ISSN 2047-3338

Evaluating the Classification Accuracy of Data Mining Algorithms for Anonymized Data

M. Sridhar[1] and Dr. B. Raveendra Babu[2]
Department of Computer Applications, R.V.R. & J.C College of Engineering, Guntur, India
[1]mandapati_s@yahoo.com, [2]rbhogapathi@yahoo.com

***Abstract*– Recent advances in hardware technology have increased storage and recording capability with regard to personal data on individuals. This has created fears that such data could be misused. To alleviate such concerns, data was anonymized and many techniques were recently proposed on performing data mining tasks in ways which ensured privacy. Anonymization techniques were drawn from a variety of related topics like data mining, cryptography and information hiding. Data is anoymized through methods like randomization, k-anonymous, l-diversity. Several privacy preserving data mining algorithms are available in literature. This paper investigates the classification accuracy of the data with and without k-anonymization to compare the efficiency of privacy preserving mining. The classification accuracy is evaluated using k nearest neighbor, J48 and Bagging.**

***Index Terms*– Privacy Preserving Data Mining, k-Anonymous, k Nearest Nneighbor, J48 and Bagging**

I. INTRODUCTION

TO maintain the privacy of the data during data mining process, the data is anonymized to preserve privacy. Several techniques of privacy-preserving data mining are available in literature [1], [2], [3]. Most privacy computations methods use some form of data transformation to ensure privacy preservation. Usually such methods reduce representation granularity to reduce privacy. This reduction leads to some loss in data management or mining algorithms' effectiveness, a trade-off between information loss and privacy. Some commonly used anonymization techniques are given below:

The randomization method: The randomization method was generally used with regard to distorting data by probability distribution for surveys which include an evasive answer bias due to privacy concerns [4, 5]. Randomization is a technique for privacy-preserving data mining where noise is added to data to cover record's attribute values. [6, 7]. The added noise is large that individual record values are not recovered. Hence, techniques derive aggregate distributions from perturbed records. Data mining techniques are developed later to work with such aggregate distributions.

The k-anonymity model and l-diversity: The k-anonymity model was developed due to the possibility of indirect record identification from public databases where combinations of record attributes are used to identify individual records. In k-anonymity method, data representation granularity is reduced with techniques like generalization and suppression. This granularity is reduced such that any record maps onto at least k other data records. In generalization, attribute values are generalized to reduce representation granularity. For example, a date of birth can be generalized to a range like year of birth, in order to reduce identification risks. In the suppression method, the attribute value is removed. It is clear that these methods lower identification risks through the use of public records while reducing the transformed data applications accuracy.

The l-diversity model handles weaknesses in the k-anonymity model. As protecting identities to the k-individual level is not the same as protecting corresponding sensitive values, when there is homogeneity of sensitive values inside a group. To do this, a concept of sensitive values, intra-group diversity is promoted within the anonymization scheme [8]. Hence, the l-diversity technique was proposed to maintain minimum group size of k, and also to focus on maintenance of sensitive attributes diversity. The t-closeness model is an improvement on the l-diversity concept. In [9], a t-closeness model was proposed using property that distance between distributions of sensitive attribute inside an anonymized group could not be different from global distribution by more than a threshold. The Earth Mover distance metric is used to quantify distance between two distributions. Further, t-closeness approach is more effective than other privacy preserving data mining methods for numeric attributes.

Distributed privacy preservation: In some cases, individual entities will derive aggregate results from data sets partitioned across entities. The goal in distributed methods for privacy-preserving data mining is allowing computation of aggregate statistics over a data set without compromising individual data set privacy, within different participants. Thus, participants may collaborate to obtain aggregate results, without trusting each other with regard to distribution of own data sets. Such partitioning can be horizontal where records are distributed across multiple entities or vertical where attributes are distributed across multiple entities. While individual entities may not share entire data sets, they could consent to limited sharing through various protocols. These methods ensure

privacy for every individual entity, while getting aggregate results over entire data.

Downgrading Application Effectiveness: In some cases, though data may be un-available, application output like association rule mining, classification or query processing may led to privacy violations. This has resulted in research regarding downgrading application effectiveness by either data/application modifications. Examples of such techniques are association rule hiding [10], classifier downgrading [11], and query auditing [12].

