www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 10 October 2017, Page No. 22537-22544
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i10.02

Data Mining and its Techniques

Hardik Uppal

Tata Consultancy Services
Department of Electronics Technology, GNDU, Amritsar

Abstract

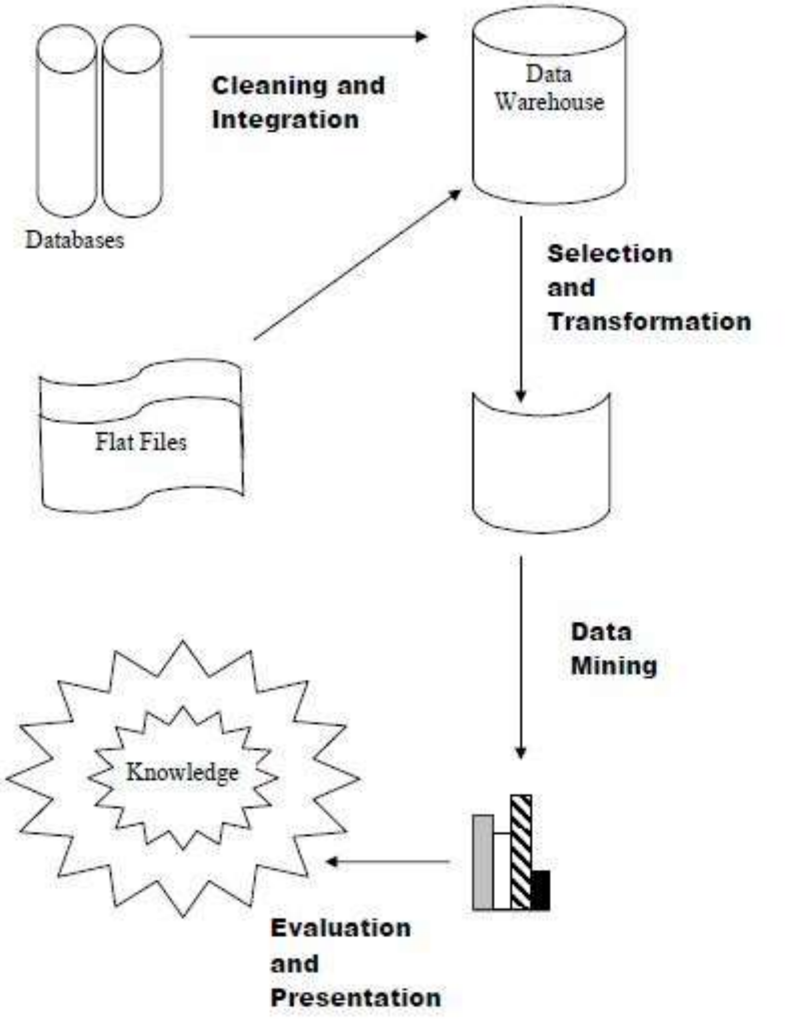
With the evolution of technology, the speed at which we gather data has increased exponentially. The aim of organizing and providing meaningful access to data is the task of of prime importance. We can see a tremendous potential in it as the influx of data brings with it an exponential growth in knowledge. The process through which all the data is traversed and meaningful patterns discovered is known as Data Mining or Knowledge Discovery. This review provides an insight into the most common techniques in the field such as Prediction, Clustering, Association, Sequential Pattern detection etc.

Keywords: *Data mining, Data mining techniques, Association, Clustering, Sequential Pattern, Prediction, Classification, Decision trees*

Introduction

The Last Decade saw the rise of internet and social media and hence the rate and speed at which the information became available has been overwhelming. There has been an uncontrollable growth in databases, which has created a void, which needs to be filled with new technology. This new technologies should be able to use information and knowledge more efficiently. Therefore, Data mining Techniques (DMT) has become an important research area [16]. With tremendous opportunities who have capabilities to unlock the information embedded within this data, it introduces new challenges. The topics included are: what data mining is, what type of challenges it faces, what type of problems it can address , and why it developed [5]. DMT have formed a branch of applied artificial intelligence (AI), where the aim is not only to find pattern and underlying information in a specific data set but also to generalize it beyond available data. Data mining is not just now an active research area but also has found its way in corporate houses with many large companies recognizing that it will have an impact on their performance. Major elements of data mining are:

