Scientific Journal of Impact Factor (SJIF): 3.134
ISSN (Online): 2348-4470
ISSN (Print) : 2348-6406
International Journal of Advance Engineering and Research Development
Volume 2, Issue 1, January -2015

FOCUSED PRIVACY PRESERVING OF DATA USING DECISION TREE INDUCTION CLASSIFIER

Komal Arunjeet Kaur[1] and Dr. Lalita Bhutani[2]

[1]Department of Computer Science & Engineering,SVIET,komalarunjeetkaur@gmail.com
[2]Department of Computer Science & Engineering,SVIET,lalita.b@rediffmail.com

Abstract- *Data mining is a recent advance technology with great potential to help companies focus on the crucial information in the data they have collected about the behavior of their customers. In this paper we are trying to use Data mining as a strong tool for security purposes. This paper covers collection of information within the data that queries and reports can't effectively reveal. Privacy-preserving data publishing addresses the problem of disclosing sensitive data when mining for useful information. In already existing privacy models, differential privacy provides one of the strongest privacy guarantees. In this paper, we have resolved issues like Experimental determination and optimization of the way of sharing information without leaking personal data. Continuous monitoring was being done and with the help of DTI (Decision tree induction classifier) we have classified the upcoming events and made data sharing more reliable. More over the location based identification is also included to ensure the overall security in the future also by resolving this chronic problem.*

Keywords: *Decision Tree Induction (DTI), Decision Tree (DT), Cumulative Sum (CUSUM).*

I. INTRODUCTION

Data mining, or knowledge discovery, is the computer-facilitated process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They cleanse databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

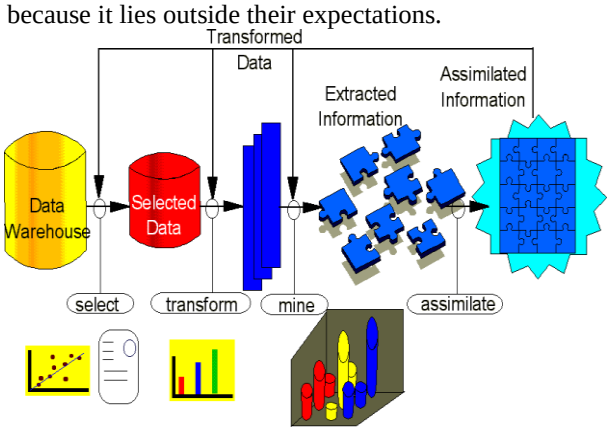

Figure 1. Process of Data Mining

Data mining derives its name from the similarities between searching for decisive data in a large database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides. In typical data mining systems, the mining procedures require computational intensive computing units for data analysis and comparisons [1]. Data mining can be seen as a process for extracting hidden and valid knowledge from huge databases [11]. A computing platform is, therefore, needed to have efficient access to, at least, two types of resources: data and computing processors. For small scale data mining tasks, a single desktop computer, which contains hard disk and CPU processors, is sufficient to fulfill the data mining goals. For example, regression analysis was used at some point in time in the Vietnam war to predict the possible mortar attacks [12]. Indeed, many data mining algorithm are designed for this type of problem settings. Data mining an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use [2]. For medium scale data mining tasks, data are typically large (and possibly distributed) and cannot be fit into the main memory. Specific uses of data mining include:

- *Market segmentation* - Identify the common characteristics of customers who buy the same products from your company.

- *Customer churn* - Predict which customers are likely to leave your company and go to a competitor.
- *Fraud detection* - Identify which transactions are most likely to be fraudulent.
- *Direct marketing* - Identify which prospects should be included in a mailing list to obtain the highest response rate.
- *Interactive marketing* - Predict what each individual accessing a Web site is most likely interested in seeing.
- *Trend analysis* - Reveal the difference between a typical customer this month and last.

Large databases exist today due to the rapid advances in communication and storing systems. Each database is owned by a particular autonomous entity, for example, medical data by hospitals, income data by tax agencies, financial data by banks, and census data by statistical agencies. Moreover, the emergence of new paradigms such as cloud computing increases the amount of data distributed between multiple entities. These distributed data can be integrated to enable better data analysis for making better decisions and providing high-quality services. For example, data can be integrated to improve medical research, customer service, or homeland security.

