International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Streaming Data Clustering by Incremental Affinity Propagation

Jismy Muhammed

Computer Science and Engineering, ICET, Mahatma Gandhi University, Muvattupuzha, Kerala, India

Abstract: ***This paper introduce based on clustering. Affinity Propagation (AP) clustering has been successfully used in a lot of clustering problems. This paper considers how to apply AP in incremental clustering problems in streams of data . First, we point out the difficulties in Incremental Affinity Propagation (IAP) clustering, and then propose two strategies to solve them. Two IAP clustering algorithms are proposed. They are IAP clustering based on K-Medoids (IAPKM) and IAP clustering based on Nearest Neighbor Assignment (IAPNA). Traditional AP clustering is also implemented to provide benchmark performance. Experimental results show that IAPKM and IAPNA can achieve comparable clustering performance with traditional AP clustering on all the data sets. Popular labeled data sets, real world time series and a video, datastreams are used to test the performance of IAPKM and IAPNA. Both the effectiveness and the efficiency make IAPKM and IAPNA able to be well used in incremental clustering tasks.***

Keywords: Affinity propagation, incremental clustering, K-medoids, nearest neighbor assignment

1. Introduction

Clustering, or cluster analysis, is an important subject in data mining. It aims at partitioning a data set into some groups, often referred to as clusters, such that data points in the same cluster are more similar to each other than to those in other clusters.

There are different types of clustering. However, most of the clustering algorithms were designed for discovering patterns in static data. Now days, more data, e.g., blogs, Web pages, video surveillance, etc., are appearing in dynamic manner, known as data streams. data stream is an ordered sequence of points x1; . . . ; xn that must be accessed in order and that can be read only once or a small number of times. The data stream and online or incremental models are similar in that they both require decisions to be made before all the data are available.

Therefore, incremental clustering, evolutionary clustering, and data stream clustering are becoming important topics in data mining. The dynamic data, or data streams, include their high volume and potentially unbounded size, sequential access, and dynamically evolving nature. In this paper, we extend a recently proposed clustering algorithm, affinity propagation (AP) clustering, to handle dynamic data. The goal of this paper is to propose a 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'

Extensions of AP clustering in dynamic environment have been discussed. The new object is assigned to an exemplar if fit criterion is satisfied. Otherwise, it is put into a reservoir. When the size of reservoir is big enough, traditional AP is re-implemented to empty reservoir. AP clustering was recomputed nearly for every newly observed data point. This obviously didn't work when real-time performance was required. Therefore, they improved the efficiency of streaming AP clustering by limiting the numbers of recomputing. Proposed a semi-supervised based incremental AP clustering, IAP is realized by adjusting similarity matrix. The similarities between objects with same label are set much larger than usual, and those with different label are set much smaller. Clustering performance was improved by both label-based similarity matrix adjusting and ID learning principle.

In this paper, we point out that the difficulty of extending AP in dynamic data clustering is that, the pre-existing objects Have established some 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. This problem will be discussed in a further step in this paper. And then, two strategies will be proposed to overcome this problem.

2. Related Work

A data stream is defined as an ordered sequence of data points that can be read only once or a small number of times in [4], and data stream clustering is expected to be a single-pass algorithm. It's assumed that only one object is observed at each time step. Incremental clustering is defined as follows: for an orderly given data set, maintain a collection of clusters such that when some new objects are arriving, either assign them into the current clusters, or create a new cluster. Divided the existing incremental clustering algorithms into two classes: Single Pass Incremental

Algorithms (SPIAs) and Cluster Center Adding Algorithms (CCAAs). Most of the existing incremental clustering algorithms are SPIAs, where new objects are added at each iteration and cluster centers are redefined accordingly. In CCAAs, the number of clusters is not fixed.

2.1 Affinity Propagation Clustering

Exemplar-based clustering is one of the most important clustering algorithms. It is realized by firstly picking out some special objects that called exemplars, and then associating each left object to its nearest exemplar.

2.2 Incremental AP Clustering

Incremental AP clustering is still a difficult problem. The difficulty in incremental AP clustering is that: after affinity propagation, the first batch of objects has 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 .

