View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Librarians' Digital Library

DRTC Workshop on
Semantic Web
8th – 10th December, 2003
DRTC, Bangalore

Paper: K

Data Mining and Clustering Techniques

I. K. Ravichandra Rao

Professor and Head

Documentation Research and Training Center

Indian Statistical Institute

Bangalore

ikrrao@hotmail.com

Abstract

Data mining techniques are most useful in information retrieval; some of these techniques are classification, association rules and clustering. An attempt has been made here to discuss these techniques.

1. Introduction

Data mining, a synonym to "knowledge discovery in databases" is a process of analyzing data from different perspectives and summarizing it into useful information. It is a process that allows users to understand the substance of relationships between data. It reveals patterns and trends that are hidden among the data. It is often viewed as a process of extracting valid, previously unknown, non-trivial and useful information from large databases. Data mining systems can be classified according to the kinds of databases mined, the kinds of knowledge mined, the techniques used or the applications. Three important components of data mining systems are databases, data mining engine, and pattern evaluation modules.

Data mining engine is ideally consists of a set of functional modules for tasks such as characterization, association, classification, cluster analysis, and evolution. On the other hand, pattern evaluation module typically employs certain measures and interacts with the data mining modules so as to focus the search towards unknown patterns. For example, the measures for association rules are "support" and "confidence". The support is the percentage of task-relevant data tuples for which rule patterns appears. The confidence is an estimate of the strength of the implication of the rule. The measures vary from techniques to techniques.

Data are usually associated with classes or concepts. For example, in an electronic shop, classes of items for sale may be cell phone, computer, TV, etc. The concept of customers includes rich and poor classes of customers. In analysis of sales, for example, it may be useful to describe individual classes and concepts in summarized, concise, and yet precise terms. Such descriptions of a class or concept are called class / concept descriptions. These descriptions can be derived via

- Data characterization: by summarizing the data of the class under study (-- target class) in general terms
- Data discrimination: by comparison of the target class with one or a set of comparative classes (-- contrasting classes)
- Both data characterization and discrimination

The main objective of this lecture note is to discuss a few important methods and techniques of data mining with emphasis on clustering techniques.

2. Data Mining Techniques

Classification is a most important and frequently used technique in data mining. It is a process of finding a set of models that describe and distinguish data classes or concepts. The derived model may be represented in various forms such as classification (IF-THEN) rules, decision tree, neural networking, etc.

A decision tree is a flowchart like tree structure when each node denotes a test on an attribute value where each branch represents an outcome of the test, and tree leaves represent classes. Decision trees can be easily converted to classification rules. A neural network when used for classification is typically a collection of neuron-like

processing units with weighted connections between the units. While learning classification rules the system has to find the rules that predict the class from the prediction attributes. So firstly the user has to define conditions for each class, the data mine system then constructs descriptions for the classes.

Once classes are defined the system should infer rules that govern the classification. There fore the system should be able to find the description of the each class. The description should only refer to the prediction attributes of the training set so that the positive examples should satisfy the description. A rule is said to be correct if its description covers all the positive examples and none of the negative examples of a class.

2.1. Association Analysis

Association analysis is the discovery of association rules sharing attribute-value conditions that occur frequently together in a given set of data. It is widely used in the context of analysis of "transaction data." Association rules are of the form (3)

X => Y

i.e. $A_1 \wedge A_2 \wedge \ldots \wedge A_m \Rightarrow B_1 \wedge B_2 \wedge B_3 \ldots B_n$, where A_i (for $i \in \{1, 2, \ldots m\}$) and B_J (for $j \in \{1, 2, 3, \ldots n\}$) are attribute-value pairs. This rule is interpreted, as "database tuples that satisfy the conditions in X are also likely to satisfy the conditions in Y." For example,

age (x, "25 ….. 35") ^ income (x, 15k …. 25k) =>

buys (x, "Cell phones'): support 2%, confidence 60%.

i.e. the rule indicates that of the customers under study, 2% are 25 to 35 years of age with an income of 15k to 25k and have purchased a cell phone in a shop; there is a 60% confidence or certainty that a customer in the said age and income group will purchase a cell phone.

