Post Anonymization Techniques in Privacy Preserved Data Mining

A. K. Ilavarasi
Assistant Professor
Department of Computer Science and Engineering,
Sona College of Technolgy,
Salem, India.

D. Jeniffa
PG Scholar
Department of Computer Science and Engineering,
Sona College of Technolgy,
Salem, India.

Dr. B. Sathiyabhama
Head Of the Department
Department of Computer Science and Engineering,
Sona College of Technolgy,
Salem, India.

ABSTRACT
Privacy preserving data mining deals with the effectiveness of preserving privacy and utility of the data. Privacy becomes a key concern when the medical data is published for research purposes. Anonymization techniques can be used to transform the dataset into less specific values before publishing to overcome the security breaches. Privacy preservation may reduce the utility value of data. Classification helps to improve the utility of the anonymized data. We propose a model in which a multi-decision tree classifier is built on the anonymized dataset to improve the utility. Multi-decision tree classifier is constituted by Improved ID3 based ADABOOST classifier. The proposed approach is different as the decision tree built is multi-decision tree and as it is constructed on the anonymized dataset. It is proved to be better than the pure decision tree classifier as the multi-decision tree classifier has accuracy better than and training duration shorter than the normal ID3 based ADABOOST classifier.

Keywords
Data Privacy, Anonymization, Classification.

1. INTRODUCTION
Number of benefits can be obtained by storing and sharing information. Each record in a table corresponds to one individual. The attribute categories are: (i) Attributes that clearly identify individuals known as explicit identifiers, e.g., Name, Address. (ii) Attributes whose values when taken together can potentially identify an individual known as quasi-identifiers, e.g., Zip-code, Birth-date. (iii) Attributes that are considered as sensitive, e.g., Disease.

The micro data (e.g., medical data) need to be published for research and other purposes. There is a need to prevent the sensitive information of individuals from being disclosed, when releasing the privacy sensitive micro data. Two types of information disclosure identified in [1]: When an individual is linked to a record in the released table, Identity disclosure occurs and When new information about some individuals is revealed, attribute disclosure occurs. The objective while releasing a table is to limit the disclosure risk, while maximizing the benefit. To achieve this, the data need to be anonymized before release.

This paper integrates anonymization techniques with classification techniques of data mining. The idea is to build a multi-decision tree classifier which is constituted by Improved ID3 based ADABOOST classifier on the dataset to which k-anonymity with t-closeness technique is applied. Performance is evaluated to prove that the multi-decision tree classifier built on the anonymized data is better than pure decision tree classifier. Hence, t-closeness principle is applied to preserve privacy and the multi-decision tree classifiers are constructed on the anonymized data thereby ensuring utility preservation.

2. RELATED WORK
The combination of generalization and suppression reduces the granularity representation of data by making it less specific, is the privacy definition for anonymity given by Latanya Sweeney in [2], [6]. Linking attack is possible by which an adversary can link a public database (e.g., voter's list) with the released data to get the personal information about an individual [10]. Two attacks were identified [7]: the Homogeneity attack and the Background Knowledge attack. K-anonymity protects against identity disclosure but it does not provide sufficient protection against the attribute disclosure [7].

K-anonymity is extended to L-Diversity in which for each sensitive attribute **e**ach equivalence class must have minimum l well-represented values and l different sensitive values [7]. This is insufficient to prevent attribute disclosure [8]. Two attacks were identified [8]: (i) Skewness Attack, satisfying l-diversity does not prevent attribute disclosure when the overall distribution is skewed and (ii) Similarity Attack, important information can be learned by the adversary when the sensitive attribute values are distinct but semantically similar in an equivalence class.

T-Closeness requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold *t*) [8]. This limits the amount of individual-specific information an observer can learn. The Earth Mover Distance metric [3] is used to measure the distance between the two distributions.

Naive Bayes and decision tree classifiers were built over partially specified data by Zhang et al. in [6]. The classifiers are built on a mixture of partially and fully specified data.

Two classification-aware data anonymization methods which combine global attribute generalization and local value suppression was given in [13]. Instead of by privacy requirement, the attribute generalization is determined by the data distribution. Based on the normalized mutual information, the generalization levels are optimized for preserving classification capability. Privacy requirement k (IACk) or data distributional constraints (IACc) determine the value suppression. IACk anonymizes data that supports better classification models than the data anonymized by a

benchmark utility-aware data anonymization method, and is faster.

