International Journal of Database Theory and Application
Vol.9, No.6 (2016), pp.239-246
http://dx.doi.org/10.14257/ijdta.2016.9.6.24

Stability Threshold-based Affinity Propagation and Its Application

Limin Wang[1,2], Yizhang Wang[1,2], Xuming Han[3*]and Qiang Ji[1,2]

[1]School of Management science and information engineering, Jilin University of finance and economics, Jilin, 130117, China
[2]Jilin Province Key Laboratory of Internet Fintech, Jilin University of Finance and Economics, Changchun 130117, China
[3]School of Computer Science and Engineering Changchun University of Technology,
Jilin, 130117, China
**Corresponding author:hanxvming@163.com*

Abstract

Given the performance of original affinity propagation algorithm is greatly affected by preference (P), stability threshold-based affinity propagation clustering algorithm (STAP) is proposed in this paper, including stability threshold to obtain the state of convergence when getting real class number and capture the corresponding P, and it take S-type function as damping factor to accelerate the convergence speed of STAP clustering algorithm. Besides it is successfully applied in the financial evaluation of public companies. The simulation experimental results show that, comparing the traditional affinity propagation clustering algorithm, STAP clustering algorithm can obtain high precision and fast convergence rate to improve clustering performance.

Keywords: *Affinity propagation; Stability threshold; Convergence factor*

1. Introduction

Affinity propagation is a new clustering algorithm appeared in the journal Science by Frey in 2007. Currently, the algorithm has been successfully applied to image segmentation [1-3], image search [4], gene identification [5-6], text clustering [7], determining the optimal air route [8] and so on. In recent years, many scholars have proposed a variety of improved methods, for instance, Givoni proposed that affinity propagation was coupled with hierarchical clustering applied in dynamic simulation of HIV mutant strains [9]. Qasim proposes that affinity propagation was combined with semi-automatic text concept maps [10]. Jihong Yu proposed that using the vector space model to calculate similarity applied to three-dimensional model ships for uniform view space projection [11]. Wenshuai Wang proposed a new data stream clustering along with affinity propagation called SAPStream algorithm.

We propose an improved affinity propagation using stability threshold to optimize parameter P more accurately，and accelerate the convergence speed with S-type function as convergence factor. Then we introduce STAP index to the financial evaluation in the field of public companies for improvement of stock investment.

2. Affinity Propagation Clustering Algorithm

Affinity propagation takes as input a collection of real-valued similarities between data points. For points x_i and x_k , $s(i,k) = -\|x_i - x_k\|$ [11]. Later, the median of similarities is chosen as shared values–this value can be varied to produce different numbers of clusters,

ISSN: 2005-4270 IJDTA

we call it preferences (*P*) [12]. Then, there are two kind of messages exchanged deciding which points are exemplars, and which point it belongs to [13]. For (1), the "responsibility", sent from data point i to candidate exemplar point *k*, reflects the accumulated evidence for how well-suited point k is to serve as the exemplar for point i. The "availability", sent from candidate exemplar point k to i, reflects the accumulated evidence for how appropriate it would be for point i to choose k as its exemplar in (2) [14]. To limit the influence of strong incoming positive responsibility, the "self-availability" $a(i,k)$ is updated as (3) [15]. For point i, the value of k that maximizes $a(i,k)+r(i,k)$ either identifies point i as an exemplar. When updating the messages [16], λ (damping factor) is set to times its value of previous value plus $1-\lambda$ times updated iteration in (4) and (5). All of it did not change until 10 iterations [17].

$$r(i,k)=s(i,k)-\max_{k's.t.k'\neq k}\left\{a(i,k')+s(i,k')\right\} \tag{1}$$

$$a(i,k)=\min\left\{0,r(k,k)+\sum_{i's.t.i'\notin\{i,k\}}\max\left\{0,r(i',k)\right\}\right\} \tag{2}$$

$$a(i,k)=\sum_{i's.t.i'\neq k}\max\left\{0,r(i',k)\right\} \tag{3}$$

$$r(i,k)^{(t+1)}=\lambda r(i,k)^{t}+(1-\lambda)r(i,k)^{(t-1)} \tag{4}$$

$$a(i,k)^{(t+1)}=\lambda a(i,k)^{t}+(1-\lambda)a(i,k)^{(t-1)} \tag{5}$$

3. Stability Threshold-Based Affinity Propagation Clustering Algorithm

Traditional clustering is achieved upon the whole set of attributes or variables, and therefore only capable of discovering global information [18]. Nevertheless, affinity propagation clustering algorithm has many advantages include identifying outliers, but it cannot obtain optimal preferences. We take some methods to improve it as follow.

