Security for Data mining and its challenges

Mula Malyadri[1], **B. Laxmaiah**[2]
1 & 2 :Assistant Professor, Dept of CSE, CMR Technical campus, Medchal, TS, India

Abstract—

In this paper we first look at data mining applications in safety measures and their suggestions for privacy. After that we then inspect the idea of privacy and give a synopsis of the developments particularly those on privacy preserving data mining. We then present an outline for research on confidentiality and data mining.

Keywords: clustering, Data-Driven Approach, Model Driven, privacy in Distributed mining

INTRODUCTION

The amount of data being generated and stored is growing exponentially, due in large part to the continuing advances in computer technology. This presents tremendous opportunities for those who can unlock the information embedded within this data, but also introduces new challenges. In this chapter we discuss how the modern field of data mining can be used to extract useful knowledge from the data that surround us. Those that can master this technology and its methods can derive great benefits and gain a competitive advantage. In this introduction we begin by discussing what data mining is, why it developed now and what challenges it faces, and what types of problems it can address. In subsequent sections we look at the key data mining tasks: prediction, association rule analysis, cluster analysis, and text, link and usage mining. Before concluding we provide a list of data mining resources and tools for those who wish further information on the topic.

DATA mining is the procedure of posing questions and taking out patterns, often in the past mysterious from huge capacities of data applying pattern matching or other way of thinking techniques. Data mining has several applications in protection together with for national protection as well as for cyber protection. The pressure to national protection includes aggressive buildings, demolishing dangerous infrastructures such as power grids and telecommunication structures. Data mining techniques are being examined to realize who the doubtful people are and who is competent of functioning revolutionary activities. Cyber security is concerned with defending the computer and network systems against fraud due to Trojan cattle, worms and viruses. Data mining is also being useful to give solutions for invasion finding and auditing. While data mining has several applications in protection, there are also serious privacy fears. Because of data mining, even inexperienced users can connect data and make responsive associations. Therefore we must to implement the privacy of persons while working on practical data mining. In this paper we will talk about the developments and instructions on privacy and data mining. In particular, we will give a general idea of data mining, the different types of threats and then talk about the penalty to privacy. This paper is organized as follows. Section 2 talks about data mining for safety applications. Section 3 explains the overview of privacy. Section 4 discusses different aspects of data mining on. Directions are provided in section 5 and section 6 gives the conclusion of this paper or work done on the paper

Motivation and Challenges

Data Mining developed as a new discipline for several reasons. First, the amount of data available for mining grew at a tremendous pace as computing technology became widely deployed. Specifically, high speed networks allowed enormous amount of data to be transferred and rapidly decreasing disk costs permitted this data to be stored cost-effectively. The size and scope of these new datasets is remarkable. According to a recent industry report (International Data Corporation 2007), in 2006 161 Exabyte's (161 Billion Gigabytes) of data were created and in 2010 988 Exabyte's of data will be created. While these figures include data in the form of email, pictures and video, these and other forms of data are increasingly being mined. Traditional corporate datasets, which include mainly fixed-format numerical data, are also quite huge, with many companies maintaining Terabyte datasets that record every customer transaction

Overview of Data Mining Tasks

The best way to gain an understanding of data mining is to understand the types of tasks, or problems, that it can address. At a high level, most data mining tasks can be categorized as either having to do with prediction or description. Predictive tasks allow one to predict the value of a variable based on other existing information. Examples of predictive tasks include predicting when a customer will leave a company (Wei and Chiu 2002), predicting whether a transaction is fraudulent or not (Fawcett and Provost 1997), and identifying the best customers to receive direct marketing offers (Ling and Li 2000). Descriptive tasks, on the other hand, summarize the data in some manner. Examples of such tasks include automatically segmenting customers based on their similarities and differences (Chen et al. 2006) and finding associations between products in market basket data (Agrawal and Srikant 1994). Below we briefly describe the major predictive and descriptive data mining tasks. Each task is subsequently described in greater detail later in the chapter.

Classification and Regression

Classification and regression tasks are predictive tasks that involve building a model to predict a target, or dependent, variable from a set of explanatory, or independent, variables. For classification tasks the target variable usually has a small number of discrete values (e.g., “high” and “low”) whereas for regression tasks the target variable is continuous. Identifying fraudulent credit card transactions (Fawcett and Provost 1997) is a classification task while predicting future prices of a stock (Enke and Thawornwong 2005) is a regression task. Note that the term “regression” in this context should not be confused with the regression methods used by statisticians (although those methods can be used to solve regression tasks).

