Message Passing in Clusters using Fuzzy Density based Clustering

M. Suganya* and S. Nagarajan

School of Computing, SASTRA University, Thanjavur - 613401, Tamil Nadu, India; sugucsc23@gmail.com, Nagarajan@cse.sastra.edu

Abstract

The Communication between objects in the cluster takes place through message passing techniques. Affinity Propagation (AP) clustering has been used successfully in a lot of clustering problems, which deals with static data. By applying Fuzzy DENCLUE, similarity between objects and its exemplar is improved. A dynamic variant of AP clustering called Incremental Affinity Propagation with K-Medoid (IAPKM) is used along with the Fuzzy Density based clustering (DENCLUE) method. In Fuzzy DENCLUE, number of clusters is reduced and randomness in the form of noise is removed. The experiment is performed in Net Beans 8.0.2 using jdk1.7 as a language. The evaluation result defines that the effectiveness and efficiency of IAPKM and Fuzzy DENCLUE can achieve comparable performance. Compared with IAPKM and Fuzzy DENCLUE, Sum of similarity and accuracy is increased in Fuzzy DENCLUE which is based on the parameters. The current approach achieves higher accuracy rate which is 10% greater than the former approach.

Keywords: Exemplar, Fuzzy DENCLUE, Incremental Affinity Propagation based on K-Medoid, Message Passing, Sum of Similarities

1. Introduction

In Data mining, Clustering plays a major role. Grouping similar objects into single class is called a cluster. Various types of clustering algorithms are designed for static data such as K-Medoid, K-neighbor, etc. Nowadays, several data appear in the dynamic manner such as web pages, blogs, etc. Some of the dynamic techniques in data mining are incremental clustering, evolutionary clustering. AP clustering is an exemplar-based method that realized by assigning each data point to its closest exemplar, where exemplars are recognized by passing messages such as responsibility and availability on bipartite graph. DENCLUE stands for density based clustering which clusters the data based on density attractor. Membership function is to represent fuzzy elements that map every element of the universe of discourse X to the interval [0, 1]. Triangular membership function is one of the functions that can be used in fuzzy set.

In1, Clustered framework contains a huge number of data sets along with attributes which surveys the earlier clustering algorithm by theoretical and logical view. In2, clustering the data is important to process the sensory signal and to detect the pattern. Affinity Propagation solves the issue by taking similar data point sets as input and exchanging the real valued message to get a good quality exemplar set. In3, main objective of the data stream clustering model is to cluster the huge amount of data set with limited requirement of memory and time using constant factor approximation algorithm. In4, the Incremental clustering algorithm is used to maintain the dynamic data set clusters efficiently. In5, the problem of
clustering real time data is solved by classical k-means clustering algorithm. A current clustering structure is maintained by analyzing the arriving data in an online mode. In a, Preference value is found out by the Adaptive affinity propagation method and also it defines the space of preference to optimize the clustering result. This method is applied in document clustering to achieve better clustering results.

In b, the intrinsic clustering is performed using Message Passing Clustering (MPC) thereby improving the clustering performance. MPC considers both global and local structure which offers better performance compared to that of hierarchical clustering. In a, Fuzzy algorithm handles large amount of relational dataset as data chunks to process it sequentially, which provides better result in real time environment. In a, the performance of AP algorithm is improved using Semi-supervised clustering technique that includes named exemplars and unnamed exemplars. By using this, similarity matrix can be modified and useless samples can be removed. In a, the mobility of an unseen object is studied by the robot and the environment interface. Images are grouped into clusters with similar features by using an art clustering algorithm.

In a, Introduced Dynamic threshold in DENCLUE determines the number of partitions by combining density attractor. For modifying the parameters of fuzzy model, Back propagation method is used. In a, Pre-known information is described using Incremental Affinity Propagation clustering. In order to determine the usefulness of the method text clustering is described to evaluate the given data set. In a, for noise elimination a well-organized fast face recognition method has been proposed which is a combination of Affinity Propagation (AP) and Linear Discriminate Analysis (LDA). In a, Applying AP in incremental clustering problems is discussed. Two IAP clustering algorithms are projected here, namely IAP clustering based on K-Medoid (IAPKM) and Incremental Affinity Propagation based on Neighbor Assignment (IAPNA). IAPKM dynamically clusters the objects using message passing techniques such as responsibility and availability. IAPNA is similar to the former approach, in addition the newly arriving objects forms the cluster individually. The novel approach uses an IAPKM method for better performance. In Existing system, two phases were done.

