Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 6, July, 2021: 7862-7869

Research Article

Elective Ensembling Methods (Eem) On Classification Using Data Anonymization For Health Care Data

Dr P.Chandra Kanth[1], Dr K.V.Nagendra[2], Dr K. Sankar[3], Dr N. Krishna Kumar[4]

ABSTRACT

Enhancing the classification performance mostly used ensemble classification techniques. Research studies shows that classification through ensembling techniques shows the good classification concert in dynamic model representation in data anonymization approach. This paper we propose a elective ensembling methods based on the dynamic model data anonymization (EEM-DM-DA). This proposed technique enable to understanding the numerous trials met in privacy preserving data mining and also support us to discover best appropriate technique for numerous data modification techniques. **The proposed** anonymization technique can simultaneously disturb attributes presenting in the elected dataset. This can increase the diversity among different classifiers. Tentative stage of EEM-DM-DA is compared with the existing ensemble methods on maximum UCI data sets, where the SVM classification algorithm is used to train the ensemble classifiers. Proposed EEM-DM-DA technique results provides competitive solution for elective ensemble Method.

Keywords: Ensemble Methods, dynamic model data anonymization, classification, Support Vector machine, Privacy preserving

INTRODUCTION

Anonymized data is a type of information sanitization in which data anonymization tools encrypt or remove personally identifiable information from datasets for the purpose of preserving a data subject’s privacy. This reduces the risk of unintended disclosure during the transfer of information across boundaries and facilitates evaluation and analytics post-anonymization. Ensemble methods of classification allows in our research to develop a classifier that contains dynamic definitions of criteria on attributes this is called ensemble classification in our work. If we do not apply ensemble classification, the classifier will become obsolete (waste) for the upcoming new data in the data streams as the characteristics on the stream data change dynamically. Improving the classification performance widely used for Ensemble techniques

Recently, Medical services has shifted from treatment to prevention, there is a growing interest in smart healthcare that can provide users with healthcare services anywhere, at any time,

[1,2,3,4] Associate Prof, Dept. of Computer Science & Engineering Audisankara College of Engineering & Technology, Gudur
E-Mail Address: chandrakanthc4u@gmail.com

using information and communications technologies. With the development of the smart healthcare industry, there is a growing need for collecting large-scale personal health data for improving the smart healthcare services. Such health data can be a valuable asset to prevent critical disease. But there exist serious privacy problems if sensitive information of an individual user is leaked to outside users. So Privacy Preserving in Health Care is the biggest challenge in smart healthcare data storage environment.

Anonymization method aims at making the individual record be indistinguishable among a group records by utilizing techniques of generalization and suppression [9]. Privacy has become crucial in knowledge based applications. Proper integration of individual privacy is important for data processing operations. This privacy based data processing is vital for sectors like Healthcare, Pharmaceuticals, Research, and Security Service Providers, to call a couple of Different attributes in a data set may play different roles in either facilitating identification or facilitating sensitive information release.

An Effective data anonymization scheme using rotation was contributed to establish a good balance between the data utility and privacy [7]. This data anonymization scheme also focused on the scalability and efficiency under the release of datasets [9,10]. This data anonymization scheme is also determined to be highly resistant over the reconstruction attacks. This data anonymization scheme was determined to ensure maximum classification accuracy over the perturbed datasets with an effective rate of privacy preservation under the task of classification. Then, a Sensitive data anonymization with multiple iterative k-anonymity was proposed for the purpose of anonymization the attribute values in order to guard the data efficiently [11].

