Privacy Preservation for Data Mining Security Issues

D. Ganesh,
Research Scholar
Bharathiyar University, Coimbatore

S.K. Mahendran, Ph.D
Director
SVS Institute of Computer Applications, Coimbatore

ABSTRACT

The development in data mining technology brings serious threat to the individualinformation. The objective of privacy preserving data mining (PPDM) is to safeguard the sensitive information contained in the data. The unwanted disclosure of the sensitive information may happen during the process of data mining results. In this paper we identify four different types of users involved in mining application i.e. data source provider, data receiver, data explorer and determiner decision maker].We differentiate each type of user's responsibilities and privacy concerns with respect to sensitive information. We'd like to provide useful insights into the study of privacy preserving data mining.

Keywords

Anonymization, DataMining, Sensitive Information, Privacy Preserving data Mining, Provenance.

1. INTRODUCTION

DATA mining has attracted more and more attention in recent years, probably because of the popularity of the"big data" concept. Data mining is the process of discoveringinteresting patterns and knowledge from large amounts ofdata [1]. As a highly application-driven discipline, data mininghas been successfully applied to many domains, such asbusiness intelligence, Web search, scientific discovery, digital libraries, etc. The Process of KDD.

The term "data mining" is often treated as a synonym foranother term "knowledge discovery from data" (KDD) which highlights the goal of the mining process. To obtain useful knowledge from data, the processes which are performed in aniterative wayi.e.Data preprocessing, Data transformation, Data mining and Pattern evaluation (Refer Fig. 1a).

1.1 Privacy Preserving Data Mining [PPDM]

Despite that the information discovered by data mining canbe very valuable to many applications; people have
Shown increasing concern about the other side of the coin, namelythe privacy threats posed by data mining [2]. The objective of PPDM is to safeguard sensitive informationfrom unsolicited or unsanctioned disclosure, and meanwhile, preserve the utility of the data. The consideration of PPDM is two-fold. First, sensitive raw data, such as individual's ID card number and cell phone number, should not be directly used for mining. Second, sensitive mining results whose disclosure will result in privacy violation should be excluded. After the pioneering work of [3], [4], numerous studies on PPDM have been conducted.

1.2 User Role-based Methodology

Current models and algorithms proposed for PPDM mainly focus on how to hide that sensitive information from certain mining operations. However, as depicted in Fig. 1a, the whole KDD processes involve multi-phase operations. In this paper, we investigate the privacy aspects of data mining by considering the whole knowledge-discovery process. We present an overview of the many approaches which can help to make proper use of sensitive data and protect the security of sensitive information discovered by data mining. We use the term "sensitive information" to refer to privileged or proprietary information that only certain people are allowed to see and that is therefore not accessible to everyone. If sensitive information is lost or used in any way other than intended, the result can be severe damage to the person or organization to which that information belongs. The term "sensitive data" refers to data from which sensitive information can be extracted. Throughout the paper, we consider the two terms "privacy" and "sensitive information" are interchangeable. In this paper, we develop a user-role based methodology to conduct the review of related studies. Based on the stage division in KDD process (see Fig. 1A), we can identify four different types of users, namely four user roles, in a typical data mining scenario (see Fig. 1B):

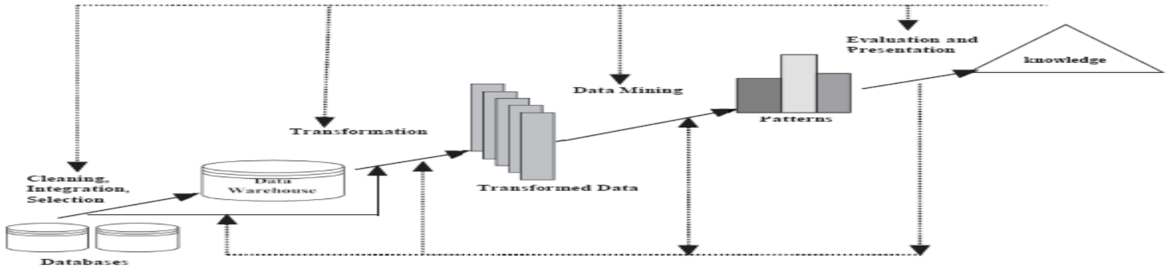


Fig: 1a An Overview of KDD Process

- Data Source provider: the user who owns some data that aredesired by the data mining task.
- Data Receiver: the users who collects data from data providers and then publish the data to thedata miner.
- Data Explorer: the user who performs data mining taskson the data.
- Determiner: the user who makes decisions based onthe data mining results in order to achieve certain goals.

1.3 Data Mining Scenario

In the data mining scenario depicted in Fig.1B, a user represents either a person or an organization. Also, one user can play multiple roles at once. For example the customer plays the role of data source provider, and the retailer plays the roles of data receiver, data explorerand determiner [who makes decision].By differentiating the four different user roles, we can explore the privacy issues in data mining in a principled way. All users care about the security of sensitive information, but each user role views the security issue from its own perspective. Here we briefly describe the privacy concerns of each user role. Detailed discussions will be presented in following sections.

