International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 3, 2013. ISSN: 2348-6600
Reference ID: IJCS-019 PAGE NO: 099-104

An Overview of Data Mining Techniques and Clustering concepts

Dr.S.S.Dhenakaran [#1], M.Sathish Kumar [*2]

[#1] Assistant professor, Department of Computer science and Engineering,
Alagappa University, Karaikudi.
India.
[*2] Department of Computer science and Engineering,
Alagappa University, Karaikudi.
India.

Abstract— **This paper provides an overview of Data mining is used to extract meaningful information and to develop significant relationships among variables stored in large data set/data warehouse. In the case study reported in this paper, a data mining approach is applied to extract knowledge from a data set. We present techniques for the discovery of patterns hidden in large data sets, focusing on issues relating to their feasibility, usefulness, effectiveness, and scalability. Fast retrieval of the relevant information from databases has always been a significant issue. There are many techniques are developed for this purpose; In among data clustering is one of the major technique. The process of creating vital information from a huge amount of data is learning. It can be classified into two such as supervised learning and unsupervised learning. Clustering is a kind of unsupervised data mining technique. It describes the general working behavior, the methodologies followed by these approaches and the parameters which affect the performance of these algorithms. In classifying web pages, the similarity between web pages is a very important feature. The main objective of this paper is to gather more core concepts and techniques in the large subset of cluster analysis.**

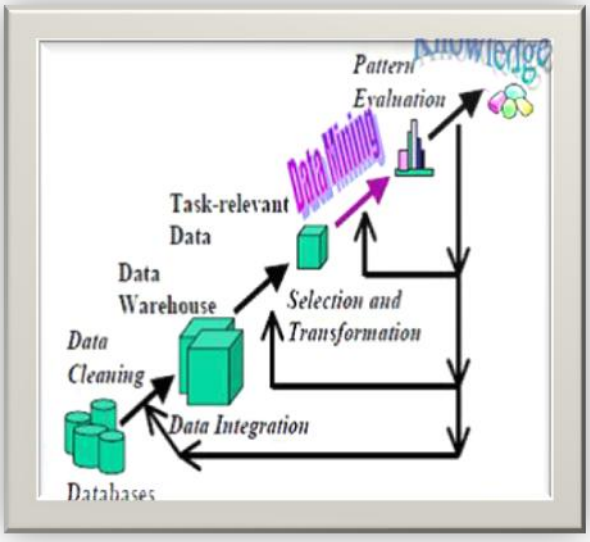
Keywords: Data mining Clustering, Unsupervised Learning Web Pages, Classifications

I. INTRODUCTION

Data mining is a process to extract the implicit information and knowledge which is potentially useful and people do not know in advance, and this extraction is from the mass, incomplete, noisy, fuzzy and random data [2].

The essential difference between the data mining and the traditional data analysis (such as query, reporting and on-line application of analysis) is that the data mining is to mine information and discover knowledge on the premise of no clear assumption [1]

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.



Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 3, 2013. ISSN: 2348-6600
Reference ID: IJCS-019 PAGE NO: 099-104

Fig.1: Data mining is the core of Knowledge Discovery Process

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure (Figure 1.1) shows data mining as a step in an iterative knowledge discovery process. The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge.

II. Data Mining Algorithms and Techniques

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is

acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis.

The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

Types of clustering methods

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

Predication

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict.

Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 3, 2013. ISSN: 2348-6600
Reference ID: IJCS-019 PAGE NO: 099-104

necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.

Types of regression methods

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

Association rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behaviour analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally

very large and a high proportion of the rules are usually of little (if any) value.

Types of association rule

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

Neural networks

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries.

Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

III. Clustering concepts

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achievessimplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis.

From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing,medical diagnostics and many others.

Clustering is the subject of active research in several fields such as statistics, pattern recognition, and machine learning. This survey focuses on clustering in data mining. Data mining adds to clustering the complications of very large datasets with very many attributes of different types. This imposes unique computational requirements on relevant clustering algorithms. A variety of algorithms have recently emerged that meet these requirements and were successfully applied to real-life data mining problems.

Classification of Clustering Algorithms,Categorization of clustering algorithms is neither straightforward, nor canonical.