The problem of classification over anonymized data is addressed in [11]. It models the generalized attributes of anonymized data as uncertain information. The generalized value of an anonymized record r is released along with the statistics collected from records in the same equivalence class. This extra information, released with the anonymized data, supports to compute the expected values of important functions for data analysis such as dot product and square distance. SVM and k-nearest neighbor classification were used in [11]. Many classification algorithms can be extended to handle anonymized data.

3. PROPOSED WORK

The availability of medical data helps to prevent medical errors and enhance patient care. The objective is to provide data privacy for sharing medical data and to utilize the shared data for data mining tasks. Data privacy is provided by applying anonymization techniques and classifiers are built on those anonymized data for utility purpose.

The proposed approach builds multi-decision tree classifiers on the data anonymized using k-anonymity with t-closeness. The multi-decision tree classifier is constituted by an Improved ID3 based ADABOOST classifier and is proved to be better than pure decision tree classifier.

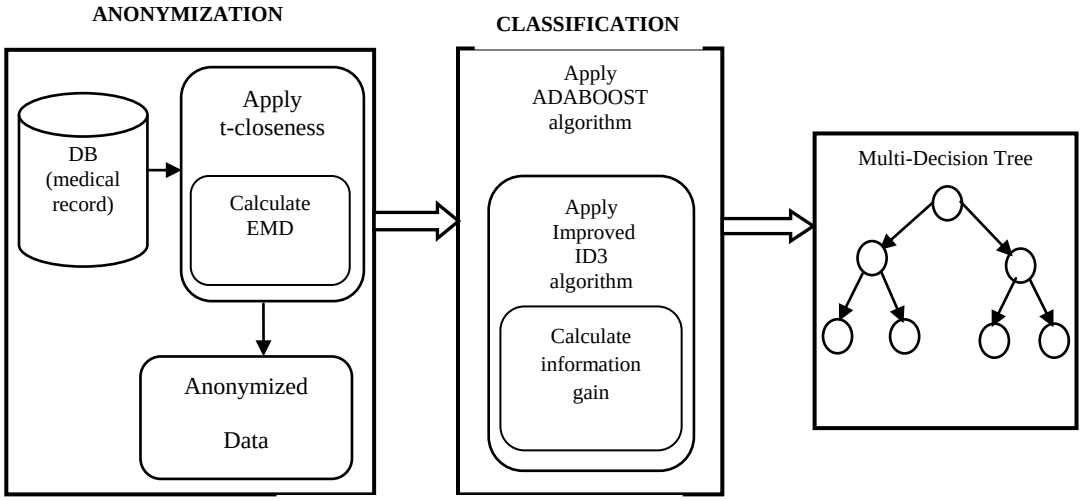


Fig 1: System Architecture

3.1 K-anonymity

K-Anonymity is a combination of Generalization, in which the quasi identifier attributes are generalized and Suppression, which removes the explicit identifiers [5]. It requires each individual in the database to be indistinguishable from k−1 others. The anonymity is specified by k to indicate that certain characteristics and combinations of tuples must be found in the data to match at least k individuals.

K-Anonymity Table T is k anonymous with respect to attributes X1,…,Xd if every unique tuple (x1,…,xd) in the projection of T on X1,…,Xd occurs at least k times.

3.2 T-Closeness

If the distance between the distribution of a sensitive attribute in an equivalence class and in the whole table is no more than a threshold t, the class is said to have t-closeness. In a t-closeness table, all equivalence classes have t-closeness [8]. The distance between the two distributions P and Q can be measured using the Earth Mover Distance (EMD) metric [3].

Given two distributions P = (p_1, p_2, ..., p_m) and Q = (q_1, q_2,..., q_m), the problem is to measure the distance between P and Q. Two well-known distance measures used in [11] are,

The variational distance defined as:

$$\mathbf{D}[\mathbf{P},\mathbf{Q}] = \sum_{i=1}^{m} \frac{1}{2} |p_i - q_i| \quad (1)$$

And the Kullback-Leibler (KL) distance defined as:

$$\mathbf{D}[\mathbf{P},\mathbf{Q}] = \sum_{i=1}^{m} p_i \log \frac{p_i}{q_i}$$
$$= H(P) - H(P,Q) \quad (2)$$

where H(P) is the entropy of P and H(P,Q) is the cross-entropy of P and Q.