An objective measure for association rules of the form x => y is rule support, representing the percentage of transactions from a transaction database that the given rule satisfies; i.e. P (X $\cup$ Y), where X $\cup$ Y indicates that a transaction contains both X and Y – the union of items sets X and Y (in the context of set theory it is X $\cap$ Y!). Another objective measure for association rules is confidence, which assesses the degree of certainty of the identified association, i.e., P (Y|X) – probability that a transaction containing X also contains Y. Thus, support and confidence are defined as:

$$\text{support } (X \Rightarrow Y) = P(X \cup Y) = \frac{\textit{No. of tuple containing both A \& B}}{\textit{Total no. of tuples}}$$

$$\text{confidence } (X \Rightarrow Y) = P(Y \mid X) = \frac{\textit{No.of tuples containing both A \& B}}{\textit{No. of tuples containing A}}$$

2.2 Data Mining Standards

There are several established and emerging standards related to data mining.

These standards are for different components of the data mining systems. For instance, they are for (1)

- Models – to represent data mining and statistical data; for producing, displaying and for using the models, for analyzing and mining remote and distributed data.

- Attributes – to represent the cleaning, transforming and aggregating of attributes used as input in the models

- Interfaces – to link to other languages and systems

- Setting – to represent the internal parameters required for building and using the models

3. Cluster Analysis

The concept of clustering has been around for a long time. It has several applications, particularly in the context of information retrieval and in organizing web resources. The main purpose of clustering is to locate information and in the present day context, to locate most relevant electronic resources. The research in clustering eventually led to automatic indexing --- to index as well as to retrieve electronic records. Clustering is a method in which we make cluster of objects that are some how similar in characteristics. The ultimate aim of the clustering is to provide a grouping of similar records. 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 formed. The term “class” is in fact frequently used as synonym to the term “cluster”.

In database management, data clustering is a technique in which, the information that is logically similar is physically stored together. In order to increase the efficiency of search and the retrieval in database management, the number of disk accesses is to be minimized. In clustering, since the objects of similar properties are placed in one class of objects, a single access to the disk can retrieve the entire class. If the clustering takes place in some abstract algorithmic space, we may group a population into subsets with similar characteristic, and then reduce the problem space by acting on only a representative from each subset. Clustering is ultimately a process of reducing a mountain of data to manageable piles. For cognitive and computational simplification, these piles may consist of "similar" items.

There are two approaches to document clustering, particularly in information retrieval; they are known as term and item clustering. Term clustering is a method, which groups redundant terms, and this grouping reduces, noise and increase frequency of assignment. If there are fewer clusters than there were original terms, then the dimension is also reduced. However semantic properties suffer. There are many different algorithms available for term clustering. These are cliques, single link, stars and connected components.

Cliques require all items in a cluster to be within the threshold of all other items. In

single link clustering the strong constraint that every term in a class is similar to every other term is relaxed. The rule to generate single link clusters is that any term that is similar to any other term in the cluster can be added to the cluster. The star technique selects a term and then places in the class all terms that are related to that term (i.e. in effect a star with the selected term as the core). Terms not yet in classes are selected as new seeds until all terms are assigned to a class. There are many different classes that can be created using the star technique.

Item clustering, on the other hand, assists the user in identifying relevant items**.** It is used in two ways:

1 Directly find additional items that might not have been found by the query and to serve as a basis for visualization of the Hit file. Each item cluster has a common semantic basis containing similar terms and thus similar concepts.

1. To assist the user in understanding the major topics resulting from a search, the items retrieved to be clustered and used to create a visual (e.g., graphical) representation of the clusters and their topics. This allows a user to navigate between topics, potentially showing topics the user had not considered. The topics are not defined by the query but by the text of the items retrieved.

