International Journal of Advanced Research in Management and Social Sciences

ISSN: 2278-6236
Impact Factor: 5.313

A SHORT REVIEW OF CLUSTERING TECHNIQUES

Saibal Dutta*

Sujoy Bhattacharya*

Abstract: *Data mining has been applied successfully in various research area and takes an important role in the business domain. This paper examines the several clustering techniques based on the basis of cluster policy and method, and exhibits the steps for clustering process. The paper discusses some of the important concepts regarding data type, feature selection, and cluster evolution. The results indicate that overall clustering techniques can be divided into the seven groups, namely Distance based, Density based, Model based, Grid based, Kernel based, Spectral based, Hierarchical based. This paper will serves as a guideline for industry and academic world.*

*Indian Institute of Technology, Kharagpur, West Bengal, India

ISSN: 2278-6236
Impact Factor: 5.313

1. INTRODUCTION:

Clustering is the part of data mining process, applied to the business world for discovering customer insight and useful information. Cluster analysis defines as the process of dividing large data set into a number of groups that share similar characteristics. The concept of clustering is well utilized for automatically finding groups and has been successfully applied in many domains, e.g. biology and medicine, psychology, and business. Several literature reviews have been published in last few decades. This research paper examined previous reviewed papers and summarized in tabular form. The purpose of this paper is to present the summary of clustering techniques. The clustering techniques describe elaborately in several literature (Jain and Dubes, 1988). The methodology of the cluster analysis consists of seven basic steps.

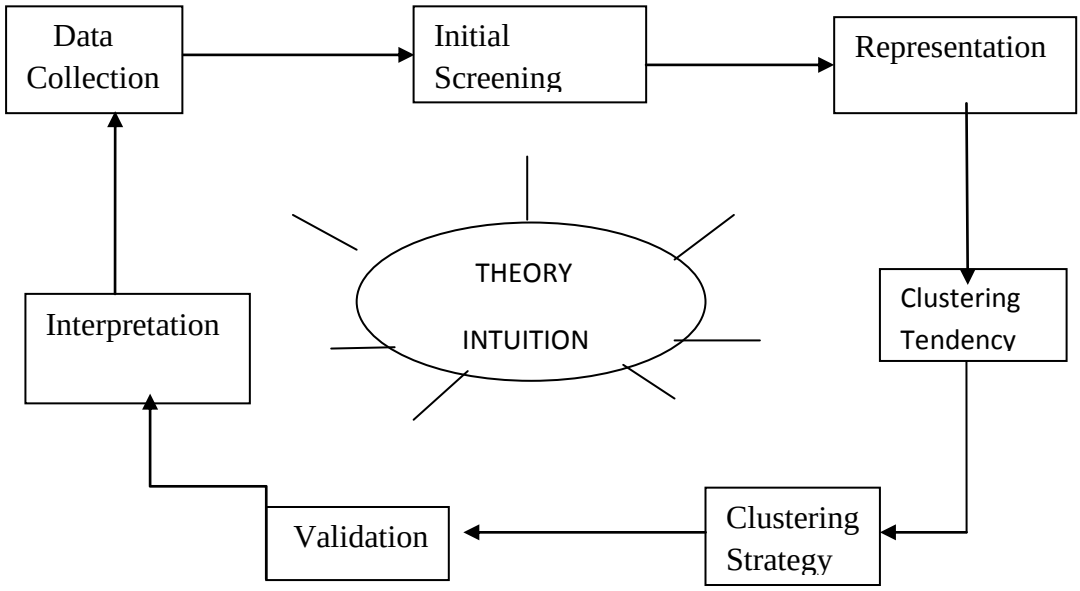

Fig.1 Steps of the clustering process

- Data Collection: Data collection is the first and most important step in the analysis. The data collection processes can be divided into two categories i) Primary data collection ii) Secondary data collection. The amount and type of mainly depend on the research objective and a number of variables. The literature suggests that the number of data collection should be five to ten times of the variables.

- Initial screening: This step is necessary to handle the raw data before the formal analysis. Researchers generally do normalization and used other visual tools like Cnernoff faces, Andrew's plot etc.
- Representation: Researcher generally transforms the data into a suitable form. This transformation helps to extract the feature from the data. This step includes a selection of proximity index and performing multidimensional scaling.
- Cluster tendency: This step is indispensable for selection of appropriate cluster algorithms. Researcher selects an algorithm based on the nature of data, the number of variables, different clustering criteria.
- Clustering Strategy: there is no any straight forward rule for selection or any universal accepted best algorithm. It is quite difficult to choose clustering techniques for a particular dataset as the performance of any clustering techniques varies from the nature of the data set to another data set. There is no such clustering techniques available that will always work well for every set of that data set.
- Validation: This step is required for choosing the optimum number of clusters and also helpful for comparing the performance of several algorithms. The cluster validity indices broadly classified into three categories i) external indices ii) internal indices iii) relative indices
- Interpretation: The interpretation of cluster result is fully depends upon the subject matter expert and domains of application