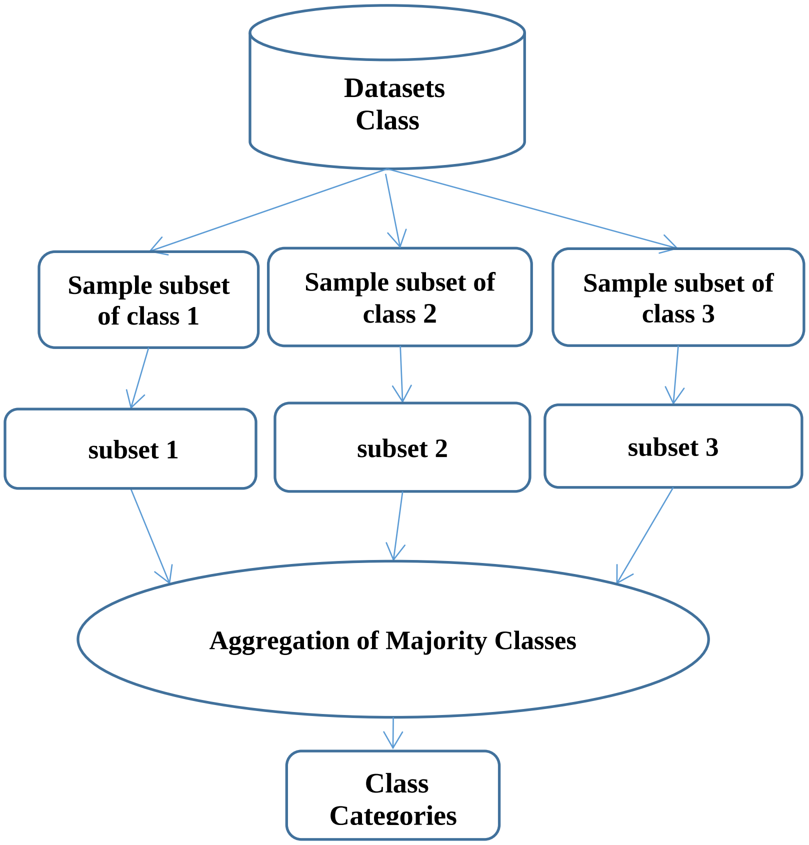

Fig : Aggregate Ensemble Classification

Ensemble models have been used extensively in credit scoring applications and other areas because they are considered to be more stable and, more importantly, predict better than single classifiers [1,2,4]. They are also known to reduce model bias and variance [5,6]. The objective of this article is to compare the predictive accuracy of four distinct datasets using two ensemble classifiers (Gradient boosting(GB)/Random Forest(RF)) and two single classifiers (Logistic regression(LR)/Neural Network(NN)) to determine if, in fact, ensemble models are always better[6,8]. My analysis did not look into optimizing any of these algorithms or feature engineering, which are the building blocks of arriving at a good predictive model. I also decided to base my analysis on these four algorithms because they are the most widely used methods.

DATA SETS DESCRIPTION

"For comparing the performances, experiments on four datasets from various real domains were conducted. These data sets are available on UCI machine learning repository" [3] and its details are described in Table 1. All these data sets contain private information which is to be protected from disclosure.

ELECTIVE ENSEMBLING METHODS BASED ON THE DYNAMIC MODEL DATA ANONYMIZATION (EEM-DM-DA*)*

In this proposed EEM-DM-DA Scheme, the data are clustered into multiple numbers of homogeneous clusters. Then, the method of data processing for the purpose of privacy preservation is enforced over each and every data chunks that are fixed in the size [10]. Further, the covariance matrix is determined for each of the clusters using the merits of characteristics derived from each cluster. Then, the generation of the covariance matrix is initiated for each of the corresponding geometric rotational clusters after the generation of covariance of matrices [12,14]. Once the covariance matrix is generated, the eigenvectors related to each individual covariance matrix are determined by partitioning them based on Equation (1)

$$C(M_i) = E(M_i) * \Delta(M_i) * E(M_i)^T \quad (1)$$

In this context, $E(M_i)$ corresponds to the eigenvectors estimated for each covariance matrix $C(M_i)$. The Eigenvectors form an axis system based on the property of orthogonality, since the determined covariance matrix is positive semi-definite in nature. Thus, the resultant matrix of Eigenvectors corresponding to each of the covariance matrix relates to a homogenous cluster that possess the characteristics of an orthogonal matrix since the rows and columns are orthonormal in characteristics. Hence, $E(M_i)$ is significant in preserving the association $E(M_i) * E(M_i)^T = E(M_i)^T * E(M_i) = I$, where $E(M_i)^T$ represents the transpose matrix of $E(M_i)$ and I as the identity matrix. This orthogonal property of $E(M_i)$ relates to each of the specific homogenous clusters that possess the complete set of properties involved in matrix rotation

***Confusion Matrix*:**

Category	Class 1	Class 2
Class 1 "Yes"	True Positive $FS_S \rightarrow FS_S$	False Negative $FS_S \rightarrow FN_S$
Class 2 "No"	False Positive $FS_{NS} \rightarrow FS_S$	True Negative $FS_{NS} \rightarrow FS_{NS}$

True Positive: Estimation of Selected Feature Set Considered correctly as selected Feature Set.

