IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 23, 2018, accepted May 29, 2018, date of publication June 5, 2018, date of current version June 26, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2844158

Educational Sensitive Information Retrieval: Analysis, Application, and Optimization

XIYUAN WANG[1] AND YONG WANG[2]
[1]School of Humanities, Xidian University, Xi'an 710071, China
[2]School of Cyber Engineering, Xidian University, Xi'an 710071, China

Corresponding author: Yong Wang (wangyong@mail.xidian.edu.cn)

This work was supported in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2017ZDJC-28, in part by the Science Research Plan in Shaanxi Province of China under Grant 2016GY-085, in part by the Shaanxi Province Science Project under Grant 2015KRM086, in part by the Shaanxi Province Science Fund under Grant 2015R008, in part by the Fundamental Research Funds for the Central Universities under Grant RW160104, and in part by the 111 Project under Grant B16037.

ABSTRACT Social networking and cloud services will collect a large number of high-dimensional and complex data for third party statistical analysis and data mining. Although data analysis is beneficial to users and external parties, they constitute a serious privacy risk disclosure of user sensitive information. First of all, the authors use random forest, decision tree, and support vector machines methods with a discriminant component model to analyze the data set of the new curriculum reform in Shaanxi Province and make an accurate prediction of the impact on bad habits of students' comprehensive performance. Second, though various kinds of data transformation for privacy protection have been achieved and applied, there is little research designing two preference classification tasks simultaneously. In order to prevent the privacy leakage of sensitive information in the data analysis, the data space projection algorithm based on multiclass discriminant is used to deal with the linear and nonlinear. Clearly, the role of sensitive task and insensitive task can be exchanged. The goal is to preserve the statistical properties of one preference classification as far as possible and to realize the data security of the other preference classification.

INDEX TERMS Sensitive information, discriminant component, analysis information, extraction classification.

I. INTRODUCTION

The rapid growth of data generated in our society presents important challenges to database system design. In addition to the need for efficient information systems to manage high dimensional and complex data generated by multiple devices, a large amount of data is creating new privacy and social concerns. In fact, as the amount of data acquisitions are greatly increased in the past few decades, many users of personal devices (such as smart mobile phone, smart watch, wristbands) created, is a large part of the user specific data. According to Patrick [1] technical report, the typical American office workers produce about 1800000 trillion bytes of data per year (5000 MB / day). User-generated data is usually collected by third parties to improve the user experience, providing personalized service for the secondary analysis. Although data analysis is beneficial to users and external parties, they pose a serious privacy risk and leak sensitive information from users. With the continuous growth of personal data, the risk of information disclosure is growing.

Some work [2], [3] surrounds the privacy needs of users and the focus is on designing effective privacy protection solutions for data analysis. Sharing user data can cause serious privacy problems because it contains a behavioral pattern that may reveal sensitive personal information when a third party mines data. For a recent target event, the store can view the history of personal purchases to predict whether a customer has changed purchase behavior, it may show that a customer experiences a major life event such as pregnancy. Therefore, the design of privacy protection is crucial while protecting the user's sensitive information by enabling the application of the technology to obtain meaningful results.

The other work [4] is designed to address social needs that are emerging in our society. Rich data and user interaction for modern applications (such as social networks, location-based services) create opportunities for users to have a new social role. The increasing popularity of social networks, such as Facebook, Google+, and rich user interaction in spatial databases, has created new opportunities and challenges for users. Users use social network to share their location, as well as their preferences, such as restaurants, museums and other recent works [5]–[7] as the user's social information and road network conditions. (such as traffic route recommendation process). For example, Yoon *et al.* [7] makes use of such social information by integrating the similarity between user profiles to provide a better route recommendation. Although social information determines the user's line usage, the query task is still for a single user. A new problem emerges from such an environment, and how to determine a set of routes that users can share. The problem of finding a preferred path for multiple users is solved [8], which minimizes the distance from the whole group. Despite the efforts to establish model of user interests, this solution does not provide the required flexibility when users have different preferences or mobility constraints.

Privacy leaking can also occur when data is deliberately disclosed in an anonymous manner. This shows that security and anonymity may not be enough to provide privacy protection. For this reason, before the data is shared to the public space, a new data privacy protection mechanism is needed to protect the privacy of the user. The core of this idea is to perform some form of transformation on the original data before publishing the data, and reduce the data dimension by projecting the data into a suitable linear subspace. Although some works have been done on data projection transformations [9], [10], almost all of them focus only on improving the information cognition. However, when the current privacy issue becomes more important, privacy reservation should be incorporated into the design of the algorithm in order for the projection transformation to be applied to the actual system.

There are two main contributions in this paper. First, data mining and discriminant analysis are used to determine which attributes play a more important role in information extraction. By exploring the potential connections of students' information, the main factors that affect students' performance are obtained. The research results will help improve the educational concept and management level. Second, the direct use of raw data in data mining will be faced with the risk of privacy leakage. The main focus of this part is the classification problem. Two classification problems are defined: one is insensitive information task; the other is sensitive information task. The design goal is to produce high recognition accuracy for insensitive information tasks while allowing low privacy accuracy.

II. SENSITIVE INFORMATION MODEL

For achieving privacy vulnerabilities, it is necessary to define what leaks privacy for specific data set. There are different levels of privacy. Certain levels are determined by the personal or by the privacy policies or laws. An optimal privacy model result is defined by the Dalenius [11]. Fung *et al.* [12] believes that even if the attacker obtains any background knowledge from other sources, he should not allow illegal attackers to use the published data to learn any other target knowledge. A secure public data mining model is shown in figure 1.

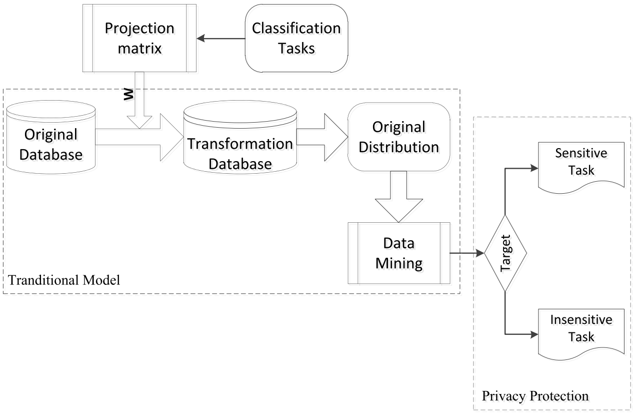

FIGURE 1. Structure of Information extraction.

Considers two categories of privacy models:

1) There is a threat to privacy when an attacker is able to establish a relationship between the sensitive attributes in the published data sheet and the record owner.

2) The published data should provide as less useful information as possible to the attacker besides background knowledge.

When an attacker learns information about personal sensitive attributes, a privacy model based on sensitive attributes is needed. In other words, it is impossible to gain additional knowledge through an attacker's observation of the camouflaged data set. Micro data privacy can be understood as the prevention of member exposure, that is, the attacker does not know whether a person is included in the database.

Once you decide to protect data privacy threats, it is time to decide what kind of privacy technologies to use to implement an algorithm based on these technologies.

III. SENSITIVE INFORMATION HIDING TECHNIQUES

The concept of disguising data sets is called anonymization. The implementation of the original data that will satisfy the specified privacy requirements results in the publication of the modified data set. There are four general categories for anonymization [13], 1) generalization and suppression, 2) perturbation, 3) cryptographic, and 4) Other techniques. Most of the techniques created to protect privacy belong to one or more of these categories, and have some drawbacks.

