Asian Journal of Information Technology 15 (19): 3708-3711, 2016
ISSN: 1682-3915

Model Framework for Sensitive Data Preservation Using Fuzzy

[1]S. Dhanalakshmi, [2]J. Abdul Samath and [3]M.S. Irfan Ahmed
[1]Computer Applications and Software Systems, Sri Krishna Arts and Science College Coimbatore, India
[2]Government Arts College Udumalpet, Tamil Nadu, India
[3]Computer Applications Nehru Institute of Technology, Coimbatore, India

Abstract: Research on privacy preserving data mining focus on protecting the sensitive data by transforming the data set and also allow using the data for mining purpose. These techniques help to maintain the accuracy of the results as that of original data set. In the proposed research the data set transformation for protecting the sensitive data is done using the fuzzy member functions. We use various fuzzy member functions such as triangular member function, trapezoidal member function and sigmoid member function to convert the original dataset into perturbed dataset. The accuracy of our research is evaluated by comparing the original data set and perturbed data set on applying to various classification algorithms. The results show that it conserves privacy of sensitive data and also produce valid results.

Key words: Privacy preserving, classification, fuzzy member function, data transformation, triangular

INTRODUCTION

Rapid growth in Information technologies makes information access easier among various organizations. The information available may contain sensitive information about the individuals which has to be protected when collaboration mining is done. This leads to an area of research called privacy preserving data mining. Privacy preserving techniques falls in to three main categories they are randomization, anonymization and cryptographic techniques. The first technique randomization uses the value distortion method which draw a random value from some distribution is added to the original value. The reconstruction method is used to calculate the distribution of the original data values (Agrawal and Aggarwal, 2001). The second technique anonymization makes the individual record indistinguishable among a group of records. Various solutions has been proposed using anonymization techniques like k-anonymity, l-diversity and t-closeness (Li *et al*., 2007; Sweeney, 2002). The third cryptographic technique is an encryption based approach uses the concept of secure multiparty computation (Lindell and Pinkas, 2000, 2009).

In our reseasrc we use fuzzy logic as a privacy preserving technique. Fuzzy logic is a range-to-point control. The outputs of fuzzy control are derived from fuzzification of input values. A crisp input value will be converted into different fuzzy value using member functions called fuzzification (Bai and Wang, 2006).

Literature review: Classification is a data mining model and its goal is to accurately predict the target class for the given data. Many privacy preserving techniques has been derived for classification. Liu *et al*. (2008) discussed about the applicability of perturbation based privacy preserving data mining for real-world data. Random response technique to handle multiple attributes to conduct privacy preserving classification was proposed by Du and Zhan (2003) and Wang *et al*. (2005) proposed a secure data integration of multiple databases for classification analysis which satisfy the k-anonymity requirement. SVD-based perturbation method for privacy preserving classification was proposed by Li and Wang, (2012). Fuzzy C-Means logic based perturbation was used for privacy preserving clustering (Cano *et al*., 2010; Kumar *et al*., 2011).

MATERIALS AND METHODS

The first step in the proposed work is to preprocess the data set. We apply the Z-score normalization technique to preprocess the data set. The technique normalizes the attribute values for a specified range. It converts the data into normal distribution with mean = 0 and variance = 1. The formula of z-score is:

Corresponding Author: S. Dhanalakshmi, Computer Applications and Software Systems, Sri Krishna Arts and Science College, Coimbatore, India

$$z-score=\frac{(Value-Mean)}{Standard\ deviation}$$

Next, we identified the sensitive attributes from the data set and applied various fuzzy member functions to the data set which results in transformed data set which perturbs the sensitive data in the original data set.

In our research we used various fuzzy member functions like triangular, trapezoid and sigmoid member functions to the original dataset for transformation which helps in preserving the sensitive numeric data. Triangular member function is defined by limit a, an upper limit c and a value b where a<b<c:

$$Triangle(x:a,b,c)=\begin{cases}0 & x<a\\ \frac{x-a}{b-a} & a\le x\le b\\ \frac{c-x}{c-b} & b\le x\le c\\ 0 & x>c\end{cases}$$

Trapezoidal member function is defined by four parameters {a, b, c, d} as follows:

$$Trapezoid(x:a,b,c,d)=\begin{cases}0 & x<a\\ \frac{x-a}{b-a} & a\le x\le b\\ 1 & b\le x\le c\\ \frac{d-x}{d-c} & c\le x\le d\\ 0 & x>d\end{cases}$$

Sigmoid member function is defined by two parameters {a,c} as follows:

$$Sig(x:a,c)=\frac{1}{1+e^{-a}(x-c)}$$

Where c is the center of the function and a control the slope. After applying the fuzzy member functions the original data set is converted into perturbed dataset. Finally we tried to evaluate the accuracy of the perturbed data set as that of original data set by checking the result based on classification accuracy. We build various classification models like naïve-bayes and K-NN on both original and perturbed data set.

