International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
Volume 4 Issue 1 January - July 2014
International Manuscript ID : ISSN22308849-V4I1M1-012014

AN EMPIRICAL REVIEW ON UNSUPERVISED CLUSTERING ALGORITHMS IN MULTIPLE DOMAINS

Aditi Chawla
M.Tech. Research Scholar
Department of Computer Science and Engineering
Punjab Technical University
Jallandhar, Punjab, India

Navneet Kaur
Assistant Professor
Department of Computer Science and Engineering
RIMT Institute of Engineering and Technology
Mandi Gobindgarh, Punjab, India

ABSTRACT

Data mining refers to the investigation of the huge quantities of data sets stored in computers. Data mining is also called exploratory data analysis in multiple streams. Masses of information produced from money registers, from examining, from subject particular databases all around the organization, are investigated, examined, lessened, and reused. Quests are performed crosswise over diverse models proposed for anticipating deals, promoting reaction, and benefit. Traditional factual methodologies are principal to information mining. Mechanized AI strategies are additionally utilized. Information mining obliges ID of an issue, alongside accumulation of information that can prompt better understanding and machine models to give factual or different method for investigation. Information comes in, perhaps from numerous sources. It is incorporated and put in some normal information store. A piece of it is then taken and preprocessed into a standard organization. This arranged information is then moved to an information mining calculation which handles a yield as standards or some other sort of patterns. In this manuscript, we have analyzed the clustering algorithms for multiple applications.

Keywords - Data Mining, Clustering, Dynamic Clustering

INTRODUCTION

Numerous analytic computer models have been used in the domain of data mining. The standard model types in data mining include normal regression for prediction, logistic regression for classification, neural networks, and decision trees. These techniques are well known in the academic and research domains.

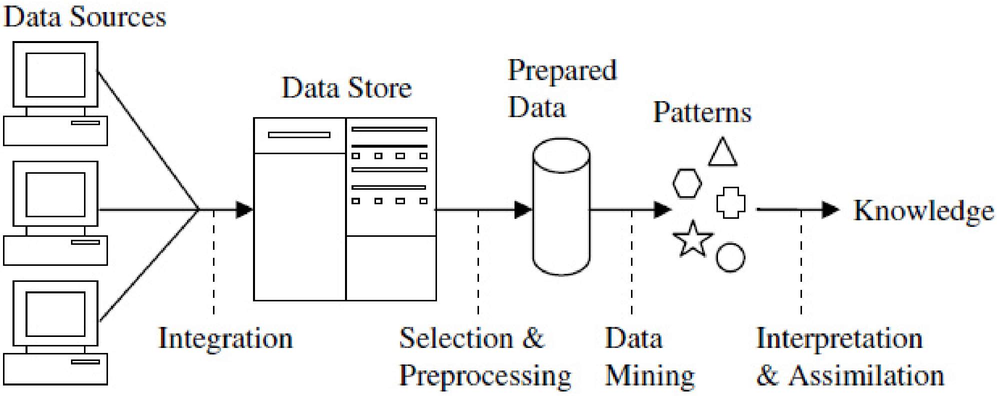


Figure 1: Data Mining and Knowledge Discovery Process

Data mining needs the identification of a problem, with collection of data that can lead to better understanding and computer models to deliver statistical or other means of analysis. This may be assisted by visualization tools, that display data, or through fundamental statistical analysis, such as correlation analysis. Data mining tools need to be versatile, scalable, capable of accurately predicting responses between actions and results, and capable of automatic implementation. Versatile refers to the proficiency of the tool to be applied in a wide variety of models. Scalable tools refer that if the tools works on a small data set, it should also work on larger data sets. Automation is useful, but its application is relative. Some analytic functions are often automated, but human setup prior to implementing procedures is required. In fact, analyst judgment is critical to successful implementation of data mining. Proper selection of data to include in searches is critical. Data transformation also is often required. Too many variables produce too much output, while too few can overlook key relationships in the data. Fundamental understanding of statistical concepts is mandatory for successful data mining.

