Data Mining and Medical Informatics

Credit: R. E. Abdel-Aal

Contents

• Introduction to Data Mining: Definition, Functions, Scope, and Techniques
• Data-based Predictive Modeling
  Neural and Abductive Networks
• Data Mining in Medicine
  Motivation and Applications
• Experience at KFUPM
• Summary
The Data Overload Problem

- Amount of data doubles every 18 months!:
  - NASA’s Earth Orbiting System sends 4,000,000,000,000 bytes a day
  - One fingerprint image library contains 200,000,000,000,000 bytes

- Data warehouses, data marts, ... of historical data
- The hidden information and knowledge in these mountains of data are really the most useful
- “Drowning in data but starving for knowledge”? "Siftware"

The Data Pyramid

- Data
- Information (Data + context)
- Knowledge (Information + rules)
- Wisdom (Knowledge + experience)

How many units were sold of each product line?
What was the lowest selling product?
What made it that unsuccessful?
How can we improve it?
What is wrong with conventional statistical methods?

- Manual hypothesis testing:
  Not practical with large numbers of variables
- User-driven... User specifies variables, functional form and type of interaction:
  User intervention may influence resulting models
- Assumptions on linearity, probability distribution, etc. May not be valid
- Datasets collected with statistical analysis in mind Not always the case in practice

Recent advances in computers made data mining practical

- Cheaper, larger, and faster disk storage:
  You can now put all your large database on disk
- Cheaper, larger, and faster memory:
  You may even be able to accommodate it all in memory
- Cheaper, more capable, and faster processors:
- Parallel computing architectures:
  Operate on large datasets in reasonable time
  Try exhaustive searches and brute force solutions
Data Mining: Some Definitions

- Knowledge Discovery in Databases (KDD)

- The use of tools to extract ‘nuggets’ of useful information & patterns in bodies of data for use in decision support and estimation

- The automated extraction of hidden predictive information from (large) databases

Data Mining Functions

- Clustering into ‘natural’ groups (unsupervised)
- Classification into known classes; e.g. diagnosis (supervised)
- Detection of associations; e.g. in basket analysis: ”70% of customers buying bread also buy milk”
- Detection of sequential temporal patterns; e.g. disease development
- Prediction or estimation of an outcome
- Time series forecasting
Data Mining Scope

- **Finance and business:**
  - Loan assessment, Fraud detection, Market forecasting
  - Basket analysis, Product targeting, Efficient mailing
- **Engineering:**
  - Process modeling and optimization
  - Machine diagnostics, Predictive maintenance
- **Internet:**
  - Text mining, Intelligent query answering
  - Web access analysis, Site personalization
- **Medical Informatics**

Data Mining Techniques (box of tricks)

- **Statistics**
- **Linear Regression**
- **Visualization**
- **Cluster analysis**
  - Older, Data preparation, Exploratory

- **Decision trees**
- **Rule induction**
- **Neural networks**
- **Abductive networks**
  - Newer, Modeling, Knowledge Representation
Data-based Predictive Modeling

1. Develop Model With Known Cases
   - Attributes, X
   - F(X)
   - Diagnosis, Y
   - Determine F(X)

2. Use Model For New Cases
   - Attributes, X
   - Diagnosis (Y)
   - Y = F(X)

Modeling by Supervised Learning

- Y = F(X): true function (usually not known) for population P
- 1. Collect Data: “labeled” training sample drawn from P
   - 57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0
   - 78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
   - 69,F,180,0,115,85,40,22,0,0,0,0,1,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
   - 18,M,165,0,110,80,41,5,4,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
- 2. Training: Get G(X); model learned from training sample,
  Goal: $E<(F(X)-G(X))^2> \approx 0$ for future samples drawn from P
  - Not just data fitting!
- 3. Test/Use:
  - 71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
Data-based Predictive Modeling by supervised Machine learning

- Database of solved examples (input-output)
- Preparation: cleanup, transform, add new attributes...
- Split data into a training and a test set
- Training:
  Develop model on the training set
- Evaluation:
  See how the model fares on the test set
- Actual use:
  Use successful model on new input data to estimate unknown output