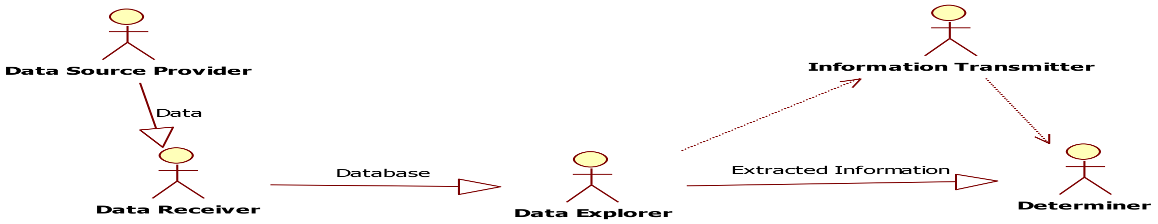


Fig:1b.User Role based methodology in Typical

Data Source Provider: The major concern of a data source provider iswhether he can control the sensitivity of the data he providesto others. On one hand, the provider should be able to makehis very private data, namely the data containing informationthat he does not want anyone else to know, inaccessible to the data receiver. On the other hand, if the data source provider has toprovide some data to the data receiver, he wants to hidehis sensitive information as much as possible and get enoughcompensation for the possible loss in privacy.

Data Receiver: The data collected from data source providers maycontain individual’s sensitive information. Directly releasingthe data to the data explorer will violate data source provider’s privacy, hence data modification is required. On the other hand, the datashould still be useful after modification; otherwise collectingthe data will be meaningless. Therefore, the major concern ofdata receiver is to guarantee that the modified data containno sensitive information but still preserve high utility.

Data Explorer: The data explorer applies mining algorithmsto the data provided by data receiver and he wishto extract useful information from data in a privacy-preservingmanner. As introduced in PPDM, it covers two typesof protections, namely the protection of the sensitive datathemselves and the protection of sensitive mining results. Withthe user role-based methodology proposed in this paper, weconsider the data receiver should take the major responsibilityof protecting sensitive data, while data explorer can focus on how to hide the sensitive mining results from untrusted parties.

Determiner: As shown in Fig. 1B, a determiner who makes decision canget the data mining results directly from the data explorer, orfrom some Information Transmitter. It is likely that the information transmitter changes the mining results intentionally orunintentionally, which may cause serious loss to the determiner. Therefore, what the determiner who makes the decision concerns is whetherthe mining results are credible.In addition to investigate the privacy-protection approachesadopted by each user role, in this paper we emphasize acommon type of approach, namely game theoretical approach,that can be applied to many problems involving privacyprotection in data mining. The rationality is that, in the datamining scenario, each user pursues high self-interests in termsof privacy preservation or data utility, and the interests ofdifferent users are correlated. Hence the interactions amongdifferent users can be modeled as a game. By using methodologiesfrom game theory, we can get useful implicationson how each user role should behavior in an attempt to solvehis privacy problems.

2. DATASOURCEPROVIDER

2.1 Concerns of Data Source Provider

A data source provider owns some data from which valuable information can be extracted. In the data mining scenario depicted in Fig. 1b, there are actually two types of data source providers: one refers to the data source provider who provides data to data receiver, and the other refers to the data receiver whocollect and provides data to data explorer. To differentiate the privacy protecting methods adopted by different user roles, here in this section, we restrict ourselves to the ordinary data source provider, the one who owns a relatively small amount of data which contain only information about himself. Data reporting information about an individualare often referred to as “Micro data” [5]. If a data provider reveals hisMicro data to the data receiver, his privacy might be comprised due to the unexpected data breach or exposure of sensitive information. Hence, the privacy concern of a data sourceprovider is whether he can take control over what kind of andhow much information other people can obtain from his data.To investigate the measures that the data source provider can adopt to protect privacy, we consider the following three situations:

1) If the data source provider considers his data to be very sensitive, that is, the data may reveal some information that he doesnot want anyone else to know, the data source provider can just refuseto provide such data, so that he can prevent his sensitivedata from being stolen by the data receiver who collects the data.

2) Realizing that his data are valuable to the data receiver (as well as the data explorer), thedata source provider needs to know how to negotiate with the data receiver, so that he will get enough compensation for anypossible loss in privacy.

3) If the data source provider can neither prevent the access to hissensitive data nor make a gainful deal with the data receiver,the data source provider can distort his data that will be fetched bythe data receiver, so that his original information cannot be easilydisclosed.

