The International Journal Of Engineering And Science (IJES)
|| Volume || 5 || Issue || 3 || Pages || PP -48-57 || 2016 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Analysis on Fuzzy based Privacy Preserved Classification

[1]Selva Rathna S., [2]Dr. T.Karthikeyan
[1]*Department of Computer Science, Mamonmaniam Sundaranar University, Tirunelveli, Tamil Nadu*
[2]*P.S.G. Arts & Science College, Bharathiyar University, Coimbatore, Tamil Nadu, India*

----------**ABSTRACT**----------
In this research, various data mining classification algorithms are applied on Fuzzy based Privacy preserved database to analyse their performance to identify the suitable classifier. Fuzzy based member ship functions like Bell shape, S- Shape and PI shape member ship functions are applied on standard database to generate Privacy preserved database. Further, various classifiers are applied on the sanitised database and the results are compared. WEKA tool is used for analysing classification algorithm on privacy preserved database generated using various fuzzy member ship function. This analysis will help to develop new Fuzzy Based privacy preserving classification algorithms.

Keywords *- Classification, Clustering, fuzzy Logic, membership Function, Privacy Preserving Data Mining, WEKA.*

Date of Submission: 01 March 2016 Date of Accepted: 20 March 2016

I. INTRODUCTION

Privacy preserving Data mining means hiding the sensitive values of a database and performing data mining operations on the sanitised database. Effective preservation techniques will enable the data base owner to prepare privacy preserved data base which can be further used by a third party for any data mining operations like clustering, classification, association rule mining etc. A number of techniques have been proposed for modifying or transforming data to preserve privacy which are effective without compromising security. In Section 2, earlier researches related to Privacy preserving data mining and use of Fuzzy logic techniques in maintaining privacy of the sensitive data is discussed. In Section 3, various classification algorithms in data mining is discussed. In Section 4, analysis of classification algorithm on Fuzzy based privacy preserved data using various fuzzy approaches is discussed. In Section 5, results of the analysis are presented. In Section 6, conclusion of the analysis is given.

II. PRIVACY PRESERVED DATA MINING

2. Privacy Preserving Data Mining Survey

2.1 Privacy preserving Data mining

Malik [1], describes the current scenario of Privacy preserving data mining and propose some future research directions. In Shweta [2], all Cryptography and Random Data Perturbation methods techniques of PPDM is studied. Chris [3] illustrates the application of certain techniques for preserving privacy on experimental dataset, and reveals their effects on the results.

2.2. Fuzzy Based Privacy Preserved Data Mining

Mukkamala[4] compared a set of fuzzy-based mapping techniques in terms of their privacy-preserving property and their ability to retain the same relationship with other fields. Jian [5] proposed a method to extract global fuzzy rules from distributed data with the same attributes in a privacy-preserving manner. Cano [6] proposed a fuzzy c-regression method to generate synthetic data which allows third parties to do statistical computations with a limited risk of disclosure Kasugai [7] studies the applicability of fuzzy k-member clustering to privacy preserving pattern recognitionl. k-member clustering is a basic technique for achieving k-anonymization, in which data samples are summarized so that any sample is indistinguishable from at least k - 1 other samples. Tanak [8] proposed a secure framework for privacy preserving fuzzy co-clustering for handling both vertically and horizontally distributed cooccurrence matrices.

2.3. Privacy Preserved Classification Algorithms

Yang [9] proposed simple cryptographic approach of classification that is efficient even in a many-customer setting. It provides strong privacy for each customer, and does not lose any accuracy as the cost of privacy. Olvi [10] proposed a novel privacy-preserving nonlinear support vector machine (SVM) classifier which is public but does not reveal the privately-held data, has accuracy comparable to that of an ordinary SVM classifier based on the entire data. An algebraic-technique-based scheme is introduced by Zhang [11] which can build classifiers more accurately but disclose less private information. Benjamin [12] find a k-anonymization, not necessarily optimal in the sense of minimizing data distortion, which preserves the classification structure. Zhoujia [13] put forward a framework to synthesize and characterize existing PPDDM protocols so as to provide a standard and systematic approach of understanding PPDDM-related problems, analyzing PPDDM requirements and designing effective and efficient PPDDM protocols.

