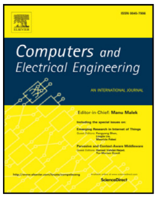
Computers and Electrical Engineering 000 (2018) 1–12

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng



Hybrid approach of improved binary particle swarm optimization and shuffled frog leaping for feature selection[☆]

S.P. Rajamohana[a,b,*], K. Umamaheswari[a,b]

[a] *Department of Information Technology, Coimbatore, Tamilnadu, India*
[b] *PSG College of Technology, Coimbatore, Tamilnadu, India*

ARTICLE INFO

Article history:
Received 11 August 2016
Revised 1 February 2018
Accepted 8 February 2018
Available online xxx

Keywords:
Review spam classification
Feature subset selection
Naive Bayes
kNN and SVM

ABSTRACT

Currently, the masses are interested in sharing opinions, feedbacks, suggestions on any discrete topics on websites, e-forums, and blogs. Thus, the consumers tend to rely a lot on product reviews before buying any products or availing their services. However, not all reviews available over internet are authentic. Spammers manipulate the reviews in their favor to either devalue or promote products. Thus, customers are influenced to take wrong decision due to these spurious reviews, i. e., spammy contents. In order to address this problem, a hybrid approach of improved binary particle swarm optimization and shuffled frog leaping algorithm are proposed to decrease high dimensionality of the feature set and to select optimized feature subsets. Our approach helps customers in ignoring fake reviews and enhances the classification performance by providing trustworthy reviews. Naive Bayes (NB), K Nearest Neighbor (kNN) and Support Vector Machine (SVM) classifiers were used for classification. The results indicate that the proposed hybrid method of feature selection provides an optimized feature subset and obtains higher classification accuracy.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

In current times, the amount of content available to the user on the internet is rapidly increasing [1]. While purchasing the product or availing services customers generally tend to make a decision relying solely on the information available in the review sites [2]. However, there is a limited quality control for these available data. This limitation invites people to post spurious reviews on the websites in order to either promote or demote the products [3]. Such individuals are known as opinion spammers. The positive spam reviews about a product may lead to financial gains and would help to increase the popularity of the product [4]. Similarly, negative spam reviews are posted with the intention of defaming a product or services [5]. Recently, the problem of spam or fake reviews has been on the rise, and many such cases have been released in the news. Hence, there arises a necessity of finding the authenticity of these reviews. Feature selection (FS) is a technique in which a subset of features are selected from the original dataset [6]. It is mainly used to build more robust learning models and to reduce the processing cost. The main purpose of feature selection is to reduce the number of features to increase both the performance of the model and the accuracy of classification [7]. FS can be examined as a search into a state space. Thus, a full search can be performed in all the search spaces traversed. However, this approach is not feasible in case of a very large number of features. Hence, a heuristic search deliberates those features, which have not yet been

[☆] Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. O. Bayat.

[*] Corresponding author at: Department of Information Technology, Coimbatore, Tamilnadu, India.
E-mail addresses: spr@ity.psgtech.ac.in (S.P. Rajamohana), uma@ity.psgtech.ac.in (K. Umamaheswari).

https://doi.org/10.1016/j.compeleceng.2018.02.015
0045-7906/© 2018 Elsevier Ltd. All rights reserved.

Please cite this article as: S.P. Rajamohana, K. Umamaheswari, Hybrid approach of improved binary particle swarm optimization and shuffled frog leaping for feature selection, Computers and Electrical Engineering (2018), https://doi.org/10.1016/j.compeleceng.2018.02.015

selected at each iteration, for evaluation. A random search creates random subsets within the search space that can be evaluated for importance of classification performance. Due to their randomized nature, meta-heuristics such as particle swarm optimization (PSO), evolutionary algorithms (EA), bat algorithm (BA), ant colony optimization (ACO) and genetic algorithm [8,9] are widely used for feature selection. When the feature space is high dimensional, selecting the optimal feature subset using traditional optimization methods have not proven to be effective. Therefore, meta-heuristic algorithms are used extensively for the appropriate selection of features. Two types of feature selection methods, namely the filter method and wrapper method can be incorporated for selecting subset of features. The filter model analyzes the intrinsic properties of data without involving the use of any learning algorithms [9] and can perform both subset selection and ranking. Though ranking involves identifying the importance of all the features, this method is more specifically used as a pre-process method since it selects redundant features. The wrapper model unlike other filter approaches considers the relationship between features [10]. This method initially uses an optimizing algorithm to generate various subsets of features and then uses a classification algorithm to analyze the subsets generated.

