Integrated Intelligent Research (IIR) International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 02 December 2013, Page No.60-63
ISSN: 2278-2419

Multidimensional Suppression for K-Anonymity in Public Dataset Using See5

B.Sathiya

Assistant Professor, M.A.M.School of Engineering, Trichy.

Abstract-Data mining is a well known technique for many applications. Privacy is very important while handling public datasets. Generalization and suppression are the two methods that are widely used for annonymizing datasets. In generalization the values are replaced with less specific but semantically consistent values and in suppression technique the value is never released. So generalization is applied for many areas since suppression may reduce the accuracy of results if not properly used. Though generalization requires manually generated domain hierarchy taxonomy for every quasi identifier in the dataset on which K-Anonymity is performed. . K-Anonymity is one of the best methods for deidentification of large public datasets. Previous K-anonymity algorithms such as kACTUS that using classifiers C4.5 for building their decision tree. In kACTUS(K-Anonymity of Classification Trees Using Suppression) efficient multidimensional suppression is performed without need for manually produced domain hierarchy trees using C4.5 algorithm In this paper better classifier such as see5 algorithm is used with kACTUS that will produce better results than previous one.

I. INTRODUCTION

THE world now deals with variety and enormous amount of datasets that contain person specific information. While dealing with large datasets so much attention need to take for protecting important details. Many algorithms are used nowadays for privacy. Data mining is the process of analyzing data from different perspective and summarizing it into useful information. Census data, voter details, customer details, medical records are some of the widely used large datasets for analysis. Before releasing this datasets some information should be deidentified like persons social security number, address, phone number. The privacy preserving data mining deals with these issues.The data should not be identified by others through released attributes. The researchers won't get accurate results when important details are deidentified. Laws such as the Health Insurance Portability and Accountability Act (HIPAA), Gramm-Leach-Bliley Act, and California's pending SB 1386 identity-protection law regulate what data companies can share.

While dealing with privacy K-anonymity is used for micro data protection anonymity concept is actually given by samarati and sweeny. K-anonymity concept is that each tuple in the micro data table released be indistinguishably related to no fewer than k respondents[1]. Generalization and suppression are the two methods that widely used in k-anonymity algorithms for deidentification of personal data. G*eneralization* involves replacing (or recoding) a value with a less specific but semantically consistent value. *Suppression* involves not releasing a value at all [2].

Generalizations perform substituting attribute values with semantically consistent but less precise values. For example, the street name is replaced by city name which occurs in more records so that the identification of a specific individual is more difficult. Even though Generalization maintains the correctness of the data at the record level the results in less specific information that may affect the accuracy of machine learning algorithms applied on the k-anonymous data set. Different systems use various methods for selecting the attributes and records for generalization as well as the generalization technique [3][9].In Suppression scheme removing a certain attribute value and replacing occurrences of the value with a special value "?," indicating that any value can be placed instead. It may reduce the quality of the data if not properly used. Because of this reason 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 [6]. One major drawback of existing generalization techniques is that manually generated domain hierarchy trees are required for every quasi-identifier attribute of data sets before k-anonymity can be applied [8], [7], [10], [11], [12], [13].kACTUS uses the wrapping approach for deidentification of data. The automatically induced tree is used by kACTUS for applying k-anonymity on dataset. Previously C4.5 algorithm is used for tree generation by kACTUS. While using See5 algorithm instead of C4.5 it produce better results in the way of smaller decision tree and faster than C4.5 and lesser memory usage.

II. RELATED WORKS

Verykios et al. [15] 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 (e.g., classification [13]);

4. Data or rule hiding referring to whether the PPDM Method hides the raw or the aggregated data; and Finally,

5. Privacy preservation, referring to the type of Technique that is used for privacy preservation: Heuristic, cryptography or reconstruction-Based (i.e., perturbing the data and reconstructing the distributions to perform mining .

