Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 8 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Performance Evaluation of K-Anonymized Data

By J. Paranthaman & Dr. T. Aruldoss Albert Victoire
University College of Engineering, India

Abstract - Data mining provides tools to convert a large amount of knowledge data which is user relevant. But this process could return individual's sensitive information compromising their privacy rights. So, based on different approaches, many privacy protection mechanism incorporated data mining techniques were developed. A widely used micro data protection concept is k-anonymity, proposed to capture the protection of a micro data table regarding re-identification of respondents which the data refers to. In this paper, the effect of the anonymization due to k-anonymity on the data mining classifiers is investigated. Naïve Bayes classifier is used for evaluating the anonymized and non-anonymized data.

Keywords : *data mining, privacy-preserving data mining, k-anonymity, naïve bayes.*

GJCST-C Classification : *D.2.m*

Strictly as per the compliance and regulations of:

Performance Evaluation of K-Anonymized Data

J. Paranthaman [α] & Dr. T. Aruldoss Albert Victoire [σ]

Abstract - Data mining provides tools to convert a large amount of knowledge data which is user relevant. But this process could return individual's sensitive information compromising their privacy rights. So, based on different approaches, many privacy protection mechanism incorporated data mining techniques were developed. A widely used micro data protection concept is k-anonymity, proposed to capture the protection of a micro data table regarding re-identification of respondents which the data refers to. In this paper, the effect of the anonymization due to k-anonymity on the data mining classifiers is investigated. Naïve Bayes classifier is used for evaluating the anonymized and non-anonymized data.

Keywords : data mining, privacy-preserving data mining, k-anonymity, naïve bayes.

I. Introduction

Data mining technology provides tools to transform large amount of data into knowledge useful to the user [1]. Knowledge extracted from data mining is expressed as association rules, decision trees or clusters, permitting one to locate interesting patterns/regularities in data which facilitates decision making [2]. Such knowledge discovery can inadvertently return individual sensitive information compromising their privacy. They could also reveal business information, compromising free competition. So confidential personal information disclosure and that of sensitive information should be prevented [3].

Great effort was recently devoted to overcoming privacy preserving problems in data mining leading to many data mining techniques with privacy protection mechanisms. Sanitization techniques were proposed to hide sensitive items/patterns based on removing reserved information or by noise insertion into data. Privacy preserving classification procedures thwart data miners from using classifier to predict sensitive data. Additionally, privacy preserving clustering techniques which distort sensitive numerical attributes were also suggested while retaining general features/clustering analysis [4].

Privacy issue also includes commercial concerns. Organizations collect individuals' information for particular needs, but different departments might need to share such information. Then, each organization/unit must ensure that individual privacy is not violated, nor sensitive business information revealed [5]. Confidentiality is a major issue in mass data collection. Privacy needs could be due to law or due to motivated business interests. But some data sharing situations could lead to mutual gain. Research – scientific, economic or market oriented – is a key database utility. The medical field gains by data pooling for research and even for competing businesses with mutual interests. Increasing confidentiality issues ensure it is impossible to attain any potential gain.

Privacy-preserving data mining use algorithms on confidential data that are to be unknown even to the algorithm operator. PPDM has twofold considerations. First, names and addresses which are sensitive raw data identifiers should be modified/trimmed from the original database, to ensure that data recipient does not compromise another's privacy. Second, sensitive knowledge from a database mined with data mining algorithms should be kept out as such knowledge can also compromise data privacy [6]. Users' personal information and information concerning their collective activity are two major privacy preservation dimensions. The former is called individual privacy preservation and the latter collective privacy preservation.

Privacy-preserving data mining is split into 2 parts: data hiding and rule hiding. Data hiding converts data or designs new computation protocols to ensure that private data is private during/after data mining ensuring recovery of data patterns/models while the capable of discovery. Additive perturbation, multiplicative perturbation, and secure multi-party computation techniques come into this category. Rule hiding, in contrast transforms the database to ensure masking of sensitive rules while underlying patterns can be discovered [7].

Privacy preservation protects individual identifications and sensitive relationships [8]. An emerging micro data protection is the *k-anonymity concept*, recently proposed as a property to capture micro data table protection regarding respondent's re-identification which data referred to [9]. *k*-anonymity demands that micro data table tuple be released and be related to a specific number of *k* respondents. An interesting *k*-anonymity aspect is its connection to protection techniques preserving data bonafides. K-anonymity concept captures on the private table PT yearning for release, a main requirement followed by statistical community and by data releasing agencies, that released data be related to a specific number of respondents. The private table attribute set is available externally and hence capable of linking is called *quasi-identifier*. The stated requirement is translated in the *k*-anonymity requirement below, stating that all released

Author α : University College of Engineering, India.
E-mail : paran_2013@rediffmail.com

tuples should be related to a certain number of *k* respondents.