NOTION OF CLUSTERING AND RELATED ASPECTS

Clustering is the important knowledge discovery technique with numerous applications, such as marketing and customer segmentation. Clustering group data into sets in such a way that the intra-cluster similarity is maximized and while inter-cluster similarity is minimized. Clustering is the form of unsupervised learning that examines data to find groups of items that are similar. For example, an insurance company might group customers according to income, age, types of policy purchased, prior claims experience in a fault diagnosis application, electrical faults might be grouped according to the values of certain key variables.

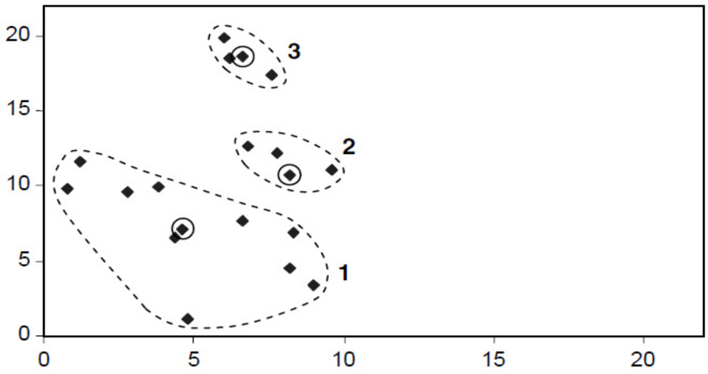


Figure 2: Clustering of Data

Many earlier clustering algorithms focus on the numerical data whose inherent geometric properties can be exploited naturally to define distance functions between data points. However, much of the data existed in the databases is categorical, where attribute values can't be naturally ordered as numerical values. Due to the special properties of categorical attributes, the clustering of categorical data seems more complicated than that of numerical data. To overcome this problem, several data-driven similarity measures have been proposed for categorical data. The behavior of such measures directly depends on the data.

Clustering is the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

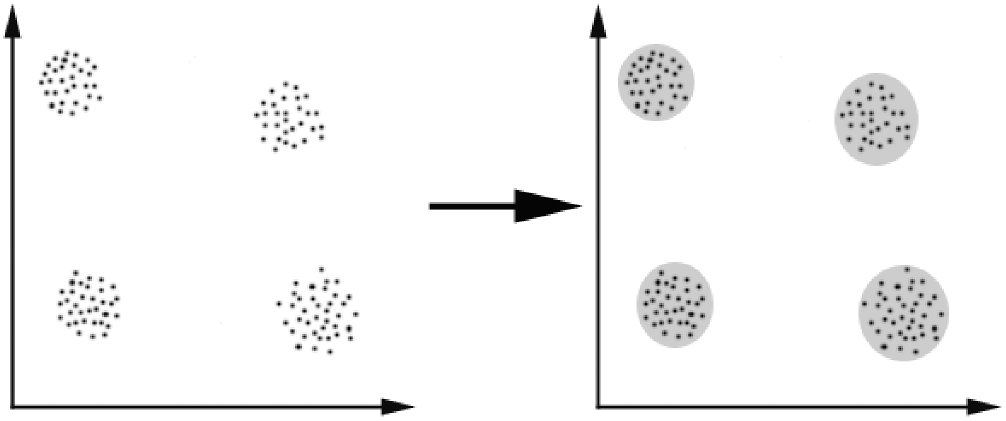


Figure 3: Formation of Clusters

Here, we easily identify the 4 clusters into which the data can be divided; the similarity criterion is *distance*: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance).This is called *distance-based clustering*.

Another kind of clustering is *conceptual clustering*: two or more objects belong to the same cluster if this one defines a concept *common* to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

MAJOR FOCUS IN CLUSTERING

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (*data reduction*), in finding "natural clusters" and describe their unknown properties (*"natural" data types*), in finding useful and suitable groupings (*"useful" data classes*) or in finding unusual data objects (*outlier detection*).

TAXONOMY

PARTITIONAL CLUSTERING

Partition-based methods construct the clusters by creating various partitions of the dataset.So, partition gives for each data object the cluster index pi. The user provides the desired number of clusters M, and some criterion function is used in order to evaluate the proposed partition or the solution. This measure of quality could be the average distance between clusters; for instance, some well-known algorithms under this category are k-means, PAM and CLARA. One of the most popular and widely studied clustering methods for objects in Euclidean space is called k-means clustering. Given a set of N data objects x_i and an integer M number of clusters. The problem is to determine C, which is a set of M cluster representatives c_j, as to minimize the mean squared Euclidean distance from each data object to its nearest centroid.

