Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data

Min Zeng[1,2], Beiji Zou[1,2], Faran Wei[1,2], Xiyao Liu[1,2], Lei Wang[1,2]*

[1] School of Information Science and Engineering, Central South University, Changsha 410083, People's Republic of China

[2] Mobile Health Ministry of Education-China Mobile Joint Laboratory, Central South University, Changsha 410083, People's Republic of China

Corresponding author: Lei Wang

E-mail address: zengmin1990@163.com (Min Zeng), bjzou@csu.edu.cn (Beiji Zou), franwee@163.com(Faran Wei), lxyzoewx@csu (Xiyao Liu), wanglei@csu.edu.cn (Lei Wang)

***Abstract*—Diabetes, vertebral column pathologies and Parkinson's disease are three common diseases which have high prevalence and brought great trouble and pain to billions of patients. Computer aided diagnosis can support decision making of physicians. However, imbalanced nature of data sets hampered the mining of medical resources. In this study, we proposed a powerful preprocessing method by combining Synthetic Minority Oversampling Technique (SMOTE) with Tomek links technique and then is applied to the imbalanced medical data sets of the three diseases. By using 8 classifiers, we compared the experimental results with those of using only SMOTE technique to evaluate the effectiveness of this method. The results show that the method of SMOTE combined with Tomek links technique is much superior compared with that of using only SMOTE. The performances are evidently better, with 31, 27, 30 out of a total of 32 evaluation metrics are improved for diabetes, Parkinson's disease, and vertebral column, respectively.**

Keywords-imbalanced medical data; SMOTE; Tomek links; diabetes; vertebral column pathologies; Parkinson's disease

I. INTRODUCTION

Diabetes, vertebral column pathologies and Parkinson's disease brought great trouble and pain to a great number of patients. These diseases are causing serious harm to people's health and living quality and are bringing a heavy burden to the family and our society. Accurate diagnosis of these diseases is vital to better improve the patients' quality of life.

Machine learning techniques are developing rapidly and are used for the detection and prediction of these common diseases [1-3]. However, medical data are normally collected over a long period of time and thus we often encounter imbalanced data sets. Data imbalance means that there is not an even distribution of samples between the different classes. The imbalanced nature of medical data often hampered the mining of medical resources. Although great progress has been achieved in machine learning, it remains a challenging task to construct efficient algorithms that learn from imbalanced data. Several methods have been proposed to deal with imbalanced data and data sampling method is one of the most frequently used preprocessing technique [4].

Synthetic Minority Oversampling Technique (SMOTE), as a effective data sampling algorithm, has been applied to analyze imbalanced medical data [5-11]. The results in Refs. [5-8] showed that improved performance was obtained by using SMOTE method compared with those of using raw data sets. Other work [9-11] demonstrated that SMOTE method is superior or at least comparable to conventional random sampling techniques. These researches improved the performance of classifiers by using the SMOTE method. We suggest that the performance can be further improved by combining SMOTE with data cleaning techniques.

In this work, we proposed a preprocessing technique of combining SMOTE with Tomek links to address the problem of imbalanced medical data. Tomek links, as a data cleaning technique, were effectively applied to remove the samples, which generated by the SMOTE method, near the boundary of classification. By combining Tomek links technique, the boundary between different classes can be easily identified. To the best of our knowledge, such a combined technique has never been utilized in the treatment of medical data. Our results show that much better performance is obtained by using the SMOTE with Tomek links than using only SMOTE. Based on previous researches [5] and our experimental results we conclude that the combined method using SMOTE as a data sampling technique combining Tomek links as a data cleaning technique is a powerful preprocessing algorithm to address imbalanced medical classification data.

II. METHODOLOGY

Synthetic Minority Over-sampling Technique (SMOTE), a kind of oversampling technique, was proposed by Chawla et al. [12]. The key idea is to find K-nearest neighbors which defined as the K elements belong to the minority class for each minority class sample x_i and then randomly selects one $\hat{x}_i$ of these neighbors. By using

978-1-4673-7755-3/16/$31.00 ©2016 IEEE

interpolation theory we can generate a new sample x_{new} as follows:

$$x_{new} = x_i + (\hat{x}_i - x_i) \times \delta \quad (1)$$

where δ is a random value between 0 and 1.

