THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

Security Aspects of Data Mining With Its Applications

Aman Bhaskar[1], Kavita[2], Sahil Verma[3], Atul Malhotra[4]

[1]Research Scholar,LovelyProfessional University
amanbhaskar113114@gmail.com
[2,3,4]Department of Computer Science and Engineering,
LovelyProfessionalUniversity, Phagwara, India
[2]kavita.21914@lpu.co.in, [3]sahil.21915@lpu.co.in, [4]atul.18011@lpu.co.in

*Corresponding author: kavita.21914@lpu.co.in

Abstract

Human civilizations survive on two most essential things i.e. Knowledge and Energy. In search of energy, humans gain knowledge from experience learning, gathering information from others and then analyzing outcomes. Information is an important asset for working on various real-life applications, such as credit scoring for banks, website optimization, to improve user interface, its experience and for card fraud detection, etc. With the rise in data mining technologies, there is serious risk for security of individual venerable information. The privacy risk which gets introduced during operations such as data collection, publishing and information delivering could be reduced using privacy-preserving data mining ayielding area of research. Here, the paper discusses data mining application and security issues with data mining algorithms.

Keywords: data mining, data collection, privacy-preserving data mining.

1. Introduction

As the popularity of "big data" concepts prevails in recent years it results in data mining to attract more and more attention. Big data can be defined as its four properties i.e., large volume, variety, velocity and veracity which is usually called four V's model. Data mining needs clean and static data for its operation, which must be applied to the data before applying the various machine learning algorithms of classification and prediction. Machine learning as a computational part of data mining and analytics serves as a basic tool for retrieving information from raw facts, recognizing patterns in correlated data and based on some hypothesis, predict some outcome. The aim of data mining referred asthe knowledge discovery process is to find previously unknown patterns. Hence it supports decisions making for the development of businesses. Machine learning algorithms are classified into four types referred to as supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.

The steps involved in data mining are as follows:-

- Data exploration: It includes data cleaning which is referred to as a process of eliminating noise, incompleteness, and inconsistency in the data source. The next step is to integrate data

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

from several sources. And finally, we transform the data uniformly this is due to the presence of heterogeneity in data. Here we look at the summary data, apply sampling and intuitions.

- Pattern recognition: Data mining is a process that helps to find meaningful patterns in a huge amount of data depository. It involves methods to find useful features to represent the data. It identifies the truly emphasizing patters using those features which represent information and understandably show them.

- Deployment: Once we discovered the patterns we analyze and use the above knowledge for finding the desired outcome. This is through predicting the new results from later results. Another way to exploit the outcome is package, price and advertise the product in case of business applications.

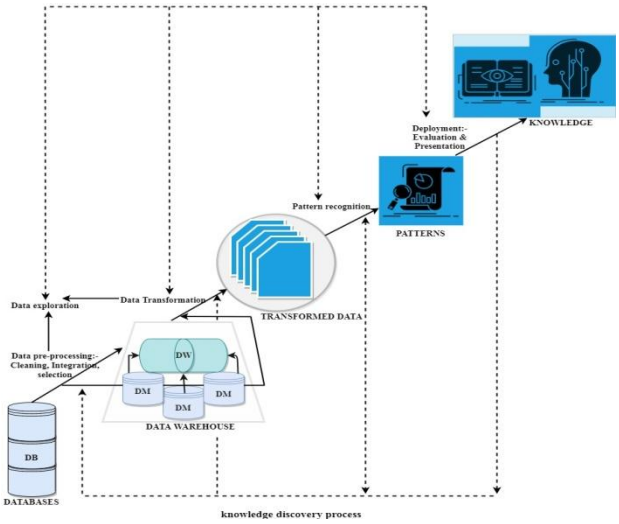

Figure 1. Showing data mining as a knowledge discovery process.