The best way to gain an understanding of data mining is to understand the types of tasks, or problems, that it can address. At a high level, most data mining tasks can be categorized as either having to do with prediction or description. Predictive tasks allow one to predict the value of a variable based on other existing information. Examples of predictive tasks include predicting when a customer will leave a company (Wei and Chiu 2002), predicting whether a transaction is fraudulent or not (Fawcett and Provost 1997), and identifying the best customers to receive direct marketing offers (Ling and Li 2000). Descriptive tasks, on the other hand, summarize the data in some manner. Examples of such tasks include automatically segmenting customers based on their similarities and differences (Chen et al. 2006) and finding associations between products in market basket data (Agrawal and Srikant 1994). Below we briefly describe the major predictive and descriptive data mining tasks. Each task is subsequently described in greater detail later in the chapter.

Data mining technology is not only composed by efficient and effective algorithms, executed as standalone kernels. Rather, it is constituted by complex applications articulated in the non-trivial interaction among hardware and software components, running on large scale distributed environments. This last feature turns out to be both the cause and the effect of the inherently distributed nature of data, on one side, and, on the other side, of the spatiotemporal complexity that characterizes many DM applications. For a growing number of application fields, Distributed Data Mining (DDM) is therefore a critical technology. In this research paper, after reviewing the open problems in DDM, we describe the DM jobs on Grid environments. We will introduce the design of Knowledge Grid System.

By analyzing three different approaches, we have provided some definitions of DDM Systems. They pose different problems and have different benefits. Existing DDM systems can in fact be classified in one of these approaches

The simplest model for a DDM system only takes into account the distributed nature of data, but then relies on local and sequential DM technology. Since in this system the focus is solely posed in the location of data, we refer to this model as data-driven

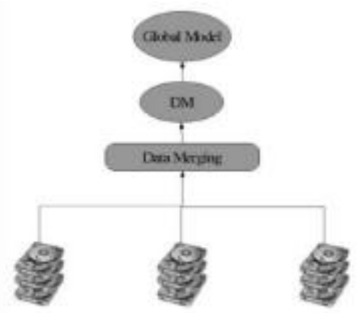

Figure 1 : Data-Driven Approach for Distributed Data Mining

In this model, data are located in different sites which do not need to have any computational capability. The only requirement is to be able to move the data to a central location in order to merge them and then apply sequential DM algorithms. The output of the DM analysis, i.e. the final knowledge models are then either delivered to the analyst' location or accessed locally where they have been computed

The process of gathering data in general is not simply a merging step and depends on the original distribution. For example data can be partitioned horizontally – i.e. different records are placed in different sites – or vertically – i.e. different attributes of the same records are distributed across different sites. Also, the schema itself can be distributed, i.e. different tables can be placed at different sites. Therefore when gathering data it is necessary to adopt the proper merging strategy.

Model-driven: A different approach is the one we call model-driven. Here, each portion of data is processed locally to its original location, in order to obtain partial results referred to as local knowledge models. Then the local models are gathered and combined together to obtain a global model.

Also in this approach, for the local computations it is possible to reuse sequential DM algorithms, without any modification. The problem here is how to combine the partial results coming from the local models. Different techniques can be adopted, based on voting strategies or collective operations, for example. Multi-agent systems

may apply meta-learning to combine partial results of distributed local classifiers

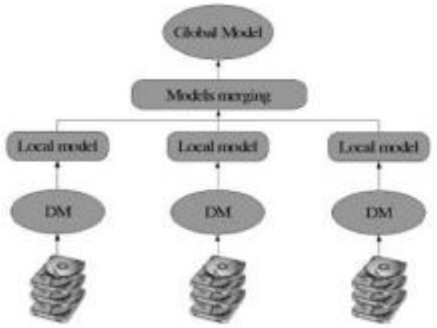

Figure 2 : Model-Driven Approach for Distributed Data Mining

The draw-back of the model-driven approach is than it is not always possible to obtain an exact final result, i.e. the global knowledge model obtained may be different from the one obtained by applying the data-driven approach (if possible) to the same data. Approximated results are not always a major concern, but it is important to be aware of that. Moreover, in this model hardware resource usage is not optimized. If the heavy computational part is always executed locally to data, when the same data is accessed concurrently, the benefits coming from the distributed environment might vanish due to the possible strong performance degradation.

