A Review of Data Mining Techniques

Anjli negi, Dr. Varun jaiswal
Student of computer science and engineering, facility of computer science and engineering
Anjalin098@gmail.com, computitionalvarun@gmail.com

Abstract: Data mining is the process of discovering new information from vast dataset in form of interesting patterns, and combination of rules. There is need of a powerful technique for better interpretation for such a vast amount of data. Data mining tools predict future trends and behaviors. This paper present the data mining techniques, some of the tasks of data mining and the real world applications of data mining.

Keywords: *Data mining, Genetic Algorithm, mining tasks, Neural Network, SVM, Bayesian Classification, decision tree, K-nearest neighbor classification.*

*****

I. Introduction

Data mining is an important subfield of computer science. It is the process which performs operations on a large dataset to discovers the pattern, which uses the method of artificial intelligence, machine learning, statistics and database systems[1]. The goal of data mining is to extract information from a huge dataset and convert it into an understandable form for further use[2]. It is an analytic process which is designed to discover data in the search for consistent patterns, systematic relationship between variables and then validates the findings by applying the detected patterns on new data[3]. It is also a process of extracting hidden pattern or information from a large amount of data. Data mining technologies are used in various other fields, some of the data mining applications are in marketing, fraud detection, finance, biological data analysis, telecommunication industry, retail industry, financial data analysis, in the medical field, and in other scientific applications.

Data mining is the important part of the “Knowledge Discovery in Databases” (KDD). It is also known as knowledge mining from data, knowledge extraction or data/ pattern analysis.

There are mainly seven steps in KDD process which are selection, preprocessing, transformation, data mining,interpretation, and evaluation.

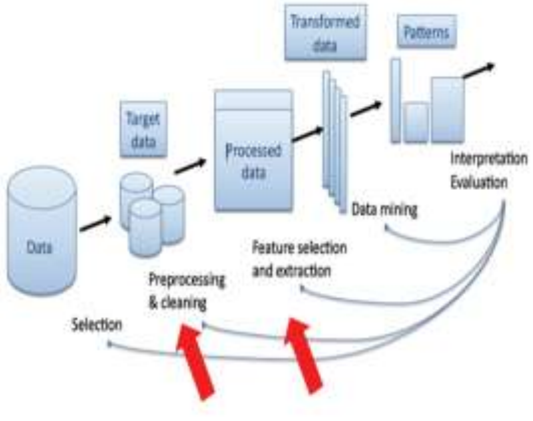

Fig: Data mining process

II. Data Mining Tasks

As there is various type of patterns in a large database, so there are distinct and diverse tasks of data mining. Different type of methods is required to find different kinds of patterns. Task in data mining can be classified into the following type based on the kinds of patterns. These tasks are classification, prediction, clustering, association, summarization, visualization and trend analysis.

1. **Classification**: It is the technique which works on a set of pre-classified data to develop a model that can classify the population of records at large. It is one of the most commonly used technique of data mining. A classification model is developed by analyzing the relationship between the attributes and classes of objects in training data[4]. These model are used to classify future objects and develop the better understanding of classes of objects in the datasets. For example: from patients diagnosed data a classification model can be build, which can help in diagnosing the new patients disease based on the available patients diagnosis data. Classification is one kind of predictive modeling. It is the process in which new objects are assigned to the predefined classes [5]. It includes two steps first is the model construction which defines the class labels and second is model usages to classification object in future.

2. **Clustering**: clustering is the process of identifying similar class’s objects and grouping them together. The objects are clustered on the basis of two rules: the first one is that intra-class similarities are maximized and inter-class similarities are minimized based on attributes of objects. Once the clusters are defined the objects are labeled based on cluster and class description is described based on some common features[6]. For example, a bank can cluster its customers based on income, age, residence etc. and common features are used to describe the group. The clusters are useful to understand the customers and help in providing suitable policies and customer services. Clustering uses various schemes to cluster objects into groups. These schemes are distance based, partition based, hierarchal, model-based and density based. And some of the clustering algorithms are k-mean, k-medoids, expectation maximization, self-organization and competitive learning[4].

3. **Association**: association is helpful in finding frequent item sets among large datasets. This type of finding is helpful businesses to make decisions such as customer shopping pattern analysis, catalog design[7]. It is helpful in finding the connection between objects. Such kind of connection is named as association rules. Association rule discloses the connection between objects. For example, the rule set [onion, potatoes] [hamburger] find in the data of sales of the supermarket then it means that customer buy onion and potatoes together, and he or she likely to buy the hamburger.
4. **Summarization**: it is the abstraction and generalization of data. A set of appropriate data is generalized or summarized to form a smaller set of data which can give an outline of data. For example, calls can be summarized into the local call, STD call, ISD calls, etc.[8].
5. **Trend analysis:** Now a lot of data is available among which time series data is also available which is gathered over time. For example stock market data, customer money transaction, sales market data, etc. this type of data is observed as objects having on attribute as time, and objects are pictures of entities whose value changes over time[9]. Trend analysis finds the interesting pattern in the growth history of the object. One example of trend analysis is the identification of the pattern in an object evolution.

