ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **9** (2020), no.9, 6623–6629
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.9.19 Spec. Issue on CAMM-2020

A NOVEL MULTI-COMBINED METHOD FOR HANDLING MEDICAL DATASET WITH IMBALANCED CLASSES PROBLEM

CHING-HSUE CHENG[1] AND YUN-CHUN WANG [2]

ABSTRACT. The imbalanced datasets are quite common in medical research, which has impacted on classification accuracy. Diseases are always life-threatening, and any small bias may harm the patient; hence the medical field has a relatively low tolerance of misjudgment. In order to accurately identify diseased and non-disease individuals, this study proposes a novel multi-combined method to deal with the medical dataset with imbalanced classes problem. The proposed method applies the synthetic minority oversampling technique (SMOTE) to deal with the minority class and then used the particle swarm optimization (PSO) to select important attributes. Finally, the five cost-sensitive classifiers based on the MetaCost method are employed to classify the medical datasets. The experimental results indicate that the random forest can achieve the best accuracy. The best false positive rate is the LibSVM classifier, and the best false-negative rate occurs in the LibSVM classifier combined with the PSO attribute selection, SMOTE, and MetaCost. At last, the results can assist physicians in identifying diseases, and then the patients can be early diagnosed and treated for increasing their chances of survival.

1. INTRODUCTION

In the medical field, it is almost impossible to get complete balanced data in the real world. When a class has a large of samples, and another class has only

[1]*corresponding author*

2010 *Mathematics Subject Classification.* 68W99,92C42.

Key words and phrases. Imbalanced classes, Synthetic minority oversampling technique (SMOTE), Particle swarm optimization (PSO), MetaCost.

a few samples, it means that the dataset is imbalanced[1]. The main source of the collected medical data depends on the patient's medical information, patients diagnosed with the disease are much less than those without the disease, and it leads to the medical datasets with imbalanced classes. In addition, the different prevalence of the disease is also a factor in the data with imbalanced classes. In past research, learning classifiers are mostly assumed to treat the balanced classes during development, but the assumptions are the opposite of most medical datasets. The existence of imbalanced datasets has always plagued the medical field, such as the imbalanced classes always cause the analysis results to tend to most class, which makes the high accuracy[2]; on the contrary, the accuracy of the minority class is low. The actual medical environment has a greater concern on the minority class, which is the prevalence of the disease. The main reason is that minority class usually has higher meaning[3].

In order to overcome the problems caused by the imbalanced classes in classification, the researchers adopted the resampling method to change their sample size and cost-sensitive learning (CSL) considering the different misclassification costs of each sample[4]. The resampling method is mainly divided into oversampling[5] for sample replication in minority class and undersampling for sample reduction in the majority class. The common oversampling method currently used in an imbalanced dataset is the synthetic minority oversampling technique proposed (SMOTE) by Chawla[5], SMOTE creates synthetic samples on minority class to balance the dataset. Particle Swarm Optimization (PSO) algorithm is based on the concept of the group. The fundamental idea of the PSO algorithm is simulating the social behavior of birds and fish.

The cost of misjudgment is very high in the diagnostic process, especially false negative judgments. False-negative judgment is an unhealthy person to be classified as a healthy person. This misjudgment will lead to delayed treatment, a patient's deteriorating condition, and even death. CSL can deal with imbalanced problems by calculating the cost of misclassification [6]. Medical workers hope to use data mining to solve the mystery in the medical field, explore more useful information, and improve the physician-patient relationship.

2. PROPOSED METHOD

2.1. **Experimental Material.** The Breast Cancer Wisconsin (Diagnostic) Dataset was obtained from the UCI that is a public dataset website[7]. The data set will use fine-needle aspiration (FNA) to diagnose breast lumps and establish different attributes according to the characteristics of the cells. After preprocessing, there were a total of 30 attributes, a class at-tribute, and 569 records in the dataset. The class attribute is a binary classification problem, and the content of the class attribute is a malignant or benign tumor, respectively. The number of malignant tumor records was 212, and the number of benign tumor records was 357.

