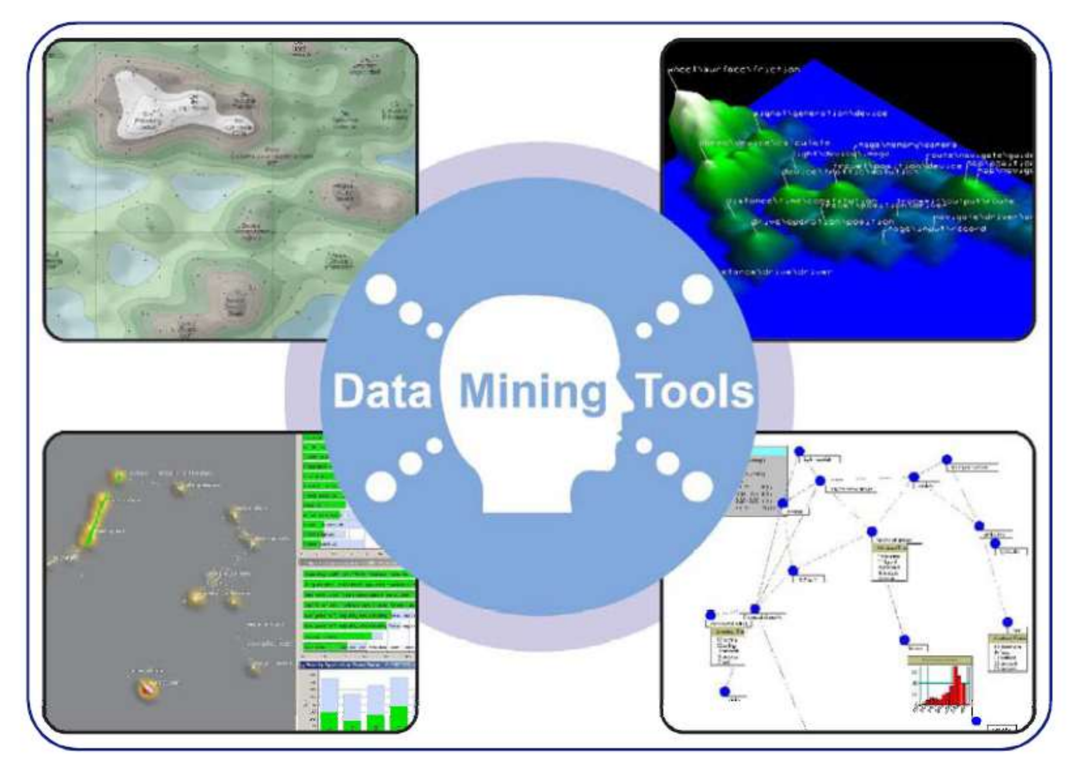
Data mining tools

Student: Oltean Andreea (gr 253)

Data mining

Content

- Data mining tools
- DM Tool Comparisons
- Data Mining Tools for Analytics Applications
- WEKA



Data Mining Tools

- **Data mining tools**

  - describes a category of software applications and methodologies designed to help businesses understand and make sense of their data.

  - provide statistical data models (classification or clustering studies, linear regression, and current or predictive modeling) and utilize visualization functions to support the analysis of massive quantities of data stored by customer organizations.

  - predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions

- **Data Mining Tools vs. Data Mining Applications**

  - data mining tools contain numerous methods that can be applied universally to any basic business problem.

  - data mining applications are typically more customized, operating on a specific business problem.

- **Why Choose Data Mining Tools?**

  - provide users with a platform for uncovering, converting and analyzing private or corporate data; flexibility, thorough technique and large margin for accuracy.

- **Standards of Data Mining Tools**

  - guideline for ensuring consistent results from data mining tools is the “Cross-Industry Standard Process for Data Mining” (CRISP-DM).

