Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

Automatic Coding of Occupations

Arne Bethmann, Malte Schierholz, Knut Wenzig, and Markus Zielonka[1]

Abstract

Occupational coding in Germany is mostly done using dictionary approaches with subsequent manual revision of cases which could not be coded. Since manual coding is expensive, it is desirable to assign a higher number of codes automatically. At the same time the quality of the automatic coding must at least reach that of the manual coding. As a possible solution we employ different machine learning algorithms for the task using a substantial amount of manually coded occupations available from recent studies as training data. We assess the feasibility of these methods by evaluating performance and quality of the algorithms.

Key Words: Occupation Coding; Machine Learning; Naïve Bayes; Bayesian Categorical

1. Introduction

In recent years several German large-scale panel studies demonstrated the demand for the coding of open-ended survey questions on respondents' occupations (e. g. the National Educational Panel Study (NEPS), the German Socio-Economic Panel (SOEP), and the Panel Study “Labour Market and Social Security” (PASS)). So far occupational coding in Germany is mostly done semi-automatically, employing dictionary approaches with subsequent manual coding of cases which could not be coded automatically.

Since the manual coding of occupations generates considerably higher costs than automatic coding, it is highly desirable from a survey cost perspective to increase the proportion of coding that can be done automatically. At the same time the quality of the coding is of paramount importance calling for close scrutiny. The quality of the automatic coding must at least reach that of the manual coding if survey cost is not to be traded for survey error. From a total survey error perspective this would free resources formerly spent on the reduction of processing error and offer the opportunity of employing those resources to reduce other error sources.

In contrast to dictionary approaches, which are mainly used for automatic occupational coding in German surveys, we employ two machine learning algorithms (i. e. *Naïve Bayes* and *Bayesian Multinomial*) for the task. Since we have a substantial amount of manually coded occupations from recent studies at our disposal we use these as training data for the automatic classification. This enables us to evaluate the performance as well as the quality—and hence the feasibility—of machine learning algorithms for the task of automatic coding of open-ended survey questions on occupations.

2. Algorithms

So far machine learning algorithms have only been applied for occupational coding by a few institutions (see e. g. Thompson, Kornbau, and Vesely 2012). The approach we use here has been implemented by Schierholz (2014) using two survey datasets collected by the German Institute for Employment Research (IAB).

[1] Arne Bethmann, Institute for Employment Research, Regensburger Straße 104, Nuremberg, Germany, 90478 (arne.bethmann@iab.de); Malte Schierholz, Mannheim Centre for European Social Research, University of Mannheim, Mannheim, Germany, 68131, (malte.schierholz@mzes.uni-mannheim.de); Knut Wenzig, German Institute for Economic Research, Berlin, Germany, 10108 (kwenzig@diw.de); Markus Zielonka, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, Bamberg, Germany, 96047 (markus.zielonka@lifbi.de)

The task of coding open-ended survey questions on respondents' occupations using an automated machine learning approach can be loosely described by the following three step procedure:

1. Find prior category assignments in the training data for each case in the data to be coded based on the text string from the open-ended survey question q_i and possibly additional covariates x_i

2. Estimate correctness probabilities $\hat{P}_{cor}$ for every occupation category c_j given the information in the training data:
$$\hat{P}_{cor}(c_j|q_i, x_i)$$

3. Assign one (or more) categories to each case based on the correctness probabilities

The first algorithm used is based on the well-known Naïve Bayes (NB) approach:

$$\hat{P}_{cor}\left(c_j\middle|q_i, x_i\right) \propto \hat{P}\left(c_j\right) \times \hat{P}\left(x_i\middle|c_j\right) \times \hat{P}\left(q_i\middle|c_j\right)$$
$$\propto \hat{P}(c_j) \times \hat{P}\left(x_i\middle|c_j\right) \times \prod_{v=1}^{V}(0.95\hat{P}(T_v|c_j) + (1-0.95)\hat{P}(T_v))^{w_{iv}}$$