Architecture-driven: In order to be able to control the performance of the DDM system, it is necessary to introduce a further layer between data and computation. As show in below Figure, before starting the distributed computation, we consider the possibility of moving data to different sites with respect to where they are originally located, if this turns out to be profitable in terms of performances. Moreover, we introduce a communication layer among the local DM computations, so that the global knowledge model is built during the local computation. This allows for arbitrary precision to be achieved, at the price of a higher communication overhead. Since in this approach for DDM the focus is on optimized resource usage, we refer to this approach as the architecture-driven.

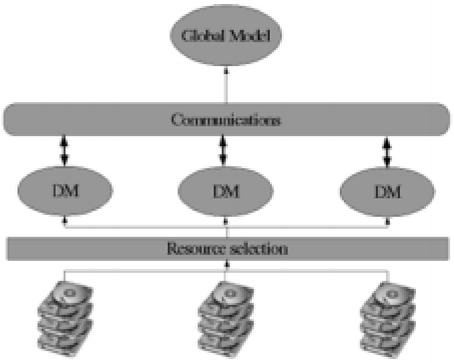

Figure 3 : Architecture-Driven Approach for Distributed Data Mining

The higher flexibility of this model and the potentially higher performance that it is possible to achieve, are payed in terms of the higher management effort that it is necessary to put in place. A suitable scheduling policy must be devised for the resource selection layer. Moreover, DM sequential algorithms are not reusable directly and must be modified or redesigned in order to take advantage of the communication channel among the different DM computations.

Data mining is fitting a key technology for identifying doubtful activities. In this section, data mining will be discussed with respect to use in both ways for non real-time and for real-time applications. In order to complete data mining for counter terrorism applications, one wants to gather data from several sources. For example, the subsequent information on revolutionary attacks is wanted at the very least: who, what, where, when, and how; personal and business data of the possible terrorists: place of birth, religion, education, ethnic origin, work history, finances, criminal record, relatives, friends and associates, and travel history; unstructured data: newspaper articles, video clips, dialogues, e-mails, and phone calls. The data has to be included, warehoused and mined. One wants to develop sketches of terrorists, and activities/threats. The data has to be mined to take out patterns of possible terrorists and forecast future activities and goals. Fundamentally one wants to find the "needle in the haystack" or more suitably doubtful needles among probably millions of needles. Data integrity is essential and also the methods have to SCALE. For several applications such as urgent situation response, one needs to complete real-time data mining. Data will be incoming from sensors and other strategy in the form of nonstop data streams together with breaking news, videocassette releases, and satellite images. Some serious data may also exist in caches. One wants to quickly sift through the data and remove redundant data for shortly use and analysis (non-real-time data mining). Data mining techniques require to meet timing restriction and may have to stick the quality of service (QoS) tradeoffs among suitability, accuracy and precision. The consequences have to be accessible and visualized in real-time. Additionally, alerts and triggers will also have to be employed. Efficiently applying data mining for safety applications and to develop suitable tools, we need to first find out what our present capabilities are. For instance, do the profitable tools balance? Do they effort only on particular data and limited cases? Do they carry what they assure? We require a balanced objective study with display. At the same time, we also require to work on the large picture. For instance what do we desire the data mining tools to carry out? What are our end consequences for thepredictable future? What are the standards for achievement? How do we assess the data mining algorithms? What test beds do we construct? We require both a near-term as well as longer-term resolutions. For the future, we require to influence present efforts and fill the gaps in a objective aimed way and complete technology transfer. For the longer-term, we require a research and development diagrams. In summary, data mining is very helpful to resolve security troubles. Tools could be utilized to inspect audit data and flag irregular behavior. There are many latest works on applying data mining for cyber safety applications, Tools are being examined to find out irregular patterns for national security together with those based on categorization and link analysis. Law enforcement is also using these kinds of tools for fraud exposure and crime solving.

PRIVACY SUGGESTIONS

We require finding out what is meant by privacy before we look at the privacy suggestions of data mining and recommend efficient solutions. In fact different society-ties have different ideas of privacy. In the case of the medical society, privacy is about a patient finding out what details the doctor should discharge about him/her. Normally employers, marketers and insurance corporations may try to find information about persons. It is up to the individuals to find out the details to be given about him. In the monetary society, a bank customer finds out what financial details the bank should give about him/her. Additionally, retail corporations should not be providing the sales details about the persons unless the individuals have approved the release. In the case of the government society, privacy may get a whole new significance. For example, the students who attend my classes at AFCEA have pointed out to me that FBI would gather data about US citizens. However FBI finds out what data about a US citizen it can provide to say the CIA. That is, the FBI has to make sure the privacy of US citizens. Additionally, permitting access to individual travel and spending data as well as his/her web surfing activities should also be provided upon receiving permission from the individuals. Now that we have explained what we signify by privacy, we will now check up the privacy suggestion of data mining. Data mining provides us "facts" that are not clear to human analysts of the data. For instance, can general tendency across individuals be calculated without enlightening details about individuals? On the other hand, can we take out highly private relations from public data? In the former case we require to protect the person data values while enlightening the associations or aggregation while in the last case we need to defend the associations and correlations between the data

