Data Science Campus

UNECE HLG-MOS ML Project Classification & Coding Theme Report

Claus Sthamer

Claus.Sthamer@ONS.gov.uk

The C&C Theme

- C&C, along with Edit & Imputation and Imagery themes, is part Work Package 1

- C&C objective: This theme was selected from the GSBPM as one of the processes suitable for Machine Learning

Classification & Coding

In the terminology of machine learning, classification is considered an instance of supervised learning, i.e., learning where a training set of correctly identified observations is available.”

https://en.wikipedia.org/wiki/Statistical_classification

Challenge of manual C&C?

- Very labour intensive
- Repetitive
- Lengthy process – delay in statistical output
- Inconsistent – level of experience of coders might differ
- Deterministic/rule-based/word matching systems are difficult to build and maintain, rules/reference text entries have to be updated frequently

Pilot Studies – How can Machine Learning help?

1.	**BLS – USA**	Survey of Occupational Injuries and Illnesses	Workplace Injury – SOC, OIICS, 6 codes
2.	**Stats Canada**	Canadian Community Household Survey	Occupation & Industry – NAICS, NOC
3.	**Statistics Norway**	New Companies for the Central Coordination Register	Standard Industrial Code – SIC
4.	INEGI – Mexico	Household Income and Expenditure	Occupation & Economic activity - SCIAN, SINCO
5.	Statistiek Vlaaderen – Belgium	Sentiment of Twitter Data	Positive/Negative
6.	SORS – Serbia	Labour Force Survey	Economic Activity – NACE
7.	Statistics Poland	Web scraped food products	Food description - ECOICOP
8.	IMF	Catalogue of Time Series - CTS	

Pilot Studies Objectives

Quality – Efficiency – Timeliness - Accuracy

Pilot Studies - Insights

- Data Requirements: Golden Data Set – Ground Truth
- Algorithms:
  - SVM & XGBoost, Random Forest, FastText performed well
  - Neural Network best result
- IT hardware:
  - Normal desktop/laptop used by 5 pilot studies
  - Cloud computing used by one
  - 4 x Graphical Processing Units with 3584 cores used to run a neural network
- Quality Measures used
  - Accuracy
  - Recall
  - F1-score

Value added by ML for C&C

- Auto-coding can be achieved, but not for 100%
- ML/Human work together to get best results
  - Prediction threshold – auto-coding where ML is ‘sufficiently’ confident
  - Human coding of minority classes and low confident predictions, >95% Accuracy
- Faster processing than manual
- Increased data consistency

Challenges/Blockers

- 3 of 8 are in production

- The other 5:
  - High demand for analysts
  - Volume and quality of training data
  - Accuracy
  - Not enough resource to progress the pilot study
  - Cost of IT systems - prevent usage of Neural Networks

Data Science Campus

Thank you