In the first phase, Affinity Propagation (AP) for clustering static data with a set of exemplars is formed. AP clustering is an exemplar-based method that realized by assigning each data point to its closest exemplar, where exemplars are recognized by passing messages such as responsibility and availability on bipartite graph. These messages are collectively called as Affinity. The sum of Similarity plays a vital role, in which similarity between objects and its nearest exemplars are calculated.

Two categories of message passing for finding optimal set of exemplar are Responsibility are \( r(o, e) \) and Availability \( a(o, e) \). Responsibility refers to how object \( o \) send a request to select \( e \) as its exemplar.

\[
 r(o, e) \leftarrow s(o, e) - \max_{j \neq i} \{a(o, e_j') + s(o, e_j')\} \quad (1)
\]

Availability refers to how object \( j \) accepts \( i \) as its object set.

\[
 a(o, e) \leftarrow \min_{j \neq i} \{0, r(o, e) + \sum_{j \neq i} \max\{0, r(o, e_j')\}\} \quad (2)
\]

The clustering result after performing responsibility and availability is given by this formula,

\[
 \hat{C}_t = \arg \max_j \{a(o, e_j) + r(o, e_j)\} \quad (3)
\]

AP clustering can be used in several applications such as document clustering, face recognition. Even though
it supports static data, nowadays most applications deal with dynamic data.

In the Second phase, Incremental Affinity Propagation based on the K-Medoid algorithm is used. A dynamic variant of AP clustering called Incremental Affinity Propagation (IAP) is used. The initial batch of objects is clustered by Traditional AP. K-Medoid is implemented for new objects to adjust the clustering result. IAPKM can be used to quickly modify the clustering result which makes enough to be in a dynamic manner. Even though it supports dynamic data it cannot adjust the number of clusters according to new objects. Responsibility for IAPKM can be defined as,

\[ r_{km}(o_i, e_j) = \arg \max_{j \in E} \{ s(o_i, e_j) \} \] (4)

Here, \( E \) is the present exemplar set. Equation (4) defines the similarity between node \( i \) and exemplar \( j \). Availability can be defined as,

\[ a_{km} = \arg \max_{q, \hat{c}_k=\hat{c}_q} \left\{ \sum_{i=1}^{n} f(c) \cdot s(i, c) \right\} \] (5)

Equation (5) defines aggregated sum of similarities, which explains when SS (Sum of Similarity) is maximum then the particular object belongs to cluster \( j \).

3. Proposed System

The proposed system has been designed by combining IAPKM and Fuzzy DENCLUE method in clusters. In this work we are proposing Fuzzy DENCLUE approach. The working has been carried out in two phases.

In the first phase, the Fuzzy DENCLUE method reduces the number of clusters into a single cluster and removes randomness in the form of noise, generated from IAPKM approach. DENCLUE is the density based clustering algorithm in which single cluster is formed to set of exemplars. Based on density attractor, single cluster is formed in which all objects are placed. By applying triangular function in DENCLUE helps to remove noise from the cluster.

3.1 Algorithm 1: Fuzzy DENCLUE

**Input:** \( C_{t-1} \), \( \epsilon \), Minpts;

**Output:** \( X_t \);

**Step 1.** Initialize \( \epsilon \) and \( \text{Minpts} \) value;

**Step 2.** If Neighbor size is greater than \( \text{Minpts} \), then add the unclassified cluster to the neighbor;

**Step 3.** Otherwise, remove the noise from the cluster;

**Step 4.** Apply triangular function in DENCLUE;

**Step 5.** Calculate Predistance and PredistanceMembership function to find out the neighbor;

**Step 6.** Repeat step 5 until the cluster \( X_t \) is formed.

Where, \( C_{t-1} \) is the clustering output of IAPKM and \( X_t \) is the clustering output of Fuzzy DENCLUE. \( \epsilon \) ranges from 0.5 to 1.

In the second phase, effectiveness and efficiency measures are evaluated for the Fuzzy DENCLUE method then compared with IAPKM. Performance measures for the Incremental Affinity Propagation based on K-Medoid and Fuzzy DENCLUE can be experimentally compared. Performance measures such as,

- Sum of Similarities.
- Normalized Mutual Induction.
- Accuracy.
- Number of Iterations.
- Memory Usage.

The sum of Similarities measures the similarity between objects and its nearest exemplar.

\[ SS = \sum_{i=1}^{n} s(i, \hat{c}) \] (6)

Where, \( i \) is the object and \( \hat{c} \) is the nearest exemplar.

