Lifelong semi-supervised tabular learning

Letter of Intent for IBS Young Scientist Fellowship - Sundong Kim (Dec 13, 2020)

Keywords: Tabular learning, Lifelong learning, Deep learning, Semi-supervised learning, Representation learning, Data augmentation, Hierarchical classification, Noisy data

Purpose

% Please describe the reason and purpose for selecting your research theme (No more than 250 characters)

Despite being the most common data type in real-world, deep learning and AI methods for tabular data remain under-explored. By studying lifelong DNN modeling to tabular data domains, my research will bring fundamental advances to applied data mining.

Research theme

% Please describe the objectives and content of your research idea, expected research outcomes and their potential academic and societal impact. (No page limit)

Deep neural networks (DNNs) have shown notable success with images, text, and audio. For these, canonical architectures that efficiently encode the raw data into meaningful representations, fuel the rapid progress. One data type that has yet to see such success with a canonical architecture is tabular data. Despite being the most common data type in real-world AI (as it is composed of any categorical and numerical features), deep learning for tabular data remains under-explored, with variants of ensemble decision trees (DTs) still dominating most applications [1]. DT-based approaches have certain benefits such as i) they are representationally efficient, ii) highly interpretable, iii) fast to train.

However, it is worth exploring deep learning for tabular data. First, end-to-end modeling with DNN alleviates the need for feature engineering, a key aspect of DT-based tabular learning. Modern machine learning tries to avoid feature engineering since it is labor-intensive that takes advantage of human ingenuity and prior knowledge. Second, we expect huge performance improvement by merging self- and semi-supervised techniques proposed in the natural language processing (NLP) or computer vision (CV) domain. To incorporate recent advancements in tabular settings, an end-to-end DNN model is essential since these methods are difficult to apply to the DT-based model. Last, DNN allows diverse applications such as generative modeling or domain adaptation. In the next few years, I would like to study an end-to-end tabular learning framework, starting from augmentation strategies for self-, and semi-supervised learning. Then I will extend this framework to lifelong learning (aka. continual learning) settings by making it robust to concept drifts. With this configuration, the proposed tabular learning framework can be easily deployable in the industrial setting. Along this line, I will consider wild setups such as data imbalance, extreme classification, and robustness to input noise while designing the framework.

The main part of the proposal is organized into five components: 1) Data augmentation and representation learning, 2) semi-supervised learning, 3) lifelong learning, 4) hierarchical classification, 5) robust learning to mitigate input noise and adversarial attacks. Each section includes in which context the components needed for lifelong semi-supervised tabular learning. In addition to introducing the contribution of recent studies, I considered how to generalize and develop those studies to make them useful in practical applications. After introducing the main contents, I will conclude the letter of intent with reasons why I would like to continue my research with the IBS data science group.

1. Data augmentation and representation learning for tabular data

Representation learning (or feature learning, embedding learning) is a set of techniques that allows a machine learning system to automatically discover the latent features that best represent the data point. Finding effective data representation is crucial, as machine learning methods' performance is heavily dependent on the choice of data representation on which they are applied. The major direction in representation learning is to use self-supervised learning with data augmentations. The basic idea of self-supervised learning is to automatically generate supervisory signals to learn representations, since annotating all dataset is no longer available in the era of data flooding.

In the domain of computer vision (CV) and natural language processing (NLP), there have been huge advances in self-supervised representation learning [1], since input data often have spatial or semantic relationships between pixels or words, which naturally led to diverse data augmentation strategies for supporting self-supervision. For example, image augmentation techniques such as geometric transformations, flipping, color modification, cropping, rotation, noise injection, and random erasing are taken for granted in every application. Besides, augmentation strategies for NLP include word substitution by thesaurus and word-embedding, random noise injection, synthetic-tree manipulation, back translation, and masked language model. These allow self-supervised learning frameworks to learn data representations effectively and such advances significantly raise the performance of so-called downstream tasks, such as face recognition, and question and answering. However, most existing works with self-supervised learning are applicable only for images or natural language due to their data augmentation scheme, which in general does not support tabular data. The major reason why these methods are difficult to apply in the tabular dataset is that the correlation structure among features in tabular data is unknown and varies across different datasets. In other words, there is no "common" correlation structure in tabular data. This makes the self-, and semi-supervised learning in tabular data more challenging.

Recently, self-supervised tabular learning has begun to be proposed in the form of blocking some features as an alternative to data augmentation [3-5]. Their idea is to corrupt some features and recover them. While recovering corrupted features (pretext task), the representation of each data sample is learned. They showed the feasibility of end-to-end self-supervised learning for tabular settings. Nevertheless, a very limited form of augmentation (i.e., corruption) is only applicable to these methods and there is a long way to go. Promising self-supervised contrastive learning frameworks in the image domain such as SimCLR [6] requires diverse augmentation strategies, in which contrastive learning is an approach to formulate the task of finding similar and dissimilar things, and it requires diverse ranges of augmentation strategies; from weak to strong.

