Description of the problems submitted to the 13th Montreal Industrial Problem Solving Workshop

1. Unit Load Device forecasting (Air Canada)

At Air Canada, we use Unit Load Devices (ULDs) to move both freight and baggage on our flights. As the schedule changes, we need to determine the number of ULDs we need for each aircraft and destination (based on our flight schedule) in order to predict empty ULD movements. Imbalances between stations happen; therefore, empty ULDs need to be moved. Not having enough ULDs results in baggage and freight being unable to move. Having too many ULDs results in an overpayment of rental fees.

The objective is to build an algorithm to identify the number of ULDs required per flight and destination, based on the airline schedule. The data available will include: (a) the flight schedule, (b) the aircraft types for the flights, (c) the historical number of ULDs by flight, (d) the ULD types that fit on each aircraft, and (e) the historical mix of freight versus baggage. The goal is to forecast the ULD quantity per type and flight for the next few months and also to predict the mix between freight and baggage.

2. Interactions between the events in the life of an insured client (Beneva)

Beneva wishes to analyze the time during which a client is on disability leave and receives insurance benefits before going back to work (the event of interest). There are, however, other events (called concurrent events) such as (a) the client's referral to government programs such as the RRQ and the CPP), (b) the proposal of a lump-sum payment to the client, or (c) a passing. Each of these events can be analyzed through a survival analysis model while viewing the other concurrent events as censorings. This method assumes that the events are independent, which is not the case.

Rather than using Kaplan-Meier curves, one could use cumulative incidence curves, which give an estimate of the marginal probability of an event given that concurrent events are present. The probabilities of these events, also called competitive risks, can be computed thanks to classical survival methods such as Cox models or cumulative incidence models and to random forests in survival analysis.

During the workshop the team will explore various methods while taking into account several criteria such as (a) missing data, (b) right censoring and left censoring, and (c) time-varying characteristics. A data sample will be provided to the team in order to test the proposed methods on real-world data.

3. Evaluation of a method for creating synthetic tabular data (Desjardins)

Data analysis and use is at the heart of two strategic goals of the Mouvement Desjardins, which seem hard to reconcile: (1) the creation of value for our members and clients, achieved through advanced analytics using any relevant information, and (2) data security, which entails the protection of confidential data.

Desjardins has developed new Artificial Intelligence (AI) approaches that allow the creation of synthetic data unrelated to any member or client of Desjardins but having the same statistical properties as the confidential data. These new approaches could serve to develop analytical models for value creation without compromising data security.

The proposed challenge is to put to the proof the new approach for making tabular data and to evaluate its performance. In particular Desjardins would like the following questions to be answered. (1) Is it possible to reconstruct a real-world observation starting from synthetic data? (2) Is it possible to tell the real-world observations from the synthetic data?

A data set consisting of identified real-world observations and synthetic data will be provided to the team. In order to answer the above questions, another data set will be provided for the evaluation of models put forward by the team.

4. Online methods and learning rates (Ericsson)

Online supervised learning methods can be seen as a class of techniques that allow a machine learning (ML) model to learn from data that is coming on the fly and cannot be stored anywhere [1]. Therefore the model can be trained only one observation at a time (or a few, in the best-case scenario). Clearly this situation makes methods based on the use of data in batches, such as the stochastic gradient descent method, unusable.

Although various online techniques exist, they all have one point in common: the user must specify a hyperparameter known as the learning rate. The value of this parameter is not easy to determine, however. If it is too large, the model will favor future observations while quickly forgetting past data. If it is too small, the model will favor past data and, consequently, have difficulties with learning from future observations. Some knowledge about the data can help at times but, for now, choosing this value remains an art. Hence it would be highly desirable to have an automatic technique for tackling this situation and to provide some guidance (at least). Ideally we would aim at a method following the same philosophy as the ADAM method [2].

The data will be provided in the shape of time series at the beginning of the workshop. Time series are regularly collected and processed at Ericsson and there is a high interest in online learning methods. The approach we would like to design in this workshop could eventually have interesting and useful applications in the field of Telecom AI.

References

[1] HOI, Steven C. H.; SAHOO, Doyen; LU, Jing; and ZHAO, Peilin. Online learning: A comprehensive survey. (2021). *Neurocomputing*. 459, 249-289. Research Collection School Of Computing and Information Systems.