3.1 The Optimization of Preferences

Table 1 shows results under different preferences to analyze the impact that preferences on clustering performance. Table 1 demonstrates that preferences have great impact on Silhouette (*Sil*), number of clusters (*NC*), time of algorithm. Choosing adaptive preference may gain optimal number of clustering. STAP algorithm can solve the optimization of preferences to search the class space more accurately.

Table 1. Experimental Results of Wine by AP under Different *P* Value

P	Time	Iterations	*Sil*	*NC*
1	0.255050	89	-0.3544	12
2	0.181141	117	-0.2727	11
3	0.218614	166	-0.2319	7
4	0.290893	256	-0.243	8
5	0.39452	376	-0.1825	6
6	0.657784	689	-0.3172	6
…	…	…	…	…

16	7.862692	9367	0.0678	3

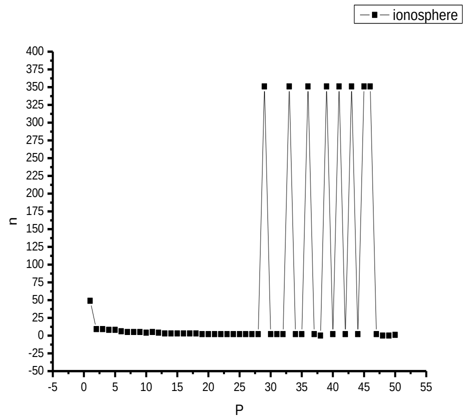


Figure 1. *NC* of Ionosphere under Different *P* Value

Figure 1 shows that *NC* of the Ionosphere increases as preferences change in three stages, clustering number in state1 repeats rarely and convergence of it is quick called convergence phase. State 2, which is named stabilization phase, gets better *NC* and accuracy, results from state 3 distort. In the experiment of large number of data sets can be found that the characteristic in three stages.

We present a technology of preferences optimization that using stability threshold to measure steady state of iteration and quantifying it through mathematical modeling. Stability threshold is the repetition degree of clustering results. The larger the repetition degree is, the more stable the algorithm iteration. At this time it can achieve excellent clustering performance under the corresponding preference in the class space.

Selection of stability threshold is very important, the sample and dimensions of data sets have great influence on the clustering results of AP, Therefore putting forward a kind of linear function about sample and dimensions as a limit of stable threshold, *SN* is sample, *Dim* is dimension:

$$\lim_{ST \to +\infty} ST = \frac{SN}{n \cdot Dim} \tag{6}$$

The value of the stability threshold is mainly determined by the sample and dimension. The greater the number of samples and the smaller the dimension are, the greater the degree of dynamic search is, and vice is true. This complies with the principle of the AP algorithm clustering. *n* is proportional to ratio of the sample and dimension, the degree of fitting of them is better when default *n* = 5.

Linear function proposed in this paper about the stability factor, shock factor and weak stability factor as a stable threshold function. Stability factor is the clustering results of repetitions, shock factor is the clustering results of oscillation frequency, and weak stability factor is decreasing number of the step length is one. Weak stability factor is used to more accurately get the steady state of data sets with fewer shocks. *SF* means stable factor, *CF* is shock factor, and *WSF* means weak stability factor:

$$ST = SF - CF - WSF \tag{7}$$

3.2. The Accelerating Technology with S-type Function

Given the convergence rate of AP for large data and high-dimensional data is slow, we propose a method to accelerate technology, which accelerates technology with S-type convergence factor. For AP clustering algorithm, the larger the damping coefficient is, the slower convergence speed is [19]. Appropriate damping coefficient reduces the algorithm running time. Default value: a=1, b=1, d=1, net=4:

$$f(net) = a + b / (1 + \exp(-d \cdot net)) \tag{8}$$

In this paper, the STAP algorithm clustering algorithm is as follows:

(1) Enter the similarity matrix $S\,(i, j)$.