1.1 History

Data mining starts its journey in the past 30 years. At the beginning of the 1960s, data mining is referred to as statistical analysis. At the end of the 1980s, the trend goes more logistic with coming up with new techniques such as fuzzy logic, heuristics, and neural networks. With the machine learning evolution during the past several years, it has proved its role in building learning-based models from processed data, based on which quality information is extracted and used for yielding new patterns, predictions of new data sets results in quicker and accurate decision making.[1]

1.2 Statistics versus data mining

Statistics have the same usual way as that of data mining with the same results. Regression is used in statistics quite often to create predictive models from large stores of historical data. So what is the main difference between both these terms? The main difference

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

between them is that data is meant to be built for the corporate end-users, while statistics is more used for mathematical field. We can automate the statistical methods so that the end-user of the corporative area can be free from some of the burden. Data mining do similar thing and produces specific tools that are easier to use. These tools are no longer only the main concern of data analyst but also the business user. The notion of giving potentially high predictive technologies to these users who may or may not have taken a statistical course is a threat to the system's reliability. So keeping the business users reliable and allowing them to access the predictive and visualized data analysis is the job of data mining technology. [1]

1.3Discovery versus prediction

Miners are of two types which use to fetch the information from the large databases. As we can imagine data sets as the mountains of data and we have to mine them for gold, diamonds which represent the quality data or information. So the miner who focused mainly on the gold and diamonds, not the other rubies which appear similar to them, are called farmers. And on the other hand, those which are not aware of what they need and just firing queries to the data sets randomly are called explorers. Every corporation has these types of DW users.

- Farmers: Know exactly what they need before they set out to find it. They execute small queries and retrieve small and precise piece of information.

- Explorers: They are not expectable. They usually execute large queries and sometimes as a result find nothing beneficial, except those times when they find rare and very valuable information.

Discovery: -

Finding something that we aren't looking for.There may be times when explorers are found in mining. They can be easily distinguished because they are asking the impossible:" Tell me something I didn't know but would like to know". How can we describe this as a nugget of gold when even the explorer doesn't know exactly what they want? So, discovery is sometimes are a result of such explorers when they find sometimes unexpected and accidental but valuable. [1]

Prediction: -

So the question comes in mind, how does a farmer know what they likely to be needed for example the information pictured as diamonds or veil of gold? The answer to this might be is that they know the important properties of diamond or gold because they have been discovered before. So farmers usually based on those properties operate their queries efficiently and with accuracy. [1]

1.4Data mining in the Business Process

To choose data mining technology depends on whether it would deliver real value to the business. As it must convert bottom-line profit, increased revenue, decreased cost, or return on

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

investment. Data mining needs to be more than knowledge discovery if it is to be successfully diploid in any business. Usually, some response (reaction) takes against customer or competitors in the marketplace. These customer reactions fairly define targets that relate to the business process. The target must be followed by the following attributes to be successful with data mining:

- Value: It must have some relationship to baseline business value. Predicting stopping loss through attrition or fraud, customer weight from buying habits are suitable example showing these attributes (attrition, customer buying habits) might have value, or not.

- Actionable: It states certain actions can be taken to influence target. For example "retirement age is the predictor of employee attrition". We find there is little could be done about how old employees are? On the other hand, improving health care coverage could be effective.

- Capturing actions effect: It states that if the action performed by good predictive models couldn't be used to measure effects on a customer, then there is no way to tell the exact impact of the action. Also, there is no way to have that information to be fed back into the system so that the next model could be further improved.[1]

1.5Comparing data mining algorithms

To make a comprehensive selection among data mining tools and technologies, it will be helpful to categorize areas where they are differing. One of the best ways to see the valid distinction between the algorithms is to see similarity i.e. each data miming algorithm have the following:

- Model structure: It is the structure that defines the model. For example in the case of statistical regression mathematical equations are used and in case of a decision, trees model might just be SQL queries.

