International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 10, October 2015
Available at http://internationaljournalofresearch.org

Incremental Affinity Spread Clustering Created on Message Passing

P.Krishna Chaitanya#1 & P.Anil#2

#1 Asst. Professor, Dept. Of CSE, MalineniLakshmaiah Engineering College(MLEC), Singarayakonda.Prakasam,AP

#2 PG Student, Dept. Of CSE, MalineniLakshmaiah Engineering College(MLEC), Singarayakonda.Prakasam,AP

ABSTRACT :

Affinity propagation (AP) is a clustering method that can find data centers or clusters by sending messages between pairs of data points. Seed Affinity Propagation is a novel semi supervised text clustering algorithm which is based on AP. AP algorithm couldn't cope up with part known data direct. Therefore, focusing on this issue a semi-supervised scheme called incremental affinity propagation clustering is present in the paper where pre-known information is represented by adjusting similarity matrix The standard affinity propagation clustering algorithm also suffers from a limitation that it is hard to know the value of the parameter "preference" which can yield an optimal clustering solution. This limitation can be overcome by a method named, adaptive affinity propagation. The method first finds out the range of "preference", then searches the space of "preference" to find a good value which can optimize the clustering result.

Index Terms—Affinity propagation; incremental clustering; *K*-Medoids; nearest neighbor assignment

I. INTRODUCTION

Clustering is a process of organizing objects into groups whose members are similar in some way. Cluster analysis seeks to partition a given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups [1], Therefore, a cluster is a collection of objects that are similar among themselves and dissimilar to the objects belonging to other clusters. Text Clustering is to divide a set of text into cluster , so that text within each cluster are similar in content. Clustering is an area of research, finding its applications in many fields. One of the most popular clustering method is k-means clustering algorithm. Arbitrarily k- points are generated as initial centroids where k is a user specified parameter. Each point is then assigned to the cluster with the closest centroid then the centroid of each cluster is updated by taking the mean of the data points of each cluster. Some data points may move from one cluster to other cluster. Again calculate new centroids and assign the data points to the suitable clusters. Repeat the assignment and update the centroids, until convergence criteria are met. In this algorithm mostly Euclidean distance is used to find distance between data points and centroids. Standard K-Means method has two limitations: (1) the number of cluster needs to be specified first. (2) the clustering result is sensitive to the initial cluster centers . Traditional approaches for clustering data are based on metric similarities, i.e. symmetric, nonnegative, and satisfying the triangle inequality measures. More recent approaches, like Affinity Propagation (AP) algorithm [2], can take as input also general non metric similarities. In the domain of image clustering, AP can use as input metric selected segments of images pairs [3]. AP has been used to solve a wide range of clustering problems

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 10, October 2015
Available at http://internationaljournalofresearch.org

[4] and individual preferences predictions [5].The clustering performance depends on the message updating frequency and similarity measure AP has been used in text clustering for its simplicity, good performance and general applicability. By using AP to preprocess text for text clustering. It was combined with a parallel strategy for e-learning resources clustering. But AP was used only as an unsupervised algorithm and did not consider any structural information derived from the specific documents. For text mining tasks, the vector space model (VSM), which treats a document as a bag of words and uses plain language words as features [6]. This model can represent the text mining problems directly and easily. With the increase of data set size, the vector space becomes sparse, high dimensional, and the computational complexity grows exponentially. In many practical applications, unsupervised learning is lacking relevant information whereas supervised learning needs an initial large number of class label information, which requires time and expensive human labor. [7], [8]. In recent years, semi-supervised learning has captured a great deal of attentions [9], [10]. Semisupervisedlearning is a machine learning in which the model is constructed using both labeled and unlabeled data for training—typically a small amount of labeled data and a large amount of unlabeled data

II . RELATED WORK

This section includes so far study on Affinity propagation.

2.1 Affinity Propagation

AP, a new and powerful technique for exemplar learning. It is a fast clustering algorithm especially in the case of large number of clusters and has some advantages: speed, general applicability and good performance. In brief , AP works based on similarities between pairs of data points and simultaneously considers all the data points as potential cluster centers called exemplar. It computes two kinds of messages exchanged between data points. The first one is called ―responsibility‖ and the second one is ―availability‖. Affinity propagation takes as input a collection of real-valued similarities between data points, where the similarity s(i,k) indicates how well the data point with index k is suited to be the exemplar for data point i. When the goal is to minimize squared error, each similarity is set to a negative squared error i.e Euclidean distance: For point's xi and xk,

