ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Incremental Affinity Propagation Clustering with Feature Selection

Cisy Soman, Abdul Ali

M Tech Scholar, Dept. of C.S, ICET, M.G University, Kottayam, India

Assistant Professor, Dept. of C.S., ICET, M.G University, Kottayam, India

ABSTRACT: Affinity Propagation (AP) is a clustering algorithm based on the concept of message passing. Unlike clustering algorithms such as k-means or k-medoids, AP does not require the number of clusters to be determined or estimated before running the algorithm. This paper deals with how to apply AP in incremental clustering problems. Here we use two clustering algorithms: Incremental AP based on K-Medoids (IAPKM) and Incremental AP based on Nearest Neighbour Assignment (IAPNA). However, a large number of features in the dataset may consume a lot of time. Therefore, we use a binary krill herd algorithm for feature selection, which help to reduce execution time and also increase the accuracy.

.

KEYWORDS: Affinity propagation, incremental clustering, k-medoids, feature selection

I. INTRODUCTION

Clustering is an important step in the process of data mining. Clustering data based on a measure of similarity is a critical step in scientific data analysis and in engineering systems [2]. Clustering aims at partitioning dataset in to several groups, often reffered to as clusters, such that data points in the same cluster are similar to each other than those in other clusters [1].

Different types of clustering are there in data mining. But most of them are used for mining patterns from static data. Nowadays, more and more data are available in dynamic manner like blogs, webpages. Therefore incremental clustering gaining importance in mining field. Characteristics of dynamic data include their high volume and potentially unbounded size, sequential access, and dynamically evolving nature. This imposes additional requirements on traditional clustering algorithms and make dynamic data clustering a challenge.

Affinity Propagation (AP) clustering is used to handle dynamic data. Here we use dynamic variant of AP clustering, which can achieve comparable clustering performance with traditional AP clustering by just adjusting the current clustering results according to new arriving objects, rather than re-implemented AP clustering on the whole data set. Therefore, a great deal of time can be saved, which makes AP clustering efficient enough to be used in dynamic environment.

AP clustering is an exemplar-based method that realized by assigning each data point to its nearest exemplar, where exemplars are identified by passing messages on bipartite graph. There are two kinds of messages passing on bipartite graph. They are responsibility and availability, collectively called 'affinity' by Frey and Dueck [2]. Such message-passing methods have been shown to be remarkably efficient in many hard problems that include error correction, learning in neural networks, digital signal processing and Bayesian inference in artificial intelligence.

The problem with AP clustering is that: the pre-existing objects have established certain relationships (nonzero responsibilities and nonzero availabilities) between each other after affinity propagation, while new objects' relationships with other objects are still at the initial level (zero responsibilities and zero availabilities). Objects added at different time are at the different statuses, so it's hard to find a proper exemplar set by simply continuing affinity propagation in this case.

Compared with the previous works [3], [4], another remarkable feature of our work is that the IAP clustering algorithms are proposed based on a message-passing framework. That's, each object is a node in a graph, and weighted edges between nodes correspond to pairwise similarity between objects. When a new object is observed, it will be added on the graph and then message passing is implemented to find a new exemplar set. Because that only one, or a

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

few of nodes' entering will not change the structure of the whole graph a lot, a local adjustment of availabilities and responsibilities is enough. Therefore, messages passing on graphs will reconverge quickly. Based on these features, the IAP clustering algorithms proposed in this paper don't need to re- implemented AP clustering on the whole data set, nor need to change the similarities between objects.

II. RELATED WORKS

Data analysis plays an indispensable role for understanding various phenomena. Cluster analysis, primitive exploration with little or no prior knowledge, consists of research developed across a wide variety of communities[5]. In [2], Clustering data by identifying a subset of representative examples is important for processing sensory signals and detecting patterns in data. Such "exemplars" can be found by randomly choosing an initial subset of data points and then iteratively refining it, but this works well only if that initial choice is close to a good solution. Here a method called "affinity propagation," is used, which takes as input measures of similarity between pairs of data points.

2.1 Affinity Propagation (AP) Clustering

AP clustering is an exemplar-based clustering method. It is realized by firstly picking out some special objects that called exemplars, and then associating each left object to its nearest exemplar.