HIERARCHICAL CLUSTERING

Hierarchical clustering methods build a cluster hierarchy, i.e. a tree of clusters also known as dendogram. A dendrogram is a tree diagram often used to represent the results of a cluster analysis. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down). An agglomerative clustering starts with one-point clusters and recursively merges two or more most appropriate clusters. In contrast, a divisive clustering starts with one cluster of all data points and recursively splits into nonoverlapping clusters.

DENSITY-BASED AND GRID-BASED CLUSTERING

The key idea of density-based methods is that for each object of a cluster the neighbourhood of a given radius has to contain a certain number of objects; i. e. the density in the neighborhood has to exceed some threshold.

The shape of a neighborhood is determined by the choice of a distance function for two objects. These algorithms can efficiently separate noise. DBSCAN and DBCLASD are the well-known methods in the density based category. The basic concept of grid-based clustering algorithms is that they quantize the space into a finite number of cells that form a grid structure. And then these algorithms do all the operations on the quantized space. The main advantage of the approach is its fast processing time, which is typically independent of the number of objects, and depends only on the number of grid cells for each dimension. Famous methods in this clustering category are STING and CLIQUE.

OUTLIERS

An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid. However, in large samples, a small number of outliers is to be expected.

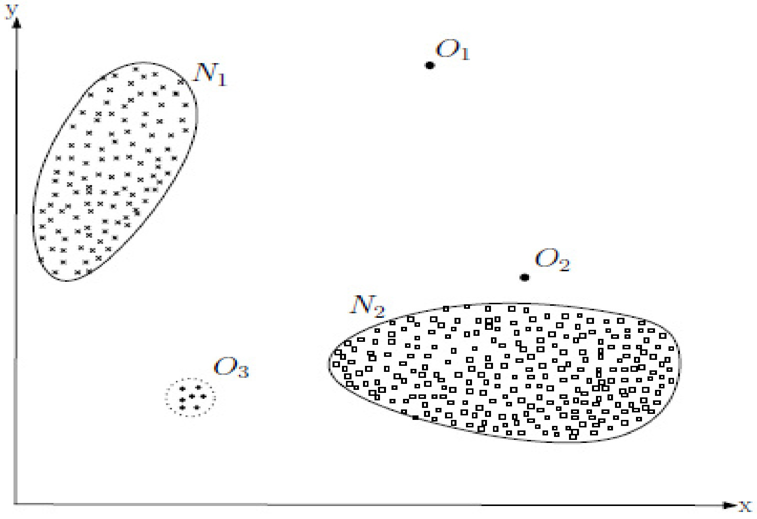


Figure 4: Outliers in two-dimensional dataset

OUTLIER DETECTION

Most outlier detection techniques treat objects with K attributes as points in ℜK space and these techniques can be divided into three main categories. The first approach is distance based methods, which distinguish potential outliers from others based on the number of objects in the neighborhood. Distribution-based approach deals with statistical methods that are based on the probabilistic data model. A probabilistic model can be either a priori given or automatically constructed using given data. If the object does not suit the probabilistic model, it is considered to be an outlier. Third, density-based approach detects local outliers based on the local density of an object's neighborhood. These methods use different density estimation strategy. A low local density on the observation is an indication of a possible outlier.

Distance-based approach

In Distance-based methods outlier is defined as an object that is at least dmin distance away from k percentage of objects in the dataset. The problem is then finding appropriate d_{min} and k such that outliers would be correctly detected with a small number of false detections. This process usually needs domain knowledge.

Definition: A point x in a dataset is an outlier with respect to the parameters k and d, if no more than k points in the dataset are at a distance d or less from x.

To explain the definition by example we take parameter k = 3 and distance d. Here are points x_i and x_j be defined as outliers, because of inside the circle for each point lie no more than 3 other points. And x′ is an inlier, because it has exceeded number of points inside the circle for given parameters k and d.

