Improvements in Clustering Using Affinity Propagation: A Review

Aniket Bombatkar[1], Thaksen Parvat[1]
[1]Dept. of Computer Engineering
Sinhgad Institute of Technology Lonavala
S.P.P.U. Pune, India
aniketb778@gmail.com, pthaksen.sit@sinhgad.edu

ABSTRACT: *Data Analysis is getting more and more important in Today's world. If data of a person gets lost then, it will be a loss of his/her identity so maintain a big or huge amount of data without loosing, is a challenge. Data clustering can do that because it can maintain a huge amount of data by dividing it into various clusters.*

Clustering is separation of similar data from dissimilar content. This paper reviews the data clustering techniques used to separate large amount of data. Motivation for this paper comes from an enormous amount of data in clusters and finding a small errors in them. The review shows that Incremental clustering is becoming a significant problem because the data is generating in an enormous amount. The data can be of various types like it could be in the form of text, spatial data, images, sequence data, data in the form of streams, multimedia data. To manage such kinds of data, data mining has various techniques. Affinity Propagation (AP) is one of the methods that has use in much incremental clustering problems. Most recent approaches reduce the data content by various methods such as compressed model that combines horizontal compression with vertical compression. The review covers incremental clustering, Affinity propagation to maintain a large amount of data and other methods regarding clustering. The analysis shows a trend towards reducing a data into small size so the data analyzation will become accessible. Data is information and to take out the knowledge from that information is important, so that mining of data will become efficient.

Keywords: Incremental Clustering, Affinity Propagation, Data Mining, Intrusion Detection

Received: 26 February 2015, Revised 20 March 2015, Accepted 27 March 2015

1. Introduction

With the increasing amount of data, the management of data is becoming more and more difficult. So to manage that data, data analysts are using different techniques of data mining so that to mine that data will be feasible. Data mining is an approach to discovering interesting patterns from a large amount of data. Clustering is making a group of a set of objects into multiple clusters. In clustering the objects that are in the cluster should have the same similarity, and they should be very dissimilar to the objects in another cluster.

Partitioning methods can do clustering; Hierarchical methods, Density-based methods, Grid-based clustering, these are some methods that used in clustering However it entirely depends on how to cluster data. To find mutually exclusive clustering of

unique shape partitioning method is useful. To store data at multiple levels, consider Hierarchical methods. Density-based clusters can find arbitrarily shaped clusters and many more. Such as we can use grid based methods for fast processing time, it depends on the grid size.

1.1 Intrusion Detection System

The larger and larger amount of data is generating because of the network communication. The social media, social networking sites making a tremendous amount of profit through all the communication. The management of data is also creating it is importance and with the increasing amount of network communication the risk is also increasing, so various techniques of Intrusion detection system(IDS) are coming to the market. Intrusion detection systems can be host-based intrusion detection system (HIDS), or it could be Network intrusion detection system (NIDS). HIDS monitors log files and system calls and NIDS monitors network.

Abnormal behaviors should be discovered at an initial stage therefore research on anomaly detection is going on rapidly than traditional signature based detection. Training data should be secure by IDS because it is on a large scale, so an effective technique to minimize time cost for detection is to compress the volume of the model. To compress the volume of model Chen et.al. [1] Proposed a compressed model that combines horizontal compression with vertical compression. To compress large training data, this model plays a significant role of compressing data.

Intrusion detection system is an important research topic with many potential applications. Because of the larger amount of communication, the Ids design is getting much more complicated.[1] To discover abnormal behaviors at it is an early stage the anomaly detection method can used, because anomaly detection is more modern and efficient than signature based detection. Anomaly detection has the advantage to detect unknown attacks because it uses heuristic learning on historical training data. To find out a systematic solution on training data the compressed model is used.[1][22] The compressed model is useful for efficiently and effectively handling the problems.

There are some attacks that are harmful to our system. The attacks are as follows –

• Probing	In probing the attacker attacks on system vulnerability for that it scans the whole system that is probe attack Ipsweep, portsweep, and nmap are the attack types of probing. Eg: port scanning
• Denial of Service (DOS)	The user who is unauthorized to access that system try to access the system remotely such as guessing password, The DOS attacks can access the system and make changes to them Process table, land, and Apache are the attacks of Denial of Service (DOS).
• Remote to Local (R2L)	Attacker who wants to access the system but don't have an account on that system (victim's machine) so he tries to gain access to that machine that is r2l attack. Multihop, guess_password are some attack types of r2l.
•User to Root (U2R)	In the user to root, the attacker is having a local access such as an employ account, and he tries to get access to privileges of that company. Rootkit, buffer_overflow are some of the attack types of U2R.

