Authorship Attribution: Comparison of Single-Layer and Double-Layer Machine Learning

Jan Rygl and Aleš Horák

Natural Language Processing Centre
Faculty of Informatics, Masaryk University, Brno, Czech Republic
{xrygl,hales}@fi.muni.cz

Abstract. In the traditional authorship attribution task, forensic linguistic specialists analyse and compare documents to determine who was their (real) author. In the current days, the number of anonymous documents is growing ceaselessly because of Internet expansion. That is why the manual part of the authorship attribution process needs to be replaced with automatic methods. Specialized algorithms (SA) like delta-score and word length statistic were developed to quantify the similarity between documents, but currently prevailing techniques build upon the machine learning (ML) approach.

In this paper, two machine learning approaches are compared: Single-layer ML, where the results of SA (similarities of documents) are used as input attributes for the machine learning, and Double-layer ML with the numerical information characterizing the author being extracted from documents and divided into several groups. For each group the machine learning classifier is trained and the outputs of these classifiers are used as input attributes for ML in the second step.

Generating attributes for the machine learning in the first step of double-layer ML, which is based on SA, is described in detail here. Documents from Czech blog servers are utilized for empirical evaluation of both approaches.

Keywords: double layered machine learning, authorship attribution, similarity of documents.

1 Introduction

Since medieval times, authorship attribution is one of the topics of forensic linguistics. The task is to ascertain whether two documents were written by the same author. Despite the fact that authors frequently hide behind pseudonyms, it is usually possible to obtain a collection of documents that can be with certainty ascribed to the analyzed author. A credible method for judging the authorship of two documents is needed to identify anonymous writers [1,2].

Authorship identification studies were initiated by the work of Mosteller and Wallace [3]. Nowadays, machine learning, information retrieval, and natural language processing are utilized to solve the authorship identification problem. One of the most important approaches are similarity-based models (proposed by [4]).

The process of identification consists of two steps: 1) Document properties are quantified and authors' characteristics (similarities according to these properties) are

P. Sojka et al. (Eds.): TSD 2012, LNCS 7499, pp. 282–289, 2012.

calculated. 2) The characteristics are utilized as attributes for the machine learning (ML). A classifier is trained for the decision whether two documents are written by the same author.

In this paper, two machine learning approaches are compared: Single-layer ML, where the results of heuristic algorithms are used as input attributes for the machine learning, and Double-layer ML with the heuristics being replaced by an another layer of ML. Two-layered ML has been used in other fields, such as biology [5], but it has never been applied to authorship identification. The experiments and evaluation of such application are presented further in this text.

2 Machine Learning

Support Vector Machines method (*SVM*) achieves the best results in comparisons of machine learning approaches to authorship verification problem [6]. For the purposes of experiments the *Orange* [7] implementation of *SVM* was selected. *Orange* is a data mining tool and its implementation of *SVM* is based on the *LibSVM* [8] and *LIBLINEAR* [9] libraries.

The model was trained to estimate the probability of the same authorship instead of class labels (SVM predicts only class label without probability information. The LIBSVM implementation is extending SVM to give probability estimates [10]). The resulting classifier indicates that two documents have the same authorship if the probability of the same authorship is greater or equal to 50 %.

2.1 Single-Layer ML

The single-layer ML process consists of two components: *heuristic algorithms* and *machine learning*.

1. The first component uses five heuristic algorithms (each heuristic represents similarity of two documents based on the selected author's characteristic – these characteristics are described in Section 3). To verify the same authorship of two documents, the heuristics compare differences between documents according to corresponding authors' characteristics and return five values in the interval $\langle 0, 1 \rangle$. The values describe the similarity of two documents (1 means identical documents, 0 is for entirely different authors' styles).
2. The second component, ML, learns a classifier that accepts 5-tuples of numbers. The value returned by the classifier is the probability of the same authorship.

2.2 Double-Layer ML

The double-layer machine learning has three phases: 1) Comparison of attributes of documents, 2) the first layer of ML, and 3) the second layer of ML. The whole process is illustrated by an example in Figure 1.