Asian Journal of Engineering and Applied Technology (AJEAT)
Vol.2.No.1 2014 pp 46-54
available at: www.goniv.com
Paper Received :05-03-2014
Paper Published:28-03-2014
Paper Reviewed by: 1. John Arhter 2. Hendry Goyal
Editor : Prof. P.Muthukumar

DATA MINING CLUSTER ANALYSIS

R. Jayalakshmi[*1] and A. J. Abdur Rakhib[2]
[1,2]Assistant Professor
[1,2]Department of Computer Science & Applications
[1,2]Sri Vidya Mandir Arts and Science College
[1,2]Uthangarai,Krishnagiri(DT)
[*1]jayasvm179@gmail.com

ABSTRACT

Data mining is the process of extracting hidden analytical information from large databases using multiple algorithms and techniques. This technology allows companies to focus on the most important information in their data warehouses. Generally organizations collect and process huge amount of data. Data mining techniques can be applied rapidly on existing software and hardware platforms to increase the value of existing information resources, and can be integrated with new products and systems. Previously data analysis process often involved manual work and during which interpretation of data was slow, costly, and highly immanent. The goal is to provide a self-contained review of the concepts and the mathematics underlying clustering techniques. The chapter begins by providing measures and criteria that are used for determining whether two objects are similar or dissimilar. Then the clustering methods are presented, divided into: hierarchical, partitioning, density-based, model-based, grid-based, and soft-computing methods. Following the methods, the challenges of performing clustering in large data sets are discussed. Finally, the chapter presents how to determine the number of clusters.

Keywords - Clustering, K-means, Intra-cluster homogeneity,

1. INTRODUCTION

Clustering and classification are both fundamental tasks in Data Mining. Classification is used mostly as a supervised learning method, clustering for unsupervised learning (some clustering models are for both). The goal of clustering is descriptive, that of classification is predictive. Since the goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is intrinsic. In classification tasks, however, an important part of the assessment is extrinsic, since the groups must reflect some reference set of classes.

Data mining technology is used for verifying sample data, analyze their paths, to verify and validate their modules for specific general and business utilization. For example an insurance industry uses these techniques for finding out their rate of risks for their customers. In general application of data mining tools are in the area of marketing, fraud fortifications, and observations. Different types of algorithms and tools are used for data evaluation and they are also providing different results depending on specific algorithms.

The purpose of data mining is to discover valid novel, possibly helpful and understandable relationships and patterns in the existing data. Data mining is to extract useful information; this is also known with different names like knowledge extraction, information discovery, information harvesting, and data pattern processing. The name “data mining” is mainly used by database researchers, statisticians, and business communities. Knowledge Discovery in Databases (KDD) is usually used to denote the process of finding out useful information from database, where data mining gives the detailed information in this process.

The processes in the KDD, such as data grounding, data collection, data clean-up, and proper understanding of the outcome of the data mining method, make sure that useful information is resulting from the data. The KDD have different steps.

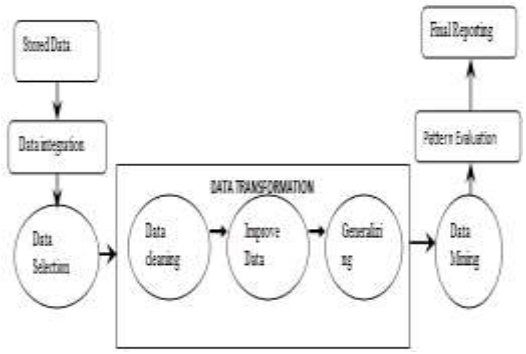


Figure 1.: KDD Process

- Stores data: Where data are stored in any data repository.
- Data integration: Multiple or heterogeneous data sources are integrated into a single unit.
- Data selection: Data retrieve from the database as needed for the KDD method.
- Data transformation: It is a process of data normalization where data are transformed and joined together into a form that is appropriate for mining process. Sub stages of this data transformation are,
  - Data cleaning: It handles noisy, erroneous and irrelevant data.
  - Improve data: Improve quality of data by adding new information, missing values to available data.
  - Generalizing data: Applying operations on data in order to prepare for a Data mining: This is a very important step in mining process. Here intelligent methods are applied in order to extracts data patterns.
- Pattern evaluation: This process is to identify the truly interesting patterns that are presented in knowledgebase.
- Knowledge presentation: It is final reporting of KDD process. Where visualization and knowledge representation techniques are used to represent the mined knowledge to the user.

