Authorship Attribution

Performance of various features and classification methods

İlker Nadi Bozkurt
Computer Science Department
Bilkent University
Ankara, Turkey
bozkurti@cs.bilkent.edu.tr

Özgür Bağlıoğlu
Computer Science Department
Bilkent University
Ankara, Turkey
ozgurb@cs.bilkent.edu.tr

Erkan Uyar
Computer Science Department
Bilkent University
Ankara, Turkey
euyar@cs.bilkent.edu.tr

***Abstract*—Authorship attribution is the process of determining the writer of a document. In literature, there are lots of classification techniques conducted in this process. In this paper we explore information retrieval methods such as tf-idf structure with support vector machines, parametric and nonparametric methods with supervised and unsupervised (clustering) classification techniques in authorship attribution. We performed various experiments with articles gathered from Turkish newspaper Milliyet. We performed experiments on different features extracted from these texts with different classifiers, and combined these results to improve our success rates. We identified which classifiers give satisfactory results on which feature sets. According to experiments, the success rates dramatically changes with different combinations, however the best among them are support vector classifier with bag of words, and Gaussian with function words.**

Keywords- Authorship attribution, feature reduction, classifier feature reationship, text categorization, parametric nonparametric calssifiers.

I. INTRODUCTION

Authorship attribution (AA) is the process of attempting to identify the likely authorship of a given document, given a collection of documents whose authorship is known [1]. Authorship attribution becomes an important problem as the range of anonymous information increases with fast growing Internet usage worldwide. Applications of authorship attribution include plagiarism detection (e.g. college essays), deducing the writer of inappropriate communications that were sent anonymously or under a pseudonym (e.g. threatening or harassing e-mails), as well as resolving historical questions of unclear or disputed authorship [2]. Authorship attribution is the way of determining the aouthor of a text when it is unclear who wrote it. It is useful when two or more people claim to have written something or when no one is willing (or able) to state that she or he wrote the piece.

Authorship attribution is a kind of classification problem. But it is different from text classification, because style of writing is also important in authorship attribution as well as text content which is the only factor used in text categorization. Also, with different data (e.g. books, articles), the classifiers and feature sets may behave differently. Also in authorship attribution, the feature set is not deterministic as in text categorization. So, these differences make authorship attribution task more challenging.

The organization of the paper is as follows: In section 2, we briefly mention related work that is done in the area of authorship attribution. In section 3, we explain the steps of our authorship attribution process, in addition to our feature sets. In section 4, we explain the classification methods that we used in our experiments and present the results of these experiments. Section 5 ends the paper with a summarization of work and conclusion.

II. RELATED WORK

There are hundreds of researches conducted about this subject in the last 10 years. With the increasing amount of documents in Internet, and as most of the writings are anonymous, authorship attribution becomes important. The researches are focused on different properties of texts. There are two different properties of the texts that are used in classification: the content of the text and the style of the author. Stylometry - the statistical analysis of literary style - complements traditional literary scholarship since it offers a means of capturing the often elusive character of an author's style by quantifying some of its features [3]. Most stylometric studies employ items of language and most of these items are lexically based.

The usefulness of function words in Authorship attribution is examined by Argamon and Levitan[15]. The authors conducted experiments with support vector machine classifiers in twenty novels and they obtained success rates above 90%. They concluded that, using function words is a valid and good approach in authorship attribution.

According to last researches in 2001, Stamatatos, Fakotakis, Kokkinakis [4] have measured a success rate of %65 and %72 in their study for authorship recognition, which is an implementation of Multiple Regression and Discriminant Analysis.

Also in 2003, Joachim Diederich and his collaborators conducted experiments with support vector classifiers and detected author with %60-80 success rates with different parameters [5].

Kjell [14] performed experiments with neural networks and Bayesian classifiers in this area and obtained about 80-90% success.

In 2006, the effect of word sequences in authorship attribution is studied [16]. The researchers aimed to consider both stylistic and topic features of texts. In this work the documents are identified by the set of word sequences that

combine functional and content words. The experiments are done on a dataset consisting of poems using naive Bayes classifier; the researchers claim that they achieved good results. However because the dataset is obtained from poems, the classification success is expected because of the special structure of poems.