In this paper, the effect of the anonymization due to k-anonymity on the data mining classifiers is investigated. The data is anonymized for different granularity. Naïve Bayes classifier is used for evaluating the anonymized and non-anonymized data. The following sections deal with related works, methods, experimental results and discussion.

II. Related Works

K-anonymity of Classification Trees Using Suppression (kACTUS), a new method to achieve k-anonymity was proposed by Kisilevich et al [10]. kACTUS performs efficient multi-dimensional suppression where values are suppressed by certain records based on other attribute values, without manually-produced domain hierarchy trees. kACTUS identifies attributes with reduced influence on data records classification suppressing them to comply with k-anonymity. kACTUS was evaluated for accuracy on ten separate datasets compared to other k-anonymity generalization and suppression methods. Results proved that kACTUS' predictive performance is better than current k-anonymity algorithms. TDS, TDR and kADET accuracies on average are lower than kACTUS in 3.5%, 3.3% and 1.9% respectively inspite of manually defined domain tree usage. Accuracy goes up to 5.3%, 4.3% and 3.1% respectively when domain trees are left unused.

A new data record anonymizing method was proposed by Aggarawal et al [11], where data records quasi-identifiers are first clustered with cluster centers then being published. To ensure data records privacy, a constraint that clusters contain a pre-specified number of data records was imposed. This technique has a bigger choice for cluster centers than k-Anonymity. In most cases, it releases more information without privacy compromises. Clustering is through a constant-factor approximation algorithm. This algorithm set is for anonymization problem where performance does not depend on anonymity parameter k. Extended algorithms ensure that a fraction of points remain unclustered through deletion from anonymized publication. Release of a fraction of database records ensures that data published for analysis is useful as it has less distortion.

A new globally optimal de-identification algorithm satisfying k-anonymity criterion suiting health datasets was developed and evaluated by El Emam et al [12]. It was empirically compared to OLA (Optimal Lattice Anonymization) and to Datafly, Samarati, and Incognito, three existing k-anonymity algorithms, on six public, hospital, and registry datasets for different values of k and suppression limits. Precision, discernability metric, and non-uniform entropy, three information loss metrics were compared, and each algorithm's performance speed was evaluated. The Datafly and Samarati algorithms ensured higher information loss than OLA and Incognito; OLA was quicker regularly than Incognito in locating a globally optimal de-identification solution.

An (α, k)-anonymity model to protect data's identification and relationship to sensitive information was proposed by Wong et al [13]. The properties of (α, k)-anonymity model were discussed. That the optimal (α, k)-anonymity problem is NP-hard is proved. The (α, k)-anonymity problem had an optimal global recoding method being presented. A more scalable and less data distortion local-recoding algorithm was proposed next, and its effectiveness/efficiency was proved by experiments.

III. Materials and Methods

A total of 22 attributes with 8124 tuples is in the mushroom data set with each tuple recording physical characteristics of a single mushroom. A poisonous or edible classification label is provided to each tuple. The numbers of edible and poisonous mushrooms in the dataset include 4208 and 3916, respectively.

a) K-Anonymity

Data refers to person-specific information conceptually organized as rows (or records) and columns (or fields) with each row being termed a tuple having a relationship among values set linked to a person. Tuples in a table is not necessarily unique. An attribute is a column denoting a field/semantic category of information which could be a set of possible values; hence, an attribute is also a domain. Attributes are unique within a table. In a table, each row is an ordered n-tuple of values <d1, d2, …,dn> so that each value dj is in the domain of the j-th column, for j=1, 2, …, n where n is the column number, A relation corresponds with this tabular presentation in mathematical set theory, the difference being the absence of column names [9].

Let B(A1,…,An) be a table with finite tuples. The finite attributes set of B are {A1,…,An}. All attributes are to be identified by the data holder in private information that can link external information. Such attributes not only include name, address, and phone number as explicit identifiers, but also include attributes that when combined can uniquely identify individuals through birth date and gender. Such attributes set is called a quasi-identifier. In anonymity, linking should be prevented on publicly available data so that private and public data and are candidates for linking; hence, such attributes include a quasi-identifier where attributed disclosure should be controlled. Data holders can easily identify such attributes.