A. GENERALIZATION AND SUPPRESSION

Many researchers study hiding technology how to misconstrue the data. Typical techniques are generalization and suppression. Generalization can be used to cover a sub-range of value to replace the exact value. For example, 5 may become 3∼7 or replacement symbol with a more general term. Or "house number" becomes "street number". Suppression can use symbols instead of exact values, such as triangle or star.

The privacy model based on generalization and suppression methods can be extended to other fields such as k-anonymity, l-diversity and t-closeness. K-anonymity [14] requires that each quasi identifier attribute in an anonymous table is indistinguishable only by relying on the other at least $k - 1$ records in the data set. Based on generalization and suppression, the information about the k-anonymous table is real, which is essentially different from the traditional privacy preserving techniques such as data exchange and additive noise. In practice, if the attacker has the background knowledge in the field, k-anonymity cannot guarantee the privacy of users. First, the owner of the database may be difficult to determine which attributes are available or unavailable in the external table. The second restriction is that the k-anonymous model always supposes that there is a specific pattern of incursion, whereas in the actual scenario the attacker has no reason not to try other methods.

Sowmyarani *et al.* [15] proposed l-diversity to elaborate the shortcoming of k-anonymity by introducing a collection of examples whose quasi-identifier(QI) values are the same. With regard to each sensitive attribute value, we can receive at least l different values. Latest research shows that l-diversity due to poor robustness leads to bias. A similar attack is not enough to prevent it from being disclosed. Zhang *et al.* [16] improved t-closeness to focus on holding the distance between the distributions of sensitive attributes in a quasi-identifier group while no more than the threshold. It has adopted the k-nearest neighbor algorithm to choose similar QI values, and it can preserve privacy of sensitive data well meanwhile maintaining high data utility.

B. PERTURBATION

The perturbation method makes it impossible for a data server to recover accurate records or learning, and this limitation poses some challenges. Because the method does not rebuild the initial data value and just the distribution, it is necessary to develop a new algorithm to mine the underlying data using these reconstructed distributions. This means that, for example, classification, association rule mining, or clustering, it is necessary to develop a new distributed data mining algorithm to deal with each individual data problem. An efficient and effective data perturbation method is proposed by Ankleshwaria [17] that aims to protect privacy of sensitive attribute and getting data clustering with minimum information loss. It focuses on data perturbation by geometric transformation and noise addition to preserve privacy of sensitive attributes. They extended existing MOA framework in which, each tuple of data stream is independently treated. All distributed data mining algorithms implicitly handle each dimension independently. In general, a lot of related information about data mining algorithm is hidden in the correlation between attributes. Because perturbation method is used to deal with different attributes independently, this means that the inherent defects of the distributed data mining algorithms in the multi-dimensional records are the existence of implicit information loss. In addition, suppose that the attacker accesses multiple independent samples from the initial data with the same distribution. In these circumstances, principal component analysis can reconstruct the transformation data from the original to current. The attacker can estimate the transformation matrix from the original data to the current data, and then cancel the perturbation imposed on the original data.

C. CRYPTOGRAPHIC METHODS

Another way to hide sensitive information is to develop cryptographic technology. The solution has become very popular [18] for two main reasons: First of all, cryptography provides a suitable privacy model and includes complete proofs and exact quantization methods. The purpose of encryption is to make the original information after encryption becomes unreadable, only the user who holds the decryption key to recover the encrypted text. After encryption, because the structure of the original message is destroyed to form a ciphertext, the resulting ciphertext looks completely random and cannot obtain valuable information. Secondly, a data mining algorithm based on encryption algorithm toolbox to realize sensitive information security. Cryptography does not protect the output of computation [19]. Instead, it simply keeps from privacy disclosure in the computing process. Therefore, it lacks a complete solution to deal with privacy preserving data mining problems.

D. OTHER TECHNIQUES

The basic idea of randomized response [20] is that no matter whether the data from user is true or false, the central location will not have a better prediction probability than the threshold. The basic characteristic is that the respondents adopt a random answer to the questions investigated to avoid the direct response to the problem without any protection. Therefore, both the privacy and the privacy of the respondents are protected, and the real information is obtained. Another technique in [20] is to construct constrained clusters in a data set and generate pseudo data by means of condensation. This method compresses the data into a predefined number of packets, and ensures that each packet has certain statistical properties. The size of each packet is at least k to indicate the ability of privacy protection. The higher the level of each packet, the stronger the ability to protect privacy. It should be noted that a large number of data records are compressed into a statistical packet, thus losing a lot of useful information.

IV. PUBLISHED DATA MINING AND DISCRIMINANT COMPONENT ANALYSIS

Shaanxi province has implemented of the new curriculum reform for five years since 2012. In order to comprehensively understand the comprehensive quality of ordinary middle school students after the new curriculum reform in Shaanxi Province, the government selected three middle schools in southern Shaanxi, northern Shaanxi and Central Shaanxi area as the sampling objects. Through the school statistics and student survey, the present situations of comprehensive quality of general middle school students were investigated [21].

The government hopes to improve the quality of education and the level of teaching management on the basis of scientific analysis of existing problems.

Data preprocessing is performed according to the data source before we extract useful data information. Because of their large amount of data (usually several GB or more) and possibly from multiple heterogeneous data sources, real world data is vulnerable to noise, loss, and mismatched data. Poor quality data will result in poor quality mining results. We use data mining preprocessing algorithms [22] to implement data filtering to eliminate noise and correct inconsistencies in data, and merge three different datasets by using data integration.

There are many reasons why the data is inaccurate, so the dataset has an incorrect attribute value. The use of the data collection instrument may not be correct. An artificial or computer error may occur when data is entered. When a user does not wish to submit personally sensitive information (for example, by selecting the default birth year as "2000"), the user can intentionally submit incorrect data values for the required fields. This is considered to be disguised as lost data. The objective of this section is to use a data set that is collected and stored by the education department of Shaanxi Province to extract valuable information in the data and to explore the relationship between students' performance and other social factors. This dataset consists of 744 instances and each contains 32 attribute features as shown in table 1.

The initial data is represented mathematically as M-dimensional vector $\mathbf{x}$ i.e. $\mathbf{x} = \left[x^1, \cdots, x^m \cdots, x^M\right]^T$, where m represents the number of components extracted from the original data space. For given set of training vectors: $\mathbf{X} = \left\{\mathbf{x}_1, \cdots, \mathbf{x}_n \cdots, \mathbf{x}_N\right\}$ map to a given target y_n. In the classification model, the percentage of correct classification (PCC) is usually used to evaluate the performance. $R(i) = 1$ if $y_i = \tilde{y}_i$, else $R(i) = 0$. So we obtain $PCC = \frac{1}{N}\sum_{i=1}^{N} R(i)$.

In this paper, we use the core curriculum of mathematics and Chinese as the basis for the evaluation of student performance. According to the above agreement, we designed two kinds of academic record classification models:

1> Binary label classification—fail if Grade3<2, else pass;

2> Five level classification—excellent, good, adequate, pass, fail (the Grade3 value)

According to the model, we can get the figure 2.

All reports in this section are simulated using Rapid-Miner software. Rapid-Miner has a lot of data mining algorithm and analysis function. It is often used to solve various problems, such as marketing response rate, customer segmentation, customer loyalty and lifetime value, asset maintenance, resource planning, predictive maintenance, quality management, social media monitoring and sentiment analysis and typical commercial case. We can use zero code operation of client software and graphical development environment to

TABLE 1. Students' related attributes.

