International Journal of Advances in Engineering Research http://www.ijaer.com
(IJAER) 2012, Vol. No. 4, Issue No. III, September **ISSN: 2231-5152**

MOST COMMONLY USED TECHNIQUES IN DATA MINING

Omkar Singh Lodhi

ABSTRACT

Data mining has emerged in last few years a growing and major area in the field of research. Data mining is able to tell us important things that we didn't know or what is going to happen next? The technique that is used to perform these feats in data mining is called mining techniques. In this research paper our proposal work for description of some methods in data mining to handle general query mechanism along with suitable data structures to keep track of patterns identified during the data mining process. These algorithms are also useful in industries, companies, banking, medical etc. to increase the performance.

Mathematical models are useful in providing a unified theoretic approach and therefore it is a need to study different data mining technique under this approach. Using these approaches the strength and weakness of the data mining technique would come out in scientific and analytic way for computer professionals, data analyst and information providers.

INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. It sits at the common frontiers of several fields including Data Base Management, Artificial Intelligence, Machine Learning, Pattern Recognition, and Data Visualization. From a statistical perspective it can be viewed as computer automated exploratory data analysis of large complex data sets. In spite of the somewhat exaggerated hype, this field is having a major impact in business, industry, and science.

EVOLUTION OF DATABASE TECHNOLOGY

The major reason that data mining has attracting and become a very important field in information industry in recent years is because of the wide availability of huge amounts of data and the imminent use need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications such as ranging from business management, production control, and market analysis, to engineering design and science exploration.

WHAT IS DATA MINING?

Data Mining (DM) is at best a vaguely defined Field; its definition largely depends on the background and views of the definer. Here are some definitions taken from the DM literature:

"Data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". - Fayyad.

"Data mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions". – Zekulin

"Data Mining is a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data". – Ferruzza.

"Data mining is the process of discovering advantageous patterns in data". – John

"Data mining is a decision support process where we look in large data bases for unknown and unexpected patterns of information". – Parsaye

Thus data mining can be viewed a simply an essential step in the process of knowledge discovery in data base. Knowledge discovery processes consist of following steps:

1. **Data Cleaning** – the removal of noise and inconsistent data
2. **Data Integration** – The combination of multiple sources of data
3. **Data Selection** – The data relevant for analysis is retrieved from the database
4. **Data Transformation** – The consolidation land transformation of data into forms appropriate for mining by performing summary of data.
5. **Data Mining** – The use of intelligent methods to extract patterns from data.
6. **Pattern Evaluation** – identification of patterns that are interesting
7. **Knowledge Presentation** – visualization and knowledge representation techniques are used to present the extracted or mined knowledge to the end user.

Steps 1 to 4 are different forms of data preprocessing, where the data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

ROLE OF DATA PREPROCESSING IN WEB USAGE MINING

Data preprocessing is also an essential step in web usage mining which is the application of data mining techniques to discover usage patterns from web data, in order to understand and better serve the needs of web based applications. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis.

TECHNIQUES IN DATA MINING

A data warehouse is a subject-oriented, integrated, non - volatile and time- variant collection of data in support of management's decisions processes like data mining fetch the hidden predictive information from data warehouse. Data mining predicts future trends and behavior which makes businesses upbeat, knowledge-driven decisions. The most frequently used techniques in data mining are:

Clustering:—Process of managing objects into groups whose members have similar property in some way. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters.

Decision trees: Tree-shaped structures that represent sets of decisions. These decisions are helpful in generating rules for the classification.

Naïve Bayes: The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Naive Bayes can often outperform more sophisticated classification methods.

Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset k-nearest neighbor technique.

The *k*-means algorithm: The k-means algorithm is a simple iterative method to partition a given dataset into a user specified number of clusters.

Association Analysis: Association analysis useful in finding hidden relationship in large data warehouse. This hidden relationship can be representing in the form of frequent item sets or association rules. For example the Apriori Algorithm is an influential algorithm for mining frequent item sets for boolean association rules.

Rule Induction: The extraction of useful if-then rules from data based on statistical significance.

In data mining classifier is the important tool. Systems that construct classifiers get input as a set of classes, each class consist of small number of classes and having fixed set of attribute values and give output which predict the class that belongs with a new class .

Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data.