3.2.1 EMD for Numerical Attributes

The values of numerical attribute will be ordered. Let {v_1, v_2...v_m} be the attribute domain, where v_i is the i^{th} smallest value.

Ordered Distance. The distance between values v_i and v_j is based on the number of values between them in total order,

$$Ordered\ dist\ (v_i, v_j) = \frac{|i-j|}{m-1} \quad (3)$$

If the value of r_i is formally assumed as $r_i = p_i - q_i$, (i=1,2..m), the distance between P and Q can be calculated as:

$$D[P,Q] = \frac{1}{m-1}(|r_1| + |r_1 + r_2| + \cdots + |r_1 + r_2 + \cdots r_{m-1}|)$$
$$= \frac{1}{m-1} \sum_{i=1}^{i=m} \sum_{j=1}^{j=i} r_j \quad (4)$$

3.2.2 EMD for Categorical Attributes

For the categorical attributes a total order often does not exist. Two distance measures are considered for the categorical attributes in [8] are,

Equal Distance. As the ground distance between two values of a categorical attribute is defined to be 1, for each point p_i-q_i>0, the extra need to be moved to some other point. Thus,

$$\mathbf{D}[\mathbf{P}, \mathbf{Q}] = \sum_{i=1}^{m} \frac{1}{2} |p_i - q_i| \quad (5)$$

Hierarchical Distance. The distance between values v_1,v_2 of a categorical attribute is based on the minimum level to which v_1 and v_2 are generalized to the same value according to the domain hierarchy.

Given a domain hierarchy and distributions P and Q, the extra of a leaf node that corresponds to element i can be defined as:

$$Extra(N) = \begin{cases} p_i - q_i & if\ N\ is\ a\ leaf \\ \sum_{C \in Child\,(N)} extra(C) & otherwise \end{cases} \quad (6)$$

where *child*(*N*) is the set of all leaf nodes below node *N*. Here, the sum of *extra* values for nodes at the same level is 0. Two other functions defined for *internal nodes*:

$$\text{Pos\_extra(N)} = \sum_{\text{C} \in \text{Child (N)} \wedge \text{extra (C)} > 0} |\text{extra(C)}| \quad (7)$$

$$\text{Neg\_extra(N)} = \sum_{c \in Child(N) \wedge extra(C) < 0} |extra(C)| \quad (8)$$

Here the cost of movings between *N*'s children branches is,

$$cost(N) = \frac{height(N)}{H} \min(pos\_extra(N), neg\_extra(N)) \quad (9)$$

The earth mover's distance can now be written as:

$$\text{D[P, Q]} = \sum_{\text{N}} \text{cost(N)} \quad (10)$$

where N is a non-leaf node.

Table 1: Original table

	Name	*Zip code*	*Age*	*Disease*
1	Chirag	441105	29	Gastric Ulcer
2	Sanchit	442308	22	Gastritis
3	Chaudhari	441107	27	Stomach Cancer
4	Mukherjee	441211	43	Gastritis
5	Faarooq	441207	52	Flu
6	Abulas	441200	47	Bronchitis
7	Bhupendra	442308	30	Bronchitis
8	Nikola	441103	36	Pneumonia
9	Rasal	442308	32	Stomach Cancer

In the above Table1, 'Name' is the explicit identifier, 'Disease' is the sensitive attribute and the quasi identifier attributes are Zip-code and Age. When the Table1 is anonymized, the explicit attribute is suppressed and the values of quasi-identifiers are generalized.

Table2: 0.278-closeness w.r.t. Disease

	Zip code	*Age*	*Disease*
1	4411**	≤40	Gastric Ulcer
2	4411**	≤40	Stomach Cancer
3	4411**	≤40	Pneumonia
4	4412**	≥40	AIDS
5	4412**	≥40	Flu
6	4412**	≥40	Heart Disease
7	4423**	≤40	AIDS
8	4423**	≤40	Heart Disease
9	4423**	≤40	Stomach Cancer

The original Table1 is anonymized using k-anonymity and to the anonymized data t-closeness is applied where the value of t is 0.278 to obtain Table 2.

3.3 Multi-Decision Tree Classifiers

Advanced ID3 algorithm in [12] is used to build multi-decision tree classifier. An ADABOOST classifier based on multi-decision tree is used to improve the classification accuracy of using pure decision tree classifier.

