Survey on Clustering based Categorical Data Protection

Amrutha HJ1, Anu A Kittur2, Chaitra MS3, Gowri M4, Sowmya SR5

1,2,3,4BE Student, Department of Information Science and Engineering
5Professor, Dept. of ISE, Dayananda Sagar Academy of Technology & Management, Karnataka, India

Abstract - The amount of publicly accessible datasets is rising every day in the present age. Improving data privacy erore becomes mandatory. This has become a major reason why prolonged research has been undertaken to deliver effective fortification techniques that obstruct the revelation of entities in the datasets by conserving the data utility. A comprehensive attachment for categorical data protection is carried out by applying clusters to the dataset and then safeguarding every data segment.

Key Words: Categorical Data, Clustering, Data mining, Data privacy

1. INTRODUCTION

Providing the requisite privacy is the main agenda to protect the data or information. All the clients who entered the data would expect their data to be protected. Data mining is a method in which it transforms the base data to finished data.

It is approach that calls for and examines the vast quantity of dat collected to obtain trends. Categorical data can also be known as statistical data consisting of categorical values.

There are three major attributes to reflect when consideringa dataset, namely confidential, identifiers and Quasiidentifiers. Quasiidentifiers are pieces of information with some degree of uncertainty that are not by themselves distinct identifiers.

In the case of confidential attributes, it includes information of employment, health issues or religion. Clustering can be defined as the process in which the abstract objects become an interconnected class of objects within the set. The study of clustering takes into account in applications such as market survey, data-analysis, pattern recognition and image processing.

Protection approaches are tested on the basis of two important measures they threaten the loss and disclosure of information.

The information loss is calculated by comparing the statistical parameter between the anonymousone and the original data table. Security approaches can be classified into two general categories: disruptive and non-perturbatory. Perturbative is a technique for changing the attribute's sensitive value via a new value.

NonPerturbative technique does not change the attribute's sensitive value, rather it attribute's sensitive value, rather it suppresses or deletes certain datasets.

2. METHODOLOGY

2.1 Subtractive Clustering:

The currently in effect subtractive clustering approach can be used only for numerical data that cannot be used for data with categorical values. Many cluster grids have a maximum value in the conventional mountain-clustering process. But this mountain clustering approach can sometimes trigger the computation's increasing complexity, so one subtractive method to clustering has been proposed. This approach can be used only in numerical data since there is no natural ordering of the categorical data. Though clustering using kmeans gives better efficiency, subtractive clustering is powerful.

2.2 Robust Hierarchical Clustering (RHC):

Hierarchical clustering is the popular unsupervised technique used for the Metabolomics data. In the case of conventional hierarchical clustering system, it is highly reactive to outliers and if there is the existence of misleading clustering tests, those outliers exist. Two Stage Generalized S-estimator (TSGS) is used to robustify hierarchical clustering which allows use of the covariance matrix.

There are 3 major steps in robust hierarchical data segmentation methodology.

1. Estimation of Robust covariance matrix:

The biggest hurdle here is to estimate an appropriate matrix of correlation or dispersion at a time in the presence of cell-wise anomalies or outliers in case-wise and cell-wise.

2. Robust evaluation of correlation matrix based on dissimilarity using the TSGS covariance matrix.

3. Estimate of RHC proposed with TSGS dispersion matrix.

2.3 Decision Tree Categorical Value Clustering

Data breakdown methods add noise to the data to avoid correct confidential values being revealed. Categorical values
of attributes are clustered in the beginning, and these clusters are then used in the later stages to create noise.

Categorical value clustering and disruption technique of the decision-tree disturbs a non-class categorical feature of a dataset. Therefore, we apply it once for each non-class attribute specified on the original dataset to agitate all non-class categorical attributes. Every time a dataset is generated with one disturbed attribute within it. Lastly, we construct a dataset (combining all disturbed data sets) where each non-class categorical attribute is disturbed and all other attributes s are not disturbed.

2.4 Outlier Diagnosis:
Outlier is one that does not adhere to the pattern in the dataset or any other feature expected. This may be diagnosed using anomaly detection methods. These phenomena can also be called outliers, novelties, noise, or variations.

They come in three different types:
1. Supervised anomaly detection
2. Semi supervised anomaly detection
3. Unsupervised anomaly detection

Unmonitored detections of anomalies identify anomalies in an unlabeled test data data set under which the data collection standard of events is considered normal by searching for instances that appear to conform to the rest of the data set at least.

