International Journal of Innovative Computing, Information and Control
Volume **15**, Number **5**, October **2019**

ICIC International ©2019 ISSN 1349-4198
pp. 1989–2006

A HYBRID FILTER-WRAPPER FEATURE SELECTION APPROACH FOR AUTHORSHIP ATTRIBUTION

Jianbin Ma[1,*], Bing Xue[2,*] and Mengjie Zhang[2]

[1]College of Information Science and Technology
Hebei Agricultural University
No. 289, Lingyusi Street, Baoding 071001, P. R. China
*Corresponding author: majianbin@hebau.edu.cn

[2]School of Engineering and Computer Science
Victoria University of Wellington
P.O. Box 600, Wellington 6140, New Zealand
*Corresponding author: Bing.Xue@ecs.vuw.ac.nz; Mengjie.Zhang@ecs.vuw.ac.nz

Received December 2018; revised April 2019

Abstract. *Many criminals make use of the convenience of anonymity in the cyber world to conduct inappropriate or illegal activities. Authorship attribution aims to identify the most likely author from potential suspects for evidence collection and forensic investigation. Authorship attribution is typically achieved by employing classification algorithms to identify the author based on various writing-style features. However, not all features are useful (relevant) and irrelevant or redundant features may even deteriorate the classification accuracy and slow down the processing time. Feature selection as important data processing techniques can solve this problem, but they have not been investigated in authorship attribution. In this paper, we propose a novel hybrid filter-wrapper feature selection approach to authorship attribution tasks, where a rich set of writing-style features, including syntactic features, lexical features, and structural features, is extracted in order to include all available useful information. In the proposed approach, a correlation based filter feature selection method is used to filter out irrelevant features and then a particle swarm optimization based wrapper method is proposed for feature selection to further remove redundant features, select only relevant features. Experiments on real-life Blog and E-mail datasets show that the proposed approach can improve the classification performance by selecting only a small subset of features, and achieve better classification performance than filter only, wrapper only, and a commonly used wrapper method (linear forward selection).*

Keywords: Feature selection, Filter-wrapper, Particle swarm optimization, Authorship attribution, Forensic

1. **Introduction.** The rapid popularity of Internet technologies and applications provides a new way to share information over the Internet without limitation by time and space. Online mediums such as E-mail, Blog, Web Forum and Chatting Room are extensively used to distribute information with the advent of World Wide Web. Unfortunately, inappropriate or illegal activities such as antisocial information, fraud information, pornographic information, terroristic threatening information and gambling information appear on the Internet, which have strongly disturbed people's daily life [1]. A common nature of online messages is anonymity. People usually do not need to provide their real identity information such as name, age, gender, and address [2]. Therefore, it is challenging to identify cybercrime's real identity in anonymous cyberspace.

DOI: 10.24507/ijicic.15.05.1989

Lots of studies have been developed on authorship attribution based on analyzing authors' writing-style features for forensic purposes [3, 4, 68]. Most previous contributions accumulated various writing-style features and employed classification algorithms such as support vector machines (SVMs) and K-nearest neighbors (KNN) to identify authors in authorship attribution problem [2, 5, 6, 7, 8, 9, 10]. Various writing-style features including lexical features, syntactic features, structural features, and content-specific features were experimented separately or the features were combined together. Experiments were done to test the effect of different feature combinations. However, not all the writing-style features are effective. There are a large number of relevant, irrelevant and redundant features. Irrelevant and redundant features are not useful for classification and they may even reduce the classification performance due to large search space, which is termed "the curse of dimensionality" [11]. In the previous studies on authorship attribution, little attention has been given on feature selection methods that can select a subset of features and eliminate the irrelevant and redundant ones to increase the classification performance.

Feature selection is the process of selecting a subset of relevant features and removing irrelevant and redundant features. Feature selection could reduce the number of features, shorten the training time, simplify the learned classifiers, and/or improve the classification performance [12]. Especially, in authorship attribution applications for forensic purpose, it is important to assure the quality of the evidence. So, feature selection that is used for selecting effective features and increasing the identification performance is a key technique for providing convincing and credible evidence to support the authorship attribution results.

Feature selection techniques have two major approaches. One is the filter approach and the other is the wrapper approach. In filter approaches, an attribute (or attribute subset) is evaluated by only using intrinsic properties of the data [13]. Filter approaches are independent of a learning algorithm and they have the advantage of being fast and general. On the other hand, the wrapper approaches are those that use a classifier in order to assess the quality of a given attribute subset. Wrapper approaches have the advantage of achieving greater accuracy than filter approaches but with the disadvantage of being more time-consuming. So, this study tries to develop a hybrid filter-wrapper feature selection approach to take advantages of both the filter and wrapper feature selection approaches. The hybrid mechanism is more feasible in real authorship attribution applications, which usually involve a large number of features.

Correlation feature selection (CFS) [14, 15] attempts to find a subset of features such that the features in the subset are highly correlated with the class label but uncorrelated with each other. Freeman et al. [16] compared the performance of common filter measures and draw the conclusion that CFS performs quite well overall, particularly for the KNN classifier. So, we chose correlation as a filter feature selection approach in the paper.

Evolutionary computation (EC) techniques are well known for their global search ability. Particle swarm optimization (PSO) [17, 18, 19] is an efficient evolutionary computation technique. PSO optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. Compared with genetic algorithms (GAs) and genetic programming (GP), PSO is easier to implement, has fewer parameters, computationally less expensive and can converge quickly [20]. PSO has been applied to feature selection in classification widely. In our study, PSO is adopted in a wrapper based feature selection method. A PSO based feature selection method is proposed to reduce the irrelevant and redundant features.

The overall contribution of this paper is to propose a hybrid filter-wrapper feature selection approach for authorship attribution to select a smaller number of writing-style

features and achieve better classification performance than using all writing-style features. In order to achieve the overall contribution, the following four objectives will be investigated.

Objective 1. Extract an extensive set of writing-style features including syntactic features, lexical features, and structural features by taking account of English and Chinese language characteristics.

Objective 2. Employ a CFS based filter method to remove irrelevant and redundant features, and investigate whether the CFS method can select a smaller number of features and achieve better classification performance than using all features.

Objective 3. Develop a PSO based wrapper method for feature selection, and investigate whether this method can select a smaller number of features and achieve better classification performance than using all features selected by CFS.

Objective 4. Propose a hybrid filter-wrapper approach/system based on CFS and PSO for feature selection, and investigate whether this hybrid filter-wrapper approach can select a smaller number of features, and achieve better classification performance than the above methods and a benchmark wrapper method.

The remainder of the paper is organized as follows. Section 2 provides background information. Section 3 describes the writing-style feature extraction method, and the proposed hybrid filter-wrapper feature selection approach. Section 4 provides experiment design and Section 5 presents our experimental results and discussions. Section 6 provides conclusions and future work.

2. **Related Works.**

2.1. **Related work on writing-style features.**

Traditionally, fingerprints, blood, hair, witness testimony, written documents, shoe prints, videotapes and photographs are used to identify criminals. However, traditional tools and techniques may no longer be applicable in cyber criminals. Stylistics and the study of Stylometric features show that individuals can be identified by their relatively consistent writing-styles [21]. The writing-styles included unusual dictions, frequency of certain word usages and character usages, choice of rhymes, and habits of hyphenations.

