ANALYSIS ON MACHINE LEARNING ALGORITHMS AND ITS TRENDS, APPLICATIONS USING DATA MINING

G. Ravishankar,
Assistant Professor,
Hindusthan College of Arts & Science, Coimbatore, TamilNadu , India

Abstract

Data mining can be viewed as a result of the natural evolution of information technology. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. There is formerly a blast in the amount of data which is produced and have accessed. Data mining can unearth significant information and widespread utilizations in business, medical, telecommunication wireless and ubiquitous computing. In this paper a survey is done on machine learning techniques, issues and trends and applications used for data mining.

Keywords: Machine Learning, Supervised, Un-Supervised, Applications.

1. Introduction

Data is a very important asset of any organization. Company has to store all its transaction related data for its future use in business. The digital revolution provided relatively inexpensive data storage devices, which have helped the organization to store all related transaction data in the form of large information systems. Now a day because of internet usage the way transaction taking place within the organizations has completely changed. At a click of a button we can transfer the data from one part of the world into another part. Internet opened lot of opportunities for the organization to do business. Increased business opportunities create more number of possible transactions and volume of data growth. Databases today can range in size into the terabytes, more than 1, 000, 000, 000, 000 bytes of data. Within the masses of data lies hidden information of strategic importance. The quantity of data in the world roughly doubles every year[1] Tremendous data growth in the organizational databases gives the major difficulty to retrieve the hidden and useful information, which may be used for decision-making. We need a unique technique, which will work effectively to retrieve the hidden and useful decision-making information even in the midst of data growth in the databases. Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. To mine the hidden and useful information we have to take the available dataset through the process of data mining. It‟s not a

single step. It contains various groups of interlinking steps which will help us to find the useful information for decision making. Data mining searches databases to find hidden patterns and predict information to increase the business in the organization. Data mining Life Cycle We have to do the following steps to solve a data mining. Define the problem: To have the successful data mining application, the organization has to come up with a precise formulation of the problem they are trying to solve. A focused problem statement usually results in the best payoff. Data collection and selection: The organization has to use the right data for mining. data collection and selection step identifies the related data sources and acquires it. From the collected data source data selection process selects the subset of data to mine. Data preprocessing: Data cleaning It fills in the missing data and correcting the invalid data into a valid one. It finds the outliers data and removes the inconsistencies in the data source. Data integration: It combines data from different data sources into a single mining database. Data transformation: It converts the source data into a common format for processing. Data reduction: It is a process of discarding unwanted parameters from the data. So that the data volume will be less at the same time it will not suffer on the quality of the information. Data discretization: It is a part of data reduction process. It replaces the numerical attributes with the nominal attributes. Figure 1 represented into data mining process.

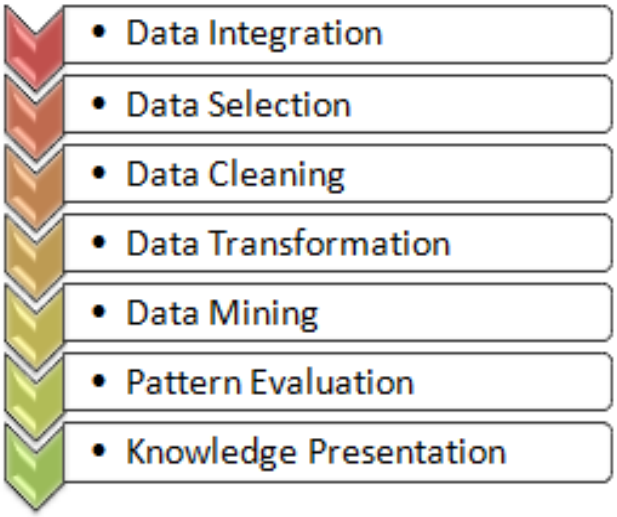

Figure 1: Data Mining Process

Steps in Knowledge Discovery (Or) Data Mining Process: 1. Data integration - Where multiple data sources may be combined): First of all the data are collected and integrated from all the different sources. 2. Data selection - Where data relevant to the analysis task are retrieved from the database. We may not all the data we have collected in the first step. So in this step we select only those data which we think useful for data mining. 3. Data cleaning - To remove noise and inconsistent data): The data we have collected are not clean and may contain errors, missing values, noisy or inconsistent data. So we need to apply different techniques to get rid of such anomalies. 4. Data transformation - Where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance. The data

even after cleaning are not ready for mining as we need to transform them into forms appropriate for mining. The techniques used to accomplish this are smoothing, aggregation, normalization etc. 5. Data mining - An essential process where intelligent methods are applied in order to extract data patterns. Data mining techniques are applied to discover the interesting patterns. 6. Pattern evaluation - To identify the truly interesting patterns representing knowledge based on some interestingness measure. This step involves visualization, transformation, removing redundant patterns etc from the patterns we generated. 7. Knowledge presentation - Where visualization and knowledge representation techniques are used to present the mined knowledge to the user. This step helps user to make use of the knowledge acquired to take better decisions.

