Classifying Unbalanced Pattern Groups by Training Neural Network

Bo-Yu Li, Jing Peng, Yan-Qiu Chen, and Ya-Qiu Jin

Key Laboratory of Wave Scattering and Remote Sensing Information,
(Ministry of Education), Fudan University, Shanghai 200433, China
liboyu@Fudan.edu.cn

Abstract. When training set is unbalanced, the conventional least square error (LSE) training strategy is less efficient to train neural network (NN) for classification because it often lead the NN to overcompensate for the dominant group. Therefore, in this paper a dynamic threshold learning algorithm (DTLA) is proposed as the substitute for the conventional LSE algorithm. This method uses multiple dynamic threshold parameters to gradually remove some training patterns that can be classified correctly by current Radial Basis Function (RBF) network out of the training set during training process, which changes the unbalanced training problem into a balanced training problem and improves the classification rate of the small group. Moreover, we use the dynamical threshold learning algorithm to classify the remote sensing images, when the unbalanced level of classes is high, a good effect is obtained.

1 Introduction

Unbalanced training set refers to number unbalance of training patterns or probability distribution unbalance of different groups [1]. In this paper, we will stress the problem of number unbalance of training patterns, we call the group that has more training patterns the dominant group (large class), similarly, we call the group that has few training patterns the non-dominant group(small class). Unbalanced training set can be observed in many applications. For example, in the task of image classification, we often obtain a training set that includes many positive examples and few negative examples because of different cost to obtain them; In defective product detection, most products will be good and only a few are defects. For these unbalanced data-sets, how to train ANN efficiently and impartially to classify every group will be immediate problem we must face. The conventional training method theoretically leads NN to ignore the small group because the conventional least square error (LSE) training algorithm will overcompensate for the dominant group, particularly when the separability of data-set is poor [1], [2]. As a consequence, although NN classifier probably achieves a good performance for the dominant group, it probably has a very low percentage of correct classification for small groups.

In this paper, the dynamic threshold learning algorithm (DTLA) is introduced to train neural network to solve this problem. In literatures [4], [5], a

J. Wang et al. (Eds.): ISNN 2006, LNCS 3972, pp. 8–13, 2006.

single threshold training method has been suggested to train RBF NN [3] for pattern recognition based on balanced data-set. However, it can not deal with the unbalanced data set because single threshold can not characterize the difference between large classes and small classes when there exist multiple classes in the data set. Therefore, Multi-threshold method is used to characterize these difference in the unbalanced data sets. Furthermore, new adjusting criteria are suggested to adjust the thresholds dynamically. Using this method, the number of training patterns that have large quantities could be reduced, consequently transform the unbalanced data set into a balanced one. Thus balances the training process. To prevent harming the performance of the NN classifier being trained and balance the training process, in the implementing we adjust the thresholds to remove some correctly recognized patterns out of the training set gradually during the training process. This method will result in the good results.

The DTLA method can make neural network classifier to care the small group effectively, which often includes some more important information of images, by changing the unbalanced situation into the balanced one. That is to say, we can use NN to capture the objective in which we are really interested even if its prototypes are very little. In this paper, the satellite images are especially focused on. Compared with the other training methods, experiment results show DTLA training method can stress the small group effectively.

2 The Dynamic Threshold Learning Algorithm (DTLA)

According to the analysis above, in order to train RBF NN impartially, the key problem is how to define suitable thresholds as controllers to remove recognized training patterns from the training set. Assuming each class corresponds to an output unit of the NN, during the training process, the recognized training patterns always have large output values (approaching to 1) on the corresponding output unit of the NN, while have small output values on the other output units(approaching to -1). Therefore, an upper threshold and a lower threshold should be considered to decide whether an input pattern should be ignored or not. Moreover, during the training iteration, large classes usually have more recognized patterns than that of small classes. Therefore, different thresholds are adopted for different output unit of the RBF NN, and the recognized training patterns of different groups are controlled to be removed from the training set sequentially. In most cases, large group have more recognized training patterns. If these patterns could be removed from the training set, it will help to restrain the overcompensation caused by large classes and force the RBF classifier to care the non-dominant class.

Assuming each output unit of RBF NN corresponds to a single group, the RBF NN has the same number of the output unit as the number of the groups. In order to simplify the presentation, in this paper, we do not distinguish the concept of the class and its corresponding output unit of NN. For example, the thresholds of the ith output unit of NN refers to the threshold of ω_i class.