PSYCHOLOGY AND EDUCATION (2021) 58(1): 4262-4266
An Interdisciplinary Journal

AUTHORSHIP VERIFICATION USING MODIFIED PARTICLE SWARM OPTIMIZATION ALGORITHM

N.Selvaganesh[1], Sharmila D[2] , A V Prabu[3]

[1]Assistant Professor, Department of Computer Science andEngineering, PSNACollege of Engineering, Dindigul, Tamil Nadu.
[2]Assistant Professor, Department of Information Technology, M.Kumarasamy College of Engineering, Karur, Tamil Nadu.
[3]Dept of ECE, Associate Professor,Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India.

ABSTRACT

Digital forensics is the study of recovery and investigation of the materials found in digital devices, mainly in computers. Forensic authorship analysis is a branch of digital forensics. It includes tasks such as authorship attribution, authorship verification, and author profiling. In Authorship verification, with a given a set of sample documents D written by an author A and an unknown document d, the task is to find whether document d is written by A or not. Authorship verification has been previously done using genetic algorithms, SVM classifiers, etc. The existing system creates an ensemble model by combining the features based on the similarity scores, and the parameter optimization was done using a grid search. The accuracy of verification using the grid search method is 62.14%. The time complexity is high as the system tries all possible combinations of the features during the ensemble model's construction. In the proposed work, Modified Particle Swarm Optimization (MPSO) is used to construct the classification model in the training phase, instead of the ensemble model. In addition to the combination of linguistic and character features, Average Sentence Length is used to improve the verification task accuracy. The accuracy of verification has been improved to 63.38%.

Keywords
Authorship Verification, One-class problem, Profile-based method, MPSO algorithm, Ensemble model.
Article Received: 20 September 2020, Revised: 30 November 2020, Accepted: 18 December 2020

1. Introduction

Digital forensics locates the evidence located on computers, mobile phones, and networks [1]. Digital forensics branches are computer forensics, network forensics, forensic data analysis, and mobile device forensics [2]. Tasks in digital forensics include Authorship Verification (AV), Author clustering, and Text classification.

Text classification is the task of classifying a document into one or more classes. It can be done either manually or algorithmically. The manual category of the documents is widely used in the library science. Algorithmic classification of documents is used in information science and computer science. It is done either according to the subject or the content in the text document.
The process of determining whether a given unknown document x is written by the same author who had written the given set of known documents D is known as Authorship Verification (AV). It can be viewed either as a mono-class classification or as a multi-class classification task. It can be used for tasks like intrinsic plagiarism detection.

The classical approach of authorship verification involves four steps [3]. Documents, articles or online messages composed by potential creators are gathered from the web. Subsequent to gathering the reports, the unstructured writings are spoken to as a vector of composing style highlights. The preparation information is utilized to prepare the arrangement model. Created model is utilized to anticipate the origin of the obscure reports.

2. RELATED WORKS

The techniques for verifying the authorship of an unknown document are based on information sources used, classification approach and learning methods used.

There are two approaches to perform the authorship verification task based on information sources. These are extrinsic method and intrinsic method. The extrinsic method transforms a one-class classification task to a binary classification task [4]. It requires information from the external resources in addition to the information about the set of known documents written by the author and the unknown document for which the authorship verification is to be done. The authorship verification tasks based on extrinsic method have performed well. One of the best examples is the Imposter method (IM) proposed by Koppel and winter (2014) [5]. So as to choose whether X and Y were composed by same creator, a lot of "sham" records were methodicallly delivered. In the event that X is adequately more like Y than to any of the created impostors, at that point both X and Y were composed by a similar creator. The estimation of report comparability relies upon choosing irregular subsets of highlights utilized while looking at the records.

In the inherent technique, just the data about the arrangement of realized archives composed by the creator and the obscure report for which the initiation check is to be done are required [6]. Lately, the creation check

undertakings dependent on inborn strategy has likewise performed well. The inborn technique proposed by Ferry et al. (2014) [7]used CART calculation (Classification and RegressionTrees) to choose if obscure report is composed by An or not. Another model is the Profile-Based Method for Authorship Verification [8] proposed by Potha and Stamatatos, in view of the natural strategy, as it doesn't need any outer data to choose whether the obscure archive is composed by a similar creator.