To transfer the successes of self-supervised contrastive learning from image to tabular domains, proposing applicable and proper pretext tasks and augmentations for tabular data is critical. I am currently leading a project on advancing customs' targeting systems, and we are developing augmentation strategies for customs import declaration data, which is a tabular format. Given that domain, we start from the simple idea of weak augmentation in the form of scaling-up and scaling-down the quantities. Moderate augmentation strategy may replace the value of a categorical variable with a similar one, or control numeric values. A straightforward example is to change user ID, which is not crucial in determining the item's tariff changes. An example of strong augmentation is to switch the commodity code to a similar one, a change of commodity code may lead to some changes in the final tariff rate.

My research will generalize to any tabular format. The tentative augmentation approach for general tabular data consists of two steps. First, find the pre-trained embeddings for each feature and values, according to the data structure and value distribution of each feature. After that, generate an infomax matrix between features, and between their feature values. With this information, the augmentations of various intensities can be created. After showing its efficacy, we will contribute our tabular augmentation modules to widely-used frameworks such as Pytorch or Tensorflow.

2. Toward semi-supervised learning

We can extend our task to task agnostic semi-supervised learning (SSL) by fine-tuning the model towards labeled samples. More specifically, a SSL model can be trained by learning a self-supervised learning objective for unlabeled samples and cross-entropy loss for labeled samples. This training method can use the labeled sample's information and the unlabeled sample's information, so if there is sufficient unlabeled data, the model can be effectively trained.

VIME is a recently proposed methodology that applies this framework to tabular data [5]. This method also performs classification on the labeled sample in the middle of guessing the hidden value after masking some columns for the unlabeled sample. In this method, after masking some columns for unlabeled samples, classification for labeled data is performed as well as mask prediction for unlabeled data. However, there is still a limitation that the self-supervised learning methodologies for tabular data currently only use simple pretext tasks such as random masking, and this is a bit far from the techniques used in current state-of-the-art contrastive learning frameworks.

In practice, the SSL methodology can be effectively applied to improve the performance of the Customs Service's AI-based targeting system. Due to the large trade volumes, only a small number of goods can be physically inspected, and remaining items are cleared without being inspected. Currently, the system operates in a fully-supervised way that uses only the inspected items' information. If we successfully deploy the SSL method, the AI-based targeting system can more precisely operate and protect our nation from unexpected hazardous goods.

Another way to tackle this problem is by using semi-supervised anomaly detection. Anomaly detection is the process of identifying data points that deviate from a dataset's normal behavior. Critical and hazardous goods would be marked as high anomalous scores. The problem setting we encountered in the actual customs is very challenging. Only 5-10% of the data is labeled, and a very small portion of those labeled data are marked as anomalies. Besides, we don't have any clue that uninspected items are normal. Recently, several papers have been published on semi-supervised anomaly detection, but some additional assumptions have been added compared to the actual settings we encountered.

Deep-SAD [8] measures the anomaly score as the distance from a predefined point. The encoder model pulls normal data closer to this point, and keeps anomalies away from the point. However, all unlabeled samples are assumed to be normal and pulled towards the point without any other treatment. GOAD [9] pulls original data and augmentations together into a predefined point. For augmenting tabular data, they generate random matrices then apply affine transformations. With a loss function that collects normal data into a single point, it classifies which affine transformation is used for each data point. With this self-supervision, the anomaly score is refined. However, such affine transition may not be applied to categorical input in the desired form.

With our representation learning scheme, several clusters can occur among cleared (normal) goods, but these two approaches simply pull all normal dataset into a single point. We plan to overcome these limitations by designing a novel anomaly detection algorithm. Tentative approach will narrow the difference in classification results between the data point and its augmentations. The model does not necessarily pull all points into a single point, instead using multiple centroids. This can be achieved by tweaking the mutual information concept from our image clustering [10] and supervised contrastive learning technique [11].

3. Toward static prediction setup to the lifelong learning setting

What if we develop the semi-supervised tabular leading model, would it be applicable to live environments? Research on semi-supervised learning mainly concentrates on the static setting, in which a model is trained from large training batches and evaluated with a held-out test set. However, there is an ambush called domain shift or concept drift in an online setting. The underlying data distribution changes over time and traditional approaches will fail to detect frauds.

