Improving efficiency in assignment and quality control of NACE codes combining innovative methodologies with human expertise

Athanassia Chalimourda[1], Lorenz Helbling[2], Mathias Constantin[2]

[1]Swiss Federal Statistical Office, Statistical Methods, Neuchâtel, Switzerland

[2] Swiss Federal Statistical Office, Business Registers Data, Neuchâtel, Switzerland

Abstract

NOGAuto is an assistance system designed to support coding experts in assigning NACE codes to establishments, based on descriptions of their economic activity. This innovative system, developed by the Business Registers Data Section of the Swiss Federal Statistical Office (SFSO) uses Natural Language Processing and Supervised Machine Learning. It exploits the hierarchical structure of the nomenclature générale des activités économiques, NOGA, the Swiss six digits NACE version. Bridging domain and methodological expertise from the Business Registers Data and the Statistical Methods Sections, we employ a series of quality measures to assess the overall and by-class performance of NOGAuto by comparing the agreement of its predictions to existing NOGA codes. Classes of economic activities are inherently imbalanced; we therefore include measures which account for class imbalance like the balanced accuracy. For the 21 NOGA-Sections in our current test set, NOGAuto achieves overall accuracy, balanced accuracy and Cohen´s Kappa of 90%, 87% and 89% respectively. These values are slightly lower at the next lower NOGA level consisting of 88 classes. By-class performance is assessed by measures like precision and recall. While the corresponding work is still in progress, we show in examples of NOGA-Sections how by-class measures combined with the prediction probability can be used to distinguish areas where automatic classification works well from areas where the expert should rather complete the coding task. Although NOGAuto was originally developed to assist experts in their coding work, a further application is planned in the context of quality control of existing codes. In a first use case, NOGAuto helps to limit the effort of detecting misclassifications in a series of about 50'000 codes assigned to activity descriptions in French and German, two of the four official languages in Switzerland. Codes which deviate from NOGAuto predictions are prioritized for a review of their coding, thus streamlining the quality control process. Integrating an innovative system into statistical production is a challenging task. High standards on quality must be met while allowing for progress in innovative methodologies. Continuous monitoring of adequate global and by-class performance measures helps to combine these seemingly contradictory aspects. We show how connecting NOGAuto with other expert-systems, like the automatic translation service DeepL and an SFSO-internal rule-based classification tool fosters efficiency and user-friendliness. The coding experts with feedback on the final code and recorded comments support the development of NOGAuto, thus continuously improving efficiency and quality throughout the coding process.

Keywords: automatic classification, NOGAuto, quality measures, gradient boosting machine

1. Introduction

We briefly describe the NOGAuto assistance system in the following section, *Automatic Classification of Economic Activities*. Supporting the coding process, currently performed

completely manually, results in a reduction of time required, standardization of the coding process and reduction of the interpretation bias due to coding experts with various degrees of experience. In Section *Overall Quality: Global Performance Measures*, we quantify the overall performance of NOGAuto for the NOGA levels Section and Division. These measures should be monitored at every enhancement in NOGAuto or change in the pipeline around it. In the following section, *Quality Measures in Decision Making: Performance by Class*, we describe how by-class quality measures guide the decision which predictions of NOGAuto to trust and which should be double-checked by the coding expert. Although NOGAuto was originally developed to assist the work of coding experts, in section *Quality Control of Existing Codes and Activity Descriptions* we show its usefulness in the context of quality control. In this use case NOGAuto helps to reduce the effort required to detect misclassifications in a series of about 50'000 codes assigned to activity descriptions in French and German. In section *Interaction with Experts and Expert-Systems*, we show how connecting NOGAuto with other expert systems, like an SFSO-internal rule-based classification tool and the automatic translation service DeepL fosters user-friendliness and efficiency.

2. Automatic Classification of Economic Activities

In Switzerland, enterprises can submit their data in one of three national languages (i.e., German, French, Italian) or English. We chose the French Dataset for the design and the training of the algorithm, because of its size and the good overall representation of the NOGA categories. Additionally, the performance of open-source Natural Language Processing (NLP) routines was better in processing French than German.

