ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 2 Issue Special 1 Jan-March 2014
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014)

Accuracy Enhancement of Mining Association Rules

[I]Sindhujaa.N, [II]Anees.M, [III]Dr. P.S.K.Patra
[I]Dept. of CSE, Agni College of Tech., Anna University,Chennai,Tamilnadu, India
[II]Assistant Professor, Dept. of CSE, Agni College of Tech., Anna University,Chennai,Tamilnadu, India
[III]Head of the Dept., Dept. of CSE, Agni College of Tech., Anna University,Chennai,Tamilnadu, India

Abstract
A new anonymization algorithm called Non-homogeneous generalization with Sensitive Value Distributions (NSGVD) has with make use of data mining algorithm as association rule been devised. This algorithm helps to generate minimum anonymity and diversity parameters along with an information loss measure. In the experiments, using eight datasets and four different classification algorithms, it is shown that classifiers induced from data generalized by NSGVD tend to be more accurate than classifiers induced using state of the art anonymization algorithms.

Keywords
Classification, t-closeness, Frequency Distribution, Privacy Preserving Data Publishing,Anonymization

I. Introduction

An vast quantity of privately owned records that express individuals, interests, activities, and demographics. The records often include sensitive data and may violate the privacy of the users if published. The information is suitable for very important resource for many systems and corporations that may improve their services and performance by remind novel and potentially useful data mining models. One of the common practice for releasing such confidential data without violating privacy and apply some regulations and policies for the data usage. These type of regulations usually entail data distortion operations such as generalization . The challenge with this approach is the data leakage can still occur and the data and the resulting data mining models may become nearly useless after excessive distortion[8]. The upcoming research field of Privacy Preserving Data Publishing (PPDP) is targeting this challenge [8]. PPDP also aims at developing techniques that allow publishing data while minimizing distortion for maintaining utility on one hand and ensuring that privacy is preserved on the other. In this paper we present a new privacy-preserving data publishing(PPDP) method, which is shown to preserve the predictive utility of supervised classification algorithms that are qualified on the published data. The analytical utility is measured by the classification accuracy of the induced classification models when applied to new previously unseen data. A directly related research area is Privacy Preserving Data Mining (PPDM) that was initiated in 2000 by [1]. PPDM algorithms seek at anonymizing data towards its release for specific data mining goals so that the data utility is maximized, one thing, and its privacy is preserved on the other thing. The developed PPDM algorithms are tailored to specific data mining tasks and algorithms. In PPDP on the other hand the correct purposes of the data release are unknown and it is needed to anonymize the data using utility measures that are not under attack to a specific data mining algorithm. It is usual to distinguish between the types of attributes in the database table that needs to be published (see [9]):

Identifiers - Attributes that uniquely identify an individual;

Quasi-identifiers - Publicly-accessible attributes that do not identify a person, but some combinations of their values might yield unique identification (e.g., gender, age, and zipcode);

Sensitive information -Attributes of private nature, such as medical, financial data etc.

Other non-sensitive attributes - It cannot be used for identification since they are unlikely to be accessible to the adversary and do not represent information of sensitive nature. A common practice in PPDP and PPDM is to remove the identifiers and to generalize or suppress the quasi-identifiers in order to protect the sensitive data of individuals from being revealed. Generalization means that the original values of quasi-identifiers are replaced with less specific values but in case of control no values are released. The sensitive data is usually retained unchanged. In the past years, several models were suggested for maintaining privacy when disseminating data. Most approaches evolved from the basic model of k-anonymity [10]. In the model the practice is to remove the identifiers and generalize the quasi-identifiers as until each generalized record is the same from at least k - 1 other generalized records when projected on the quasi-identifiers. Therefore, an adversary who wishes to trace a record of a specific person in the anonymized table, will not be able to trace that person's record to subsets of less than 'k' anonymized records.

II. An algorithm for Non-homogeneous Generalization with Sensitive Value Distribution (NSGVD)

A. The Algorithm

Non-homogeneous generalization algorithm with Sensitive Value Distribution (NSVDist) (see Algorithm) generates for each record $R_n \in T$ corresponding generalized record R_n which is the closure of R_n and k - 1 additional records in T; the subset of T that includes R_n and the additional k - 1 records is denoted B_n. Selection of the k - 1 additional records in B_n is guided by two policy[3]:

1. It relates to the generalized quasi-identifiers
2. It also relates to the sensitive distribution:
 (a) Trying to decrease the resulting information loss IL of B_n.
 (b) Making sure that the variety of Bn is at least l.

The selection is accepted in a greedy manner. The k - 1 records that will be second-hand to mask a given record $R_n \in T$ are chosen one at a time where in each stage select a record that complies with the variety constraint and reduces the resulting information loss. The operation of the algorithm is free of the choice of information loss measure and the entropy measure [11].)