- Search: It states that with an exponential increase in storage size over some time, there is in the same manner algorithms are amends and modifies model. For example, genetic algorithms search through random permutation and genetic recombination. Similarly, the NN search through link-weight via backpropagation algorithm is used.

- Validation: It states that a valid model determines by an algorithm. For example, the decision tree uses cross-validation to determine the optimal level of growth of the tree, whereas NN doesn't have any validation technique to determine termination.[1]

Table 1: Comparison table of Data Mining Algorithms [1]

Algorithm	Structure	Search	Validation
CART	Binary Tree	Split chosen by entropy or Gini metric	Cross-validation
CHAID	Multiway split tree	Split chosen by Chi-	Validation performed

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

		square test and Bonferroni adjustment	at time of split selection
Neural network	Forward propagation network with non-linear thresholding	Backpropagation of errors	Not applicable
Genetic algorithm	Not applicable	Survival of fittest on mutual and genetic crossover	Usually crossover
Role induction	"If then rule"	Add new constraint to rule retain it if it matches the "interestingness criterion based on accuracy and coverage	Chi-square test with statistical significance cutoff
Nearest neighbor	Distance of prototype in n-dimension feature space	Usually, there is no search	Cross-validation used for reported accuracy rates but not for algorithm termination

2. Literature Review

2.1 Applications of data mining

C. Dorn et. al.[2] tell us about the concepts enabled in various internet-based collaboration and virtual community of applying several supervising, mining and statistical tools to study team formation activity and interaction between the individual entities. The paper addresses in detail the team building issue that is independent of direct interactions to express sufficient team connectivity. The paper has two discoveries based on genetic algorithms and another on simulated annealing. It helps to get optimum team structuring resulting in balancing both the acceptable skill coverage and required team connectivity. At last the self-adjusting mechanism they proposed focuses to find the best combinations of direct interactions and recommendations to establish connectivity.

Y. Zhu and Z. Feng [5] states that the increase in location monitoring methods there is a vast hike in the production of trajectory data, as it is required for tracking the moving object location traces. A timestamp of geographical location can be seen as such a trace and in a pattern represent a trajectory. Some of the trajectory data mining applications are path discovery, location prediction, movement behavior analysis, and so on. However, people may also suffer if these data collected and used inappropriately. On the other hand trajectory, data mining could reduce the cost and increase the scalability of supervision and monitoring activities. Concerning about the trajectory data mining, the paper review extensively several existing studies.

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

S. Pan et. al.[6] proposed atechnique to utilize the consumer-related data and mining electricity consumption data to make perfect e-grocery home delivery. The paper shows concern about e-grocery loss of logistic efficiency due to the customer's attrition, high rate of failed delivery. They estimate the attrition probability of consumer to optimize the success rate of delivery and improve transportation. The paper achieves its top targets as it gains a 3-20% reduction in traveling and theoretically raises the success rate of 1st round delivery by 18-26% approx. The social impact of the paper is to reduce the impact on the surroundings of bulk shipping and it improves once experience on e-grocery shopping. At last proposed approach has limitations such as data privacy and security, time slot selection scenario is not demonstrated and finally computational capacity.

Anita Bai et. al.[7] introduces an alternate technique to frequent dataset mining known as High-Utility itemset mining (HUIM). In this paper, they mainly proposed three things i.e. HUI mining based algorithm based on the selective database projection which reduces the scanning time of the database making it more efficient. The second thing is an efficient data format referred to as HUI-RTPL which is best to represent data-consuming low memory. The third thing is the data structure they have proposed i.e. Trail-Count list and selective database projection utility list to minimize the search space for the HUI mining algorithm. The paper shows an algorithm generates less dimensional data projection and instances which validate the upper bound over memory consumption.