2. Machine Learning

2.1 Machine Learning Algorithms Used in Data Mining

This is the algorithm part of the data mining process. It provides computers with the ability to learn without being explicitly programmed. This taxonomy or way of organizing machine learning algorithms is useful because it forces us to think about the the roles of the input data and the model preparation process and select one that is the most appropriate for our problem in order to get the best result.

2.1.1 Supervised Learning

Input data is called training data and has a known label or result. A model is prepared through a training process where it is required to make predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data.

2.1.2 Unsupervised Learning

Input data is not labeled and does not have a known result. A model is prepared by deducing structures present in the input data. This may be to extract general rules. It may through a mathematical process to systematically reduce redundancy, or it may be to organize data by similarity.

2.1.3 Semi-Supervised Learning:

Input data is a mixture of labelled and unlabelled examples. There is a desired prediction problem but the model must learn the structures to organize the data as well as make predictions.

3. Trends & Applications

A. Data mining in business

Business is a diverted field with several general areas of specialization such as accounting or financial analysis. Almost any neural network application would fit into one business area or financial

analysis. There is some potential for using neural networks for business purposes, including resource allocation and scheduling. There is also a strong potential for using neural networks for database mining that is, searching for patterns implicit within the explicitly stored information in databases. There is a marketing application which has been integrated with a neural network system. The Airline Marketing Tactician is a computer system made of various intelligent technologies including expert systems. A feed forward neural network is integrated with the AMT and was trained using back-propagation to assist the marketing control of airline seat allocations. The adaptive neural approach was amenable to rule expression. Additionally, the application's environment changed rapidly and constantly, which required a continuously adaptive solution. The system is used to monitor and recommend booking advice for each departure. Such information has a direct impact on the profitability of an airline and can provide a technological advantage for users of the system.

B. Telecommunication Industry

The telecommunication industry provides a lot of services in addition to telephone service i.e. fax, pager, cellular phone, Internet messenger, images, e-mail, computer and web data transmission etc. This has created a great demand for data mining in order to help understand the business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of service. The telecommunications industry generates and stores a tremendous amount of data. These data include call detail data, which describes the calls that traverse the telecommunication networks; network data, which describes the state of the hardware and software components in the network; and customer data, which describe the telecommunication customers.

C. Fight against Terrorism

After 9-11 attacks, many countries imposed new laws against fighting terrorism. These laws allow intelligence agencies to effectively fight against terrorist organizations. USA launched Total Information Awareness program with the goal of creating a huge database of that consolidate all the information on population. Similar projects were also launched in European countries and rest of the world. This program faced several problems, a. The heterogeneity of database, the target database had to deal with text, audio, image and multimedia data. b. Second problem was scalability of algorithms. The execution time increases as size of data (which is huge). For example, 230 cameras were placed in London, to read number plates of vehicles. An estimated 40,000 vehicles pass camera every hour, in this way the camera must recognize 10 vehicles per second, which poses heavy load.

D. Biological Data Analysis

Recently, the collection of biological data has seen exponential increase due to improvements in existing technologies and the introduction of new ones such as the microarrays. These technological advances have assisted the conduct of large scale experiments and research programs. The explosive growth in the amount of biological data demands the use of computers for the organization; the maintenance and the analysis of these recent data mining development tools assume importance in the analysis of data in order to gain new biological insights. The field of bioinformatics has many applications in the modern day world, including molecular medicine, industry, agriculture, stock farming, and comparative studies intense order calculations.

E. Data Mining in Health Care Management

Nowhere in the field of science is the need for tools to deal with uncertainty more critical than in medicine, as disease diagnosis involves several levels of imprecision and uncertainty. A single disease may manifest itself quite differently in different patients and with different disease status. Further, a single symptom may be indicative of different diseases, and the presence of several diseases in a single patient may disrupt the expected symptom pattern of any of them.

F. Ubiquitous Data Mining

Accessing and analyzing data from a ubiquitous computing device offer many challenges. For example, UDM introduces additional cost due to communication, computation, security, and other factors. So one of the Human-computer interaction is another challenging aspect of UDM. Visualizing patterns like classifiers, clusters, associations and others, in portable devices are usually difficult. The small display areas offer serious challenges to interactive data mining environments.

G. Time Series Data Mining

Another important area in data mining centers on the mining of time series and sequence-based data. Simply put, this involves the mining of a sequence of data, which can either be referenced by time (time-series, such as stock market and production process data), or is simply a sequence of data which is ordered in a sequence. In general, one aspect of mining time series data focuses on the goal of identifying movements or components which exist within the data (trend analysis).

4. Issues of Data Mining

One of the key issues raised by data mining technologies is not a business or technological one, but social one. Some of the issues are address below:

A. Security and social issues Today, Security [7] is an important issue with any data collection that is shared and/or is intended to be used for strategic decisionmaking. When data is collected for customer profiling, user behavior understanding, correlating personal data with other information, etc., large amounts of sensitive and private information about individuals or companies is gathered and stored. This becomes controversial given the confidential nature of some of this data and the potential illegal access to the information. Moreover, data mining could disclose new implicit knowledge about individuals or groups that could be against privacy policies, especially if there is potential dissemination of discovered information. Another issue that arises from this concern is the appropriate use of data mining. Due to the value of data, databases of all sorts of content are regularly sold, and because of the competitive advantage that can be attained from implicit knowledge discovered, some important information could be withheld, while other information could be widely distributed and used without control.