RESULTS AND DISCUSSION

To evaluate the accuracy of our transformation technique we used the UCI data set extracted from Pima Indians diabetes database. The diabetes data set has 768 instances which has eight continuous attributes and a class label. We apply the fuzzy member functions to the sensitive attribute of the diabetes data set. Then we apply the classification models like Naïve Bayes and K-NN on both original and perturbed data set. Cross-validation is evaluated in order to estimate the statistical performance of a learning data set. It is mainly used to estimate how accurately a model will perform in practice. The classification accuracy is depicted in Table 1. It shows that accuracy of data mining results is preserved in transformation data sets.

Table 1: Classification accuracy on original and perturbed datasets

Classifier	Accuracy (%)
Naive bayes on original data set	75.52
Naive bayes on triangle FMF perturbed data set	74.08
Naive bayes on trapezoid FMF perturbed data set	74.87
Naive bayes on sigmoid FMF perturbed data set	75.51
K-NN on original data set	68.24
K-NN on triangle FMF perturbed data set	67.71
K-NN on trapezoid FMF perturbed data set	67.71
K-NN on sigmoid FMF perturbed data set	66.15

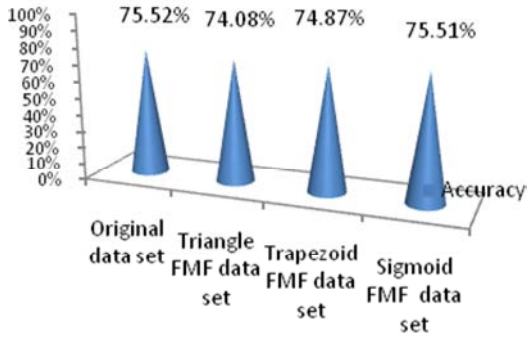

Fig. 1: Naive bayes classifier accuracy on original and perturbed dataset

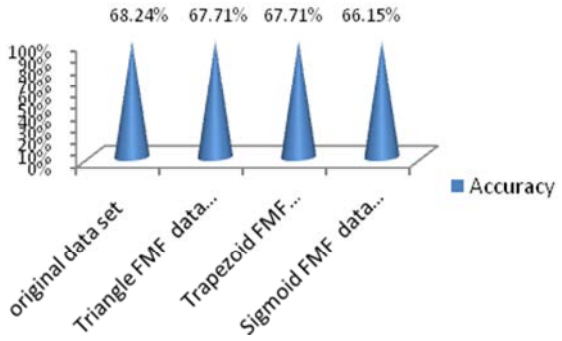

Fig. 2: K-NN classifier accuracy (%) on original and perturbed dataset

Performance criteria are determined in order to fit the learning task type. Criteria like accuracy, precision and recall are determined for classification tasks. Figure 1 and 2 shows the classification accuracy on both classifier

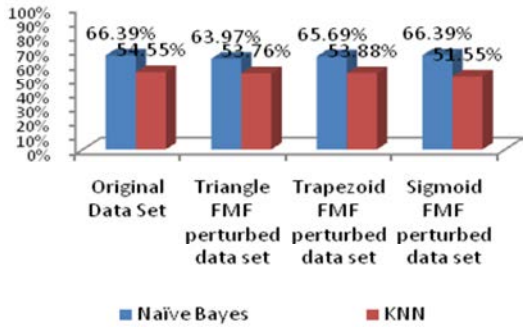

Fig. 3: Class precision (predicted tested-positive)

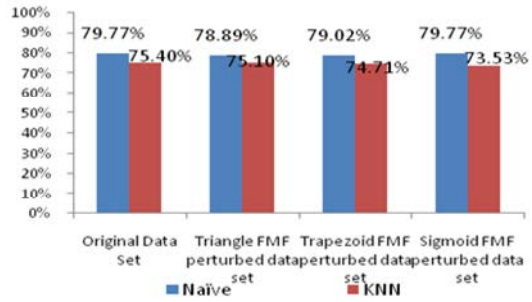

Fig. 4: Class precision (predicted as tested-negative)

