International Journal of Latest Trends in Engineering and Technology
Vol.(12)Issue(4), pp.052-056
DOI: http://dx.doi.org/10.21172/1.124.10
e-ISSN:2278-621X

A SURVEY ON USE OF DATA MINING TECHNIQUE IN DIFFERENT DOMAIN

Urvashi Sangwan[1]

Abstract- Data mining is a process which finds useful patterns from large amount of data. Data mining is a knowledge discovery that extracts useful information. It has been a major advance in machine learning artificial agent systems and decision making in the expert systems. The last decade the researcher has surveyed most of the techniques and applications that used in different fields in our life such as manufacturing education engineering and business. The paper discusses few of the data mining techniques algorithms and some of the domains which have adapted in data mining technology to improve their businesses education healthcare and found excellent results.
Keywords: Data mining Techniques; Data mining algorithms; Data mining applications.

1. OVERVIEW OF DATA MINING

The development of Information Technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process knowledge mining from data knowledge extraction or data /pattern analysis.

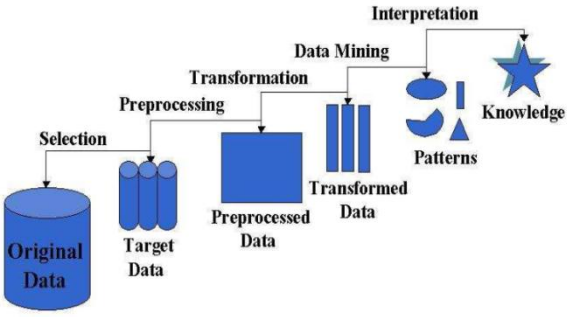

Figure 1. Knowledge discovery Process

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 of their businesses.

Three steps involved are

- Exploration
- Pattern Identification
- Deployment

Exploration:
In the first step of data exploration data is cleaned 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.

2. DATA MINING ALGORITHMS AND TECHNIQUES

2.1. 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

[1] Assistant professor Vaish college of Engg Rohtak

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[1].

Types of classification models:

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

2.2. 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

2.3. 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 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
- Multiplicative Linear Regression
- Non Linear Regression
- Multivariate Nonlinear Regression

2.4. 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 [2].

Types of association rule

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

2.5. 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 forecasting needs[3].

Types of neural networks

- Back Propagation

3. USES OF DATA MINING IN DIFFERENT DOMAIN

3.1 Cloud computing using hadoop

This paper combines the latest cloud computing technology to improve the traditional data mining algorithm and uses Hadoop platform to improve the parallel processing ability of the algorithm. The K-Means algorithm relies on the initial k-value and the initial center point and is combined with the Hadoop platform features. Before the K-Means algorithm clusters the Hadoop platform is used to sample the initial data and the neighborhood density is used to determine the initial center point. Then cluster again. Based on the previous analysis of the defect of K-Means algorithm The initial k value and the center point are determined by the sample and density and the defect of specifying the k value and the initial center point in the initial stage is solved. The K-Means algorithm will be improved MapReduce and the ability to process data in parallel using Hadoop will improve the scalability of the K-Means algorithm. Finally the algorithm proved to be more scalable during the experiment[4].

3.2. Software analytics

This paper claims that a new field of empirical software engineering research and practice is emerging: data mining using/used-by optimizers for empirical studies or DUO. For example data miners can generate models that are explored by optimizers. After collecting data about software projects and before making conclusions about those projects there is a middle step in empirical software engineering where the data is interpreted. When the data is very large and/or is expressed in terms of some complex model of software projects then interpretation is often accomplished in part via some automatic algorithm. Data miners“slice” data such that similar patterns are found within each division.Optimizers “zoom” into interesting regions of the data using a model to fill in missing details about those regions. For requirements engineers we can find the least cost mitigations that enable the delivery of most requirements. For project managers we can apply optimizers to software process models to find options that deliver more code in less time with fewer bugs. For developers our optimizers can tune data miners looking for ways to find more bugs in fewer lines of code (thereby reducing the human inspection effort required once the learner has finished Genetic Algorithms Different evolution multi-objective evolutionary algorithms Particle Swarm Optimization (PSO) Genetic programs etc For software analytics we could try to learn data miners that find the highest priority bugs after the fewest tests found in the smallest methods in code that is most familiar to the current human inspector. Such a data miner would return the most important bugs and easiest to fix (thus reducing issue resolution time for important issues. For project management when crowd sourcing large software projects we could allocatetasks to programmers in order to minimize development time while maximizing work assignments to programmers that have the most familiarity with that area of the code. For refereeing new research results in SE the tools described here could assign reviewers to new results in order to minimize the number of reviews per reviewer while maximizing the number of reviewers who work in the domain of that paper we anticipate several years where DUO is explored by a minority of software analytics researchers. That said by 2025 we predict that DUO will be standard practice in software analytics[5].

