Comparison and Analysis of Anonymization Techniques for Preserving Privacy in Big Data

Johny Antony P
Research Scholar
NGM College, Pollachi, Tamilnadu, India.

Dr. Antony Selvadoss Thanamani
Head, Department of Computer Science
NGM College, Pollachi, Tamilnadu, India.

Abstract

Modern technology and networking generates huge volume of data. Privacy of data is a crucial issue and a topic for significant research. Data publishing faces the problem of deciding how to publish useful data while preserving privacy-sensitive information according to the privacy requirements of data holders. According to the concept of the privacy protection, it is defined as such the accessing of published data must not allow the unwanted users to identify anything about the targeted individuals. This paper presents a classification and analysis of various anonymization techniques for privacy preservation like k-anonymity, l-diversity, t-closeness, differential privacy, slicing.

Keywords: k-anonymity, l-diversity, t-closeness, differential privacy, slicing, data utility

1. INTRODUCTION

Present information technology creates vast amount of data characterized by velocity, volume, veracity. Sharing and dissemination of this data gives rise to the violation of privacy of individuals who are the subjects of the data. Privacy protection is one of most important issue in big data processing. Customer privacy is an issue that attracts
considerable attention from both academia and IT industry [2]. The question that comes to the mind is that can share the data while protecting the privacy and at the same time providing the data utility. It opens new avenues for research and study as privacy is a right of an individual. The primary goal is to extract the hidden wisdom and knowledge from the huge amount of data at the same time sensitive data should not be misused. Despite the use of big data for innovation and insights, the massive amount of data can breach the privacy of users. In order to preserve privacy of data several mechanisms have been proposed and developed in the recent years. It is a great challenge to keep balance between data utility and data privacy. [3]

This paper presents the basic models of privacy preservation, their comparative study and performance with regard to execution time, data utility and privacy preservation.

2. RELATED WORK


3. BASIC MODEL IN PRIVACY PRESERVATION

3.1. k-anonymity

k-anonymity is one of the basic privacy preservation model. In the k-anonymity, every published record has to be indistinguishable from at least (k-1) others on its QI attribute. The "quasi-identifiers" are the attributes available to an adversary. It is defined as: A table T satisfies k-anonymity if for every tuple \( t \in T \) there exist \( k-1 \) other tuples \( t_{i1}, t_{i2}, \ldots, t_{ik-1} \in T \) such that \( t[C] = t_{i1}[C] = t_{i2}[C] = \ldots = t_{ik-1}[C] \) for all \( C \in Q \). [17].

3.2. l-diversity

l-diversity is a group based anonymization model that assists to preserve the privacy of data through reducing the granularity of a data representation using generalization and suppression. In l-diversity, an equivalence class is said to have l-diversity if there is at least l "well-represented" value for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity. It is defined as: A q block is \( \ell \)-diverse if it contains at least \( \ell \) “well-represented” values for the sensitive attribute S. A table is \( \ell \)-diverse if every q block is at least \( \ell \)-diverse [6,3].
3.3. t-closeness

t-closeness is another group based privacy model that extends the l-diversity model. It treats the values of an attribute distinctly, and considers the distribution of data values of the attribute to preserve the privacy. It uses the Earth Mover Distance (EMD) function to compute the closeness between two distributions of sensitive values. It is defined as: An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is not more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness [5].

3.4. Differential Privacy

Differential Privacy offers one way forward that to extract insights from a database while guaranteeing that no individual can be identified. It achieves the guarantee of privacy by adding noise to answer to the queries. The amount of noise added must be large enough to conceal the effect of individuals and small enough that does not distort the genuineness of the answer. It is defined as: Let databases (D, D’) differing only in one row, meaning one is a subset of the other and the larger database contains just one additional row. A randomized function K gives ε-differential privacy if for all data sets D and D’ differing on at most one row, and all S ⊆ Range(K) [7].

\[ Pr[K(D) \in S] \leq \exp(\varepsilon) \times Pr[K(D') \in S] \]

3.5. Slicing

It is a technique that partitions data horizontally and vertically. The basic idea is to break the association between cross columns but to preserve the association within each column. Slicing preserves data more accurately. [14].

