Data Mining
(*Big Data Analytics*)

Andrew Kusiak
Intelligent Systems Laboratory
2139 Seamans Center
The University of Iowa
Iowa City, IA 52242 - 1527
andrew-kusiak@uiowa.edu
http://user.engineering.uiowa.edu/~ankusiak/
Tel. 319-335 5934
Fax. 319-335 5669

What is Data Mining?

A process

- Domain understanding
- Data selection
- Data cleaning, e.g., data duplication, missing data
- Preprocessing, e.g., integration of different files
- Pattern (knowledge) discovery
- Interpretation (e.g.,visualization)
- Reporting

Illustrative Applications

General category

- Prediction of equipment faults
- Determining a stock level
- Process control
- Fraud detection
- Genetics
- Disease staging and diagnosis
- Decision making

Manufacturing
Service industry
Healthcare
E-Applications
Pharmaceutical Applications

What is Knowledge Discovery?

Process

Data Set

Learning

Knowledge

E.g., Excel, Access, Data Warehouse

Learning Systems (1/2)

Knowledge discovery

- Classical statistical methods
  (e.g., discriminant analysis)
- Modern statistical techniques
  (e.g., *k*-nearest neighbor, Bayes theorem)
- Neural networks
- Support vector machines
- Decision tree algorithms
- Decision rule algorithms
- Learning classifier systems

Black box tools

Learning Systems (2/2)

- Association rule algorithms
- Text mining algorithms
- Meta-learning algorithms
- Inductive learning programming
- Sequence learning

Neural Networks

Definition

- Based on biology
- Inputs transformed via a network of simple processors
- Processor combines (weighted) inputs and produces an output value
- Obvious questions: What transformation function do you use and how are the weights determined?

Types of Decision Trees

- CHAID: Chi-Square Automatic Interaction Detection
  - Kass (1980)
  - n-way splits
  - Categorical variables
- CART: Classification and Regression Trees
  - Breimam, Friedman, Olshen, and Stone (1984)
  - Binary splits
  - Continuous variables
- C4.5
  - Quinlan (1993)
  - Also used for rule induction

Text Mining

- Mining unstructured data (free-form text) is a challenge for data mining
- Usual solution is to impose structure on the data and then process using standard techniques, e.g.,
  - Simple heuristics (e.g., unusual words)
  - Domain expertise
  - Linguistic analysis
- Presentation is critical

Yet Another Classification

Examples

- Supervised
  - Regression models
  - k-Nearest-Neighbor
  - Neural networks
  - Rule induction
  - Decision trees
- Unsupervised
  - k-means clustering
  - Self organized maps

Supervised Learning Algorithms

Characteristics

- kNN
  - Quick and easy
  - Models tend to be very large
- Neural Networks
  - Difficult to interpret
  - Training can be time consuming
- Rule Induction
  - Understandable
  - Need to limit calculations
- Decision Trees
  - Understandable
  - Relatively fast
  - Easy to translate into SQL queries

Knowledge Representation Forms

Examples

- Decision rules
- Trees (graphs)
- Patterns (matrices)

DM: Product Quality Example

Training data set

Product ID	Process param 1	Test_1	Process param_2	Test_2	Quality D
1	1.02	Red	2.98	High	Good_Quality
2	2.03	Black	1.04	Low	Poor_Quality
3	0.99	Blue	3.04	High	Good_Quality
4	2.03	Blue	3.11	High	Good_Quality
5	0.03	Orange	0.96	Low	Poor_Quality
6	0.04	Blue	1.04	Medium	Poor_Quality
7	0.99	Orange	1.04	Medium	Good_Quality
8	1.02	Red	0.94	Low	Poor_Quality

Decision Rules

Rule 1. IF (Process_parameter_1 < 0.515) THEN (D = Poor_Quality);
[2, 2, 50.00%, 100.00%][2, 0][5, 6]

Rule 2. IF (Test_2 = Low) THEN (D = Poor_Quality);
[3, 3, 75.00%, 100.00%][3,0][2, 5, 8]

