Report for Project 4: Genetic & Evolutionary Feature Selection

Linyuan Zhang

***Abstract*—In this project, a Genetic & Evolutionary Feature Selection (GEFeS) is developed for a simple Author Identification System which only takes the unigram as its potential features. The feature selection method uses a steady-state genetic algorithm in this project. The text data, the extracted features of the text, and the baseline models are provided by Dr. Dozier. The relationship between the number of features, the accuracy of models, and the most consistent features are explored in this project.**

***Index Terms*—Feature selection, SSGA, GEFeS, rbf-svm, lsvm.**

I. Introduction

AUTHORSHIP identification is the process of identifying the author of a given text from a set of suspects. It can be beneficial for various tasks and areas including bibliometrics, information retrieval, and plagiarism detection [1]. One major subtask of the authorship identification problems is the extraction of the most appropriate feature sets for representing the style of an author [2]. There are several features that have been proposed, including lexical, which could be the character or word frequencies, structural, which includes the paragraph lengths or words per sentence, and domain-specific, like the keywords in a document, etc. The feature used in this project is the unigram feature. The text sample is divided into characters. According to the ASCII code table, each printable character is treated as a feature. Therefore, there are 95 features. Efficient feature selection in authorship attributes is key for successful identifications. Feature selection is the process where the system automatically or manually selects those features which contribute most to the prediction variable. The feature selection method uses the steady-state genetic algorithm (SSGA).

The training and evaluation dataset used in this project is CASIS-25 dataset, which consists of online blog entries. It contains the first 25 authors of CASIS-1000 dataset. There are four samples for each author and the size of each sample is small. Therefore, there are 100 text samples in this dataset.

In the following report, the performance of baseline models and the model with feature selection are compared. The selected features are discussed as well.

II. Methodology

In this section, the mechanisms of the new unigram feature extractor, the SSGA, and the procedure of GEFeS are explained.

A. Unigram Feature Extractor

Firstly, the 100 text samples are read in. Each text sample's file name represents its author information. The number of occurrences of the 95 ASCII code printable characters are counted for each text sample. Then the counting results are saved in a list with the length of 95. The index of the list represents a printable character in the order of the ASCII code and the value stored in each index represents the number of occurrences in the text samples. After the above-mentioned steps, 100 lists, which represent the unigram features of the 100 text sample, are collected. Each list is written to a file with its corresponding author information.

B. Baseline Models

The three baseline models provided by Dr. Dozier are rbf-svm, linear-svm (lsvm), and MLP. In this project, only the first two support vector machine models, which are used as the baseline since the MLP model requires too much running time in the laptop to run for enough time to collect data.

C. Steady-State Genetic Algorithm (SSGA)

An SSGA uses a (μ+1) replacement strategy where one offspring replaces the parent with the worst performance [3]. In this algorithm, when an offspring is generated, no matter whether the performance of this offspring is good or bad, it replaces the parent in the current generational population which has the lowest fitness value. Tournament selection and uniform crossover are used in this SSGA.

D. Procedure of GEFeS

Firstly, an individual with the length of 95 is randomly generated as the feature mask set, which contains 0 and 1 as the first generation in the SSGA. Each feature list multiples the feature mask set to get the new dataset. The following is an example of how to apply a mask to the feature list.

The example feature list: [[213, 6, 9, 35, 90], [156, 14, 65, 8, 11]].

The mask set: [0, 1, 0, 0, 1].

Applying the mask to feature list: [[213 * 0, 6 * 1, 9 * 0, 35 * 0, 90 * 1], [156 * 0, 14 * 1, 65 * 0, 8 * 0, 1 * 11]].

The new dataset: [[0, 6, 0, 0, 90], [0, 14, 0, 8, 11]].

The new dataset is used to train and evaluate the rbf-svm and lsvm models one by one. The accuracy value of a model is returned to the SSGA to evolve the population of the feature mask.

L. Zhang is with the Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849, USA. L. Zhang's e-mail: lzz0035@auburn.edu.

