Data Collection

——Improvement of Existing Data and Models——

"Data Mining: Methods and Applications"

1

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

2

Improvement of Existing Data and Models

- **Using Existing Data and Models**
- **Scenarios**
  - It may be difficult to find new datasets because the application is too novel or non-trivial for others to have produced datasets.
  - simply adding more data may not significantly improve the model’s accuracy anymore.
- Re-labeling or cleaning the existing data may be the faster way to increase the accuracy.

Task	Techniques
Improve Data	Data Cleaning [158]–[166]
	Re-labeling [119]
Improve Model	Robust Against Noise [167]–[171]
	Transfer Learning [172]–[178]

3

Improvement of Existing Data and Models

- **Improving Existing Data:** the data can be noisy and the labels may be incorrect.
  - **Data Cleaning:** To Introduce in Data Pre-process
  - **Re-labeling**
    - If the labels are noisy, then the model accuracy plateaus from some point and does not increase further, no matter how many more labeling is done.
    - Repeated labeling using workers of certain individual qualities can significantly improve model accuracy where a straightforward round robin approach(轮询调度算法) already give substantial improvements, and being more selective in labeling gives even better results.

4

Improvement of Existing Data and Models

- **Improving Models**
  - *Robust Against Noise and Bias* : there is a large number of noisy or even adversarial labels and a relatively smaller number of clean labels. Simply discarding the noisy labels will result in reduced training data, which is not desirable for complex models.
  - *Transfer Learning* : When there is not enough training data or time to train from scratch, a common technique is to start from an existing model that is well trained (also called a *source task*), one can incrementally train a new model (a *target task*) that already performs well.
    - AlexNet, VGGNet
    - TensorFlow Hub, AutoML
    - **Problems**: What to transfer, How to transfer, and When to transfer.

5

Contents

6

Thanks !