12} = \sqrt[m]{\sum_{i=1}^{p} |(x_i(2) - x_i(1))|^m}$$

where |x| denotes the absolute value of x, (i.e. the size, without the sign).

C. Standardization

If the values of the variables are in different units then it is likely that some of the variables will take vary large values, and hence the "distance" between two cases, on this variable, can be a big number. Other variables may be small in value, or not vary much between cases, in which case the difference in this variable between the two cases will be small. Hence in the distance metrics considered above, are dependent on the choice of units for the variables involved. The variables with high variability will dominate the metric. We can overcome this by standardizing the variables.

D. Similarity Measure

A measure of the similarity (or closeness) between two cases must take its highest value when the two cases have identical values of all variables, (i.e. when the two cases are coincident in multivariable space). The similarity measure must decrease monotonically as the case variable values increase, i.e. as the distance between the cases increases. Hence, any monotonically decreasing function of distance will be a possible similarity measure. If we want the similarity measure to take value 1 when the cases are coincident, then we might consider:

D. Particle Swarm Optimization

PSO was first proposed by Kennedy and Eberhart [9]. The main principle behind this optimisation method is communication. In PSO there is a group of particles that look for the best solution within the search area. If a particle finds a better value for the objective function, the particle will communicate this result to the rest of the particles. All the particles in the PSO have "memory" and they modify these memorized values as the optimisation routine advances. The recorded values are: velocity (*V*), position (*p*), best previous performance (*pbest*) and best group performance (*gbest*). The velocity describes how fast a particle should move from its current position which contains the coordinates of the particle. The last two parameters are the recorded best values that have been found during the iterations. A simple PSO algorithm is expressed as [9]:

II. LITERATURE SURVEY

Lei Jiang defined a work on "Hybrid Adaptive Niche to Improve Particle Swarm Optimization Clustering Algorithm",. This paper use adaptive niche particle swarm algorithm to improve clustering And it is also studied the impact of different fitness optimization function to clustering data [1]. Shuai Li, Xin-Jun Wang, Ying Zhang defined a work on "X-SPA: Spatial Characteristic PSO Clustering Algorithm with Efficient Estimation of the Number of Cluster". In this paper author unifies Particle Swarm optimization (PSO) algorithm and Bayesian Information Criterion (BIC), proposes a numeric clustering algorithm. Chaos and space characteristic ideas are involved in the algorithm to avoid local optimal problem. Furthermore, BIC is also involved to provide an efficient estimation of the number of cluster [2].

Rehab F. Abdel-Kader presented a work on "Genetically Improved PSO Algorithm for Efficient Data Clustering. The proposed algorithm combines the ability of the globalized searching of the evolutionary algorithms and the fast convergence of the k-means algorithm and can avoid the drawback of both. The performance of the proposed algorithm is evaluated through several benchmark datasets. The experimental results show that the proposed algorithm is highly forceful and outperforms the previous approaches such as SA, ACO, PSO and k-means for the partitioned clustering problem [3]. Surat Srinoy presented a work on "Combination Artificial Ant Clustering and K-PSO Clustering Approach to Network Security Model", His paper presents natural based data mining algorithm approach to data clustering. Artificial ant clustering algorithm is used to initially create raw clusters and then these clusters are refined using k-mean particle swarm optimization (KPSO). KPSO that has been developed as evolutionary-based clustering technique [4].

Shafiq Alam has defined a work on "Particle Swarm Optimization Based Hierarchical Agglomerative Clustering The proposed algorithm exploits the swarm intelligence of cooperating agents in a decentralized environment. The experimental results were compared with benchmark clustering techniques, which include K-means, PSO clustering, Hierarchical Agglomerative clustering (HAC) and Density-Based Spatial

Clustering of Applications with Noise (DBSCAN). The results are evidence of the effectiveness of Swarm based clustering and the capability to perform clustering in a hierarchical agglomerative manner [5]. Alireza Ahmadyfard presented a novel work on "Combining PSO and k-means to Enhance Data Clustering". In this paper we propose a clustering method based on combination of the particle swarm optimization (PSO) and the k-mean algorithm. PSO algorithm was showed to successfully converge during the initial stages of a global search, but around global optimum, the search process will become very slow. On the contrary, k-means algorithm can achieve faster convergence to optimum solution [6].

