Clustering based Categorical Data Protection

Sowmya S.R, Manjunath S.S

Abstract: ***At present, the number of publicly available datasets is increasing day by day. It is therefore imperative to improve the confidentiality of the data, which has become one of the main reasons for an investigation. Extended to provide effective protection techniques that hinder the disclosure of entities in datasets while preserving the usefulness of the data. A systematic approach to categorical data protection is achieved by applying groups to the dataset and then protecting each group. In this paper, we present a survey and analysis on clustering techniques. The analysis of grouping techniques can result in confidential data or outliers in groups, and effective protection methods for such groups.***

Keywords-Clustering, Categorical Data, privacy, Data mining.

I. INTRODUCTION

The most important reason to protect your data is to guarantee the security of all information that is stored. When it comes to customers, ensuring that their data is stored in the safest possible way is the minimum that most people expect from the companies in which they invest time or money.

Data mining is a process used by companies to convert raw data into useful information. Data mining involves exploring and analysing large blocks of information to obtain significant patterns and trends. Categorical data is the type of statistical data that consists of categorical variables or data that has been converted to that form, for example as grouped data. There are three types of attributes in each data set: identifiers, quasi-identifiers, and confidential attributes. Quasi-identification data is pieces of information that are not unique identification data with any lack of clarity. Confidential characteristics contain sensitive information for the respondent, such as salary, religion or health status.

Clustering is the process of making a group of abstract objects in comparable object classes. Clustering analysis is widely used in many applications such as market research, pattern recognition, data analysis and image processing. Protection methods are evaluated using two measures, that is, loss of information and the risk of disclosure.

Revised Manuscript Received on December 12, 2019.

***** Correspondence Author

Sowmya S.R, Assistant Professor, Department of Information Science and Engineering. Dayananda Sagar Academy of Technology and Management, Bengaluru, srs.is.08@gmail.com

Dr. Manjunath S.S, Professor, Department of Computer Science and Engineering. Academy of Technology& Management Excellence college of Engineering, manjunath.dsatm@gmail.com

The loss of information is measured by comparing the statistical parameter between the anonymous data table and the original. Protection methods can be divided into two general categories: perturbative and non-perturbative. Perturbative is a technique for changing the value of the sensitive characteristic to a new value. The non-perturbative technique does not change the value of the sensitive attribute, but removes or removes certain data sets.

II. LITERATURE SURVEY

2.1 Subtractive Clustering:

The subtractive clustering method that is in effect can now only be used for numeric data; it cannot be applied to categorical data. But this method of mountain clustering can sometimes lead to increasing computational complexity, hence this approach has been proposed. This method can only be used for numeric data, because categorical data does not have a natural order. Although the k-means offers better performance, subtractive clustering is efficient.

2.2 Robust Hierarchical Clustering (RHC):

For metabolomics data, the successful way of grouping is hierarchical clustering. Traditional hierarchical clustering algorithms are very sensitive to outliers and cause misleading clustering results in the presence of those outliers. The hierarchical clustering algorithm can be strengthened by using the covariance matrix of the two-stage generalized S-estimator (TSGS).

The robust hierarchical cluster method has 3 main steps.

1. Estimation of the robust covariance matrix: The main challenge here is to simultaneously estimate a sufficient correlation or a covariance matrix in the presence of atypical values of cells or atypical values of cells and cases.
2. Robust correlation based on inequality Estimation of the matrix using the covariance matrix TSGS
3. Estimation of the proposed RHC using the TSGS correlation matrix

2.3 Decision Tree Categorical Value Clustering

Data perturbation techniques add noise to the data to prevent confidential values from being accurately disclosed.

First categorical attribute values are grouped and then, in later phases, these groups are used to add noise.

The technique of perturbation and grouping of categorical values based on the decision tree also known as Decision Tree based categorical value clustering and perturbation technique (DETECTIVE) disrupts a non-categorical categorical characteristic of a data set. Therefore, we apply all categorical attributes that are not a class in the original data set for each categorical attribute that is not a class. Each time it produces a set of data with a disturbed characteristic. Finally, we produce a data collection

(which combines all perturbed data sets) in which every categorical feature that does not belong to the class is perturbed and all other features are not perturbed.

2.4 Outlier Diagnosis:

Anomaly is the one that does not fit the expected pattern or any other element of the data set. This can be diagnosed using anomaly detection methods. These anomalies can also be called outliers, novelties, noise or deviation.

They are of three types:
1. Supervised anomaly
2. Unsupervised
3. Semi supervised anomaly detection.

Detection methods for unsupervised abnormalities detect abnormalities in an unlabelled test dataset in most instances of the dataset that is considered normal when searching for instances that seem to fit less with the rest of the dataset.

2.4.1 Outlier Detection Techniques:

A. *Statistical outlier detection:* Certain kind of statistical distribution is used that computes the parameters by assuming all data points have been generated by a statistical distribution.

B. *Depth based outlier:* Depth based outlier detection search novelty at the border of the data space.They are independent of statistical data distributions.

C. *Deviation based outlier:* The data elements are distributed as a sparse matrix in data set and due to this it creates confusion over data analysis. Some points get deviated from normal points are declared as outliers.

