Machine Learning with Low-Resource Data from Psychiatric Clinics

Hongmin W. Du[1], Neil De Chen[2], Xiao Li[3(✉)], and Miklos A. Vasarhelyi[1]

[1] Accounting and Information Systems Department, Rutgers University, Piscataway, NJ 08854, USA
hd255@scarletmail.rutgers.edu, miklosv@business.rutgers.edu
[2] School of Medicine, Saint Luis University, St. Louis, MO 63103, USA
[3] Department of Computer Science, University of Texas at Dallas, Richardson, TX, USA
xiao.li@utdallas.edu

Abstract. Amidst the rapid growth of big data, the success of machine learning is critically tethered to the availability and quality of training data. A pertinent challenge faced by this symbiotic relationship is the issue of "low-resource data," characterized by insufficient data volume, diversity, and representativeness, and exacerbated by class imbalances within datasets. This study delves into the intersection of machine learning and big data, exploring innovative methodologies to counteract the challenges of data scarcity. Focusing on psychiatric clinic data, marked by subjectivity and inconsistency, we outline the unique challenges posed by the nature of data in this domain. To address these challenges, we explore the potential of data augmentation-using transformations or operations on available data-and transfer learning, where knowledge from a pre-trained model on a large dataset is transferred to a smaller one. Through a comprehensive exploration of these methodologies, this research aims to bolster the effectiveness of machine learning in low-resource environments, with a vision of advancing the digital landscape while navigating inherent data constraints.

Keywords: Small Data · Medical Data · Machine Learning · Low Resource Data · Data Augmentation

1 Introduction

In the era of exponential data growth, the role of machine learning has emerged as a pivotal force in extracting valuable insights and knowledge from vast and complex datasets, commonly referred to as big data. This symbiotic relationship between machine learning and big data has led to significant advancements across numerous domains, ranging from healthcare and finance to marketing and autonomous systems. However, amidst the remarkable progress achieved, a persistent challenge looms large, casting a shadow on the efficacy of machine

W. Wu and J. Guo (Eds.): COCOA 2023, LNCS 14462, pp. 455–465, 2024.
https://doi.org/10.1007/978-3-031-49614-1_34

learning algorithms-the scarcity of adequate training data and the pervasive issue of class imbalance within datasets. This challenge, synonymous with the term "low-resource data," has garnered substantial attention and constitutes a paramount concern in the field of machine learning.

Despite the undeniable potential of machine learning to revolutionize decision-making and predictive modeling, its effectiveness crucially hinges upon the quality and quantity of training data available. While the digital landscape is inundated with data, a considerable portion of it remains inadequate in terms of volume, diversity, and representativeness. This limitation not only impedes the development of accurate and robust machine learning models but also diminishes their potential to generalize to new, unseen instances. Concurrently, the uneven distribution of class labels within datasets, commonly referred to as class imbalance, further compounds the predicament by skewing the learning process in favor of the majority class, often at the expense of the minority classes. Consequently, these inherent challenges collectively underscore the pressing need to devise innovative strategies and methodologies that address the intricacies of low-resource data and mitigate the associated adversities.

In this research endeavor, we delve deep into the multifaceted domain of machine learning and big data, centering our focus on the intersection between model performance and the scarcity of training data. By analyzing the manifold implications of low-resource data scenarios and the intricate nature of class imbalance, we aim to unearth novel approaches that enable the harnessing of invaluable insights even from data-scarce environments. Through a comprehensive exploration of methodologies ranging from transfer learning and active learning to data augmentation and synthetic data generation, we strive to not only elucidate the intricacies of these strategies but also evaluate their effectiveness in ameliorating the performance of machine learning models when confronted with the challenges of low-resource data.

As we navigate through the unknown areas of this research, our ultimate goal is to contribute to the arsenal of tools and techniques that empower machine learning practitioners to transcend the constraints imposed by data scarcity and class imbalance. By doing so, we envision a future where the promise of machine learning remains steadfast, unswayed by the inherent limitations of data availability, and continues to pave the way for unprecedented advancements in our ever-evolving digital landscape.

For example, in the analysis of psychiatric clinics data, we face a difficult situation that the available data is not large enough due to the following.

- Psychiatry is a highly subjective field when it comes to gathering patient data, characterized by a scarcity of numerical information and marked inconsistency in the documentation of symptoms, signs, concerns, and the progression of cases across different psychiatrists. The lack of standardized language, apart from diagnostic criteria, presents a significant challenge for Natural Language Processing (NLP) to effectively capture and analyze the data in a coherent manner.