2.2 Approaches to Privacy Protection

Limit the Access: A data source provider provides his data tothe receiver in an active way or a passive way. By “active”we mean that the data source provider willingly opts in a surveyinitiated by the data receiver, or fill in some registration formsto create an account in a website. By "passive" we meanthat the data, which are generated by the source provider's routineactivities, are recorded by the data receiver, while the data source provider may even have no awareness of the revealing of hisdata. When the data source provider provides his data actively, he cansimply ignore the receiver's demand for the facts thathe consider very sensitive. If his data are passively provided tothe data receiver, the data source provider can take some measuresto limit the receiver's access to his sensitive data. Also, thedata sourceprovider can utilize various security tools that are developedfor Internet environment to protect his data. Many of thesecurity tools are designed as browser extensions for ease of use. Based on their basic functions, current security tools canbe categorized into the following three types:

Anti-tracking extensions: Knowing that valuable information can be removed from the data produced by user's online activities, Internet companies have a strong motivation to track the user's movements on the Internet. When browsing the Internet, a user can utilize an anti-tracking extension to block the trackers from collecting the cookies. Popular anti-tracking extensions include Disconnect, Do Not Track Me4, Ghostery, etc. A major technology used for anti-tracking is called Do Not Track (DNT), which enables users to opt out of tracking by websites they do not visit. A user's opt-out preference is signaled by an HTTP header field named DNT: if DNT=1, it means the user does not want to be tracked (opt out). Two U.S. researchers first created a prototype addon supporting DNT header for the Firefox web browser in 2009. Later, many web browsers have added support for DNT.DNT is not only a technology but also a policy framework forhow companies that receive the signal should respond. TheW3C Tracking Protection Working Group is now trying to standardize how websites should response to user's DNTrequest [6].

Advertisement and script blockers: This type of browserextensions can block advertisements on the sites, and killscripts and widgets that send the user's data to some unknownthird party. Example tools include Ad Block Plus6, NoScript7, FlashBlock8, etc.

Encryption Tools: To make sure a private online communicationbetween two parties cannot be intercepted by thirdparties, a user can utilize encryption tools, such as MailCloak9and TorChat10, to encrypt his emails, instant messages, or othertypes of web traffic. Also, a user can encrypt all of his internettraffic by using a VPN (virtual private network) service.There is no guarantee that one's sensitivedata can be completely kept out of the reach of treacherousdata collectors, making it a habit of clearing online traces andusing security tools does can help to reduce the risk of privacydisclosure.

Trade Privacy for Benefit: In some cases, the data sourceprovider needs to make a trade-off between the loss of privacyand the benefits brought by participating in data mining. Forexample, by analyzing a user's demographic information andbrowsing history, a shopping website can offer personalizedproduct recommendations to the user. The user's sensitivepreference may be disclosed but he can enjoy a better shopping experience. Driven by some benefits, e.g. a personalized service or monetary incentives, the data source provider may be willing to provide his sensitive data to a reliable data receiver, who promises the provider's sensitive informationwill not be revealed to an unauthorized third-party. If the provider is able to predict how much benefit he can get, he can logically decide what kind of and how many sensitive datato provide. For example, suppose a receiver asks the data source provider to provide information about his age, gender, occupation and annual salary. And the receiver tells the data source provider how much he would pay for each data item. If the data source provider considers salary to be his sensitive information, then based on the prices offered by the receiver, he chooses one of the following actions: i) not to report his salary, if he thinks the price is too low; ii) to report a fuzzy value of his salary, e.g. "less than 10,000 dollars", if he thinks the price is just acceptable; iii) to report an accurate value of his salary, if he thinks the price is high enough. For this example we can see that, both the privacy preference of data source provider and the incentives offered by data receiver will affect the data source provider's decision on his sensitive data. On the other hand, the data receiver can make profit from the data collected from data source providers, and the profit heavily depends on the quantity and quality of the data. Hence, data source providers' privacy preferences have great influence on data receiver's profit. The profit plays an important role when data receiver decides the incentives. That is to say, data receiver's decision on incentives is related to data source provider'sprivacy preferences. Therefore, if the data source provider wants to obtain satisfying benefits by "selling" his data to the data receiver, he needs to consider the effect of his decision on data receiver's benefits (even the data explorer's benefits), which will in turn affects the benefits he can get from the receiver. In the data-selling scenario, both the seller (i.e. the data source provider)and the buyer (i.e. the data receiver) want to get more benefits, thus the interaction between data source provider and data receiver can be formally analyzed by using game theory. Also, the sale of data can be treated as an auction, where mechanism design theory can be applied [7].

Provide False Data: As discussed above, a data source provider can take some measures to prevent data receiver from accessing his sensitive data. However, a disappointed fact that we have to admit is that no matter how hard they try, Internet users cannot completely stop the unwanted access to their personainformation. So instead of trying to limit the access, the data sourceprovider can provide false information to those treacherous data receiver. The following three methods can help anInternet user to falsify his data: Using "sock puppets" to hide one's true activities. A sock puppet is a false online identity though which a member of an Internet community speaks while pretending to beanother person, like a puppeteer manipulating a hand puppet.By using multiple sockpuppets, the data produced by oneindividual's activities will be deemed as data belonging todifferent individuals, assuming that the data receiver does nothave enough knowledge to relate different sock puppets to one specific individual. Using a fake identity to create phony information. In2012, Apple Inc. was assigned a patient called "Techniques to pollute electronic profiling" which can help to protectuser'sprivacy [8]. This patent discloses a method for polluting theinformation gathered by "network eavesdroppers" by making a false online identity of a principal agent, e.g. a service subscriber. A browser extension called Mask Me, which was release by the online privacy company Abine, Inc. in 2013, can help the user to create and manage aliases (or Masks) of these personal information. Users can use these aliases just like they normally do when such information is required, while the websites cannot get the real information. In this way, user's privacy is protected.

