Authorship Attribution with GatorCAAT: Genetic and Evolutionary Feature Selection (GEFeS)

Rahul Alapati
Department of Computer Science and Software Engineering
Auburn University
rza0037@auburn.edu

Sutanu Bhattacharya
Department of Computer Science and Software Engineering
Auburn University
szb0134@auburn.edu

Abstract—**Authorship Attribution can be viewed as a categorization problem. Feature selection is an important factor that has a major impact on the classification accuracy. To determine the most effective features that help discriminate between different authors, we present Genetic and Evolutionary Feature Selection (GEFeS) Algorithms. In this paper, classification algorithms like SVM and Genetic Algorithms (GA) are used to find the classification accuracy of SVM over the CASIS-25 and SEC Sports Writers datasets. SVM is a supervised learning method and in this case a Linear kernel is used for classification. GA is a search and optimization method. We have used Binary and Real coded GAs to improve the classification accuracy of the SVM Classifier.**

Keywords—Authorship Attribution, GatorCAAT, Genetic and Evolutionary Feature Selection, SVM, CASIS-25, SEC Sports Writers

I. Introduction

Support Vector Machines (SVMs) are a set of related supervised learning methods used for classification and regression. Given a set of training examples, each marked as belonging to a particular class, SVM classifies the data by determining a set of support vectors that are members of the training set that outline a hyperplane in the feature space. SVM provides a generic mechanism that fits the hyperplane surface to the training data using a kernel function [1]. The user may select a kernel function (e.g. linear, radial basis function ,polynomial or sigmoid) for the SVM during the training process that selects support vectors along the surface of this function.

Feature selection is the process of removing irrelevant and redundant features, which helps in improving the performance of classification algorithms like SVM. It is also known as attribute selection or feature reduction.

Genetic and Evolutionary Feature Selection (GEFeS) and Genetic and Evolution Feature Selection and Weighting (GEFeWS) are feature selection techniques used to evolve a near-optimal/optimal subset of features in order to maximize the classification accuracy [2]. In order to perform feature selection, a genetic algorithm is used to evolve a set of feature weights. In the set of feature weights, there is a weight that corresponds to each value in the feature vector. In GEFes, a feature weight is a binary value i.e. either 0 or 1, whereas in GEFeWs, a feature weight is a real value between 0 and 1. The set of feature weights are then used as a feature mask, to select the most effective features.

The organization of the paper is as follows: In section II, methodology is elaborated. In section III, the experiments and results are discussed in detail. Section IV ends the paper with discussion on the breakdown of work among the two authors.

II. Methodology

A. Data Collection

As a part of the data collection process, we copy pasted 56 Football Game Recaps of Florida Gators for the first 9 weeks, from 10 daily newspapers [3-12] within the state of Florida written by 16 different sports writers. No recaps were available for Week 8, as it was a bye week for the Gators.

B. Feature Extraction

After acquiring the game recaps, we generate the raw and normalized feature vectors using the Feature Extractor developed as a part of Assignment-1 [13]. In addition to the raw and normalized feature vectors, we apply Tf-idf term weighting to our raw feature vectors, in order to prevent the very frequent characters in our character unigram to shadow the frequencies of rarer yet interesting characters. Tf-idf means term-frequency times inverse document-frequency. We use the scikit learn python library to apply the Tf-idf term weighting to our raw feature vectors [14].

C. Genetic and Evolutionary Feature Selection

To improve upon the baseline accuracy of Linear Support Vector Machine (LSVM), we have used the following GEFeS methods [15]:

1) Steady State Genetic Algorithm

The following is the Steady State GA, which is used to select important and relevant features:

Step 1. Initialize an initial population of 25 binary coded feature masks. Each feature mask consists of 95 feature weights corresponding to the 95 features in a feature vector.

Step 2. Evaluate the fitness of the 25 binary coded feature masks with the accuracy of LSVM as its evaluation metric. The feature mask is applied on the training set and the validation set before training the LSVM.