When items in the database have been clustered, it is possible to retrieve all of the items in a cluster, even if the search statement did not identify them. When the user retrieves a strongly relevant item, the user can look at other items like it without issuing another search. When relevant items are used to create a new query (i.e., relevance feedback), the retrieved hits are similar to what might be produced by a clustering algorithm.

How ever, term clustering and item clustering in a sense achieve the same objective even though they are the inverse of each other. The objective of both is to determine additional relevant items by a co-occurrence process. For all of the terms within the same cluster, there will he significant overlap of the set of items they are found in. Item clustering is based upon the same terms being found in the other items in the cluster. Thus the set of items that caused a term clustering has a strong possibility of being in the same item cluster based upon the terms. For example, if a term cluster has 10 terms in it (assuming they are closely related), then there will be a set of items where each item contains major subsets of the terms. From the item perspective, the set of items that has the commonality of terms has a strong possibility to be placed in the same item cluster.

3.1 Definitions

In this section some frequently used terms are defined.

3.1.1 Cluster

A cluster is an ordered list of objects, which have some common objects. The objects belong to an interval [a,b].

3.1.2 Distance between Two Clusters

The distance between two clusters involves some or all elements of the two clusters. The clustering method determines how the distance should be computed. The

distance between two points is taken as a common metric to assess the similarity among the components of a population. The most commonly used distance measure is the **Euclidean metric** which defines the distance between two points p = (p1,p2,……) and q = (q1,q2, ….) as

$$d = [\ \sum(p_i - q_i)^2]^{1/2}$$

3.1.3 Similarity Measures

A similarity measure SIMILAR (D_i, D_j) can be used to represent the similarity between two documents i and j. Typical similarity generates values of 0 for documents exhibiting no agreement among the assigned indexed terms, and 1 when perfect agreement is detected. Intermediate values are obtained for cases of partial agreement

3.1.4 Threshold

The lowest possible input value of similarity required joining two objects in one cluster. A threshold T(J) is given for the Jth variable ($1 \leq J \leq N$). Cases are partitioned into clusters so that within each cluster the Jth variable has a range less than T(J). The thresholds should be chosen fairly large, especially if there are many variable. The procedure is equivalent to converting each variable to a category variable (using the thresholds to define the categories) and the clusters are then cells of the multidimensional contingency table between all variables.

3.1.5 Similarity Matrix

Similarity between objects calculated by the function SIMILAR (D_i,,D_j), represented in the form of a matrix is called a similarity matrix.

3.1.6 Cluster Seed

First document or object of a cluster is defined as the initiator of that cluster i.e. every incoming object's similarity is compared with the initiator. The initiator is called the cluster seed.

3.2 Characteristics of the Classes

A well-defined semantic definition should exist for each class. There is a risk that the name assigned to the semantic definition of the class could also be misleading. In some systems numbers are assigned to classes to reduce the misinterpretation that a name attached to each class could have. A clustering of items into a class called "computer" could mislead a user into thinking that it includes items on main memory that may actually reside in another class called "hardware."

The size of the classes should be within the same order of magnitude. One of the primary uses of the classes is to expand queries or expand the resultant set of retrieved items. If a particular class contains 90 per cent of the objects, that class is not useful for either purpose. It also places in question the utility of the other classes that are distributed across 10 per cent of the remaining objects.

Within a class, one object should not dominate the class. For example, assume a thesaurus class called "computer" exists and it contains the objects "microprocessor," "286-processor,"

"386- processor" and "Pentium." If the term "microprocessor" is found 85 per cent of the time and the other terms are used 5 per cent each, there is a strong possibility that using "microprocessor" as a synonym for "286- processor" will introduce too many errors. It may be better to place, microprocessor" into its own class.