Though there is no fixed feature set that is optimized and is applicable to all situations in all domains, previous studies [2, 6, 10, 66] focused on lexical, syntactic, structural and content-specific features. Each type of these features is described below.

Lexical features are word-based or character-based statistical measures on lexical variation. Some studies extracted lexical features including average sentence length, average paragraph length, vocabulary richness, word-length distribution, total number of words, total number of characters, and frequency of special characters (such as @, #, %, $). Ma et al. [1, 22] selected frequent words as the features set. Stamatatos [23] extracted the 1,000 most frequent words. Madigan et al. [24] used all the words that appeared at least twice in the corpus. Tsuboi and Matsumoto [25] used sequential word patterns and word n-grams with $n = 2$ and 3 from each sentence as the feature sets in authorship attribution studies. Iqbal et al. [21, 26] mined write-prints called frequent pattern for E-mail authorship attribution. Gómez-Adorno et al. [65] extracted n-grams including character n-grams, word n-grams and n-grams of POS tags to solve the task of cross-topic authorship attribution.

Syntactic features consist of all purpose function words, punctuations, and part-of-speech (POS) tags and hyphenations. Syntactic features extraction methods are language-dependent procedures. Function words were used to attribute the author of "The Fed-

eralist Papers" by Mosteller and Wallace [27]. Burrows [28] proposed the common high-frequency words (at least 50 strong). Zheng et al. [2] used 303 function words and 8 punctuations. Stamatatos et al. [29, 30] used style markers to analyze the text performed by an already existing natural language processing (NLP) tool using three Stylometric levels, i.e., token-level, phrase-level, and analysis-level measures for Greek language. Abbasi and Chen [10] extracted punctuations and POS tags as syntactic features. Ding et al. [4] extracted 150 function words and 9 punctuation marks. Soler-Company and Wanner [67] studied the relevance of "deep linguistic", i.e., syntactic and discourse, and solved the author and gender identification tasks.

Structural features are used to measure the overall structure and layout of documents. For instance, number of lines, number of sentences, number of paragraphs, number of sentences per paragraph, and number of words per paragraph were extracted as structural features by Zheng et al. [2]. Vel et al. [6] suggested structural features for E-mail authorship attribution. They used not only the general structural features but also specific features to e-mails such as the presence of greetings and farewell remarks and their position within the E-mail body. Abbasi and Chen [10] used the structural features such as the use of various file extensions, fonts, sizes, and colors.

Content-specific features refer to features that are dependent on specific application domains. Zheng et al. [2] used 11 content-specific features particularly for the "for-sale" topic such as "deal", "obo", and "sale". A more comprehensive list of stylistic features including idiosyncratic features was used in Abbasi and Chen [10].

Many studies [2, 21] focused on the impact of different features and feature combinations. However, the writing-style features are predefined and there are many irrelevant and redundant writing-style features. Therefore, an effective feature selection method which can select a subset of features and eliminate the irrelevant and redundant ones is needed to improve the authorship attribution performance.

2.2. **Related work on hybrid filter-wrapper feature selection.** Feature selection algorithms can be divided into two categories according to their evaluation functions: a filter approach and a wrapper approach. The filter approach employs an evaluation measure to assess the feature subset. Some well-known feature set evaluation measures are distance-based measures [31], probability-based measures [32], mutual information-based measures [33, 34], consistency measures [35, 36], correlation feature selection (CFS) [14, 15], and neighborhood-graph based measures [37]. The wrapper approach uses a learning algorithm to assess the quality of feature subsets. Some algorithms such as Sequential Forward Selection (SFS) [38], Sequential Backward Selection (SBS) [39], Genetic Programming (GP) [40], Genetic Algorithms (GAs) [41], Ant Colony Optimization (ACO) [42] and Particle Swarm Optimization (PSO) [11, 20, 43] have been employed to enhance the search ability. Xue et al. [11] showed that PSO was an effective search technique for feature selection problems.

Since simple heuristics measures are assumed to be faster than most learning algorithms, filter approaches are known to be generally faster than wrapper approaches. Since wrappers use a learning algorithm in the evaluation function, they are known to generally outperform filters approaches in terms of accuracy [44].

Hybrid filter-wrapper feature selection methods can take advantages of both filter and wrapper approaches, which is helpful in improving computational speed and prediction accuracy. Hu et al. [45] proposed a hybrid filter-wrapper method for short-term load forecasting (STLF) feature selection, which first used the Partial Mutual Information (P-MI) based filter method to filter out most of the irrelevant and redundant features and subsequently applied a firefly algorithm (FA) based wrapper method to further reducing

the redundant features without degrading the accuracy. Bermejo et al. [13] proposed an algorithm that iteratively alternated between filter ranking and wrapper feature subset selection. The wrapper only analyzes a few top ranked features. Bermejo et al. [46] presented a stochastic algorithm based on the greedy randomized adaptive search procedure (GRASP) meta-heuristic. GRASP is a multi-start constructive method which constructs a solution in its first stage, and then runs an improving stage over that solution. Yang et al. [47] proposed a filter method (information gain, IG) and a wrapper method (genetic algorithms, GAs) for feature selection in microarray data sets. Leung and Hung [48] proposed a multiple-filter-multiple-wrapper (MFMW) method that made use of multiple filters and multiple wrappers to improve the accuracy and robustness of the classification. Sebban and Nock [49] proposed a hybrid filter/wrapper method of feature selection using information theory. Zhu et al. [50] presented a hybrid wrapper and filter feature selection algorithm for a classification problem using a memetic framework. It incorporates a filter ranking method in the local search to improve classification performance and accelerate the search in identifying the core feature subsets. Gunasundari and Janakiraman [51] proposed a PSO hybridized with sequential forward selection (SFS) and sequential backward selection (SBS) algorithm in feature selection for liver cancer data. Marwa et al. [69] proposed a tri-objective hybrid filter-wrapper evolutionary algorithm for feature selection that optimized two filter objectives and one wrapper objective.

2.3. **Related work on PSO for feature selection.** As an EC technique, PSO has recently gained more attention for solving feature selection problems [20]. Xue et al. [43] proposed an archive based PSO feature selection method which introduced an external archive to store promising solutions obtained during the search process. Xue et al. [11] presented three new initialization strategies and three new personal best and global best updating mechanisms in PSO to develop feature selection methods. Xue et al. [52, 53] focused on multi-objective PSO for feature selection.

Daliri [54] proposed a feature selection strategy using binary PSO for the diagnosis of different medical diseases. The SVMs are used for the fitness function of the binary PSO. The results are better than traditional feature selection methods, namely, the F-score and the information gain. Compared to GAs, the proposed method shows a higher accuracy in most datasets.

Chen et al. [55] proposed an improved PSO using the opposite sign test (OST), which increased population diversity in PSO, and avoided local optimal trapping by improving the jump ability of flying particles. Liu et al. [56] formulated four rules by introducing the mechanism for survival of the fittest. A modified multi-swarm PSO (MSPSO) to solve discrete problems was designed. MSPSO consists of a number of sub-swarms and a multi-swarm scheduler that can monitor and control each sub-swarm using the rules.