Target	Attribute	Description	Value
Insensitive Data	school	student's school	numeric: from 1 to 3
	sex	student's sex	binary: 'F' - female or 'M' - male
	age	student's age	numeric: from 13 to18
	famsize	family size	binary: 'L' - less or equal to 3 or 'G' - greater than 3
	Mjob	mother's Occupation	‘Civil servant’, ’Teacher’, ’employee ’, ’Medical worker’, ‘at home’, ‘other’
	Fjob	father's Occupation	‘Civil servant’, ’Teacher’, ’employee ’, ’Medical worker’, ‘at home’, ‘other’
	traveltime	home to school travel time	numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour
	studytime	weekly study time	numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours
	schoolsup	extra educational support	binary: yes or no
	internet	Internet access at home	binary: yes or no
	freetime	free time after school	numeric: from 1 - very low to 5 - very high
	famrel	quality of family relationships	numeric: from 1 - very bad to 5 - excellent
	paid	extra paid classes within the course subject	binary: yes or no
	goout	going out with friends	numeric: from 1 - very low to 5 - very high
Sensitive Data	address	student's home address	
	Medu	mother's education	numeric: from 1 to 5
	Fedu	father's education	numeric: from 1 to 5
	id	certificate of identification	
	failures	number of past class failures	numeric: n if 1<=n<3, else 4
	absences	number of school absences	numeric: from 0 to 93
	health	current health status	numeric: from 1 - very bad to 5 - very good
	Dalc	workday alcohol consumption	numeric: from 1 - very low to 5 - very high
	Walc	weekend alcohol consumption	numeric: from 1 - very low to 5 - very high
	income	Monthly household income	numeric: from 1 - 3000 yuan to 5 - 20000 yuan
	Grade1	Score grade for grade1	numeric: from 1 to 5
	Grade2	Score grade for grade2	numeric: from 1 to 5
	Grade3	Score grade for grade3	numeric: from 1 to 5

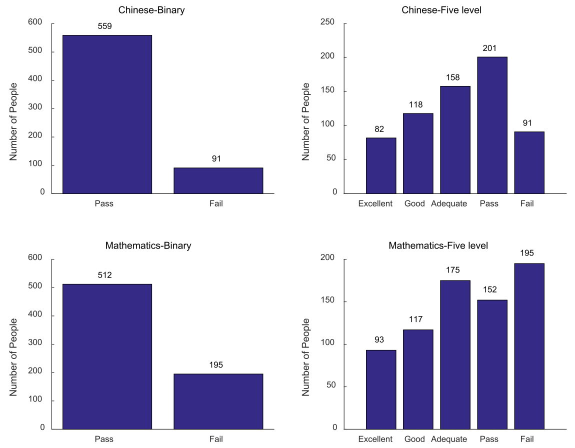

FIGURE 2. **The value distribution of grade attributes.**

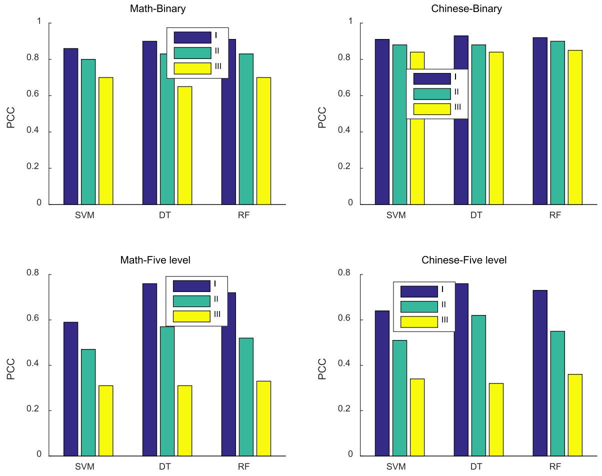

FIGURE 3. **Classification results of different input schemes.**

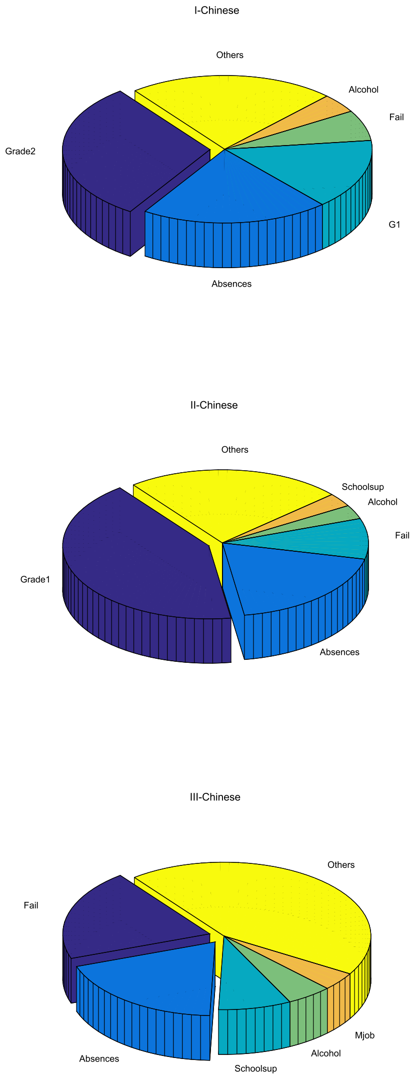

FIGURE 4. **Attributes' contribution rate.**

the design the analysis process. Based on the Rapid-Miner software platform, we use Random Forest (RF), Decision Tree (DT), Support Vector Machines (SVM) data mining algorithms to analyze the statistical performance and predictive ability of different algorithms for students' performance.

We intuitively think that grade one and grade two records have a greater impact on students' performance. Therefore, we design three kinds of input data schemes for each data mining model.

(I) All attributes excluding Grade3

(II) All attributes excluding Grade3 and Grade2

(III) All attributes excluding Grade3, Grade2 and Grade1

For each input scheme, we divide the data into 10 groups for a cross validation of 30 runs. For each mining algorithm, we randomly select one of the 10 sets of data as a training sequence to generate a classifier to verify other data. When this process is finished, the simulation set to be evaluated contains the entire dataset. Based on the 95% confidence interval of each simulation, the average results are shown in figure 3.

As we expect, the (I) scheme can achieve the best results. With the continuous reduction of grade records, the prediction performance will be gradually reduced, and the worst results can be obtained when the (III) scheme is used. By contrast, DT can achieve better classification results if the condition is better. Generally, the nonlinear method SVM can achieve robust performance when conditions are poor. The reason for this result is that there is no correlation among a large amount of input data.

In order to analyze which attributes play a more important role in knowledge extraction, we only list the top five attributes that are most relevant to classification performance. As it can be seen from Figure 4, five of the most important attributes account for the entire impact from 56% to 76%. Grade2 accounted for the largest proportion for (I) scheme. At the same time, Grade1 is the most important attribute for the (II) scheme. Failure rates and absenteeism rates are the most important factors that have been associated with past performance when no previous grade records are available for the (III) scheme. Other factors, such as school background, family background, and social background, have little contribution to prediction ability. However, we cannot ignore their role (for instance, Alcohol).