III. PROPOSED APPROACH

The proposed work is improvement of the data clustering algorithm with effect of Particle Swarm Optimization. Here the improvement is expected in terms of accuracy and the efficiency. The proposed work will be affected with 3 dimensions of improvement in data clustering. As the first dimension the data will be segmented according to the time series. Such as in case of temperature the time series will be according to the seasonal parameters. Once the data is segmented, The clustering algorithm will be implemented on it.

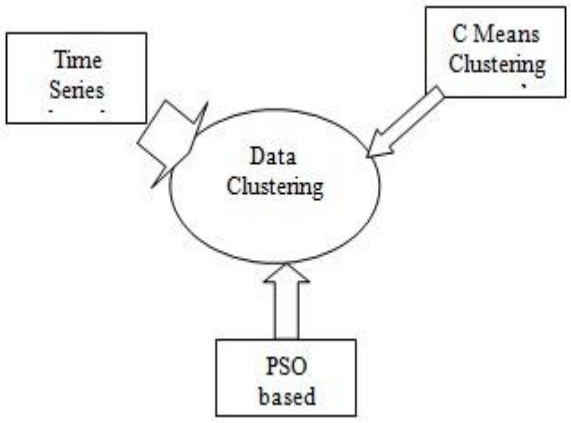

Figure 2: Three Dimensional Clustering

The clustering algorithm will perform the actual data categorization according to the data values. Here either C-Means or the K-Means algorithm will be implemented based on the dataset. This work will be done only once. The clustering algorithm is the seed of the proposed algorithm.
Once the clustering done the refinement over the dataset is performed by the PSO. The PSO will use the output of the clustering algorithm and generate a Hybrid algorithm. The work of PSO is first to find the best fitness function and then pass this clustered dataset on this fitness function. The fitness function will be generated respective to the cluster as well as the whole dataset. The fitness function will accept a particular data item and verify that whether it should belong to the cluster or not after examine the cluster data. To implement the PSO some parameters are required. These parameters will be decided according to the dataset and the deficiencies in the clustered dataset.

The basic algorithm for the PSO that will be implemented is given as under

1: Initialize a population array of particles with random positions and velocities on D- dimensions in the search space.
2: For each particle, evaluate the desired optimization fitness function in *D* variables.
3: Compare particle's fitness evaluation with its $pbest_i$. If current value is better than $pbest_i$, then set $pbest_i$ equal to the current value, and p_i equal to the current location x_i in D- dimensional space.
4: Identify the particle in the neighborhood with the best success so far, and assign its index to the variable g.
5: Change the velocity and position of the particle according to the equation (3)
6: If a criterion is met (usually a sufficiently good fitness or a maximum number of iterations), exit.
7: If criteria are not met, go to step 2

The presented work is implemented for the weather dataset under the different attributes. These attributes includes the temperature, rainfall, humidity etc. The process on this dataset is performed under a three layered approach. The algorithmic process is given as under

Table 1: Proposed Algorithm

Step 1:	Accept Whole Dataset As Input
Step 2.	Arrange the Dataset According to Years
Step 3:	Find the Instance Count for Each Year
Step 4:	Find the ThresholdCount Value for whole Dataset
Step 5:	Remove the Year Dataset Having Count(Year)<ThresholdCount
Step 6:	Represent the Remaing as Training Dataset
Step 7:	Select N-points as initial centroids
Step 8:	Repeat
a.	Form N-clusters by assigning each point to its closest centroid
b.	Recompute the centroid of each cluster
Step 9:	Until centroid does not change
Step10:	Run PSO on initial clusters generated by Clustering Algorithms
a.	Initialize the Particles (Clusters)
b.	Initialize Population size and maximum iterations
c.	Initialize clusters to input data
d.	Evaluate fitness value and accordingly find personal best and global best position
Step 11:	Start with an initial set of particles, typically randomly distributed throughout the design space.
Step 12:	Calculate a velocity vector for each particle in the swarm.
Step 13:	Update the position of each particle, using its previous position and the updated velocity vector.
Step 14:	Repeat From Step 10 and Exit on reaching stopping criteria (maximum number of iterations).

