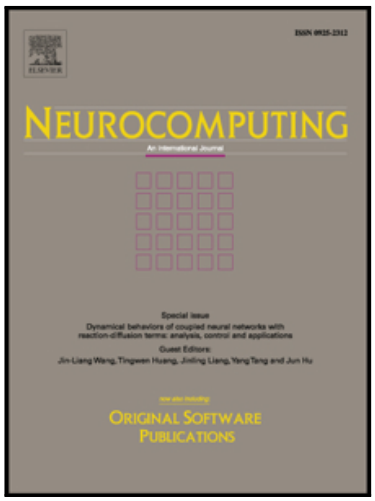

A cost-sensitive semi-supervised learning model based on uncertainty

Hongyu Zhu , Xizhao Wang

PII: S0925-2312(17)30680-X
DOI: 10.1016/j.neucom.2017.04.010
Reference: NEUCOM 18343

To appear in: *Neurocomputing*

Received date: 21 November 2016
Revised date: 31 March 2017
Accepted date: 3 April 2017

Please cite this article as: Hongyu Zhu , Xizhao Wang , A cost-sensitive semi-supervised learning model based on uncertainty, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.04.010

A cost-sensitive semi-supervised learning model based on uncertainty

Hongyu Zhu and Xizhao Wang
Big Data Institute, College of Computer Science and Software Engineering,
ShenZhen University, ShenZhen, China
Email: xizhaowang@ieee.org 870784502@qq.com

Abstract. Aiming at reducing the total cost in cost-sensitive learning, this paper introduces a semi-supervised learning model based on uncertainty of sample outputs. Its central idea is (1) to categorize the samples which are not in training set into several groups based on the uncertainty-magnitude of their outputs, (2) to add the group of samples which have the least uncertainty together with their predicted labels in the original training set, and (3) to retain a new classifier for total cost reduction. The ratio of costs between classes and its impact on learning system improvement are discussed. Theoretical analysis and experimental demonstration show that the model can effectively improve the performance of a cost-sensitive learning algorithm for a certain type of classifiers.

Keywords: Uncertainty, Cost-Sensitive, Semi-supervised learning, Sample selection, Extreme Learning Machine

1. **Introduce**

Technology of classification which belongs to a topic of machine learning has been widely applied in a lot of domains such as pattern recognition, knowledge discovery, intelligent control, network security, gene engineering, bioinformatics, and so on. From literature we can find many approaches to designing classifiers such as decision tree induction [1,2], hypothesis version spaces [3], Bayesian networks [4,5], evolutionary computing [6], logistic regressions [7,8], support vector machines [9,10], neural networks [11], and deep learning technique [12,13]. The most important index for evaluating a designed classifier is the generalization ability, i.e., the rate of correctly classifying samples which are not in training set.

The concept of cost-sensitive classification can be introduced into the process of classifier design [14, 15, 16, 17, 18]. The “cost-sensitive” refers to that a class (feature, or object) has the different cost in comparison with another class (feature, or object respectively) in classification process. For example, the cost of wrongly classifying a patient as non-cancer class is much bigger than the cost of wrongly classifying the patient as cancer class. From the viewpoint of loss function, cost-sensitive classifier learning is to minimize a cost-loss function but cost-insensitive learning is to maximize the correct rate of classification. Under some certain conditions, these two objective functions of optimization can be equivalent.