INTERNATIONAL JOURNAL FOR TRENDS IN ENGINEERING &TECHNOLOGY
VOLUME 27 ISSUE 1 – MARCH 2018 - ISSN: 2349 – 9303

Data Mining and Its Techniques

C.Sangeetha [1]
[1] PhD Scholar, Department of Computer Science
CMS College of Science and Commerce, Coimbatore,Tamilnadu
cgss369@gmail.com

Dr.V.Chitraa[2]
[2]Associate Professor, Department of Computer Science
CMS College of Science and Commerce,
Coimbatore,Tamilnadu
chitramanikam@gmail.com

Abstract— In this paper, the concept of data mining was summarized and its significance towards its methodologies and techniques was illustrated. Which attract attention from researchers in from various fields which includes Data mining Techniques and its Applications are an important kind of pattern which occur frequently in many fields.

Keywords— Data Mining, Data mining Techniques, Data mining applications, Knowledge discovery process, Pattern, Sequential Patterns.

1 INTRODUCTION

Data mining refers to extracting or mining the knowledge from large amount of data. The term data mining is appropriately named as "Knowledge mining".

Data collection and storage technology has made it possible for organizations to accumulate huge amounts of data at lower cost. Exploiting this stored data, in order to extract useful and actionable information, is the overall goal of the generic activity termed as data mining.

2 Data Mining

Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development. Three steps involved are Exploration Pattern identification and Deployment

Exploration: In the first step of data exploration data is cleaned and transformed into another form, and transformed into another form, and important variables and then nature of data based on the problem are determined.

Pattern Identification: Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction.

Deployment: Patterns are deployed for desired outcome.

Data mining is an interdisciplinary subfield of computer science which involves computational process of large data sets' patterns discovery. The goal of this advanced analysis process is to extract information from a data set and transform it into an understandable structure for further use. The methods used are at the juncture of artificial intelligence, machine learning, statistics, database systems and business intelligence. Data Mining is about solving problems by analyzing data already present in databases[2]

Data mining is also stated as essential process where intelligent methods are applied in order to extract the data patterns.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

3 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.[3]

3.1 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

3.2 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

3.3 Regression

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 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
- Non linear Regression
- Multivariate Nonlinear

3.4 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 behavior 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

3.5 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 fore casting needs.

Types of neural networks

- Back Propagation

3.6 Artificial Neural Networks

Non-linear predictive models that learn through training and resemble biological neural networks instruction.

3.7. Decision trees

Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square. Automatic Interaction Detection(CHAID). CART and CHAID are decision tree techniques used for Classification of a dataset. They provide a setoff rules that Can apply to a new (unclassified) dataset to predict which records will have a given out come. CART segments a dataset bycreating2-waysplits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

3.8. Genetic algorithms

Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts combination ,mutation, and natural selection in a design based on the concepts of natural evolution

3.9 Nearest neighbor method

A technique that classifies a record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1), sometimes called the k-nearest neighbor technique.

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

Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

4.Data Mining Applications

Data mining is a relatively new technology that has not fully matured. Despite this, there are a number of industries that are already using it on a regular basis. Some of these organizations include retail stores, hospitals, banks, and insurance companies. Many of these organizations are combining data mining with such things as statistics, pattern recognition, and other important tools. Data mining can be used to find patterns and connections that would otherwise be difficult to find. This technology is popular with many businesses because it allows them to learn more about their customers and make smart marketing decisions. Here is overview of business problems and solutions found using data mining technology.[4]

5.Pattern Mining

Pattern mining concentrates on identifying rules that describe specific patterns within the data. Mining frequent patterns is probably one of the most important concepts in data mining. A lot of other data mining tasks and theories stem from this concept. To identify the pattern to improve the quality of IT Services pattern is identify the patterns. This research work focuses on various types of pattern mining frame works for Domain Driven Data mining.

CATEGORIES OF SEQUENTIAL PATTERNMINING

Sequential pattern can be partitioned into three categories. Periodic patterns, Statistically patterns, and Approximate patterns.

5.1 Periodic Patterns

This model is not flexible and it is unable to find patterns whose occurrences are asynchronous[8]. Periodicity detection in time series database is an important data mining problem and has a number of applications. For example, “The gold prices increase every weekend” is a periodic pattern. As this model is restrictive it may fail to detect some interesting pattern if its occurrence is misaligned due to noise events.

