Wise Instance Selection Algorithms Help Reduce Annotation Work in Multi-Task Multi-Class Urinary Tract Dilation Prediction

[1,2] Yining Hua, MS; [1,3]Hsin-Hsiao Scott Wang*, MD; [1,4]Michael Lingzhi Li*, PhD
{[1]Boston Children's Hospital; [2]Harvard T.H. Chan School of Public Health; [3]Harvard Medical School; [4]Harvard Business School}, MA

INTRODUCTION

We explore methods to optimize annotation work efficiency for developing real-life clinical prediction models on the task of predicting urinary tract dilation (UTD) system classification using infant hydronephrosis ultrasound reports.

METHODS

Study Settings and Data Curation A retrospective analysis of Boston Children's Hospital urology radiology records identified a cohort of infants (n = 2479) aged 0-90 days with early ultrasounds. The research team analyzed the report and images to label 11 UTD classification outcomes as absence/presence of the UTD feature, or missing kidney.

Algorithm designs We study previously published as well as novel algorithms to aid in instance selection for training multitask and multiclass prediction models in a semi-supervised learning manner. In detail, we design (1) *Temperature Scaling + Entropy*: An optimal temperature parameter is calculated to scale the logits of the model after each round of fine-tuning using the logits and actual labels in the validation set. This temperature is used to scale prediction logits in the next iteration. The rest follows the Entropy method. (2) *K-Means + Entropy*: Unlabeled data points are transformed into feature vectors. The model computes their entropy score based on their task logits for each data point. K-Means clustering partitions the data into *k* clusters. Within each cluster, instances are ranked by their entropy scores. Top *n/k* instances with the highest entropy are chosen from each cluster for model retraining. In addition, we implement (3) *Least Confidence*, and (4) *Entropy* in a multitask and multiclass setting.

Experiment Settings We use the Bio_ClinicalBERT encoder (the classifier adds 11 linear layers to the encoder) and calculate the mean performance across ten runs for each task. For each experiment, 50 instances were selected per iteration for the first 14 iterations and 250 instances for the subsequent iterations. We compare the algorithms to (5) *Random* sampling.

RESULTS & DISCUSSION

All algorithms show improvements over more sample labels and plateaued after adding around 400-450 samples (Figure 1). The lower performance on the validation set compared to the test set for *Entropy*-based algorithms suggests that these algorithms are effectively pinpointing difficult instances for subsequent training iterations. All algorithms show more stabilized performance across ten experiments compared to *Random* sampling, showcasing the importance of carefully designed instance selection algorithms. *KMeans + Entropy* stands out as the top performer, likely due to its simultaneous focus on identifying the most challenging and different instances for subsequent model retraining, which simultaneously maximizes information uncertainty and density in the selected cases. As a subsequent step in our research, we intend to design more end-to-end solutions, such as multitask multiclass entropy. Our aim is to further enhance predictive performance while reducing the need for manual annotations. We will also explore how the number of instances selected for each round influences the effectiveness of the algorithms.

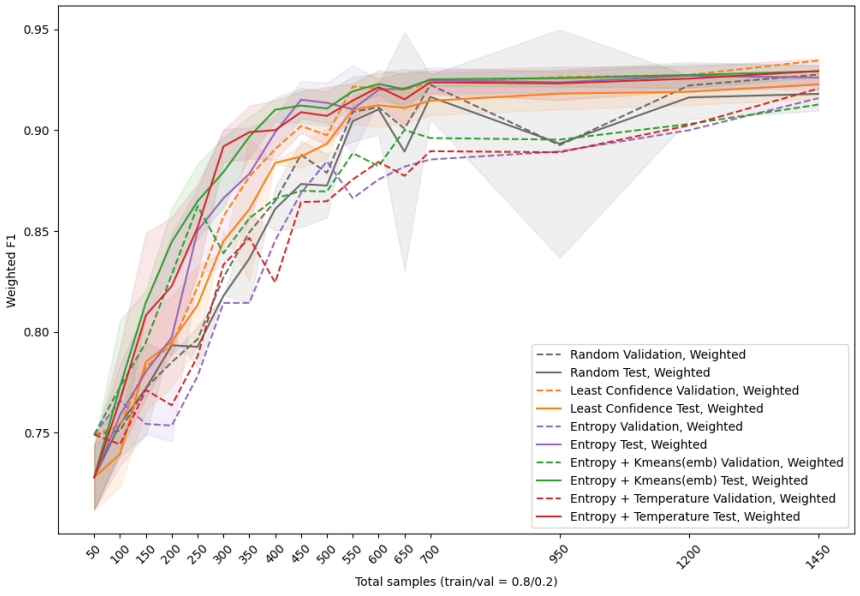

Fig. 1. Algorithm performances, trimmed to <=1450 samples (approximate convergence point of validation and test performance).

CONCLUSION

Thoughtfully crafted instance selection algorithms can significantly increase annotation work efficiency in multi-task, multi-class UTD prediction, yielding substantial performance improvements over random sampling. Incorporating these instance selection algorithms into real-world clinical prediction research, especially in contexts where labeling is resource-intensive, has the potential to streamline model development and enhance predictive performance.