AI Assisted Data Labeling with Interactive Auto Label

Michael Desmond, Michelle Brachman, Evelyn Duesterwald, Casey Dugan, Narendra Nath Joshi, Qian Pan, Carolina Spina

IBM Thomas J. Watson Research Center
1101 Kitchawan Rd,
Yorktown Heights,
NY 10598
mdesmond@us.ibm.com

Abstract

We demonstrate an AI assisted data labeling system which applies unsupervised and semi-supervised machine learning to facilitate accurate and efficient labeling of large data sets. Our system (1) applies representative data sampling and active learning in order to seed and maintain a semi-supervised learner that assists the human labeler (2) provides visual labeling assistance and optimizes labeling mechanics using predicted labels (3) seamlessly updates and learns from ongoing human labeling activity (4) captures and presents metrics that indicate the quality of labeling assistance, and (5) provides an interactive auto labeling interface to group, review and apply predicted labels in a scalable manner.

Introduction

Labeling data is an important but tedious activity that is often necessary during the development of supervised machine learning applications. It requires human labelers to examine data and assign labels in order to create training data for a machine learning task.

(Desmond et al. 2021) studied the effects of AI assistance on human labeling performance and found that labelers were significantly faster and more accurate when presented with label suggestions generated by a learner trained on previously labeled data. Based on observations from this study we designed a data labeling system which seamlessly integrates a learner into the labeling process. The learner assists the human labeler by predicting labels and optimizing the mechanics of the labeling task.

A notable feature of our system is 'Interactive Auto Labeling', a paradigm in which the human labeler can selectively apply predicted labels to slices of unlabeled data. The labeler groups and sorts unlabeled data from the perspective of the predicted labels and confidence estimates produced by the learner. The labeler can then review the data slice before deciding to apply the predicted labels. We believe this feature in particular can facilitate significant labeling productivity gains, and greatly improves the usability of auto labeling as a core feature of our tool.

System Design

Our system supports an end-to-end AI assisted labeling experience with interactive auto labeling support. The primary technical features are:

- **Semi-Supervised Learner:** AI assistance is implemented using a semi-supervised learning algorithm (Van Engelen and Hoos 2020), which learns from both labeled and unlabeled data, both of which co-exist in the data labeling context. Specifically the system applies label spreading (Zhou et al. 2004), a transductive semi-supervised algorithm that iteratively propagates label signals via an affinity graph representing the overall dataset.
- **Representative Sampling:** In order to optimally seed label spreading at the initialization of a labeling task, the system applies representative sampling. K-means partitioning is applied to the unlabeled data in order to select a set of exemplars which form an initial training set for the learner. The value of K is user defined, derived from the number of labels configured in the labeling project.
- **Active Learning:** The ongoing performance of the learner is optimized via the application of active learning (Settles 2009). The sequence of data presented to the human labeler is continually reordered using a min-margin active learning heuristic. The heuristic prioritizes the most informative or 'valuable' data from the perspective of the learner. This ensures optimal performance of the learner throughout the process. Label spreading is applied after each active learning batch is labeled.

User Experience

A seamless and engaging user experience was a priority when designing our labeling system.

- **AI Assistance:** Except for an initial bootstrapping phase, where representative examples are labeled to seed the learner, all human labeling is performed with AI assistance. When presented with data to label, the labeler is provided with a set of recommended labels as shown in Fig. 2. The entire label set is also 'decorated' using a predicted label distribution. Label recommendations and decorators help the labeler to make correct labeling decisions, and improve the overall efficiency of the labeling process.

