An Approach for Preserving Privacy and Knowledge 
In Data Mining Applications

Alaa H Al-Hamami, Suhad Abu Shehab
1, 2 College of Computer Sciences and Informatics, Amman Arab University, Amman - Jordan
1 Alaa_hamami@sau.edu.jo, 2Suhad_cis@hotmail.com

ABSTRACT
To control data mining process for different users with different privileges, two approaches were designed to protect the Database and Data Mining usage. The first approach, privacy protection of individuals by adding a white Gaussian noise to selected columns in a database to be mined for an unauthorized user. The second approach, knowledge protection, is done by encrypting the result of data mining before it appears to the unauthorized users by using Rijndael algorithm. In this case the clear answer will appear to the authorized users only. The proposed design is providing a protection for privacy and knowledge in data mining. Data Mining is used to find hidden patterns “relations” between data in huge datasets. Anyone can use data mining, can get the new knowledge.

Keywords: Data Mining, Privacy, Knowledge, Protection, and database.

1. INTRODUCTION
Data mining has emerged as a significant technology for gaining knowledge from vast quantities of data. However, there has been growing concern that use of this technology is violating individual privacy. The goal of most data mining approaches is to develop generalized knowledge, rather than identify information about specific individuals. Data mining used to find hidden patterns “relations” between data in huge datasets [1]. The data mining result “Knowledge” is something new which did not exist before in the database. Data mining knowledge is an asset for a company so it must be protected [2]. This knowledge comes from personal information about individuals in the database; in this case privacy must be protected too [3]. Data mining is one of the business intelligence techniques because it can extract valuable knowledge from huge databases and find hidden relations between data [4]. In this research the most important topic is protecting both privacy and knowledge in data mining process from unauthorized users. At the same time, the authorized user can use data mining algorithm normally to get the knowledge to use in her/his work. The more complete and accurate the data, the better the data mining results. The existence of complete, comprehensive, and accurate datasets raises privacy issues regardless of their intended use. Data mining results represent a new type of “summary data”; ensuring privacy means showing that the results (e.g., a set of association rules or a classification model) do not inherently disclose individual information. This research is executed in two separated applications; the first one is the privacy protection by adding a white Gaussian noise for selected columns “data”, for unauthorized user, to be mined before the data mining algorithm runs. Data loads into a dataset before the noise added to it. The second approach deals with knowledge protection by encrypting the result of data mining algorithm before it appears to the unauthorized user using Rijndael Encryption algorithm. For data mining process the K-means algorithm is used.

2. SCIENTIFIC BACKGROUND
In 2006, Liu et. al proposed a technique called Data perturbation for Privacy-Preserving Data Mining (PPDM) to hide or remove the sensitive and privacy information before the dataset is published to public. However mining the removed privacy information could still gather some if the original knowledge in the dataset. The six methods in PPDM are data perturbation, data swapping, k-anonymity, secure multiparty computation, distributed data mining and rule hiding. The methods can be applied to protect privacy data from different data mining techniques. For example, data perturbation can be used to generate noise data to replace the original data fields so that the privacy and sensitive information in original data can be protected. [5]. In 2008, Chain et. al proposed an approach called Clustering against mining knowledge for Anti-Data Mining (ADM) to protect the privacy data and knowledge from disclosure. The protection done by adding noise data to change the knowledge in the original dataset in order to prevent knowledge from being mined. The main focus was to make use of the random seed in the clustering technique to add a certain number of noise data or noise fields to change the clustering structure of the original data. For example, the aim is to add a noise field to change the data character and then generating a different clustering result to meet the purpose of protecting the clustering result in the original data. The ADM doses not modify the original data. On the other hand, the privacy issue was also not considered. [6]. In 2010, Tung et. al combined Liu et. al's work in 2006 with Chain et. al's work in 2008 together to protect both knowledge and privacy. This protection technique is two phase. The first phase applies data perturbation to do PPDM privacy information protection in original dataset; its purpose is to hide the values of the original data and also to analyze the correct mining result. The second phase adds the noise data to the data set protected by first phase and to use ADM to protect knowledge. The purpose is to protect the correct clustering knowledge in the protected dataset from being

http://www.cisjournal.org

Vol. 4, No. 1 Jan 2013
ISSN 2079-8407
Journal of Emerging Trends in Computing and Information Sciences
©2009-2013 CIS Journal. All rights reserved.
mined. Therefore, the dataset protected by the proposed technique can not only protect its own privacy information but also prevent the clustering knowledge from being mined by the illegal users. [7].

3. PROBLEM
Data Mining is a technique used to extract a new knowledge from existing data. Anyone can use data mining, can get the new knowledge. At the same time we cannot prevent unauthorized user from using the database but prevent him from using data mining process. The aim is to use both database and Data Mining by the authorized users only.

4. THE PROPOSED SOLUTION
Two approaches are suggested to protect data mining privacy and knowledge. A database is used as a datasets and a Gaussian noise will be added to selected columns to protect the privacy of the authorized users. An encryption method is applied to the data mining algorithm result to prevent the unauthorized user (not allowed to use data mining) from gaining the clear result (knowledge). The two approaches are as the followings:

4.1 Privacy Protection in Data Mining
The first application is concerned about the privacy protection. See figure (1) steps for privacy protection in data mining.

4.2 Knowledge protection in data mining
In this approach the authorized and unauthorized user ask for the same query in the mining process. The system will go in different direction for each user. Figure (4) shows the steps.
In case of the unauthorized user query, the system will encrypt the knowledge “result of data mining process” before its appearance. The encryption process is done using Rijndael algorithm. The Rijndael encryption algorithm is designed to replace the aging DES algorithm. Like DES, it is a block cipher. It uses 128-bit, 192-bit or 256-bit keys. This implementation encrypts 128-bit blocks and key [9].

Figure (6). The authorized user will have useful knowledge.

![Fig 4](image-url): Steps for knowledge protection in data mining process.

Figure (3) shows Unauthorized User in data mining knowledge protection.

![Fig 5](image-url): Unauthorized User in Knowledge protection in data mining.

In both applications when the user is authorized, the result of data mining process will be clear as seen in Figure (6). The authorized user will have useful knowledge.

5. CONCLUSIONS

The knowledge in data mining process comes from data, and data contain personal information about individuals. Since knowledge comes from data, then it tells something about individuals. In this case the privacy of individuals is threatened if the unauthorized user got the data mining knowledge. In this research, the two applications provide protection for both privacy and knowledge. When the data appears with noise to the unauthorized user, the privacy is protected and the first goal in this research is achieved. The encrypted knowledge appears to the unauthorized user, in this way the second goal which is knowledge protection is achieved. While the authorized user work does not affected and got the knowledge he/she needs.

REFERENCE