2. DISTANCE MEASURES

Since clustering is the grouping of similar instances/objects, some sort of measure that can determine whether two objects are similar or dissimilar is required. There are two main type of measures used to estimate this relation: distance measures and similarity measures.

Many clustering methods use distance measures to determine the similarity or dissimilarity between any pair of objects. It is useful to denote the distance between two instances x_i and x_j as: $d(x_i, x_j)$. A valid distance measure should be symmetric and obtains its minimum value (usually zero) in case of identical vectors. The distance measure is called a metric distance measure if it also satisfies the following properties:

1. Triangle inequality $d(x_i, x_k) \leq d(x_i, x_j) + d(x_j, x_k)$ $\forall x_i, x_j, x_k \in S$.
2. $d(x_i, x_j) = 0 \Rightarrow x_i = x_j$ $\forall x_i, x_j \in S$.

2.1 Minkowski: Distance Measures for Numeric Attributes

Given two p-dimensional instances, $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $x_j = (x_{j1}, x_{j2}, \ldots, x_{jp})$, The distance between the two data instances can be calculated using the Minkowski metric

$$d(x_i, x_j) = (|x_{i1} - x_{j1}|^g + |x_{i2} - x_{j2}|^g + \ldots + |x_{ip} - x_{jp}|^g)^{1/g}$$

The commonly used Euclidean distance between two objects is achieved when $g = 2$

Given $g = 1$, the sum of absolute paraxial distances is obtained, and with $g=\infty$ one gets the greatest of the paraxial distances (Chebychev metric).

The measurement unit used can affect the clustering analysis. To avoid the dependence on the choice of measurement units, the data should be stan- dardized. Standardizing measurements attempts to give all variables an equal weight. However, if each variable is assigned with a weight according to its importance, then the weighted distance can be computed as:

$$d(x_i, x_j) = (w_1 |x_{i1} - x_{j1}|^g + w_2 |x_{i2} - x_{j2}|^g + \ldots + w_p |x_{ip} - x_{jp}|^g)^{1/g}$$

2.2 Distance Measures for Binary Attributes

The distance measure described in the last section may be easily computed for continuous-valued attributes. In the case of instances described by categorical, binary, ordinal or mixed type attributes, the distance measure should be revised.

n the case of binary attributes, the distance

between objects may be calcu- lated based on a contingency table. A binary attribute is symmetric if both of its states are equally valuable. In that case, using the simple matching coefficient can assess dissimilarity between two objects:

$$d(x_i, x_j) = \frac{r + s}{q + r + s + t}$$

where q is the number of attributes that equal 1 for both objects; t is the num- ber of attributes that equal 0 for both objects; and s and r are the number of attributes that are unequal for both objects.

A binary attribute is asymmetric, if its states are not equally important (usually the positive outcome is considered more important). In this case, the denominator ignores the unimportant negative matches (t).

$$d(x_i, x_j) = \frac{r + s}{q + r + s}$$

2.3 Distance Measures for Nominal Attributes

When the attributes are *nominal*, two main approaches may be used:

1. Simple matching:

$$d(x_i, x_j) =$$

where p is the total number of attributes and m is the number of matches.

3. CLASSIFICATION OF CLUSTERING ALGORITHMS

Categorization of clustering algorithms is neither straightforward, nor canonical.

In reality, groups below overlap. For reader's convenience we provide a classification closely followed by this survey. Corresponding terms are explained below.

Clustering Algorithms

- Hierarchical Methods
  - Agglomerative Algorithms
  - Divisive Algorithms
- 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

4. CLUSTERING ALGORITHMS

Clustering algorithms can be categorized based on their cluster model, as listed above. The following overview will only list the most prominent examples of clustering algorithms, as there are possibly over 100 published clustering algorithms. Not all provide models for their clusters and can thus not easily be categorized. An overview of algorithms explained in Wikipedia can be found in the list of statistics algorithms.

