A Novel Random Subspace Method for Online Writeprint Identification

Zhi Liu, Zongkai Yang, Sanya Liu
National Engineering Research Center for E-Learning, Central China Normal University,
Wuhan, P. R. China
Email: liuzhi8673@gmail.com, lsy5918@gmail.com

Wenting Meng
Department of Computer Science, Central China Normal University,
Wuhan, P. R. China

***Abstract*—With the widespread application of computer network technology, diverse anonymous cyber crimes begin to appear in the online community. The anonymous nature of online-information distribution makes writeprint identification a critical forensic problem. But the difficulty of the task is the huge number of features in even a moderate-sized available text corpus, which causes the problem of over-training. In this paper, we proposed a novel random subspace method by constructing a set of stable classifiers to take advantage of nearly all the discriminative information in the high dimensional feature space. In the construction of base classifiers, an optimized synergetic neural network is employed to provide probabilistic information for each class. Performance results on the subset of Reuters Corpus Volume 1 (RCV1) show that the proposed random subspace method achieves the better identification performance than a single classifier and conventional random subspace methods.**

***Index Terms*—online writeprint, random subspace method, synergetic neural network, principal component analysis, linear discriminant analysis**

I. INTRODUCTION

With the rapid advancement of new media technology, some network users wish to share some attractive information via e-mail, forum, online chat room, blog, microblog, etc. Indeed, the openness and anonymity of online community makes people freely express their opinions. But cyber criminals often take this opportunity to deliver some illegal information by some anonymous ways, such as sending offensive, threatening emails, rumors, radical speeches, and so on. These activities significantly harm the interests of individuals and the whole society. Therefore, efficient automated methods for authorship identification of online texts are becoming imperative in the forensic investigation. In recent years, some researchers have paid increasing attention to authorship identification of online texts, such as verifying the authorship of emails and messages on the cyber community [1,2], plagiarism detection [3] and personal blogs [4].

Similar to the definition of biological fingerprint, the unique writing-style hidden in texts is vividly described as writeprint [5]. Online writeprint identification is the task of predicting the most likely authorship of anonymous texts by using stylistic information in language. This study can be seen as a single-label multi-class text categorization problem [6] where the candidate authors represent different classes.

The key task of writeprint identification is to draw some fine-grained features from texts for quantifying the style of an author. Character n-grams have been proved to be very effective for capturing complicated stylistic information hidden in the texts. For example, the most frequent character 4-grams of an experimental text indicate lexical (|_the|, |_to_|, |that|), word-class (|_was|, |ing_|), and punctuation usage (|,_wh|, |,_s|). Keselj et al. [7] used the fixed-length n-grams in various test collections of English, Greek and Chinese texts, achieved very good identification results and improving previously reported results. Stamatatos [8] conducted the authorship attribution experiment on the subset of Reuters Corpus Volume 1 (RCV1) by extracting variable-length n-grams (n=2, 3, 4), and achieved higher identification accuracy than word-based features. Generally, training samples consist of a large number of n-grams provide the potential to improve the discrimination of authors. But this approach considerably increases the dimensionality of the problem, and size of training texts is too small in the high dimensional feature space. The problem of high feature-to-instance ratio greatly degrades the generalization ability of the classification system. How to reduce the feature-to-instance ratio and make full use of the important information in the feature space are the two key issues in the research of online writeprint identification.

In recent years, in the field of pattern recognition, the random subspace method (RSM) as a ensemble technique [9] has become a popular method to deal with the

Corresponding author: Sanya Liu, lsy5918@gmail.com.

This work was supported by the National Key Technology R&D Program in the 12th Five-Year Plan (Grant No. 2011BAK08B03, 2011BAK08B05), Program for New Century Excellent Talents in University (Grant No. NCET-11-0654), self-determined research funds of Central China Normal University from the colleges' basic research and operation of MOE (Grant No. CCNU09A02006).

doi:10.4304/jcp.7.12.2997-3004

problem of high feature-to-instance ratio such as face recognition [10], image retrieval [11], fMRI Classification [12], and so on. But RSM is not very much investigated in the writeprint identification, except that Stamatatos developed a RSM variant which is called the exhaustive disjoint subspace method (EDS) [13] to construct the ensemble classifier to identify the author. In this method, a large feature set is divided into equally-sized disjoint feature subsets drawn at random. Each particular attribute is used exactly once by using the random sampling without replacement. Each resulting feature subset is used to train a base classifier (BC) using a learning algorithm able to provide posterior possibilities. In this way, the diversity among different BCs is improved. However, the discriminative information will become more dispersed as the result of the increasing size of feature space. Thus, the stability of the BCs constructed in the different random subspaces will be weakened. In the traditional RSM, a number of low-dimensional subspaces are generated by randomly sampling from the original high-dimensional feature space. Then, multiple classifiers constructed in these random subspaces are combined to make a final decision. Although the method contributes to improve the diversity among predictions of BCs, it cannot guarantee the discriminability of each subspace. To deal with the difficulty, Wang and Tang do random sampling in reduced principal component analysis (PCA) subspace instead and random subspace is not completely random by fixing the first several dimensions of each subspace as those largest eigenvectors [14]. This improved method is not a completely random selection method; it can improve the stability (accuracy) of BCs in comparison to the completely random subspace method. However, there exists the several same features in each subspace, it degrades the diversity among the predictions of BCs, the final result is not necessarily better than the mean accuracy of BCs.