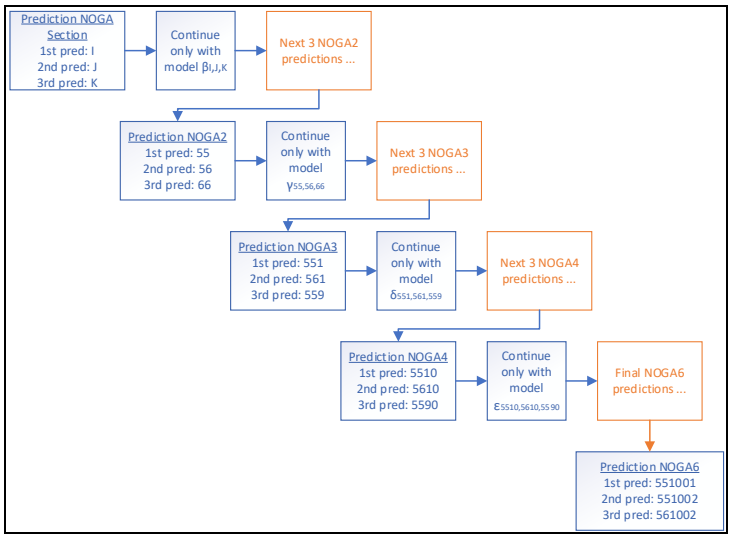
The first step in the automatic classification process involves text cleaning techniques, like eliminating numbers and special characters as well as reducing words to their roots. Word embedding methods assign numeric vectors to activity descriptions, with similar descriptions resulting to nearby vectors.

The classification task in NOGAuto is performed by a Gradient Boosting Machine (GBM), a supervised machine learning algorithm that generates and improves decision trees sequentially, where previous trees have performed poorly. The models were trained with the NOGA-Code as the dependent variable and the activity description as the only independent variable.

NOGAuto exploits the hierarchical structure of the NOGA codes, first classifying an activity description in one of the 21 categories of the level NOGA-Section, from A: Agriculture, Forestry and Fishing, to U: Extra-territorial organisations and entities. Training and classification in the

next level, the first two digits of the NOGA Code, is performed in the data subset defined by the two-digit classes that belong to the three most probable predictions of NOGA-Section. We considered it a good compromise to extend the classification of the next level to include the two NOGAuto suggestions subsequent to the first prediction, due to the observation that sometimes the actual class is the second or third guess of NOGAuto with probability close to the first one. This procedure is continued down to the lowest level of the NOGA hierarchy, as illustrated in the example of Figure 1. The impact of decisions of this kind can be measured by the global performance measures, explained in the next section.

Figure 1: Illustration of the NOGAuto classification through the NOGA hierarchical structure.



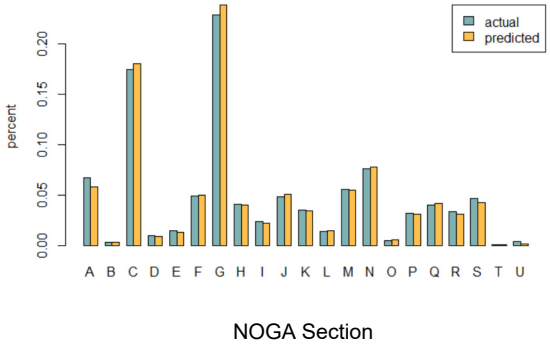
3. Overall Quality: Global Performance Measures

Global Performance Measures assess the overall quality of NOGAuto. They can also measure the impact of decisions in the extended assistance pipeline, like how to exploit better the NOGA hierarchical structure. Examples of measures for overall quality are the overall accuracy, the balanced accuracy and Cohen's Kappa. The overall accuracy is the percentage of elements for which the predicted and the actual class are the same over all elements. Overall accuracy can be dominated by larger classes, whereas smaller classes hardly contribute to its value. Since in the context of economic activities all classes are important independently of their size, the balanced accuracy is employed since it accounts for class imbalance. It is defined as the average of the agreement percentages by class with respect to the actual classes. Cohen's Kappa corrects the overall accuracy for the class agreement expected by chance. Figure 1, left, shows the values of these three global measures for the 21 classes of the NOGA-Section and the 88 Classes of NOGA-Division (the first two digits of the six-digit code) for a test set consisting of 6408 elements. Figure 1, right, compares the empirical distributions of the actual

and predicted classes for the 21 Classes of the NOGA-Section. Note that the predicted class corresponds to the class with the highest assigned probability. Since also small classes exhibit good agreement for this test set, the values of balanced accuracy and Cohen`s Kappa are close to the value of overall accuracy.