Rule 3. IF (Process_parameter_2 >= 2.01) THEN (D = Good_Quality);
[3, 3, 75.00%, 100.00%][0, 3][1, 3, 4]

Rule 4. IF (Process_parameter_1 >= 0.515) & (Test_1 = Orange) THEN (D = Good_Quality);
[1, 1, 25.00%, 100.00%][0, 1][7]

Data Mining Result

Decision Rule Metrics

Support, approximate, error

Rule 12
IF (Flow = 6) AND (Pressure = 7)
THEN (Efficiency = 81);
[13, 8, 4.19%, 61.54%] [1, 8, 4]

No of supporting objects

Support Strength Relative strength Confidence

[{ 524 },
{ 527, 528, 529, 530, 531, 533, 535, 536 },
{ 525, 526, 532, 534 }]

Supporting objects

Definitions

Rule metrics

- Support = Number of objects satisfying condition of the rule
- Strength = Number of objects satisfying condition and the decision of the rule
- Relative strength = Number of objects satisfying condition and the decision of the rule/The number of objects in the class
- Confidence = Strength/Support

Classification Accuracy

Cross-validation

Test: Leaving-one-out
Confusion Matrix

	Poor_Quality	Good_Quality	None
Poor_Quality	**3**	**1**	**0**
Good_Quality	**1**	**3**	**0**

Average Accuracy [%]

	Correct	Incorrect	None
Total	**75.00**	**25.00**	**0.00**
Poor_Quality	**75.00**	**25.00**	**0.00**
Good_Quality	**75.00**	**25.00**	**0.00**

Decision rules

Leave-one-out Approach

Rule 113

IF (B_Master >= 1634.26)
AND (B_Temp in (1601.2, 1660.22]
AND (B_Pressure in [17.05, 18.45))
AND (A_point = 0.255) AND (Average_O2 = 77)
THEN (Eff = 87) OR (Eff = 88);

[6, 6, 23.08%, 100.00%][0, 0, 0, 0, 0, 0, 0, 3, 3, 0]
[{2164, 2167, 2168}, {2163, 2165, 2166}]

Decision rules

Leave-one-out Approach

Rule 12
IF (Ave_Middle_Bed = 0) AND (PA_Fan_Flow = 18) THEN (Efficiency = 71);

[16, 10, 10.31%, 62.50%] [1, 1, 2, 10, 2,]
[{ 682 }, { 681 }, { 933, 936 },
{ 875, 876, 877, 878, 879, 880, 934, 935, 1000, 1001},
{ 881, 882 }]

Decision Rule vs Decision Tree Algorithms

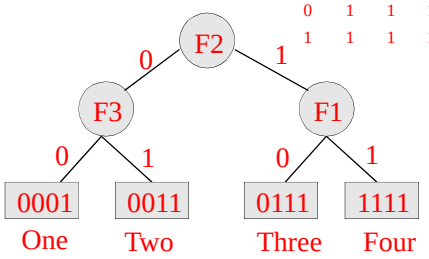
Training Data Set

F1	F2	F3	F4	D
0	0	0	1	One
0	0	1	1	Two
0	1	1	1	Three
1	1	1	1	Four

Decision Tree

Reduced number of features

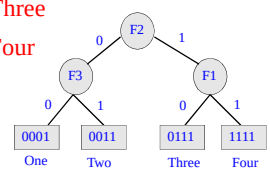
F1	F2	F3	F4	D
0	0	0	1	One
0	0	1	1	Two
0	1	1	1	Three
1	1	1	1	Four



Decision Tree

Pattern

F1	F2	F3	F4	D
0	0	0	1	One
0	0	1	1	Two
0	1	1	1	Three
1	1	1	1	Four



Decision Rules

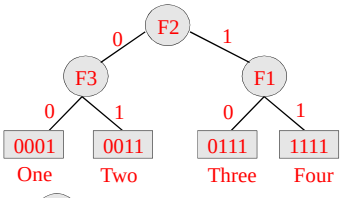
```
Rule 1. (F3 = 0) THEN (D = One);
[1, 100.00%, 100.00%][1]
Rule 2. (F2 = 0) AND (F3 = 1) THEN (D = Two);
[1, 100.00%, 100.00%][2]
Rule 3. (F1 = 0) AND (F2 = 1) THEN (D = Three);
[1, 100.00%, 100.00%][3]
Rule 4. (F1 = 1) THEN (D = Four);
[1, 100.00%, 100.00%][4]
```