Two approaches are used for association rule hiding: Distortion: In distortion, [13], entry for a given transaction is changed to another value. Blocking: In blocking [14], entry remains the same but is incomplete. Classifier downgrading modifies data so that classification accuracy is reduced, while retaining data utility for other applications. Query auditing, denies one or more queries from a sequence. The to-be-denied queries are chosen so that underlying data sensitivity is preserved.

Data anonymization techniques were under investigation recently, for structured data, including tabular, graph and item set data. They published detailed information, permitting ad hoc queries and analyses, while at the same time guaranteeing sensitive information privacy against a variety of attacks.

This paper investigates the classification accuracy of the data with and without k-anonymization to compare the efficiency of privacy preserving mining to extract the desired patterns from the dataset. The classification accuracy is evaluated using k nearest neighbor, J48 and Bagging. The proposed method is evaluated using Adult dataset. The paper is organized as follows: Section II reviews some the related works in literature, section III details the methodology, section IV gives the results and section V concludes the paper.

II. LITERATURE REVIEW

Aggarwal, et al., [15] reviewed the state-of-the-art methods for privacy. In the study the following was discussed:

- Methods for randomization, k-anonymization, and distributed privacy-preserving data mining.
- Cases where the output of data mining applications requires to be sanitized for privacy-preservation purposes.
- Methods for distributed privacy-preserving mining for handling horizontally and vertically partitioned data.
- Issues in downgrading the effectiveness of data mining and data management applications
- Limitations of the problem of privacy.

Bayardo, et al., [16] proposed an optimization algorithm for k-anonymization. Optimized k-anonymity records are NP-hard due to which significant computational challenges are faced. The proposed method explores the space of possible anonymization and develops strategies to reduce computation. The census data was used as dataset for evaluation of the proposed algorithm. Experiments show that the proposed method finds optimal k-anonymizations using a wide range of k. The effects of different coding approaches and quality of anonymization and performance are studied using the proposed method.

LeFevre, et al., [17] proposed a suite of anonymization algorithms for producing an anonymous view based on a target class of workloads. It consists of several data mining tasks and selection predicates. The proposed suite of algorithms preserved the utility of the data while creating anonymous data snapshot. The following expressive workload characteristics were incorporated:

- Classification & Regression, for predicting categorical and numeric attributes.
- Multiple Target Models, to predict multiple different attributes.
- Selection & Projection, to warrant that only a subset of the data remains useful for a particular task.

The results show that the proposed method produces high-quality data for a variety of workloads.

Inan, et al., [18] proposed a new approach for building classifiers using anonymized data by modeling anonymized data as uncertain data. In the proposed method, probability distribution over the data is not assumed. Instead, it is proposed to collect necessary statistics during anonymization and release them with anonymized data. It was demonstrated that releasing statistics is not violative of anonymity. The aim was to compare accuracy of: (1) various classification models constructed over anonymized data sets and transformed through various heuristics (2) approach of modeling anonymized data as uncertain data using expected distance functions. The experiments investigate relationship between accuracy and anonymization parameters: the anonymity requirement, *k*; the number of quasi-identifiers; and, especially for distributed data mining the data holders number. Run experiments were conducted on Support Vector Machine (SVM) classification methods and Instance Based learning (IB) methods. For implementation and experimentation Java source code of the *libSVM* SVM library, and IB1 instance-based classifier from *Weka* was used. For experiments, *DataFly* was selected as is a widely known solution. Experiments which spanned many alternatives in both local and distributed data mining settings revealed that our method performed better than heuristic approaches used to handle anonymized data.

Fung., et al., [19] presented a Top-Down Specialization (TDS) approach, a practical algorithm to determine a general data version that covers sensitive information and is for modeling classification. Data generalization is implemented by specializing/detailing information level in a top-down manner till violation of a minimum privacy requirement. TDS generalizes by specializing it iteratively from a most general state. At every step, a general (i.e., parent) value is specialized into a specific (i.e., child) value for a categorical attribute. Otherwise an interval is divided into two sub-intervals for continuous attribute. Repetition of this process occurs till specialization leads to violation of anonymity requirement. This top-down specialization is efficient to handle categorical and continuous attributes. The proposed approach exploits the idea that data has redundant structures for classification. While generalization could terminate some structures, others come to help. Experimental results show that classification

quality can be preserved for highly restrictive privacy requirements. TDS gels well with large data sets and complex anonymity requirements. This work has applicability in public and private sectors where information is shared for mutual benefits and productivity.

