DIFFERENTIAL PRIVACY IN BIG DATA ANALYTICS FOR HAPTIC APPLICATIONS

K. Prema
M.Tech Scholar, Department of CSE,
Sree Vidyanikethan Engineering College (Autonomous), Tirupati, India

Dr. A.V. Sriharsha
Associate Professor, Department of CSE,
Sree Vidyanikethan Engineering College (Autonomous), Tirupati, India

ABSTRACT

Preservation of privacy in big data has emerged as an absolute prerequisite for exchanging confidential information in terms of data analysis, validation and publishing. Many new techniques have been suggested and implemented for privacy preservation in big data but unfortunately they seems to be failing due to the nature of big data, previous methods and traditional methods like k-anonymity and other anonymization techniques have overlooked privacy protection issues leading to privacy infringement. In this work, a differential privacy protection scheme for ‘big data in body area network’ is developed. Compared with previous methods, the proposed privacy protection scheme is best in terms of availability and reliability. “Differential Privacy” (DP) prevents unwanted re-identification and other privacy threats to individuals whose personal information is present in large datasets, while providing useful access to data. Under the DP model, personal information (micro data) in a large database is not modified and released for analysts to use. Differential privacy works by inserting an intermediary piece of software between the analyst and the database. The analyst never gets to access or actually see the contents of the database; instead the intermediary acts as a privacy-protecting screen or filter, effectively serving as a privacy guard. To verify the advantages of our scheme, several experiments are conducted to show the results. Exploratory results demonstrate that, even when the attacker has full background knowledge, the proposed scheme can still provide enough interference to big sensitive data so as to preserve the privacy. Body Area Networks (BANs), collects enormous data by wearable sensors which contains sensitive information such as physical condition, location information, and so on, which needs protection.

Key words: BAN’s, Privacy, Differential Privacy, Haptic Technology
1. INTRODUCTION
The differential privacy provides information about the database while simultaneously ensuring very high levels of privacy. Differential Privacy is a method enabling analysts to extract useful answers from databases containing personal information while offering strong individual privacy protections. It aims to minimize the chances of individual identification while querying the data. The method of differential privacy is shown in below fig.1

Steps:

- Analyst sends a query to an intermediate piece of software which is called Differential privacy guard.
- Then the guard accesses the privacy impact of the query using a special algorithm.
- The guard sends the query to database and gets back a clean answer based on the data that has not been distorted in anyway.
- The privacy guard then adds the appropriate amount of noise which makes answer imprecise in order to protect confidentiality of individual information in the database and sends modified response back to the analyst.

As opposed to anonymization, data is not modified in differential privacy. Users don’t have direct access to the database. There is an interface that calculates the results and adds desired inaccuracies. It acts as a firewall. These inaccuracies are large enough that they protect privacy but small enough that the answers provided to analysts and researchers are still useful.

Differential privacy (DP) method prevents unwanted Re-identification and other privacy threats. In this model, Personal information in a large database is not modified. Differential privacy works by inserting an intermediary piece of software between the analyst and the database. The analyst never gets access to the contents of a database. The intermediary acts as a privacy-protecting screen and this serves as privacy guard. Differential privacy aims to provide accurate queries from statistical databases while minimizing the chances of identifying its records.

According to this definition, differential privacy is a condition of data release mechanism but not over the dataset itself. This means that for any two datasets that are similar to each other the differential privacy algorithm behaves approximately same on both the datasets. This scheme guarantees that presence or absence of an individual may not affect the final output of the algorithm.

1.1. An Example
Assume that a hospital has a database of patients with a potentially life-threatening disease. It is shown in Table 1. If we consider the above database of medical records D1 where each record is a pair (Name, X), X is a Boolean denoting whether a person has diabetes or not. Now a malicious user wants to find whether E has diabetes or not and if he knows in which row of E resides.

Now the malicious user is only allowed to use a particular form of query Qi that returns the partial sum of first ‘i’ rows of column X in the database. In order to find E’s
Differential Privacy in Big Data Analytics for Haptic Applications

diabetes status, the adversary executes Q5 (D1) and Q4 (D1) then computes their difference.

