ISSN: 2229-6956(ONLINE)
DOI: 10.21917/ijsc.2011.0031



# GAIN RATIO BASED FEATURE SELECTION METHOD FOR PRIVACY PRESERVATION

**R. Praveena Priyadarsini[1], M.L.Valarmathi[2] and S. Sivakumari[3]**
[1, 3]*Department of Computer Science and Engineering, Avinashilingam Deemed University for Women, Tamil Nadu, India*
E-mail: [1]praveena.priya04@gmail.com, [3]hod_cse_au@yahoo.co.in
[2]*Department of Computer Science and Engineering, Government College of Technology, Tamil Nadu, India*
E-mail: ml_valarmathi@rediffmail.com

***Abstract:***
***Privacy-preservation is a step in data mining that tries to safeguard sensitive information from unsanctioned disclosure and hence protecting individual data records and their privacy. There are various privacy preservation techniques like k-anonymity, l-diversity and t-closeness and data perturbation. In this paper k-anonymity privacy protection technique is applied to high dimensional datasets like adult and census. since, both the data sets are high dimensional, feature subset selection method like Gain Ratio is applied and the attributes of the datasets are ranked and low ranking attributes are filtered to form new reduced data subsets. K-anonymization privacy preservation technique is then applied on reduced datasets. The accuracy of the privacy preserved reduced datasets and the original datasets are compared for their accuracy on the two functionalities of data mining namely classification and clustering using naïve Bayesian and k-means algorithm respectively. Experimental results show that classification and clustering accuracy are comparatively the same for reduced k-anonym zed datasets and the original data sets.***

***Keywords:***
***Privacy Preservation, Data Mining, K-Anonymity, Feature Subset Selection, Gain Ratio***

## 1. INTRODUCTION

Data mining is the extraction of hidden information from large database. A key problem that arises in any mass collection of data is that of confincidality of the data.Privacy-preserving data mining (PPDM) is the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure. Privacy preservation is primarily concerned with protecting against disclosure of individual data records. Most traditional data mining techniques analyze and model the data set statistically, in aggregation, while privacy preservation is primarily concerned with protecting against disclosure individual data records There are many basic privacy preservation techniques like suppression, summarization, cryptography and randomization[1]. The k-anonymity is a model for protecting privacy which was proposed by Latanya Sweeney et.al [2].In the k-anonymity approach generalization techniques are applied in order to mask the exact values of attributes .For example, a quantitative attribute such as the age may only be specified to a range. This is referred to as attribute generalization. By defining a high enough level of generalization on each attribute it is possible to guarantee k-anonymity. [3]

In this paper we propose an improved method for achieving privacy preservation using feature ranking method where the utility of the datasets are not affected. Feature ranking method Gain Ratio is used to rank the attributes of high dimensional datasets like Adult and Census. The low ranking attributes are filtered to form new reduced data subsets. K-anonymization privacy preservation technique is then applied on reduced datasets and the original datasets. The privacy preservation of these anonymoized reduced datasets are tested using two functionalities of data mining namely classification and clustering using naïve Bayesian and k-means algorithm respectively.

The rest of this paper is organized as follows Section2 presents related work and section 3 gives the proposed methodology section 4 gives the data set description and preprocessing done section 5 presents the dimensionality reduction techniques used section 6 presents the data mining algorithms used to test the privacy preservation. Section 8 discusses the results and comparisons. Section 9 presents the conclusions and future enhancements.

## 2. RELATED WORK

Alexandre et al in his work has described Privacy-preserving data mining (PPDM) as the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure. Privacy preservation is primarily concerned with protecting against disclosure of individual data records. Most traditional data mining techniques analyze and model the data set statistically, in aggregation, while privacy preservation is primarily concerned with protecting against disclosure individual data records [1]. Aggarwal C. C et al presents that Real data sets are usually extremely high dimensional, and this makes the process of privacy preservation extremely difficult both from a computational and effectiveness point of view. The curse of dimensionality becomes especially important when adversaries may have considerable background information, as a result of which the boundary between pseudo-identifiers and sensitive attributes may become blurred. In recent years, it has been observed that many privacy-preservation methods such as k-anonymity and randomization are not very effective in the high dimensional case [9]. A. Friedman et al indicates that the k-Anonymitty model makes two major assumptions: 1.The database owner is able to separate the columns of the table into a set of quest-identifiers, which are attributes that may appear in external tables the database owner does not control, and a set of private columns, the values of which need to be protected. We prefer to term these two sets as public attributes and private attributes, respectively. 2. The attacker has full knowledge of the public attribute values of individuals, and no knowledge of their private data. The attacker only performs linking attacks. [10] Sweeney et al has provided a formal foundation for anonymity problem against linking and for the application of generalization and supervision towards its solution. They have also define quasi identifiers as attributes that can be exploited for linking.and k-

anonymization as characterizing the degree of protection of data with respect to linking error. [4].

