The Survey Statistician, 2021, Vol. 84, 12–17.

Ask the Experts

Machine Learning from the Perspective of Official Statistic

Marco J. H. Puts[1] **and Piet J. H. Daas**[2]

[1]Statistics Netherlands, The Netherlands, m.puts@cbs.nl
[2]Statistics Netherlands, The Netherlands, pjh.daas@cbs.nl

Abstract

Artificial Intelligence methods enable the extraction of information from large amounts of data. Machine learning is a sub field of Artificial Intelligence. In this paper, Artificial Intelligence and Machine Learning are introduced and discussed in the context of applying them to produce official statistics. The five quality dimensions of statistical output are used to identify the challenges for their application. The paper ends with a list of the most important research topics that need to be studied to enable the successful application of those methods for official statistics.

Keywords: Artificial Intelligence, Big data, Data science, Output quality.

1 Introduction

Before focusing on Machine Learning (ML), we will start by introducing the field of Artificial Intelligence (AI). AI is still a young field of science, with its official birth in the 1950s, of which ML is a sub field (Russel and Norvig, 2003). In the 1950s the main focus of AI research was on intelligence, especially mimicking human intelligence, and self-learning. The most important topics studied during that period make that clear. These topics were, rephrased in a modern context,: Simulating the human brain, Natural Language Processing, Artificial Neural Networks, Theory of complexity of functions, Self-improving algorithms, Abstracting from sensory data and Randomness and creativity algorithms (McCarthy et al., 1955). The research in the area of self-improving algorithms laid the foundation for ML.

In the beginning of AI research, the founding fathers were really optimistic. They expected to reach human intelligence within a couple of years (Mitchell, 2019). However, this goal has still not been achieved today. Clearly, human intelligence is much more complex than originally expected. Nevertheless, algorithms that evolved from the field of AI are currently being used in a whole range of areas for a multitude of applications and, although many scholars would argue that the field of AI as such has failed (see, for instance, Sowa, 2014), these algorithms are very successful (by)products (see, Royal Society, 2017, and Sejnowski, 2020).

Essentially, ML algorithms work by discovering structure in data, i.e. ’learn from data’. This work was inspired by statistical methods and elaborated upon by computer scientists (Russel and Norvig, 2003). ML algorithms can learn in a number of ways. The most familiar ones are i) by training them with labelled examples, so-called supervised learning, or ii) by training them with unlabeled examples, known as unsupervised learning. In the latter case, structures are discovered by comparing the items to one-another in the data set. Advances in ML were for a long time predominantly of a technical nature; which is not unexpected considering the predominant involvement of computer scientists. Over time, ML algorithms were able to discover more and more complex relations in data. This work lay the foundation of many of the well-known successful applications of ML such as image and speech recognition, spam detection and fraud detection (Sarker, 2021). In all these cases, ML algorithms are trained on large numbers of examples that enables them, by identifying specific structures - also known as features - in the data, to successfully perform specific tasks, such as predicting classes. We are aware that these tasks also sometimes fail which have resulted in some remarkable errors (Mitchell, 2019). But in this paper we will focus on the successes.

From the above it is clear that ML algorithms are able to successfully extract information from a whole range of data sources, including those used by statisticians. However, applying ML learning algorithms to produce official statistics is still challenging (Yung et al., 2014). The most important reasons and issues are discussed in Section 2 of this paper. We will do this by looking at the quality of the output of ML algorithms because we found that this identifies the most important issues.

2 Quality of Statistics

The high quality standards of official statistics are an important reason that affect the application of ML in this area. Let it be clear that we agree that high quality standards are essential and need to be adhered, but for ML it raises a number of challenges. In this section we will identify those hurdles. To enable this we use the five dimensions defined for the output quality of official statistics in the Quality Assurance Framework of the European Statistical System (ESS, 2019). The dimensions are:

- Relevance
- Timeliness and Punctuality
- Accessibility and Clarity
- Coherence and Comparability
- Accuracy and Reliability

The dimensions Relevance and Timeliness and Punctuality are not discussed here as they do not differ when ML algorithms are applied in comparison to any other method used in producing statistics. Relevance is about the usefulness and value of the statistics for the user which is obviously not different in the ML case. Timeliness and Punctuality focus on the point in time when the results have to be published. This is also no different for ML. The real challenges are in the other dimensions of quality.

