k-Anonymization using Multidimensional Suppression for Data De-identification

Snehal M. Nargundi
IT Department
MIT College of Engineering
Pune, India

Rashmi Phalnikar
IT Department
MIT College of Engineering
Pune, India

ABSTRACT
As searching methods have advanced the increased risk of privacy disclosure makes it important to protect privacy of user during data publishing. Many of the algorithms used for the data de-identification are not efficient because resulted dataset can easily linked with the public database and it reveals the users identity. One of the method uses for protecting the privacy of user is to apply anonymization algorithms. TDS and TDR using generalization of method to anonymized the dataset. Major drawback as these algorithm is they requires a manually generated domain hierarchy taxonomy for every quasi-identifier in the data set on which k-anonymity has to be performed. Therefore, in this paper we propose new approach which will makes use of suppression based k-anonymization method to allow data publisher to de-identify datasets and in this method only certain attributes from record are suppressed based on values other attributes. As suppression method is used in algorithm, it does not required manually created taxonomy tree of quasi-identifiers. We applied this algorithm on 3 different data sets to evaluate its accuracy as compared to other k-anonymity generalization algorithms. It is found that predicative performance of this algorithm is better than existing generalization methods. This method is expected to provide privacy and accuracy measures to data publishers.

General Terms
Data Mining

Keywords
Privacy Preservation, Data Mining, Data De-identification, PPDM, k-Anonymization, Suppression

1. INTRODUCTION
Data mining algorithms are used for extracting the hidden knowledge from the large databases. Privacy preserving data mining is the new research area which deals with the protecting the sensitive data and knowledge. Nowadays, many organizations are collecting large amount of data that contains person specific information. Due to advances in data mining algorithms have increased the disclosure risks of sensitive data. Therefore, providing security to sensitive data against unauthorized access is the major goal of the data holder to protect the identity of person. Therefore the important task is to preserve the privacy of user while applying data mining algorithms on this kind of data. For instance, consider example medical research where data mining algorithms are applied on patient's data to find out the different disease patterns. Most methods used for privacy preservation removes the identifiers from the dataset and perform some transformation on remaining dataset and then this private dataset can be given to researchers. But these type data sets can be linked with the public database which reveals the identity of person.

Table 1: De-identified patient database

DOB	Sex	Zip code	Disease
1/21/76	Male	53715	Heart Disease
4/13/86	Female	53715	Hepatitis
2/28/76	Male	53703	Bronchitis
1/21/76	Male	53703	Broken Arm
4/13/86	Female	53706	Flu
2/28/76	Female	53706	Hang Nail

In Table 1 database table storing patient records are de-identified by removing name of patients and this table can be used for medical research work but this table can be linked with public database table (i.e. Table 2) which stores the voter's registration information. By matching DOB, Sex and Zip code one can easily find that "John has Heart Disease"

Table 2: Voter's Registration database

Name	DOB	Sex	Zip code
John	1/21/76	Male	53715
Smith	1/10/81	Female	55410
Kedar	10/1/44	Female	90210

2. RELATED WORK
Most of algorithm uses generalization techniques which requires generalization taxonomy tree for data de-identification as in [1],[2],[3],[4].[6]. Fung et al. [2] presented method of generalization for classification using k-anonymity: the "top-down specialization (TDS)" algorithm. TDR starts from the most general state of the table and specialized it by assigning specific values to attributes. Fung et al. [3] presented modified algorithm which is called as "Top-Down Refinement" (TDR). TDR is capable of suppressing a categorical attribute with no taxonomy tree. In TDR each refinement increases the information and decreases the anonymity since records with specific values are more distinguishable. Friedman et al. [4] presented kADET, k-anonymization is done during growing phase of a decision

tree therefore output of kADET is anonymous decision tree rather than an anonymous data set.

Kisilevich et al. [19] have proposed a multidimensional suppression approach, called kACTUS, for classification-aware anonymization. kACTUS makes use of a decision tree, i.e. C4.5 [18], as a base for deciding multi-dimensional regions to be suppressed. The pioneering work in multidimensional generalization has been proposed by Lefevre et al. [20], called Mondrian.