3. DATARECEIVER

3.1. Concerns of Data Receiver

A data receiver collects data from data source providers in order to support the subsequent data mining operations [see

Fig 1B]. The original data collected from data source providers usually contain sensitive information about individuals. If the data receiver doesn’t take sufficient precautions before releasing the data to public or data miners, those sensitive information may be disclosed, even though this is not the receiver’s original intention. It is necessary for the data receiver to modify the original data before releasing them to others. The data modification process adopted by data receiver, with the goal of preserving privacy and utility simultaneously, is usually called privacy preserving data publishing (PPDP).Extensive approaches to PPDP have been proposed in last decade. Fung et al. have systematically summarized and evaluated different approaches in their frequently cited survey [9]. Also, Wong and Fu have made a detailed review of studies on PPDP in their monograph. To differentiate with their work, in this paper us mainly focus on how PPDP is realized in two emerging applications, namely social networks and location-based services. To make our review more self-contained, in next subsection we will first briefly introduce some basics of PPDP, e.g. the privacy model, typical anonymization operations, information metrics, etc., and then we will review studies on social networks and location-based services respectively.

3.2 Approaches to Privacy Protection

Basics of PPDP: PPDP mainly studies anonymization approaches for publishing useful data while preserving privacy. The original data is assumed to be a private table consisting of multiple records. Each record consists of the following 4 types of attributes:

Identifier (ID): Attributes that can directly and uniquely identify an individual, such as name, ID number and mobile number.

Quasi-identifier (QID): Attributes that can be linked with external data to re-identify individual records, such as gender, age and zip code.

Sensitive Attribute (SA): Attributes that an individual wants to conceal, such as disease and salary.

Non-sensitive Attribute (NSA): Attributes other than ID, QID and SA. Before being published to others, the table is anonymized, that is, identifiers are removed and quasi-identifiers are modified. As a result, individual’s identity and sensitive attribute values can be hidden from adversaries. Typical privacy models [9] includesk-anonymity (for preventing record linkage), l-diversity (for preventing record linkage and attribute linkage), t-closeness (for preventingattribute linkage and probabilistic attack), epsilon-differential privacy (for preventing table linkage and probabilistic attack), etc.Among the many privacy models, k-anonymity and its variants are most widely used. The idea of k-anonymity is to modify the values of quasi-identifiers in original data table, so that every tuple in the anonymized table is indistinguishable from at least k − 1 other tuples along the quasi-identifiers. The anonymized table is called a k-anonymous table. Fig. 3 shows an example of 2-anonymity. Intuitionally, if a table satisfies k-anonymity and the adversary only knows the quasiidentifier values of the target individual, then the probability that the target’s record being identified by the adversary willnot exceed 1=k. To make the data table satisfy the requirement of a specified privacy model, one can apply the following anonymization operations [9]:

Generalization: This operation replaces some values with a parent value in the taxonomy of an attribute. Typical generalization schemes including full-domain generalization, sub tree generalization, multidimensionalgeneralization, etc.

Suppression: This operation replaces some values witha special value (e.g. a asterisk ‘*’), indicating that there placed values are not disclosed. Typical suppression schemes include record suppression, value suppression, cell suppression, etc.

Anatomization: This operation does not modify the quasi-identifieror the sensitive attribute, but de-associates the relationship between the two. Anatomization-based method releases the data on QID and the data on SA in two separate tables.

Permutation: This operation de-associates the relationship between a quasi-identifier and a numerical sensitive attribute by partitioning a set of data records into groups and shuffling their sensitive values within each group.

Perturbation: This operation replaces the original data values with some synthetic data values, so that the statistical information computed from the perturbed data does not differ significantly from the statistical information computed from the original data. Typical perturbation methods include adding noise, swapping data, and generating synthetic data. The anonymization operations will reduce the utility of data. The reduction of data utility is usually represented byinformation loss: higher information loss means lower utility of the anonymized data. Various metrics for measuring information loss have been proposed, such as minimal distortion, discernibility metric, the normalized average equivalence class size metric, weighted certainty penalty, information-theoretic metrics, etc. A fundamentalproblemPPDP is how to make a tradeoff between privacy and utility. Given the metrics of privacy preservation and Information loss, current PPDP algorithms usually take agreedy approach to achieve a proper trade-off: multiple tables, all of which satisfy the requirement of the specified privacymodel, are generated during the anonymization process; andthe algorithm outputs the one that minimizes the information loss.