In statistical quality control, the CUSUM (or cumulative sum control chart) is a sequential analysis technique. It is typically used for monitoring change detection [6]. It referred to a "quality number" θ, by which it meant a parameter of the probability distribution; for example, the mean. And, devised Cumulative Sum **(**CUSUM) as a method to determine changes in it, and proposed a criterion for deciding when to take corrective action. When the CUSUM method is applied to changes in mean, it can be used for step detection of a time series. A few years later, George Alfred Barnard developed a visualization method, the V-mask chart, to detect both increases and decreases in θ [5].

II. PRIVACY PRESERVING

Privacy preserving has originated as an important concern with reference to the success of the data mining. Privacy preserving data mining (PPDM) deals with protecting the privacy of individual data or sensitive knowledge without sacrificing the utility of the data. People have become well aware of the privacy intrusions on their personal data and are very reluctant to share their sensitive information. This may lead to the inadvertent results of the data mining. Within the constraints of privacy, several methods have been proposed but still this branch of research is in its infancy. Privacy models are as follows:

- **Non-interactive:** In this database owner first anonymizes the raw data and then releases the anonymized version for data analysis. Database is sanitized and released. In a non interactive framework, a database owner first anonymizes the raw data and then releases the anonymized version for data analysis. Once the data are published, the data owner has no further control over the published data. This approach is also known as privacy preserving data publishing (PPDP) [4].
- **Interactive:** Multiple questions asked/answered adaptively. In an interactive framework, a data miner can pose queries through a private mechanism, and a database owner answers these queries in response. In an interactive framework, a data miner can pose queries through a private mechanism, and a database owner answers these queries in response.

The success of privacy preserving data mining algorithms is measured in terms of its performance, data utility, level of uncertainty or resistance to data mining algorithms etc. Privacy-preserving data publishing addresses the problem of disclosing sensitive data when mining for useful information [8]. However no privacy preserving algorithm exists that outperforms all others on all possible criteria [3]. Differential privacy [10] has recently received considerable attention as a substitute for partition-based privacy models for PPDM. However, so far most of the research on differential privacy concentrates on the interactive setting with the goal of reducing the magnitude of the added noise [14,13,15] releasing certain data mining results [7] ,or determining the feasibility and infeasibility results of differentially-private mechanisms [5,10,2]. Effects of data leakage :

- **Invasion of privacy:** The intrusion into the personal life of another, without just cause, which can give the person whose privacy has been invaded a right to bring a lawsuit for damages against the person or entity that intruded. However, public personages are not protected in most situations, since they have placed themselves already within the public eye, and their activities (even personal and sometimes intimate) are considered newsworthy, i.e. of legitimate public interest.
- **Trust issues:** Enable collaboration/communication. Social paradigm: small village, big city, Dynamic and open environments. To initiate and build trust we should have to use Formal models, Type of trust data, users, system components, Context dependent, bi-directional, asymmetric, Direct evidence and second-hand recommendations.
- **Ruin reputation:** Maintaining a reputation is hard. Failure to do so can be catastrophic. The insurance industry is proficient at helping clients defend against executive risks or loss of property. More challenging, however, is safeguarding an organization's reputation.

III. DECISION TREE INDUCTION

Decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. It is a flowchart like structure in which internal node represents a "test" on an attribute (e.g. whether a coin flip

comes up heads or tails), each branch represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes). The path from root to leaf represents classification rules.
In decision analysis a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values of competing alternatives are calculated.
Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. If in practice decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Another use of decision trees is as a descriptive means for calculating conditional probabilities.
Decision trees, influence diagrams, utility functions, and other decision analysis tools and methods are taught to undergraduate students in schools of business, health economics, and public health, and are examples of operations research or management science methods.
They are easily understandable. They build a model (made up by rules) easy to understand for the user.

