IJEMR –May 2013-Vol 3 Issue 5 - Online - ISSN 2249–2585 - Print - ISSN 2249-8672

Data Mining – Partition based clustering approach for improving the similarity and dissimilarity between clusters

*M. Dhevendran

* Head of the Department and Asst. Professor in Computer Science,Meenaakshi Ramasamy Arts and Science College, Thathanur, Ariyalur Dt. Pincode-621804

"Data mining is the process of discovering meaningful new correlation, patterns and trends by sifting through large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques."

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

Data partitioning in data mining is the division of the whole data available into two or three non-overlapping sets: the training set , the validation set , and the test set . If the data set is very large, often only a portion of it is selected for the partitions. Partitioning is normally used when the model for the data at hand is being chosen from a broad set of models. The basic idea of data partitioning is to keep a subset of available data out of analysis, and to use it later for verification of the model.

For example, a company that sale a variety of products may need to know about the sale of all of their products in order to check that what product is giving extensive sale and which is lacking. This is done by data mining techniques. But if the system clusters the products that are giving less sale then only the cluster of such products would have to be checked rather than comparing the sales value of all the products. This is actually to facilitate the mining process

A cluster is an ordered list of objects, which have some common characteristics. The criterion for checking the similarity is implementation dependent. Clustering is often confused with classification, but there is some difference between the two. In classification the objects are assigned to pre defined classes, whereas in clustering the classes are also to be defined.

Precisely, Data Clustering is a technique in which, the information that is logically similar is physically stored together. In order to increase the efficiency in the database systems the numbers of disk accesses are to be minimized. In clustering the objects of similar properties are placed in one class of objects and a single access to the disk makes the entire class available.

Types of Clustering Methods

There are many clustering methods available, and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired, the known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset. In general, clustering methods may be divided into two categories based on the cluster structure which they produce. The non-hierarchical methods divide a dataset of N objects into M clusters, with or without overlap.

These methods are sometimes divided into partitioning methods, in which the classes are mutually exclusive, and the less common clumping method, in which overlap is allowed. Each object is a member of the cluster with which it is most similar; however the threshold of similarity has to be defined. The hierarchical methods produce a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains. The hierarchical methods can be further divided into agglomerative or divisive methods. In agglomerative methods, the hierarchy is build up in a series of N-1 agglomerations, or Fusion, of pairs of objects, beginning with the un-clustered dataset. The less common divisive methods begin with all objects in a single cluster and at each of N-1 steps divide some clusters into two smaller clusters, until each object resides in its own cluster.

Some of the important Data Clustering Methods are described below.

❖ Partitioning Methods

The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster.
Single Pass: A very simple partition method, the single pass method creates a partitioned dataset as follows:

Make the first object the centroid for the first cluster. For the next object, calculate the similarity, S, with each existing cluster centroid, using some similarity coefficient.

If the highest calculated S is greater than some specified threshold value, add the object to the corresponding cluster and re determine the centroid; otherwise, use the object to initiate a new cluster. If any objects remain to be clustered, return to step 2.

As its name implies, this method requires only one pass through the dataset; the time requirements are typically of order O (NlogN) for order O (logN) clusters. This makes it a very efficient clustering method for a serial processor. A disadvantage is that the resulting clusters are not independent of the order in which the documents are processed, with the first clusters formed usually being larger than those created later in the clustering run

- Hierarchical Agglomerative methods

The hierarchical agglomerative clustering methods are most commonly used. The construction of a hierarchical agglomerative classification can be achieved by the following general algorithm.

Find the 2 closest objects and merge them into a cluster

Find and merge the next two closest points, where a point is either an individual object or a cluster of objects.

If more than one cluster remains, return to step 2

Individual methods are characterized by the definition used for identification of the closest pair of points, and by the means used to describe the new cluster when two clusters are merged.

There are some general approaches to implementation of this algorithm, these being stored matrix and stored data, are discussed below

In the second matrix approach , an N*N matrix containing all pairwise distance values is first created, and updated as new clusters are formed. This approach has at least an O(n*n) time requirement, rising to O(n3) if a simple serial scan of dissimilarity matrix is used to identify the points which need to be fused in each agglomeration, a serious limitation for large N.

The stored data approach required the recalculation of pairwise dissimilarity values for each of the N-1 agglomerations, and the O(N) space requirement is therefore achieved at the expense of an O(N3) time requirement.

❖ The Single Link Method (SLINK)

The single link method is probably the best known of the hierarchical methods and operates by joining, at each step, the two most similar objects, which are not yet in the same cluster. The name single link thus refers to the joining of pairs of clusters by the single shortest link between them.

❖ The Complete Link Method (CLINK)

The complete link method is similar to the single link method except that it uses the least similar pair between two clusters to determine the inter-cluster similarity (so that every cluster member is more like the furthest member of its own cluster than the furthest item in any other cluster). This method is characterized by small, tightly bound clusters.

- The Group Average Method

The group average method relies on the average value of the pair wise within a cluster, rather than the maximum or minimum similarity as with the single link or the complete link methods. Since all objects in a cluster contribute to the inter –cluster similarity, each object is , on average more like every other member of its own cluster then the objects in any other cluster.

- Text Based Documents

In the text based documents, the clusters may be made by considering the similarity as some of the key words that are found for a minimum number of times in a document. Now when a query comes regarding a typical word then instead of checking the entire database, only that cluster is scanned which has that word in the list of its key words and the result is given. The order of the documents received in the result is dependent on the number of times that key word appears in the document.