III. FUZZY BASED PRIVACY PRESERVED CLASSIFICATION

3. Fuzzy Based Privacy Preservation

3.1. Fuzzy member ship function

Various methods and algorithms were used for privacy preservation in data mining process. In this paper, fuzzy based privacy preservation is considered for analysis. In this approach, a fuzzy membership function is applied on the original data to hide the sensitive data which will generate the sanitized data with fuzzification methods. The shapes of various fuzzy membership functions are plotted in Fig.2 The following membership functions are used in this paper for converting the original data into privacy preserved data.

3.1.1. S-Fuzzy Membership Function

The S-Shaped function takes three parameters as input and produces the modified membership plane or property plane. By using S-Shaped membership function the shape of the input is modified i.e. fuzzy domain values are modified. The formula for S-Shaped membership function is given in Equation (1).

$$f(x,a,b,c,d) = = \begin{cases} 0, & x \le a \\ 2\left(\frac{x-a}{b-a}\right)^2, & a \le x \le \frac{a+b}{2} \\ 1-2\left(\frac{x-a}{b-a}\right)^2, & \frac{a+b}{2} \le x \le b \\ 1, & b \le x \le c \end{cases} \quad (1)$$

S-Shaped function has *a* ,*b* and *c* as input parameters where *a* = minimum value, *c* = maximum value and *b= (a + c) / 2* for given range. For any point *i, j*, membership value $\mu_{x(i,j)}$ is computed using a, b & c.

3.1.2. Z-Fuzzy Membership Function

The Z-Shaped function takes three parameters as input and produces the modified membership plane or property plane. By using Z-Shaped membership function the shape of the input is modified i.e. fuzzy domain values are modified. The formula for Z-Shaped membership function is given in Equation (2)

$$f(x,a,b,c,d) = \begin{cases} 1, & x \le a \\ 1-2\left(\frac{x-a}{b-a}\right)^2, & a \le x \le \frac{a+b}{2} \\ 2\left(\frac{x-a}{b-a}\right)^2, & \frac{a+b}{2} \le x \le b \\ 0, & b \le x \le c \end{cases} \quad (2)$$

Z-Shaped function has *a* ,*b* and *c* as input parameters where *a* = minimum value, *c* = maximum value and *b= (a + c) / 2* for given range. For any point *i, j*, membership value $\mu_{x(i,j)}$ is computed using a, b & c.

3.1.3. PI-Fuzzy Membership Function

The PI-Shaped function takes four parameters *a, b, c* and *d* as input and produces the modified membership plane or property plane. Parameter *a* and *d* located in the feet and *b* and *c* is located in the top of the curve. PI membership function is the product of *S* membership function and *Z* membership function. The formula of PI fuzzy membership function is as given in Equation (2).

$$f(x,a,b,c,d) = \begin{cases} 0, & x \le a \\ 2\left(\frac{x-a}{b-a}\right)^2, & a \le x \le \frac{a+b}{2} \\ 1-2\left(\frac{x-a}{b-a}\right)^2, & \frac{a+b}{2} \le x \le b \\ 1, & b \le x \le c \\ 1-2\left(\frac{x-c}{d-c}\right)^2, & c \le x \le \frac{c+d}{2} \\ 2\left(\frac{x-d}{d-c}\right)^2, & \frac{c+d}{2} \le x \le d \\ 0, & x \ge d \end{cases} \quad (3)$$

3.1.4. Bell-Fuzzy Membership Function

The Bell-Shaped function takes three parameters *a*, *b* and *c* as input and produces the modified membership plane or property plane. *b* is always positive and *c* is located in the center of the curve. The formula of PI fuzzy membership function is as given in Equation (3)