III. METHODOLGY

A. Dataset

The 'Adult' dataset used for evaluation is obtained from UCI Machine Learning Repository [20]. The dataset contains 48842 instances, with both categorical and integer attributes from the 1994 Census information. The Adult dataset contains about 32,000 rows with 4 numerical columns. The columns and their ranges are: age {17 – 90}, fnlwgt {10000 – 1500000}, hrsweek {1 – 100} and edunum {1 – 16}. The age column and the native country are anonymized using the principles of k-anonymization. Table I and II show the original data and the modified attribute data. The dataset is classified using 10 fold cross validation the original and the K anonymized dataset.

B. k Nearest Neighbor

In the *k* Nearest Neighbor classification, the algorithm finds a set of *k* objects in the training set that are the closest to the input and classifies the input based on the majority of the class in that group [21]. The main elements required for this approach are: a set of labeled objects, distance metrics and number of *k*. Following is the *k*-nearest neighbor classification algorithm:

Input: D with *k* sets of training objects x, y and test object x', y'

Process: compute distance $d\ x', x$ between test object and every object.
Select $D_z \subseteq D$, set of k closest training objects to test object.

Output: $y' = \arg\max_{v} \sum_{x_i, y_i \in D_z} I\ v = y_i$
Where v is the class label, I is a indicator function.

C. J48

J48 algorithm is a version of the C4.5 decision tree learner [22]. Decision tree models are formed by J48 implementation. The decision trees are built using greedy technique and analyzing the training data. The nodes in the decision tree evaluate the significance of the features. Classification of input is performed by following a path from root to leaves in the decision tree, resulting in a decision of the inputs class. The decision trees are built in a top-down fashion by selecting the most suitable attribute every time.

The classification power of each feature is computed by an information-theoretic measure. On choosing a feature, subsets of the training data is formed based on different values of the selected feature. Process is iterated for every subset until majority of the instances belong to the same class. Decision tree induction generally learns a high accuracy set of rules which ensure accuracy but produce excessive rules. Thus, the trees are pruned to generalize a decision tree to gain balance of flexibility and accuracy. J48 utilize two pruning methods; subtree replacement and subtree raising. In subtree replacement, nodes are replaced with a leaf, working backwards towards the root. In subtree raising, a node is moved upwards towards the root of the tree.

Table I: The Original Attributes Of Adult Dataset

age	native-country	Class
39	United-States	<=50K
50	United-States	<=50K
38	United-States	<=50K
53	United-States	<=50K
28	Cuba	<=50K
37	United-States	<=50K
49	Jamaica	<=50K
52	United-States	>50K
31	United-States	>50K
42	United-States	>50K

Table II: The K-Anonymous Dataset

age	native-country	Class
adult	United-States	<=50K
middle aged	United-States	<=50K
adult	United-States	<=50K
middle aged	United-States	<=50K
adult	US-oth	<=50K
adult	United-States	<=50K
middle aged	African	<=50K
middle aged	United-States	>50K
adult	United-States	>50K
adult	United-States	>50K

D. Bagging

Bagging improves classification [23] by combining models learned from different subsets of a dataset. Overfitting and

variance is considerably reduced on application of bagging. The instability in the classifier is used to perturb the training set producing different classifiers using the same learning algorithm. For a training set A of size t, n number of new training sets A_i ($t' < t$) is generated. The subsets are generated by sampling of the instances from A uniformly and by replacement. Some instances are repeated in each A_i due to sampling with replacement and are called bootstrap samples. The n number of models are fitted using all the n number of subsets (bootstrap samples). The output is obtained by averaging the above result.

IV. RESULT AND DISCUSSION

The classification accuracy obtained from k nearest neighbor, J48 and Bagging is tabulated in Table III and is shown in Fig. 1. It is seen that the classification accuracy does not change considerably and are within manageable limits for all the classifiers. The variation in the classification accuracy is not more than 0.4% for any classifier.

