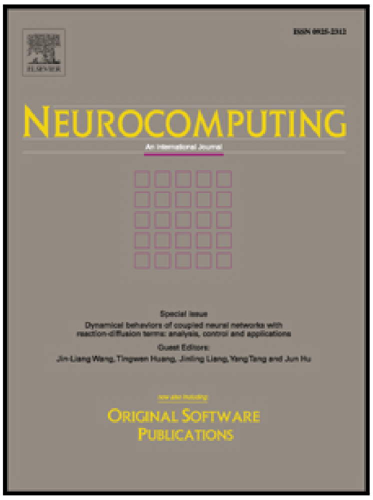
Communicated by Dr. Nianyin Zeng

Accepted Manuscript



Margin & Diversity based Ordering Ensemble Pruning

Huaping Guo, Hongbing Liu, Ran Li, Changan Wu, Yibo Guo,
Mingliang Xu

PII: S0925-2312(17)31180-3
DOI: 10.1016/j.neucom.2017.06.052
Reference: NEUCOM 18651

To appear in: *Neurocomputing*

Received date: 1 November 2015
Revised date: 26 October 2016
Accepted date: 24 June 2017

Please cite this article as: Huaping Guo, Hongbing Liu, Ran Li, Changan Wu, Yibo Guo, Mingliang Xu, Margin & Diversity based Ordering Ensemble Pruning, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.06.052

Available online at www.sciencedirect.com

ScienceDirect

Neurocomputing 00 (2017) 1–13

Journal Logo

Margin & Diversity based Ordering Ensemble Pruning

Huaping Guo[a,*], Hongbing Liu[a], Ran Li[a], Changan Wu[a], Yibo Guo[b], Mingliang Xu[b,*]

[a]*School of Computer and Information Technology, Xinyang Normal University, Xinyang, Henan, P. R. China*
[b]*School of Information Engineering, Zhengzhou University, Zhengzhou, Henan, P. R. China*

Abstract

Ensemble pruning is a technique used to improve ensemble performance and reduce the ensemble size by selecting an optimal or sub-optimal subset as the final ensemble for prediction. In this research, using example margin and ensemble diversity, we prove that the ensemble pruning method should focus more on the following two factors: (1) examples with small absolute margin and (2) classifiers that correctly classify more examples and contribute larger diversity. By incorporating ensemble members in a decreasing order based on the MDM, sub-ensembles are formed such that users can select the top T ensemble members for predictions. Compared to the original ensemble and other state-of-the-art ensemble pruning methods, the proposed method shows better performance in terms of accuracy.

Keywords:
Example margin, ensemble diversity, ensemble pruning

1. Introduction

Over the last decade, ensemble of multiple learning machines has been a very popular research topic in the machine learning and data mining. The basic idea is to construct multiple classifiers from original data and then aggregate their predictions when classifying examples with unknown classes. Theoretical and empirical results show that an ensemble has the potential to increase classification accuracy beyond the level reached by an individual classifier alone [1]. Dietterich stated [2] "A necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of its individual members is if the classifiers are accurate and diverse".

Many approaches have been proposed to create ensemble members that have both high accuracy and high diversity, and these can be grouped into four categories, involving: (1) manipulation of the data set [3, 4], (2) manipulation of the features [5, 6, 7, 8], (3) manipulation of the algorithms [11], and (4) manipulation through a combination of approaches [3, 4, 9, 10]. Bagging [3] and boosting [4], the most widely used and successful ensemble learning methods, fall into the first category. In bagging, individual classifiers are learned on data sets obtained by randomly sampling from the original training set and, through randomly disturbing the training set, the learned classifiers obtain high accuracy and sufficient diversity. Unlike bagging, boosting is an iterative learning process. For each iteration, boosting adjusts the distribution of the training set such that classifiers focus more on the examples that are hardly correctly classified. The approaches involving the manipulation of features try to build individual classifiers on diverse feature spaces that are obtained by selecting subsets or by generating new features from the original features. For example, random forests [5, 6] learn each

*Corresponding author
Email addresses: hpguo_cm@163.com (Huaping Guo), liuhbing@126.com (Hongbing Liu), liran358@163.com (Ran Li), wchangan@126.com (Changan Wu), ieybguo@zzu.edu.cn (Yibo Guo), iexumingliang@zzu.edu.cn (Mingliang Xu)