IV. PROPOSED METHODOLOGIES

- To experimentally determine and optimize the way of sharing information without leaking personal details.
- To monitor continuously the data that is being shared and learning or training data set of favorable and non-favorable events.
- To apply DTI (decision tree induction) classifier to classify the upcoming events and make data sharing at more reliable.

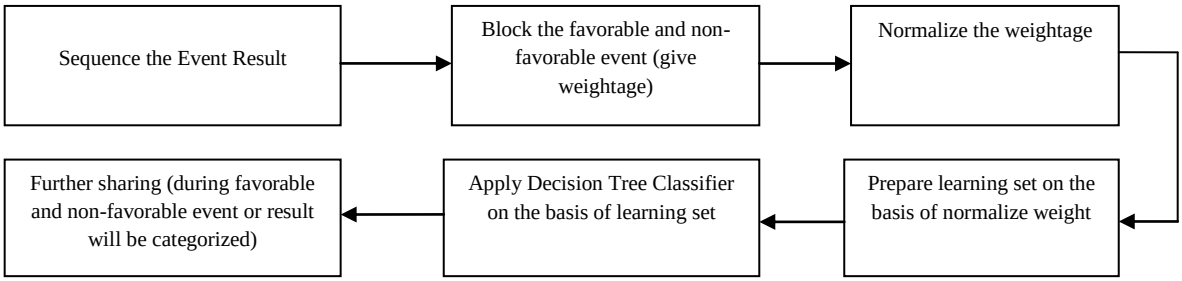

Figure 2. Represents Proposed Methodology

V. RESULTS AND DISCUSSION

Monitoring of user profile: First of all we filter the file according to Weights with the help of code that we have developed in MATLAB GUI tool. Then to Identify whether user is legitimate or attacker. The description to identify the type of user whether it is a 'Legitimate user 'or an 'Attacker' and the condition to identifying the user is if the average fluctuation is over or greater than 12 and it repeats more than 10 times (i.e. if the avg. fluctuation goes over '12' for more than 10 times for an ip address) than the user will consider as an Attacker otherwise it will be a legitimate user or normal user.

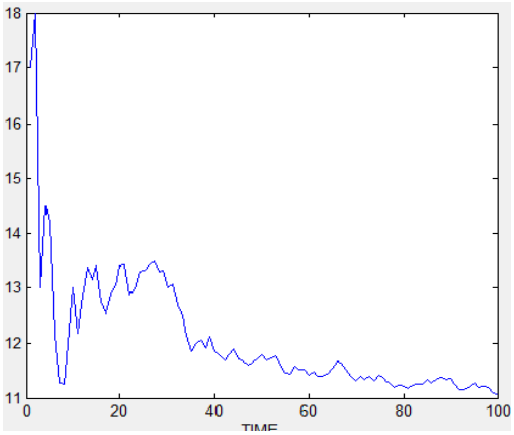

Figure 3. Average fluctuation represents Attacker

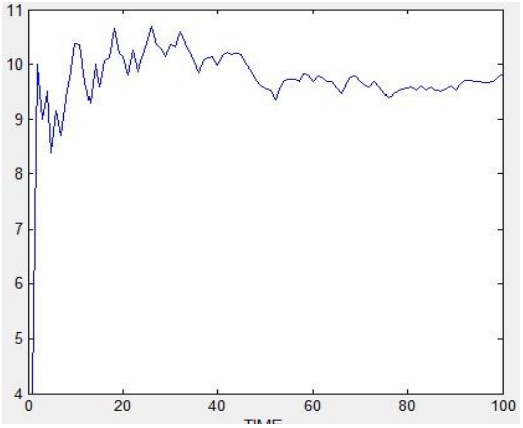

Figure 4. Average fluctuation represents Legitimate user

If the average fluctuation is over or greater than 12 then the type of user is attacker as figure 3, then the type of attack can be of two types, whether it is a DOS or DDOS according to the networks from which the packets are coming. In computing, a denial-of-service (DoS) or distributed denial-of-service (DDoS) attack is an attempt to make a machine or network resource unavailable to its intended users. Then the decision tree induction classifier is to be used. To check the accuracy of the type of user which represents zero(0) percent accuracy for attacker but not for legitimate user.
The results determined after applying CUSUM algorithm for monitoring user's profile are successfully detecting the abnormalities and abrupt changes if the attacker tries to enter into system and intent to alter the documents. In the accuracy comparison with results of the base paper our technique has shown more accuracy than the previous one. Therefore, by following this technique user data can be protected against insider theft attacks and any malicious activity can be detected. As in the analysis the average fluctuation shows the difference between the access behavior of the user and the decoy technology is also effective in confusing the attacker and making the attacker believe that it is a useful file for the attacker. Through this research we concluded that decoy technology and fog computing together can provide security to real world problems like insider data theft attacks.

