A Privacy Preserving Data Mining Technique for Preventing Data Discloser and Leakage

Lokesh Pathak
Computer science Department
Rajive Gandhi Technical University
Bhopal M.P. India

Keshav Puraswani
Computer science Department
Rajive Gandhi Technical University
Bhopal M.P. India

ABSTRACT

The data is an essential element in any business domain, discloser or leakage of sensitive and private data to others can create a number of social and financial issues. In this context the data mining is shifting towards the privacy preserving data mining. In different real world conditions the end data owner submitted their private and confidential data to a business domain. But due to cretin requirements of business intelligence and marketing research the data discloser is required. In these conditions the discloser or leakage of the actual data owner's data can create various issues. In this context the proposed work is intended to work with the privacy preserving data mining environment to preserve the data privacy. The proposed data model consists of the three main contributions first designing the noise based data transformation approach. Secondly the data model help to prevent the data discloser to another party. Third the technique by which the data publishing and it's utility in other public domain becomes feasible. Therefore the proposed work introduces a lightweight privacy preserving data model that combines data from different data sources. Include the regulated noise to entire dataset to modify the values. Process the data using data mining model for finding combined data based decisions and help to publish the data for other marketing and research purposed without disclosing the actual data values. The implementation of the proposed technique is given using JAVA technology and their performance is measured. The obtained results demonstrate the proposed work is helpful for the PPDM based data processing and publishing.

Keywords

Data mining, privacy preserving data mining, decision making, data publishing, data discloser and effects

1. INTRODUCTION

Data mining techniques enable us to analyze the bulk data and obtain the relevant patterns. These extracted patterns can helpful for making decisions, predictions and recognition. Therefore a number of institutions and organizations are accepting the data mining processes i.e. banking, finance, academics and others, but sometimes the decision making process need additional information from other industries. Therefore by combining the available data with the newly appeared data more precise decision can be made. But the data providers are worried to disclose their client's data to other industries. That can be misused by someone or can be used to abuse someone. In order to manage privacy and utility of data (in a data mining system), a privacy preserving technique is required.In this presented work a noise based data privacy preserving data model is proposed for design and implementation. The proposed model is aimed to involve an amount of noise that modify the actual data values to relevant other values. Thus the data is similar but it not contains the actual values to preserve the confidentiality of data. In addition of that to justify the noise based preserving technique a data mining algorithm is also implemented with the system. That algorithm is used to train and classify the data instances in both the data formats i.e. without mixing noise in data and after involving noise to the data. That experiment is performed for finding effect of noise in classifiers performance. If the results are varying in major amount then the application of data or utility of data is also disturbed when the privacy is managed using noise. Otherwise if a regulated amount of noise is introduced then it can help to manage privacy and data mining based decisions also. Using this concept the entire proposed work is designed and developed.

2. PROPOSED WORK

The main aim of the proposed work is to enhance the security in distributed database environment of the end client's data. In this context a data model is presented which works on the data confidentiality during utilization in data mining environment. This chapter discusses the methodology and proposed algorithm which is employed to improve the existing privacy preserving data mining environment.

2.1 System Overview

Privacy preserving data mining is sub-domain of data mining techniques where the data processing and pattern recovery aimed to preserve the confidentiality of data owner. In this context the number of parties can be involved in this process by contributing the own part of data. The data contributions either accepted in attribute basis or the data instance basis. If the data is contributed attribute basis this technique is known as the vertically partitioned data or when the contributors involve the part of data as the instances then it is known as the horizontal partitioned data. In both the techniques the data contributing parties are worried about the sensitivity, confidentiality and the privacy of end data owners. Therefore various techniques based on cryptography and noise based techniques are developed recently. But the cryptographic techniques transform the data entirely and the normal human eyes are not able to understand the data processing outcomes. But in some cases the data publishing and their decisional outcomes are required to publish publically. Therefore the technique required that provide understandable outcomes with the privacy preserving techniques. In this presented work the proposed work is aimed to provide a noise based technique that preserves the privacy of data, during submission and processing. In addition of that the processed outcomes are also required to publish without harming the privacy of the end data owner. The proposed technique is a light weight and efficient technique for preventing privacy in multiparty data mining environment. Additionally the processed data or made decisions are publishable in the public domain. This section provides the overview of the proposed work, the next section provides the detailed discussion of the proposed system.

