Imbalanced Classification for Business Analytics

Talayeh Razzaghi
University of Central Florida, USA

Andrea Otero
University of Central Florida, USA

Petros Xanthopoulos
University of Central Florida, USA

INTRODUCTION

In pattern recognition, classification is a crucial task for automated data driven knowledge discovery. The objective of classification is to separate a set of data into classes or sub-categories and then to identify the classes that a new observation belongs to according to a training set of data. The mathematical model trained by a classification algorithm is termed *classifier*. When the class size of given examples is not equal for all classes, the classification problem is known as imbalanced (Japkowicz, 2000). For instance, in a cancer diagnostic problem the main objective is to identify individuals stricken with cancer and such events are relatively rare compared to normal cases. Imbalanced classification problems are also known as *skewed class distribution problems* or as *small/ rare class learning problems* (He & Garcia, 2009; Lemnaru & Potolea, 2012; Sun, Wong, & Mohamed, 2009). In binary classification, the class with fewer examples is known as the *minority class* and the other class as the *majority class*. In many applications (e.g. fraud detection, computer intrusion detection, oil spill detection, defect product detection), detection of minority class examples is more important than the majority class. Therefore, there is a need for efficient classification algorithms to address such problems. A preferred classification algorithm is the one that yields higher identification rate on rare events especially for applications where their misclassification yields to high losses. For instance in automated credit card fraud detection, a fraud event misclassification might result in high monetary losses for the credit card vendor. On the other side misclassification of non-fraudulent events will worsen the customer satisfaction experience.

The emerging nature of imbalanced classification problems has led to the development of modified algorithms and new performance metrics. Standard performance measures such as classification accuracy are not appropriate when the data is imbalanced (N. V. Chawla, 2010; He & Garcia, 2009). In this chapter, we analyze the theoretical framework of imbalanced classification, the main algorithmic approaches proposed in the literature and some of the most prominent applications in business. These business applications include customer relationship management (CRM) (Kim, Chae, & Olson, 2013), fraud detection (Wei, Li, Cao, Ou, & Chen, 2012), and risk management (Groth & Muntermann, 2011).

BACKGROUND

Advances in science and technology accelerate the accessibility of raw data and create new opportunities for knowledge discovery. Imbalanced problems can be found in a wide variety of applications, including security surveillance (Wu, Wu, Jiao, Wang, & Chang, 2003), medical diagnosis (Mena & Gonzalez, 2009; You, Zhao, Li, & Hu,

DOI: 10.4018/978-1-4666-5202-6.ch105

2011), bioinformatics (Al-Shahib, Breitling, & Gilbert, 2005), geomatics (Kubat, Holte, & Matwin, 1998), telecommunications (Tang, Krasser, Judge, & Zhang, 2006), risk management (Ezawa, Singh, & Norton, 1996), manufacturing (Adam et al., 2011), quality estimation (Lee, Song, Song, & Yoon, 2005), and power management (Hu, Zhu, & Ren, 2008). Imbalanced classification has been studied in a number of studies (N. V. Chawla, 2010; Guo, Yin, Dong, Yang, & Zhou, 2008; He & Garcia, 2009; Su, Mao, Zeng, Li, & Wang, 2009; Sun et al., 2009). Previous works on the classification of imbalanced data (N. V. Chawla, 2010; Kubat et al., 1998; Ngai, Hu, Wong, Chen, & Sun, 2011; Su et al., 2009; Sun et al., 2009) address that many standard classification algorithms achieve poor performance. Therefore, despite the existing amounts of literature there is room for improvement and future contribution.

MAIN FOCUS

In this part, we present (1) the appropriate performance measures for imbalanced data; (2) imbalanced classification techniques and (3) the most popular business analytics applications.

Performance Measures

Classification performance measures can be obtained, directly or indirectly, from the confusion matrix. For a classification problem with k classes, the confusion matrix is a square matrix $C \in R^k$, with each of its entries c_{ij}, denoting the percentage of the samples that belong to the class i and classified to the class j. For the special case of binary classification (positive and negative), the confusion matrix is as follows:

Clearly, the confusion matrix of an ideal classifier is diagonal. In this matrix, diagonal elements represent accurately classified examples and the off-diagonal elements the misclassified data for each class. A typical performance measure for classification is the so-called accuracy, which is calculated as the correctly classified samples over the total number of training samples. Since the majority class dominates the behavior of this metric, it might not be an appropriate performance indicator for imbalanced classification problems. More specifically a naive decision rule can yield high classification accuracy with no real practical value. For this, performance measures such as sensitivity and specificity are often employed:

Table 1. Confusion matrix for binary classification problem

		Predicted	
		Prevalent	**Rare**
Actual	Prevalent	TP (True Positive)	FN (False Negative)
	Rare	FP (False Positive)	TN (True Negative)

$$Sensitivity = \frac{TP}{TP + FN},$$
$$Specificity = \frac{TN}{TN + FP}$$

Sensitivity value is driven by the correct classification of the majority class whereas specificity depends on the minority class. The plot of sensitivity versus specificity is called *Receiver Operator Characteristic* (ROC) curve and it provides a good visual representation of the classifier (Figure 1). A combined measure frequently used for imbalanced data is the geometric mean of sensitivity and specificity (often abbreviated G-mean) defined by

$$G - mean = \sqrt{Sensitivity * Specificity}$$

There are other metrics used in the literature, including precision and recall or hit rate which is the ratio of true positive to the sum of true positive and false positive (Duman, Ekinci, & Tanrıverdi, 2012) and lift which is highly related to accuracy, but it is well used in marketing practice (Ling &