IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 37-40
www.iosrjournals.org

A Survey Paper on Clustering Data using Incremental Affinity Propagation

Sreeja Ajithkumar[1], Praveen K Wilson[2]
[1]*(Department of Computer Science, College of Engineering Perumon, Kollam, India)*
[2]*(Department of Information Technology, College of Engineering Perumon, Kollam, India)*

Abstract: *Clustering is vital part of data mining technique and widely used in different applications. In this project we are focusing on Affinity propagation (AP) clustering algorithm, which is presented recently to overcome many clustering problems in different clustering applications. Many clustering applications deal with static data. AP clustering supports only static data applications; hence it becomes research problem that how to deal with incremental data (dynamic data) using AP. To solve this problem, recently proposed Incremental Affinity Propagation (IAP) clustering. Two strategies are proposed to overcome the difficulties in IAP clustering. Then two IAP clustering algorithms are proposed, namely IAP clustering based on K-Medoids (IAPKM) and IAP clustering based on Nearest Neighbor Assignment (IAPNA). Both IAPKM and IAPNA can achieve comparable clustering performance with traditional AP clustering on all the data sets. The time cost can be dramatically reduced in IAPKM and IAPNA. Then propose a feature grouping method which increases the prediction accuracy by increasing the values of precision and recall.*

Keywords: *Affinity propagation, Data streams, Incremental Affinity Propagation, K-Medoid, Nearest neighbor assignment.*

I. Introduction

Clustering or cluster analysis is an important research topic in data mining. It is the process of partitioning a set of data objects into subsets. Each subset is known as Cluster. Objects in the same cluster are more similar to each other than to those in other cluster. Clustering find a variety of applications in the image pattern recognition, structure identification in an unstructured data, web search, business intelligence. In 1955 the first clustering algorithm K-means was published. Thousands of different clustering algorithms have been proposed since then but the general purpose clustering algorithm is yet to be standardized. This is because the wide range of formats of the unstructured data.

One of the clustering method is used that can identify the data center points or clusters by simply sending the messages between data points pairs is called as Affinity Propagation(AP). In statics and data mining Affinity propagation is a clustering algorithm based on the concept of message passing between data points. Before running the Affinity propagation clustering algorithm it does not know the number of clusters to be determined. Affinity propagation finds exemplars, members of the input set that representative of clusters. Affinity Propagation is advanced algorithm which first find out the similarity between the pairs of data items that taken as an input as well as check all the data items called as exemplars. The proposition of IAPKM is inspired by combining K-Medoids and AP clustering, where AP clustering is good at finding an initial exemplar set and K-Medoids is good at modifying the current clustering result according to new arriving objects.

Affinity propagation clustering is emerging which is an “Exemplar”-based approach. It is given by the assignment of the data points to their nearest exemplar. In this paper we are extending the Affinity propagation approach to work in the streamed data environment. A new approach to handle the streamed data is proposed to adjust the clustering results as the new object arrives. This reduces the time required to apply the clustering to the whole data set. So, an efficient approach is designed to work with the dynamic data. Traditional AP clustering is also put into practice to supply standard performance. Experimental results show that IAPKM and IAPNA can realize equal clustering performance with established AP clustering on all the data sets. For now, the time cost is considerably reduced in IAPKM and IAPNA. Both the effectiveness and the efficiency make IAPKM and IAPNA gifted to be well used in incremental clustering tasks. We point out that the anonymity of extending AP in dynamic data clustering is that, the pre-existing objects have customary definite relationships (non-zero responsibilities and non-zero availabilities) between each other after affinity propagation, while new objects’ relationships with other objects are motion-less at the initial level (zero responsibilities and zero availabilities). Objects added at different time are at the different statuses, so its firm to find a good exemplar set by just progressing affinity propagation in this case.

II. Existing System

Clustering can be defined as grouping a set of objects into different classes (clusters) so that the similar objects in a particular sense get added to the same class and the objects with dissimilarities get in the different classes. Sometimes the clustering can be used to form the natural clusters based on the natural hierarchy.

Affinity propagation find a wide range of applications in clustering the images of faces, detect genes in microarray data, identify representative sentences in this manuscript, and identify cities that are efficiently accessed by airline travel. An incremental Affinity propagation algorithm was proposed aimed at streaming data by Xiangliang Zhang, Cyril Furtlehner [1] in 2008. The bipartite or factor graph is used to represent the message passing between the different local functions. The Time series data streams like stock rate i.e. data items in the real number form were clustered by J. Beringer and E. Huller Meier [2]. Affinity propagation clustering does not need the number of clusters to be specified previously as that was needed in the former approaches K-mean, instead it takes similarity value s (k, k) as an input for each data point so that the data point having maximum s (k, k) is chosen as an exemplar.

