American Journal of Applied Sciences 9 (2): 254-258, 2012
ISSN 1546-9239

A Review on Clustering and Outlier Analysis Techniques in Datamining

[1]Koteeswaran, S., [2]P. Visu and [3]J. Janet
[1,2]Department of CSE,
Vel Tech Dr. RR and Dr. SR Technical University, Chennai
[3]Professor and Head, Department of CSE, Dr. MGR University, Chennai

Abstract: Problem statement: The modern world is based on using physical, biological and social systems more effectively using advanced computerized techniques. A great amount of data being generated by such systems; it leads to a paradigm shift from classical modeling and analyses based on basic principles to developing models and the corresponding analyses directly from data. The ability to extract useful hidden knowledge in these data and to act on that knowledge is becoming increasingly important in today's competitive world. **Approach:** The entire process of applying a computer-based methodology, including new techniques, for discovering knowledge from data is called data mining. There are two primary goals in the data mining which are prediction and classification. The larger data involved in the data mining requires clustering and outlier analysis for reducing as well as collecting only useful data set. **Results:** This study is focusing the review of implementation techniques, recent research on clustering and outlier analysis. **Conclusion:** The study aims for providing the review of clustering and outlier analysis technique and the discussion on the study will guide the researcher for improving their research direction.

Key words: Datamining, data warehousing, clustering, outlier analysis, prediction, classification

INTRODUCTION

In the earlier stage, a basic scientific model which includes Newton's laws of motion, Maxwell's equations in electromagnetism are used for developing various applications in mechanical engineering and or electrical engineering. In the later stage, with the help of computers, there are huge developments in almost all engineering applications and optimization in every field of study.

Data mining is an iterative processes, these process are defined by discovery through either automatic or manual methods. Data mining is most useful in an exploratory analysis, in which, there are no predetermined notions. Data mining is the technique which search for new, valuable and nontrivial information in large volumes of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers.

The major tasks (Kantardzic, 2011) in the data mining are:

- Classification-discovery of a predictive learning function that classifies a data item into one of several predefined classes
- Regression-discovery of a predictive learning function, which maps a data item to a real-value prediction variable
- Clustering-a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data
- Summarization-an additional descriptive task that involves methods for finding a compact description for a set (or subset) of data
- Dependency modelling-finding a local model that describes significant dependencies between variables or between the values of a feature in a data set or in a part of a data set
- Change and deviation detection-discovering the most significant changes in the data set

There are tremendous volumes of data are filled in the computers and in the internet. The Government agencies, scientific institutions and businesses have all dedicated enormous resources to collecting and storing data. In the real world, only a small amount of these data is used due to the volumes are simply too large to manage, the data structures themselves are too complicated to be analysed effectively. To utilize and understand large, complex, information-rich data sets is a common problem.

Corresponding Author: Koteeswaran, S., Department of CSE, Vel Tech Dr. RR and Dr.SR Technical University, Chennai-62, Tamilnadu, India Tel: +91 9884378785

The major goal of the data mining is further reduced as prediction and the classification. The Prediction is the process which predicts unknown or future values of interest by using some variables or fields in the data set and the prediction produces the model of the system described. Classification is the process which is used for finding patterns by describing the data that can be interpreted by the human being and the classification produces new, nontrivial information based on the available data set. In order to execute these processes in the data mining requires clustering and outlier analysis for reducing as well as identifying useful dataset.

MATERIALS AND METHODS

The clustering or the cluster analysis is a set of methodologies for classification of samples into a number of groups. Therefore, the samples in one group are grouped and samples belonging to different groups are grouped as another group. The input of clustering is a set of samples and the process of clustering is to measure the similarity and or dissimilarity between given samples. The output of the clustering is a number of groups or clusters in the form of graphs, histograms and normal computer results showing group no.

