HCTL Open Int. J. of Technology Innovations and Research
HCTL Open IJTIR, Volume 8, March 2014
e-ISSN: 2321-1814
ISBN (Print): 978-1-62951-499-4

A Survey of Security Techniques and Algorithms for Data Mining

***Sudeep Panchpuri*[*], *Vivek Uniyal*[†], *Govind Kamboj*[‡]**
sudeepbeit@gmail.com

Abstract

Data mining is the process of analysing data from different prospects and summarizing it into useful information. Data mining also referred to as KDD(Knowledge Discovery in Databases).The complete goal in the data mining process is to extract information from your data set and change it into an understandable structure for further use. The knowledge that can be used to enhance revenue, cuts costs, or both. Data mining software packages are an analytical tools for analysing data. It allows users to handle data from a number

[*] M.Tech Student, Department of Computer Science & Engineering, Graphic Era University, Dehradun, India

[†] M.Tech Student, Department of Computer Science & Engineering, Graphic Era University, Dehradun, India. Email: vivek.akshay@gmail.com

[‡] Assistant Professor, Department of Computer Science & Engineering, Graphic Era University, Dehradun, India. Email: govind.kamboj@gmail.com

of dimensions or angles, categorize it, and summarize the relationships identified. Security is a wide issue in data mining. Businesses generally own info on their employees and customers and demand a mechanism to guard similarly info from theft. In this paper, we have presented a survey of security methods and techniques for data mining.

Keywords

Data Mining, Security Issues, Security Algorithms.

Introduction

Data mining is the process of analysing data from different prospects and summarizing it into useful information. Data mining also called KDD (Knowledge Discovery in Databases).The goal in the data mining process is usually to extract information coming from a data set and change it into an understandable structure for additional use. The information that can be used to boost revenue, cuts, costs or both. Data mining software is an analytical tools for analysing data. it allows users to analyse data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is a tool, not really a magic. It won't sit inside your database watching what the results are and post you e-mail to have your attention when it sees an interesting pattern. It doesn't take away the have to know your online business, to recognize important computer data, as well as to understand analytical methods. Data mining assists business analysts with finding patterns and relationships within the data - it doesn't tell you the additional value on the patterns towards organization. Furthermore, the patterns uncovered by data mining must be verified in real life. Data mining is primarily used today by companies that have a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these lenders to find out relationships among "internal" factors for instance price, product positioning, or staff skills, and "external" factors like economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

1. Data Mining Applications,

2. Credit Assessment,
3. Stock Market Prediction,
4. Fault Diagnosis in Production Systems,
5. Medical Discovery etc.

Advantages of Data Mining

Marketing/Retail

Data mining helps marketing companies build models according to historical data to calculate which will interact with the new marketing campaigns including direct mail, online marketing campaign etc. Through the results, marketers may have appropriate approach to sell profitable products to targeted customers. Data mining brings plenty of benefits to retail companies in the same manner as marketing. Through market basket analysis, an outlet may have the ideal production arrangement in a manner that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain reduced prices for particular goods that will have more customers.

Finance/Banking

Data mining gives financial institutions specifics of loan information and credit reporting. Because they build a model from historical customer's data, the bank and financial institution can determine good and bad loans. Furthermore, data mining helps banks detect fraudulent plastic card transactions to defend charge card's owner.

Manufacturing

By using data mining in operational engineering data, manufacturers can detect faulty equipments and determine optimal control parameters. E.g. semiconductor manufacturers includes a challenge that including the conditions of manufacturing environments at different wafer production plants are similar, the grade of wafer are lot the identical and a few for unknown reasons even has defects. Data mining has been signing up to determine the ranges of control parameters contributing on the manufacture of golden wafer. Then those optimal control parameters are utilized to manufacture wafers with desired quality.

Governments

Data mining helps agency by digging and analysing records of financial transaction to construct patterns which could detect money laundering or criminal activities.

Disadvantages of Data Mining

Privacy Issues

The concerns in regards to the personal privacy are actually increasing enormously recently specially when internet is booming with social support systems, e-commerce, forums, blogs. On account of privacy issues, folks are afraid of their private information is collected and found in unethical manner in which potentially causing them a lot of troubles. Businesses collect details about their customers in many ways for understanding their purchasing behaviours trends. However businesses don't last forever, some days they might be acquired by other or gone. Currently the individual information they own probably is sold with other or leak.