2. LITERATURE REVIEW:

We are living in a world where every moment we deal with data. The revolution of digital world makes a huge amount of data each of the fraction of the time. Data always plays a crucial role in understanding various situations and that information is also helpful in making the decision of any company. We generally extracted pattern from the data and those patterns are important for better decisions. Ross cited the work of noble prize winner Herbert Simon, who emphasizes the importance of the role of Pattern recognition. Classification of data is one of the primary aim of pattern recognition and generally in the literature, unsupervised classification is called clustering. Clustering is the process of grouping of multidimensional data based on some features. Grouping of the data as per business requirement is the key issue of the clustering and this makes us a challenging as

well as an opportunity us to analysis and find out meaningful result which help to take proper decision. Clustering is useful for several domains including business and marketing. The required literature review of clustering techniques has been done by many previous researchers (Berkhin, 2006; Jain et al., 1999; Murtagh, 1983; Xu & Wunsch, 2005; Cimpanu and Ferariu, 2012; Filippone et al., 2008) and showed in below table.

Cluster Policy	Key idea	Algorithm	References
Distance based	Mean centroids	K-means	(Jain & Dubes, 1988)
		Fuzzy K-means	(Jang et al., 1996)
	Medoid centers	K-medoids	(Jain & Dubes, 1988)
		PAM	(Kaufmann & Rousseeuw, 1990)
	Median centroids	CLARA	(Kaufmann & Rousseeuw, 1990)
		CLARANS	(Vijayarani & Nithya, 2011)
		K-medians	(Guha et al., 2003)
Density based	ε vicinity of fix size	DBSCAN	(Shah et al., 2012)
		SNN	(Moreira et al., 2005)
	ε vicinity of variable size	OPTICS	(Ankerst et al., 1999)
	vicinity of adaptive size	Subtractive clustering	(Jang et al., 1996;)
		AutoClass	(Beitzel et al., 2007)
	Tree-based	Decision trees	(Han & Kamber, 2006)
Model based	Neural networks	COBWEB	(Cimpanu and Ferariu, 2012)
		SOM	(Vesanto & Alhoniemi, 2000)
		GFMM	(Gabrys and Bargiela, 2000)
Grid based	Single grid	O-CLUSTER	(Ilango & Mohan, 2010)
	Multiple grids	STING	(Han & Kamber, 2006)
		Wave Cluster	(Han & Kamber, 2006)
	Adaptive grid	MAFIA	(Ilango & Mohan, 2010)
	Combined with density based policies	ASGC	(Ilango & Mohan, 2010)
		CLIQUE	(Ilango & Mohan, 2010)
		Mountain clustering	(Jang et al., 1996)

Cluster Policy	Key idea	Algorithm	References
Kernel based		Kernel K means	(Schölkopf et al., 1998)
	Kernelization of the metric		
		Kernel fuzzy c means	(Wu et al., 2003; Zhang and chen, 2003)
	Clustering in feature space	Kernel SOM	(Inokuchi and Miyamoto, 2004; Macdonald and Fyfe, 2000)
			(Qinand and Suganthan, 2004)
		Kernel neural gas	
	Description via support vector		
		Support vector clustering	(Huang et al., 2007)
Spectral based	Spectral graph theory	Spectral clustering	(Cristianini et al., 2001)
Hierarchical based		Single linkage	(Jain and Dubes,1988)
		Complete linkage	(Jain and Dubes,1988)
		Group average linkage	(Murtagh, 1983)
	Agglomerative	Ward's method	(Murtagh, 1983)
		BIRCH	(Zhang et al., 1996)
		CURE	(Guha et al., 1998)
		ROCK	(Guha et al., 2000)
		DIANA	(Kaufman and Rousseeuw, 2009)
	Divisive	MONA	(Kaufman and Rousseeuw, 2009)

PAM	Partitioning Around Medoids
CLARA	CLustering LARge Applications
CLARANS	Clustering Large Applications based upon RANdomized Search
DBSCAN	Density-based spatial clustering
SNN	Shared Nearest Neighbor Clustering
OPTICS	Ordering Points To Identify the Clustering Structure
COBWEB	Incremental system for hierarchical conceptual clustering
SOM	Self-organizing feature map
GFMM	Generalized Fuzzy Min-Max

O-CLUSTER	Hierarchical grid-based clustering model
STING	STatistical INformation Grid approach
MAFIA	Adaptive Grids in High Dimensions
ASGC	Axis-Shifted Grid-Clustering
CLIQUE	The Classical High-Dimensional Algorithm
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
CURE	Clustering Using REpresentatives
ROCK	RObust Clustering using linKs
DIANA	Divisive hierarchical clustering

Many data mining techniques have been used since last few decades in market segmentation, but still there is always a scope of improvement. Cluster evaluation or Cluster validation is the most difficult task in clustering techniques. Researchers have proposed several cluster validity index and sill is an open research area. Another important task is handling missing value and outlier of the data. Future researchers can improve the existing algorithm for better performance in market segmentation. Some examples of such data mining techniques include Kernel based method, neural network based model, Probabilistic Fuzzy c-means, Random forest, Evolutionary algorithm.