The objective is to perform effective data mining but at the same time guard individual data values and sensitive relations. Agrawal was the first to invent the word privacy preserving data mining. His early work was to initiate random values into the data or to bother the data so that the real data could be confined. The challenge is to initiate random values or agitate the values without touching the data mining results [1]. Another new approach is the Secure Multi-party Computation (SMC) by Kantarcioglu and Clifton [3]. Here, each party knows its individual contribution but not the others' contributions. Additionally the final data mining outcomes are also well-known to all. Various encryption techniques utilized to make sure that the entity values are protected. SMC was demonstrating several promises and can be used also for privacy preserving scattered data mining. It is provably safe under some suppositions and the learned models are correct; It is assumed that procedures are followed which is a semi truthful model. Malicious model is also investigated in some current work by Kantarcioglu and Kardes [4]. Many SMC footed privacy preserving data mining algorithms contribute to familiar sub-protocols (e.g. dot product, summary, etc.). SMC does have any disadvantage as it's not competent enough for very large datasets. (E.g. petabyte sized datasets); Semihonest model may not be reasonable and the malicious model is yet slower. There are some novel guidelines where novel models are being discovered that can swap better between efficiency and security. Game theoretic and motivation issues are also being discovered. Finally merging anonimization with cryptographic techniques is also a route. Before performing an evaluation of the data mining algorithms, one wants to find out the objectives. In some cases the objective is to twist data while still preserving some assets for data mining. Another objective is to attain a high data mining accuracy with greatest privacy protection. Our current work imagines that Privacy is a personal preference, so should be individually adjustable. That is, we want to make privacy protecting data mining approaches to replicate authenticity. We examined perturbation based approaches with real-world data sets and provided applicability learning to the existing approaches [5]. We found that the rebuilding of the original sharing may not work well with real-world data sets. We attempted to amend perturbation techniques and adjust the data mining tools. We also developed a new privacy preserving decision tree algorithm [6]. Another growth is the platform for privacy preferences (P3P) by the World Wide Web association (W3C). P3P is an up-and-coming standard that facilitates web sites to convey their privacy practices in a typical format. The format of the strategies can be robotically recovered and appreciated by user agents. When a user comes in a web site, the privacy policies of the web site are communicated to the user; if the privacy policies are dissimilar from user favorites, the user is notified; User can then make a decision how to continue. Several major corporations are working on P3P standards.

DIRECTIONS FOR PRIVACY

Thuraisingham verified in 1990 that the inference problem in common was unsolvable; therefore the suggestion was to discover the solvability features of the problem [7]. We were able to explain comparable results for the privacy problem. Therefore we need to inspect the involvement classes as well as the storage and time complication. We also need to discover the base of privacy preserving data mining algorithms and connected privacy ways out. There are various such algorithms. How do they evaluate with each other? We need a test bed with practical constraints to test the algorithms. Is it meaningful to observe privacy preserving data mining for each data mining algorithm and for all application? It is also time to enlarge real world circumstances where these algorithms can be used. Is it possible to build up realistic commercial products or should each association get used to products to suit their needs? Investigative privacy may create intelligence for healthcare and monetary applications. Does privacy work for Defense and Intelligence purposes? Is it even important to have privacy for inspection and geospatial applications? Once the image of my home is on Google Earth, then how much isolation can I have? I may wish for my position to be private, but does it make sense if a camera can detain a picture of me? If there are sensors all over the position, is it important to have privacy preserving surveillance? This proposes that we require application detailed privacy. Next what is the connection between confidentiality, privacy and faith? If I as a user of Association A send data about me to Association B, then imagine I read the privacy policies imposed by Association B. If I agree to the privacy policies of Association B, then I will drive data about me to Association B. If I do not concur with the policies of association B, then I can bargain with association B. Even if the website affirms that it will not distribute private information with others, do I faith the website? Note: while secrecy is enforced by the association, privacy is strong-minded by the user. Therefore for confidentiality, the association will conclude whether a user can have the data. If so, then the association can additional decide whether the user can be trusted. Another way is how can we make sure the confidentiality of the data mining procedures and outcome? What sort of access control policies do we implement? How can we faith the data mining procedures and results as well as authenticate and validate the results? How can we join together confidentiality, privacy and trust with high opinion to data mining? We need to check up the research challenges and form a research schema. One question that Rakesh Agrawal inquired at the 2003 SIGKDD panel on Privacy [2] "is privacy and data mining friends or rivals? We think that they are neither associates nor rivals. We need progresses in both data mining and privacy. We require planning flexible systems. For some applications one may have to hub entirely on "pure" data mining while for some others there may be a need for "privacy-preserving" data mining. We need flexible data mining techniques that can settle in to the changing environments. We consider that technologists, legal specialists, social scientists, policy makers and privacy advocates MUST work together.