Clustering Algorithms

- Hierarchical Methods
  - Agglomerative Algorithms
  - Divisive Algorithms

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 3, 2013. ISSN: 2348-6600
Reference ID: IJCS-019 PAGE NO: 099-104

- Partitioning Methods
  - Relocation Algorithms Probabilistic Clustering
  - K-medoids Methods
  - K-means Methods
- Density-Based Algorithms
  - Density-Based Connectivity Clustering
  - Density Functions Clustering
- Grid-Based Methods
- Methods Based on Co-Occurrence of Categorical Data
- Constraint-Based Clustering
- Clustering Algorithms Used in Machine Learning
  - Gradient Descent and Artificial Neural Networks
  - Evolutionary Methods
- Scalable Clustering Algorithms
- Algorithms For High Dimensional Data
  - Subspace Clustering
  - Projection Techniques
  - Co-Clustering Techniques

Hierarchical Clustering

Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendrogram. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows exploring data on different levels of granularity. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down) [Jain & Dubes 1988; Kaufman & Rousseeuw 1990]. An agglomerative clustering starts with one-point (singleton) clusters and recursively merges two or more most appropriate clusters. A divisive
clustering starts with one cluster of all data points and recursively splits the most appropriate cluster. The process continues until a stopping criterion (frequently, the requested number k of clusters) is achieved.

Advantages of hierarchical clustering include:

- Embedded flexibility regarding the level of granularity
- Ease of handling of any forms of similarity or distance
- Consequently, applicability to any attribute types.

Disadvantages of hierarchical clustering are related to:

- Vagueness of termination criteria
- The fact that most hierarchical algorithms do not revisit once constructed (intermediate) clusters with the purpose of their improvement.

In hierarchical clustering our regular point-by-attribute data representation is sometimes of secondary importance. Instead, hierarchical clustering frequently deals with the matrix of distances (dissimilarities) or similarities between training points. It is sometimes called connectivity matrix. Linkage metrics are constructed (see below) from elements of this matrix. The requirement of keeping such a large matrix in memory is unrealistic. To relax this limitation different devices are used to introduce into the connectivity matrix some sparsity. This can be done by omitting entries smaller than a certain threshold, by using only a certain subset of data representatives, or by keeping with each point only a certain number of its nearest neighbors. For example, nearest neighbor chains have decisive impact on memory consumption [Olson 1995]. A sparse matrix can be further used to represent intuitive concepts of closeness and connectivity. Notice that the way we process original (dis)similarity matrix and construct a linkage metric reflects our a priori ideas about the data model.

Partition Clustering Algorithm

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

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 3, 2013. ISSN: 2348-6600
Reference ID: IJCS-019 PAGE NO: 099-104

some sort of a summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In cases where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases; e.g., a cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of clusters is large, the centroids can be further clustered to produce a hierarchy within a dataset. The partition clustering algorithm splits the data points into k partition, where each partition represents a cluster. The partition is done based on certain objective functions. One such criterion function is minimizing the square error criterion which is computed as, $E = \Sigma \Sigma \mid\mid p - mi \mid\mid 2$ (2) Where p is the point in a cluster and mi is the mean of the cluster. The cluster should exhibit two properties; they are (1) each group must contain at least one object (2) each object must belong to exactly one group. The main drawback of this algorithm [3] is whenever a point is close to the center of another cluster; it gives a poor result due to overlapping of the data points. Single Pass is a very simple partition method; it creates a partitioned dataset in three steps. First, it makes the first object the centroid for the first cluster. For the next object, it calculates the similarity, S, with each existing cluster centroid, using some similarity coefficient. Finally, if the highest calculated S is greater than some specified threshold value, it adds the object to the corresponding cluster, and re determines the centroid; otherwise, it uses the object to initiate a new cluster. If any objects remain to be clustered, it returns to step 2. As its name implies, this method requires only one pass through the dataset; the time requirements are typically of the 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. K- Means Clustering Algorithm is one of the partition based clustering algorithms. The advantages of the simple K means algorithm. That it is easy to implement and works with any of the standard norms. It allows straight forward parallelization; and it is insensitive with respect to data ordering. The disadvantages of the K means algorithm are as follows. The results strongly depend on the initial guess of the centroids. The local optimum (computed for a cluster) does not need to be a global optimum (overall clustering of a data set). It is not obvious what the good number K is in each case, and the process is, with respect to the outlier.