D. *Distance based outlier:* This judges a point based on the distances to its neighbours.

E. *Density based outlier:* This uses density distribution of data points into the data set.

Approach	Outlier detection percentage
Statistical outlier detection	78
Depth based outlier detection	85
Deviation based outlier detection	80
Density based outlier detection	84

2.5 Evolutionary Optimization Approach:

An advanced data protection method is based on an evolutionary algorithm guided by a combination of information loss and disclosure risk measures. Evolutionary algorithms are aimed at finding exact or approximate solutions for search or optimization problems. This uses state-of-the-art security methods for categorical data together with this type of algorithm to obtain good protection for a specific file, simply by combining pairs of protected files or changing their values in an evolutionary way.

The algorithm is based on two fundamental genetic operators: mutation and crossing.

Mutation: In this case the bits are randomly sorted to obtain a new offspring.
Crossover: it consists of recombinant values of two chromosomes, whereby two new descendants are also obtained.

2.6 l-diversity:

Anonymity models through generalization can protect individual privacy, but often lead to loss of information. (K, l, θ): diversity minimizes the loss of information and ensures data quality. This method guarantees the privacy of the data even without identifying the knowledge of the enemies to prevent the disclosure of attributes. Here sensitive attributes are "well represented" k-anonymity change.
To find a solution from a given set of n records (k, l, θ) diversity is used so that each group contains at least k ($k \leq n$) data points, as well as at least one clearly sensitive attribute, and the sum of all distances within the cluster is minimized.

III. RESULT COMPARISION

3.1 Clustering algorithm:

Table 2: Comparative performance of clustering algorithms

Algorithm	Advantage	Disadvantage
Subtractive clustering	This approach replaced the Euclidean distance with the Hamming distance. Experimental results describe that the presented method can achieve the best clustering accuracy compared to traditional k-modes.	Unsupervised clustering is not clear.
Robust hierarchical clustering	The simulation study remarkably indicates that this proposed method significantly improves performance compared to traditional hierarchical clustering.	1. Unable to undo the last step. 2. Adequate time complexity for a limited number of entities in a data set

3.2 Outlier detection algorithm:

Table 3: Comparative analysis of performance of outlier detection 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 algorithm:

Table 4: Comparison of performance of protection methods

Algorithm	Advantage	Disadvantage
Evolutionary Optimization Approach	They adapt well to higher dimensional problems	They are robust with regard to noise evaluation functions that do not produce a sensitive result within a certain period.
L-Diversity	1. It offers a larger distribution of confidential characteristics within the group, which increases data protection. 2. Protects against the disclosure of attributes.	1. This can be redundant and consume lots of effort to achieve. 2. Susceptible to attacks such as skewness attack.

IV. CONCLUSION

This paper presents an investigation into the privacy of categorical data using clustering techniques. The SCCA algorithm generally obtains more satisfied clustering accuracy than the traditional k-mode algorithm in each data set. TSGS algorithm performs better than the existing robust assessment procedure in presence of cell-wise and case-wise outliers.

L-diversity can support the intensification of respondents' privacy, but this feature is not suitable for protecting subtle attributes. That is why the evolutionary optimization approach is an improved security method.

REFERENCES

1. L. Gu, "Subtractive clustering for categorical data," *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Changsha, 2016, pp. 1229-1232.
2. M. A. Rahman, M. M. Raman, Asma-Ul-Husna and M. N. Haque Mollah, "Robust Hierarchical Clustering for Metabolomics Data Analysis in presence of Cell-wise and Case-wise outliers," *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, Rajshahi, 2018, pp. 1-4.
3. G.S.Karthick, M.Sridhar, “Data Renovation Algorithm for Protecting Sensitive Categorical Data,” 2017 International Journal for Research in Applied Science & Engineering Technology (IJRASET).
4. Stanley R. M. Oliveira Osmar R. Za¨ıane "Privacy Preserving Clustering by Data Transformation" Stanley Oliveirawas partially supported by CNPq (Consuelo National de Desenvolvimento Cient´ıfico e Technologic) of Ministry for Science.
5. 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.
6. G. Dubey, P. Navaney, A. Singh and G. Agarwal, "Outlier Detection Using Cluster Analysis for Fixed Income Bonds," *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, 2018, pp. 14-15.
7. W. Lixia and H. Jianmin, "Utility evaluation of K-anonymous data by microaggregation," *2009 ISECS International Colloquium on Computing, Communication, Control, and Management*, Sanya, 2009, pp. 381-384.
8. Gaoming Yang, Jingzhao Li, Shunxiang Zhang and Li Yu, "An enhanced l-diversity privacy preservation," *2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Shenyang, 2013, pp. 1115-1120.
9. J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez and S. Martínez, "t-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3098-3110, 1 Nov. 2015
10. Marés 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

AUTHORS PROFILE

Sowmya S.R is a graduate in information science and engineering and has a master’s degree in Software enginerring from VTU. Areas of interest are data analytics and big data.

Dr Manjunath S.S, Professor and Head Department of Computer science and Engineering, Academy of technology and Management, Mysore. His area of research being image processing.