In this paper, we propose a novel identification method based on random subspace, the method divides feature space into the stable subspace (SS) and the unstable subspace (US) in reduced PCA space. Each random subspace is composed of two parts, one part is n_0 dimensions randomly selected from the SS and another part is n_1 dimensions randomly selected from the US. After sampling a certain granularity of subspaces, we apply linear discriminant analysis (LDA) to further compact the data in each subspace to improve their separability. Moreover, the constructed BCs are stable with satisfactory accuracy rates and the ensemble classifier covers nearly the full feature space. The experiments on the subset of RCV1 clearly demonstrate the efficiency and superiority of our method.

The organization of the paper is as follows. Section 2 describes the framework of online writeprint identification based on the proposed RSM. Section 3 gives the theoretical analysis. Experimental results and discussions are reported in Section 4. Section 5 concludes this paper.

II. A Framework of Online Writeprint Identification

Online messages, as the major channel of Web communication, are important sources for identifying the authorship of anonymous online messages. Due to the special characteristics of online short texts, traditional text classification methods are no longer suitable for writeprint identification. Some complicated methods need to be introduced. The main procedure of writeprint identification is described as follows:

- Extract n-grams features from training samples.
- Reduce the dimensionality of initial feature space.
- Perform PCA on the training samples to reduce redundancy.
- Apply the random subspace method to build the model of ensemble classifier.
- Divide the anonymous sample μ into different feature subspaces. Classify μ with the trained BCs.
- Combine all classification results using mean rule for final decision.

The framework of online writeprint identification is shown as Fig. 1.

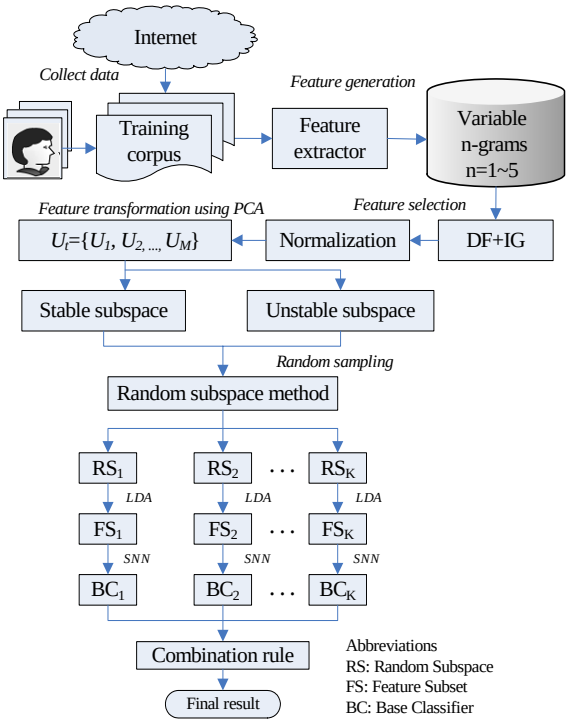

Figure 1. A framework of online writeprint identification.

A. Feature Selection

In our research, each text is represented as a vector of variable-length *n*-grams frequencies. The combination of all *n*-grams (n=1~5) considerably increases the dimensionality of the problem in comparison to word-based methods. For improving the generalization ability of identification system, the feature selection is of crucial importance. The two-step feature selection of document

frequency (DF) and information gain (IG) are adopted in this research. These two techniques are briefly explained here.

DF denotes the number of documents in which a feature occurs. This method can be used to reduce sparsity of datasets. Those features whose DF is less than the predetermined threshold are removed. IG measures the entropy required for category prediction by knowing the presence or absence of a feature in a document. Compared with DF, IG is a finer feature selection by removing less informative features.

B. Ensemble Method Based on Random Subspace

More discriminative features need to be retained for achieving a better performance in the recognition system, so dimension of features is still high after preprocessing. Inspired by the similar research on face recognition [15], a two-stage method of PCA+LDA is used to reduce feature dimension. Firstly, we apply the PCA [16] technique to compact the whole feature space using orthogonal transformation, the eigenvectors with zero eigenvalues are usually removed, and the dimension of the PCA subspace depends on the size of training set, which avoids classifier over-fitting problem. Secondly, after randomly sampling from the PCA subspace using proposed RSM, we apply LDA for maximizing the between-class scatter matrix and minimizing the whin-class scatter matrix in each random subspace to improve the discriminability of them. Thus, we can construct multiple stable BCs, and they are combined to form a more powerful classifier that makes the final decision.

Let us assume that the feature vector after preprocessing is $\boldsymbol{F}_n = \{f_1, f_2, ..., f_n\}$, and the feature set is transformed to $\boldsymbol{G}_d = \{g_1, g_2, ..., g_d\}$ by using PCA, where d is retained dimension contains nearly 99% of information of the initial feature space. Let $L(S_j)$ be a feature subset by applying LDA to compact *j* th random subspace. Then $C(L(S_j))$ as a single classifier is trained on $L(S_j)$. *K* represents the number of constructed BCs. The novel RSM is designed as follows.