1. **Decision tree (C4.5 and beyond)**

Set of D of tuples using divide-and-conquer technique C4.5 first make an initial tree as follows:

- Create a node N.
- If tuples in D belong to the same class say C, then return N as leaf node labeled with the class C.
- Otherwise, apply attribute selection method on a single attribute. This attribute is the root of the tree with one child for each splitting attribute of the method, partition D into the corresponding tuples and grow subtrees for each partition.

Here C4.5 use attribute measure: information gain, which calculates the total entropy and second approach, is gain ratio which divides the information gain by the information given by the attribute.

Decision trees prefer a splitting method and not return to splitting node.

2. **Rule set classifiers**

In decision tree information about one class is generally dispersed throughout the tree, so complex decision tree can be thorny to understand. C4.5 gives a different formalism represented by as a set of IF-THEN rules. It consists of a set of rules of the form if X and Y and Z

and ... then class A. IF- THEN rule is an expression of the form:

IF condition THEN conclusion

If condition in a rule antecedent holds true for a given tuples, we say that the rule is accepted. When there is no rule is accepted by the tuple then a fallback or default rule can be set to indicate a default class. C4.5 rulesets are created from the decision tree which is not pruned.

a. *k*-nn: *k*-nearest neighbor classification

Nearest neighbor was originally proposed by Fix and Hodges in 1952. kNN first find out k objects in the training data that are neighbor to the test object. These k training tuples are the k-nearest neighbors of the unidentified test tuple. Then assign this test tuple with the class of training tuple. Patrick and Fischer take a large view of the nearest neighbor rules consist of weighting of different types of fault and problems in which the training datasets available are not in the same magnitude as in the priori class probabilities. S.Cost and S. Salzberg proposed an algorithm, PEBLS which is based on *k*-NN classification that include relationship determine for class information. In text classification k-nearest neighbor (k-NN) classification shown to be very effective because it is object based learning algorithm. The distance from the unlabelled object to the labeled object is computed, k-nearest neighbors are recognized and choose the class label of the object with the class labels of these neighbors Algorithm for kNN as given below:

Given a training data D and a test instance w = (x' y'), the algorithm calculates the distance between test instance *w* and training instances (x, y) Є D to find out its nearest-neighbor list, D*w*. (x is the tuple value of a training instance, while *y* is its class. Likewise, x' is the tuple value of the test instance and y' is its class.

Input: M **is** the set of training instances and test instance w= (x' y')

Process: Calculate d (x**',** x**),** the distance between w and every object, (x,y) ЄD Select, Dw $\subseteq$ D the set of nearby training instances to w.

Output:

$$W' = argmax \sum_{(xi\,y)\in D} X(x = yi)$$

b. Naive Bayesian Classifier

Simple classifier called naive bayes classifier which is based on bayes theorem. One assumption called class conditional independence in this classifier i.e. the attribute value on a given class is independent of the value of the other attributes. Naive Byes classification the naive Bayesian classifier, or simple Bayesian classifier, works as follows:

1. Let S be a training set of tuples and their class labels.

2. *n* classes, *D1*, *D*2, … , *Dn*. the naïve Bayesian classifier predicts that tuple W belongs to the class *Di* if and only if *P(Di*/ w) > *P(Dj*/ w) for $1 \le j \le n,\ j=i$

The class Di for which *P (Di*/ w) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem

$$P(Di/\ w) = \frac{P(W/\ Di)P(Di)}{P(w)}$$

3. For class conditional independence

$$P\left(\frac{W}{Di}\right) = \prod_{k=1}^{m} P\left(\frac{Wk}{Di}\right) = P\left(\frac{W1}{D1}\right) \times P\left(\frac{W2}{D2}\right) \times P\left(\frac{W3}{D3}\right) \times \ldots\ldots\ldots\ldots \times P\left(\frac{Wm}{Dm}\right)$$

In other words, the predicted class label is the class Di *for* which $P(W/Di)P(Di)$ is the maximum.

c. Document Classification:

The basic principle of document classification is to classify or group a set of unlabeled documents into classes or clusters. The division of document classification is subcategories into three sets, i.e. supervised or unsupervised, hard or soft and partitioning, hierarchical or frequent item set-based. These subcategories can be shown in a tree structure as Fig. 1, which explain as follows:

1. Supervised and Unsupervised: in supervised document classification, a set of pre-define classes are available. On the other hand, in unsupervised document classification, also called document clustering, there are no pre-determined classes available. Document clustering is the process of calculating document similarities to form clusters. The documents within a cluster are similar to each other and, simultaneously, dissimilar to the documents in the other groups.

