International Journal of Innovative Computing, Information and Control
Volume **5**, Number **2**, February **2009**

ICIC International ©2009 ISSN 1349-4198
pp. **423–432**

FEATURE SELECTION USING PARTICLE SWARM OPTIMIZATION WITH APPLICATION IN SPAM FILTERING

Chih-Chin Lai[1], Chih-Hung Wu[1] and Ming-Chi Tsai[2]

[1]Department of Electrical Engineering
National University of Kaohsiung
Kaohsiung, Taiwan 81148
cclai@nuk.edu.tw; johnw@nuk.edu.tw

[2]Department of Information Management
Shu-Te University
Kaohsiung, Taiwan 82445

Received October 2007; revised March 2008

Abstract. *Using a finite set of features to determine an e-mail as spam or non-spam is a very popular but important topic. However, in most cases, the feature selection is empirically verified. In order to automatically determine the proper number of features to classify spam e-mails, this paper investigates the possibility of using a particle swarm optimization algorithm to find more relevant subset of the set of features. The selected subset contains the least number of features that most contribute to classification accuracy. The experimental results show that the proposed approach can be used to select the most proper discriminative features and to increase the performance of spam e-mails classification.*
Keywords: Feature selection, Particle swam optimization, Spam filtering

1. **Introduction.** With the explosive growth of the Internet, so comes the proliferation of unsolicited commercial e-mail (UCE), more commonly known as spam. Spam e-mails bring most users huge inconvenience such as they clutter users' mailboxes, expose unsuitable contents, waste network bandwidth, etc. According to some data in a report [8], spam e-mails cost all non-corp Internet users 255 million (US) dollars in 2006 year and the number of every user received annual spam is 2,200. Therefore, it is a big and difficult challenge problem to develop a good mechanism that automatically classifies and, if necessary, filters spam e-mails.

As of now, spam filters - computer programs that attempt to automatically scan and identify incoming e-mails - seem to be the most systematic progress against spam problem. Most commercially available filters depend on some simple techniques such as constructing white-lists of trusted senders, black-lists of known spammers, and hand-crafted rules that block e-mails containing specific words or phrases. On the other hand, in recent years considerable concern has been arisen over the adoption of machine learning techniques in the spam filtering research. From the machine learning perspective, spam filtering based on the textual content of an e-mail can be viewed as a classic case of text categorization [20], with the categories being spam or non-spam.

These machine learning-based approaches include rule learning [1], Naïve Bayes [2, 19], decision trees [5], support vector machines [9], maximum entropy model [26], or combinations of different learners [17]. The common concept behind these approaches consists of: