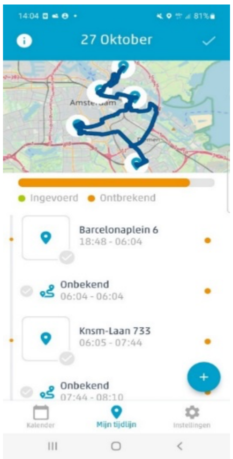
Choosing human annotators from a target population to train machine learning models

Challenge owners: Jonas Klingwort, Chris Lam, Marco Puts and Barry Schouten
(main contact, jg.schouten@cbs.nl)
Institute: Statistics Netherlands/Centraal Bureau voor de Statistiek
Keywords: Statistics, Probability, Machine Learning, Artificial Intelligence, Complex Data

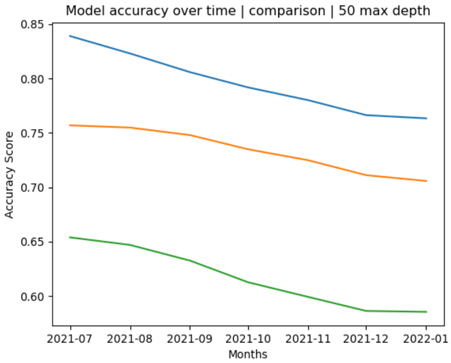


Barcelonaplein 6
18:48 - 06:04

Onbekend
06:04 - 06:04

Knsm-Laan 733
06:05 - 07:44

Onbekend
07:44 - 08:10



About CBS

CBS' statutory task is to compile statistics on a wide range of topics that are important to society and to make the outcomes publicly available. To this end, CBS gathers data from individuals and enterprises. The data gathered is then processed as statistics. CBS uses several methods to gather this data. It carries out surveys among individuals and enterprises. This is mainly done digitally, but may also be done in writing or in a face-to-face interview. In recent decades, CBS has increasingly made use of existing registers such as the Personal Records Database or the files held by the Dutch Chamber of Commerce. In addition to government records, CBS has also started using business records – data from supermarket checkout scans, for example – to calculate price developments. The main advantage of using registers is that CBS no longer needs to contact individuals and enterprises as often, which makes the survey process less intensive and time-consuming for everyone involved.

BACKGROUND

AI and ML applications are now widespread. Technology companies develop models using human annotated data. CBS sees the value in these models for complex survey and sensor data. Unlike companies, CBS emphasizes transparency and reproducibility, using international statistical classifications like HETUS (Harmonized European Time Use Survey) and COICOP(Classification of Individual Consumption by Purpose), which report on time use of individuals and household consumption and expenditures, respectively. For these applications, no pre-trained AI and ML models are available and CBS have begun training their own models, which involves working with a large number of human annotators.

In such settings, 'annotator burden' quickly may become prohibitive. Therefore, efficient sampling and invitation strategies for annotators (and households) are crucial. In other words, minimizing the number of participants and the time each spends annotating is essential.

There are four key complications impacting the quality of annotations:

1) **Feature selection.** The set of covariates with good predictive qualities must be chosen. In AI-ML terms these are usually refered to 'important features'. While there is often a large set of potential predictors, their importance is unknown before training. For example, in time-use studies one may extract location-related predictors from open-source points-of-interest (POI) databases such as OpenStreetMaps. The number of such POI's grows rapidly with a growing search radius around a location.
2) Label mismatch. The annotators can also use a priori information about observed data. For example, a household may know that it only buys certain products in a specific shop without looking at the shopping receipt.
3) **Clustering of features.** Several features may be correlated with a specific annotator, while we only have a freedom to allocate a whole annotator and not a feature. For example, all individuals living in a household likely go to the same shops and buy the similar type of products.
4) **Shift of concepts.** Both the features and behaviour change over time and models must be updated. For example, certain products or services may gradually vanish from the market whereas others emerge. But also, shops may alter the way they describe products or services, e.g. by moving towards digital receipts rather than paper receipts.

Finally, and most importantly, the choice of distance measures is entirely open. AI-ML approaches aim to maximize the separation of clusters within the same classification category to minimize miscla ssification errors. The optimal measure for achieving this is typically unknown at the start.