Table III: Classification Accuracy

Technique used	Classification Accuracy
kNN without anonymization	79.32%
J48 without anonymization	85.32%
Bagging without anonymization	85.01%
kNN with anonymization	79.56%
J48 with anonymization	85.13%
Bagging with anonymization	84.68%

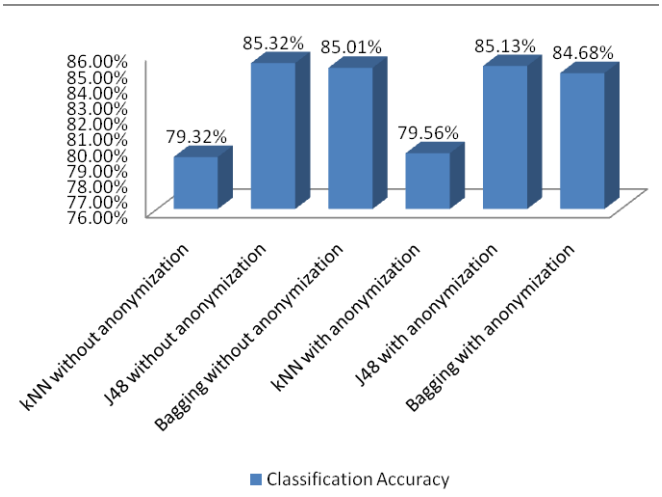

Fig. 1: Classification accuracy for various techniques

The root mean square error is shown in Fig. 2. The precision, recall and fMeasure are tabulated in Table IV and shown in Fig. 3 and Fig. 4.

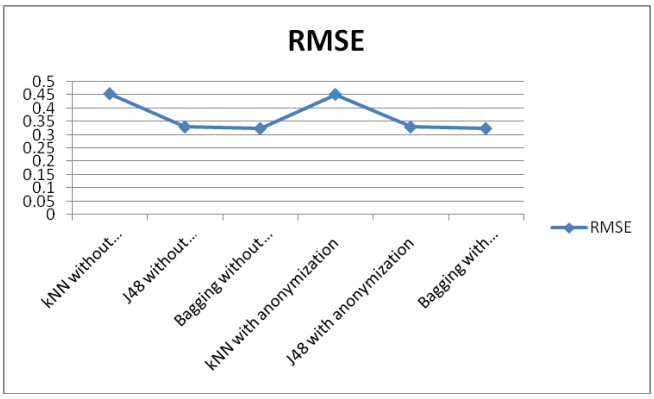

Fig. 2: The root mean square error

Table IV: Precision, Recall and fMeasure

Technique used	Precision	Recall	fMeasure
kNN without anonymization	0.791	0.793	0.792
J48 without anonymization	0.848	0.853	0.849
Bagging without anonymization	0.844	0.85	0.845
kNN with anonymization	0.796	0.796	0.796
J48 with anonymization	0.845	0.851	0.845
Bagging with anonymization	0.84	0.847	0.841

