Machine-Learning in Surveys Steps

ELIZABETH AYRES, BSMD
CATHERINE DESHAIES-MOREAULT, BSMD
FRANCE LABRECQUE, HSMD
MARIE-PIER LEMIEUX, SSMD
LAST UPDATED AUGUST 2017

Introduction

The purpose of this document is to identify machine learning (ML) techniques that have been applied at various survey steps by statistical agencies, or could potentially be applied. To illustrate the process, it was decided to link ML techniques to certain steps in the Generic Statistical Business Process Model (GSBPM). The goal is not to cover each and every step of the GSBPM, but rather to identify for which ones ML techniques could be an asset and potentially an improvement to the more traditional methodology approaches.

This document is a work in progress, and should be updated frequently as ML is used more and more widely throughout statistical agencies. Some of the ideas presented here will also covered Natural Language Processing (NLP). NLP is a key component of Artificial Intelligence (AI) where we can capture a "meaning" from an input of words (sentences, paragraphs, pages, etc.). The underlying algorithms used in NLP rely heavily on ML techniques.

What is machine learning?

According to the SAS Institute, ML is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, ML allows computers to find hidden insights without being explicitly programmed where to look. The iterative aspect of machine learning is important because as models are exposed to new data, they can independently adapt.

Machine learning algorithms can be divided into 2 major categories which have their own techniques— supervised learning (decision trees, Naïve Bayes classification, OLS regression, logistic regression support vector machines, etc.) and unsupervised learning (clustering algorithms, principal component analysis, singular value decomposition, etc.). Supervised learning analyzes the training data and produces an inferred function (Mohri 2012), which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. Unsupervised learning is the machine learning task of inferring a function to describe hidden structure from "unlabeled" data. Other categories exist and may be called differently among scientific: reinforcement learning, clustering, dimensionality reduction, etc.

Table of Contents

Generic Statistical Business Process Model (GSBPM) Version 5.0

1 **Specify needs**	**2** **Design**	**3** **Build**	**4** **Collect**	**5** **Process**	**6** **Analyse**	**7** **Disseminate**	**8** **Evaluate**
1.1 Identify needs	**2.1** Design outputs	**3.1** Build collection instrument	**4.1** Create frame and select sample	**5.1** Integrate data	**6.1** Prepare draft outputs	**7.1** Update output systems	**8.1** Gather evaluation inputs
1.2 Consult and confirm needs	**2.2** Design variable descriptions	**3.2** Build or enhance process components	**4.2** Set up collection	**5.2** Classify and code	**6.2** Validate outputs	**7.2** Produce dissemination products	**8.2** Conduct evaluation
1.3 Establish output objectives	**2.3** Design collection	**3.3** Build or enhance dissemination components	**4.3** Run collection	**5.3** Review and validate	**6.3** Interpret and explain outputs	**7.3** Manage release of dissemination products	**8.3** Agree on action plans
1.4 Identify concepts	**2.4** Design frame and sample	**3.4** Configure workflows	**4.4** Finalise collection	**5.4** Edit and impute	**6.4** Apply disclosure control	**7.4** Promote dissemination products	
1.5 Check data availability	**2.5** Design processing and analysis	**3.5** Test production system		**5.5** Derive new variables and units	**6.5** Finalise outputs	**7.5** Manage user support	
1.6 Prepare business case	**2.6** Design production systems and workflow	**3.6** Test statistical business process		**5.6** Calculate weights			
		3.7 Finalise production systems		**5.7** Calculate aggregates			
				5.8 Finalise data files			

Source: http://www23.statcan.gc.ca/standards-normes/gsbpm-msgpo/2013/index-c-eng.html

Description of ML techniques

GSBPM linked to ML techniques

1 Specify needs

This phase in the GSBPM framework outlines procedures that are largely decision based, and carried out without computational aid. As such, ML techniques don't seem to be as relevant at this point in time, except maybe in a few selected cases. The phase is triggered when a need for new statistics is identified, or feedback about current statistics initiates a review. In cases where the stakeholders are numerous or the feedback plentiful, NLP approaches could potentially be of use.

1.1 Identify needs

- To identify the needs, information is gathered from various sources, including action plans from evaluations of previous iterations of the process, from other processes, or from practices in place in other statistical organisations. Should the information gathered be voluminous, NLP could be of used to extract key concepts or commonalities.

1.2 Consult and confirm needs

- If there are multiple and conflicting needs from a vast array of stakeholders, NLP *could* be used to extract key concepts or commonalities. However, because understanding the needs is a critical aspect of the process, it is believed AI approach should only be used here as complement (if at all) and not as a replacement approach.

2 Design

2.3 Design collection

- ML techniques can help to analyze comments done by respondents in previous cycles and help us to better design questionnaires.

2.4 Design frame and sample

- Automatic coding can be used to standardize variables in survey frames (ie. current studies in Germany (Arne Bethmann, 2014) look at coding occupation; a study in New Zealand is looking into coding variables from their census)
- ML classification methods (ie. K nearest neighbors, random forests, decision trees, cluster analysis) can be used to help identify frame defects (misclassification, duplication of records, errors in information such as addresses), where errors are pooled in a separate cluster for manual review
- ML techniques could be useful in the creation of more homogenous strata based on a more complex combination of variables, and/or their orthogonal projections (ie. random forest, dimensionality reduction algorithms)

2.5 Design process and analysis

- ML techniques could be used on the paradata from previous occurrences to identify area for improvement to be considered in the design. Other applications of these techniques will be described in the specific sub-steps of the Process phase.

