International Journal of Soft Computing 10 (6): 468-475, 2015
ISSN: 1816-9503

Privacy Preserving Data Mining Using Sliced Data for Classification Technique

[1]V. Shyamala Susan and [2]T. Christopher
[1]Department of Computer Science, Government Arts College, Udumalpet, India
[2]Department of Computer Science, Government Arts College, Coimbatore, India

Abstract: Privacy preservation in data publishing is the major topic of research in the field of data security. Data publication in privacy preservation provides methodologies for publishing useful information; simultaneously the privacy of the sensitive data has to be preserved. There has been little research addressing how to effectively use the preserved data for data mining in general and for distributed data mining in particular. Here, we propose a new approach for building classifiers using Radial Basis Function (RBF) and Multiple Linear Regression (MLR) by employing sliced data as uncertain data. Use of probability distribution employed in the slicing approach was replaced by classification techniques to enable modeling for sliced data. InRBF, the sliced data is sent into the input layer, the activation function is executed by the hidden lauer and output layer produces classified data. In the same manner, MLR calculates approximate value of one or more sliced data responses on the basis of certain predictors. Results from the experiments show that these techniques show better performance in comparision with other classification approaches.

Key words: Privacy preservation data publishing, Radial Basis Function (RBF), Multiple Linear Regression (MLR), classification technique, India

INTRODUCTION

Sensitive information related to individuals, such as medical data are collected today, stored and processed in a large variety of application domains. Such data is typically used to provide better quality services to individualsenhance patient care. At the same time, privacy of the individuals to whom the data is related should be assured as well.

In order to preserve privacy, the attributes likeid, namet needs to be removed prior to data publishing. They do not guarantee privacy because when these attributes are combined together with attributes like gender, birthday, postcode, and others the sensitive information may get exposed. These attributes are referred to quasi-identifier attributes (Aggarwal and Yu, 2008; Samarati, 2001). The K-Anonymity Model (El-Emam and Dankar, 2008) is a useful approach that enables data protectionespecially with regards to individual identification.

There have been various algorithms that have been suggested (Lin and Wei, 2008; Ye *et al*., 2008; Yu *et al*., 2009) which employ cell level Generalization. Generalisation and Bucketisationare the two famous techniques of privacy preservation. These techniques eliminate the identifiers and cluster the tuples. Generalisation focuses on quasi identifiers whereas Bucketisation focuses on splitting sensitive attributes from quasi identifiers. Slicing is the technique that deals with high dimensional data and also preserves privacy.

Mining sliced data becomes more challenging and this research addresses issues related to classification over sliced dat. The proposed method is an innovative approach that enables modeling which further helps model generalized attributes of the anonymize data and then categorizes them as uncertain information. The research here focuses on RBF and MLR classifiers which necessitate computing distance functions and dot products over sliced data to classify those samples. The approach used here is one wherein each sliced value of an anonymized record r is accompanied by statistics collected from records in the same equivalence class as r. The additional information, disclosed along with anonymized data, further support exact expected values that have been computed with respect to data analysis and important functions like dot product and square distance.

Related work: Astonishingly, though little research has been done which enables investigation of data mining algorithms performance on anonymized data. An exception that stands out here is the "top-down specialization" or TDS method (Fung *et al*., 2005) where data anonymizationis done on the basis of class

Corresponding Author: V. Shyamala Susan, Department of Computer Science, Government Arts College, Tuticorin, Udumalpet, India

conditional entropy measure. Fundamentally, TDS may be considered as an anonymization technique that is tweaked around a bit in order to build accurate decision trees through the usage ofanonymized data whilst the main purpose is to build an extensive classifiers range using existing and already anonymized data. LeFevre *et al.* (2006) suggests various algorithms that enable generation of anonymous data set which effectively may be utilized over pre-defined workloads. The workload features that have been considered here are primarily algorithms that incorporate selection, projection, classification and regression.

