Big Data and Lifelong Machine Learning
Table of Content

1. Value in Big Data
2. Lifelong Machine Learning
3. Transfer Learning
4. Human Computation
Characterizing the Big Data

Volume

Variety

Velocity

Veracity
Characterizing the Big Data

- Volume
- Variety
- Velocity
- Veracity

Value
Realizing the Value in Big Data

- Life-long Machine Learning
  - Can learning continue from last time? (never ending learning)
LML Examples

Never-Ending Language Learner [Tom Mitchell et al. 2010]

- Goal:
  - attempt to create a computer system that learns over time to read the Web (24x7, forever)
  - each day:
    - extract more facts from the web to populate the initial ontology, e.g.,
      - Brazil is a country
      - Poza Rica is a city located in the country Mexico
    - learn to read (perform #1) better than yesterday
LML Examples

Explanation-based neural networks [Thrun 1996]

- Input: a stream of supervised learning tasks

- Goal:
  - using the previous n-1 tasks to acquire a better inductive bias for the n-th task

![](image)

- mapping n-th task data into new space for further learning
- training a neural network from previous n-1 tasks
Case Study

- Lifelong Microblogging Agent @ Noah’s Ark Lab
- Weibo == Microblogging like Twitter
Weibo Software Agent

- Lifelong Information Assistant
  - Built on Social Media
  - Each user can have her own Weibo Agent
  - Each community can have their own Weibo Agent
  - Weibo Agent can understand user/community’s information needs
  - Weibo Agent can help user or community to collect, analyze, disseminate information
Main Features of Weibo Agent

- Features (basic)
  - Following People
  - Re-Tweeting (Forwarding Tweets)
  - Generating Simple Comments

- Features (advanced)
  - Learn to generate Tweets based on Analysis of Weibo Trends
  - Learn to Create Tweets based on Hot Topics in Other Sources
Recommending People and Tweets via User Models

- Model 1
- Model 2
- Model 3

New people
New tweets

Weibo HAL of User 1-2-3
Semantic Matching: Similar Tweets Have Similar Comments

Our paper entitled learning to rank has been accepted by ACL.

We are lucky. Our paper has been accepted by SIGIR this year. We are going to present it.

The PC of WSDM noticed us that our paper has been accepted.

Congratulations! It is a great achievement

Great news! Please accept my congrats!

Awesome! It is a great achievement
Weibo Agent in Social Media
Case: Noah LAB Weibo Agent

角色定义："数据挖掘领域的新兵"

能力：搜索相关热点价值内容和发现关键领袖进行Follow，转贴，并简单点评。

结果：虽然还只能简单回复，微软亚洲院统计把我们小诺作为第一粉进行宣传。

正如知名IT评论人易建所说：不公知不五毛不卖萌不传销不淘宝，搞到6万粉丝，还真是难啊。

角色定义：“数据挖掘领域的新兵”

能力：搜索相关热点价值内容和发现关键领袖进行Follow，转贴，并简单点评。

结果：虽然还只能简单回复，微软亚洲院统计把我们小诺作为第一粉进行宣传。
Foundations of Lifelong Machine Learning: Transfer Learning
When distributions are different

- Part-of-Speech tagging
- Named-Entity Recognition
- Classification
When Features are different

- Heterogeneous: different feature spaces

**Source: Text**

**Apple**

The apple is the pomaceous fruit of the apple tree, species *Malus domestica* in the rose family *Rosaceae* ...

**Banana**

Banana is the common name for a type of fruit and also the herbaceous plants of the genus *Musa* which produce this commonly eaten fruit ...

**Future: Images**
Lifelong Machine Learning Transfer Learning Evaluation
[Lisa Torrey and Jude Shavlik, 2009]

![Graph showing performance vs. training with and without transfer](image)

**Fig. 2.** Three ways in which transfer might improve learning.
HTL Setting: Text to Images

- Source data: labeled or unlabeled
- Target training data: labeled

Training: Text

Apple

The apple is the pomaceous fruit of the apple tree, species Malus domestica in the rose family Rosaceae ...

Banana

Banana is the common name for a type of fruit and also the herbaceous plants of the genus Musa which produce this commonly eaten fruit ...

Testing: Images
Heterogeneous Transfer Learning

Learn latent representation for auxiliary images
Using all source data

Target images

Projected representation of target images
Experiments: Accuracy vs. # text docs
Product Recommendation as Link Prediction

- Task: predict missing links in a network
- Focus:
  - bipartite graph of users and items
Transfer Learning for Collaborative Filtering?

IMDB Database

Amazon.com
Transfer Learning in CF [Bin Li et al., IJCAI 2009]

- **Source** (Dense): Encode cluster-level rating patterns
- **Target** (Sparse): Map users/items to the encoded prototypes

### MOVIES (Auxiliary - Dense)

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>?</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>?</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>?</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

### BOOKS (Target - Sparse)

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>?</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>?</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>?</td>
<td>3</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>?</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>?</td>
<td>?</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>?</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>?</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

### Permute rows & cols

### Reduce to Groups

Cluster-level Rating Pattern Matching

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>?</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>?</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>?</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>?</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>II</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>III</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>c</th>
<th>d</th>
<th>a</th>
<th>b</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>?</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>?</td>
<td>3</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>?</td>
<td>3</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>?</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>?</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>?</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>?</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>3</td>
<td>?</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

HCC 2013 2013 华为云计算大会
Human Computation [von Ahn 2005]

- Motivation
  - Payment (AMT)
  - Altruism (clickworkers)
  - Fun (ESP Game)
  - Socialization
  - By-product

- Algorithm
  - Voting & Rank

- System Platform
  - Crowd DB
  - Social media platform

Foldit: Predicting protein structures with a multi-player online game. [Cooper et al., 2009]
E-Pet: Activity Labeling via Mobile Game

Game on Client Side

- Activity Recognition
- Context Recognition
- Sensor Data Collection
- User Label Generation
- Q & A
- Photo
- Audio
- Text

Game Log

Internet

Integrated Game Log

General Models

- Sensor Data ↔ Human Labels
- User Model
- Context Model
- Activity Model
- Emotion Model

Server and Cloud Side

Model

Align
Conclusions

● Value in Big Data: most important
● Lifelong Machine Learning
  □ Transfer Learning
  □ Human Computation
Thank you

www.huawei.com