RESEARCH PAPER | Science | Volume : 3 | Issue : 7 | July 2013 | ISSN - 2249-555X

Clustering Algorithm With Reference To TANGARA

KEYWORDS: Data mining algorithms, K-means algorithms, EM algorithm,Clustering methods etc.

Dr.K.M Nalawade
Professor, DGCC, Satara

Prof. Mrs.Rohini Ganesh Gaikwad
Mphil Research Scholar, KBPIMSR,Satara

ABSTRACT *Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. TANGARA is a data mining tools. It is contain the many machine leaning algorithms. In this paper we are studying the various clustering algorithms. Cluster analysis or clustering is the task of assigning a set of objects into groups(called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. This paper mainly focus on the comparison of the different clustering algorithms of TANGARA with respect to execution time .*

I. Introduction

Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. Three of the major data mining techniques are regression, classification and clustering. In this research paper we are working only with the clustering because it is most important process, if we have a very large database.

Researcher is using TANGARA tool for clustering. Clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters.

Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval,and bioinformatics. The reserarcher is using TANGARA data mining tools for this purpose. It provides a batter interface to the user than compare the other data mining tools.

II. What is Cluster Analysis?

Cluster analysis is a groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups.

The greater the likeness (or homogeneity) within a group, and the greater the disparity between groups, the —better‖ or more distinct the clustering.

The definition of what constitutes a cluster is not well defined, and, in many applications clusters are not well separated from one another. Nonetheless, most cluster analysis seeks as a result, a crisp classification of the data into non-overlapping groups.

However, the apparent division of the two larger clusters into three sub clusters may simply be an artifact of the human visual system. Finally, it may not be unreasonable to say that the points from four clusters. Thus, we stress once again that the definition of what constitutes a cluster is imprecise, and the best definition depends on the type of data and the desired results.[4]

Data mining often involves the analysis of data stored in a data warehouse. Three of the major data mining techniques are regression, classification and clustering. In this research paper we are working only with the clustering because it is most important process, if we have a very large database. Clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters.

III. Data Mining Tool- TANGARA:

TANAGRA [5] is a free DATA MINING software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. This project is the successor of SIPINA which implements various supervised learning algorithms, especially an interactive and visual construction of decision trees. TANAGRA is more powerful, it contains some supervised learning but also other paradigms such as clustering, factorial analysis, parametric and nonparametric statistics, association rule, feature selection and construction algorithms.

TANAGRA is an "open source project" as every researcher can access to the source code, and add his own algorithms, as far as he agrees and conforms to the software distribution license. The main purpose of Tanagra project is to give researchers and students an easy-to-use data mining software, conforming to the present norms of the software development in this domain (especially in the design of its GUI and the way to use it), and allowing to analyze either real or synthetic data. The second purpose of TANAGRA is to propose to researchers an architecture allowing them to easily add their own data mining methods, to compare their performances.

TANAGRA acts more as an experimental platform in order to let them go to the essential of their work, dispensing them to deal with the unpleasant part in the programmation of this kind of tools: the data management. The third and last purpose, in direction of novice developers, consists in diffusing a possible methodology for building this kind of software. They should take advantage of free access to source code, to look how this sort of software is built, the problems to avoid, the main steps of the project, and which tools and code libraries to use for. In this way, Tanagra can be considered as a pedagogical tool for learning programming techniques. The following figure 1 shows the GUI for Tanagra.

Volume : 3 | Issue : 7 | July 2013 | ISSN - 2249-555X

Fig 1: GUI of Tangara

IV. Dataset

For performing the comparison analysis we need the past project datasets. In this research the reasearcher considered following data set which is available on data repositories. These repositories are very helpful for the researchers. We can directly apply this data in the data mining tools and predict the result.

Source:

Georges Hébrail (georges.hebrail **'@'** edf.fr), Senior Researcher, EDF R&D, Clamart, France Alice Bérard, TELECOM ParisTech Master of Engineering Internship at EDF R&D, Clamart, France

This archive contains 2075259 measurements gathered between December 2006 and November 2010 (47 months). Out of it 65536 instances are considered.

Attribute Information:

1.date: Date in format dd/mm/yyyy
2.time: time in format hh:mm:ss
3.global_active_power: household global minute-averaged active power (in kilowatt)
4.global_reactive_power: household global minute-averaged reactive power (in kilowatt)
5.voltage: minute-averaged voltage (in volt)
6.global_intensity: household global minute-averaged current intensity (in ampere)
7.sub_metering_1: energy sub-metering No. 1 (in watt-hour of active energy). It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas powered).
8.sub_metering_2: energy sub-metering No. 2 (in watt-hour of active energy). It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.
9.sub_metering_3: energy sub-metering No. 3 (in watt-hour of active energy). It corresponds to an electric water-heater and an air-conditioner.

V. Methodology

The researcher taking the past project data from the repositories and apply it on the tools.From the entire dataset the 65536 instances are considered for testing.

Clusters are formed on the basis of attribute Sub_metering1,sub_metering2,sub_metering 3.