Authorship Verification can be viewed as a one-class or multi-class classification problem. In one-class classification, the training dataset consists of elements of a specific class; the task is to identify the elements which belong to the classification training dataset, from the collection of elements which belongs to different classes. Authorship verification is the application of one-class classification methods to stylo metric datasets [4]. Authorship Verification is a difficult problem, if modelled as a one-class classification task. The challenge is, in determining the boundaries between the class elements and the outliers without negative examples or at least exhaustive and representative positive examples [9]. Though it is difficult, one-class classification is more efficient to decide whether the unknown document is written by the same author or not. Koppel and winter (2014) [5] transformed the one-class classification problem of AV into multi-class classification problem and then calculated the similarity score.

The learning methods used in authorship verification are instance based learning and profile based learning. In instance based learning the known set of documents in the training corpus are not concatenated after pre-processing. They are kept separately and similarity is checked between each of the known documents and the unknown document [9]. It is called instance-based as it constructs the hypothesis from each document. Complexity increases with the increase indata. It is a kind of lazy learning. The time for training the model is high when using the instance based learning approach.

In profile based learning approach, after pre-processing, all the documents of known authorship in the training corpora are concatenated to form a single large corpus [9]. Then the most frequent n-grams in the corpus are identified, it is compared with the most frequent n-grams in the unknown document using the dissimilarity function. Based on the threshold obtained, the authorship of unknown document is verified using binary values [9]. The existing system is a profile based system, as all the known documents in the training document are concatenated after the pre-processing is completed. This reduces the time complexity of the training phase.

Features are used to determine the individual writing styles of each author to distinguish them from others. The features used for authorship verification are lexical, Syntactic, Semantic, Character and Application specific features. The number and types of features used for authorship identification were varied in order to determine the influence of each type of feature [10].

3.PRPOSED SYSTEM

In the proposed approach , MPSO was used for the construction of the ensemble models used for verification task.

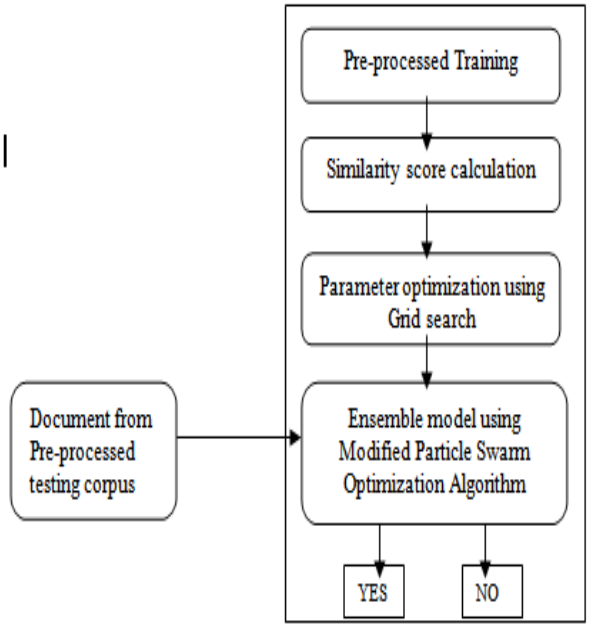

Figure 1: System Architecture of the proposed system

The system architecture of the proposed system is shown in Figure 1. The dataset collected is pre-processed and a subset of corpus is used in the training phase. All the ten features used in the proposed system are derived and a similarity score between the known and the unknown documents is calculated. Then the parameter optimization is done using the Grid search method i.e. best parameter value of each feature is identified and a MPSO model is constructed using those features. In the testing phase the testing corpus is given as input and the authorship of the unknown documents is verified using the proposed MPSO model.

The proposed methodology divided into four stages as follows:

1. Pre-processing.
2. Pair-wise similarity score calculation.
3. Parameter optimization using Grid search
4. Construction of ensemble model using MPSO algorithm.

3.1 Pre-processing

In pre-processing, first all the known documents are concatenated and only two documents are available for each

book, novel and essay, one is the concatenated known document and the other is a document for which authorship is to be identified.

3.2. Pair-wise Similarity score calculation

The similarity between the known and the unknown documents is being calculated. The ten types are used in this system are as follows:
punctuation n-grams, Average Sentence Length, character n-grams, token k-prefixes, token k-suffix n-
grams, n-prefixes k-suffixes and n % frequent tokens.

The features of Average Sentence Length(ASL) and Construction of feature vector also used to improve the accuracy of verification.
Once the feature vector is built, the Manhattan distance is calculated using the formula:

$$\text{Dist}(X,Y) = \sum_{i=1}^{n} |xi - yi| \qquad (3.3)$$

Where X is the known document D and Y is the unknown document A.