Major disadvantage of existing systems [2], [3], [4] is that they required taxonomy of quasi-identifiers present in data set which require prior knowledge about domain. It might be possible that for same attribute one can have different structure of taxonomy tree and it is difficult to decide correct structure of taxonomy tree. Existing methods of k-anonymization are inadequate because they cannot guarantee privacy protection in all cases, and often incur unnecessary information loss by performing excessive generalization.

In this paper we proposed new multidimensional approach which is based on suppression of attribute depending on values of other attributes. Performance of multi-dimensional suppression is better than single-dimensional suppression because it suppresses a certain value in all tuples without considering values of other attributes. This causes over suppression of data which may lead to loss of information. Main advantage of proposed method is it does not require taxonomy tree in anonymization process because suppression used.

3. K-ANONYMITY MODEL

The k-anonymity model was developed because of the possibility of indirect identification of records from public databases. This is because combinations of record attributes can be used to exactly identify individual records. K-anonymity is an anonymizing approach proposed by Samarati and Sweeney [1]. A dataset complies with k-anonymity protection if each individual's record stored in the released dataset cannot be distinguished from at least k-1 individuals whose data also appears in the dataset [1]. This protection guarantees that the probability of identifying an individual based on the released data in the dataset does not exceed 1/k.

The term data refers to person-specific information that is conceptually organized as a table of rows (or records) and columns (or fields). Each row is termed a tuple. *Quasi-identifier (QID)* is a set of features whose associated values may be useful for linking with another data set to re-identify the entity that is the subject of the data [1], [5].

While releasing private tables for research purpose identifiers are removed from the table to de-identify the person but still by matching quasi-identifiers from private table with public table one can easily identify the person. Therefore k-Anonymization is used to make at least k tuples similar by using generalization or suppression. *Generalization* is process 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 [1], [3].

Suppression refers to removing a certain attribute value and replacing occurrences of the value with a special value "*", indicating that any value can be placed instead [1].

Table 3: Example of Patient Data

Identifier	Quasi Identifiers			Sensitive Attribute
Name	Age	Country	Zipcode	Disease
Anthony	25	UK	83244	Diabetes
John	30	US	83244	Flu
Smith	22	US	83245	Cancer
Alice	45	JP	83245	Cancer
Bob	35	US	83248	Cancer
Charles	32	JP	83246	Flu
David	37	JP	83248	Diabetes
Donna	56	US	83247	Flu
Emily	66	UK	83246	Cancer
Johnson	65	UK	83245	Diabetes
Frank	58	US	83244	Flu
Edward	49	US	83248	Flu

While releasing above Table 3, identifiers (i.e. Name) will be removed from the table. Then we can apply Generalization and Suppression on Quasi–identifiers. After using Generalization on age attribute and Suppression on Zipcode attribute anonymous table will be as shown in Table 4. 2-Anonymous table means at least two rows are indistinguishable from each other.

Table 4: Example of 2-Anonymous table

Quasi Identifiers			Sensitive Attribute
Age	Country	Zipcode	Disease
20-30	UK	8324*	Diabetes
20-30	**US**	**8324***	**Flu**
20-30	**US**	**8324***	**Cancer**
30-40	JP	8324*	Cancer
30-40	US	8324*	Cancer
30-40	**JP**	**8324***	**Flu**
30-40	**JP**	**8324***	**Diabetes**
50-60	US	8324*	Flu
60-70	**UK**	**8324***	**Cancer**
60-70	**UK**	**8324***	**Diabetes**
60-70	US	8324*	Flu
40-50	US	8324*	Flu

4. SUPPRESSION BASED K-ANONYMITY FOR CLASSIFICATION TREES

Main goal of this paper is to generate anonymous data set where performance of classifier trained over anonymous data set is similar to classifier trained on the original dataset. We proposed an approach which works in two stages. In first, it generates classification tree and second, anonymization is applied to classification tree to generate anonymous data set.