Alcohol has a lot of bad effects on the way we live. In school days, the brain is still developing. Students who are addicted to alcohol may affect their learning ability, memory function, response and attention throughout their schooling. It was a painful experience. We guide and prevent the deterioration of the situation if we understand the main factors that influence the students' drinking.

As it can be seen from Figure 5, gender is the first factor, that is, men are more likely to drink. Secondly, children with poor grades have a higher chance of exposure to alcohol. If a child is often absent from school and has plenty of free time, he is much more likely to drink than others. There are also something easy to be ignored by parents are their family factors. A good education level of parents, a father's work, and a harmonious family can affect a child's health and growth.

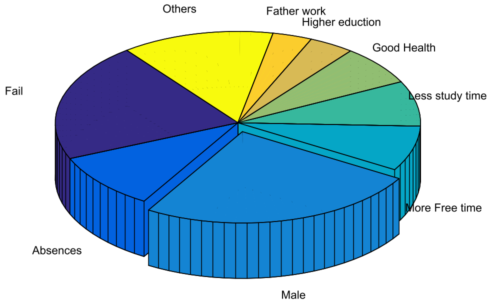

FIGURE 5. Attributes impact on Alcohol.

Education is an important part of social development. We obtain the main factors restricting the students' performance through the depth of excavation of student information. Make full use of the research results to improve the educational ideas and management level. At the same time, we should also note that the seemingly insignificant factors have great influence on students' growth. The whole society must work together to create a good educational environment. On the other hand, the students' information database contains a lot of personal privacy data. If illegal attackers use the data mining algorithm to extract personal sensitive information, it will pose a potential threat to students and families. Therefore, it is necessary to hide the users' sensitive information as far as possible without reducing the ability of insensitive information extraction.

V. CLASSIFICATION DISCRIMINANT

In the cooperative learning environment, we must put forward a feasible method to reduce the privacy leakage and ensure the security of the data. By reducing data dimension and removing some components, it prevents illegal data reconstruction. In this paper, we shall propose a new approach where sensitive information hiding can be proposed as a pair of classification transformation tasks which base on idea of Discriminant Component Analysis (DCA) for creating reduced dimensional subspaces in collaborative learning environment. We further propose the kernel function to nonlinear data extraction.

The realization scenario is sensor network or cloud in a collaborative environment. The goal is to improve the accuracy of the expected information and protect the identity information of the data owner as privacy. It is assumed that expected information and privacy are known to the system designer. Different from the existing technologies, we design a pair of sensitive and insensitive classification tasks using second classification goals. The target is to design an optimal data transformation algorithm that does not reduce the desired task performance while minimizing the performance of the undesired classification task.

The initial data is represented mathematically as M-dimensional vector $\mathbf{x}$ i.e. $\mathbf{x} = \left[x^1, \cdots, x^M\right]^T$. m represents the main components extracted from the original data space. The traditional principal component analysis(PCA) is to design transformation matrix $\mathbf{W} \in \mathbb{R}^{M\times m}$, i.e. $\mathbf{W} = \left[\mathbf{w}_1 \cdots \mathbf{w}_m\right]$, where the result of the transformation can be expressed as $\mathbf{z} = \mathbf{W}^T\mathbf{x}$. If $\mathbf{x}$ fits the assumption of Gaussian distribution, in many practical applications, the second-order statistics are usually unknown. The covariance matrix $\mathbf{R} = \left\{r_{ij}\right\}$. $r_{ij} = E\left[x^i x^j\right]$ indicates the correlation between the i-th and j-th entities. For given set of training vectors: $\mathbf{X} = \left\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\right\}$ with the zero-mean standard variance, we can get a "center-adjusted" data matrix defined as: $\bar{\mathbf{X}} = \left[\bar{\mathbf{x}}_1\ \bar{\mathbf{x}}_2 \cdots \bar{\mathbf{x}}_N\right]$, where $\bar{\mathbf{x}}_i = \mathbf{x}_i - \vec{\boldsymbol{\mu}}$, and $\vec{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{N}\mathbf{x}_i}{N}$, $i = 1, \ldots, N$. The derivation process of scatter matrix is as follows:

$$\bar{\mathbf{S}} = \bar{\mathbf{X}}\bar{\mathbf{X}}^T = \sum_{i=1}^{N}\left[\mathbf{x}_i - \vec{\boldsymbol{\mu}}\right]\left[\mathbf{x}_i - \vec{\boldsymbol{\mu}}\right]^T \tag{1}$$

The class label of the training vectors are given as the complete training dataset: $[\mathbf{X}, \mathbf{Y}] = \left\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)\right\}$, where $y_i \in \Re$ is corresponded to class labels related to training vector $\mathbf{x}_i$, $i = 1, \ldots, N$. We define L for the different number of classes, and N_ℓ as the number of training vectors contained in the l-th class, $l = 1, \ldots, L$. Therefore, the within-class scatter matrix $\mathbf{S}_W$ can be denoted as:

$$\mathbf{S}_W = \sum_{\ell=1}^{L}\sum_{j=1}^{N_\ell}\left[\mathbf{x}_j^{(\ell)} - \vec{\boldsymbol{\mu}}_\ell\right]\left[\mathbf{x}_j^{(\ell)} - \vec{\boldsymbol{\mu}}_\ell\right]^T \tag{2}$$

Moreover, the between-class scatter matrix $\mathbf{S}_B$ can be denoted as:

$$\mathbf{S}_B = \sum_{\ell=1}^{L} N_\ell\left[\vec{\boldsymbol{\mu}}_\ell - \vec{\boldsymbol{\mu}}\right]\left[\vec{\boldsymbol{\mu}}_\ell - \vec{\boldsymbol{\mu}}\right]^T \tag{3}$$

In linear discriminant analysis, the within-class scatter matrix affects the projection direction of discriminant component analysis. A high directional power of the between-class the scattering matrix $\mathbf{S}_B$ will enhance the ability to distinguish. The within-class scatter matrix $\mathbf{S}_W$ will play a negative role, if a high directivity of the noise power and discriminant power in the same direction will weaken the performance.

For supervised learning applications, a new criterion for Multiple Discriminant Analysis (MDA) is proposed [23]. The MDA aims at maximizing the following criterion:

$$J(\mathbf{W}) = \underset{\{\mathbf{W}\in\mathbb{R}^{M\times m}\}}{\arg\max} \frac{\left|\mathbf{W}^T\mathbf{S}_B\mathbf{W}\right|}{\left|\mathbf{W}^T\mathbf{S}_W\mathbf{W}\right|} \tag{4}$$

According to the ideas mentioned above, we design an algorithm that enables it to reach the desired task without revealing privacy. We describe our model as a set of classification tasks: insensitive information task (IIT) and sensitive information task (SIT). For IIT, there is a set of related labels p_i put forward the new data set: $[\mathbf{X}, \mathbf{P}] = \left\{(\mathbf{x}_1, p_1), (\mathbf{x}_2, p_2), \cdots, (\mathbf{x}_N, p_N)\right\}, p_i \in 1, \ldots, L$, where L is redefined as different number of classes for insensitive information. At the same time, we give the same data set for SIT as: $[\mathbf{X}, \mathbf{S}] = \left\{(\mathbf{x}_1, s_1), (\mathbf{x}_2, s_2), \cdots, (\mathbf{x}_N, s_N)\right\}$, $s_i \in 1, \ldots, F$, where F is defined as different number of classes for sensitive information. The goal is to design a system to predict p_i well, but not s_i.

