Machine Learning for Big Data

Texts, Signals, Images and Video

Professor Konstantin Vorontsov

Moscow Institute of Physics and Technology

December 15, 2014
SDP: Machine Learning for Big Data (2014)

Structure of research group:

→ MIPT: K.Vorontsov, V.Strijov, …
→ MSU: D.Vetrov, A.Konushin, …

Activities

→ Research
→ Development
→ Education
→ Innovation

Projects of MIPT group:

→ Texts:
  → 1. BigARTM for Topic Modeling
  → 2. Conference Hierarchy Tool

→ Signals:
  → 3. ECG Multi-Disease Diagnostics
  → 4. Human Behavior Recognition
1. Topic Modeling for big text collections
1. Topic Modeling for big text collections

Metadata:
Authors Data Time Conference Organization URL etc.

Text documents

Topics of documents

doc1:

doc2:

doc3:

doc4:
...

Words and keyphrases of topics

Topic Modeling
1. Topic Modeling for big text collections

Metadata:
Authors
Data Time
Conference
Organization
URL
etc.

Text documents

Topics of documents

doc1:

doc2:

doc3:

doc4:
...

Words and keyphrases of topics

Images

December 15, 2014 | Page 5
1. Topic Modeling for big text collections

Metadata:
Authors
Data Time
Conference
Organization
URL
etc.

Text documents

Topics of documents

doc1:

doc2:

doc3:

doc4:

...

Words and keyphrases of topics

Images

Links

Topic Modeling

December 15, 2014 | Page 6
1. Topic Modeling for big text collections
1. Topic Modeling for big text collections
1. Topic Modeling for big text collections

Metadata:
- Authors
- Data Time
- Conference
- Organization
- URL etc.

Users

Text documents

Topics of documents
- doc1:
- doc2:
- doc3:
- doc4:
- ...

Words and keyphrases of topics

Topic Modeling
1. Topic Modeling: BigARTM project

- **Challenge:** how to combine many functionalities in a single Topic Model
- **Theory:** ARTM – Additive Regularization for Topic Modeling
  (easy to understand – easy to design – easy to infer – easy to combine)
- **Implementation:** BigARTM – open-source with permissive license
- **Experiments:** multi-criteria optimization of Topic Models
1. Topic Modeling: BigARTM project

- Open-source (http://bigartm.org)
- Parallel, Distributed,
- Online, Fast, Sparse, Robust,
- Multi-modal,
- Multi-criteria,
- Multi-language,
- Semi-supervised,
- Temporal,
- Hierarchical, etc…
1. Topic Modeling: BigARTM project

Performance testing on datasets

- Enron \((W=28\,102, \; D=39\,861)\)
- Nytimes \((W=102\,660, \; D = 300\,000)\)
- Pubmed \((W=141\,043, \; D = 8\,200\,000)\)
- Wiki \((W=100\,000, \; D = 3\,665\,223)\)

\[\text{Average of speedup}\]

Intel® Xeon® CPU E5-2630 v2 @ 2.60 GHz
(12 cores + hyper threading)
1. Topic Modeling: publications


2. Hierarchical Conference Topic Model

→ The goal:
  automatic construction of
  a scientific conference program

→ EURO – European Conference
  on Operational Research:
  >3500 participants,
  >200 experts,
  24 areas, 137 streams

---

[Diagram of hierarchical conference topic model]

EURO 2013

Areas

Streams

Sessions

Documents
2. Hierarchical Conference Topic Model

▶ Visualization of inconsistencies:

▶ Green points
  – consistent submissions

▶ Red points
  – inconsistent submissions
2. Hierarchical Conference Topic Model

→ http://EUROprogramAdvisor.com

Conference program validation for EURO/INFORMS abstract collection

Title: Additive Regularization of Topic Models for Topic Selection and Sparse Factorization

Abstract:
Probabilistic topic modeling of text collections is a powerful tool for statistical text analysis. Determining the optimal number of topics remains a challenging problem in topic modeling. We propose a simple entropy regularization for topic selection in terms of Additive Regularization of Topic Models (ARTM), a multi-criteria approach for combining regularizers. The entropy regularization gradually removes insignificant and linearly dependent topics. Changing regularization coefficients during iterations allows us to find a set of solutions with almost equal likelihood but different sparsity and number of topics. Thus optimization of the number of topics is an ill-posed problem and in general has many solutions. We propose additional criteria to choose the best of them.

Search results (page 1 of 15)

- Area: Continuous Optimization
- Stream: Nonconvex Programming: Local and Global
- Area: Continuous Optimization
- Stream: Convex Optimization
- Area: Continuous Optimization
- Stream: Variational Inequalities and Bi-Level Problems
- Area: Continuous Optimization
- Stream: Nonlinear Programming
- Area: Continuous Optimization
- Stream: Nonsmooth Optimization
- Area: Continuous Optimization
- Stream: Linear and Conic Programming
- Area: Continuous Optimization
- Stream: Vector and Set-Valued Optimization
2. Hierarchical Conference Topic Model


3. ECG processing for Multi-Disease Diagnostics

The Technology of Informational Analysis of ECG-signal

1. Measuring RR-interval and amplitude of each R-peak
2. Discretization
3. Vectorization
5. Estimation (Cross-Validation, Sensitivity-Specificity, AUC)
3. ECG processing for Multi-Disease Diagnostics