Bayardo and Agrawal (2005) in their study present an optimization algorithm for k-anonymization. Existing technique therein investigates the space available for anonymization and there after develops strategies that enable reducing of computation. Census data employed for evaluating and other experimental results indicated that the technique shown here achieves optimal k-anonymizations by employing wide range of k. Investigation was also carried out on outcomes of various coding approaches as well as for performance and assessment of quality of anonymization. The experiments conducted on the real census data showed that suggested algorithm has the ability to exactly locate optimal k-anonymizations as per two representative cost measures and a wide range of k.

Data sharing inherently poses various threats that possibly lead to individual identification. Hence, it is become imperative to research the issues concerning privacy preserving data publication (Ram *et al*., 2011). Major aim of the issue is to preserve individual privacy while disclosing any information deemed useful. Any organization as such can employ and successfully pursue its policy regarding privacy preservation. However, though when two different companies exchange or share information regarding individuals that are common to both then if respective privacy policies vary, generally a breach of privacy is unless common policy regarding this has been commonly agreed upon. A solution presented in lieu of a scenario like this is one on the basis of the k-anonymity and cut-tree method for 2-party data. The study here presents an uncomplicated solution facilitating integration of nparty data while deploying dynamic programming on various subsets. Solution presented here in on the basis of thresholds for privacy and informativeness based on k-anonymity.

Kumari *et al.* (2008) present a holistic approach aimed at maximizing privacy without any loss of information as well as one with minimum overheads. Studies indicate that both l-diversity and t-closeness methods heightened computational effort to infeasible level and simultaneously increased privacy. Several techniques explain maximum loss of information loss while attaining privacy. Technique presented here addresses the above mentioned issue through the use of fuzzy set approach again which is a total paradigm shift and also presents itself as being innovative offering a new perspective to privacy problem while publishing. The technique here permits personalization while preserving privacy and also isvaluable for numerical and categorical attributes and required only duringformation of tuples.

Shang *et al.* (2010) in their study present an innovative method for select distribution of content that are encoded as documents while maintaining user privacy on the basis of an effective and innovatively advanced group key management scheme. Approach presented here is one drawn on the basis of access control policies which specifically entail and list those userswho have access to the documents or sub-documents. Based on this a broadcast document has been separatedas multiple subdocuments. Every subdocument has been encrypted using a separate key. In conformance with the modern attribute-based access control, specific policies have been formulated against access to user identity attributes. However, this approach here enables privacy preservation so that users can specifically access certain documents or subdocument, on the basis of policies without disclosing information regarding identity attributes to respective publishers. As per the approach, document publisher has no access or pertinent information regarding identity values of users, also is unaware of the policy conditions that have been verified by users additionally enables prevention of inferences being drawn regarding identity attributes values. The key management scheme presented here is based on broadcasting approach proves to be efficient here because decryption keys are not required to be sent to users along with respective encrypted documents.

Slicing: In this study, first give an example to illustrate slicing. Then formalize slicing, compare it with generalization and bucketization and discuss privacy threats that slicing can address. Table 1 shows an example microdata table and its anonymized versions using various anonymization techniques. The original table is shown in Table 1. The three QI attributes are (Age, Sex, Zipcode) and the sensitive attribute SA is Disease. A generalized table that satisfies 4-anonymity is shown in Table 2, a bucketized table that satisfies 2-diversity is shown in Table 3, a generalized table where each attribute value is replaced with the multiset of values in the bucket is shown in Table 4 and two sliced table are shown in Table 5 and 6. Slicing first partitions attributes into

Table 1: The original table

Diseases	Age	Sex	Zipcode
Dyspepsia	M	22	47906
Flu	F	22	47906
Flu	F	33	47905
Bronchitis	F	52	47905
Flu	M	54	47302
Dyspepsia	M	60	47302
Dyspepsia	M	60	47304
Gastritis	F	64	47304