2.2. **Proposed Method.** An imbalanced dataset represents the class imbalance of the dataset, and the number of data from one class is greater than that of another class. This study proposed a novel multi-combined method for overcoming the imbalanced medical dataset in classification problems. The proposed method applied the different combinations of the synthetic minority oversampling technology (SMOTE)[5], the PSO attribute selection method[8], and the cost-sensitive MetaCost[6] to handle the imbalanced medical dataset. We expect that the proposed method can improve the classification accuracy of an imbalanced dataset. The proposed procedure is shown in Figure 1, and the proposed procedure are briefly described as follows:

(1) Public Datasets: The breast cancer is collected from the UCI Machine Learning Repository (UCI) that is an online public data collection platform[7].
(2) Data Preprocessing: Breast cancer dataset is a binary classification problem. The step deletes the missing values and normalizing the data.
(3) Synthetic Minority Oversampling Technique (SMOTE): Oversampling mainly generates more samples by copying action to achieve sample balance [9]. In order to improve the problems, Chawla published Synthetic Minority Oversampling Technology (SMOTE) in 2002[5]. The method uses synthetic samples to deal with imbalance data. The main purpose of this step is to make the imbalanced dataset achieve the balance of the number of class data by the SMOTE[5] or not to perform this action.
(4) Attribute selection: The attribute selection is used to reduce the dimension of the attribute subset and keep the most important attribute. The