Tomek links, a data cleaning technique, was proposed by Ivan Tomek [13]. A Tomek link is defined as a pair of minimally Euclidian distanced neighbors (x_i, x_j) with x_i belonging to the minority class and x_j to the majority class. Let $d(x_i, x_j)$ denote the Euclidian distance between x_i and x_j. If there is no sample x_k satisfies the following condition: $d(x_i, x_k) < d(x_i, x_j)$ or $d(x_j, x_k) < d(x_i, x_j)$ then the pair of (x_i, x_j), is a Tomek link.

Choosing appropriate evaluation metrics in medical imbalanced data classification is very important. The classification performance is evaluated using metrics of accuracy (Acc), F-measure, G-mean, and the area under the receiver operating characteristic curve (AUC).

Table 1. Data sets characteristics used in this work

Name	Total instances	Number of minority	Number of majority	Attributes
Diabetes	768	268	500	9
Parkinson's disease	195	48	147	23
vertebral column	310	100	210	7

III. EXPERIMENT

In this work, we are concerned with three medical data sets of common diseases including diabetes, Parkinson's disease and vertebral column. The experimental data sets are publicly available at "The Data Mining Repository of University of California Irvine (UCI)". After dealing with the minority and majority classes for these data sets, a brief description of such data sets is summarized in table 1.

By using these medical data sets, we carried out experiments of computer aided diagnosis to these common diseases. By combining SMOTE with Tomek links technique in the preprocessing step to address the raw data, we obtained balanced data sets with clear boundary between the two classes of majority and minority. After preprocessing the imbalanced data, we employed 8 classifiers to train the data sets. These classifiers include instance-based learner (K*) , decision tree (C4.5) , ensemble algorithm (AdaBoost) , Bayesian network (BN), RBF Network, Logistic regression (LR), SVM and Logistic model trees (LMT). In order to reduce any bias due to a lucky or unlucky split, we perform 10-fold cross-validation. To demonstrate the superiority of performance in our proposed method, we investigated the evaluation metrics by using only SMOTE method and without any preprocessing procedure.

IV. RESULTS AND DISCUSSION

Tables 2-4 present the classification results. The results shown in cases 1 and 2 represent the evaluation metrics obtained by using the preprocessing method of SMOTE combined with Tomek links technique and with only SMOTE, respectively. The evaluation metrics in case 3 denotes the performance by using the raw data without any preprocessing procedure. Bold values in this table represent the maximal evaluation metrics for the respective performance of F-measure, G-mean, AUC, and accuracy. From the inspection of tables 2-4, we can see that F-measure, G-mean, AUC, and accuracy are evidently improved in cases 1 and 2 than in case 3 for all three common diseases, which showed the effects of preprocessing step.

Table 2. The performances (F-measure, G-mean, AUC, and accuracy) of 8 classifiers for diabetes. The evaluation metrics obtained by preprocessing with SMOTE combined with Tomek links, with only SMOTE and without any preprocessing procedure are denoted by cases 1, 2, and 3, respectively.

	F-Measure			G-mean			AUC			Acc (%)		
	1	2	3	1	2	3	1	2	3	1	2	3
SVM	**0.773**	0.753	0.763	**0.775**	0.753	0.762	**0.768**	0.753	0.720	**77.459**	75.290	77.344
BN	**0.783**	0.768	0.742	**0.780**	0.771	0.713	**0.864**	0.851	0.806	**78.232**	76.931	74.349
AdaBoost	**0.775**	0.754	0.738	**0.775**	0.754	0.716	**0.848**	0.821	0.801	**77.569**	75.386	74.349
K*	**0.804**	0.802	0.683	0.801	**0.806**	0.650	**0.886**	0.873	0.714	**80.332**	80.309	69.141
C4.5	**0.775**	0.758	0.736	**0.772**	0.759	0.707	**0.813**	0.793	0.751	**77.459**	75.869	73.828
RBF Network	**0.755**	0.726	0.746	**0.754**	0.726	0.732	**0.832**	0.804	0.783	**75.580**	72.587	75.391
LR	**0.781**	0.754	0.765	**0.782**	0.754	0.729	**0.864**	0.838	0.832	**78.232**	75.386	77.214
LMT	**0.792**	0.752	0.766	**0.789**	0.753	0.760	**0.859**	0.831	0.831	**79.116**	75.290	77.474