Step 3. For 4975 evaluations do the following:

a. Select two parents using Tournament Selection.
b. Create a child using Uniform Crossover of the two parents.
c. Mutate the child with a mutation rate of 0.05. In this step, we generate a random number and mutate the bit in the child if the random number is less than or equal to 0.05.
d. Evaluate the fitness of the mutated child.
e. Replace the worst feature mask in the population with the child.

At the end of 5000 evaluations, we evolved a population consisting of 25 feature masks which can be used in feature selection.

2) *Elitist Genetic Algorithm*

The following are the steps involved in feature selection using Elitist GA:

Step 1. Initialize an initial population of 25 binary coded feature masks. Each feature mask consists of 95 feature weights corresponding to the 95 features in a feature vector.

Step 2. Evaluate the fitness of the 25 binary coded feature masks with the accuracy of LSVM as its evaluation metric. The feature mask is applied on the training set and the validation set before training the LSVM.

Step 3. For 4975/24 iterations do the following:

a. Repeat until 24 children are created
    a. Select two parents using Tournament Selection.
    b. Create two children using Uniform Crossover.
    c. Mutate the children with a mutation rate of 0.05. In this step, we generate a random number and mutate the bit in the child if the random number if less than or equal to 0.05.
    d. Evaluate the fitness of the mutated child.
b. Keep the feature mark with maximum accuracy and replace the rest 24 feature masks with the newly created children.

At the end of 5000 evaluations, we evolved a population consisting of 25 feature masks which can be used in feature selection.

3) *Elitist Estimation of Distribution Algorithm*

The following are the steps involved in feature selection using Elitist EDA:

Step 1. Initialize an initial population of 25 binary coded feature masks. Each feature mask consists of 95 feature weights corresponding to the 95 features in a feature vector.

Step 2. Evaluate the fitness of the 25 binary coded feature masks with the accuracy of LSVM as its evaluation metric. The feature mask is applied on the training set and the validation set before training the LSVM.

Step 3. For 4975/24 iterations do the following:

a. Select twelve parents using Tournament Selection.
b. Create a probability distribution function (PDF).
c. Create 24 children by sampling from the PDF 24 times. Here we a create a child using 12-parent uniform crossover.
d. Mutate the children with a mutation rate of 0.05. In this step, we generate a random number and mutate the bit in the child if the random number if less than or equal to 0.05.
e. Evaluate the fitness of the mutated children.
f. Keep the feature mark with maximum accuracy and replace the rest 24 feature masks with the newly created children.

At the end of 5000 evaluations, we evolved a population consisting of 25 feature masks which can be used in feature selection.

In all the methods described above, we have used a Stratified 4 fold cross validation in case of the CASIS-25 Dataset and a Stratified 6 fold cross validation in case of the SEC Sports Writers Dataset. In case of the SEC Sports Writers Dataset, the k value has been optimized in terms of the accuracy, by varying the k value from 2 to 8 representing the number of writing samples per author.

The LSVM classifier has been implemented using the Scikit Learn python library [14].

D. *Genetic and Evolutionary Selection and Weighting*

We introduced weighting into the above mentioned GEFeS, to improve their performance in terms of feature selection. In the three GEFeS mentioned above, we initialized the initial population of 25 feature masks with real values between 0 and 1, instead of binary values. Also, we replaced the mutation step with a Gaussian Mutation. Here, we apply Gaussian mutation with mean 0 and standard deviation 1 on the child, if the random number is less than or equal to 0.05.

III. Experiments and Results

A. *Feature Selection Using GEFeS*

Here, we have analyzed the classification accuracy of LSVM with feature selection using GEFeS. The performances of the three methods in terms of best and average accuracies over 10 runs are reported in the tables below.

Table 1: Classification Accuracy on CASIS-25 Dataset

Method	**Average Accuracy**	**Best Accuracy**
LSVM (Baseline)	0.73	0.73
GEFeS SSGA	0.87	0.86
GEFeS EGA	0.84	0.83
GEFeS EEDA	0.85	0.84

Table 2: Classification Accuracy on SEC Sports Writers Dataset

Method	**Average Accuracy**	**Best Accuracy**
LSVM (Baseline)	0.64	0.64
GEFeS SSGA	0.85	0.83
GEFeS EGA	0.84	0.81
GEFeS EEDA	0.83	0.82

The following are the plots of accuracy curves with respect to the best performances of the 3 methods in the best run.