2.4.1 Outlier Detection Techniques:
A. Statistical outlier detection:
It calculates the arguments in the case of statistical distribution by imagining all the data points produced by statistical dispersion

B. Depth based outlier detection:
Depth based search originality at data space cap for outlier detection. They’re autonomous regarding statistical data distribution.

C. Distance based outlier detection:
This judges a point based on separation of neighborhoods.

D. Density based outlier detection:
It practices the distribution of data element density into the set of data.

E. Deviation based outlier detection:
The data components are scattered as a sparse matrix in the data set which creates confusion over the analysis of results. When departing from standard points some points are considered anomalies.

Table 1: Comparison table for outlier algorithms

<table>
<thead>
<tr>
<th>Approach</th>
<th>Outlier detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical outlier detection</td>
<td>70%</td>
</tr>
<tr>
<td>Depth based outlier detection</td>
<td>65%</td>
</tr>
<tr>
<td>Deviation based outlier detection</td>
<td>80%</td>
</tr>
<tr>
<td>Distance based outlier detection</td>
<td>84%</td>
</tr>
</tbody>
</table>

2.5 Evolutionary Optimization Approach
A progressive accession to protection of data is based on an evolutionary algorithm, driven by the amalgamation of loss in information and threat disclosure procedures. This algorithm is dedicated to discover precise or approximate results to simplify or explore problems. The algorithm uses two simple genetic operators: mutation and crossover. It uses state-of-the-art techniques for categorical stability.

Mutation: The pieces are randomly arranged to obtain a new offspring in case of mutation.

Crossover: Consists of 2 chromosomal recombined values which also produce two new off springs.

2.6 L-Diversity
The anonymity models through generalization can shield the confidentiality of individuals but often lead to information loss. (K,l,al)-variety diminishes knowledge loss and ensures data quality. This method ensures data privacy even without the knowledge of the opponent’s background to avoid disclosure of attributes. In this case sensitive attributes are well represented. That technique is k-anonymity modification. A definition from a set of n records (k, l, range) diversity is used in such a way that the data segment cluster includes at least k (k = n) data elements as well as at least 1 dissimilar sensitive characteristics and the sum of all intra cluster distance is reduced.
3. RESULT COMPARISON

3.1 Clustering Algorithms

Table 2: Comparison table for clustering algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Benefit</th>
<th>Drawback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subtractive clustering</td>
<td>There is an efficient method in this case using. On numerous UCI datasets, a few investigations are carried out, and some experimental results describe that the approach given can attain better clustering precision when compared to k-modes algorithm.</td>
<td>Unsupervised clustering is not clear.</td>
</tr>
<tr>
<td>Robust hierarchical clustering</td>
<td>Simulation training clearly shows that the anticipated approach improves performance considerably over conventional hierarchical clustering</td>
<td>1. The preceding step cannot be undone. 2. Complexity of time: Not suitable for large datasets.</td>
</tr>
</tbody>
</table>

Evolutionary Optimization Approach | We perform better for advanced dimensional failures. | We are robust in terms of noisy valuation functions that do not reap any sensible outcome in a given stipulated amount of time.

3.2 Outlier Algorithms

Table 3: Comparison table for outlier algorithms

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Techniques / algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computation cost</td>
<td>Cluster based</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Low</td>
</tr>
<tr>
<td>High-dimensional data</td>
<td>Applicable</td>
</tr>
<tr>
<td>Complexity</td>
<td>Less complex</td>
</tr>
</tbody>
</table>

3.3 Protection Algorithms

Table 4: Comparison table for outlier algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-Diversity</td>
<td>1. Makes distribution more robust</td>
<td>1. This can be redundant and laborious to within the category of critical attributes, thereby increasing data protection. 2. Protects from disclosing attribute.</td>
</tr>
</tbody>
</table>

4. CONCLUSIONS

In this paper, a new approach is used to deal with categorical data confidentiality using the SCCA algorithm clustering technique, which can result in more contented clustering accuracy than the obsolete kmodes algorithm on each collection. The efficiency of TSGS algorithms is greater than that of robust estimation techniques.

Ldiversity will intensify the privacy of the defendant but this function is not sufficient to protect critical attributes. Hence, evolutionary optimization strategy is a better method of defense.

REFERENCES