(2) Searching the preference, step length is 1. Initial P value of similarity matrix is the median, $i \neq j$.

(3) Updating ST, SF, CF, and WSF.

$$r(i,k)^{(t+1)} = f(net) \cdot \lambda \cdot r(i,k)^{(t)} + (1-\lambda) \cdot r(i,k)^{(t-1)}$$

$$a(i,k)^{(t+1)} = f(net) \cdot \lambda \cdot a(i,k)^{(t)} + (1-\lambda) \cdot a(i,k)^{(t-1)}$$

(4) Breaking when $ST = SN / (n \cdot Dim)$

4. Experimental Results and Analysis

Experiments of AP and STAP are in a same computer (Pentium G645 2.9 GHz CPU, 4GB). The damping coefficient of the parameters is λ=0.5, the algorithm selects the evaluation index of Silhouette-effective clustering evaluation method. As shown in table 2, different sample of data sets including high and low dimension ensure comprehensiveness and effectiveness of the simulation experiment.

Table 2. Characteristic Parameters of Experimental Data Sets

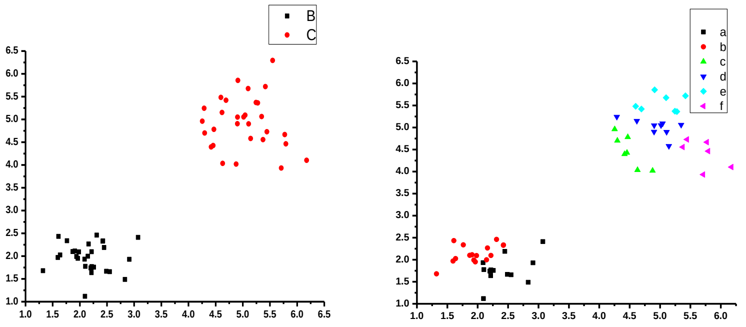
Data sets from UCI	*Samples*	*Dimensions*
Wine	178	13
Iris	150	4
Ionosphere	351	34
Seeds	210	7
Harberman	306	3
Cmc	1473	10

The affinity propagation clustering algorithm is, generally speaking, extremely effective to low dimensional datasets and cannot work on the increasing high dimensional datasets in different areas effectively. Algorithm time is too long, moreover, we uses S-type function to select the appropriate value to adjust damping coefficient and lessen the time, the experimental results as shown in following table 3, wine, iris, seeds, harberman are low dimension and less samples, the difference is very small, Ionosphere which is high-dimensional algorithm reduces time by 73.3%, in addition, the Cmc reduces time by 11%. This shows that when the sample and dimension are bigger, the algorithm running time is less, this technique is effective.

Table 3. Comparison of Time by AP and the Improved AP Algorithm

Data sets from UCI	time of AP	time of AP with S-type function
Wine	0.156582	0.150809
Ionosphere	57.204761	15.277670
Cmc	20.614013	18.342031
Harberman	0.703635	0.678900
Iris	0.121239	0.119951
Seeds	0.212192	0.206391

As Figure 2 seen, we use artificial data set to describe AP and STAP algorithm clustering performance. Artificial data set is automatically generated by the binomial distribution random data through Matlab, AP clustering result is 6 classes, and STAP is 2, STAP clustering results are closer to actual situation of the data, the recognition rate is 100%. This suggests that the STAP clustering algorithm has obvious performance.



(a) Improved AP (b) AP

Figure 2. Results of the Improved AP Clustering Algorithm

Table 4. Clustering Results between AP and STA-AP

Data sets from UCI	AP *sil*	*NC* of AP	STA-AP *Sil*	STA-AP *NC*	Known *NC*
Wine	-0.3544	12	0.3562	3	3
Iris	-0.1894	12	0.4542	3	3
Ionosphere	No results	49	-0.0092	2	2
Seeds	-0.2594	17	0.3946	3	3
Harberman	-0.2668	31	0.3959	2	2

Table 4 shows that the proposed STAP clustering algorithm is compared with the original algorithm. The Silhouette increases significantly, Wine, Iris can achieve real class number. We improve the performance of AP clustering and realize the optimization of preferences. But the algorithm also has its own shortcomings, because the stability threshold is based on the algorithm of iteration and the ratio of the sample and dimension, stability without appearing in the process of convergence is not applicable of the algorithm.

