Big Data Analytics and Healthcare

Anup Kumar, Professor and Director of MINDS Lab
Computer Engineering and Computer Science Department
University of Louisville
Road Map

• *Introduction*

• Data Sources
  – Structured EHR data
  – Unstructured EHR data

• Data Analytics Approaches
  • Processing of Structured data
  • Processing of Unstructured data

• Example Applications

• Conclusions
Big Data Applications

• Advertising and marketing
  – Customer shopping patterns
  – Response to promotional campaign
• Manufacturing
  – Maintenance of machine health
• Social Media
  – Browsing and sentiment analysis
  – Impact on buying patterns
• Email
  – Communication and interaction patterns
  – Influencing the product perception
• Government data
  – Efficient process management
Big Data Applications (cont’d)

• Stock Market
  – Stock performance prediction

• Healthcare Management
  – Patient health monitoring
  – Impact of preventive care

• Financial Institutions
  – Fraud detection and mitigation

• Weather
  – Prediction
  – Impact analysis and better disaster management
Why Big Data Analytics Now?

• Volume
  – Data generated by 2020 will be in Zettabytes \((10^{21})\)
  – Soon after that the measure will be Yottabyte \((10^{24})\) and Brontabyte \((10^{27})\)

• Variety
  – Structured (transactional data)
  – Unstructured (image, video and text data)

• Velocity
  – Rate of data generation
  – Increase in the ability to process data

• Variability/Veracity
  – Trustworthiness of data
  – Quality of data
Big Data Analytics: Benefits

• Getting to know your patient better
  – Targeted medicine
  – Accurate diagnosis
  – Predicting disease onset
• Saving money
  – Fraud detection in medical industry
  – Risk Management
  – Lower cost and better outcomes
• Real-time decision making
  – Sensor data analysis
  – Influence of social sentiment on patient health
Data Scientist: A Challenging Combination of Backgrounds

- Math and Statistics Background
- Domain Expertise
- Programming Skills
Big Data Analytics

- Big Data Analytics = Big Data + Advanced Analytics
- Advanced Analytics includes:
  - Association Rules
  - Classification and decision trees
  - Text analytics
  - Clustering
  - Regression
  - Machine learning
  - Etc....
- Deployments of Analytics
  - MapReduced / Hadoop Based Deployment
  - In-Database Deployment
- Data Analytics is applicable to any size of data
Road Map

• Introduction
• **Data Sources**
  – Structured EHR data
  – Unstructured EHR data
• Data Analytics Approaches
  • Processing of Structured data
  • Processing of Unstructured data
• Example Applications
• Conclusions
Healthcare Data Types

- EHR
- Public Health
- Social Behavioral
- Public Health
Healthcare Data

• Billing Data
  – International Classification of Diseases (ICD)

• Lab results
  – Logical Observation Identifiers Names and Codes (LOINC)

• Medication
  – National Drug Code (NDC) by Food and Drug Administration (FDA)
Healthcare Data (Cont’d)

• Clinical notes
  – Unstructured text data

• Image Data
  – Unstructured data

• Social Interaction data
  – Unstructured data
Heritage Health Prize

Improve Healthcare, Win $3,000,000.

Identify patients who will be admitted to a hospital within the next year using historical claims data. (Enter by 06:59:59 UTC Oct 4 2012)
GE Head Health Challenge
GE Challenge I

Challenge I Award

GE and the NFL will be awarding up to $10 million for two types of solutions: Algorithms and Analytical Tools, and Biomarkers and other technologies.
Road Map

- Introduction
- Data Sources
  - Structured EHR data
  - Unstructured EHR data
- Data Analytics Approaches
  - Processing of Structured data
  - Processing of Unstructured data
- Example Applications
- Conclusions
Analytic Platform

Structured EHR

Unstructured EHR

Patients / Context Feature Selection

Clustering

Classification

Recommendation

MINDS
University Of Louisville
CECS Department
Mobile Information Networks
and Distributed Systems Lab
Types of Data Analysis

• Descriptive Analytics (traditional Business Intelligence):
  – Specifies the data characteristics
  – Also known as unsupervised learning
    • How to describe the system?
    • What happened in the system and when?
    • What are the parameters in the systems?
    • What is the impact of a parameter on the system?
    • Is there any co-relation between the parameters?

• Predictive Analytics: Uses data mining and predictive modeling
  – Also know as supervised learning
    • What are the future trends?
    • What is the decision based on past history?
    • Perform what if analysis.
Implementation Options

- In-Database Analytics
- Distributed Analytics
  - Cluster and cloud computing based
  - Hadoop / MapReduce based
In-Database Analytics

• Allows analytic computation to be carried out in the database
  – Uses SQL and
  – SQL extensions

• Advantages
  – Computation is close to data and does not require data movement
  – Analytic centralization may allow easy security, data and version management
  – Client access to in-database analytics is easy
  – Higher analytics efficiency, easy usability, better database manageability

• Disadvantages
  – Vendor dependent
  – Limited data type support in databases
    • Cannot run location dependent analytics or Text analytics
  – Cost of analytics
Distributed Analytics

• Motivation
  – Large data size
  – Complex computation logic
  – Real time processing requirement
  – Cheaper hardware
  – Larger and faster storage space

• Challenges
  – How to divide and distribute data?
  – How to divide the algorithm?
  – How to manage distributed resources?