ID3 algorithm constructs decision tree using the entropy measure that generates the information gain. Shannon Entropy is given as,

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \quad (11)$$

ADABOOST is an iterative algorithm which gets different base classifiers and put them together to form a strong classifier [9]. Let C_1,C_2,…C_t represents base classifiers built in t number of iterations. Each base classifier will form appropriate weigh. Using weight voting machine, summarize the classifier results of t base classifiers.

Decision Tree ID3 is used as the algorithm of base classifier in ADABOOST classifier. The decision tree ID3 algorithm need to be improved to maximize the classification accuracy and make the training time as short as possible of ADABOOST classifier.

Through decision tree ID3 algorithm we obtain,

$$Gain(A) = Info(D) - Info_A(D)$$

$$= -\sum_{i=1}^{v} p_i \log_2(p_i) - \sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j) \quad (12)$$

We get,

$$Gain(A)' = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j) = -Info_A(D) \quad (13)$$

We consider F and T to be the amount of input by current node and $F + T = |D|$.

$$Info(D_j) = -\left(\frac{f_j}{f_j + t_j} \log_2 \frac{f_j}{f_j + t_j} + \frac{t_j}{f_j + t_j} \log_2 \frac{t_j}{f_j + t_j}\right) \quad (14)$$

$$Gain(A)' = -\sum_{j=1}^{v} \frac{f_j + t_j}{F + T} Info(D_j) \quad (15)$$

Substitute (14) in (15) we get Gain(A)' and after elimination of the the factor in Gain(A)', we get,

$$Gain(A)'' = \sum_{j=1}^{v} f_j \ln \frac{f_j}{f_j + t_j} + t_j \ln \frac{t_j}{f_j + t_j} \quad (16)$$

This can be rewritten as,

$$Gain(A)'' = -\sum_{j=1}^{v} \frac{2 f_j t_j}{f_j + t_j} \quad (17)$$

To overcome the bias in ID3 algorithm, W proportional to the number of attribute value is introduced,

$$\text{Gain (A)} = \text{Info (D)} - W\ \text{Info}_A\ (D) \quad (18)$$

The Gain(A)′′′ after the introduction of W is obtained as,

$$Gain(A)''' = -W \sum_{j=1}^{v} \frac{2 f_j t_j}{f_j + t_j} \quad (19)$$

After introducing W, the information gain value of the node of attribute A is dependent on the number of property value, thus overcoming the bias of original decision tree algorithm ID3.Multi-decision tree finds its major application where redundant data need to be pruned.

3.4 Building Multi-Decision Tree Classifiers on Anonymized Data

Classifiers are built on the data anonymized using anonymization techniques for utility purpose. The proposed work builds a multi-decision tree classifier from the training data in which the values of records have been anonymized. Here, a multi decision tree is built on the data anonymized using t-closeness.

The medical dataset is anonymized to preserve the sensitive information of an individual from adversaries. Let T denote the medical database and $\hat{f}$ be the anonymization function. Performing anonymization on T results in a less specific anonymized database T', which is obtained by applying the anonymization function, $\hat{f}$ on T.

$$T' = \hat{f}(T) \quad (20)$$

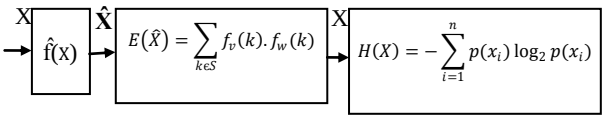
The anonymized database, T′ obtained as result of the anonymization module is the input for constructing multi-decision tree classifier. Use the formulas in the section 3.3 to compute the information gain for each node n in the anonymized database T'.

ADABOOST classifier uses the classifiers obtained by t number iterations using improved ID3 algorithm as the base classifiers. Thus improved ID3 based ADABOOST classifier is constructed. Each leaf node, n should have the most common classification value among the records as its label.

The output multi-decision tree obtained is stated as 'MDT($\hat{f}$(T))' or MDT(T'), where MDT is the multi decision tree function on the anonymized database T'. The output is the multi decision tree constructed from the anonymized data. It provides better efficiency than the pure decision tree.