- Extract, transform, and load (ETL process) data onto the data warehouse system.
- Provide meaningful data access to end users.
- Present the data as a graph or table to make it more meaningful.
- Analyse the data by application software.
- Manage and store the data in a multidimensional database system, to retrieve faster results. [15]



Literature Review

Data mining is an important area from research point of view and artificial neural network solves many problems of data mining. Artificial neural network is a biological system that helps in making predictions and pattern detection. Neural network and genetic algorithm are common data mining methods. With help of genetic algorithm, we can build a better solution by combining the good parts of other solution. Many scientific applications are dependent on these methodologies [1].

Data mining makes use of some visualization techniques, machine learning and various statistical techniques to discover or find patterns among the databases which can be easily understood by humans. There are various data mining tools like WEKA, Tanagra, Rapid miner, Orange. WEKA is acronym for Waikato Environment for Knowledge Learning. It was developed to identify information from raw data collected from agricultural domains. TANAGRA tool is used in research purposes. It is an open source application and provide helps to researchers in adding their own data mining methods. One can use data mining tools to answer “What-if” questions and help in demonstration of real effects [2].

Data mining and knowledge discovery in database (KDD) are linked. KDD helps us in making data useful. Today there is a great need of tools and computational theories which can help humans in the extraction of useful and meaningful information from the volumes of data present today. Data mining is a step in KDD process in which data analysis is done to produce hidden values in database [3].

Many difficulties in data mining research has been identified. No unifying theory for individual problems in data mining, handling high dimensional data, mining of time series and sequence data, extracting complex knowledge from complex databases, application of data mining methods to environmental and biological problems are some of the main problems [4].

There are various steps in data mining process which are data cleaning, data integration, data selection, data Pre-processing, data transformation, data mining method and interpretation or knowledge discovery. Today data is growing at a tremendous pace. This data must be stored cost-effectively and also analysis of this data should be done so as to extract meaningful information from this data. Various data mining tasks are employed for this purpose like classification, regression, cluster analysis, text analysis, and link analysis tasks [5].

Data mining can be used in business environment, weather forecasting, product design, load prediction etc. It can be viewed as a decision support system than just using traditional query languages. Data mining has become one of the most important area in database and information technology [6].

Data mining can be used for decision making in pharmaceutical industry. Data in an organization can be used for extracting hidden information rather than just for administrative purposes. User interface designed for accepting all kinds of information from the user can be scanned and list of warning can be issued on the basis of the information entered by the user. Neural network technique of data mining can be used to clinically test drugs on patients. New drugs can be generated through clustering, classification and neural network [7].

Sequential pattern mining is used to discover interesting sequential patterns from the large databases. It can be classified into two groups namely Apriori based and Pattern growth based. Pattern growth based algorithms are more efficient than Apriori based algorithms in terms of space utilization, running time complexity and scalability [8].

Association rule mining is used to uncover interesting relationships among data. Many companies want to increase their profits by mining association rules from the large datasets. Association rule is generally used in market based analysis to analyze customer buying habits. Association rule is simple and can be implemented easily so as to increase the profits [9].

Data mining applications are limited to educational contexts. Data mining approach can be applied to educational data sets. Educational data mining (EDM) has emerged to solve educational related problems. It can be used to understand student retention and attrition so as to make personal recommendation to every individual student. EDM is also applied to admissions and enrolment [10].

Techniques of Data Mining

Data mining aims to find useful patterns within datasets or databases. We use two kinds' basic models: - predictive models and descriptive model. Predictive models focuses on predicting values based on previous data and descriptive model helps understanding the underlying characteristics of the data [1].

Some of the most common methods used are

- Association
- Classification
- Clustering
- Decision tree
- Sequential patterns
- Prediction

A. *Association*

Association Rule Induction is most common form of data mining and knowledge discovery in unsupervised learning systems. Association rule analysis is discovers new and fascinating relationships hidden in large volume of data [5]. It is a descriptive model technique.