• Limitations
  – Many problems are not suitable for distributed computing
    • Sequential algorithms
    • For example, computing Fibonacci sequence
Cluster and Cloud Computing

Benefits

• Availability of large computing resource
• Huge storage space availability
• Easy distribution of data and computation logic
• Availability of more flexible distributed programming paradigm
  – MapReduce
• Effective implementation of MapReduce
  – Hadoop
  – R-Hadoop
Hadoop Based Analytics

• Allows analytics to be carried out on any type of data store
• Provides standard framework for computation
  – On cloud environment
  – On in-premises network
• Advantages
  – Vendor independent
  – Allows any type of analytics to be carried out
  – Provides flexible and adaptable architecture for implementation
• Disadvantages
  – Complex implementation
MapReduce

• Allows use of large computational resources
  – Inter cluster communication is managed by MapReduce

• MapReduce architecture supports
  – Data and task distribution
  – Fault monitoring
  – Task and data replication
  – Simple programming model

• Limitations of MapReduce
  – Cannot solve all the problems
MapReduce: A Pragmatic Approach

• It can solve many Big Data problems
  – Data filtering
  – Statistics and aggregation
  – Graph analytics
  – Decision Tree and classification
  – Clustering and recommendations
• Practical Distributed API
  – Easier to understand and use
• Higher level APIs exist
  – To reduce the complexity of programming
  – Ability to schedule multi-stage jobs
Phases in Hadoop Processing

- Data Processing with Hadoop goes through three phases
  - Map Phase
    - Processes the data and generate <key, Value> pairs
  - Shuffle Phase
    - Moves the data <key, value> pairs to appropriate processing node for reduction
  - Reduce Phase
    - Processes the data <key, value> pairs to generate final output
- Hadoop can use multiple machines for each phase
Association Rule Example

- In order to compute support
  - The number of times each product and its combinations occur in the data has to be calculated
- The original transaction file format
  - 1, milk, bread
  - 2, bread, butter
  - 3, milk, bread, butter
  - 4, milk

<table>
<thead>
<tr>
<th>Transaction ID</th>
<th>Milk (M)</th>
<th>Bread (B)</th>
<th>Butter (T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Association Rule MapReduce

Input Splitting Mapping Shuffling Reducing

M B MB
B T BT
M B T MB MT BT MBT
M

M B MB
B T BT
M B T MB MT BT MBT
M

B T BT

M,1 B,1 MB,1
B,1 T,1 BT,1
M,1 B,1 T,1 MB,1 BT,1
M,1 B,1 T,1 MB,1 BT,1 MT,1 MBT,1
M,1 B,1 T,1 MB,1 BT,1 MT,1 MBT,1

M,1 M,1 M,1
B,1 B,1 B,1
T,1 T,1
MB,1 MB,1
BT,1 BT,1
MT,1 MT,1
BMT,1

M,3 B,3 T,2 MB,2 BT,2 MT,2 BMT,1

M,1

M,1

M,1
Road Map

• Introduction
• Data Sources
  – Structured EHR data
  – Unstructured EHR data
• Data Analytics Approaches
  • *Processing of Structured data*
  • Processing of Unstructured data
• Example Applications
• Conclusions
Steps in Structured Data Analytics

- Step 1: Business domain analysis
- Step 2: Data exploration and investigation
- Step 3: Data preparation and cleaning
- Step 4: Model design and development
- Step 5: Model verification and testing
- Step 6: Analyze the output
Application for Recommendation Framework

- **Medicine**
  - Disease recommendation
  - Drug recommendation
  - Case based search

- **Marketing**
  - Cell phone companies for identifying users that may switch
  - Recommending books at Amazon
  - Recommending products on the web sites

- **Education**
  - Universities guiding students what courses to take
  - Conference organizers assigning papers to reviewers
• Recommender can use the following rule:
Road Map

• Introduction
• Data Sources
  – Structured EHR data
  – Unstructured EHR data
• Data Analytics Approaches
  • Processing of Structured data
  • Processing of Unstructured data
• Example Applications
• Conclusions
Challenges in Text Mining

- Each document text may contain large amounts of text
  - High dimensionality
  - Difficult to identify which part is important to a pattern
- Ambiguity of content due to language features
- Sematic issues
  - Words and phrases may not be semantically independent
  - May have subtle and complex relations between concepts in text
- Complexity of natural language processing
- Processing large training set
General Steps in Text Mining