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. Generalization is the most common method used for deidentification of the data in k-anonymity-based algorithms. Generalization consists of replacing specific data with a more general value to prevent individual identification; for example, the address that includes (Street, City, and State) can be replaced by (City and State) which applies to more records so that identification of a specific individual is more difficult. Wang et al presented bottom up generalization for privacy and using "information gain" metric to measure the information/privacy tradeoff. Fung et al presented top-down specialization (TDS) algorithm. It handles categorical and continuous attributes. Later he present top-down refinement (TDR) algorithm with capability of suppression using no taxonomy tree. Friedman et al present kADET, a decision tree induction algorithm that is guaranteed to maintain k-anonymity. The existing work proposes a method to achieving k-anonymity named k-anonymity of Classification Trees Using Suppression (kACTUS). In kACTUS, efficient multidimensional suppression is performed. Values are suppressed only on certain records depending on other attribute values, without the need for manually produced domain hierarchy trees. Identify attributes that have less influence on the classification of the data records and suppress them if needed in order to comply with k-anonymity. kACTUS wraps a decision tree inducer which is used to induce a classification tree from the original data set.

III. PROBLEM STATEMENT

The existing classification tree induction algorithm (such as C4.5) and is referred to as K-Anonymity of Classification Trees Using Suppression (kACTUS) [1]. The classification tree inducer is used to induce a classification tree from the original data set which indicates how the attribute values affect the target class. The classification tree can be easily interpreted by a machine in order to perform the k-anonymity process.

Table-1

OUT LOOK	TEMP RATURE	HUMIDITY	WINDY	PLAY
SUNNY	85	85	FALSE	DON'T PLAY
SUNNY	80	90	TRUE	DON'T PLAY
OVERCAST	83	76	FALSE	PLAY
RAIN	60	68	FALSE	PLAY
RAIN	64	62	FALSE	PLAY
RAIN	63	81	TRUE	DON'T PLAY
OVERCAST	75	80	TRUE	PLAY
SUNNY	68	74	FALSE	DON'T PLAY
SUNNY	75	65	FALSE	PLAY
RAIN	80	83	FALSE	PLAY
SUNNY	64	79	TRUE	PLAY
OVERCAST	61	67	TRUE	PLAY
OVERCAST	73	76	FALSE	PLAY
RAIN	81	89	TRUE	DON'T PLAY

A. Implementation Of C4.5

In this second module the obtained Dataset is displayed. The basic steps are performed. From the given Dataset the class values, Attribute names, Attribute values, and training data are evaluated and displayed. The continuous values are transformed into discrete values for tree generation. In the implementation of C4.5 all the continuous values in the Dataset are sorted and the mid value is taken. The values below the mid value are transformed as low, and above the mid value are

Transformed as high. The number of class values is calculated. The probability distribution of class values is calculated and from that *the* info (T) is gained [2]. Next for each attribute information values are calculated using this following

$$I(P) = -\sum_{i=1}^{k} p_i * \log(p_i)$$

$$I(P) = -\sum_{i=1}^{k} p_i * \log(p_i)$$

Now the information gain is calculated for each attribute in the given Dataset [2]. From the previously obtained *info* (T), and *info* (X, T) the information gain for each attribute in the given Dataset is calculated using the following formula.

Gain(X, T) = *info* (T) – *info* (X, T)

From the information gain value the root of the decision tree is obtained. The attribute which is having highest information gain value is considered as root of the decision tree.

After finding the root of the tree the sub tree are generated for the given dataset. When all the sub trees are generated the final tree structure is displayed and it is called as decision tree for the given dataset.

B. Implementation Of See5

The same Dataset which is used in the previous module is taken for this module also. The class values and the number of class values are also obtained. The continuous values are added and the average value is calculated. The values below the average are transformed as low, and the values above the average value are transformed as high.