We face two classification problems, and their classification objectives are contradictory. By projecting the data into a subspace of $\mathbf{W}$, we want to classify the IIT classification task as much as possible, and minimize the separability of SIT. According to SNR maximum criterion, we define Classification Discriminant Criterion (CDC) to search an optimal projection matrix which maximizes the ratio.

$$J_{CDC}(\mathbf{W}) = \underset{\{\mathbf{W}\in\mathbb{R}^{M\times m}\}}{\arg\max} \frac{\left|\mathbf{W}^T\mathbf{S}_{B_{IIT}}\mathbf{W}\right|}{\left|\mathbf{W}^T\mathbf{S}_{B_{SIT}}\mathbf{W}\right|} \tag{5}$$

where we denote the between-class scatter matrix for IIT as:

$$\mathbf{S}_{B_{IIT}} = \sum_{\ell=1}^{L} N_\ell^{IIT}\left[\vec{\boldsymbol{\mu}}_\ell^{IIT} - \vec{\boldsymbol{\mu}}\right]\left[\vec{\boldsymbol{\mu}}_\ell^{IIT} - \vec{\boldsymbol{\mu}}\right]^T \tag{6}$$

Similarly, for SIT the between-class scatter matrix is:

$$\mathbf{S}_{B_{SIT}} = \sum_{f=1}^{F} N_f^{SIT}\left[\vec{\boldsymbol{\mu}}_f^{SIT} - \vec{\boldsymbol{\mu}}\right]\left[\vec{\boldsymbol{\mu}}_f^{SIT} - \vec{\boldsymbol{\mu}}\right]^T \tag{7}$$

The M-dimensional vector of the original data space can be mapped to the J-dimensional vector on the corresponding inherent space $\mathbf{x} \to \phi(\mathbf{x})$ and $\mathbf{y} \to \phi(\mathbf{y})$, where $J < M$. Moreover, is defined the kernel function of $\mathbf{x}$ and $\mathbf{y}$. The data $\mathbf{X}$ in the initial space can be mapped to the corresponding intrinsic space according to the above method, i.e.

$$\begin{aligned} \mathbf{X} &= \left[\mathbf{x}_1, \quad \mathbf{x}_2, \quad \cdots, \quad \mathbf{x}_N\right] \\ \to \boldsymbol{\Phi} &= \left[\phi(\mathbf{x}_1) \quad \phi(\mathbf{x}_2) \quad \cdots \quad \phi(\mathbf{x}_N)\right] \end{aligned} \tag{8}$$

The "center-adjusted" data matrix can be denoted as: $\bar{\boldsymbol{\Phi}} = \boldsymbol{\Phi}\left[\mathbf{I} - \frac{1}{N}\mathbf{e}^T\mathbf{e}\right]$, where $\mathbf{e}$ is a full one column vector.

As described in the paper, an intrinsic learning model by using the kernel matrix $\mathbf{K} = \boldsymbol{\Phi}^T\boldsymbol{\Phi}$ related with the kernel function can be transformed into its empirical variant, where $K(\mathbf{x}_i, \mathbf{x}_j)$ is the (i, j)-th entry. Then, the center-adjusted kernel matrix is $\bar{\mathbf{K}} = \bar{\boldsymbol{\Phi}}^T\bar{\boldsymbol{\Phi}}$.

There is a $N \times m$ matrix $\mathbf{A}$ that can be expressed as $\mathbf{W} = \bar{\boldsymbol{\Phi}}\mathbf{A}$. In order to avoid ill-conditioned matrix, we introduce a ridge parameter ρ. According to formula (5), we can get

$$\begin{aligned} \mathbf{W}^T\left[\mathbf{S}_{B_{IIT}} + \rho\mathbf{I}\right]\mathbf{W} &= \mathbf{A}^T\bar{\boldsymbol{\Phi}}^T\left[\mathbf{S}_{B_{IIT}} + \rho\mathbf{I}\right]\bar{\boldsymbol{\Phi}}\mathbf{A} \\ &= \mathbf{A}^T\left[\mathbf{K}_{B_{IIT}} + \rho\bar{\mathbf{K}}\right]\mathbf{A} \end{aligned} \tag{9}$$

$$\begin{aligned} \mathbf{W}^T\left[\mathbf{S}_{B_{SIT}} + \rho\mathbf{I}\right]\mathbf{W} &= \mathbf{A}^T\bar{\boldsymbol{\Phi}}^T\left[\mathbf{S}_{B_{SIT}} + \rho\mathbf{I}\right]\bar{\boldsymbol{\Phi}}\mathbf{A} \\ &= \mathbf{A}^T\left[\mathbf{K}_{B_{SIT}} + \rho\bar{\mathbf{K}}\right]\mathbf{A} \end{aligned} \tag{10}$$

We define $\mathbf{R}_s = \mathbf{K}_{B_{IIT}} + \rho\bar{\mathbf{K}}$ and $\mathbf{R}_n = \mathbf{K}_{B_{SIT}} + \rho\bar{\mathbf{K}}$ as equivalent optimization target components and apply Rayleigh entropy to find the optimal solution.

$$\begin{aligned} J_{CDC}(\mathbf{W}) &= \max_{\mathbf{A}} \frac{\mathbf{A}^T\mathbf{R}_s\mathbf{A}}{\mathbf{A}^T\mathbf{R}_n\mathbf{A}} \\ &= \max_{\mathbf{A}} \frac{\mathbf{A}^T\left(\mathbf{R}_n^{\frac{1}{2}}\right)^T\mathbf{R}_n^{-\frac{1}{2}}\mathbf{R}_s\mathbf{R}_n^{-\frac{1}{2}}\mathbf{R}_n^{\frac{1}{2}}\mathbf{A}}{\mathbf{A}^T\mathbf{R}_n^{\frac{1}{2}}\mathbf{R}_n^{\frac{1}{2}}\mathbf{A}} \\ &\overset{\mathbf{V}=\mathbf{R}_n^{\frac{1}{2}}\mathbf{A}}{=\!=\!=\!=\!=} \max_{\mathbf{V}} \frac{\mathbf{V}^T\mathbf{R}_n^{-\frac{1}{2}}\mathbf{R}_s\mathbf{R}_n^{-\frac{1}{2}}\mathbf{V}}{\mathbf{V}^T\mathbf{V}} \\ &\overset{\mathbf{R}_{sn}=\mathbf{R}_n^{-\frac{1}{2}}\mathbf{R}_s\mathbf{R}_n^{-\frac{1}{2}}}{=\!=\!=\!=\!=\!=\!=\!=\!=} \max_{\mathbf{V}} \frac{\mathbf{V}^T\mathbf{R}_{sn}\mathbf{V}}{\mathbf{V}^T\mathbf{V}} \end{aligned} \tag{11}$$

According to the Rayleigh entropy, it can be seen that the optimization goal is to solve the maximum eigenvalue problem of $\mathbf{R}_{sn}$. Based on the generalized eigenvalue decomposition, the optimal solution of $\mathbf{R}_s\mathbf{A}_{opt} = \lambda_{\max}\mathbf{R}_n\mathbf{A}_{opt}$ is the eigenvectors corresponding to the largest eigenvalues of the matrix pair $(\mathbf{R}_s, \mathbf{R}_n)$. The solution can be obtained from the m principal eigenvectors according to the descending order of eigenvalues, i.e.

$$\mathbf{A} = \left[\mathbf{a}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_m\right] \tag{12}$$