At the training stage.

1. Normalize every component of the feature vector $\boldsymbol{F}_n$ to the range of [0,1], then the feature vector is transformed into $\boldsymbol{Nf}_n = \{nf_1, nf_2, ..., nf_n\}$.

2. Apply PCA to feature space $\boldsymbol{Nf}_n$ to compute its eigenvalues e_i and eigenvectors u_i $(i = 1, 2, ..., d)$. The eigenvectors with zero eigenvalues mostly containing noise are removed.

3. For each training sample, project every feature vector nf_i to the respective PCA subspace using $g_i = u_i^{\mathrm{T}} \cdot (nf_i - m_i)$, where m_i is the mean of all values in the *i*-th feature vector.

4. Generate *K* random subspaces $\{S_j\}_{j=1}^{K}$ in the space $\boldsymbol{G}_d$ by using randomly sampling with replacement, each random subspace S_j is spanned by $n_0 + n_1$ dimensions. The first n_0 dimensions are randomly selected from the N_0 largest eigenvectors (stable PCA subspace) and the other n_1 dimensions are randomly selected from the remaining N_1 eigenvectors (unstable PCA subspace).

5. Apply LDA to each random subspace S_j to compute its class-whin scatter matrix (S_w) and class-between scatter matrix (S_b).

6. Compute the eigenvalues λ_i and eigenvectors w_i of $S_w^{-1}S_b$, and project each subspace S_j to the LDA subspace $L(S_j)$.

7. Construct *K* classifiers $C(L(S_j))$ from the *K* respective LDA subspaces $L(S_j)$ $(j = 1, 2, ..., K)$.

At the recognition stage.

1. For each test sample, normalize every component of the combined feature vector to the range of [0,1] and project them to the respective PCA subspace.

2. Project the eigenvectors in PCA subspace to the each of *K* random subspaces and compact these random subspaces to their respective LDA subspaces

3. Feed the feature vectors in each LDA subspace to the *K* corresponding classifiers in parallel.

4. Combine the outputs of the *K* classifiers using a fusion rule to make the decision.

Compared with the traditional RSM that samples in the original feature vector directly, our method samples in the PCA subspace. The high dimension of the original feature set is greatly reduced by removing a large number of redundant information, and the relevancy between different features is nearly reduced to zero. The irrelevance between the different eigenvectors in PCA subspace indirectly contributes to the diversity among the predictions of the BCs, the diversity is one of the key-factors for the ensemble model based on RSM.

Secondly, we do not fixedly select first n_0 dimensions corresponding to n_0 largest eigenvectors for each random subspace; the selection method will make each subspace have the same n_0 component to reduce the diversity among the predictions of the BCs. Instead, each random subspace dimension is fixed as $n_0 + n_1$, in each random subspace, the first n_0 dimensions are randomly selected from SS and the other n_1 dimensions are randomly selected from the US. Thus, the BCs constructed in each random subspace has a satisfactory accuracy since the first n_0 dimensions encode much discriminative information, and the first n_0 dimensions in different random subspace are different, this method can improve the diversity among predictions of different BCs as well as classification accuracy of BCs.

Thirdly, we use the two-stage PCA+LDA method to deal with the high dimensional problem, the first-stage PCA is mainly used to avoid the problem of the high feature-to-instance ratio. The second-stage LDA is used to further compact the granularity of the data in each random subspace for maximizing S_b and minimizing S_w. The dimensionality of each random subspace is changed

to L-1 (L is number of candidate authors). Thus, the subspace dimension is much smaller than that of the original feature space, while the number of training samples remains the same. Therefore, our approach makes use of nearly all discriminative information in the feature space.

C. Settings of Base Classifier

In the construction of the BC, we use the synergetic neural network (SNN) [17] to train the BC constructed in each random subspace. The learning algorithm has the advantage of fasting convergence and self-leaning [18,19] , and it can provide rich posterior probabilities attribute to different authors. As a single classifier, SNN has shown the good performance in online writeprint identification [20], and the ensemble method with neural networks [21] has been proved to be effective in pattern recognition. In this study, we will introduce the regularization factor [22] into the algorithm. The optimized algorithm is described as follows.

At the training stage.

1. Read the feature vectors of different authors' training sets for i th subspace, and compute prototype pattern vectors v_k $(k=1,2,...,M)$ for each author by averaging the training sample vectors of each author. Then, these prototype pattern vectors representing different authors are normalized with zero-mean. The matrix composed of all prototype pattern vectors is $V_i = (v_1, v_2, ..., v_M)$.

2. Compute adjoint pattern vector $V_i^+ = (v_1^+, v_2^+, ..., v_M^+)$ for each author by using the regularization method as following formula:

$$V_i^+ = (V^{\mathrm{T}}V + rI)^{-1} \tag{1}$$

where r is the regularization factor, and I is the $M \times M$ identity matrix.

At the recognition stage.

1. Read the initial feature vector $q_k(0)$ $(k=1,2,...,M)$ representing testing sample and normalize it with zero-mean value.