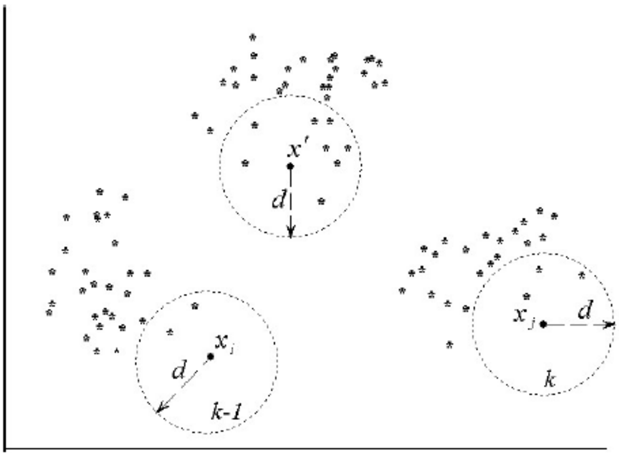


Figure 5: Illustration of outlier definition by Knorr and Ng.

DISTRIBUTION-BASED APPROACH

Distribution-based methods originate from statistics, where object is considered as an outlier if it deviates too much from underlying distribution. For example, in normal distribution outlier is an object whose distance from the average object is three times of the variance.

DENSITY-BASED APPROACH

Density-based methods have been developed for finding outliers in a spatial data. These methods can be grouped into two categories called multi-dimensional metric space-based methods and graph-based methods. In the first category, the definition of spatial neighbourhood is based on Euclidean distance, while in graph-based spatial outlier detections the definition is based on graph connectivity. Whereas distribution-based methods consider just the statistical distribution of attribute values, ignoring the spatial relationships among items, density-based approach consider both attribute values and spatial relationship.

RELATED ASPECTS OF THE CLUSTER FORMATION PROBLEM

Given a set of unclassified training data sets

- To find an efficient way of partitioning and classifying the training data into classes.

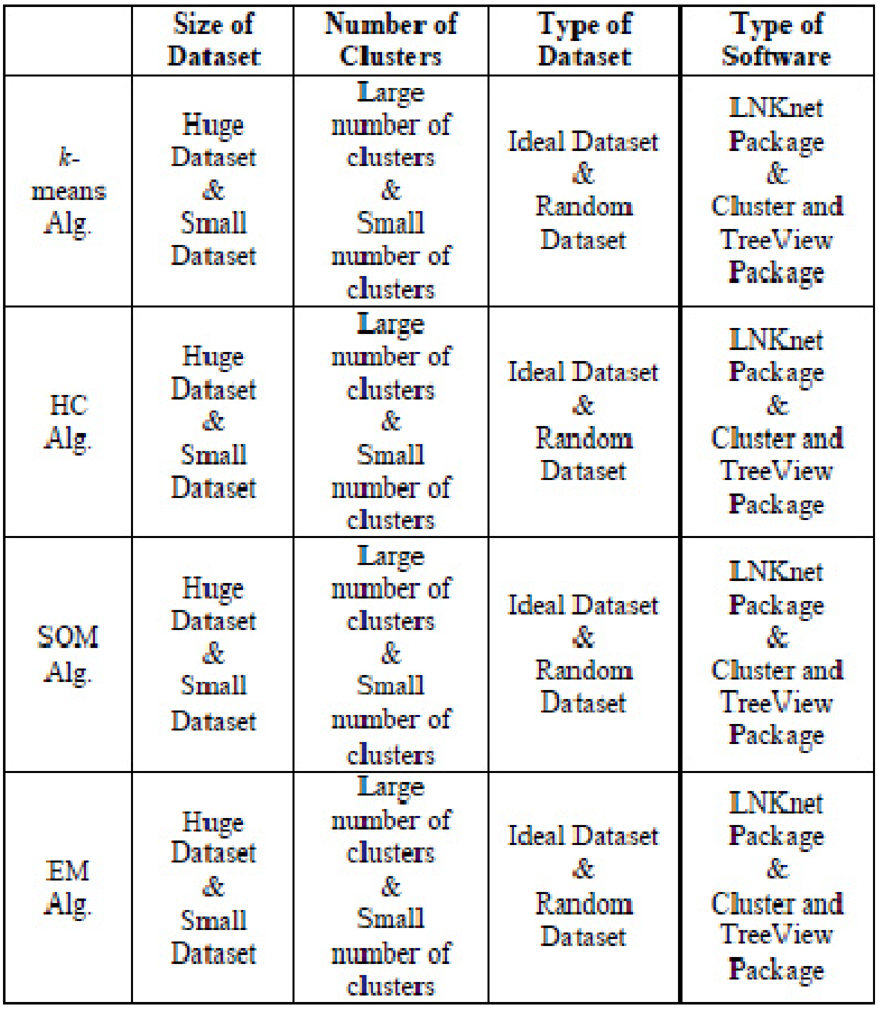
- To construct the representation that enables the category of cluster of any new example to be determined.
- Although the two subtasks are logically distinct, they are usually performed together.
- Classification learning programs are successful if the predictions they make are correct. i.e. If they agree with an externally defined classification.