Security Issues

Security is a big issue. Businesses own details about their employees and customers including SSN, birthday, payroll and etc. However how properly this post is taken care remains to be in questions. There are a lot of cases that hackers accessed and stole big data of consumers from big corporation for instance Ford Motor Credit Company, Sony with the much personal and financial information available, the loan card stolen and identity theft turn into serious issue.

Misuse of Information/Inaccurate Information

Details are collected through data mining created for the ethical purposes can be misused. These records may be exploited by unethical people or businesses for taking important things about vulnerable people or discriminate against a group of people. Moreover, data mining way is not perfectly accurate. If any mistakes are used for decision-making, it will eventually cause serious consequence.

Security in Data Mining

Data mining is the discovery of new and charismatic patterns in large datasets, is an exploding field. Lately there's been a preliminary understanding that data mining comes with a encounter on security (including a workshop on Data Mining for Security Applications.) Looking after would be the utilization of data mining to enhance security, e.g., for intrusion detection. A second aspect may be the potential security hazards posed when an adversary has data mining capabilities. Data mining is the process of model a number of relevant queries to extract information from large packages of information from the database. Data mining techniques might be applied to handle problems in database security. Alternatively, data mining techniques may also be helpful to cause security problems. The operation of securing Server Analysis Services occurs at multiple levels. You need to secure each instance of Analysis Services and its data sources to make certain only authorized users have read or read/write permissions to selected dimensions, mining models, and data sources. You will need to also secure underlying data sources to counteract unauthorized users from maliciously compromising sensitive business information. For data mining, you need a different group of permissions to develop and process models than it is advisable to view or query the models. Making predictions against a model is a form of query and does not require administrative permissions. Data mining techniques include those based on rough sets, inductive logic programming, machine learning, and neural networks, among others. Essentially one gets to some hypothesis, which can be the info extracted, from examples and patterns observed. These patterns are observed from posing a number of queries; each query may depend upon the response obtained towards the previous queries posed. Data mining techniques have applications in intrusion detection and auditing databases. In the case of auditing, the data for being mined is the large quantity of audit data. It's possible to apply data mining tools to detect abnormal patterns. For instance, suppose a staff makes an excessive volume of trips to particular country this also truth is known by posing some queries. Another query to pose is whether the worker has associations with certain people from that country. If the answer is positive, then this employee's behaviour is flagged.

Literature Survey

T. Y. Lin et al. (1996) have explained about data mining and security in their research paper **Security and Data Mining** [1]. They have discussed well developed new security concern and research problems and finally they have developed theory and rough set theory. Also they have discussed some potential applications to security problems. In this research paper did not have any security concern in fact they had just provide the opposite objective to content based meta data. Data mining can be used effectively to impose security.

Yehuda Lindell and Benny Pinkas (2008) have discussed about the basic paradigms and notation of security multiparty computation in their research paper **Secure Multi-party Computation for Privacy-Preserving Data Mining** [2]. And they also discussed their relevance to the field finally they explained the relationship between secure multi-party computation and privacy of data mining. This formal basis is desperate when designing cryptographic protocols for any task, and in particular for privacy-preserving.

Korosh Golnabi et al (2006) have explained some firewall rules using data mining techniques in their paper **Analysis of Firewall Policy Rules Using Data Mining Techniques** [3] they define the firewall problem and how much firewall rules are useful well-organized, up-to-dated or efficient to reflect network traffic's current characteristics. And also they present set of techniques and algorithms to manage the firewall policy rules. At the end of they developed a prototype and demonstrated usefulness of this approached. At the end they analysis two new types of the anomalies.

Charu C. Aggarwal and Philip S. Yu have proposed models and algorithms for privacy-preserving Data Mining in their book **Privacy-Preserving Data Mining: Models and Algorithms** [4]. In this book they explained about the field of privacy which has seen increasing rapidly in recent years because of the increases in strength to store data. They have presented some of the broad areas of privacy-preserving data mining and the basic algorithms. They explained a variety of data modification techniques like randomization and k-anonymity based techniques.

