ANALYSING THE APPLICATIONS OF KNOWLEDGE DISCOVERY IN DATA MINING.

SONAM

M. Tech, Department of computer Science and Engineering

ABSTRACT

Finding meaningful information and patterns in data is known as Knowledge Discovery in Databases. In the future decades, data mining research will continue to flourish in businesses and learning organizations alike. Data mining is an important topic of study, and we have covered a wide range of methods, methodologies, and research fields in this work. As we all know, many multinational corporations (MNCs) and major organizations have offices all over the world. Large amounts of data may be generated at any one location. We need tools known as data mining to evaluate, organize, and make decisions with such a large volume of data. These approaches will alter numerous sectors. More data mining applications are discussed in this work, as well as the data mining scope, which will be useful for future research.

KEYWORD: Data Mining, Information, Knowledge, KDD

INTRODUCTION

Data Mining (DM) is the mathematical core of the KDD process, involving the inferring algorithms that explore the data, develop mathematical models and discover significant patterns (implicit or explicit) which are the essence of useful knowledge. Advances in data gathering storage and distribution have created a need for computational tools and techniques to aid in data analysis. Data mining is the extraction of useful patterns and relationships from data sources, such as databases, texts, the web. It has nothing to do however with SQL, OLAP, data warehousing or any of that kind of thing. It uses statistical and pattern matching techniques. The concern in data mining are noisy data, missing values, static data, sparse data, dynamic data, relevance, interestingness, heterogeneity, algorithm efficiency, size and complexity of data. The data we have is often vast, and noisy, meaning that it's imprecise and the data structure is complex. This is where a purely statistical technique would not succeed, so data mining is a solution. Data mining has become a popular tool for analyzing large datasets. The efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed. The proliferation of database management systems has also contributed to recent massive gathering of all sorts of information. Information retrieval is simply not enough anymore for decision-making.

LITERATURE REVIEW

Shaker H. El-Sappagh (2013) Many healthcare executives are swamped with information, but often lack the knowledge to use it effectively. Using Knowledge Discovery in Databases (KDD), companies may transform their data into knowledge. Utilizing rapid and improved clinical decision making, organizations who use KDD methodologies discover that they may reduce healthcare expenditures while enhancing healthcare quality. Data mining and knowledge discovery approaches, applications, and process models that may be used in healthcare settings are reviewed in this article. Additionally, the difficulties of using data mining methods in a hospital setting will be explored.

Yihao Li, (2010) [11] Different types of data, environmental data, financial information, as well as mathematical data are available in the globe. The unbelievable growth of data in this era of network sharing

and information make it difficult to manually analyse, categorise and summarise the data. This study examines the foundations of Current data mining and data mining integration research to develop new technology in data mining integration uncertainty management.

Sanjuktaranijena (2015) The peer-reviewed scientific journal Data Mining and Knowledge Discovery focuses on data mining. Springer Science + Business Media is the publisher. Geoffrey 1 Webb has been the editor-in-chief since 2012. No matter how complex its techniques and applications are, the profession of data mining or knowledge discovery is founded on a very fundamental notion. The process of gathering data from several sources in order to do analysis. Data mining for hard drive recovery is often done to gather data that may be used to enhance a method or procedure. Data mining has become critical to everyday operations in a wide range of fields, including business and research. To learn more about data mining and knowledge discovery, check out the resources listed below, which will take you on a comprehensive tour of this many-faceted science. The use of data mining and knowledge discovery has a wide range of potential applications. Many AI (Artificial Intelligence), database, and statistical conferences have discussed data mining and knowledge discovery. In general, when we talk about knowledge discovery, we are talking about the process of discovering new, valid patterns. Large-scale data mining and knowledge discovery. Discovery may be split down into multiple processes, such as building a knowledge of the application domain, constructing a target data set, data cleaning and processing, discovering usable characteristics to describe the data, and then doing a search for patterns of interest in the data.

Sk.Abid Hussain (2014) [14] Data mining is a method used to discover patterns and relationships in data used to predict validity using various data analysis tools. A data description – the statistical feature The first and simplest phase of analysis in data mining characteristics; examine them using graphs and charts, and search for possible connections between variables. The collection, exploration and selection of the appropriate information is of crucial importance in the data mining process. Discovery of knowledge is the most desired and interesting outcome of IT and is the greatest example of smart computing. The job many academics and practitioners strive to achieve to find and extract information from data. Much buried knowledge awaits discovery - This is today's information wealth challenge. Discovery of know-how in databases through which genuine, new and valuable patterns from huge databases are identified.

DATA MINING

Data mining is an essential step in the knowledge discovery in databases (KDD) process that produces useful patterns or models from data (Figure 2). The terms of KDD and data mining are different. KDD refers to the overall process of discovering useful knowledge from data. Data mining refers to discover new patterns from a wealth of data in databases by focusing on the algorithms to extract useful knowledge.