5. Experimental Results and Analysis

STAP clustering algorithm was applied to 2013 in the third quarter of 98 public company financial index evaluation in this paper, we put forward the concept of STAP clustering index to measure the stock in the status of public companies in the industry. Attributes select average earnings per share, average per share capital reserve fund, average undistributed profit per share, average return on net assets and average net assets per share. All the data is from the tidal wave net.

Table 5. The Third Quarter of 2013 Financial Index of Public Companies

public company	Average earnings per share	average per share capital reserve fund	average undistributed profit per share	average return on net assets	Average net assets per share
002285	0.48	1.3155	1.09	12.81	3.76
002146	0.93	0.7711	2.72	17.92	5.18
000671	0.37	0.2526	1.32	14.09	2.6
600173	0.098	0.1077	0.79	4.82	2.03
600208	0.046	0.1797	0.72	2.34	1.98
600383	0.16	1.3094	2.7	3.01	5.32
…	…	…	…	…	…
600767	-0.083	0.1561	-0.09	-7.67	1.09
000502	0.037	0.1233	-0.08	3.4	1.08

Running STAP clustering algorithm, 98 public companies are divided into 2, STAP clustering index level one has 24 and level two has 74.

Table 6. Clustering Results of STAP on Public Companies

STAP index	Average earnings per share	average per share capital reserve fund	average undistributed profit per share	average return on net assets	average net assets per share
Level one	0.546	0.816	2.247	12.863	4.45
Level two	0.099	0.879	1.084	0.084	3.161

Table 6 shows that the level one’s average earnings per share is 5 times as much as the secondary level, the rate of average net assets per share is 153. The differences of two important indexes illustrates STAP clustering algorithm’s effect is remarkable. STAP clustering index can effectively evaluate the performance of optimal stock and can serve as an important parameter for investment. Investors should pay close attention to first level. Secondary enterprises need to adjust the strategy and promote the further development.

6. Conclusions

Original affinity propagation clustering algorithm can improve the performance through adjusting the preferences, but you cannot know the actual preference in advance. Stability threshold-based affinity propagation clustering algorithm uses the method of stability threshold to find stable state of *NC* space and the homologous preference. At the same time the S-type function is introduced into the algorithm as a convergence factor to improve convergence for single cycle. Combination of preferences optimization and accelerating technology make it better. In addition, STAP algorithm for evaluation of financial indicators of public companies also gets satisfactory results. We need further research when data structure is complicated like multi-manifold.

Acknowledgements

The research was supported by the National Science Foundation of China under grant No. 61202306, 61472049, 61402193 and 61572225, the National Society Science Foundation of China under grant No. 15BGL090, the Ministry of Education, the Foundation under grant No. 13YJAZH130, the Application Basis Foundation of Jilin Provincial Science & Technology Department under Grant No. 201215119, 20130522177JH and 2013s0101072JC, the Society Science Research Foundation of Jilin Province under Grant No. 2014B166, the support project on the excellent talent of Jilin Province in colleges and universities under grant No. 2014159.

References

[1] D. Napoleon, M. Praneesh, M. Subramanian, and S. Manhattan, “Distance Based Affinity Propagation Technique for Clustering in Remote Sensing Images”, International Journal, vol. 2, no. 3, **(2012)**, pp. 327-329.

[2] J. Tang, Q. Chen and M. Wang, “Towards optimizing human labeling for interactive image tagging”, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), vol. 9, no. 4, **(2013)**, pp. 29.

[3] S. Ding, G. Ma and Z. Shi, “A novel self-adaptive extreme learning machine based on affinity propagation for radial basis function neural network”, Neural Computing and Applications, **(2013)**, pp. 1-9.