2.2 Methodology

The proposed system architecture is demonstrated in figure 2.1. In addition of the required system components are also explained.

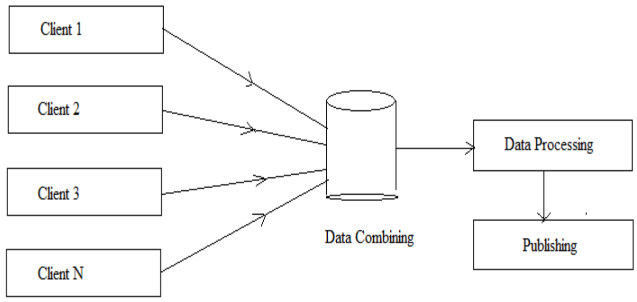

Figure 2.1 proposed system architecture

The above given diagram 2.1 contains four main components, the details about these components are given as:

2.2.1 Client

The privacy preserving data mining system requires different number of parties who are agreed to contribute the data for making the data mining based combined decisions. Therefore the data from different clients are collected to mine the data using data mining techniques. In this context to disclose the data before providing the data for mining need to be transformed to hide the actual values of end data owner.

2.2.2 Data combining

The different sources of data are needed to be process, to find the combine outcomes from the mined data. Therefore the contributed data from various sources is collected in this place to obtain a combine form of entire collected data.

2.2.3 Data processing

The data mining techniques required a central database to apply the different data mining techniques. In this phase the collected data from different data sources are processed using the data mining algorithm.

2.2.4 Data publishing

After processing the collected data using data mining and machine learning algorithms the data and the data mining outcomes are needed to be distributed to all the associated data parties. During this data outcome distribution it is required to regulate the privacy of the data.

In this context the proposed work is focused on designing a noise mixture model that accept the actual data as input and transform the available values to feel like actual values, but it is modified for public use. The figure 3.2 shows the process involved in this design.

The given diagram includes the client and a server additionally the remaining communication conduced among both the entities. First of the entire client who is agreed to collaborate the data with the other parties make a request to server for connection. At the server end, when server getting the connection request, it accept the connection and generate a random number as a session key. The session key is communicated to the just connected client. Client receives the session key and modifies the numerical values using the session key. In this context the following process is taken place as described in table 2.1.

Table 2.1 modification on data

Input: session key S, dataset D

Output: noise added dataset D

Process:

```
1.  $R_{n,m} = ReadData(D)$
2.  $for(i = 1; i \leq n; i++)$
    a.  $for(j = 1; j \leq m; j++)$
        i.  $if(R_{i,j} == number)$
            1.  $R_{i,j} = R_{i,j} + S$
        ii. End if
    b.  End for
3.  End for
4.  $D = reForm(R_{n,m})$
5.  $returnD$
```

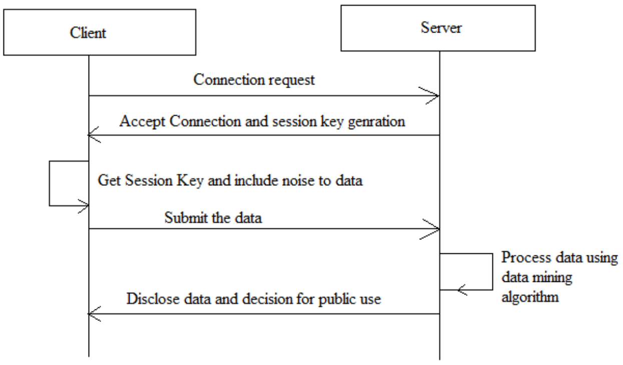
Client
Server
Connection request
Accept Connection and session key genration
Get Session Key and include noise to data
Submit the data
Process data using data mining algorithm
Disclose data and decision for public use

Figure 2.2 process involved

After adding the noise to the actual values on the dataset the dataset is reformed and that is ready to submit to the server. Server accepts the data from the client and process data using the data mining algorithm. In this experiment the data is processed using the ID3 decision tree algorithm. The decision tree algorithm produces the decision tree which is distributed to the all the connected parties for making the common decisions. Additionally the session keys are used with the obtained decisions to recover their own part of data. The ID3 decision tree algorithm is explained as:

ID3 is an easy decision tree learning algorithm residential by Ross Quinlan. The basic idea of ID3 algorithm is to accumulate the decision tree by employing a top-down, greedy search through the recognized sets to test every attribute at each tree node. In order to select the attribute that is the popular useful for classify a known sets; we establish a metric information increase.

