Cost-Sensitive Learning

Zhi-Hua Zhou

National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
zhouzh@lamda.nju.edu.cn

In conventional classification settings, the classifiers generally try to maximize the *accuracy* or minimize the *error rate*, both are equivalent to minimizing the number of mistakes in classifying new instances. Such a setting is valid when the costs of different types of mistakes are equal. In real-world applications, however, the costs of different types of mistakes are often unequal. For example, in intrusion detection, the cost of mistakenly classifying an intrusion as a normal access is usually far larger than that of mistakenly classifying a normal access as an intrusion, because the former type of mistakes will result in much more serious losses.

In cost-sensitive learning, rather than simply minimizing the number of mistakes, the goal is to minimize the *total cost*. Roughly speaking, there are two types of misclassification costs, i.e., class-dependent or example-dependent costs. The former assumes that the costs are associated with classes, that is, every class has its own misclassification cost; the latter assumes that the costs are associated with examples, that is, every example has its own misclassification cost. In most real tasks it is feasible to get the cost of misclassifying one class to another class, e.g., by querying domain experts, while only in some special tasks it is easy to get the cost for every training example. In this talk we will focus on the class-dependent misclassification costs.

The most fundamental and popular approach to cost-sensitive learning is `Rescaling`, or called `Rebalance`. This approach tries to rebalance the classes such that the influences of different classes are in proportion to their costs. For example, the `Rescaling` approach can be realized by *resampling*, where the lower-cost class examples can be under-sampled such that the number of examples of the lower-cost and higher cost classes are in proportion to their misclassification costs, respectively. In addition to *resampling*, the `Rescaling` approach can also be realized in other forms, such as *reweighting* the training examples or *threshold-moving* of the decision boundaries. Notice that `Rescaling` is an essential procedure for handling unequal costs; indeed, most cost-sensitive learning approaches can be regarded as different realizations of `Rescaling` with different base learners.

Though `Rescaling` works very well in two-class classification problems, it was found that it often fails in multi-class problems. In this talk, we will analyze why this phenomenon occurs, and introduce an updated `Rescaling` approach. Then, we will discuss on how to handle inexact cost information; this is an important and challenging problem since it is usually difficult to get exact cost information

V. Torra et al. (Eds.): MDAI 2011, LNAI 6820, pp. 17–18, 2011.

in real-world tasks, yet previous cost-sensitive learning studies assumed that exact costs of different types of misclassifications are known. We will also briefly introduce cost-sensitive face recognition, and a task involving other types of unequal costs such as feature extraction cost.

References

1. Zhou, Z.H., Liu, X.Y.: On multi-class cost-sensitive learning. Computational Intelligence 26, 232–257 (2010)
2. Zhou, Z.H., Liu, X.Y.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering 18, 63–77 (2006)
3. Liu, X.Y., Zhou, Z.H.: Learning with cost intervals. In: Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, pp. 403–412 (2010)
4. Zhang, Y., Zhou, Z.H.: Cost-sensitive face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1758–1769 (2010)
5. Liu, L.P., Yu, Y., Jiang, Y., Zhou, Z.H.: TEFE: A time-efficient approach to feature extraction. In: Proceedings of the 8th IEEE International Conference on Data Mining, Pisa, Italy, pp. 423–432 (2008)