Machine Learning approaches for coding occupations into the new national occupational classification

Keywords: *Socio-professional classification, occupations, automatic coding, fastText, Natural Language Processing*

1 Introduction

Occupational classifications are essential tools built by National Statistical Institutes in order to deal with the diversity of jobs. The French socio-professional classification (professions et catégorie socio-professionnelles, PCS, in French) enables statisticians to group individual professional status based on job contents, together with economic and institutional contexts. This classification is widely used by sociologists and statisticians in order to understand and analyse professional and social differences. Even though this classification relies on very practical questions, its structure is not obvious to the uninitiated ones. Therefore, statisticians may have difficulties for coding when they face survey data with imprecise answers. Moreover, the dictionary used today was set in 2003 (PCS2003) and needed to be updated.

In this context, in 2020, a group of experts was appointed to upgrade the French socio-professional classification (PCS2020) and, among others, to make the production process easier (see Eidelman and Chardon [1]). The use of auto-completed response tools is encouraged by the working group. Indeed, interactivity enables respondents to know what information must be provided. A list of about 6,000 jobs has been settled to produce data in those new categories. The job titles have been enriched with other contextual information such as the economic activity of the company when it was useful for the classification task. The introduction of additional data in the wording simplifies the questionnaire. Now, only 3 (out of the previously 10) other variables (status of the employer, professional position (CEO, executive, worker...), and size of the company) are required besides the textual field about occupation for determining the PCS category.

Even if this protocol is feasible for online surveys, paper surveys are excluded from this protocol. Yet, nearly 1.3 million of paper questionnaires from the French national census have to be coded according to the socio-professional classification each year. And, only 30% of respondents write, without knowing it, an occupation which belongs to the list established by the experts. How can we handle the remaining 70% ?

In this case, it is not possible nor desirable to make available a show card with 6,000 job titles. Only methods after data collection seem to be conceivable. A first solution would be to classify data thanks to a transformation of the output given by of our previous classifier (a rule-based automatic coding system, see Schuhl [2]). However, conversions would be tricky since some transformations from PCS2003 to PCS2020 are not one-to-one. Moreover, we would not take advantage of the shortening of the questionnaire because the ten contextual variables needed previously would still be necessary with the rules.

Another way is to experiment supervised machine learning algorithms to determine which variables are indispensable when using a paper-based, self-administered questionnaire to predict occupation and thus forget the expert-system which is difficult to maintain. However, we do not have any data set labeled into the new classification. A first step is to estimate the number of observations we would need to initialize the models, with the aim to reach at least 80% of good predictions. This threshold matches the accuracy estimate of the current automatic coding expert-system. That accuracy is measured during quality operations. A second step is to consider strategies to improve models each year, especially regarding the manual re-coding.

2 Minimal training set size estimation

In the French census, three kinds of occupations are coded. The questionnaire and, therefore, the collected variables are different for each one. There are :

- Current occupation for employees (PROFS)
- Current occupation for self-employed workers (PROFI)
- Previous occupation for retirees or unemployed (PROFA)

Despite the lack of data coded with the new dictionary, we used the data from 2015 to 2019 coded (automatic coding + manual re-coding) in the previous one. In this context, we were not able to estimate the ground truth accuracy of our models. Instead, we compared our predictions with the coding given by the SICORE process and based on the old dictionary, as we used this to train our model. We assume that the absolute accuracy (when training on a correctly labeled sample) would be very close to this one. Indeed, the new classification is based on the previous one, very often a new categories is a regrouping of old ones. To estimate minimal training set size, we compare models (classifiers × hyperparameters × selected features).

Let $\mathcal{M}$ be the set of models to compare, we applied the methodology below :

Algorithm 1 Pseudo-code for model selection

Randomly shuffle data (in case order)
Split data into two separated sets (train A and test E)
Set the data volume v and build k training datasets $A_1, \ldots, A_k \subset A$ where for all $i \in \{1, \ldots, k\}, |A_i| = v$. Non-overlapping training sets are better.
for $m \in \mathcal{M}$ **do**
    for $i \in \{1, \ldots, k\}$ **do**
        Train m on A_i
        Calculate the accuracy $a_i(m)$ on E with $a_i(m) = \frac{1}{|E|} \sum_{j \in E} 1_{\hat{Y}_j^{m(A_i)} = Y_j}$
    Compute the average $\overline{a}(m) = \frac{1}{k} \sum_{i=1}^{k} a_i(m)$
Choose $\hat{m} = \underset{m \in \mathcal{M}}{\operatorname{argmax}}\ \overline{a}(m)$

No matter the classifier (Random forest[3] or SVM[4] with TF-IDF embedding, fastText[5]...), the selected variables are always the same. For instance with PROFS: job title, professional position and activity of the company are relevant. Additional features do not improve accuracy. Withdrawing one of selected features would damage accuracy.

Among the classification methods we tested, fastText, which is a neural network with one hidden layer created by Facebook, is the most efficient from figure 1. This is a bag of word method based on both word and character n-gram embedding (see Mestre [6]). Sub-word information lead to be more accurate and to deal with word out of the vocabulary. We estimate we need about 5,000 observations for PROFA, 10,000 for PROFI and 70,000 for PROFS.