1.2 Incremental Clustering

Clustering is making a set of similar data points and for that we have to define a measure of similarity. Clustering is a solution to a problem of large training data. However, the new issue of the incremental cluster regarding clustering is arised, Because of the network communication and other uses the data that is generating a massive amount, so the cluster size is also increasing. Most of the applications in clustering deals with static data [2]. The solution to the problem of incremental clustering is necessary. Apply Affinity Propagation (AP) in incremental clustering Problems [2].

Clustering high dimensional data can be a challenge because conducting a cluster analysis on high dimensional is time consuming. The dimensionality of the data is not high means it is having less than ten attributes with the clusters are incrementing, and the dimensionality is also becoming high. To reduce the dimensionality of the data dimensionality reduction methods and subspace clustering methods are useful. Managing low dimensional data is not a problem but with incremental clustering the problem of high dimensional data is arising.

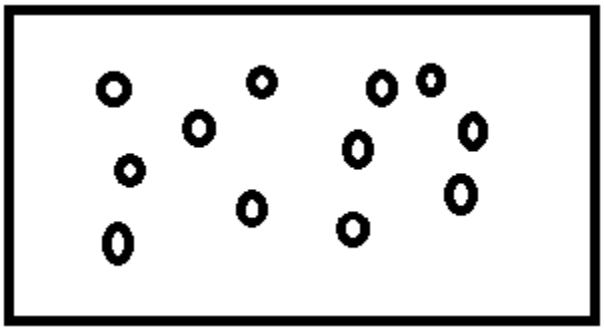

Figure 1. Data nodes

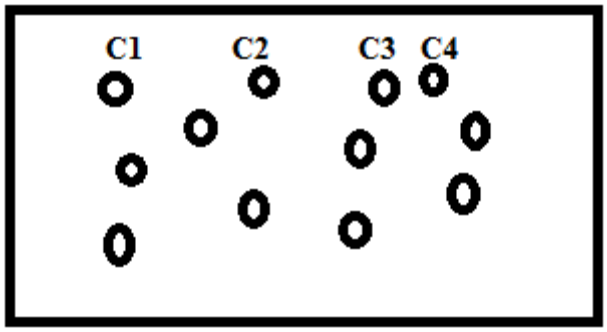

Figure 2. Clusters

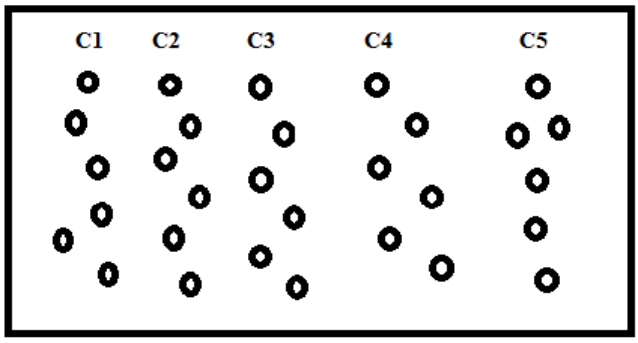

Figure 3. Incremental clustering

The data is increasing and with the data the clusters are also increasing and the cluster size grows too. The data points that are not similar to any data points and which are far from others they include as outliers. Data streaming is also large scale clusters that stores data of living streaming. The data of streaming is also increasing. Zhang et.al. [3]. applied Affinity Propagation(AP) make clusters.

1.3 Aims and Overview

This paper reviews some incremental clustering problems that are coming to the researcher. There are some solutions found on the issue of incremental clustering. AP is a better algorithm for clustering problems.

Chen et.al. employed Affinity Propagation as vertical compression to select small exemplars from training data in a large

amount[1]. Sun & Guo applied Affinity Propagation (AP) in incremental clustering problems [2]. Zhang et.al. also applied AP on Data Stream Clustering to Cluster the data[3]. Zhang et.al. have done analysis of Functional magnetic resonance imaging(fMRI) data using AP and integrated principal component analysis(PCA)[4]. On the same technology, AP Clustering in Multispectral images is done by Yang et al.