To address this issue, recent studies have begun focusing on "lifelong learning" [12]. One changes the architecture of the model dynamically over time [13], the other selects appropriate data instances to help to converge the model in time [14]. However, since most of the ML research simulates on benchmark data such as CIFAR-10, realistic research on lifelong learning in tabular data lacks so far. The customs import declaration data that we have in our hands is a longitudinal setup with concept drift, it is a good situation where I can bring this study closer to reality.

Recently, we started our research on expanding the custom targeting system to lifelong settings. We tested whether our fraud detection model DATE [15], which showed a high detection rate in the static setting, can be well adapted to the domain shift. We simulated that the model is deployed for the targeting system and used for several years. Although the DATE model is known to detect fraudulent items successfully, its working mechanism is highly dependent on historical data. Therefore, the performance drops in such countries with domain shifts. To cope with this problem, we presented an exploration strategy to inspect uncertain items and effectively combine it with the DATE model [16]. This brings insights for practical guidelines for setting model parameters in the context of customs targeting systems.

In addition to extending this effort into general semi-supervised tabular settings, there is another major direction that this work can be extended. That is to study the reinforcement learning methodology that can adaptively determine the exploration ratio. During the revision process, we received a comment that our exploration-exploitation dilemma may be relevant to multi-armed bandit theory and reinforcement learning. Toward this direction, we will devise a metric to measure the amount of domain shift and consider this value as a dataset difficulty to train the reinforcement learning frameworks. This is a promising direction that we can extend a recent study introducing the effectiveness of using semi-supervised active learning [17]. Our experiment on finding exploration ratio in a lifelong learning setting will produce meaningful results in the intersection of reinforcement learning, semi-supervised learning in online active setups.

4. Toward hierarchical classification setting

So far I discuss the necessity of a life-long semi-supervised tabular learning framework. In practice, the problem can face practical issues such as data imbalance, extreme classification, and input noises. While designing the framework, I will consider these challenging issues. In this section, I will address a problem when the number of labels in the classification issue is very high, or when the hierarchy of the label exists.

In our previous research of customs targeting by using data from four developing countries, we set our problem as multi-task learning with binary fraud labels and their corresponding revenue. On the other hand, some countries successfully track more than a hundred types of frauds. And each fraud is categorized with hierarchies. Another problem we are tackling is to classify the category of imported goods. The problem setting is even more complex. The number of product categories is more than thousands and the product hierarchy is occasionally updated. The tabular inputs also include some images or product description texts. The algorithm aims to classify products into the most relevant sub-categories in due course since customs administration must let importers know the exact tariff of the goods. There are many tricky cases, such as PC monitors and televisions have different tariff rates when they are imported.

With hundreds of thousands of categories, the classification task becomes much more challenging. Large numbers of classes makes the class imbalance issue more prominent and the system may face missing value problems. To remedy these challenges, hierarchical classification and extreme classification frameworks have been proposed among researchers [18-20], that provide systematic ways to deal with multi-class classification with a given hierarchy. We would like to extend these studies in the direction of lifelong-learning settings with dynamic environments. This study will widely impact the web domain. For example, user experiences will be improved if articles in Wikipedia or products in e-commerce are managed well.

5. Robust learning to mitigate input noise and adversarial attacks

Although developing a semi-supervised tabular learning model for a life-long prediction setting is such a huge area that these topics would already be enough for spending several years of endeavor, we cannot ignore the security concerns from our external collaborators in practice. Assuming that our proposed model is used in a sensitive application, there exist some concerns from outside that their logic will be published to the public. By knowing how the model works, it would be easier to make malicious attempts to infer the decision boundary of the system by tweaking some features [21]. For example, dishonest importers may declare their goods by fooling some information to avoid ad-valorem duties and taxes.

Small changes can confuse a model's decisions, which is a well-known problem of deep learning. Even if there is no actual confidential data leakage, the decision boundary of fraud can be inferred through the results obtained by sending multiple queries (i.e., Black-box attack). Fraudsters can make an imperceptible change of features to avoid inspection based on the inferred decision boundary. Furthermore, If the architecture and parameters of the model are leaked, importers can launch more powerful attacks by using detailed model information, including the gradient of the loss (i.e., White-box attack). Our two-stage tree-enhanced neural network model [7] is known to be robust to well-known gradient-based attacks because of the non-differentiable tree components. However, attacking strategies on boosting trees are also being studied recently [22].

With this motivation, I will study robust learning components on our lifelong learning framework by simulating adversarial attacks. I would like to see how many attacks can be defended by attaching anomaly filters on our model. The anomaly filter will reduce the size of the gradient for the input to reduce the change in the result of noise, or filter out the input that contains noise when it comes in. Adversarial attacks are the type of noise that most drastically changes the outcome of the machine learning model. The defense strategy to protect adversarial noise will be applied to various types of input noise, which results in a robust lifelong learning model.