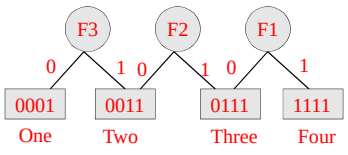
Pattern

F1	F2	F3	F4	D
0	0	0	1	One
0	0	1	1	Two
0	1	1	1	Three
1	1	1	1	Four

Decision Tree vs Rule Tree

Decision Tree





Rule Tree

Decision Rule Algorithms

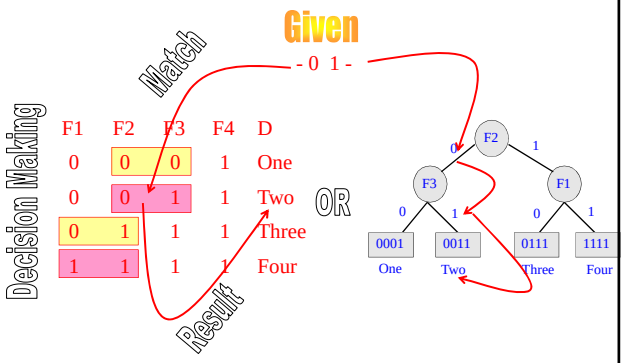
Important Class of Algorithms

Rough Set Theory

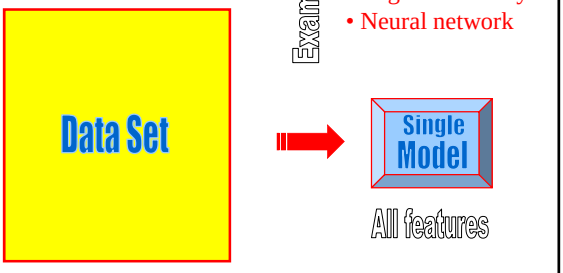
Identify
unique features of an object
rather than
commonality among all objects

Individual object vs population based approach

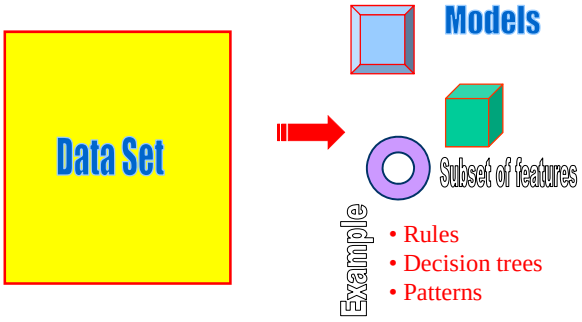
Use of Extracted Knowledge



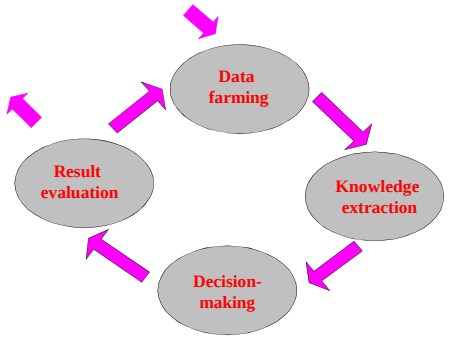
Traditional Modeling



Data Mining



Data Life Cycle



Data Mining Standards

- Predictive Model Markup Language (PMML)
  - The Data Mining Group (www.dmg.org)
  - XML based (DTD)
- Java Data Mining API spec request (JSR-000073)
  - Oracle, Sun, IBM, …
  - Support for data mining APIs on J2EE platforms
  - Build, manage, and score models programmatically
- OLE DB for Data Mining
  - Microsoft
  - Table based
  - Incorporates PMML

Summary

- Data mining algorithms support a new paradigm: Identify what is unique about an object
- DM tools to enter new areas of information analysis