Ghamisi and Benediktsson [57] proposed a feature selection method that integrated GAs and PSO. SVMs were used to evaluate the fitness value. Wang et al. [58] proposed a feature selection algorithm based on improved PSO and rough set theory. Chakraborty [59] compared the performance of PSO with that of GAs with a fuzzy set based fitness function. The results show that PSO performs better than GAs in terms of the classification performance. Lin et al. [60] and Huang and Dun [61] proposed feature selection algorithms which could search for the best feature subset and optimize the parameters in SVMs. Tran et al. [70] proposed a variable-length PSO representation for feature selection.

Most authorship attribution studies accumulate various writing-style features and test different feature combinations on effect of experimental results. However, little attention has been given on feature selection methods that can select a subset of features and

eliminate the irrelevant and redundant ones. A variety of PSO based feature selection methods have been proposed and shown that PSO is an efficient search technique for feature selection. However, little work has been conducted for combining filter feature selection and PSO based wrapper feature selection.

3. The Proposed Method.

3.1. Correlation feature selection (CFS).

The CFS algorithm evaluates and ranks feature subsets rather than individual features. CFS measure evaluates subsets of features on the basis of the following hypothesis: "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other" [14, 15]. The CFS algorithm is a heuristic for evaluating the merit of subset of features. The following equation gives the heuristic merit of a feature subset S consisting of k features.

$$Merit_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{1}$$

where $Merit_S$ is the heuristic "merit" of a feature subset S containing k features. $\overline{r_{cf}}$ is the average feature-class correlation. $\overline{r_{ff}}$ is the average feature-feature inter-correlation.

Based on Equation (1), irrelevant features should be ignored because they will have low correlation with the class label. Redundant features should be screened out as they will be highly correlated with one or more of the remaining features [14].

3.2. Particle swarm optimization.

Particle swarm optimization (PSO) is an evolutionary computation technique proposed by Kennedy and Eberhart [17, 18] and was first intended for simulating social behaviors such as bird flocking or fish schooling.

PSO is a computational method that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. In PSO, a population is called swarm and the candidate solutions are called particles in the swarm. The particles move in the search space to search for the optimal solutions by updating the position and velocity of each particle in the population. The current position of particle i is represented as $x_i = \{x_{i1}, x_{i2}, \ldots, x_{iD}\}$, where D is the dimensionality of the search space. The velocity of particle i is represented as $v_i = \{v_{i1}, v_{i2}, \ldots, v_{iD}\}$. During the movement, each particle updates its position and velocity according to its best known position and the entire swarm's best known position. The best known position of the particle is denoted as the personal best *pbest*, and the entire swarm's best known position is recorded as the global best *gbest*. The position and velocity of each particle are updated by the following equations.

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \tag{2}$$

$$v_{id}^{t+1} = w \times v_{id}^{t} + c_1 \times r_{1i} \times \left(p_{id} - x_{id}^{t}\right) + c_2 \times r_{2i} \times \left(p_{gd} - x_{id}^{t}\right) \tag{3}$$

where t denotes the tth iteration. $d \in D$ denotes the dth dimension in the search space. w is inertia weight. c_1 and c_2 are acceleration constants. r_1 and r_2 are random values uniformly distributed in $[0, 1]$. p_{id} and p_{gd} are the elements of *pbest* and *gbest* in the dth dimension. The velocity is limited by a predefined maximum velocity $v_{\max}$ and $v_{id}^{t+1} \in [-v_{\max}, v_{\max}]$.

3.3. Overview.

The framework of the proposed hybrid filter-wrapper feature selection approach to authorship attribution, as shown in Figure 1, can be divided into four steps.

Step 1. Messages Collection

Let us suppose there are inappropriate or illegal messages that cause bad effect. Investigators need to collect a set of online messages written by potential suspects to profile the writing styles of each author. By preliminary investigation, investigators should narrow

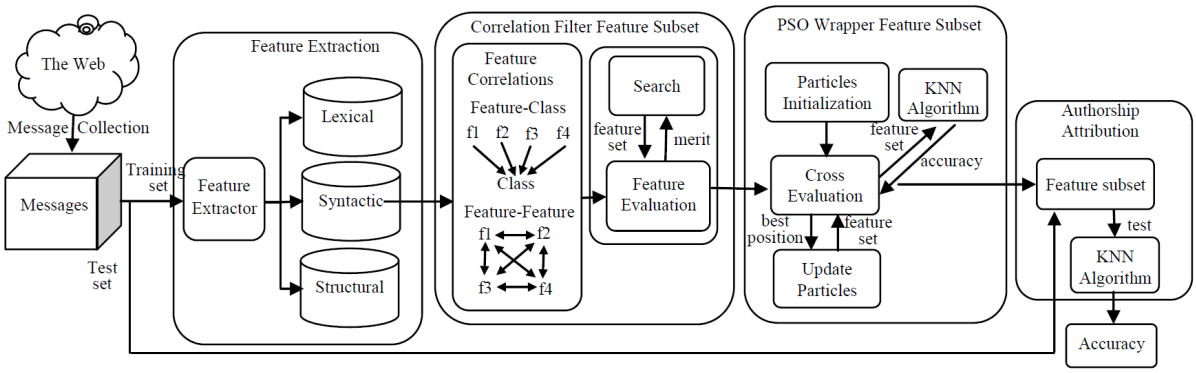


FIGURE 1. The framework of the proposed hybrid filter-wrapper feature selection approach to authorship attribution

down their suspect list. In our research, we divide the online message into two portions. One portion is training set which is used to find optimal feature subset. Then, the training set is used to obtain the authorship attribution model. Another portion is the test set, which is used to validate the effectiveness of our method.

Step 2. Feature Extraction

Online messages on the Web are in unstructured text format. Based on the predefined writing-style features, the feature extractor analyzes the messages and extracts the writing-style features in textual online messages [2]. The messages are transformed to a vector of writing-style features. In this paper, a rich set of writing-style features including lexical features, syntactic features, and structural features is extracted by taking account of English and Chinese language characteristics (details in Section 3.4).

Step 3. Hybrid Filter-Wrapper Feature Selection

In feature extraction step, the feature set might contain hundreds or even thousands of features. Feature selection has two objectives. The first is to identify relevant features, which can increase the performance. The second is to identify a small set of features with minimum redundancy, which can reduce the computational cost without reducing the performance. To tackle these two objectives, a hybrid filter-wrapper feature selection approach to authorship attribution is proposed. A CFS method is used to eliminate a large number of irrelevant and redundant features. Then, with the reduced feature subset obtained by the filter method, a PSO based wrapper method is applied to further reducing the irrelevant and redundant features and increasing the performance. So, the proposed hybrid filter-wrapper approach takes advantages of the filter method's efficiency and wrapper method's accuracy to complement each other's shortcomings [45].

Step 4. Authorship Attribution

The authorship attribution model is developed after training on training set at the reduced feature subset. The inappropriate or illegal online messages are attributed to the authorship by the authorship attribution model, which provide evidence for investigators.

