Universidade de Aveiro

Departamento de Electrónica, Telecomunicações e Informática

Algorithmic Information Theory (2016/17)

Lab work nº 2 — Due: 4 Nov 2016

1 Introduction

Consider the problem of determining the similarity between a target text, t, and some reference texts, r_i. For example, each r_i could be a collection of texts written by known authors and t could be a text whose authorship needs to be determined. The traditional approach to solve this classification problem begins with feature extraction and selection. The features obtained are then fed to a function that maps the feature space onto the set of classes and performs the classification. One of the most difficult parts of this problem is how to choose the smallest set of features that retains enough discriminant power to tackle the problem.

The representation of the original data by a small set of features can be seen as a form of lossy data compression. This suggests a question: can data compression be explicitly used to approach classification problems, removing the need for a separate feature extraction stage? The answer is affirmative. It is possible to adopt an information theoretic approach to classification that bypasses the feature extraction and selection stage. In other words, compression algorithms can be used to measure the similarity between files.

The idea is the following. For each class, represented by the reference texts r_i, we create a model that is a good description of r_i. By a "good description" we mean a model that requires fewer bits to describe r_i than most models or, in other words, that it is a good compression model for the members of the class "r_i". Then, we assign the class to t corresponding to the model that requires less bits to describe (compress) t.

2 Work to be done

1. Develop a program, named `similarity`, that accepts two files: one, with a collection of texts representing the class r_i (for example, representing a certain author); the other,

with the text under analysis, t. Modeling should be performed using the finite-context models implemented in the previous Lab Work. Other parameters, such as the order of the context model and the parameter α of the probability estimator, should also be provided to the program. The program should report the number of bits needed to compress t using the model computed from r_i.

2. As example, you should use the well-known authorship dispute on the Federalist Papers (see `https://en.wikipedia.org/wiki/The_Federalist_Papers`). The 85 texts are available in the moodle (FedPapers.zip). Of course, you may apply your program to other authorship attribution cases that you may find interesting...

3. Elaborate a small report, where you describe all the steps and decisions taken in the work, as well as the relevant results that were obtained.