Normalized Mutual information computes the mutual information between initial cluster and clustering results.

\[ NMI = \frac{I(C_i, \hat{C}_i)}{\sqrt{H(C_i) \cdot H(\hat{C}_i)}} \] (7)

Where, \( I(C_i, \hat{C}_i) \) represent mutual information between initial cluster and clustering results and \( H(.) \) represents information entropy.

Accuracy is a measure to represent the effectiveness of the clustering result.

\[ Accu = \frac{\sum_{i=1}^{n} \delta(C_i, \text{map}(C_i))}{n} \] (8)
These three are the effectiveness measures of the cluster. Efficiency can be calculated in terms of Number of Iterations and Memory Usage. Number of iterations represents the total amount of iterations taken to form the clustering result. Memory usage represents memory needed for forming the cluster.

4. Experimental Results

We predict the cellular Localization site of proteins using yeast dataset as shown in Table 1. Yeast set consists of 9 attributes and 10 categories.

| Data Set   | Yeast   |
|------------|---------|
| Number of attributes | 9       |
| Number of Categories  | 10      |
| Usage       | Partly  |
| Cluster Name: Cluster_2 |
| Cluster Size: 5 |
| Cluster Medoid: 2 |

Each attribute defines information related to protein localization in which 8 attributes are predictive and remaining one is the name of the category.

4.1 IAPKM

Incremental affinity propagation based on K-method results in clusters of data in which each cluster consists of objects with exemplar id. 16 clusters are formed in which each cluster consists of cluster size and Medoid. The cluster method defines best exemplar among the entire exemplar id. The Table 2 shows the clustering result of IAPKM.

| ID = 2 | ID = 79 | ID = 81 | ID = 127 | ID = 142 |
|--------|---------|---------|----------|----------|
| 0.3978 | 0.3922  | 0.3490  | 0.4357   | 0.7566   |
| 0.5824 | 0.8829  | 0.5191  | 0.8189   | 0.4304   |
| 0.7076 | 0.6038  | 0.6083  | 0.6488   | 0.4762   |
| 1.4228 | 1.6725  | 0.9392  | 0.9422   | 1.1535   |
| 1.5296 | 1.3670  | 1.3125  | 1.3228   | 1.3927   |
| 1.4276 | 1.3895  | 1.5035  | 1.5359   | 1.1009   |
| 0.9927 | 0.8289  | 1.3087  | 1.5267   | 0.9008   |
| 0.4915 | 0.1265  | 0.6422  | 0.1192   | 0.8517   |
| 1.1657 | 1.1754  | 0.7092  | 0.6551   | 0.7322   |
| 0.9752 | 1.2130  | 1.5092  | 0.7798   | 0.7546   |

Cluster Name: Noise
Cluster Size: 3
Cluster Medoid: 164

In the table, ID represents exemplar and cluster size defines the number of exemplar in the cluster. Among this exemplar set cluster Medoid choose the best exemplar.

4.2 Limitations of IAPKM

Randomness occurred in IAPKM is in the form of noise which is generated along with the set of clusters. If data with an invalid set of attribute occurs, then it is added to noise cluster.

| ID = 164 | ID = 163 | ID = 165 |
|----------|----------|----------|
| 123456   | 0.002546 | 215603   |
| 201345   | 0.102364 | 125648   |
| 124563   | 1.456213 | 231546   |
| 124569   | 0.451236 | 124502   |
| 784563   | 0.001256 | 151200   |
| 123645   | 1.02364  | 254630   |
| 102365   | 1.456123 | 264651   |
| 126312   | 1.231456 | 235646   |
| 456312   | 0.748954 | 254521   |
| 102345   | 0.123645 | 789523   |

Cluster Size: 165

The Table 3 shows the invalid set of attributes which is placed in noise cluster.

4.3 Fuzzy DENCLUE

DENCLUE is the density based clustering algorithm in which single cluster is formed to set of exemplars. Based on density attractor, single cluster is formed in which all objects are placed. By applying Fuzzy concepts in DENCLUE, randomness occurred in IAPKM is removed. Randomness is defined as noise occurred in a cluster, which is an invalid range of attributes.

| ID = 165 | ID = 164 | ID = 1 |
|----------|----------|--------|
| 215603   | 102345   | 0.421185 |
| 125648   | 456312   | 1.082854 |
| 231546   | 126312   | 0.629526 |
| 124502   | 102365   | 1.341007 |
| 151200   | 123645   | 0.978398 |
| 254630   | 784563   | 1.464798 |
| 264651   | 124569   | 1.332644 |
| 235646   | 124563   | 0.360659 |
| 254521   | 201345   | 0.602224 |
| 789523   | 123456   | 0.937226 |

Cluster Size: 165
In Table 4, the number of clusters is reduced into a single cluster without noise.