Table 2. Comparison of Privacy Models

<table>
<thead>
<tr>
<th>No</th>
<th>Model</th>
<th>Merits</th>
<th>Demerits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Randomization</td>
<td>Simple method that can be easily implemented</td>
<td>Difficulty for multiple attributes and categorical attributes</td>
</tr>
<tr>
<td>2</td>
<td>k-anonymity</td>
<td>Easy to implement, Chance for Re-identification is less when the value of k is high.</td>
<td>It fails in preventing the background knowledge and homogeneity attacks, Suffers from attribute linkage and record linkage, Long processing time, Utility may be compromised that any query returns minimum of k matches.</td>
</tr>
<tr>
<td>3</td>
<td>l-diversity</td>
<td>Reduce the data set into summary form. Sensitive attribute would have at most same frequency.</td>
<td>Depends upon the range of sensitive attributes. For l diverse, there should be l different values of sensitive attribute. It is prone to skewness and similarity attack and may not prevent</td>
</tr>
<tr>
<td></td>
<td></td>
<td>attribute disclosure. Vulnerable to homogeneity attach and back ground knowledge attack.</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>t-closeness</td>
<td>Prevent data from skewness attack. Complex computational procedure to enforce t-closeness. It looses the co relation between different attributes since each attribute is generalized separately. Utility is damaged when t is very small</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Differential Privacy</td>
<td>Most suitable for big data. Provides strongest privacy guarantee. Data utility may be reduced. Data miner is only allowed to pose aggregate queries. Probability of attacking both the databases by adversary is not taken into consideration.</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Slicing</td>
<td>Randomization on sensitive attributes. Prevents attribute disclosure. Utility and risk measure is not matched. It may break association between attributes.</td>
<td></td>
</tr>
</tbody>
</table>

**Architecture**

4. EXPERIMENTAL RESULTS

Experiments were carried out on adult data set taken form UC Irvine Machine Learning Repository - UCI Machine Learning. (https://archive.ics.uci.edu/ml/datasets.html). The data set contains 48842 instances with 14 attributes both categorical and integer. The data contains sensitive and non sensitive (quasi identifier) attributes. The data was cleansed and formatted and made into sets of 40000, 80000, 160000, 320000 and 640000 with random replication. The
experiments are conducted on a machine with Intel® Core TM i5-2120 CPU @ 3.30 GHZ, 4 GB RAM, Window 7, JAVA –JDK 8.0.

The objective of the experiment is to find out performance metrics such as execution time, data utility and privacy of the various privacy preservation models applied to big data.

4.1. Execution Time

The following table shows the execution time – the time taken by the algorithm to perform the task by various models with different data size.

<table>
<thead>
<tr>
<th>Sl. No</th>
<th>Models/ Data Size</th>
<th>40000</th>
<th>80000</th>
<th>160000</th>
<th>320000</th>
<th>640000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>k-anonymity</td>
<td>105</td>
<td>200</td>
<td>410</td>
<td>800</td>
<td>1450</td>
</tr>
<tr>
<td>2</td>
<td>l-diversity</td>
<td>130</td>
<td>240</td>
<td>430</td>
<td>860</td>
<td>1600</td>
</tr>
<tr>
<td>3</td>
<td>t-closeness</td>
<td>170</td>
<td>300</td>
<td>600</td>
<td>1300</td>
<td>2700</td>
</tr>
<tr>
<td>4</td>
<td>Differential privacy</td>
<td>160</td>
<td>280</td>
<td>590</td>
<td>1250</td>
<td>2500</td>
</tr>
<tr>
<td>5</td>
<td>Slicing</td>
<td>150</td>
<td>275</td>
<td>560</td>
<td>1200</td>
<td>2400</td>
</tr>
</tbody>
</table>

![Execution Time Graph](image-url)
4.2. Data Utility & Complexity

Data utility is measured by the accuracy of the queries MIN, MAX, COUNT on the original data and the transformed data after applying the privacy preserving techniques.

<table>
<thead>
<tr>
<th>Sl. No</th>
<th>Models</th>
<th>Data Utility</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>k-anonymity</td>
<td>low</td>
<td>Very Low</td>
</tr>
<tr>
<td>2</td>
<td>l-diversity</td>
<td>high</td>
<td>Low</td>
</tr>
<tr>
<td>3</td>
<td>t-closeness</td>
<td>high</td>
<td>Very high</td>
</tr>
<tr>
<td>4</td>
<td>Differential privacy</td>
<td>medium</td>
<td>high</td>
</tr>
<tr>
<td>5</td>
<td>Slicing</td>
<td>medium</td>
<td>high</td>
</tr>
</tbody>
</table>

5. CONCLUSION

This paper gives a view about the basic model of privacy preservation and its effect when applied to big data. It also present the merits and demerits of each model for preserving privacy in data. An experimental result is also given in relation to execution time, implementation complexity and data utility.

REFERENCE