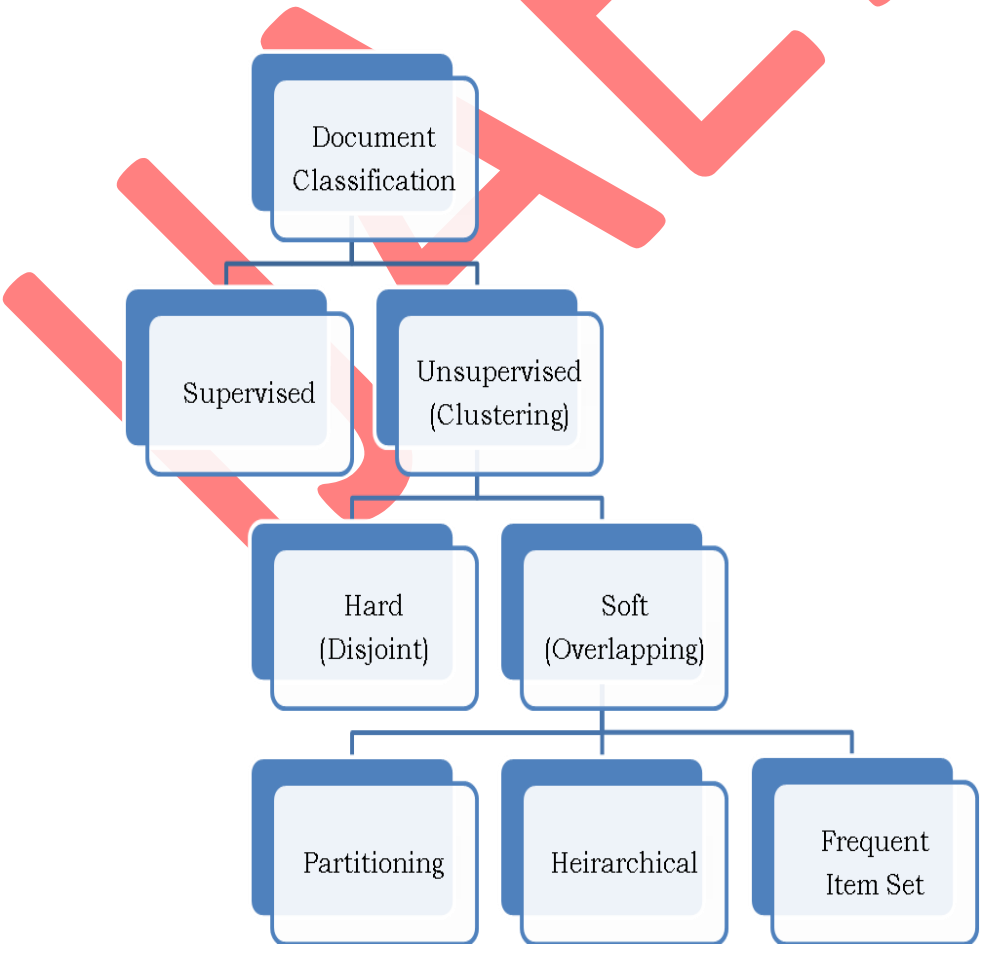

Fig. 1. A tree structure with three types of document classification

2. Hard and Soft: hard clustering algorithms compute the hard assignment and produce a set of disjoint clusters. Soft clustering algorithms compute the soft assignment and generate a set of overlapping clusters. For instance, a document discussing Natural language and Information Retrieval should be assigned to both of the clusters "Natural language" and "Information Retrieval".

3. Partitioning, Hierarchical, and Frequent itemset-based: for document clustering, partitioning-based methods exclusively partition the set of documents into a number of clusters by moving documents from one cluster to another, such as k-means and Bisecting k-means. Hierarchical-based document clustering is to build a hierarchical tree of clusters, whose leaf nodes represent the subset of a document collection, like Hierarchical Agglomerative Clustering (HAC) and Unweighted Pair Group Method with Arithmetic Mean (UPGMA). Besides, a new category of document clustering, namely "frequent itemset-based clustering," has been extensively developed, including FIHC, HFTC, TDC, and F2IDC. Frequent itemset-based clustering methods use frequent itemsets generated by the association rule mining and further cluster the documents according to these extracted frequent itemsets. These methods reduce the dimensionality of term features efficiently for very large datasets, thus they can improve the accuracy and scalability of the clustering algorithms. An advantage of frequent itemset- based clustering method is that each cluster can be labeled by the obtained frequent itemsets shared by the documents in the same cluster. Moreover, the organization of clusters generated by frequent itemset-based clustering methods could be a flat set or a hierarchical tree of clusters.

d. **CLUSTERING**

Clustering is important technique in exploratory data analysis. It finds out the useful pattern and correlation between attributes here we discussed EM algorithm for clustering.