Table 2: The generalized table

Diseases	Sex	Age	Zipcode
Dyspepsia	*	[20-52]	4790*
Flu	*	[20-52]	4790*
Dyspepsia	*	[20-52]	4790*
Dyspepsia	*	[20-52]	4790*
Gastritis	*	[54-64]	4730*
Flu	*	[54-64]	4730*
Flu	*	[54-64]	4730*
Bronchitis	*	[54-64]	4730*

Table 3: The bucketized table

Diseases	Sex	Age	Zipcode
Flu	M	22	47906
Dyspepsia	F	22	47906
Bronchitis flu	F	33	47905
	F	52	47905
Gastritis	M	54	47302
Flu	M	60	47302
Dyspepsia	M	60	47304
Dyspepsia	F	64	47304

Table 4: Multiset-based generalization

Diseases	Sex	Age	Zipcode
Dysp	M:1, F:3	22:2.33:1.52:1	47905:2, 47906:2
Flu	M:1, F:3	22:2.33:1.52:1	47905:2, 47906:2
Flu	M:1, F:3	22:2.33:1.52:1	47905:2, 47906:2
Bron	M:1, F:3	22:2.33:1.52:1	47905:2, 47906:2
Flu	M:3, F:1	54:1.60:2.64:1	47302:2, 47304:2
Dysp	M:3, F:1	54:1.60:2.64:1	47302:2, 47304:2
Dysp	M:3, F:1	54:1.60:2.64:1	47302:2 ,47304:2
Gast	M:3, F:1	54:1.60:2.64:1	47302:2, 47304:2

columns. Each column contains a subset of attributes. This vertically partitions the table. For example, the sliced table in Table 5 contains 2 columns: the first column contains (age, sex) and the second column contains (Zipcode, disease). The sliced table shown in Table 4 contains 4 columns where each column contains exactly one attribute. Slicing also partition tuples into buckets. Each bucket contains a subset of tuples. This horizontally partitions the table. For example, both sliced tables in Table 4 and 5 contain 2 buckets, each containing 4 tuples.

Within each bucket, values in each column are randomly permutated to break the linking between different columns. For example, in the first bucket of the sliced table shown in Table 5, the values ((22,M), (22, F), (33, F), (52, F)) are randomly permutated and the values ((47906, dyspepsia), (47906, flu), (47905, flu), (47905, bronchitis)) are randomly permutated so that the linking between the two columns within one bucket is hidden.

Table 4: One-attribute-per-column slicing

Diseases	Sex	Age	Zipcode
Flu	F	22	47906
Flu	M	22	47905
dysp	F	33	47906
Bron	F	52	47905
Flu	M	54	47302:2, 47304:2
dysp	F	60	47302:2, 47304:2
dysp	M	60	47302:2, 47304:2
gast	F	64	47302:2, 47304:2

Table 5: The sliced table

Diseases	Sex	Age	Zipcode
Flu	M	22	47905
Dysp	F	22	47906
Bron	F	33	47905
Flu	F	52	47906
Gast	M	54	47304
Flu	M	60	47302
Dysp	M	60	47302
Dysp	F	64	47304

MATERIALS AND METHODS

The research here focuses on Radial Basis Function (RBF) and Multiple Linear Regression (MLR) classifiers which necessitate computing distance functions and dot products over sliced data to classify those samples prior to execution of classification steps what is needed first is finding dot product instance based sliced data to perform better classifier, these steps as follows: According to the Taylor theorem know that for any differentiable function g (X) and for random variable X with E(X) = μ_x and Var(X) = σ_x^2 can approximate g(X) around μ_x as:

$$g(X) \sim g(\mu_X)+(X-\mu_X)g^{I}(\mu_X)+1/2(X-\mu_X)^2 g^{II} \quad (1)$$

Such an approximation provides a first order approximation of the expected value of g (X) as:

$$E(g(X)) \sim E(g(\mu_X)+(X-\mu_X)g^{I}(\mu_X)) = g((\mu_X)) \quad (2)$$

Also can obtain a second order approximation of the expected value of g(x) as:

$$E(g(X)) \sim E(g(\mu_X)+(X-\mu_X)g^{I}(\mu_X))+ E\frac{(X-\mu_X)^2 g^{II}(\mu_X)}{2} \quad (3)$$