3.5 Mathematical model

Generalization is applied to quasi-identifiers in order to achieve privacy protection. By carefully releasing statistics of the attributes, they can be reconstructed by finding the expected values of the kernel function. The reconstructed data is precise enough to perform classification resulting in a decision tree classifier.



where,

$\hat{X}$ is the anonymized quasi-identifier

E ($\hat{X}$) is the expected value of kernel function (anonymized)

$f_v(k)$ and $f_w(k)$ denote the probability mass functions of generalized values

H(X) is the entropy of X and $p(x_i)$ is a probability a tuple in X belongs to class C_i

3.4.1 Modeling Anonymized Data as Uncertain Data

Anonymized data can be used for classification purposes by modeling it as uncertain data. Calculate the kernel functions for anonymized data instances and compute the "expected" values of the kernel functions with the imprecise attributes [11].

Given generalized instances gen ($\mathbf{x}_i$) and gen ($\mathbf{x}_j$), our goal is to calculate E(K(gen($\mathbf{x}_i$), gen($\mathbf{x}_j$))) for anonymized data instances for common kernel functions. Define X_d = gen($\mathbf{x}_i$)^{T}gen($\mathbf{x}_j$) (i.e., X_d is the random variable that represents the dot product of two anonymized instances) and X_e = $\|\text{gen}(\mathbf{x}_i) - \text{gen}(\mathbf{x}_j)\|^2$ is the random variable that represents the square distance between two anonymized instances.

According to the Taylor theorem [4], for any differentiable function $g(X)$, and for random variable X with $E(X) = \mu_X$ and $Var(X) = \sigma^2_X$, $g(X)$ can be approximated around μ_x as:

$$g(X) \sim g(\mu_X) + (X - \mu_X) g'(\mu_X) + \frac{1}{2}(X - \mu_X)^2 g''(\mu_X) \quad (21)$$

The first order approximation of the expected value of $g(X)$ is obtained as,

$$E(g(X)) \sim E(g(\mu_X) + (X - \mu_X)\, g'(\mu_X)) = g(\mu_X) \quad (22)$$

Since each of the above kernel functions is of the form $g(X_d)$ or $g(X_e)$, a first order approximation of the expected value of the kernel function results can be obtained as $g(E(X_d))$ or $g(E(X_e))$. So, to compute E(K(gen($\mathbf{x}_i$), gen($\mathbf{x}_j$))), it is necessary to compute $E(X_d)$ = E(gen($\mathbf{x}_i$)^{T}gen($\mathbf{x}_j$)) and $E(X_e)$ = $E(\|\text{gen}(\mathbf{x}_i) - \text{gen}(\mathbf{x}_j)\|^2)$.

Let $E(X^t_d)$ and $E(X^t_e)$ denote expected dot product and expected square Euclidean distance on the t^{th} attribute respectively. Then, $E(X_d)$ and $E(X_e)$ can be formulated as a summation of attribute-wise expected values:

$$E(X_d) = \sum_{t=1}^{m} E(X^t_d) \quad (23)$$

$$E(X_e) = \sum_{t=1}^{m} E(X^t_e) \quad (24)$$

3.4.1.1 Numerical Quasi-identifiers

For numerical quasi-identifiers, first calculate E(gen($\mathbf{x}_i$)[t]^{T}gen($\mathbf{x}_j$)[t]) for i≠ j as follows:

$$E(X^t_d) = E(gen(x_{ij})[t].gen(x_j)[t]) = E(gen(x_{ij})[t]).E(gen(x_j)[t]) \quad (25)$$

Also, compute $E(\|gen(\mathbf{x}_i)[t] - gen(\mathbf{x}_j)[t]\|^2)$ (i.e., $E(X_e)$) as follows:

$$E(X^t_e) = E((gen(x_i)[t])^2) - 2E(gen(x_i)[t]).E(gen(x_j)[t]) + E\left((gen(x_j)[t])^2\right) \quad (26)$$

To support classification, $E((gen(\mathbf{x}_i)[t])^2)$ and $E((gen(\mathbf{x}_i)[t]))$ need to be evaluated for numerical quasi-identifiers.

