International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

Survey on Clustering based Categorical Data Protection

Amrutha HJ[1], Anu A Kittur[2], Chaitra MS[3], Gowri M[4], Sowmya SR[5]

[1,2,3,4]BE Student, Department of Information Science and Engineering
[5]Professor, 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 erefore 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 attachement 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 diagonalised 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 atleast .

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

Approach	Outlier detection
Statistical outlier detection	78%
Depth based outlier detection	85%
Deviation based outlier detection	80%
Distance based outlier detection	84%

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 a 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 COMPARISION

3.1 Clustering Algorithms

Table 2: Comparison table for clustering algorithms

Algorithm	Benefit	Drawback
Subtractive clustering	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.	Unsupervised clustering is not clear.
Robust hierarchical clustering	Simulation training clearly shows that the anticipated approach improves performance considerably over conventional hierarchical clustering	1. The preceding step cannot be undoed. 2. Complexity of time: Not suitable for large datasets.
	within the category of critical attributes, thereby increasing data protection 2. Protects from disclosing attribute.	achieve. 2. Prone to attacks such as skewness attack.
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

Parameters	Techniques / algorithms.		
	Cluster based	*Distance based*	*Density based*
Compu-tation cost	Low	Low	High
Efficiency	Very efficient	Efficient	Efficient
High-dimensional data	Applicable	Applicable	Applicable.
Complexity	Less complex	Moderately Complex	Highly complex

3.3 Protection Algorithms

Table 4: Comparison table for outlier algorithms

Algorithm	Advantage	Disadvantage
L-Diversity	1. Makes distribution more robust	1. This can be redundant and laborious to

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

[1]H. Zhao and Z. Qi, "Hierarchical Agglomerative Clustering with Ordering Constraints," 2010 Third International Conference on Knowledge Discovery and Data Mining, Phuket, 2010, pp. 195-199. doi:10.1109/WKDD.2010.123

[2] Lei Gu, "A novel locality sensitive k-means clustering algorithm based on subtractive clustering," 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, 2016, pp. 836-839. doi:10.1109/ICSESS.2016.7883196

[3] Jiang Chundong, Jia Haipeng, Du Taihang, Zhang Lei and Chunbo Jiang, "Evolutionary algorithm and its application in structural topology optimization," 2008 27th Chinese Control Conference,Kunming,2008, pp.10-14. doi:10.1109/CHICC.2008.4605057

[4] Mar J., Torra V. (2012) Clustering-Based Categorical Data Protection. In: Domingo-Ferrer J., Tinnirello I . (eds) Privacy in Statistical Databases PSD 2012.Lecture Notes in Computer Science,vol 7556.Springer,Berlin,Heidelberg

[5] Wanliang Fu, "Multi-media data mining technology for the systematic framework," 2012 IEEE International Conference on Computer Science and Automation Engineering, Beijing, 2012, pp. 570-572. doi:10.1109/ICSESS.2012.6269531

[6] H. C. Mandhare and S. R. Idate, "A comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS),Madurai,2017,pp.931-935.

[7] B. M. Varghese and U. A., "Recursive Decision Tree Induction Based on Homogeneousness for Data Clustering," 2008 International Conference on Cyberworlds, Hangzhou,2008,pp.754-758.
doi:10.1109/CW.2008.56

[8] Han Jianmin, Cen Tingting and Yu Juan, "An l-MDAV microaggregation algorithm for sensitive attribute l-diversity," 2008 27th Chinese Control Conference, Kunming, 2008, pp. 713-718. doi:10.1109/CHICC.2008.4605421

[9] S. Banerjee, A. Choudhary and S. Pal, "Empirical evaluation of K-Means, Bisecting K-Means, Fuzzy C-Means and Genetic K-Means clustering algorithms," 2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), Dhaka, 2015, pp. 168-172. doi:10.1109/WIECON-ECE.2015.7443889

[10] Fayyoumi and O. Nofal, "Applying Genetic Algorithms on Multi-level Micro-Aggregation Techniques for Secure Statistical Databases," 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA),Aqaba,2018,pp.1-6.
doi: 10.1109/AICCSA.2018.8612813