Rule induction on a data base can be tedious task where the pattern are identified and then driven out from the data and then an accuracy and significance describe the possibility of it occurring again. An association rule is about relationship between two disjoint item sets X and Y as X-> Y.

Association technique is mostly applied in **Market based analysis**. Let's say if an item is bought what are the chances that another item is bought with it, such as:

- If bagels are purchased then cream-cheese is purchased 90% of the time and this pattern occurs in 3% of all shopping baskets.
- If live plants are purchased from a hardware store then plant fertilizer is purchased 60% of the time and these two items are bought together in 6% of the shopping baskets.[17]

B. *Clustering*

Clustering is a descriptive task where data is divided into groups of similar objects. Each group is called cluster [5]. Different objects are paired with similar to them groups known as clusters. This technique may simplify but normally does lose some finer details. Clustering algorithms can be applied to web analysis, computational biology, CRM, marketing and medical diagnosis [18].

Clustering Techniques

Clustering technique can be broadly divided into Hierarchical and Non - Hierarchical Clustering.

- *Hierarchical Clustering*: - In Hierarchical clustering, we try to find hierarchy in clusters. Few smaller clusters come together to form larger clusters of similar properties. There are 2 common algorithms used in hierarchical clustering
    - Agglomerative clustering technique: - It is a bottom up technique which starts with n clusters. Each cluster contains one record. The clusters which are similar or nearest to each other are paired together in one cluster. This process is successively repeated until we are left with one cluster which contains all the data.
    - *Divisive Hierarchical Clustering:* - It is a top down clustering method and works opposite to agglomerative clustering. It stars with One cluster which contains all the records and then split that cluster into smaller clusters until cluster of a single record remains [18].
- *Non-hierarchical clustering:* - This technique is used when the relationship between the clusters is indeterminable. The user controls the clustering process as to how many clusters should be formed. This method is a lot faster than hierarchical clustering. There are two methods *single pass* and *reallocation.* The difference between the two is that in Single pass method it goes through the database only once assigning data to clusters whereas in reallocation it can go through databases multiple times and hence creating better clusters.

C. *Sequential Pattern*

Sequential pattern is a technique of discovering relevant sequential patterns among large databases with user specified minimum support [13]. It find out frequent sub-sequences from a given set of data sequences. The sequential pattern problem was first introduced in 1995 by Agrawal and Srikant [13] and was defined as:

"*Given a database of sequences, where each sequence consists of a list of transactions ordered by transaction time and each transaction is a set of items, sequential pattern mining is to discover all sequential patterns with a user-specified minimum support, where the support of a pattern is the number of data-sequences that contain the pattern*."

Some basic concepts [8] of Sequential Pattern Mining are:

1. An itemset is the non-empty subset of items. Let I={i1,i2…,iN} is the set of items where each item is associated with some attribute. Itemset with k items is called p-*itemset*.
2. A Sequence α= {A1,A2,…,Al} is an ordered collection list of item sets. An item set Ai where (1 ≤ i ≤ l) in a sequence is called a transaction.
   Number of transaction in a sequence is called the *length* of the sequence. A sequence with length *l* is called *l-sequence.* If there are two given sequences α={A1,a2,…,An} and β ={B1,B2,…,Bm} (n<=m) then α is called a subsequence of β denoted by α⊆β, if there exist integers 1≤ i1< i2<…< in≤m such that A1⊆Bi1, A2⊆Bi2,…, An⊆Bin .
3. *Sequential Pattern Mining Algorithm:*
   Sequential pattern mining algorithm can be divided into two parts. They are Apriori-based and Pattern growth based.

I. **Apriori-Based algorithm**

The algorithms depend on apriority property which states that "if a sequence S is not frequent then none of the super-sequences of S will be frequent". Key features of Apriori-based algorithm are Breadth first search, Generate and Test, Multiple scans of the database. Some algorithms are GSP, SPADE, SPAM .
1) GSP (Generalized Sequential Pattern): This algorithm makes passes to the data multiple times. Steps involved in GSP are Candidate Generation and Candidate Pruning Method. The outline of the method [8] is:

- Initially every item in database is of length-1.
- -for each level (sequences of length-p) do
- Scan the database to collect support count for each candidate sequence
- Using Apriori generate candidate length-(p+1) sequences from length-p frequent sequences
- Repeat until no frequent sequence or no candidate could be found.