Age	Sex	Zipcode	Disease
5	Female	12000	HIV
9	Male	14000	dyspepsia
6	Male	18000	dyspepsia
8	Male	19000	bronchitis
12	Female	21000	HIV
15	Female	22000	cancer
17	Female	26000	pneumonia
19	Male	27000	gastritis
21	Female	33000	Flu
24	Female	37000	pneumonia

Original table

Age	Sex	Zipcode	Disease
[1,10]	People	1****	HIV
[1,10]	People	1****	dyspepsia
[1,10]	People	1****	dyspepsia
[1,10]	People	1****	bronchitis
[11,20]	People	2****	HIV
[11,20]	People	2****	cancer
[11,20]	People	2****	pneumonia
[11,20]	People	2****	gastritis
[21,60]	People	3****	flu
[21,60]	People	3****	pneumonia

2-Anonimous Table

Fig : 3 An example of 2 – anonymity, where QID ={Age,Sex,Zipcode}

3.2 Privacy-Preserving Publishing of Social Network Data:

Social networks have gained great development in recentyears. Aiming at discovering interesting social patterns, socialnetwork analysis becomes more and more important. Tosupport the analysis, the company who runs a social networkapplication sometimes needs to publish its data to a thirdparty. However, even if the truthful identifiers of individualsare removed from the published data, which is referred toas naıve anonymized, publication of the network data maylead to exposures of sensitive information about individuals, such as one's intimate relationships with others. Therefore, thenetwork data need to be properly anonym zed before they arepublished. A social network is usually modeled as a graph, wherethe vertex represents an entity and the edge represents therelationship between two entities. Thus, PPDP in the contextof social networks mainly deals with anonymizing graph data, which is much more challenging than anonymizingrelational table data.[10] have identified the following threechallenges in social network data anonymization: First, modeling adversary's background knowledge aboutthe network is much harder. For relational data tables, a smallset of quasi-identifiers are used to define the attack models.While given the network data, various information, such asattributes of an entity and relationships between differententities, may be utilized by the adversary. Second, measuring the information loss in anonymizing socialnetwork data is harder than that in anonymizing relationaldata. It is difficult to determine whether the original networkand the anonymized network are different in certain properties of the network.Third, devising anonymization methods for social network data is much harder than that for relational data. Anonymizinga group of tuples in a relational table does not affect othertuples. However, when modifying a network, changing onevertex or edge may affect the rest of the network. Therefore "divide-and-conquer" methods, which are widely applied torelational data, cannot be applied to network data.

4. DATA EXPLORER

4.1. Concerns of Data Explorer

The primary concern of data explorer is how to prevent sensitive information from appearing in the mining results. To perform a privacy-preserving data mining.The data explorer usually needs to modify the data he got from the data receiver. As a result, the decline of data utility is inevitable. Similar to data receiver, the data miner also faces the privacy-utility trade-off problem. But in the context of PPDM, quantifications of privacy and utility are closely related to the mining algorithm employed by the data explorer who mines the data.

4.2 Approaches to Privacy Protection

Extensive PPDM approaches have been proposed. These approaches can be classified by different criteria, such as data distribution, data modification method, data mining algorithm, etc. Based on the distribution of data, PPDM approaches can be classified into two categories, namely approaches for centralized data mining and approaches for distributed data mining. Distributed data mining can be further categorizedinto data mining over horizontally partitioned data and data mining over vertically partitioned data. Based on the technique adopted for data modification, PPDM can be classified into perturbation-based, blocking-based, swapping based, etc.

4.2.1. Privacy-preserving association rule mining:

Association rule mining is one of the most important data mining tasks, which aims at finding interesting associations and correlation relationships among large sets of data items. A typical example of association rule mining is market basket analysis, which analyzes customer buying habits by finding associations between different items that customers place intheir "shopping baskets"[1]. These associations can help retailers develop better marketing strategies. The problem of mining association rules can be formalized as follows [1]. Given a set of items I = {i1; i2; · · · ;im}, and a set of transactions T = {t1; t2; · · · ; tn}, where each transaction consists of several items from I. An association rule is an implication of the form: A ⇒ B, where A ⊂ I, B ⊂ I, A/= ∅, B /= ∅, and A∩B/= ∅. The rule A ⇒ B holds in the transaction set T with support s, where s denotes the percentage of transactions in T that contain A∪B. The rule A⇒ B has confidence c in the transaction set T, where c is the percentage of transactions in Tcontaining A that also contain B.

4.2.2.Privacy-preserving Classification:

Classification is a form of data analysis that extracts models describing important data classes [1]. Data classification can be seen as a two-step process. In the first step, which is called learning step, a classification algorithm is employed to build a classifier (classification model) by analyzing a training set made up of tuples and their associated class labels. In the second step, the classifier is used for classification, i.e. predicting categorical class labels of new data. Typical classification model include decision tree, Bayesian model, support vector machine, etc.

4.2.3. Decision Tree:

A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) represents a class label [1]. Given a tuple X, the attribute values of the tuple aretested against the decision tree. A path is traced from the root to a leaf node which holds the class prediction for the tuple.

4.2.4. Naive Bayesian classification:

Naive Bayesian classification is based on Bayes' theorem of posterior probability. It assumes that the effect of an attribute value on a given class is independent of the values of other attributes. Given a tuple, a Bayesian classifier can predict the probability that the tuplebelongs to a particular class.Vaidya et al study the privacy-preserving classification problem in a distributed scenario, where multi-parties collaborate to develop a classification model, but no one wants to disclose its data to others [18]. Based on previous studies on secure multi-party computation, they propose different protocols to learn naive Bayesian classification models from vertically partitioned or horizontally partitioned data.

