Recommendation System for Privacy-Preserving Education Technologies

Shasha Xu and Xiufang Yin

Zhengzhou Preschool Education College, Zhengzhou, Henan 450000, China

Correspondence should be addressed to Shasha Xu; xushasha1984@126.com

Received 2 March 2022; Revised 7 March 2022; Accepted 17 March 2022; Published 16 April 2022

1. Introduction

Artificial intelligence-based educational techniques have advanced significantly in recent years, and their applications in various academic fields have increased. Implementing artificial intelligence in education encompasses a broad range of intelligent instructional and evaluation methods, including intelligent tutoring systems, intelligent performance assessment, intelligent virtual agents, talking robots, humanized chatbots, and any other approach based on artificial intelligence [1]. These classroom innovations can benefit a diverse range of students, particularly those with disabilities. Thanks to new intelligence technologies, these students now have a more flexible and personalized educational solution.

In general, artificial intelligence can be combined with other methods (e.g., speech recognition, machine vision, and disability assistant) to develop advanced tutor systems that can help students learn more effectively [2]. Furthermore, approaches based on artificial intelligence can be used to create adaptive and personalized learning systems that are tailored to the unique characteristics of each individual student.

Nevertheless, as AI models reflect the data over which they were trained, data that may have privacy or other sensitivities associated with it, they are vulnerable. This work proposes a privacy-preserving recommendation system that uses differential privacy in this spirit. Differential privacy [4] is a technology that allows researchers and database analysts to acquire useful information from databases that contain people’s personal information without disclosing the unique identify of the persons who have provided the information. Achieving this can be accomplished by including the bare minimum of distractions in the information provided by the database system. The amount of distraction introduced is significant enough to protect privacy while still allowing for the provision of information to analysts to continue to be valid. Differential privacy, in its most basic sense, is the process of forming data anonymously by deliberately adding noise into a dataset. Data analysts are capable of doing any and all possible (functional) statistical analysis without revealing any personal information.
Specifically, this study presents an innovative privacy-preserving recommendation system for educational technologies. It is a fully automated intelligent system that can categorize trainees based on their requirements and special skills. The abilities of each student are automatically categorized into one of several categories. Using a directed acyclic graph machine learning method, the model uses differential privacy in order to protect the private information of each individual learner. Also, an intelligent module based on collaborative filtering offers personalized real-time privacy recommendations.

Afterward, in Section 2, we learn about the proposed system’s technique. Exemptions for applying the proposed method are outlined in Section 3. Section 4 concludes by summarizing the findings and drafting the following potential directions for the work.

2. Proposed Methodology

A directed acyclic graph (DAG) [5] is used to express a probabilistic representation of the data structure created from the model and their putative independence. The classification method is then utilized to validate the whole combined probability distributions in the DAG [5]. The goal is to categorize an X sample into one of the supplied categories C₁, C₂, . . . , Cᵣ using a probability model constructed according to Bayes theory in order to get the desired result. Overall, this is a first-level classification based on probabilities rather than predictions, a fact that has been demonstrated experimentally to be more useful, faster, and more efficient. In this case, projections are made to a certain extent, and the goal is to keep costs as low as possible. Each category is distinguished by a probability distribution that has occurred in the past. We assume that the sample X belongs to a class Ci, and we calculate the probability [5, 6] using the definitions and Bayes theory, respectively. To put it in another way, the initial step in the procedure is to understand how the pupils are dependent on one another and then assign probabilities to them, insuring how likely it is that their ability will change over time. As a result, the proposed system incorporates prior knowledge gathered from the model into the model learning process through a probabilistic representation of the data structure that arises for each learner, hence, enhancing the overall effectiveness of the system. A further consideration is the uncertainty in the model parameters that have been generated, which may be caused by noise such as a random or deceptive evaluation procedure, among other things.