In this work K Means, EM clustering and Hierarchical clustering are used for comparison with respect to execution time .

Sr no	Cluster Method	Input	Target
1	K-Means	One or More Continuous	NOne
2	EM-Clustering	One or More Continuous	NOne
3	HAC	One or More Continuous	Possibly one discrete attribute
4	Kohonen--SOM	One or More Continuous	None
5	CT	One or More continuous/ Discrete	One or more continuous
6	CTP	One or More continuous/ Discrete	One or more continuous

Table 1 : Clustering Method supported in TANGARA

In this Paper the first 3 algorithms are considered.

VI. K-MEANS Clustering

In data mining, k-means clustering [6] is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

K-means (Macqueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster.

These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid.

When no point is pending, the first step is completed and an early group age is done. At this point we need to recalculate k new centroids as bar centers of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

VII. EM Clustring

EM algorithm [3] is also an important algorithm of data mining. We used this algorithm when we are satisfied the result of k-means methods. an expectation– maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM [8] iteration alternates between performing an expectation (E) step, which computes the expectation of the log likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

The result of the cluster analysis is written to a band named class indices. The values in this band indicate the class indices, where a value '0' refers to the first cluster; a value of '1' refers to the second cluster, etc.

VIII.Hirarchical Clustering

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

Volume : 3 | Issue : 7 | July 2013 | ISSN - 2249-555X

- Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram.

IX. Result

The three algorithms namely K-means Clustering, EM-Clustering and hierarchical clustering are studied on data set size 65536*9. The clusters are formed on the basis of three attributes as Sub_metering1, sub_metering2, sub_metering3.

Computation Time	2356 ms
Allocated Memory	2259KB

Table 2 :Data Source processing

No.Of Clusters: 8		
Algorithm	Dist.Normalization	Computation Time
K-Means	Variance	6818 ms
EM-Clustering		5554 ms
Hierarchical clustering Algorithm	Variance	78 ms

No.Of Clusters: 16		
Algorithm	Dist.Normalization	Computation Time
K-Means	Variance	15007 ms
EM-Clustering		13853 ms
Hierarchical clustering Algorithm	Variance	125 ms

Table 3 : Result Set for Clustering

X. Conclusion

Tanagra is open source software designed primarily for research use. Tanagra is free, which is definitely a positive for most companies. If our company has the ability to program any extensions required, then Tanagra may be adequate, as it is open source, and we can add any extra functionality we want.

TANGARA contains various clustering algorithms, as CT,CTP,Kmeans,EM-clustering,EM-Selection,HAC,Kohonen-SOM,LVQ,Neighbourhood Graph,VARCLUS,VARHCA, VAR-KMEANS.

The EM-Selection is based on EM-Clustering, LVQ & Neighbourhood graph are the supervised clustering methods. VARCLUS,VARHCA & VARKmeans are based on latent variables. CT & CTP shows the tree structure of clustering. The main algorithm of clustering KMeans,EM-clustering & Hierarchical Clustering algorithms are considered for execution. Each clustering result also shows the result in ANOVA table which test the F-Test for significance test. With the help of this result anybody can interpreted the result for their data/ application.

It has been observed that the there is variation in computation time for same dataset and same attributes for different clustering algorithm. All 3 algorithm takes continuous attribute as a input parameter and none of the target attribute.

From the above result set, concluded that the Kmeans algorithm takes more time for execution as compare to EM-Clustering and EM-clustering requires more time for execution as compare to HCA. Also if the number of clusters are vary the execution time will be varies.

But in all TANGARA GUI is very user friendly, which does not require detail understanding of all data mining technologies. Along with the data Mining technologies it supports statistical methods such as correlation ,regression,Various test..etc.

REFERENCE [1] Yuni Xia, Bowei Xi —Conceptual Clustering Categorical Data with Uncertainty‖ Indiana University – Purdue University Indianapolis,Indianapolis, IN 46202, USA | [2] Sanjoy Dasgupta —Performance guarantees for hierarchical clustering‖ Department of Computer Science and Engineering University of California, San Diego | [3] A. P. Dempster; N. M. Laird; D. B. Rubin —Maximum Likelihood from Incomplete Data via the EM Algorithm‖ Journal of the Roya Statistical Society. Series B (Methodological), Vol. 39, No. 1. (1977), pp.1-38. | [4] Narendra Sharma , Aman Bajpai, Mr. Ratnesh Litoriya-Comparison the various clustering algorithms of WEKA Tool", Department of computer science, Jaypee University of Engg. & Technology,International Journal of Emerging Technology and Advanced Engineering,(ISSN 2250-2459, Volume 2, Issue 5, May 2012,73-80) | [5] http://eric.univ-lyon2.fr/~ricco/tanagra | [6] Jinxin Gao, David B. Hitchcock —James-Stein Shrinkage to Improve K-meansCluster Analysis‖ University of South Carolina,Department of Statistics November 30, 2009 | [7] http://rapid-i.com/ | [8] Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. Computational statistics and data analysis, 14:315–332 | [9] 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. |