4.2.5. Support Vector Machine.

Support Vector Machine (SVM) is widely used in classification [1]. SVM uses a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension, SVM searches for a linear optimal separating hyper plane (i.e. a "decision boundary"separating tuples of one class from another), by using support vectors and margins (defined by the support vectors).[31] propose a solution for constructing a global SVM classification model from data distributed at multiple parties, without disclosing the data of each party [17]. They consider the kernel matrix, which is the central structure in a SVM, to be an intermediate profile that does not disclose anyinformation on local data but can generate the global model.

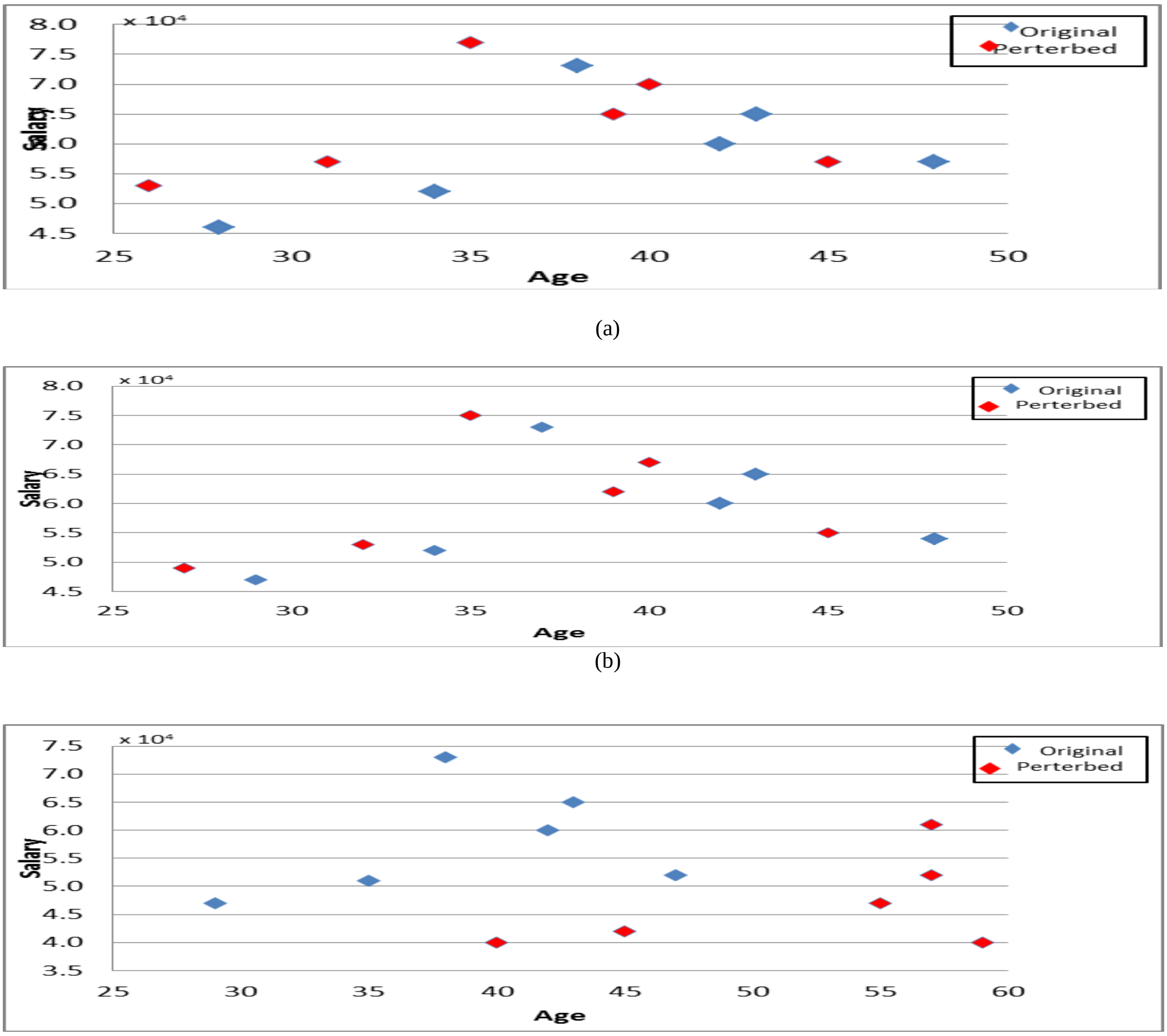
5. DETERMINER

5.1. Concerns of Determiner

The ultimate goal of data mining is to provide useful information to the determiner, so that the determiner can choose a better way to achieve his objective, such as increasing sales of products or making correct diagnoses of diseases. At a first glance, it seems that the determiner has no responsibility for protecting privacy, since we usually interpret privacy as sensitive information about the original data owners (i.e. data source providers). Generally, the data explorer, the data receiver and the data source provider himself are considered to be responsible for the safety of privacy. However, if we look at the privacy issue from a wider perspective, we can see that the determiner also has his own privacy concerns. The data mining results provided by the data explorer are of high importance to the determiner. If the results are disclosed to someone else, e.g. a competing company, the determiner may suffer a loss. That is to say, from the perspective of determiner, the data mining results are sensitive information. On the other hand, if the determiner does not get the data mining results directly from the data explorer, but from someone else which we called information transmitter, the determiner should be skeptical about the credibility of the results, in case that the results have been distorted.

