Authorship Attribution and Scalability

Exploring Gradient Boosting

Laurens Kuiper (s4467299)

December 12, 2018

Abstract

This report presents results of using authorship attribution methods on weblog data. Blogs are regularly updated web pages, typically run by individuals, in an informational or conversational style. Scalability limitations of conventional methods were tested on selections of blogs from 50-150 authors from a publicly available dataset, and an ablative study on the extracted features was performed. The selections were constructed carefully as to not include authors with a low amount of blogs, or a low median blog length. Finally, the performance of a gradient boosting classifier in combination with the conventional methods, inspired by top submissions in various Kaggle competitions, was tested to verify whether this can improve authorship attribution results.

1 Introduction

Authorship attribution is the task of distinguishing texts written by different authors from a set of candidate authors, and finds use in historical and forensic applications. When considering hundreds of authors, and thousands of texts, the problem goes far beyond human potential limits. By extracting textual features from each document it becomes a computational/statistical problem, which can be approached using various machine learning methods. The effectiveness of current techniques varies depending on the context of the task. Included in the variables that affect the performance of authorship attribution systems are the size of the set of candidate authors, the size of the dataset, and the length of the texts.

At the end of the twentieth century the emphasis of many websites on the Internet slowly shifted towards user-generated content, and participation. This had lead to a constant influx of electronic text documents and authors. Many of these authors post several short documents a day in the form of messages, tweets and blogs, sometimes anonymously. The challenges that these developments bring, as well as the flexibility of machine learning approaches, have lead to computational authorship attribution finding more and more applications.

The central research question of this report is how well computational authorship attribution techniques scale with the amount of candidate authors, and to what extent collecting more data improves effectiveness. Finally, experiments with XGBoost, a gradient boosting classifier, in combination with another classifier, will be conducted. The aim is to design a system that scales well with the amount of candidate authors, even when document length is relatively short. All experiments were done on a publicly available dataset containing blog post data collected in 2004, found on Kaggle .

2 Related Work

Before machine learning methods became the dominant approach to authorship attribution, the set of candidate authors was usually kept small [2], and for good reason. Since then, classification and feature extraction methods have improved, and have allowed for larger dataset sizes. Notable feature sets include character N-grams, embedding representations [7], and punctuation use [6]. The most popular and often most effective classifiers support vector machines (SVM) and (convolutional) neural networks [2][4].

Due to recent developments on the internet, the average size of documents has decreased, while the size of the set of candidate authors has increased. For classification models to work well on small documents collecting more data always help, as well as clever feature extraction [5]. In general, when the number of candidate authors increases, the performance of the classifier decreases, but the type of features that are extracted can counteract this effect [3].

On Kaggle, an on-line community for data scientists and machine learners, various machine learning competitions are often held, including classification challenges. These challenges bring out creative solutions. Clever features are extracted, and classifiers are applied in ingenious ways. One classifier that stands out is XGBoost [1], an open source gradient boosting library, which works especially well on tabular data. It makes an appearance in many winning submissions. Even though its speciality is tabular data, it can even be used in image labelling challenges. The approach is often to train another classifier first, and fit XGBoost on the output of the first classifier to fine-tune the prediction.

Our approach is similar. We will first train an SVM, before applying XGBoost. Not only to improve classifier performance, but to improve scalability.

3 Methodology

In terms of machine learning, authorship attribution is a supervised classification task. With classification, the aim is to fit a model on training instances, along with their associated class labels. Here, the training instances are documents, and the class labels are the corresponding authors of the documents.

3.1 Dataset

All experiments were carried out dataset containing blog data gathered from blogger.com in August 2004. The dataset is publicly available on Kaggle.com, and contains $681,288$ posts from $19,320$ bloggers.

[EXPLORATORY DATA ANALYSIS WITH PLOTS TO COME HERE]

The dataset was filtered to deal with outliers etc...

3.2 Feature Extraction

Different feature sets will be discussed here. Many features are being worked on, but nothing is set in stone yet. They will be reported here when they are, and experiments are being done.

3.3 Classification

Classification with SVM will be discussed here, as well as re-classification with XGBoost. However, same as with feature extraction, this is not set in stone, and therefore it is not productive to write about it yet. 10-fold cross-validation will be used.

3.4 Evaluation

Both *precision* and *recall* will be reported, but the model will optimize F_1 score during training.

4 Results

5 Discussion

6 Conclusion

References

[1] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016.

[2] Hitschler, Julian and van den Berg, Esther and Rehbein, Ines. Authorship Attribution with Convolutional Neural Networks and POS-Eliding. In *Proceedings of the Workshop on Stylistic Variation*, pages 53–58. Association for Computational Linguistics, 2017.

[3] Kapočiūtė-Dzikienė, Jurgita and Šarkutė, Ligita and Utka, Andrius. The Effect of Author Set Size in Authorship Attribution for Lithuanian. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)* , pages 87–96. Linköping University Electronic Press, Sweden, 2015.

[4] Sari, Yunita and Vlachos, Andreas and Stevenson, Mark. Continuous N-gram Representations for Authorship Attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 267–273. Association for Computational Linguistics, 2017.

[5] Schwartz, Roy and Tsur, Oren and Rappoport, Ari and Koppel, Moshe. Authorship Attribution of Micro-Messages. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1880–1891. Association for Computational Linguistics, 2013.

[6] Solorio, Thamar and Hasan, Ragib and Mizan, Mainul. A Case Study of Sockpuppet Detection in Wikipedia. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 59–68. Association for Computational Linguistics, 2013.

[7] Vajjala, Sowmya and Banerjee, Sagnik. A study of N-gram and Embedding Representations for Native Language Identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 240–248. Association for Computational Linguistics, 2017.