In order to assess the overall performance, the following criteria of the DAG algorithm were used [7–9]:

1. Overall accuracy (OvAc): this metric reflects the proportion of correctly identified samples in relation to the total number of test samples in a given period.
2. Average accuracy (AvAc): this indicator displays the average accuracy of the different categories.
3. Kappa rate: using the following function, we can determine how well the truth map and the final categorization map agree on various statistical criteria:

\[
K = \frac{P_o - P_e}{1 - P_e}
\]

\[
= 1 - \frac{1 - P_0}{1 - P_e},
\]

where \(P_o\) is the correlation between actual agreement and \(P_e\) is the theoretical likelihood of random agreement.

4. McNemar test: to evaluate the significance of categorization accuracy derived from different methodologies, a McNemar test was used:

\[
z_{12} = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}}
\]

where \(f_{ij}\) samples accurately categorized in classification and \(i\) mistakenly classified in the other one \(j\).

5. Coefficient of determination, \(R^2\): use it to express correlation between two variables in percentage terms. The coefficient of determination is a measure of the degree to which the values of \(X\) and \(Y\) are correlated and calculated as follows:

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2},
\]

where \(Y_i\) are the actual values of the dependent variable, \(\bar{Y}_i\) have been calculated based on our best estimates for this dependent variable, and \(Y\) is computed by taking the observed data and averaging it the number of observations.

6. Root relative squared error (RRSE): in order for a model to be considered successful, the absolute correlation between predicted and actual values must be equal to zero:

\[
RRSE = \frac{\sum_{i=1}^{n} (P_{(ij)} - T_{ij})^2}{\sum_{j=1}^{n} (T_j - \bar{T})^2},
\]

where \(P_{(ij)}\) is the anticipated value for a simple hypothesis that the algorithm generates \(j\) and \(T_j\) and \(T\) are the desired value for the simple hypothesis \(j\), with the following connection being used to determine:

\[
\bar{T} = \frac{1}{n} \sum_{j=1}^{n} T_j.
\]
data, curators who need to provide statistics, and adversaries who want to retrieve the sensitive data—can all be solved through differential privacy. This reverse engineering is a type of privacy breach [3, 4, 10].

Finally, an intelligent recommendation memory-based approach was used to measure user privacy and compute the similarity between users [11, 12]. Finding persons with similar interests may be accomplished using the locality-sensitive hashing, which utilizes the closest neighbor algorithm in a linear time frame. A set of privacy restrictions is then proposed based on the k most comparable users and their related user-item matrices. Easy construction and usage, easy facilitation of new data, content-independent of the items being recommended, and effective scalability with co-rated goods are some of the advantages that this technique has to offer [13].

An abstract illustration of the proposed architecture is presented in Figure 1, which depicts as part of a flowchart the basic steps of how the proposed system works.

3. Experiments

A preliminary exam for categorizing pupils’ ability in their various level departments is the subject of this scenario. Students take this simple examination to determine their fitness to continue in higher-level education studies. It includes a set of questions or exercises evaluating skill or knowledge based on a scientific standard that can identify the real learning abilities of each learner [14–16]. Specifically, the preliminary test includes psychometric questionnaires and the purpose is to detect misunderstandings, ambiguities, disabilities, or other learning difficulties that may have the students.

The outcomes of this preliminary test are the dataset used by the classification algorithm. The dataset is used to contain ten questions that come from 350 volunteer students. Table 1 presents the statistical analysis of the preliminary test used in this study.

The questionnaire is satisfactorily reliable in measuring the determination of students’ moods and corresponding abilities and can be used for further processing by the proposed learning system [17, 18].