Most of the studies are conducted with one or two classifiers and with limited feature sets. We do not know a comprehensive study in this field. Our study differs from others by conducting many tests with various feature sets and classifiers.

III. STEPS OF AUTHORSHIP ATTRIBUTION

A complete authorship attribution process consists of gathering texts which are the observations to be classified in some sense; a feature extraction mechanism that computes numerical or symbolic information from the observations; and a classification or description scheme that does the actual job of classifying or describing observations, relying on the extracted features [6]. The feature extraction phases of our project are as follows:

A. Dataset Gathering

First, a crawler is used to download documents from Milliyet newspaper's web site. We downloaded all writings of Milliyet columnists from 2001 to 2005. This program acquires the certain dates with the site's default formatting procedure as follows:

http://www.milliyet.com.tr/Year/Month/Date/yazar/writer name.html

With this procedure all column writings between 2001-2005 writers are downloaded. After downloading all these writers, we came to the step of parsing the data to gather pure text and author of the article. Parsing of all pages is done with the help of open source java library HTMLParser [7]. With the help of this library, all HTML tags are cleaned and the data are saved as XML file structure. The structure of this file is as follows:

<DOC>

<DOCNO>X</DOCNO>

<AUTHOR>writer</AUTHOR>

<TEXT>article</TEXT>

</DOC>

After creating this XML structure, the dataset is ready to be processed. This XML file consists of 25559 articles from varying numbers of different authors. However since the number of articles of authors are different, we only used authors that had written more than 500 articles to our dataset. The number of authors that satisfies this criterion is 18 among 34 column writers. The other writers are discarded. And the experiments are done according to these data which have 500 articles from 18 different writers.

B. Feature Extraction

After clearing and acquiring pure data, another fundamental step is to find distinctive features from this data. In authorship attribution not only the content (i.e. the text itself) is important as in IR systems, but also stylometry and other features that define the characteristics of a writer. The features that are extracted from this XML file are as follows:

- Stylometry

The statistical analysis of style, stylometry, is based on the assumption that every author's style has certain features being accessible to conscious manipulation. Therefore they are considered to provide a reliable basis for the identification of an author [5]. The features that are specific to stylometry are as follows: number of sentences in an article, number of words in an article, average number of words in a sentence, average word length in an article, vocabulary size of author (word richness), number of periods, number of exclamation marks, number of commas, number of colons, number of semicolons, number of incomplete sentences [8]. These features can be used together with one classifier.

- Vocabulary Diversity

Measuring the "richness" or "diversity" of an author's vocabulary is also used as a discriminating feature.

- Bag of Words

As Information Retrieval literature, all words (stopwords excluded) are used in document vector which is called vector space model.

- Frequency of function words

The function words (particle, pronoun, conjunction) are used as a discriminating feature of authors. The function words are extracted by the help of data gathered from Turkish Language Association (TDK) that has a list of different types of function words.

After extracting feature sets, now we will describe the classifiers and their performances on different feature sets.

IV. METHODS AND EXPERIMENTAL RESULTS

Various methods are used in the experiments. These methods are Gaussian classifiers which is a Bayesian classifier and parametric method, Parzen windows, histogram methods and k-nearest-neighbor methods which are non-parametric methods, support vector machines which is a non-bayesian classifier and k-means clustering algorithm and neural network approach. We also used principal component analysis (PCA) in conjunction with some of the above methods to reduce the size of the feature space. In the following, the experiments and their evaluation is explained in detail. Because neural network model gives very bad results about 20-30%, we decided not to use it. Also for the experiments we divided our feature sets into

3 parts: bag of words, function words, and stylometric features (also vocabulary diversity included in stylistic features).

A. *Histogram Method*

We used 4 features in histogram method, because we wanted to see the results on a small feature space and if performance of method is good, this feature space would be increased in future experiments. Another reason of using small feature space is the correlation (and parallelism) between some features. We removed the correlated features and used only one of the features that are correlated. The features we used in this method are average number of words in a sentence, average word length, number of different words and number of incomplete sentences.