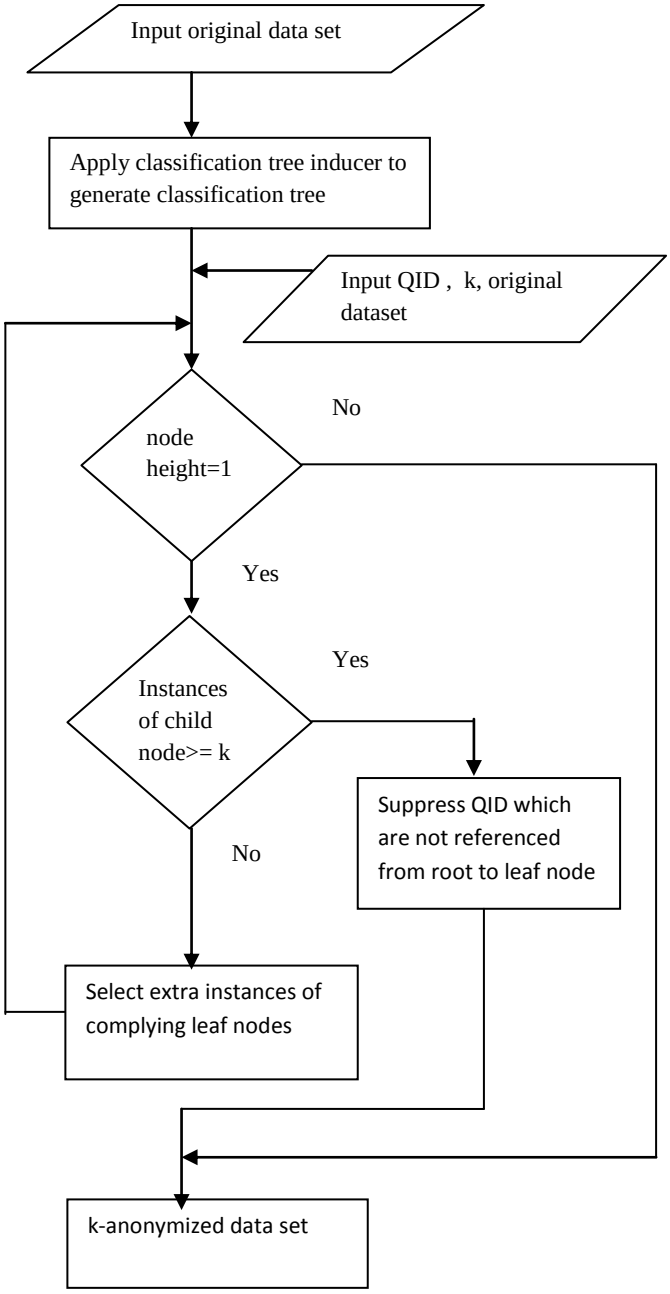


Fig 1: Stages in Algorithm

4.1 Algorithm:

4.1.1 Stage 1: Generate classification tree

In this stage, classification tree is generated form original dataset. Various decision tree inducers can be applied for to generate classification tree. We used C4.5 algorithm to generate classification tree for k-anonymization process. In this classification tree is trained over only on quasi-identifiers and each internal node of classification tree represents one attribute which is quasi-identifier.

4.1.2 Stage 2: Apply k-anonymization

In this stage, generated classification tree is given as an input for anonymization method. Following are the steps in anoymization process.

1. Find list of all nodes with height equal to 1.
2. For every node in that list check the number of instances associated with every child of that node. If instances are greater than k (threshold value) then suppress all quasi-identifiers (QID) attributes that are not referenced in one of the nodes along the path from the root.
3. If value of k is less than k then prune that leaf node or add extra instances of complying node.
4. Repeat step 1 till tree is not fully prune

From generated classification tree, every path from root to leaf node is considered as a Classification rule. Every leaf node has some instances associated with it. In step 1 find all nodes from generated classification tree with height = 1. Step 2 find children where numbers of instances are greater than k and suppress all quasi-identifiers that are not reference in one of the nodes along the path from the root to leaf node. If number of instances is less than k then prune that node and add that instances to that node to non-complying node.

