April 2012

Optimization through Bayesian Classification on the k-Anonymized Data

L Mohana Tirumala

Department of CSE, MVGR College of Engineering, Vizianagaram, India, mohanatirumala@gmail.com

S. Srinivasa Rao

Department of CSE, MVGR College of Engineering, Vizianagaram, India, siringi35@yahoo.co.in

Follow this and additional works at: <https://www.interscience.in/ijcsi>

Part of the [Computer Engineering Commons](#), [Information Security Commons](#), and the [Systems and Communications Commons](#)

Recommended Citation

Tirumala, L Mohana and Rao, S. Srinivasa (2012) "Optimization through Bayesian Classification on the k-Anonymized Data," *International Journal of Computer Science and Informatics*: Vol. 1 : Iss. 4 , Article 4.

DOI: 10.47893/IJCSI.2012.1046

Available at: <https://www.interscience.in/ijcsi/vol1/iss4/4>

This Article is brought to you for free and open access by the Interscience Journals at Interscience Research Network. It has been accepted for inclusion in International Journal of Computer Science and Informatics by an authorized editor of Interscience Research Network. For more information, please contact sritampatnaik@gmail.com.

Optimization through Bayesian Classification on the k-Anonymized Data

L Mohana Tirumala & S. Srinivasa Rao

Department of CSE, MVGR College of Engineering, Vizianagaram, India

E-mail : mohanatirumala@gmail.com, siringi35@yahoo.co.in

Abstract - Privacy preserving in Data mining & publishing, plays a major role in today networked world. It is important to preserve the privacy of the vital information corresponding to a data set. This process can be achieved by k-anonymization solution for classification. Along with the privacy preserving using anonymization, yielding the optimized data sets is also of equal importance with a cost effective approach. In this paper Top-Down Refinement algorithm has been proposed which yields optimum results in a cost effective manner. Bayesian Classification has been proposed in this paper to predict class membership probabilities for a data tuple for which the associated class label is unknown.

Keywords - Privacy preserving, classification, anonymization, Bayesian classification.

I. INTRODUCTION

Advancement in technology has facilitated the world in storing large data sets. The collection of the data sets related to an individual from various sources has paved a path to increased necessity to preserve the privacy of individual information. An example given [4] is that a Sensitive medical record was uniquely linked to a named Voter record in a publicly available voter list through the shared attributes of Zip, Birthdate and Sex. When the sources are cross-examined, protection of individual sources does not guarantee privacy.

Hence a huge research is carried out in the field of privacy preserving in data mining. To adopt privacy, a technique k-anonymity is used to attain privacy protection by revealing the information safety and validating the truthfulness of an individual's record. It is also necessary to obtain an optimized result & to minimise the complexity in extracting the desired results. For this TOP-DOWN Refinement algorithm is used. Predicting Class membership probability is also necessary when the data is large and class label is unknown. This feature cannot be satisfied by Top-Down Refinement algorithm. For this purpose Bayesian Classification has been proposed.

Advantages of our approach are:

- 1) It is simpler than decision tree and selected neural network classifier in performance.

- 2) It has exhibited high accuracy and speed when applied to large databases.
- 3) Bayesian classifiers have the minimum error rate in comparison to all other classifiers.

The rest of the paper is organized as follows: Section –II depicts about the related work and section – III discusses about the existing problems and in section – IV explains about Bayesian classification. The experimental results are discussed in section – V and section – VI concludes the paper.

II. RELATED WORK

A well studied technique for privacy preservation and to prevent record linkage through QID is k-anonymity [5,6,7,8,9,10,12,13]. The k-anonymity model assumes that each record represents a distinct individual. If several records in a table represent the same record owner, a group of k records may represent fewer than k record owners, and the record owner may be under protected.

Bottom-up generalization was studied [2,3]. It improves the k-anonymization. It starts from the original data which violates k-anonymity. It does generalization operation at each step according to a search metric similar to information lost per each gain of privacy (ILPG). After each step size of group increases. This operation contains until all groups get minimum size. It normally selects critical generalizations, with which the

size of the minimum group can be increased. When there are no critical generalizations, it considers other generalizations. In this process some of the information can be lost for maintaining privacy.

Top-Down refinement algorithm was studied [1]. Instead of masking the most specific table bottom-up, masked values are refined top-down starting from

the most masked table. Different types of categorical attributes are handled. The user can stop the Refinement at any time and have a table satisfying the anonymity requirement.

We considered Fig.1 and TABLE1 [1].

Fig. 1 : Taxonomy Tree

Table 1 : Refinement of Table

EDUCATION	SEX	WORK_hrs	CLASS	# OF RECs
9 th	M	30	0Y3N	3
10 th	M	32	0Y4N	4
11 th	M	35	2Y3N	5
12 th	F	37	3Y1N	4
Bachelors	F	42	4Y2N	6
Bachelors	F	44	4Y0N	4
Masters	M	44	4Y0N	4
Masters	F	44	3Y0N	3
Doctorate	F	44	1Y0N	1
		TOTAL:	21Y13N	34

N: Income $\leq$ 50K, Y: Income $>$ 50K

Output is shown below.