4.4 Comparison Results
The effectiveness and efficiency measure for IAPKM and Fuzzy DENCLUE can be evaluated (Figure 1).

Figure 1. Comparison of IAPKM and Fuzzy DENCLUE in terms of Sum of Similarity (SS).

Figure 2. Comparison of IAPKM and Fuzzy DENCLUE in terms of effectiveness and efficiency measures.

The Figure 2 shows the Fuzzy DENCLUE method that achieves improved performance measures compared to IAPKM.

5. Conclusion

In this paper, clustering of static data using Affinity Propagation (AP) method is defined. It also focuses on Incremental Affinity Propagation based on K-Medoid (IAPKM) method for incrementally clustering the dynamic data. To overcome the limitations in IAPKM, the Fuzzy DENCLUE algorithm has been proposed in this paper. DENCLUE is the density based clustering algorithm for dynamically clustering the data using density attractor. By applying the Fuzzy concept such as a triangular membership function in DENCLUE the noise was removed and also the number of clusters was reduced. This method has attained a higher level of performance than the existing algorithm. Future work will focus on the reduction of range in performance measures that can be analyzed by Optimization approach.

6. References

1. Xu R, Wunsch D. Survey of clustering algorithms. IEEE Transactions on Neural Networks. 2005; 16(3):645–78.
2. Frey BJ, Dueck D. Clustering by passing messages between data points. Science. 2007; 315(5814):972–6.
3. Guha S, Meyerson A, Mishra N, Motwani R, O’Callaghan L. Clustering data streams: Theory and practice. IEEE Transactions on Knowledge and Data Engineering. 2003; 15(3):515–28.
4. Charikar M, Chekuri C, Feder T, Motwani R. Incremental clustering and dynamic information retrieval. SIAM Journal on Computing. 2004; 33(6):1417–40.
5. Beringer J, Hullermeier E. Online clustering of parallel data streams. Data and Knowledge Engineering. 2006; 58(2):180–204.
6. He Y, Chen Q, Wang X, Xu R, Bai X, Meng X. An adaptive affinity propagation document clustering. IEEE 27th International Conference on Informatics and Systems (INFOS); 2010.
7. Geng H, Deng X, Ali H. A new clustering algorithm using message passing and its applications in analyzing microarray data. IEEE Proceedings of 4th International Conference on Machine Learning and Applications; 2005.
8. Labroche N. New incremental fuzzy c medoids clustering algorithms. IEEE 2010 Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS); 2010.
9. Yang C, Bruzzone L, Guan R, Lu L, Liang Y. Incremental and decremental affinity propagation for semisupervised clustering in multispectral images. IEEE Transactions on Geoscience and Remote Sensing. 2013; 51(3):1666–79.
10. Ott L, Ramos F. Unsupervised incremental learning for long-term autonomy. IEEE International Conference on Robotics and Automation (ICRA); IEEE; 2012.
11. He J, Pan W. A DENCLUE based approach to neuro-fuzzy system modeling. IEEE 2nd International Conference on Advanced Computer Control (ICACC); 2010.
12. Shi X, Guan R, Wang L, Pei Z, Liang Y. An incremental affinity propagation algorithm and its applications for text clustering. International Joint Conference on Neural Networks (IJCNN’2009); IEEE; 2009.
13. Du C, Yang J, Wu Q, Zhang T. Face recognition using message passing based clustering method. Journal of Visual Communication and Image Representation. 2009; 20(8):608–13.
14. Sun L, Guo C. Incremental affinity propagation clustering based on message passing. 2014.
15. Han J, Kamber M, Pei J. Data mining. Southeast Asia edition: Concepts and Techniques. Morgan Kaufmann; 2006.
16. Senthilmarasu S, Hemalatha M. A genetic algorithm based intuitionistic fuzzification technique for attribute selection. Indian Journal of Science and Technology. 2013; 6(4):4336–46.
17. Jafarzadeh H, Torkashvand RR, Asgari C, Amiry A. Provide a new approach for Mining Fuzzy Association Rules using Apriori Algorithm. Indian Journal of Science and Technology. 2015; 8(8):707–14.