$$s(i,k)=-||xi-xk||2 \quad (1)$$

The two kinds of messages are exchanged between data points, and each takes into account a different kind of competition. Messages can be combined at any stage to decide which points are exemplars and, for every other point, which exemplar it belongs to. The ―availability‖ a(i, k) message is sent from candidate exemplar point j to point i and it reflects the accumulated evidence for how appropriate it would be for point i to choose point j as its exemplar. The responsibility‖ r(i, k) message is sent from data point i to candidate exemplar point j and it reflects the accumulated evidence for how well-suited point j is to serve as the exemplar for point i. At the beginning, the availabilities are initialized to zero: a(i ,j) = 0.Then, the responsibilities are computed using the rule

$$r(i,k) \leftarrow s(i,k) - \max\{ a(i,k') + s(i,k')\} \quad (2)$$

The availabilities are zero in the first iteration, r(i,k) is set to the input similarity between point i and k as its exemplar, minus the largest of the similarities between point i and other candidate exemplars. This update is data-driven and does not take into account how many other points favor each candidate exemplar. In later iterations, when some points are effectively assigned to other exemplars, their availabilities will drop below zero. These negative availabilities will decrease the effective values of some of the input similarities s(i,k′) , removing the corresponding candidate exemplars from competition. For k = i, the responsibility r(k,k) is set to the input preference that point k be chosen as an exemplar, s(k,k), minus the largest of the similarities between point i and all other candidate exemplars. This self-responsibility reflects that point k is an exemplar.

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 10, October 2015
Available at http://internationaljournalofresearch.org

a(i,k) ←min{0, r(k,k) + ∑ max{0, r(i' ,k)}} (3)

The availability a(i ,k) is set to the self responsibility r(k,k) plus the sum of the positive responsibilities candidate exemplar k receives from other points. For a good exemplar only the positive portions of incoming responsibilities are added to explain some data points well regardless of how poorly it explains other data points Negative self responsibilityr(k,k) indicates that point k is currently better suited as belonging to another exemplar rather than being an exemplar itself, the availability of point k as an exemplar can be increased if some other points have positive responsibilities for point k being their exemplar. The total sum is threshold to limit the influence of strong incoming positive responsibilities so that it cannot go above zero. The ―self-availability‖ a(k,k) is updated differently:

a(i,k) ← ∑ max { 0, r(i', k)} (4)

This message reflects that point k is an exemplar sent to candidate exemplar k from other points. The above update rules require only local and simple computations that are easily implemented in eq. (3) and messages need be exchanged between pairs of points with known similarities Availabilities and Responsibilities can be combined to identify exemplars at any point during affinity propagation. For point i, the value of k that maximizes a(i,k) + r(i,k) either identifies point i as an exemplar if k = i, or identifies the data point that is the exemplar for point i. After changes in the messages fall below a threshold , the message-passing procedure may be terminated after a fixed number of iterations. To avoid numerical oscillations that arise in some circumstances, it is important that they be damped when updating the messages. Each message is set to l times its value from the previous iteration plus 1 – λ times its prescribed updated value, where the damping factor l is between 0 and 1. Each iteration of affinity propagation consists of : 1. Updating all responsibilities given the availabilities. 2. Updating all availabilities given the responsibilities and 3. Combining availabilities and responsibilities to monitor the exemplar decisions and terminate the algorithm. 2.1.1 Disadvantages of Affinity Propagation 1.It is hard to know the value of the parameter preferences which can yield an optimal clustering solution. 2.When oscillations occur, AP cannot automatically eliminate them. 2.2 Seeds Affinity Propagation Seeds Affinity Propagation is based on AP method. The main new features of the new

Each iteration of affinity propagation consists of

1. Updating all responsibilities given the availabilities.

2. Updating all availabilities given the responsibilities and

3. Combining availabilities and responsibilities to monitor the exemplar decisions and terminate the algorithm.

III INCREMENTAL AP CLUSTERING

AP clustering has been successfully used in a series of problems, e.g., face recognition , fMRI data analysis , and document Clustering . However, most of the applications deal with static data. Incremental AP clustering is still a difficult problem. 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 timestep are at the different statuses, so it is not likely to find the correct exemplar set by simply continuing affinity propagation in this case.

Algorithm 1 IAPKM

Input: U*t*−1, **c***t*−1, **X***t*;

Output: c*t*;

Steps:

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 10, October 2015
Available at http://internationaljournalofresearch.org

1: Assign each new object to the current exemplars,and the label vector of all the newobjects is indicated by $\mathbf{c}*t-1$;

2: $\mathbf{U}t = \mathbf{U}t-1 \cup \mathbf{X}t$, $\mathbf{c}t = [\mathbf{c}t-1\ \mathbf{c}*t]$;

3: Message-passing continues according to equation (6) and equation (7);

4: Repeat Step 3 till convergence, $\mathbf{c}t$is saved.