5.2. Approaches to Privacy Protection

To deal with the first privacy issue proposed above, i.e. to prevent unwanted disclosure of sensitive mining results, usually the determiner has to resort to legal measures. For example, making a contract with the data explorer to forbid the miner from disclosing the mining results to a third party. Tohandle the second issue, i.e. to determine whether the received information can be trusted, the determiner can utilize methodologies from data provenance, credibility analysis ofWeb information, or other related research fields. In the rest part of this section, we will first briefly review the studies on data provenance and web information credibility, and then present a preliminary discussion about how these studies can help to analyze the credibility of data mining results.



(a)

(b)

(c)

Fig. 5.Examples of geometric data transformation.

Red circles represent original data and blue circles represent perturbed data. Data are perturbed in 3 ways: (a) translation; (b) scaling; (c) rotation.

5.*2.1.* ***Data Provenance***:If the determiner does not get the data mining results directly from the data explorer, he would want to know how the results are delivered to him and what kind of modification may have been applied to the results, so that he can determine whether the results can be trusted. This is why “provenance” is needed. The term provenance originally refers to the chronology of the ownership, custody or location of a historical object. In information science, a piece of data is treated as the historical object, and data provenance refers to the information that helps determine the derivationhistory of the data, starting from the original source [11].Twokinds of information can be found in the provenance of thedata: the ancestral data from which current data evolved andthe transformations applied to ancestral data that helped toproduce current data. With such information, people can betterunderstand the data and judge the credibility of the data. Since 1990s, data provenance has been extensively studiedin the fields of databases and workflows. Several surveys are now available [11]. Present taxonomyof data provenance techniques. The following five aspects areused to capture the characteristics of a provenance system: Application of provenance. Provenance systems may beconstructed to support a number of uses, such as estimatedata quality and data reliability, trace the audit trail ofdata, and repeat the derivation of data, etc.Subject of provenance. Provenance information can becollected about different resources present in the dataprocessing system and at various levels of detail.

Representation of provenance: There are mainly two typesof methods to represent provenance information, oneis annotation and the other is inversion. The annotationmethod uses metadata, which comprise of the derivationhistory of the data, as annotations and descriptions aboutsources data and processes. The inversion method usesthe property by which some derivations can be invertedto find the input data supplied to derive the output data.

Provenance dissemination:They summarize the key components of a provenancemanagement solution, discuss applications for workflowprovenance, and outline a few open problems for databaserelatedresearch. As Internet becomes a major platform for informationsharing, provenance of Internet information has attracted someattention. Researchers have developed approaches for informationprovenance in semantic web and social media.Hartig proposes a provenance model thatcaptures both the information about web-based data accessand information about the creation of data [12]. In this model, an ontology-based vocabulary is developed to describe theprovenance information. Moreau reviews research issuesrelated to tracking provenance in semantic web from the followingfour aspects: publishing provenance on the web; usingsemantic web technologies to facilitate provenance acquisition, representation, and reasoning; tracking the provenance of RDF (resource description framework)-based information; trackingthe provenance of inferred knowledge [13]. Barbier and Liu study the information provenance problem in social media.They model the social network as a directed graph G (V; E; p), whereVis the node set and E is the edge set. Each node inthe graph represents an entity and each directed edge representsthe direction of information propagation [14]. An informationpropagation probability p is attached to each edge, Based on the model they define. The information provenance problems follows: given a directed graph G(V;E; p), with knownterminals T ⊆V , and a positive integer constant k ∈Z+,identify the sources S ⊆V , such that |S| ≤ k, and U (S; T)is maximized. The function U (S; T) estimates the utility ofinformation propagation which starts from the sources S andstops at the terminals T. To solve this provenance problem, one can leverage the unique features of social networks, e.g.user profiles, user interactions, spatial or temporal information, etc. Two approaches are developed to seek the provenance of information. One approach utilizes the network information to directly seek the provenance of information, and the other approach aims at finding the reverse flows of informationpropagation. There are still many problems tobe explored in future study.[17]

Web Information Credibility: Because of the lack ofpublishing barriers, the low cost of dissemination and the laxcontrol of quality, credibility of web information has becomea serious issue. Identify the following fivecriteria that can be employed by Internet users to differentiatefalse information from the truth [15].
Authority: the real author of false information is usuallyunclear.
Accuracy: false information does not contain accuratedata or approved facts.
Objectivity: false information is often prejudicial.
Currency: for false information, the data about its source, time and place of its origin is incomplete, out of date, ormissing.
Coverage: false information usually contains no effectivelinks to other information online. TheCurrent research usually treats rumoridentification as a classification problem, thus the following two issues are involved: Preparation of training data set[16]. Current studies usuallytake rumors that have been confirmed by authoritiesas positive training samples.[18].
Considering the huge amountof messages in microblogging networks, such trainingsamples are far from enough to train a good classifier.Building a large benchmark data set of rumors is in urgentneed feature selection. Various kinds of features can be used to characterize the micro blogging messagesit is still quite difficult to automatically identifying false information on the Internet. It is necessary to incorporate methodologies from multiple disciplines, such as nature language processing, data mining, machine learning, social networking analysis, and information provenance, into the identification procedure.