2.1 Accessibility and Clarity

This dimension indicates the need to define and fully understand the process by which results are obtained. Obviously, this is a problem for some ML algorithms and it touches the topic of explainable AI (Gunning et al., 2019). Making clear how the results are exactly obtained is challenging for many ML algorithms, in particular for Deep learning and other neural network based methods, since some of them are essentially a 'black box'. In cognitive science and neuroscience researchers work on a very similar problem. They are trying to understand the functionality of the human brain: a vast neuronal network which can also be seen as a (giant) black box. In the 1980s, David Marr (1982) describes how such a non-transparent device should be studied. According to him, to fully understand the process by which the brain processes data to come to a conclusion, it should be investigated at three levels. These are: i) Computational theory, ii) Representation and algorithm, and iii) (hardware) Implementation. His suggestions provide valuable insights regarding the accessibility and clarity of ML algorithms.

Computational theory is concerned with the goal, the appropriateness and the logic of the process used. ML models learn or need to learn some aspects of the physical world. For instance, when we want to determine if a person on social media has, based on the messages he/she produces, COVID-19, we need to realize that this person creates its utterances in a natural language on a platform. It is the combination of the structure of the language, the meaning of the words, the limitations of the social media platform (e.g. the maximum number of words allowed in a message) and the intention of the person for sharing the message that specifies the problem the algorithm has to solve. Understanding the decisions made by a ML algorithm requires a complete comprehension of the above-mentioned properties in the 'learning from data', including its context, process.

Representation and algorithms concern the input, the output and the ML algorithm used to perform the task. What is important here is the design pattern (among others the learning method used, transformations of data,...) used in creating the model. It focuses on the entire process from the input, metadata, architecture and output perspective. The whole processing pipeline needs to be reviewed and fully understood. The choice of the algorithm is important. On the one end, choosing a more transparent and simpler algorithm may result in a better explainable model, but maybe the problem at hand is so complex that a more complex approach, like a (deep) neural network, is actually required to achieve a better result.

Implementation concerns the physical implementation of the ML algorithm. At this level, all things mentioned above are combined and implemented. Of course, the implementation consists of a program, but this program is limited by the hardware architecture and software programming language(s) used. Regarding the hardware, ML algorithms can be implemented on single core processors, multi core processors, Graphics Processing Units, Field Programmable Arrays and maybe, in the (near) future, on quantum computers. With this, we also need to make implementation decisions, and these are, of course, crucial for our understanding of the process being developed. Without understanding there is no transparency. Regarding the software, ML algorithms are often implemented in C/C++ or Java to make them run as fast as possible. Other programming languages, such as R and Python, use these implementations by deploying them. As mentioned before, the choice of algorithm is important. However, the choice for a less explainable model makes it more challenging to determine what the model has actually learned. Models can be studied in a mechanistic way (open box), by investigating what happens inside the model, or in a more functional way (black box). We suggest that it is always good to use a functional approach, maybe extended with a mechanistic approach, and determine what happens when the model is 'fed' small chunks of data.

2.2 Coherence and Comparability

In the context of ML this dimension is about how well the model is able to give a stable result over time and the correlations it has found. We will first talk about stability over time of the model, followed by a short discussion on correlation and causation.