In clustering, there is no externally defined notion of correctness. There are a huge number of ways in which a training set could be partitioned. Some of these are better than others. The classical methods suggest members of a cluster should resemble each other more than resemble members of other classes. Hence a good partition should

- Maximise similarity within classes
- Minimise similarity between classes.

Clustering is a well-studied data mining problem that has found applications in many areas. For example, clustering can be applied to a document collection to reveal which documents are about the same topic. The objective in any clustering application is to minimize the inter-clusters similarities and maximize the intra-cluster similarities. There are different clustering algorithms each of which may or may not be suited to a particular application.

The traditional clustering paradigm pertains to a single dataset. Recently, attention has been drawn to the problem of clustering multiple heterogeneous datasets where the datasets are related but may contain information about different types of objects and the attributes of the objects in the datasets may differ significantly. A clustering based on related but different object sets may reveal significant information that cannot be obtained by clustering a single dataset.



	Size of Dataset	Number of Clusters	Type of Dataset	Type of Software
k-means Alg.	Huge Dataset & Small Dataset	Large number of clusters & Small number of clusters	Ideal Dataset & Random Dataset	LNKnet Package & Cluster and TreeView Package
HC Alg.	Huge Dataset & Small Dataset	Large number of clusters & Small number of clusters	Ideal Dataset & Random Dataset	LNKnet Package & Cluster and TreeView Package
SOM Alg.	Huge Dataset & Small Dataset	Large number of clusters & Small number of clusters	Ideal Dataset & Random Dataset	LNKnet Package & Cluster and TreeView Package
EM Alg.	Huge Dataset & Small Dataset	Large number of clusters & Small number of clusters	Ideal Dataset & Random Dataset	LNKnet Package & Cluster and TreeView Package

Table 1 : Comparative Analysis of assorted clustering algorithms

CONCLUSION

Cluster Formation or basically grouping is the procedure of total the set of articles in such a way, to the point that protests in the same assembly called group that are more comparable in some sense or an alternate to one another than to those in different aggregations or Clusters. It is an unmistakable and required errand of exploratory information mining, and a basic system for factual information examination utilized as a part of numerous fields, including machine taking in, example difference, picture dissection, data recovery, and bioinformatics. Group investigation itself is not one particular calculation, however the general assignment to be explained. It could be attained by different calculations that vary altogether in their idea of what constitutes a group and how to productively discover them. Famous thoughts of groups incorporate aggregations with little separations around the Cluster parts, thick ranges of the information space, interims or specific measurable disseminations. Clustering can hence be figured as a multi-objective enhancement issue. A huge measure of examination work is under methodology all around the globe in different calculations. In this examination work, we have proposed a novel algorithmic approach and model that makes utilization of the numerical establishment and evolutionary methodology for the shaping of Clusters in productive and successful behavior regarding execution time and cohorted outcomes. What's to come extent of the examination work can stretched out to the cross breed methodology. The mixture methodology makes utilization of two or more algorithmic methodologies to be consolidated in single plan to get the ideal effects. The mixture methodology can make utilization of the ground dwelling insect state enhancement or hereditary calculation to get the ideal outcomes. In the event that the displayed calculation is executed to the emphases with hereditary algorithmic methodology, the best result might be attained. Later on work, the bunch structuring could be coordinated with best first pursuit of the heuristic quest strategies for the evacuation of clamor.