Algorithm 1 presents IAPKM. Traditional AP clustering is implemented on the first batch of objects $\mathbf{U}t-1$, and the clustering result is $\mathbf{c}t-1$. When a new batch of objects $\mathbf{X}t$are arriving, assign each new object to the currentexemplars. Renew available data set to $\mathbf{U}t$, and renew label vector $\mathbf{c}t-1$ to $\mathbf{c}t$. Then *K*-Medoids is implemented to modify the clustering result till to the end.Additionally, *K*-Medoids can modify the clustering result quickly, which makes IAPKM efficient enough to be used in dynamic environment. A disadvantage of IAPKM is that the number of clusters can not be adjusted according to the new arriving objects. That's because that traditional *K*-Medoids can't adjust the number of clusters automatically. In next section, we will propose an IAP clustering algorithm according to the second strategy, where the number of clusters can be adjusted automatically.

IV.Partition Adaptive Affinity Propagation

Affinity propagation exhibits fast execution speed and finds clusters with low error rate when clustering sparsely related data but its values of parameters are fixed. Partition adaptive affinity propagation can automatically eliminate oscillations and adjust the values of parameters when rerunning affinity propagation procedure to yield optimal clustering results, with high execution speed and precision [20] The premise is that both AP and AAP are a message communication process between data points in a dense matrix. The time spent is in direct ratio to the number of iterations. During each iteration of AP, each element r(i, k) of the responsibility matrix must be calculated once and each calculation must be applied to N-1 elements, where N is the size of the input similarity matrix, according to Eq.(2).And each element of the availability matrix can be calculated in the same way. During an iteration of AAP, the convergent of K is detected but the execution speed is still much lower than AP. Partition adaptive affinity propagation (PAAP). This modified algorithm can eliminate oscillations by using the method of AAP. Our adaptive technique consists of two parts. One is called fine adjustment, another is coarse adjustment. Fine adjustment is used to decrease the values of parameter "preference "slowly, and coarse adjustment is used to rapidly decrease the values of preference correspondingly. The original similarity matrix is decomposed into sub-matrices to gain higher execution speed [21] [22], when executing our method. PAAP can yield optimal clustering solutions on both dense and sparse datasets. Assuming that Cmax is the expected maximal number of clusters, Cmin is the expected minimal number of clusters, and K (i) is the number of clusters in the iteration, and maxits is the maximal number of iterations. λstep and Pstep are the adaptive factors as in AAP. The PAAP algorithm goes as follows:

Algorithm PAAP algorithm:

1. Execute AP procedure, get the number of clusters:K(i).
2. If K(i) ≤ K(i + 1), then go to step 4. Else, count == 0, then go to step 3.
3. λ←λ + λstep , then go to step 1. If A > 0.85, then p← p+ pstep ,s(i,i) ←p, Else go to step 1.
4. If │Cmax - K(i) │ > CK, then Astep == ─20 * │K(i) - Cmin │Go to step 6. Else, delay 10 iterations and then go to step 5.
5. If K(i) ≤ K(i + 1), then count == cout + 1, Astep == count *Pstep. Go to step 6 or Else, go to step 1.
6. p == p + Astep, then s(i, i) ← p.

7. If i == maxits or K(i) ~ Cmin, the algorithm terminates. Else, go to step 1. PAAP can find the true or better number of clusters with high execution speed on dense or sparse datasets , meanwhile, it can automatically detect the number oscillations and eliminate them. This verified that both acceleration technique and partition technique are effective. If Kpart s and A step (acceleration factor) is well chosen, the average number of iteration can be reduced effectively.

4.1 Advantages of partition Adaptive affinity propagation.

1. PAAP improved approach based on affinity propagation. It can automatically escape from the oscillation and adjust values of parameters λ and p.

V. CONCLUSION

In this survey, various clustering approaches and algorithms in document clustering are described. . A new clustering algorithm which combines Affinity Propagation with semi supervised learning ,i.e Seeds Affinity Propagation algorithm is present. In comparison with the classical clustering algorithm k-means, SAP not only improves the accuracy and reduces the computing complexity of text clustering and but also effectively avoids being trapped in local minimum and random initialization. Whereas Incremental Affinity Propagation integrates AP algorithm and semi-supervised learning. The labeled data information is coded into similarity matrix. The Adaptive Affinity Propagation algorithm first computes the range of preferences, and then searches the space to find the value of preference which can generate the optimal clustering results compare to AP approach which cannot yield optimal clustering results because it sets preferences as the median of the similarities. The area of document clustering has many issues, which need to be solved. We hope , the paper gives interested readers a broad overview of the existing techniques. As a future work, improvement over the existing systems with better results which offer new information representation capabilities with different techniques can be attempted.