Table 3. The performances (F-measure, G-mean, AUC, and accuracy) of 8 classifiers with 3 cases (the same as Table 3) for Parkinson's disease

	F-Measure			G-mean			AUC			Acc (%)		
	1	2	3	1	2	3	1	2	3	1	2	3
SVM	0.851	0.841	**0.856**	0.865	0.849	**0.908**	**0.849**	0.843	0.747	85.252	84.192	**87.180**
BN	**0.833**	0.814	0.810	**0.840**	0.817	0.723	**0.935**	0.918	0.871	**83.453**	81.443	80.000
AdaBoost	**0.910**	0.828	0.846	**0.911**	0.827	0.808	**0.952**	0.917	0.889	**91.007**	82.818	85.128
K*	**0.935**	0.931	0.900	0.935	**0.937**	0.847	**0.991**	0.987	0.968	**93.525**	93.127	89.744
C4.5	**0.884**	0.880	0.804	**0.888**	0.880	0.726	**0.889**	0.883	0.769	**88.489**	87.973	80.513
RBF Network	0.829	**0.834**	0.830	**0.841**	**0.841**	0.806	**0.883**	0.878	0.862	83.094	83.505	**84.103**
LR	0.831	0.828	**0.866**	**0.831**	0.828	0.818	0.924	**0.927**	0.883	83.094	82.818	**86.667**
LMT	**0.899**	0.893	0.856	**0.900**	0.894	0.825	**0.962**	0.951	0.835	**89.928**	89.347	86.15

Table 4. The performances (F-measure, G-mean, AUC, and accuracy) of 8 classifiers with 3 cases (the same as Table 3) for vertebral column

	F-Measure			G-mean			AUC			Acc (%)		
	1	2	3	1	2	3	1	2	3	1	2	3
SVM	**0.812**	0.811	0.770	**0.822**	0.818	0.786	**0.815**	0.814	0.704	**81.390**	81.220	78.710
BN	**0.825**	0.809	0.772	**0.833**	0.818	0.734	**0.874**	0.861	0.853	**82.630**	80.976	76.452
AdaBoost	**0.835**	0.830	0.803	**0.846**	0.842	0.765	0.912	**0.917**	0.894	**83.623**	83.171	80.000
K*	**0.867**	0.854	0.835	**0.883**	0.873	0.800	**0.976**	0.960	0.896	**86.849**	85.610	83.226
C4.5	**0.871**	0.824	0.812	**0.871**	0.824	0.795	**0.890**	0.858	0.838	**87.097**	82.439	81.613
RBF Network	**0.853**	0.846	0.805	**0.856**	0.849	0.769	**0.904**	0.895	0.871	**85.360**	84.634	80.323
LR	**0.861**	0.859	0.855	**0.862**	0.859	0.830	**0.944**	0.941	0.934	**86.104**	85.854	85.484
LMT	**0.866**	0.851	0.856	**0.868**	0.853	0.828	0.926	**0.935**	0.929	**86.601**	85.122	85.484

By using combined SMOTE and Tomek links technique, the evaluation metrics are further steadily improved than those of using only SMOTE. For diabetes and vertebral column, nearly all evaluation metrics of 8 classifiers are better than those obtained by using only SMOTE. Only three of them are slightly lower, with the G-mean of K* lowered from 0.806 to 0.801 for diabetes and AUC of AdaBoost and Logistic model trees from 0.917 to 0.912 and from 0.935 to 0.926, respectively, for vertebral column. For Parkinson's disease, there is also a dramatic enhancement of the evaluation metrics although the effect is not as good as that of diabetes and vertebral column. Compared with results with only SMOTE, 27 out of a total of 32 evaluation metrics are improved for Parkinson's disease.

Let us discuss some details on the improving evaluation metrics of the combined technique in tables 2-4. The largest improvement is from G-mean of Parkinson's disease with an increase from 0.827 (with only SMOTE) to 0.911 (combined technique) for the AdaBoost classifier. For diabetes, the largest enhancement is contributed by F-measure, increasing from 0.752 to 0.792 with a classifier of Logistic model trees. Both F-measure and G-mean increase 0.047 with a classifier of C4.5 for vertebral column. Moreover, the evidently improved evaluation metrics originate from different classifiers. Technique of combining the SMOTE and Tomek links have superior performance in a variety of classifiers, which means that the combined technique applies for different environment including the data and classifiers. The average performance of using combined SMOTE and Tomek links technique is better than that of using only SMOTE and the standard deviation of is smaller than that of using only SMOTE. Thus we concluded that the performance of combining SMOTE and Tomek links technique is much better than the performance of using only SMOTE.