The Neural Network (NN) Approach

<table>
<thead>
<tr>
<th>Input Layer</th>
<th>Hidden Layer</th>
<th>Output Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td>Neurons</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td>Actual: 0.65</td>
</tr>
<tr>
<td><strong>Stage</strong></td>
<td></td>
<td>0.60</td>
</tr>
</tbody>
</table>

Independent Input Variables (Attributes)

Weights

Transfer Function

Dependent Output Variable

Error back-propagation

Error: 0.05
Limitations of Neural Networks

- **Ad hoc approach** for determining network structure and training parameters - Trial & Error?
- Opacity or **black-box** nature gives poor explanation capabilities which are important in medicine

\[ F(x) \approx G(x) \]

G(x) is ‘distributed’ in a maze of network weights

- Significant inputs are not immediately obvious
- When to stop training to avoid over-fitting?
- Local Minima may hinder optimum solution

Self-Organizing Abductive (Polynomial) Networks

- Network of polynomial functional elements - not simple neurons
- No fixed *a priori* model structure. Model evolves with training
- Automatic selection of: Significant inputs, Network size, Element types, Connectivity, and Coefficients
- Automatic stopping criteria, with simple control on complexity
- Analytical input-output relationships

**Double** Element:

\[ y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 + w_5 x_1 x_2 + w_6 x_1^3 + w_7 x_2^3 \]
Data Mining in Medicine

Medicine revolves on Pattern Recognition, Classification, and Prediction

**Diagnosis:**
- Recognize and classify patterns in multivariate patient attributes

**Therapy:**
- Select from available treatment methods; based on effectiveness, suitability to patient, etc.

**Prognosis:**
- Predict future outcomes based on previous experience and present conditions
Need for Data Mining in Medicine

- Nature of medical data: noisy, incomplete, uncertain, nonlinearities, fuzziness ⇒ Soft computing
- Too much data now collected due to computerization (text, graphs, images,...)
- Too many disease markers (attributes) now available for decision making
- Increased demand for health services: (Greater awareness, increased life expectancy, ...)
  - Overworked physicians and facilities
- Stressful work conditions in ICUs, etc.

Medical Applications

- Screening
- Diagnosis
- Therapy
- Prognosis
- Monitoring
- Biomedical/Biological Analysis
- Epidemiological Studies
- Hospital Management
- Medical Instruction and Training
Medical Screening

- Effective low-cost screening using disease models that require easily-obtained attributes: (historical, questionnaires, simple measurements)
- Reduces demand for costly specialized tests (Good for patients, medical staff, facilities, …)
- Examples:
  - Prostate cancer using blood tests
  - Hepatitis, Diabetes, Sleep apnea, etc.

Diagnosis and Classification

- Assist in decision making with a large number of inputs and in stressful situations
- Can perform automated analysis of:
  - Pathological signals (ECG, EEG, EMG)
  - Medical images (mammograms, ultrasound, X-ray, CT, and MRI)
- Examples:
  - Heart attacks, Chest pains, Rheumatic disorders
  - Myocardial ischemia using the ST-T ECG complex
  - Coronary artery disease using SPECT images
Diagnosis and Classification

ECG Interpretation

- R-R interval
- QRS duration
- AVF lead
- QRS amplitude
- S-T elevation

- SV tachycardia
- Ventricular tachycardia
- LV hypertrophy
- RV hypertrophy
- Myocardial infarction

Therapy

- Based on modeled historical performance, select best intervention course:
  e.g. best treatment plans in radiotherapy
- Using patient model, predict optimum medication dosage: e.g. for diabetics
- Data fusion from various sensing modalities in ICUs to assist overburdened medical staff
Prognosis

- Accurate prognosis and risk assessment are essential for improved disease management and outcome
  - Examples:
    - Survival analysis for AIDS patients
    - Predict pre-term birth risk
    - Determine cardiac surgical risk
    - Predict ambulation following spinal cord injury
    - Breast cancer prognosis

Biochemical/Biological Analysis

- Automate analytical tasks for:
  - Analyzing blood and urine
  - Tracking glucose levels
  - Determining ion levels in body fluids
  - Detecting pathological conditions
**Epidemiological Studies**

Study of health, disease, morbidity, injuries and mortality in human communities

- Discover patterns relating outcomes to exposures
- Study independence or correlation between diseases
- Analyze public health survey data
- Example Applications:
  - Assess asthma strategies in inner-city children
  - Predict outbreaks in simulated populations