The Clustering is a well-established technique for data interpretation. It usually requires prior information, e.g., about the statistical distribution of the data or the number of clusters to detect. “Clustering” attempts to identify natural clusters in a data set. It does this by partitioning the entities in the data such that each partition consists of entities that are close (or similar), according to some distance (similarity) function based on entity attributes. Conversely, entities in different partitions are relatively far apart (dissimilar). An example of cluster is shown in Fig. 1.

Existing clustering algorithms such as K-means, PAM, CLARANS, DBSCAN, CURE and ROCK are designed to find clusters that fit some static models. For example, K-means, PAM and CLARANS assume that clusters are hyper-ellipsoidal or hyper-spherical and are of similar sizes. The DBSCAN assumes that all points of a cluster are density reachable and points belonging to different clusters are not. However, all these algorithms can breakdown if the choice of parameters in the static model is incorrect with respect to the data set being clustered, or the model did not capture the characteristics of the clusters (e.g., size or shape). Because the objective is to discern structure in the data, the results of a clustering are then examined by a domain expert to see if the groups suggest something.

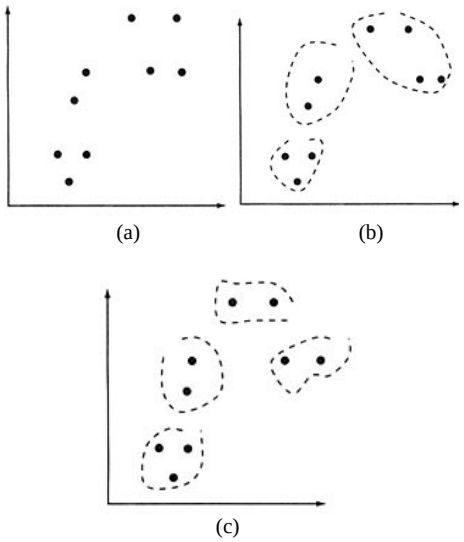

Fig. 1: Cluster analysis process (a) Initial data (b) Output in three clusters (c) Output in four clusters

For example, crop production data from an agricultural region may be clustered according to various combinations of factors, including soil type, cumulative rainfall, average low temperature, solar radiation, availability of irrigation, strain of seed used and type of fertilizer applied. Interpretation by a domain expert is needed to determine whether a discerned pattern- such as a propensity for high yields to be associated with heavy applications of fertilizer-is meaningful, because other factors may actually be responsible (e.g., if the fertilizer is water soluble and rainfall has been heavy). Many clustering algorithms that work well on traditional data deteriorate when executed on geospatial data (which often are characterized by a high number of attributes or dimensions), resulting in increased running times or poor-quality clusters. For this reason, recent research has cantered on the development of clustering methods for large, highly dimensioned data sets, particularly techniques that execute in linear time as a function of input size or that require only one or two passes through the data. Recently developed spatial clustering methods that seem particularly appropriate for geospatial data include partitioning, hierarchical, density-based, grid-based and cluster-based analysis. Hierarchical methods build clusters through top-down (by splitting) or bottom-up (through aggregation) methods. Density-based methods define clusters as regions of space with a relatively large number of spatial objects; unlike other methods, these can find arbitrarily-shaped clusters. Grid-based methods divide space into a raster tessellation and clusters objects

based on this structure. Model-based methods find the best fit of the data relative to specific functional forms. Constraints-based methods can capture spatial restrictions on clusters or the relationships that define these clusters.

An input to a cluster analysis can be described as an ordered pair (X, s), or (X, d), where X is a set of descriptions of samples and s and d are measures for similarity or dissimilarity (distance) between samples, respectively. Output from the clustering system is a partition $A = \{G_1, G_2, \ldots, G_N\}$ where G_k, $k = 1, \ldots, N$ is a crisp subset of X such that Eq. 1 and 2:

$$G_1 \cup G_2 \cup \ldots, \cup G_N = X \quad (1)$$

And:

$$G_1 \cap G_2 \cap \ldots, \cap G_N = \emptyset \quad (2)$$

The $G_1, G_2 \ldots G_n$ are the clusters.