Let $E(X_d^t)$ and $E(X_e^t)$ denotes the expected value and square Euclidean distance on the tth attributes such as age, sex, zipcode and diseases in the sliced data. Then, formulate $E(X_d)$ and $E(X_e)$ as the summation of attribute wise expected values:

$$E(X_d) = \sum_{t=1}^{m} E(X_d^t) \tag{4}$$

$$E(X_e) = \sum_{t=1}^{m} E(X_e^t) \tag{5}$$

These values of $E(X_d^t)$ and $E(X_e^t)$ can be calculated for both numerical and categorical attributes, Numerical quasi identifiers first calculate $E(slic(X_i)[t]^T$. $E(slic(X_j)[t])$ for i and j:

$$E(X_d^t) = E(slic(x_i)[t] \times slic(x_j)[t]) \tag{6}$$

$$E(X_e^t) = E(slic(x_i)[t]^2 - 2E(slic(x_i)[t] \times E(slic(x_j)[t]) + E(slic(x_j)[t]^2) \tag{7}$$

Hence, depending on long and efficiently estimation of first two moments of gas is performed, the knowledge regarding probability distribution function is not necessary. Let us consider in this case here knowing some range r = (1. 35) is its expected value and its variance is necessary (second moment calculation is possible from variance and expected value). Categorical attributes values representing probability of the values are equal. Therefore, assuming that the slic $((x_i)(t))$ and slic $((x_j)(t))$ are from the same domain S, $E(slic(x_j)(t)\, slic(x_j)(t))$ can be calculated for i¦j as follows: Let $f_v(v)$ and $f_w(w)$ denotes the probability mass function of sliced instances values $V = slic(x_i)(t)$ and $W = slic(x_j)(t)$:

$$E(X_d^t) = Pr(V = M) = \sum_{K \in S} Pr(V = k) \times Pr(W = k) \tag{8}$$

$$\sum f_V(k) \times f_w(k) \tag{9}$$

The square distance of:

$$E(X_d^t) = Pr(V \neq W) = 1 - E(X_d^t) \tag{10}$$

Finally, need to define the probability mass function of an original value V∈S so that $E(E_e)$ and $E(X_d)$ are defined on arbitrary pairs of generated and original values:

$$f_v(k) \begin{cases} 1 \text{ if } v = k \\ 0 \text{ otherwise} \end{cases} \tag{11}$$

Radial Basis Functions (RBF): A Radial Basis Function (RBF) neural network has three layers, input, hidden and output layers. In the approach used here, generalized value of every sliced record r has been accompanied as input dataset samples statistics that have been collected and derived from records in the same equivalence class as r. The additional information, disclosed along with sliced data, is provided enabling accurate computation of expected values for vital functions like data analysis like dot product and square distance Neurons part of the hidden layer possess radial basis transfer functions where the outputs are inversely proportional to distance from neuron's center. Hidden units execute a radial basis function. Inputs are represented by sliced record and output a class r in the sliced data which is illustrated in Fig. 1. Hidden units match theneural network subclasses. Hidden units ascertain network's classification accuracy.

The RBF function enables allocation of each RBF neuron to respond to each of sub-spaces of a pattern class r, formulated by the clustered training samples. On the basis of the same, learning inherent in the hidden layerusually is configured to ascertain the problem of finding clusters along with parameters using some means that are characterized by functional optimization. RBF network sets the constant of the weight between input layer and hidden layer and updates the weight between hidden layer and output layer. The hidden-to-output weights are usually 0 or 1; for each hidden unit, a weight of 1 is used for the connection to the output which is true for it and all other connections are given weights of 0. Input layer comprises of source neurons which connect the network and its environment. Hidden layer neurons are usually associated with the centerwhich helps in assessing network's behavior structure. Output layer presentsnetwork response to input layer's activation pattern which serves the purpose of being the summation unit. The Kernel based radial biases function $K(slic(x_i), slic(x_j))$ is used in this research computed by the ith hidden neurons is maximum when the input vector is near the center(the mean)of that neuron. Various kinds of radial basis functions are present like:

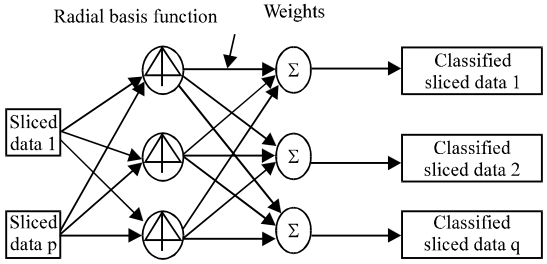

Fig. 1: RBF neural network architeeture

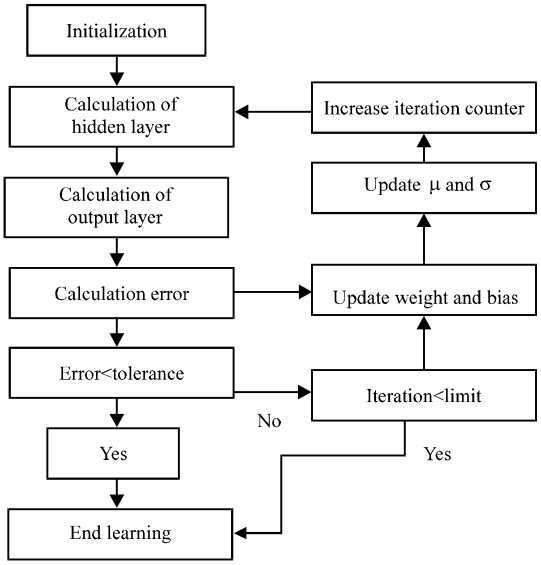

Fig. 2: Learning the sensitive attribute (Target: Occupation) vs classifiers

$$K(slic(x_1), slic(x_j)) = \exp(-\gamma x_e) \quad (12)$$

A scertaining hidden layer neurons is relatively significant as this influences network's complexity as well asits generalizing capability. Center position in hidden layer also considerably influences performance of the network; hence determining centers optimal locations is inherently another relevant task. RBF networks training procedure is inclusive of centers optimization for every neuron. Thereafter which selection of weights between hidden and output layers must be carried out appropriately. Lastly, in the RBF network training procedure bias values added along with every output are determined. In the model every, RBF neuron in the RBF network inserts into the state space an open basin of attraction around its center whichleads to introduction of a stable fixed point. The processing procedure of the RBF Neural Network (RBFNN) is illustrated in Fig. 2. The output of the f: $R_n \neg R$ of the network is thus:

$$f_n(x) = \sum_{m=1} W_{mn} K(Slic(x_i), Slic(x_j)) \quad (13)$$

where, W_{mn} is the connection weight of the mth RBF unit to the nth output node. The x is the input vector that is sliced data:

$$z_n(x) = \frac{1}{1+e^{-f_n(x)}} \quad (14)$$

Multiple Linear Regression (MLR): Regression analysis is generally employed for predicting value of one or more responses from any given set of predictors which are sliced data results. Alternatively, it may also be utilizedfor estimating linear association between predictors and responses. Predictors may be either continuous or categorical or sometimes even a mixture of both. First revisiting the multiple linear regression model or MLR for a particular dependent variable is performed thereafter which it moves on where more than one response is measured for every sample unit. Simple linear regression model, a single response measurement Y is related to a single predictor (covariate, regressor) X for each observation. The critical assumption of the model is that the conditional mean function is linear:

$$E(Y|X) = \alpha + \beta X \quad (15)$$

This leads to the following "multiple regression" mean function:

$$E(Y|X) = \alpha + \beta_1 X_1 + ... + \beta_p X_p \quad (16)$$

Where:

X_1 = Source data of the sliced data with instances upto p predictions
α = Intercept
β_j = Slopes or coefficients

The observation of the class labels varies based on the responses in the sliced data, the above step is further improved to specify how the responsesvary around their mean values. This leads to a model of the form:

$$Y_m = \alpha + \beta_1 X_{m1} + ... + \beta_p X_{mp} + l_m \quad (17)$$

That is similar to:

$$Y_m = E(Y|X_m) + \in_m \quad (18)$$

where, $Y_{m,n}$ for the nth predictor variable measured for the mth observation. For linear algebraic, it is written as:

$$Y_m = \beta_m + \quad (19)$$

where, β is the matrix vector product of the nthpredictor variable measured for the mth observation. In order to estimate β_m take a least squares approach that is analogous and did in the simple linear regression case. That is want to minimize:

$$\sum_m (Y_m) - a - \beta_1 X_{m,1} - ... \beta_p X_{m,p})^2 \quad (20)$$

Over all possible values of the intercept and slopes. It is a fact that this is minimized by setting:

$$\hat{\beta} = (X'X)\ X\ Y \quad (21)$$

And $(X^1X)^{-1}$ are symmetric matrices is p+1 dimensional vector. The linear regression with multiple model the instance of the sliced data table is defined as:

$$Slic(X_i, X_j) = X_i^T X_j \quad (22)$$

The fitness values are:

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}(X'Y) \quad (23)$$

And there sidues are:

$$\hat{r} = Y - \hat{Y} = (1 - X(X'X)^{-1}X)Y \quad (24)$$

The error standard deviation is estimated as:

$$\hat{\sigma} = Y\sqrt{\sum_m \frac{r_m^2}{(n-p-1)}} \quad (25)$$

The variance of $\hat{\alpha}, \hat{\beta}_1, ..., \hat{\beta}_p$ are the diagonal elements of the standard error matrix:

$$\hat{\sigma}^2(X'X)^{-1} \quad (26)$$

For every p response variable the very same predictors set in the sliced data must be employed. At any given point of time model's goodness of t and respective diagnostics are performed for only one regression model. Thereafter, which the system establishes particularly statistical information regarding quasi-identifiers so that it is sufficient enough for accurate evaluation of expected value of kernel functions (or square distance for RBF classification). Equipped with the information, an extension has been suggested here with respect to existing slicing methods, so that additionally along with generalized values of quasi-identifiers, sliced data may also comprise of one new attribute per quasi-identifier. The new attribute containsinformation regarding expected value and variance for numerical quasi-identifiers and probability mass function for categorical quasi-identifiers.

RESULTS AND DISCUSSION

Adult data set from the UC Irvine machine learning repository which is comprised of data collected from the US census is used for evaluation. The data set is described in Table 6. Tuples with missing values are eliminated and there are 45,222 valid tuples in total. The adult data set contains 15 attributes in total. An experiment, obtaining two data sets from the adult data set is explained. The first data set is the "OCC-7" data set which includes seven attributes: QI = ({Age, Workclass, Education, Marital-Status, Race, Sex}) and S = Occupation. The second data set is the "OCC-15" data set which includes all 15 attributes and the sensitive attribute is S = Occupation. Note that do not use salary as the sensitive attribute because salary has only two values 50<50 K which means that even 2-diversity is not achievable when the sensitive attribute is salary. In the "OCC-7" data set, the attribute that has the closest correlation with the sensitive attribute occupation is gender with the next closest attribute being education. In the "OCC-15" data set, the closest attribute is also Gender but the next closest attribute is salary.

