5. Analysis and Results

Algorithm in this paper are based on similarity matrix S. The measure of similarity between two objects is also an important problem in data mining. In this paper, negative square root of Euclidean distance is adopted. A problem of Euclidean distance is that some features with large amplitude often cover the effect of other features. Therefore, a preprocessing is used to normalize the original data set

$$x_i^j = \frac{x_i^j - x_{min}^j}{X_{\max}^j - x_{min}^j} \quad (7)$$

Where x_i^j is the j^{th} feature of object *i* and $X_{\max}^j = \max(x_1^j, x_2^j, \ldots x_{Mt}^j)$, $X_{\min}^j = \min(x_1^j, x_2^j, \ldots x_{Mt}^j)$. s(i, j) gives similarity between object *i* and object *j*. It is defined as

$$s(i, j) = -\sqrt{\|x_i - x_j\|^2} \quad (8)$$

Another parameter needs to be specified is the preference *p*. Generally, larger preference p generates larger number of clusters. Frey et al. suggests it should be the median, or minimum value of similarities[5]. Therefore, the following procedure is employed to determine the value of p

$$p = \min_{i,j}[s(i,j)] - pc \cdot Mt \quad (9)$$

Where Mt is the number of current available objects. pc, short for preference coefficient, is a constant determined by the initial batch of objects. Varying pc and running traditional AP clustering on the first batch of objects, when the number of exemplars is proper, the corresponding value of pc is stored and used in the following incremental clustering.

Accuracy is a measure to state the effectiveness of the clustering algirithm. It is calculated as:

$$Accu = \frac{\sum_{i=1}^{n} \delta\left(c_i, \text{map}\left(\bar{c}_i\right)\right)}{n} \quad (10)$$

where c_i is the real label of object *i*, and where $\bar{c}_i$ is the real label of object *i*, and $\bar{c}_i$ is the obtained clustering label. $\delta(i, j) = 1$, if $i = j$; $\delta(i, j) = 0$, otherwise. Function map() matches true class label and the obtained cluster label.

Experiment results on the four unlabled data sets are shown in Table 2

Table 2: Comparison of Accuracy

Data Set	Method	First	Second
CAR	IAPSDC	0.74	0.74
	IAPKM	0.30	0.30
	IAPNA	0.61	0.61
WINE	IAPSDC	0.91	0.90
	IAPKM	0.91	0.91
	IAPNA	0.90	0.90
WDBC	IAPSDC	0.89	0.89
	IAPKM	0.89	0.89
	IAPNA	0.89	0.90

6. Conclusion

In this paper we have proposed a clustering algorithm IAPSDC to use Incremental Affinity Propagation for the streamed data. IAPSDC when compared to the AP, IAPKM and IAPNA give comparable results. Two popular unlabeled datasets are used to evaluate the IAPSDC. Results of the experiments show the effectiveness of the IAPSDC. The proposition IAPSDC is inspired by the combination of Incremental Affinity Propagation and Streamed Data Clustering. Affinity propagation is very effective in finding out the initial exemplar set which can later be used to cluster the streamed data points coming as an ordered sequential data. Streamed clustering is a branch of incremental data clustering. Some other incremental clustering problems are also of great importance.

Additionally, some other problems such as how to measure similarity between objects, and how to extract features from time series and labelled data set are also of great importance. However, that is not the focus of the paper. It can be a future scope of this paper.

References

[1] Leilei Sun, Chonghui Guo, "Incremental Affinity Propagation Clustering Based on Message Passing," IEEE Transactions On Knowledge And Data Engineering Vol:Pp No:99 Year 2014
[2] X. Zhang, C. Furtlehner, and M. Sebag, "Frugal and Online Affinity Propagation," Proc. Conf. francophone sur l'Apprentissage (CAP '08), 2008.
[3] J. Beringer and E. Hullermeier, "Online Clustering of Parallel Data Streams," Data and Knowledge Engineering, vol. 58, no. 2, pp. 180-204, Aug. 2006.
[4] J. Beringer and E. Hullermeier, "Online Clustering of Parallel Data Streams," Data and Knowledge Engineering, vol. 58, no. 2, pp. 180-204, Aug. 2006.
[5] B.J. Frey and D. Dueck, "Response to Comment on 'Clustering by
[6] 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] 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.
[9] L. Ott and F. Ramos, "Unsupervised Incremental Learning for Long-term Autonomy," Proc. 2012 IEEE Int. Conf. Robotics and Automation (ICRA '12), pp. 4022-4029, May 2012,
[10] Adil M. Bagirov, Julien Ugon, Dean Webb, " Fast modifiedglobal k-means algorithmforincrementalclusterconstruction," Pattern Recognition 44 (2011) 866–876
[11] A.K. Jain, "Data Clustering: 50 Years Beyond K-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651-666, June 2009.