We can get the projected data by using CDC:

$$\begin{aligned} \mathbf{y} &= \mathbf{W}^T\phi(\mathbf{x}) \\ &= \mathbf{A}^T\bar{\boldsymbol{\Phi}}^T\phi(\mathbf{x}) = \mathbf{A}^T\vec{\mathbf{k}}(\mathbf{x}) \end{aligned} \tag{13}$$

where $\bar{\boldsymbol{\Phi}}$ is "center-adjusted" training vector set, $\phi(\mathbf{x})$ is test vector set.

$$\begin{aligned} \vec{\mathbf{k}}(\mathbf{x}) &= \bar{\boldsymbol{\Phi}}^T\phi(\mathbf{x}) \\ &= \left[\phi(\mathbf{x}_1)^T\phi(\mathbf{x}) \quad \phi(\mathbf{x}_2)^T\phi(\mathbf{x}) \quad \cdots \quad \phi(\mathbf{x}_N)^T\phi(\mathbf{x})\right]^T \\ &= \left[K(\mathbf{x}_1, \mathbf{x}) \quad K(\mathbf{x}_2, \mathbf{x}) \quad \cdots \quad K(\mathbf{x}_N, \mathbf{x})\right]^T \end{aligned}$$

The detail of the proposed scheme is shown in Table 2.

Discussion 1: The above algorithm presents a nonlinear projection by using nonlinear transformation $\phi(\mathbf{x})$. In linear collaborative learning environments, we can extend this approach easily to linear subspace with linear transformation

TABLE 2. The main steps of classification discriminant scheme.

Initialization step:

$\mathbf{x}$: an M–dimensional vector for initial data

M : attribute components for the original vector space

N : the number of the set of training vectors

L : the number of different classes

N_ℓ : the number of training vectors for the l-th class

$\mathbf{W}$: the optimal projection matrix

We describe our model as a group of classification tasks: insensitive information task(IIT) and sensitive information task(SIT).

Main step:

Using the idea of Multiple Discriminant Analysis to construct the optimization target based on Classification

Discriminant Criterion: $J_{CDC}(\mathbf{W}) = \arg\max_{\{\mathbf{W}\in\mathbb{R}^{M\times m}\}} \frac{\left|\mathbf{W}^T\mathbf{S}_{B_{IIT}}\mathbf{W}\right|}{\left|\mathbf{W}^T\mathbf{S}_{B_{SIT}}\mathbf{W}\right|}$

The M-dimensional vector of the original data space can be mapped to the J-dimensional vector on the corresponding inherent space $\mathbf{x}\to\phi(\mathbf{x})$; linear product of corresponding intrinsic vectors constructs the kernel function $K(\mathbf{x},\mathbf{y})$.

Get the between-classes scatter matrices for IIT and SIT.

$$\mathbf{S}_{B_{IIT}} = \sum_{\ell=1}^{L} N_\ell^{IIT}\left[\vec{\boldsymbol{\mu}}_\ell^{IIT}-\vec{\boldsymbol{\mu}}\right]\left[\vec{\boldsymbol{\mu}}_\ell^{IIT}-\vec{\boldsymbol{\mu}}\right]^T$$

$$\mathbf{S}_{B_{SIT}} = \sum_{f=1}^{F} N_f^{SIT}\left[\vec{\boldsymbol{\mu}}_f^{SIT}-\vec{\boldsymbol{\mu}}\right]\left[\vec{\boldsymbol{\mu}}_f^{SIT}-\vec{\boldsymbol{\mu}}\right]^T$$

Equivalent transformation of optimization target

$$J_{CDC}(\mathbf{W}) = \max_{\mathbf{A}} \frac{\mathbf{A}^T\mathbf{R}_s\mathbf{A}}{\mathbf{A}^T\mathbf{R}_n\mathbf{A}}$$

The m principal eigenvectors from generalized eigenvalue decomposition of $\mathbf{R}_s\mathbf{A}_{opt} = \lambda_{\max}\mathbf{R}_n\mathbf{A}_{opt}$.

$$\mathbf{A} = \left[\mathbf{a}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_m\right]$$

The kernel projected data can be represented by dimension-reduced vector:

$$\mathbf{y} = \mathbf{A}^T\vec{\mathbf{k}}(\mathbf{x})$$

Publish or share the projected data $\mathbf{y}$.

$\phi(\mathbf{x}) = \mathbf{I}_{M\times M}$. The transformed test data uses linear projection as: $\mathbf{y} = \mathbf{W}^T\mathbf{x}$.

Discussion 2: m is the number of principal eigenvectors extracted from the eigenvalue decomposition which play a critical impact on system performance. We define $\mathbf{W}^T\mathbf{S}_{B_{IIT}}\mathbf{W}$ as signal variance and $\mathbf{W}^T\mathbf{S}_{B_{SIT}}\mathbf{W}$ as noise variance. It is easily shown that the generalized eigenvalue decomposition of $\mathbf{R}_s\mathbf{A} = \lambda\mathbf{R}_n\mathbf{A}$ has the following form: $\mathbf{A} = \left[\mathbf{A}_s\ \mathbf{A}_n\right]$, where $\mathbf{A}_s = \left[\mathbf{a}_1 \cdots \mathbf{a}_d\right]$ is signal subspace, $\mathbf{A}_n = \left[\mathbf{a}_{d+1} \cdots \mathbf{a}_M\right]$ is noise subspace, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > \lambda_{d+1} = \cdots = \lambda_M = \sigma^2$, where σ^2 is the noise power. When $m > d$, the performance degradation is due to the absence of the full use of the signal subspace. On the contrary, when $m < d$, the introduction of the additional noise subspace leads to performance penalties. Optimal way of exploiting this property would be to find the set of vectors that $m = d$, i.e. the full signal subspace is orthogonal to the noise subspace.

Discussion 3: The reconstruction of sensitive information is one of the important indicators to measure the performance of privacy protection. Let the M-dimensional vector $\hat{\mathbf{x}}$ defines the optimal estimate of $\mathbf{x}$ with the projection matrix $\mathbf{W}$, i.e. $\hat{\mathbf{x}} = \mathbf{W}\mathbf{W}^T\mathbf{x}$. According to mean square error criterion:

$$\min_{\mathbf{W}\in\mathbb{R}^{M\times m}} E\left[\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|^2\right] = \min_{\mathbf{W}\in\mathbb{R}^{M\times m}} \left(\mathbf{I}-\mathbf{W}\mathbf{W}^T\right)\bar{\mathbf{S}}\left(\mathbf{I}-\mathbf{W}\mathbf{W}^T\right) \tag{14}$$

Therefore, the reconstruction error can be expressed as

$$\begin{aligned} err &= tr\left(\left[\mathbf{I}-\mathbf{W}\mathbf{W}^T\right]\bar{\mathbf{S}}\left[\mathbf{I}-\mathbf{W}\mathbf{W}^T\right]\right) \\ &= tr\left(\left[\mathbf{I}-\mathbf{V}_s\mathbf{V}_s^T\right]\left[\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T\right]\left[\mathbf{I}-\mathbf{V}_s\mathbf{V}_s^T\right]\right) \\ &= tr\left(\mathbf{V}_n\boldsymbol{\Lambda}\mathbf{V}_n^T\right) \\ &= \sum_{i=d+1}^{M}\lambda_i \end{aligned} \tag{15}$$

The result of this discussion is consistent with the conclusions of *Discussion 2*.