Most clustering algorithms are based on the following four popular approaches:

- Partitioning methods
- Hierarchical clustering
- Iterative square-error partitioned clustering
- Density based clustering

Partitioning methods: Given a database of n objects or data tuples, a partitioning method constructs $k(\leq n)$ partitions of the data, where each partition represents a cluster. That is, it classifies the data into k groups, which together satisfy the following requirements:

- Each group must contain at least one object
- Each object must belong to exactly one group

Notice that the second requirement can be relaxed in some fuzzy partitioning techniques. Such a partitioning method creates an initial partitioning. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. Representative algorithms include k-means, k-medoids, CLARANS and the EM algorithm.

Hierarchical clustering methods: Hierarchical techniques organize data in a nested sequence of groups, which can be displayed in the form of a dendrogram or a tree structure. A hierarchical method creates a hierarchical decomposition of a given set of data objects. Hierarchical methods can be classified as agglomerative (bottom-up) or divisive (top-down), based on how the hierarchical decomposition is formed. AGNES and DIANA are examples of agglomerative and divisive methods, respectively.

Iterative square-error partitioned clustering methods: Square-error partitioned algorithms attempt to obtain that partition which minimizes the within-cluster scatter or maximizes the between-cluster scatter. These methods are nonhierarchical because all resulting clusters are groups of samples at the same level of partition. To guarantee that an optimum solution has been obtained, one has to examine all possible partitions of N samples of n-dimensions into K clusters (for a given K), but that retrieval process is not computation ally feasible.

Density-based clustering methods: Most partitioning methods cluster objects based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty in discovering clusters of arbitrary shape. Other clustering methods have been developed based on the notion of density. Their general idea is to continue growing a given cluster as long as the density (the number of objects or data points) in the "neighbourhood" exceeds a threshold. Such a method is able to filter out noises (outliers) and discover clusters of arbitrary shape. Representative algorithms include DBSCAN, OPTICS and DENCLUE.

RESULTS AND DISCUSSION

The Table 1 shows the samples from the class room in which the number of students passed in the examination is listed out. These samples are grouped as failed students, mere pass and distinction students. The Table 2 shows the clustered result in which the three clusters are shown with the corresponding samples. The graphical representation of the result of clustering technique is shown in the Fig. 1.

Ramadhan *et al*. (2005) proposes a Classification of Techniques for Web Usage Analysis, in which the author provides an application that offer useful insights to address crucial points such as user interests into a particular page, server load balancing, Web site reorganization, clustering of similar browsing patterns,

Geetha and Narayanan (2008) studied various clustering method applied for video data base segmentation and retrieval, the study includes shot segmentation, key frame extraction, feature extraction, clustering, indexing and video retrieval-by similarity, probabilistic, transformational, refinement and relevance feedback.

Table 1: Samples of result (pass % of students)

Number of students	Marks
20	10
30	20
42	30
50	40
65	50
11	60
22	70
34	80
54	90
23	100

Table 2: Samples after clustering

Cluster	Number of students	Marks
Cluster 1	20	10
	30	20
	42	30
Cluster 2	50	40
	65	50
	11	60
	22	70
Cluster 3	34	80
	54	90
	23	100

Indrayanti *et al*. (2010) proposed acoustic model adaptation which explained in three models as: (1) to analyze pronunciation variant with knowledge-based and data-derived methods (2) to align knowledge-based and data-derived results in order to list frequently mispronounced phones with their variants; (3) to perform a state-clustering procedure with the list obtained from the second step. The author expressed that the proposed method achieved an average gain in Hit + Rejection (the percentage of correctly accepted and correctly rejected utterances by the system as the human raters do) of 2.9 points and 2 points for native and non-native subjects, respectively, when compared with the system without adaptation. Average gains of 12.7 and 6.2 points for native and non-native students in Hit + Rejection were obtained to the acoustic model adaptation.