REFERENCES

[1] Achtert, E.; Böhm, C.; Kröger, P. (2006). "DeLi-Clu: Boosting Robustness, Completeness, Usability, and Efficiency of Hierarchical Clustering by a Closest Pair Ranking". LNCS: Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science 3918: 119–128. doi:10.1007/11731139_16. ISBN 978-3-540-33206-0.

[2] Aditya Desai, Himanshu Singh, VikramPudi, 2011. DISC: Data-Intensive Similarity Measure for Categorical Data, Pacific-Asia Conferences on Knowledge Discovery Data Mining

[3] Andre Baresel, HarmenSthamer, Michael Schmidt,2002. Fitness Function Design to improve Evolutionary Structural Testing

[4] Andrew L.Nelson, Gregory J.Barlow, Lefteris Doitsidis,2008 .Fitness Functions in Evolutionary Robotics: A Survey and Analysis

[5] Can, F.; Ozkarahan, E. A. (1990). "Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases".ACM Transactions on Database Systems15 (4): 483. doi:10.1145/99935.99938

[6] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek (2011). "Density-based Clustering".WIREs Data Mining and Knowledge Discovery1 (3): 231–240. doi:10.1002/widm.30.

[7] He Zengyou, Xu Xiaofei, Deng Shenchun, 2002. Squeezer: An Efficient Algorithm for Clustering Categorical Data,Journal of Computer Science and Technology,Vol. 17, No. 5,pp 611-624

[8] He Zengyou, Xu Xiaofei, Deng Shenchun, 2003. Discovering Cluster Based Local Outliers,Article Published in Journal Pattern Recognition Letters, Volume 24. Issue 9-10,pp 1641-1650,01 June 2003

[9] He Zengyou, Xu Xiaofei, Deng Shenchun, 2006. Improving Categorical Data Clustering Algorithm by Weighting Uncommon Attribute Value Matches, ComSIS Vol.3,No.1

[10] Jerzy Stefanowski, 2009, Data Mining - Clustering, University of Technology, Poland

[11] Lloyd, S. (1982). "Least squares quantization in PCM". IEEE Transactions on

Information Theory28 (2): 129–137. doi:10.1109/TIT.1982.1056489.

[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In EvangelosSimoudis, Jiawei Han, Usama M. Fayyad. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96).AAAI Press. pp. 226–231. ISBN 1-57735-004-9.

[13] Microsoft academic search: most cited data mining articles: DBSCAN is on rank 24, when accessed on: 4/18/2010

[14] M.Davarynejad, M.-R.Akbarzadeh-T, N.Pariz,2007. A Novel Framework for Evolutionary Optimization: Adaptive Fuzzy Fitness Granulation, IEEE Conference on Evolutionary Computation, pp 951-956,2007

[15] MihaelAnkerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander (1999). "OPTICS: Ordering Points To Identify the Clustering Structure". ACM SIGMOD international conference on Management of data.ACM Press. pp. 49–60.

[16] R. Ng and J. Han. "Efficient and effective clustering method for spatial data mining". In: Proceedings of the 20th VLDB Conference, pages 144-155, Santiago, Chile, 1994.

[17] R.Ranjini, S.AnithaElavarasi, J.Akilandeswari.2012. Categorical Data Clustering Using Cosine Based Similarity for Enhancing the Accuracy of Squeezer Algorithm

[18] S Roy, D K Bhattacharyya (2005). "An Approach to find Embedded Clusters Using Density Based Techniques".LNCS Vol.3816.Springer Verlag. pp. 523–535.

[19] ShyamBoriah, VarunChandola, Vipin Kumar, 2008. Similarity Measures for Categorical Data: A Comparative Evaluation, SIAM International Conference on Data Mining-SDM

[20] Tian Zhang, Raghu Ramakrishnan, MironLivny. "An Efficient Data Clustering Method for Very Large Databases." In: Proc. Int'l Conf. on Management of Data, ACM SIGMOD, pp. 103–114.

[21] Z. Huang. "Extensions to the k-means algorithm for clustering large data sets with categorical values". Data Mining and Knowledge Discovery, 2:283–304, 1998.

[22] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 17 December 2008

[23] Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93.p. 207.doi:10.1145/170035.170072. ISBN 0897915925.

[24]Barnett, V. and Lewis, T.: 1994, Outliers in Statistical Data. John Wiley &Sons., 3rd edition.