Figure 2: Left: the global performance measures for the 21 Classes of NOGA-Section and the 88 Classes of NOGA-Division (the first two digits of the NOGA code). Right: visual comparison between the actual and predicted classes for the 21 classes of NOGA-Section.

Performance Measures	NOGA – Section	NOGA – Division
Accuracy	0.90	0.88
Balanced Accuracy	0.87	0.86
Cohen's Kappa	0.89	0.87



4. Quality Measures in Decision Making: Performance by Class

In the previous section we presented measures for overall quality of the NOGAuto assistance system. In this section we show how we could use by-class quality measures to decide whether to accept the NOGAuto prediction or involve human expertise. Examples of such measures are precision, recall and their harmonic mean, the F1-score.

Precision, or *positive predictive value*, is defined as the proportion of correctly predicted elements in a given class out of all elements predicted in that class, whether correctly or not. Correctly predicted elements are also referred to as *true positives* (TP), while elements that are incorrectly predicted in a given class are also referred to as *false positives* (FP).

$$precision = \frac{\#TP}{\#TP + \#FP}$$

High precision implies that the model makes accurate predictions with a low false positive rate.

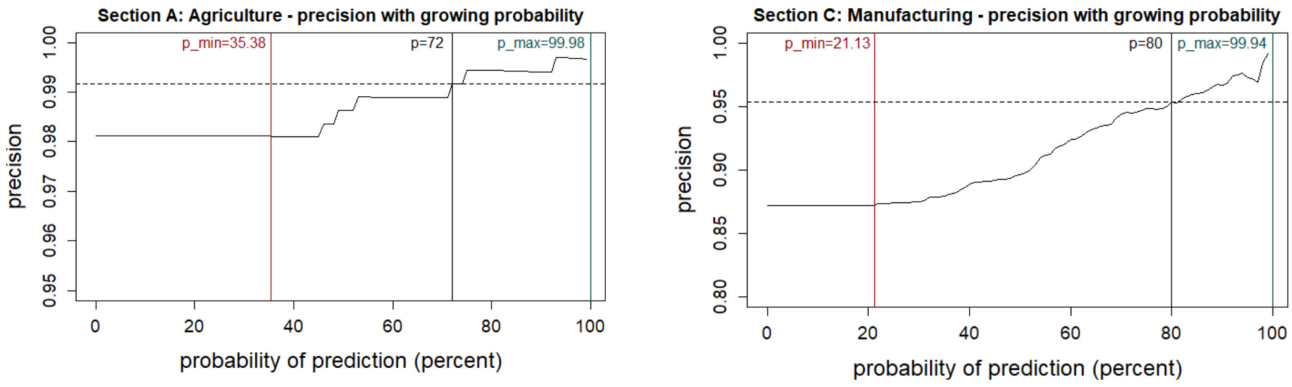
Recall is defined as the number of true positives divided by the number of all positives to an actual class ($\#TP + \#FN$). A high recall value indicates that the model is effectively capturing the relevant elements of a class and has a low false negative rate. The $F1$-*score* is calculated as the harmonic mean of precision and recall, giving equal weight to both measures. A high $F1$-score shows a good balance between precision and recall making it an additional measure for overall classification performance.

Since we are interested in elements for which we can trust the NOGAuto prediction, we focus on precision, which we combine with the element's prediction probability.