Fig.1 Data Mining

Based on figure 1, KDD process consists of iterative sequence methods as follows

1. Selection: Selecting data relevant to the analysis task from the database

2. Preprocessing: Removing noise and inconsistent data; combining multiple data sources

3. Transformation: Transforming data into appropriate forms to perform data mining

4. Data mining: Choosing a data mining algorithm which is appropriate to pattern in the data; Extracting data patterns

5. Interpretation/Evaluation: Interpreting the patterns into knowledge by removing redundant or irrelevant patterns; Translating the useful patterns into terms that human understandable

DATA MINING TASKS

Define six main functions of data mining:

1. Classification is finding models that analyze and classify a data item into several predefined classes

2. Regression is mapping a data item to a real-valued prediction variable

3. Clustering is identifying a finite set of categories or clusters to describe the data

4. Dependency Modeling (Association Rule Learning) is finding a model which describes significant dependencies between variables

5. Deviation Detection (Anomaly Detection) is discovering the most significant changes in the data

6. Summarization is finding a compact description for a subset of data

Data mining has two primary objectives of prediction and description. Prediction involves using some variables in data sets in order to predict unknown values of other relevant variables Description involves finding human understandable patterns and trends in the data.

DATA MINING TECHNIQUES

There are several major data mining techniques have been developing and using in data mining projects. The art of data mining has been constantly evolving [6]. There are several innovative and intuitive techniques that have emerged that fine-tune data mining concepts in a bid to give companies more comprehensive insight into their own data with useful future trends. Many techniques are employed by the data mining experts, some of which are listed below:

1. Classification: This analysis is used to retrieve important and relevant information about data, and metadata. This data mining method helps to classify data in different classes.

2. Clustering: Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.

3. Regression: Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

4. Association Rules: This data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set.

5. Outer detection: This type of data mining technique refers to observation of data items in the dataset which do not match an expected pattern or expected behavior. This technique can be used in a variety of domains, such as intrusion, detection, fraud or fault detection, etc. Outer detection is also called Outlier Analysis or Outlier mining.

6. Sequential Patterns: This data mining technique helps to discover or identify similar patterns or trends in transaction data for certain period.

TREND IN DATA MINING TECHNIQUES

The data collected dates from 2010, until August 2018. The trend for the keywords; Data mining, Decision tree, Artificial neural network, Clustering, Association rule, Artificial intelligence, Bioinformatics, Customer relationship, Fuzzy logic, and their applications, are shown in Table 1.

Table 1 2010–2018 DMT keywords trends.

Keyword	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
Data mining	1	0	6	5	8	8	12	16	10	66
Decision tree	0	0	1	0	0	2	0	1	2	06
Artificial neural network	1	1	2	1	2	2	2	0	2	13
Clustering	0	0	1	0	0	3	0	1	0	05
Association rule	0	0	0	1	0	0	0	2	1	04
Artificial intelligence	0	0	0	0	0	0	1	0	1	02
Bioinformatics	0	0	0	0	0	0	0	3	0	03
Customer relationship management	0	0	0	1	0	0	0	0	0	01
Fuzzy logic	0	0	0	1	1	0	0	1	0	03
Total	2	1	10	9	11	15	15	24	16	103

Knowledge-based systems and their applications

Visualizing Data Mining Model

The fundamental goal of data visualization is to provide the user a general sense of the data mining model's construction. Most of the time, while doing data mining, we are getting information from repositories that has being stored in a concealed format. Users will have a tough time with this. As a result, we are able to deliver the highest levels of knowledge and trust thanks to this depiction of the data mining approach. A

considerably greater leap must be taken to turn a system's output into an actionable solution to a business issue since the user has no prior knowledge of what the data mining process has revealed. Predictive and Descriptive data mining models [1,2,6,45] are the two varieties. Predictive models use known values to create predictions about unknown data values. Example: classification, regression, time series analysis, and prediction, amongst other things. It investigates the data's attributes using a descriptive model to find patterns or correlations. For instance, clustering, summarization, association rules, and sequence discovery are examples of advanced analytics techniques.

Many data mining applications are targeted at predicting the future condition of the data in the present timeframe. An attribute's future state may be predicted by looking at its present and previous states. While this may seem like a simple strategy, it is really a sophisticated learning approach since the classes are already determined before the target data is ever examined. It is necessary to learn functions that map each data point to its corresponding real-valued prediction variable before doing regression analysis on the data. The value of a characteristic is analyzed as it changes over time in a time series analysis. Many statistical approaches for analyzing time-series data are utilized in time series analysis, such as auto regression algorithms and so on. Long-memory time series modeling and ARIMA (II) modeling both utilize it. As a result of clustering, data from multiple sources may be analyzed independently of one another. Unsupervised learning and segmentation are other names for it. Data is segmented or partitioned into groups or clusters. Domain specialists examine the data's behavior to determine the clusters. Data splitting into discrete groups of identical tuples is called segmentation, and it is a method utilized in very specialized contexts. The method of presenting the distilled essence of data is known as summarization. The association rule determines whether certain properties are linked together. It takes two steps to mine association rules: All frequent groups of items are found. The frequent item sets are used to generate strong association rules. Finding sequence patterns in data is referred to as sequence discovery. Using this sequence, we can figure out what is going on.