5. Illustrative Example

Proposed algorithm is illustrated on patient dataset which contains information related to patient with attributes i.e. name, city, state, zipcode, age, gender, marital status and disease. We consider the value of k=2 and quasi-identifiers are {city, state, zipcode, age, gender, marital status}

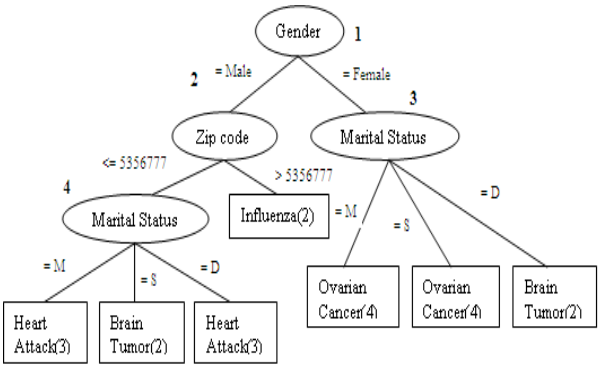
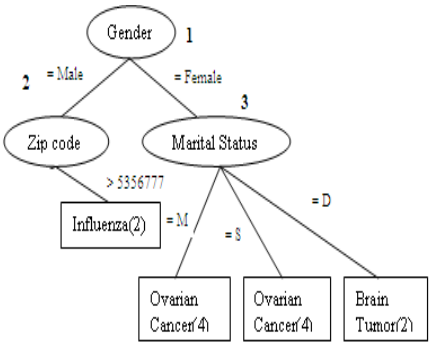


Fig 2: Classification Tree

Classification tree generated in stage 1 is as shown in Fig 2. Each leaf node represents number of instances associated with that node. Anonymization process is applied to this classification tree. Here value of k-anonymity threshold is k=2 therefore resultant anonymized dataset will contain at least two rows identical.

In stage two internal node height =1 gets selected thus we have three nodes 2, 3 and 4. Select node randomly say ,node 4 all children's of it comply with k-anonymity therefore suppress all quasi-identifiers which are not referenced from root to leaf node i.e. from root node to leaf node we have three quasi identifiers {gender, zip code , marital status} therefore remaining quasi-identifiers {city, state, age }will suppressed. After suppression extra instances will be used for non-complying children and prune that node. After pruning tree will be as follow.



Now select another node randomly which is at height=1 and repeat above steps till there is node at height=1. At the end of algorithm dataset is in 2-anonymous form

6. Measuring information loss in suppression

Previous generalization is not enough to produce tables with adequate privacy protection. Suppression is necessary to produce tables with sufficient privacy protection. We have not used generalization since the domain consistency is important for classification. Suppression produces missing values, which can be handled by most classification methods. k is a parameter of privacy requirement. Normally, the larger the k is the better protection for privacy. k is usually determined by users as an input parameter for a program. However, a large k may distort the data distribution and make models on the data useless.

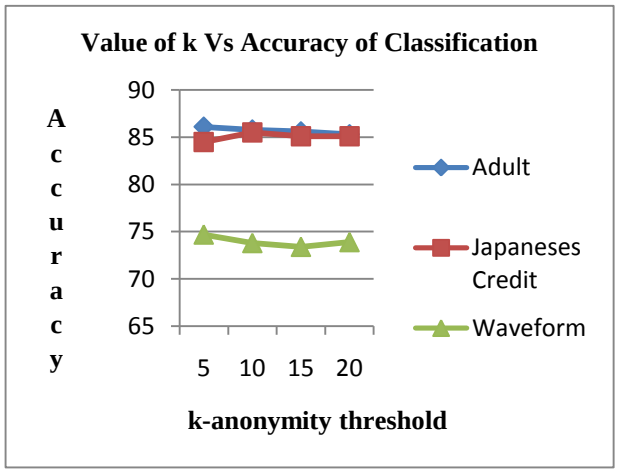
7. EXPERIMENTAL EVALUATION

We used WEKA 3.6.3 environment to test the performance of the algorithm. We used J48 as a classifier which is java version C4.5 to generate tree. In this experimental study, we used 3 data sets which were selected from the UCI Machine Learning Repository [7] and they are widely used for evaluating learning algorithms. We used various values of k-anonymity threshold to find out effect of suppression on classification task table 2 represents the values of k and classification accuracy in percentage.

Table 5: Value of k and Accuracy

Data Set	k-anonymity threshold			
	K=5	K=10	K=15	K=20
Adult	86.1	85.8	85.6	85.3
Japanese credit	84.5	85.5	85.1	85.1
Waveform	74.7	73.8	72.4	71.9

We analyze the effect of the value of k(anonymity level) on the accuracy. Accuracy results obtained by the proposed algorithm on different values of k for various data sets are shown in figure increasing the anonymity level decreases accuracy.



