Suppression of Multidimensional Data Using K-Anonymity

[1]Mrs.M.Aruna Safali, [2]Mr.T.Bala Murali Krishna, [3] Mr. G Sai Chaitanya Kumar

[1] Asst.Professor, Dept. of CSE, NRI Institute of Technology, Vijayawada, A.P., India.,
arunachowdary_maddineni@yahoo.co.in

[2] Assoc. Professor, Dept. of CSE, Paladugu Parvathi Devi College of Engg & Tech, Vijayawada, A.P. ,India.,
balu_thati@yahoo.com

[3] Asst.Professor, Dept. of CSE, Paladugu Parvathi Devi College of Engg & Tech, Vijayawada, A.P. ,India.,
sai.dmca@gmail.com

Abstract

Many applications that employ data mining techniques involve mining data that include private and sensitive information about the subjects. One way to enable effective data mining while preserving privacy is to anonymize the data set that includes private information about subjects before being released for data mining. One way to anonymize data set is to manipulate its content so that the records adhere to k-anonymity. Two common manipulation techniques used to achieve k-anonymity of a data set are generalization and suppression. Generalization refers to replacing a value with a less specific but semantically consistent value, while suppression refers to not releasing a value at all. A new method for achieving k-anonymity named K-anonymity of Classification Trees Using Suppression (kACTUS). In kACTUS, efficient multidimensional suppression is performed, i.e., values are suppressed only on certain records depending on other attribute values, without the need for manually produced domain hierarchy trees.

Keywords: Privacy-preserving data mining, k-anonymity, decision trees

1 Introduction

Knowledge Discovery in Databases (KDDs) is the process of identifying valid, novel, useful, and understandable patterns from large data sets. Data Mining (DM) is the core of the KDD process, involving algorithms that explore the data, develop models, and discover significant patterns. Many of these applications involve mining data that include private and sensitive information about users. For instance, medical research might be conducted by applying data mining algorithms on patient medical records to identify disease patterns. A common practice is to deidentify data before releasing it and applying a data mining process in order to preserve the privacy of users.

However, private information about users might be exposed when linking deidentified data with external public sources. For example, the identity of a 95-year-old patient may be inferred from deidentified data that include the patients' addresses, if she is known as the only patient at this age in her neighborhood. This is true even if sensitive details such as her social security number, her name, and the name of the street, where she lives, were omitted. To avoid such situations, privacy regulations

were promulgated in many countries. The data owner is required to omit identifying data so that to assure, with high probability, that private information about individuals cannot be inferred from the data set that is released for analysis or sent to another data owner. At the same time, omitting important fields from data sets, such as age in a medical domain, might reduce the accuracy of the model derived from the data by the DM process. Privacy rules have affected significantly their ability to perform retrospective, chart-based research. Privacy-preserving data mining (PPDM) deals with the trade-off between the effectiveness of the mining process and privacy of the subjects, aiming at minimizing the privacy exposure with minimal effect on mining results.

K-anonymity is an anonymizing approach proposed by Samarati and Sweeney. A data set complies with k-anonymity protection if each individual's record stored in the released data set cannot be distinguished from at least k- 1 individuals whose data also appear in the data set. This protection guarantees that the probability of identifying an individual based on the released data in the data set does not exceed 1/k. Generalization and suppression are the most common methods used for deidentification of the data in k-anonymity-based algorithms .Generalization consists of substituting attribute values with semantically consistent but less precise values. For example, the month of birth can be replaced by the year of birth which occurs in more records so that the identification of a specific individual is more difficult. Suppression can drastically reduce the quality of the data if not properly used. This is the reason why most k- anonymity-related studies have focused on generalization. Quasi-identifier is a set of features whose associated values may be useful for linking with another data set to Reidentify the entity that is the subject of the data kACTUS was specifically designed to support classification, but can be extended to support other data mining methods. The new algorithm we developed, kACTUS, wraps a decision tree inducer which is used to induce a classification tree from the original data set. The wrapping approach refers to using a known algorithm as a black box so that only knowledge of the interface is needed. While the wrapping approach was used in other domains it was not used for anonymizing data sets, and presents important advantages for this application. Another advantage of the Wrapping approach is that it does not require any revision of existing algorithms for adjusting it to k-anonymity. The automatically induced tree is used by kACTUS to apply k-anonymity on the data set kACTUS generates a k-anonymous data set that can be used by external users that may utilize any mining algorithm for training a classifier over the anonymous data set. The output of our algorithm is an anonymous data set which can be transmitted to the data consumers for further mining. The kACTUS algorithm takes the suppression approach to anonymize the data set.