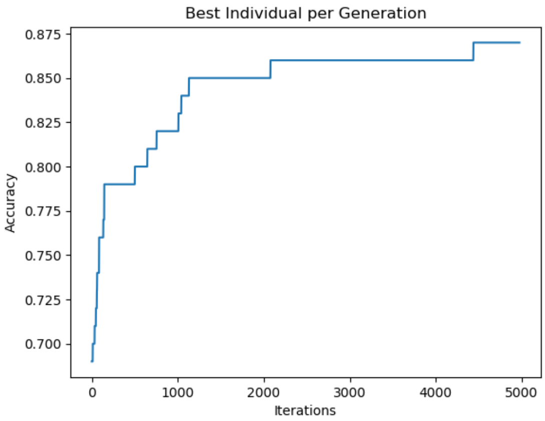

Figure 1: Classification Accuracy of LSVM with Feature Selection using SSGA over CASIS-25 Dataset

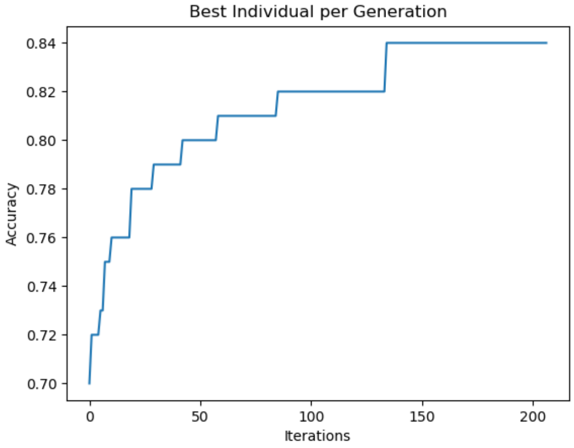

Figure 2: Classification Accuracy of LSVM with Feature Selection using EGA over CASIS-25 Dataset

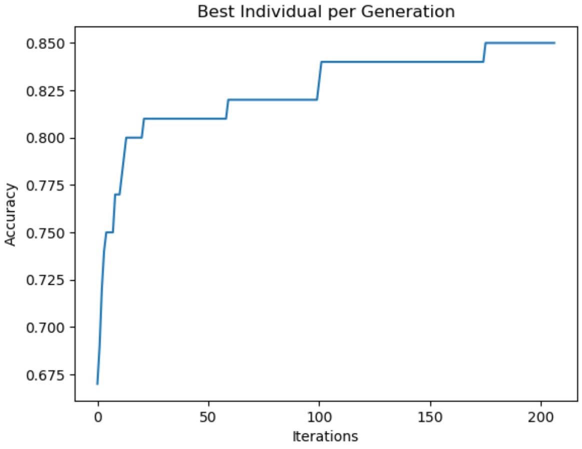

Figure 3: Classification Accuracy of LSVM with Feature Selection using EEDA over CASIS-25 Dataset

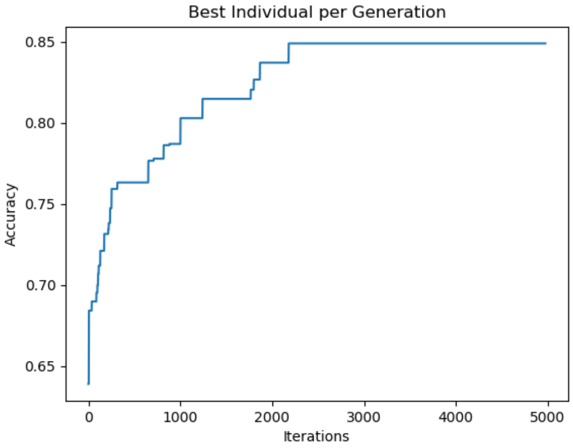

Figure 4: Classification Accuracy of LSVM with Feature Selection using SSGA over SEC Sports Writers Dataset