2. Related Work

2.1 AP in Vertical Compresssion

To achieve high efficiency of classification in IDS Chen et. al. [1] proposed a compressed model. In compressed model, OneR classifier is used for horizontal compression and Affinity propagation (AP) employed as vertical compression. AP used in many clustering problems. Sun & Guo [2] 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 [1]. Chen et. al. [1] improved efficiency of the model with the increasing accuracy so that the model will detect intrusions. Compressed model is built using training data.

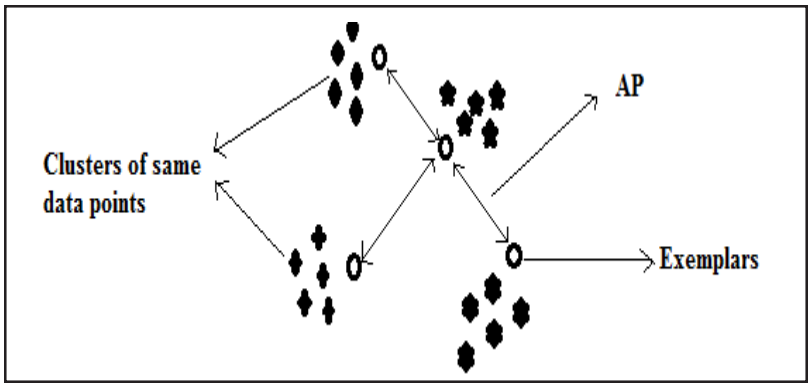

Figure 4. Exemplars and Clusters

The classification by use of compressed model can speed up detection procedure efficiently[1].Compressed model that based on training dataset considered SVM and KNN methods to identify parameters of AP for model of compression[1]. The compressed model runs 184 times faster than the traditional model without model compression [1]. Affinity propagation is a way better algorithm in data stream clustering, AP is used to cluster the data with the best representatives. AP gives guarantee of clustering optimality to select exemplars [1][2][3]. AP clusters multispectral images [4].

MapReduce parallelization approach compress the large scale of training data. The compression of data using MapReduce can do the data compression. For applying MapReduce, first MapReduce parallel processing for OneR compresses training data, then MapReduce parallel processing for AP compresses the reversed transposed matrix and in the end the compression takes place. Chen et. al. [1] proposed a compressed model and checked the results using KDD99 and CDMC2012 dataset on KNN and SVM classifiers.

2.2 Incremental Clustering of AP

Clustering is a main topic in data mining. However, with clustering the problem of incremental is arising now a days [3][4]. The figure 1,2,3 gives the idea of incremental clustering. Clustering algorithms are discovered and designed to identify patterns in static data [2] However Vedic Mathematics offers an entirely unique and a new approach for pattern recognition [7].

Sun & Guo [2] have extended a clustering algorithm AP to handle dynamic data. Affinity Propagation (AP) is an exemplar based method. In AP, exemplars are recognized by passing the messages on a bipartite graph. There is a difficulty in expanding AP in dynamic clustering of data; that is the objects that are established have a particular relationship to each other. After AP, While new relationships of objects are at the initial level objects add at different statuses and different time, so it was hard to get proper exemplary set by continuing the AP.

Sun & Guo discussed the problem for that they have proposed strategies. They have proposed IAP clustering algorithms that based on message passing framework.

Sl. no	Author Name	Problem	Solution	Algorithm
1.	Compressed model (Chen et.al. 2014)	Intrusion detection in large training data was time consuming	Compressed model (Horizontal compression+Vertical compression).	
2.	Incremental AP clustering (Sun & Guo 2014)	Incremental clustering	Affinity Propagation (AP) is Applied.	Incremental affinity propagation based on K-medoids (IAPKM) & Incremental affinity propagation based on Nearest Neighbor Assignment (IAPNA).
3.	Data stream clustering with AP (Zhang et.al. 2014)	Difficulty in clustering best representatives of data & Handling evolving patterns	Employed Affinity Propagation (AP)	STRAP
4.	Combining PCA with AP (Zhang et.al. 2011)	Computational load creates load creates difficulties for fMRI	Combined Principal component analysis (PCA) + Supervised affinity propagation clustering (SAPC)	
5.	FS-AP (Yang et.al. 2010)	Multispectral images were difficult to classified with high accuracy and efficiency	Fuzzy-statistics based affinity propagation (FS-AP)	