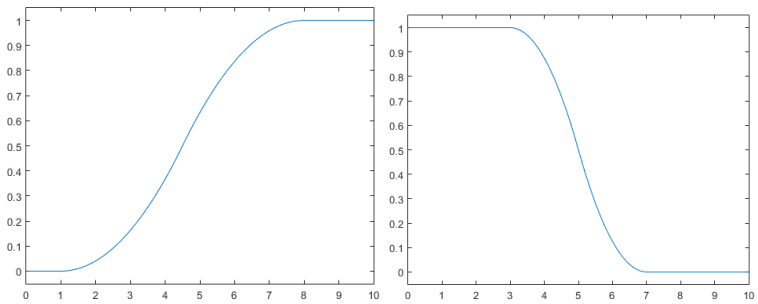
$$f(x,a,b,c,d) = \frac{1}{1+\left|\frac{x-c}{a}\right|^{2b}} \quad (4)$$

3.1.5. Elliptical Type 2 - Fuzzy Membership Function

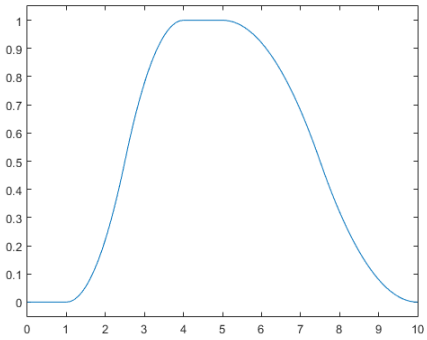
Elliptical Type 2 Fuzzy system with two inputs can be written as

$$y = q\frac{\sum_{j=1}^{J}\sum_{i=1}^{I}\underline{W_{ij}}f_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{I}\underline{W_{ij}}} + (1-q)\frac{\sum_{j=1}^{J}\sum_{i=1}^{I}\overline{W_{ij}}f_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{I}\overline{W_{ij}}} \quad (5)$$

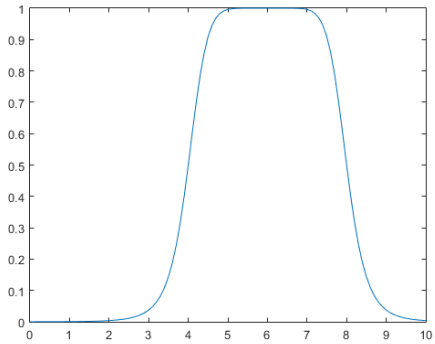
where $\underline{W_{ij}} = \underline{\mu_{1i}}\,\underline{\mu_{2j}}$ and $\overline{W_{ij}} = \underline{\mu_{1i}}\,\underline{\mu_{2j}}$ and the parameter *q* is the weighting parameter which reflects the sharing of the contribution of the upper and the lower MFs.



S- Fuzzy membership Z- Fuzzy membership



PI- Fuzzy membership



Bell shape Fuzzy membership

Figure 2 : Shapes of Fuzzy membership functions

3.2. Fuzzy based Privacy preserved Classification

3.2.1. Bayes Net Classifier

BayesNet learns Bayesian networks made in nominal attributes (numeric ones are prediscretized) and no missing values (any such values are replaced globally). Bayes Nets or Bayesian networks are graphical representation for probabilistic relationships among a set of random variables.

Given a finite set X={X_1…Xn} of discrete random variables where each variable Xi may take values from a finite set represented by Val (X_i). A Bayesian network is an annotated directed acyclic graph (DAG) G that encodes a joint probability distribution over X. The nodes of the graph correspond to the random variables X1… Xn. The links of the graph represent to the direct influence from one variable to the other. If there is a directed relationship from variable Xi to variable Xj , variable Xi will be a parent of variable Xj . Each node is annotated with a conditional probability distribution (CPD) that represents P (Xi | Pa (Xi)) where Pa (Xi) denotes the parents of Xi in G. [5]. The pair (G, CPD) encodes the joint distribution P(X1…Xn). A unique joint probability distribution over X from G is factorized using Equation (5):

$$P(X1…Xn) = \Pi i\ (P\ (Xi \mid Pa\ (Xi))) \quad \ldots\ldots \quad (5)$$

3.2.2. Naïve Bayes:

The Naive Bayes algorithm is based on conditional probabilities. NB uses Bayes Theorem that calculates a probability by counting the frequency of values and combinations of values in the historical data. Naive Bayes is a simple technique and highly scalable for constructing classifiers. An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. It uses Bayes theorem which provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c) using Equation (6)