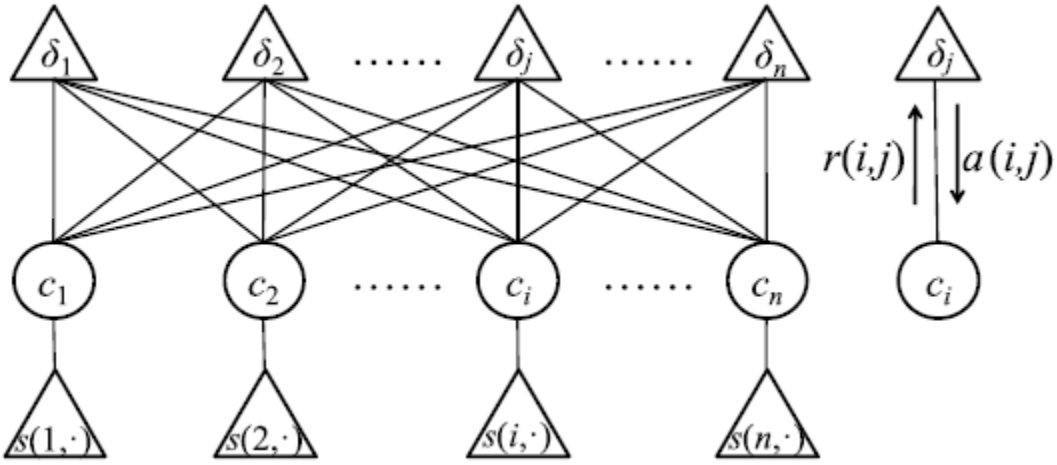

Fig 1. Factor graph of AP clustering

In Fig.1 , triangle nodes represent function nodes, circle nodes represent variable nodes. Object function is the sum of all the triangle nodes. There are two kinds of message passing on

graph. They are responsibilities and availabilities. Responsibility r(i,j) is sent from variable node ci to function node ∂j. It indicates how strongly object i wants to choose candidate exemplar j as its exemplar. Responsibility:

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

Availability a(i, j) sent from function node ∂j to variable node ci, reflects how well suited it would be for point i to choose point j as its exemplar. Availability:

$a(i,j) = \min\left\{0, r(j,j) + \sum_{i',i' \neq \{i,j\}} \max\{0, r(i',j)\}\right\}$

Responsibilities and availabilities update as above equations till convergence, then clustering result ĉ = (ĉ1,…,ĉn) can be obtained by

$\hat{c}i = \arg\max \{a(i,j) + r(i,j)\}$.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

III. INCREMENTAL AP CLUSTERING

The difficulty in incremental AP clustering is that: after affinity propagation, the first batch of objects have established certain relationships (nonzero responsibilities and nonzero availabilities) between each other, while new objects' relationships with other objects are still at the initial level (zero responsibilities and zero availabilities). Objects arriving at different time step are at the different statuses, so it is not likely to find the correct exemplar set by simply continuing affinity propagation in this case. So we are using two algorithms to overcome the problem.

3.1 Incremental AP Clustering Based on K-Medoids (IAPKM)

In this[1], a new subsequent clustering algorithm is designed that is not sensitive to the previous responsibilities and availabilities. That is, the first batch of objects are clustered by traditional AP clustering, when new objects arrive, new clustering algorithm is implemented to adjust the current clustering result.

The advantage of combining AP and K-medoids in an incremental clustering task is that: AP clustering is good at finding an initial exemplar set, while k-medoids is good at modifying the clustering result according to the new arriving objects.

3.2 Incremental AP clustering Based on Nearest Neighbour Assignment (IAPNA)

Here[1], all data points are put at the same status by reasonable assignment. That is, when new objects are coming, the relationships between new objects and other objects are assigned the proper values.

Then message passing procedure continues till convergence. Sometimes, the data points within the same cluster will be far away from the exemplar. In that case, we will specify a threshold value of the similarity. Above that value, new clusters will be formed.

3.3 IAP with Feature Selection

The dataset for the IAP clustering may contain many attributes i.e features. The processing of all these features consume more time. Therefore we use binary krill herd algorithm [6] to select relevant features based on an accuracy factor. Here we take different combinations of features and select the one with highest accuracy and store it in a global position. Only the features in the global position will be used for clustering and all other features will be removed. Thus we can reduce the processing time and increase accuaracy.

IV.SIMULATION RESULTS

We can compare the accuracy and time of the IAP clustering with and without feature selection using Weka tool [7]. Inorder to use Weka tool, the dataset should be in arff (Attribute Relation File Format).