Table 6: The "OCC-7" data set description

Attributes	Type	No. of values
Age	Continuous	74
Workclass	Categorical	8
Final-weight	Continuous	NA
Education	Categorical	16
EducationNum	Continuous	16
Maritalstatus	Categorical	7
Occupation	Categorical	14
Relationship	Categorical	6
Race	Categorical	5
Sex	Categorical	2
CapitalGain	Continuous	NA
Capital-loss	Continuous	NA
Hoursperweek	Continuous	NA
Country	Categorical	41
Salary	Categorical	2

Evaluate the quality of the sliced anonymized data for classifier learning which has been used in (Carlisle *et al*., 2007). In all of the existing research use the Weka software package to evaluate the classification accuracy for Decision Tree C4.5 (J48) and Naive Bayes (NB). In our proposed research, we use RBF and MLR Classifier for classification of sliced data.

Learning the sensitive attribute: In this experiment, build a classifier on the sensitive attribute which is "Occupation".

Learning a QI attribute: In this experiment, build a classifier on the QI attribute "Education".

The results of the learning sensitive attribute,can take different l values {5, 8, 10} for various classifiers accuracy is illustrated in Fig. 3. It shows that the performance accuracy of the proposed classifier such as the RBF and MLR performs better for sliced data table than the weka tool based classifier for learning sensitive attributes.

The results of the Learning a QI attribute can take different l values {5, 8, 10} for various classifiers accuracy is illustrated in Fig. 4. It shows that the performance

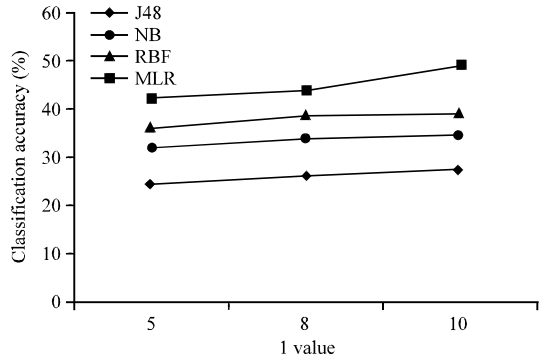

Fig. 3: Learning the sensitive attribute (Target: occupation) vs classifiers

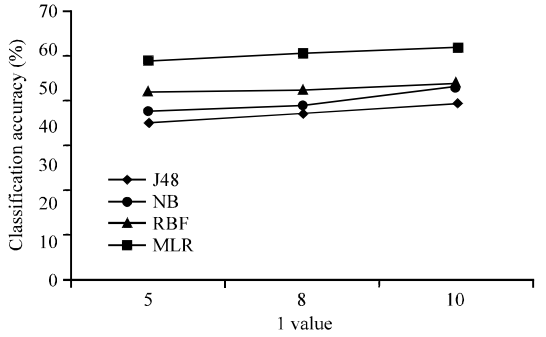

Fig. 4: Learning a QI attribute (Target: education) vs classifiers

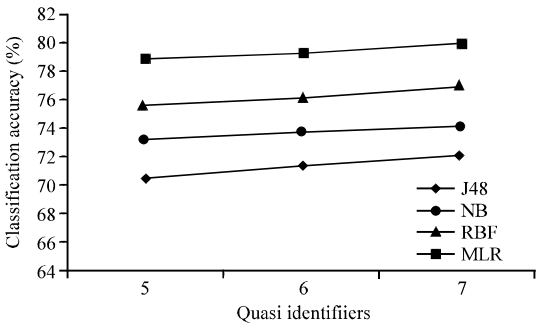

Fig. 5: Learning quasi identifiers vs classifiers

accuracy of the proposed classifier such as the RBF and MLR performs better for sliced data table than the weka tool based classifier for learning a QI attribute in education field.

The results of the learning quasi identifier can take different l values {5, 8, 10} value for various classifiers accuracy is illustrated in Fig. 5. It shows that the performance accuracy of the proposed classifier such as the RBF and MLR performs better for sliced data table than the weka tool based classifier for learning quasi identifiers.