[2] D.P. Kingma, L.J. Ba, “Adam: A Method for Stochastic Optimization,” *3rd International Conference for Learning Representations*, San Diego, (2015).

5. Balancing time-series data for machine learning (Hitachi)

Machine Learning (ML) based models, while achieving incredible success in many applications, are largely dependent on the data they are trained on. ML models can achieve good accuracy when their training dataset is balanced, i.e., comprises a comparable number of exemplary samples from all scenarios. In many real-world applications, constructing such a balanced dataset is difficult for reasons such as frequency of occurrences, difficulty in collecting data, missing or corrupt information, etc. In such cases the dataset suffers from imbalance, i.e., there is a high level of discrepancy between the numbers of samples in distinct classes. When an ML model is trained on such a data set, the model tends to prefer the majority class(es), at the cost of neglecting the minority class(es). This imbalance issue is a challenge for any ML model and balancing an imbalanced dataset is thus an active research field.

The most widely practiced data-balancing techniques involve resampling. There are ways to increase the samples representing the minority class(es) (up-sampling) or reducing the number of samples representing the majority class(es) (down-sampling). These techniques (such as – SMOTE [1]) require a tuning parameter such as a threshold on the degree of balance and a stopping criterion for the iterative method of correcting the imbalance through resampling. Typically the choice of these parameters depends on the application domain and the user expertise. Ericsson wishes the workshop participants to explore an automatic approach for optimizing the parameters (i.e., degree of balance, stopping criterion) while taking into account the time to rebalance and the memory required.

A labeled dataset for the protection of an electric power transmission line will be provided, where the data is comprised of time series of electrical signals and the labels indicate whether a fault has occurred within a zone of interest. There are 50 000 time series examples in the dataset, each of which having 576 time samples and 27 features. The training data was generated using a power system simulator such as PSCAD or ATPDraw. The task is to find an automatic method for balancing the data set, leading to an accurate and fast classification of electrical faults. Note that the impact of false positives (predicting that a fault has occurred within a zone of interest when actually it has occurred outside) is more detrimental than the impact of false negatives (predicting that the fault has occurred outside a zone when actually it has occurred inside).

Reference

[1] Chawla, Nitesh V., et al. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research* 16 (2002): 321-357.

6. Alarm pattern recognition (Hitachi)

Alarm systems are a vital support tool for power grid operators and an essential part of the interface between automatic and manual control. Alarms play an important role in detecting and mitigating abnormal situations that require operator attention. A well configured alarm system can greatly facilitate operations; a poorly configured alarm system can distract from important information and increase the operator workload. The number of alarms in power systems is increasing rapidly, a situation made worse by a lack of guidelines for creating consistent and informative alarms. During a disturbance, the quantity and rate of alarms are too high for operators to understand and act on them and the alarm system may hinder an operator's ability to deal with the situation.

A key challenge is that alarms are often coupled, i.e., one alarm is accompanied by a sequence of related alarms. Due to the nature of the alarm processing system, it is difficult to determine which alarm is the originating alarm or to understand the process by which alarms are connected. The problem to be solved is to analyze a large alarm time series dataset to identify which alarms follow other alarms or can be predicted from them. This problem is challenging because (a) the causal relationships within the power grid may not be stable, (b) the time stamps of alarms may be inaccurate, i.e., even if Alarm A occurs before Alarm B, sometimes Alarm B may arrive to the alarm system before Alarm A, and (c) the dataset is messy (e.g., unrelated alarms may occur within sequences of related alarms).

The dataset provided consists of 1.5 million alarms collected over a period of eight months. Each alarm has the following information: an alarm text generally describing the alarm (e.g., brief short form description + limit value + violation value), an object identifier that generated the alarm, a general geographical area to which the object belongs (i.e., substation), a data type (e.g., discrete or continuous measurement), an object class describing the type of measurement (e.g., voltage, current, etc.), and a state transition value (e.g., normal to low limit or high limit 1 to high limit 2). In addition to the raw data set, periods of alarm floods (~4500) have been pre-identified. Alarm floods arise when a very high alarm rate (up to several hundred alarms in a few seconds) results from disturbances in the power network or misconfigured alarms. For a very small number of these floods there is an indication of the type of disturbance.

Based on this information, the task is to identify causal/process-based relations between the alarms in the real data. This could lead to a real-time tool for assisting power grid operators in disturbance problem solving.