3.3.Data mining in IOT

Internet of Things (IoT) accumulates bulk of data from heterogeneous devices implanted with sensors. This data is accumulated over a period of time from sensory devices and is maintained on a server. To take optimal decisions in realtime meaningful information need to be extracted out from the data accumulated. Numerous data mining (DM) techniques are available for analyzing the data and then to make future predictions based on the discovered information. The number of devices connections in IoT is expected to reach 25 to 30 billion by 2020 and so as the new applications of IoT are going to emerge. Therefore an efficient and fast DM technique is required to make predictions and take decisions in real time to maintain the goodwill of IoT in society. This paper first discusses three different DM techniques: classification clustering and association based mining and their possible combinations. Later this work highlights the applications of using a particular DM technique in IoT. Finally a comparative analysis is made among each DM technique on the basis of its precision accuracy and recall value. This will lead to identity the best DM technique that can be applied in IoT. There are in total 12 combinations possible from these 3 techniques[6]. For example clustering technique can be combined individually either with classification or association based mining or it can be combined with other two together indifferent order

3.4. Bit Combination

A frequent item set mining algorithm based on bit combination is proposed in this paper. Frequent item set mining algorithm based on bit combination is an algorithm that searches for possible frequent item set by transforming data into binary bit representation and adding data representing the combination of regulatory elements step by step and then mining frequent item set by bit and calculation. In the process of data mining the algorithm is optimized by pruning preprocessing and frequent item set culling. Because the recursive method used by traditional frequent item set mining algorithms such as FP-Growth algorithm can' effectively parallel a large number of data the greatest advantage of this algorithm is that it facilitates the parallel computation of data and provides a new idea for improving the efficiency of frequent item set mining. The parallel acceleration of the algorithm is realized by using Open MP technology to verify the parallel feasibility of the algorithm in this paper. Apriori algorithm and FP-growth algorithm Soybean promoter data will be used as input data to mine frequent patterns among regulatory elements in this experiment. The regulatory element of promoter data is composed of short DNA sequences. It binds specifically with transcription factors. It controls the start time and expression degree of gene expression. Like "switch" it usually binds with transcription factors to control gene activity The promoter data are stored with binary bits in this algorithm and are expressed through the combination of control elements. The frequent item set of the promoter data is calculated by bitwise and and the computing efficiency can be improved by pruning and preprocessing. This algorithm breaks away from the recursive method adopted by the classical FP-growth algorithm in the process of data mining. It is fit for the parallel processing. It provides a new idea for improving the efficiency of data mining.

However this algorithm also has shortcomings. After adding parallel because of the imbalance of promoter data satisfying support thread allocation is not the optimal solution so that the parallel performance improvement has not achieved the desired results but compared with serial algorithm the computational efficiency has been significantly improved.

3.5.Health care.

In this paper we have used the UCI machine learning repository Cleveland heart disease database having 303 instance and 76 attributes. For the proposed method we have used the Informa- tion gain concept for selection of best attribute and processes the selected features using weka and python. This paper identifies the gap of research on prediction of heart disease based on python Anaconda navigator spyder and weka platform on which we have much emphasized. The various techniques processes which have used to train the model of heart datasets such as feature selection numpy pandas library decision tree classifier KNN classifier entropy gini- index confusion matrix.The result shows that decision tree classifier is most effective and appropriate for prediction of UCI repository Cleveland heart dataset. The result shows that decision tree classification technique holds good for the analysis of medical data classification especially for heart diseases. For better performance and more accuracy deep learning techniques can be applied for the diagnosis of heart disease[8].

3.6. User behavior.

The aim of this paper is to do analysis of 7 data mining algorithms and their variants for determining the best appropriate prediction of people interest in facebook pages. For this we will analyze various algorithms such as KStar LWL IBK J48 Naive Bayes Bayes Net Decision Table. Their performances have been evaluated by using a dataset with 10172 instances of 6 attributes each. Among all the utilized algorithms Bayes Net demonstrates the best execution in regard to the accuracy. However in respect to the computational time KStar demonstrates the best execution instead of Bayes Net. Social Networking sites are the massive source for big data that gives the opportunity to study about the human behaviors and interactions in different perspectives. Acquire posts from selected facebook pages. Extract interactive data such as likes comments replies posts name and URLs (page names) texts from posts[9]. Analyze acquired data to explore human behavioral interest. Apply matching techniques on training and test data set for identifying the people interests. Weka tool is applied for data mining. In future we intend to research individuals' enthusiasm to look for progressively precise expectation and make a regressional investigation to proposed the connection between computational time and accuracy.

