改进 NB 算法在垃圾邮件过滤技术中的研究

刘月峰，苑江浩，张晓琳

（内蒙古科技大学 信息工程学院，内蒙古 包头 014010）

摘　要：朴素贝叶斯（NB）是一种简单高效的分类算法，且在垃圾邮件过滤中得到广泛应用，但是其属性间独立性的假设在一定程度上影响了分类效果.针对这一问题，提出一种改进的 NB 算法——FOA-NB 算法.该算法将 NB 算法与果蝇优化算法（FOA）相结合，根据不同特征属性对分类的影响程度赋予不同的权值，通过 FOA 对权值进行优化，得到全局最优特征权向量，该算法在保留 NB 算法的简洁高效的优点的同时，通过权值优化获取更加具有决策性的特征属性，从而提高垃圾邮件过滤的正确率和召回率.通过仿真实验与 NB 算法、加权贝叶斯(WB)进行对比，结果表明 FOA-NB 算法使得垃圾邮件过滤效果得到明显改善，正确率和召回率均有所提高，且提高幅度约为 5%.

关键词：垃圾邮件；朴素贝叶斯；特征权重优化；果蝇优化算法

Improved NB Algorithm Research in Spam Filtering Technology

LIU Yue-feng, YUAN Jiang-hao, ZHANG Xiao-lin

(School of Information Engineering, Inner Mongolia University of Science and Technology, Baotou 014010,China)

Abstract：Naive Bayes (NB) is a simple and efficient classification algorithm, and it is widely used in spam filtering, but because of the independence between the attributes of the hypothesis which has to some extent affected in classification effect. To solve this problem, the FOA-NB algorithm is proposed which is an improved NB algorithm. The algorithm of NB algorithm and Fruit fly optimization algorithm (FOA) combination, according to the different feature attributes of the influence degree of the classification given different weights, to optimize the weights by FOA and get the global optimal feature weight vector, the algorithm in the NB algorithm retains the advantage of simple and efficient at the same time, by optimization of the weights to obtain attributes which have more decision-making, so as to improve the spam filtering correct rate and recall rate. Through the simulation experiment with NB algorithm, Weighted Bayesian (WB), the results show that the FOA-NB algorithm makes the spam filtering effect has been improved significantly, and the correct rate and recall rate are improved，and the increase of about 5%.

Key words：spam email；naive bayes；optimization of feature weight；fruit fly optimization algorithm

作者简介：

刘月峰　男，(1977-)，博士研究生，副教授.研究方向为机器学习、文本分类.

苑江浩(通讯作者)　男，(1992-)，硕士研究生.研究方向为机器学习、数据挖掘.E-mail:qiji_2014@163.com.

张晓琳　女，(1966-)，博士，教授.研究方向为数据库、大数据隐私保护.