In order to calculate the generalization R_n of the record R_n, $n \in [N]$, compute a set B_n & R_n and additional k - 1 records, so that the diversity of B_n is at least l and its information loss is as small as achievable. The set B_n is initialized only to R_n . Then adding to it one additional record at a time until its size comes to k . In order to confirm the variety constraint and maintain a frequency vector F of length $|A_{M+1}|$, at each stage F(q) equals the number of records in B_n whose sensitive value is the q^{th} value in A_{M+1}. The

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

vector is initialized in Line 3 for the initial set Bn. The loop that outfit the greedy selection and review all records that were not selected yet. Particularly, since B_n will ultimately be of size k it will be l-diverses. Hence, focus only on records whose sensitive value appears in B_n strictly less than bk = lc times. All the records choose the one R_i, whose addition to B_n would yield a set B_n U {Ri} of minimal Information Loss(IL) . The function that calculates the information loss of a given set of records. Selecting the record, attach it to B_n and revise the vector F accordingly. Finally when B_n includes R_n & additional k - 1 records, set R_n to be the end of B_n. The algorithm given below can be viewed from [3]. Because this paper mainly based on [3] with association rules.

Algorithm - Non-homogeneous generalization with Sensitive Value Distribution.
Input: A table T = {R_1; : : : ;R_N}, anonymity parameter k, diversity parameter `.
Output: A non-homogeneous (k; `)-anonymization T = {R_1; : : : ;R_N} with sensitive value distribution.
1: for all 1 < n < N do
2: Set B_n = {R_n}.
3: Set F(q) = 0 for all q ∈ A_{M+1} \ {R_n(M + 1)} and F(q) = 1 for q = R_n(M + 1).
4: while |B_n| < k do
5: Among all records R_i ∈ T \ B_n for which F(R_i(M + 1)) < [k/l], find one that minimizes I L(B_n U {Ri}).
6: Add the selected R_i to B_n and set F(R_i(M + 1)) = F(R_i(M + 1))+ 1.
7: end while
8: R_n = B_n.
9: end for
10: Return T = {R_1,…..R_N}.

III. Experimental Results

Table 1 gives the information on the number of records in each dataset, the number of records that each dataset have and the list of statistically relevant quasi-identifiers. Out of the eight datasets that used for our evaluation; hence in that dataset did not relate the 10-fold cross validation methodology and consequently, the accuracy values reported for that dataset are an average over the p- autonomous samples. Then the statistically related quasi-identifiers were identifiers by applying on each dataset the Weka software (version 3.68) [12] operator. This method which depends on greedy hill climbing & backtracking search chooses a subset of attributes having the highest analytical value along with a less degree of redundancy. The exchange between the anonymity level k of the training data and testing accuracy of the 4 evaluated classifiers, in each of the eight datasets. The leftmost column of plots in each outline shows the classifier accuracy when the training data was anonymized with the diversity parameter l = 1, while the rightmost column shows the results with a higher diversity parameter. We got each plot in each of those figures involvess four curves, representing the sequential anonymization algorithm, Mondrian algorithm, the privacy aware information sharing algorithm and NSVDist algorithm. In all output accuracy ranking table have eight rows corresponding to the eight data sets & ten columns: the first four columns shows the accuracy of a classifier that was trained on the

Table 1: Datasets

Datasets	Reco-rds	Quasi Identifiers
Heart	5000	13: age, sex, chest pain, chol, trestbps, fbs, restecg, thalach, exang, oldpeak, slow, ca, thal
Stomach cancer	297	11: ID, Age, Sex, Infection_with_Helicobacter_pylori, Fasting_Blood_sugar, A_Diet_with_high_Salty_and_Smoked Foods, Alcohol_intake, PeRnicious_ anemia, Gastritis_Type, Family_ history, Stomach_polyps.
Breast cancer	297	11:ID, Age, Childbearing, Hormones_ type, high_fat_diet, alcohol_intake, Obesity, Personal_history_of_breast_ cancer, Family_history, environmental_ factors, breast_lumpfitype.
Lung cancer	297	11: ID, Age, Sex, Hoarseness, Recurring_Inflammation, Radon, Tobacco_Smoking, Asbestos, Marijuana, Family_history, Chest_ pain_type.
HIV	5000	14 : ID, Age, Sex, Depression, Diarrhea, Thrush, Lipodystrophy, Lactic Acidosis, BuRning & Tingling of the Feet & Hands, Weight Loss, Sinus Infection, Vomitting, Number_ of_major_vessels_colored_by_ flourosopy, Fatigue.
Yeast	1484	4: seq, alm, erl, pox.
Ecoli	336	6: seq, mcg, gvh, lip, alm1, alm2.

