“Scalable Distributed Online Machine Learning Framework for Realtime Analysis of Big Data”

Hiroyuki MAKINO, NTT Software Innovation Center
Objective of Jubatus

- To satisfy “Scalable”, “Realtime”, and “Profound analysis” for Big Data.
Use Case: Social stream analysis

- Social stream analysis for marketing research
  - We want to know reputation or mood from their voice.
- Too hard to go over each tweet
  - We need machine learning to classify tweets automatically in realtime.

More than 8000 tweets/sec

Realtime twitter analysis with Jubatus

Classifies into 1600 companies

Demonstration
See you in the poster session
Performance

**The number of servers and throughput**

- Throughput and **linear scalability**
  
  Classification Tests: Classifying tweets into 1600 companies automatically
  
  - Throughput: Classifies all the Japanese tweets (6700 tweets/sec) with **1 server**

- Accuracy and learning time
  
  - Achieving the accuracy as good as batch processing based learning.
  
  - 90% accuracy in 1 sec. with 16 servers

---

*SPAM classification task (Dataset: LIBSVM webspam)*

* 200,000 features / approx. 30 features (per tweet) = 6,700 tweets/sec.
KEY POINT 1: Jubatus Architecture

- **User process**
  - Implemented using Jubatus client API (C++/Java/Python/PHP/Ruby clients)
  - Acquires input data and sends requests to servers via proxies

- **Proxy process (JubaKeeper)**
  - Relays clients’ requests to servers

- **Server process (JubaServer)**
  - Performs the training and prediction processing, and learning model synchronization
  - Increases performance linearly with the number of servers

- **ZooKeeper process**
  - Manages distributed coordination, such as process alive check and leader selection
KEY POINT 2: MIX technique

- Deterioration in the consistency of data synchronization is considered to be a loss of data to be trained, and it results in a decrease in accuracy.
- MIX technique can increase robustness by exchanging the intermediate results among servers loosely.

MIX technique

**MIX calculator**
[Calculation for aggregated result based on intermediate results]

**MIX Protocol controller**
[Regulating aggregated result]

**Membership manager**
[Leader election, members addition or deletion]
# Open Source Software

- **Jubatus OSS website**
  - [http://jubat.us](http://jubat.us)
  - Github [https://github.com/jubatus/jubatus](https://github.com/jubatus/jubatus)

---

<table>
<thead>
<tr>
<th>Supported machine learning engines</th>
<th>Description</th>
<th>Algorithms</th>
<th>Usecases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Linear classification</strong></td>
<td>Classify input data into given categories</td>
<td>Perceptron, Passive Aggressive (PA), Confidence Weighted Learning (CW), AROW, Normal HERD (NHERD)</td>
<td>Spam mail, Twitter classification</td>
</tr>
<tr>
<td><strong>Recommendation</strong></td>
<td>Recommend similar data as input data</td>
<td>Inverted Index, LSH, MinHash</td>
<td>Item recommendation, Advertisement</td>
</tr>
<tr>
<td><strong>Regression</strong></td>
<td>Estimate output value for input data</td>
<td>SVR using PA</td>
<td>Power consumption estimation, Stock price estimation</td>
</tr>
<tr>
<td><strong>Statistics</strong></td>
<td>Calculate statistics data, such as sum, max, min, average, standard deviation, etc</td>
<td></td>
<td>Sensor monitoring, Data anomaly detection</td>
</tr>
<tr>
<td><strong>Graph mining</strong></td>
<td>Extract a centrality and shortest path of the given graph structure</td>
<td>Centrality computation (PageRank), Shortest path search</td>
<td>Social community analysis, Network traffic analysis</td>
</tr>
</tbody>
</table>