TABLE I
ACCURACY VALUES OF THREE BASELINE MODELS.

	rbf-svm	lsvm	MLP
Original	0.70	0.62	0.65
Without TFIDF	0.70	0.62	0.64
Without Standardization	0.46	0.48	0.61
Without Normalization	0.51	0.52	0.59

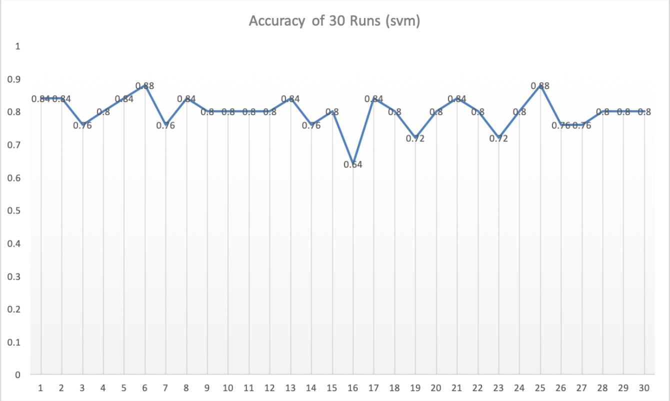

Fig. 1. Accuracy values of 30 runs for rbf-svm.

III. EXPERIMENTS

In this section, more details about the baseline model and GEFeS implementations procedure are provided. The results of the implementations are also analyzed. The combination of population size and the value of k which is used in the Tournament Selection is 15 and 8, respectively, to make the highest fitness. The mutation rate is 0.01. Each model runs with GEFeS for 30 runs. And the best accuracy value and its corresponding feature mask set are collected for each run. The data which are collected by the above step is written to two Excel files for the two models respectively. The number of evolutionary cycles is equal to the max evaluation minus the size of the population then divided by the size of children's population. The max evaluation is 4000 in this project.

A. Baseline Model

The original accuracy values of the rbf-svm, linear svm, and MLP, the accuracy values without using TFIDF, without standardization, and without normalization are listed in Table I.

B. GEFEs with rbf-svm Model

The accuracy values of the 30 runs are shown in Fig. 1. The highest accuracy value is 0.88 while the lowest accuracy value is 0.64. The mean of the accuracy over 30 runs is 0.8 and its standard deviation (STD) is 0.05, which indicates that the accuracy value is stable. Comparing the accuracy of this model, i.e. 0.8, with the accuracy of baseline, i.e. 0.7, the accuracy is obviously improved by involving the feature selection method. The accuracy and the number of features being present in each run is shown in Fig. 2. There is no obvious relationship between the accuracy and the number of features being present.

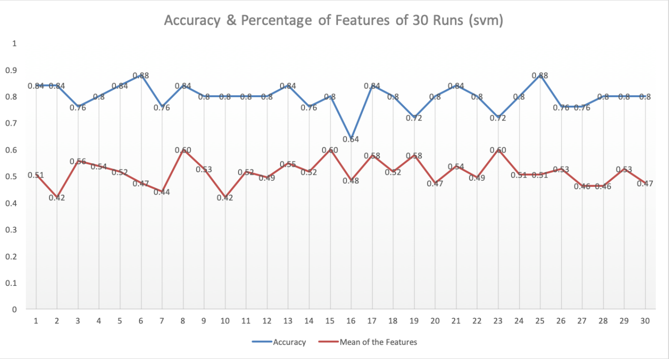

Fig. 2. Accuracy values and percentage of features of 30 runs for rbf-svm.

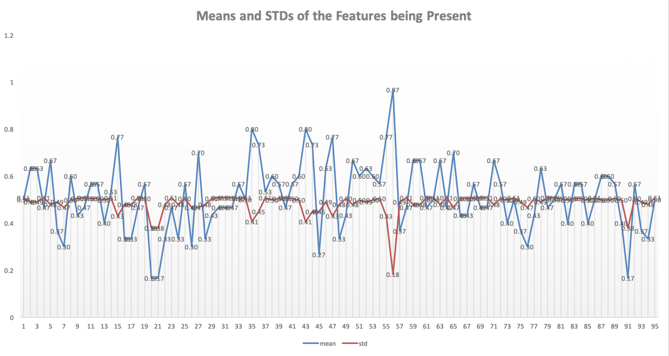

Fig. 3. Means and STDs of the features being present for rbf-svm.

The means and STDs of the features being present over the 30 runs of the GEFEs with rbf-svm model are shown in Fig. 3. When the percentage of a feature is being present over the 30 runs is higher than 0.7, the feature can be regarded as a consistent feature. According to Fig. 3, features 15, 27 35, 36, 43, 44, 47, 55, 56, and 65 are the most consistent features.

C. GEFEs with Linear-svm (lsvm) Model

The accuracy values of the 30 runs are shown in Fig. 4 for lsvm. The highest accuracy value is 0.92 while the lowest accuracy value is 0.68. The mean of the accuracy over 30 runs is 0.82 and its STD is 0.05, which indicates that the accuracy value is stable. Comparing the accuracy of this model, i.e. 0.8, with the accuracy of baseline, i.e. 0.62, the accuracy is obviously improved by involving the feature selection method. The accuracy and the number of features being present in each run are shown in Fig. 5. There is no obvious relationship between the accuracy and the number of features being present.