Success percentages in histogram method		
# of bins (per feauture)	Training Set	Test Set
10	42.2	28.4
30	91.3	16.4
Various size	53.8	31.6

Table 1. Success rates of histogram method.

According to these bin sizes with high number of bins the data is memorized as seen in the success rates of about 90%. However, the results of various parameters (# of bins) are unsatisfactory, so we concluded that histogram method is not useful for stylistic features. From this experiment, we have concluded that our features are not linearly discriminative, because histogram method works with dividing classes into small bins.

B. *K-nearest neighborhood method and Parzen Windows*

For K-nearest neighbor classifier, the experiments are done again with 18 different writers with varying values of the parameter k. In the classification both stylistic features and function words are used. The first set of experiments are conducted by stylistic features, the second set is conducted by function words. According to the experiments using stylistic features, results are in table 2. Although, the training results are satisfactory, test results are not good.

K-nn classifier success percentages (%)					
K=	3	5	7	9	11
Training Data	71,8	68,2	65,9	63,9	63,1
Test Data	46,0	49,1	49,5	50,0	50,4

Table 2. Success percentages with K-nn classifier and stylistic features.

According to table 2 and confusion matrices ((i,j) entry of the confusion matrix holds the number of class i articles that are classified as class j) that we observed, it is clear that stylistic features may be distinctive for some writers, but for others they are not distinctive; so we have to extend our features or use different features and classifiers to combine and get better results. And with different sizes of k, error rates on test data didn't change much. But 50% error rates are too high for classification and this method is no better. Because this method is nonparametric, the training data is memorized by the classifier and in test data these features may change and success is expectedly decreased. So this method is also not suitable for high dimensional data.

The same features (stylistic) are tested also with parzen windows. Training error is 51 % and test error is about 58 %. So these features are also no good with parzen windows.

According to second set of experiments conducted with function words, the results are not satisfactory either and worse than the first set of experiments. Also the function words didn't perform well with parzen windows where the results are about 25% percentage. Lastly, non-parametric methods are not successive in discriminating these kinds of high dimensional text data.

To conclude; from these two sets of results, we can say that nonparametric methods are not good classifiers (parzen windows, histogram method, and k-nn classifier) in authorship attribution. Note that, k-nn classifier and parzen window experiments are conducted by the help matlab library Prtools [13].

K-nn classifier success percentages (%)					
K=	3	5	7	9	11
Training Data	44,5	41,0	39,0	36,4	34,8
Test Data	20,7	23,0	23,6	24,7	25,0

Table 2. Success percentages with K-nn classifier and function words.

C. *Bayes Classifier*

We used normal densities in our Bayesian classifier. We tested this classifier with both the stylometry feature set and the function words. For both feature sets we had 18 different writers each with 500 articles. Half of the data is used for training and the other half is used in testing.

For the stylometry feature set we used and compared two different approaches for our Bayes classifier. First we used same arbitrary covariance matrix for each class (each writer is a different class), which results in a linear discriminant function. Secondly, we used separate covariance matrices for each class. In this case the resulting discriminant function is quadratic. On all our experiments with Bayes classifiers we used equal prior probabilities for each author since we have an equal number of articles from each one. The mentioned stylometric features that are used in these experiments are shown in the table 3.

We tested the classifier both on training and test data. Table 4 shows the success percentages of the two different approaches on training and test data.

Number of Sentences	Number of words	Average Sentence Length
Average Word Length	Number of Different Words	Number of Periods
Number of Commas	Number of Colons	Number of Semiolons
Number of Exclamation Marks	Number of Incomplete Sentences	Number of Question Marks

Table 3. Stylistic features.

