Big data analytics and automated decision making in medicine and healthcare

Computer Aided Diagnosis

Dr. Oliver Faust

1 School of Science & Engineering
Habib University

2014
Outline

1. Introduction
   - Problem: Information overload
   - Solution: Big data analytics

2. Materials and methods
   - A methodology for data analytics
   - Conclusions
The case for big data

Science and observation
Analytical science observes nature and analyzes the data from these observations.

Medicine and biomedical engineering
Far from being an exception, biomedical data acquisition is on the forefront in the quest for more data. It strives towards extracting relevant information that could accurately diagnose a disease or predict its prognosis.
Big, as in more and better, data

More data sources
- Medical facilities are equipped with new scanners while older ones are still used in parallel.
- Mobile physiological monitoring equipment. For example, a mobile phone with a health application.

Better data quality
- Higher resolution.
- Specialized targeted applications.
- Combination of different signal acquisition techniques.
Example: X-ray evolution

Data size:

- A simple Digital Radiography (DR) such as a chest X-ray takes up about **30 MB** (Megabyte).
- In Computed Tomography (CT), a full study takes **more than one GB** (Gigabyte).
- PET-CT combines CT with about **5 MB** of Positron Emission Tomography (PET) data.
Medical data is **Big**

In 2008 it has been estimated that all kinds of medical images occupied 1250 Petabytes.

**Truly very big**

In comparison, the giant Internet search company Google used an estimated disk space of 20 to 200 Petabytes in August 2009.
Problem: Information overload

Medical practitioners face:

- Fatigue;
- Lag of attention;
- Optical illusions;
- Stress:
  - Time pressure;
  - Decision pressure;
  - Responsibility.
Solution: Big data analytics

Analytics for practitioner support
Highlight important parts of the data to support the practitioner.

+ Low system complexity;
+ Large margin of error – the brain is fault tolerant;
- Human decision making;
- Lag of traceability.

Computer aided diagnosis

+ Traceability;
+ Automation – riding Moors law;
- Small margin of error;
- High system complexity.
Diseases

Need definition
Detect symptoms of a disease in data.

These diseases can be:

- **Cardiovascular Disease** – The killer:
  - Heart attack;
  - Stroke.

- **Diabetes** – Paving the way for the killer diseases:
  - Diabetic retinopathy – A leading cause of blindness;
  - Diabetic foot – Walking becomes painful.

- **Epilepsy** – Thunderstorm in the brain.
What is the nature of the problem?

### Unsupervised

There is no previous knowledge on the data. A possible approach is:
- Find structure in the data;
- Characterize the structure.

### Supervised

Medical data, such as body temperature, is used for disease diagnosis since Milena.
- Feature extraction;
- Automated decision making.
Introduction

Materials and methods

A methodology for data analytics

Conclusions

Blockdiagram

Data

Offline:
- Pre-processing
- Feature extraction
- Feature assessment
- Offline classification
- Classification assessment

Online:
- Pre-processing
- Calculation of selected features
- Online classification

Best Features

Training Information
Data

- One dimensional signals:
  - Electrocardiogram [9];
  - Heart Rate [3, 17, 15, 18, 9, 8, 14];
  - Electroencephalogram [7, 10, 12, 10, 13, 6].

- Two dimensional signals:
  - Ultrasound [2, 4, 1, 5];
  - X-ray and mammography;
  - Thermography;
  - Fundus images of the retina [11, 16].
Fundus image with signs of Diabetes retinopathy

- Blood Vessels
- Microaneurysms
- Optic Disc
- Exudate
- Haemorrhage
- Macula
Ultrasound images of Carotid plaque

Carotid images. (a) Symptomatic (AF). (b) Asymptomatic (AS).
Data example plots: (a) left normal $s_n(n)$, (b) right alcohol $s_a(n)$.
A methodology for data analytics

Introduction

Materials and methods

Conclusions

Blockdiagram

Data

Offline

Pre-processing

Feature extraction

Feature assessment

Offline classification

Classification assessment

Pre-processing

Online

Calculation of selected features

Online classification

Best Features

Training Information

Computer Aided Diagnosis
Preprocessing

Unfiltered Normal ECG

Denoised Normal ECG

Unfiltered LBBB ECG

Denoised LBBB ECG
A methodology for data analytics

**Introduction**

**Materials and methods**

**A methodology for data analytics**

**Conclusions**

---

**Blockdiagram**

- **Offline**
  - Pre-processing
  - Feature extraction
  - Feature assessment
  - Offline classification
  - Classification assessment