To determine the majority approving way to systematize a learning set, what we need to do is to minimize the questions asked (i.e. minimizing the depth of the tree). Thus, we require several functions which can establish which query provides the popular balanced splitting. The in sequence gain metric is such a function.

2.3 Entropy

In order to describe information gain precisely, we need discussing entropy first. Let's suppose, with no loss of generalization, that the resultant decision tree categorize occurrence into two categories, we'll call them P (positive) and N (negative)

Given a set S, containing these optimistic and depressing targets, the entropy of S connected to this Boolean organization is:

P (positive): proportion of positive examples in S

P (negative): proportion of negative examples in S

2.4 Information Gain

As we talk about previous to, to minimize the conclusion tree depth, while we navigate the tree passageway, we require selecting the mostly favorable attribute for splitting the tree node, which we can basically imply that the attribute with the classically entropy reduce is the best choice.We elucidate information gain as the predictable decrease of entropy related to exacting feature when splitting a decision tree node.

The information gain, Gain(S, A) of an feature A,We can use this notion of gain to rank attribute and to build conclusion trees anywhere at every node is located the feature with greatest gain between the features not yet measured in the pathway from the root.The intention of this ordering is:

- To create small result trees so that records can be recognized behind simply a few decision tree opening.
- To match a hoped for simplicity of the procedure of decision worshippers

The ID3 algorithm can be summarized as follows:

- receive all unused feature and count their entropy regarding test samples
- Choose feature for which entropy is minimum (or, consistently, information gain is maximum)
- Make node containing that feature

The algorithm is as follows:

Table 2.2 ID3 Decision Tree

Input: Examples, Target Attribute, Attributes
Output: Decision Tree
Process: • Create a root node for the tree If all examples are positive, Return the single-node tree Root, with label = +. If all examples are negative, Return the single-node tree Root, with label = -. If number of forecast feature is empty, then Return the single node tree Root, with label = most ordinary value of the target feature in the examples. • Otherwise Begin • A = the feature that best classifies examples. • Decision Tree feature for Root = A. • For every possible value, v_i, of A, Add a new tree branch below Root, corresponding to the test A = v_i. Let Examples(v_i) be the subset of examples that have the value v_i for A • If Examples(v_i) is empty Then below this new branch add a leaf node with label = most common target value in the examples • Else beneath this new branch add the sub-tree ID3 (Examples(v_i), Target Attribute, Attributes – {A}) • End • Return Root

3. RESULTS ANALYSIS

This chapter provides the explanation of the conducted experiments. Additionally during the experiments the obtained performance is also provided in this chapter. Therefore the evaluated parameters and their line graphs are explained in this chapter.

3.1 Aim of Experiments

The main aim of the conducted experiments is to find the variations in the different performance parameters over the same data with the two different utilities. Therefore two different scenarios of experiments are demonstrated in the results analysis.

3.1.1 Utilizing the dataset with the ID3 and obtain the performance

In this scenario the ID3 decision tree algorithm is implemented and the combined dataset is processed using decision tree.

3.1.2 Utilizing the same dataset with add-on purpose based noise to hide the confidentiality of data

The deviation is tried to measure on modified attribute values with respect to the actual data and actual algorithm.

3.2 Accuracy

In order to measure the data utility of modified data the algorithm is compared with both kinds of data format. Therefore accuracy of algorithm is measured for differentiating the made loss before and after modification of data. The accuracy of a developed data models are computed using the following formula:

$$\text{accuracy} = \frac{\text{total correctly classified instances}}{\text{total instance to classify}} \text{X100}$$

According to the above given formula the accuracy is the ratio of total correctly recognized data instances and the total instances to recognize. Thus the accuracy is measurement of correctness of data mining algorithms recognition ability. The figure 3.1 and table 3.1 shows the obtained accuracy of the ID3 classifier for the pure dataset and dataset including the privacy based noise.