Intuitively, we expect that in a class with high precision, the predictions of NOGAuto have a high probability. Figure 3, left, shows that this is the case in the example of NOGA-Section *A: Agriculture, Forestry and Fishing*. Section A has precision over 0.98. The lowest prediction probability of elements predicted to be in it is 35.4%, while the highest is close to 100%, see Figure 3, left, marked by red and green lines respectively. Due to a false positive rate as low as less than 2%, we could accept all predictions attributed to section A in this example. Further restricting the set of predicted elements to those with higher probabilities results in subsets of section A with higher precision. For example, a probability threshold of 72% defines a subset with precision over 0.99, see Figure 3, black vertical line.

In comparison, the precision of section C: Manufacturing/Production of Goods is with 0.87 lower than for section A predictions, see Figure 3, right. The lowest probability in this set is 21.1%. In order to define subsets with higher than the overall precision, a higher threshold to the prediction probability than for section A is needed. For example, an element needs a prediction probability of at least 80% in order to belong to a subset with precision higher than 0.95. We remind that the entirety of section A predictions exhibited a precision over 0.98.

Figure 3: Left: Evolution of the precision with growing prediction probability for the NOGA Section A: Agriculture, Forestry and Fishing. Right: Evolution of the precision with growing prediction probability for the NOGA Section C: Manufacturing/Production of Goods



Determining the thresholds on the prediction probabilities for a large-scale production environment is current work in progress.

5. Quality Control of Existing Codes and Activity Descriptions

Although NOGAuto is originally developed to assist coding experts, we use it for an additional use case in the context of quality control of 50'000 codes and economic activity descriptions from an external administrative database. The original idea was to prioritize for review codes

that deviate from the NOGAuto predictions, thus streamlining the quality assurance process. We were interested in exploring the added value of the innovative NOGAuto system, currently under development, in this quality control context.

Table 1 shows examples where the actual and the predicted code differed, where the actual code was double-checked by coding experts. After the Unit ID (UID) the actual code and its official description are given, followed by the available description in the external database. The available description was the only input for the code prediction of NOGAuto. The official description of the predicted code is given in the last column. We observe that the available description fits better to the predicted code than to the actual code in the database. This suggests that the available description may only be part of the information used by the external data provider to define the actual code.

Table 1: Examples of codes and the available descriptions in the database where the actual code (second column) is different from the NOGAuto prediction.

UID	code	code description	available description	prediction	code description
xxxxx	014100	Dairy cow farming	Removals and transports	494200	Removal transport
xxxxx	014900	Breeding of other animals	Carpentry (installation of windows and doors)	433200	Installation of windows and doors
xxxxx	309201	Manufacturing of bicycles	Bicycle repairs and trade	952900	Repair of other consumer goods

Furthermore, we examined if it is possible to automatically find mismatches between the code and the available description in the database indicating an error in either of the two. Indeed, we can find obvious mismatches, see Table 2 for a few examples. They are found by applying the following two conditions: first, the two codes, actual and predicted, belong to different economic sectors, for example *Agriculture* as opposed to *Services*. Second, the prediction probability is over 90%, indicating that NOGAuto is rather sure for the predicted code.

Table 2: Examples of obvious mismatches between the code and its available description in the database.

UID	code	code description	available description	prediction	code description
xxxxx	711101	Architectural firms	Construction planning	411000	Development of construction projects
xxxxx	711101	Architectural firms	Human Resources	783000	Temporary employment agency
xxxxx	711101	Architectural firms	Real estate expertise	683100	Procurement of properties for third parties

