Automatic Coding of Occupations

Using Machine Learning Algorithms for Occupation Coding in Several German Panel Surveys

Arne Bethmann[1] Malte Schierholz[2]
Knut Wenzig[3] Markus Zielonka[4]

VI European Congress of Methodology
Utrecht University
24 July 2014

[1]Institute for Employment Research (IAB) [2]Mannheim Centre for European Social Research (MZES)
[3]German Institute for Economic Research (DIW) [4]Leibniz Institute for Educational Trajectories (LIfBi)

Agenda

1. Problem
2. Algorithms & Data
3. Preliminary Results
4. To do

Problem

High quality occupation coding is …

- … mostly “manual labour” and therefore expensive
- … complex to organise and time consuming
- … to be done repeatedly for different native coding schemes
- … hard to monitor by the researcher since often done by the field institute

⇒ (Partial) Automation could free up ressources which would be more efficiently invested in other aspects of survey quality

Examples

Partially Automated

- GESIS Coding Guidelines
- ALWA study (IAB)
- National Survey on Public Health (Stat. Sweden)
- Cascot (Warwick Institute for Employment Research)
- G-CODE (Stat. Canada)

Machine Learning

- I & O Autocoder (American Community Survey)
- COBS (Stat. Netherlands)

$\Rightarrow$ So far no ML-based coding in Germany

Algorithms

Task: Assign appropriate occupation code to participant's answer to open-ended occupation question

1. Find prior category assignments for text string (q_i – and possibly additional covariates x_i) in training data
2. Estimate correctness probabilities ($\hat{P}_{cor}$) for every occupation category (c_j)

$$\hat{P}_{cor}(c_j|q_i, x_i)$$

3. Assign one (or more) categories to answer

Algorithms

Complex Learning Problem

- Many categories – 1,286 using KldB2010
- Messy input strings due to misspelling etc.

Naive Bayes

- Use frequencies of single words in conjunction with every category to estimate probabilities
- Naive: Words occur inependently of one another

Bayesian Multinomial

- Model uncertainty when only very few identical verbatim answers are available in the training data