6. CONCLUSION

How to protect sensitive information from the security threats brought by data mining has become a hot topic in recent years. In this paper we review the privacy issues related to data mining by using a user-role based methodology. We differentiate four different user roles that are commonly involved in data mining applications, i.e. data source provider, data receiver, data explorer and determiner. Each user rolehas its own privacy concerns; hence the privacy-preserving approaches adopted by one user role are generally different from those adopted by others. For data source provider, his privacy-preserving objective is toeffectively control the amount of sensitive data revealedto others. To achieve this goal, he can utilize securitytools to limit other’s access to his data, sell his data atauction to get enough compensation for privacy loss, orfalsify his data to hide his true identity. For data receiver, his privacy-preserving objective is torelease useful data to data miners without disclosingdata source provider’s identities and sensitive information aboutthem. To achieve this goal, he needs to develop properprivacy models to quantify the possible loss of privacyunder different attacks, and apply anonymizationtechniquesto the data. For data explorer, his privacy-preserving objective is to getcorrect data mining results while keep sensitive informationundisclosed either in the process of data mining orin the mining results. To achieve this goal, he can choosea proper method to modify the data before certain miningalgorithms are applied to, or utilize secure computationprotocols to ensure the safety of private data and sensitiveinformation contained in the learned model. For determiner, his privacy-preserving objective is tomake a correct judgment about the credibility of the data mining results he's got. To achieve this goal, he can utilize e-provenance techniques to trace back the history of thereceived information, or build classifier to discriminatetrue information from false information. To achieve the privacy-preserving goals of different usersroles, various methods from different research fields are required. We have reviewed recent progress in related studies, and discussed problems waiting to be further investigated.We hope that the review presented in this paper can offer researcherdifferent insights into the issue of privacy-preservingdata mining, and promote the exploration of new solutions tothe security of sensitive information.

7. REFERENCES

[1] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques. Morgan kaufmann, 2006.

[2] L. Brankovic and V. Estivill-Castro, "Privacy issues in knowledge discovery and data mining," in Australian institute of computer ethics conference, 1999, pp. 89–99.

[3] R. Agrawal and R. Srikant, "Privacy-preserving data mining," SIGMOD Rec., vol. 29, no. 2, pp. 439–450, 2000.

[4] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in Advances in Cryptology CRYPTO 2000. Springer, 2000, pp. 36–54.

[5] V. Ciriani, S. D. C. Di Vimercati, S. Forest, and P. Samarati, "Microdata protection," in Secure Data Management in Decentralized Systems. Springer, 2007, pp. 291–321.

[6] R. T. Fielding and D. Singer, "Tracking preference expression (dnt).w3c working draft," 2014.[Online]. Available: http ://www.w3.org/TR/2014/WD- tracking-dnt –20140128

[7] D. C. Parkes, "Classic mechanism design," Iterative Combinatorial Auctions: Achieving Economic and Computational Efficiency. Ph. D. dissertation, University of Pennsylvania, 2001.

[8] S. Carter, "Techniques to pollute electronic profiling," Apr. 26 2007, us Patent App. 11/257,614. [Online]. Available:

[9] B. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Computing Surveys (CSUR), vol. 42, no. 4, p. 14, 2010.

[10] B. Zhou, J. Pei, and W. Luk, "A brief survey on anonymization techniques for privacy preserving publishing of social network data," ACM SIGKDD Explorations Newsletter, vol. 10, no. 2, pp. 12–22,2008.

[11] Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science," ACM Sigmod Record, vol. 34, no. 3, pp. 31–36, 2005.

[12] O. Hartig, "Provenance information in the web of data." in LDOW, 2009.

[13] L. Moreau, "The foundations for provenance on the web," Foundations and Trends in Web Science, vol. 2, no. 2–3, pp. 99–241, 2010.

[14] G. Barbier, Z. Feng, P. Gundecha, and H. Liu, "Provenance data in social media," Synthesis Lectures on Data Mining and Knowledge Discovery, vol. 4, no. 1, pp. 1–84, 2013.

[15] M. Tudjman and N. Mikelic, "Information science: Science about information, misinformation and disinformation," Proceedings of Informing Science+ Information Technology Education, pp. 1513–1527, 2003.

[16] M. J. Metzger, "Making sense of credibility on the web: Models for evaluating online information and recommendations for futureresearch," Journal of the American Society for Information Science and Technology, vol. 58, no. 13, pp. 2078–2091, 2007. [

[17] J. Vaidya, H. Yu, and X. Jiang, "Privacy-preserving svm classification," Knowledge and Information Systems, vol. 14, no. 2, pp. 161–178, 2008.

[18] J. Vaidya, M. Kantarcıo˘glu, and C. Clifton, "Privacy-preserving naïve bayes classification," The VLDB Journal The International Journal on Very Large Data Bases, vol. 17, no. 4, pp. 879–898, 2008.