META-LEARNING OF TEXTUAL REPRESENTATIONS

Jorge G. Madrid, Hugo Jair Escalante
INAOE, Puebla, Mexico
jg.madridp@gmail.com, hugojair@inaoep.mx

ABSTRACT

Non-experts in Machine Learning research have an increasing demand for off the shelf systems in a wide number of applications, these are expected to perform at least as well as one build by a human expert. Some state-of-the-art methods are already available, nonetheless none of them concentrate on the challenges of Natural Language Processing. This work will provide a step towards an autonomous text supervised classification method applying meta-learning. The experiments presented in this report contemplate one of the fundamentals parts in the design of a text classification pipeline: the selection of the representation vector for a document given a text corpus.

1 Introduction

Recent machine learning success relies on the knowledge of human-experts that accordingly to their experience design and test multiple models extensively. However, in many domains there isn't always an expert available, hence the increasing demand for easy-to-use automated solutions. Text classification is one of the most studied tasks in Natural Language Processing (NLP), this is because of the number of applications that can be approached as text classification problems (e.g. sentiment analysis, topic labeling, spam detection, and author profiling). Many techniques have been developed over the last decades, despite all this progress an NLP expert is still needed for the determination of a classification pipeline that includes one or many of such techniques.

In this work we take a step towards the automated selection of the full pipeline for text classification by focusing on pre-processing methods. Unlike typical tabular data, language provides an unstructured and rich source of information, which features to extract is one of the most studied questions in NLP, the selection of an adequate representation is fundamental for successfully solving a classification task and usually has a greater impact than other parts of the pipeline such as the classification model. Our work is one of the first to approach text classification via meta-learning. Furthermore, we perform experiments of larger scale than previous work (in 81 publicly available corpora) and include more meta-data by extracting information not only from pipelines but also from the *raw* text.

2 Related Work

In the context of text mining, few work have explored the automated selection of different parts of classification pipelines. Similar to our work, Yogamata & Smith presented the first automated method for selecting a representation for text classification problems [1], in this work text representations are searched with Bayesian Optimization, their search space was very limited, using only word n-grams and tested with few datasets, nevertheless, they outperformed every linear classifier reported until their publication date. Other works have explored different meta-learning approaches for text classification in small-scales [2, 3].

On the other hand, meta-learning has been studied for a while in machine learning context [4, 5, 6, 7, 8]. But it is only recently that it has become a mainstream topic, this mainly because of its successes in several tasks. For instance, [9] successfully used a set of meta-features to warm-start a hyper-parameter optimization technique in the popular state-of-the-art AutoML solution *Autosklearn*. Similarly, for this work it is expected that our proposed set of meta-features from text corpora will provide an accurate description of them which could be used to perform an analogous process for finding a text representation in a determined search space.

3 Method

We present a method that takes as input the labeled raw text from a corpus and automatically selects a representation based on results from previous experiments, then a fixed model is used for classifying unseen documents. Our method comprises 2 stages, an offline phase where it *learns how to learn* and the predicting phase.

A human-expert uses knowledge acquired in the past when a new task is presented, equivalently, *meta-learning* imitates this reasoning. Our method applies meta-learning to learn from the performance of different representations on a number of corpora. Specifically, we defined 73 meta-features to characterize 81 text corpora and performed an exhaustive search for the performance of 60 representations. A *knowledge base* is built associating the performance of each representation with a task, described by the vector of meta-features. Traditionally, meta-features extract meta-data from a dataset such as statistics of its distribution or simple characteristics like the number of classes and attributes, in our proposed set we contemplate this type of features as well as other attributes extracted directly from the raw text, like those proposed in [10].

After the offline phase, for a new task the same meta-features are extracted and compared with the prior knowledge, we tested 4 strategies that leverage learned experiences and make predictions in a new task. (1) Using directly the representation with best performance of the *nearest* corpus, (2) predicting the representation as a classification problem, (3) predicting the performance for every representation, and (4) predicting the rank of each representation. After the representation is chosen an SVM classifier with linear kernel is used in every case to train and make predictions with the new corpus.