4.1 Accuracy Comparison

Here the accuracy of clustering with feature selection (accuracy new) and without feature selection (accuracy old) is compared. From the figure it is clear that accuracy of clustering with feature selection is better than without feature selection.

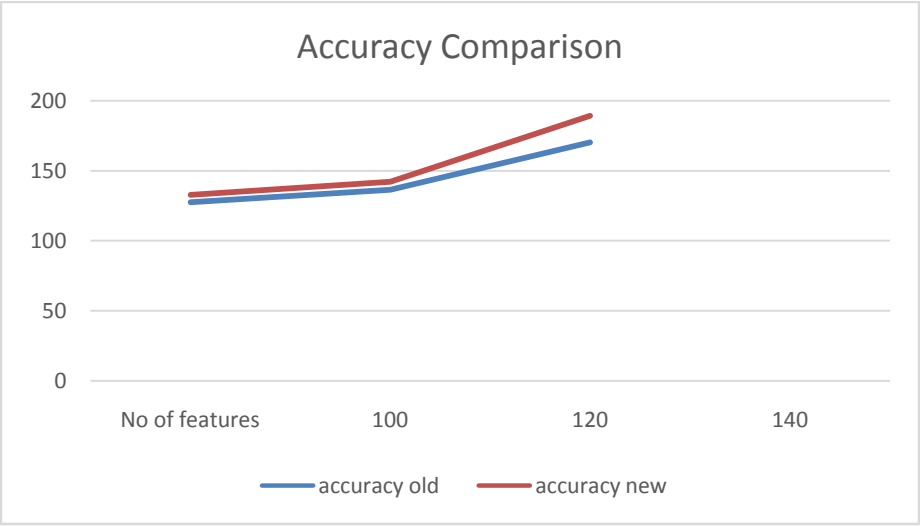

Fig 4.2 Accuracy Comparison

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

4.2 Time Comparison

Here the time of clustering with feature selection (time new) and without feature selection (time old) is compared. The feature selection removes the irrelevant features, which helps in reducing the execution time. As we can see from the graph that the time will be always is less with feature selection than without feature selection.

Fig 4.2 Time Comparison

V. CONCLUSION

In this paper, we discussed about how to apply AP in the incremental clustering task. It is an efficient method for clustering dynamic data. From the simulation results, we understand that by applying feature selection we can reduce the execution time and increase the accuracy. The error rate can also be reduced by removing irrelevant features.

REFERENCES

1.Leilei Sun and ChonghuiGuo "Incremental Affinity Propagation Clustering Based on Message Passing" , IEEE Transactions on Knowledge and Data Engineering, Vol.26 , No.11, November 2014B

2..J. Frey and D. Dueck, "Clustering by Passing Messages between Data Points," Science, vol. 315, no. 5814, pp. 972-976, Feb. 2007

3.M. Charikar, C. Chekuri, T. Feder, and R. Motwani, "Incremental Clustering and Dynamic Information Retrieval," Proc. ACM Symp.Theory of Computing (STOC '97), pp. 626-635, 1997

4.C. Du, J. Yang, Q. Wu, and T. Zhang, "Face Recognition Using Message Passing Based Clustering Method," J. Visual Comm. And Image Representation, vol. 20, no. 8, pp. 608-613, Nov. 2009.

5.R. Xu and D. Wunsch, "Survey of Clustering Algorithms," IEEE Trans. Neural Networks, vol. 16, no. 3, pp. 645-677, May 2005.

6.Douglas Rodrigues , Lu´ıs A. M. Pereira, Jo˜ao P. Papa , and Silke A. T. Weber "A Binary Krill Herd Approach for Feature Selection" 2014 22nd International Conference on Pattern Recognition.

7.Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham, "Weka: Practical Machine Learning Tools and Techniques with Java Implementations"

BIOGRAPHY

Cisy Somanis aM Tech scholar in the Computer Science department, ICET, MG University. She received Bachelor of Technology (B Tech) degree in 2013 from M.G University, Kottayam, Kerala. Her research interests are data mining and databases.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Abdul Ali is an Assistant professor in the Computer Science department, ICET, M.G University. He received B-Tech degree in 2007 from M. G. University, Kottayam, Kerala. He received M Tech degree in 2010 from M S university, Trinulveli. His research interests are image processing and networking.