Chris Clifton et al have introduced numerous applications for privacy preserving distributed Data Mining in their research paper **Tools for Privacy Preserving Distributed Data Mining** [5]. They provide a solution for privacy, desired results , data distribution , constraints on collaboration and cooperative computing etc. is toolkit of components which can combined for application of privacy-preserving data mining and showed how they can be used to figure out certain privacy-preserving data mining problems.

Godswill Chukwugozie Nsofor (2006) described five different predictive data mining techniques and compared those techniques in their thesis **A Comparative Analysis of Predictive Data-mining Techniques** [6]. They compares 5 different predictive techniques of data mining on 4 uniquely different data sets: Boston housing, collinear, airline and simulated data sets. At the last they analysis the various data preprocessing techniques that were used to process the four data sets and some of the data sets were seen to unique qualities.

Qiang Yang and Xindong Wu (2005) identified some challenging issues in their research paper **10 Challenging Problems in Data Mining Research** [7]. They summarize 10 most important problems in data mining research. These problems are sampled from a small, albeit important, segment of the community. Finally they summarize the 10 problems.

Charu C. Aggarwal and Philip S. Yu discussed about the privacy-preserving data mining methods in their research paper **A General Survey of Privacy-Preserving Data Mining Models and Algorithms** [8]. In this paper they implemented a review of the state-of-the-art methods for privacy. They discussed methods for randomization, k-anonymization, and distributed privacy-preserving data mining. They also discussed some cases in which the out put of data mining applications needs to be disinfect for privacy preservation purposes and they also explained the computational and theoretical limits combined with privacy preservation over high dimensional data sets.

Benny Pinkas describe research in secure distributed computation, which was done as a part of a bigger body of research in the theory of cryptography, has get remarkable results in their research paper **Cryptographic Techniques for Privacy-preserving Data Mining** [9]. In this paper they showed how non trusting parties can jointly compute function of their different inputs while they ensure that no party can learn anything. These results showed that can be applied to any function and these results also describe their efficiency, and validate their importance to privacy preserving computation of data mining methods. They describe the brief definition of security and the generic developments for the two - party and multi-party scheme.

Wenke Lee and Salvatore J. Stolfo discussed about developing general and systematic rule for intrusion detection in their research paper **Data Mining Approaches for Intrusion Detection** [10]. The main concept of this research were to use of data mining methods to observe consistent and useful patterns of system features that explain program and user behaviour, and use the set of applicable system features to enumerate classifiers that can verify anomalies and known intrusion. They also provide a review on two general algorithms. In this paper they proposed a systematic framework that apply data mining methods for intrusion detection.

Yehuda Lindell and Benny Pinkas (2000) introduced the concept of privacy preserving data mining in their paper **Privacy Preserving Data Mining** [11]. They proposed a model where two parties owning private databases wish to run a data mining algorithm on the union of their databases, without notify any useless information. They present a solution that is more powerful than generic algorithm. They demonstrate this on ID3, an algorithm extensively used and implemented in many applications.

Sheng Zhong (2004) studied about the privacy, integrity and incentive compatibility in their thesis **Privacy, Integrity and Incentive-Compatibility in Computations with Un-trusted Parties** [12]. In their thesis they explained about the privacy compatibility in computation with un-trusted parties. They also study about the secure multi-party computation. They presented a mix network custom fit for election system, with a significant speed-up over previous work. They designed and analysed an efficient algorithms for distributed mining and also present to protect the data integrity in storage services. They also introduced a significant VDOT, a new cryptographic primitive. Finally they recommended a way to add incentive discussions to the studies of secure multi-party computation.

Rakesh Agrawal and Ramakrishnan Srikant discussed about the development of techniques that consolidate privacy concerns in their research paper **Privacy-Preserving Data Mining** [13]. They addressed the following question, can they developed accurate model without access to actual information in individual data records? They considered the detailed case of building a decision-tree classifier from training data. The resulting data records seem very distinct from the original records and the distribution of data values are also differ from the original.

Murat Kantarcoglu and Chris Clifton (2003) described about the data mining and their privacy concern in their research paper **Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data** [14]. In this paper they explained privacy concern may restrict the parties from directly sharing the data and also restrict the some types of information. This paper also addresses the secure mining of association rules. The methods consolidate cryptographic techniques to decrease the information shared, although adding little overhead to the mining task.