From the example, Q5 (D1) =3 and Q4 (D1) =2 so their difference is 1. If we construct database D2 by replacing E1 with E0 then this malicious user will be able to distinguish D2 from D1 by computing Q5 – Q4 for each dataset. If the adversary requires receiving values Qi via ε- differentially private algorithm then they cannot distinguish between two datasets.

The advantages of differential privacy over anonymization are:

- The original data set is not modified at all. There is no need for suppression or generalization.
- Distortion is added to the results by mathematical calculations based on the type of data, type of questions etc.

Assume that a hospital has a database of patients with a potentially life-threatening disease. The database contains a detailed record of the treatments each patient has undergone, dates and times for appointments, and records of prescriptions. It also contains general information about the patients, including places where they have lived in the past for five or more years. The hospital has deployed a DP guard for this database that keeps an eye out for patient privacy.

A researcher believes that the disease is more likely to manifest itself in people who have lived for long periods of time in certain regions, and wants to find out if the information in the database confirms this hypothesis. The researcher connects to the Differential Privacy guard and requests, for each town located in the suspected regions, the number of patients with the disease. The guard poses the question to the database and, when the answer comes back, it discovers that there are a significant number of individuals that lived in eight towns in the regions the researcher asked about, that there is one additional town, “Smallville,” for which the number of patients is one, and a number of other towns for which the number is zero. We will call the single patient from Smallville, Bob, and use Bob’s case to illustrate how DP works.

If, for instance, the researcher does other work at the hospital and because of this work has access to other, less detailed, patient records that show Bob recently moved in from Smallville, and she knows the town is located thousands of miles away from the hospital, she may reasonably conclude that Bob has the illness because it is unlikely that two people from such a small town would move such a long distance and end up in the same city and hospital at more or less the same time.

Whether the researcher finds this out by accident or through a deliberate analysis is not relevant; the important thing is that Bob’s privacy has been breached because he has been identified.

To avoid this situation, the DP guard will introduce a random but small level of inaccuracy, or distortion, into the results it serves to the researcher. Thus, instead of reporting one case for Smallville, the guard may report any number close to one. It could be zero, or ½ (yes, this would be a valid noisy response when using DP), or even -1.

The researcher will see this and, knowing the privacy guard is doing its job, she will interpret the result as “Smallville has a very small number of cases, possibly even zero.” In fact, and in order to maintain privacy, the guard may also report non-zero (but equally small) numbers for some of the towns that really have zero cases.
2. EXISTING SYSTEM

2.1. Existing Algorithm
1. **Input:** ECG datasets
2. **Output:** Transformed ECG datasets
3. Extract feature vector from scanned ECG dataset
4. Divide ECG feature equivalence class
5. Initialize interference threshold record of the characteristics
6. **While** There are data in equivalence class that have not been processed **do**
7. Select subroutine and add noise
8. **While** Differential privacy protection model is not satisfied **do**
9. Add noise to important features
10. **end** while
11. Update interference threshold record of the characteristics.
12. **end** while

**Algorithm:** ECG privacy protection algorithm

3. PROPOSED SYSTEM

3.1. Proposed Algorithm
1. **Input:** Heart Disease datasets
2. **Output:** Cleansed Heart disease datasets
3. Calculate information gain ratio
4. Sum up tuples row wise
5. Calculate mean and standard deviation
6. **While** Data is in normal distribution **do**
7. Perform data perturbation
8. **While** Differential privacy protection model is not satisfied **do**
9. Perform data perturbation to important features
10. **end** while
11. **end** while

**Algorithm:** ECG privacy protection algorithm

3.2. Algorithm Explanation
1. We input Heart disease dataset to the algorithm.
2. We get cleansed Heart disease dataset as output.
3. We now calculate Information gain ratio (IGR) by using the following formula
   \[ IGR (D, x) = \frac{IG}{IV} \]
   Where IG is Information Gain and IV is intrinsic values
(i) Formula we used to calculate Information Gain is as follows

\[
InfoGain (D, x) = H(D) - \sum_{y \in w(x)} \frac{|x \in D | x_i = y|}{|D|} \times H(\{x \notin T | x_i = y\})
\]
(ii) Formula we used to calculate intrinsic value is as follows:

\[
\text{Intrinsic Value } IV(D, x) = - \sum_{v \in \text{val}(x)} \frac{|\{x \in D \mid x_i = v\}|}{|D|} \log_2 \left( \frac{|\{x \in D \mid x_i = v\}|}{|D|} \right)
\]