3.1. Step 1: Classification Process and Results. This model’s probabilistic values and the abovementioned statistical analysis of the questions [19] map each student’s reply to the DAG as a pair of variables based on these criteria [6] in the form $B = (\langle G, \Theta \rangle)$, where $G$ is the nodes $X_1, X_2, \ldots, X_n$. In this form, each question in the questionnaire is represented as a probability value, along with its corresponding edge (the answers to each question). Graph $G$ conveys the assumption that each variable $X_i$ is independent of the inheritance assumed by $G$. $\Theta$ identifies the parameters of the network. Specifically, this set contains the parameter $\theta_{X_i|\pi} = P_B(x_i|\pi)$ for each $X_i$ implementation of $X_i$ in the condition $\pi$, for the set of $X_i$ parents in $G$. Therefore, $B$ defines a unique probability distribution over the variables, namely [5, 7, 20],

$$
P_B = (X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P_B(\pi_i)
$$

(6)

There are three internally distinct paths linking two vertices $u$ and $v$ such that neither of them has the same orientation, or there are two directed cycles with a common vertex if there is a strong component, that is, neither a cycle or a single vertex. Number of predicted components is capped above [5, 20, 21]:

$$
2 \left( \begin{array}{c} n \\ 2 \end{array} \right) n \sum_{i=1}^{n} \sum_{j=1}^{n} \left( \begin{array}{c} n \\ i \end{array} \right) ! l_{1}^{i+1} \left( \begin{array}{c} n \\ j \end{array} \right) ! l_{1}^{j+1} \left( \begin{array}{c} n \\ k \end{array} \right) ! l_{1}^{k+1} + \left( \begin{array}{c} n \\ 1 \end{array} \right) n \sum_{i=2}^{n} \sum_{j=2}^{n} \left( \begin{array}{c} n \\ i \end{array} \right) ! l_{1}^{i+1} \left( \begin{array}{c} n \\ j \end{array} \right) ! l_{1}^{j+1} \leq \frac{3}{n} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \lambda_{i=j=k} + \frac{2}{n} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \lambda_{i=j} = O(n^{-1}).
$$

(7)
Based on the Markov inequality, there are no such components. So, we can bound the expected number of cycles of length larger than \( \omega \) by

\[
\sum_{k=0}^{n} \binom{n}{k} (k-1)!p^k \leq \frac{\prod_{k=0}^{k-1} (n-i)}{\sum_{k=0}^{n} \frac{\lambda^k}{k}} = O(\lambda^n). 
\]

To compute the expectation of \( X_n \), we have

\[
\mathbb{E}[X_n] = \sum_{k=3}^{n} (n-k)! \cdot p^k. 
\]

It follows that

\[
\lim_{n \to \infty} \mathbb{E}(X_n) = \lim_{n \to \infty} \sum_{k=3}^{n} \frac{\lambda^k}{k} \leq \sum_{k=0}^{n} \lambda^k = -\log(1-\lambda) - \frac{\lambda^2}{2} = a(\lambda). 
\]

The \( r \)th factorial moment of \( X_n \) is

\[
\mathbb{E}[(X_n)_r] = \hat{\theta} \sum_{k_1=3}^{n-k_1} \sum_{k_2=3}^{n-k_1} \cdots \sum_{k_r=3}^{n-k_1-k_2-\cdots-k_{r-1}} \binom{n}{k_1,k_2,\ldots,k_r,n-k_1-k_2-\cdots-k_r}^r (k_1-1)!p^k. 
\]

The Hessian matrix of second-order partial and cross-

\[
H(\Lambda) = 
\begin{bmatrix}
\frac{\partial^2 \ell}{\partial \theta_1^2} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_k} \\
\frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_k} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 \ell}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_k^2}
\end{bmatrix}
\]

In order to optimize the problem, we use bordered Hessian:

<table>
<thead>
<tr>
<th>Quest</th>
<th>Mean</th>
<th>S</th>
<th>Sδ</th>
<th>( r_g )</th>
<th>( R^2 )</th>
<th>Cronbach ( \alpha )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>3.425</td>
<td>1.659</td>
<td>5.456</td>
<td>0.799</td>
<td>0.887</td>
<td>0.879</td>
</tr>
<tr>
<td>Q2</td>
<td>3.376</td>
<td>1.544</td>
<td>5.433</td>
<td>0.711</td>
<td>0.806</td>
<td>0.806</td>
</tr>
<tr>
<td>Q3</td>
<td>3.125</td>
<td>1.355</td>
<td>5.562</td>
<td>0.798</td>
<td>0.890</td>
<td>0.811</td>
</tr>
<tr>
<td>Q5</td>
<td>2.788</td>
<td>1.678</td>
<td>6.226</td>
<td>0.542</td>
<td>0.651</td>
<td>0.870</td>
</tr>
<tr>
<td>Q7</td>
<td>3.115</td>
<td>1.454</td>
<td>5.987</td>
<td>0.794</td>
<td>0.874</td>
<td>0.806</td>
</tr>
<tr>
<td>Q8</td>
<td>3.089</td>
<td>1.599</td>
<td>5.998</td>
<td>0.789</td>
<td>0.799</td>
<td>0.798</td>
</tr>
<tr>
<td>Q9</td>
<td>3.341</td>
<td>1.473</td>
<td>5.887</td>
<td>0.801</td>
<td>0.888</td>
<td>0.783</td>
</tr>
<tr>
<td>Q10</td>
<td>3.184</td>
<td>1.932</td>
<td>5.752</td>
<td>0.732</td>
<td>0.801</td>
<td>0.797</td>
</tr>
</tbody>
</table>

Table 2 presents the results of the classification process:

For each variable (response), a probability value is generated, revealing the degree to which it is interdependent with its class and hence the direction in which each question has an effect. In other words, a first classification of the responses into distinct categories can define the options and skills of each student. In this example, based on the

\[
\begin{array}{c|ccccc}
\hline
\text{Quest} & \text{Mean} & \text{S} & \text{Sδ} & \text{\( r_g \)} & \text{\( R^2 \)} & \text{Cronbach \( \alpha \)} \\
\hline
Q1 & 3.425 & 1.659 & 5.456 & 0.799 & 0.887 & 0.879 \\
Q2 & 3.376 & 1.544 & 5.433 & 0.711 & 0.806 & 0.806 \\
Q3 & 3.125 & 1.355 & 5.562 & 0.798 & 0.890 & 0.811 \\
Q5 & 2.788 & 1.678 & 6.226 & 0.542 & 0.651 & 0.870 \\
Q7 & 3.115 & 1.454 & 5.987 & 0.794 & 0.874 & 0.806 \\
Q8 & 3.089 & 1.599 & 5.998 & 0.789 & 0.799 & 0.798 \\
Q9 & 3.341 & 1.473 & 5.887 & 0.801 & 0.888 & 0.783 \\
Q10 & 3.184 & 1.932 & 5.752 & 0.732 & 0.801 & 0.797 \\
\hline
\end{array}
\]
questionnaire, three classes were used (theoretical direction, positive direction, and technological direction), where the students were classified based on their answers and the algorithm of the DAG used.

3.2. Step 2: Differential Privacy and Results. On the contrary, in order to protect an individual who is deciding to allow their data to be included in the repositories that proposed the method, we use differential privacy. Let $q$ be a counting query. Trying to protect privacy by adding noise results in [3, 23–25],

$$M(x) = q(x) + \text{noise}. \quad (14)$$

The Laplace distribution with scale parameter $b > 0$ (assuming position parameter 0) is defined as the distribution with probability density function:

$$\text{Lap}(x | b) = \frac{1}{2b} e^{-|x|/b}. \quad (15)$$

So, it turns out

$$\Pr[M(x) \in S] = \Pr[M(x) \in S|\text{enoughnoise}] + \Pr[M(x) \in S|\text{notenoughnoise}] \leq e^\epsilon \Pr[M(y) \in S] + \delta. \quad (16)$$

The 1-sensitivity of a function $f$ is calculated as

$$\Delta = \max_{x-x'} |f(x) - f(x')|. \quad (17)$$

For example, compare the $x$ database with the test scores and the query for the average score:

$$q(x) = \frac{\sum_{i=1}^{n} x_i}{n}. \quad (18)$$

If we use a neighborhood type relationship such as $|x_i - x_i'| \leq x_{max}$, then the sensitivity of the question will be

$$\Delta = \max_{x-x'} |q(x) - q(x')| = \frac{1}{n} \max_{x-x'} |x_i - x_i'| i \in [1, n]. \quad (19)$$