Shanmugam *et al*. (2011) proposes a modified K-Means clustering algorithm, called "Fast SQL K-Means" for Medical Image Segmentation. The author concluded that the method proposed by the author takes less than 10 sec to cluster an image size of 400×250 (100K pixels), whereas the running time of direct K-Means is around 900 sec. Since the entire processing is done with database, additional overhead of import and export of data is not required.

Clustering is applied for variety of application, which includes web mining (Ramadhan *et al*., 2005), image processing. This study review various clustering technique in the data mining and image processing. Generally, the Existing clustering methods has few drawbacks, which are speed and quality of segmentation. A faster method is required especially for effective, accurate and scalable clinical analysis and diagnosis. Geetha and Narayanan (2008) reviews the content based video retrieval using clustering methods.

Sujaritha and Annadurai (2011) proposes Expectation Maximization (EM) model for fitting Maximum A Posteriori (MAP) algorithm which segments the image by utilizing the pixel's color and texture features and the captured neighborhood relationships among them. This method enhances the smoothness towards piecewise-homogeneous segmentation and reduces the edge-blurring effect. The results of this algorithm simultaneously calculates the model parameters and segments the pixels iteratively in an interleaved manner. Finally, it converges to a solution where the model parameters and pixel labels are stabilized within a specified criterion.

In recent days, the clustering applies swarm intelligence. The swarm intelligence is behavioral approach learned from animal and insects. Particle Swarm Optimization, Fireflies, Artificial Bee Colony, Ant Colony Optimization (ACO) are few swarm intelligence algorithms. The detailed study of ACO based clustering is available in Sumathi *et al*. (2010) and Mohan and Baskaran (2011; 2012).

CONCLUSION

The introduction to the clustering, various clustering methods are briefly explained in this study. The basic clustering process is explained with an example. The clustering methods which is implemented or proposed in the last few years are listed on each category of clustering methodology. Therefore, the objective of the study is to provide basic review on clustering technique is fulfilled.

REFERENCES

Geetha, P. and V. Narayanan, 2008. A survey of content-based video retrieval. J. Comput. Sci., 4: 474-486. DOI: 10.3844/jcssp.2008.474.486

Indrayanti, L., Y. Chisaki and T. Usagawa, 2010. Acoustic model adaptation for indonesian language utterance training system. J. Comput. Sci., 6: 1334-1340. DOI: 10.3844/jcssp.2010.1334.1340

Kantardzic, M., 2011. Data Mining: Concepts, Models, Methods and Algorithms. 2nd Edn., John Wiley and Sons, Oxford, ISBN: 1118029135, pp: 520.

Mohan, B.C. and R. Baskaran, 2011. Reliable transmission in network centric military network. Eur. J. Sci. Res., 50: 564-574.

Mohan, B.C. and R. Baskaran, 2012. A survey: Ant Colony Optimization based recent research and implementation on several engineering domain. Expert Syst. Appli., 39: 4618-4627. DOI: 10.1016/j.eswa.2011.09.076

Ramadhan, H., M. Hatem, Z. Al-Khanjri and S. Kutti, 2005. A classification of techniques for web usage analysis. J. Comput. Sci., 1: 413-418. DOI: 10.3844/jcssp.2005.413.418

Sujaritha, M. and S. Annadurai, 2011. A new modified gaussian mixture model for color-texture segmentation. J. Comput. Sci., 7: 279-283. DOI: 10.3844/jcssp.2011.279.283

Shanmugam, N., A.B. Suryanarayana, S. Tsb, D. Chandrashekar and C.N. Manjunath, 2011. A novel approach to medical image segmentation. J. Comput. Sci., 7: 657-663. DOI: 10.3844/jcssp.2011.657.663

Sumathi, C.P., R.P. Valli and T. Santhanam, 2010. An application of session based clustering to analyze web pages of user interest from web log files. J. Comput. Sci., 6: 785-793. DOI: 10.3844/jcssp.2010.785.793