Zhang etal. Proposed a streaming AP clustering algorithm. In their work new object is allocate to an exemplar if fit principle is satisfied. Otherwise, it is put into reservoir. When the size of reservoir is big enough, traditional AP is re-implemented to empty reservoir. Ott et al. pointed out that, in Zhang et al.'s work, AP clustering was recomputed almost for every recently observed data point. This visibly didn't work when real-time performance was required. Consequently, they better the competence of streaming AP clustering by limiting the numbers of recomputing. Shi et al. proposed a semi-supervised based incremental AP clustering, and applied it in text clustering.

The Author, X.H. Shi, [3] (Et. Al), Aim in a semi-supervised system called incremental affinity propagation clustering is proposed in the paper. In the scheme, the pre-known information is represented by fiddle with similarity matrix. Moreover, an incremental revise is applied to strengthen the prior knowledge. To observe the effectivity of the method we use it to text clustering problem and describe the specific method accordingly. The method is practical to the benchmark dataset Reuters-21578. Numerical results show that the proposed method performs exceptionally well on the data set and has generally compensation over two other generally used clustering methods.

The Author, Adil M. Bagirov Julien Ugon [4] (ET. Al) Aim in, the k-means algorithm and its differences are known to be fast clustering algorithms. Though they are responsive to the choice of starting points and are incompetent for unravel clustering problems in large data sets. Of late incremental approaches have been developed to determine difficulties with the alternative of starting points. The global k-means and the modified global k-means algorithms are based on such an approach. They iteratively add one cluster Centre at a time. Numerical experiments show that these algorithms very much improve the k-means algorithm. This makes both algorithms time overwhelming and reminiscence demanding for clustering even reasonably great datasets. In this paper, a new version of the modified global k-means algorithm is proposed. We set up supplementary cluster function to produce a set of starting points two-faced in different parts of the dataset.

Geng [5]. proposed a new clustering algorithm which outperforms hierarchical clustering in some aspects. Hwang [1]. proposed a signal transduction mock-up for clustering and sense functional modules in protein-protein interaction networks. As well some traditional clustering methods can also be well made clear in a message-passing manner. Lead-to of AP clustering in self-motivated environment have been thrashed out by many researchers. B.J.Frey and D.Dueck [6]explained an algorithm termed "affinity propagation" (AP) as a talented option to traditional data clustering procedures. We show that heuristic for the p-median difficulty frequently obtain clustering solutions with inferior error than AP and create these solutions in similar computation time.

Xiangliang Zhang [1], A new Data Clustering algorithm, Affinity Propagation undergoes from its quadratic difficulty in function of the number of data items. Some extension of Affinity Propagation was proposed aspire at online clustering in the data stream framework. Initially the case of increase defined items or weighted items are handled using Weighted Affinity Propagation (WAP). Secondly, Hierarchical AP achieves distributed AP and uses WAP to merge the sets of exemplars learned from subsets. Based on these two building blocks, the third algorithm performs Incremental Affinity Propagation on data streams. The paper legalizes the two algorithms both on standard and on real-world datasets.

The weighted fuzzy c-mean clustering algorithm (WFCM) and weighted fuzzy c-mean-adaptive cluster number (WFCM-AC) [1] are extension of traditional fuzzy c-mean algorithm to stream data clustering algorithm. Clusters in WFCM are generated by renewing the centers of weighted cluster by iteration. On the other hand, WFCM-AC generates clusters by applying WFCM on the data & selecting best K± initialize center. In this paper we have compared these two methods using KDD-CUP'99 data set. We have compared these algorithms with respect to number of valid clusters, computational time and mean standard error.

F.R. Kschischang, B.J. Frey, and H.A. Loeliger [7] present a generic message-passing algorithm, the sum-product algorithm, that operates in a factor graph. Following a single, simple computational rule, the sum-product algorithm computes—either exactly or approximately—various marginal functions derived from the global function. A wide variety of algorithms developed in artificial intelligence, signal processing, and digital communications can be derived as specific instances of the sum-product algorithm, including the forward/backward algorithm, the Viterbi algorithm, the iterative "turbo" decoding algorithm, Pearl's belief propagation algorithm for Bayesian networks, the Kalman filter, and certain fast Fourier transform (FFT) algorithms.

To achieve high efficiency of classification in IDS Chen et. al. proposed a compressed model. In compressed model, one classifier is used for horizontal compression and Affinity propagation (AP) employed as vertical compression. AP used in many clustering problems. Sun & Guo used AP in incremental clustering problems. AP clustering based on K-medoids. Feature selection or Attribute reduction are the modern methods to improve detection efficiency. Chen et. al improved efficiency of the model with the increasing accuracy so that the model will detect intrusions. Compressed model is built using training data.