• Text Splitting
  – Split text into bag of words using text tokenization
  – Disadvantage: often loses semantic meaning
• Text Preprocessing
  – Removal of numbers
  – Removal of punctuation marks
  – Text case conversion as needed
• Feature selection
  – Determine nGrams necessary
  – Stop word removal (can use pre-specified list or generic list)
  – Stemming (identify word by its root)
General Steps in Text Mining (Cont’d)

• Determining the weighting of individual words
  – Term Frequency-Inverse Document Frequency (TF-IDF)
• Creating a Term Document Matrix
  – Terms and frequency of each word in a document
    • Simple TD
    • TF-IDF
    • Latent Semantic Indexing matrix
Stop Words

- Most common words in English that do not contribute to classification, clustering or association are:
  - Articles – a, an, the
  - Conjunctions – and, or...
  - Prepositions – as, by, of ...
  - Pronouns – you, she, he, it ...
- Text documents are high-dimension data
  - Removal of stop words acts as technique for dimensionality reduction
- Other non-context related words can also be removed
Stemming

• The process for reducing inflected (or sometimes derived) words to their stem, base or root form
  – Typically achieved by removing – ing, - s, -er -ed etc.
  – For example: “mining”, “miner”, “mines”, “mined”
  – Stemmed word “mine”

• Common Algorithms are
  – Porter’s Algorithm
  – KSTEM Algorithm
  – Snowball Stemming
Steps in Association Mining

- Loading the data
- Text preprocessing (as needed)
  - Cleaning
  - Punctuation removal
  - Number removal
  - Stop word removal
  - Stemming
- Building term document matrix
- Finding frequent term association
Road Map

• Introduction
• Data Sources
  – Structured EHR data
  – Unstructured EHR data
• Data Analytics Approaches
  • Processing of Structured data
  • Processing of Unstructured data
• *Example Applications*
• Conclusions
Impact of Data Driven Features

Figure 2: AUC significantly improves as complementary data driven risk factors are added into existing knowledge based risk factors. A significant AUC increase occurs when we add first 50 data driven features.

Applications of Patient Similarity

- Heart Failure Prediction
- Likelihood of Diabetic onset
- Disease recommendation
- Medicine recommendation
Medical Imaging

Analysis Outcomes

• Modality Classification
• Image-based Retrieval
• Case-based Retrieval
Modality Classification

- Compound or multipane images
  - Radiology
  - Ultrasound
  - Magnetic Resonance
  - Computerized Tomography
  - X-Ray, 2D Radiography
  - Angiography
  - PET
  - Combined modalities in one image
  - Visible light photography
    - Dermatology, skin
    - Endoscopy
    - Other organs

- Diagnostic images
  - Printed signals, waves
    - Electromyography
    - Electrocardiography
    - Electroencephalography
  - Microscopy
    - Light microscopy
    - Electron microscopy
    - Transmission microscopy
    - Fluorescence microscopy
    - 3D reconstructions

- Generic biomedical illustrations
  - Tables and forms
  - Program listing
  - Statistical figures, graphs, charts
  - Screenshots
  - Flowcharts
  - System overviews
  - Gene sequence
  - Chromatography, Gel
  - Chemical structure
  - Mathematics, formulae
  - Non-clinical photos
  - Hand-drawn sketches
Image Query

• Image-based Retrieval
  – Given a query image and find the most similar images

• Case-based Retrieval
  – Given a case description, details of the symptoms, tests including images
  – Find similar cases including images with case descriptions
Genetic Data

• Human genome is composed of DNA with four building blocks
  – A, T, C, G
• Contains three billion pairs of bases of A, T, C, G
• Size of human genome is 3GB
Genome Wide Association Studies (GWAS)

- Identifying common genetic factors that influence health and diseases
- Compare DNA of patients with disease and similar people without disease
- Single nucleotide polymorphisms (SNPs) are DNA sequence variations that occurs when a single nucleotide in genome sequence differs between individuals
Epidemiology Data

• Source
  – Surveillance Epidemiology and End Results (SEER) Program at NIH

• Usage
  – Understanding the disparity in diseases related to race, age and gender
  – Information correlation with other data sources such as pollution, climate and socio economic
  – Can use predictive analysis for various disease trends
Social Networks for Patients

- PatientsLikeMe
  - Has more than 200,000 patients and is tracking more than 1,600 diseases
  - www.patientslikeme.com
Big Data Analytics: Barriers

- Cost of analytics
- Lack of skilled talent
- Difficult to architect a Big Data Solutions
- Big data scalability issues
- Limited capability of existing database analytics
- Tangible business justification
- Lack of understanding of Big Data benefits
Concluding Remarks

• Better diagnosis
• Better health care delivery
• Better value for patient, provider and payer
• Better innovation
• Better living