2.) SPADE (Sequential Pattern Discovery using Equivalence classes): It is an Apriori based vertical format sequential pattern algorithm [12]. The sequences in database are given in vertical order rather than horizontal order. It consist of id list pairs *(sequence-id, timestamp).* The first value denotes customer sequence and second value stand for transaction in it. This algorithm can make use of Breadth first and Depth first search. Lattice theoretic approach is used to divide the original search space into smaller sub-lattices which can be processes in main memory independently. SPADE algorithm reduces the input-output costs as well as computational costs.
3.) SPAM (Sequential Pattern Mining): SPAM [14] uses a vertical bit-map data structure representation of a database. It integrates the concept of GSP, SPADE, Free-span algorithms. The sequence tree is traversed in depth first manner and thus the performance is increased. Merging cost is reduced but it takes more space compared to SPADE and thus there is a space-time-tradeoff.

II**. Pattern Growth algorithm**

In Apriori-based algorithms the candidate generation was exponential. When database were large, candidate generation and candidate pruning method suffered greatly. This lead to Pattern growth algorithms in which candidate generation and candidate pruning were avoided. Pattern growth algorithms are complex algorithms to develop, test and maintain but these algorithms are faster when the database size is large. The key features of pattern based algorithms are partitioning of search space, tree projection, depth-first traversal. Some pattern-based algorithms are FREESPAN, PREFIXSPAN.

1) *FREESPAN (Frequent pattern-projected Sequential Pattern Mining):* This algorithm was developed to reduce candidate generation and testing of Apriori. Frequent Items are used to recursively project the sequence database into projected database and make use of the projected database to confine the search and growth of subsequence fragments [8,14]. The size of each projected database reduces with recursion.

2) PREFIXSPAN (Prefix-projected Sequential pattern mining): This algorithm [11] looks for the frequent items while going through database in one iteration and hence, is the fastest algorithm. It makes use of a divide and search technique, and reduces the size of projected database with reduction in effort of candidate subsequence generation. The cost of memory space may be high due to creation of huge number of projected sub-databases. It is a DFS based approach. It is one of the efficient pattern growth method and performs more efficiently than GSP and Freespan. Prefixspan makes the search space smaller. Longer sequential patterns are built on smaller counterparts.