2 RELATED WORK

PPDM is a relatively new research area that aims to prevent the violation of privacy that might result from data mining operations on data sets .PPDM algorithms modify original data sets so that privacy is preserved even after the mining process is activated, while minimally affecting the mining results quality. Verykios classified existing PPDM approaches based on five dimensions:

1. Data distribution, referring to whether the data are centralized or distributed.
2. Data modification, referring to the modifications performed on the data values to ensure privacy. There are different

possible operations such as aggregation (also called generalization) or swapping.
3. Data mining algorithms referring to the target DM algorithm for which the PPDM method is defined.
4. Data or rule hiding referring to whether the PPDM method hides the raw or the aggregated data and finally.
5. Privacy preservation.
According to the above dimensions: it deals with centralized databases (dimension 1), on which suppression of the data (dimension 2) is performed. The DM algorithm that this study is targeting is classification (dimension 3), and part of the raw data is hidden (dimension 4). We use the k-anonymity method which is a heuristic-based technique (dimension 5).

One of the PPDM techniques is k-anonymity. The k-anonymity concept requires that the probability to identify an individual by linking databases does not exceed 1/k. The "top-down specialization (TDS)" algorithm handles both categorical and continuous attributes. TDS starts from the most general state of the table and specializes it by assigning specific values to attributes until violation of the anonymity may occur. More recently, Fung presented an improved version of TDS which is called "Top-Down Refinement" (TDR). In addition to the capabilities of TDS, TDR is capable of suppressing a categorical attribute with no taxonomy tree. They use a single dimension recoding , i.e., an aggressive suppression operator that suppresses a certain value in all records without considering values of other attributes so that data that might adhere to k-anonymity might be also suppressed.
This "over suppression" reduces the quality of the anonymous data sets.

3 METHODS

The kACTUS algorithm consists of two main phases: In the first phase, a classification tree is induced from the original data set; in the second, the classification tree is used by a new algorithm developed in this study to k-anonymize the data set.

3.1 Deriving the Classification Tree

In this phase, we are employing a decision tree inducer (denoted by CTI) to generate a decision tree denoted by CT. The tree can be derived using various inducers. We concentrate on top-down univariate inducers which are considered the most popular decision tree inducers and include the well-known algorithms Top-down inducers are greedy by nature and construct the decision tree in a top-down recursive manner (also known as divide and conquer). Univariate means that the internal nodes are split according to the value of a single attribute.

3.2 K-Anonymity Process

In this phase, we use the classification tree that was created in the first phase to generate the anonymous data set. We assume that the classification tree complies with the following properties:
1. The classification tree is univariate, i.e., each internal node in the tree refers to exactly one attribute.
2. All internal nodes refer to a quasi-identifier attributes.
3. Assuming a top-down inducer, the attributes are sorted (from left to right) according to their significance for predicting the class (where the rightmost relates to the least significant attribute).
4. Complete Coverage: Each instance is associated with exactly one path from root to leaf.

3.3 kACTUS **Algorithm**

kACTUS **Algorithm**
Input:
1: KT - k-anonymity threshold
2: ST - swapping threshold
3: OD - original dataset
4: CT - classification tree
5: NQI - non-quasi-identifier set
Output:
6: AD - anonymous dataset
7: **procedure** Main(*KT*, *ST*, *OD*, *AD*, *CT*, *NQI*)
8: *CS* ← *OD*
9: *AD* ←∅
10: non-complying-node-set ←∅
11: **for all** root-node in root(CT) **do**
12: **while** height(root − node) > 0 **do**
13: longest-node ← get-longest-node(root-node)
14: **if** height(longest-node) > 1 **then**
15: PerformAnonymization(longest-node, KT, ST, CS, AD)
16: **else** _ the longest node is the root node
17: **if** count-instances(longest-node, CD) ≥ KT **then**
18: move-complying-node(longest-node, CS, AD)
19: remove-leaf-nodes(longest-node)
20: **end if**
21: non-complying-node-set ← non-complying-node-set _longest-node
22: **end if**
23: **end while**
24: **end for**
25: **if** get-total-instance-count(non-complying-node-set) ≥ KT **then** move-root-non-complying-nodes(non-complying-node-set, CS, AD, NQI)
26: **end if**
27: **end procedure**

4 EXPERIMENTAL EVALUATIONS

In order to evaluate the performance of the proposed method for applying k-anonymity to a data set used for classification tasks, a comparative experiment was conducted on benchmark data sets. Specifically, the experimental study has the following goals:

1. To compare the obtained classification accuracy to the original accuracy without applying k-anonymity
2. To compare the proposed method to existing k-anonymity methods in terms of classification accuracy.
3. To investigate the sensitivity of the proposed method to different classification methods