3.4. **Writing-style features.** Table 1 shows our predefined writing-style features. In view of English and Chinese language characteristics, we extract 583 writing-style features for Chinese datasets and 383 writing-style features for English datasets. There are 3 specific structural features for E-mail datasets.

Syntactic features are divided into two categories: function words and punctuations features. 500 function words features for Chinese datasets and 300 function words features for English datasets are extracted. The function words reflect the author's preference or habit for usage of some specific words. The feature values of function words are calculated

Table 1. Writing-style features

Group	Feature types	Features
group 1	Syntactic features	1. Function words (500 features for Chinese datasets, 300 features for English datasets)
group 2	Syntactic features	2. Chinese and English punctuations (30 features for Chinese datasets, 8 features for English datasets) 3. Part-of-speech features (12 features for Chinese datasets, 8 features for English datasets)
	Lexical features	4. Number of distinct punctuations/total number of punctuations 5. Total number of characters (C) 6. Number of digital characters/C 7. Number of lowercase letters/C 8. Number of uppercase letters/C 9. Number of space/C 10. Number of tab/C 11. Mean sentence length in terms of words 12. Mean sentence length in terms of characters 13. Number of special characters/C:<,>% | { }[ ] / \@ # ∼ + − * = # & ˆ$ (21 features) 14. Number of alphabets (A-Z)/C (26 features for English datasets) 15. Total number of words (M) 16. Number of distinct words/M 17. Hapax legomena 18. Hapax dislegomena
	Structural features	19. Total number of lines (L) 20. Total number of sentences 21. Total number of paragraphs 22. Mean paragraph length in terms of words 23. Mean paragraph length in terms of characters 24. Number of blank lines/L 25. Number of indents/C 26. Has a greeting acknowledgment 27. Uses a farewell acknowledgment 28. Contains signature text (26-28 for E-mail datasets)

by traditional *tf-idf* Formula [62].

$$F\left(t,\vec{d}\right) = tf\left(t,\vec{d}\right) \times \log(N/n_t + 0.01) \tag{4}$$

where $F\left(t,\vec{d}\right)$ is the feature value of function word t in document $\vec{d}$, $tf\left(t,\vec{d}\right)$ is the frequency of function word t in document $\vec{d}$, N is the total number of documents, and n_t is the number of documents that contain function word t.

The web text is inputted by keyboard. To save time, authors always ignore the difference of Chinese and English punctuations. Chinese web texts have more punctuations categories than English web texts. So, we extract 30 punctuations for Chinese dataset and 8 punctuations for English dataset. The feature value of a punctuation feature is the ratio of number of the punctuations to the total number of punctuations in the document. The part-of-speech features reflect the preference for word class usage. For example, some authors always use exclamation; however, some authors hardly ever. POS tagging software is used to tag each word of web text. 12 part-of-speech features for Chinese dataset and 8 part-of-speech features for English dataset are extracted. The feature values of part-of-speech features are the ratio of the number of the part-of-speech to total number of part-of-speech in the document.

Lexical features involve character-based and word-based features. We extracted character-based and word-based lexical features used in [2, 5, 21]. The feature values of features indexed 6-10, 13-14 and 16 are normalized to $[0, 1]$ by dividing the maximum of the feature in the document. The feature values of features indexed 5 and 15 are the ratio of number of features to the maximum of the features in the training set. Vocabulary richness is the richness of author's vocabulary. Hapax legomena and Hapax dislegomena are the features used for once-occurring and twice-occurring words. Feature indexed 14 is removed from Chinese dataset, because Chinese dataset has few English letters. In total, we extracted 34 lexical features for the Chinese datasets and 60 lexical features for the English datasets.

Structural features represent the way an author organizes the layout of the web text. We extracted 10 structural features used in [2, 5]. Features indexed 26-28 are specific for E-mail datasets and the values of the features are binary values. The feature values of features indexed 19-21 are normalized to $[0, 1]$ by dividing the maximum of features in the training set.

In this paper, our works are not domain-specific, and the content of datasets selected (described in Section 4) does not aim at specific application domains. So, we have not extracted *content-specific features* (mentioned in 2.1).

3.5. **CFS.** Equation (1) calculates the correlation merit of a feature subset. The purpose of feature selection is to find a feature subset that has the highest evaluation value. If there are n features, then there are 2^n possible subsets. So, effective heuristic search strategy is needed.

In this paper, best first search [15] was used to search the feature subset. The best first search method starts with an empty set of features and generates all possible single feature expansions. The search progresses forward through the search space by adding single features. To prevent the best first search from exploring the entire feature subset search space, the search will terminate if five consecutive fully expanded subsets show no improvement over the current best subset [14].

In consideration of feature selection in the next step, all the writing-style features were divided into two groups as shown in Table 1. The two groups were filtered by CFS method separately and merged together afterwards. The majority of writing-style features are function words features and there are a large number of irrelevant and redundant features. If all the features are mixed together and filtered by the CFS method, some relevant features might be filtered out by mistake. The purpose of our design is to remove irrelevant and redundant features as much as possible and protect against removing relevant features by mistake. The procedure of feature filtering for different group features is shown in Figure 2.

3.6. **A PSO based feature selection method.** After the CFS step, a PSO based feature selection method was adopted to further remove the irrelevant and redundant features and increase the performance.

In this paper, a continuous PSO algorithm is presented for solving feature selection problem. The PSO algorithm for solving the feature selection problems considers each particle based on three key vectors, namely position (x_i), velocity (v_i), and training examples (e_j). $x_i = \{x_{i1}, x_{i2}, \ldots, x_{iN}\}$ denotes the ith position vector in the swarm, where N is the total number of features, and x_{id} is the position value of the ith particle with respect to the dth dimension. $v_i = \{v_{i1}, v_{i2}, \ldots, v_{iN}\}$ denotes the ith velocity vector in the swarm, where v_{id} is the velocity value of the ith particle with respect to the dth dimension. $e_j = \{f_{j1}, f_{j2}, \ldots, f_{jN}\}$ denotes the jth example in the training examples set, where N is the total number of features, and f_{jd} is the feature value of the jth example with respect

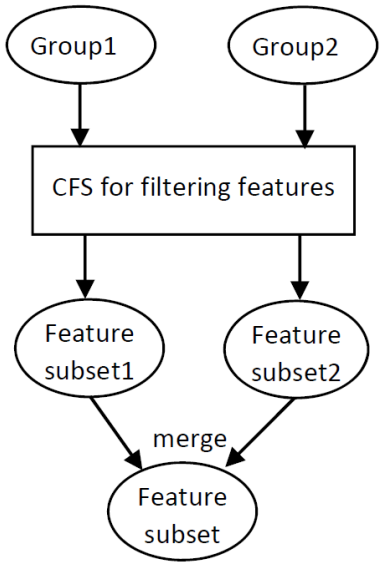


FIGURE 2. The procedure of feature filtering for different group features

to the dth dimension. Each training example e_j has its class label. The dimension of position vector is equal to the number of features in training examples. Each element of position vector corresponds to one feature in training examples and determines whether the feature is selected or not.