B. User interface issues The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user. Good data visualization eases the interpretation of data mining results, as well as helps users better understand their needs. Many data exploratory analysis tasks are significantly facilitated by the ability to see data in an appropriate visual presentation. There are many visualization ideas and proposals for effective data graphical presentation. However, there is still much research to accomplish in order to obtain good visualization tools for large datasets that could be used to display and manipulate mined knowledge. The major issues related to user interfaces and visualization are "screen real-estate", information rendering, and interaction. Interactivity with the data and data mining results is crucial since it provides means for the user to focus and refine the mining tasks, as well as to picture the discovered knowledge from different angles and at different conceptual levels.

C. Mining methodology issues These issues pertain to the data mining approaches applied and their limitations. Topics such as versatility of the mining approaches, the diversity of data available, the dimensionality of the domain, the broad analysis needs (when known), the assessment of the knowledge discovered, the exploitation of background knowledge and metadata, the control and handling of noise in data, etc. are all examples that can dictate mining methodology choices. For instance, it is often desirable to have different data mining methods available since different approaches may perform differently depending upon the data at hand. Moreover, different approaches may suit and solve user's needs differently

D. Performance issues Many artificial intelligence and statistical methods exist for data analysis and interpretation. However, these methods were often not designed for the very large data sets data mining is dealing with today. Terabyte sizes are common. This raises the issues of scalability and efficiency of the data mining methods when processing considerably large data. Algorithms with exponential and even medium-order polynomial complexity cannot be of practical use for data mining. Linear algorithms are usually the norm. In same theme, sampling can be used for mining instead of the whole dataset. However,

concerns such as completeness and choice of samples may arise. Other topics in the issue of performance are incremental updating, and parallel programming. There is no doubt that parallelism can help solve the size problem if the dataset can be subdivided and the results can be merged later. Incremental updating is important for merging results from parallel mining, or updating data mining results when new data becomes available without having to re-analyze the complete dataset.

Conclusion

The objective of this survey work is to provide a study of different machine learning techniques that can be employed in automated Data Mining process. In this paper analyzed and some techniques, methods for enhance the machine learning process. Each one method or technique have some performance ratio not only the advantages and also have some drawbacks within that. In future work will choose any one algorithm which is most secure and suitable to do better accuracy for data mining process and then apply some enhancement within that to proof much better than the old performance.

References:

[1] Ms. Ishtake S.H , Prof. Sanap S.A. “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, International J. of Healthcare & Biomedical Research, Volume: 1, Issue: 3, April 2013.

[2] Chaitrali S. Dangare Sulabha, “ Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques”, International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012.

[3] Jyoti Rohilla, Preethi, “ Heart Disease Prediction Using Data Mining Techniques”, International Journal of Computer Science and Mobile Computing, A Monthly Journal of Computer Science and Information Technology,ISSN 2320–088X , IJCSMC, Vol. 4, Issue. 7, July 2015.

[4] Abhishek Taneja, “Heart Disease Prediction System Using Data Mining Techniques”, Oriental Journal Of Computer Science & Technology, ISSN: 0974-6471 December 2013, Vol. 6, No. (4).

[5] Nidhi Bhatla, Kiran Jyoti , “An Analysis of Heart Disease Prediction using Different Data Mining Techniques”, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8, October - 2012 ISSN: 2278-0181.

[6] Jyoti Soni, Ujma Ansari, Dipesh Sharma, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction”, International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.

[7] Shadab Adam, Pattekari and Asma Parveen, “ Prediction System For Heart Disease Using Naive Bayes”, International Journal of Advanced Computer and Mathematical Sciences, ISSN 2230-9624. Vol 3, Issue 3, 2012.

[8] D. Ramesh ,B Vishnu Vardhan, O Subhash Chander Goud," Density Based Clustering Technique on Crop Yield Prediction"IJEEE,2014

[9] Mohammad Motiur Rahman,Naheena Haq, Rashedur M Rahman, "Application of Data Mining Tools for Rice Yield Prediction on Clustered Regions of Bangladesh", IEEE, 2014

[10] S.Veenadhari, Dr. Bharat Misra, Dr. CD Singh," Machine learning approach for forecasting crop yield based on climatic parameters", ICCCI -2014

[11] D Ramesh , B Vishnu Vardhan," Data Mining Techniques and Applications to Agricultural Yield Data" , IJARCCE, 2013

[12] José R. Romero , Pablo F. Roncallo , Pavan C. Akkiraju , Ignacio Ponzoni , Viviana C. Echenique, Jessica A. Carballido,"Using classification algorithms for predicting durum wheat yield in the province of Buenos Aires", ELSVIER, 2013.

[13] Yunous Vagh,Jitian Xiao,"A data mining perspective of dual effect of Rainfall and Temperature on Wheat Yield", ECU, 2012.