III. Proposed System

There are diverse types of clustering. On the other hand, nearly every one of the clustering algorithms was intended for determine outline in static data. This induces additional requirements to traditional clustering algorithms to summarily method and précis the massive amount of continually arriving data. It also needs the ability to get used to changes in the data distribution, the aptitude to detect up and-coming clusters and distinguish them from outliers in the data and the ability to join old clusters or abandon end ones. Clustering or cluster psychiatry is a vital subject in data mining. It plans at separator a dataset into some groups often referred to as clusters such that data points in the same cluster are additional analogous to each other than to those in other clusters.

AP clustering is an exemplar-based method that realized by handover each data point to its nearest exemplar, where exemplars are recognized by passing messages on bipartite graph. There are two kinds of messages passing on bipartite graph. They are responsibility and availability, jointly called 'affinity'. AP clustering can be seen as a request of belief propagation, which was pretend by Pearl to grip deduction problems on probability graph. The objective of this paper is to offer an active alternative of AP clustering, which can attain comparable clustering recital with traditional AP clustering by just regulating the current clustering results according to new arriving objects, somewhat than re-implemented AP clustering on the whole dataset. We lengthen a recently proposed clustering algorithm, affinity propagation (AP) clustering, to grip dynamic data. Several experiments have shown its reliable advantage over the preceding algorithms in static data. Incremental affinity propagation clustering is used for the purpose of clustering dynamic data. To solve the difficulties of Incremental affinity propagation(IAP) algorithm propose two algorithms namely, Incremental affinity propagation (IAP) clustering K-Medoids (IAPKM) and Incremental affinity propagation (IAP) clustering based on Nearest Neighbor assignment (IAPNA). The time cost is dramatically reduced in IAPKM and IAPNA. Then propose a feature grouping method. The feature grouping increases the prediction accuracy by increasing the values of precision and recall.

IV. Conclusion

In this paper, we consider how to apply AP in incremental clustering task. First point out the difficulty in IAP clustering, and then propose two plan of action to solve it. Correspondingly, two IAP clustering algorithms, IAPKM and IAPNA, are proposed.

The proposition of IAPKM is inspired by combining K-Medoids and AP clustering, where AP clustering is good at finding an initial exemplar set and K-Medoids is good at modifying the current clustering result according to new arriving objects. Experimental results show the correctness of this idea. By combing K-Medoids and AP clustering, we cannot only extend AP to competent an incremental clustering task, but also improve the clustering performance of AP clustering. IAPNA is realized by a technique called nearest neighbor assignment. The proposition of NA is based on such an idea that "if two objects are similar, they should not only be clustered into the same group, but also have the same statuses". Both the two ideas are significant, and will be very helpful in dynamic clustering design. The feature grouping increases the prediction accuracy by increasing the values of precision and recall.

References

[1] X. Zhang, C. Furtlehner, and M. Sebag, "Frugal and Online Affinity Propagation," Proc. Conf. Francophone sur l'Apprentissage (CAP '08), 2008.

[2] J. Beringer and E. Hullermeier, "Online Clustering of Parallel Data Streams," Data and Knowledge Eng., vol. 58, no. 2, pp. 180-204, Aug. 2006.

[3] X.H. Shi, R.C. Guan, L.P. Wang, Z.L. Pei, and Y.C. Liang, "An Incremental Affinity Propagation Algorithm and Its Applications for Text Clustering," Proc. Int'l Joint Conf. Neural Networks (IJCNN '09), pp. 2914-2919, June 2009.

[4] A.M. Bagirov, J. Ugon, and D. Webb, "Fast Modified Global kMeans Algorithm for Incremental Cluster Construction," Pattern Recognition, vol. 44, no. 4, pp. 866-876, Nov. 2011.

[5] H. Geng, X. Deng, and H. Ali, "A New Clustering Algorithm Using Message Passing and Its Applications in Analyzing Microarray Data," Proc. Fourth Int'l Conf. Machine Learning and Applications (ICMLA '05), 2005.

[6] 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.

[7] F.R. Kschischang, B.J. Frey, and H.A. Loeliger, "Factor Graphs and the Sum-Product Algorithm," IEEE Trans. Information Theory, vol. 47, no. 2, pp. 498-519, Feb. 2001.

[8] T.W. Liao, "Clustering of Time Series Data: A Survey," Pattern Recognition, vol. 38, no. 11, pp. 1857-1874, Nov. 2005.

[9] M. Mezard, "Where Are the Exemplars," Science, vol. 315, no. 5814, pp. 949-951, Feb. 2007.

[10] H. Geng, X. Deng, and H. Ali, "A New Clustering Algorithm Using Message Passing and Its Applications in Analyzing Microarray Data," Proc. Fourth Int'l Conf. Machine Learning and Applications (ICMLA '05), 2005.