IV. RESULTS

The presented clustering approach is implemented on weather dataset in mat lab 7.8 environment. The initial dataset contains about 10000 instances and having 7 attributes. The dataset is containing the data between years 1970 and 2010. As the first layer the time series segmentation is performed and the data values of specific years are pruned as these are found less effective. The dataset remained after the pruning process is shown in table 2 with description of specific years.

Table 2: Pruned Dataset

1971, 1972, 1974, 1976, 1977, 1978, 1979, 1981, 1984, 1986, 1987, 1988, 1991, 1993, 1994, 1996, 1998, 2000, 2001, 2003, 2005, 2009

Initial data contains 10000 instances and after the pruning process about 6700 records left. To perform the pruning, the time series segmentation is been implemented in this work. After the pruning process, the C-Menas clustering is performed over the dataset. In this work, we have divided the complete dataset in 4 clusters under the different attributes. Figure 3, 4, 5 are showing the outcome of 1st cluster under three attributes respectively i.e. temperature, rain and humidty.

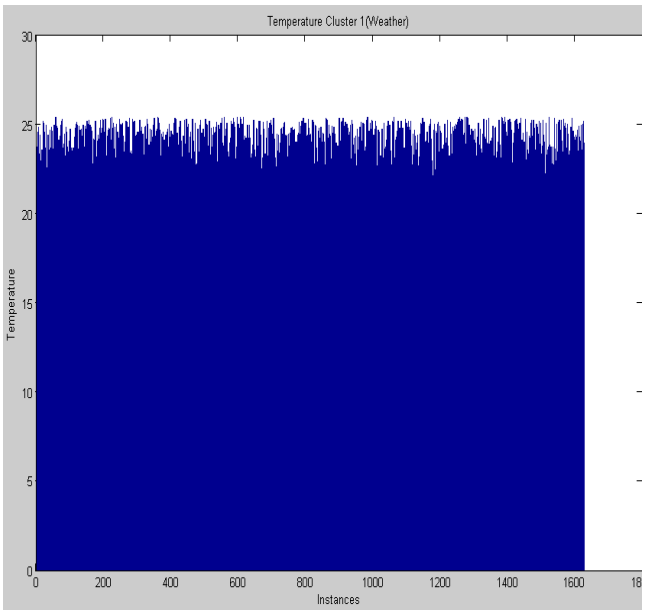

Figure 3: Cluster 1(Temperature)

As we can see figure 3 is showing the outcome of clustering on temperature dataset. As we can see in this figure, the data obtained in cluster 1 is shown. This particular cluster is having about 1600 instances and contains the temperature values between 22 to 25.

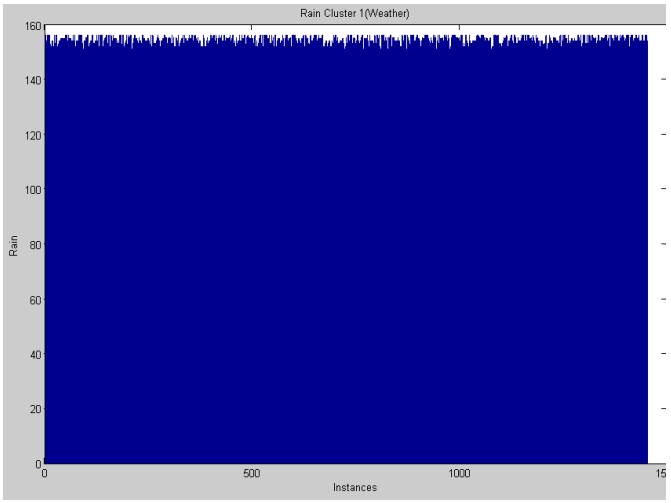

Figure 4: Cluster 1(Rain)

Figure 4 is showing the outcome of clustering on Rain dataset. As we can see in this figure, the data obtained in cluster 1 is shown. This particular cluster is having about 1400 instances and contains the temperature values between 150 to 160. Out of 6700 instances, only 1400 instances are having that much of rainfall.