Concept Drift. Stability of the way by which a concept is measured is very important for any model-based statistics. Because of this stationary nature, a model is able to detect changes in the number of occurrences of the concept in the population. However, when the detection of the concept is affected - because of changes in the way the concept is expressed by the population - the outcomes of the model are also affected. When this is happening, the measurements can no longer be trusted. Suppose we measure the sentiment on social media by looking at the ratio of a limited number of words with a positive and negative connotation. The measurement of this concept can be affected by a change in the use of (one or more of) the words in a totally different context. This leads to a phenomenon called concept drift, i.e. the concept originally measured can no longer be measured in exactly the same way (Gama et al., 2014). It is important to be able to discern both type of changes: i) those resulting from an actual change in behaviour of the population included and those due to concept drift. The latter should, where possible, be corrected for. Increasing the amount of data to train the model and including new, more recently, classified cases certainly help (Daas and Van der Doef, 2020).

Correlation and causation are very important topics when applying ML. In principle ML algorithms pick up the relation between the values of variables in the data set and those of the target variable. Because of this most algorithms simply look for variables that correlate with the target variable. If this is a spurious or causal relation it is not considered. This shows a very big problem in ML (and statistics), the difference between correlation and causation. The fact that we find a correlation between two variables does not mean that there is a causal link between them and, hence, we can use one as a proxy for the other. Proving a causal relationship between two variables is quite hard. There are, however, some ways to get a clue over the causality between two variables (Pearl, 2009). Downside of some of these approaches are that they require to randomize, experiment or intervene in the data being used. When the data is given, which is for instance the case in many big data sources, such approaches cannot be used. However, natural experiments also occur and may provide new insights, for instance by comparing the situation before and during COVID. More on this topic can be found in Pearl (2009) and Schöllkopf et al. (2021).

2.3 Accuracy and Reliability

This dimension is about how far the finding is from the true value in the population. Bias and variance are the major components that affect accuracy. For official statistics it is important to have an unbiased estimator. If large amounts of data are being used, which is often the case when ML algorithms are applied, variance is usually not the major concern. Bias can be introduced in each step of the statistical process. When ML algorithms are used the most obvious causes of bias are the annotated data set used for training (and testing), the representativity of this data set and misclassification bias introduced by the model developed.

Annotated data and representativity. An important aspect in creating a ML-based model is the data used to train (and test) the model and how it was obtained. Many decisions are made in this process that may introduce a bias. Basically, designing an annotated data set is comparable with designing a sample survey; errors similar to those affecting the Total Survey Error (Groves and Lyberg, 2010) apply to the training data used. The inter-comparability and perceptual and perception bias of human annotators are also important contributors [1]. Also, the better the structures (features) in the training data represents those of the target population, the better the model is expected to perform. Within AI, this is known as the *closed world assumption* (Russel and Norvig, 2003): the model works the best in exactly the same context as it was trained. Any deviations of this context can lead to biases. However, since the data selection process will obviously affect the features included, this is far from a trivial task. It may even be an advantage to over represent particular examples (i.e. features or classes) in the training data. This is similar to taking a stratified sample in survey design with the aim to over represent particular groups of units. This is especially important when studying rare cases (i.e. small areas).

Misclassification will introduce a bias when the ratio between the false positives (type I error) and false negatives (type II error) of the trained model deviates from the actual data. This can be corrected

[1] Perceptual bias refers to biases introduced by the perceptual system (e.g. visual illusions) and perception bias refers to the social context, where the annotator could be biased.

for by posing a constraint on it (Meertens, 2021). In addition, binary classifiers trained on a certain proportion of positive items can also introduce a bias when the model is applied to (real-world) data with a different proportion of positive items. Since the latter proportion is generally unknown, it often is the target variable, a maximum likelihood estimator has been developed by which the true proportion of positive items in data sets can accurately be determined (Puts and Daas, 2021).