True Negative: Estimation of Non-Selected Feature Set Considered correctly as Non-Selected Feature Set.

False Positive: Estimation of Non-Selected Feature Set Considered incorrectly as Selected Feature Set.

False Negative: Estimation of Selected Feature Set Considered incorrectly as Non-Selected Feature Set.

Input: Random Sample

Output: Predicted Values

Algorithm:

Step 1: Import data set

Step 2: splitting dataset into Train & Test

Step 3: Features of sampling

Step 4: Training the SVM Classification Model

Step 5: Predicting the Results

Step 6: confusion Matrix & Accuracy

Step 7: Real values & Predicted Values

Step 8: Visualizing Results

EXPERIMENTAL RESULTS AND DISCUSSIONS

The experiments of the proposed ***EEM-DM-DA*** scheme is conducted using a PC with an Intel Core i7 CPU 3.40 GHz with 8.00 GB RAM for quantifying its predominance over the compared ensemble classification methods used for investigation. The experiments are conducted through 10-fold cross validation under which the input dataset is divided into mutually disjoint folds of 10 sets. One out of the 10-folds are used for the purposed for testing and the remaining 9 folds are utilized for the objective of training. This 10-fold cross validation-based experiments are iterated for 10 times and the average results of the complete 10 folds are determined as results and documented. The experiments of the proposed ***EEM-DM-DA*** is conducted using WEKA version

3.7.11, which is the open source platform machine learning toolkit and known for its compatibility over the other machine learning mechanisms implemented in Java. This experiment of the proposed ***EEM-DM-DA*** is conducted using the datasets of *Spambase, Diabetes, and Transfusion*: Identification Dataset with possible default factors that are unique to the design and implementation of the ensemble classification methods.

Table 1: Classification Accuracy determined for the proposed *EEM-DM-DA* scheme under different feature sets

Dataset	No.of Instances	No.of Classes
SpamBase	4601	2
Diabetes	768	2
Transfusion	748	2

Table 2 : Accuracy

Dataset	Ensemble Classifier	Anonymous SVM ensemble
SpamBase	90.99	2
Diabetes	76.82	2
Transfusion	74.8	2

Table 3: Training Time Classifier

Dataset	Ensemble Classifier	Anonymous SVM ensemble
SpamBase	13.35	17.28
Diabetes	1.51	2.09
Transfusion	3.39	4.24

Table 2 highlights the F-Measure value and second maximum F-Measure value under the *Spambase, diabetes, Transfusion*: Data sets determined under the integration of .and .feature datasets, respectively. It is inferred that the optimal F-Measure value of the proposed ***EEM-DM-DA*** scheme under Spambase, diabetes, Transfusion: Gene Identification data set is visualized at 0.982, when the features .are integrated with L2NO-ELM of ensemble classification scheme. It is also confirmed that the second optimal predictive performance of the proposed ***EEM-DM-DA*** scheme under Spambase, diabetes, Transfusion data set is visualized at 0.988, when the features .are integrated with L2NO-ELM of ensemble classification scheme. Likewise, the optimal predictive performance of the proposed ***EEM-DM-DA*** scheme under *Spambase, diabetes, Transfusion* set is visualized at 0.981, when the features .are integrated with L2NO-ELM of ensemble classification scheme.