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad \ldots \quad (6)$$

Where $P(c \mid X)=P(x_1 \mid c) * P(x_2 \mid c \ldots\ldots * P(x_n \mid c) \times P(c)$
And

- $P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

3.2.3. AdaBoost:

Ada Boost is short for "Adaptive Boosting" which refers to a particular method of training a boosted classifier in the form of

$$F_T(x) = \sum_{t=1}^{T} f_t(x) \quad \ldots \quad (7)$$

where each f_t is a weak learner that takes an object x as input and returns a real valued result indicating the class of the object. The sign of the weak learner output identifies the predicted object class and the absolute value gives the confidence in that classification.

Each weak learner produces an output, hypothesis $h(x_i)$, for each sample in the training set. At each iteration t, a weak learner is selected and assigned a coefficient α_t such that the sum training error E_t of the resulting t-stage boost classifier is minimized.

$$E_T = \sum_i E[F_{t-1}(x_i) + \alpha_t h(x_i)]$$

Here $F_{t-1}(x)$ is the boosted classifier that has been built up to the previous stage of training,

$E(F)$ is some error function and

$f_t(x) = \alpha_t h(x_i)$ is the weak learner that is being considered for addition to the final classifier.

3.2.4. Random Tree

Random forests is a notion of the general technique of random decision forests that are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1, \ldots, x_n$ with responses $Y = y_1, \ldots, y_n$, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For $b = 1, \ldots, B$:

1. Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.
2. Train a decision or regression tree f_b on X_b, Y_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x')$$

3.2.5. Simple CART

Classification and Regression Tree Analysis, CART, is a simple yet powerful analytic tool that helps determine the most "important" (based on explanatory power) variables in a particular dataset, The process of computing classification and regression trees can be characterized. It involves four basic steps and uses formula (8):

- Specifying the criteria for predictive accuracy
- Selecting splits
- Determining when to stop splitting
- Selecting the "right-sized" tree.

$$F(x) = \sum_{i=1}^{M} c_m I(x \in R_m) \quad \ldots \quad (8)$$

$\{R_m\}_1^M$ are subregions of the input variable space and x is a vector of input variables.

c_m are the estimated values of the outcome(y) in region R_m.

CART tries to minimize

$$e(T) = \sum_{i=1}^{N} [y_i - \sum_{m=1}^{M} c_m I(x \in R_m)]^2$$

with respect to c_m and R_m

3.2.6. K Star

K* is a simple, instance based classifier, similar to KNearest Neighbour (K-NN). It differs from other instance-based learners in that it uses an entropy-based distance function. The benefits are that it provides a consistent approach to handling of real valued attributes, symbolic attributes and missing value. New data instances, x, are assigned to the class that occurs most frequently amongst the k-nearest data points, y_j , where j = 1, 2…k. Entropic distance is then used to retrieve the most similar instances from the data set.

The K* function can be calculated as:

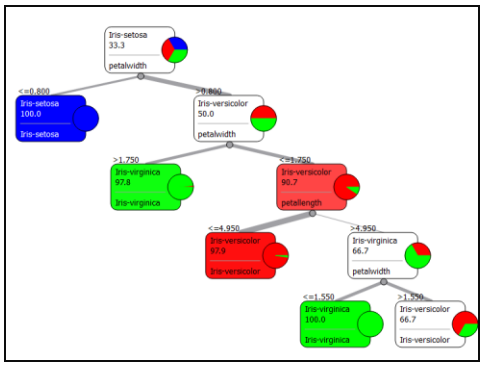
$$K*(yi\ ,\ x) = -ln\ P*(yi\ ,\ x) \quad \ldots (9)$$

Where P* is the probability of all transformational paths from instance x to y.