Decision about the Single/multiple class: Whether an object can be assigned to multiple classes or just one must be decided at creation time. This is a tradeoff based upon the specificity and partitioning capability of the semantics of the objects. Given the ambiguity of language in general, it is better to allow an object to be in multiple classes rather than limited to one.

4. Basic Clustering Step

4.1 Preprocessing and feature selection

Most clustering models assume that n-dimensional feature vectors represent all data items. This step therefore involves choosing an appropriate feature, and doing appropriate preprocessing and feature extraction on data items to measure the values of the chosen feature set. It will often be desirable to choose a subset of all the features available, to reduce the dimensionality of the problem space. This step often requires a good deal of domain knowledge and data analysis.

4.2 Similarity measure

Similarity measure plays an important role in the process of clustering where a set of objects are grouped into several clusters, so that similar objects will be in the same cluster and dissimilar ones in different cluster. In clustering, its features represent an object and the similarity relationship between objects is measured by a similarity function. This is a function, which takes two sets of data items as input, and returns as output a similarity measure between them.

4.3 Clustering algorithm

Clustering algorithms are general schemes, which use particular similarity measures as subroutines. The particular choice of clustering algorithms depends on the desired properties of the final clustering, e.g. what are the relative importance of compactness, parsimony, and inclusiveness? Other considerations include the usual time and space complexity. A clustering algorithm attempts to find natural groups of components (or data) based on some similarity. The clustering algorithm also finds the centroid of a group of data sets. To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroids. The output from a clustering algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster (2).

4.4 Result validation

Do the results make sense? If not, we may want to iterate back to some prior stage. It may also be useful to do a test of clustering tendency, to try to guess if clusters are present at all; note that any clustering algorithm will produce some clusters regardless of whether or not natural clusters exist.

4.5 Result interpretation and application.

Typical applications of clustering include data compression (via representing data samples by their cluster representative), hypothesis generation (looking for patterns in the clustering of data), hypothesis testing (e.g. verifying feature correlation or other data properties through a high degree of cluster formation), and prediction (once clusters have been formed from data and characterized, new data items can be classified by the characteristics of the cluster to which they would belong).

5. Clustering Techniques

Traditionally clustering techniques are broadly divided into hierarchical and partitioning. Hierarchical clustering is further subdivided into agglomerative and divisive.

a) ***Agglomerative:*** Start with the points as individual clusters and, at each step, merge the most similar or closest pair of clusters. This requires a definition of cluster similarity or distance.

b) ***Divisive:*** Start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain. In this case, we need to decide, at each step, which cluster to split and how to perform the split.

Hierarchical techniques produce a nested sequence of partitions, with a single, all-inclusive cluster at the top and singleton clusters of individual points at the bottom. Each intermediate level can be viewed as combining two clusters from the next lower level (or splitting a cluster from the next higher level). The result of a hierarchical clustering algorithm can be graphically displayed as tree, called a dendogram. This tree graphically displays the merging process and the intermediate clusters. For document clustering, this dendogram provides a taxonomy, or hierarchical index. The traditional agglomerative hierarchical clustering procedure is as follows:

5.1 Profile Algorithm

It is one of the simplest algorithms. The profile technique simultaneously plots several variables. It is useful in giving a feeling for the numbers without commitment to any mode of analysis. It is especially useful in clustering – it suggests possible clusters of similar variables. It is sometimes necessary before clustering to decide the weights to be given to different variables and profiles may suggest reasonable weights. Profiles are best described as histogram as each variable connected between variables by identifying cases usually the case name ignored in plotting a single histogram.

5.2 Steps in Profile algorithm

Step 1. Choose a symbol for each case, preferably one or two characters preferably mnemonic so that the case can readily be identified from its symbol.

Bangalore – BA, Kolkatta – KO, Mumbai – MU, Chennai – CH, etc.

Step 2A. For each variable, plot the cases along a horizontal line, identifying each case by its symbol. If a number of cases have identical values, their symbols should be placed vertically over this value as in a histogram.