According to the above equation, the differential privacy mechanism will be

$$M(x) = \frac{\sum_{i=1}^{n} x_i}{n} + \text{Lap}\left(\frac{x_{\text{max}}}{n}\right). \quad (20)$$

Finally,

$$\Pr[\text{notenoughnoise}] \leq \Pr[\text{Lap}(\mu, b) \leq 2] + \Pr[\text{Lap}(\mu, b) \leq 1] = \frac{1}{2} \exp\left(\frac{2 - \mu}{b}\right) + \frac{1}{2} \exp\left(\frac{1 - \mu}{b}\right) \leq \exp\left(\frac{2 - \mu}{b}\right). \quad (21)$$

To prove that the proposed differential privacy system is secure against level 2 attacks, we need to prove that it does not allow distance calculation. Specifically, assuming that a DRE $E$ is used to encrypt the DB to get $E(DB)$, a level 2 attacker with $A = <E(DB), P, 1>$ can retrieve DB if $P$ contains at least $d + 1$ points $x_i (1 \leq i \leq d + 1)$ so that the set of vectors $|x_i - x_j| (2 \leq j \leq d + 1)$ is linearly independent. A hash function used by the Distributed Hash Table (DHT) to assign file ownership to network nodes which generates a key of 256 bits, which is enough to withstand the level 2 attack on the DRE. This system’s encryption function hides the distance between two points in a database table; therefore, it must be determined which of the two points is closest to a query point $q$, and it must also be implemented [4, 26]:

$$d(p_1, q) \geq d(p_2, q) \sqrt{\|p_1\|^2 - 2p_1 \cdot q + \|q\|^2} \geq \sqrt{\|p_2\|^2 - 2p_2 \cdot q + \|q\|^2} - 2p_1 \cdot q + \|q\|^2 - \|p_2\|^2 + 2(p_1 - p_2) \cdot q \geq 0. \quad (22)$$

The attacker cannot increment the estimate of $P$ to diminish the likelihood of a collision and rehash the attack as, within the proposed design, the item maintenance encryption is not remotely retrievable as [28, 29]

$$f(p_1', p_2') = \sqrt{p_1' \cdot q + p_1' - 2(p_1' \cdot p_2') + p_2' \cdot p_2' - d(p_1', p_2'). \quad (23)$$
To put it another way, if the encryption $E$ (i.e., $E$ is DRE), then a computing technique $f$ such that, for all points in time, the differential privacy function $ET$ (i.e., $ET$ is DRE) cannot be remotely retrieved, $p_1$ and $p_2$, and any encryption key $K_1$; it holds that $a_1 = E(p_1, K_1)$ and $a_2 = E(p_2, K_1)$; we have $f(a_1, a_2) = d(p_1, p_2)$. That is, the distance $d(p_1, p_2)$ may be determined from the encrypted values $a_1$ and $a_2$ regardless of the encryption key.

3.3. Step 3: Recommendation System. Finally, an intelligent recommendation memory-based approach was used to measure user privacy and compute the similarity between users. It is a neighborhood-based collaborative filtering approach to produce recommendations [11–13]:

$$r_{u,i} = \text{aggr}_{u' \in U} r_{u',i}.$$  \hfill (24)

The top $N$ most comparable users to user $u$ who share the same level of privacy as user $i$ are denoted by $U$. The aggregation function includes

$$r_{u,i} = \frac{1}{N} \sum_{u' \in U} r_{u',i}r_{u'j,i} = k \sum_{u' \in U} \text{simil}(u, u') r_{u',j,i},$$  \hfill (25)

where $r_u$ is the average privacy of user $u$ for all the users rated by $u$.