There is no objectively "correct" clustering algorithm, but as it was noted, "clustering is in the eye of the beholder." The most appropriate clustering algorithm for a particular problem often needs to be chosen experimentally, unless there is a mathematical reason to prefer one cluster model over another. It should be noted that an algorithm that is designed for one kind of model has no chance on a data set that contains a radically different kind of model.[4] For example, k-means cannot find non-convex clusters.[4]

4.1 Connectivity based clustering (hierarchical clustering)

Connectivity based clustering, also known as *hierarchical clustering*, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don't mix.

Connectivity based clustering is a whole family of methods that differ by the way distances are computed. Apart from the usual choice of distance functions, the user also needs to decide on the linkage criterion (since a cluster consists of multiple objects, there are multiple candidates to compute the distance to) to use. Popular choices are known as single-linkage clustering (the minimum of object distances), complete linkage clustering (the maximum of object distances) or UPGMA ("Unweighted Pair Group Method with Arithmetic Mean", also known as average linkage clustering). Furthermore, hierarchical clustering can be agglomerative (starting with single elements and aggregating them into clusters) or divisive (starting with the complete data set and dividing it into partitions).

These methods will not produce a unique partitioning of the data set, but a hierarchy from which the user still needs to choose appropriate clusters. They are not very robust towards outliers, which will either show up as additional clusters or even cause other clusters to merge (known as "chaining phenomenon", in particular withsingle-linkage clustering). In the general case, the complexity is $\mathcal{O}(n^3)$, which makes them too slow for large data sets. For some special cases, optimal efficient methods (of complexity $\mathcal{O}(n^2)$) are known: SLINK[5] for single-linkage and CLINK[6] for complete-linkage clustering. In the data mining community these methods are recognized as a theoretical foundation of cluster analysis, but often considered obsolete. They did however provide inspiration for many later methods such as density based clustering.

Linkage clustering examples

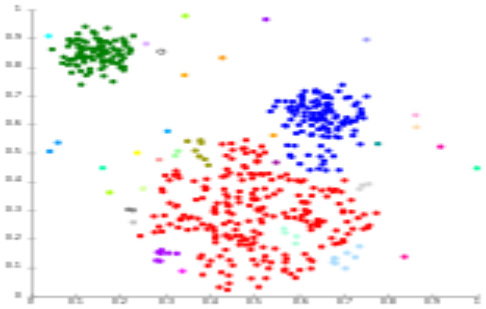


Figure 2. Single-linkage on Gaussian data. At 35 clusters, the biggest cluster starts fragmenting into smaller parts, while before it was still connected to the second largest due to the single-link effect.

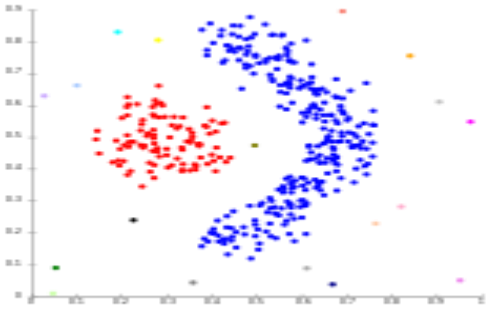


Figure 3. Single-linkage on density-based clusters. 20 clusters extracted, most of which contain single elements, since linkage clustering does not have a notion of "noise".

4.2 Centroid-based clustering

Main article: k-means clustering

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, *k*-means clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

The optimization problem itself is known to be NP-hard, and thus the common approach is to search only for approximate solutions. A particularly well known approximative method is Lloyd's algorithm,[7] often actually referred to as "*k-means algorithm*". It does however only find a local optimum, and is commonly run multiple times with different random initializations. Variations of k-means often include such optimizations as choosing the best of multiple runs, but also restricting the centroids to members of the data set (k-medoids), choosing medians (k-medians clustering), choosing the initial centers less randomly (K-means++) or allowing a fuzzy cluster assignment (Fuzzy c-means).

Most k-means-type algorithms require the number of clusters - k - to be specified in advance, which is considered to be one of the biggest drawbacks of these algorithms. Furthermore, the algorithms prefer clusters of approximately similar size, as they will always assign an object to the nearest centroid. This often leads to incorrectly cut borders in between of clusters (which is not surprising, as the algorithm optimized cluster centers, not cluster borders).