Discussion 4: The determinant criterion based on formula (5) is invariant to any linear transformation. For any nonsingular transformation $\mathbf{R}$, we draw a conclusion that $J_{CDC}(\mathbf{W}) = J_{CDC}(\mathbf{W}\mathbf{R})$. It is shown that $J_{CDC}(\mathbf{W})$ depends only on the subspace spanned by $\mathbf{W}$ and does not depend on the individual column vector of the linear transformation of the matrix $\mathbf{W}$.

VI. EXPERIMENTS AND EVALUATION

In order to pay attention to the harm of students' performance, we design a more effective classification prediction algorithm. This data set consists of 744 instances and each contains 32 attribute features [21]. The insensitive task is to identify usage, social, gender, internet, free time and study time attributes (six utility classes) for each student. The sensitive task is to identify user identity. We use cross validation to test our results. Cross validation data are randomly divided into multiple subsets or groups, and Monte Carlo simulation is carried out for training and testing. Part of the data is used as the training set and the other as the test set.

Figure 6 compares the classification accuracy of linear and nonlinear subspace projection algorithms with the increase of the projective dimension. The original data is at the top left of the graph, and the optimal algorithm should be as close as possible to the top right of the graph. As it can be seen from the graph, the performance of the system is poor when the dimension of the projection matrix is small. With the increasing of the dimension, the performance is improved gradually. When the dimension exceeds a certain threshold, the performance becomes worse.

The proposed classification algorithm can execute well in sensitive privacy preserving tasks while losing as little as

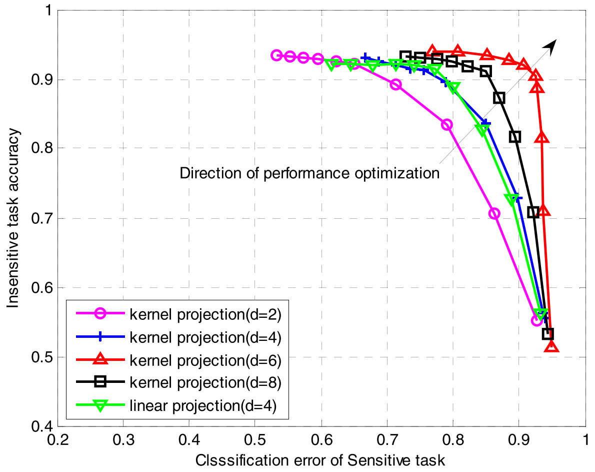

FIGURE 6. Performance of classification data projection.

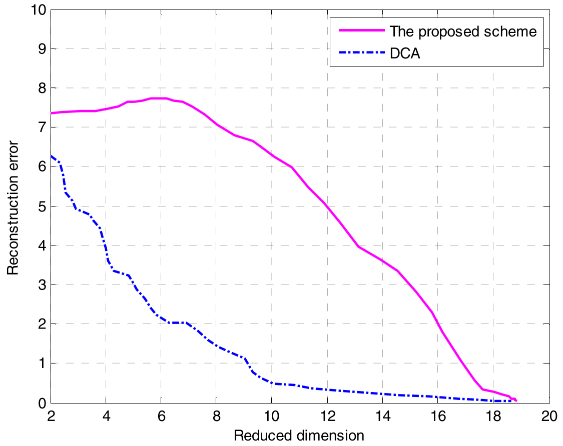

FIGURE 8. Performance of reconstruction error for privacy protection.

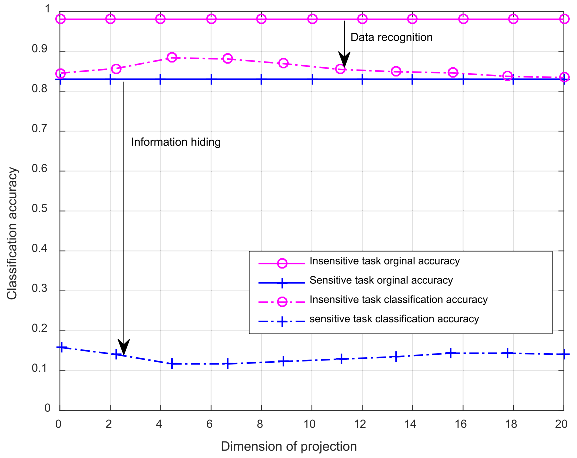

FIGURE 7. Capability of information hiding and data recognition.

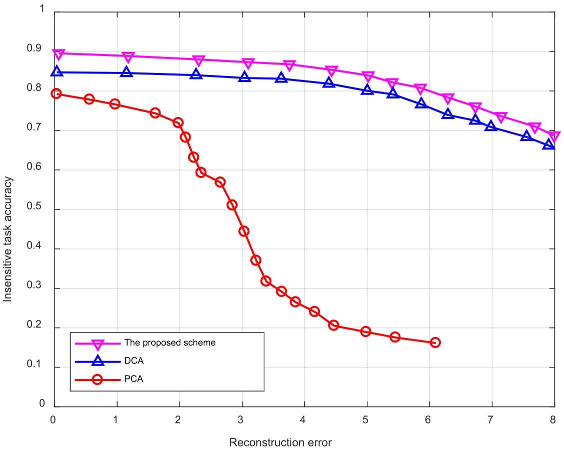

FIGURE 9. Insensitive accuracy of different schemes

possible the performance of insensitive tasks. As it can be seen from Figure 7, the original data before the classification projection has very good classification accuracy for sensitive task and insensitive task. When the data are transformed by using the classification discriminant, the insensitive task still maintains excellent classification accuracy and the sensitive task has better privacy protection ability.

Spatial dimension plays a pivotal role in determining the performance of privacy preserving classification projection schemes. When the dimension is small, the performance is poor due to the loss of more useful information. When the dimension is large, the performance is poor due to the introduction of more noise space. In figure 8, the performance is optimal when m=6. The simulation results verify the conclusion of Discussion 2.

At the same time, it should be noted that sensitive task and insensitive task can be exchanged between each other. We may design a new classification projection matrix to suppress insensitive task while preserving sensitive task as much as possible. In some applications, we need to identify the identity of the user and ignore the information that has nothing to do with the users' identity. The role of sensitive task and insensitive task can be exchanged.

Reconstruction error (RE) can be used to characterize the loss of the original data projected onto the subspace. RE is an important indicator of system security if an illegal attacker attempts to restore the data or classify sensitive tasks through the classification projection matrix. It can be seen from Figure 8 that the proposed scheme is more effective than DCA. The simulation results are consistent with Discussion 3.

The aim of PCA is to find the optimal subspace to reconstruct the original data with minimum loss. Therefore, PCA should be able to cast the initial data into a low dimensional space for better privacy. When the space projection dimension is reduced appropriately, DCA can selectively keep back the important information of the desired task. The above analysis is verified in Figure 9. In this paper, the performance of the proposed scheme and DCA scheme is relatively stable, and the proposed scheme is slightly better. PCA is the worst performance due to utility-privacy tradeoff.

For comparison, four methods are investigated in the Table 3. In terms of privacy, experiments have shown that

TABLE 3. Performance comparison of different schemes.