3.4.1.2 Categorical Quasi-identifiers

Dot product of two categorical values represents the probability that the values are equal. Therefore, assuming that gen($\mathbf{x}_i$)[t] and gen($\mathbf{x}_j$)[t] are from the same domain S, E(gen($\mathbf{x}_i$)[t]^{T}gen($\mathbf{x}_j$)[t]) can be calculated for i≠ j as follows:

Let the probability mass functions of generalized values V=gen($\mathbf{x}_i$)[t] and W = gen($\mathbf{x}_j$)[t] denoted as $f_V(v)$ and $f_W(w)$.

$$E(X_d^t) = P_r(V - W)$$
$$= \sum_{k \in S} P_\gamma(V = k).P_\gamma(W = k)$$
$$= \sum_{k \in S} f_v(k).f_w(k) \quad (27)$$

The square distance $E(X^t_e)$ can be derived from $E(X^t_e)$ as follows:

$$E(X_e^t) = P_r(V \neq W) = 1 - E(X_d^t) \quad (28)$$

Finally, define the probability mass function of an original value $v \in S$, so that $E(X_e)$ and $E(X_d)$ are defined on arbitrary pairs of generalized and original values:

$$f_v(k) = \begin{cases} 1 & if\ v = k \\ 0 & otherwise \end{cases} \quad (29)$$

3.4.2 Releasing QI-statistics with Anonymized Data

Statistical characteristics of each attribute are used to compute the expected value of the required kernel function results for classifier construction.

Table 3: Sample anonymized DataSet with Statistics of QIs

R″	A_1	S_1
r_1″	≤40	μ_X=35.6,
		σ^2_X=0.22
r_2″	≥40	μ_X=35.6,
		σ^2_X=0.22
r_3″	≤40	μ_X=27.6,
		σ^2_X=20.22

Table 3. illustrates the proposed idea. Instead of releasing R', the data holder adds one new attribute S_i per quasi-identifier A_i. If A_i is categorical, then the corresponding S_i stores a probability mass function obtained from the equivalence class of each record. On the other hand, if A_i is numerical, then S_i contains the mean and variance of A_i within the equivalence class of each record.

4. RESULTS AND DISCUSSION

The experimental results for the quality of data and efficiency of the algorithm are discussed using various metrics.

The data quality of the data anonymized using k-anonymity without t-closeness and k-anonymity with t-closeness are shown based on the average size of the equivalence classes generated by the anonymization algorithm.

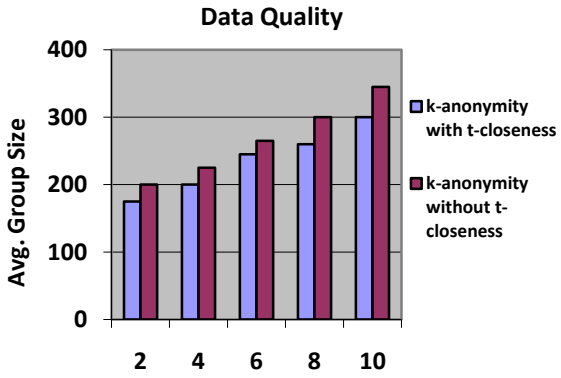


Fig 2: Quality of data anonymized using k-anonymity with and without t-closeness

Figure2 shows that the data quality of k-anonymous tables without t-closeness is slightly better than k-anonymous tables with t-closeness. This is because t-closeness requirement provides extra protection to sensitive values as the cost is decreased.

Through experiments, the accuracy rates of ten base classifiers are found to range from 79% to 91%. After the integration of ADABOOST algorithm to the base classifiers above, the accuracy rate of classification was 96.15%. The classifier rate has been greatly improved after improving the decision tree algorithm ID3 and integrating it with the ADABOOST algorithm.

Table 4: Training Duration

Algorithm	Duration (seconds)
Decision Tree ID3 based ADABOOST Classifier	49
Improved Decision Tree ID3 based ADABOOST Classifier	35

From the above table, we can easily find that the time of the ADABOOST classifier algorithm which is based on the improved decision tree algorithm ID3 has been decreased a lot compared to the ADABOOST classifier algorithm which is based on the original decision tree algorithm ID3.