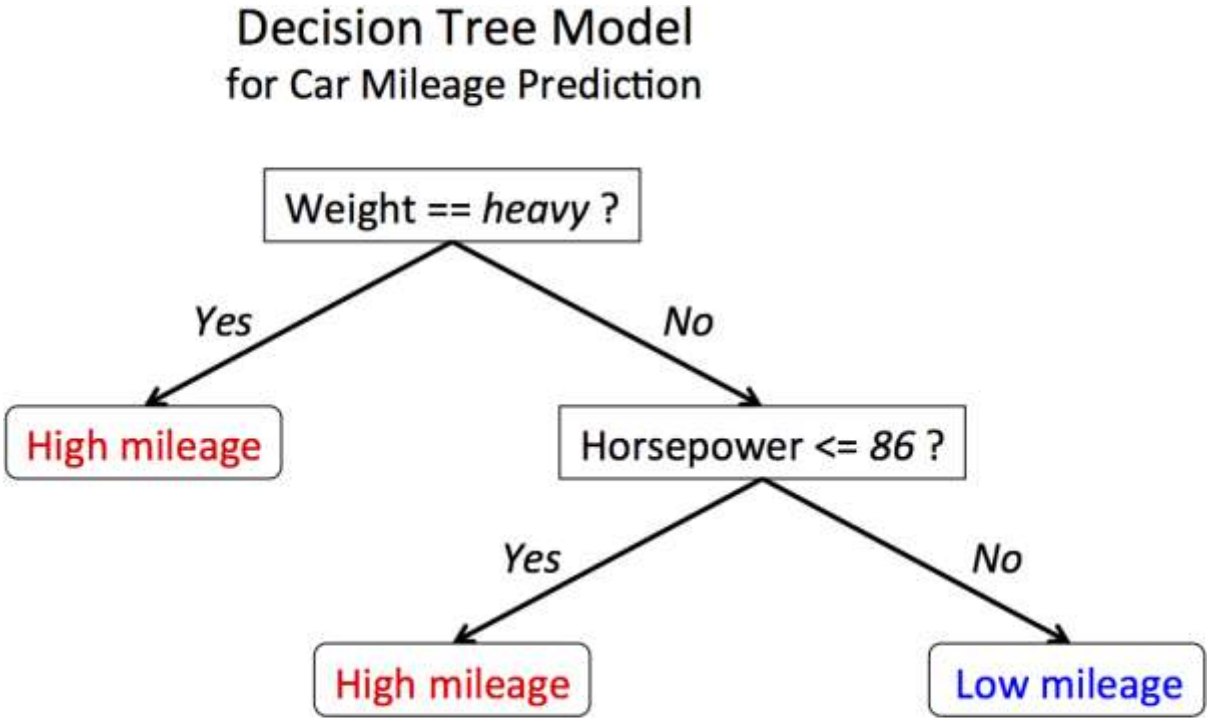
D. *Classification and Regression*

These fall under predictive model.Classification basically means examining an objects and then allotting it to pre defined classes.It requires us to build a model or a training set according to which the classification is to occur

E. Predictive Data Mining Algorithms

A basic knowledge of data mining algorithms is essential to know when each algorithm is relevant to about advantages and disadvantages of each algorithm and use of the algorithms to solve real-world problems.

1.) Decision Trees: Most popular of learning algorithms for classification tasks is decision tree algorithm. An attribute is represented by the internal nodes of the decision tree and terminal nodes are labeled with a class value., Until a leaf node is reached, the branch that matches the attribute value is followed. Predicted value for the example is the class value assigned to the leaf node. Let's take an example in which decision tree on their automobile mileage if their automobile weight is "heavy" or it is not "heavy" but mileage is "low" if the horsepower is "<=86" with not "heavy" automobiles.



2.)*Rule Based Classifiers:* Rule sets are generated from rule based classifiers. A rule covers a specific situation or instance. First, if rule is satisfied directly then it is assigned the class associated to the rule. If it satisfy more than one rule, the final decision is based on a voting scheme. Rule-based classifiers are similar to decision tree learners. It has similar comprehensibility, computation time and expensive power. This can be converted to Decision tree, where one branch satisfy the Rule and second does not. Some rule-based learners such as C4.5Rules (Quinlan 1993) operate this way, other rule learners, such as RIPPER (Cohen 1995), generate rules directly.

3.)*Artificial Neural Networks:* Artificial Neural Networks (ANNs) or "Neural Net" were developed to imitate the complex network neurons make in our brain and function just like a "Human Brain" does. It could perform both regression and classification tasks. A neural net consists of input layer and output layers between which you have a hidden layer where computation occurs.

The input layer to output layer is traversed as follows. First, the input values are chosen from the features of the training example, and is inputted to neural net. These values are then assigned specific weights and further fed to the next node. Now weighted sum is passed through a non-linear activation function and then the resulting value is passed to the next layer, where this process is repeated, until the final value(s) are resulted. The ANN learns by changing the values of the weights according to which our final values are varied. Now this process is repeated till the output value start coming close to the actual outputs in training set. Now there are many well known algorithms to modify the weights being assigned to the inputs like Back propagation algorithm, genetic algorithm for neural networks.

Future work

The definition of data mining technique is not complete as other methodologies such as social science methodologies were not included in the survey. Research technology that is often used in social studies are qualitative questionnaires and statistical methods. In future DMT might help in social science to develop method to investigate specific problems related to human behaviour and psyche.

DMT is an interdisciplinary research topic and would grow with more widespread application. New insights into the problems associated with DMT may be offered by the integration of methodologies and cross-disciplinary research. Many new industries are entering with passion and vigour to explore this field. Several advances have been made in many sectors but still there is an open scope for improvement. Few are listed.