3 Conclusions

From the above its clear that ML is an interesting application for official statistics. An overview study conducted in 2018 at NSI's worldwide, lists a large number of potential applications of ML algorithms of which the majority are ideas (Beck, Dumpert and Feuerhake, 2018). In a more recent report, the most mentioned applications of ML for survey data are stratification and outlier detection (Yung et al., 2014). In the area of administrative and big data, the better scalability, the less sensitivity to outliers and erroneous data and the ability to capture non-linear relationships are mentioned as the major advantage of applying ML compared to traditional statistical methods (Yung et al., 2014). Because ML algorithms are particularly well suited to extract information from texts and images, it is clear that in these areas they should certainly be applied within the realm of official statistics. Examples of experimental statistics making use of these kind of sources can be found in (Daas et al., 2020). However, to fully enable the use of ML algorithm in official statistics a number of challenges need to be solved. According to us, the following topics need to be studied within the realm of official statistics:

- Methodology concerning the human annotation of data
- Sampling the population to obtain representative training sets
- Using stratification in the context of Machine Learning
- Data structure engineering and selection to increase the transparency of models
- Reducing spurious correlations
- Methodology for studying causation
- Correcting the bias caused by the ML model
- Dealing with concept drift (representativity over time)

For a number of those topics, expertise is available at methodology and statistics departments at Universities and National Statistical Institutes. For some, however, cooperation between ML experts and statisticians is required. These could, for instance, be studied in a joint European or Global research project.

References

Beck, M., Dumpert, F. and Feuerhake, J. (2018) *Machine Learning in Official Statistics*. `https://arxiv.org/abs/1812.10422`.

Daas, P., Maslankowski, J., Salgado, D., Quaresma, S., Tuoto, T., Di Consiglio, L., Brancato, G., Righi, P., Six, M., Weinauer, M. and Kowarik, A. (2020) *Revised Version of the Methodological report. Deliverable K9.* ESSnet Big Data II.

Daas, P. and Van der Doef, S. (2020) Detecting Innovative Companies via their Website. *Statistical Journal of IAOS*, **36**, 1239-1251.

ESS (2019) *Quality Assurance Framework of the European Statistical System, version 2.0*. ESS Quality Assurance Framework.

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M. and A. Bouchachia (2014) A Survey on Concept Drift Adaptation. *ACM Computing Survey*, **46**, 1-37.

Groves, R. and Lyberg, L. (2010), Total Survey Error: Past, Present, and Future. *The Public Opinion Quarterly*, **74**, 849-879.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. and Yang, G-Z. (2019) XAI-Explainable Artificial Intelligence. *Science Robotics*, **4**.

Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT press.

McCarthy, J., Minsky, M., Rochester, N. and Shannon, C. (1955) *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. `http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf`.

Meerten Q. (2021) *Misclassification Bias in Statistical Learning*. PhD Thesis University of Amsterdam.

Mitchell, M. (2019) *Artificial Intelligence, A Guide for Thinking Humans*. Pelican Books, UK.

Pearl, J. (2009), Causal Inference in Statistics: An Overview. *Statistics Surveys*, **3**, 96-146.

Puts, M. and Daas, P. (2021) Unbiased Estimations Based on Binary Classifiers: A Maximum Likelihood Approach. *Paper for the Symposium on Data Science and Statistics*.

Royal Society (2017) *Machine learning: The Power and Promise of Computers that Learn by Example*. The Royal Society, London.

Russel, S. and Norvig, P. (2020) *Artificial Intelligence - A Modern Approach, 4th edition*. Pearson, Boston.

Sarker, I.H. (2021), Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, **2**, 1-21.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N., Kalchbrenner, N., Goyal, A. and Bengio, Y. (2021), Toward Causal Representation Learning. In: *Proceedings of the IEEE*, **109**, 612-634.

Sejnowski, T.J. (2020) The Unreasonable Effectiveness of Deep Learning in Artificial Intelligence. *Proceedings of the National Academy of Science* **48** 30033–30038.

Sowa, J. (2014) Why Has Artificial Intelligence Failed? And How Can It Succeed? *Computación y Sistemas*, **18**, 433-437.

Yung, W., Karkimaa, J., Scannapieco, M., Barcaroli, G., Zardetto, D., Ruiz Sanches, J.A., Braaksma, B., Buelens, B. and Burger, J. (2014) *The Use of Machine Learning in Official Statistics*. UNECE.