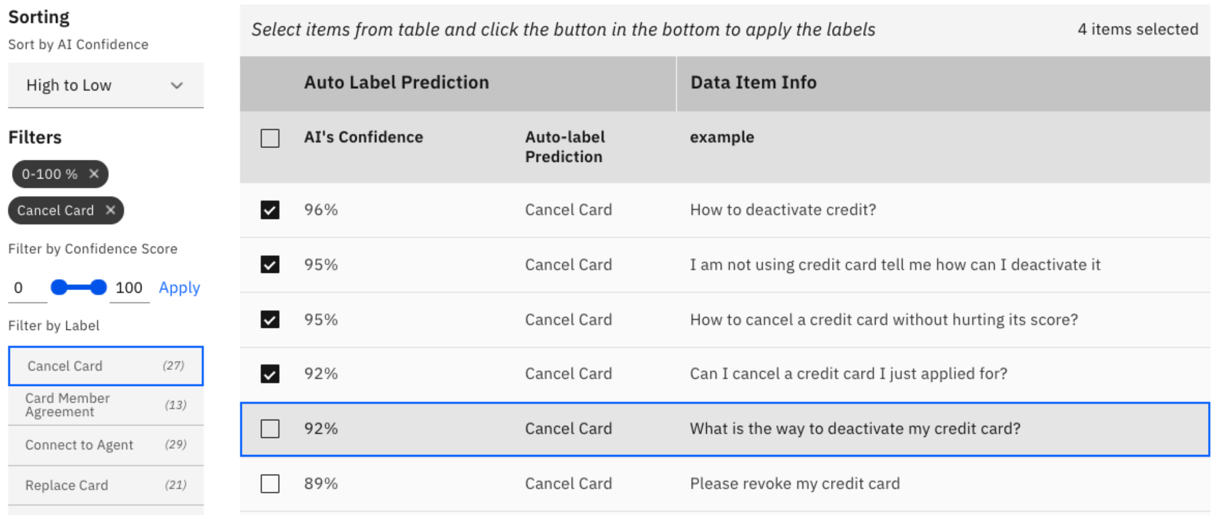

Figure 1: The Interactive Auto Labeling Interface: The labeler can group and sort unlabeled data based on the prediction and confidence of the learner. The labels can then be checked for correctness and applied all together.

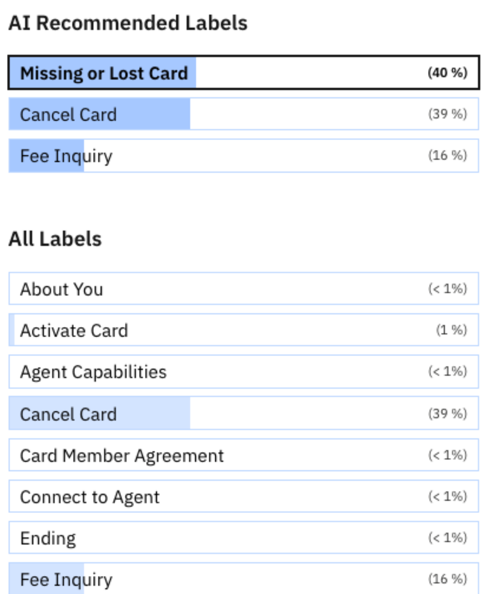

Figure 2: During labeling the system highlights the three most probable label predictions. The remaining labels are also decorated based on a predicted probability distribution.

- **Metrics & Assessment:** It is important for the labeler to understand the performance of the learner, particularly when auto labeling is an option. To aid in the understanding of the learner's performance, the labeler is presented with a set of metrics (Fig. 3). The agreement metric compares the top label predicted by the learner, with the label selected by the human labeler. Agreement is presented as a rolling average and may be considered as a proxy for the accuracy of the learner. The system also calculates the confidence of the learner on previously labeled data, and the learners prediction confidence on the remaining unlabeled set.
- **Interactive Auto labeling:** The interactive auto labeling interface, shown in Fig. 1, is where the labeler reviews predicted labels, and can apply them to unlabeled data as a batch. Data is grouped by predicted label and sorted by the learners confidence using an intuitive system of filters. Essentially the labeler can slice unlabeled data from the perspective of the learner, and auto apply predicted labels to each slice, thus greatly improving the efficiency of the labeling process.

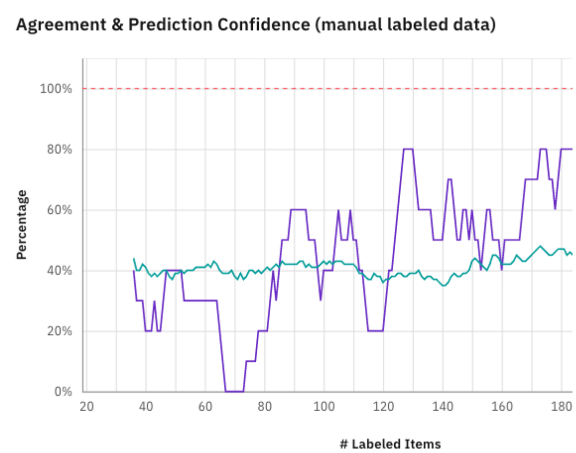

Figure 3: Agreement between the learner's predictions and labelers selected labels (purple) is tracked along with learner confidence (green).

Demonstration

The demonstration will showcase an end to end AI assisted labeling experience, with a particular focus on the interactive auto-labeling feature.

References

Desmond, M.; Muller, M.; Ashktorab, Z.; Dugan, C.; Duesterwald, E.; Brimijoin, K.; Finegan-Dollak, C.; Brachman, M.; Sharma, A.; Joshi, N. N.; et al. 2021. Increasing the Speed and Accuracy of Data Labeling Through an AI Assisted Interface. In *26th International Conference on Intelligent User Interfaces*, 392–401.

Settles, B. 2009. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Van Engelen, J. E.; and Hoos, H. H. 2020. A survey on semi-supervised learning. *Machine Learning*, 109(2): 373–440.

Zhou, D.; Bousquet, O.; Lal, T. N.; Weston, J.; and Schölkopf, B. 2004. Learning with local and global consistency. In *Advances in neural information processing systems*, 321–328.