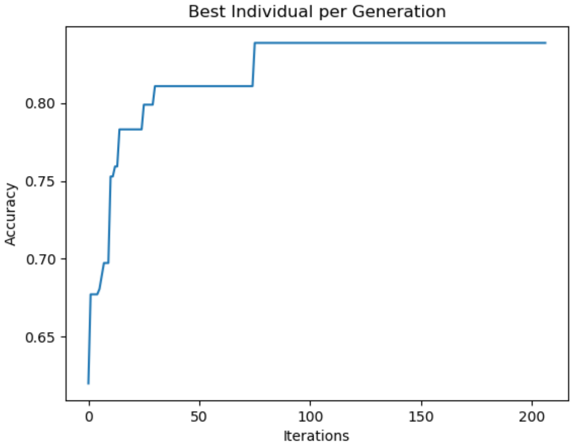

Figure 5: Classification Accuracy of LSVM with Feature Selection using EGA over SEC Sports Writers Dataset

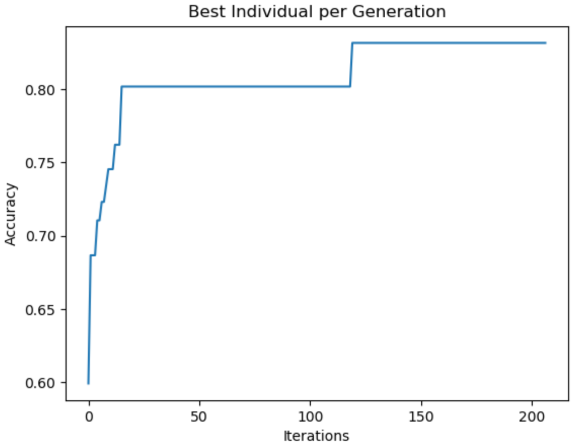

Figure 6: Classification Accuracy of LSVM with Feature Selection using EEDA over SEC Sports Writers Dataset

In all the figures shown above, for every iteration we plot the individual with the maximum accuracy. For Steady State GA we will have 5000 iterations and for Elitist GA and Elitist EDA we will have 207 iterations (4975/24).

From the results shown above, we can determine the impact of different GEFeS on the classification accuracy of LSVM. When benchmarked on the CASIS-25 dataset, the GEFeS result in an increase of 14% in the accuracy when SSGA is used, an increase of 11% when EGA is used and an increase of 12% when EEDA is used.

Whereas on the SEC Sports Writers dataset, the GEFeS result in an increase of 21% in the accuracy when SSGA is used, an increase of 20% when EGA is used and an increase of 19% when EEDA is used.

B. *Feature Selection using GEFeWS*

For improving the accuracy of the GEFeS, we evolved real weights with values between 0 and 1 in the feature mask. The use of real coded GA in the place of binary coded GA, results in an improvement in the accuracy of all the three GEFeS over both the CASIS-25 and SEC Sports Writers Datasets. The tables below report the performances of the three methods in terms of best and average accuracies over 10 runs.

Table 3: Classification Accuracy on CASIS-25 Dataset

Method	Average Accuracy	Best Accuracy
LSVM (Baseline)	0.73	0.73
GEFeWS SSGA	0.91	0.89
GEFeWS EGA	0.92	0.90
GEFeWS EEDA	0.96	0.94

Table 4: Classification Accuracy on SEC Sports Writers Dataset

Method	Average Accuracy	Best Accuracy
LSVM (Baseline)	0.64	0.64
GEFeWS SSGA	0.854	0.83
GEFeWS EGA	0.88	0.85
GEFeWS EEDA	0.89	0.86

The following are the plots of accuracy curves with respect to the best performances of the 3 GEFeWS methods in the best run.