THE EM ALGORITHM

Dempster et al. (1977) have given the name EM algorithm. Historical perspective of the EM algorithm can be found in McLachlan and Krishnan (1997) Model-based clustering methods attempt to optimize the fit between the mathematical model and given data. Each cluster can be represented mathematically by a parametric probability distribution. So we cluster the data using a finite mixture density model of k probability distribution Now finding the parameter estimates of probability distributions which can be best fit in the data The EM(Expectation-Maximization) algorithm is a recursive refinement algorithm that can be used for finding the parameter estimates. It can be shown as an extension of the *k*-means approach, in which object is assigned to that cluster whose cluster mean is similar to that object. In EM algorithm new means are calculated by weighted measures. Then object is assigned to cluster according to weighted measures that represent the membership probability.

The algorithm is described as follows:

1. Initially guess the parameter vector. Randomly selecting *d* objects to represent the

cluster means or centers.

2. Recursively process the parameters (or clusters) based on the following two steps:

(a) Expectation Step: Assign each object xi to cluster *Ck* with the probability

$$P(w_i \in N_d) = p(N_d / w_i) = \frac{p(N_d)p(w_i / N_d)}{p(w_i)}$$

Where $p(w_i / N_d) = D(n_d, E_d(w_i))$ follows the normal (i.e., Gaussian) distribution around mean n_d with expectation E_d.

(b) Maximization Step: Use the probability estimates from above to re-estimate (or refine) the model parameters. For example,

$$n_d = \frac{1}{m}\sum_{i=1}^{m} \frac{w_i P(w_i \in N_d)}{\sum_j P(w_i \in N_j)}$$

Neal and Hinton (1998) presented for improving the convergence rate of the EM algorithm.

CONCLUSION

Information plays a major role in every field. Data mining is a tool that exploits to discover patterns from raw data, extraction of useful information stored. Data mining is wide area that assimilates techniques from various fields including pattern recognition, artificial intelligence, database systems and machine learning. Here we explain few data mining algorithms which are used to perform data analysis tasks in different fields. These algorithms employed in fraud detection, intrusion detection, Health care and finance for extraction of useful information. This attempt could add additional information in the field of data mining and new idea would come out for the development of efficient algorithms in the field of knowledge discovery.

REFERENCES

[1] . Han J. Micheline Kamber “Data mining concept and techniques” Academic press 2001.
[2] . A.K. Poojari “Data Mining Techniques”, University press 2002.
[3] . Andrea Marin (2006) : Hidden Markov Models applied to data mining, BISS 2006
[4] . G. Goulbourne, F. Coenen and P. Leng (April 2000) : Algorithms for computing association rules using a partial-support tree Knowledge-Based Systems.
[5] . L. Kaufman and P. J. Rousseeuw(1990) : Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons.
[6] . Nong Ye, Yebin Zhang, and Connie M. Borror(March 2004): Robustness of the Markov-Chain Model for. CyberAttack Detection, IEEE Transactions on Reliability.