Lior Rokach et al has proposed data mining privacy by decomposition (DMPD) and employs a genetic algorithm for searching for optimal feature set partitioning. The search is guided by k-anonymity level constraint and classification accuracy. Both are incorporated into the fitness function. They also show that the new approach significantly outperforms existing suppression-based and generalization-based methods that require manually defined generalization trees. In addition, DMPD can assist the data owner in choosing the appropriate anonymity level.[11].Zhiqiang Yang et al presents Naive Bayes classifiers that have been used in many practical applications. They greatly simplify the learning task by assuming that attributes are independent given the class. Although independence of attributes is an unrealistic assumption, naive Bayes classifiers often compete well with more sophisticated models, even if there is modest correlation between attributes. NaiveBayes classifiers have significant advantages in terms of simplicity, learning speed, classification speed, and storage space. They have been used, for example, in text classification and medical diagnosis[2]. Fukunaga, K et al has used K-means clustering is one of the most widely used techniques for statistical data analysis. Researchers use cluster analysis to partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers[3]

## 3. PROPOSED METHODOLOGY

The attributes of the high dimentional datasets adult, census are ranked using gain ratio attribute ranking method. The low ranking attributes are filtered to form new reduced data subsets. K-anonymization privacy preservation techniques are applied on both original and reduced datasets. The accuracy of the privacy preserved datasets and the original datasets are compared on the two functionalities of data mining namely classification and clustering using naïve Bayesian and k-means algorithm respectively. The classification and clustering accuracy for the privacy preserved reduced datasets and the original data sets compared. Fig.1 show the methodology used in this work.

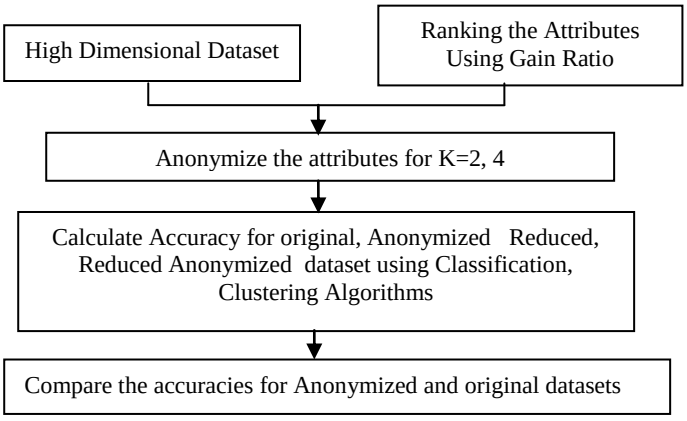


Fig.1. Flowchart of Methodology

## 4. DATASET DESCRIPTIONS

The dataset used in this work are Adult dataset and Census dataset available on UCI Machine Learning Repository [16].Adult predicts whether the income exceeds $50K/yr. It has a size of 3,755KB. Census dataset contains weighted census data extracted from the 1994 and 1995 population surveys conducted by the US Census Bureau. It has a size of 50,800KB. Table.1 shows the dataset information for both the datasets.

Table.1. Dataset Information

| Dataset | No. of records | No. of attributes |
|---|---|---|
| ADULT | 32561 | 15 |
| CENSUS | 99763 | 42 |

### 4.1 PREPROCESSING OF ADULT AND CENSUS DATASET

In order to improve the quality of the data, accuracy and efficiency of the mining process the adult dataset undergoes a preprocessing step. In adult attributes like fnlwgt, capital gain, capital loss, hours per week are removed since they are not considered as relevant attribute for privacy preservation in data mining. Thus reducing the number of attributes to 10. The adult test dataset is then resampled by 5% and all the missing values are removed. In census dataset the less sensitive attributes like wage per hour, enroll in edu inst last wk, capital gain, capital loss, dividends from stock, live in house one year ago, migration prev res in sunbelt, fill inc questionnaire for veterans admin, veterans benefits, instance weight are removed since they are not considered as relevant attribute for privacy preservation in data mining. So the number of attributes is reduced to 32. The census test dataset is resampled by 10% and all the missing values are removed. Table.2 shows the information about the preprocessed datasets.