Now find the *Gini* value for each class values and then find the *Gini* for each attribute in the Dataset find the *GiniIndex* for each attribute using the following formula

$$Gini\ Index = 1 - \sum_{j} p_j^2$$

From the calculated GiniIndex values the root of the decision tree is obtained. The attribute which is having higher *GiniIndex* value is taken as root of the decision tree for the given dataset.The sub trees are generated as like previous model except the tree is stopped when the information T is 1

C. Comparison

The decision trees generated by both the algorithms are taken into comparison. The decision tree generated by See5 is having less computational complexity when compared to C4.5.

International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 02 December 2013, Page No.60-63
ISSN: 2278-2419

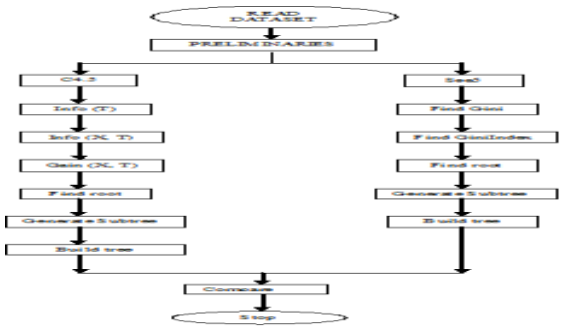
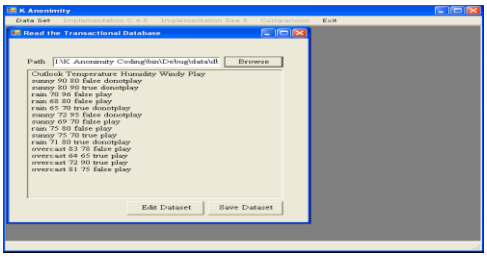


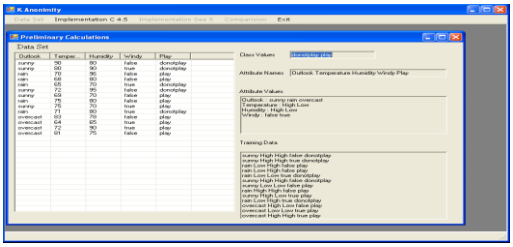
Fig.1.Overall representation of comparison between C4.5 and See5.

D. Results

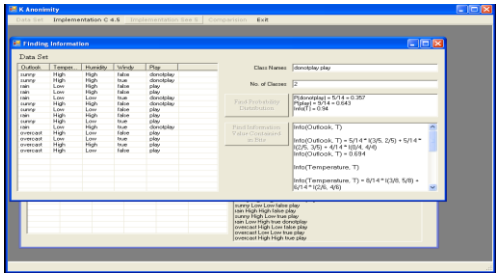
Initially the dataset is read for further processing.



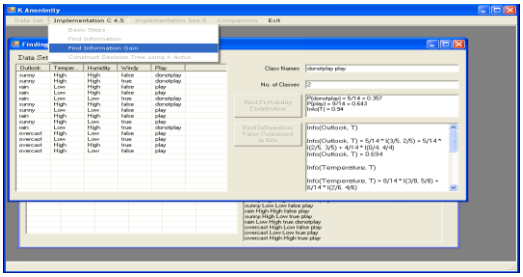
The class values, attribute names, attribute values, training data are displayed



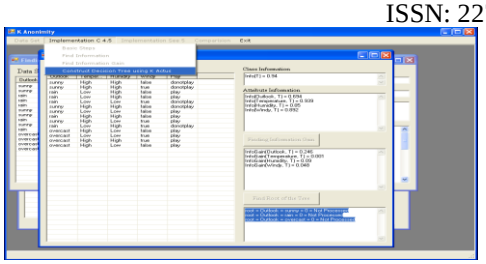
Find information value for given attributes.



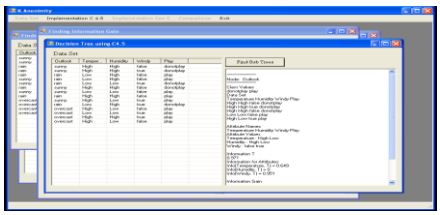
Find the information gain values.