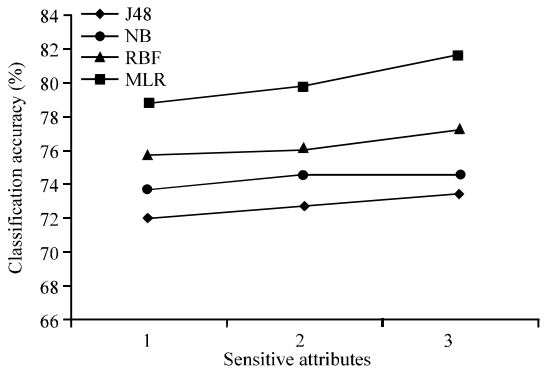

Fig. 6: Learning sensitive attributes vs classifiers

The results of the Learning sensitive attributes can take different values {1, 2, 3} for various classifiers accuracy is illustrated in Fig. 6. It shows that the performance accuracy of the proposed classifier such as the RBF and MLR performs better for sliced data table than the weka tool based classifier for learning a QI attribute in education field.

CONCLUSION

In this study, presents a new approach called slicing to privacy-preserving microdata publishing. Slicing overcomes the limitations of generalization and bucketization and pre-serves better utility while protecting against privacy threats. In this study, classification models such as RBF and Multiple Linear Regression (MLR) used sliced data. The RBF and MLR classifier calculates expected values of functions that are imperative overall like kernel based product and square distance by modeling sliced data as uncertain data. Experiment results indicate that RBF and MLR based classifier carefully disclose quasi-identifiers statistics and built models which precisely handle sliced data which in turn could directly be implemented during the process of privacy preservation data mining, without considerably degrading performance, privacy and classification accuracy.

REFERENCES

Aggarwal, C.C. and P.S. Yu, 2008. Privacy-Preserving Data Mining Models and Algorithms. Springer-Verlag.

Bayardo, R.J., R. Agrawal, 2005. Data Privacy through Optimal k-Anonymization. Proceedingsofthe ICDE Conference, pp: 217-228.

Carlisle, D.M., M.L. Rodrian and C.L. Diamond, 2007. California inpatient data reporting manual, medical information reporting for California. 5th Edn. Tech. rep., Office of Statewide Health Planning and Development.

El-Emam, K. and F.K. Dankar, 2008. Protecting privacy using k-anonymity. J. Am. Medical Informatics Association, 15 (5): 627-637.

Fung, B., K. Wang and P. Yu, 2005. Top-down specialization for information and privacy preservation. In ICDE'05, Tokyo, Japan, pp: 205-216.

Kumari, V.V., S.S. Rao, K.V.S.V.N. Raju, K.V. Ramana and B.V.S. Avadhani, 2008. Fuzzy based approach for privacy preserving publication of data. Int. J. Computer Sci. Network Security, 8 (1): 115-121.

LeFevre, K., D.J. DeWitt and R. Ramakrishnan, 2006. Workload-aware anonymization. In KDD'06, Philadelphia, PA, USA, pp: 277-286.

Lin, J.L. and M.C. Wei, 2008. An Efficient Clustering Method for k-Anonymization. PAIS, pp: 46-50.

Ram, P.R., S., Kvsvn Raju and V. ValliKumari, 2011. A Dynamic programming approach for privacy preserving collaborative data publishing. Int. J. Computer Applications (0975-8887) Vol. 22, No. 4.

Shang, N., M. Nabeel, F. Paci and E. Bertino, 2010. A privacy-preserving approach to policy-based content dissemination. In Data Engineering (ICDE). IEEE 26th International Conference, pp: 944-955.

Samarati, P., 2001. Protecting Respondent's Privacy in Microdata Release. TKDE, 13 (6): 1010-1027.

Ye, Y., Q. Deng, C. Wang, D. Lv, Y. Liu and J. Feng, 2008. BSGI: An effective algorithm towards Stronger l-Diversity. DEXA, pp: 19-32.

Yu, J., J. Han, J. Chen and Z. Xia, 2009. TopDown-KACA, An Efficient Local-Recoding Algorithm for k-Anonymity. IEEE Gr.C, pp: 727-732.