The output of Manhattan distance is transformed into the likeness score using the following equation:

$$\text{sim}(X, Y) = 1 / (1+\text{dist}(X,Y)) \quad (3.4)$$

3.3 Parameter optimization using Grid search

In order to increase a accuracy, parameter optimization is carried out to choose the best parameters. The acceptance threshold Θ is calculated separately for each feature category. It is used to classify the problem P with Yes or No based on the classification function:

$$\text{Classify} = \begin{cases} \text{Yif Sp} > 0 \\ \text{Notherwise} \end{cases} \qquad (3.5)$$

Where Sp denotes the similarity score of an unknown document. The value of Θ is chosen so that Equal Error Rate (EER) is gotten for the issues in the corpus C (the bogus positive rate is equivalent to bogus negative rate) during the grouping of preparing issues. The limit Θ isn't really situated at the convergence purpose of the two likelihood capacities, yet near them. The EER is picked as the standards to decide esteems for Θ, since the exhibition measure in existing framework gives equivalent loads to bogus positive and bogus negative [10]. For the reasonable corpora in existing framework, the edge is resolved as the middle of all comparability scores.

The above advances are rehashed for each element in Fi and the entirety of its potential blends for the boundaries n and k. The exactnesses of all conceivable boundary mixes of each highlights is gotten and afterward for each element, the boundary esteem that prompts most extreme precision is acquired and put away as the model M.

3.4 Construction of the Ensemble model using Modified Particle Swarm Optimization Algorithm

From the model M made utilizing boundary advancement, group models are made utilizing MPSO calculation. In MPSO the intellectual part of the overall PSO is splitted into two distinct segments.

The principal part is acceptable experience segment. The molecule has memory about its recently visited best position, which is like the overall PSO. The subsequent segment is the awful experience part. It causes the molecule to have memory about its recently visited most noticeably terrible position. To figure the speed of the molecule, the terrible experience part is likewise thought about [11].

$$V_{i+1} = \omega*V_i + c_1*R1(p_{best} - S_i) + c_2*R_2(S_i - p_{worsti}) + c_3*R3(g_{besti} - S_i) \quad (3.6)$$

The position update equation is same as general PSO algorithm:

$$Si+1 = S_i + V_{i+1} \quad (3.7)$$

Where ω ->The inertia weight
C1,C2,C3 → The acceleration coefficients
p_{worst} -> the worst position of the particle.
R1, R2, R3 denotes uniformly distributed random numbers varies in the range (0, 1).

Algorithm of Modified PSO method [11] are:

Step 1 :
Select the number of particles, generations, tuning acceleration coefficients c_1, c_2, and c_3 random variables R1, R2, R3 to start optimal solution searching
Step 2 :
Initialize the particle position and velocity
Step 3 :
Select the particle's individual best value for each generation
Step 4 :
Choose the particle's global best value
Step 5 :
Choose the particle's separate worst value
Step 6 :
Apprise the particle separate best p_{best}, global best g_{best}, particle worst P_{worst} in the velocity equation and find updated velocity.
Step 7 :
Apprise the new velocity in Eq (3.7) and get the location of the particle.
Step 8 :
Repeat all steps till the required accuracy is attained.

In the proposed technique, the quantity of particles is set as 20. The quantity of ages is changes between 10 to 100. Every molecule is a gathering of haphazardly chosen

highlights signified as 1 if the component is available and 0 on the off chance that it is absent in the molecule.

4. RESULTS AND DISCUSSION

In Authorship Verification task, one of the difficulties is to locate a normalized, reasonable and publically accessible corpus. Skillet sorts out many shared assignments on creation attribution, initiation check and creator bunching. Skillet likewise gives publically open corpora to these undertakings. The dataset utilized in the current framework is the English corpus utilized for PAN 2013, 2014 [6]:

- PAN 2014: Training corpora with one hundred and ninety seven English essays and seventy five novels are used.
- The testing corpora with two hundred and thirty English essays and two hundred and thirty English text books are used.