To find out how many individuals a released tuple matches, needs a combination of released data and available data externally, along with analysis of other possible attacks. Such a direct determination is

tough for data holders who release information. That data holders know which data in PT appear externally is assumed and also what constitutes a quasi-identifier but external data specific values cannot be assumed. Thus, if RT(A1,...,An) be a table and QIRT be associated quasi-identifier, RT can satisfy k-anonymity only if values of each sequence in RT[QIRT] appear with k occurrences in RT[QIRT] [9].

K anonymity guarantee is that an attacker will be unable to link private information with groups of less than k individuals, ensured by making sure that every public attribute values combination in the release is in at least k rows. The k-anonymity privacy model was studied intensively in a public data releases context where database owner want to ensure that nobody will be able to link database information to individuals from whom it was collected. This method could also provide anonymity in other contexts like anonymous message transmission and location privacy.

b) Naive Bayes Classifier

Classifiers predict class membership probabilities like probability of a given term to belong to a particular class. Common classification algorithm of Bayesian is the Naïve Bayesian classifier with accuracy and speed when applicable to huge dataset. A brief summation of the classifier is given below as Naïve Bayesian classifiers are extensively used.

Let *D* be a training documents set and associated class labels. Each document is represented by an n-dimensional attribute vector, $V = (v_1, ..., v_n)$. $C_1, ..., C_m$ represents *m* classes. The classifier predicts by matching test document to class with the highest posterior probability. Naïve Bayesian classifier predicts that document V belongs to the class C_i if $P(C_i|V) > P(C_j|V)$ for $1 \leq j \leq m, j \neq I$ Maximizing $P(C_i|V)$, the class C_i for which $P(C_i|V)$ is maximized is called maximum posteriori hypothesis. By Bayes theorem,

$$P(C_i|V) = \frac{P(V|C_i)P(C_i)}{P(V)}$$

As P(X) is constant for all classes, only $P(V|C_i)P(C_i)$ needs maximization. If class prior probabilities are unknown, then it is thought that classes are equally likely, and then only $P(V|C_i)P(C_i)$ is maximized.

But it is computationally expensive to compute $P(V|C_i)$. Naïve assumption of class conditional independence is made to reduce computation.

$$P(V|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

To predict V class label, $P(V|C_i)P(C_i)$ is evaluated for each class $C_{i,\text{ with the}}$ classifier predicting that class label of V document is class C_i if $P(V|C_i)P(C_i) > P(V|C_j)P(C_j)$ for $1 \leq j \leq m, j \neq i$. The predicted class label is class C_i for which $P(V|C_i)P(C_i)$ is maximum. This classifier's empirical study compared to a decision tree revealed that it is comparable in some domains. Bayesian classifiers has minimum error rate of all classifiers.

IV. Results and Discussion

Experiments are conducted for different levels of k-anonymity (5, 10,…, 45, 50). The anonymized data is classified using Naïve Bayes classifier. The following Figures and Tables give results for classification, precision and recall.

Table 1 : Classification Accuracy for different levels of K-anonymity

K-Anonymity Level	Classification Accuracy
No anonymization	0.958272
K=5	0.954333
K=10	0.94707
K=20	0.9371
K=25	0.934515
k=30	0.923929
k=35	0.917528
k=40	0.914082
k=45	0.908419
k=50	0.907189

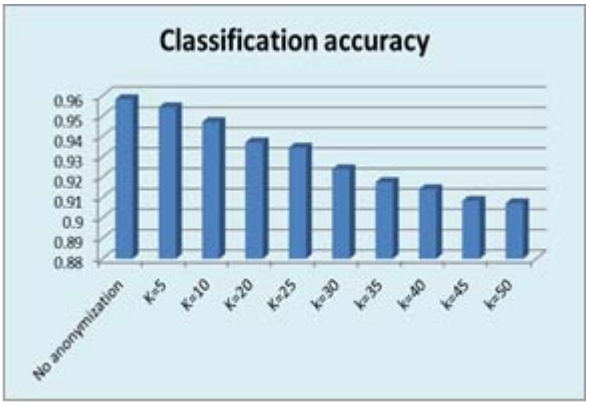

Figure 1 : Classification Accuracy for Different Levels of K-anonymity

Figure 1 reveals that classification accuracy decreases when k-anonymity level increases. Table 2 and Figure 2 show precision and recall for different levels of k-anonymity.