- Diagnosis does not always dictate the treatment plan; more frequently, it serves as a coding tool. However, diagnoses can undergo changes quite easily. Moreover, many medications are employed in off-label capacities, deviating from algorithmic guidelines. As a consequence, this diversity leads to a wide array of distinct approaches.
- Medical records in the field of psychiatry are often unreliable, with numerous patients receiving incorrect diagnoses that do not align with current diagnostic criteria. There is a prevalence of inaccurate or misleading information that may not be rectified by subsequent psychiatrists, leading to a lack of proper sanitization of the records.
- There is no evidence indicating that superior documentation or clustering of patient data directly results in improved patient outcomes. This raises an ethical dilemma, as it could be perceived that such practices primarily benefit hospital administration financially, potentially exacerbating the challenges patients face in accessing proper care.

The aforementioned facts underscore the challenging nature of researching psychiatric clinic data. Given this context, what methodologies can be employed? In this concise article, we explore potential techniques that could be applied to the analysis of psychiatric clinic data. By comparing these techniques, our aim is to identify a more effective approach. Subsequently, in future research endeavors, we aspire to extract valuable insights from psychiatric clinic data, benefiting both psychiatrists and their patients.

2 Data Augmentation

One of the ways for dealing with low-resource data is so called data augmentation. What is data augmentation? This involves artificially increasing the size of the training set by creating new data instances through transformations or operations on available data. The following are three examples.

In [9], data augmentation is employed to deal with the task of image classification. They compared and analyzed multiple methods of data augmentation, including classical image transformations like rotating, cropping, zooming, histogram based methods, and operations at Style Transfer and Generative Adversarial Networks, together with the representative examples. They also presented their own method of data augmentation based on image style transfer. Those methods generate the new images of high perceptual quality, which can be used to pre-train the given neural network in order to improve the training process efficiency. Finally, the three medical case studies are carried out to validate proposed method. They are skin melanomas diagnosis, histopathological images and breast magnetic resonance imaging (MRI) scans analysis. The image classification is utilized to provide helpful information for a diagnose. In such medical case studies, the data deficiency is a very important relevant issue. Moreover, they discussed the advantages and disadvantages of discussed methods.

In the study of computer vision, augmented data are still images. Considered operations include horizontally flipping, random cropping, tilting, and altering

the color channels of the original images. Since the content of the new image is still the same, the label of the original image is preserved. This situation is changed in training networks for NLP tasks such as Machine Translation. Given a source and target sentence pair (S, T), one may need to alter it in a way that preserves the semantic equivalence between S and T. For low-resource language pairs, this is difficult to do. In [4], a novel data augmentation approach is proposed to target low-frequency words by generating new sentence pairs containing rare words in new, synthetically created contexts. Their method is found to have improved translation quality.

For text classification tasks, four simple but powerful operations are synonym replacement, random insertion, random swap, and random deletion. In [20], data augmentation techniques with those operations are employed in the study of five text classification tasks. They demonstrate strong results for smaller datasets and improve performance for both convolutional and recurrent neural networks. Across five datasets, with mentioned data augmentation techniques, using only 50% of the available training set achieved the same accuracy as normal training with all available data. Moreover, they give suggested parameters for practical use.

All above three examples contain helpful information for the analysis of psychiatric clinics data.

- Psychiatric clinics data is in text format. The task of identifying proper diagnosis is equivalent to doing text classification. Therefore, those operations mentioned in [20] may be useful.
- However, we have to select operations carefully because label (i.e., diagnosis) should be preserved, such as summarization and rearrange ordering of sentences and consider those in the case of skin melanomas diagnosis in [9].
- To preserve the label, we may treat psychiatric clinics data as data pairs of patient's suboptimal statement and doctor's diagnosis with treatment. Then, employ techniques in [4] to augment data pairs.

3 Transfer Learning

Transfer learning is a simple and powerful method. It can be used to boost model performance of low-resource neural machine translation. This technique includes taking a pre-trained model (usually trained on a large dataset) and fine-tuning it on a smaller dataset. The idea is that the pre-trained model has already learned useful features from the larger dataset, which can be applied to the smaller dataset.

Existing transfer learning methods for neural machine translation are simply transfer knowledge from a parent model to a child model once via parameter initialization. It has been showed that the encoder-decoder framework for neural machine translation is very effective in large data scenarios, but much less effective for low-resource languages. In [21], a transfer learning method is presented. This method significantly improves BLEU (Bilingual Evaluation Understudy) scores across a range of low-resource languages. The key idea is to first train a

high-resource language pair of encoder-decoder (the parent model), then transfer the learned parameters to the low-resource pair (the child model) to initialize and constrain training.