Density based Clustering Algorithm

The density based algorithm allows the given cluster to continue to grow as long as the density in the neighbor hood exceeds a certain threshold [4]. This algorithm is suitable for handling noise in the dataset. The following points are enumerated as the features of this algorithm: it handles clusters of arbitrary shape, Handles noise, needs only one scan of the input dataset, and the density parameters to be initialized. DBSCAN, DENCLUE and OPTICS [4] are examples of this algorithm.

Density-Based Connectivity

The crucial concepts of this section are density and connectivity, both measured in terms of local distribution of nearest neighbors. The algorithm DBSCAN (Density Based Spatial Clustering of Applications with Noise) targeting low-dimensional spatial data is the major representative of this category. Two input parameters and Min Pts are used to define: 1) An -neighborhood ( ) { | (,) } N x = y_ X d x y _ of the point x, 2) A core object (a point with a neighborhood consisting of more than Min Pts points) 3) A concept of a point y density-reachable from a core object x (a finite sequence of core objects between x and y exists such that each next belongs to an - neighborhood of its predecessor) 4) A density-connectivity of two points x, y (they should be density-reachable from a common core object).

Density Functions

Hinneburg & Keim [1998] shifted the emphasis from computing densities pinned to data points to computing density functions defined over the underlying attribute space. They proposed the algorithm DENCLUE (DENsity-based

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 3, 2013. ISSN: 2348-6600
Reference ID: IJCS-019 PAGE NO: 099-104

CLUstEring). Along with DBCLASD, it has a firm mathematical foundation. DENCLUE uses a density function. f D(x) = y€ D f (x, y) (1) That is the superposition of several influence functions. When the f-term depends on x y, the formula can be recognized as a convolution with a kernel. Examples include a square wave function f (x, y) = Θ(||x − y || σ) equal to 1, if the distance between x and y is less than or equal to 0, and a Gaussian influence function f(x,y) = e –||x-xy||2 /2 σ2 .This provides a high level of generality: the first example leads to DBSCAN, the second one to k-means clusters! Both examples depend on parameter σ. Restricting the summation to D = {y :| | x – y || < k σ }c X enables a practical implementation. DENCLUE concentrates on the local maxima of the density functions called density-attractors, and uses the flavor of the gradient hill-climbing technique for finding them. In addition to center-defined clusters, arbitrary-shape clusters are defined as continuations along sequences of points whose local densities are no less than the prescribed threshold £. The algorithm is stable with respect to outliers and authors show how to choose parameters £ and σ. DENCLUE scales well, since at its initial stage it builds a map of hyper-rectangle cubes with an edge length 2 σ. For this reason, the algorithm can be classified as a grid-based method. The advantages of this density function in clustering are, that it does not require a-priori specification of the number of clusters, is able to identify noise data while clustering, and the DBSCAN algorithm is able to find arbitrarily sized and arbitrarily shaped clusters.

IV. Conclusion

Data mining is a hot topic of the computer science research in recent years, and it has a extensive applications in various fields. Data mining technology is an application oriented technology. It not only is a simple search, query and transfer on the particular database, but also analyzes, integrates and reasons these data to guide the solution of practical problems and find the relation between events, and even to predict future activities through using the existing data.

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

References

[1] Ming-Syan Chen, Jiawei Han, Philip S yu. Data Mining: An Overview from a Database Perspective[J]. IEEE Transactions on Knowledge and Data Engineering, l996, 8(6):866-883.

[2] R Agrawal ,T 1 mielinski, A Swami. Database Mining: A Performance Perspective[J]· IEEE Transactions on Knowledge and Data Engineering, 1993,12:914-925.

[3] P. IndiraPriya, Dr. D.K.Ghosh ,"**A Survey on Different Clustering Algorithms in Data Mining Technique"** International Journal of Modern Engineering Research (IJMER) www.ijmer.com Vol.3, Issue.1, Jan-Feb. 2013 pp-267-274 ISSN: 2249-6645

[4] P. Berkhin, 2002. Survey of Clustering Data Mining Techniques. Ttechnical report, AccrueSoftware, San Jose, Cailf.

[5] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques" Elsevier Publication.

A. Jain, R. Duin, and J. Mao, "Statistical pattern recognition: A review," IEEE Trans. Pattern Anal. Mach.Intell., vol. 22, no. 1, pp. 4–37, 2000.