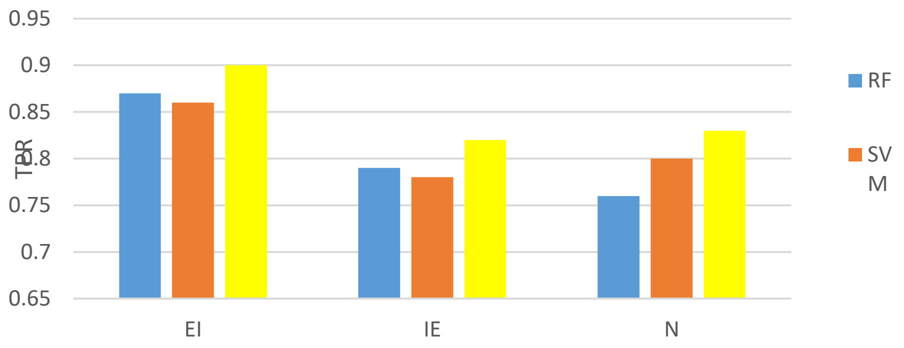

Fig 1 : True Positive Rate

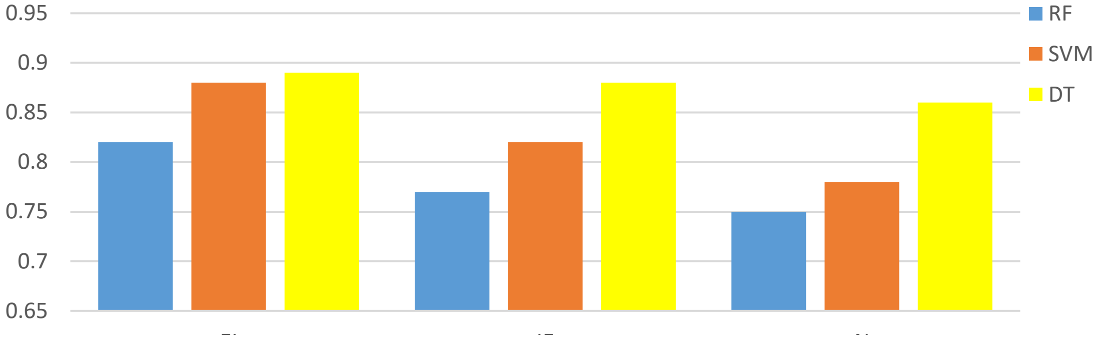

Fig 2 : Accuracy on Feature sets

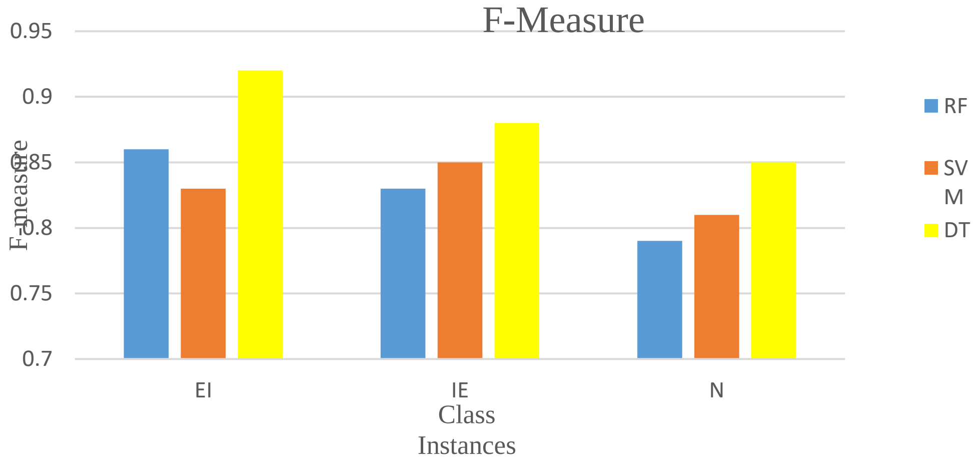

Figure 3: Proposed *EEM-DM-DA* scheme –Recall value under different datasets used for ensemble classification

Similarly, the recall of the proposed ***EEM-DM-DA*** ensemble classification scheme was confirmed to be excellent with all the three KELM, RELM and L2NO-ELM base classifiers compared to the Adaboost and Bagging ensemble classification methods under the investigation with SJCS and Spambase, Diabetes,Transfusion datasets. The recall of the proposed ***EEM-DM-DA*** ensemble classification scheme on an average was confirmed to be superior by 11%, 13% and 16%, remarkable to the Adaboost and Bagging ensemble classification with KELM, RELM and L2NO-ELM base classifiers under the investigation with SJCS and Spambase, Diabetes, Transfusion datasets[13,15].