Fig. 2 is a toy example to demonstrate such a problem, where data come from the computational experiments. traditional AP clustering is implemented on the first batch of objects. Responsibilities and availabilities converge in Fig. 2d. The clustering result is shown in Fig. 2e. New objects, represented by triangle nodes, arrive in Fig. 2f. In order to illustrate, the exemplar at the upper left corner is denoted by point A, the triangle point nearest to point A is denoted by point B. Among the five new arriving objects, point B is a particularly generated object, which is the center of the upper left cluster. Therefore, interclass similarity can be increased by choosing point B as the exemplar, instead of point A. The left four are randomly generated. It can be observed that, as an old object, point A has accumulated a lot of supports (represented by arrow lines directing at point A) in Fig. 2f

In this paper, we propose two strategies to overcome such a problem: 1) Design new subsequent clustering algorithm that is not sensitive to the previous responsibilities and availabilities. That is, the first batch of objects is clustered by traditional AP clustering, when new objects arrive, new clustering algorithm is implemented to adjust the current clustering result. 2) Put all the data points at the same status by reasonable assignment. That is, when new objects are coming, the relationships between new objects and the other objects are assigned the proper values. Then message passing procedure continues till convergence. Correspondingly, two IAP clusteirng algorithms are proposed in the following paper.

3. Methodologies

3.1 Incremental AP Clustering Based on K-Medoids

In this we propose an incremental AP clustering algorithm according to the first strategy. K-Medoids is chosen to be used as the subsequent clustering algorithm. K-Medoids is also an exemplar-based clustering algorithm, and has been widely used. Due to its simplicity, many dynamic clustering variants of K-Medoids have been proposed. The algorithm proposed in this section is named Incremental AP clustering Based on K-Medoids (IAPKM).

The rationality 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 new arriving objects. the final exemplar set is usually found around the initial exemplar set. Therefore, the final clustering performance of K-Medoids largely depends on the initial exemplar set. However, this problem can be overcome by AP clustering. AP clustering can find a good exemplar set automatically.

IAPKM consists of two basic steps: AP clustering is implemented on the initial batch of objects, and K-Medoids is employed to modify the current clustering result according to the new arriving objects. In order to combine K-Medoids with AP clustering, K-Medoids is introduced in a message-passing manner. Fig. 1 is a toy example to illustrate IAPKM. In Figs. 1f,1g, 1h, and 1i, new message-passing algorithm implemented

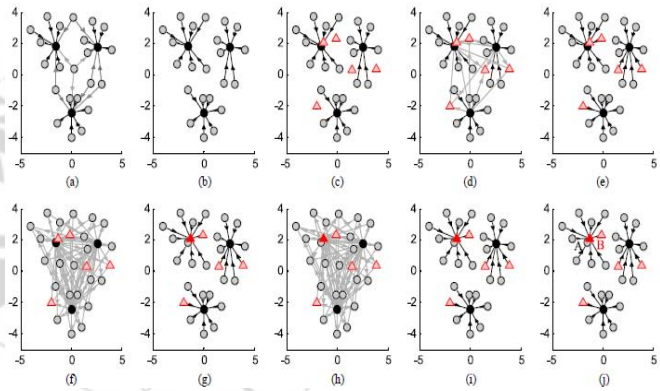

Figure 1: A toy example to illustrate IAPKM. Traditional AP clustering is implemented on the initial batch of objects, and the clustering result is shown in (b). New objects arrive in (c). Each new object decides which exemplar it belongs to in (d) and (e). New message-passing algorithm is implemented in (f)_(i). The final clustering result is shown in (j).

3.2 Incremental AP Clustering Based on Nearest Neighbor Assignment

In this section, an incremental AP clustering algorithm is proposed according to the second strategy. A technique of Nearest-neighbor Assignment (NA) is employed to construct the relationships (values of responsibilities and availabilities) between the new arriving objects and the previous objects. NA means that the responsibilities and availabilities of the new arriving objects should be assigned referring to their nearest neighbors. NA is proposed based on such a fact that if two objects are similar, they should not only be clustered into the same group, but also have the same relationships (responsibilities and availabilities). However, most of the current algorithms utilize the former part only.

Fig. 2 is a toy example to illustrate IAPNA. From Fig. 2d, it can be observed that, after nearest neighbor assignment, point B has established nonzero relationships with other points by copying responsibilities and availabilities from point A. Other new objects also established nonzero relationships according to their nearest neighbors[3]. In Fig. 2j, point B becomes the exemplar of the upper left cluster, which indicates IAPNA also achieves a better exemplar set.projected databases, and then construct an FP-tree from each of these smaller databases. IAPNA can discovery new cluster automatically.