J.Shao et. al.[8] provides a technique to prove rules containing range association from numerical features. Intelligible classification and data summary models are derived from these above-mentioned rules. The paper is based on producing rules similar to association rule mining but differ in search conducted through rule outcomes. This approach gives convincing rules not limits to fixed rules, which are results of association rule mining to create models. The research experiment shows promising results and more efficient than traditional rule mining techniques like C4.5 and RIPPLE in both summarizing and classifying data. At last, they show interest in studying categorical attributes as a part of their findings for splitting ranges into sub-ranges using range based rule-formation.

Wil van derAalst[9]introduces the process mining concept providing a collection of methods to convert the raw data into valuable information, updated notions, predictions, and suggestions. The essential tool of process mining plays an important role for both data scientists and business analysts. The author wants to visualize process mining the same as spreadsheets which works the same with events as later does with data. The paper discusses the superiority of process mining over spreadsheet techniques by mentioning features such as discovery processes, analyze bottleneck, operational process support. Also paper expresses the implementation of process mining is not limited to the application field but can be used in a large domain of industries. The paper illustrates 20 commercially available process mining tools also uses a strong dataset.

X. Chen et. al. [10] used both large-scale data mining and qualitative analysis methods to implement a multi-label classification algorithm to classify student's problems. The paper concentrate on engineering student's twitter posts to know queries and problems in their educational experiences. For quantitative analysis, they collected 25000 tweets based on engineering student's college life and used 35000 tweets to train the classifier which detects

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

student problems. Results show how informal data from social media can be used to depict student's experiences. The paper resolves the major drawbacks of manual quantitative analysis and large-scale computational analysis of human textual content.

YanliBai[11] states the reason for choosing data mining as its superior performance, wide application, and ease of usage. The paper is about the wireless sensor networks cleaning using the methodology of talent management data cleaning based on mining technology. Specific forms of the wireless sensor networks are analyzed and the clustering model is used for data cleaning. The author proposed a cluster-based replication record deletion algorithm to verify the accuracy of cleansing methods. The outcomes of the research technique depict correct and effective results. At last imperfection in data mining technology open up in-depth study in the follow-up work.

2.2 Security in data mining

L. Xu et. al.[3] provides a wider view of the privacy issue related to data mining and also search various other techniques that can help to protect sensitive information. The paper discusses privacy concerns for four different types of users involved in data mining, particularly identified for this purpose are data provider, data collector, data miner, and decision-maker. The paper also reviews the game-theoretical approaches, proposed to analyze the interaction among users in a mining scenario. The paper provides an intuitive study of PPDM (Privacy-Preserving Data Mining) by classifying the responsibilities of individual users having own valuation of sensitive information concerning their privacy.

The security Privacy-Preserving objective and measures for the above-mentioned users are as follows: -

- Data Provider: - His target is to strictlymonitor the quantity of vulnerable data exposed publicly. And another main aim is to use security mechanisms to authorize data accessibility, use the auction method to retrieve from loss of privacy or hide his identity by falsifying his data.

- Data Collector: - His target is to exposeessential data without compromising provider’s identities and vulnerable information. And another main aim is to produce good security policies to maintain a record of possible loss after different attacks.

- Data Miner: - His target is to gather data mining outcomes while storingvulnerable information confidential. And another main aim is to update data and then send modified data for applying data mining algorithms, or use security procedures to protect sensitive information inside the classifier.

- Decision Maker: - His target is to decide which correct judgmentregarding data mining outcomes is reliable. And another main aim is to utilize the original technique of collectingoutcomes or constructlearning model to classify information.