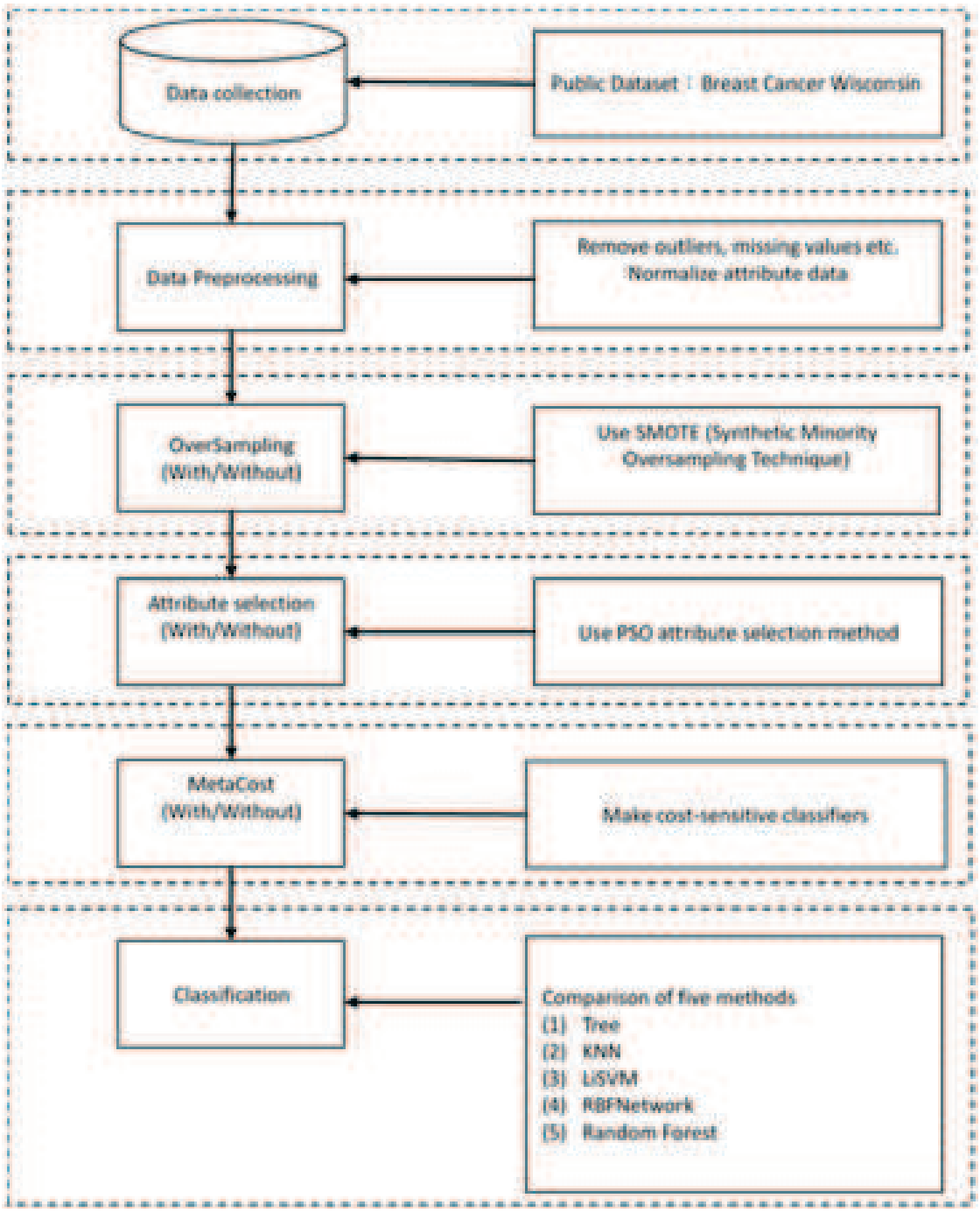

FIGURE 1. Proposed procedure

PSO algorithm was proposed by Kennedy and Eberhart [8]. The main idea is to simulate the behavior of birds search for food and find the best solution for continuous search. The research uses the PSO algorithm-based attribute selection method to perform attribute selection actions, or not.

(5) MetaCost: Type 1 error means that you are healthy, but the diagnosis result shows that you are sick. Type 2 error means that you are sick, but the diagnosis result shows that you are healthy. The cost of "type 2 error" is higher than the cost of "type 1 error" in the medical dataset, so the study tries to use MetaCost to improve the classification accuracy. In order to compare whether the method is effective, the step will also not execute MetaCost at the same time.

(6) Classification: The study uses five classification methods (Decision tree[10], K-nearest neighbors (KNN)[11], LibSVM[12], Radial Basis Function (RBF) Network[13], Random Forest[14]) for dataset to explore and verify the best classification method for the imbalanced medical dataset.

3. EXPERIMENT AND RESULT

Accuracy: The study uses five classification algorithms to analyze the dataset in order to find out the best hybrid classification method for the medical dataset. Also, the study compares the accuracy of different classification algorithms. The results are shown in Table 1. The best accuracy of breast cancer dataset is 96.20, which occurs in a random forest. Therefore, the study concluded that random forest could improve the classification accuracy of imbalanced medical datasets. It is the best algorithm for binary classification of imbalanced datasets.

False-positive rate (Type 1 error): The definition of false-positive rate (FPR) in the medical environment is the patient is healthy, but the diagnosis result shows that the patient has an illness. The lower value means lesser misjudgment rate and higher quality. The results are shown in Table 1. The breast cancer dataset has the three best FPR with a value of 0.0000. First, the best FPR is LibSVM; next, it is LibSVM with MetaCost. Finally, it is LibSVM with PSO attribute selection and MetaCost.

False-negative rate (Type 2 error): The definition of false-negative rate (FNR) in the medical environment is the patient is an illness, but the diagnosis result shows the patient is healthy. The lower value means lesser misjudgment