Table.2. Preprocessed Dataset Information

| Dataset | No. of records | No. of attributes |
|---|---|---|
| ADULT | 519 | 10 |
| CENSUS | 470 | 32 |

## 5. DATA MINING ALGORITHMS USED

### 5.1 NAIVE BAYESIAN ALGORITHM

This classifier simply computes the conditional probabilities of the different classes given the values of attributes and then selects the class with the highest conditional probability. If an instance is described with n attributes ai (i=1…n),then the class that instance is classified to a class v from set of possible classes V according to a Maximum a Posteriori (MAP) Naive Bayes classifier is,

$$v = \arg\max P(v_j)^n \pi_{i-1} p(a_i \mid v_j) \quad (1)$$

Eq.(1) gives conditional probability obtained from the estimates of the probability mass function using training data. The class probability is not used in these experiments, since no prior phoneme distribution information is available, and thus we are implementing Maximum Likelihood (ML) classification.

This Bayes classifier minimizes the probability of classification error under the assumption that the sequence of points is independent. [9]

### 5.2 K-MEANS ALGORITHM

K-means is one of the simplest unsupervised learning algorithms and a non-hierarchical approach that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. A very common measure is the sum of distances or sum of squared Euclidean distances from the mean of each cluster. K-Means training starts with a single cluster with its center as the mean of the data. This cluster is split into two and the means of the new clusters are iteratively trained. These two clusters are again split and the process continues until the specified number of clusters is obtained.[7]

## 6. FEATURE SUBSET SELECTION

Feature subset selection is of great importance in the field of data mining. The high dimension data makes testing and training of general data mining tasks difficult. Feature selection is the problem of choosing a small subset of features that ideally is necessary and sufficient to describe the target concept [13]. The terms features, variables, A goal of feature selection is to avoid selecting too many or too few features than is necessary. If too few features are selected, there is a good chance that the information content in this set of features is low. On the other hand, if too many (irrelevant) features are selected, the effects due to noise present in (most real-world) data may over shadow the information present. Hence, this is a tradeoff which must be addressed by any feature selection method [14]. In this paper filter feature subset approach namely Gain ratio has been used to rank the attributes of the datasets used.

### 6.1 GAIN RATIO

Gain ratio (GR) is a modification of the information gain that reduces its bias. Gain ratio takes number and size of branches into account when choosing an attribute. It corrects the information gain by taking the intrinsic information of a split into account. Intrinsic information is entropy of distribution of instances into branches (i.e. how much info do we need to tell which branch an instance belongs to). Value of attribute decreases as intrinsic information gets larger. [7].

$$\text{Gain ratio}\left(\text{Attribute}\right) = \frac{\text{Gain}\left(\text{Attribute}\right)}{\text{Intrinsic\_info}\left(\text{Attribute}\right)} \quad (2)$$

### 6.2 FEATURE SELECTION USING GAIN RATIO

The adult dataset is of high dimension. Ranking method is used to select a subset of 7 attribute from the original dataset of 10 attributes. Among those attribute we have considered age, work class, occupation, relationship, sex, native country and income for gain ratio. Among these attributes "age", "occupation", "sex" are considered as quasi attributes.

Attributes that are less specific are removed to form the census dataset of 32 attributes. Age, class of worker, detailed industry recode, detailed occupation recode, education, major industry code, major occupation code, sex, state of previous residence, num persons worked for employer, family members under 18,country of birth father, country of birth self, own business or self employed, weeks worked in year, instance weight are the selected attributes for gain ratio method. Among these attributes "age", "class of worker", "detailed industry recode", "detailed occupation recode", "education" are considered as quasi attributes.