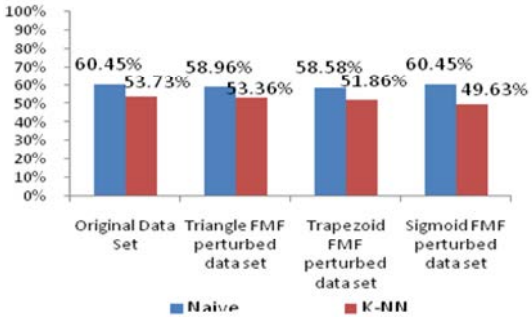

Fig. 5: Class recall (prediction tested-positive)

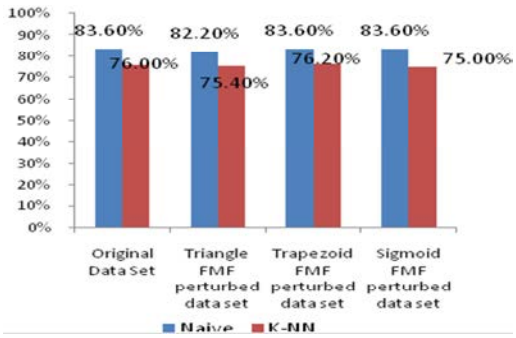

Fig. 6: Class recall (prediction tested-negative)

Naves Bayes and K-NN. The dataset contains 768 examples with 8 dimensions and a class label with values as tested-positive and tested-negative. Class precision and recall percentage are shown from Fig. 3-6.

CONCLUSION

In the proposed research we have implemented fuzzy member functions as a transformation technique for preserving the sensitive attributes. The interesting results show that accuracy of mining result is maintained in the transformation data set as that of original data set. Naïve Bayes and K-NN classification model was implemented using Rapid miner tool for diabetes data set. We measure the performance of our methodology using various performance criteria like accuracy, class precision and recall.

REFERENCES

Agrawal, D. and C.C. Aggarwal, 2001. On the design and quantification of privacy preserving data mining algorithms. Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, May 21-24, 2001, ACM, Santa Barbara, California, ISBN: 1-58113-361-8, pp: 247-255.

Bai, Y. and D. Wang, 2006. Fundamentals of Fuzzy Logic Control-Fuzzy Sets Fuzzy Rules and Defuzzifications. In: Advanced Fuzzy Logic Technologies in Industrial Applications. Ying, B., Z. Hanqi and D. Wang (Eds.). Springer, London, England, ISBN: 978-1-84628-468-7, pp: 17-36.

Cano, I., S. Ladra and V. Torra, 2010. Evaluation of information loss for privacy preserving data mining through comparison of fuzzy partitions. Proceedings of the 2010 IEEE International Conference on Fuzzy Systems (FUZZ), July 18-23, 2010, IEEE, Barcelona, Spain, ISBN: 978-1-4244-6919-2, pp: 1-8.

Du, W. and Z. Zhan, 2003. Using randomized response techniques for privacy-preserving data mining. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2003, Washington, DC, USA, pp: 505-510.

Kumar, P., K.I. Varma and A. Sureka, 2011. Fuzzy based clustering algorithm for privacy preserving data mining. Int. J. Bus. Inf. Syst., 7: 27-40.

Li, G. and Y. Wang, 2012. A privacy-preserving classification method based on singular value decomposition. Int. Arab J. Inf. Technol., 9: 529-534.

Li, N., T. Li and S. Venkatasubramanian, 2007. T-closeness: Privacy beyond k-anonymity and l-diversity. Proceedings of the IEEE 23rd International Conference on Data Engineering ICDE 2007, April 15-20, 2007, IEEE, Istanbul, Turkey, ISBN: 1-4244-0802-4, pp: 106-115.

Lindell, Y. and B. Pinkas, 2000. Privacy Preserving Data Mining. In: Advances in Cryptology-CRYPTO 2000. Mihir, B. (Ed.). Springer Berlin Heidelberg, Berlin, Germany, ISBN: 978-3-540-67907-3, pp: 36-54.

Lindell, Y. and B. Pinkas, 2009. Secure multiparty computation for privacy-preserving data mining. J. Privacy Confidentiality, 1: 59-98.

Liu, L., M. Kantarcioglu and B. Thuraisingham, 2008. The applicability of the perturbation based privacy preserving data mining for real-world data. Data Knowl. Eng., 65: 5-21.

Sweeney, L., 2002. k-Anonymity: A model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl. Based Syst., 10: 557-570.

Wang, K., B.C. Fung and G. Dong, 2005. Integrating Private Databases for Data Analysis In: Intelligence and Security Informatics. Paul, K., M. Gheorghe, F. Roberts, D.Z. Daniel and F.Y. Wang et al. (Eds.). Springer Berlin Heidelberg, Berlin, Germany, ISBN: 978-3-540-25999-2, pp: 171-182.