The suggested technique determines the cosine similarity between two users in a neighborhood-based approach [4, 6, 30]:

$$\text{simil}(x, y) = \cos \left( \vec{x}, \vec{y} \right)$$

$$= \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \|\vec{y}\|}$$

$$= \frac{\sum_{i \in I_x} r_{x,i} f_{x,i} r_{y,i} f_{y,i}}{\sqrt{\sum_{i \in I_x} r_{x,i}^2 f_{x,i}^2} \sqrt{\sum_{i \in I_y} f_{y,i}^2}}.$$  \hfill (26)

Figure 2 shows the performance results of the proposed method.

As seen in the information supplied above, these findings demonstrate a solid solution to the challenging problem of grouping students to execute tailored educational programs. With the widespread usage of intelligent approaches such as those used in this study, small and heterogeneous student groups can form with members of each group sharing comparable characteristics of student ability, learning difficulties, and psychosocial and cognitive profile. By quickly managing the student potential in their class, as well as being aware of each group’s unique characteristics such as their interests, unique experiences, learning rhythms, and...
learning styles, the teacher can easily manage the student potential of their class and offer high-quality education, taking into account the specific educational needs and capabilities of each group. In addition, the algorithm may be utilized in traditional classrooms and digital or e-learning programs, facilitating the teaching role, as it can compensate for challenges in multicriteria grouping and differentiation of students in a wide range of subject areas. Additionally, it can be utilized with many pupils and produce results in a short amount of time, assuming that the required data is available. Another presumption supporting this idea is that the amount of data that can be regarded as quantitative data or the number of evaluable criteria that come from a comprehensive evaluation of a student is limitless. Finally, each student’s talents are automatically classified into various groups. They were applying an AI technique known as directed acyclic graph learning. Each learner’s private information is protected using differential privacy in the model. Collaborative filtering’s intelligent module provides customized real-time privacy advice.

4. Conclusions

This study presented an innovative recommendation system for privacy-preserving education technologies. It is a hybrid intelligent computing system that can create learning programs based on the unique needs of each learner. It is based on advanced machine learning techniques for performing high-level privacy-preserving analyses to create learning repositories adapted to the trainees’ skills and experiences. The instructional material of educational systems may be successfully rearranged depending on assessment criteria using this novel and privacy-preserving approach. Specifically, using machine learning and differential privacy, this study provides a directed acyclic graph approach to automatically classify each student into a category based on their skills. Next, the model takes advantage of differential privacy. This technology makes it possible to gather relevant information from databases containing the personal information of individuals without disclosing sensitive identification about each individual. Personalized real-time data are also provided by an intelligent suggestion process based on collaborative filtering.

The proposed intelligent system achieved remarkable results in all cases of evaluation, always taking into account the modeling difficulties and uncertainty introduced by the subjective learning system. An important innovation is related to using privacy-preserving recommendations capable of solving multidimensional and complex problems. Also, an exciting finding is emerged from this research related to the possibility of applying to truly unstructured data, techniques, and methodologies and derived from fully theoretical computing, with fully exploitable and realistic results. Furthermore, the proposed method uses a neighborhood-based methodology to determine the cosine similarity between two users as a significantly innovative approach.

Humans are prone to errors or biases that might skew results while doing repetitive tasks such as reading and analyzing open-ended survey replies and other text data. A few simple steps are required for natural language processing (NLP)-powered tools to be taught to the language and criteria of the educational process. So, once they get the machines up and running, they perform far better than humans could ever hope to accomplish. To keep up with the changing marketplace or the language of their education, NLP can be used to investigate and extend the model, which will allow the automated system to take full advantage of modeling learning systems’ wider dependencies with greater accuracy and efficiency. Also, text analysis on a large scale on a variety of papers, internal systems, emails, social media data, online reviews, and more will be made possible by NLP technology. Data can be processed in a matter of seconds or minutes, compared to the days or weeks it would take to analyze manually.

Data Availability

The data used to support the findings of the study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

References