[7] . R. Ng and J. Han.(1994) : Efficient and effective clustering method for spatial data mining, In Proceedings of the 20th VLDB Conference.
[8] . Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. Machine Learning.
[9] . Langley, P. (1993). Induction of recursive Bayesian classifiers. Proceedings of the Eighth European Conference on Machine Learning.
[10] . O.M. San et al. An alternative extension of the k-means algorithm for clustering categorical data Int. J. Appl. Math. Comput. Sci., 2004.
[11] . MacQueen J.B. (1967): Some methods for classification and analysis of multivariate observations. 5-th Symp. Mathematical Statistics and Probability.
[12] . Ralambondrainy H. (1995): A conceptual version of the kmeans algorithm. Pattern Recogn. Lett.
[13] . Huang Z. (1998): Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining Knowl. W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, and K. Y. Yip, –Efficient clustering of uncertain data, in IEEE International Conference on Data Mining (ICDM) 2006.
[14] . M. Chau, R. Cheng, B. Kao, and J. Ng, –Data with uncertainty mining: An example in clustering location data in Proc. of the Methodologies for Knowledge Discovery and Data Mining, Pacific-Asia Conference (PAKDD 2006).
[15] . A. C and Y. PS, –A framework for clustering uncertain data streams, in Proceedings of IEEE 24rd International Conference on Data Engineering, 2008.
[16] . Y. Xia and B. Xi, –Conceptual clustering categorical data with uncertainty, in IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 2007.
[17] . McLachlan, G.J. and Krishnan, T. (1997). The EM Algorithm and Extensions. Wiley, New York.
[18] . Neal, R.M. and Hinton, G.E. (1998). A view of the EM algorithm that justifies incremental,sparse, and other variants. In Jordan, M.I., editor, Learning in Graphical Models.
[19] . Bradley, P.S., Fayyad, U.M., and Reina, C.A. (1998). Scaling EM (expectation maximization) clustering to large databases.
[20] . Z. Yu and H. Wong, –Mining uncertain data in low- dimensional subspace, in International Conference on Pattern Recognition (ICPR) 2006.
[21] . C. Chui, B. Kao, and E. Hung, –Mining frequent itemsets from uncertain data, in Proc. of the Methodologies for Knowledge Discovery and Data Mining, Pacific-Asia Conference (PAKDD) 2007.
[22] . P.Ponniah, –Data Warehousing Fundamentals-A comprehensive guide for IT professionals, Ist ed.,second reprint, ISBN-81-265-0919-8, Glorious Printers: New Delhi India, 2007.
[23] . An Introductionto Data Mining, Review: http://www.thearling.com/text/dmwhite/dmwhite.htm
[24] . A Tutorial on Clustering Algorithms, Review http://home.dei.polimi.it/matteucc/Clustering/tutorial_html
[25] . Naive Bayes Classifier Review: http://www.statsoft.com/textbook/naive-bayes-classifier/
[26] . Pang-Ning Tan,Michael Steinbach,Vipin Kumar, –An Introduction to Data Mining, ISBN : 0321321367. Addison-Wesley, 2005.

[27] . XindongWu · Vipin Kumar et all, –Top 10 algorithms in data mining Knowl Inf Syst (2008).
[28] . Murthy, S. & Salzberg, S. (1995), Lookahead and pathology in decision tree induction,in C. S. Mellish, ed., Proceedings of the 14th International Joint Conference on Artificial Intelligence', Morgan Kaufmann.
[29] . Utgo, P. E. (1997), `Decision tree induction based on efficient tree restructuring', Machine Learning.
[30] . Shafer, J., Agrawal, R. & Mehta, M. (1996), Sprint: a scalable parallel classier for data mining, in `Proceedings of the 22nd International Conference on Very Large Databases (VLDB)'.
[31] . Freitas, A. A. & Lavington, S. H. (1998), Mining Very Large Databases with Parallel Processing, Kluwer Academic Publishers.
[32] . Kearns, M. & Mansour, Y. (1998), A fast, bottom-up decision tree pruning algorithm with near-optimal generalization, in J. Shavlik, ed., `Machine Learning: Proceedings of the Fifteenth International Conference', Morgan Kaufmann Publishers,Inc.
[33] . Friedman, J., Kohavi, R. & Yun, Y. (1996), Lazy decision trees, in `Proceedings of the Thirteenth National Conference on Artificial Intelligence.
[34] . Quinlan, J. R. & Rivest, R. L. (1989), `Inferring decision trees using the minimum description length principle', Information and Computation.
[35] . Mehta, M., Rissanen, J. & Agrawal, R. (1995), MDL- based decision tree pruning, in U. M. Fayyad & R. Uthurusamy, eds, `Proceedings of the first international conference on knowledge discovery and data mining.
[36] . Wallace, C. & Patrick, J. (1993), `Coding decision trees', Machine Learning.
[37] . Biao Qin, Yuni Xia et al. –A Rule-Based Classification Algorithm for Uncertain Data IEEE International Conference on Data Engineering 2009.
[38] . Y. Yang. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In SIGIR-94, 1994.
[39] . Eui-Hong (Sam) Han et al. Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification.
[40] . E. A. Patrick and F. P. Fischer, III, "A generalized k- nearest neighbor rule," Inform. Contr., vol. 16, pp. 128- 152, Apr. 1970.
[41] . Dennis l. Wilson, Asymptotic properties of nearest neighbor rules using edited data IEEE transactions on systems, man, and cybernetics, vol. Smc-2, no. 3, July 1972.
[42] . Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifiers. Proceedings of the Tenth National Conference on Artificial Intelligence.