All estimations $\hat{P}$ in the formulas above are calculated as relative frequencies from the training data. The first formula is derived from Bayes' theorem and the doubtful, "naïve" assumption that covariates X and text strings Q (with their respective realisations x_i and q_i) are stochastically independent given target categories c_j. Because exact matches to verbatim answers are often not found in the training data, strings q_i are split into multiple words T_v. $\hat{P}\left(q_i\middle|c_j\right)$ is then calculated as the product of estimated probabilities that a term T_v is used by a respondent if category c_j is correct. The other algorithm is based on a conjugate Bayesian analysis to the multinomial distribution (BMN). The posteriori expectation then evaluates to
$$\hat{P}_{cor}\left(c_j\middle|q_i\right) = (1-\omega)\hat{P}\left(c_j\right) + \omega\hat{P}\left(q_i\middle|c_j\right)$$

Hereby, $\hat{P}_{cor}$ is a weighted mean of the relative frequency for each category and the relative frequency of the category given that the exact same answer was given in the training data. Weights ω are larger when the answer q_i appears more often in the training data:

$$\omega = \frac{\#\{q_i\}}{\#\{q_i\} + 0.5}$$

A major difficulty for automatic coding is the fact that most jobs are rare in the population and many different verbatim formulations are possible to specify the same job code. Consequently, the ML-estimator $\hat{P}_{cor}\left(c_j\middle|q_i\right) = \frac{\#\{c_j,q_i\}}{\#\{c_j\}}$ will be equal to one if the answer q_i appears only a single time in the training set. This is obviously not desired because if it were true we would code every answer q_i into category c_j. BMN adresses this by downweighting the ML-estimator to a more reasonable value.

A weighted averaged is also included in the NB formula, but weights are not a function of term frequencies T_v and the desired downweighting is not done well. An advantage of NB is instead that answers q_i are split into words $\#\{T_v\}$ which makes it possible to predict categories when no exact matching answers are found. More detailed descriptions about these algorithms can be found in Schierholz (2014).

3. Data

The key ingredient to good machine learning results is a large amount of high quality data. For the analyses conducted here we use data from the German National Educational Panel Study (NEPS; Blossfeld, Roßbach, and

Maurice 2011). The data consist of coded answers from diverse open-ended questions on occupations and occupational aspirations in all starting cohorts of the NEPS. Coding was conducted using the recently updated German “Klassifikation der Berufe 2010” (KldB2010; Bundesagentur für Arbeit 2011). So far the coding process makes extensive use of dictionaries and automated suggestion systems developed at the NEPS Research Data Center:

All open-ended answers were coded manually. But human coders were supported with computer generated suggestions derived from previously coded material from previous waves, the official classification terms and key words delivered by the Federal Employment Agency (BA). In cases of nearly identical strings only one suggestion has been presented. In a second loop an expert supervisor proved the results. The manual steps are supported by the search engine capabilities of the German Federal Employment Agency. Despite of the computer driven suggestion system the whole process still relies on manual coding and proving to ensure high data quality (Munz, Wenzig, and Bela forthcoming).

This data from the NEPS with more than 300.000 already coded answers are an ideal source for training data (where one uses the coded results as input information for the algorithms) and test data (where the existing codes are used to check the results of the coding algorithms).

4. Preliminary Results

In order to compare classification performance we ran both algorithms on subsets of the training data with varying sample sizes. Figure 1 shows the results for both algorithms using 300,000 cases in the training data and a sample of 7,500 cases as the test data to be coded.[2]

The x-axis indicates the *production rate* meaning the percentage of the test cases that were assigned automatically. For these production rates the corresponding *cumulative agreement rates* are plotted on the y-axis. The cumulative agreement rates represent the percentage of correctly coded cases, meaning that the automatic classification of the algorithm and the manual classification according to the training data are in agreement. The agreement is expected to drop with an increasing production rate. This is due to the algorithm coding easy cases first.[3] Increasing the production rate then forces the algorithm to also classify cases with more uncertainty and hence increases the amount of classification error.

Regarding high quality classification the BMN approach delivers considerably better results. At 50 % production rates the agreement rate for BMN is about 94 % and a little under 90 % for NB. Up to a production rate of about 80 % BMN performs better than the NB algorithm. At this point both algorithms provide an agreement rate of approx. 83 %. At higher production rates NB performs a little better giving an agreement rate of about 73 % compared to only 68 % for BMN. While the agreement rates are generally lower for smaller training datasets, the relative performance of the two algorithm is very similar across all sample sizes.