VI. CONCLUSION

In this paper, we have created a model which can detect the insider theft attacks. The algorithm which we have used for detecting the abnormality in access behavior of the user has given accurate results and is more efficient than previous algorithms. With the help this security mechanism we can identify the user access pattern and also fog computing helps in finding the location of the user if it is detected as an attacker. In fog computing all the logs are cached at a location near to the user, therefore it is easier to retrieve the logs and identify the user behavior. The detection criteria should be such that the abnormality or any anomaly is detected accurately. By increasing the number of user's cases we can get more accurate results. The system should be able to recognize the pattern generated earlier when the legitimate user had accessed the file system. For these reasons proper learning should be provided to the system so that it could detect the abrupt changes in behavior of the user if it is not authorized or is an insider. More over the location based identification is also included to ensure the overall security in the future also by resolving this chronic problem.

VII. REFRENCES

[1] V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-Art in Privacy Preserving Data Mining," ACM SIGMOD Record, vol. 3, no. 1, pp. 50-57, Mar. 2004.
[2] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE "Data Mining with Big Data" Page(s): 97 - 107 , January 2014.
[3] Malik, M.B. ; Ghazi, M.A. ; Ali, R. "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects" Computer and Communication Technology (ICCCT),Page(s):26-32,(2012) .
[4] B.C.M. Fung, K. Wang, R. Chen, and P.S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments", ACM Computing Surveys, vol. 42, no. 4, pp. 1-53, June 2010.
[5] Barnard, G.A. "Control charts and stochastic processes", Journal of the Royal Statistical Society. B (Methodological) (21, number 2): 239–71, (1959).
[6] Grigg et al.; Farewell, VT; Spiegelhalter, DJ "The Use of Risk-Adjusted CUSUM and RSPRT Charts for Monitoring in Medical Contexts", Statistical Methods in Medical Research **12** (2): 147–170, (2003).
[7] K. Chaudhuri, C. Monteleoni, and A. Sarwate, "DifferentiallyPrivate Empirical Risk Minimization", J. Machine Learning Research, vol. 12, pp. 1069-1109, July 2011.
[8] Noman Mohammed, Dima Alhadidi, Benjamin C.M. Fung, Senior Member, IEEE, and Mourad Debbabi "Secure Two-Party Differentially Private Data Release for Vertically Partitioned Data". Page(s): 59 – 71,(2014).
[9] N. Mohammed, R. Chen, B.C.M. Fung, and P.S. Yu, "Differentially Private Data Release for Data Mining", Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '11), Pages 493-501, (2011).
[10] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi. "Discovering Data Mining from Concept to Implementation". Prentice Hall PTR, New Jersey 07458, USA, Page 195, (1998).
[11] Page, E. S. "Continuous Inspection Scheme". Biometrika 41 (1/2): 100–115. pp. 100-115, (June 1954).
[12] V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-Art in Privacy Preserving Data Mining", ACM SIGMOD Record, vol. 3, no. 1, pp. 50-57, Mar. 2004.
[13] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis", Proc. Theory of Cryptography Conf. (TCC '06), Pages 265-284 ,(2006).
[14] I. Dinur and K. Nissim, "Revealing Information while Preserving Privacy", Proc. ACM Symp. Principles of Database Systems (PODS '03), Pages 202-210, (2003).
[15] A. Roth and T. Roughgarden, "Interactive Privacy via the Median Mechanism", Proc. ACM Symp. Theory of Computing (STOC '10), Pages 765-774, 2010.