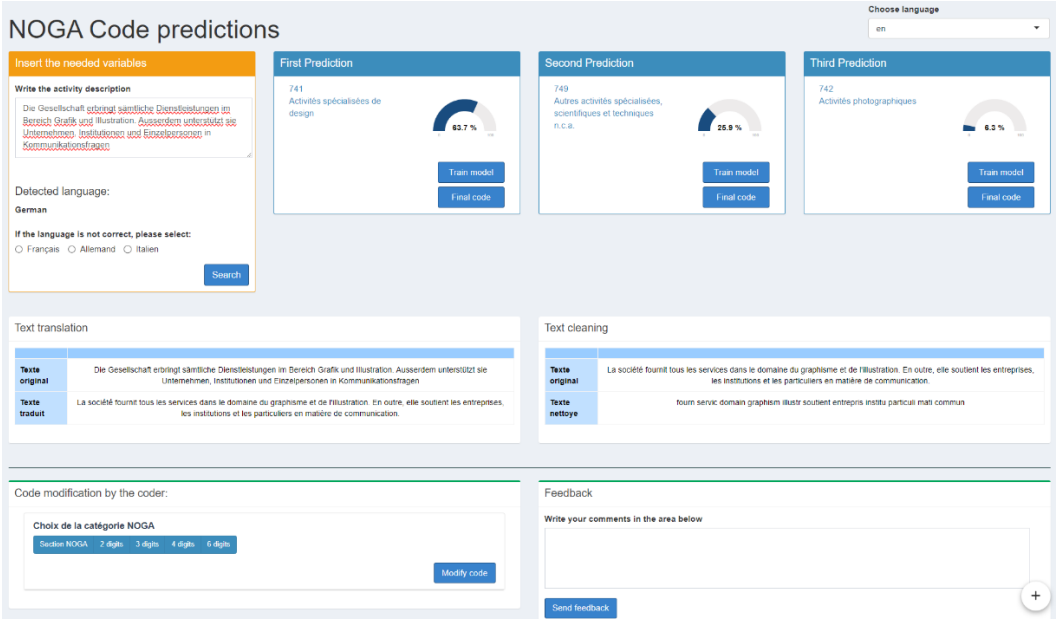
6. Interaction with Experts and Expert Systems

NOGAuto is deployed in two ways, first as an application for the coding experts and second for predicting codes and control of activity descriptions. The application allows the experts to interact with the system, viewing not only the first but also predictions with lower probabilities.

They are able to register an alternative code in the feedback window in case of disagreement. This valuable information is stored for further evaluation and quality assurance purposes.

NOGAuto stops the automatic classification in an intermediate level when, for example, requirements in the class precision and the element's prediction probability are not fulfilled. Even in a semi-automated coding process, time is saved by automatically coding parts of the code that belong to higher levels of the NOGA hierarchy. The application is additionally connected to an SFSO-internal rule-based classification website which allows the expert to have a direct access to the description of the predicted codes. DeepL is used via an API to translate activity descriptions from German into French.

Figure 4: A screenshot of the NOGAuto application



7. Conclusions

In this work we presented how we can improve efficiency in assignment and quality control of NOGA codes with the help of NOGAuto, an innovative system, developed in the Swiss FSO by the Business Registers Data Section with the methodological support of the Statistical Methods Section. We studied quality from various points of view. These stretch from evaluating the overall performance of the assistance system, to punctual, by class quality measures for deciding whether to let NOGAuto predict a six digits code or stops at a higher level to ask for the expert's help.

The GBM models in the core of NOGAuto have been trained with descriptions in French and Italian. Connecting NOGAuto to the translation service DeepL enabled us to use existing models on translations of activity descriptions in German.

Although NOGAuto is currently under development and adaption in order to safely be integrated in a large-scale production environment, it can already be useful for quality control of lists of existing codes and their corresponding activity descriptions. With two straightforward conditions we could detect obvious mismatches between codes and their descriptions in an external administrative database. Our analyses suggested that the information available in the external database was only partially used to determine the code in question, whereas it was the only input for the NOGAuto prediction.

When interacting with the NOGAuto application, the experts retain full control of the coding task. Their feedback on the final code and the recorded comments support the development of NOGAuto, enabling continuous improvement in efficiency and quality throughout the coding process.

References

Helbling, L., Constantin, M., Marx, D., Duc Sfez, C. (2023). Le projet d'innovation des données « NOGAuto ». *Experimental Statistics, Swiss Federal Statistical Office.* NOGAuto | FSO - Experimental statistics (admin.ch).

Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232.

Hastie, T., Tibshirani, R., Friedman, J.H. (2009). 10. Boosting and Additive Trees. *The Elements of Statistical Learning* (2nd ed.). New York, Springer, 337-384. ISBN 978-0-387-84857-0

Cohen, A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.

Solokova, M., & Lapalme, G (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45, 427-437.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5), 1-26. https://doi.org/10.18637/jss.v028.i05