Evaluation of minority class oversampling (SMOTE) on imbalanced CVD data

Dr. S. Kowalczyk[1], Dr. M. Pawlak[1], Dr. T. Michalski[1], Dr. A. Kaczmarek[1]*

[1] Department of Clinical Medicine and Health Sciences, Medical University of Lublin, Lublin, Poland

Abstract: In this work, we employ the minority oversampling technique (SMOTE) to generate instances of the minority class in an imbalanced Coronary Artery Disease dataset. We firstly analyze the public dataset published by Roohallah Alisadehsani, a dataset used for non-invasive prediction of CAD. We perform feature selection to exclude attributes unrelated to Coronary Artery Disease risk. The generation of new samples is performed using SMOTE, a technique commonly employed in machine learning tasks. We design Artificial Neural Networks, Decision Trees, and Support Vector Machines to classify both the original dataset and the augmented. The results demonstrate that data augmentation may be beneficial in specific cases, but it is not a panacea, and its application in a specific dataset should be carefully examined.

Keywords: Minority class oversampling, SMOTE, Imbalanced medical datasets, Artificial Neural Networks, Machine Learning, Coronary Artery Disease

1. Introduction

The imbalanced data set is defined as two-class data set in which one class (called majority) has an overwhelming number of instances than the other class (called minority). The classification problem for imbalanced data is exciting and challenging to researchers because most standard data mining methods claim their assumption for balanced data but are not applicable for imbalanced one [1].

The application of the SMOTE technique is widely used in economics, industry, and other fields. It has proven a useful technique to increase the number of data. However, it is not a panacea in all situations. Medical data are a special kind of data; they derive from human beings; therefore, applying oversampling techniques is challenging. In essence, the generation of hypothetical patient cases is risky, and its implementation must be strictly thought.

In this work, we investigate the ability of specific machine learning algorithms to improve their accuracy and reduce their False Positives and False Negatives with augmented medical data. We show that the public dataset Allisadehsani [2], is heavily imbalanced.

The prementioned dataset is commonly used for the prediction of CAD. The prediction and diagnosis of Coronary Artery Disease (i.e., the stenosis of the coronary arteries of our heart), is challenging. Diagnostic tests and symptoms (Angina) are not trustworthy enough for a doctor to directly diagnose CAD. Therefore, the most common practice to confirm or deny the presence of the disease is the invasive Coronary Angiography [3]. Therefore, several machine learning, deep learning, and more generally, artificial intelligence techniques have been proposed to solve the issue. The absence of balanced, correctly labeled, and complete datasets is making the task even more challenging.

We extract specific features of the dataset and perform several experiments on the mentioned dataset. We show that the SMOTE technique can be useful to particular algorithms and cases, as it can improve the classifiers' capability to recognize the False Negatives and False Positives.

2. Materials and Methods

2.1. SMOTE

The technique called SMOTE was introduced by Chawla [4]. SMOTE is an oversampling approach for the increase of the minority class instances. The minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with duplicated real data entries. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors of a record are randomly chosen. Our implementation currently uses five nearest neighbors.

The main idea behind SMOTE 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 of these neighbors. Utilizing interpolation theory, new samples are generated, therefore avoiding the duplication of random instances.

2.2 The imbalance of the dataset

From the 59 attributes of the dataset, we use 22 attributes to generate a version of this set. Some attributes do not play an important role in the Coronary Artery Disease diagnosis problem [5], and thus they were excluded from this version. The final attributes are:

1. Typical Angina with crisp 0,1 values
2. Atypical Angina with crisp 0,1 values
3. Dyspnea with crisp 0,1 values
4. Asymptomatic with crisp 0,1 values
5. Male with crisp 0,1 values
6. Female with crisp 0,1 values
7. Age under 40 with crisp 0,1 values
8. Age 40-50 with crisp 0,1 values
9. Age 50-60 with crisp 0,1 values
10. Age above 60 with crisp 0,1 values
11. Previous Stroke with crisp 0,1 values
12. Ex-Smoker with crisp 0,1 values
13. Current Smoker with crisp 0,1 values
14. Arterial Hypertension with crisp 0,1 values
15. Dyslipidemia with crisp 0,1 values
16. Obesity with crisp 0,1 values
17. Diabetes with crisp 0,1 values
18. Chronic Kidney Failure with crisp 0,1 values
19. ECG Normal with crisp 0,1 values
20. ECG Abnormal with crisp 0,1 values
21. ECHO Normal with crisp 0,1 values
22. ECHO abnormal with crisp 0,1 values
23. CAD/NO CAD with categorical values "YES" or "NO"

Figure 1 illustrates the number of Healthy and Diseased instances, proving the strong imbalance of the dataset. The number of healthy instances is 87, whereas the number of diseased is 217. Practically that means that if a classifier labels all instances as diseased, it automatically achieves 71.3% accuracy since 71.3% of the instances are, in fact, diseased.

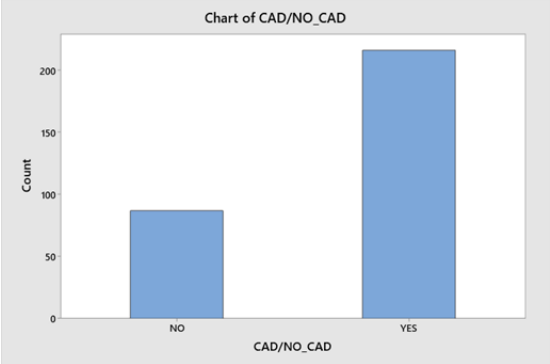

Figure 1. Chart of Healthy and Diseased instances in the dataset

2.3 The classifiers' characteristics

The architectures and parameters of ANN (Artificial Neural Network) and trees are given in Table 1. We specify the parameters of the networks we developed, for example, the Learning Rate, the Momentum, the Weight Decay, etc.

Table 1. Classifiers' parameters

Classifier / Parameters	Epochs / Batch Size	Hidden Layers / Size	Other
ANN SA1	150 / 32	512	LR: 0.1, Mom:0.1, Decay: Yes
ANN SA2	150 / 32	1024	LR: 0.1, Mom:0.1, Decay: Yes
ANN SA3	150 / 32	1024	LR: 0.01, Mom:0.1, Decay: Yes
Spegasos	250 / 32	-	Lambda: 1e-4, Hinge Loss
Random Forest	150 / 32	Unlimited	-
RepTree	Unspecified / 32	Unlimited	-

3. Results

We provide the results in accuracy and incorrectly classified instances of each classifier in each case. The most vital factor and evaluation criteria are the actual reduction of False Positives or False Negatives and not an improvement in classification accuracy. Besides, an accuracy improvement may, in fact, come from the generated instances (if they are correctly classified) and not from an actual learning improvement of the classifier. Table 2 illustrates the results of the classifiers.

Table 2. Results on original and on augmented data

Classifier / Metrics	Augmented Data Accuracy	Augmented Data Incorrectly Classified Instances (total)	Augmented Data actual Incorrectly Classified Instances (total)	Original Data Accuracy	Original Data Incorrectly Classified Instances (total)
ANN SA1	82.81	11 (64)	8 (61)	82.22	8 (45)
ANN SA2	**81.25**	**12 (64)**	**10 (62)**	**82.22**	**8 (45)**
ANN SA3	**76.56**	**15 (64)**	**10 (59)**	**68.8**	**14 (45)**
Spegasos	81.25	12 (64)	12 (64)	77.7	10 (45)
Random Forest	**89.06**	**7 (64)**	**7 (64)**	**84.44**	**7 (45)**
RepTree	85.9	9 (64)	9 (64)	75.55	11 (45)

4. Conclusions

RandomForest classifier outmatches the others, achieving 84.44% accuracy and only seven incorrectly classified instances in the original dataset. Augmentation on the dataset did not improve the actual mistakes; however, it improved its accuracy (89.06%). In essence, the improvement in accuracy comes from the correct classification of the generated instances and does not depict an actual improvement.

ANN SA3 classifier was benefited from the application of SMOTE. Not only did the accuracy improved from 68.8% to 76.56%, but also achieved an actual reduction of mistakes. The incorrectly classified instances in the original dataset were 14, whereas the actual incorrectly classified instances in the augmented data were only 10. The classifier classified incorrectly five generated instances.

RepTree was also benefited from the data augmentation, reducing its actual mistakes from 11 to 9, and improving its accuracy by +10.45%. Spegasos and ANN SA1 were not benefited from data augmentation, despite their accuracy being slightly improved.

The main conclusion is that the imbalance we observe in the dataset is transferred to the performance of the classifiers. As expected, data augmentation is not always beneficial, and its application should be deeply examined and considered before any action.

References

1. Q. Gu, Z. Cai, L. Ziu, Classification of imbalanced data sets by using the hybrid re-sampling algorithm based on isomap, in: LNCS, Advances in Computation and Intelligence, 5821, 2OO9, pp. 287–296.
2. R. Alizadehsani, J. Habibi, M.J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, Z.A. Sani, A data mining approach for diagnosis of coronary artery disease, Computer Methods and Programs in Biomedicine, 111 (2013) 52-61.
3. P.H. Brubaker, Coronary Artery Disease: Essentials of Prevention and Rehabilitation Programs, Human Kinetics Publishers, 2002.
4. Nitesh V. Chawla, SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research 16 (2002) 321–357
5. CAD. Information available at http://www.nhlbi.nih.gov/health/healthtopics/topics/cad/ (accessed 24.02.12).