International Journal of Innovative
Computing, Information and Control
Volume **7**, Number **11**, November **2011**

 ISSN 1349-4198
pp. **6247–6259**

IGF-BAGGING: INFORMATION GAIN BASED FEATURE SELECTION FOR BAGGING

Gang Wang[1,2], Jian Ma[3] and ShanLin Yang[1,2]

[1]School of Management
[2]Ministry of Education Key Laboratory of Process Optimization and Intelligent Decision-Making
Hefei University of Technology
No. 193, Tunxi Road, Hefei 230009, P. R. China
wgedison@gmail.com; slyang@mail.hf.ah.cn

[3]Department of Information Systems
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
isjian@cityu.edu.hk

Received July 2010; revised January 2011

Abstract. *Bagging is one of the older, simpler and better known ensemble methods. However, the bootstrap sampling strategy in bagging appears to lead to ensembles of low diversity and accuracy compared with other ensemble methods. In this paper, a new variant of bagging, named IGF-Bagging, is proposed. Firstly, this method obtains bootstrap instances. Then, it employs Information Gain (IG) based feature selection technique to identify and remove irrelevant or redundant features. Finally, base learners trained from the new sub data sets are combined via majority voting. Twelve datasets from the UCI Machine Learning Repository are selected to demonstrate the effectiveness and feasibility of the proposed method. Experimental results reveal that IGF-Bagging gets significant improvement of the classification accuracy compared with other six methods.*
Keywords: Ensemble learning, Bagging, Feature selection, Information gain

1. **Introduction.** Research in the area of machine learning and data mining has achieved significant progress in the concept of learning from labeled instances. Although many efficient methods have been proposed, they have been limited to simple concepts or problems. Furthermore, numerous results suggest that learning more difficult concepts tends to be extremely difficult. Among the research directions, they have evolved to address these difficulties, which is ensemble learning [1,2]. A good understanding of how to build more sophisticated ensemble methods and exploit various possibilities of extracting information from the environment will move us to be closer to achieving the original intent of machine learning and data mining [2].

Ensemble learning is a machine learning paradigm where multiple learners are trained to solve the same problem. In contrast to ordinary machine learning methods which try to learn one hypothesis from training data, ensemble learning tries to construct a set of hypotheses and combine them to use [3]. Learners composing an ensemble are usually called base learners. Ensemble methods have been approved theoretically and empirically to demonstrate the advantages over the individual base learner. Bagging [4] and boosting [5,6] are two popular ensemble methods to enforce weak base learners. The effectiveness of such methods comes primarily from the diversity caused by re-sampling the training set.

In practice, there are two basic requirements on the base learners for ensemble creation: diversity, i.e., the candidate base learners should be as diverse as possible, and accuracy,