The means and STDs of the features being present over the 30 runs of the GEFEs with lsvm model are shown in Fig. 6. When the percentage of a feature is being present over the 30 runs is higher than 0.7, the feature can be regarded as a consistent feature. According to Fig. 6, features 14, 19, 35, 36, 43, 50, 51, 52, 53, 56, 72, and 87 are the most consistent features.

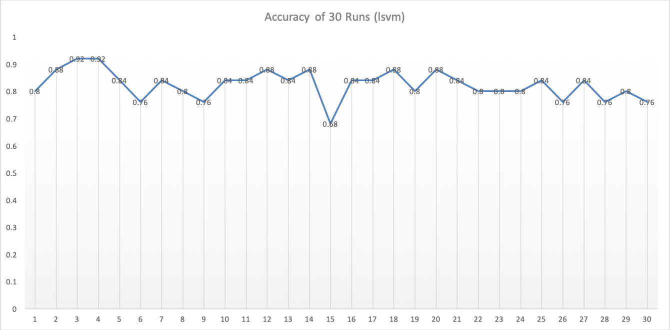

Fig. 4. Accuracy values of 30 runs for lsvm.

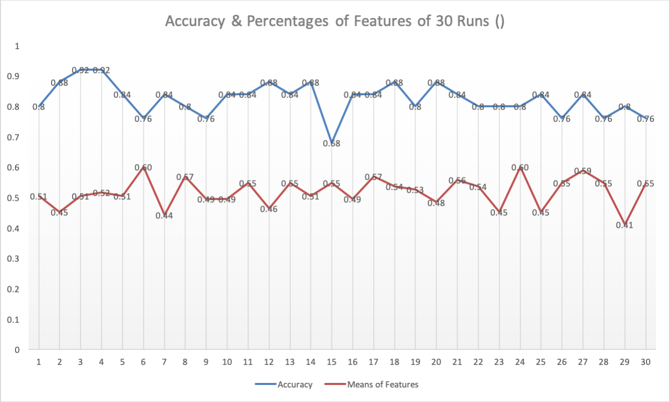

Fig. 5. Accuracy values and percentages of features of 30 runs for lsvm.

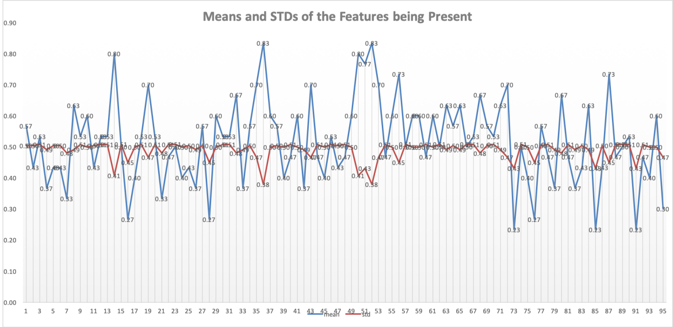

Fig. 6. Means and STDs of the features being present for lsvm.

IV. Results

The performances of the GEFEs with both rbf-svm model and lsvm model are improved. The involving of feature selection method brings the benefit to the Author Identification System since the feature selection reduces the misleading data. Less misleading data indicates improved modeling accuracy. The overlapping of the most consistent features in the two models are features 35, 36, 43, and 56.

The student t-test is used to analyze the means of the accuracy values of the two models. The null hypothesis (H0) is $\mu1 = \mu2$, and the alternative hypothesis (HA), is $\mu1 \neq \mu2$. The p-value of t-test is 0.0484, which is smaller than 0.05. Therefore, the null hypothesis should be rejected. There is statistically significant difference between the accuracy values of the rbf-svm model and lsvm model.

TABLE II
Breakdown of the Work

Work	Name
Project 4	Linyuan Zhang

V. Breakdown of the Work

The breakdown of the work for this report is shown in Table II

References

[1] A. Rexha, M. Kröll, H. Ziak, and R. Kern, "Authorship identification of documents with high content similarity," *Scientometrics*, vol. 115, no. 1, pp. 223–237, 2018.

[2] J. Houvardas and E. Stamatatos, "N-gram feature selection for authorship identification," in *International conference on artificial intelligence: Methodology, systems, and applications*. Springer, 2006, pp. 77–86.

[3] F. Vavak and T. C. Fogarty, "Comparison of steady state and generational genetic algorithms for use in nonstationary environments," in *Proceedings of IEEE international conference on evolutionary computation*. IEEE, 1996, pp. 192–195.