Initially, the position values x_{id} and velocity values v_{id} are generated randomly and uniformly as continuous sets of values between $[0, 1]$. A threshold β is used to compare with the value x_{id} in the position vector. If $x_{id} > \beta$, feature d is selected; otherwise, feature d is not selected.

The fitness function, which is to minimize the classification error rate obtained by the selected features, is used in the proposed algorithms and shown by Formula (5).

$$Fitness = ErrorRate = \frac{FP + FN}{TP + TN + FP + FN} \tag{5}$$

where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, respectively.

Since the range of feature values varies widely, in some machine learning algorithms, the objective function will not work properly without normalization. Normalization is a method used to standardize the range of independent features and normalize the feature vectors to a unit vector. Formula (6) is used to normalize the feature vectors of training examples.

$$\hat{\vec{e_j}} = \frac{\vec{e_j}}{\|\vec{e_j}\|} \tag{6}$$

where $\vec{e_j}$ is one feature vector. $\|\vec{e_j}\|$ is the length of $\vec{e_j}$. $\vec{e_j}$ is transformed to unit vector $\hat{\vec{e_j}}$.

In this paper, we define the following terms to facilitate readability (particularly for Sections 4 and 5). PSOS represents that PSO is used for feature selection and the threshold $\beta > 0$. PSOS can be applied on all the original available features as a wrapper method. They can also be applied on the features selected by CFS, which is the proposed hybrid filter-wrapper approach, named CFS-PSOS.

Algorithm 1 shows the pseudo-code of the proposed PSO based feature selection (PSOS) algorithm.

Algorithm 1. Pseudo-Code of PSOS

```
begin
    divide database into Training set and Test set;
    randomly initialize the position and velocity of each particle in the swarm;
    while Maximum Iterations is not reached do
        for i = 1 to Population Size do
            select features according to threshold β value;
            normalize the Training set according to Formula (6);
            evaluate the fitness of each particle on Training set according to
            Formula (5);
        end
        for i = 1 to Population Size do
            update the velocity of particle i;
            update the position of particle i;
        end
    end
    return the swarm's best known position (gbest);
    adjust feature values for Test set by the position value of gbest
    according to Formula (5);
    calculate the accuracy of the selected feature subset on the Test set;
    return the test classification accuracy;
end
```

4. **Experiment Design.** In this section, experiments are designed to examine the performance of our proposed hybrid filter-wrapper feature selection approach on authorship attribution. The overall experimental objective is to verify whether our method can select a smaller number of writing-style features and achieve better results than using all writing-style features. Several experiments are performed to test the overall experimental objective. (1) The first experiment is to test whether CFS method can select a smaller number of features and achieve better classification performance than using all features. (2) The second experiment is to test whether PSO based feature selection method (PSOS) can select a smaller number of features and achieve better performance than using all features selected by CFS. Experimental results of CFS-PSOS with different β values are tested. (3) The third experiment is to compare the experimental results of the different feature selection methods (CFS and PSOS) with the proposed hybrid filter-wrapper feature selection approach (CFS-PSOS) and test whether CFS-PSOS is effective. (4) The fourth experiment is to compare our proposed hybrid filter-wrapper feature selection approach (CFS-PSOS) with a traditional wrapper based feature selection method.

The traditional wrapper based feature selection method is linear forward selection (LFS), which is adopted as a benchmark technique to compare. LFS is a feature selection technique to reduce the number of attributes expansions in each forward selection step, which can reduce computational cost and even increase the accuracy compared to standard forward selection method [63].

Two real-life datasets including Blog and E-mail were used in the experiments. The detailed information of the two datasets was shown in Table 2. There are not public Chinese datasets for online messages. So, the Blog dataset written in Chinese consists of Blog information collected from 12 most popular Bloggers on the website http://blog.sina.com.cn. We selected 200 instances for each author. The E-mail dataset written in English was

Table 2. Description of the two datasets

Dataset	Language	No. of authors	Average no. of instances per author	#features
Blog	Chinese	12	200	583
E-mail	English	15	100	386

gained from Enron E-mail Dataset [64], which contains 200,399 E-mails of about 150 employees of Enron Corporation. We randomly selected 15 authors from the Enron E-mail Dataset. For each author, we selected 100 E-mails. "#features" represents the number of features.

For the two datasets, the instances in each dataset were randomly divided into two sets: 70% as the training set, and 30% as the test set [11, 20]. The classification accuracy was used to evaluate the experimental results.

In the wrapper feature selection method, a simple and commonly used learning algorithm, K-nearest neighbors (KNN) was chosen in the experiments and k was set to 5. To obtain more accurate performance, 10-fold cross-validation is applied to evaluating the performance of selected feature subset on training set [11, 20].

In all the PSO-based algorithms, the population size is 30, the maximum iteration is 100, the inertia weight $w = 0.7298$, and the acceleration constants $c_1 = c_2 = 1.49618$. These values are common settings in some literature such as [11, 20], and are proved to be effective in their experiments. The parameter β was used to adjust the number of selected feature subset. The particles in swarm are initialized randomly. To avoid the effect of particles initialization on experimental results and achieve more reasonable results, we initialize 30 different swarms randomly. Each of the PSO based methods has been run 30 times independently. Since CFS is a deterministic method, the CFS is run once.

The benchmark technique, LFS is experimented on Waikato Environment for Knowledge Analysis (Weka). All the parameters in LFS were set to the defaults. The evaluation classifier was KNN and k was set to 5 to keep consistent with CFS-PSOS for fair comparisons.

5. Experimental Results and Discussion.

5.1. The experimental results of CFS.

In this experiment, we tested whether CFS could achieve better performance than using original features. In Table 3, "All" shows the results of using all available features for authorship attribution. "Acc" represents the accuracy. The experimental results of CFS are shown in Table 3.

According to Table 3, the numbers of original features of Blog and E-mail datasets are 583 and 386 respectively, while the numbers of CFS features of Blog and E-mail datasets are only 93 and 25 respectively which are reduced to 15.95% and 6.48% of the original features. The Acc of All on Blog and E-mail dataset are 45.95% and 56.67% respectively. The Acc of CFS on Blog and E-mail dataset are increased by 30.21% and

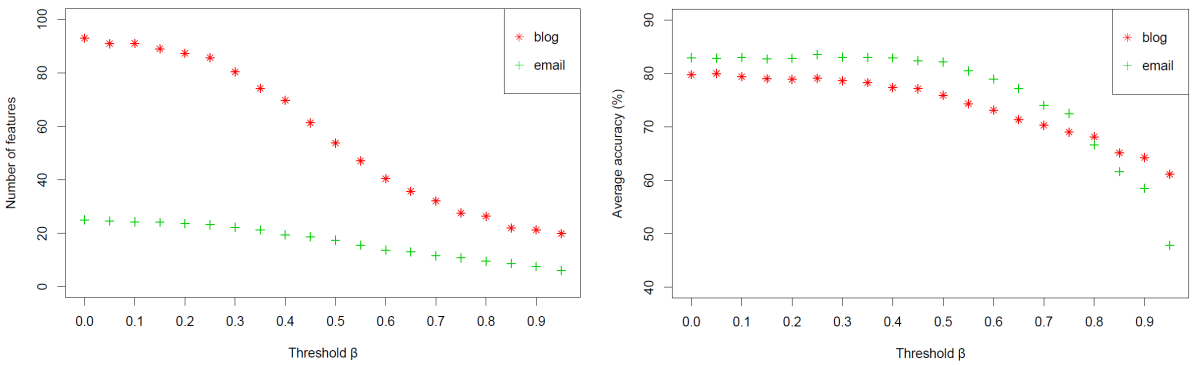
Table 3. Experimental results of CFS

Database	All		CFS	
	#features	Acc (%)	#features	Acc (%)
Blog	583	45.95	93	76.16
E-mail	386	56.67	25	81.33

24.66% respectively compared with All. The experimental results show that CFS can select a smaller feature subset and achieve better performance than using all features.

