Correlation between Privacy Preserving Data Publishing and Feature Selection Stability

Mohana chelvan P[1], Dr. Perumal K[2]

[1]Assistant Professor, Department of Computer Science,
Hindustan College of Arts and Science, Chennai – 603 103, India

[2]Associate Professor, Department of Computer Applications,
Madurai Kamaraj University, Madurai – 625 021, India

Abstract

Data mining is the technique of getting useful information from huge amount of available data stored in organizations. In these days, microdata published are high-dimensional. Feature selection is an important dimensionality reduction technique for data mining. Selection stability is the robustness of feature selection algorithms for the small perturbations of data sets. Selection stability is an important criterion for the feature selection in data mining. Earlier researches have been in the direction that selection stability is algorithmic dependent. But recently researches proved that selection stability is data dependent but not completely algorithmic independent. Privacy preserving data publishing is the publishing of public data to the private parties for research purposes after the perturbation of data to preserve privacy. In privacy preserving data publishing, we modify the data in some way in order to preserve the privacy of the data. This perturbation affects the selection stability and also the utility of the data. This paper finds the relationship between privacy preserving data publishing and feature selection stability.

Keywords: data mining, feature selection stability, Jaccard index, privacy preserving data publishing, k-anonymity, l-diversity, t-closeness, slicing

1.INTRODUCTION

There will be high-dimensional microdata produced by organizations involving e-commerce, e-governance, etc. Data mining is the extraction of useful information from the archived transactional data of organizations for getting edge over the competitors. Feature selection is an important dimensionality reduction technique in data mining which selects the subset of relevant features. Feature selection improves accuracy, efficiency and model interpretability. Microdata publishing techniques including *k*-anonymity, *l*-diversity, *t*-closeness and slicing will perturb the data in order to preserve the privacy of data. This paper is related with the impact of the privacy preserving data publishing techniques on feature selection stability and accuracy in data mining. It has been found that there will be no valuable contribution of research work to find the relationship between privacy preserving data publishing and feature selection stability.

This paper is organized as follows. Section 2 gives an account on data perspective nature of feature selection stability. Section 3 gives information about Jaccard index. Section 4 explains privacy preserving data publishing. Section 5 explains about privacy preserving approaches. Section 6 explains privacy threats. Section 7 elaborates on microdata publishing techniques. Section 8 gives experimental results and section 9 gives conclusion.

2.DATA PERSPECTIVE NATURE OF FEATURE SELECTION STABILITY

Due to the advancements in Information Technology, online collection of microdata about individuals as high-dimensional datasets is the everyday activity. In feature selection, only subset of relevant attributes is obtained as a dimensionality reduction technique. If there will be a small perturbation or addition of new samples, there should be selection of similar set of features in feature selection. Otherwise, it will create confusion in researchers mind about the conclusion of their research experiment and also lower their confidence. Earlier research work is in the direction that selection stability is mostly algorithmic dependant. But recent researches have proved that selection stability is dataset dependant but not completely algorithmic independent. Data variation will affect the selection stability [1, 2].

3.JACCARD INDEX

There are various measures for selection stability. Jaccard index, aims to evaluate the amount of overlap between two set of feature indices, while Pearson's and Spearman's correlations aim to measure the consistency of weights or ranks of the features in the two lists. In the experiments, Jaccard index is used to measure feature selection stability. Given different results R = $\{R_1, R_2 ... R_l\}$ corresponding to *l* different folds of the data set D, its stability can be assessed by the amount of overlap between the sets in R. The Jaccard Index is to evaluate the stability for subsets of results that contain selected features' indices by evaluating the amount of overlap between the subsets [3]. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. The following equation shows a Jaccard Index for two selected subsets.

$$S_J(R_{i,} R_j) = \frac{|R_i \cap R_j|}{|R_i \text{ U } R_j|}$$

$$S_J(R) = \frac{2}{l(l-1)} \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} S_J(R_{i,} R_j)$$

The Jaccard Index S_J returns a value in the interval of [0,1] where 0 means the feature selection results are not stable and 1 means the results are identical, hence very stable.