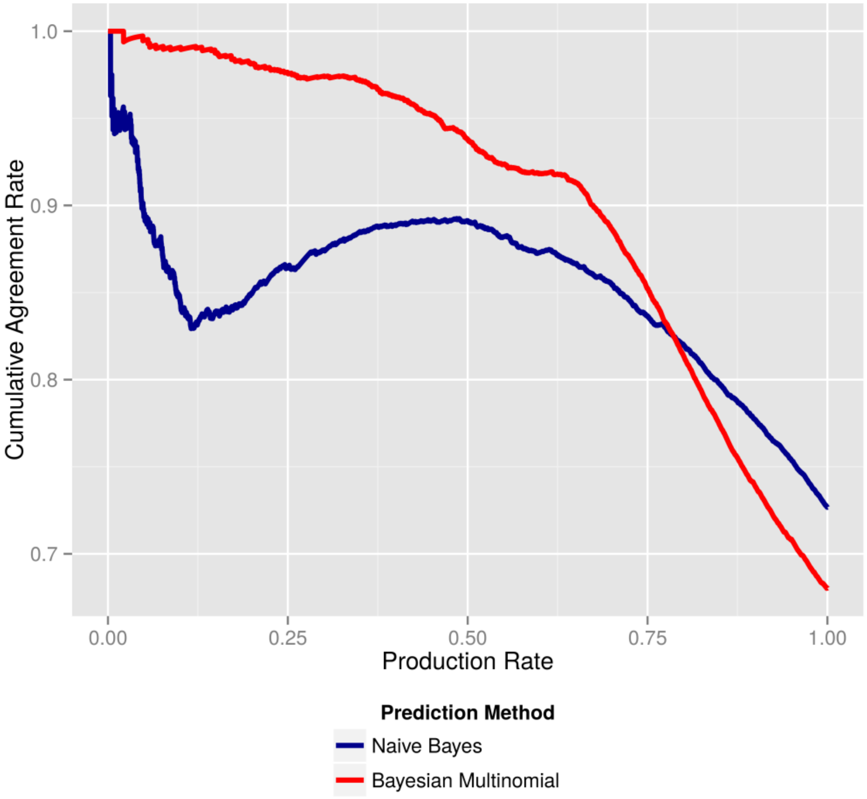
It is often not the level of agreement at a certain production rate that is of interest for the evaluation of the coding algorithm, but rather the production rate that could be achieved at a fixed agreement rate. For the automatic coding of the American Community Survey (ACS) only an agreement rate as high as 95 % is considered acceptable (Thompson, Kornbau, and Vesely 2012). In table 1 we compare the performance of both algorithms at the 95 % and the—somewhat gentler—90 % agreement level.

At both levels NB performs far worse than BMN. With production rates consistently under 5 % NB seems unusable for practical purposes. Although when looking at figure 1 it seems that slightly lower levels of fixed agreement rates would most likely yield acceptable production rates.

[2] At the moment this is the result for a single sample. Cross-validation checks are planned and will be reported in the future.

[3] Meaning cases with high correctness probabilities for any of the classes.

Figure 1
Production Rates



NEPS data; $\boldsymbol{N_{Test} = 7,500}$***;*** $\boldsymbol{N_{Training} = 300,000}$

Table 1
Production Rates

	Fixed Agreement Rates			
	90 %		95 %	
$N_{Training}$	NB	BMN	NB	BMN
25,000	0.5	51.5	0.3	36.4
50,000	0.9	55.6	0.1	38.1
100,000	1.1	60.2	0.3	41.9
300,000	4.9	67.4	3.1	45.8

NEPS data; $\mathrm{N_{Test}} = 7{,}500$; NB = Naïve Bayes; BMN = Bayesian Multinomial

Performance at the different sample sizes develops as expected: larger datasets produce better classification results. Clearly, for practical applications sample sizes of the training data should be in the high six digit range, at least for the scenario analysed here.[4]

[4] Results for different sample sizes of the test data to be coded were estimated, but since there was no considerable variation they are not shown here.