As ensemble classifiers produce more accurate results, the approach is quite suitable for Privacy-Preserving Classification of Homogeneously Distributed Data and the same is proved experimentally. However, few conclusions about data at other sites can be easily derived from the classifiers released by those sites and privacy can be breached. Our proposed approach of k=anonymous SVM classifier ensemble overcomes this disadvantage and preserves privacy to a greater extent. Also, unlike traditional privacy protection techniques such as data swapping and adding noise, information preserved using k-anonymization remains truthful[16].

CONCLUSION

The proposed ***EEM-DM-DA*** scheme integrated data anonymization and ensemble classification method suitable for potential determination of medical data sets. This ***EEM-DM-DA*** scheme not only prevents the leakage of sensitive data but also concentrates on the task of classification without any alteration in the dataset by deriving the benefits of Effective Ensemble Method based Data Anonymization. In addition, the mean absolute error and standard deviation of the proposed ***EEM-DM-DA*** scheme was estimated to be considerably reduced by 12% and 14% independent to the utilized SJCS and Spam base, Diabetes, Transfusion datasets. In the near future, it is also planned to formulate an improved data Anonymization ensemble classification scheme for studying its suitability and applicability in the process of identifying medical datasets.

REFERENCES

[1] Baesens, Bart, et al. "Benchmarking state-of-the-art classification algorithms for credit scoring." *Journal of the operational research society* 54.6 (2003): 627-635.

[2] Lessmann, Stefan, et al. "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research." *European Journal of Operational Research* 247.1 (2015): 124-136.

[3] Demšar, Janez. "Statistical comparisons of classifiers over multiple data sets." *The Journal of Machine Learning Research* 7 (2006): 1-30.

[4] Kim, Myoung-Jong, Sung-Hwan Min, and Ingoo Han. "An evolutionary approach to the combination of multiple classifiers to predict a stock price index." *Expert Systems with Applications* 31.2 (2006): 241-247.

[5] Tsai, Chih-Fong, and Yu-Chieh Hsiao. "Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches." *Decision Support Systems* 50.1 (2010): 258-269.

[6] Godase, Abhijeet, and Vahida Attar. "Classifier ensemble for imbalanced data stream classification." *Proceedings of the CUBE International Information Technology Conference*. 2012.

[7] Kotecha, Radhika, Vijay Ukani, and Sanjay Garg. "An empirical analysis of multiclass classification techniques in data mining." *2011 Nirma University International Conference on Engineering*. IEEE, 2011.

[8] Han, Jiawei, and Micheline Kamber. "Pei. Data mining concepts and techniques." *MK* (2011).

[9] Matwin, Stan. "Privacy-preserving data mining techniques: survey and challenges." *Discrimination and Privacy in the Information Society*. Springer, Berlin, Heidelberg, 2013. 209-221

[10] Shanthi, A. S., and M. Karthikeyan. "A review on privacy preserving data mining." *2012 IEEE International Conference on Computational Intelligence and Computing Research*. IEEE, 2012.

[11] Aggarwal, Charu C., and S. Yu Philip. "A general survey of privacy-preserving data mining models and algorithms." Privacy-preserving data mining. Springer, Boston, MA, 2008. 11-52

[12] Saranya, K., K. Premalatha, and S. S. Rajasekar. "A survey on privacy preserving data mining." *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*. IEEE, 2015.

[13] Wang, Ke, and Benjamin CM Fung. "Anonymizing sequential releases." *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006

[14] Fung, Benjamin CM, et al. "Privacy-preserving data publishing: A survey of recent developments." *ACM Computing Surveys (Csur)* 42.4 (2010): 1-53.

[15] Sweeney, Latanya. "Achieving k-anonymity privacy protection using generalization and suppression." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002): 571-588.

[16] Malik, Majid Bashir, M. Asger Ghazi, and Rashid Ali. "Privacy preserving data mining techniques: current scenario and future prospects." *2012 Third International Conference on Computer and Communication Technology*. IEEE, 2012

.