Step 2B. The horizontal scale for each variable is initially set so that the minima for different variables coincide and the maxima coincide, approximately

Step 2C. Vertical positions of the horizontal scales for each variable are assigned so that similar variables are in adjacent rows.

Step 3. A profile for connecting the symbols for the case in the various horizontal draws each case scales, one for each variable.

Step 4. Rescale and reposition the variables to make the case profile smoother.

Step 5. Clusters of cases will correspond to profiles of similar appearance and clusters of variables will be positioned closely together.

5.3 Simple Agglomerative Clustering Algorithm

This method involves:

1. Compute the similarity between all pairs of clusters, i.e., calculate a similarity matrix whose ij^{th} entry gives the similarity between the i^{th} and i^{th} clusters.

2. Merge the most similar (closest) two clusters, considering some thresholds.

3. Update the similarity matrix to reflect the pair wise similarity between the new clusters and the original clusters.

4. Repeat steps 2 and 3 until only a single cluster remains.

In contrast to hierarchical techniques, partitioned clustering techniques create a one-level partitioning of the data points. If K is the desired number of clusters, then partitioned approaches typically find all K clusters at once. There are a number of partitioned techniques, but we shall only describe the K-means algorithm. It is based on the idea that a center point can represent a cluster. In particular, for K-means we use the notion of a centroid, which is the mean or median point of a group of points. Note that a centroid almost never corresponds to an actual data point. The basic K-means clustering technique is presented below.

5.3 Basic K-means Algorithm for finding K clusters

This method involves:

1. Select K points as the initial centroids.

2. Assign all points to the closest centroid.

3. Recompute the centroid of each cluster.

4. Repeat steps 2 and 3 until the centroids don’t change.

6. Evaluation of Cluster Quality

For clustering, two measures of cluster "goodness" or "quality" are used. One type of measure compares different sets of clusters without reference to external knowledge and is called an *internal quality* measure. The "overall similarity" measure is based on the pair wise similarity of documents in a cluster. The other type of measures evaluates how well the clustering is working by comparing the groups produced by the clustering techniques to known classes. This type of measure is called an *external quality* measure. One external measure is entropy], which provides a measure of "goodness" for un -nested clusters or for the clusters at one level of a hierarchical clustering. Another external measure is the F-measure. It is more oriented toward measuring the effectiveness of a hierarchical clustering. There are many different quality measures and the performance and relative ranking of different clustering algorithms can vary substantially depending on which measure is used. However, if one clustering algorithm performs better than other clustering algorithms on many of these measures, then we can have some confidence that it is truly the best clustering algorithm for the situation being evaluated

7. Conclusion

Clustering has a number of applications in every field of life. We are applying this technique whether knowingly or unknowingly in day-to-day life. One has to cluster a lot of thing on the basis of similarity either consciously or unconsciously. So the history of data clustering is old as the history of mankind. In computer field also, use of data clustering has its own value. Especially in the field of information retrieval data clustering plays an important role. Now the importance of clustering is being seen in the current digital environment especially in information retrieval, image indexing and searching, data mining, networking, GIS, web searching and retrieval etc.

Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models, such as demographic-based customer segmentation. Additional analyses using standard analytical and other data mining techniques can determine the characteristics of these segments with respect to some desired outcome. For example, the buying habits of multiple population segments might be compared to determine which segments to target for a new sales campaign.

For example, a company that sells 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 fewer sales 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.

8. References

1. Grossman, Robert. L., Hornick, Mark. F. and Meyer, Gregor. "Data Mining Standards Initiatives." (Communications of the ACM, Vol. 45, No. 8, 2002, 59-61)

2. Hartigan, John A, "Clustering Algorithms". 1975. John Wiley. New York.

3. Han Jiawei and Kamber, Micheline. "Data Mining: Concepts and Techniques". 2001. Morgon Kaufmann. Sanfransico, CA.