Table 2
Correlations: Status and Prestige Measures

	ISEI-08		SIOPS-08		% valid	
$N_{Training}$	NB	BMN	NB	BMN	NB	BMN
25,000	0.904	0.968	0.922	0.973	78	54
50,000	0.907	0.966	0.928	0.973	80	58
100,000	0.917	0.967	0.937	0.973	82	61
300,000	0.929	0.964	0.945	0.971	85	68

NEPS data; $\mathrm{N_{Test}} = 7{,}500$; NB = Naïve Bayes; BMN = Bayesian Multinomial

For many applied research questions the correct occupation code is not the top priority. It would often suffice to have a reasonably precise account of a person's socio-economic status or occupational prestige. The scales used for these purposes are often derived from the occupation code (e. g. ISEI or SIOPS derived from ISCO). Although occupations might differ in many aspects similar occupations often also provide similar levels of prestige or socio-economic status. Therefore even cases where the automatic coding provided a wrong classification may produce correct status and prestige measures as long as the occupation code that was assigned leads to the same ISEI or SIOPS score.

We assessed the quality of the ISEI and SIOPS measures for the automatic codings of both algorithms by correlating the derived scores with those derived from the manual coding.[5] NB as well as BMN yielded quite good results with correlations ranging between 0.904 for NB at a sample size of 25,000 up to 0.971 for BMN at 300,000 cases in the training data (see table 2).

Only cases that had a valid code in the manually and the automatically coded data where used for the analysis. For BMN this meant considerably less cases than for NB and might explain some of the advantage that BMN had over NB. The cases not coded by BMN would most likely be cases with high uncertainty. Leaving these uncoded would mean less classification error among the coded cases and hence better correlations for the derived status and prestige measures.

The quality improvement with increasing size of the training data sample is far less pronounced than what was found for the agreement rates. For BMN it is even non-existent.

5. Conclusion & Further Research

From these preliminary results we would conclude that automatic coding of open-ended survey questions on occupations seems feasible, provided that you have a large enough sample of high quality training data. Both algorithms perform acceptable although BMN has an advantage over NB when high fixed agreement rates (>= 90 %) are desired. In cases where only derived prestige and status scales are needed even smaller datasets might be sufficient to provide acceptable estimates.

We labeled our results "preliminary" for several reasons. First, we have to do more analyses with data and algorithms discussed here in order to validate our results and check their robustness. Therefore cross-validation checks will be one of the next tasks.

Secondly, the algorithms might be improved further. Apart from testing additional machine learning algorithms (e. g. *Random Forests* or *Support Vector Machines*), we will try to optimize NB and MNB further. One idea is to incorporate distance measures (e. g. Levenshtein distance) into the algorithms in order to make better use of the

[5] For these analyses we first recoded the KldB2010 codes of both the automatic and the manual coding to ISCO08. This works reasonably well due to the KldB2010 being developed with the ISCO08 in mind. Then the ISCO08 codes were used to derive the ISEI and SIOPS scores. A native ISCO08 coding would have been preferable but was not available in the data.

information in the training data. First trials have yielded some promising results. Further, we will try to develop a sensible form of preprocessing of the data in order to get cleaner text strings, which in turn should reduce the amount of noise in the data and lead to more precise estimates of the correctness probabilities.

Lastly, one of the main long-term aims of the project is to develop a best practice for automatic occupation coding which should then be applied to several large-scale surveys in Germany.

References

Blossfeld, H.-P., H.-G. Roßbach, and J. von Maurice, eds. (2011). "Education as a Lifelong Process. The German National Educational Panel Study (NEPS)", *Zeitschrift für Erziehungswissenschaft - Special Issue* 14. VS Verlag für Sozialwissenschaften.

Bundesagentur für Arbeit, ed. (2011). Klassifikation der Berufe 2010. Nürnberg.

Munz, Manuel, Knut Wenzig, and Daniel Bela (forthcoming). "String coding in a generic framework".

Schierholz, Malte (2014). "Automating Survey Coding for Occupation". FDZ-Methodenreport, 10/2014. Nürnberg.

Thompson, M., M. E. Kornbau, and J. Vesely (2012), "Creating an Automated Industry and Occupation Coding Process for the American Community Survey", unpublished report.