Data Collection

——Data Labeling——

"Data Mining: Methods and Applications"

1

Contents

2

Data labeling

- **The goal of data labeling is to label individual examples.**
- The following categories for understanding the data labeling landscape.
  - *Use existing labels:* An early idea of data labeling is to exploit any labels that already exist.
  - *Crowd-based:* The next set of techniques are based on crowdsourcing.
  - *Weak labels:* While it is desirable to generate correct labels all the time, this process may be too expensive.
  - Each labeling approach can be further categorized as follows: *Machine learning task* and *Data type*.

3

Data labeling

A classification of data labeling techniques. Some of the techniques can be used for the same application. For example, for classification on graph data, both self-labeled techniques and label propagation can be used.

Category	Approach	Machine learning task	Data types	Techniques
Use Existing Labels	Self-labeled	classification	all	[92]–[96]
		regression	all	[97]–[99]
	Label propagation	classification	graph	[100]–[102]
Crowd-based	Active learning	classification	all	[103]–[109]
		regression	all	[110]
	Semi-supervised+Active learning	classification	text	[111], [112]
			image	[113]
			graph	[114]
	Crowdsourcing	classification	all	[50], [54], [115]–[122]
		regression	all	[123]
Weak supervision	Data programming	classification	all	[3], [124]–[127], [127]–[130]
	Fact extraction	classification	text	[131]–[142]

4

Data labeling

- **Utilizing existing labels**
  - *Classification* : For semi-supervised learning techniques for classification, the goal is to train a model that returns one of multiple possible classes for each example using labeled and unlabeled datasets.
  - *Regression* : Relatively less research has been done for semi-supervised learning for regression where the goal is to train a model that predicts a real number given an example.
  - *Graph-based Label Propagation* : Graph-based label propagation techniques also start with limited sets of labeled examples, but exploit the graph structure of examples based on their similarities to infer the labels of the remaining examples.

5

Data labeling

- **Utilizing existing labels**

Figure 1. Our structured labeling approach allows people to group data in whatever way makes sense to them. By seeing the resulting structure, people can gain a deeper understanding of the concept they are modeling. Here, the user sees an uncategorized page (top left) and can drag it to an existing group (right), or create a new group for it. The thumbnails (bottom left) show similar pages in the dataset to help the user gauge whether creating a new group is warranted.

6

Data labeling

- **Crowd-based techniques**
  - *Active Learning* : Active learning focuses on selecting the most "interesting" unlabeled examples to give to the crowd for labeling.
    - *Uncertain Examples* Uncertainty Sampling is the simplest in active learning and chooses the next unlabeled example that the model prediction is the most uncertain.
    - *Decision Theoretic Approaches* Another line of active learning performs decision-theoretic approaches. Decision theory is a framework for making decision under uncertainty using states and actions to optimize some objective function.

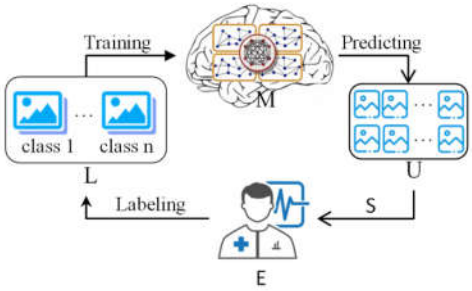


FIGURE 1. The active learning model in the method.

7

Data labeling

- *Regression* Active learning techniques can also be extended to regression problems.
- *Self and Active Learning Combined* The data labeling techniques we consider are complementary to each other and can be used together. A key observation is that the two techniques solve opposite problems where semi-supervised learning finds the predictions with the highest confidence and adds them to the labeled examples while active learning finds the predictions with the lowest confidence (using uncertainty sampling, query-by-committee, or density-weighted method) and sends them for manual labeling.