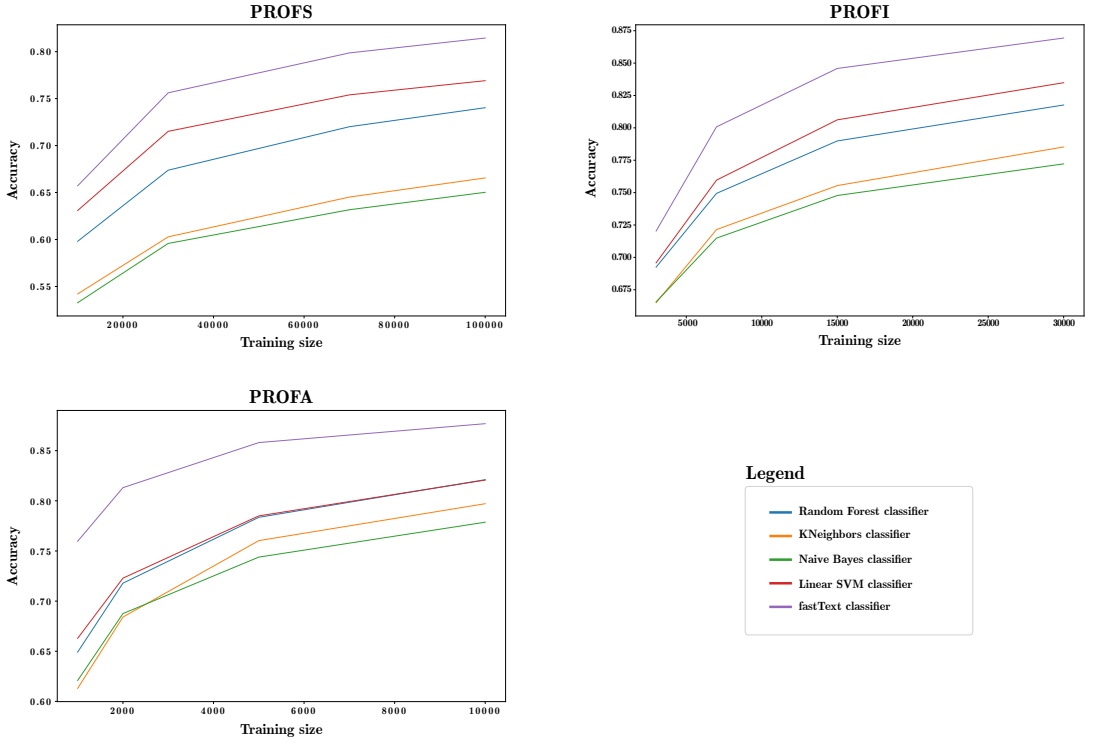

Figure 1: Classifier performance as a function of training volume

3 Considerations for a new automatic coding strategy

Each year, 90% of the paper questionnaires are automatically coded by SICORE, and the remaining 10%, which cannot be coded automatically are re-coded manually. We may benefit from that and ask for manual re-coding some uncertain predictions and then retrain the models. To determine which observations x it would be optimal to re-code, we compute :

$$I(x) = p_1(x) - p_2(x)$$

where $p_1(x)$ and $p_2(x)$ are the first and the second highest class membership scores, respectively. Indeed, the last activation function of fastText neural network is the softmax function therefore we have a normalized score for each K class at our disposal. Empirically, the 10% lowest $I(x)$ are better to send than the 10% lowest $p_1(x)$.

As we can see in Figure 2, models, which are not updated, leads to a deterioration of the performance over time because new occupations are not included in the model. Conversely, if we integrate data each year from initial dataset (see section 2) but also previous and current manual re-coding campaigns (the lowest $I(x)$), we can reach significantly higher accuracy levels.

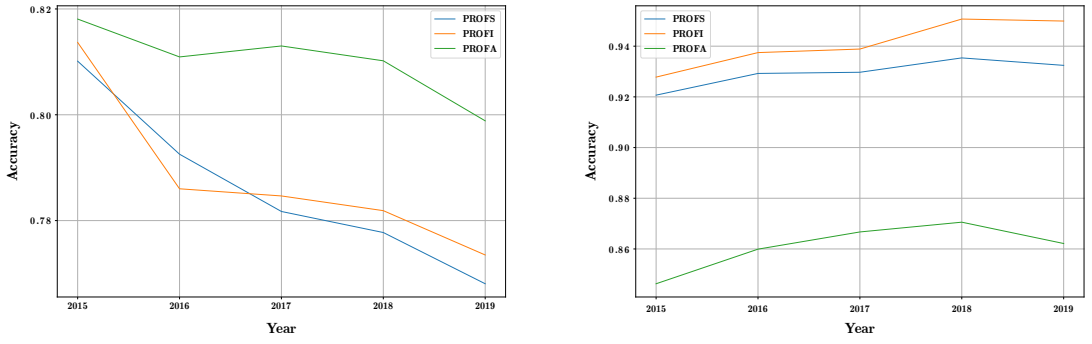

Figure 2: Importance of annual retraining from the accuracy point of view

4 Conclusion

A machine learning-based approach to solve the problem of occupation coding of paper questionnaire could be beneficial. Firstly, this method could provide higher accuracy than the current used expert-system tool. Secondly, it could be more cost-efficient because we would not need an "aggregator" who update system-expert rules each year after the manual re-coding experience. Moreover, a recommendation system could be developed from those models and help during the manual re-coding. Finally, machine learning could be more flexible since we could adjust the volume sent to manual re-coding.

References

[1] Alexis Eidelman and Olivier Chardon. La rénovation de la nomenclature socio-professionnelle (2018-2019). 2019.

[2] Pierrette Schuhl. Sicore, the insee automatic coding system. In *Bureau of the Census 1996 Annual Research Conference and Technology Interchange*. Citeseer, 1996.

[3] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[4] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[5] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[6] Maria Mestre. Fasttext: stepping through the code, Aug 2018. URL https://medium.com/@mariamestre/fasttext-stepping-through-the-code-259996d6ebc4.