Privacy Preserving in Data Mining by Normalization

Syed Md. Tarique Ahmad  
Dept. of Computer Science  
Pacific University, Udaipur  
Rajasthan, India

Shameemul Haque  
Dept. of Computer Science  
King Khalid University, Abha, KSA

Prince Shoeb Khan  
Dept. of Computer Science  
King Khalid University, Abha, KSA

ABSTRACT
Extracting previously unknown patterns from massive volume of data is the main objective of any data mining algorithm. In current days there is a tremendous expansion in data collection due to the development in the field of information technology. The patterns revealed by data mining algorithm can be used in various domains like Image Analysis, Marketing and weather forecasting. As a side effect of the mining algorithm some sensitive information is also revealed. There is a need to preserve the privacy of individuals which can be achieved by using privacy preserving data mining. In this paper we use min-max normalization approach for preserving privacy during the mining process. We clean the original data using min-max normalization approach before publishing. For experimental purpose we have used k-means algorithm and from our results it is obvious that our approach preserves both privacy and accuracy.

Keywords
Clustering, K-Means, Accuracy, Privacy, Min-Max Normalization, Normalization.

1. INTRODUCTION
Since we are in an era of information explosion, it is very important to be able to find out useful information from massive amounts of data. Consequently, various data mining techniques have been developed. Data mining is often applied to fields such as marketing, sales, finance, and medical treatment. Besides, the rapid advance in Internet and communications technology has led to the emergence of data streams. Due to the consecutive, rapid, temporal and unpredictable properties [1,2] of data streams, the study of data mining techniques has transformed from traditional static data mining to dynamic data stream mining.

In recent years, enabled by the rapid development of various telecommunication technologies, many companies have improve their competitive edge by forming strategic alliances or information outsourcing, one after another. Consequently, many companies frequently expose private data while engaging in data analysis activities, which has led to grave threats to data privacy. For example, online marketing companies usually employ information technology outsourcing with a data mining company for cluster mining, in order to earn greater profits and to find the best target groups of customers. Therefore, how to preserve private data without disclosure while obtaining an accurate mining result in the process of mining will become increasingly difficult, which in turn has led to the development of Privacy-Preserving Data Mining techniques.

Nonetheless, traditional Privacy-Preserving Data Mining is not applicable in a data stream environment which requires dynamic updating. For example, for a massive amount of income data, the execution efficiency of traditional methods can no longer respond to user demand. Furthermore, the potential infinite number of data streams plus limited memory space has constrained the traditional methods from obtaining the mining result with accuracy. In view of the above-mentioned issues, studies on Privacy-Preserving Data Stream Mining in recent years have become one of the important issues in the field of data mining.

Several privacy preserving algorithms have been proposed and are used nowadays. In this paper, we propose a new method using min-max normalization for preserving data through data mining. In general, min–max normalization is used as a preprocessing step in data mining for transformation of data to a desired range. Our purpose is to use it for preserving privacy through data mining. We use K-means clustering to validate the proposed approach and validate for accuracy.

The rest of the paper is organized as follows: Section 2 provides an overview of literature review carried out in clustering techniques; Section 3 elaborates the implementation of min-max normalization and K-mean clustering techniques in our proposed system. Experimental results and simulations are tabulated and compared in Section 4 and finally in Section 5, we arrive to an overall conclusion from our work.