2. Compute the initial values of order parameters ξ_k s in the middle layer of SNN as the following formula:

$$\xi_k(0) = v_k^+ \cdot q(0) \tag{2}$$

3. Conduct the competition among the order parameters ξ_k s by performing the following discrete evolution equation:

$$\begin{gathered} \xi_k(n+1) - \xi_k(n) = \frac{l}{D} \cdot (\lambda_k - D + B\xi_k^2(n)) \cdot \xi_k(n) \\ D = (B+C)\sum_k \xi_k^2(n),\ \lambda_k = B = C \end{gathered} \tag{3}$$

where l represents the step length of iteration, which determines the stability of SNN, and the attention parameter λ_k is obtained by computing the correlation coefficient between prototype pattern and the testing pattern as the following formula:

$$\lambda_k = \frac{\sum_{i=1}^{N}(v_{ki} - \overline{v}_{ki})(q_i - \overline{q}_i)}{\sqrt{\sum_{i=1}^{N}(v_{ki} - \overline{v}_{ki})^2 \sum_{i=1}^{N}(q_i - \overline{q}_i)^2}} \tag{4}$$

In practice, the inversion of the $l \times l$ matrix $V^{\mathrm{T}}V$, may pose some numerical difficulties, especially in the high-dimensional feature space. Besides being computationally complex, the matrix may be nearly singular. We add a small regularization factor r along the main diagonal to avoid the singularity of computing $V^{\mathrm{T}}V$ in sparse data. In the evolution process of ξ_k s, we select 50 as the maximum iterations to output probabilistic information corresponding to each category. The posterior probabilities are computed by the following formula:

$$P_j(C(L(S_i)), x, c_k) = \frac{\xi_k}{\sum_{k=1}^{M} \xi_{k'}} \tag{5}$$

D. Combination Rule

Provided the posterior probabilities of the constituent classifiers, the ensemble system computes probability of the anonymous sample attribute to each author. Then, a combined decision is achieved by averaging the estimated probabilities for all authors according to the following mean rule:

$$P(E(C(L(S_i)), mean), x, c) = \frac{1}{K}\sum_{i=1}^{K} P_i(C(L(S_i)), x, c_k) \tag{6}$$

where x represents a anonymous sample, c_k is the k th class, K is the number of the BCs. The mean rule has a low sensitivity to the classification error of different BCs according to [23]. To complete the classification model, the identification system automatically chooses the class that maximizes the posterior probability for x as follows:

$$label(ensemble, x) = \arg\max_{k \in L}(P(ensemble, x, c_k)) \tag{7}$$

E. Diversity Measures

The effectiveness of a classifier ensemble is indirectly indicated by the diversity among the predictions of the BCs as well as the accuracy of the individual BCs. So far, many measures [24] have been proposed to compute the diversity of an ensemble. We use the *entropy* measure as following:

$$entropy = \frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{L} -\frac{N_j^i}{K} log_L(\frac{N_j^i}{K}) \tag{8}$$

where k is the number of BCs, T is the total number of test samples and N_j^i is the number of BC that assign text i to class j, The higher the entropy of an ensemble, the more diverse the predictions of the different BCs.

III. Theoretical Analysis

It is assumed that the training sample $(\boldsymbol{x}, p)$ is independently drawn from the underlying probability distribution P_1 , where p represent the posterior probability indicator of sample $\boldsymbol{x}$. For example, $p = (0.62,\ 0.13,\ ...,\ 0.10)$ indicates that $\boldsymbol{x}$ attributes to 1-

th author as the probability of 1-th author is the maximum value in all indicators. The feature space is denoted by F'. $h(\boldsymbol{x},F)$ is an BC constructed in the feature subspace F, where F is randomly sampled from the whole feature space F' by proposed RSM. Then, the expectation result of the ensemble classifier is

$$h_E(\boldsymbol{x},F) = E_F h(\boldsymbol{x},F) \tag{9}$$

where $E_F h(\boldsymbol{x},F)$ represents the expectation output of $h(\boldsymbol{x},F)$ over F.

Let $(\boldsymbol{X}, p')$ be a testing sample drawn from the probability distribution P_2, and independent of the training set, where p' is the probabilistic results (soft labels) of the test sample $\boldsymbol{X}$ attributing to candidate authors. The mean error of the individual BC $h(\boldsymbol{x},F)$ is deduced as following:

$$\begin{aligned} Err(f) &= E_F E_{p',\boldsymbol{X}}(p' - h(\boldsymbol{X},F))^2 \\ &= E_F p'^2 + E_F E_{\boldsymbol{X}} h^2(\boldsymbol{X},F) - 2(E_{p'} p') \cdot (E_{\boldsymbol{X}} h_E(\boldsymbol{X},F)) \\ &= E_F p'^2 + E_{\boldsymbol{X}} E_F h^2(\boldsymbol{X},F) - 2(E_{p'} p') \cdot (E_{\boldsymbol{X}} h_E(\boldsymbol{X},F)) \end{aligned} \tag{10}$$

And the average classification error by the classifier ensemble is deduced as following:

$$\begin{aligned} Err(f_E) &= E_{p',\boldsymbol{X}}(p' - h_E(\boldsymbol{X},F))^2 \\ &= E_{p'} p'^2 + E_{\boldsymbol{X}} h_E^2(\boldsymbol{X},F) - 2(E_{p'} Y) \cdot (E_{\boldsymbol{X}} h_E(\boldsymbol{X},F)) \\ &= E_{p'} p'^2 + E_{\boldsymbol{X}}[E_F h(\boldsymbol{X},F)]^2 - 2(E_{p'} p') \cdot (E_{\boldsymbol{X}} h_E(\boldsymbol{X},F)) \end{aligned} \tag{11}$$