The performance of the proposed Authorship verification system is evaluated using the accuracy measure:

$$\text{Accuracy} = \frac{\text{numberofcorrectlyclassifieddocuments}}{\text{Totalnumberofdocumentsinthecorpus}} \quad (4.1)$$

The classification model generated using the MPSO algorithm has the highest accuracy compared to the ensemble model generated by the grid search, the results are shown in table 1

Table 1 Accuracies of Classification Models

Classification Model	Features	Accuracy
Ensemble model generated by Grid Search	1, 3, 4, 5, 9	62.14%
Ensemble model generated by MPSO algorithm	1,6,10	63.38%

From the table, it is evident that the first feature (punctuation n-gram) performs well in both the existing and the proposed methods. The time complexity of construction of ensemble model using MPSO algorithm and evaluating its performance is 23 minutes, whereas the time complexity in existing system is 54minutes. The proposed method has a low time complexity compared to that of the existing system. The accuracy has also increased from 62.14 % in the existing system to 63.38% in the proposed method.

5. CONCLUSION AND FUTURE WORK

Authorship verification is a task of deciding whether two documents originate from same author. It is used in plagiarism detection. The existing system performs Authorship verification task by optimizing the parameters and constructing ensemble models using grid search method. The drawback of grid search method is it's high time complexity.

In the proposed strategy, authorship verification task is finished utilizing the classification model developed utilizing Modified PSO. Every molecule is alloted with arbitrary position and speed, which are refreshed in every age. In MPSO calculation every molecule has the memory of its best and furthermore most exceedingly terrible position. The primary preferred position of the proposed strategy is its time intricacy. The time taken for advancement of grouping model and check task by the proposed system is low, appeared differently in relation to that of the current structure. Ordinary Sentence Length, a lexical part is added to improve the exactness of the order. The proposed order model was had a go at using a sub-set of PAN 13, PAN 14 English language datasets. The accuracy of grouping in proposed model was improved in the seventh time itself.

REFERENCES

[1] Simson L. Garfinke (2013), 'Modern Crime Often Leaves an Electronic Trail. Finding and Preserving that Evidence Requires Careful Methods as well as Technical Skill', In Proceedings of Sigma'13, The Scientific Research Society, vol.101, pp. 370-377.

[2] Damshenas M, Dehghantanha A, Mahmoud R (2014), 'A Survey on Digital Forensics Trends', International Journal of Cyber-Security and Digital Forensics (IJCSDF), vol.3, pp. 209-234.

[3] Nirkhi S, R.V.Dharaskar (2013), 'Comparative Study of Authorship Identification Techniques for Cyber Forensics Analysis', International Journal of Advanced Computer Science and Applications, vol. 4, No.5, pp. 32-35.

[4] Koppel M, Schler J(2004). 'Authorship Verification as a One-class Classification Problem', In Proceedings of the twenty-first International conference on Machine learning, ICML '04, pp. 489-95.

[5] Koppel M, Winter Y(2014), 'Determining if Two Documents are by the Same Author', JASIST 2014, vol.65, pp. 178-87.

[6] Stamatatos E, Daelemans W, Verhoeven B, Potthast M, Stein B, Juola P (2014), 'Overview of the Author Identification task at PAN 2014', In Proceedings of CLEF 2014 Conference, vol. 1180, pp. 877-97.

[7] Frery J, Largeron C, Juganaru-Mathieu M (2014), 'UJM at CLEF in Author Identification Notebook for PAN at CLEF 2014', In proceedings of CLEF

14, vol.1180, pp. 1042-48.

[8] Potha N, Stamatatos E(2014), 'A Profile-based Method for Authorship Verification', In Proceedings of 8th Hellenic Conference on Artificial Intellegence, SETN 2014, Springer International Publishing, pp. 313-326.

[9] Potha N, Stamatatos E(2014), 'A Profile-based Method for Authorship Verification', In Proceedings of 8th Hellenic Conference on Artificial Intellegence, SETN 2014, Springer International Publishing, pp. 313-326.

[10] Halvani O, Winter C, Pflug A(2016), 'Authorship Verification for Different Languages, Genres and Topics', In Proceedings of DFRWS Europe, vol.16, pp. 33-43.

[11] S N Deepa, G.Sugumaran (2011), 'Model Order Formulation of a Multivariable Discrete System Using a Modified Particle Swarm Optimization Approach', In Proceedings of Swarm and Evolutionary computation, vol.1, pp. 204-212.

[12] R. Sathish, R. Manikandan, S. Silvia Priscila, B. V. Sara and R. Mahaveerakannan, "A Report on the Impact of Information Technology and Social Media on Covid–19," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 224-230, doi: 10.1109/ICISS49785.2020.9316046.