Table -1 The results of the breast cancer Wisconsin (diagnostic) dataset

		Tree	KNN	LibSVM	RBF Network	Random Forest
No SMOTE	Accuracy	93.50(1.82)	95.54 (1.30)	62.74 (0.11)	94.15 (1.83)	96.20(1.24)
	Type I	0.0494	0.0320	0.0000*	0.0417	0.0225
	Type II	0.0913	0.0658	1.0000	0.0868	0.0640
SMOTE	Accuracy	93.40(1.46)	95.14(1.08)	64.48(2.41)	94.42(1.78)	96.13(1.25)
	Type I	0.0618	0.0425	0.7088	0.0466	0.0372
	Type II	0.0700	0.0546	0.0048	0.0648	0.0402
No SMOTE + MetaCost	Accuracy	93.44(1.61)	95.58(1.38)	62.74(0.11)	93.23(1.64)	95.88(1.30)
	Type I	0.0492	0.0305	0.0000*	0.0520	0.0239
	Type II	0.0932	0.0672	1.0000	0.0941	0.0702
SMOTE + MetaCost	Accuracy	93.54(1.38)	95.20(1.11)	57.46(4.68)	93.67(1.56)	95.99(1.16)
	Type I	0.0647	0.0408	0.8335	0.0600	0.0376
	Type II	0.0644	0.0552	0.0208	0.0666	0.0425
PSO + No SMOTE	Accuracy	93.73(1.74)	95.48(1.36)	63.06(0.51)	94.64(1.54)	95.78(1.30)
	Type I	0.0474	0.0334	0.0018	0.0353	0.0266
	Type II	0.0886	0.0649	0.9883	0.0845	0.0686
PSO+SMOTE	Accuracy	93.80(1.46)	94.96(1.26)	67.42(2.57)	95.65(1.16)	96.10(1.25)
	Type I	0.0580	0.0419	0.6438	0.0356	0.0382
	Type II	0.0659	0.0588	0.0108	0.0513	0.0397
PSO + No SMOTE + MetaCost	Accuracy	93.81 (1.61)	95.38(1.41)	63.02(0.34)	93.82(1.49)	95.60(1.38)
	Type I	0.0469	0.0342	0.0000*	0.0461	0.0277
	Type II	0.0872	0.0655	0.9924	0.0884	0.0715
PSO+ SMOTE+ MetaCost	Accuracy	93.67 (1.43)	95.01 (1.39)	60.59 (4.16)	94.90 (1.20)	95.87 (1.20)
	Type I	0.0599	0.0421	0.7878	0.0477	0.0396
	Type II	0.0667	0.0576	0.0039	0.0543	0.0429

rate and higher quality. The results are shown in Table 1. The best FNR of the breast cancer dataset is 0.0039, which occurs when LibSVM is used with PSO attribute selection, SMOTE, and MetaCost.

4. CONCLUSION

This study has proposed a novel multi-combined method for handling the imbalanced medical dataset. The proposed method combined SMOTE with the PSO at-tribute selection method and MetaCost. It can solve the imbalanced data problems in medical diagnosis and reduce the misjudgment rate. The results can

assist doctors in diagnosing diseases so that the patients can be diagnosed and treated early to in-crease the chance of survival. The method of this study brings different directions of thinking to the classification of imbalanced data problems in medicine. The experiment only employed Wisconsin (diagnostic) breast cancer dataset from the UCI learning repository; hence we will be necessary to expand scope of experiments in the future in order to make the classification more accurate and reliable.

REFERENCES

[1] N. JAPKOWICZ: *Learning from imbalanced data sets: a comparison of various strategies*, AAAI workshop on learning from imbalanced data sets, **68** (2000), 10-15.

[2] X. WAN, J. LIU, W.K. CHEUNG, T. TONG: *Learning to improve medical decision making from imbalanced data without a priori cost*, BMC medical informatics and decision making, **14**, art.no. 111, (2014),

[3] P. YANG, L. XU, B.B. ZHOU, Z. ZHANG, A.Y. ZOMAYA: *A particle swarm based hybrid system for imbalanced medical data sampling*, BMC genomics, **10**(S3) (2009). https://doi.org/10.1186/1471-2164-10-S3-S34

[4] C.X. LING, V.S. SHENG: *Cost-sensitive learning and the class imbalance problem*, Encyclopedia of Machine Learning Springer, 2008, 231-235.

[5] N.V. CHAWLA, K.W. BOWYER, L. O. HALL, W. P. KEGELMEYER: *SMOTE: synthetic minority over-sampling technique*, Journal of artificial intelligence research, **16** (2002), 321-357.

[6] P. DOMINGOS: *Metacost: A general method for making classifiers cost-sensitive*, in Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999, 155-164.

[7] C. BLAKE: *UCI repository of machine learning databases*, 1998. http://www.ics.uci.edu/ mlearn/MLRepository.html

[8] J. KENNEDY, R. EBERHART: *Particle swarm optimization*, Proceedings of ICNN'95-International Conference on Neural Networks, **4** (1995), 1942-1948.

[9] L. ABDI, S. HASHEMI: *To combat multi-class imbalanced problems by means of over-sampling techniques*, IEEE transactions on Knowledge and Data Engineering, **28**(1) (2015), 238-251.

[10] J. R. QUINLAN: *Induction of decision trees*, Machine learning, **1**(1) (1986), 81-106.

[1,2]DEPARTMENT OF INFORMATION MANAGEMENT, NATIONAL YUNLIN UNIVERSITY OF SCIENCE AND TECHNOLOGY, YUNLIN, TAIWAN