Classification of CPP - Application of a Multilayer Neural Network

Almiro Moreira, Ana Carmona, M. Conceição Ferreira, David Santos, Rui Alves
Statistics Portugal

ABSTRACT

One of the most challenging problems in dealing with Census data is the classification of open answers, such as the job classification. The 2011 Portuguese Census data regarding individual jobs were limited to web answers collected through an open answer corresponding to more than 2.5 million cases. The rest of data were collected on paper and processed with OCR (Optical Character Recognition) and, due to their particular characteristics, were left aside in this work. The Standard Occupational Classification (SOC) or CPP classification (in Portuguese) is used as a standard to classify jobs in categories.

The CCP is the set of all professions existing in Portugal and their respective functional description, aggregated by professional groups. It is a fundamental instrument for statistics on occupations, both in terms of observation, analysis, consolidation of series and statistical technical coordination, and for statistical comparability at European and international level at all these common levels. The classification of occupations is relevant not only for the Census but also for other more regular statistical operations such as the Employment Survey (IE) or the Living Conditions and Income Survey (ICOR), for example.

In this work we use a 1-digit classification of the CPP of the 2011 Census (Large Group levels) in a multiclass classification problem (10 classes) by applying a multilayer neural network. Word Embeddings have been used, as a type of word representation that allows words with similar meaning to have a similar representation. Roughly speaking, word embedding, transforms text into numbers. Therefore, a technique like word embedding is used to map words or phrases from a vocabulary to a corresponding vector of real numbers. The algorithm used to learn word embedding was Embedding Layer.

Results show that that after evaluating the classes predicted in the test data, we find out that this model has an accuracy of 90%.