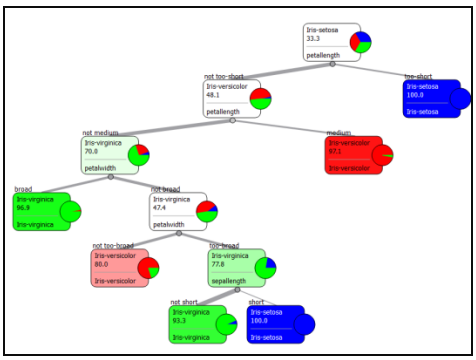
IV. RESULTS

4. Fuzzy Based Privacy Preserved Classification

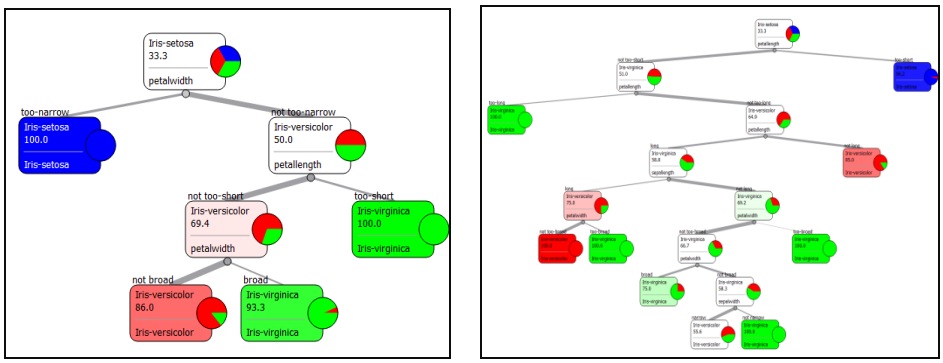
Various membership functions were applied on IRIS data to generate privacy preserved data base. Bell Shape, PI Shape, S Shape, Z shape, Elliptical Type 2 Fuzzy membership functions were applied on IRIS data. The resultant data is processed in Orange Tool for generating classification trees and the classification trees generated were presented in Fig.1.



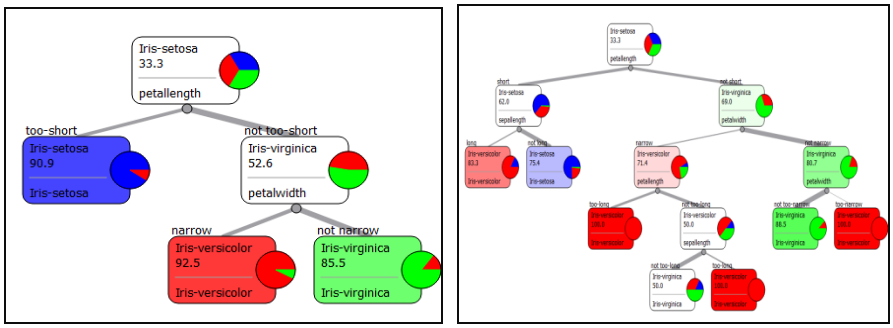
a.IRIS



b. IRIS – Bell



c.IRIS - Elliptical Type 2 **d.IRIS – PI**



e.IRIS - S **f.IRIS – Z** .

Figure 1 : Classification Tree for IRIS Data set

The performance of various classification algorithms are evaluated by comparing Mean absolute error, Number of instances correctly classified and Root Relative Squared Error. IRIS data set is sanitized using Bell, S-Shape, PI Shape, Z shape and ET2 fuzzy member ship functions to create privacy preserved data. Further the preserved data is used in WEKA tool to apply various classification algorithms on the data. Based on the analysis, the comparison of Mean absolute error, Number of instances correctly classified and Root Relative Squared Error is tabulated in Table 1, 2 & 3. The comparison is plotted in Figure 2, 3 &4.

