International Journal of Innovative
Computing, Information and Control
Volume **6**, Number **11**, November **2010**

ICIC International ©2010 ISSN 1349-4198
pp. 5299–5310

AN ENHANCED ADABOOST ALGORITHM WITH NAIVE BAYESIAN TEXT CATEGORIZATION BASED ON A NOVEL RE-WEIGHTING STRATEGY

Huanling Tang[1,2], Jun Wu[1], Zhengkui Lin[1] and Mingyu Lu[1]

[1]School of Information Science and Technology
Dalian Maritime University
Dalian 116026, P. R. China
thl01@163.com; wujunas8@gmail.com
dalianjx@163.com; lumingyu@tsinghua.org.cn

[2]Shandong Institute of Business and Technology
Yantai, Shandong, 264005, P. R. China

Received August 2009; revised January 2010

Abstract. *AdaBoost cannot improve the performance of naive Bayesian (NB) categorization because NB classifier is relatively stable. In order to boost NB text classifier, a new re-weighting strategy for training examples is proposed. That is, the weight of the training examples is updated considering (a) whether the training example is misclassified by the current base classifier; and (b) the disagreement caused by those previous base classifiers for this training example, which is estimated by vote entropy. Thus, the misclassified examples with larger disagreement will be more informative and should be assigned higher weights in next iteration. Moreover, the confidence of NB base classifier is determined both in related to the training error rate and to its contribution to increase diversity among those base classifiers. Using the above strategies, an enhanced algorithm termed BoostVE is presented. Theoretical analyses prove that the upper training error bound for BoostVE is better than that for AdaBoost. Experimental results show BoostVE is effectively to improve the performance of NB text categorization.*

Keywords: AdaBoost, Re-weighting strategy, Vote entropy, Confidence, Naive Bayesian, Text categorization

1. **Introduction.** AdaBoost, the adaptive boosting algorithm introduced by Freund and Schapire [1,2] is generally considered as an effective boosting algorithm in many previous studies [1-8]. It maintains a set of weights over the original training set and updates these weights at each iteration to construct a new base classifier. The re-weighting rule is to increase the weight of examples that are incorrectly classified by the base classifier and decrease the weight of examples that are correctly classified.

AdaBoost can significantly improve the accuracy of an unstable classifier algorithm such as decision tree learning and neural network learning [1-5]. The so-called unstable is that if small changes to the training set causes large changes in the learned classifier. However, AdaBoost cannot increase or even decrease the accuracy of the stable naive Bayesian (NB) classifier [9,10]. A possible reason is that the instability of classifiers is a vital factor in the performance of boosting, but the NB classifier is relatively stable with small perturbation of training data. AdaBoost with NB classifier cannot generate multiple base classifiers with sufficient diversity.

For the above problem, most researchers try to use more complex classifier, such as tree or network style classifiers, instead of original NB classifier. For instance, Ting and Zhang introduced tree structures into NB classification [9]. The work of Ref. [10] proposed a