Data Mining

Lesson 10

Engineering data mining tasks

MSc in Computer Science
University of New York Tirana
Assoc. Prof. Dr. Marenglen Biba

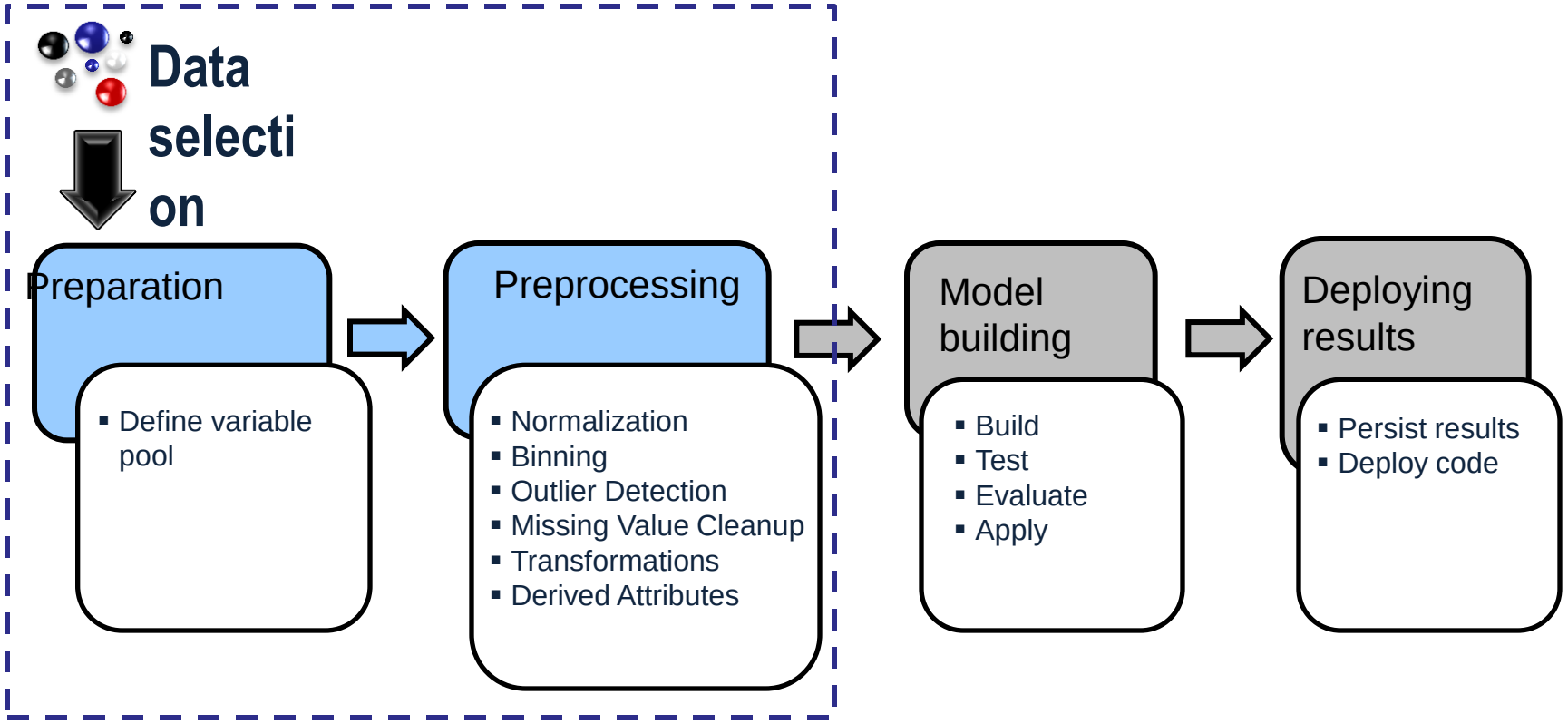
Data Mining: Content

- Introduction to data mining and machine learning
- Inductive learning
- Decision trees
- Rule induction
- Instance-based learning
- Bayesian learning
- Neural networks
- Support vector machines
- Other machine learning models
- Engineering data mining tasks

Engineering data mining tasks

- Two parts:
  - Credibility: Evaluating what's been learned – Lesson 6
  - Transformations: Engineering the input and output
    - Useful input transformations in Weka
    - Text classification in Weka (Some papers published with Cohort 1 and 2)
    - Anomaly detection in Oracle

Step 2: Data Selection, Preparation, and Preprocessing



Real data mining tasks

- Until now we have examined a large number of machine learning methods:
    - decision trees,
    - decision rules,
    - linear models,
    - instance-based schemes,
    - numeric prediction techniques,
    - clustering algorithms, and
    - Bayesian networks.

- All are sound, robust techniques that are eminently applicable to practical data mining problems.

- But successful data mining involves far more than selecting a learning algorithm and running it over your data.