Security Algorithms for Data Mining

The Randomization Method

It's a way of privacy-preserving data mining through which noise is added to the information so as to mask the attribute values of records. The noise added is sufficiently large to ensure individual record values can't be recovered. Therefore, techniques are designed to derive aggregate distributions on the perturbed records. Subsequently, data mining techniques could be developed to be able to use these aggregate distributions. In this section, I will discuss the randomization opportunity for privacy preserving data mining. The randomization method may be traditionally employed in the context of distorting data by probability distribution for methods for example surveys who have an evasive answer bias as a consequence of privacy concerns. It has additionally been extended towards problem of privacy preserving data mining .The strategy of randomization is actually a follows. Consider a set of data records denoted by $R = r_1.......r_N$. For record $r_i \in R$, we noise component which is drawn from the probability distribution $TY(y)$.These noise components are drawn independently, and are denoted s1...sN. Thus, the new set of distorted records are denoted by $r_1 + s_1...r_N + s_N$. we denote this new set of records by $U_1...U_N$. On the whole, the assumption is how the variance on the added noise is just right, so the original record values is not easily guessed from the distorted data. Thus, the initial records can't be recovered, however the distribution in the original records can be recovered. Thus, if R be the stochastic variable denoting your data distribution to the original record, T function as stochastic variable describing the noise distribution, and U function as variate denoting the last record, we have now:

$$U = R + T$$

$$R = U - T$$

Now, we observe that N instantiations in the probability distribution U are known, When the distribution T is well know publicly. For a big enough number of values of N, the distribution U might be approximated closely simply using a number of methods for example kernel density estimation. By subtracting T through the approximated distribution of U, you are able to approximate the initial probability distribution R. Used, it's possible to combine the whole process of approximation of U with subtraction with the distribution T from U using a selection of iterative methods for example those discussed in earlier paper. Such iterative methods routinely have a greater accuracy compared to the sequential solution of first approximating U then subtracting T from that. For example, the EM method proposed in research paper shows many optimal properties in approximating the distribution of R.

The k-anonymity Model

The k-anonymity model was evolved because of the prospect of indirect identification of records from common databases. This is as long as combinations of record attributes can be used to absolutely identify individual records. In the k-anonymity method, we recede the granularity of data representation with the need of techniques such as generalization and suppression. This granularity is recede sufficiently that any obsessed record maps onto at least k other records in the data. The l-diversity model was constructed to handle a few weaknesses in the k-anonymity model as protecting identities to the level regarding k-individuals is not the like as protecting the corresponding sensitive values, especially when there is analogy of sensitive values within a group. To do so, the perception of intra-group diversity of sensitive values is developed within the anonymization scheme. In lots of applications, the results records are manufactured available by simply removing key identifiers such as name and social-security numbers from personal records. However, other varieties of attributes (often known as pseudo-identifiers) can be utilized to be able to accurately identify the records. As an example, attributes including age, zip-code and sex can be bought in public records such as census rolls. When these attributes can also be found in a very given data set, they are often utilized to infer the identity on the corresponding individual. A combination of these attributes can be very powerful, given that they enable you to reduce the options to some small number of individuals. In k-anonymity techniques, we reduce the granularity of representation of those pseudo-identifiers by using techniques for instance generalization and suppression. Inside technique of generalization, the attribute values are generalized to some zero in order to scale back the granularity of representation. One example is, the birth date may be generalized to your range for example year of birth, in order to reduce the risk of identification. Within the technique of suppression, the additional value with the attribute is taken away completely. It is clear that such methods lessen the risk of identification if you use public record information, while lowering the accuracy of applications about the transformed data. As a way to lessen the risk of identification, the k-anonymity approach requires that every tuple within the table take distinguish ability related to no less than k respondents.

Distributed Privacy Preservation

In many cases, particular entities may want to derive aggregate results from data sets whatever are partitioned across these entities. Such partitioning may be horizontal (when the records are distributed across multiple entities) or vertical (when the attributes are distributed across multiple entities). While the particular entities may not desire to division their entire data sets, they may consent to finite information sharing with the way of a variety of protocols. The overall effect of such techniques is to maintain privacy for all particular entity, although deriving aggregate results over the entire data.