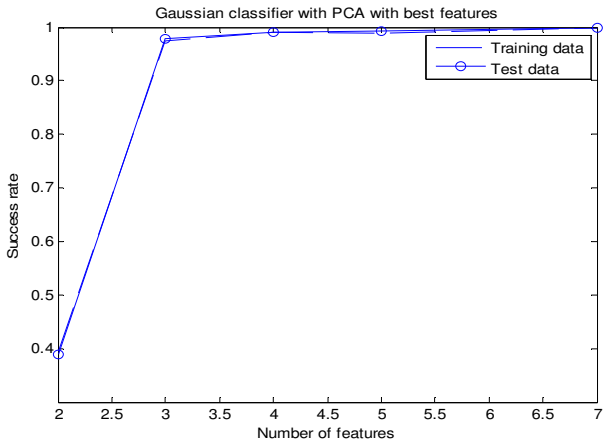
Bayes Classifier Success Percentages (%)		
	Same Arbitrary Covariance Matrix For Each Class	Arbitrary Covariance Matrix For Each Class
Training Data	67,2	74,3
Test Data	60,8	66,9

Table 4. Success percentages of Bayes classifier on stylometry feature set

Despite the positive correlation between some of the features, the above results show the benefits of using stylometry in authorship attribution considering the fact that with our 18 different writers success of a random classifier would be about 0.05.

We used the Bayes classifier with Gaussian density also with the function words feature set. In this dataset we have the count values of function words in each article of each writer. We have 3135 different function words in total which we obtained from Turkish Language Association. Apparently number of features is high for this dataset. Moreover some of these words appear very rarely, suggesting a way of decreasing the number of features.

To decrease the number of features, we took all the function words, removed the infrequent ones (words with maximum count value less than 2 are removed). This step reduced the size of the function words to 476. Then we applied PCA to reduce the size of the feature space furthermore. Surprisingly, the Gaussian classifier with a very small number of features from the result of PCA gave very high success rates. The following figure shows the success rate for this case for various values of feature space size.



The reason behind attaining %100 success rate with just 7 features show that among the more than 3000 function words, may be that there is a very small but distinctive set of words for our dataset.

D. *k-means Clustering*

We used k-means clustering method with the function words data set. We evaluated the validity of the resulting clusters using the following consistency measure:

$$consistency = \frac{1}{n}\sum_{j=1}^{n}\frac{\#\{i \mid GT(i) = GT(j),\ i = 1,\ldots,n\}}{n}$$

In the above formula n is the number of samples in the cluster, GT(i) is the ground truth group to which sample i of that group belongs. j indexes the samples in the cluster. We calculate the consistency of each cluster and then find the average consistency for measuring the validity of the resulting clustering structure.

With the whole set of function words the consistency of the clustering method is 0.23 which is not very good. After removing the infrequent words and reducing the size of the feature space to 100 with principal component analysis (PCA), we applied the k-means algorithm again. This time the consistency of the clustering method is 0.47 which is a significant improvement over the previous case but still is not acceptable. Our experiments with k-means clustering algorithm with the stylometry feature set did not give good results either.

E. *Combination of Classifiers*

In this step, results of classifiers such as Bayesian classifier, k-nearest neighbors, classifier with stylistic features are combined. However the result of this combination doesn't increase our success rates significantly. As an example although, bayesian gives about 60% success rates the combinations gives 56% success rates with median combiner. So combination of classifiers is not a good way in authorship attribution with these classifiers. From these results, we concluded that these classifiers are parallel in predictions, e.g. when one gives false results, others gives these results and success rate doesn't significantly increases, the combination on the contrary decreases.

F. *Support Vector Machines*

After stemming&tokenizing phase, document representations are extracted by using all words of column writers (stopwords are eliminated, stems of words are found). For support vector classifier, we have used vector space model for representing documents [9]. Also for weighting terms the following tf-idf approach is applied.

$w_i = tf_i * \log (D / df_i)$ [9]

where

tf_i = term frequency (term counts) or number of times a term i occurs in a document. This accounts for local information.

df_i = document frequency or number of documents containing term i.

D = number of documents in a database.