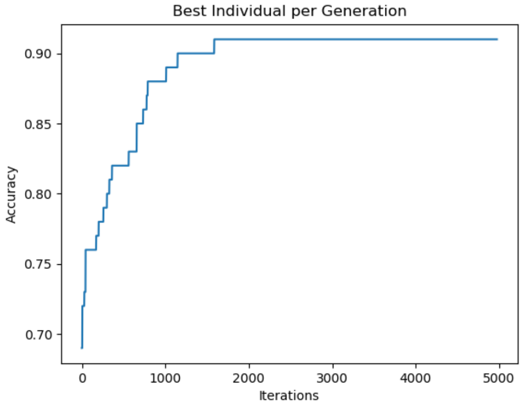

Figure 7: Classification Accuracy of LSVM with Feature Selection using SSGA over CASIS-25 Dataset

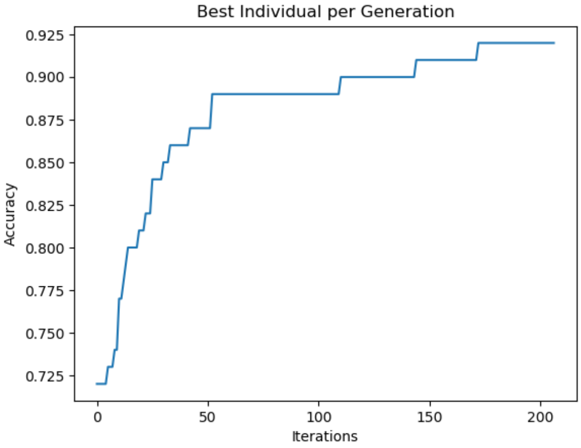

Figure 8: Classification Accuracy of LSVM with Feature Selection using EGA over CASIS-25 Dataset

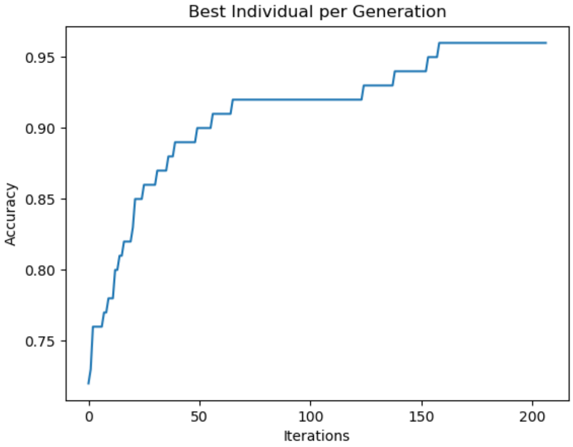

Figure 9: Classification Accuracy of LSVM with Feature Selection using EEDA over CASIS-25 Dataset

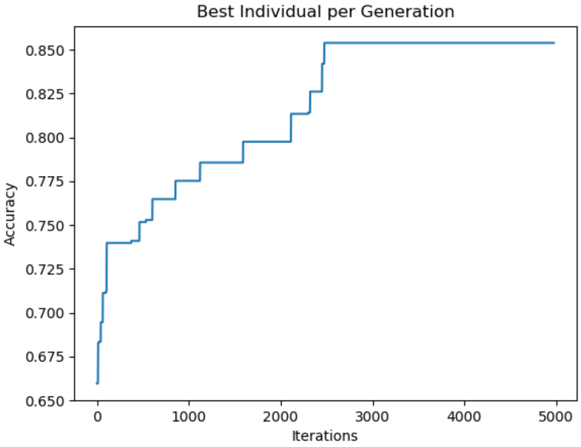

Figure 10: Classification Accuracy of LSVM with Feature Selection using SSGA over SEC Sports Writers Dataset

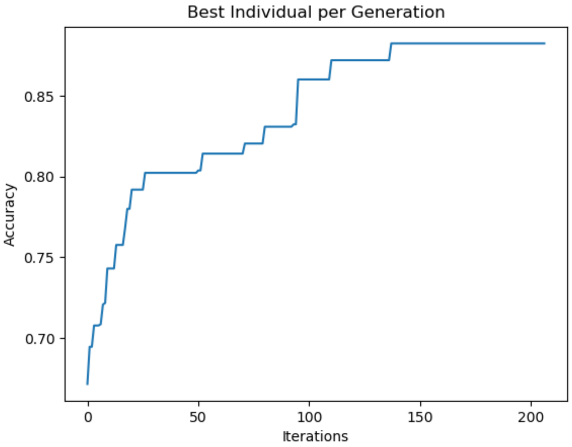

Figure 11: Classification Accuracy of LSVM with Feature Selection using EGA over SEC Sports Writers Dataset