3.7. Agriculture

In this paper our focus is on what opportunities data mining provide for e-government in Afghanistan. Moreover this study gives some use-cases of data mining techniques for improving agriculture sector in Afghanistan. One of the applications of data mining techniques in e-governance is in the agriculture sector. It can be used to analyze soil predict yield predict climate etc.to find patterns over the historical data. Neural networks: this data mining technique is well suited to tackle problems such as predictions and pattern recombination. • Rule Induction: based on the statistical significance the extraction of useful if-then rules from data is called rule induction. Rule induction method has the capability to use retrieved cases for predictions.

• Case-based reasoning (CBR): this method is very simple. In order to forecast a future situation or to make a correct decision such system check the closest past analogs of the present situation and selects the same solution which was the right one in those past situations therefore this method is also called the nearest neighbor method. • Decision Trees: it has a tree-shaped structure and a set of decisions. These decisions generate rules for the classification of the data set As an example of successful implementation of data mining techniques in other countries specially developing countries we can mention usage of SVM technique to predicting yields fuzzy algorithms for crops management Integrated Component Analysis (ICA) for weather broadcasting K-means algorithm for soil classification interpolation methods for estimating suitability of soils for

maze regression models to estimating percentage of organic matter and predicting soil-water-retaining and ANN algorithms to discover and predict crop diseases[10].

3.8 Education

As an interdisciplinary Field of study Educational Data Mining (EDM) applies machine-learning statistics DataMining (DM) psycho-pedagogy information retrieval cognitive psychology and recommender systems methods andtechniques to various educational data sets so as to resolve educational issues. EDM is concerned with analyzing data generated in an educational setup using disparate systems. Its aim is to develop models to improve learning experience and institutional effectiveness.

The International Educational DataMining Society is the EDM as ``an emerging discipline. One of the pre-processing algorithms of EDM is known as Clustering. It is an unsupervised approach for analyzing data in statistics machine learning pattern recognition DM and Bioinformatics[7]. Data clustering enables academicians to predict student performance associate learning styles of different learner types and their behaviors and collectively improve upon institutional performance.

So far we see that subject specific research has been done but what about domain specific? For instance how do institutions employ or apply data mining methods to improve institutional effectiveness

4. CONCLUSION

Data mining has importance regarding finding the patterns forecasting discovery of knowledge etc. in different domains. Data mining techniques and algorithms such as classification clustering etc. helps in finding the patterns to decide upon the future trends in different domains 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[11].

5. REFERENCES

[1] Jiawei Han and Micheline Kamber (2006) Data Mining Concepts and Techniques published by Morgan Kauffman 2nd ed.
[2] Dr. Gary Parker vol 7 2004 Data Mining: Modules in emerging fields CD-ROM.
[3] Crisp-DM 1.0 Step by step Data Mining guide from http://www.crisp-dm.org/CRISPWP-0800.pdf.
[4] 2019 IEEE International Conference on Power Intelligent Computing and Systems (ICPICS)Research of Data Mining Algorithms Based on Hadoop Cloud Platform by Xiangqin Li1 Yurong Hu2 Chuanjun Luo
[5] Better Software Analytics via "DUO":Data Mining Algorithms Using/Used-by Optimizers by Amritanshu Agrawal _ Tim Menzies _Leandro L. Minku _ MarkusWagner _ Zhe Yu
[6] Performance Analysis of Data Mining Techniques in IoT by sahil verma kavita and isha
[7] A Systematic Review on Educational Data Mining by Maizatul Akmar Ismail
[8] Predictive Analysis of Rapid Spread of Heart Disease with Data Mining Radhanath Patra and Bonomali Khuntia
[9] Frequent Item S et Mining Algorithm Based on Bit Combination 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analytics by Renpeng Zhao jun lu and Kailong Zhou Tutut Herawan and ashish dutt
[10] 1st International Conference on Advances in Science Engineering and Robotics Technology 2019 (ICASERT 2019) Data Mining Techniques for Predicting User Interest in Facebook Pages: A ComparisonNusrat Jahan Farin_ Morium Aktery Puthi Royy and Mohammad Shorif Uddin
[11] Data-Mining Opportunities in E-Government: Agriculture Sector of Afghanistan Mursal Dawodi Jawid Ahmad Baktash and Tomohisa Wada