Table 1 : Comparison of Mean Absolute Error

Mean absolute error	Original	IRIS-Bell	IRIS-S	IRIS-PI	IRIS-Z	IRIS-ET2
BayesNet	0.045	0.045	0.072	0.095	0.136	0.044
Naïve Bayes	0.034	0.050	0.076	0.101	0.145	0.050
KSTAR	0.043	0.074	0.093	0.120	0.178	0.080
ADABoost	0.069	0.203	0.293	0.262	0.389	0.171
Simple Cart	0.437	0.077	0.735	0.114	0.197	0.049
Random Tree	0.053	0.057	0.065	0.074	0.135	0.033

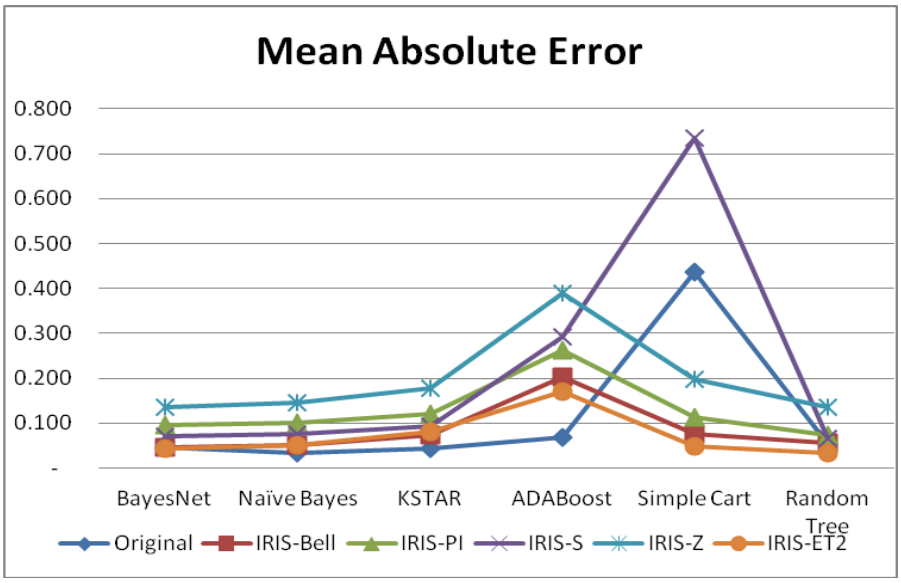


Figure 2 : Mean Absolute Error

Table 2 : Comparison of Correctly Classified Instances

Correctly Classified Instances	Original	IRIS-Bell	IRIS-S	IRIS-PI	IRIS-Z	IRIS-ET2
BayesNet	139	143	132	140	124	146
Naïve Bayes	144	142	132	138	125	146
KSTAR	142	141	128	135	126	144
ADABoost	143	132	99	103	98	130
Simple Cart	143	140	132	141	115	144
Random Tree	138	137	135	138	120	144