6. **Prediction:** It is the method that determines the relationship between the independent and dependent variables. It associated to the regression technique[8]. Independent variables are those which are already known and dependent variables are those which we want to predict. It uses the continued valued function to predict the unknown and missing values. Prediction is used in various fields some of the applications of prediction are credit approval, target marketing, medical diagnoses and fraud detection.

7. **Visualization:** It is the process of presenting the data into a form so that users are able to understand the difficult patterns. It is used in the combination with other techniques to provide the clear understanding of discovered pattern and to understand the relationship between the patterns or data[10]. Example of visualization model are 3D graphs, hyper-graph and SEENET etc.

Data Mining Techniques

Data mining techniques categorized into various types based on which kind of data to be exposed, which kind of knowledge to be discovered and which kind of techniques to be used. Data mining adopts its techniques from various areas such as neural network, machine learning, database systems and visualization[11].

1. **Support Vector Machine (SVM)**: it is supervised learning technique. It uses the labeled training data to generate the input and output mapping function. The mapping functions are two type first one is classification that classify data and second is regression analysis that estimate the desired output [12]. For classification nonlinear kernel is used which convert the input data into higher dimensional data. The foundation of SVM was laid by Vapnik and it became more popular due to some of the attractive features[13]. It can parallel reduce the empirical classification error and maximizes the geometrical margin it is a reason that SVM is called maximum margin classifier.For classifying of data, it builds a hyperplan by maximizing the margin between two classes. It takes a number of examples and assigns them to one or more category according to the condition it belongs and builds a model, according to which new data can be classified.
2. **Neural Network:** It is a collection of input and output units which are connected to each other, every connection have a weight associated with it. During the learning phase, it learns by adjusting the weights to predict the correct class level of the input[14]. In neural the network, there is a remarkable ability to extract meaning from complex data and can mine patterns and detect trends that are too complex to understand by human and computer techniques. Neural network are best in identify pattern from data so this method is appropriate for prediction and weather forecasting.Some of the applications of the neural network are handwriting recognition, real-world business applications, recognition, for training a computer to pronounce English and facial expression.
3. **Decision tree:** Decision treeis an algorithm used for classification problems. It is a supervised learning algorithm. It works for both continuous and categorical dependent variables. In this algorithm, we divide the population into two or more identical sets. It classifiesthe instances by arranging them based on feature values to make as separate groups as possible[15]. The decision tree is executed with the separate recursive examination in branches to build a tree for prediction. In a decision tree, building process inputs are divided into two or more subgroups and steps are repeated until a complete tree is built. A decision tree includes a root node, branches and leaf nodes[1]. Each node represents a feature in an instance to be classified, and each branch represents a value that the node can assume.During the training phase, single best rule is determined by considering different descriptors limits for separating the compounds according to their activity to build a decision tree.
4. **Bayesian Classification:** It is supervised learning algorithm and also a statistical method for classification. It is based on Bayes Theorem. It is named after Thomas Bayes (1702-1761), who proposed the Bayes Theorem[16]. Set of random

variables and their conditional dependencies were represented by using directed acyclic graph (DAG). Bayesian network is divided into two phase first is to learn the DAG structure of network and second is to determine its parameters. There is two type of probabilities: Posterior probability [P(H/X)] and Prior probability [P(H)]. Where X is data tuple and H is some hypothesis. According to Bayes theorem: P(H/X)=P(X/H)P(H)/P(X)[17]. It offers practical learning algorithm, prior knowledge and observed data can be combined. It is robust to noise in input data and it calculate explicit probabilities for hypothesis. It can be used in various applications like text classification, spam filtering, hyper recommender system and in online applications. It is mostly used for situation when dimensions of the inputs is high. It uses maximum likelihood for parameter estimation.