## 7. K-ANONYMITY

Let $T(A_1,\ldots A_N)$ be a table and $Q1_T$ be the quasi identifiers associated with it. T is said to satisfy k-anonymity if for each quasi-identifier QI $\in$ $Q1_T$ each sequence of values in T[QI] appears at least with k occurrences in T[QI]. Each release of data must be such that every combination of values of quasi identifiers can be matched to at least k individuals. [5]

### 7.1 K-ANONYMIZED UNREDUCED DATASET

The quasi identifiers considered for the k-anonymity in census dataset are age, class of worker, detailed industry recode, detailed occupation recode, education. The quasi identifiers selected for adult datasets are age, marital status and relationship. Then these datasets are anonymized for the k values 2, 3 and 4.

### 7.2 K-ANONYM IZED REDUCED DATASET

The reduced dataset of both adult and census dataset obtained using ranking method applied on original dataset is anonymized for various values of k, k = 2, 3, 4, thus we get k non-distinguishable records.

## 8. RESULTS AND COMPARISON

The experiments were conducted using open source software WEKA [17] the results and are recorded as follows

### 8.1 CLASSIFICATION OF ANONIMIZED UNREDUCED DATASETS USING NAÏVE BAYES ALGORITHM

The preprocessed adult and census datasets are taken and the quasi identifiers are selected in order to perform k-anonymizaton. The accuracy obtained after classification using naïve Bayesian algorithm is tabulated and shown in Table.3.

Table.3. Classification result for anonymized unreduced datasets

| DATA SETS | ACCURACY% | | |
|---|---|---|---|
| | K=2 | K=3 | K=4 |
| ADULT | 82.2736 | 82.2736 | 82.2736 |
| CENSUS | 82.766 | 82.766 | 82.766 |

## 8.2 CLUSTERING OF ANONIMIZED UNREDUCED DATASETS USING K-MEANS ALGORITHM

The preprocessed adult and census datasets are taken and the quasi identifiers are selected in order to perform k-anonymizaton.. The k-anonymization is performed for the k value k = 2,3and 4 for both the datasets. The accuracy obtained after clustering using k-means is tabulated in Table.4.

Table.4. Clustering result for anonymized original datasets

| ANONYMIZED ORIGINAL DATASET | ACCURACY% | |
|---|---|---|
| | ADULT | CENSUS |
| K=2 | 57.22 | 50.43 |
| K=3 | 58.39 | 58.09 |
| K=4 | 57.42 | 56.383 |

## 8.3 CLASSIFICATION OF REDUCED DATASETS USING NAÏVE BAYES ALGORITHM

In reduced subset of both adult and census dataset considering quasi identifiers like age, marital status and relationship for adult and quasi identifiers like age , class of worker , detailed occupation recode, detailed industry recode and education for census are anonymized for values k=2,3 and 4. The accuracy obtained after classification using naïve bayes algorithm is shown in Table.5.

Table.5. Classification result for reduced datasets

| GAIN RATIO REDUCED DATASET | CLASSIFICATION ACCURACY% | |
|---|---|---|
| | ADULT | CENSUS |
| K=2 | 78.0347 | 79.4239 |
| K=3 | 78.8054 | 81.1728 |
| K=4 | 81.5029 | 81.5844 |

## 8.4 CLUSTERING ON REDUCED DATASETS USING K-MEANS:

The reduced subset for both adult and census dataset are taken and the reduced datasets are anonymized for values k=2, 3 and 4. The accuracy obtained after clustering using k-means is shown in Table.6.

Table.6. Clustering result for census dataset and Adult Dataset

| GAIN RATIO REDUCED DATASET | CLUSTERING ACCURACY% | |
|---|---|---|
| | ADULT | CENSUS |
| K=2 | 57.23 | 51.96 |
| K=3 | 52.03 | 57.21 |
| K=4 | 53.77 | 51.34 |

## 8.5 COMPARISONS OF CLASSIFICATION AND CLUSTERING RESULTS

The classification accuracies are compared for the original datasets, anonymized datasets and reduced datasets. The classification and clustering accuracies are compared for both the datasets on original and reduced privacy preserved version. The Fig.2 shows the comparison for Clustering and classification accuracy for original, reduced and k-anonymized census dataset