- **Online**
  - Pre-processing
  - Calculation of selected features
  - Online classification

---

Dr. Faust

Computer Aided Diagnosis
The aim of feature extraction is to extract significant measures from the preprocessed physiological measurements. In general, there are two different types of feature extraction:

- Linear Features;
- Nonlinear Features.
Algorithms used for linear feature extraction:

- Statistical – Mean Variance;
- Fourier transform coefficients;
- Discrete wavelet coefficients;
- Entropy measures;
- Principal Component Analysis;
- Recurrence plots.
**Linear Feature example: Fourier transform**

![Graphs](image)

**Fig. 3.** Typical AR Spectrum graphs, (a) normal subject and (b) diabetic subject.
Nonlinear Features

Algorithms used for nonlinear feature extraction:

- Largest Lupanov Exponent (LLE);
- Correlation Dimension (CD);
- Hurst exponent (H);
- Approximate entropy (ApEn);
- Fractal dimension (FD);
- Phase space plot;
- Recurrence plots (RP).
Electroencephalogram of different sleep stages

(a) Result of recurrence plot at sleep 0 stage. (b) Phase space plot at sleep 0 stage.
Electroencephalogram of different sleep stages

(a) Result of recurrence plot at sleep 1 stage. (b) Phase space plot at sleep 1 stage.
Electroencephalogram of different sleep stages

(a) Result of recurrence plot at sleep 2 stage. (b) Phase space plot at sleep 2 stage.
Electroencephalogram of different sleep stages

(a) Result of recurrence plot at sleep 3 stage. (b) Phase space plot at sleep 3 stage.
Electroencephalogram of different sleep stages

(a) Result of recurrence plot at sleep 4 stage. (b) Phase space plot at sleep 4 stage.
Electroencephalogram of different sleep stages

(a) Result of recurrence plot at sleep 5 stage. (b) Phase space plot at sleep 5 stage.
A methodology for data analytics

**Introduction**

**Materials and methods**

**Conclusions**

**Blockdiagram**

**Data**

**Offline**

- Pre-processing
- Feature extraction
- Feature assessment
- Offline classification
- Classification assessment

**Online**

- Pre-processing
- Calculation of selected features
- Online classification

**Best Features**

**Training Information**

Dr. Faust Computer Aided Diagnosis
Analysis Of Variance (ANOVA)

ANOVA test yields the so called \textit{p-value}, which provide a statistical measure of the feature quality.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Sleep 0</th>
<th>Sleep 1</th>
<th>Sleep 2</th>
<th>Sleep 3</th>
<th>Sleep 4</th>
<th>Sleep 5</th>
<th>‘p’-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>FD</td>
<td>$-1.45 \pm 0.19$</td>
<td>$-1.60 \pm 0.21$</td>
<td>$-1.39 \pm 0.20$</td>
<td>$-1.74 \pm 0.11$</td>
<td>$-1.81 \pm 0.05$</td>
<td>$-1.55 \pm 0.20$</td>
<td>0.000</td>
</tr>
<tr>
<td>ApEn</td>
<td>$0.97 \pm 0.40$</td>
<td>$0.68 \pm 0.57$</td>
<td>$0.91 \pm 0.47$</td>
<td>$0.45 \pm 0.53$</td>
<td>$0.28 \pm 0.41$</td>
<td>$0.83 \pm 0.51$</td>
<td>0.000</td>
</tr>
<tr>
<td>CD</td>
<td>$6.03 \pm 0.86$</td>
<td>$6.34 \pm 0.60$</td>
<td>$5.96 \pm 0.67$</td>
<td>$6.29 \pm 0.67$</td>
<td>$6.42 \pm 0.54$</td>
<td>$6.30 \pm 0.59$</td>
<td>0.000</td>
</tr>
<tr>
<td>LLE</td>
<td>$0.91 \pm 0.34$</td>
<td>$0.95 \pm 0.21$</td>
<td>$0.84 \pm 0.31$</td>
<td>$1.10 \pm 0.30$</td>
<td>$0.98 \pm 0.28$</td>
<td>$1.02 \pm 0.41$</td>
<td>0.063</td>
</tr>
<tr>
<td>$H$</td>
<td>$0.51 \pm 0.20$</td>
<td>$0.39 \pm 0.19$</td>
<td>$0.58 \pm 0.24$</td>
<td>$0.25 \pm 0.07$</td>
<td>$0.23 \pm 0.07$</td>
<td>$0.43 \pm 0.23$</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Introduction