anonymized training data with a representative anonymity parameter k = 50 and diversity l=1; the next 4 columns gives the accuracy when the diversity parameter was set to a higher value; and the last two columns give the two baseline values. The best value among the results with l 1 is highlighted and so is the best value among the results with l > 1 the second group of four columns. In addition, every plot includes 2 reference baselines. The classification and accuracy based majority rule and the accuracy of the classifier that was trained on the original dataset records. Each and every point on curvature violates the average over ten independent samples and over ten training-test partitions whenever the dataset had no training-test partition. Finally, it gives another short and snappy look at the results of the above described sequence of experiments. Hence calculated the proposed anonymization methodology with different classification algorithms. For each dataset and classification algorithm carried out an assessment procedure that consisted of the following steps:
1. If the dataset had no available training-test partition and applied on it the 10-fold cross-validation [12]. In our work, the training set is used to generate the published anonymized data that may be accessible by anyone to induce a classification model the test set on the other hand represents data that is unknown at the time

of performing the anonymization and it is accessible only to the user of the classification model [3].
2. Perform (k, l) - anonymizations of the training set for various settings of k and l, using four algorithms.
3. For each and every setting of k and l trained a classifier on each of the 4 resulting anonymized tables that using different classification algorithms.
4. In addition to that, compute the accuracy of a classifier based on the majority rule. This type of classifier gives the maximum possible level of privacy. And also compute the accuracy of a classifier that was qualified on the original training. The standard classification algorithms cannot be applied directly on generalized tables since they contain non-specific values such as numeric intervals or subsets of nominal values [3]. Hence it is essential to convert the anonymized tables into tables with specific values after that only apply the classification algorithm on those non generalized tables. According to the privacy in both non-homogeneous and homogeneous anonymizations that are l-diverse require that none of the sensitive values in those multisets appear in frequency that is greater than 1=l [3]. Assume that the adversary will effort to gain knowledge on the sensitive values of some of the individuals behind the masking values in the multiset of his target records in order to know many more information on the sensitive value of his target record. Adopting that strategy that will be able to assume the sensitive value of his target record with certainty once he gains knowledge of the sensitive values of all individuals whose sensitive value differs from that of his target record [3]. The combination of the k anonymity and l-diversity conditions imply the same lower bound on the number of individuals for which the adversary needs to learn the sensitive information in both anonymization models.

IV. Conclusion

This paper offered a new privacy-preserving data publishing(PPDP) algorithm called NSVDist (Non-homogeneous generalization with Sensitive Value Distributions) with makes use of data mining algorithm as association rules. That algorithm is based on non-homogeneous anonymization of the quasi-identifiers, coupled with the generalization of the sensitive values into frequency distributions [3]. Since that the algorithm is categorized by smaller information losses than leading anonymization algorithms that the proposed algorithm allows the owner of the data to discharge the data in a more secure form while expecting the data miner to tempt accurate classification models. Our experimental results make sure that the hypothesis in many cases. These findings propose that the structure of non-homogeneous anonymizations which allows lower information loss might be more adequate than homogeneous anonymizations.
Directions for future enhancements include the following:
(a) In this paper, we studied the simplest case of a single sensitive attribute, which is also a classification attribute. The proposed approach to non-homogeneous anonymization can be extended to more general cases like disjoint or partially overlapping sets of several sensitive and classification attributes [3].
(b) Extending the NSVDist algorithm for the case of a sequential release of data attributes [4,5,6]. In the sequential release scenario, several releases of the same table are published over a period of time. The ultimate goal is to save the private information from adversaries who examine the entire sequential release[3].

References

[1] *Agrawal, R. and Srikant, R. 2000. Privacy-preserving datamining. In The ACM SIGMOD International Conference on Data Management (SIGMOD). 439-450.*
[2] *Bayardo, R. and Agrawal, R. 2005. Data privacy through optimal k-anonymization. In InteRnational Conference on Data Engineering (ICDE). 217-228.*
[3] *Mark Last, Tamir Tassa, Alexandra Zhmudyak, Erez Shmueli Improving Accuracy of Classification Models Induced From Anonymized datasets.In Business Intelligence in Risk Management. Vol 256. 2014. pages 138-161*
[4] *Matatov, N., Rokach, L., and Maimon, O. 2010. Privacy preserving data mining: a feature set partitioning approach. Information Sciences 180, 14, 2696-2720.*
[5] *Shmueli, E., Tassa, T., Wasserstein, R., Shapira, B., and Rokach, L. 2012. Limiting disclosure of sensitive data in sequential releases of databases. Information Sciences 191, 98-127.*
[6] *Wang, K. and Fung, B. 2006. Anonymizing sequential release. In The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). 414-423.*
[7] *Fung, B., Wang, K., Chen, R., and Yu, P. 2010. Privacy-preserving data publishing: A survey on recent developments. ACM Computing Surveys (CSUR) 42, 1-53.*
[8] *Bayardo, R. and Agrawal, R. 2005. Data privacy through optimal k-anonymization. In InteRnational Conference on Data Engineering (ICDE). 217-228.*
[9] *Sweeney, L. 2002. k-Anonymity: A model for protecting privacy. International JouRnal on Uncertainty, Fuzziness and Knowledge-based Systems 10, 5, 557-570.*
[10] *Gionis, A. and Tassa, T. 2009. k-Anonymization with minimal loss of information. IEEE Transactions on Knowledge and Data Engineering 21, 206-219.*
[11] *Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. 2009. The weka data mining software: an update. ACM SIGKDD Explorations Newsletter 11, 1, 10-18.*
[12] *Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., and Euler, T. 2006. Yale: rapid prototyping for complex data mining tasks. In The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). 935-940.*