5.2 Approximate Pattern Mining Techniques

Mining of Closed Sequential Patterns

RaminAfshar proposed a frequent closed subsequence mining approach CloSpan[3] that mines large sequences efficiently. This algorithm produces number of efficiently search pruning techniques. The algorithm makes use of hash technique that has two steps to carry out efficient optimization of the search space: 1) it create a superset of joint common sequences known as the LS set, and keeps the set in prefix order and 2) then it performs post-pruning to eliminate non-closed sequences. It works in following manner:

- Classification is performed on each set of item and carries out the elimination of not frequent items and sequences that are empty.
- The Clospan method is recursively applied on the prefix search tree in depth first search manner and builds the prefix sequence corresponding to it. Lastly, it removes the free sequences.
- Then is uses a hash index on the projected database size and only the sequences whose projected database size is same as that of current sequence are tested.

This algorithm performs well for exact pattern matching problems that is not suitable for approximate pattern matching

problems. As the dataset size increases, execution time of the algorithm also increases rapidly.

5.3 Sequence Mining in Domain Categories

Mohammed J. Zaki proposed cSPADE[6]algorithm for mining frequent sequences. It is an efficient algorithm based on a number of syntactical limitations. They are size of the sequences, limiting the min or max gap on consecutive sequence elements, to put a time slot on acceptable sequences and searching sequences that are predictive of one or multiple classes, even rare ones. This algorithm methodically searches the sequence grid formed by the subsequence relation, from the simplest items to the nearly particular frequent sequences in a depth-first (or breadth-first) manner. It requires preprocessing of data in a special format, as it is based on syntactical constraints. For large database more time is needed for pre-processing the data as a result the performance degrades.

5.4 Motifs Mining using Random Projections

J. Buhler has proposed an algorithm based on random projections of input sub strings[7]. It carries out a number of tests on a basic iterant. The Projection algorithm has two parts:

- A random projection is selected and hashed with each l-mer x in the input sequences to its hash bucket with every test.
- The required pattern is searched in a hash bucket that has adequate entries by applying an order of improving steps
-

This algorithm is more productive in searching required pattern in simulated data, but it needs improvement in time, in space and maybe for future more complex biological data and real time.

6. Conclusion and Future scope

Data mining, Data Techniques and Applications is the computer-assisted process and analyzing enormous sets of data and then extracting the meaning of the data. It is applied effectively not only in business environment but also in other fields such as weather forecast, medicine, transportation, healthcare, insurance, government.

The paper theoretically shows the three types of sequential pattern model and some properties of it. These models fall into three categories called periodic pattern, statistically pattern, and approximate pattern. The first model is rigid but provides full periodicity and partial periodicity. In contrast, in partial periodicity, sometime points contribute to the cyclic behavior of a time series. Use of information gain as new metric helps to find surprising patterns which comparing with the frequent patterns demonstrates the superiority of surprising patterns. The third model, approximate sequential pattern, provides a means to verify noise. But still being and active research area in the data mining field.

Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc., in different business domains. Data mining techniques and algorithms such as classification, clustering etc., helps in finding the patterns to decide upon the future trends in businesses to grow. Data mining has wide application domain almost in every industry where the data is generated that's why data mining is considered one of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in Information Technology.

7. REFERENCES

[1]Jiawei Han and ichelineKamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman,2nded.

[2]Joseph, zernik," Data Mining as a civic Duty – Online Public prisoners registration systems", International Journal on social media

[2

[3]Dr. Gary Parker, vol 7, 2004, Data Mining: Modules in emerging fields, CD-ROM.

[4]Crisp-DM 1.0 Step by step Data Mining guide fromhttp://www.crisp-dm.org/CRISPWP-0800.pdf.

[5]X.Yan,J.Han, and R.Afshar," CloSpan: Mining Closed Sequential Patterns in Large Datasets," Proc.SIAMInt'l Conf. DataMining(SDM),2003.

[6] M.J. Zaki, "Sequence Mining in Categorical Domains: Incorporating Constrains," Proc. Ninth Int'l Conf. Information and Knowledge Management (CIKM), pp. 442-429,2000.

[7] . Buhler and M. Tompa, "Finding Motifs Using Random Projections,"

[8]Zhao Q., and Bhowmick S. S., Sequential Pattern Mining: A Survey, Technical Report, CAIS, Nanyang Technological