Materials and methods

A methodology for data analytics

Conclusions

Blockdiagram

Data

Offline

Pre-processing

Feature extraction

Feature assessment

Offline classification

Classification assessment

Online

Pre-processing

Calculation of selected features

Online classification

Best Features

Training Information

Dr. Faust

Computer Aided Diagnosis
Classification

The aim of classification is to obtain an objective decision on the underlying data. For computer aided diagnosis systems, this objective decision is concerned with whether or not the underlying data comes from a diseased person, and if a disease was detected a corollary decision is needed on the nature of the disease.

- Support Vector Machine (SVM);
- ADAboost (Adaptive Boosting);
- Decision Tree (DT);
- K-Nearest Neighbour (KNN);
- Artificial Neural Network (ANN);
- Probabilistic Neural Network (PNN).
Classification assessment

This is the most important classification criteria, because it determines the quality of the analysis system.

- Accuracy;
- Sensitivity;
- Specificity;
- Positive Predictive Value;
- Confusion Matrix;
- Receiver operating characteristic.
### Cross validation electrocardiogram example

#### Table III. Ten fold cross-validation results for the DT classifier.

<table>
<thead>
<tr>
<th>Fold</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold 1</td>
<td>0.9596</td>
<td>0.9861</td>
<td>0.9373</td>
<td>0.9640</td>
</tr>
<tr>
<td>Fold 2</td>
<td>0.9632</td>
<td>0.9838</td>
<td>0.9286</td>
<td>0.9630</td>
</tr>
<tr>
<td>Fold 3</td>
<td>0.9632</td>
<td>0.9840</td>
<td>0.9292</td>
<td>0.9632</td>
</tr>
<tr>
<td>Fold 4</td>
<td>0.9601</td>
<td>0.9833</td>
<td>0.9271</td>
<td>0.9638</td>
</tr>
<tr>
<td>Fold 5</td>
<td>0.9532</td>
<td>0.9831</td>
<td>0.9239</td>
<td>0.9625</td>
</tr>
<tr>
<td>Fold 6</td>
<td>0.9671</td>
<td>0.9844</td>
<td>0.9314</td>
<td>0.9658</td>
</tr>
<tr>
<td>Fold 7</td>
<td>0.9640</td>
<td>0.9822</td>
<td>0.9217</td>
<td>0.9629</td>
</tr>
<tr>
<td>Fold 8</td>
<td>0.9642</td>
<td>0.9855</td>
<td>0.9361</td>
<td>0.9642</td>
</tr>
<tr>
<td>Fold 9</td>
<td>0.9601</td>
<td>0.9844</td>
<td>0.9308</td>
<td>0.9648</td>
</tr>
<tr>
<td>Fold 10</td>
<td>0.9744</td>
<td>0.9867</td>
<td>0.9414</td>
<td>0.9700</td>
</tr>
<tr>
<td>Average</td>
<td>0.96291</td>
<td>0.98435</td>
<td>0.93075</td>
<td>0.96442</td>
</tr>
</tbody>
</table>
Conclusion, off-line system

Find a suitable algorithm structure through off-line processing and modeling.

- Know what you are looking for – Define the need;
- Make the data accessible through preprocessing;
- Extract features, many more than needed;
- Test the features and select the best features;
- Test automated decision algorithms, select the best algorithm.
Conclusion, on-line system

Implement the on-line system to benefit the target audience.

- Clean the data through pre-processing;
- Extract the best features;
- Employ the most accurate classification algorithm to get the best decision support;
- Visualize the decision support result;
- Distribute the decision support result.


References VIII


Thesis

Data volume is nothing without analytics.

Questions? Comments?