Using Web Data to derive the Economic Activity of Enterprises

Manveer Mangat

UNECE Machine Learning for Official Statistics Workshop 2023, June 2023

www.statistik.at

Independent statistics for evidence-based decision making

STATISTICS AUSTRIA
The Information Manager

Research Objectives

1. Construct a hierarchical classification model to predict NACE level 1-5 codes of enterprises on the basis of their scraped websites
2. Propose evaluation measures which are more suitable to asses the performance of hierarchical models than the standard evaluation metrics

Hierarchical Classification

Data Set

Data: pairs of enterprise and URL links (enterprises part of the ICT survey in 2019-2021)

- URL linking done on the basis of the **Statistical Business Registry (SBR) (*Ground Truth*)**
- When VAT/CID of an enterprise found on a scraped webpage -> webpage linked to the respective enterprise -> we obtain its NACE 1-5 code from the SBR (NACE codes available on **all** levels)

-> provides approx. 72k pairs

Processing of scraped webpages:

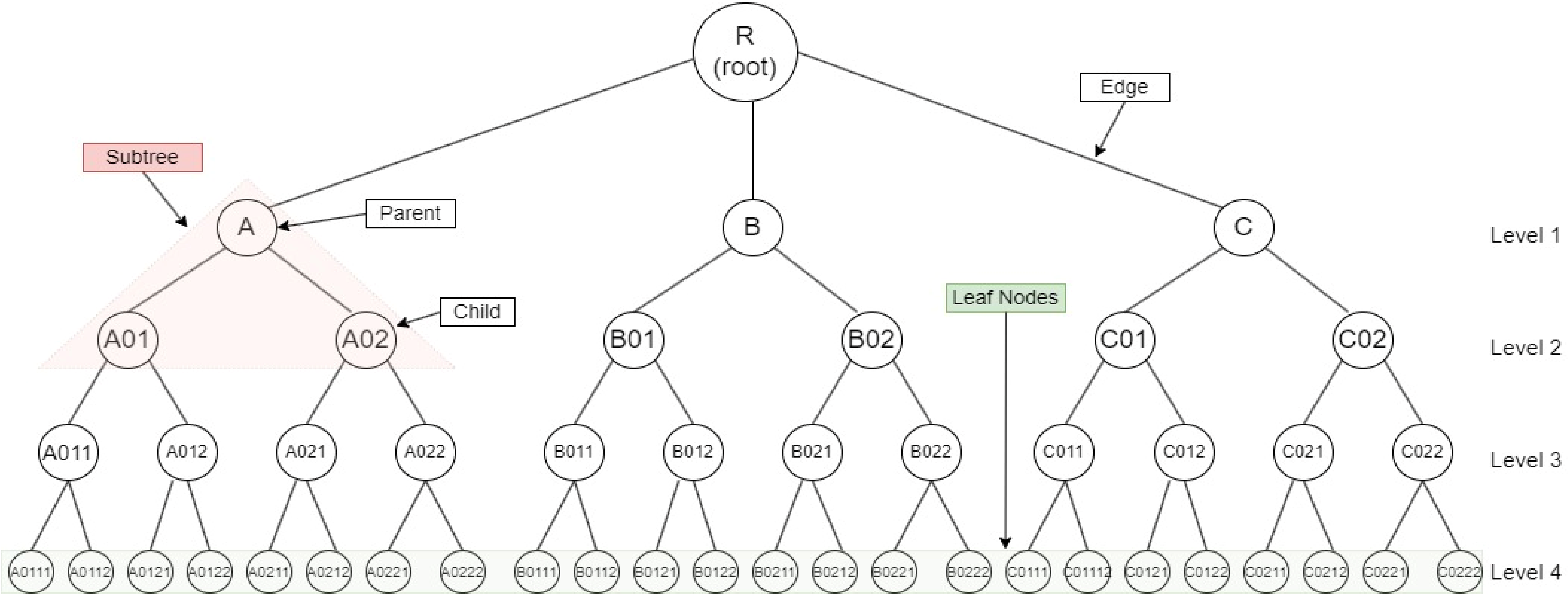
- Text on the landing page and sub-pages containing certain key-words in the link are scraped
- Only text elements are kept (Removal of digits and punctuations, Removal of characters not part of the German dictionary)
- Resulting text processed in the following way:
  1. Each word transformed using the "German morphological lexicon" (http://www.danielnaber.de/morphologie/)
  2. Stemming

