Data Collection

——Decision and Future Challenge——

"Data Mining: Methods and Applications"

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

How to Decide which Data Collection Techniques to Use When

- For one specific application scenario, data collection can be conducted.
- **Key Point 1**: It is not always easy to determine if there is enough data and labels.
- **Key Point 2**: How the labeling techniques tradeoff accuracy and scalability.

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**

Interesting Future Research Challenge

- **Data Evaluation:** how to evaluate whether the right data was collected with sufficient quantity.
- **Performance Tradeoff :** While traditional labeling techniques focus on accuracy, there is a recent push towards generating large amounts of weak labels. We need to better understand the tradeoffs of accuracy versus scalability to make informed decisions on which approach to use when.
- **Crowdsourcing** : Despite the many efforts in crowdsourcing, leveraging humans is still a non-trivial task.
- **Empirical comparison of techniques:** Although we showed a flowchart on when to use which techniques, it is far from complete, as many factors are application-specific and can only be determined by looking at the data and application.
- **Generalizing and integrating techniques:** We observed that many data collection techniques were application or data type specific and were often small parts of a larger research.

5

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**

6

References

- Yuji Roh, Geon Heo, Steven Euijong Whang，A Survey on Data Collection for Machine Learning（A Big Data - AI Integration Perspective），*IEEE Transactions on Knowledge and Data Engineering*，DOI 10.1109/TKDE.2019.2946162

Thanks !