According to the definition of the variance, we have

$$D_F(h(\boldsymbol{X},F)) \geq 0 \tag{12}$$

Thus,

$$\begin{aligned} Err(h) - Err(h_E) &= E_{\boldsymbol{X}} E_F h^2(\boldsymbol{X},F) - E_{\boldsymbol{X}}[E_F h(\boldsymbol{X},F)]^2 \\ &= E_{\boldsymbol{X}}[E_F h^2(\boldsymbol{X},F) - (E_F h(\boldsymbol{X},F))^2] \\ &= E_{\boldsymbol{X}} D_F h(\boldsymbol{X},F) \\ &\geq 0 \end{aligned} \tag{13}$$

From the formula (13), it can be deduced that the mean-squared error of the ensemble classifier is smaller than the average error of the individual classifier. The reason is that the ensemble of multiple classifiers reduces the generation error. The effectiveness of the ensemble depends on the difference between $E_F h^2(\boldsymbol{X},F)$ and $(E_F h(\boldsymbol{X},F))^2$. The more diverse individual predictor $h(\boldsymbol{X},F)$ is the better performance the ensemble system can achieve. The individual predictor $h(\boldsymbol{X},F)$ have larger diversity by using proposed RSM, since the random subspaces select discriminative information from different portions of the reduced PCA subspace.

Let us assume the average mean-squared classification error of all the individual classifiers $h(\boldsymbol{X},F)$ trained in each random subspace is similar to a single classifier trained in the whole PCA space. This can be true when the granularity of the selected subspace is sufficiently large. The size of each selected subspace is nearly half of the full space in our research, the granularity is enough for covering all discriminative information, and the drop of accuracy for each individual BC may be well compensated in the aggregation process.

IV. Experiment Details

A. Dataset

Reuter Corpus Volume 1 (RCV1) is a large online dataset for the English language. The texts in RCV1 are short (approximately 2Kbytes - 8Kbytes). The top 50 author authors (with respect to total size of texts) in the dataset have been utilized for online writeprint identification [8,25]. In this study, the criterion for selecting the authors was the topic of the available text samples. Hence, after removing all duplicate texts found the R-measure [26], the top 50 authors of texts labeled with at least one subtopic of the class CCAT (corporate/industrial) were selected. The training set consists of 2,500 texts (50 per author) and the testing set includes other 2,500 texts (50 per author) non-overlapping with the training set. Some brief information about the dataset is summarized in Table I.

B. Results

In the preprocessing, we use the two-stage feature selection method to remove a large number of redundant features. In the first step, DF is used as a coarse feature reduction technique with the threshold of 1% on document vectors. In the second step, IG is used for fine feature selection; the number of features retained is selected to be 10,000. Then we apply PCA to the reduced feature space to get a compacted feature space. The specific variation of feature dimension using two-stage feature selection and PCA is provided as Table II.

TABLE I.
THE INFORMATION DESCRIPTION OF DATASET IN THIS STUDY AND REPORTED ACCURACY RESULTS SO FAR

Information	Training set	Testing set
Authors	50	50
Texts per author	50	50
Words per text	502.33	509.56
Characters per text	3,078.82	3,127.11

As can be seen from Table 2, many sparse features in initial feature space are removed by using DF, and more informative features are selected by using IG. Then the reduced features are compacted with PCA. In the experiment, the proposed random subspace method (PRSM) is conducted in the PCA subspace instead of initial feature space.

TABLE II.
THE RETAINED DIMENSIONS IN DIFFERENT FEATURE SELECTION STEPS

Feature selection	Initial features	DF	IG	PCA
Dimension	328,048	101,292	10,000	2,500

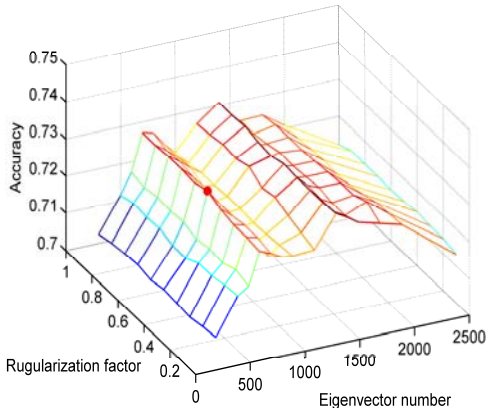

Figure 2. The accuracy of a single classifier using different number of largest eigenvectors in the feature subspace transformed by PCA+LDA subspace. The red point mark denotes the best accuracy using the largest 700 eigenvectors, in the point, the value of regularization factor is 0.5.