5. **K-nearest neighbor classification (KNN):**It locates the k nearest instances to the query instance and determines its class by identifying the single most frequent class label. K-nearest neighbor classifier, looks in training set for group of k objects that are close to the test object, and based on that assign the label[12]. There are three elements of this approach first is a set of labeled objects, second is distance between objects and third is the value of k (number of nearest neighbors). To classify the new object, the distance between the new object and labeled object is calculated, its k-nearest neighbors are identified and then the label of these nearest neighbors are used to determine the class label of the new object[18].
6. **Genetic Algorithm:** It is a search heuristic that mimics the process of natural selection. It is useful algorithm to generate the solution to optimization and search problems. It is part of evolutionary algorithms that generate the solutions to optimization algorithms using techniques like mutation, crossover, inheritance and selection[19]. The father of original genetic algorithm was John Holland who invented it in early 1970's[20]. Genetic algorithm explore the historical information to direct the search into the region of better performance within the search space.

Conclusion

In this paper, detailed study of data mining has presented with various studies like tasks of data mining, techniques, and applications. Data mining is a very dynamic research area and development area that is reaching maturity. Data mining is the computer based process of analyzing enormous sets of data and then extending the meaning of the data. Data mining tools can answer questions that traditionally takes much time to resolve due to a vast amount of data.

REFERENCES

[1] Chen, M.-S., J. Han, And P.S. Yu, Data Mining: An Overview From A Database Perspective. Knowledge And Data Engineering, Ieee Transactions On, 1996. 8(6): P. 866-883.

[2] Fayyad, U., G. Piatetsky-Shapiro, And P. Smyth, From Data Mining To Knowledge Discovery In Databases. Ai Magazine, 1996. 17(3): P. 37.

[3] Rygielski, C., J.-C. Wang, And D.C. Yen, Data Mining Techniques For Customer Relationship Management. Technology In Society, 2002. 24(4): P. 483-502.

[4] Kaur, R., Et Al., An Overview Of Database Management System, Data Warehousing And Data Mining, 2013, Ijarcce.

[5] Zaki, M.J., Spade: An Efficient Algorithm For Mining Frequent Sequences. Machine Learning, 2001. 42(1-2): P. 31-60.

[6] Jain, A.K., M.N. Murty, And P.J. Flynn, Data Clustering: A Review. Acm Computing Surveys (Csur), 1999. 31(3): P. 264-323.

[7] Karaolis, M., Et Al. Association Rule Analysis For The Assessment Of The Risk Of Coronary Heart Events. In Engineering In Medicine And Biology Society, 2009. Embc 2009. Annual International Conference Of The Ieee. 2009. Ieee.

[8] Ngai, E.W., L. Xiu, And D.C. Chau, Application Of Data Mining Techniques In Customer Relationship Management: A Literature Review And Classification. Expert Systems With Applications, 2009. 36(2): P. 2592-2602.

[9] Granger, C.W.J. And P. Newbold, Forecasting Economic Time Series2014: Academic Press.

[10] Shaw, M.J., Et Al., Knowledge Management And Data Mining For Marketing. Decision Support Systems, 2001. 31(1): P. 127-137.

[11] Han, J., M. Kamber, And J. Pei, Data Mining: Concepts And Techniques2011: Elsevier.

[12] Kotsiantis, S.B., I. Zaharakis, And P. Pintelas, Supervised Machine Learning: A Review Of Classification Techniques, 2007.

[13] Singh, V. And K. Chaturvedi. Entropy Based Bug Prediction Using Support Vector Regression. In Intelligent Systems Design And Applications (Isda), 2012 12th International Conference On. 2012. Ieee.

[14] Rumelhart, D.E., B. Widrow, And M.A. Lehr, The Basic Ideas In Neural Networks. Communications Of The Acm, 1994. 37(3): P. 87-93.

[15] Quinlan, J.R., Induction Of Decision Trees. Machine Learning, 1986. 1(1): P. 81-106.

[16] Lunn, D., Et Al., The Bugs Book: A Practical Introduction To Bayesian Analysis2012: Crc Press.

[17] Korb, K.B. And A.E. Nicholson, Bayesian Artificial Intelligence2010: Crc Press.

[18] Wu, X., Et Al., Top 10 Algorithms In Data Mining. Knowledge And Information Systems, 2008. 14(1): P. 1-37.

[19] Srinivas, M. And L.M. Patnaik, Genetic Algorithms: A Survey. Computer, 1994. 27(6): P. 17-26.

[20] Pulido, M., P. Melin, And O. Castillo. Optimization Of Type-2 Fuzzy Integration In Ensemble Neural Networks For Predicting The Us Dolar/Mx Pesos Time Series. In Ifsa World Congress And Nafips Annual Meeting (Ifsa/Nafips), 2013 Joint. 2013. Ieee.

[21] Agrawal, R. And R. Srikant. Fast Algorithms For Mining Association Rules. In Proc. 20th Int. Conf. Very Large Data Bases, Vldb. 1994.