Data engineering

- Attribute selection
- Attribute discretization
- Data transformation
- Data cleansing

Data Engineering in Weka

- Decision Tree learning
  - Credit (Australian banks)
- Normal dataset
  - Correctly Classified Instances 594 86.087 %
  - Incorrectly Classified Instances 96 13.913 %
  - Mean absolute error 0.1924
  - Root mean squared error 0.3313

- Replace Missing Values with Weka filters
- Discretize all numeric attributes with supervised method
  - Correctly Classified Instances 599 86.8116 %
  - Incorrectly Classified Instances 91 13.1884 %
  - Mean absolute error 0.1887
  - Root mean squared error 0.3206
- Smaller Tree
    - 16 leaves compared to 30
    - 22 nodes compared to 40

Data Engineering in Weka

- Glass
- DTrees
- Normal Dataset
  - Correctly Classified Instances 143 66.8224 %
  - Incorrectly Classified Instances 71 33.1776 %
  - Mean absolute error 0.1026
  - Root mean squared error 0.2897
  - Relative absolute error 48.4507 %
  - Root relative squared error 89.2727 %
- Discretized dataset
  - Correctly Classified Instances 158 73.8318 %
  - Incorrectly Classified Instances 56 26.1682 %
  - Mean absolute error 0.102
  - Root mean squared error 0.2462
  - Relative absolute error 48.1836 %
  - Root relative squared error 75.8548 %

Data Engineering in Weka

- Soybean
- Normal dataset
  - Correctly Classified Instances 625 91.5081 %
  - Incorrectly Classified Instances 58 8.4919 %
  - Kappa statistic 0.9068
  - Mean absolute error 0.0135
  - Root mean squared error 0.0842
  - Relative absolute error 14.0484 %
  - Root relative squared error 38.4134 %
  - Number of Leaves : 61
  - Size of the tree : 93
- Replace missing values
  - Correctly Classified Instances 631 92.3865 %
  - Incorrectly Classified Instances 52 7.6135 %
  - Kappa statistic 0.9164
  - Mean absolute error 0.0091
  - Root mean squared error 0.0779
  - Relative absolute error 9.4795 %
  - Root relative squared error 35.5461 %
  - Number of Leaves : 55
  - Size of the tree : 85

Remove misclassified instances in Weka

- Decision Tree learning
  - Credit (Australian banks)
  - Remove misclassified with Weka filters
    - Choose DT as algorithms in the filter
  - Number of instances from 660 to 623.
  - Accuracy with DT 98.5% compared to 86% without remove.
  - Accuracy with NBayes: 85.39% compared to 77.68% without remove.

Practical attribute selection in Weka

- Decision Tree learning
  - Credit (Australian banks)
    - Full-training set
    - InfoGainAttributeEval
    - Ranker
    - Results
      - Only 7 attributes are enough
      - Acc. 85.94% compared to 86.04%.
        - » Much simpler trees
        - » Faster training

- Conclusion: Attribute selection works!

Selecting attributes

- In Select Attributes in the Weka menu
  - Ranked attributes:
  - 0.425709 9 A9
  - 0.213511 11 A11
  - 0.156286 10 A10
  - 0.110235 15 A15
  - 0.110022 8 A8
  - 0.107525 6 A6
  - 0.05371 14 A14
- ....

- Then in Filters
  - Choose InfoGain
  - Apply 0.09 the threshold in the Ranker

Feature Selection

- Autoprice
- Select attributes in the Weka menu
  - Search Method:
  - Greedy Stepwise (forwards).
  - Start set: no attributes
  - Merit of best subset found: 0.902

  - Attribute Subset Evaluator (supervised, Class (numeric): 16 class):
  - CFS Subset Evaluator
  - Including locally predictive attributes

  - Selected attributes: 2,3,5,7,8,11,12 : 7
  - normalized-losses
  - wheel-base
  - width
  - curb-weight
  - engine-size
  - compression-ratio
  - horsepower

Feature Selection

- Autoprice
- Multilayerperceptron
- With all attributes
  - Correlation coefficient 0.8994
  - Mean absolute error 1746.4859
  - Root mean squared error 2672.8403
  - Relative absolute error 37.7849 %
  - Root relative squared error 45.2067 %

- With only the selected attributes
  - Correlation coefficient 0.912
  - Mean absolute error 1612.4005
  - Root mean squared error 2415.0072
  - Relative absolute error 34.884 %
  - Root relative squared error 40.8459 %

Feature Selection

- Autoprice
- Support Vector Machine
- With all attributes
  - Correlation coefficient 0.8826
  - Mean absolute error 1818.2199
  - Root mean squared error 2825.4293
  - Relative absolute error 39.3368 %
  - Root relative squared error 47.7875 %

- With only the selected attributes
  - Correlation coefficient 0.8956
  - Mean absolute error 1722.6805
  - Root mean squared error 2761.8501
  - Relative absolute error 37.2699 %
  - Root relative squared error 46.7122 %

Text Mining with Weka

- Classification of text documents

Readings for this part

- Data Mining. Witten and Frank
  - Chapter 7, except section 7.5