7. Long-term value of a client in the insurance industry (Intact)

To determine the long-term value of a client in the insurance industry, it is critical to understand how different clients change over time. Within the context of goods insurance these transitions could be new claims, traffic violations, but also some complex changes such as moving or adding or substituting vehicles. This results in a huge number of transitions and possible states: thus evaluating all combinations is complex or impossible. Building a stochastic model of all the transitions is also very costly. These challenges prevent us from estimating the long-term value of a client in a correct and representative way. To solve this problem, we have tried (without much success) to create a dense representation of clients so as to simplify the transition process, but unfortunately we then lose, in general, the interpretability afforded by the client's information.

During the workshop we would like to model the client's evolution over time by using a Markov chain, since this approach has been used in other sectors (such as the banking sector). The Markov chain approach, however, entails a large quantity of possible states, if one takes into account all combinations of vehicles, addresses, product choices, etc. Therefore, to limit the number of states, one must reduce the dimensionality of the data set, especially considering the large number of vehicles and postal codes.

Intact will provide the team members with an anonymized data set containing insured clients and some insurance variables. The data set will also contain snapshots of clients' situations at specific times. Transitions can be observed by comparing consecutive snapshots. To ensure confidentiality the data set will be extracted from the period 2006-2010 and the address, the postal code, and other confidential variables will not be included in the data set. Thus the team will not study transitions related to addresses but the methodology developed by the team will be relevant for address changes as well. This methodology will enable Intact to model efficiently several transition types in order to compute a long-term value that is representative.

8. Determining the right moment for suggesting the creation of an account (Radio-Canada)

At CBC/Radio-Canada our mission is to serve the Canadian public. One of our strategic objectives is the personalization of our digital services. Our goal is to offer services in which each Canadian will be able to recognize herself or himself and to highlight the diversity of voices, communities, and viewpoints that is the richness of our country.

From this perspective a better knowledge of our users is essential: this is why we wish to encourage the users of our platforms to identify themselves. This encouragement must be carried out in a responsible manner.

In our opinion statistical and machine learning methods could help us predict the right moments for suggesting an account creation to a user visiting one of our platforms.

This model could maximize the probability of an account creation while mitigating the annoyance caused by an unsolicited interruption.

Our first challenge comes from the fact that our digital services contain hundreds of thousands of pages, videos, and articles. Therefore we must cluster the pages in order to analyze and interpret a user's digital browsing.

Our second challenge is the phenomenon of *class imbalance* when considering the creation of accounts.

We are currently working with a sample corresponding to a month of data. This sample contains anonymized click-by-click data for the set of all visits to our web platforms.

References

Our strategic plan 2019-2024
Our data policy

9. Predicting the demand for spare parts (Société de transport de Montréal)

The goal of inventory management for spare parts is to avoid inventory shortages and limit inventory costs. At any point in time there must be the right spare parts in the right place, in all the organization maintenance centres, while minimizing the storage cost. An increase in the accuracy of predicted demand will result in a better performance of the Société de transport de Montréal (STM) regarding inventory management and asset maintenance. Contrary to a department store, which can remove from its catalogue an unprofitable product, the STM cannot remove from its catalogue a part that is not replaced often. To maintain the quality of its assets the STM must ensure that all the spare parts in its catalogue are available.

The classical predictive models give good results when the product use is relatively stable. In the case of parts whose demand profiles are more lumpy and sporadic, these models are not as efficient and statistical analyses of product use may help in obtaining predictions. Researchers such as Croston, Boylan, and Syntetos have studied such statistical models and the STM wishes to evaluate their performance on its parts catalogue.

On the other hand the parts in the STM catalogue may be clustered in categories. Some parts are critical for operations; some have very stable profiles and others are seldom used. The STM has already defined clusters within its catalogue, based on demand profiles, but is wondering how to improve the current classification. How many categories should there be? How does one evaluate the distance between the demand profiles of two spare parts? What recommendations can one find in the scientific literature? Each demand profile has an impact on the parts management and the storage strategies. Therefore the STM must attempt to predict the demand for each cluster of spare parts.

The team’s task will consist of clustering spare parts according to their demand profiles, determining a demand model for each cluster, and evaluating the predictions made by the model. The spare parts included in the study will be those appearing in the catalogue for buses.