Hence, apart from efficient sampling there is also the need for effective sampling. Effective sampling then means that: 1) all potential features are exploited, 2) that the feature space is fully covered in the sampling, and 3) that changes in features are also included. The complications, thus, require a conceptual framework. In which both preferences for sampling annotators as numbers of annotators can be motivated.

CHALLENGE DESCRIPTION

The proposed challenge is to develop and demonstrate a conceptual framework that addresses the aforementioned issues within the constraints of cost and burden. The expected end result consists of two parts:

1. A framework addressing both aspects: how annotators are chosen and how many annotators should be invited.
2. A demonstration of the framework through case studies, using data provided during the challenge. The demonstration may be based on simulations.

Below, the two dirrections are elaborated.

Sampling framework

AI-ML train data for now are divided into two types. The first is the provision of break/intervention points that demarcate sections of data with the same label (e.g. the segmentation of travel into stops and trips). The second is the provision of labels that form the basis to classification of each section (e.g. the transport modes of a trip or the purpose of a stop). The steps are called identification and classification.

Next to the two types of train data, there are three tasks for annotators. The first is feature selection, the second is training and the third is updating or retraining. Under the first task, there are only general ideas about relevant (and available) features. Human annotators are invited for this purpose, which is known as the "human-in-the-loop" paradigm. Annotators may be asked about how they make decisions, *i.e.* on what basis they demarcate data sections and on what basis did they annotate these sections. This is particularly relevant to discern features that are being employed by annotators that are not available for a ML model. These may point at unobserved information that leads to noise in train data. Such noise will limit the maximal performance of models, which will be important in monitoring convergence of performance. Under the second task, there is not yet a trained model or optimized set of decision rules, but the features have been selected. These features are, in general, not (necessarily) the same as those used by annotators. Under the third task, annotators evaluate predictions of a pre-trained model or set of decision rules and adjust them if needed. The three tasks differ greatly in how annotators become involved. For simplicity, it is assumed in the challenge that each task is equally costly/time-consuming.

As is true for any form of sampling, setting up an efficient design requires basic prior knowledge of associations. In the case of annotators and features, it demands for a baseline selection of relevant features and relevant person/household characteristics. So, four stages in creating a sampling strategy may be distinguished: a feature selection stage, a preliminary stage where no prior knowledge is available about accuracy/performance, a training stage where such baseline knowledge is available, and an updating stage where a trained model or optimized set of rules is modified.

In the challenge, efficiency and efficacy of a sampling strategy need to be defined in terms of explicit measures. The most obvious measures are all kinds of measures based on false positives and false negatives, i.e. the classification performance of AI-ML predictions.

The feature space itself is not yet defined at the start of the training-updating process. To be more precise, the choice of distance measure and the mapping to the classification categories are open.
In the challenge, all stages and types as well as all open choices need to be considered.

Case studies

The demonstration of the framework can be done at the hand of two realistic case studies:

1. Stop-track segmentation including travel mode and travel purpose prediction: This case study is relevant for smart travel surveys employing location tracking data. Location data need to be split into series of stops and tracks. One or more purposes need to be linked to a stop. One or more modes need to be assigned to a track. In this case study the focus is on the prediction of stop purpose. Annotated data are available from a large-scale field study.
2. Text extraction on receipts: This case study is relevant for household budget surveys where participants may scan tickets or upload e-tickets of purchases within a specified time frame. Products on the tickets need to be classified to pre-specified categories. This case study focusses on the classification of products to specified categories. Annotated data are available from a large-scale field study.

ORGANIZATION

The problem posers have a statistics and/or data science background and have worked extensively with the two case studies. They are available during the challenge for consultation. A Statistics Netherlands slack channel will be opened for the team to facilitate communication. A git repository will be created to open access to the case study data.

The challenge team may divide itself across the three different sampling stages (feature, training, updating) and/or across framework and demonstration. The results and recommendations of the challenge team likely are of high relevance to Statistics Netherlands. The team is invited to present their results after the challenge also to stakeholders.