4.PRIVACY PRESERVING DATA PUBLISHING

Data mining is indispensable for business organizations for making strategic decisions. Data mining has been mostly done by persons who are not working in the organizations and so preserving privacy of data is very important. Microdata involving individuals like medical data are published for their utility in research works but revealing of private data could affect the reputation of the organization and also will lead the heavy financial losses. Privacy preserving data publishing is to protect privacy of data by some way before publishing it to third party for data mining.

5.PRIVACY PRESERVING APPROACHES

Several privacy preserving approaches in data publishing such as randomization, sampling, suppression, swapping and perturbation have been designed for microdata publishing [9]. Suppression replaces the identifying attributes with values like '*'. Generalization transforms the Quasi Identifier (QI) attributes in each bucket into "less specific but semantically consistent" values so that tuples in the same bucket cannot be distinguished by their Quasi Identifier attributes. Generalization fails on high dimensional data due to the curse of dimensionality and also it causes too much information loss due to the uniform-distribution assumption. Bucketization first partitions tuples in the table into buckets and then separates the quasi identifiers with the sensitive attribute by randomly permuting the sensitive attribute values in each bucket. Bucketization has been mainly used for anonymizing high-dimensional data.

6.PRIVACY THREATS

The threats for microdata publishing in data mining will be identity disclosure, membership disclosure and attribute disclosure. Anonymising data would result in better protection from these threats. Identity disclosure is concerned with disclosure of identifying attributes of the individual person. For protecting attribute disclosure, matching multiple buckets was important [4]. Membership information would infer an identity of an individual through various attacks and if the selection criteria were not a sensitive attribute value, then it would lead to have a membership disclosure [5].

7.MICRODATA PUBLISHING TECHNIQUES

7.1 k-anonymity

The k-anonymity model was developed because of the possibility of indirect identification of records from public databases using Quasi-Identifier attributes. In this method, we reduce the granularity of data representation with the use of techniques such as generalization and suppression [6]. In the k-anonymity technique every combination of values of quasi-identifier attributes can be indistinguishably matched to at least k respondents. The values of the attributes are discretized into intervals for quantitative attributes or grouped into different sets of values for categorical attributes [6]. However, this technique is insufficient to prevent attribute disclosure. Two attacks on *k*-anonymity are the homogeneity attack and the background knowledge attack.

7.2 l-diversity

The l-diversity model was designed to handle some weaknesses in the k-anonymity model in which the model does not protecting the sensitive attributes corresponding to the Quasi-Identifier attributes, especially when there is homogeneity of sensitive values within a group. To prevent the attacks on k-anonymity such as homogeneity attack and background knowledge attack, the concept of intra-group diversity of sensitive values is promoted within the anonymization scheme with the bucketization technique. In the technique of l-diversity, there will be not only maintenance of the minimum group size of k, but also focuses on maintaining the diversity of the sensitive attributes [7]. If there is 'l' 'well represented' values for sensitive attribute then that class is said to have l-diversity. The simplest understanding of 'well represented' would be to ensure there are at least l distinct values for the sensitive attribute in each equivalence class [7].

7.3 t-closeness

The l-diversity model treats all values of a given attribute in a similar way irrespective of its distribution in the data. But in the case for real data sets, the attribute values may be much skewed. Often, an adversary may use background knowledge of the global distribution to make inferences about sensitive values in the data. The t-closeness model uses the property that the distance between the distributions of the sensitive attribute within an anonymized group should not be different from the global distribution by more than a threshold t [8]. The t-closeness approach tends to be more effective than many other privacy-preserving data mining methods for numeric attributes. 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 no more than a threshold t.

7.4 Slicing

Generalization loses considerable amount of information, especially for high dimensional data. Bucketization does not prevent membership disclosure and does not apply for data that do not have a clear separation between quasi-identifying attributes and sensitive attributes. In the slicing technique, we will divide the data both horizontally and vertically. This technique preserves better data utility than generalization. Partitioning attributes into columns will protect privacy by breaking the association of uncorrelated attributes and preserve data utility by preserving the association between highly-correlated attributes [4, 5]. Slicing is better suited for high-dimensional data.