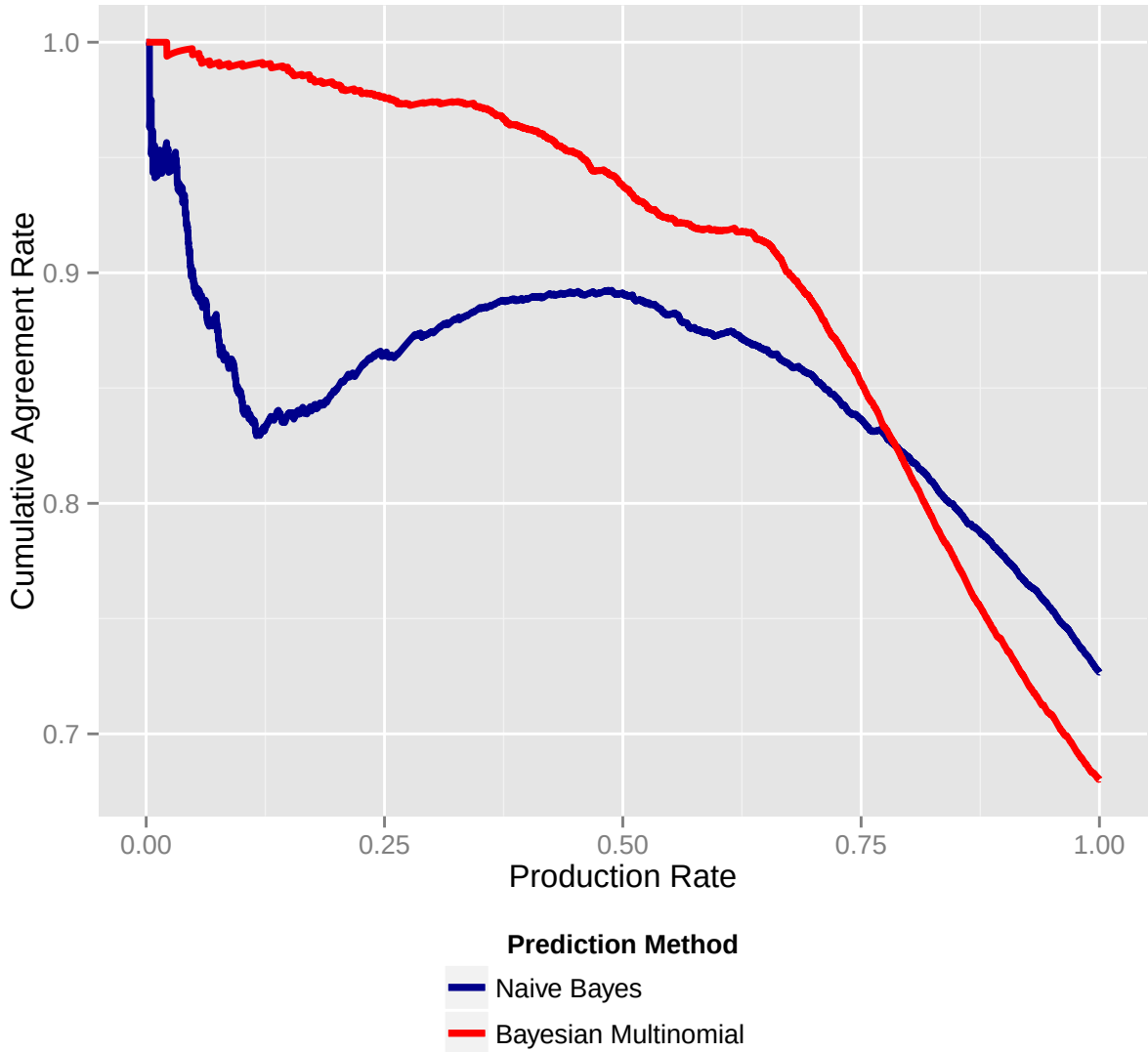
Data

National Educational Panel Study (NEPS)

Longitudinal survey conducted in Germany on the development of competencies, educational processes, educational decisions, and returns to education in formal, nonformal, and informal contexts throughout the life span

- 300,000+ occupation coded (KldB2010, ISCO08) verbatim answers
- High quality manual coding with close scientific supervision

Agreement vs Production Rates



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

Production Rates

	Fixed Agreement Rate			
	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; N_{Test} = 7,500

NB = Naive Bayes

BMN = Bayesian Multinomial

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; N_{Test} = 7,500

NB = Naive Bayes

BMN = Bayesian Multinomial

Summary

Direct classification

- Naive Bayes better algorithm at 100% production rate but Bayesian Multinomial better when high agreement rates are desired
- Larger training data improves results substantially

Derived indices

- Show very high correlations in all cases
- Bayesian Multinomial slightly better but does cherry-picking

$\Rightarrow$ Promising preliminary results. Further research needed ...

To do

Algorithms

- More analyses and cross-validation
- Covariates
- Additional algorithms
- Messy strings

Application

- Live coding system
- Apply to other datasets:
  - German Socio-Economic Panel Study (SOEP)
  - Panel Study “Labour Market and Social Security” (PASS)
- Find best practice
- (Develop R package)

Contact

Arne Bethmann arne.bethmann@iab.de

Malte Schierholz malte.schierholz@mzes.uni-mannheim.de

Backup

Estimate for every respondent i the probability

$$\hat{P}_{kor}(c_j | verbatimanswer_i, professionalstatus_i)$$

that job category $c_j, j = 1, \ldots, 1286$ is correct

Different algorithms available:

- *Official Dictionary*: Match verbatim to database with official job names
- *Naive Bayes*: Split verbatim answers into words, naive assumption of conditional independence to break down high-dimensional space
- *Bayesian Multinomial*: Model uncertainty when only very few identical verbatim answers are available in the training data
- *Combined Method*: Combine previous algorithms with multiple indicators from a database

Naive Bayes

Well known algorithm for text classification

$$\hat{P}_{kor}(c_j|q_i, x_i) \propto \hat{P}(c_j) \times \hat{P}(x_i|c_j) \times \hat{P}(q_i|c_j)$$
$$\propto \hat{P}(c_j) \times \hat{P}(x_i|c_j) \times \prod_{v=1}^{V} (\lambda \hat{P}(T_v|c_j) + (1-\lambda)\hat{P}(T_v))^{w_{iv}}$$

A verbatim answer q_i often consists of only a single word that appears very few times in our small training data.

Bayesian Multinomial

More frequent answers with less coding ambiguity should have higher probabilities

$$\hat{P}_{kor}(c_j|q_i) = \omega\hat{P}(c_j|q_i) + (1-\omega)\hat{P}(c_j)$$

with greater weights ω when the answer q_i appears more often in the training data:

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

Formulas are motivated by the conjugate Bayesian model for the multinomial distribution

$\hat{P}$ are relative frequencies