In [8], a novel transfer learning method, namely ConsistTL, for neural machine translation is proposed. ConsistTL is able to continuously transfer knowledge from the parent model to the child model during the training of the child model. Specifically, the child model learning each instance under the guidance of the parent model, that is, for each training instance of the child model, ConsistTL constructs the semantically-equivalent instance for the parent model and encourages prediction consistency between the parent and child for this instance. Experimental results demonstrate that ConsistTL results gives significant improvements over strong transfer learning baselines.

A unified framework is introduced in [14] that converts all text-based language problems into a text-to-text format, which explores the landscape of transfer learning techniques for NLP. With this framework, a systematic study is made in the same paper to compare retraining objectives, architectures, unlabeled data sets, transfer approaches, and other factors on dozens of language understanding tasks.

A new idea for unsupervised domain adaptation via a remold of Prototypical Networks is introduced in [11]. The goal is to learn an embedding space and perform classification via a remold of the distances to the prototype of each class. They present Transferrable Prototypical Networks for adaptation such that the prototypes for each class in source and target domains are close in the embedding space and the score distributions predicted by prototypes separately on source and target data are similar.

If we want to use this technique to analyze psychiatric clinics data, then we may need to do the following.

- Identify a machine learning model for analysis of psychiatric clinics data.
- Find a medical data with high resource which can have the same machine learning model. This seems very hard task since psychiatry is so different from other medical fields.

4 Few-Shot/Zero-Shot Learning

This is a technique where the model is designed to make accurate predictions given only a few or no examples. This approach often involves meta-learning where the model is learning the structure or meta-knowledge across different tasks, so it learns a prior over models that is useful for new tasks. Let us first look at a few examples.

In [16], Prototypical Networks are proposed for the problem of few-shot classification. Given only a small number of examples of each new class, a classifier must generalize to new classes not seen in the training set. By computing distances to prototype representations of each class, Prototypical Networks learn a metric space in which classification can be performed. Compared to

recent approaches for few-shot learning, their approach is simpler and achieve excellent results. Actually, they provide an analysis showing that some simple design decisions can yield substantial improvements over recent approaches; those approaches involve complicated architectural choices and meta-learning.

Meta-learning is a framework to address the challenging few-shot learning setting. It leverages a large number of similar few-shot tasks in order to learn how to adapt a base-learner for a new task when for the new task, only a few labeled samples are available. In [17], a novel few-shot learning method, called meta-transfer learning, is proposed. This method learns to adapt a deep neural networks for few shot learning tasks. Here, by meta, it means to train multiple tasks, and by transfer it means that learning is achieved by scaling and shifting functions of deep neural network weights for each task.

Deep neural networks is successful in the large data domain, but perform poorly on few-shot learning tasks if a classifier has to quickly generalize after seeing very few examples for the class. Generally speaking, gradient-based optimization in high capacity classifiers needs many iterative steps over many examples in order to perform well. In [15], a meta-learner model LSTMbased is proposed to learn the exact optimization algorithm used to train another learner neural network classifier in the few-shot regime. This meta-learning model is competitive with deep metric-learning techniques for few-shot learning.

In [7], an effort is made on prompts for pre-trained language models. It has shown great performance in bridging the gap between pre-training tasks and various downstream tasks. Especially, prompt tuning freezes pre-trained language models and only tunes soft prompts, gives an efficient and effective solution for adapting large scale pre-trained language models to downstream tasks.

This technique is similar to transfer learning. They transfer information obtained from high-resource data to low-resource data. Therefore, we may see it as a variation of transfer learning. The difference is that the parent's model and the child model have closer relationship.

5 Active Learning

This is a special case of machine learning where a learning algorithm can actively choose the data it wants to learn from. It's particularly useful when unlabeled data may be abundant or easy to collect, but labeling data is costly, time-consuming, or requires expert knowledge. Thus, it looks like another variation of transfer learning. We may find three examples in [3,5,6].

In text classification, labels are usually expensive and the data is often characterized by class imbalance. This gives a challenge in Real world scenarios for active learning. In [3], a large-scale empirical study is presented on active learning techniques for BERT-based classification, and a diverse set of AL strategies and datasets is addressed.

In active learning, a small subset of data is selected for annotation such that a classifier learned on the data is highly accurate. Usually, selection is done by

using heuristics. To improve the effectiveness of such methods, an effort in [5] is made by introducing a novel formulation which reframes the active learning as a reinforcement learning problem and explicitly learning a data selection policy. Here, the policy takes the role of the active learning heuristic.