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

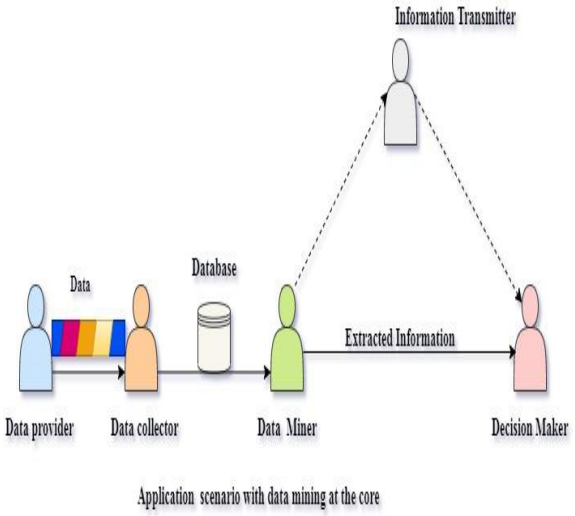

Figure 2. Showing data mining stakeholders**.** [3]

A.Monreale et. al.[4] provides the necessary concern about privacy in our society, particularly when an individual wants to utilize, advertise analyze securely. Moreover information and big data emphasizing inspire to detect, collect and study social data providing private and sensitive data of any person in great depth. Hence we couldn't implement privacy preservation by simply authenticity i.e. d-identification only. In the paper, they proposed the PRIVACY-BY-DESIGN procedure to produce methodology to handle risks having unlikely, unlawful outcomes of privacy breach. Also, the data mining of social data mining and big data analytics has not been compromised in incorporating the essential privacy requirements from the beginning.

3. Conclusion

Data mining is beneficial for both social development and economic development. With the advancement in computational power of the computer system and introduction of artificial intelligence, it has played its role in the background but still, there is a lot scope of improvement in the data mining technology. As the paper discussed earlier there is privacy concern i.e. sensitive information integrity must be maintained in the discovering knowledge from the data mining process. Moreover, there is a vast scope of data mining in the field of business, finance, daily life, research and development, education, security, and forecasting, etc. At last, the paper mainly concentrates on the mining aspect and terms related to it, hence there is a lack of exposure to the machine learning algorithms and their use in AI. The main reason for this is to limit the content to the topic only. The paper gives a review of ten research papers on the related topic, having applications and security issues of mining algorithms.

References:

[1] Alex Berson Stephen J.Smith: Data Warehousing Data Mining & OLAP, Vol 7, Tata McGraw-Hill, 2010.

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

[2] Christoph Dorn, Florian Skopik, Daniel Schall, SchahramDustdar: Interaction mining and skill-dependent recommendations forMulti-objective team composition, Data & Knowledge Engineering, Vol. 70, July 2011.
[3] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, And Yong Ren: Information security in Big Data: Privacy and Data Mining, IEEE Access, Vol.2, October 2014.
[4] Anna Monreale, Salvatore Rinzivillo, Francesca Pratesi, Fosca Giannotti and Dino Pedreschi: Privacy-by-design in big data analytics and social mining, Vol.10, EPJ Data Science,September 2014.
[5] Zhenni Feng and Yanmin Zhu: A Survey on Trajectory Data Mining: Techniques and Application, Vol. 4,IEEE Access, April 2016.
[6] Shenle Pan, VaggelisGiannikas, Yufei Han, Etta Grover-Silva, and Bin Qiao: Using customer-related data to enhance e-grocery home delivery, Vol. 117, Industrial Management and Data Systems, October 2017.
[7] Anita Bai, Parag S. Deshpande, And Meera Dhabu: Selective Database Projections Based Approach for Mining High-Utility Itemsets, Vol. 6,IEEEAccess,January 2018.
[8] Jianhua Shao, Achilleas Tziatzios: Mining range associations for classification and characterization, Vol. 118, Data & Knowledge Engineering, October 2018.
[9] Wil van derAalst: Spreadsheets for business process management, Vol. 105,Business Process Management Journal, February 2018.
[10] Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan: Mining Social Media Data for Understanding Students' Learning Experiences, Vol. 6, IEEE Transactions On Learning Technologies, September 2019.
[11] Yanli Bai: Data cleansing method of talent management data in wireless sensor networks based on data mining technology, Vol. 33,SpringerNature,Eurasip Journal on Wireless Communications and Networking, February 2019.