- Healthcare: - Patient data could be monitored with the help of IOT devices such as fitness bands and smart-watches and the doctors could have a regular update on the vitals if needed.
- Home Appliances: - Several devices are in market which help you automate your home appliances and control them. These devices record huge amount of data which could be leveraged to help make better products with better efficiency and more over help curtailing the excessive wastage of natural resources.
- Social campaign/surveys: - Survey would be easier and better if DMT could help make more sense of data. New trends and patterns could be found when data mining is applied on the survey data and more information on how the people in masses perceive news/events.

Conclusions

The crucial thing for making right decision is having right information. The problem of collecting data, which used to be a major concern for most organizations, is almost resolved. Organizations will be competing in generating information from data and not in collecting data in millennium. It has been indicated by industry surveys that over 80 percent of Fortune 500 companies believe that data mining would be a critical factor for business success by the year 2000. For sure in the coming future DATA mining will be one of the main competitive focuses of organizations. Many issues remain to be resolved and much research has to be done, though progresses are continuously being made in the data mining field. [15]

ACKNOWLEDGEMENT

I would like to thanks my friends and family for supporting me through this paper. My interest in this field was ignited when I started with machine learning in an online course. While reading and searching about this topic we learnt about various important and interesting facts.

REFERENCES

1. Nikita Jain, Vishal Srivastava, "DATA MINING TECHNIQUES: A SURVEY PAPER",IJRET: International Journal of Research in Engineering and Technology, Volume: 02 Issue: 11 | Nov-2013
2. Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam, "A Study of Data Mining Tools in Knowledge Discovery Process "International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue- 3,July 2012
3. Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", AI magazine,1997
4. QIANG YANG, XINDONG WU, "10 CHALLENGING PROBLEMS IN DATA MINING RESEARCH", International Journal of Information Technology & Decision Making Vol. 5, No. 4 (2006) 597–604
5. G. Weiss and B. Davison,"Data Mining", in Handbook of Technology Management, John Wiley and Sons, expected 2010.
6. Kalyani M Raval, "Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering,Volume 2, Issue 10, October 2012
7. Jayanthi Ranjan, "APPLICATIONS OF DATA MINING TECHNIQUES IN PHARMACEUTICAL INDUSTRY", Journal of Theoretical and Applied Information Technology, 2005 - 2007 JATIT.
8. Chetna Chand, Amit Thakkar, Amit Ganatra, "Sequential Pattern Mining: Survey and Current Research Challenges", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012
9. Irina Tudor,"Association Rule Mining as a Data Mining Technique",Vol. LX No. 1/2008

10. Richard A. Huebner, "A survey of educational data-mining research", Research in Higher Education Journal.
11. V. Uma, M. Kalaivany, Aghila, ”Survey of Sequential Pattern Mining Algorithms and an Extension to Time Interval Based Mining Algorithm”, International Journal of Advanced Research in Computer Science and Software Engineering
12. M. Zaki, “SPADE: An efficient algorithm for mining frequent sequences”, Machine Learning, 2001.
13. R. Agrawal and R. Srikant, “Mining Sequential Patterns”. In Proc. of the 11th Int'lConference on Data Engineering, Taipei, Taiwan, March 1995.
14. J. Han, G. Dong, B. Mortazavi-Asl, Q. Chen, U. Dayal and M.-C. Hsu, “FreeSpan: Frequent pattern-projected sequential pattern mining”, Proc. 2000 International Conference of Knowledge Discovery and Data Mining (KDD’00), pp. 355-359, 2000.
15. Sang Jun Lee and Keng Siau, “A review of data mining techniques”. Industrial Management & Data Systems 101/1[2001]41- 46
16. Fayyad, Djorgovski, & Weir, 1996
17. Data Mining and Data Warehousing(MAKAUT) - Dr. Bikramjit Sarkar Assistant Professor Dept. of Computer Science and Engineering ,Dr. B. C. Roy Engineering College Jemua Road, Fuljhore, Durgapur
18. Review of Data Mining Techniques (IJSWS)-2014 by Sandeep Panghal and Priyanka Yadav .