5.2. **The experimental results of CFS-PSOS with different β values.** In CFS-PSOS, CFS can filter out most of irrelevant and redundant features. Then, PSOS is applied on the features selected by CFS and further reduces the irrelevant and redundant features. The threshold β for PSOS can be used to control the number of selected features. If more features are selected, there are still irrelevant and redundant features in the feature subset. However, if fewer features are selected, some features with the discrimination ability will be deleted. How many features PSOS selected are appropriate? Whether or not a smaller number of selected features can achieve better performance than using all features? To verify these issues, we set 20 different β values. The experiment on PSOS was run for 30 (runs) $\times$ 20 (β values) $\times$ 2 (datasets) = 1200 times. The numbers of features with different β values in CFS-PSOS are shown in Figure 3(a). The average accuracies with different β values in CFS-PSOS are shown in Figure 3(b). According to Figure 3(a), the number of selected features on both Blog and E-mail datasets becomes smaller with a larger threshold β. However, according to Figure 3(b), the average accuracy on Blog dataset does not change obviously in the range of threshold β between 0 and 0.25. The average accuracy on E-mail dataset has slight oscillations in the range of threshold β between 0 and 0.5, and the average accuracy reaches maximum when the threshold β is 0.25. The maximum average accuracy on E-mail dataset reaches 83.5%, which is higher than the Acc of CFS on E-mail dataset (shown in Table 3). In general, the average accuracy on E-mail dataset keeps stable in the range of threshold β between 0 and 0.5. The average accuracies on Blog and E-mail dataset fall sharply after the threshold β exceeds a certain value. So, setting appropriate threshold β is important. Higher threshold β will remove some useful features, and the performance will be worse. Smaller threshold β will reserve more features, and there are still irrelevant and redundant features. We can draw the conclusion that PSOS can remove a few irrelevant and redundant features selected by CFS, select a smaller number of features and achieve better (or comparable) performance than using all features on the appropriate range of the threshold β values.

In summary, according to Table 3 and Figure 3, the proposed hybrid filter-wrapper feature selection approach (CFS-PSOS) can select a smaller feature subset and achieve better classification performance than using all the available features.



(a) The number of features (b) The average accuracy

FIGURE 3. The performance with different threshold β values in CFS-PSOS

5.3. **Comparisons of experimental results between CFS, PSOS and CFS-PSOS.** Our proposed hybrid filter-wrapper feature selection approach (CFS-PSOS) can take advantages of both filter and wrapper feature selection approaches, and improve computational speed and prediction accuracy simultaneously. To test the effectiveness of the hybrid filter-wrapper feature selection method, it is compared with the filter method only (CFS) and a wrapper method using PSOS on all available features. PSOS was run 30 times. "Asize", "Aacc" and "Atime" represent the average number of features, average accuracy and average computational time in the 30 runs respectively. The CFS was run once since it is a deterministic method. The numbers of features, accuracy and computational time are represented by "Asize", "Aacc" and "Atime", too. All the threshold β for Blog and E-mail dataset was set to 0.25 based on Section 5.2. The experimental results are shown in Table 4.

Table 4. Experimental results of different feature selection methods (#features, Blog dataset: 583, E-mail: 386)

Method	Asize		Aacc (%)		Atime (Minutes)	
	Blog	E-mail	Blog	E-mail	Blog	E-mail
CFS	93	25	76.16	81.33	0.15	0.13
PSOS	526.6	355.2	66.01	72.38	181.5	15.37
CFS-PSOS	85.7	23.23	79.08	83.50	27.11	2.60

From Table 4, we can see that the CFS method used much less time than the PSOS method. The number of selected features relates to threshold β in PSOS. The PSOS method at the given threshold $\beta = 0.25$ value selected much more features than the CFS method. Since the features selected by the PSOS method have a large number of irrelevant and redundant features, the average accuracy of the PSOS method is worse than that of the CFS method. However, as can be seen from Figure 3, the β values can significantly influence the number of features and the classification performance. Therefore, by using a good β value, the PSOS method may be able to further reduce the number of features and improve the classification performance. The CFS-PSOS method outperformed PSOS not only on average number of features but also on average accuracy and computational time. The CFS-PSOS method can select smaller number of features and achieve better performance than the CFS method. The computational time of the CFS-PSOS method is more than that of the CFS method. However, performance is more important than computational time for authorship attribution. So, the proposed hybrid filter-wrapper approach can take advantages of both the filter and wrapper methods and achieve better results.

5.4. **Comparisons between CFS-PSOS with LFS.** LFS is selected as a benchmark technique to compare. The comparison between the proposed hybrid filter-wrapper feature selection approach (CFS-PSOS) and LFS is shown in Table 5.

Asize, Aacc represent the average number of features, average accuracy in the 30 runs in CFS-PSOS respectively. LFS is a deterministic method and it is run once. The number

Table 5. Comparisons of experimental results between CFS-PSOS and LFS

Method	Asize		Aacc (%)	
	Blog	E-mail	Blog	E-mail
LFS	18	22	68.88	81.63
CFS-PSOS	85.7	23.23	79.08	83.50

of features and accuracy are represented by Asize, Aacc, too. All the threshold β for Blog and E-mail dataset was set to 0.25 according to Section 5.2.

From Table 5, we can see that Aacc of the proposed CFS-PSOS method on Blog dataset is better than that of the LFS method obviously. Asize of the proposed CFS-PSOS method on Blog dataset is more than that of the LFS method. For E-mail dataset, the CFS-PSOS method outperformed the LFS method in terms of Aacc slightly. Asize of the proposed CFS-PSOS method is a little more than that of the LFS method. The results suggest that the LFS method has removed many useful features and resulted in lower performance. For authorship attribution application, performance is the primary criterion. So, the proposed hybrid filter-wrapper feature selection approach achieves better experimental results for authorship attribution than LFS technique.

6. **Conclusions.** In this paper, a hybrid filter-wrapper feature selection approach (i.e., CFS-PSOS) on authorship attribution was proposed. Firstly, a correlation based filter feature selection method was used to remove most irrelevant and redundant writing-style features. Then, PSO based wrapper feature selection method was employed to further remove irrelevant and redundant writing-style features. Experiments showed that the CFS-PSO feature selection method could select a smaller number of features and achieve better (or comparable) performance than using all features on the appropriate range of threshold β values, outperformed CFS on average accuracy and the number of features, and outperformed PSOS in terms of the number of features, the classification performance, and the computational time. Compared with the traditional linear forward selection (LFS), the proposed CFS-PSOS approach achieved better performance for authorship attribution tasks.