A Final mask table after applying TOP-DOWN Refinement Algorithm:

Fig. 2 : After Top-Down Refinement

Table – 2. : Result after Top-Down Refinement

EDUCATION	SEX	WORK_hrs	CLASS	# OF RECs
JUNIOR SEC	Any SEX	[1-37]	0Y7N	7
11 th	Any SEX	[1-37]	2Y3N	5
12 th	Any SEX	[1-37]	3Y1N	4
Bachelors	Any SEX	[37-99]	4Y2N	6
Bachelors	Any SEX	[37-99]	4Y0N	4
GRAD School	Any SEX	[37-99]	8Y0N	8
		TOTAL	21Y13N	34

This output shows that the Top-Down Refinement algorithm can improve the privacy over Bottom-up generalization, but cannot predict the class membership probability. To achieve this we proposed Bayesian classification.

III. THE PROBLEM

A data owner can release person specific data table $T(D_1, \dots, D_m, \text{class})$ to the public for modelling the class label Class. A record has the form $(V_1, \dots, V_m, \text{cls})$ where V_i is a domain value for D_i and cls is a class for Class. $\text{att}(v)$ denotes the attribute of a value v . The data owner can release some attributes to the public without compromising the privacy of an individual to the sensitive information. These attributes are called Quasi identifiers, represented by QID. A data recipient wants to predict class membership of a data tuple.

IV. BAYESIAN CLASSIFICATIONS

Bayesian classification can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Let X is a data tuple. In Bayesian terms, X is considered “evidence”. It is described by measurements made on a set of n

attributes. Let H be some hypothesis, such as that the data tuple X belongs to a specified class C .

$$P(H/X) = P(X/H) P(H)/P(X)$$

Here $P(H/X)$ is the posterior probability of H conditioned on X . For example, data tuples is confined to customers described by the attributes age, working hours and income respectively, and that X is a 45 years old customer with an income of 40,000 and working hours 8. Suppose that H is the hypothesis that our customer will buy a computer.

Similarly, $P(X/H)$ is the posterior probability of X conditioned on H .i.e., it is the probability that a customer X is 45 years old and earns 40000 and working hours is 8 hours, given that we know the customer will buy a computer.

$P(X)$ is the prior probability of X . It is the probability that a person from our set of customers is 45 years old, earns 40000, and working hours is 8.

Algorithm for Bayesian Classification:

- 1) Initialize Data partition D , which is a set of training tuples and their associated class labels. Let X represents n dimensional attribute vector.
- 2) Find the value

$$\begin{aligned}
 P(X/C_i) &= \prod_{k=1}^n P(x_k/C_i) \\
 &= P(x_1/C_i) * P(x_2/C_i) * \dots * P(x_n/C_i).
 \end{aligned}$$

- 3) Find the value $P(C_i) = P(X/C_i) P(C_i)$ // Similarly find the probabilities for the all Classes in the given database
- 4) If $P(X/C_i) P(C_i) > P(X/C_j) P(C_j)$ then Return Class C_i as maximum predict variable
- 5) Else
Return Class C_j as maximum predict variable

Algorithm describes our Bayesian Classification process. Line 1 describes the database D contains training set of tuples and their associated class labels. As usual, each tuple is represented by an n -dimensional attribute vector, $X=(x_1, x_2, \dots, x_n)$, depicting n measurements made on the tuple from n attributes, respectively, $A_1, A_2, \dots, A_n$. Line 2 describes computing the value of the $P(X/C_i)$. Line 3 describes computing the value of the posterior probability $P(C_i)$ of the Class C_i . Similarly it finds the remaining posterior probabilities in the Classes in given database. Line 4 describes, if the Posterior probability of the Class C_i is greater than the Posterior probability of the Class C_j then returns Class C_i . Line 5 describes the else condition of the above condition, it returns Class C_j .

Example: The data tuples are described by the attributes Education, Sex, Work_hrs. The class label attribute, income has two distinct values (namely, {yes, no}). Let C_1 correspond to the class income > 50K = yes and C_2 correspond to the income ≤ 50K = no.

The tuple we wish to classify is $X=(\text{Education}=11^{\text{th}}, \text{Sex}=\text{M}$ and $\text{Work_hrs}=37)$

We need to maximize $P(X/C_i)P(C_i)$ for $i=1,2$.

$$P(\text{income} = \text{yes}) = 21/34 = 0.617$$

$$P(\text{income} = \text{no}) = 13/34 = 0.382$$

To compute $P(X/C_i)$ for $i=1,2$, we compute the following conditional probabilities.