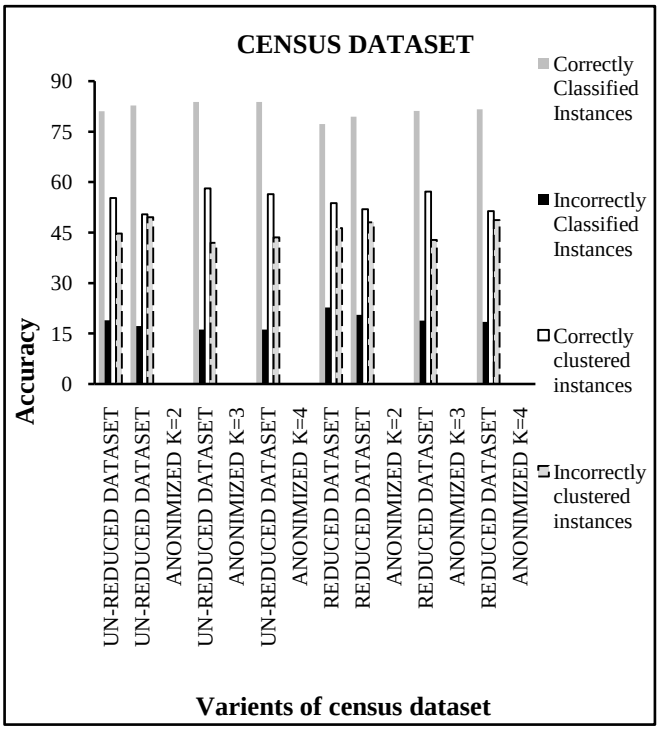


Fig.2. Comparisons of Clustering and classification accuracy for original, reduced and k-anonymized census dataset

The results show that the classification accuracy of original anonymized census dataset varies from is about 82% for K value2, 3, 4.For the reduced anonymized census dataset K=2, 3, 4 the accuracies varies from 79-81%. Thus, it's incurred that accuracies remain almost the same for both original and reduced, privacy preserved census datasets, for classification using naive bayes algorithm. This shows that the utility of the dataset is unaffected by the attribute reduction and privacy preservation.

Clustering with original census dataset the clustering accuracies vary from 50-56% while for the reduced anonymized census dataset the clustering accuracies varies from 51-57%.This shows that even though the attributes are reduced and the dataset is annonimized for privacy preservation the clustering accuracies does not vary much from the unreduced and non annonimized datasets.

The comparison of classification accuracies and clustering accuracies of adult dataset for original privacy preserved and reduced privacy preserved adult dataset is shown in Fig.3.

From the Fig.3 it can be incurred that accuracies for classification using Naive Bayes algorithm on original adult dataset the accuracies is about 82% for all k = 2,3,4 anonymized values. For the reduced anonymized dataset the accuracies vary from 78-81%.

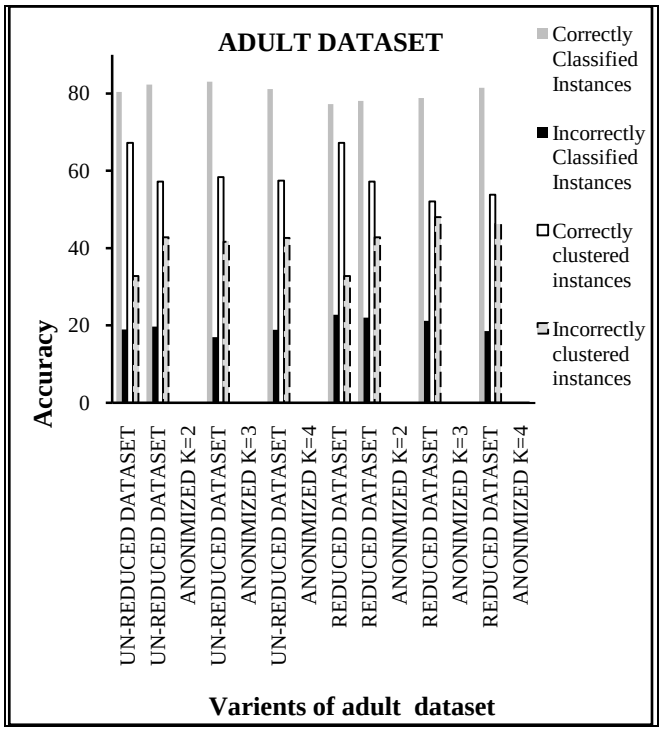


Fig.3. Comparison of Clustering and classification accuracy for original, reduced and k-anonymized adult dataset

There is only a minor variation in accuracy percentage for both original and reduced adult dataset for classification. Thus the reduction of attributes and annonymization does not affect the prediction accuracies of the dataset using naive bayes algorithm.