Fig. 2 shows the accuracy of a single classifier constructed in the PCA+LDA subspace (the subspace is transformed by two-stage PCA+LDA) using different number of largest eigenvectors and different value of regularization factor. We observe that the single classifier has the highest accuracy of 73.12% using the largest 1400 eigenvectors, and the single classifier achieves the second highest accuracy of 72.84% using the largest 700 eigenvectors. The regularization factor is used for finely adjusting parameter r in formula (1); an appropriate r can effectively avoid the singularity of computing the adjoint patterns in SNN. Let the number of the largest eigenvectors be the variable on the interval [0, 2500], the accuracy rate of the single classifier shows a monotonically increasing trend on the interval [0, 700], but the accuracy rate of the single classifier has an unstable variation process on the interval [700, 1400], and the accuracy rate of the single classifier decreases monotonically on the interval [1400, 2500]. The phenomenon described above can be used as the important basis for dividing the PCA subspace. Thus, we select 700 dimensions as the splitting point to divide the PCA space into the SS and US, the regularization factor r is set as 0.5 in each SNN in parallel; the eigenvectors in the interval [0, 700] constitute the stable PCA subspace and the eigenvectors in the interval [700, 2500] constitute the unstable PCA subspace. We select 500 dimensions from the SS and 700 dimensions from the US, we generate K=40 random subspaces for experiments. We compare the proposed random subspace method (PRSM) with the traditional completely random subspace method (CRSM) and semi-random subspace method (SRSM). In each group, the final results are obtained by averaging the results from 10 random experiments.

First, for each random subspace, we randomly select its 1200 dimensions from the preserved 2500 eigenvectors by using PCA. A BC is then trained on the selected eigenvectors. Fig. 3 demonstrates the result of combining 40 BCs using mean rule. We can see that with the CRSM, the accuracy of each BCs is low, ranging from 60% to 70%. However, these weak classifiers are greatly enforced with mean rule, and 72.48% accuracy is achieved. The result shows that the BCs constructed in different random subspace are complementary of one another.

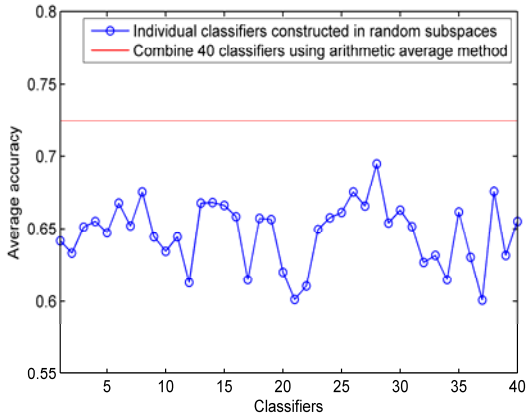

Figure 3. The average accuracy of combining 40 individual classifiers constructed in the random subspaces using mean rule. All the 1200 dimensions of each random subspace are randomly selected.

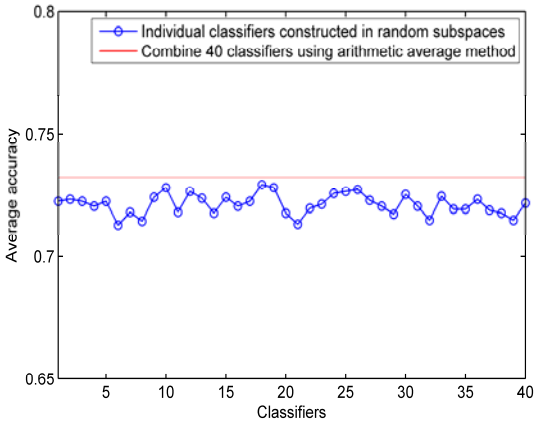

Figure 4. The average accuracy of combining 40 individual classifiers constructed in the random subspaces using mean rule. The first 500dimensions of each random subspace are fixed as the 500 largest eigenvectors, and the other 700 dimensions are randomly selected.

Second, SRSM is used for improving the classification ability of each weak BC. We fix the first 500 dimensions of each random subspace as the largest 500 eigenvectors, and randomly select the other 700 dimensions from remaining eigenvectors. As described in Fig. 4, the BCs constructed in the random subspaces are improved significantly since the much discriminative information is contained in the first 500 largest eigenvectors, but the same part in each subspace leads to a low diversity among the predictions of different BCs, so the multiple classifiers system achieves the accuracy rate of only 73.24%. The result shows that the first 500 largest eigenvectors contribute to the stability of each BC, but other 700 dimensions randomly selected from the remaining 2000 eigenvectors are not enough to improve the diversity among the predictions of different BCs, and the classification results of the BCs are similar since the same part is contained in the each subspace.

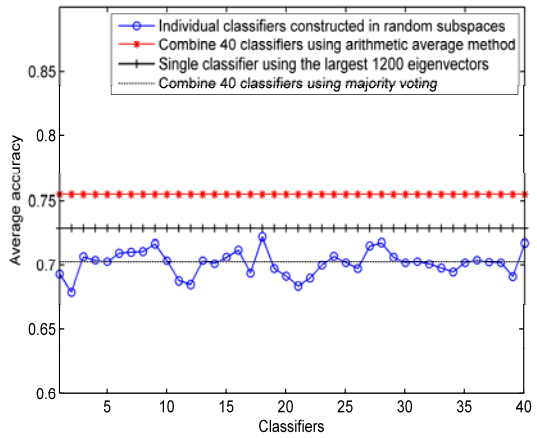

Figure 5. The average accuracy of combining 40 individual classifiers constructed in the random subspaces using mean rule. The first 500 dimensions of each random subspace are randomly selected from SS, and the other 700 dimensions are randomly selected from US.