K-means has a number of interesting theoretical properties. On the one hand, it partitions the data space into a structure known as a Voronoi diagram. On the other hand, it is conceptually close to nearest neighborclassification, and as such is popular in machine learning. Third, it can be seen as a variation of model based classification, and Lloyd's algorithm as a variation of the Expectation-maximization algorithm for this model discussed below.

k-Means clustering examples

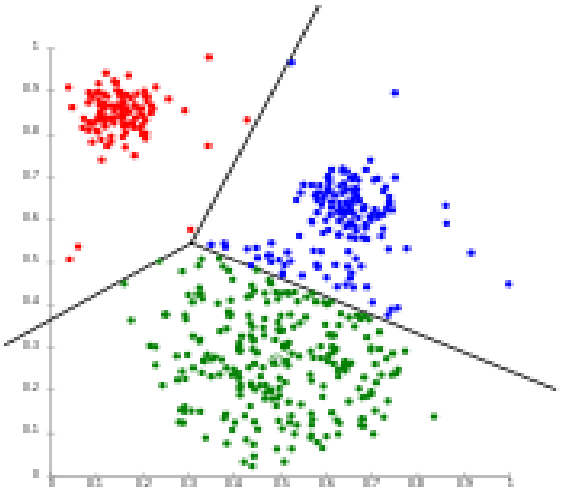


Figure 4: K-means separates data into Voronoi-cells, which assumes equal-sized clusters (not adequate here)

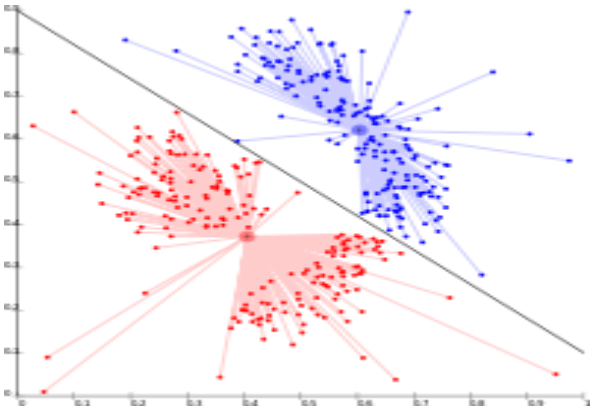


Figure 5:K-means cannot represent density-based clusters

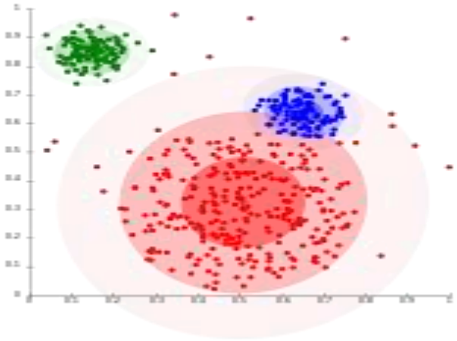


Figure 7:On Gaussian-distributed data, EM works well, since it uses Gaussians for modelling clusters

4.3 Distribution-based clustering

The clustering model most closely related to statistics is based on distribution models. Clusters can then easily be defined as objects belonging most likely to the same distribution. A convenient property of this approach is that this closely resembles the way artificial data sets are generated: by sampling random objects from a distribution.

While the theoretical foundation of these methods is excellent, they suffer from one key problem known asoverfitting, unless constraints are put on the model complexity. A more complex model will usually be able to explain the data better, which makes choosing the appropriate model complexity inherently difficult.

One prominent method is known as Gaussian mixture models (using the expectation-maximization algorithm). Here, the data set is usually modelled with a fixed (to avoid overfitting) number of Gaussian distributions that are initialized randomly and whose parameters are iteratively optimized to fit better to the data set. This will converge to a local optimum, so multiple runs may produce different results. In order to obtain a hard clustering, objects are often then assigned to the Gaussian distribution they most likely belong to; for soft clusterings, this is not necessary.

Distribution-based clustering produces complex models for clusters that can capture correlation and dependence between attributes. However, these algorithms put an extra burden on the user: for many real data sets, there may be no concisely defined mathematical model (e.g. assuming a Gaussian distribution is a rather strong assumption on the data).