Active learning methods rely on being able to learn and update models from small amounts of data. Recent advances in deep learning are notorious for their dependence on large amounts of data. This difference makes deep learning is difficult to be used in active learning. However, in [6], authors combine recent advances in Bayesian deep learning into the active learning framework in a practical way and develop an active learning framework for high dimensional data, which is a task extremely challenging.

6 Self-supervised Learning

This is a type of machine learning where the model generates its own supervised learning signals from the input data itself. It is a method of training where the labels for the training data are automatically generated from the data itself, without any human annotation. For example, a model might be trained to predict the next word in a sentence, and the learned word embeddings can then be used for a task like sentiment analysis. One paradigm for self-supervised learning is from few labeled examples while making best use of a large amount of unlabeled data, that is, unsupervised pre-training followed by supervised fine-tuning.

For self-supervised learning from images, the goal is to construct image representations. They are semantically meaningful via pretext tasks that do not require semantic annotations. A lot of pretext tasks lead to representations that are covariant with image transformations. However, in [10], authors argue that semantic representations should be invariant under such transformations. Moreover, they develop Pretext-Invariant Representation Learning that learns invariant representations based on pretext tasks.

Self-supervised learning is learning from few labeled examples while making best use of a large amount of unlabeled data. [1], authors proposed an approach by using big (deep and wide) networks during pretraining and fine-tuning. They found that for their approach, the fewer the labels, the more this approach (task-agnostic use of unlabeled data) benefits from a bigger network.

In [2], a new language representation model is introduced in [2]. This new model is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. It is a pretrained model, which can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks

In [12], a new global logbilinear regression model is proposed. This model combines the advantages of the two major model families in the literature, global matrix factorization and local context window methods. It efficiently leverages statistical information by training only on the nonzero elements in a word-word co occurrence matrix.

7 Multi-task Learning

This involves training a model on multiple related tasks at the same time, with the aim that learning from each task with high-resource data can help improve performance on the other task with low-resource data. Examples can be found in [13,18,19].

8 Conclusion

Summarizing the above six techniques, we found that there are only two ways to deal with low-resource data for our particular use case.

- Data Augmentation: Generate more data by data operations.
- Transfer Learning in Wide Sense: Transfer the information or parameters from analysis of high-resource data to low-resource data. In Transfer Learning, it transfers from a parent model to a child model. In Few-Shot/Zero-Shot Learning, it transfer from previous task to new task. In Active Learning, it transfers from unlabeled data to labeled data. In Self-Supervised Learning, it transfers from own unlabeled data to labeled data. In Multi-Task Learning, it transfers one to another data where they are processed together.

How to use these techniques in analysis of psychiatric clinic data? We have discussed at the ends of Sects. 2 and 3, respectively.

Other than selection of machine learning techniques, selection of research goal is also important for analysis of psychiatric clinics data. The following suggestions stem from via the point of a psychiatrist.

If we are interested in creating a tool, then we may consider the following:

- Develop a tool capable of analyzing a patient's historical records, past medical information, and the patient's current daily note (encompassing both Subjective and Objective aspects of the visit). Utilize this information to generate a "recommended diagnosis" primarily for coding purposes-assigning a corresponding code to a specific diagnosis for billing purposes. The goal is to alleviate the documentation burden on physicians. While similar tools already exist in certain documentation systems, there's potential to enhance and refine this approach in comparison to the existing solutions. However, it remains uncertain whether surpassing the capabilities of systems like Epic is achievable.

If we are interested in clustering, then we may consider the following:

- Would recommend limiting scope to a single type of diagnosis and looking for clusters within that diagnosis, ie possible clusters of major depressive disorder.
- A lot of psychiatry involves trial-and-error for final medication selection and titration; one hypothesis is that there are subtypes which are as yet invisible to our diagnostic criteria. Identifying these subtypes and predicting them

would result in fewer trials of medications before achieving optimal control. In this case, if you have enough data, I would try repeating your analysis but only on patients with a single psychiatric diagnosis (ie, major depressive disorder); but many patients will have enough diagnoses to make me cry
- Remember that all pilots should start with small scope, your overall dataset is small for NLP but covers a massive diagnostic space in psychiatry

If you want to be spicy, then we may consider the following:

- Look at patient outcome differences between patients seen by NPs (nurse practitioners) versus MD/DO (fully trained physician)
- There is a lot of "bad psychiatry" being practiced in the US because of NPs being significantly cheaper to hire - lot of existing data on poorer outcomes but not limited to psychiatry
- Can also look at telehealth vs. in-person outpatient outcomes I am assuming you can not obtain a dataset from another medical specialty; if possible, would look into datasets with more standardized vocabulary (if you are fixed on the use of NLP) - i.e., radiology, pathology, results of endoscopies, etc. For example, in radiology, it might be very nice to take an image's READ and try to predict the IMPRESSION (the radiologist's tldr), which would save some time, and you already essentially have the training data if you have a bunch of these

In conclusion, the invaluable insights provided by the psychiatric expert offer a significant opportunity to elevate the trajectory of future research focused on psychiatric patient datasets. The suggested approach not only sheds light on the complexities of mental health data but also presents a framework to enhance the quality and depth of these datasets.

By incorporating the guidance of an up and coming psychiatric, future research endeavors can adopt a more holistic and clinically informed perspective. The emphasis on capturing the subjective experiences of patients, alongside objective observations, serves to paint a comprehensive picture of their mental health journey. This nuanced approach not only humanizes the data but also enables a more accurate representation of the intricate interplay between symptoms, concerns, and treatment progress.

Furthermore, the call for standardized language and categorization aligns with the broader objective of creating cohesive and interoperable datasets. This standardization not only facilitates meaningful comparisons across different patient cases but also streamlines data integration, thus bolstering the potential for advanced analyses and insights.

However, these suggestions come with the recognition that the ethical dimension remains of paramount importance. Ensuring patient privacy, informed consent, and protection against biases must remain at the forefront of research endeavors. By weaving these considerations into the fabric of future studies, researchers can pave the way for responsible innovation in psychiatric healthcare data.

Incorporating the recommended approach into future research not only has the potential to advance our understanding of mental health but also contributes to the ongoing dialogue between the medical and research communities. The collaborative integration of clinical expertise and data-driven insights promises to yield datasets that are not only robust but also imbued with empathy-a vital combination in the pursuit of meaningful breakthroughs in psychiatric care.

Acknowledgement. We appreciate very much to Guanghua Wang for his help in collecting references.

References

1. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. In: Advances in Neural Information Processing Systems, vol. 33. NeurIPS (2020)
2. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT 2019, pp. 4171–4186. Minneapolis, Minnesota (2019)
3. Ein-Dor, L., et al.: Active learning for BERT: an empirical study. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7949–7962. Association for Computational Linguistics (2020)
4. Fadaee, M., Bisazza, A., Monz, C.: Data augmentation for low-resource neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pp. 567–573 Vancouver, Canada (2017)
5. Fang, M., Li, Y., Cohn, T.: Learning how to active learn: a deep reinforcement learning approach. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 595–605, Copenhagen, Denmark (2017)
6. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 1183–1192. PMLR (2017)
7. Gu, Y., Han, X., Liu, Z., Huang, M.: PPT: pre-trained prompt tuning for few-shot learning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, pp. 8410–8423 (2022)
8. Li, Z., Liu, X., Wong, D.F., Chao, L.S., Zhang, M.: ConsistTL: modeling consistency in transfer learning for low-resource neural machine translation. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 8383–8394 (2022)
9. Mikołajczyk, A., Grochowski, M.: Data augmentation for improving deep learning in image classification problem. In: 2018 International Interdisciplinary PhD Workshop (IIPhDW) (2018). https://doi.org/10.1109/IIPHDW.2018.8388338
10. Misra, I., van der Maaten, L.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6707–6717 (2020)
11. Pan, Y., Yao, T., Li, Y., Wang, Y., Ngo, C.-W., Mei, T.: Transferrable prototypical networks for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2239–2247 (2019)

12. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. Doha, Qatar (2014)
13. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog **1**, 9 (2019)
14. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 1–67 (2020)
15. Ravi, S., Larochelle, H.: Optimization as a Model for Few-Shot Learning, ICLR (2017)
16. Snell, J., Swersky, K., Zemel, R.,: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems, vol. 30. NIPS (2017)
17. Sun, Q., Liu, Y., Chua, T.-S., Schiele, B.: Meta-transfer learning for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 403–412 (2019)
18. Thoppilan, R., et al.: LaMDA: language models for dialog applications. arXiv:2201.08239 (2022)
19. Touvron, H., et al.: LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 (2023)
20. Wei, J., Zou, K.: EDA: easy data augmentation techniques for boosting performance on text classification tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 6382–6388. Hong Kong, China (2019)
21. Zoph, B., Yuret, D., May, J., Knight, K.: Transfer learning for low-resource neural machine translation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1568–1575. Austin, Texas (2016)