Finally, PRSM is used to improve the overall performance of the combined classifier. In the method, we first divide the reduced PCA subspace into two parts: the SS consisting of the largest 700 eigenvectors and the US consisting of the remaining 1800 eigenvectors. In this experiment, we randomly select n0=500 dimensions from SS and randomly select n1=700 dimensions from US. So the each subspace consists of the $n_0 + n_1$ dimensions. As can be seen in Fig. 5, we observe that the ensemble system of combining 40 classifiers using mean rule achieves a better accuracy rate of 75.48% than conventional RSMs. Moreover, the result is better than the best reported results for the same text corpora (see Table 1). In the aspect of combination method, the performance using mean rule is much better than that of using majority voting, it demonstrates the integration of probabilistic information outputted from multiple SNNs in parallel can provide a more accurate decision in comparison to the majority voting. Compared with CRSM, the stability of different BCs is improved since the accuracy of every BC ranges from 67% to 73%. The mean accuracy of different BCs is lower without fixing the largest n0 eigenvectors, but the PCA subspace is divided into the SS and US, the different subsets from the two subspaces effectively improves the diversity among the predictions of different BCs. Although the unstable PCA subspace contains less discriminative information, the remaining 700 dimensions provide some disturbance factors to increase the diversity among the BCs. On the other hand, it also demonstrates that the combination of multiple stable individual classifiers outperforms an optimal single classifier as shown in Fig. 1.

Table III shows the comparison results of PRSM and traditional RSMs: CRSM and SRSM. The constructed BCs by SRSM achieve the better mean accuracy than other methods. But in terms of the variance of the BCs, the SRSM is lowest due to the less difference between $E_F h^2(\boldsymbol{X}, F)$ and $(E_F h(\boldsymbol{X}, F))^2$ as illustrated in section 4. In spite of achieving a highest variance among different BCs, the mean accuracy of BCs is very low. So PRSM achieves a trade-off between the mean accuracy of BCs and diversity among the predictions of different BCs.

TABLE III.
THE COMPARATIVE RESULTS OF PRSM AND TRADITIONAL RSMS (CRSM, SRSM) IN IDENTIFYING 50 AUTHORS

Method	Ensemble accuracy	Mean accuracy of BCs	Variance of BCs ($\times 10^{-4}$)	Diversity
CRSM	0.7248	0.64	3.12	0.35
SRSM	0.7324	0.72	0.12	0.12
PRSM	0.7548	0.70	1.24	0.22

The average accuracy rates and the diversity of the multiple classifiers system are reported in Table IV. In general, using more classifiers leads to a better accuracy rate. But when the number K of BCs is enough to a certain extent, the overall accuracy will tend to be stable. Even if we increase the classifiers' number, the accuracy will not be improved much. But too many BCs in parallel greatly increase the system burden, and the diversity among different BCs will decrease. From Table 4, we can see that the accuracy of the ensemble is gradually improved as the number of BCs increases, but the change of diversity is different from the accuracy, the diversity begins to decrease when the number of BCs reaches 70. The case indicates that the generation of too many random subspaces may lead to a poor complementary ability of BCs. Instead, the performance of these weak BCs is a factor which needs to be considered, since the generation of some unstable BCs may weaken the whole performance of an ensemble. In this study, 40 BCs are selected as a trade-off between the overall performance and the computational cost.

TABLE IV.
THE AVERAGE ACCURACY RATES COMBINING DIFFERENT NUMBER OF BCS

Classifiers number (K)	10	20	30	40	50
Accuracy (%)	73.24	74.48	75.16	75.48	75.36
Diversity	0.18	0.21	0.23	0.22	0.21
Classifiers number (K)	60	70	80	90	100
Accuracy (%)	75.52	75.84	76.12	76.08	76.04
Diversity	0.22	0.24	0.23	0.22	0.22

V. CONCLUSION

In this paper, we have developed a novel random subspace approach for the fusion of multiple diverse classifiers for online writeprint identification. The ensemble mechanism at the feature level captures the richest information. But it is more difficult because: (i) the high feature-to-instance ratio problem and (ii) high redundancy of feature space. Our approach overcomes both problems. By respectively using random sampling in the two separate PCA spaces (stable subspace and unstable subspace), and feature vectors in each subspace have a better discriminability by applying LDA. A set of stable classifiers are constructed covering nearly all the important information in the high dimensional feature space. The combination of multiple classifiers produces a

better performance than a single classifier and the traditional RSMs. Experiments results on the benchmark dataset show that our algorithm can effectively identify the authorship of texts with an acceptable accuracy rate.

ACKNOWLEDGMENT

This work was supported by the National Key Technology R&D Program in the 12th Five-Year Plan (Grant No. 2011BAK08B03, 2011BAK08B05), Program for New Century Excellent Talents in University (Grant No. NCET-11-0654) and self-determined research funds of CCNU from the colleges' basic research and operation of MOE (No. CCNU09A02006). We would like to thank Prof. Stamatatos at University of the Aegean who provided the valuable dataset for this research, and anonymous reviewers for their comments which helped improve this paper.