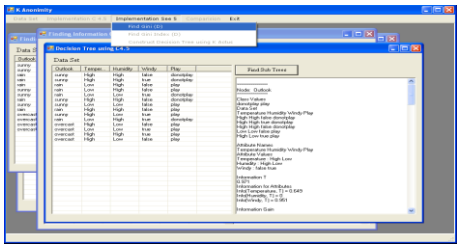
The decision tree construction begins.



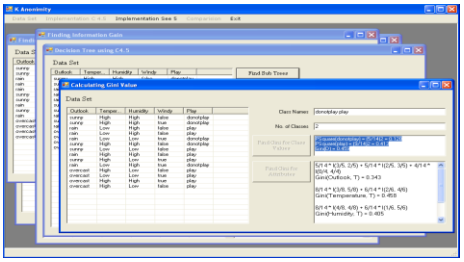
Each attribute is processed and the sub trees are generated.



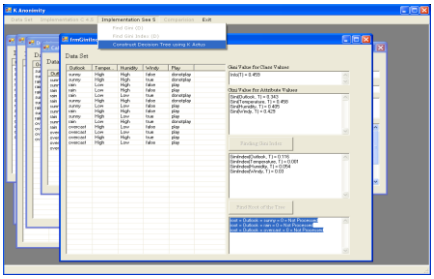
The implementation of See5 begins



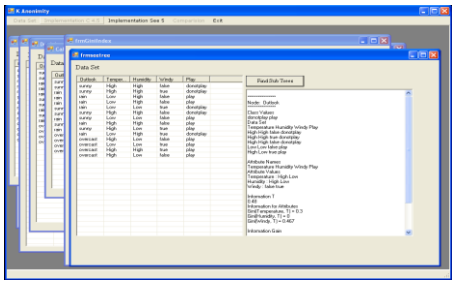
Calculate the gini for attributes



The tree construction phase begins.

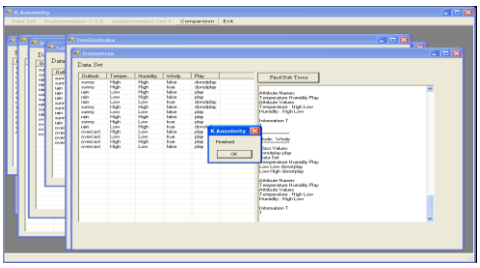


Each attribute is processed and the sub trees are generated



A message box is displayed to shown that the sub tree generation finished

International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 02 December 2013, Page No.60-63
ISSN: 2278-2419



IV. CONCLUSION

This project work proposes 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. This proposed method can be improved with 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] Slava Kisilevich, Lior Rokach, Yuval Elovici, "Efficient Multidimensional Suppression for K-Anonymity," IEEE Transactions on Knowledge and Data Engineering, vol.22, no. 3, pp. 334-347, March 2010.

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

[3] 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.

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

[5] M.S. Wolf and C.L. Bennett, "Local Perspective of the Impact of the HIPAA Privacy Rule on Research," Cancer-Philadelphia Then Hoboken, vol. 106, no. 2, pp. 474-479, 2006.

[6] P. Samarati and L. Sweeney, "Generalizing Data to Provide Anonymity When Disclosing Information," Proc. 17th ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems, vol. 17, p. 188, 1998.

[7] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertainty, Fuzziness, and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.

[8] L. Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression," Int'l J. Uncertainty, Fuzziness, and Knowledge-Based Systems, vol. 10, no. 5, pp. 571-588, 2002.

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

[10] K. Wang, P.S. Yu, and S. Chakraborty, "Bottom-Up Generalization: A Data Mining Solution to Privacy Protection," Proc.Fourth IEEE Int'l Conf. Data Mining, pp. 205-216, 2004.

[11]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.

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

[13]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.

[14] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient Full Domain k-Anonymity," Proc. 2005 ACM SIGMOD, pp. 49-60, 2005.

[15]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.