Algorithm	Insensitive task (%)	Sensitive task (%)
Random guess	15.82	10.54
Discriminant component analysis [23]	85.53	16.22
Compressive privacy[9]	86.20	13.48
The proposed two classification projection transformation (d=2)	84.89	13.67
The proposed two classification projection transformation (d=6)	88.21	11.74

two classification projection transformations are the most effective in providing privacy. By reducing data dimension and removing some components, privacy accuracy is almost at the best level or close to random guess. In terms of utility, the compressive privacy and the proposed methods show significant increase in utility accuracy in experiments. When $d < m$, the performance degradation is due to the absence of the full use of the signal subspace. Optimal way of exploiting this property would be to find the set of vectors that $d = m$, i.e. the full signal subspace is orthogonal to the noise subspace.

The simulation results confirm that the compressive privacy does provide excellent privacy protection. But this privacy protection is achieved with a large amount of information lost. Privacy Preserving Based on the two classification projection transformation adopts maximizing insensitive information while minimizing sensitive information. The key to its realization is to find an optimization criterion suitable for the goal. This paper extends such a goal to the SNR criterion. The result is achieved by considering the maximization of insensitivity-to-sensitivity ratio.

VII. CONCLUSION

We use the algorithm of discriminant component analysis to predict the influence of students' performance, and to analyze the main factors for non-dominant components. In order to protect the sensitive information in the data analysis process, we propose the classification of sensitive tasks and insensitive task. The multiclass discriminant model is used to construct the optimal projection matrix, which can suppress the leakage of sensitive information while preserving the characteristics of insensitive information.

In the current research, DCA is mainly used to improve the accuracy of insensitive information recognition, and achieved the desired results. At the same time, the current work of projection transformation mainly focuses on how to design optimal weights to improve information recognition. However, the privacy preserving learning mode should include maximizing the expected information based on the minimization of privacy. In all the above work, due to the lack of second reference classification problems, the performance of these methods can not match the method proposed in this paper. In the application of identity authentication, we also construct a new projection matrix by exchanging roles of sensitive tasks and insensitive tasks. It can preserve the identity information as far as possible while ignoring useless information.

REFERENCES

[1] T. Patrick, "Has big data made anonymity impossible?" *Technol. Rev. Manchester NH*, vol. 116, no. 4, pp. 64–66, 2013.

[2] B. Luca and L. Xiong, "Mining frequent patterns with differential privacy," *Proc. VLDB Endowment*, vol. 6, no. 12, pp. 1422–1427, 2013.

[3] B. Luca and X. Li, "A two-phase algorithm for mining sequential patterns with differential privacy," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, San Francisco, CA, USA, 2013, pp. 269–278.

[4] B. Luca and X. Li, "On differentially private longest increasing subsequence computation in data stream," *Trans. Data Privacy*, vol. 9, no. 1, pp. 73–100, 2016.

[5] H. Su, K. Zheng, J. M. Huang, H. Jeung, L. Chen, and X. F. Zhou, "Crowdplanner: A crowd-based route recommendation system," in *Proc. IEEE 30th Int. Conf. Data Eng. (ICDE)*, Mar. 2014, pp. 1144–1155.

[6] H. Wang *et al.*, "A real-time route recommendation system," *Proc. VLDB Endowment*, vol. 7, no. 13, pp. 1549–1552, 2014.

[7] H. Yoon, Y. Zheng, X. Xie, and W. Woo, "Smart itinerary recommendation based on user-generated GPS trajectories," in *Ubiquitous Intelligence and Computing* (Lecture Notes in Computer Science), vol. 6406. Berlin, Germany: Springer, Oct. 2010, pp. 19–34.

[8] T. Hashem, T. Hashem, M. Ali, and L. Kulik, "Group trip planning queries in spatial databases," in *Advances in Spatial and Temporal Databases—SSTD* (Lecture Notes in Computer Science), vol. 8098. Berlin, Germany: Springer, 2013, pp. 259–276.

[9] S. Y. Kung, "Compressive privacy: From informationestimation theory to machine learning," *IEEE Signal Process. Mag.*, vol. 34, no. 1, pp. 94–112, Jan. 2017.

[10] S. Y. Kung, "A compressive privacy approach to generalized information Bottleneck and privacy funnel problems," *J. Franklin Inst.*, vol. 355, no. 4, pp. 1846–1872, 2018.

[11] T. Dalenius, "Towards a methodology for statistical disclosure control," *Statistik Tidskrift*, vol. 15, pp. 429–444, Jan. 1977.

[12] B. C. M. Fung, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey on recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, pp. 1–53, 2010.

[13] V. Kumari and S. Chakravarthy, "Cooperative privacy game: A novel strategy for preserving privacy in data publishing," *Hum. Centric Comput. Inf. Sci.*, vol. 6, p. 12, Jul. 2016.

[14] H. Park and K. Shim, "Approximate algorithms with generalizing attribute values for k-anonymity," *Inf. Syst.*, vol. 35, no. 8, pp. 933–955, Dec. 2010.

[15] C. N. Sowmyarani, G. N. Srinivasan, and K. Sukanya, "A new privacy preserving measure: p-Sensitive t-closeness," in *Proc. Int. Conf. Comput. Adv. Intell. Syst. Comput.*, vol. 174, 2013, pp. 57–62.

[16] J. Zhang, J. Xie, J. Yang, and B. Zhang, "A t-closeness privacy model based on sensitive attribute values semantics bucketization," *J. Comput. Res. Develop.*, vol. 51, no. 1, pp. 126–137, 2014.

[17] T. Ankleshwaria and J. S. Dhobi, "Geometric data perturbation approach for privacy preserving in data stream mining," *Eng. Universe Sci. Res. Manage.*, vol. 6, no. 4, pp. 1–6, 2014.

[18] G. N. Rao, M. S. Harini, C. R. Kishore, "A cryptographic privacy preserving approach over classification," in *Proc. 48th Annu. Conv. Comput. Soc. India-Vol II. Adv. Intell. Syst. Comput.*, vol. 249. Cham, Switzerland: Springer, 2014.

[19] A. Rahmani, A. Amine, and R. H. Mohamed, "A multilayer evolutionary homomorphic encryption approach for privacy preserving over big data," in *Proc. IEEE Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discovery*, Oct. 2014, pp. 19–26.

[20] B. H. Patell and A. N. Shah, "Overview of Privacy preserving techniques and data accuracy," *Int. J. Adv. Res. Comput. Sci. Manage. Stud.*, vol. 3, no. 1, pp. 135–140, 2015.

[21] Shaanxi Provincial Department of Education, "Shaanxi provincial education development statistical communique," Tech. Rep., 2016.

[22] J. Lu, A. Hales, D. Rew, and M. Keech, "Timeline and episode-structured clinical data: Pre-processing for Data Mining and analytics," in *Proc. IEEE Int. Conf. Data Eng. Workshops*, May 2016, pp. 64–67.

[23] S.-Y. Kung, "Discriminant component analysis for privacy protection and visualization of big data," *Multimedia Tools Appl.*, vol. 76, no. 3, pp. 3999–4034, 2017.

XIYUAN WANG received the B.S. degree in radio electronics and the M.S. degree in microelectronic technology from Xidian University in 1998 and 2003, respectively. She is currently pursuing the Ph.D. degree. She is currently a Lecturer with Xidian University. Her current work concerns cooperative communication and information query and management.

YONG WANG received the B.S. degree in electronic mechanics, the M.S. degree in computer science, and the Ph.D. degree in signal and information processing from Xidian University in 2001, 2004, and 2009, respectively. He is currently an Associate Professor with the School of Cyber Engineering, Xidian University. His current work concerns adaptive antenna design and cooperative communication.