8.EXPERIMENTAL RESULTS

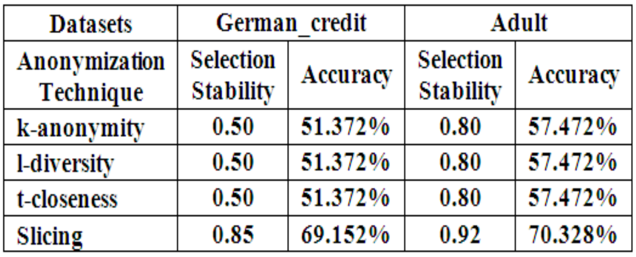
The datasets used in the experiments are german-credit and adult datasets available on UCI Machine Learning

Repository. In these experiments, weka software was used for feature selection. Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. From the unperturbed datasets, the features are selected using the attribute evaluator CfsSubsetEval and the search method is BestFirst. Table 1 gives the overall accuracy and the accuracy of selected features after applying feature selection technique on the datasets. The datasets are evaluated using the attribute evaluator InfoGainAttributeEval to obtain the ranked attributes based on information gain for the identification of quasi attributes. Table 2 gives the results for selection stability and accuracy for selected features after applying the techniques k-anonymity, *l*-diversity, *t*-closeness and slicing.

Table 1. Summary of accuracy for german_credit and adult datasets before perturbation

Datasets	**german_credit**	**adult**
Overall accuracy before perturbation	70%	75.919%
Accuracy for selected features before perturbation	70%	75.919%

Table 2. Summary of selection stability and accuracy of german_credit and adult datasets after applying anonymization techniques



Datasets	German_credit		Adult	
Anonymization Technique	Selection Stability	Accuracy	Selection Stability	Accuracy
k-anonymity	0.50	51.372%	0.80	57.472%
l-diversity	0.50	51.372%	0.80	57.472%
t-closeness	0.50	51.372%	0.80	57.472%
Slicing	0.85	69.152%	0.92	70.328%

In the experiments, selection stability and accuracy are at minimum level in the *k*-anonymity, *l*-diversity and *t*-closeness techniques but improved in slicing technique. Selection stability for *k*-anonymity, *l*-diversity and *t*-closeness are almost same and the level of perturbation of data in these techniques is similar. But in the slicing technique there will be small amount of perturbation and so it has better selection stability and data utility.

9.CONCLUSION

This paper gives an overview of feature selection stability and its importance in data mining. It also discusses about various privacy preserving data publishing techniques. It also gives an account of how the data publishing techniques affect the selection stability and accuracy in privacy preserving data mining. In future, the research will be in the direction of development of privacy preserving algorithms to perturb the data resulting in better feature selection stability and data utility during data mining.

References

[1]. Salem Alelyani, Huan Liu, The Effect of the Characteristics of the Dataset on the Selection Stability 1082-3409/11 IEEE DOI 10.1109/Inrenational Conference on Tools with Artificial Intelligence.2011.167, 2011

[2]. Salem Alelyani, Zheng Zhao, Huan Liu, A Dilemma in Assessing Stability of Feature Selection Algorithms, 978-0-7695-4538-7/11, IEEE DOI 10.1109/ International Conference on High Performance Computing and Communications. 2011.99, 2011

[3]. Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. Knowledge and Information Systems, 12(1):95–116, May 2007.

[4]. B.Vani, D.Jayanthi, (2013), "Efficient Approach for Privacy Preserving Microdata Publishing Using Slicing" IJRCTT.

[5]. Tiancheng Li, Jian Zhang, Ian Molloy ,(2012),"Slicing: A New Approach for Privacy Preserving Data Publishing" IEEE Transaction on KDD.

[6]. Charu C. Aggarwal, (2005), ''On k-Anonymity and the Curse of Dimensionality", Proceedings of the 31st VLDB Conference, Trondheim, Norway, pp.901-909.

[7]. Ashwin Machanavajjhala , Daniel Kifer,Johannes Gehrke, Muthuramakrishnan Venkita Subramanian, (2006)," ℓ-Diversity : Privacy Beyond K-Anonymity", Proc. International conference on Data Engineering.(ICDE),pp.24.

[8]. Anil Prakash, Ravindar Mogili ,(2012),''Privacy Preservation Measure using t-closeness with combined l-diversity and k-anonymity", International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARC SEE)Volume 1, Issue 8,pp:28-33

[9]. Yan Zhao, Ming Du, Jiajin Le, Yongcheng Luo,(2009), " A Survey on Privacy Preserving Approaches in Data Publishing" in the First International Workshop on Database Technology and Applications