Expectation-Maximization (EM) clustering examples

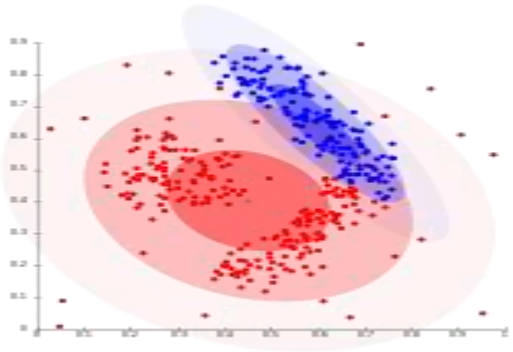


Figure 8:Density-based clusters cannot be modeled using Gaussian distributions

4.4 Density-based clustering

In density-based clustering,[8] clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points.

The most popular[9] density based clustering method is DBSCAN.[10] In contrast to many newer methods, it features a well-defined cluster model called "density-reachability". Similar to linkage based clustering, it is based on connecting points within certain distance thresholds. However, it only connects points that satisfy a density criterion, in the original variant defined as a minimum number of other objects within this radius. A cluster consists of all density-connected objects (which can form a cluster of an arbitrary shape, in contrast to many other methods) plus all objects that are within these objects' range. Another interesting property of DBSCAN is that its complexity is fairly low - it requires a linear number of range queries on

the database - and that it will discover essentially the same results (it is deterministic for core and noise points, but not for border points) in each run, therefore there is no need to run it multiple times. OPTICS[11] is a generalization of DBSCAN that removes the need to choose an appropriate value for the range parameter ε, and produces a hierarchical result related to that of linkage clustering. DeLi-Clu,[12] Density-Link-Clustering combines ideas from single-linkage clustering and OPTICS, eliminating the ε parameter entirely and offering performance improvements over OPTICS by using anR-tree index.

The key drawback of DBSCAN and OPTICS is that they expect some kind of density drop to detect cluster borders. Moreover, they cannot detect intrinsic cluster structures which are prevalent in the majority of real life data. A variation of DBSCAN, EnDBSCAN,[13] efficiently detects such kinds of structures. On data sets with, for example, overlapping Gaussian distributions - a common use case in artificial data - the cluster borders produced by these algorithms will often look arbitrary, because the cluster density decreases continuously. On a data set consisting of mixtures of Gaussians, these algorithms are nearly always outperformed by methods such as EM clustering that are able to precisely model this kind of data.

Mean-shift is a clustering approach where each object is moved to the densest area in its vicinity, based on kernel density estimation. Eventually, objects converge to local maxima of density. Similar to k-means clustering, these "density attractors" can serve as representatives for the data set, but mean-shift can detect arbitrary-shaped clusters similar to DBSCAN. Due to the expensive iterative procedure and density estimation, mean-shift is usually slower than DBSCAN or k-Means.

Density-based clustering examples

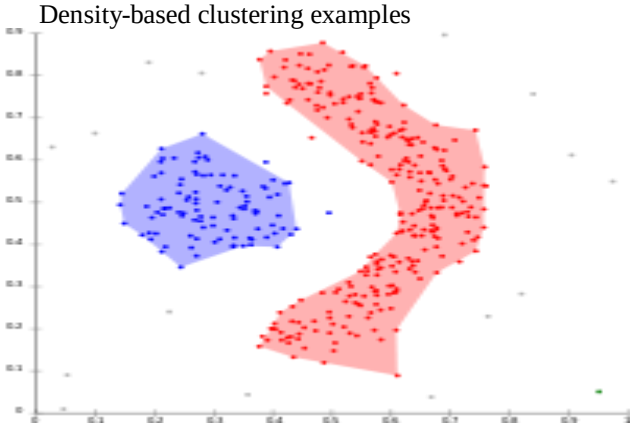
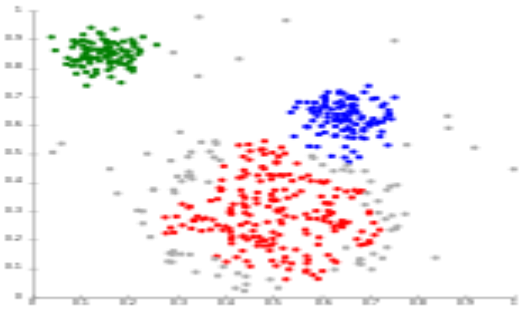


Figure 8: Density-based clustering with DBSCAN.