3 Build

This stage consists mostly of implementing what was decided in stage 2. As such, the ideas described under the *Design* stage are implemented here.

3.5 Test production system

- AI or ML led quality assurance process could be used to test the production system. ML techniques could be used in test suite optimisation (id duplicate/similar and unique test cases), for log analytics or tracability (identify scenarios to achieve test coverage), etc.

3.6 Test statistical business process

- In case of pilot testing, ML techniques could be used on paradata to detect anomalies.

4 Collect

4.1 Create frame

- Techniques used to design the frame could be used (see 2.4). Machine learning in the context of record linkage (see also 5.1) could be used to bring various information together to create a frame. In a quality assurance process, ML could be used on the frame to detect anomalies.

4.2 Set up collection

- Classification tree: Predict non-respondents during collection by using paradata

4.3 Run collection

- Bayesian Hierarchical Modeling: Parameters are given at first to the model and by using Monte Carlo sampling (MCMC) the model can update himself by giving him daily data. It could be used to prioritize cases to be called during NRFU and reduce de non-response risk during data collection.

5 Process

5.1 Integrate data

- As part of record linkage, ML algorithms could be used. Feature extraction techniques could be used in record linkage preprocessing steps. Unsupervised method such as K-means clustering and supervised methods (probit model, support vector machine and decision tree) are currently under investigation to help minimizing manual effort in probabilistic linkage and to streamline preprocessing steps. Record linkage is an area where ML techniques could be very useful.

5.2 Classify and code

- Powerful supervised ML algorithms for automatic classification (Bethmann 2014), such as neural networks, random forests, SVM, could be used to diminish manual workload and potential human mistakes. Classification and automatic coding are steps that are a natural fit for using ML techniques and likely the area where the gain could be the most important.

5.3 Review and validate

- Outlier detection can be done using unsupervised ML techniques (Breton 2016), such as clustering with nearest neighbors or isolation forests.

5.4 Edit and impute

- ML classification techniques (cluster analysis, random forests, regression trees, projection pursuit analysis, etc.) can be used not only to identify outliers, but to create imputation donor pools as well
- EuroStat has been investigating the possibility of applying ML techniques to editing and imputation since January 2000 (EuroStat, 2007). Explored various techniques (neural networks such as MLP, CMM, and SOM, as well as SVM), and created a framework for the evaluation and validation of these methods. A prototype software was created.
- A further investigation by EuroStat looked at how to use ML techniques (neural network classifier and a Bayesian network classifier) to impute missing categorical data resulting from non-response (Rey del Castillo, 2012).
- Select sets of matching variables preceding nearest neighbor imputation (Statistics New Zealand)

5.6 Calculate weights

- Using ML techniques to create homogeneous group for non-response adjustment.

5.7 Calculate aggregates

- Aggregates needed that were not planed as estimation domains at first can be created by reclassifying the data using ML techniques.

6 Analyse

6.1 Prepare draft outputs

- ML classification techniques applied on time series data could help streamline the preparation of seasonal adjustment options.

6.2 Validate outputs

- ML techniques used in the context of fraud detection could be used to detect outlying patterns in output where extra validation may be needed.

6.3 Interpret and explain outputs

- Using paradata from electronic questionnaires (EQ) in neural networks. It can help us better understand respondents' behavior.

ML classification techniques could be used to identify pattern and gain analytical insights worth sharing as part of dissemination process. 6.4 Apply disclosure control

- ML classification techniques can be used to gauge data utility based on the classification error (Mivule, 2013).

7 Disseminate

7.4 Promote dissemination products

- ML techniques used in the context of marketing personalisation and automatic recommendation could be used to enhance product promotion initiatives.

7.5 Manage user support

- Supervised methods can be used to classify users' requests and send automated answers: more effective management of user support.

8 Evaluate

8.1 Gather evaluation inputs

- Using Naïve Bayes and Support Vector Machine classification algorithm to do sentiment analysis on comments received from evaluation (Tripathy, 2015 & Medhat 2014).

References

Bethmann, Schierholz, Wenzig, and Zielonka (2014). Automatic Coding of Occupations Using Machine Learning Algorithms for Occupation Coding in Several German Panel Surveys (draft).

Breton and al. (2016). Research indices using web scraped data: May 2016 update.

Medhat, Hassan and Korashy (2014), *Sentiment analysis algorithms and applications: A survey*, Ain Shams Engineering Journal

Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012) *Foundations of Machine Learning*, The MIT Press

Mivule and Turner (2013), *A Comparative Analysis of Data Privacy and Utility Parameter Adjustment, Using Machine Learning Classification as a Gauge*, Procedia Computer Science

Tripathy, Agrawal and Rath (2015). *Classification of Sentimental Reviews Using Machine Learning Techniques*, Procedia Computer Science

Training:

https://www.cbs.nl/-/media/_pdf/2016/41/big-data-masterclass-and-datacamp-2015.pdf

General challenges:

https://www.cbs.nl/-/media/_pdf/2016/28/big-data-and-methodological-challenges-in-official-statsitics.pdf

Journal:

Journal of machine learning research (free and each processing is specific to a conference related to ML): http://www.jmlr.org

Machine Learning: http://www.springer.com/computer/ai/journal/10994

Website:

Outline of machine learning on wiki: https://en.wikipedia.org/wiki/Outline_of_machine_learning

Microsoft have an incredible website that describe their machine learning techniques used by Azure: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-what-is-machine-learning

Look also at their cheat-sheet:

microsoft-machine-learning-algorithm-c