3. CONCLUSION:

This paper provides a comprehensive literature review of clustering techniques and clearly indicates that the overall clustering techniques can be classified into seven groups. This paper also describes the different steps for performing the segmentation techniques. This research work will definitely helpful and provide future direction to the future researcher.

REFERENCES:

1. Ankerst M., Breunig M., Kriegel H.P., Sander J. (1999) OPTICS: Ordering Points To Identify the Clustering Structure. *Proceedings of the ACM SIGMOD International Conference on Management of Data SIGMOD* '99, 49–60, Philadelphia,.
2. Beitzel S.M., Jensen E.C., Lewis D.D., Chowdhury A., Frieder O. (2007),Automatic Classification of Web Queries Using Very Large Unlabeled Query Logs. *ACM Transactions on Information Systems*, 25, 2, Article 9, New York,.
3. Berkhin, P. (2006). A survey of clustering data mining techniques. In Grouping multidimensional data (pp. 25-71). *Springer* Berlin Heidelberg.

4. Cimpanu, C., & Ferariu, L. (2012) Survey of data clustering algorithms.
5. Cristianini, N., Shawe-Taylor, J., & Kandola, J. (2001). Spectral kernel methods for clustering. *Advances in neural information processing systems*, 14, 649-655.
6. Filippone, M., Camastra, F., Masulli, F., & Rovetta, S. (2008). A survey of kernel and spectral methods for clustering. Pattern recognition, 41(1), 176-190.
7. Gabrys, B., & Bargiela, A. (2000). General fuzzy min-max neural network for clustering and classification. *Neural Networks, IEEE Transactions* on, 11(3), 769-783.
8. Guha S., Meyerson A., Mishra N., Motwani R., O'Callaghan L.(2003) Clustering Data Streams: Theory and Practice. *IEEE Transactions on Knowledge and Data Engineering*, 15, 3, 515–528.
9. Guha, S., Rastogi, R., & Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. *Information systems*, 25(5), 345-366.
10. Han, J., Kamber, M., & Pei, J. (2006). Data mining: concepts and techniques. Morgan kaufmann.
11. Huang, J. H., Tzeng, G. H., & Ong, C. S. (2007). Marketing segmentation using support vector clustering. *Expert Systems with Applications*, 32, 313–317.
12. Ilango, M. R., & Mohan, V. (2010). A survey of grid based clustering algorithms. International *Journal of Engineering Science and Technology*, 2(8), 3441-3446.
13. Inokuchi, R., & Miyamoto, S. (2004, July). LVQ clustering and SOM using a kernel function. In Fuzzy Systems, *2004. Proceedings. 2004 IEEE International Conference on* (Vol. 3, pp. 1497-1500). IEEE.
14. Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Prentice-Hall, Inc..
15. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR*), 31(3), 264-323.
16. Jang J.S.R., Sun C.T., Mizutani E., Neuro-Fuzzy and Soft Computing. A Computational Approach to Learning and Machine Intelligence. Prentice Hall, 1996.
17. Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). Wiley-Interscience.
18. MacDonald, D., & Fyfe, C. (2000). The kernel self-organising map. In Knowledge-Based Intelligent Engineering Systems and Allied Technologies, 2000. *Proceedings. Fourth International Conference on* (Vol. 1, pp. 317-320). IEEE.

19. Moreira A., Santos M., Carneiro S.(2005)Density-based Clustering Algorithms – DBSCAN and SNN. University of Minho, Portugal,
20. Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal,* 26(4), 354-359.
21. Qin, A. K., & Suganthan, P. N. (2004, August). A Novel Kernel Prototype-Based Learning Algorithm. In *ICPR* (4) (pp. 621-624).
22. Schölkopf, B., Smola, A., & Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5), 1299-1319.
23. Shah G.H., Bhensdadia C.K., Ganatra A.P.(2012) An Empirical Evaluation of Density Based Clustering Techniques. *International Journal of Soft Computing and Engineering* IJSCE, 2, 1, 216–223.
24. Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *Neural Networks, IEEE Transactions on,* 11(3), 586-600.
25. Vijayarani S., Nithya S. (2011) An Efficient Clustering Algorithm for Outlier Detection. *International Journal of Computer Applications*, 32, 7, 22–27.
26. Wu, Z. D., Xie, W. X., & Yu, J. P. (2003, September). Fuzzy c-means clustering algorithm based on kernel method. *In Computational Intelligence and Multimedia Applications, 2003. ICCIMA 2003. Proceedings. Fifth International Conference on* (pp. 49-54). IEEE.
27. Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3), 645-678.
28. Zhang, D. Q., & Chen, S. C. (2003, June). Kernel-based fuzzy and possibilistic c-means clustering. *In Proceedings of the International Conference Artificial Neural Network* (pp. 122-125).
29. Zhang, T., Ramakrishnan, R., & Livny, M. (1996, June). BIRCH: an efficient data clustering method for very large databases. *In ACM SIGMOD Record* (Vol. 25, No. 2, pp. 103-114). ACM.