For clustering with K-Means algorithm the accuracies for the unreduced dataset is about 57-58% for anonymization using all the three values of K. The reduced dataset anonymized with K value 2, 3, 4 the clustering accuracy vary about 53-57%. Thus the clustering accuracy is almost the same for both original anonymized and reduced anonymized adult dataset.

## 9. CONCLUSION AND FUTURE ENHANCEMENTS

The goal of this work is to provide privacy for the datasets while reducing the dimensionality using gain ratio method. The adult dataset and census dataset available on UCI machine learning repository were used for experiments. The k-anonymized original and reduced datasets are compared for accuracy on both data mining task classification and clustering. The obtained results that showed the accuracy level remained the same for k-anonymized original datasets and reduced datasets for the both data mining functionalities .This shows that the utility of both the datasets are not affected by both dimensionality reduction and privacy preservation using K-annonimization technique. As future enhancement different classification and clustering algorithms may be used .Also other data mining task like associations, regression and prediction may be to study the effect of k-anonymity on the datasets.

## REFERENCES

[1] Alexandre Evfimievski and Tyrone Grandison, "*Privacy-Preserving Data Mining*" Encyclopedia of Database Technologies and Applications, 2007

[2] Zhiqiang Yang and Sheng Zhong, "Privacy-Preserving Classification of Customer Data without Loss of Accuracy", *Proceedings of the 5th SIAM International Conference on Data Mining*, Newport Beach, CA, pp. 21-23, 2005.

[3] Fukunaga. K "*Introduction to Statistical Pattern Recognition*", Academic Press, London, 1990.

[4] L. Sweeney "k-Anonymity: A Model for Protecting Privacy", *International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems*, Vol. 10, No.5, pp. 557-570, 2002.

[5] P. Samarati and L. Sweeney "K-Anonymity and its Enforcement through Generalization and Suppression", *Proceedings of the IEEE symposium on Research in Security and Privacy*, 1998.

[6] C.C. Aggarwal, "*On K-Anonymity and the Curse of Dimensionality", Proceedings of International Conference on Very Large Data Bases (VLDB)*, pp. 901-909, 2005.

[7] J. Han and M. Kamber, "*Data Mining Concepts and Techniques*", Morgan Kaufmann, 2001.

[8] S. Kotsiantis, D. Kanellopoulos and P. Pintelas, "Data Preprocessing for Supervised Learning", *International Journal of Computer Science*, Vol. 1, No.2, ISSN 1306-4428, pp. 111-117, 2006.

[9] J. Ye, R. J. Povinelli, and M. T. Johnson, "Phoneme classification using naive bayes classifier in reconstructed phase space", *Proc. of IEEE Signal Processing Society 10th Digital Signal Processing Workshop*, pp. 37- 40, 2002.

[10] Aggarwal C. C "On k-anonymity and the curse of dimensionality", *Proceedings of 31st International Conference on Very Large Data Bases*, 2004.

[11] Friedman, R. Wolff and A. Schuster, "Providing k-Anonymity in Data Mining", *The International Journal on Very Large Data Bases*, Vol. 17, No. 4, pp. 789-804, 2008.

[12] Nissim Matatov, Lior Rokach, and Oded Maimon "Privacy-preserving data mining: A feature set partitioning approach", *Information Sciences*, Vol. 180, No. 14, pp.2696–2720, 2010.

[13] Kira, K. and Rendell, L.A.,"A practical approach to feature selection", *Proceedings of the Ninth International Workshop on Machine Learning*, pp. 249–256, 1992.

[14] S. Piramuthu "Evaluating feature selection methods for learning in data mining applications", *European Journal of Operational Research*, Vol. 156, pp.483–494, 2004.

[15] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "k-Anonymity", *In Security in Decentralized Data Management, Springer*, 2006.

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

[17] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten, "The WEKA Data Mining Software: An Update", *ACM SIGKDD Explorations Newsletter*, Vol.11, No.1, pp. 10-18, 2009.