Cost-Sensitive Learning

Victor S. Sheng
New York University, USA

Charles X. Ling
The University of Western Ontario, Canada

INTRODUCTION

Classification is the most important task in inductive learning and machine learning. A classifier can be trained from a set of training examples with class labels, and can be used to predict the class labels of new examples. The class label is usually discrete and finite. Many effective classification algorithms have been developed, such as naïve Bayes, decision trees, neural networks, and so on. However, most original classification algorithms pursue to minimize the error rate: the percentage of the incorrect prediction of class labels. They ignore the difference between types of misclassification errors. In particular, they implicitly assume that all misclassification errors cost equally.

In many real-world applications, this assumption is not true. The differences between different misclassification errors can be quite large. For example, in medical diagnosis of a certain cancer, if the cancer is regarded as the positive class, and non-cancer (healthy) as negative, then missing a cancer (the patient is actually positive but is classified as negative; thus it is also called "false negative") is much more serious (thus expensive) than the false-positive error. The patient could lose his/her life because of the delay in the correct diagnosis and treatment. Similarly, if carrying a bomb is positive, then it is much more expensive to miss a terrorist who carries a bomb to a flight than searching an innocent person.

BACKGROUND

Cost-sensitive learning takes costs, such as the misclassification cost, into consideration. It is one of the most active and important research areas in machine learning, and it plays an important role in real-world data mining applications. A comprehensive survey (Turney, 2000) lists a large variety of different types of costs in data mining and machine learning, including misclassification costs, data acquisition cost (instance costs and attribute costs), active learning costs, computation cost, human-computer interaction cost, and so on. The misclassification cost is singled out as the most important cost, and it has also been mostly studied in recent years (e.g., (Domingos, 1999; Elkan, 2001; Zadrozny & Elkan, 2001; Zadrozny et al., 2003; Ting 1998; Drummond & Holte, 2000, 2003; Turney, 1995; Ling et al, 2004, 2006b; Chai et al., 2004; Sheng & Ling, 2006)).

Broadly speaking, cost-sensitive learning can be categorized into two categories. The first one is to design classifiers that are cost-sensitive in themselves. We call them the direct method. Examples of direct cost-sensitive learning are ICET (Turney, 1995) and cost-sensitive decision tree (Drummond & Holte, 2000, 2003; Ling et al, 2004, 2006b). The other category is to design a "wrapper" that converts any existing cost-insensitive (or cost-blind) classifiers into cost-sensitive ones. The wrapper method is also called cost-sensitive meta-learning method, and it can be further categorized into thresholding and sampling. Here is a hierarchy of the cost-sensitive learning and some typical methods. This paper will focus on cost-sensitive meta-learning that considers the misclassification cost only.

Cost-sensitive learning:

- Direct methods
  - ICET (Turney, 1995)
  - Cost-sensitive decision tree*s* (Drummond & Holte, 2003; Ling et al, 2004, 2006b)
- Meta-learning
  - Thresholding
    - MetaCost (Domingos, 1999)
    - CostSensitiveClassifier *(CSC in short)* (Witten & Frank, 2005)

- Cost-sensitive naïve Bayes (Chai et al., 2004)
- Empirical Threshold Adjusting (*ETA* in short) (Sheng & Ling, 2006)

o Sampling
- Costing (Zadronzny et al., 2003)
- Weighting (Ting, 1998)

MAIN FOCUS

In this section, we will first discuss the general theory of cost-sensitive learning. Then, we will provide an overview of the works on cost-sensitive learning, focusing on cost-sensitive meta-learning.

Theory of Cost-Sensitive Learning

In this section, we summarize the theory of cost-sensitive learning, published mostly in (Elkan, 2001; Zadrozny & Elkan, 2001). The theory describes how the misclassification cost plays its essential role in various cost-sensitive learning algorithms.

Without loss of generality, we assume binary classification (i.e., positive and negative class) in this paper. In cost-sensitive learning, the costs of false positive (actual negative but predicted as positive; denoted as (*FP*), false negative (*FN*), true positive (*TP*) and true negative (*TN*) can be given in a cost matrix, as shown in Table 1. In the table, we also use the notation *C(i, j)* to represent the misclassification cost of classifying an instance from its actual class *j* into the predicted class *i*. (We use 1 for positive, and 0 for negative). These misclassification cost values can be given by domain experts. In cost-sensitive learning, it is usually assumed that such a cost matrix is given and known. For multiple classes, the cost matrix can be easily extended by adding more rows and more columns.

Note that *C(i, i)* (*TP* and *TN*) is usually regarded as the "benefit" (i.e., negated cost) when an instance is predicted correctly. In addition, cost-sensitive learning is often used to deal with datasets with very imbalanced class distribution (Japkowicz, 2000; Chawla et al., 2004). Usually (and without loss of generality), the minority or rare class is regarded as the positive class, and it is often more expensive to misclassify an actual positive example into negative, than an actual negative example into positive. That is, the value of *FN* or *C(0,1)* is usually larger than that of *FP* or *C(1,0)*. This is true for the cancer example mentioned earlier (cancer patients are usually rare in the population, but predicting an actual cancer patient as negative is usually very costly) and the bomb example (terrorists are rare).

Given the cost matrix, an example should be classified into the class that has the minimum expected cost. This is the minimum expected cost principle (Michie, Spiegelhalter, & Taylor, 1994). The expected cost *R(i|x)* of classifying an instance *x* into class *i* (by a classifier) can be expressed as:

$$R(i \mid x) = \sum_j P(j \mid x) C(i, j), \quad (1)$$

where *P(j|x)* is the probability estimation of classifying an instance into class *j*. That is, the classifier will classify an instance *x* into positive class if and only if:

$$P(0|x)C(1,0) + P(1|x)C(1,1) \leq P(0|x)C(0,0) + P(1|x)C(0,1)$$

This is equivalent to:

$$P(0|x)(C(1,0)-C(0,0)) \leq P(1|x)(C(0,1)-C(1,1))$$

Table 1. An example of cost matrix for binary classification

	Actual negative	Actual positive
Predict negative	C(0,0), or TN	C(0,1), or FN
Predict positive	C(1,0), or FP	C(1,1), or TP