REFERENCES

[1]S.Deelersanduwatanamongkol,"Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance,‖ International Journal of Electrical and Computer Engineering 2:4 2007

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

[3] B.J. Frey and D. Dueck, "Non-Metric Affinity Propagation for Un-Supervised Image Categorization," Proc. 11th IEEE Int'l Conf. Computer Vision (ICCV '07), pp. 1-8, Oct. 2007.

[4] L. Michele, Sumedha, and W. Martin, "Clustering by Soft- Constraint Affinity Propagation Applications to GeneExpression Data," Bioinformatics, vol. 23, no. 20, pp. 2708-2715, Sept. 2007.

[5] T.Y. Jiang and A. Tuzhilin, "Dynamic Micro Targeting: Fitness- Based Approach to Predicting Individual Preferences," Proc. Seventh IEEE Int'l Conf. Data Mining (ICDM '07), pp. 173-182, Oct. 2007.

[6] F. Sebastiani, ―Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, pp. 1-47, 2002

[7] F. Wang and C.S. Zhang, "Label Propagation through Linear Neighbourhoods," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 1, pp. 55-67, Jan. 2008.

[8] Z.H. Zhou and M. Li, "Semi-Supervised Regression with Co- Training Style Algorithms," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 11, pp. 1479-1493, Aug. 2007.

[9] A. Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with CoTraining,‖ Proc. 11th Ann. Conf. Computational Learning Theory, pp. 92- 100, 1998.

[10] Z.H. Zhou, D.C. Zhan, and Q. Yang, "Semi-Supervised Learning with Very Few Labeled Training Examples," Proc. 22nd AAAI Conf. Artificial Intelligence, pp. 675-680, 2007

[11] Renchu Guan, Xiaohu Shi, Maurizio Marchese, Chen Yang, and Yanchun Liang ―Text Clustering with Seeds Affinity Propagtaion" IEEE Transactions on Knowledge and data Engineering , VOL. 23, NO. 4, APRIL 2011

[12] H.F. Ma, X.H. Fan, and J. Chen, "An Incremental Chinese Text Classification Algorithm Based on Quick Clustering," Proc. 2008 Int'l Symp. Information Processing (ISIP '08), pp. 308- 312, May 2008.

[13] Y. Xiao, and J. Yu, "Semi-Supervised Clustering Based on Affinity Propagation, " Journal of Software, Vol. 19, No. 11, November 2008, pp. 2803-2813.

[14] C. J. van Rijsbergen, Information Retrieval, 2nd edition, Butterworth, London, pp. 23- 28, 1979.

[15] K.J. Wang, J.Y. Zhang, D. Li, X.N. Zhang, and T. Guo, "Adaptive Affinity Propagation Clustering," ActaAutomaticaSinica, vol. 33, no. 12, pp. 1242-1246, Dec. 2007

[16] FAQ of Affinity Propagation Clustering: http://www.psi.toronto.edu/affinitypropagation/faq.html

[17] K.J. Wang, J.Y..Zhang, and D. Li. "Adaptive Affinity Propagation Clustering." Acta Automatic Sinica, 33(12): 1242-1246, 2007

[18] P.J. Rousseeuw, Silhouettes: "a graphical aid to the interpretation and validation of cluster analysis", Computational and Applied Mathematics. (20),53-65, 1987

[19] S. Dudoit, J. Fridlyand. "A predictionbased resampling method for estimating the number of clusters in a dataset". Genome Biology,3(7): 0036.1-0036.21, 2002

[20] Changyin Sun, Chenghong Wang, Su Song,Yifan Wang "A Local Approach of Adaptive Affinity Propagation Clustering for Large Scale Data" Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, June 14- 19, 2009

[21] Guha, S., Rastogi, R., Shim, K., "CURE: an efficient clustering algorithm for large databases," Inf.Syst., 26(1): 35-58, 2001.

[22] Ding-yin Xia, Fei Wu, Xu-qing Zhang, Yue-ting Zhuang, " Local and global approaches of affinity propagation clustering for large scale data," J Zhejiang UnivSCi A, , pp.1373 1381,2008.

Guide Profile:

Mr.P.KRISHNA CHAITANYA working as

Asst Professor in Department of Computer Science and Engineering at MalineniLakshmaiah Engineering College(MLEC), Singarayakonda. He completed his M.E in year 2010.

Student Profile:

Mr.P.ANIL was born in AndhraPradesh,India. He receivedB.tech Degree from JNTU Kakinada ,Malinenilakshmaiah Engineering college prakasam district I am pursuing M.TechDegree in CSE from JNTU Kakinada