Every time when a new object is identified it will be added to that particular graph after that the message passing will work to find a new exemplar set. Newly identified objects can be added batch by batch and number of clusters will be automatically adjusted. Sun & Guo [2] have proposed incremental AP clustering based on K-medoids (IAPKM). The reason for combining K-medoid and AP id that, AP is good at finding initial exemplar set and K-medoids is useful for modifying clustering results as per the arrival of objects. The second is incremental AP clustering based on Neighbor Assignment (IAPNA). The neighbor assignment technique uses to make relationships between the previous objects and new arriving objects. AP clustering can be used in a dynamic environment because IAPKM and IAPNA achieve higher clustering performance in compare with traditional AP clustering at the same movement the complexity of computation also can be reduced.

2.3 STRAP with AP model

Data streaming is a data that gathers through telephone records, webcams, online transactions. This kind of data is continuous, to maintain that data select best representatives from clusters of streaming data. Zhang et. al. [3] presented a STRAP algorithm for clustering of data stream with a comprehensive analysis in empirical and theoretical manner. AP is good for clustering optimality to select exemplars. The STRAP algorithm combines statistical change point detection test with AP. Every time the test detects a change the clustering model is rebuilt. STRAP confronts coming items to the existing AP model, by storing outliers in reservoir and it monitors the ratio of outliers by the use of PH change point detection test. Data streaming means incrementing the data continuously, so the clusters are incrementing too. For that, AP is good.

2.4 PCA and SAPC for fMRI

Analysis of clustering is guaranteed data-driven method for analysis if fMRI i.e. functional magnetic resonance imaging time

series data. As the computational load increases, it creates practical problems for clustering analysis. To overcome this issue Zhang et. al. [4] have proposed a novel approach, which integrates Principal component analysis (PCA) with Supervised affinity propagation clustering(SAPC). In this technique, functional magnetic resonance imaging data is initially processed by PCA to get a preliminary brain activation image. SAPC then detects different functional patterns of brain. PCA is a principal component analysis, to improve the quality of work KPCA is a better technique to use. KPCA is a kernel principal component analysis. KPCA is simply an extension for PCA. Kaung et.al. [8] have proposed support vector machine(SVM) model for combining KPCA with Genetic Algorithm(GA) for intrusion detection.

2.5 AP based on Fuzzy Statistics

Incremental Affinity propagation (AP) clustering has a fast execution speed and finds small errors in clusters for large datasets. Yang et. al. [5] proposed a novel based on fuzzy statistical similarity measure (FSS) which extracts information in multispectral imaginary. It simultaneously takes all data points into consideration to consider data points as exemplars.

Conclusion

Affinity Propagation is magnificent at clustering problems. This paper reviews about clustering and how AP is useful in most of the clustering problems. AP in vertical compression to find small representative exemplars from large training data [1]. Data streaming clustering can be effectively by AP[3]. Principal component analysis is excellent but to achieve effectivity KPCA is useful.

References

[1] Aljarah, I., Ludwig, S. A. (2013). Towards a scalable intrusion detection system Based on parallel PSO clustering using MapReduce. *In*: Proceeding of the fifteenth annual conference companion on Genetic and evolutionary computation conference Companion (p. 169–170). ACM.

[2] Elshoush, H. T., Osman, I. M. (2011). Alert correlation in collaborative intelligent Intrusion detection systems - A survey. Applied Soft Computing, 11 (7), 4349–4365.

[3] Frey, B. J., Dueck, D. (2007). Clustering by passing messages between data points. *Science*, 315 (5814), 972–976.

[4] Tieming Chen a, Ñ↑, Xu Zhang a., Shichao Jin b., Okhee Kim b Efficient classification using parallel and scalable compressed model and its application to intrusion detection.

[5] Leilei, Sun.,Chonghui Guo.,(2014). Incremental Affinity Propagation Clustering Based on Message Passing *Ieee Transactions On Knowledge And Data Engineering,* p 99.

[6] Xiangliang Zhang, Cyril Furtlehner, Cécile Germain-Renaud,Michèle Sebag Data Stream Clustering with Affinity Propagation *IEEE Transactions On Knowledge And Data Engineering*, 26 (7), JULY 2014

[7] Jiang Zhang., Xianguo Tuo., Zhen Yuan., Wei Liao., and Huafu Chen. (2011). Analysis of fMRI Data Using an Integrated Principal Component Analysis and Supervised Affinity Propagation Clustering Approach IEEE Transactions On Biomedical Engineering, 58 (11), November.