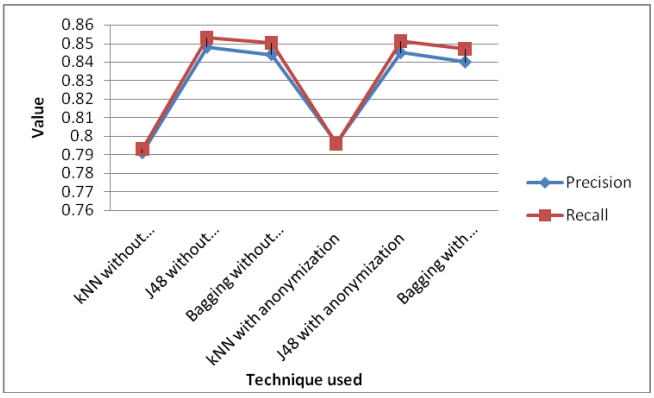

Fig. 3: Precision and Recall for different techniques

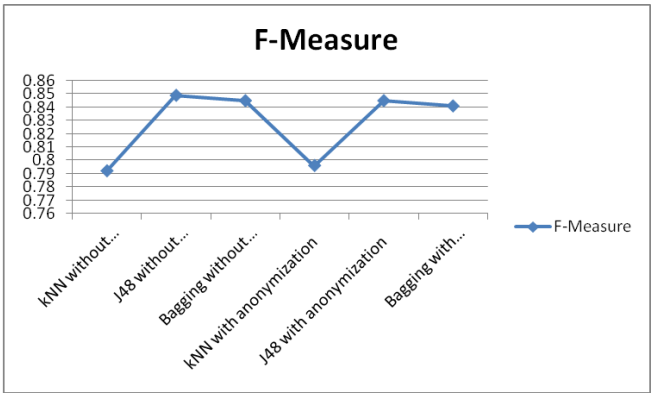

Fig. 4: fMeasure for different techniques

V. CONCLUSION

In this paper, it was proposed to compare the classification accuracy of k nearest neighbor, J48 and Bagging with anonymized and without anonymized dataset. The adult datset was used for evaluating the classification accuracy and the dataset was K-anonymized. The experimental results demonstrate that the classification accuracy of the classifiers is not diminished due to anonymization of the data.

REFERENCES

[1] Verykios V. S., Elmagarmid A., Bertino E., Saygin Y.,, Dasseni E.: Association Rule Hiding. *IEEE Transactions on Knowledge and Data Engineering*, 16(4), 2004.

[2] Sweeney L.: Privacy-Preserving Bio-terrorism Surveillance. *AAAI Spring Symposium, AI Technologies for Homeland Security*, 2005.

[3] Newton E., Sweeney L., Malin B.: Preserving Privacy by De-identifying Facial Images. *IEEE Transactions on Knowledge and Data Engineering, IEEE TKDE*, February 2005.

[4] Liew C. K., Choi U. J., Liew C. J. A data distortion by probability distribution. ACM TODS, 10(3):395-411, 1985.

[5] Warner S. L. Randomized Response: A survey technique for eliminating evasive answer bias. Journal of American Statistical Association, 60(309):63–69, March 1965.

[6] Agrawal R., Srikant R. Privacy-Preserving Data Mining. Proceedings of the ACM SIGMOD Conference, 2000.

[7] Agrawal D. Aggarwal C. C. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. ACM PODS Conference, 2002.

[8] Machanavajjhala A., Gehrke J., Kifer D., and Venkitasubramaniam M.: l-Diversity: Privacy Beyond k-Anonymity. ICDE, 2006.

[9] Li N., Li T., Venkatasubramanian S: t-Closeness: Privacy beyond k-anonymity and l-diversity. ICDE Conference, 2007.

[10] Verykios V. S., Elmagarmid A., Bertino E., Saygin Y.,, Dasseni E.: Association Rule Hiding. IEEE Transactions on Knowledge and Data Engineering, 16(4), 2004.

[11] Moskowitz I., Chang L.: A decision theoretic system for information downgrading. Joint Conference on Information Sciences, 2000.

[12] Adam N., Wortmann J. C.: Security-Control Methods for Statistical Databases: A Comparison Study. ACM Computing Surveys, 21(4), 1989.

[13] Oliveira S. R. M., Zaiane O., Saygin Y.: Secure Association-Rule Sharing. PAKDD Conference, 2004.

[14] Saygin Y., Verykios V., Clifton C.: Using Unknowns to prevent discovery of Association Rules, ACM SIGMOD Record, 30(4), 2001.

[15] C. C. Aggarwal and P. S. Yu (eds.), Privacy-Preserving Data Mining: Models and Algorithms (Springer, New York, 2008).

[16] Bayardo R. J., Agrawal R.: Data Privacy through Optimal k-Anonymization. *Proceedings of the ICDE Conference*, pp. 217–228, 2005

[17] LeFevre K., DeWitt D., Ramakrishnan R.: Workload Aware Anonymization. KDD Conference, 2006.

[18] A. Inan, M. Kantarcioglu, and E. Bertino. Using anonymized data for classification. In ICDE, 2009.

[19] Fung B., Wang K., Yu P.: Top-Down Specialization for Information and Privacy Preservation. ICDE Conference, 2005.

[20] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[21] Fix E, Hodges JL, Jr (1951) Discriminatory analysis, nonparametric discrimination. USAF School of Aviation Medicine, Randolph Field, Tex., Project 21-49-004, Rept. 4, Contract AF41(128)-31, February 1951

[22] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. Morgan Kaufmann, San Fransico, CA.

[23] L. Breiman. Bagging predictors. Machine Learning, 24(2):123–140, 1996.