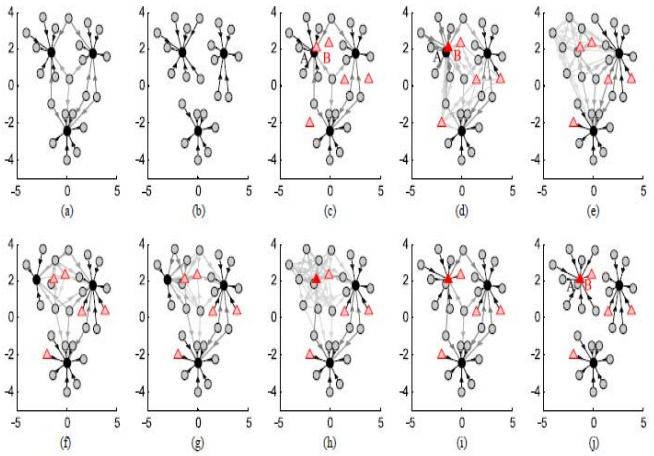

Figure 2: IAPNA can discovery new cluster automatically. New objects arrive in (c), they all come from a new cluster. NA is implemented in (d), and message-passing continues in (e)_(h). IAPNA reconverges in (i), and the final result is shown in (j).

4. Computational Experiments

This section presents evaluation results of IAPKM and IAPNA on popular labeled data sets, EEG signals and a video streaming,data streaming etc. The goal of this paper is to propose an incremental variant of AP clustering, which can achieve comparable clustering performance with traditional AP by adjusting the current clustering results according to new arriving objects. Therefore, traditional AP clustering is implemented to provide benchmark clustering performance.

The computational experiments on data set demonstrate that, IAPKM and IAPNA can achieve higher, or at least comparable, clustering performance than traditional AP clustering. At the same time, the computational complexity is dramatically reduced, which makes AP clustering able to be used in dynamic environment.

4.1 Test By Data Streams

Generally, the clustering performance of an algorithm is evaluated by external dispersity and internal dispersity. The sum of similarities is one of the most widely used criteria. In some cases, different clustering result can obtain comparable external dispersity and internal dispersity.Therefore, we use labeled data streams to evaluate the proposed algorithms in this section. An advantage is that we cannot only evaluate the clustering algorithms by dispersity, but also by some other indicators, e.g., mutual information, clustering accuracy.

4.2 Test by Real World Time Series

In this section, IAPKM and IAPNA are used to discovery patterns in real world time series. The test problem is pattern recognition in epileptic EEG signals. The epileptic EEG signals are provided by Epilepsy Center of the University. Hospital of Freiburg, which is very popular in the research of epileptic seizure prediction. The EEG signals were acquired using a Neurofile NT digital video EEG system with 256 Hz sampling rate. Following Chisci's work the width of time window is set to contain 512 data points, corresponding to a window length of 2 second

4.3Test by Video Streaming

IAPKM and IAPNA are used to clustering video streaming. Similarly, traditional AP clustering is implemented to provide benchmark clustering performance.The test video is cut from a BBC documentary The Secret Life of Chaos, which is very popular and can be found in many video websites. In this experiment, only a segment of it, from 3'30" to 12'50", is used. This segment introduces the contribution of Alan Turing to chaos theory.The topic of this paper is to extend AP in incremental clustering task. Therefore, the supposed scenario is that, the first 6 minutes of the video segment has been clustered by traditional AP clustering, then how to obtain the real time clustering result as the video streaming pours in gradually.

In this experiment, every 2 second a frame of the video is sampled. That's, the number of the first batch of objects is 180 (660/2), then every 2 seconds, a new frame pours in the model, the clustering result is renewed according toIAPKM and IAPNA. In this paper, color histogram is used to extract features of frames, and euclidean distance is used to compute similarity between frames.Two hundred seconds later, the clustering result has been renewed 100 times. The current clustering result is shown in Fig. 3. From Fig. 3, we can see that, the video segment can be divided into seven categories. The clustering results of IAPKM and IAPNA are almost the same. Exemplars identified by IAPKM have been marked in Fig. 3a,Three other frames of each category are randomly selected and also displayed. Exemplars identified by IAPKM are marked in Fig. 3b, corresponding frames are not given in this paper as the results of IAPKM and IAPNA are nearly the same.