5. CONCLUSION

Data mining, the central activity in the process of knowledge discovery in databases (KDD), which is concerned with finding patterns in data. We live in a world where vast amounts of data are collected daily. Such data is an important to need so data mining play the important role. Data mining can meet this need by providing tools to discovery knowledge from data. Now days we can see that data mining use in any area. So we observe that much kind of concepts and technique used in bio-logical and environment problems. And try the removed complicated and hard type of data. This paper highlights on biological sequences problem such as protein and genomic sequences and other biological segments such as cancer prediction. In environment presents Degradation of Land and Vegetation, Water Pollution, Air and Noise Pollution, Noise and Vibration and environmental tool also discuss. Data mining algorithms, tools and concepts used in these problems Such as MATLAB, WEKA, SWIISPORT , Clustering , Biclstering and any other thing in this survey.

REFERENCES

[1] Hui xiong, Xiaofeng HE ,Chris ding ,Ya Zhang, Vipin kumar, Stephen r.holbrook.

[2] Gerasimos Hatzidamianos, Sotiris Diplaris, Ioannis Athanasiadis, Pericles A. Mitkas.

[3] Shreyas Sen, Seetharam Narasimhan, and Amit Konar[Engineering Letters, 14:2, EL_14_2_8 (Advance online publication: 16 May 2007].

[4] Sunita Soni and O.P.Vyas[International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.1, February 2012].

[5] Basheer M. Al-Maqaleh *and* Hamid Shahbazkia [International Journal of Computer Applications (0975 – 8887) Volume 41– No.18, March 2012].
[6] Wafa Mokharrak, Nedhal Al Khalaf, Tom Altman[Department of Computer Science and Engineering, University of Colorado Denver, Denver, Colorado, United States of America].
[7] Pengyi Yang, Li Tao, Liang Xu, and Zili Zhang[P. Wen et al. (Eds.): RSKT 2009, LNCS 5589,pp. 200–207, 2009. _c Springer-Verlag Berlin Heidelberg 2009].
[8] David Page and Mark Craven[Dept. of Biostatistics and Medical Informatics and Dept. of Computer Sciences].
[9] Yifeng Li and Alioune Ngom[School of Computer Science, University of Windsor, Windsor, Ontario, Canada].
[10] K.Muralidharan [II Year B.Tech Information Technology Karpagam Institute of Technology COIMBATORE – 21].
[11] J-S. Lai , F.Tsai**[**International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XXXIX-B2, 2012 XXII ISPRS Congress, 25 August –01 September 2012, Melbourne, Australia].
[12] Cohen AM, Hersh WR, A survey of current work in biomedical text mining, Briefings in Bioinformatics, 2005, 6: 57-71
[13] Langley, P., Lessons for the computational discovery of scientific knowledge. In Proceedings of First International Workshop on Data Mining Lessons Learned, 2002, p. 9-12.
[14] Yang, Q., Xindong Wu, 10 Challenging Problems in Data Mining Research, International Journal of Information Technology & Decision Making, Vol. 5, No. 4 (2006) 597–604.
[15] M Allaby. *Basics of Environmental Science*. Routledge, London, 1996.
[16] http://www.theiia.org/ intAuditor/ itaudit/archives/ 2006/august/data-mining-101-tools-and-techniques/
[17]http://www.wisegeek.com/what-are-the-ost-important-data-mining-concepts.htm
[18]http://www.wisegeek.com/what-are-the-different-types-of-data-mining-technology.htm
[19]http://www.wisegeek.com/what-are-the-different-types-of-data-mining-analysis.htm
[20] Chung, H. M., Gray, P. (1999), "Special Section: Data Mining". Journal of Management Information Systems, (16:1), 11-17.
[21] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, R (1996). "The KDD Process for Extracting Useful Knowledge from Volumes of Data,"