[8] Chen Yang., Lorenzo Bruzzone., Fellow., IEEE, Fengyue Sun., Laijun Lu., Renchu Guan., Yanchun Liang. (2010). A Fuzzy-Statistics-Based Affinity Propagation Technique for Clustering in Multispectral Images IEEE *Transactions On Geoscience And Remote Sensing*, 48 (6), June.

[9] Longing Cao., Senior Member., IEEE, Huaifeng Zhang., Member, IEEE, Yanchang Zhao., Member, IEEE, Dan Luo., Chengqi Zhang., Senior Member., IEEE Combined Mining: Discovering Informative Knowledge in Complex Data IEEE Transactions On Systems, Man, and Cybernetics—*PART B: CYBERNETICS*, 41 (3), June 2011.

[10] www.vedicmaths.org

[11] Fangjun Kuanga, b., Weihong Xu, c. Siyang Zhangb A novel hybrid KPCA and SVM with GA model for intrusion detection

[12] Frey, B. J., Dueck, D. (2007). Clustering by passing messages between data points. *Science,* 315 (5814), 972–976.

[13] Chang, C. C., Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* (TIST), 2 (3), 27.

[14] Kurdziel, M., Boryczko, K. (2013). Finding exemplars in dense data with affinity propagation on clusters of GPUs. *Concurrency and Computation: Practice and Experience*, 25 (8), 1137–1152.

[15] Liao, H. J., Tung, K. Y., Richard Lin, C. H., et al. (2013). Intrusion detection system: *A comprehensive review. Journal of Network and Computer Applications*, 36 (1), 16–24.

[16] Zhang, J. P., Chen, F. C., Liu, L. X., et al. (2013). Online stream clustering using density and affinity propagation algorithm. In 2013 4th IEEE international conference on s*oftware engineering and service science (ICSESS)* (p. 828–832). IEEE.

[17] Wei, M., Xia, L., Jin, J., et al. (2012). Research of intrusion detection based on clustering analysis. *In*: Proceedings of the 2012 international conference on cybernetics and informatics (p. 1973–1979). New York: Springer.

[18] Wang, F., Chawla, S., Surian, D. (2013). Latent outlier detection and the low precision problem. *In*: Proceedings of the ACM *SIGKDD workshop on outlier detection and description* (p. 46–52). ACM.

[19] Srinivas, M., Andrew, H. S. (2003). Feature selection for intrusion detection using neural networks and support vector machines. In: Annual meeting of the transportation research board.

[20] Han, J., Kamber, M., Pei, J. (2011). Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, p. 444.

[21] Liao, T.W. (2005). Clustering of Time Series Data: A Survey, *Pattern Recognition*, 38 (11), p. 1857-1874, November.

[22] Geng, H., Deng, X., Ali, H. (2005). A New Clustering Algorithm Using Message Passing and its Applications in Analyzing Microarray Data, *In*: Proceedings Fourth Int'l Conf. *Machine Learning and Applications* (ICMLA '05).

[23] Nicolas, L. (2010). New Incremental Fuzzy c medoids Clustering Algorithms, *In*: Proceedings 2010 Annual Meeting of the *North American on Fuzzy Information Processing Society* (NAFIPS '10), p. 1-6, July.

[24] He, Y., Chen, Q., Wang, X., Xu, R., Bai, X., Meng, X. (2010). An Adaptive Affinity Propagation Document Clustering, *In*: Proceedings the 7th Int'l Conf. Informatics and Systems (INFOS '10), p. 1-7, March.

[25] A triangle area based nearest neighbors approach intrusion detection Chih-Fong Tsai", Chia-YingLin

[26] Chang, C. C., Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* (TIST), 2 (3), 27.

[27] Daniela, B., Kavé, S., Margin, M. (2010). A signal processing view on packet Sampling and anomaly detection. In IEEE *INFOCOM* (p. 713–721).

[28] Davis, J. J., Clark, A. J. (2011). Data preprocessing for anomaly-based network intrusion detection: A review. *Computers & Security*, 30 (6), 353–375.

[29] Meng, Y. (2012). Adaptive false alarm filter using machine learning in intrusion detection. Practical Applications of Intelligent Systems. Berlin, Heidelberg: Springer, p. 573–584.