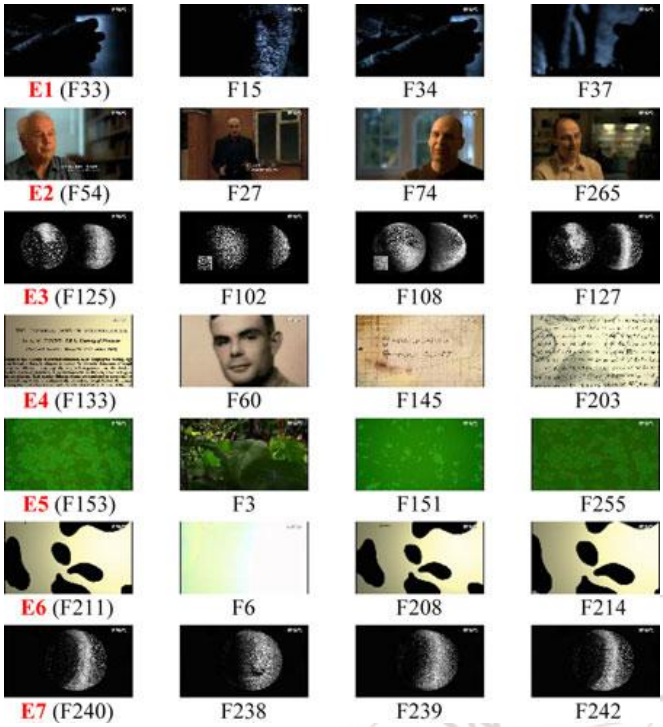

Figure 3: Represent exemplars identified by IAPKM. For each category, three other frames are displayed.

4. Conclusion

In this paper, we consider how to apply AP in incremental clustering task. We first point out the difficulty in IAP clustering,and then propose two strategies to solve it. Correspon dingly,two IAP clustering algorithms, IAPKM and IAPNA, are proposed. Five popular labeled data sets ,real world time series data streams and are used to evaluate the performance of IAPKM and IAPNA. Experimental results validate the ffectiveness of IAPKM and IAPNA. And also apply the two ideas to other dynamic data clustering tasks, such as streaming data clustering,

5. Acknowledgement

The author would like to thank Chithra Rani, Assistant Professor, Department of Information technology, Ilahia College of Engineering and Technology, Muvattupuzha for her moral and technical support.

References

[1] L. Ott and F. Ramos, "Unsupervised Incremental Learning for Long-Term Autonomy," Proc. IEEE Int'l Conf.Robotics and Automation(ICRA '12), pp. 4022-4029, May 2012.
[2] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques ,third ed., p. 444, Morgan Kaufmann, 2011.
[3] L. Nicolas, "New Incremental Fuzzy c Medoids Clustering Algorithm Proc. Ann. Meeting of the North Am. Fuzzy Information Processing Society (NAFIPS '10), pp. 1-6, July 2010.
[4] A.K. Jain, "Data Clustering: 50 Years Beyond K-Means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651-666, June 2009.
[5] B.J. Frey and D. Dueck, "Response to Comment on 'Clustering by Passing Messages between Data Points," Science, vol. 319, no. 5864, pp. 726a-726d, Feb. 2008.
[6] X. Zhang, C. Furtlehner, and M. Sebag, "Frugal and Online Affinity Propagation," Proc. Conf. Francophone sur l'Apprentissage (CAP '08), 2008.
[7] M. Mezard, "Where Are the Exemplars," Science, vol. 315 no. 5814, pp. 949-951, Feb. 2007.
[8] B.J. Frey and D. Dueck, "Clustering by Passing Messages between Data Points," *Science,* vol. 315, no. 5814, pp. 972-976, Feb. 2007.
[9] J.S. Yedidia, W.T. Freeman, and Y. Weiss, "Constructing Free- Energy Approximations and Generalized Belief Propagation Algorithms," IEEE Trans. Information Theory, vol. 51, no. 7 pp. 2282-2312, July 2005.

Author Profile

Jismy Muhammed received the Bachelor of Technology degree in Computer Science and Engineering from Mahatma Gandhi University, Kerala. She is currently doing Master of technology degree in Computer Science and Engineering with Specialization in Information Systems from Mahatma Gandhi University, Kerala.