This study extracted a small number of writing-style features. In the future, we will extract larger number of features and find feature sets that are specific for particular datasets. We will conduct further works on feature selection methods that combine other filter approaches with PSO to reduce the number of features, increase the performance, and compare the effectiveness with the proposed CFS-PSOS approach in this study. We will investigate PSO based multi-objective feature selection method that can maximize classification performance and minimize the number of features simultaneously. We also intend to investigate different parameter settings of PSO such as initial position and velocity of particles, inertial weight and acceleration constants on experimental results.

Acknowledgments. This work was supported by grants from Hebei Agricultural University (ZD201702, LG201707).

REFERENCES

[1] J. Ma, Y. Li and G. Teng, CWAAP: An authorship attribution forensic platform for Chinese web information, *Journal of Software*, vol.9, no.1, pp.11-19, 2014.

[2] R. Zheng, J. Li, H. Chen and Z. Huang, A framework for authorship identification of online messages: Writing-style features and classification techniques, *Journal of the American Society for Information Science and Technology*, vol.57, no.3, pp.378-393, 2006.

[3] M. Brennan, S. Afroz and R. Greenstadt, Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity, *ACM Trans. Information and System Security (TISSEC)*, vol.15, no.3, pp.1-12, 2012.

[4] S. H. H. Ding, B. C. M. Fung and M. Debbabi, A visualizable evidence-driven approach for authorship attribution, *ACM Trans. Information and System Security (TISSEC)*, vol.17, no.3, pp.1-30, 2015.

[5] O. de Vel, Mining E-mail authorship, *Proc. of ACM International Conference on Knowledge Discovery and Data Mining (KDD2000)*, 2000.

[6] O. de Vel, A. Anderson, M. Corney and G. Mohay, Mining E-mail content for author identification forensics, *ACM Sigmod Record*, vol.30, no.4, pp.55-64, 2001.

[7] O. de Vel, A. Anderson, M. Corney and G. Mohay, Multi-topic E-mail authorship attribution forensics, *Proc. of ACM Conference on Computer Security – Workshop on Data Mining for Security Applications*, 2001.

[8] R. Zheng, Y. Qin, Z. Huang and H. Chen, Authorship analysis in cybercrime investigation, *Proc. of the 1st International Symposium on Intelligence and Security Informatics*, pp.59-73, 2003.

[9] A. Abbasi and H. Chen, Visualizing authorship for identification, *Proc. of IEEE International Conference on Intelligence and Security Informatics*, pp.60-71, 2006.

[10] A. Abbasi and H. Chen, Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace, *ACM Trans. Information Systems (TOIS)*, vol.26, no.2, pp.1-29, 2008.

[11] B. Xue, M. J. Zhang and W. N. Browne, Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms, *Applied Soft Computing*, vol.18, pp.261-276, 2014.

[12] M. Dash and H. Liu, Feature selection for classification, *Intelligent Data Analysis*, vol.1, no.1, pp.131-156, 1997.

[13] P. Bermejo, L. de la Ossa, J. A. Gámez and J. M. Puerta, Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking, *Knowledge-Based Systems*, vol.25, no.1, pp.35-44, 2012.

[14] M. A. Hall, *Correlation-Based Feature Selection for Machine Learning*, Ph.D. Thesis, The University of Waikato, 1999.

[15] M. A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, *Proc. of the 17th International Conference on Machine Learning*, pp.359-366, 2000.

[16] C. Freeman, D. Kulic and O. Basir, An evaluation of classifier-specific filter measure performance for feature selection, *Pattern Recognition*, vol.48, no.5, pp.1812-1826, 2015.

[17] J. Kennedy and R. C. Eberhart, Particle swarm optimization, *Proc. of IEEE International Conference on Neural Networks*, vol.4, pp.1942-1948, 1995.

[18] J. Kennedy and R. C. Eberhart, A discrete binary version of the particle swarm algorithm, *Proc. of IEEE International Conference on Systems, Man, and Cybernetics – Computational Cybernetics and Simulation*, vol.5, pp.4104-4108, 1997.

[19] Y. H. Shi and R. C. Eberhart, A modified particle swarm optimizer, *Proc. of IEEE International Conference on Evolutionary Computation*, pp.69-73, 1998.

[20] B. Xue, M. Zhang and W. N. Browne, Particle swarm optimization for feature selection in classification: A multi-objective approach, *IEEE Trans. Cybernetics*, vol.43, no.6, pp.1656-1671, 2013.

[21] F. Iqbal, H. Binsalleeh, B. C. Fung and M. Debbabi, Mining writeprints from anonymous e-mails for forensic investigation, *Digital Investigation*, vol.7, no.1, pp.56-64, 2010.

[22] J. Ma, Y. Li, G. Teng and Y. Zhang, An authorship attribution forensic method for web information, *ICIC Express Letters*, vol.7, no.9, pp.2609-2613, 2013.

[23] E. Stamatatos, Authorship attribution based on feature set subspacing ensembles, *International Journal on Artificial Intelligence Tools*, vol.15, no.5, pp.823-838, 2006.

[24] D. Madigan, A. Genkin, D. D. Lewis, S. Argamon, D. Fradkin and L. Ye, Author identification on the large scale, *Proc. of the Joint Meeting of the Interface and the Classification Society of North America*, 2005.

[25] Y. Tsuboi and Y. Matsumoto, Authorship identification for heterogeneous documents, *IPSJ SIG Notes*, pp.17-24, 2002.

[26] F. Iqbal, R. Hadjidj, B. C. M. Fung and M. Debbabi, A novel approach of mining write-prints for authorship attribution in E-mail forensics, *Digital Investigation*, vol.5, pp.S42-S51, 2008.

[27] F. Mosteller and D. L. Wallace, *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, 1964.

[28] J. F. Burrows, Word-patterns and story-shapes: The statistical analysis of narrative style, *Literary and Linguistic Computing*, vol.2, no.2, pp.61-70, 1987.

[29] E. Stamatatos, N. Fakotakis and G. Kokkinakis, Automatic text categorization in terms of genre and author, *Computational Linguistics*, vol.26, no.4, pp.471-495, 2000.

[30] E. Stamatatos, N. Fakotakis and G. Kokkinakis, Computer-based authorship attribution without lexical measures, *Computers and the Humanities*, vol.35, no.2, pp.193-214, 2001.

[31] K. Kira and L. A. Rendell, The feature selection problem: Traditional methods and a new algorithm, *Proc. of the 10th National Conference on Artificial Intelligence*, vol.2, pp.129-134, 1992.

[32] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*, Prentice-Hall London, 1982.