In order to decide which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a measure of dissimilarity between sets of observations is required. In most methods of hierarchical clustering, this is achieved by use of an appropriate metric (a measure of distance between pairs of observations), and a linkage criterion which specifies the dissimilarity of sets as a function of the pair wise distances of observations in the sets.

Similarity-based Cluster definition

A cluster is a set of objects that are “similar”, and objects in other clusters are not “similar.” A variation on this is to define a cluster as a set of points that together create a region with a uniform local property, e.g., density or shape.

Measures (Indices) of Similarity and Dissimilarity
The notion of similarity and dissimilarity (distance) seems fairly intuitive. However, the quality the quality of a cluster analysis depends critically on the similarity measure that is used and, as a consequence, hundreds of different similarity measures have been developed for various situations. The discussion here is necessarily brief.

Attribute types and Scales
The proximity measure (and the type of clustering used) depends on the attribute type and scale of the data.

Binary	Two values, e.g., true and false.
Discrete	A finite number of values, or integers, e.g., counts.
Continuous	An effectively infinite number of real values, e.g., weight.

Case Study

Clusters of computer systems have been built and used for over a decade. Pfister defines a cluster as "a parallel or distributed system that consists of a collection of interconnected whole computers,that is utilized as a single, unified computing resource". In general, the goal of a cluster is to make it possible to share a computing load over several systems without either the users or system administrators needing to know that more than one system is involved.

We describe the architecture of the clustering extensions to the Windows NT operating system. Windows NT clusters provide three principal user visible advantages: improved availability by continuing to provide a service even during hardware or software failure. If any component in the system, hardware or software fails the user may see degraded performance, but will not lose access to the service Increased scalability by allowing new components to be added as system load increases. Lastly, clusters simplify the management of groups of systems and their applications by allowing the administrator to manage the entire group as a single system.

In Windows NT Environment, all the nodes (the terminals) that are using the same resources are accumulated in one cluster. So the similarity among the cluster members is the usage of the same kind of resources at one time. Now only one group of monitoring program can monitor the entire cluster as one single node.Manually, clustering can also be established by running a clustering program at a node. Now this node will be the first one in the cluster and all the other nodes that will be entered should use the same resources.
Windows NT Clusters are, in general, shared nothing clusters. This means that while several systems in the cluster may have access to a device or resource, it is effectively owned and managed by a single system at a time.

Members of a cluster are referred to as nodes or systems. The Cluster Service is the collection of software on each node that manages all cluster specific activity. Cluster service is a separate, isolated set of components.

Services in a Windows NT cluster are exposed as virtual servers. Client Workstations believe they are connecting with a physical system, but are in fact, connecting to a service which may be provided by one of several systems. Clients create a TCP/IP session with a service in the cluster using a known IP address. In the event of a failure the cluster service will "move" the entire group to another system. The client will detect a failure in the session and reconnect in exactly the same manner as the original connection. The IP address is now available on another machine and the connection will be quickly re-established. The following things work on top of clustering in Windows NT Environment.

The Node Manager Handles cluster membership, watches the health of other cluster systems.

Configuration Database Manager maintains the cluster configuration database. Resource Manager/Failover Manager makes all resource management decisions and initiates appropriate actions, such as startup, restart and failover.

Event Processor connects all of the components of the Cluster Service, handles common operations and controls Cluster Service initialization. Communications Manager manages communications with all other nodes of the cluster.
Global Update Manager - provides a global update service that is used by other components within the Cluster Service.

Creating a Cluster:When a system administrator wishes to create a new cluster, the administrator will run a cluster installation utility on the system to become the first member of the cluster. For a new cluster, the database is created and the initial cluster member is added. The administrator will then configure any devices that are to be managed by the cluster software. We now have a cluster with a single member. In the next step of clustering each node is added to the cluster by means of similarity on the basis of the resources used. The new node automatically receives a copy of the existing cluster database.

Joining a Cluster:Following a restart of a system, the cluster service is started automatically. The system configures and mounts local, non-shared devices. Cluster-wide devices must be left offline while booting because another node may be using them. The system uses a 'discovery' process to find the other members of the cluster.

Leaving a Cluster: When leaving a cluster, a cluster member will send a Cluster Exit message to all other members on the cluster, notifying them of its intent to leave the cluster.The exiting cluster member does not wait for any responses and immediately proceeds to shutdown all resources and close all connections managed by the cluster software.

Sending a message to the other systems in the cluster when leaving saves the other systems from discovering the absence and having to go to a regroup effort to re-establish the membership.

References

➢ Athman Bouguettaya "On Line Clustering", IEEE Transaction on Knowledge and Data Engineering Volume 8, No. 2, April 1996.

➢ Euripides G.M. Petrakis and Christos Faloutsos "Similarity Searching in Medical Image Databases", IEEE Transaction on Knowledge and Data Engineering Volume 9, No. 3, MAY/JUNE 1997.

➢ Rob Short, Rod Gamache, John Vert and Mike Massa "Windows NT Clusters for Availability and Scalability" Microsoft Online Research Papers, Microsoft Corporation.

➢ Jim Gray "QqJim Gray's NT Clusters Research Agenda" Microsoft Online Research Papers, Microsoft Corporation.

➢ Bruce Moxon "Defining Data Mining, The Hows and Whys of Data Mining, and How It Differs From Other Analytical Techniques" Online Addition of DBMS Data Warehouse Supplement, August 1996.

➢ Willet, Peter "Parallel Database Processing, Text Retrieval and Cluster Analyses" Pitman Publishing, London, 1990.

➢ The Challenges of Clustering High Dimensional Data(Michael Steinbach, Levent Ertöz, and Vipin Kumar)