Abstract title

Feature Engineering and Quick Prototyping of PPI Classifiers

Abstract authors

Francisco Carrero García (1)
Jose Maria Gomez Hidalgo (1)
Manuel Maña Lopez (2)
Jacinto Mata Vazquez (2)

Abstract centers-organizations

(1)
Departamento de Sistemas Informáticos
Escuela Superior Politécnica
Universidad Europea de Madrid
Villaviciosa de Odon, 28670 Madrid, SPAIN
{francisco.carrero,jmgomez}@uem.es

(2)
Departamento de Ingeniería Electrónica, Sistemas Informáticos y Automática
Escuela Politécnica Superior
Universidad de Huelva
Carretera Huelva - La Rábida
Palos de la Frontera, 21071 Huelva, SPAIN
manuel.mana@diesia.uhu.es, mata@uhu.es

Abstract

One of the most relevant steps in learning-based Text Classification tasks is the modeling of the task, which is the definition of a suitable set of attributes, amenable of being used by effective learning algorithms. In fact, the learning step is conveniently supported by a number of machine Learning libraries like WEKA and others. Our work is focused on the analysis of the most suitable attributes for a number of Text Classification tasks. We have developed a framework and software library, JTLib, which allows together the analysis, modeling and fast prototyping of classification systems, supporting both the experimentation phase and the development of functional system prototypes. The library provides the essentials of Text Classification currently not provided by WEKA, and in fact, it is a complement to it.

This library is being used in two R&D projects, Isis and Sinamed [1], whose objective is to enhance Information Access in the medical domain through the improvement and utilization of Text Classification tasks, like Text Categorization, Automated Text Summarization, and Biological Entity Recognition.

1. Document indexing. After processing the training data, a representation based on the selected attributes is obtained and configured into the WEKA ARFF format. Our set of attributes consists of the most relevant words (unigrams), as well as the most relevant pairs (bigrams) and trios (trigrams) of words. Each n-gram becomes an

binary attribute.

2. Dimensionality reduction. During the iterative process, we searched for the n-tuples with higher and lower correlation coefficient to build the attribute vector, and tried with several combinations of amounts of unigrams, bigrams and trigrams.
3. Classifier learning. After several experiments with different Machine Learning algorithms, such as Naïve Bayes, C4.5 decision tree and Adaboost, Adaboost with Naïve Bayes showed to be the most effective. Then, we continued our experiments only with the latter, considering different attributes vectors.
4. Evaluation of text classifiers. The linear increment on the amount of n-tuples used increases precision and F-Measure, but results in a poorer recall. The increment in the amount of bigrams and trigrams produces a higher recall, but with lower precision and F-measure. The experiments performed with the training set proved that increasing the number of attributes would produce similar results, but not necessarily better.

The participation of our team in the PPI task of the Biocreative competition has primary served us as a proof-of-concept for our systematic approach to feature engineering in text classification tasks. We believe we have obtained reasonable results with respect to the effort we have invested in the competitions, moreover not considering external resources apart from the documents themselves.

[1] Buenaga, M.; Maña, M.; Gachet, D. and Mata, J. The SINAMED and ISIS Projects: Applying Text Mining Techniques to Improve Access to a Medical Digital Library. 10th European Conference, ECDL 2006, Alicante, Spain, September 17-22, 2006. 548-551.