$$P(\text{Education} = 11^{\text{th}} / \text{income} = \text{yes}) = 2/21 = 0.095$$

$$P(\text{Education} = 11^{\text{th}} / \text{income} = \text{no}) = 3/13 = 0.230$$

$$P(\text{Sex} = \text{M} / \text{income} = \text{yes}) = 6/21 = 0.2857$$

$$P(\text{Sex} = \text{M} / \text{income} = \text{no}) = 10/13 = 0.769$$

$$P(\text{Work_hrs} = 37 / \text{income} = \text{yes}) = 3/21 = 0.142$$

$$P(\text{Work_hrs} = 37 / \text{income} = \text{no}) = 1/13 = 0.076$$

Using the above probabilities, we obtain

$$\begin{aligned}
 P(X/\text{income}=\text{yes}) &= P(\text{Education}=11^{\text{th}} / \text{income}=\text{yes}) \times \\
 &P(\text{Sex}=\text{M} / \text{income}=\text{yes}) \times P(\text{Work_hrs}=37 / \text{income}=\text{yes}) \\
 &= 0.095 \times 0.2857 \times 0.142 = 0.0038
 \end{aligned}$$

Similarly,

$$\begin{aligned}
 P(X/\text{income}=\text{no}) &= P(\text{Education}=11^{\text{th}} / \text{income}=\text{no}) \times \\
 &P(\text{Sex}=\text{M} / \text{income}=\text{no}) \times P(\text{Work_hrs}=37 / \text{income}=\text{no}) \\
 &= 0.230 \times 0.769 \times 0.0134 = 0.0134
 \end{aligned}$$

To find the class C_i , that maximizes $P(X/C_i) P(C_i)$, we compute

$$P(X/\text{income}=\text{yes}) P(\text{income}=\text{yes}) = 0.0038 \times 0.617 = 0.00234$$

$$P(X/\text{income}=\text{no}) P(\text{income}=\text{no}) = 0.0134 \times 0.382 = 0.00511$$

Therefore, the Bayesian Classifier predicts income=no for tuple X_i .

V. EXPERIMENTAL RESULTS

The Experiments were carried out on Adult data set from UCI Machine learning repository[11]. Model graphs are shown below.

Fig. 3 : Bayesian Classification

Fig. 4 : Decision Tree

From the above graphs, DT4, DT7 and DT10 are having equal probabilities for YES and NO. For these data tuples prediction of class label is difficult in Decision Tree. Where as it is much easier in Bayesian Classification.

Hence, it is clear that predicting class labels in Bayesian Classification is more accurate than Decision tree.

VI. CONCLUSION

In this paper, we defined a novel approach for preserving the privacy of an individual's data in a most cost effective manner and derived much optimum result set. It is also observed that the information loss is also too low when compared to existing techniques. Bayesian Classification is proposed for predicting class membership probabilities.

REFERENCES

- [1] Benjamin C.M.Fung, Ke Wang, and Philip S.Yu, "Anonymizing classification data for privacy preservation" in proc. Of the IEEE Transactions On Knowledge And Engineering, vol 19, NO 5,may 2007.
- [2] K. Wang, P. Yu, and S. Chakraborty, "Bottom-up generalization: a data mining solution to privacy protection." in Proc. of the 4th IEEE International Conference on Data Mining (ICDM), November 2004.
- [3] Benjamin C.M.Fung, Ke Wang, and Philip S.Yu "Privacy-preserving data publishing: a survey of recent developments"
- [4] L.Sweeney. "k-Anonymity – A model for projecting privacy" International Journal on uncertainty, fuzziness and knowledge-based systems, 10(5):557-570,2002.
- [5] L.Sweeney. "Achieving k-anonymity privacy protection using generalization and suppression" International Journal on Uncertainty, Fuzziness and Knowledge-based systems,10(5):571-588,2002.
- [6] M.ErcanNergiz and C.Clifton. "Thoughts on k-anonymization". In Proc of 22ndIntl.Conf on Data Engineering(ICDEW06) , IEEE Computer Society,2005.
- [7] Machanavajjhala A., Gehrke J., Kifer D., " l-diversity : privacy beyond k-anonymity ". Proceedings of the 22nd IEEE Intl.Conf. on Data Engineering,2006.
- [8] Ninghui Li, Tiancheng Li and Suresh . V." t-Closeness: Privacy beyond k-anonymity and l-diversity". ICDE 2007, 23rd IEEE Intl.Conf. on Data Engineering , 2007
- [9] Kristen LeFevre, David J. DeWitt and RaghuRamakrishna. "Multidimensional k-anonymity" Dept. Of Computer Sciences Technical Report 1521,2005
- [10] Kristen LeFevre ,David J,DeWitt and Raghu Ramakrishnan "Incognito : Efficient Full Domain k-Anonymity" . In Proceedings of ACM SIGMOD'05, USA, 2005
- [11] D.J.Newman, S.Hettich, C.L.Blake and C.J.Merz, "UCI repository of machine learning Database", available at [www.ics.uci.edu/~mllearn/ MLRepository.html](http://www.ics.uci.edu/~mllearn/MLRepository.html), University of California,Irvine,1998.
- [12] Diur. I., Nissim. K.,Revealing information while preserving privacy, In Proc ACM, San Diego, CA., pp202-210, 2003.
- [13] Pierangela S., Sweeney L, Protecting Privacy when Disclosing Information k-Anonymity and its Enforcement through Generalization and Suppression, Intl. Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10(5), pp 571-588,2002.