(iii) We now calculate information Gain Ratio as follows:

4. Sum up tuples row wise
5. Calculate Mean and Standard deviation by using the following formula:

\[
\text{Mean } \bar{x} = \frac{\sum x}{N}
\]

\[
\text{Standard deviation } = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}
\]

In this data set the mean is \( \frac{723172}{1212} = 596.6766 \) and the standard deviation is \( = 62.6783 \)
6. If the data is in Normal distribution then we perform the following step
7. We add noise to the data, adding noise to data involves many mechanisms out of which we used Gaussian Mechanism to noise. The function we used for adding Gaussian noise is:

\[
p_G(z) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}
\]

Because of limits of differential privacy protection model the noise generated with too much of intensity will be discarded.
8. While differential privacy protection model is not satisfied then add noise to important features.

4. RESULTS
The results generated through this algorithm are shown in Fig 3(a), Fig 3(b). As the graphs generated before and after perturbations are approximately same we can say that differential privacy scheme is satisfied.

5. FIGURES AND TABLES

![Figure 1 Working of differential Privacy](image)

http://www.iaeme.com/IJCET/index.asp 15  
editor@iaeme.com
### Table 1 (a) Information Gain values

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Information Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>0.9047</td>
</tr>
<tr>
<td>Cp</td>
<td>1.7344</td>
</tr>
<tr>
<td>Fbs</td>
<td>0.6061</td>
</tr>
<tr>
<td>Restecg</td>
<td>1.0879</td>
</tr>
<tr>
<td>Exang</td>
<td>0.9116</td>
</tr>
<tr>
<td>Slope</td>
<td>1.2941</td>
</tr>
<tr>
<td>Ca</td>
<td>1.5571</td>
</tr>
<tr>
<td>Thal</td>
<td>1.2456</td>
</tr>
<tr>
<td>Num</td>
<td>1.8459</td>
</tr>
</tbody>
</table>

### Table 2 (b) Intrinsic values

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Intrinsic value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>0.9045</td>
</tr>
<tr>
<td>Cp</td>
<td>1.7341</td>
</tr>
<tr>
<td>Fbs</td>
<td>0.2111</td>
</tr>
<tr>
<td>Restecg</td>
<td>0.0893</td>
</tr>
<tr>
<td>Exang</td>
<td>0.9115</td>
</tr>
<tr>
<td>Slope</td>
<td>1.2939</td>
</tr>
<tr>
<td>Ca</td>
<td>1.5571</td>
</tr>
<tr>
<td>Thal</td>
<td>1.2455</td>
</tr>
<tr>
<td>Num</td>
<td>1.8418</td>
</tr>
</tbody>
</table>

### Table 3 (c) Information Gain Ratio values

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Information Gain Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>1.0002</td>
</tr>
<tr>
<td>Cp</td>
<td>1.0002</td>
</tr>
<tr>
<td>Fbs</td>
<td>2.8712</td>
</tr>
<tr>
<td>Restecg</td>
<td>12.1825</td>
</tr>
<tr>
<td>Exang</td>
<td>1.0001</td>
</tr>
<tr>
<td>Slope</td>
<td>1.0002</td>
</tr>
<tr>
<td>Ca</td>
<td>0.9981</td>
</tr>
<tr>
<td>Thal</td>
<td>1.0001</td>
</tr>
<tr>
<td>Num</td>
<td>1.0022</td>
</tr>
</tbody>
</table>
6. CONCLUSION
In this paper, we proposed a differential privacy protection model for sensitive big data in BANs, which significantly reduces the risk of privacy exposure and greatly ensures the availability of data. Even in the case where the attacker has full knowledge of the background, it can still produce enough interference to data, which can make it unable to find matching. In this we have the concept of dynamic noise thresholds to demonstrate the relationship between the added noise and data set size, making our scheme more suitable for processing sensitive big data in BANs. At last, several experiments are conducted to show the performance of our scheme. Experimental results reveal that our scheme offers promising protection effects. Even in the case where the attacker has full knowledge of the background, it can still produce enough interference to data, which can make it unable to find matching a certain victim from the Heart disease data to protect the privacy.
REFERENCES


Differential Privacy in Big Data Analytics for Haptic Applications