Distributed Algorithms for k-Anonymity

Many times, you should maintain k-anonymity across different distributed parties. A k-anonymous protocol for data that's vertically partitioned across two parties is described. The broad idea is designed for both the parties to recognize the quasi-identifier to generalize on the same value before release. The generalization shall be performed before release. a considerable case where each site can be a customer which owns exactly one tuple from your data. It is assumed how the data record has both sensitive attributes and quasi-identifier attributes. The solution uses encryption on the sensitive attributes. The sensitive values can be decrypted only if therefore are in least k records with the same values for the quasi-identifiers. Thus k-anonymity is maintained. The challenge of k-anonymity is also crucial in the context of hiding identification has gone south distributed location based services. In such cases, k-anonymity in the user-identity is maintained regardless if the positioning details are released. Such location information is often released if a user may send some text at any point from the given location. The same issue arises in the context of communication protocols the location where the anonymity of senders (or receivers) might need to be protected. A note is considered to become sender k-anonymous, if it is guaranteed that an attacker can at most limit the identity in the sender to k individuals. Similarly, some text is claimed to be receiver k-anonymous, if it's guaranteed that attacker can at the most limit the identity with the receiver to k individuals.

References

[1] T.Y.Lin, T.H.Hinke, D.G.Marks and B.Thuraisingham, Security And data mining , Database Security IX Status and Prospects Edited by D.L.Spooner, S.A.Demurjian and J.E.Dobson ISBN 0 412 72920 2,1996.

[2] Yehuda Lindell and Benny Pinkas, Secure Multiparty Computation for Privacy-Preserving Data Mining, Department of Computer Science, Bar-Ilan University, Israel, May 6, 2008.

[3] Korosh Golnabi, Richard K. Min, and Latifur Khan, Analysis of Firewall Policy Rules Using Data Mining Techniques.

[4] Charu C. Aggarwal and Philip S. Yu, Privacy Preserving Data Mining: Models and Algorithms, Springer, e-ISBN: 978-0-387-70992-5. Available at: http://adrem.ua.ac.be/sites/adrem.ua.ac.be/files/Privacy_Preserving_Data_Mining.pdf

[5] Chris Clifton, Murat Kantarcioglu, Jaideep V aidya, Xiaodong Lin and Michael Y . Zhu, Tools for Privacy Preserving Distributed Data Mining, SIGKDD Explorations, Volume 4, Issue 2.

[6] Godswill Chukwugozie Nsofor, A Comparative Analysis of Predictive Data-Mining Techniques, The University of Tennessee, Knoxville , August, 2006.

[7] QIANG YANG and XINDONG WU, 10 CHALLENGING PROBLEMS IN DATA MINING RESEARCH, International Journal of Information Technology & Decision Making Vol. 5, No. 4, World Scientific Publishing Company, 2006.

[8] Charu C. Aggarwal and Philip S. Yu, A General Survey of Privacy-Preserving Data Mining Models and Algorithms. Available at: http://www.polyteknisk.dk/related_materials/9780387709918_Chapter_1.pdf

[9] Benny Pinkas, Cryptographic techniques for privacy-preserving data mining, SIGKDD Explorations, Volume 4, Issue 2.

[10] Wenke Lee and Salvatore J.Stolfo, Data Mining Approaches for Intrusion Detection.

[11] Yehuda Lindell and Benny Pinkas, Privacy Preserving Data Mining, M. Bellare (Ed.): CRYPTO 2000, LNCS 1880, Springer-Verlag Berlin Heidelberg, 2000.

[12] Sheng Zhong, Privacy, Integrity and Incentive-Compatibility in Computations with Un-trusted Parties, Yale University, 2004.

[13] Rakesh Agrawal and Ramakrishnan Srikant, Privacy-Preserving Data Mining, IBM Almaden Research Center 650 Harry Road, San Jose, CA 95120.

[14] Murat Kantarcoglu and Chris Clifton, Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2003.

[15] Advantages and Disadvantages of Data Mining, Available at: http://www.zentut.com/data-mining/advantages-and-disadvantages-of-data-mining/

[16] Aman Sagar, Sanjeev Kumar, Palladium in Cryptography: The Advancement in Data Security, HCTL Open International Journal of Technology Innovations and Research, Volume 7, January 2014, ISSN: 2321-1814, ISBN: 978-1-62951-250-1.