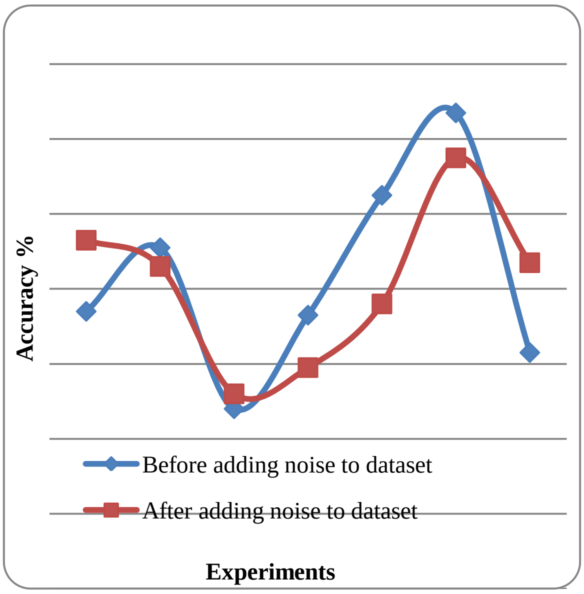

Figure 3.1 accuracy %

Table 3.1 accuracy %

Experiment number	**Before adding noise to dataset**	After adding noise to dataset
1	87.4	89.3
2	89.1	88.6
3	84.8	85.2
4	87.3	85.9
5	90.5	87.6
6	92.7	91.5
7	86.3	88.7

The table 3.1 contains the observations of the accuracy before and after noise inclusion with the dataset. Additionally their visual representation is given using figure 3.1. To demonstrate the accuracy of both the datasets the line graph contains the different number of experiments in X axis and the obtained corresponding accuracy percentage in Y axis. According to the obtained results the accuracy of the learning algorithm is fluctuating after including the noise on the data set, as compared to the without noise of data. But the variation on the accuracy is not much significant because it is varying between 1-5%.

3.3 Error Rate

The error rate of a classification algorithm describes the fraction of data instances are not recognized correctly. Thus it is the ratio between the total data samples are supplied for classification and the total incorrectly classified data instances. In this context the following formula is used:

$$\text{error rate} = \frac{\text{total incorrectly classified data}}{\text{total data passed for classification}}$$

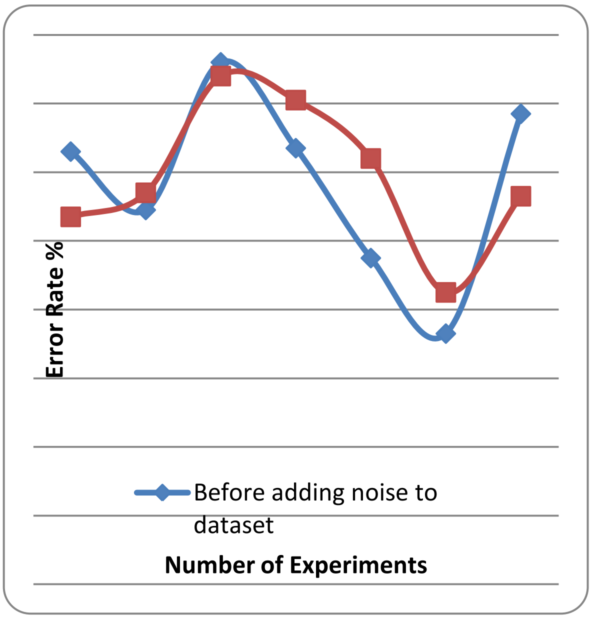

Figure 3.2 Error rate %

The error rate of ID3 algorithm for both the experimental scenarios is computed i.e. noisy and normal datasets. The error rate of algorithm is measured in terms of percentage. The Figure 3.2 shows different conducted experiments in X axis and Y axis contains the observed error rate values. The observation of error rate percentage is also reported in table 3.2. According to the obtained error rate, the performance of algorithm is not much varying after modification of data. Thus proposed methodology is acceptable for hiding security with low loss with the learning algorithms.

3.4 Memory Usages

The space complexity of algorithm is termed here as the memory usages of algorithm. That is calculated using the total amount of main memory assigned and the total free space during the process execution. In java technology the memory of particular process is computed using the following formula:

$$\text{memory usages} = \text{total allocated space} - \text{free space}$$

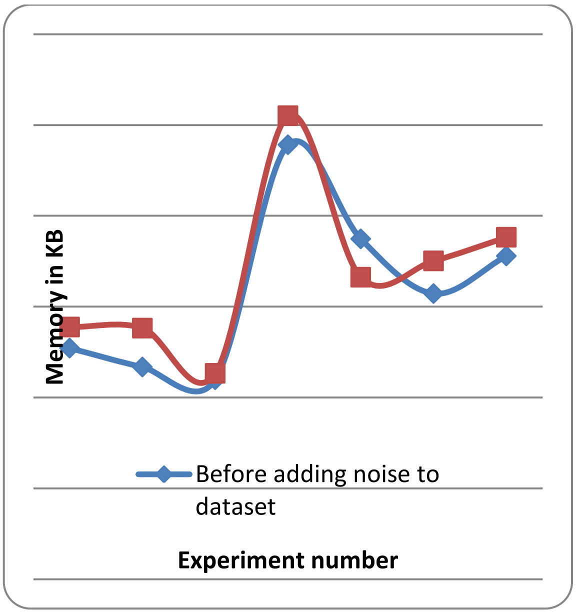

Figure 3.3 memory usages in KB

Table 3.3 memory usages in KB

Experiment number	**Before adding noise to dataset**	**After adding noise to dataset**
1	1272	1388
2	1168	1382
3	1098	1133
4	2391	2551
5	1873	1662
6	1572	1752
7	1779	1882

The experiment with the basic dataset and noise added data set is performed using the ID3 decision tree algorithm. The obtained memory usages values are reported in table 3.3. The measurement of memory usages is given here in terms of KB (kilobyte). The demonstrated results shows the noise added data take additional memory as compared to the normally classified data.

3.5 Time Consumption

It is the difference between producing input to the system and obtaining outcome from the system. Therefore the time required to process the algorithm for generation of outcome is known as time consumption of algorithm. That is computed using the following formula:

$$\text{time consume} = \text{algorithm stop time} - \text{algorithm start time}$$

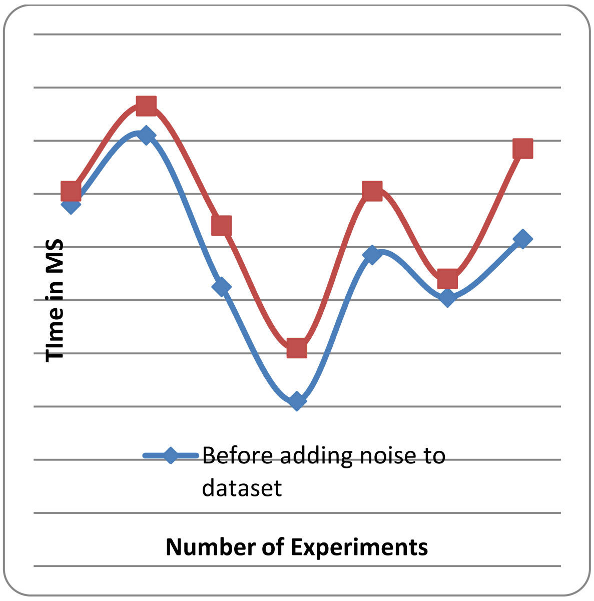

Figure 3.4 time consumption in MS

The time consumption of the ID3 algorithm for noisy data and normal data is reported using table 3.4 and figure 3.4. The table contains the values observed for processing the data in terms of MS (milliseconds). The given values in table are used to prepare line graph as given in figure 3.4. According to the given results the processing time of data is mostly similar in most of the experiments. Therefore the proposed privacy preserving data mining technique for utility based data model is acceptable.

Table 3.4 time consumption in MS

Experiment number	**Before adding noise to dataset**	**After adding noise to dataset**
1	856	861
2	882	893
3	825	848
4	782	802
5	837	861
6	821	828
7	843	877

4. CONCLUSION & FUTURE WORK

The proposed work is intended to provide security and preserve the data owner's privacy in a data mining environment. In this context a data mining model is presented which keep preserve the data confidentiality in a centralized database. Based on the system design and the experimental facts the conclusion of the work is described in this chapter, additionally the future extension of the work is also reported.

4.1 Conclusions

The data mining and its applications are increasing day by day. In addition of that it's extension in privacy based data mining applications also becoming popular. Therefore in this presented work the privacy preserving data mining is the focused domain of study. The main aim of the proposed work is to provide the following outcomes using the system:

1. Study of the privacy preserving data model to secure the end data owner's privacy against the misuse of the submitted data
2. The prepared privacy preserving data model can provides or publish the data for other experiment purpose without disclosing the data actual values but the utility of data remains consistent
3. To offer a data model that provide a simple and low cost implementable for managing security and privacy preserving by simply adding the noise on data

In order to handle above mentioned issues and expectations the proposed work is provide a data mining model on which user provide data to any valid data party. Additionally the data collector can use or share the data with other institutions to mining and obtain the decisions by combining the disclosed data with other parties shared data. After implementation of noise based model the authenticity or validity of application based on data is also measured with the help of a supervised learning algorithm, more specifically the ID3 decision tree algorithm. The results demonstrates the utility of data remains preserved after including noise on the data for preventing the privacy issues in the disclosed amount of data.