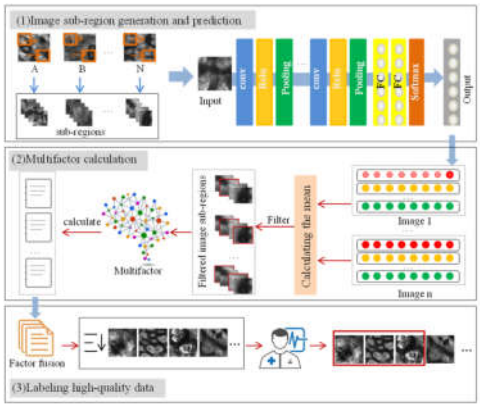


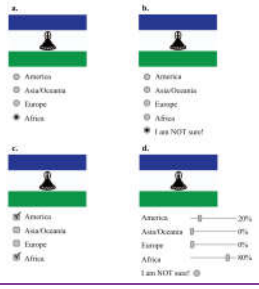
FIGURE 2. The processes that compose the method.

8

Data labeling

- **Crowd-based techniques**
  - *Crowdsourcing* : the crowdsourcing techniques here are more focused on running tasks with many workers who are not necessarily labeling experts.
    - *User Interaction* :A major challenge in user interaction is to effectively provide instructions to workers on how to perform the labeling.
    - *Quality control :* A simple way to ensure quality is to repeatedly label the same example using multiple workers and perhaps take a majority voting at the end.
    - *Scalability* : Scaling up crowdsourced labeling is another important challenge.

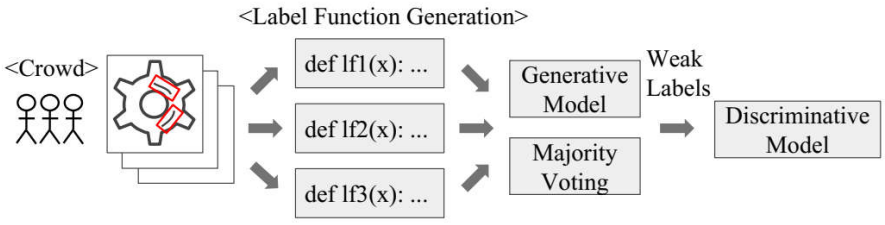
Fig. 1. Three crowd workers are hired to categorize the breed of a dog as Scotch, Yorkshire, or Australian. They express their single-option and Cumulative crowd labels using checked boxes and confidence bars, respectively. While the average score of Cumulative labels correctly indicates higher chance for Australian, the majority of single-option labels incorrectly suggests Yorkshire as the truth.



9

Data labeling

- **Weak Supervision** : it is mostly the case that there is not enough labeled data.
  - *Data Programming*: As data labeling at scale becomes more important especially for deep learning applications, data programming has been proposed as a solution for generating large amounts of weak labels using multiple labeling functions instead of individual labeling.



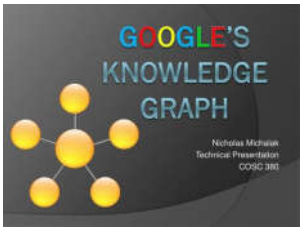
A workflow of using data programming for a smart factory application. In this scenario, Sally is using crowdsourcing to annotate defects on component images. Next, the annotations can be automatically converted to labeling functions. Then the labeling functions are combined either into a generative model or using majority voting. Finally, the combined model generates weak labels that are used to train a discriminative model.

10

Data labeling

- **Weak Supervision**
  - *Fact Extraction:* Another way to generate weak labels is to use fact extraction. Knowledge bases contain facts that are extracted from various sources including the Web. A fact could describe an attribute of an entity (e.g., Germany, *capital*, Berlin).
  - Knowledge Base: Freebase, Google Knowledge Graph and YAGO.

FREEBASE



YAGO

11

Contents

- **Background**
- **Data Acquisition**
- **Data Labeling**
- **Improvement of Existing Data and Models**
- **How to Decide which Data Collection Techniques to Use When**
- **Interesting Future Research Challenge**

12

Thanks !