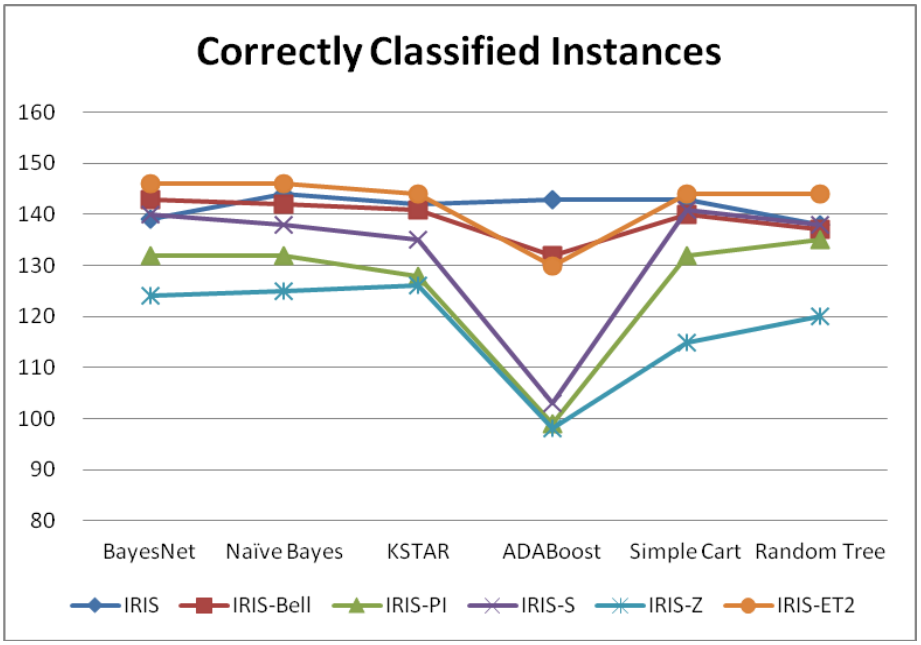
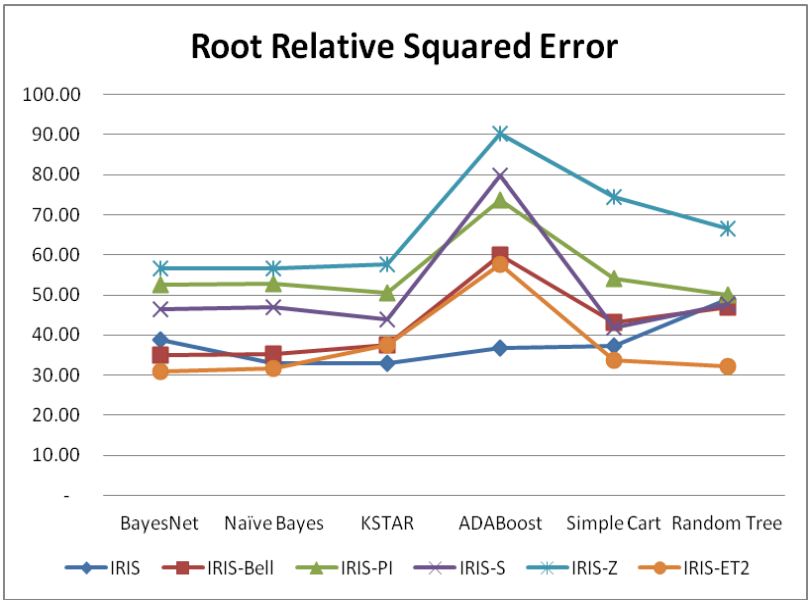


Figure 3 : Comparision of Correctly classified Instances

Table 4 : Comparison of Root Relative Squared Error

Root Relative Squared Error	Original	IRIS-Bell	IRIS-S	IRIS-PI	IRIS-Z	IRIS-ET2
BayesNet	38.78	34.89	52.60	46.37	56.69	30.82
Naïve Bayes	32.88	35.36	52.81	46.98	56.67	31.81
KSTAR	32.98	37.61	50.47	43.93	57.71	37.58
ADABoost	36.69	59.89	73.59	79.79	90.29	57.55
Simple Cart	37.17	43.16	54.21	41.81	74.36	33.63
Random Tree	48.99	47.07	50.00	47.81	66.57	32.22



V. CONCLUSION

As per the results found in the analysis, Bayes Net classification methods performed well based on Number of correctly classified instances, Mean absolute error, Root relative Squared Error. Random Tree method also have good performance based on Root relative Squared Error. Also, this research helped to identify the performance of various fuzzy membership algorithms in Privacy preserved classification algorithms. The comparision of analysis of the data shows that Bell member ship function have good performance in classification. S-Shaped Member function also perform well while comparing the Mean absolute error. This analysis helped to identify the best classification algorithms which can be applied for Privacy preserving data mining. This will lead to further researches on developing new algorithms for performing Privacy preserved classification with better performance than existing algorithms.

