Volume-3, Issue-7, July-2014 • ISSN No 2277 - 8160

Research Paper | Computer Science

Analysis of Various Database Using Clustering Techniques

Khushbu Patel Gujarat Technological University

ABSTRACT

Data mining basically refers to information elicitation from data warehouses. One of the data mining techniques is clustering. Data clustering is a process of putting similar data into groups. Cluster analysis results usually strongly depend on the clustering algorithm selected. Cluster analysis can be used to group samples and to develop ideas about the multivariate data set at hand. This paper review three of the most representative off-line clustering techniques: K-means clustering, Hierarchical clustering and Filtered clustering. Performance of clustering techniques in this paper are measured and compared. This comparison is done by using data mining tool (i.e. WEKA).

KEYWORDS : Data clustering, Filtered clustering, Hierarchical clustering, K-means clustering, Taxonomy

1. INTRODUCTION

Data Clustering is considered an approach for finding similarities in data and putting similar data into groups. The idea of data grouping, or clustering, is simple in its nature and is close to the human way of thinking; whenever we are presented with a large amount of data, we usually tend to summarize this huge number of data into a small number of groups or categories in order to further facilitate its analysis.

Moreover, most of the data collected in many problems seem to have some inherent properties that lend themselves to natural groupings. Nevertheless, finding these groupings or trying to categorize the data is not a simple task for humans unless the data is of low dimensionality (two or three dimensions at maximum.) This is why some methods in soft computing have been proposed to solve this kind of problem. Those methods are called "Data Clustering Methods" and they are the subject of this paper.

Clustering is an unsupervised Machine Learning technique of finding patterns in the data, i.e., these algorithms work without class attributes. Classifiers, on the other hand, are supervised and need a class attribute.

2. CLUSTERING TECHNIQUES

The most representative off-line clustering techniques are reviewed:

- K-means (or Hard C-means) Clustering
- Filtered Clustering
- Hierarchical clustering

As mentioned earlier, data clustering is concerned with the partitioning of a data set into several groups such that the similarity within a group is larger than that among groups. This implies that the data set to be partitioned has to have an inherent grouping to some extent; otherwise if the data is uniformly distributed, trying to find clusters of data will fail, or will lead to artificially introduced partitions. Another problem that may arise is the overlapping of data groups. Overlapping groupings sometimes reduce the efficiency of the clustering method, and this reduction is proportional to the amount of overlap between groupings.

K-means clustering

This technique has been applied to a variety of areas, i.e. image and speech data compression, data preprocessing for system modeling using radial basis function networks, and task decomposition in heterogeneous neural network architectures. This algorithm relies on finding cluster centers by trying to minimize a cost function of dissimilarity (or distance) measure.

Hierarchical clustering

Hierarchical algorithms find successive clusters using previously established clusters. These algorithms usually are either agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

Filtered clustering

Filtered clustering meta-clusterer offers the user the possibility to apply filters directly before the clusterer is learned. This approach eliminates the manual application of a filter in the Preprocess, since the data gets processed on the fly. Useful if one needs to try out different filter setups.

3. OBJECTIVE

To study which clustering algorithm works best on different types of datasets by applying variety of clustering methods like k-means, hierarchical and filtered clustering on three different datasets and to analyze it by comparing resultant datasets.

4. EXPECTED RESULTS

To find the best algorithm for different categories of datasets as every result will not be similar when same algorithms are applied on different datasets like Large-Large ,Large-Moderate, Large-Small, Moderate-Large, Moderate-Moderate, Moderate-Small, Small-Large, Small-Moderate, Small-Small.

5. METHODOLOGY

Data Collection

Data collection refers to describe a process of preparing and collecting data for specific purpose. The purpose of data collection is to obtain information to keep on record, to make decisions about important issues, to pass information on to others. Primarily, data is collected to provide information regarding a specific topic.

Data collection often contains the following activity.

1. Pre collection activity – Agree goals, target data, definitions, methods
2. Collection – data collection
3. Present Findings – usually involves some form of sorting analysis and/or presentation.

Prior to any data collection, pre-collection activity is one of the most crucial steps in the process. It is often discovered too late that the value of their interview information is discounted as a consequence of poor sampling of both questions and informants and poor elicitation techniques. After pre-collection activity is fully completed, data collection in the field, whether by interviewing or other methods, can be carried out in a structured, systematic and scientific way.

Selection of three databases.

For comparing certain results and to find suitable algorithm for specific databases clustering technique is used. We will be selecting three different databases in order to conclude certain results and by applying different algorithms we will find best algorithm for specific dataset.

Plants
Data has been extracted from the USDA plants database. It contains all plants (species and genera) in the database and the states of USA and Canada where they occur.