Feature Selection

- After pre-processing scraped text contains > 3 Mio words
- Feature Selection Method: Combine a global (DFS) and a local feature selection (OR) score function to select a set of features for each class (Uysal 2016)
- Select
  - 500 words for classes on NACE level 1
  - 200 words for classes on NACE level 2
  - 100 words for classes on NACE level 3
  - 80 words for classes on NACE level 4
  - 50 words for classes on NACE level 5

  -> use one-hot-encoding method to obtain the feature vector $f_i \epsilon \mathbb{R}^{i*n_i}$ for each NACE level i $\epsilon \{1, \dots, 5\}$, where n_i is the number of all classes on NACE level i, for each enterprise (weighted by the term-frequency inverse document frequency)

Hierarchical Structure and Classification Approaches



1. Flat Classification Approach
2. Global Classification Approach
3. Local Classification Approach: i. local classifier per node, ii. local classifier per level, iii. local classifier per parent node

Hierarchical Classification Model

- ***Local classifier per parent node approach***: each parent node *only* trained to distinguish between its child nodes
  - **Advantage:** class consistency e.g. predictions like A, B01, A011, C0112 not possible
  - **Disadvantage**: error propagation

  Implemented using XGBoost (Chen et al. 2015)

- Local classifier only constructed for a parent node if:
  1. The parent node has at least 2 children
  2. There are at least 2 enterprises that can be assigned to each of the child nodes

  -> if one of the conditions violated, prediction *not available* beyond the regarded parent node

-> NACE codes might not be available beyond a certain level

Evaluation of Hierarchical Models

Hierarchical Performance Measures

1. Standard evaluation measures (precision, recall, accuracy) (*flat metrics*)
   - **Disadvantage**: does not account for relationship between true and predicted value
     e.g. : True NACE 3 code A021, Model 1: A028, Model 2: C093
     -> both models perform equally poorly according to the standard evaluation measures
2. Distance based adjustment of standard evaluation metrics (Sun and Lim 2001)
3. Semantics based adjustment of standard evaluation metrics (Sun and Lim 2001)
   -> overall precision and recall obtained by computing the Macro- or
   Micro-Average
4. Hierarchical variation of flat metrics (Kiritchenko et al. 2006)

Tailored Hierarchical Performance Measure for NACE Code Classifications

- Regarded evaluation measures give equal weights to every class
- **Proposal**: weight a class according to the *size* (number of employees) of the enterprises that are contained in that class

$$\rho_l = \frac{1}{\sum_{i=1}^{n_l} s_{li}} \sum_{i=1}^{n_l} s_{li}\, p_{li}$$

$p_{li} \in \{Pr, Re, Ac\}$ evaluation measure at level $l \in \{1, \dots, 5\}$ for the class $i \in \{1, \dots, n_l\}$,
s_{li} number of employees -> weighted evaluation measure available for each level l ->
Take (weighted) average for an overall evaluation value

Evaluation

Table 1: Performance at NACE level 1

NACE	PR	RE	AC	PR^{CS}	RE^{CS}	AC^{CS}	PR^{CD}	RE^{CD}	AC^{CD}
C	0.820	0.862	0.900	0.884	0.919	0.935	0.933	0.950	0.960
F	0.874	0.855	0.953	0.860	0.842	0.949	0.954	0.946	0.981
G	0.843	0.854	0.903	0.897	0.915	0.936	0.942	0.947	0.961
H	0.902	0.848	0.980	0.892	0.830	0.978	0.964	0.942	0.992
N	0.908	0.823	0.969	0.916	0.837	0.972	0.967	0.933	0.988

Table 2: Overall Performance

Metric	Measure	Micro-Average	Macro-Average
-	Pr	0.74	0.71
-	Re	0.76	0.63
CS	Pr	0.84	0.8
CS	Re	0.86	0.75
CD	Pr	0.82	0.71
CD	Re	0.85	0.67