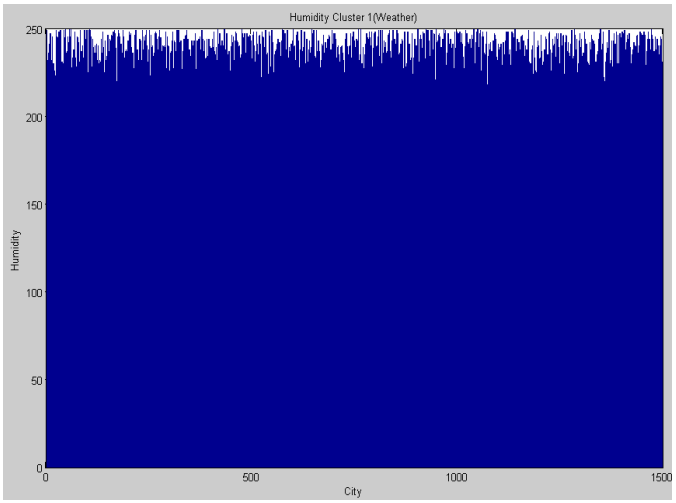

Figure 5: Cluster 1(Humidity)

As we can see figure 5 is showing the outcome of clustering on Humidity dataset. As we can see in this figure, the data obtained in cluster 1 is shown. This particular cluster is having about 1500 instances and contains the temperature values between 220 to 250.

V. CONCLUSION

The presented work is about to improve the clustering process by implementing a three layered approach. The first layer is the filtration layer to identify the most appropriate dataset on which clustering will be performed. On the second layer, the actual clustering process is performed. Finally, the PSO is used to adjust the boundary values and to optimize the outcome of clustering process.

REFERENCES

[1] Jiang Lixin Ding, "Hybrid Adaptive Niche to Improve Particle Swarm Optimization Clustering Algorithm", 978-1-4244-9953-3/11 ©2011 IEEE

[2] Shuai Li, Xin-Jun Wang, "X-SPA: Spatial Characteristic PSO Clustering Algorithm with Efficient Estimation of the Number of Cluster", Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 978-0-7695-3305-6/08 © 2008 IEEE

[3] Rehab F. Abdel-Kader, "Genetically Improved PSO Algorithm for Efficient Data Clustering", 2010 Second International Conference on Machine Learning and Computing, 978-0-7695-3977-5/10 © 2010 IEEE

[4] Surat Srinoy, "Combination Artificial Ant Clustering and K-PSO Clustering Approach to Network Security Model", IEEE 2006

[5] Shafiq Alam, Gillian Dobbie, Patricia Riddle, M. Asif Naeem," Particle Swarm Optimization Based Hierarchical Agglomerative Clustering ",2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

[6] Alireza Ahmadyfard, Hamidreza Modares ," Combining PSO and k-means to Enhance Data Clustering ",2008 Internatioal Symposium on Telecommunications, 978-1-4244-2751-2/08 ©2008 IEEE

[7] [7] Dongyun Wang, "Clustering Research of fabric deformation comfort using bi-swarm PSO algorithm", World Congress on Intelligent Control and Automation July 6-9 2010, Jinan, China

[8] [8] Junyan Chen, "Research on Application of Clustering Algorithm Based on PSO for the Web Usage Pattern", 1-4244-1312-5/07© 2007 IEEE

[9] [9] Jinxin Dong, Minyong Qi, "A New Clustering Algorithm Based on PSO with the Jumping Mechanism of SA", 2009 Third International Symposium on Intelligent Information Technology Application

[10] Brendan J. Frey and Delbert Dueck," Clustering by Passing essages Between Data Points" University of Toronto *Science* ,Vol **315**, 972–976, February 2007.

[11] Bryan Conroy and Yongxin Taylor Xi. "Semi Supervised Clustering Using Affinity Propagation". August 24, 2009.

[12] Supporting Material available at http://www.psi.toronto.edu/affinitypropagation

[13] Delbert Dueck and Brendan J. Frey. "Non-metric affinity propagation for unsupervised image categorization". In IEEE Int. Conf. Computer Vision (ICCV), 2007