---

**Hospital Management**

- Optimize allocation of resources and assist in future planning for improved services

Examples:
  - Forecasting patient volume, ambulance run volume, etc.
  - Predicting length-of-stay for incoming patients
Medical Instruction and Training

- Disease models for the instruction and assessment of undergraduate medical and nursing students
- Intelligent tutoring systems for assisting in teaching the decision making process

Benefits:

- Efficient screening tools reduce demand on costly health care resources
- Data fusion from multiple sensors
- Help physicians cope with the information overload
- Optimize allocation of hospital resources
- Better insight into medical survey data
- Computer-based training and evaluation
The KFUPM Experience

Medical Informatics Applications

- Modeling obesity (KFU)
- Modeling the educational score in school health surveys (KFU)
- Classifying urinary stones by Cluster Analysis of ionic composition data (KSU)
- Forecasting patient volume using Univariate Time-Series Analysis (KFU)
- Improving classification of multiple dermatology disorders by Problem Decomposition (Cairo University)
Modeling Obesity
Using Abductive Networks

- Waist-to-Hip Ratio (WHR) obesity risk factor modeled in terms of 13 health parameters
- 1100 cases (800 for training, 300 for evaluation)
- Patients attending 9 primary health care clinics in 1995 in Al-Khobar
- Modeled WHR as a categorical variable and as a continuous variable
- Analytical relationships derived from the continuous model adequately ‘explain’ the survey data

Modeling Obesity:
Categorical WHR Model

- WHR > 0.84: Abnormal (1)
- Automatically selects most relevant 8 inputs

<table>
<thead>
<tr>
<th>Predicted</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (250)</td>
<td>248</td>
<td>1</td>
</tr>
<tr>
<td>0 (51)</td>
<td>2</td>
<td>49</td>
</tr>
</tbody>
</table>

Classification Accuracy: 99%
Modeling Obesity: Continuous WHR - Simplified Model

- Uses only 2 variables: Height and Diastolic Blood Pressure
- Still reasonably accurate: 88% of cases had error within ±10%
- Simple analytical input-output relationship
- Adequately explains the survey data

\[ WHR = 0.704 + 0.133 y \]
\[ y = 0.368 x_2 + 0.18 x_1^2 + 0.164 x_2^2 - 0.294 x_1 x_2 \]

\[
\begin{align*}
WHR_{\text{est}} &= 0.033668 \times \text{DBP} + 1.1292 \times 10^{-4} \times \text{DBP}^2 \\
&+ 0.10035 \times \text{HGT} - 2.304 \times 10^{-8} \times \text{HGT}^2 \\
&- 2.8765 \times 10^{-4} \times \text{HGT} \times \text{DBP} - 9.1357
\end{align*}
\]

Modeling the Educational Score in School Health Surveys

- 2720 Albanian primary school children
- Educational score modeled as an ordinal categorical variable (1-5) in terms of 8 attributes: region, age, gender, vision acuity, nourishment level, parasite test, family size, parents education
- Model built using only 100 cases predicts output for remaining 2620 cases with 100% accuracy
- A simplified model selects 3 inputs only:
  - Vision acuity
  - Number of children in family
  - Father’s education
Classifying Urinary Stones by Cluster Analysis of Ionic Composition Data

- Classified 214 non-infection kidney stones into 3 groups
- Clustering with only the 3 radicals had 94% agreement with an empirical classification scheme developed previously at KSU, with the same 3 variables

Forecasting Monthly Patient Volume at a Primary Health Care Clinic, Al-Khobar Using Univariate Time-Series Analysis

- Used data for 9 years to forecast volume for two years ahead

Error over forecasted 2 years: Mean = 0.55%, Max = 1.17%
Improving classification of multiple dermatology disorders by Problem Decomposition (Cairo University)

- Standard UCI Dataset
- 6 classes of dermatology disorders
- 34 input features
- Classes split into two categories
- Classification done sequentially at two levels

- Improved classification accuracy from 91% to 99%
- About 50% reduction in the number of required input features

Summary

- Data mining is set to play an important role in tackling the data overload in medical informatics
- Benefits include improved health care quality, reduced operating costs, and better insight into medical data
- Abductive networks offer advantages over neural networks, including faster model development and better explanation capabilities