Table 2 : Precision and Recall

K-Anonymity Level	Precision	Recall
No Anonymization	0.961	0.957
K=5	0.957	0.953
K=10	0.950	0.946
K=20	0.940	0.936
K=25	0.937	0.933
K=30	0.925	0.923
K=35	0.919	0.917
K=40	0.915	0.913
K=45	0.909	0.908
K=50	0.908	0.906

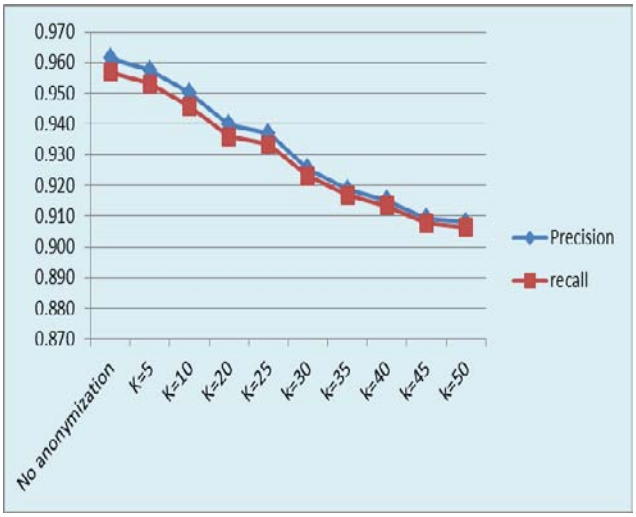

Figure 2 : Precision and Recall for Different Levels of K-anonymity

It is observed from the experimental results that with the increase in the anonymity the performance of the classifiers degrades proportionately. Further work is required to define privacy preserving methods which reduce the negative performance of the classifiers.

V. Conclusion

Privacy-preserving data mining's basic idea was extending data mining techniques to work with sensitive information masked modified data. What was at issue here was how to modify data and how to recover data mining result from it. Solutions were linked to data mining algorithms under study. This paper investigated anonymization effect due to k-anonymity on the data mining classifiers. Data is anonymized for different granularity. Naïve Bayes classifier evaluated anonymized and non-anonymized data with results showing that anonymity increase lead to proportional degradation of classifier performance.

References Références Referencias

1. K. R. Venugopal, K. G. Srinivasa and L. M. Patnaik, 2009," Soft Computing for Data Mining Applications", Studies in Computational Intelligence, Volume 190, Springer-Verlag Berlin Heidelberg.
2. Joyce Jackson, 2002, Data Mining: A Conceptual Overview Communications of the Association for Information Systems (Volume 8), pp. 267-296.
3. ArisGkoulalas-Divanis, YucelSaygin, Vassilios S. Verykios, (2011), Transactions on Data Privacy, 1st ECML/PKDD Workshop on Privacy and Security issues in Data Mining and Machine Learning, Volume 4 Issue 3, pp. 127 – 128.
4. GayatriNayak, A Survey On Privacy Preserving Data Mining: Approaches And Techniques International Journal of Engineering Science and Technology (IJEST) International Journal of Engineering Science and Technology (IJEST).
5. ArisGkoulalas-Divanis, YucelSaygin, Vassilios S. Verykios, (2011), Transactions on Data Privacy, 1st ECML/PKDD Workshop on Privacy and Security issues in Data Mining and Machine Learning, Volume 4 Issue 3, pp. 127 – 128.
6. Ge, X., & Zhu, J. (2011). Privacy preserving data mining. *New Fundamental Technologies in Data Mining,* 535-560.
7. Bhaduri, K., Das, K., & Kargupta, H. (2007). Peer-to-peer data mining, privacy issues, and games. In *Autonomous Intelligent Systems: Multi-Agents and Data Mining* (pp. 1-10). Springer Berlin Heidelberg.
8. V. Ciriani, S. De Capitani di Vimercati, S. Foresti and P. Samarati, (2007), *k*-Anonymity, Springer US, Advances in Information Security, http://www.springerlink.com/content/ht1571nl63563x16/fulltext.pdf
9. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.
10. Kisilevich, S., Rokach, L., Elovici, Y., & Shapira, B. (2010). *European Patent No. EP 2228735*. Munich, Germany: European Patent Office.
11. Aggarwal, G., Feder, T., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D., & Zhu, A. (2006, June). Achieving anonymity via clustering. In *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 153-162). ACM.
12. El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., ...& Bottomley, J. (2009). A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association, 16*(5), 670-682.
13. Wong, R. C. W., Li, J., Fu, A. W. C., & Wang, K. (2006, August). (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 754-759). ACM.

14. Chotirat "Ann" Ratanamahatana, Dimitrios Gunopulos, 'Scaling up the Naive Bayesian Classifier', Computer Science Department University of California Riverside, CA 92521 1-909-787-5190.