4.1 Experimental Process

Fig. 1 is a graphic representation of the experimental process that was conducted. Unshaded boxes represent data sets. The main aim of this process is to estimate the generalized accuracy (i.e., the probability that an instance was classified correctly). First, the data set (box 1) was divided into training (box 3) and test sets (box 4) using five iterations of twofold cross validation (box 2). On each iteration, the data set is randomly partitioned into two equal-sized sets S1 and S2 such that the algorithm is evaluated twice: on the first evaluation, S1 is the training set and S2 the test set, and vice versa the second time. We apply (box 5) the k-anonymity method on the training set and obtain a new anonymous training set (box 6). An inducer is trained (box 7 over the anonymous training set to generate a classifier (box 8). Finally, the classifier is used to estimate the performance of the algorithm over the test set (box 9). Note that in the kACTUS algorithm, the classifier can use the test set as is. In kACTUS, suppression is performed on the training set, i.e., some values are converted to missing values and all others left on their original values. Many of the existing induction algorithms are capable to train from data set with missing values.

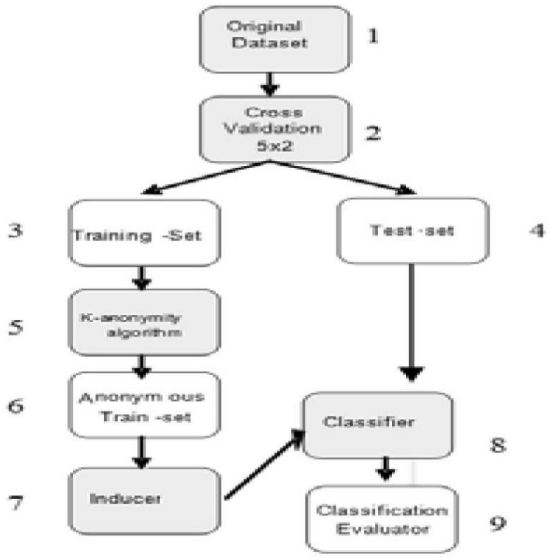

Fig1

4.2 Comparing to Suppression Methods

It is expected that generalization methods outperform suppression methods because in the former case, additional knowledge (the generalization tree) is provided to the K-anonymity algorithm and the algorithm has the choice for not entirely suppressing the datum but gently generalizes its value. In fact, one of the features of TDR is to automatically decide if it is better to perform suppression or generalization. In this section, we examine if the improved performance of kACTUS is not derived from the fact that our algorithm uses suppression approach, while other use generalization. Specifically, in this section, all generalization algorithms are provided with one level of generalization trees

4.3 Discussions

The advantages of the new kACTUS algorithm, as observed from the experimental study, can be summarized as following: . kACTUS is capable of applying k-anonymity on a given table with no significant effect on classification accuracy. . kACTUS scales well with large data sets and anonymity levels. . When compared to the state-of-the-art k-anonymity methods, kACTUS anonymized data can be used to induce classifiers which are of an equivalent or slightly higher degree of accuracy.

5 CONCLUSION

We presented a new method for preserving the privacy in classification tasks using k-anonymity. The proposed method requires no prior knowledge regarding the domain hierarchy taxonomy and can be used by any inducer. The new method also shows a higher predictive performance when compared to existing state-of-the-art methods. Additional issues to be studied further include: Examining kACTUS with other decision trees inducers; revising kACTUS to overcome its existing drawbacks; extending the proposed method to other data mining tasks (such as clustering and association rules) and to other anonymity measures (such as l-diversity) which respond to different known attacks against k-anonymity, such as homogeneous attack and background attack.

REFERENCES

[1] M. Kantarcioglu, J. Jin, and C. Clifton, "When Do Data Mining Results Violate Privacy?" Proc. 2004 Int'l Conf. Knowledge Discovery and Data Mining, pp. 599-604, 2004.

[2] L. Rokach, R. Romano, and O. Maimon, "Negation Recognition in Medical Narrative Reports," Information Retrieval, vol. 11, no. 6, pp. 499-538, 2008.

[3] V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-Art in Privacy Preserving Data Mining," ACM SIGMOD Record, vol. 33, no. 1, pp. 50-57, 2004.

[4] A. Agrawal and R. Srikant, "Privacy Preserving Data Mining," ACM SIGMOD Record, vol. 29, no. 2, pp. 439-450, 2000.

[5] B. Gilburd, A. Schuster, and R. Wolff, "k-TTP: A New Privacy Model for Large-Scale Distributed Environments," Proc. 10th ACM SIGKDD, pp. 563-568, 2004.