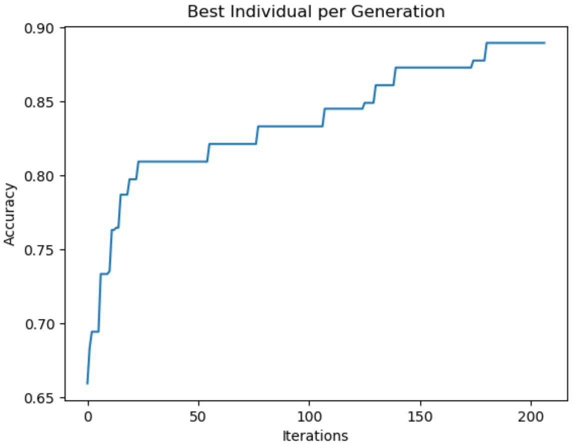

Figure 12: Classification Accuracy of LSVM with Feature Selection using EEDA over SEC Sports Writers Dataset

In the case of GEFeS the feature masks were applied on the raw feature vectors and then the selected features were transformed using tf-idf, standardization and normalization. But, in the case of GEFeWS, the feature masks are applied on the transformed vectors (i.e. after applying tf-idf, standardisation and normalization) to achieve higher accuracies.

When GEFeWS feature masking was applied on the raw feature vectors, we could only achieve an accuracy of 73% for SSGA over SEC Sports Writers Dataset. We also tried to improve the population size to 100 in the case of SSGA and could only achieve an accuracy of 86% over CASIS-25 Dataset. When we tried to increase the mutation rate to 0.2, we could only an accuracy of 85% for SSGA over CASIS-25 Dataset. Finally, we achieved the maximum accuracy when using the real coded GA with Gaussian Mutation in GEFeWS.

From the results shown above, we can determine the impact of different GEFeWS on the classification accuracy of LSVM with GEFeS. When benchmarked on the CASIS-25 dataset, the GEFeWS result in an increase of 4% in the accuracy when SSGA is used, an increase of 8% when EGA is used and an increase of 11% when EEDA is used.

Whereas on the SEC Sports Writers dataset, the GEFeS result in an increase of 0.4% in the accuracy when SSGA is used, an increase of 4% when EGA is used and an increase of 6% when EEDA is used.

IV. Breakdown of Work

Rahul was responsible for the collection of game recaps from newspapers. He was also responsible for implementing the baseline versions of Elitist GA and Elitist EDA. He also proposed improvements such as GEFeWS for Elitist GA and Elitist EDA.

Sutanu was responsible for implementing the baseline version of Steady State GA. He also implemented the improvised version of Steady State GA using GEFeWs and GEFeS with increased population size and mutation rate.

References

[1] Huang CL, Wang CJ. A GA-based feature selection and parameters optimizationfor support vector machines. Expert Systems with applications. 2006 Aug 1;31(2):231-40.

[2] Williams HC, Carter JN, Campbell WL, Roy K, Dozier GV. Genetic & evolutionary feature selection for author identification of html associated with malware. International Journal of Machine Learning and Computing. 2014 Jun 1;4(3):250.

[3] http://www.staugustine.com/.

[4] http://www.ocala.com/.

[5] https://www.alligator.org/.

[6] https://www.orlandosentinel.com/.

[7] http://www.tampabay.com/.

[8] https://www.palmbeachpost.com/.

[9] https://www.local10.com/.

[10] https://www.cbssports.com/.

[11] https://www.tennessean.com/.

[12] https://www.alligatorarmy.com/.

[13] R Alapati, S Bhattacharya, S Dsouza. "Authorship Attribution with GatorCAAT: Data Collection and Feature Extraction".

[14] http://scikitlearn.org/.

[15] Dozier GV. "Assignment 3: Genetic & Evolutionary Feature Selection." Computational Intelligence and Adversarial Learning. Auburn University, AL. 25 Oct. 2018. Lecture.