[33] D. J. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.
[34] G. Brown, A new perspective for information theoretic feature selection, *Proc. of International Conference on Artificial Intelligence and Statistics*, pp.49-56, 2009.
[35] H. Almuallim and T. G. Dietterich, Learning with many irrelevant features, *Proc. of National Conference on Artificial Intelligence*, pp.547-552, 1991.
[36] M. Dash and H. Liu, Consistency-based search in feature selection, *Artificial Intelligence*, vol.151, no.1, pp.155-176, 2003.
[37] X. F. He, D. Cai and P. Niyogi, Laplacian score for feature selection, *Proc. of Advances in Neural Information Processing Systems*, pp.507-514, 2005.
[38] A. W. Whitney, A direct method of nonparametric measurement selection, *IEEE Trans. Computers*, vol.C-20, no.9, pp.1100-1103, 1971.
[39] T. Marill and D. M. Green, On the effectiveness of receptors in recognition systems, *IEEE Trans. Information Theory*, vol.9, no.1, pp.11-17, 1963.
[40] K. Neshatian and M. J. Zhang, Dimensionality reduction in face detection: A genetic programming approach, *Proc. of International Conference on Image and Vision Computing*, pp.391-396, 2009.
[41] B. Chakraborty, Genetic algorithm with fuzzy fitness function for feature selection, *Proc. of the 2002 IEEE International Symposium on Industrial Electronics*, vol.1, pp.315-319, 2002.
[42] H. Ming, A rough set based hybrid method to feature selection, *Proc. of International Symposium on Knowledge Acquisition and Modeling*, pp.585-588, 2008.
[43] B. Xue, A. K. Qin and M. J. Zhang, An archive based particle swarm optimisation for feature selection in classification, *Proc. of IEEE Congress on Evolutionary Computation*, pp.3119-3126, 2014.
[44] Ö. Uncu and I. B. Türkşen, A novel feature selection approach: Combining feature wrappers and filters, *Information Sciences*, vol.177, no.2, pp.449-466, 2007.
[45] Z. Y. Hu, Y. K. Bao, T. Xiong and R. Chiong, Hybrid filter-wrapper feature selection for short-term load forecasting, *Engineering Applications of Artificial Intelligence*, vol.40, pp.17-27, 2015.
[46] P. Bermejo, J. A. Gámez and J. M. Puerta, A grasp algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets, *Pattern Recognition Letters*, vol.32, no.5, pp.701-711, 2011.
[47] C.-H. Yang, L.-Y. Chuang and C. H. Yang, IG-GA: A hybrid filter/wrapper method for feature selection of microarray data, *Journal of Medical and Biological Engineering*, vol.30, no.1, pp.23-28, 2010.
[48] Y. Leung and Y. Hung, A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification, *IEEE/ACM Trans. Computational Biology and Bioinformatics (TCBB)*, vol.7, no.1, pp.108-117, 2010.
[49] M. Sebban and R. Nock, A hybrid filter/wrapper approach for feature selection using information theory, *Pattern Recognition*, vol.35, no.4, pp.835-846, 2002.
[50] Z. Zhu, Y.-S. Ong and M. Dash, Wrapper-filter feature selection algorithm using a memetic framework, *IEEE Trans. Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol.37, no.1, pp.70-76, 2007.
[51] S. Gunasundari and S. Janakiraman, A hybrid PSO-SFS-SBS algorithm in feature selection for liver cancer data, *Power Electronics and Renewable Energy Systems*, pp.1369-1376, 2015.
[52] B. Xue, L. Cervante, L. Shang, W. N. Browne and M. J. Zhang, A multiobjective particle swarm optimisation for filter-based feature selection in classification problems, *Connection Science*, vol.24, nos.2-3, pp.91-116, 2012.
[53] B. Xue, M. J. Zhang and W. N. Browne, Multi-objective particle swarm optimisation (PSO) for feature selection, *Proc. of the 14th Annual Conference on Genetic and Evolutionary Computation*, pp.81-88, 2012.
[54] M. R. Daliri, Feature selection using binary particle swarm optimization and support vector machines for medical diagnosis, *Biomedizinische Technik/Biomedical Engineering*, vol.57, no.5, pp.395-402, 2012.
[55] L. F. Chen, C. T. Su and K. H. Chen, An improved particle swarm optimization for feature selection, *Intelligent Data Analysis*, vol.16, no.2, pp.167-182, 2012.
[56] Y. N. Liu, G. Wang, H. L. Chen, H. Dong, X. D. Zhu and S. J. Wang, An improved particle swarm optimization for feature selection, *Journal of Bionic Engineering*, vol.8, no.2, pp.191-200, 2011.

[57] P. Ghamisi and J. A. Benediktsson, Feature selection based on hybridization of genetic algorithm and particle swarm optimization, *Geoscience and Remote Sensing Letters*, vol.12, no.2, pp.309-313, 2015.

[58] X. Y. Wang, J. Yang, X. L. Teng, W. J. Xia and R. Jensen, Feature selection based on rough sets and particle swarm optimization, *Pattern Recognition Letters*, vol.28, no.4, pp.459-471, 2007.

[59] B. Chakraborty, Feature subset selection by particle swarm optimization with fuzzy fitness function, *Proc. of the 3rd International Conference on Intelligent System and Knowledge Engineering*, vol.1, pp.1038-1042, 2008.

[60] S. W. Lin, K. C. Ying, S. C. Chen and Z. J. Lee, Particle swarm optimization for parameter determination and feature selection of support vector machines, *Expert Systems with Applications*, vol.35, no.4, pp.1817-1824, 2008.

[61] C.-L. Huang and J.-F. Dun, A distributed PSO-SVM hybrid system with feature selection and parameter optimization, *Applied Soft Computing*, vol.8, no.4, pp.1381-1391, 2008.

[62] A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets*, Cambridge University Press, Cambridge, 2011.

[63] M. Gütlein, E. Frank, M. Hall and A. Karwath, Large-scale attribute selection using wrappers, *Proc. of IEEE Symposium on Computational Intelligence and Data Mining*, pp.332-339, 2009.

[64] W. W. Cohen, *Enron Email Dataset*, http://www.cs.cmu.edu/∼./enron/, 2015.

[65] H. Gómez-Adorno, J. P. Posadas-Durán, G. Sidorov and D. Pinto, Document embeddings learned on various types of n-grams for cross-topic authorship attribution, *Computing*, vol.100, no.7, pp.741-756, 2018.

[66] M. Yang, X. J. Chen, W. T. Tu, Z. Y. Lu, J. Zhu and Q. Qu, A topic drift model for authorship attribution, *Neurocomputing*, vol.273, pp.133-140, 2018.

[67] J. Soler-Company and L. Wanner, On the role of syntactic dependencies and discourse relations for author and gender identification, *Patter Recognition Letters*, vol.105, pp.87-95, 2018.

[68] J. B. Ma, B. Xue and M. J. Zhang, A profile-based authorship attribution approach to forensic identification in Chinese online messages, *Proc. of 2016 Pacific Asia Workshop on Intelligence and Security Informatics*, pp.33-52, 2016.

[69] H. Marwa, B. Slim, H. Chih-Cheng and B. S. Lamjed, A multi-objective hybrid filter-wrapper evolutionary approach for feature selection, *Memetic Computing*, pp.1-16, 2018.

[70] B. Tran, B. Xue and M. Zhang, Variable-length particle swarm optimisation for feature selection on high-dimensional classification, *IEEE Trans. Evolutionary Computation*, 2018.