V. CONCLUSION

Effective predictions of common diseases are fulfilled by combining SMOTE with Tomek links technique for the preprocessing the imbalanced medical data. By using this combined technique, we preprocessed the imbalanced data of common diseases for diabetes, vertebral column pathologies and Parkinson's disease to arrive at a relative balance. To evaluate the algorithm of combined SMOTE and Tomek links, we compared the classification performance with that of using only SMOTE technique by using 8 classifiers. Experiments are carried out to show the effects of the combined preprocessing technique. We compared the classification performances including F-measure, G-mean, AUC, and accuracy. The results

showed that evaluation metrics with combined SMOTE and Tomek links are much more improved than those of without any preprocessing procedure. Moreover, the performances are evidently better than those of with only SMOTE, which is considered to be a good preprocessing method in some recent literature [5]. Compared with results of only SMOTE, 31, 27, 30 out of a total of 32 evaluation metrics are improved for diabetes, Parkinson's disease, and vertebral column, respectively. Evidently improved evaluation metrics coming from different classifiers indicated that the combination of SMOTE and Tomek links technique can be applied to different environment including the data and classifiers.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grants No. 61173122, Key Project of Natural Science Foundation of Hunan Province of China (12JJ2038) and Natural Science Foundation of Hunan Province of China (09JJ6102).

REFERENCES

[1] E. Salzsieder, L. Vogt, K. D. Kohnert, P. Heinke, P. Augstein, Model-based Decision support in Diabetes Care, Computer methods and programs in biomedicine 102 (2011) 206–218.

[2] F. Calle-Alonso, C. J. Pérez, J. P. Arias-Nicolás, J. Martín, Computer-aided diagnosis system: A Bayesian hybrid classification method, Computer methods and programs in biomedicine 112 (2013) 104-113.

[3] M. Hariharan, K. Polat, R. Sindhu, A new hybrid intelligent system for accurate detection of Parkinson's disease, Computer methods and programs in biomedicine 113 (2014) 904–913.

[4] H. He, E. A. Garcia, Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering 21 (2009) 1263-1284.

[5] E. M. Karabulut, T. Ibrikci, Effective Automated Prediction of Vertebral Column Pathologies Based on Logistic Model Tree with SMOTE Preprocessing, Journal of medical systems 38 (2014) 1-9.

[6] Y. Wang, M. A. Simon, P. Bonde, B. U. Harris, J. J. Teuteberg, R. L. Kormos, J. F. Antaki, Decision tree for adjuvant right ventricular support in patients receiving a left ventricular assist device, The Journal of Heart and Lung Transplantation 31 (2012) 140-149.

[7] T. Sun, R. Zhang, J. Wang, X. Li, X. Guo, Computer-aided diagnosis for early-stage lung cancer based on longitudinal and balanced data, PloS one 8 (2013) e63559

[8] H. Irshad, L. Roux, D. Racoceanu, Multi-channels statistical and morphological features based mitosis detection in breast cancer histopathology, In Engineering in Medicine and Biology Society, 35th Annual International Conference of the IEEE, 2013, 6091-6094.

[9] K. J. Wang, B. Makond, K. M. Wang, An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data, BMC medical informatics and decision making 13 (2013) 124.

[10] L. M. Taft, R. S. Evans, C. R. Shyu, M. J. Egger, N. Chawla, J. A. Mitchell, SN. Thornton, B. Bray, M. Varner , Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery, Journal of biomedical informatics 42 (2009) 356-364.

[11] H. L. Yin, T. Y. Leong, A model driven approach to imbalanced data sampling in medical decision making, Studies in health technology and informatics, 160(2010) 856-860.

[12] 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) 341-378.

[13] I. Tomek, Two modifications of CNN, IEEE Transactions on Systems, Man and Cybernetics 6 (1976) 769-772.