Hierarchical Versions: hPR=0.74 hRE=0.59

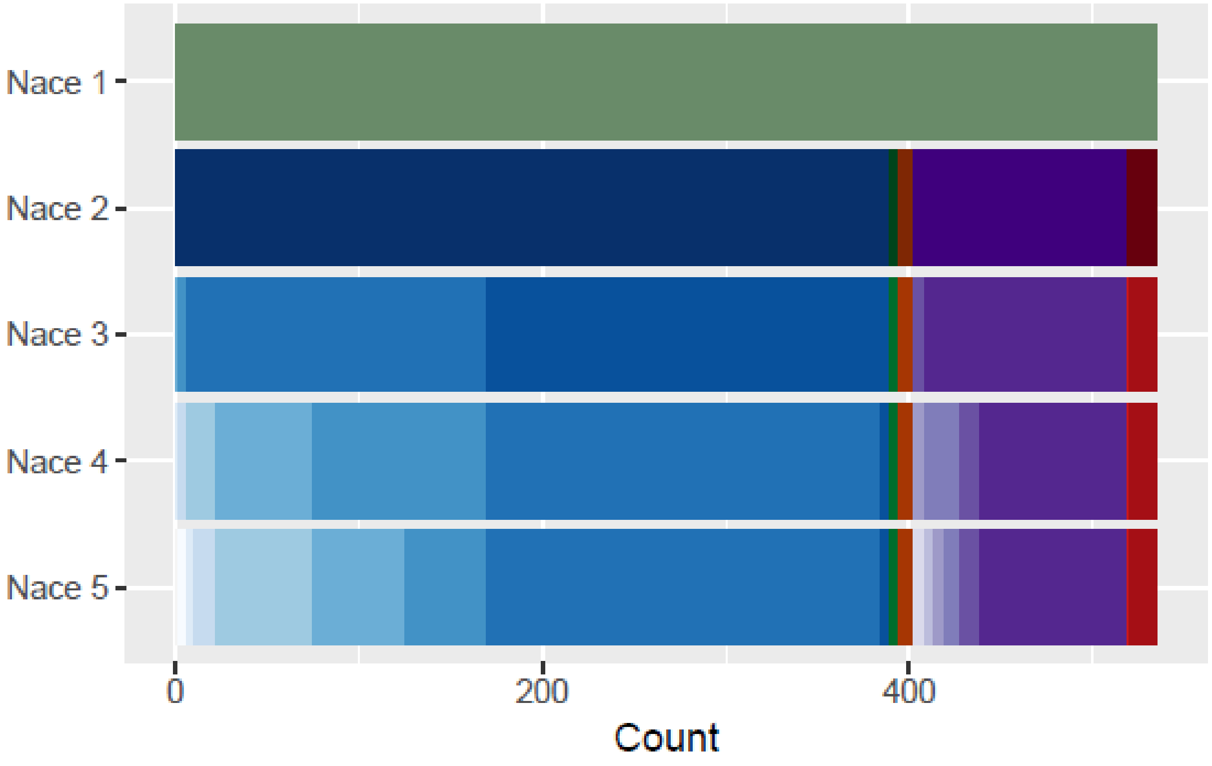
Evaluation

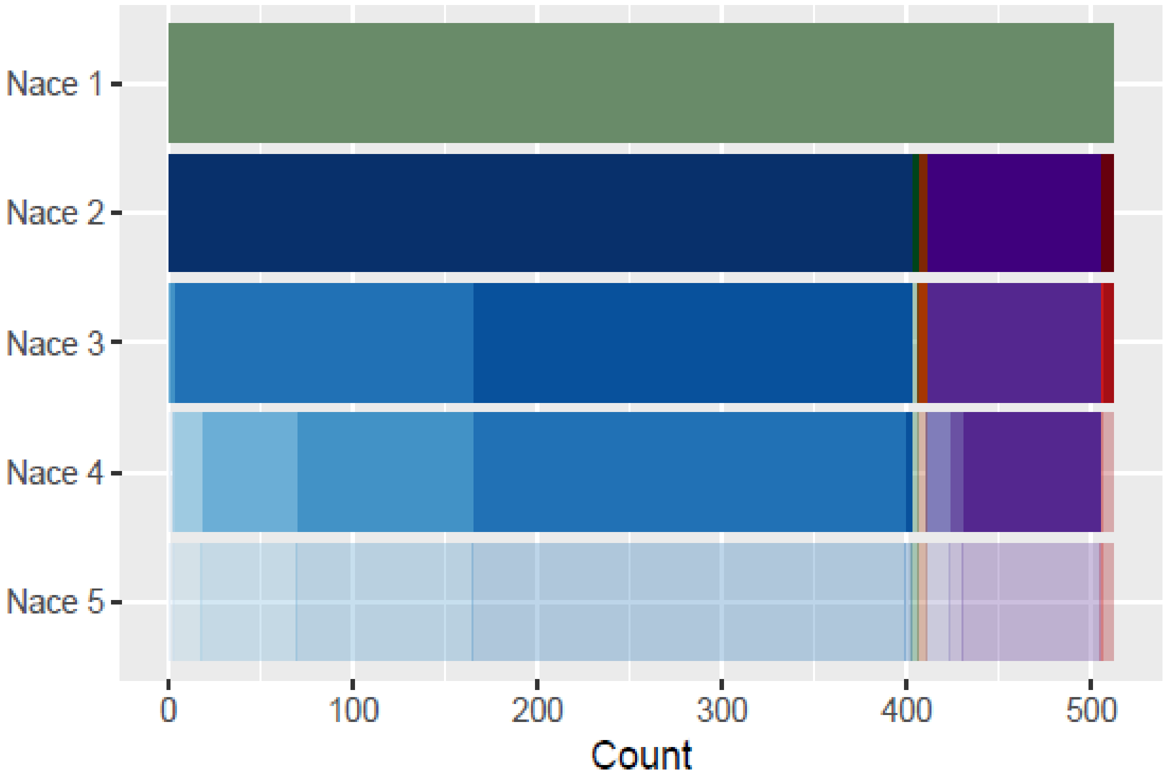
Table 3: Performance in terms of weighted precision, recall and accuracy at 1) each level separately (NACE 1- 5) and 2) over the whole category space (Weighted Average)

Metric	NACE 1	NACE 2	NACE 3	NACE 4	NACE 5	Weighted Average
PR	0.86	0.78	0.70	0.46	0.14	0.68
PR^{CS}	0.89	0.87	0.83	0.56	0.15	0.77
PR^{CD}	0.95	0.91	0.79	0.45	0.12	0.76
RE	0.85	0.74	0.68	0.47	0.16	0.67
RE^{CS}	0.88	0.83	0.80	0.56	0.17	0.75
RE^{CD}	0.94	0.90	0.77	0.46	0.16	0.75
AC	0.93	0.97	0.96	0.69	0.18	0.86
AC^{CS}	0.95	0.98	0.97	0.69	0.18	0.87
AC^{CD}	0.97	0.99	0.97	0.69	0.18	0.88

Visual Evaluation of class H

Distribution of the actual vs predicted NACE level 1-5 codes with NACE level 1 code ***H***





References

Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. 2015. *“Xgboost: extreme gradient boosting.”* R package version 0.4-2 1 (4): 1–4.

Kiritchenko, Svetlana, Stan Matwin, Richard Nock, and A Fazel Famili. 2006. *“Learning and evaluation in the presence of class hierarchies: Application to text categorization.”* In Advances in Artificial Intelligence: 19th Conference of the Canadian Society forComputational Studies of Intelligence, Canadian AI 2006, Qu´ebec City, Qu´ebec, Canada, June 7-9, 2006. Proceedings 19, 395–406. Springer.

Sun, Aixin, and Ee-Peng Lim. 2001. *“Hierarchical text classification and evaluation.”* In Proceedings 2001 IEEE International Conference on Data Mining, 521–528. IEEE.

Uysal, Alper Kursat. 2016. *“An improved global feature selection scheme for text classification.”* Expert systems with Applications 43:82–92.

Please address queries to

Manveer MANGAT

Manveer.Mangat@statistik.gv.at

STATISTIK AUSTRIA

Guglgasse 13, 1110 Wien

Independent statistics for evidence-based decision making

STATISTICS AUSTRIA

The Information Manager