7.1 The Effect of k on the Accuracy

Table 5 represents value of k and classification accuracy. It is found that increasing the value of k decreases classification accuracy. Therefore information gain is less from anonymous data. Therefore value of k must be chosen properly for classification

7.2 Comparison of Suppression Methods

As existing algorithms are using generalization trees. But if generalization is absent then the performance of proposed system is better that existing generalization algorithm. As suppression is done only on tuples by examining other attributes it does not degrade the performance of classification.

8. CONCLUSION

This paper represents new method for privacy preserving in classification tasks using suppression based k-anonymity is presented. The proposed method requires no prior knowledge regarding the domain hierarchy taxonomy and suppression attributes is based on values of other attributes to avoid over suppression. Performance of this method is good as compared to existing methods which are used to avoid attacks against k-anonymity.

9. REFERENCES

[1] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, pp. 557-570, 2002.

[2] B.C.M. Fung, K. Wang, and P.S. Yu, "Top-Down Specialization for Information and Privacy Preservation," Proc. 21st IEEE Int'l Conf. Data Eng. (ICDE '05), pp. 205-216, Apr. 2005.

[3] B.C.M. Fung, K. Wang, and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 5, pp. 711-725, May 2007.

[4] A. Friedman, R. Wolff, and A. Schuster, "Providing k-Anonymity in Data Mining," Int'l J. Very Large Data Bases, vol. 17, no. 4, pp. 789-804, 2008.

[5] S.V. Iyengar, "Transforming Data to Satisfy Privacy Constraints," Proc. Eighth ACM SIGKDD, pp. 279-288, 2002.

[6] L. Tiancheng and I. Ninghui, "Optimal K-Anonymity with Flexible Generalization Schemes through Bottom-Up Searching," Proc. Sixth IEEE Int'l Conf. Data Mining Workshops, pp. 518-523, 2006.

[7] A. Asuncion and D.J. Newman, "UCI Machine Learning Repository," School of Information and Computer Science, Univ. of California, http://mlearn.ics.uci.edu/MLRepository.html, 2007.

[8] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan. (2005). Incognito: efficient full-domain K-anonymity. In Proceedings of the 2005 ACMSIGMOD international conference on Management of data(SIGMOD '05). ACM, New York, NY, USA, pp. 49-60.

[9] R.J. Bayardo and R. Agrawal. (2005). Data Privacy through Optimal k-Anonymization. In Proceedings of the 21st International Conference on Data Engineering (ICDE '05). IEEE Computer Society, Washington, DC, USA, pp. 217-228.

[10] P. Samarati, "Protecting Respondents' Identities in Microdata Release," IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov./Dec. 2001.

[11] E. Bertino, B.C. Ooi, Y. Yang, and R.H. Deng, "Privacy and Ownership Preserving of Outsourced Medical Data," Proc. Int'l Conf. Data Eng., vol. 21, pp. 521-532, 2005.

[12] G. Aggarwal, A. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, "Approximation Algorithms for k-Anonymity," J. Privacy Technology, 2005.

[13] A. Meyerson and R. Williams, "On the Complexity of Optimal k-Anonymity," Proc. 23rd ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems, pp. 223-228, 2004.

[14] Z. Yang, S. Zhong, and R.N. Wright, "Privacy-Preserving Classification of Customer Data without Loss of Accuracy," Proc.Fifth Int'l Conf. Data Mining, 2005.

[15] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee, "Toward Privacy in Public Databases," Proc. Theory of Cryptography Conf., pp. 363-385, 2005.

[16] L. Sweeney, "Datafly: A System for Providing Anonymity in Medical Data," Proc. IFIP TC11 WG11.3 11th Int'l Conf. Database Security XI: Status and Prospects, pp. 356-381, 1997.

[17] J.R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[18] I.H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools. Morgan Kaufmann, 2005.

[19] S. Kisilevich, L. Rokach, Y. Elovici, B. Shapira, Efficient multidimensional suppression for k-anonymity, IEEE Transaction on Knowledge and Data Engineering 22 (3) (2010) 334–347

[20] K. LeFevre, D.J. DeWitt, R. Ramakrishnan, Mondrian multidimensional k-anonymity, International Conference on Data Engineering (ICDE 06), IEEE Computer Society, 2006, p. 25.