4 Experiments and results

Since we are exploring the effectiveness of the proposed meta-features and of 4 different meta-learning strategies the objective of these experiments is to get as close as possible to the optimal in our bounded search space of 60 representations, the whole process is done without any human intervention and with a simple *search* method (i.e. best first). We compare each strategy against the best representation found and with the average of all representations for each corpus. We evaluated each strategy in a leave-one-out setting, thus every corpus is treated as unseen by our method. Table 1 shows the average result for each strategy, Figure 1 compares the performance of 1 strategy in 9 selected corpora.

Method	Best	(1)	(2)	(3)	(4)	Mean
Avg Accu	0.77	0.74	**0.75**	0.74	**0.75**	0.68
Avg Rank	1.00	14.20	8.78	12.32	**8.31**	30.30
# of 1s	81	17	**28**	4	14	0

Table 1: Average accuracy [0,1] and average rank [60,1] of different strategies in 81 corpus, the last row indicates the number of times the best representation was predicted. (1) Nearest corpus, (2) classification, (3) performance regression, (4) rank prediction.

These 4 simple strategies clearly outperform a random representation and while in terms of average ranking they could be closer to the optimal, the average accuracy of (2) and (4) strategies was only 2% behind the best. (2) also found the best representation 35% of the time. Results show strong evidence that our meta-learning approach finds relations between corpora and pipeline performances that exploits prior knowledge for the autonomous classification of texts.

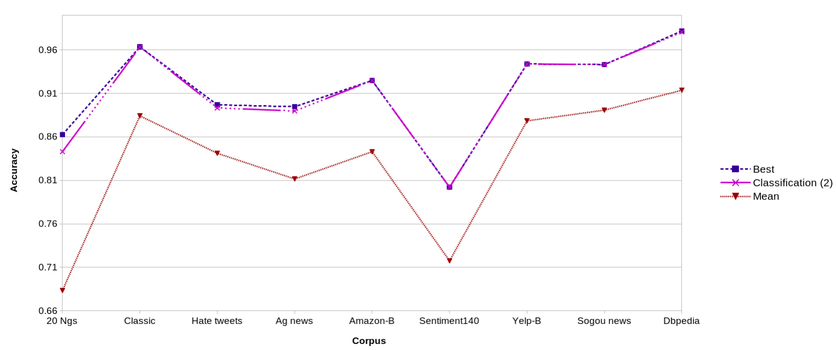

Figure 1: Accuracy of (2) in 9 selected corpora.

5 Conclusion and future work

We introduce a method that takes as input a corpus and without human intervention builds a model to solve a text classification task focusing on the selection of a vector representation and based mainly on meta-learning techniques. The results show empirically that with this approach we are able to characterize tasks and approximate an optimal representation. We will extend this work by *warm-starting* a state-of-the-art optimization algorithm with our proposed meta-learning setting allowing us to expand the search space and ideally finding pipelines that perform better than those designed by humans. We will compare our final method with state-of-the-art solutions for benchmark datasets along with other AutoML methods not focused on text classification.

References

[1] Dani Yogatama and Noah A Smith. Bayesian optimization of text representations. *arXiv preprint arXiv:1503.00693*, 2015.

[2] Wai Lam and Kwok-Yin Lai. A meta-learning approach for text categorization. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 303–309. ACM, 2001.

[3] Juan Carlos Gomez, Stijn Hoskens, and Marie-Francine Moens. Evolutionary learning of meta-rules for text classification. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 131–132. ACM, 2017.

[4] Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

[5] Yoshua Bengio. Gradient-based optimization of hyperparameters. *Neural computation*, 12(8):1889–1900, 2000.

[6] James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems*, pages 2546–2554, 2011.

[7] Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18(2):77–95, 2002.

[8] Joaquin Vanschoren. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018.

[9] Matthias Feurer, Jost Tobias Springenberg, and Frank Hutter. Initializing bayesian hyperparameter optimization via meta-learning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[10] David Pinto. On clustering and evaluation of narrow domain short-text corpora. *PhD. UPV*, 2008.