Variations of RR-intervals and R-amplitudes carry information about the functioning of not only the heart, but all the systems of the body, and can be used for the diagnosis at any stage of the disease [V.M.Uspenskiy, 2008]
3. ECG processing for Multi-Disease Diagnostics

The results of our cross-validation experiments

Data set:

- 20,000 ECGs
- 5-7 minutes each
- 60 Gb total size
- 40 diseases

<table>
<thead>
<tr>
<th>disease</th>
<th>cases</th>
<th>AUC, %</th>
<th>spec, % (sens=95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>femoral head necrosis</td>
<td>327</td>
<td>99.19 ± 0.10</td>
<td>96.6 ± 1.76</td>
</tr>
<tr>
<td>cholelithiasis</td>
<td>277</td>
<td>98.98 ± 0.23</td>
<td>94.4 ± 1.54</td>
</tr>
<tr>
<td>coronary heart disease</td>
<td>1262</td>
<td>97.98 ± 0.14</td>
<td>91.1 ± 1.86</td>
</tr>
<tr>
<td>gastritis</td>
<td>321</td>
<td>97.76 ± 0.11</td>
<td>88.3 ± 2.64</td>
</tr>
<tr>
<td>hypertensive disease</td>
<td>1891</td>
<td>96.76 ± 0.09</td>
<td>84.7 ± 1.99</td>
</tr>
<tr>
<td>diabetes</td>
<td>868</td>
<td>96.75 ± 0.19</td>
<td>85.3 ± 2.18</td>
</tr>
<tr>
<td>benign prostatic hyperplasia</td>
<td>257</td>
<td>96.49 ± 0.13</td>
<td>80.1 ± 3.19</td>
</tr>
<tr>
<td>cancer</td>
<td>525</td>
<td>96.49 ± 0.28</td>
<td>82.2 ± 2.38</td>
</tr>
<tr>
<td>nodular goiter thyroid</td>
<td>750</td>
<td>95.57 ± 0.16</td>
<td>73.5 ± 3.41</td>
</tr>
<tr>
<td>chronic cholecystitis</td>
<td>336</td>
<td>95.35 ± 0.12</td>
<td>74.8 ± 2.46</td>
</tr>
<tr>
<td>biliary dyskinesia</td>
<td>714</td>
<td>94.99 ± 0.16</td>
<td>70.3 ± 4.67</td>
</tr>
<tr>
<td>urolithiasis</td>
<td>649</td>
<td>94.99 ± 0.11</td>
<td>69.3 ± 2.14</td>
</tr>
<tr>
<td>peptic ulcer</td>
<td>779</td>
<td>94.62 ± 0.10</td>
<td>63.6 ± 2.55</td>
</tr>
</tbody>
</table>
3. ECG processing for Multi-Disease Diagnostics


4. Physical Activity and Behavior Recognition

- **Goal:** to reveal sudden changes in user behavior

- **Data:** multidimensional time series captured from a wearable device.

- **Solution:** automatic generation of a Deep Learning network
4. Physical Activity and Behavior Recognition

- Human motion tracker on mobile phone
- Wearable sensing system
4. Physical Activity and Behavior Recognition

- Time series segmentation for Jogging and Skipping
  - Manually
  - Automatically
4. Physical Activity and Behavior Recognition

Classification results

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>Jog</th>
<th>Walk</th>
<th>Up</th>
<th>Down</th>
<th>Sit</th>
<th>Stand</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jog</td>
<td>490</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.99</td>
</tr>
<tr>
<td>Walk</td>
<td>0</td>
<td>622</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0.99</td>
</tr>
<tr>
<td>Up</td>
<td>1</td>
<td>2</td>
<td>154</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0.95</td>
</tr>
<tr>
<td>Down</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>124</td>
<td>0</td>
<td>0</td>
<td>0.95</td>
</tr>
<tr>
<td>Sit</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>79</td>
<td>1</td>
<td>0.95</td>
</tr>
<tr>
<td>Stand</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>65</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Data set:
1M examples
200 points each
4. Physical Activity and Behavior Recognition


1. **Fedor Chervinskii.** EEG Classification

2. **Alvis Logins.** TOUCH: In-Memory Spatial Join by Hierarchical Data-Oriented Partitioning

3. **Rustem Feyzkhanov.** Email filters generator.

4. **Sergei Kasatkin.** Determination of the type of human activity based on the data from the accelerometer

5. **Ekaterina Kotenko, Alexandra Kudryashova.** NDVI calculation for satellite images

6. **Mikhail Matrosov.** Short-term forecasting of musical compositions.

7. **Roman Prilepskiy.** Text detection.

8. **Oleg Urzhumtsev.** Dictionary builder.

9. **Irina Zhelavskaya.** Automatic Filters Generator for Gmail.

10. **Sergey Voronov.** Topic model for filtering scientific papers.
Questions?

Contacts:

Moscow Institute of Physics and Technology

➔ Konstantin Vorontsov
e-mail: voron@forecsys.ru

➔ Vadim Strijov
e-mail: strijov@ccas.ru
URL: http://www.ccas.ru/strijov