CONCLUSION

In this paper we have examined data mining applications in security and their implications for privacy. We have examined the idea of privacy and then talked about the developments particularly those on privacy preserving data mining. We then presented an agenda for research on privacy and data mining. Here are our conclusions. There is no collective definition for privacy, each organization must clear-cut what it indicates by privacy and develop suitable privacy policies. Technology only is not adequate for privacy; we require Technologists, Policy expert, Legal experts and Social scientists to effort on Privacy. Some well acknowledged people have believed 'Forget about privacy" Therefore, should we follow research on Privacy? We trust that there are attractive research problems; therefore we need to carry on with this research. Additionally, some privacy is better than nil. One more school of consideration is tried to avoid privacy destructions and if destructions take place then put on trial. We need to put into effect suitable policies and check up the legal aspects. We need to undertake privacy from all directions

REFERENCES

[1] Agrawal, R., Srikant, R.: Privacy-Preserving Data Mining. In: SIGMOD Conference, pp.439–450 (2000)

[2] Agrawal, R.: Data Mining and Privacy: Friends or Foes. In: SIGKDD Panel (2003)

[3] Kantarcioglu, M., Clifton, C.: Privately Computing a Distributed k-nn Classifier. In: Bou-licaut, J.-F., Esposito,

F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS, vol. 3202,279–290. Springer, Heidelberg (2004)

[4] Kantarcioglu, M., Kardes, O.: Privacy-Preserving Data Mining Applications in the Mali-cious Model. In: ICDM Workshops, pp. 717–722 (2007)

[5] Liu, L., Kantarcioglu, M., Thuraisingham, B.M.: The applicability of the perturbation based privacy preserving data mining for real-world data. Data Knowl. Eng. 65(1), 5–21 (2008)

[6] Liu, L., Kantarcioglu, M., Thuraisingham, B.M.: A Novel Privacy Preserving Decision Tree. In: Proceedings Hawaii International Conf. on Systems Sciences (2009)

[7] Thuraisingham, B.: One the Complexity of the Inference Problem. In: IEEE Computer Se-curity Foundations Workshop (1990) (also available as MITRE Report, MTP- 291)

[8] Thuraisingham, B.M.: Privacy constraint processing in a privacy-enhanced database man-agement system. Data Knowl. Eng. 55(2), 159–188 (2005)

[9] Clifton, C.: Using Sample Size to Limit Exposure to Data Mining. Journal of Computer Security 8(4) (2000)

Author Details

Mula Malyadri, Assistant Professor, working in the dept. of CSE in CMR TECHNICAL CAMPUS, Medchal, Hyderabad. He is having 10 years of experience in teaching for UG and PG engineering students. He received B.E.(ECM) from Koneru Lakshmaiah College of Engineering(KLCE), M.Tech(CSE) from JNTUH, and Pursuing Ph.D.(CSE) from OPJS University. He published papers in international conferences and journals. His research interests are Data Mining and Ware Housing. He is also member of several technical organizations

Bagam Laxmaiah was received Bachelor of Science in Computer Science from Kakatiya University, 2003,Received Master of Computer Application from Kakatiiya University, 2008 and Received Master of Technology in Computer Science & Web Technologies from JNTU Hyderabad, 2011.He is currently working as an Asst.Professor, Department of Computer Science & Engineering in CMR Technical Campus, Kandlakoya, Medchal, Hyderabad. He is having 10 years of experience in teaching for UG and PG engineering students. A Member of the Indian Society for Technical Education, a Member of International Association of Engineers and Scientists.