A rule-based approach was investigated to detect fake reviews in which the unexpected rules were defined to detect unusual behaviors of reviewers [11]. The study used an dataset available from Aamazon to identify spam activities. The N-gram method was applied to detect negative deceptive opinion [12]. Gold standard negative spam dataset which contains 400 reviews of 20 hotels in Chicago was used. The unigram and bigram features were trained by Support Vector Machine (SVM) classifiers. The results revealed that, the N-gram based SVM classifier achieved 86% accuracy in surpassing human judges. Two kinds of N-gram methods namely the character n gram (BON) and the word n-gram (BOW) were proposed to detect fake reviews [5]. Naive Bayes (NB) classifier was used for classifying both positive and negative reviews. The experimental results showed that the NB classifier achieved better results for positive reviews. Further, the SVM method was found to show better results in classifying deceptive and truthful negative reviews. The authors claimed that the BON showed better robustness when compared to BOW as it provided superior results with a small training dataset.

The content duplication technique was preferred for identifying the fake review [13]. Both duplicate and near-duplicate reviews were considered in training data set. Furthermore, two different techniques for spam detection were considered in the test dataset. The authors illustrated the content-based features which include 3 categories of reviews. Firstly, similarity of a review with the author's and other reviews on the target products. They also elucidate reviewer's centric features based on the burst patterns. The Probabilistic language model was developed to generate a similarity score between the reviews [14]. This approach evaluates the possibility of one review that are derived from the other. To detect the content similarity, they compared a couple of reviews by Kullback–Leibler. In addition to that Kullback–Leibler divergence measure calculates the spam score for every review. SVM was chosen for spam classification to classify both spam and ham reviews. They have achieved 81% precision in their method for detecting spam reviews.

Stylometric features, characterized either as lexical or syntactic representation were used for identifying review spam. While the lexical features represent the character or word-based features, the syntactic feature denotes the reviewers writing style at each sentence level. Graph-based methodology, the graph comprising three nodes: namely the review, the reviewer and store was applied for detecting review spammers [15,16]. It establishes the inter-relationships between two nodes, which is achieved by evaluating following: the credibility of the reviewer, the honesty of the reviews and the reliability of the store. In this case agreement score is calculated based on the user rating. The reliability of the store depends on the credibility of its reviewer's comments.

The existing works investigated the traditional feature selection techniques such as bag of words, bag of nouns, linguistic features, weighted PCA, keyword spotting and the machine learning algorithm for reviewing spam classification. However, till date no attempts have been made to use hybrid evolutionary algorithms for reviewing spam classification. The evolutionary algorithms have been applied for different applications such as scheduling, power system, and wireless sensor networks. This is the first study that utilizes evolutionary algorithms for classifying reviews into spam and ham. FS plays a major role in classification. Hence, lot of researchers primarily focus on statistical measures to choose the features. However, these methods do not furnish an appropriate solution space. The search space size has increased exponentially corresponding to the number of features in a given data set. Traditional feature selection techniques involve larger number of features. Although all of them are not required during classification, substantial number of irrelevant and redundant features tend to affect the overall performance of the classifier.

2. Proposed model

The proposed methodology uses evolutionary algorithms for FS in order to obtain the feature subset for achieving better accuracy of classification and identification of fake reviews. It consists of four phases namely, preprocessing, feature extraction and feature subset selection using hybrid iBPSO and SFLA and classification. The block diagram of the proposed system is illustrated in Fig. 1.

2.1. Data preprocessing

The data preprocessing phase consists of four phases- tokenization, stop words removal, stemming, and SentiWordNet. First, tokenization process is applied to convert the strings into tokens. Hence, each document is divided into tokens. After the tokenization process, the stop words are eliminated from the dataset. Following this stemming is applied to select the

Download English Version:

https://daneshyari.com/en/article/6883471

Download Persian Version:

https://daneshyari.com/article/6883471

Daneshyari.com