Chapter 35
The Use of Prediction Reliability Estimates on Imbalanced Datasets:
A Case Study of Wall Shear Stress in the Human Carotid Artery Bifurcation

Domen Košir
University of Ljubljana, Slovenia & Httpool Ltd., Slovenia

Zoran Bosnić
University of Ljubljana, Slovenia

Igor Kononenko
University of Ljubljana, Slovenia

ABSTRACT

Data mining techniques are extensively used on medical data, which is typically composed of many normal examples and few interesting ones. When presented with highly imbalanced data, some standard classifiers tend to ignore the minority class which leads to poor performance. Various solutions have been proposed to counter this problem. Random undersampling, random oversampling, and SMOTE (Synthetic Minority Oversampling Technique) are the most well-known approaches. In recent years several approaches to evaluate the reliability of single predictions have been developed. Most recently a simple and efficient approach, based on the classifier's class probability estimates was shown to outperform the other reliability estimates. The authors propose to use this reliability estimate to improve the SMOTE algorithm. In this study, they demonstrate the positive effects of using the proposed algorithms on artificial datasets. The authors then apply the developed methodology on the problem of predicting the maximal wall shear stress (MWSS) in the human carotid artery bifurcation. The results indicate that it is feasible to improve the classifier's performance by balancing the data with their versions of the SMOTE algorithm.

DOI: 10.4018/978-1-4666-2455-9.ch035

INTRODUCTION

Increase of the stroke risk is induced by many factors: age, systolic and diastolic hypertension, diabetes, cigarette smoking, high levels of cholesterol, arrhythmia, etc. Changes of the geometrical vessel dimensions in the region of the carotid artery bifurcation certainly affect the blood flow and may lead to stenosis process (Schulz & Rothwell, 2001).

The stenosis is a narrowing of the inner surface (lumen) of the blood vessel. Carotid artery stenosis is usually caused by the cholesterol plaque buildup. The plaque makes the blood flow to become faster and more turbulent. Irregular blood flow can cause pieces of plaque to break off and block smaller arteries in the brain. The pieces of plaque can partially or completely restrict blood flow to parts of the brain which that vessel supplies. The risk of this happening is especially high in patients with arrhythmia.

The common carotid artery supplies the neck, head and brain with oxygenated blood. In the neck it bifurcates into the internal and external carotid artery. The blood flow in this section was simulated using a 3D model in order to analyze the influence of geometric parameters on maximum wall shear stress (MWSS) in the human carotid artery bifurcation (Radović & Filipović, 2010).

We transformed the regression problem of predicting the MWSS value into two classification problems by setting two thresholds for wall shear stress values. We try to predict the levels MWSS using the 3D model's geometric parameters, but both classification datasets (mwss95 and mwss99) suffer from the class imbalance problem.

Big imbalance in data can cause some classifiers to perform poorly. Imbalanced data is common in real world problems, such as image analysis (Kubat, Holte & Matwin, 1998), fraud detection (Fawcett & Provost, 1996), text classification (Zheng, Wu & Srihari, 2004) and medicine (Mac Namee, Cunningham, Byrne & Corrigan, 2002; Cohen, Hilario, Sax, Hugonnet & Geissbuhler, 2006). When the majority examples heavily outnumber the minority examples some classifiers tend to ignore the minority class. Classification accuracy measure, however, does not consider this. For instance, a simple classifier that always predicts the majority class would show a 99% classification accuracy when presented with a dataset that consists of 99% majority examples and 1% minority examples. This classifier would of course be useless. In this study, we focus more on the informative AUC value (Area Under the ROC Curve) instead of relying on classification accuracy.

Several already existant approaches enable even the imbalance-sensitive classifiers to be able to successfully predict minority examples. Some of the proposed solutions focus on the algorithmic level – they modify existing classifiers and present new algorithms that are not sensitive to imbalanced learning data. Other approaches focus on the data. They modify the data itself in order to soften the ratio between the numbers of majority and minority examples.

The imbalance in data can be, for example, countered by reducing the number of majority examples by randomly removing majority examples from the dataset (random undersampling) or by replicating minority examples (random oversampling). Random undersampling and random oversampling are very straightforward algorithms that can be used to change the numbers of majority and minority examples. The effects of data undersampling and oversampling were extensively studied in the last decade (Estabrooks & Japkowitz, 2001; Chawla, Japkowitz & Koltz, 2004).

A new algorithm called SMOTE (Chawla, Bowyer, Hall & Kegelmeyer, 2002) was introduced in 2002 (see Algorithm 1). Instead of deleting or duplicating random examples in the dataset, this algorithm generates synthetic examples using the existing minority examples. For every synthetic example a minority example and one of its nearest neighbors are used to generate a new minority example. Several researchers used this algorithm