2. LITERATURE SURVEY
The study of Privacy-Preserving Data Mining techniques started extensively since 2000 [3], covering development approximately in two categories: Perturbation-Base technique [3, 4] and Secure Multi-Party Computation Base technique [5, 6]. The main idea of Perturbation-Based technique involves increasing a noise in the raw data in order to perturb the original data distribution and to preserve the content of hidden raw data. Geometric Data Transformation Methods (GDTMs) [7] is one simple and typical example of data perturbation technique, which perturbs numeric data with confidential attributes in cluster mining in order to preserve privacy. Nonetheless Kumari et al. [4] proposed a privacy preserving clustering technique of Fuzzy Sets, transforming confidential attributes into fuzzy items in order to preserve privacy. Furthermore, the largest issue encountered when implementing a perturbation technique is the inaccurate mining result from a perturbed data. In view of this issue, the technique of Random-data perturbation introduced by Agrawal and Skrikant [3] was the first study addressed. Whereas the technique derives the original data distribution using a random noise for data distribution, and constructs a result similar to the original data, it finally use this similar result to execute mining. This method could construct a more accurate data mining model, while reducing mining errors. In addition, usually the perturbation technique that has higher privacy preservation comes with a lower level of mining accuracy, whereas most of the perturbation techniques today belong to the one-size-fits-all and are relatively inflexible. To resolve this issue, Liu and Thuraisingham [8] developed the two-phase perturbation technique which frames different intervals
3. PROPOSED SYSTEM

In this paper we put forward an approach for privacy preserving using min-max Normalization.

3.1 Min-Max Normalization

Min-max normalization performs a linear transformation on the original data. For mapping a value, v of an attribute A from range \([\min_A, \max_A]\) to a new range \([\text{new}_{\min A}, \text{new}_{\max A}]\), and the computation is given by:

\[
\frac{v - \min A}{\max A - \min A} \left( \text{new}_{\max A} - \text{new}_{\min A} \right) + \text{new}_{\min A}
\]

Where \(v\) is the new value in the required range.

The main advantage of Min-Max normalization is that it preserves the relationships between the original data values [18].

Table 1 is the sample data set used for experiment. Table 2 is the corresponding normalized values for the “Age” attribute after applying min-max normalization. The steps involved in our approach can be summarized in the form of a procedure as shown below. Fig. 1 shows the below diagram for the proposed system.

![Diagram](image-url)
3.2 Procedure Steps

![Diagram of Procedure Steps]

**Fig 2: Step by Step Procedure**

<table>
<thead>
<tr>
<th>S No.</th>
<th>Name</th>
<th>Age</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Salman</td>
<td>2</td>
<td>M</td>
</tr>
<tr>
<td>2</td>
<td>Sonia</td>
<td>10</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>Bushra</td>
<td>20</td>
<td>F</td>
</tr>
<tr>
<td>4</td>
<td>Rajeev</td>
<td>25</td>
<td>M</td>
</tr>
<tr>
<td>5</td>
<td>Sharmi</td>
<td>12</td>
<td>F</td>
</tr>
<tr>
<td>6</td>
<td>Dhoni</td>
<td>30</td>
<td>M</td>
</tr>
<tr>
<td>7</td>
<td>Abu</td>
<td>20</td>
<td>M</td>
</tr>
</tbody>
</table>

**Table 1. Original data**

<table>
<thead>
<tr>
<th>S No.</th>
<th>Name</th>
<th>Age</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Salman</td>
<td>10</td>
<td>M</td>
</tr>
<tr>
<td>2</td>
<td>Sonia</td>
<td>33</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>Bushra</td>
<td>62</td>
<td>F</td>
</tr>
<tr>
<td>4</td>
<td>Rajeev</td>
<td>76</td>
<td>M</td>
</tr>
<tr>
<td>5</td>
<td>Sharmi</td>
<td>39</td>
<td>F</td>
</tr>
<tr>
<td>6</td>
<td>Dhoni</td>
<td>90</td>
<td>M</td>
</tr>
<tr>
<td>7</td>
<td>Abu</td>
<td>62</td>
<td>M</td>
</tr>
</tbody>
</table>

**Table 2. Normalized Data**

3.3 Clustering Technique

In our methodology, in order to check for the effectiveness of min-max normalization on the data partitioning techniques, we used K-means clustering algorithm. In K-means, the objects are clustered based on attributes into ‘n’ number of clusters where ‘n’ is a positive integer. The central idea of this Clustering is to minimize the sum of squares of the distance between data and corresponding cluster centroid in that data set. The clustering process must be carried out until it gets stabilized. Then, the objects are grouped based on the inter-relative distance among each object and the centroid.