REFERENCES

[1] O. de Vel, A. Anderson, M. Corney, G. Mohay, "Mining E-mail content for author identification forensics," ACM SIGMOD Record, 30(4), pp.55-64, 2001.
[2] A. Abbasi, H. Chen, "Applying authorship analysis to extremist-group web forum messages," IEEE Intelligence Systems, 20(5), pp.67-75, 2005.
[3] H.V. Halteren, N.H.J. Oostdijk, "Linguistic profiling of texts for the purpose of language verification," In Proceedings of the 20th International Conference on Computational Linguistics, pp. 966-972, 2004.
[4] A. Abbasi, and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," ACM Trans. Inf. Syst, 26(2), pp. 1–29, 2008.
[5] J. Li, R. Zheng, H. Chen, "From fingerprint to writeprint, Communication of the ACM," 49(4), pp.76-82, 2006.
[6] F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys, 34(1), pp.1-47, 2002.
[7] V. Keselj, F. Peng, N. Cercone, C. Thomas, "N-gram based author profiles for authorship attribution," In Proceedings of the Conference Pacific Association for Computational Linguistics, pp.255-264, 2003.
[8] J. Houvardas, E. Stamatatos, "N-gram feature selection for authorship identification," In Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications, pp.77-86, 2006.
[9] T. Kam, Ho, "The random subspace method for constructing decision forests," IEEE Trans. on PAMI, 20(8), pp.832-844, 1998.
[10] X. Wang, X. Tang, "Random sampling LDA for face recognition," In Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition, pp.259-265, 2004.
[11] D. Tao, X. Tang, "Random sampling based SVM for relevance feedback image retrieval," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.647-652, 2004.
[12] L.I. Kuncheva, J.J. Rodriguez, C.O. Plumpton, D.E.J. Linden, S.J. Johnston, "Random subspace ensembles for fMRI classification," IEEE Transactions on Medical Imaging, 29(2), pp.531-542, 2010.
[13] E. Stamatatos, "Ensemble-based author identification using character N-Grams," In Proceedings of TIR'06, pp.41-46, 2006.
[14] X. Wang, X. Tang, "Random sampling for subspace face recognition," International Journal of Computer Vision, 70(1), pp.91-104, 2006.
[15] P.N. Belhumeur, J. Hespanda, and D. Kiregeman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," IEEE Trans. on PAMI, 19(7), pp.711-720, 1997.
[16] J. Peng, Y. Ma, "Integer wavelet image denoising method based on principle component analysis," Journal of Software, 7(5), pp.982-989, 2012.
[17] H. Haken, *Synergetics*, Springer-Verlag, Berlin, 1977.
[18] T. Zhao, H.L. Tang, H. Ip, F. Qi, "On relevance feedback and similarity measure for image retrieval with synergetic neural nets," Neourocomputing, Vol. 51, pp.105-124, 2003.
[19] D. Hu, F. Qi and J. Liu, "Recognition of objects with skew distortion based on synergetics," Pattern Recognition Letters, 20(3), pp.255-265, 1999.
[20] S. Liu, Z, Liu, J. Sun and L. Liu, "Application of synergetic neural network in online writeprint identification," International Journal of Digital Content Technology and its Applications, 5(3), pp.126-135, 2011.
[21] Y. Gu, Y. Shi, J. Wang, "Efficient intrusion detection based on multiple neural network classifiers with improved genetic algorithm," Journal of Software, 7(7), pp.1641-1648, 2012.
[22] V.A. Postnov, "Use of Tikhonov's regularization method for solving identification problem for elastic systems," Mechanics of Solids, 45(1), pp.51-56, 2010.
[23] D.M.J. Tax, M. Breukelen, R. P. W. Duin, J. Kittler, "Combining multiple classifiers by averaging or by multiplying?" Pattern Recognition, 33(9), pp.1475-1485, 2000.
[24] L. Kuncheva, C. Whitaker, "Measures of diversity in classifier ensembles," Machine Learning, 51(2), pp.181-207, 2003.
[25] E. Stamatatos, "Author identification using imbalanced and limited training texts," In Proceedings of the 4th International Workshop on Text-based Information Retrieval, pp. 237-241, 2007.
[26] D.V. Khmelev, W.J. Teahan, "A repetition based measure for verification of text collections and for text categorization," In Proceedings of the 26th ACM SIGIR conference on Research and development in informaion retrieval, pp. 104-110, 2003.

Zhi Liu is a Ph.D. Candidate at National Engineering Research Center for E-learning (NERCEL), Central China Normal University (CCNU), Wuhan, China. His research interests are in machine learning, data mining, intelligent system and knowledge service.

Zongkai Yang received the B.E. and M.E. degrees from Huazhong University of Science and Technology (HUST), Wuhan, China, in 1985 and 1988,respectively, and the Ph.D. degree from Xi'an Jiaotong University, Xi'an, China, in 1991. From 1991 to 1993. He is currently a professor in NERCEL, CCNU. His research interests include signal processing, network communication, and information technology.

Sanya Liu received the B.E. and M.E. degrees in 1996 and 1999, and received the Ph.D. degree in 2003 from HUST. He devoted himself to his postdoctoral research in Xiamen University from 2003 to 2005. Currently, he is a professor in NERCEL, CCNU. His research interests include artificial intelligence, and computer application.