After representing each document in vector space model, the documents are ready to be processed by SVMLight classifier tool gathered from [10] [11]. We conducted k-fold experiments with setting k as 2, 5 and 10. By using bag of words as a feature set, the results are as follows:

Success percentages in SVM with bag of words	
K-fold	Success Rate (%)
2	91.2
5	95.1
10	95.7

Table 5. Success percentages of SVM.

The success rate is very high for bag of words. The most important reason behind this success may be authors generally writing about different topics by using different diversity of vocabulary. So, this difference in vocabulary may affect their writing styles. According to these high results we can say that SVM works well with high dimensional data as indicated in [12]. Although SVM gives high success rate for bag of words, it has a high computational burden. That is because of the very large size of the feature space. On a larger data set, the computation of the solution may not be feasible.

V. CONCLUSION

In this work on authorship attribution, we used different feature sets with our data set, which are function words, stylometry feature set and bag of words; and performed experiments on these feature sets using different classifiers such as Bayes classifiers with Gaussian density, support vector machines, histogram, k-nearest neighbor method and Parzen windows and k-means clustering. We used Principal Component Analysis (PCA) wherever reduction of the feature space size is necessary. We obtained best results with Gaussian classifiers on the function words feature set after applying PCA, agreeing with the results Argamon and Levitan[15]. Gaussian classifiers on the stylometry feature set also worked well obtaining around % 60 success rates. Support vector machine classifier is also seen as a very good classifier for authorship attribution obtaining a success rate around %95 on bag of words feature set.

In this work we didn't examine whether the classification errors occur in the same documents among different classifiers. Also we didn't compare the classification errors of the authors using different classifiers. By examining these error rates we can get an insight of the styles of the authors and can better understand why some classifiers work well on some feature sets but not so well on others.

ACKNOWLEDGMENT

We thank Dr. Selim Aksoy for various helpful discussions, to Dr. İlyas Çiçekli for helping us to obtain function words set of Turkish Language Association and Dr. Fazli Can for his help in data gathering. Also thanks to Tubitak for their fellowship which made this work possible.

REFERENCES

[1] Ying Zhao Justin Zobel. "Searching with Style: Authorship Attribution in Classic Literature".

[2] Sanderson J. and Simon G., "Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation".

[3] Laan, N.M. "Stylometry and Method. The Case of Euripides", Literary and Linguistic Computing, 10, 271-278, (1995).

[4] Stamatatos E, Fakotakis N., and Kokkinakis G. "Computer- Based Authorship Attribution without lexical measures". Computers and Hummanities, 2001 pp.193-214.

[5] Diederich Joachim, Kindermenn Jörg, Leopold Edda, and Pass Gerhard. "Authorship attribution with Support Vector Machines" . Applied Intelligence. 2003 pp.109-123.

[6] Pattern Recognition. Wikipedia. http://en.wikipedia.org/wiki/Pattern_recognition

[7] HTMLParser Java Library. http://htmlparser.sourceforge.net/

[8] Diri Banu , Amasyalı M.Fatih. Automatic Author Detection for Turkish Texts. ICANN 2003.

[9] The classic vector space model.http://www.miislita.com/term-vector/term-vector-3.html .

[10] http://svmlight.joachims.org

[11] Joachims T., Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, Schölkopf B. and Burges C. and Smola A. (ed.), MIT-Press, 1999.

[12] Gavrishchaka Valeriy V., Ganguli Supriya B. "Support vector machine as an efficient tool for high-dimensional data processing: Application to substorm forecasting". Journal of Geophsical Research. 2001. pp. 911-940.

[13] PrTools. Matlab based toolbox www.prtools.org.

[14] Bradley Kjell. Authorship Attribution of Text Samples using Neural Networks and Bayesian Classifiers.

[15] Shlomo Argamon, Levitan Shlomo. Measuring the usefullness of function words for Authorship Attribution. *Proceedings of ACH/ALLC Conference 2005* in Victoria, BC, Canada, June 2000.

[16] Rosa Maria Coyotl- Morales, Luis Villasenor- Pineda, Manuel Montes-y-Gomez and Paolo Rosso.Authorship Attribution using Word Sequences.