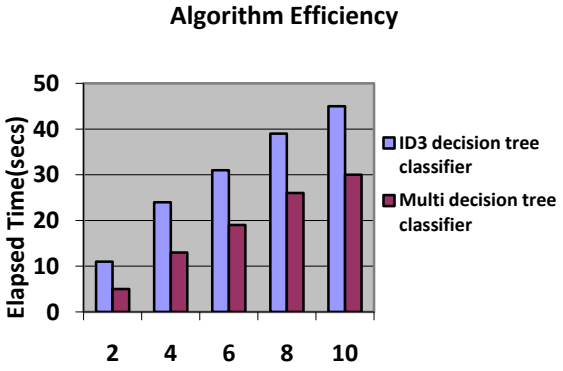


Fig 3: Algorithm Efficiency of original and multi-decision tree classifiers

We compare the efficiency of the ID3 decision tree classifier and multi-decision tree classifier. The above shows the running times with fixed k values. It is observed that the improved ID3 based ADABOOST (Multi) decision tree classifiers run faster than the ID3 decision tree classifier.

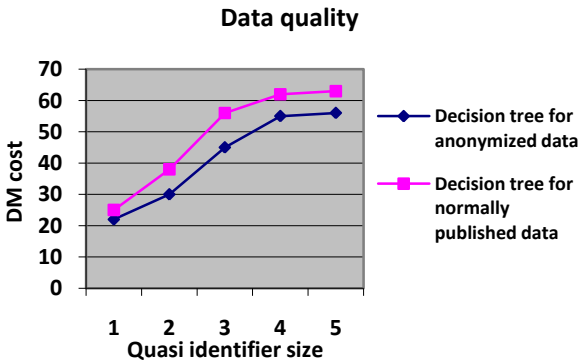


Fig 4. Data Quality of decision tree for normally published and anonymized data

The above graph measures the data quality using the discernibility metric cost for the classification based on the quasi-identifier size. The data quality of the decision tree classifier built on anonymized data is observed to be nearly equal to the decision tree classifier built on normally published data.

5. CONCLUSION

In this paper, a framework in which the privacy is preserved while the data utility is maintained by building a multi-decision tree classifier is proposed. K-anonymity with t-closeness is used as the anonymization technique. For utilizing the anonymized data, Multi-decision tree classifier which is a Improved Decision Tree ID3 based ADABOOST classifier is built and proved to be better than the original decision tree classifiers.

6. REFERENCES

[1] Duncan, G.T., and Lambert, D. 1986. Disclosure-limited data dissemination. In: Journal of the American Statistical *Association*, pp. 10-28.

[2] Sweeney, L. 1997. Guaranteeing anonymity when sharing medical data, the Datafly system. In: Proceedings of the American Medical Informatics Association, Annual Symposium.

[3] Rubner, Y., Tomasi, C., and Guibas, L.J. 2000. The earth mover's distance as a metric for image retrieval. In: *International Journal of Computer Vision.*

[4] Bickel, P., and Doksum, K. 2000. Mathematical Statistics-Basic Ideas and Selected Topics, second edition, Prentice Hall.

[5] Sweeney, L. 2002. Achieving k-anonymity privacy protection using generalization and suppression. In: International Journal of Uncertainty, Fuzziness, and knowledge-based Systems, vol. 10, pp. 571-588, IEEE.

[6] Zhang, J., Kang, D.K., Silvescu, A., and Honavar, V. 2006. Learning accurate and concise naive bayes classifiers from attribute value taxonomies and data. In: Knowl. Inf. Syst., vol. 9, pp.157-179.

[7] Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. 2006. l-diversity: Privacy beyond k-anonymity. In: ICDE '06, Atlanta, GA,USA.

[8] Ninghui, Li., Tiancheng, Li., and Venkatasubramanian, S. 2007. t-closeness: Privacy beyond k-anonymity and l-diversity. In: ICDE '07, pp.106-115, Istanbul, Turkey.

[9] Hatami, N., and Ebrahimpour, R. 2007. Combining multiple classifiers:diversity with boosting and combining by stacking [J]. International Journal of Computer Science and Network Security, pp.127-131.

[10] Yan ZHU., and Lin PENG. 2007. Study on K-anonymity Models of Sharing Medical Information. Beijing, China, IEEE.

[11] Ali Inan., Murat Kantarcioglu., and Elisa Bertino. 2009. Using Anonymized Data for Classification. In: International Conference on Data Engineering , IEEE.

[12] Duan Xiaochen., and Hong Xue. 2011. Multi-Decision-Tree Classifier in Master Data Management System. IEEE.

[13] Jiuyong Li, Jixue Liu, Muzammil Baig and Raymond Chi-Wing Wong. 2011. Information based data anonymization for classification utility. In: Data & Knowledge Engineering 70 (2011) 1030–1045, Elsevier publication.