4. RESULTS AND SIMULATIONS

In this paper, we have used min-max normalization to achieve privacy and accuracy during data mining and accuracy is tested using K-means clustering. Here the computations for min-max normalization of sample data, k-means clustering and effectiveness calculations are carried out in C++.

We have also tested the efficiency of our approach on a real time dataset, “adult-dataset” from UCI data repository [19]. This data set comprises of about 32561 records with 12 attributes namely age, work class, education, marital status, occupation, relationship, race, sex, capital gain and capital loss, hours per week and native country. For experimental purpose, we used only age as the key attribute to carry out normalization in our work.

The clustering of data before and after normalization for 2-clusters is given in the figures Figure 3 and Figure 4 respectively. Similarly, those for 3-clusters are provided in the figures Figure 5 and Figure 6. Table 3 and Table 4 describe the clustering of data before and after min-max normalization for 2-clustering and 3-clustering respectively.

Table 5 and Figure 7 summarize the comparisons among Fuzzy S-Shaped approach, Shearing Noise addition approach and our proposed min-max normalization approach.
Fig 3: Snapshot for 2 clusters After Normalization.

Fig 4: Snapshot for 2 clusters before Normalization.

Fig 5: Snapshot for 3 clusters before Normalization.

Fig 6: Snapshot for 3 clusters after Normalization.

Table 3. Results of 2 clusters

<table>
<thead>
<tr>
<th>K=2</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before Normalization</td>
<td>{2,4,10,12,3,11}</td>
<td>{20,30,25}</td>
</tr>
<tr>
<td>After Normalization</td>
<td>{10,16,33,39,13,36}</td>
<td>{62,90,76}</td>
</tr>
</tbody>
</table>

Table 4. Results of 3 clusters

<table>
<thead>
<tr>
<th>K=3</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before Normalization</td>
<td>{2,4,3}</td>
<td>{10,12,11}</td>
<td>{20,30,25}</td>
</tr>
<tr>
<td>After Normalization</td>
<td>{10,16,13}</td>
<td>{33,39,36}</td>
<td>{62,90,76}</td>
</tr>
</tbody>
</table>

Table 5. Results of 3 clusters

<table>
<thead>
<tr>
<th>Original Data</th>
<th>Min-Max Normalization</th>
<th>Fuzzy S-Shaped</th>
<th>Shearing Noise (=10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>10</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>0.0025</td>
<td>33</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>0.0102</td>
<td>44</td>
</tr>
<tr>
<td>10</td>
<td>33</td>
<td>0.1632</td>
<td>110</td>
</tr>
<tr>
<td>11</td>
<td>36</td>
<td>0.2066</td>
<td>121</td>
</tr>
<tr>
<td>12</td>
<td>39</td>
<td>0.2551</td>
<td>132</td>
</tr>
<tr>
<td>20</td>
<td>62</td>
<td>0.7449</td>
<td>220</td>
</tr>
<tr>
<td>25</td>
<td>76</td>
<td>0.9362</td>
<td>275</td>
</tr>
<tr>
<td>30</td>
<td>90</td>
<td>1</td>
<td>330</td>
</tr>
</tbody>
</table>

From the above comparisons we observe that data transformation approach based on shearing, scales the values and scatters them over a large range. On the other hand, Fuzzy approach based on S-Shaped membership function narrows down the range of values to (0, 1). The produced results of both the approaches prove evidently the duplication of data for privacy preservation. Our approach overcomes this limitation as the normalized values lie in the same range as the actual range of the attribute. Thus, the distortion of data
for sake of privacy will not be revealed to the analyst or data-users, at the same time preserving privacy. Fig. 3 shows the comparison graph.

![Comparison Graph](image-url)

**Fig 5: Comparison Graph.**

5. CONCLUSION
In this paper we have dealt with min-max normalization based data transformation to preserve data privacy. This approach transforms the original data to privacy-preserved data maintaining the inter relative distance among the data. Experiments have proven that performing k-means clustering on the distorted data produces similar clustering results as original data. Thus we have succeeded in achieving both accuracy and privacy. We have tested the technique for numerical data set. The future scope of this paper is to extend the same over categorical data.

6. REFERENCES