[4] L. Xie, Q. Tian and W. Zhou, “Fast and accurate near-duplicate image search with affinity propagation on the Image Web”, Computer Vision and Image Understanding, vol. 124, **(2014)**, pp. 31-41.

[5] C. Yang, S. Liu and L. Bruzzone, “A Feature-Metric-Based Affinity Propagation Technique for Feature Selection in Hyper spectral Image Classification”, **(2013)**.

[6] R. Akulenko and V. Helms, “DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples”, Human molecular genetics, **(2013)**.

[7] L. Xie, Q. Tian and W. Zhou, “Fast and Accurate Near-Duplicate Image Search with Affinity Propagation on the Image Web”, Computer Vision and Image Understanding, **(2014)**.

[8] H. Zhu, S. Ding and H. Zhao, “Attribute Granulation Based on Attribute Discernibility and AP Algorithm”, Journal of Software, vol. 8, no. 4, **(2013)**, pp. 834-841.

[9] I. Qasim, J. W. Jeong and J. U. Heu, “Concept map construction from text documents using affinity propagation”, Journal of Information Science, **(2013)**.

[10] J. Yu, X. Bai and J. Lv, “Improved similarity affinity propagation”, Mini-Micro Systems, vol. 34, no. 003, **(2013)**, pp. 603-604.

[11] W. Wang and G. Chen, “Semi-supervised affinity propagation-based data stream clustering algorithm”, Computer Engineering and Applications, vol. 49, no. 8, **(2013)**, pp. 7-8.

[12] N. M. Arzeno and H. Vikalo, “Semi-Supervised Affinity Propagation with Soft Instance-Level Constraints”, Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 37, no. 5, **(2015)**, pp. 1041-1052.

[13] Y. Kokkinos and K. G. Margaritis, “Confidence ratio affinity propagation in ensemble selection of neural network classifiers for distributed privacy-preserving data mining”, Neurocomputing, vol. 150, **(2015)**, pp. 513-528.

[14] A. Sakellariou, D. Sanoudou and G. Spyrou, "Combining multiple hypothesis testing and affinity propagation clustering leads to accurate, robust and sample size independent classification on gene expression data", BMC bioinformatics, vol. 13, no. 1, **(2012)**, pp. 270.

[15] V. A. Vakorin, A. R. McIntosh and B. Mišić, "Exploring Age-Related Changes in Dynamical Non-Stationary in Electroencephalographic Signals during Early Adolescence", PloS one, vol. 8, no. 3, **(2013)**.

[16] J. Nie, G. Li and D. Shen, "Development of cortical anatomical properties from early childhood to early adulthood", Neuroimaging, **(2013)**.

[17] J. Zhao, M. Ma and R. Wang, "A Novel Image Retrieval Algorithm Based on K-Neighbor Semi-Supervised Affinity Propagation Algorithm",Journal of Computational and Theoretical Nano science, vol. 10, no. 9, **(2013)**, pp. 2282-2287.

[18] W. Chengduan, "Applying Bi-clustering Algorithm in Customer Segmentation for High-Value Customers", International Journal of Database Theory and Application, **(2015)**, pp.39-46.

[19] F. T. Fontbona, V. Muñoz and B. Lopez, "Solving large immobile location-allocation by affinity propagation and simulated annealing application to select which sporting event to watch. Expert Systems with Applications", **(2013)**, pp. 4593–4599.

Authors

Limin Wang, Female, born in 1975. Received her Ph.D. degree in ma Jilin University in 2007. Now she is a professor, membership of China Computer Federation. Recently several years, she has published more than 60 papers. Her current research interests include machine learning, data mining and finance engineering.

Yizhang Wang, Male, born in 1990, M.S degree candidate in Jilin University of finance and economics. He has presented 3 papers in domestic and international journals of Data Mining and Machine Learning. His research interests include machine learning, data mining and finance engineering

Xuming Han, Male, born in 1971. Received his Ph.D. degree in Jilin University. Now she is a professor, membership of China Computer Federation, his main research interests include machine learning, data mining, evolutionary algorithms and intelligent computation.

Qiang Ji, Male, born in 1989, M.S degree candidate in Jilin University of Finance and Economics, his main research interests including data mining, machine learning and intelligent computation.