The implementation of the proposed system is performed on JAVA technology and using NetBeans IDE. Additionally the performance of the system measured in different performance parameters to justify the obtained results. The acquired performance based conclusion of system is reported in table 4.1.

Table 4.1 performance summary

S. No.	Parameters	Remark
1	Accuracy	The accuracy of the noise added data is fluctuating but is simply between 1-5 % from the actual data. Thus method is acceptable for the proposed work.
2	Error rate	Low error rate of the system is in control and measured influence of noise is between 1-5%.
3	Memory usage	The amount of memory usages is depends on the size of data passed for processing and generating the tree
4	Time consumption	The less time difference is measured between both the experimental scenario

According to given observations in the performance summary table 4.1. The accuracy of the classifiers is deviates with respect to basic data format. That is limited in a range between 1-5% and the amount of main memory increases with the size of data. Thus the proposed system is acceptable for real world application usages.

4.2 Future Work

The proposed work is aimed to provide a privacy preserving data mining model. That is used for publish the data confidential data for other application usages without disturbing the utility of data. Additionally to handle the confidentiality of data a small amount of random noise is included with the data. The proposed concept is promising for different privacy preserving data mining applications. The following future extension is proposed for future work:

1. The proposed work provides a simple linear manner to include the noise on data which is theoretically less secure therefore some advancement noise add-on manner is required.
2. The proposed work is extended with a real world application to provide understanding of system more appropriately
3. Currently simple rule based technique is used to evaluate the fluctuation on decision making by adding the noise in near future the work is evaluated on any opaque data model such as neural network and SVM for decision making utility measurement.

5. REFERENCES

[1] Ji-Young Lim, Woo-Cheol Kim, Hongchan Roh, Sanghyun Park, “A Practical Database Security Model Using PurposeBased Database Access Control and Group Concept”, Proceedings of the 2nd International Conference on Emerging Databases (EDB2010)

[2] Introduction to Data Mining and Knowledge Discovery, Dunham, M. H., Sridhar, S., “Data Mining: Introductory and Advanced Topics”, Pearson Education, New Delhi, 1st Edition, 2006.

[3] Phridvi Raj MSB., GuruRao CV (2013) Data mining – past, present and future – a typical survey on data streams. INTER-ENG Procedia Technology 12, pp. 255 – 263

[4] Veepu Uppal and Gunjan Chindwani, “An Empirical Study of Application of Data Mining Techniques in Library System”, International Journal of Computer Applications (IJCA), Volume 74– No.11, July 2013.

[5] Meenakshi and Geetika, “Survey on Classification Methods using WEKA", International Journal of Computer Applications, Vol. 86, No.18, January 2014.

[6] S. Archana and Dr. K. Elangovan, “Survey of Classification Techniques in Data Mining”, International Journal of Computer Science and Mobile Applications, Volume 2 Issue 2, February 2014.

[7] Han, Jiawei, Jian Pei, and Micheline Kamber, Data mining: concepts and techniques, Elsevier, 2011.

[8] Vahida Attar, Pradeep Sinha, and Kapil Wankhade, A fast and light classifier for data streams, Evolving Systems, 1:199–207, 2010.

[9] S. B. Kotsiantis, I. D. Zaharakis and P. E. Pintelas, “Machine learning: a review of classification and combining techniques”, Artificial Intelligent Rev (2006) 26:159–190

[10] “Fundamental Security Concepts”, https://cryptome.org/2013/09/infosecurity-cert.pdf

[11] Philip S. Antón, Robert H. Anderson, Richard Mesic, Michael Scheiern, “Finding and Fixing Vulnerabilities in Information Systems”, © Copyright 2003 RAND

[12] Anor F.A. Dafa-Alla, Eun Hee Kim, Keun Ho Ryu, Yong Jun Heo, “PRBAC: An Extended Role Based Access Control for Privacy preserving Data mining”, Proceedings of the Fourth Annual ACIS International Conference on Computer and Information Science ICIS’05 of IEEE, ’05.