ACKNOWLEDGEMENTS

We thank all those who have supported us in this research.

REFERENCES

[1] Malik, M.B., Ghazi, M.A., Ali, R., (2012), Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects, *Third International Conference on Computer and Communication Technology* (ICCCT),. pp: 26 – 32

[2] Shweta Taneja, Shashank Khanna, Sugandha Tilwalia, Ankita, (2014), A Review on Privacy Preserving Data Mining: Techniques and Research Challenges, *International Journal of Computer Science and Information Technologies*, Vol. 5 (2) , 2014, pp: 2310-2315

[3] Chris, C., K. Murat, V. Jaideep, L. Xiadong and Y.Z. Michael, (2002). Tools for privacy preserving distributed data mining. *ACM SIGKDD Explorations* Newslett., 4: 28-34.

[4] Mukkamala, R., Ashok, V.G. (2011) Fuzzy-based Methods for Privacy-Preserving Data Mining Eighth International Conference on Information Technology: New Generations (ITNG), pp: 348 – 353

[5] Jiang, J. and Umano, M. (2014), Privacy preserving extraction of fuzzy rules from distributed data with different attributes, Joint 7th International Conference on and Advanced Intelligent Systems (ISIS), 15th International Symposium on Soft Computing and Intelligent Systems (SCIS), 2014, pp : 1180-1185

[6] Cano, I. ; Torra, V., (2009) Generation of synthetic data by means of fuzzy c-Regression . IEEE International Conference on Fuzzy Systems, 2009. FUZZ-IEEE, pp: 1145 – 1150

[7] Kasugai, H. ; Kawano, A. ; Honda, K. ; Notsu, A., (2013), A study on applicability of fuzzy k-member clustering to privacy preserving pattern recognition, IEEE International Conference on Fuzzy Systems (FUZZ), 2013, pp:1-6

[8] Tanaka, D.; Oda, T.;Honda, K.;Notsu, A., (2014), Privacy preserving fuzzy co-clustering with distributed cooccurrence matrices Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS), 2014 and 15th International Symposium on Advanced Intelligent Systems (ISIS), pp: 700-705

[9] Zhiqiang Yang; Sheng Zhong; Rebecca N. Wright (2005), Privacy-Preserving Classification of Customer Data without Loss of Accuracy, Proceedings of the Fifth SIAM International Conference on Data Mining, Newport Beach, CA, April 21–23, 2005.

[10] Olvi L. Mangasarian; Edward W. Wild; Privacy-Preserving Classification of Horizontally Partitioned Data via Random Kernels, Data Mining Institute Report 07-03, October 2007

[11] Nan Zhang.; Shengquan Wnag.; Wei Zhao.; A new scheme on privacy preserving data classification, Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005, Pages 374-383

[12] Benjamin C.M. Fung, Ke Wang, and Philip S. Yu Anonymizing Classification Data for Privacy Preservation, IEEE Transactions on Knowledge and Data Engineering, Vol. 19, NO. 5, May 2007

[13] Zhuojia Xu .; Victoria Univ., Classification of Privacy presering Distrubuted data mining protocols

About Authors

S. Selva Rathna received M.C.A degree in 2000 and M.Tech in Information Technology in 2010 in Manonmaniam sundaranar university, Tirunelveli, Tamil Nadu, India. She is presently doing Ph.D in Computer Science in Manonmaniam sundaranar university, Tirunelveli, Tamil Nadu, India.. She is highly interested in topics like data mining, data ware housing, privacy preservation, image processing etc. Her 14 years of experience in Oracle data base has supported her a lot in successful completion of this paper.

Dr. T.Karthikeyan has completed his Ph.D degree in Computer Science and presently working as Associate Professor in P.S.G. Arts and Science College, Coimbatore, Tamil Nadu, India. His extensive knowledge in Data mining, Image processing, Image mining, Security and privacy preserving data mining has supported a lot in conceptualizing this research