PERFORMANCE ANALYSIS

The KDD process's data mining phase employs a variety of iterative data mining approaches, including those used in the previous step. The intended usage of a framework characterizes learning awareness aims. Objectives may be classified into two categories: confirmation and disclosure. By providing confirmation, the system is just validating the customer's hypotheses. The framework now finds new projects on its own, thanks to the disclosure. Furthermore, we divide the aim of revelation into prognosis and depiction, in which the system identifies strategies for introduction to a customer in a structure acceptable to man. In this piece, we are primarily concerned with the dissemination of gleaned data.

Structure recognition, AI techniques, and estimations: Initiating the process of making things happen. Typically, techniques for a specific technique use a representation of the real essential model that is based on a grouping of a few clearly acquired other possibilities, such as polynomials, splines, part and reason capabilities, and limited Boolean capacities, etc. Thus, the computations will be compared in the adaptation integrity paradigm for evaluating the model's fit or in the research approach for discovering a good match in general. ”

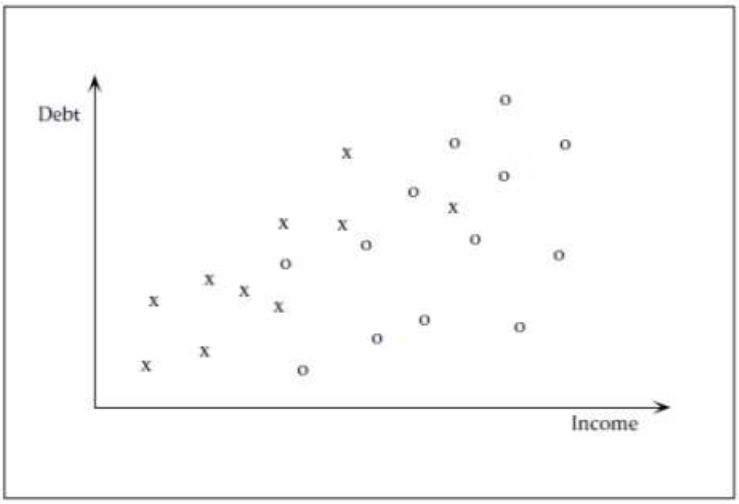

Figure 2 basic data set with two classes utilized

As you can see in the image, the dataset is made up of 23 fictitious examples. A person who has just received a loan from a certain bank will not be interested in any of the points on the outline. A person's remuneration is discussed in turn; the vertical community evaluates their whole commitment. The information has been divided into two categories: x refers to customers who have had credit issues, and refers to customers who are making excellent progress with the bank.

CONCLUSION

Data mining is the growth of an area with an extensive history. The innovation of word takes place in 1990s. The origin of Data mining is vestige back by the side of three unit lines. First is the artificial intelligence, second is the statistics and third is the machine learning. Artificial intelligence (AI) is based on heuristics; it tries to utilize human like thinking procedure to statistical jobs. Without specific effort, your mind is building clusters and associations. When you see a man and a woman walking close to each other,, you just know that they are either related or a couple. You see a woman coming out of a certain shop and you immediately associate her with the image the shop portrays. Data mining systems just make it easier for us to handle large amounts of data. Almost everything that is done in data mining can be done manually by a human but that would just take tremendously longer.

REFERENCE

1. Agrawal, R., and Psaila, G., “Active Data Mining. In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)”, American Association for Artificial Intelligence, Menlo Park, Vol.1, No. 6, Nov 1996.
2. U. Fayyad et.al “Knowledge Discovery and Data Mining: Towards a Unifying Framework” Published in KDD 1996
3. Mr. S. P. Deshpande and Dr. V. M. Thakare, ‘Data Mining System and Applications: A Review,’International Journal of Distributed and Parallelsystems (IJDPS) Vol.1, No.1, September 2010.
4. Larose, D.T., Discovering knowledge in data: an introduction to data mining, JohnWiley and Sons, 2005. Maimon O., and Rokach, L. Data Mining by Attribute Decomposition with semiconductors manufacturing case study, in Data Mining for Design and Manufacturing: Methods and Applications, D. Braha (ed.), Kluwer Academic Publishers, pp. 311–336, 2001
5. Apte, C., and Hong, S. J. 1996. Predicting Equity Returns from Securities Data with Minimal Rule Generation. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 514–560. Menlo Park, Calif.: AAAI Press.

6. Sk.Abid Hussain, Knowledge Discovery Process through Data Mining: A Technical Approach for Data Analysis, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 7, July 2014
7. Yihao Li,, DATA MINING: CONCEPTS, BACKGROUND AND METHODS OF INTEGRATING UNCERTAINTY IN DATA MINING, south central E-journal, 2010
8. L. A. Kurgan and P. Musilek, "A survey of Knowledge Discovery and Data Mining Process," The Knowledge Engineering Review, vol. 21, no. 1, pp. 1-24, 2006.
9. F. Weiping and W. Yuming, "The Development of Data Mining," International Journal of Business and Social Science, vol. 4, no. 16, pp. 157-162, 2013.