6. Concluding remarks

In the next few years, I would like to study a lifelong semi-supervised tabular learning framework. Tentative research topics include (but are not limited to):

1) Data augmentation and representation learning for tabular data
2) Semi-supervised classification and outlier detection
3) Strategy to cope with domain shift in a lifelong learning setting
4) Multi-class classification with hierarchies

5) Robust learning to mitigate input noise and attacks

This foundational research will also have practical applicability that the research outcome will contribute to data mining and machine learning communities. We hope to contribute our research to the applied domain, such as the web, e-commerce, and non-profit community as well. Currently, I am leading a collaboration between IBS and the World Customs Organization. Under the name of IBS, my research contributes to advancing data analytics capabilities in customs organizations. The DATE algorithm that I developed last year is being applied to improve the Korean Customs Service's targeting system for import clearance, and it is included in training materials for customs officers and data scientists from the partner countries of the World Customs Organization. Also, the hierarchical classification algorithm to be created during the YSF will be applied to the navigation service for classifying the imported goods.

The IBS data science group is the best place to work on this topic since IBS supports challenging research and our group already has close collaboration with data providers and domain experts. Providing foundational algorithms in lifelong tabular learning will also benefit other group members to solidify their data science and computational social science research.

Besides, advancing deep tabular learnings comes with novel augmentations and wide-and-deep models, which necessitates huge computation cost. Training deep models often takes a few weeks with dozens of GPUs and the price of those computing machines is beyond the budget of individual PIs. Computation resources in our group will be a great help in conducting research.

Last, the IBS data science group has strong student support. I had the privilege of working with excellent students from KAIST as well as collaborators from five different countries. Currently, four graduate students and six undergraduate interns are under my direct supervision. Continuing research mentorship would be a good opportunity for building my career.

With the generous support of the YSF problem, I hope to be able to conduct this research at IBS.

References

[1] LightGBM: A Highly Efficient Gradient Boosting Decision Tree. NeurIPS 2017 [Link]
[2] Self-Supervised Representation Learning. Online article [Link]
[3] TaBERT: Learning Contextual Representations for Natural Language Utterances and Structured Tables. ACL 2020 [Link]
[4] TabNet: Attentive Interpretable Tabular Learning. arXiv:1908.07442 [Link]
[5] VIME: Extending the Success of Self- and Semi-supervised Learning to Tabular Domain. NeurIPS 2020 [Link]
[6] SimCLR: A Simple Framework for Contrastive Learning of Visual Representations. [Link]
[7] MixMatch - A Holistic Approach to Semi-Supervised Learning. NeurIPS 2019 [Link]
[8] Deep-SAD: Deep Semi-Supervised Anomaly Detection. ICLR 2020 [Link]
[9] GOAD: Classification-Based Anomaly Detection for General Data. ICLR 2020 [Link]

[10] Mitigating Embedding and Class Assignment Mismatch in Unsupervised Image Classification. ECCV 2020 [Link]
[11] CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances. NeurIPS 2020 [Link]
[12] Lifelong Machine Learning. Morgan & Claypool, 2018 [Link]
[13] Lifelong Learning with Dynamically Expandable Networks. ICLR 2018 [Link]
[14] Carpe Diem: Seize the Samples Uncertain "at the Moment" for Adaptive Batch Selection. CIKM 2020 [Link]
[15] DATE: Dual Attentive Tree-aware Embedding for Customs Fraud Detection. KDD 2020 [Link]
[16] Take a Chance: Managing the Exploitation-Exploration Dilemma in Customs Fraud Detection via Online Active Learning. arXiv:2010.14282 [Link]
[17] Consistency-based Semi-supervised Active Learning: Towards Minimizing Labeling Cost. ECCV 2020 [Link]
[18] Extreme Classification Workshop. ICML 2020 [Link]
[19] Hyperbolic Interaction Model for Hierarchical Multi-Label Classification. AAAI 2020 [Link]
[20] Coherent Hierarchical Multi-Label Classification Networks. NeurIPS 2020 [Link]
[21] GRACE: Generating Concise and Informative Contrastive Sample to Explain Neural Network Model's Prediction. KDD 2020 [Link]
[22] Robust Decision Trees Against Adversarial Examples. ICML 2019 [Link]

Additional remarks

% Please describe how utilizing the Center's infrastructure can create synergy. (No more than 250 characters)

Computation resources in the IBS data science group will be a great help in conducting deep learning research. Also, I will continue the privilege of continuing research mentorship to excellent students and international collaborations.