Water Treatment Plants
Faults in a urban waste water treatment plant.
Sponge
Marine Sponges,
O.HADROMERIDA (DEMOSPONGIAE.PORIFERA)
Derive nine datasets from the above three databases.

To derive different datasets we have set specific parameters like to derive large-large dataset we will keep attributes and tuples both constant, while in large-moderate tuples will remain constant and attribute list will vary to moderate similarly in large-small attribute will vary to small. Similarly to derive other three datasets tuples will vary to moderate and attribute list will vary to large, moderate and small and to derive last three datasets tuples will vary to small and attribute list to large, small and moderate.

(i.e. Large-Large ,Large-Moderate, Large-Small)
(i.e. Moderate-Large, Moderate-Moderate, Moderate-Small)
(i.e. Small-Large, Small-Moderate, Small-Small)

Run different algorithms on above nine datasets.
To select best algorithm for all three selected datasets, different algorithms like k-means, filtered and hierarchical will be applied to conclude certain results.

Collect results by applying various run parameters.
To get satisfied result accuracy of all techniques is to be taken into consideration, so proper parameter will generate new result.

Compare the results.
By applying certain algorithms specific results will be analyzed in order to find best algorithm so that we can conclude by comparing the final results which algorithm is best for specific data.

6. IMPLEMEMTATION AND RESULTS OF CLUSTERING TECHNIQUES

Having introduced the different clustering techniques and their basic mathematical foundations, we now turn to the discussion of these techniques on the basis of a practical study. This study involves the implementation of each of the three techniques introduced previously, and testing each one of them on a different dataset. Each clustering algorithm is presented with the training data set, and as a result nine clusters are produced for each technique. The data in the evaluation set is then tested against the found clusters and an analysis of the results is conducted. The following sections present the results of each clustering technique, followed by a comparison of the three techniques.

Table 1.1 - Sample Datasets

No of Instance	Large	Moderate	Low
Large	Plants (20000-70)	Sponge Data Set (76-45)	Water Treatment Plant (500-38)
Moderate	Plants (10000-70)	Sponge Data Set (50-45)	Water Treatment Plant (250-38)
Low	Plants (5000-70)	Sponge Data Set (20-45)	Water Treatment Plant (100-38)

Table 1.2 - Comparison of all Datasets

Dataset	No of instances	K-Mean Clustering		Hierarchical Clustering	Filtered Clustering	
		No of iterations	RMSE	Clustered Instances	No of iterations	RMSE
Large-Large	Plants	13	147385.0	X	X	X
Large-Moderate	Sponge	2	711.0667032163744	0 74 (97%) 1 2 (3%)	2	711.0667032163744
Large-Low	Water-Treatment-Plants	18	252.245578305058	X	18	252.245578305058
Moderate-Large	Plants	3	78742.0	X	3	78742.0
Moderate-Moderate	Sponge	3	418.0421455938697	0 29 (59%) 1 20 (41%)	3	418.0421455938697
Moderate-Low	Water-Treatment-Plants	10	159.46419618909374	0 249 (100%)	10	159.46419618909374
Low-Large	Plants	6	37416.0	X	6	37416.0
Low-Moderate	Sponge	6	163.6187657828283	0 8 (42%) 1 11 (58%)	6	163.6187657828283
Low-Low	Water-Treatment-Plants	16	92.46644227003816	0 99 (100%)	18	92.46644227003816

8. CONCLUSION

The clustering algorithm takes a data set and sorts them into groups, so we can make conclusions based on what trends we see within these groups. Clustering differs from classification and regression by not producing a single output variable, which leads to easy conclusions, but instead requires that we observe the output and attempt to draw our own conclusions. From the further studies we come to conclusion that:-

1. For Large no. of data Filtered Clustering algorithm is good.
2. For moderate no. of data Hierarchical clustering is good.
3. For low no. of data K-mean clustering is good.

REFERENCES

[1] A.A. Abdelmalek, E. Zakaria, S. Michel, Evaluation of text clustering methods using WordNet, The International Arab Journal of Information Technology 7 (4) (2010). | [2] Li C. and Biswas G., "Unsupervised Learning with Mixed Numeric and Nominal data," IEEE | Transactions on Knowledge and Data Engineering, vol. 14, no. 4, pp.673-690, 2002. | [3] Jiawei Han and. Micheline Kamber. "Data Mining: Concepts and Techniques. Second | Edition." | [4] Box, G.E.P. and Cox, D.R., 1964. An analysis of transformations. Journal of the Royal | Statistical Society (B) 26, 211-252. | [5] Calinski, T. and Harabasz, J., 1974. A dendrite method for cluster analysis. Communications | in Statistics 3, 1-27. |