- **Comparing Data Mining Tools**

  - when choosing data mining tools, it is important to keep in mind the following elements:

    a) the type of platform that the data mining tool supports/complements

    b) the algorithms included

    c) any decision trees and neural networks

    d) data input and model output options

    e) visualization capabilities …and modeling automation methods.

DM software/tools

- **Commercial Software**

1. SPSS Clementine (PASW Modeler)
2. Salford : CART, MARS, TreeNet, RF
3. SAS
4. KXEN
5. SAS Enterprise Miner
6. MATLAB- Armada 1.3.2
7. SQL Server - Analysis Services
8. DBminer Technology Inc.: DBMiner
9. Megaputer Intelligence: PolyAnalyst
10. Statistica Data Miner
11. Insightful Miner

- **Free/Open Source Data Mining Software**

1. RapidMiner
2. R
3. Weka
4. KNIME
5. C4.5/C5.0
6. Cubist
7. GritBot
8. ODBCHook
9. Orange

DM Tool Suite Comparisons

- **SPSS Clementine**

  - was the first data mining suite to use the graphical programming in the 1980s.

  Pros:

  - Good variety of data mining algorithms.

  - Very powerful optimal parameter search routines built into many of the data mining algorithms (automatic trials of different parameter sets).

  - Power meta-learning models can be built, in which the results of one modeling algorithm can be easily streamed as input to another modeling algorithm.

  - Powerful (but proprietary) internal scripting language (CLEM) for creating complex variable processing.

  - Moderately easy to use.

  Cons:

  - Relatively little descriptive statistical or parametric statistical analysis capabilities are available directly in the tool.

  - Relatively poor descriptive or output graphics forms.

- **STATISTICA Data Miner**

  - this uniqueness is defined primarily by in terms of the many things it does well, and the completeness in facilitating all tasks of the data mining project

  Pros:

  - Provides the richest combination of parametric statistical and machine learning data mining algorithms

  - Relatively easy to use graphical programming user interface.

  - Provides tools for all common data mining tasks.

  - Highly flexible tools for model output.

  - Powerful tools for reduction of dimensionality.

  - Powerful customization options based on the industry standard VB language.

  Cons:

  - Lift charts are not easily available for evaluation of neural net models.

- **KXEN**

- is one of tool suite that provide an implementation of a support vector machine (SVM).

Pros:

- is clearly the most accurate data mining tool available today.

- Various combinations and transforms of existing variables are automatically created and included in the analysis as derived predictor variables.

Cons:

- A clean data set must be submitted to the Consistent Coder of KXEN in the form of one record per entity to be modeled.

- There are no data preparation tools to help you put the data in this form (although many data preparation steps required by parametric statistical or neural nets/decision trees are not necessary with SVMs).

- No coincidence (or "confusion") matrix is available for binary output, from which precision and recall values can be calculated. You can create one, if you can determine the correct threshold to use on the decimal output to convert it to binary predicted values.

- **Insightful Miner**

  - This tool suite may be the best one available for a company that would like to use ordinary business analysts to do relatively simple data mining projects.

  - it provides a rich assortment of data mining and statistical data mining algorithms (but not nearly as rich as does STATISTICA Data Miner).

  Pros:

  - Excellent tools for data import/export, data exploration and data cleansing tasks, and reduction of dimensionality prior to modeling.

  - Even though it does not employ a graphical programming interface, it is relatively easy to use by non data miners.

  - The most complete general purpose data mining suite available, and it is relatively inexpensive.

  Cons:

  - A relatively low level of automation.

  - No scripting interface for coding of complex problems.

  - No model exporting capabilities.

Which tool would be best for you?

→One way is to match the tool to the data scenarios for which you plan to use it.

Data Scenarios :

- **Scenario #1.** If the company has access to (or is willing to hire) people with statistical expertise, the best tool will be one that statisticians understand and can use effectively:

  STATISTICA Data Miner , SPSS Clementine (in conjunction with SPSS Stat package)

- **Scenario #2**. If data preparation must done by hand inside the data mining package, then the best tools would include:

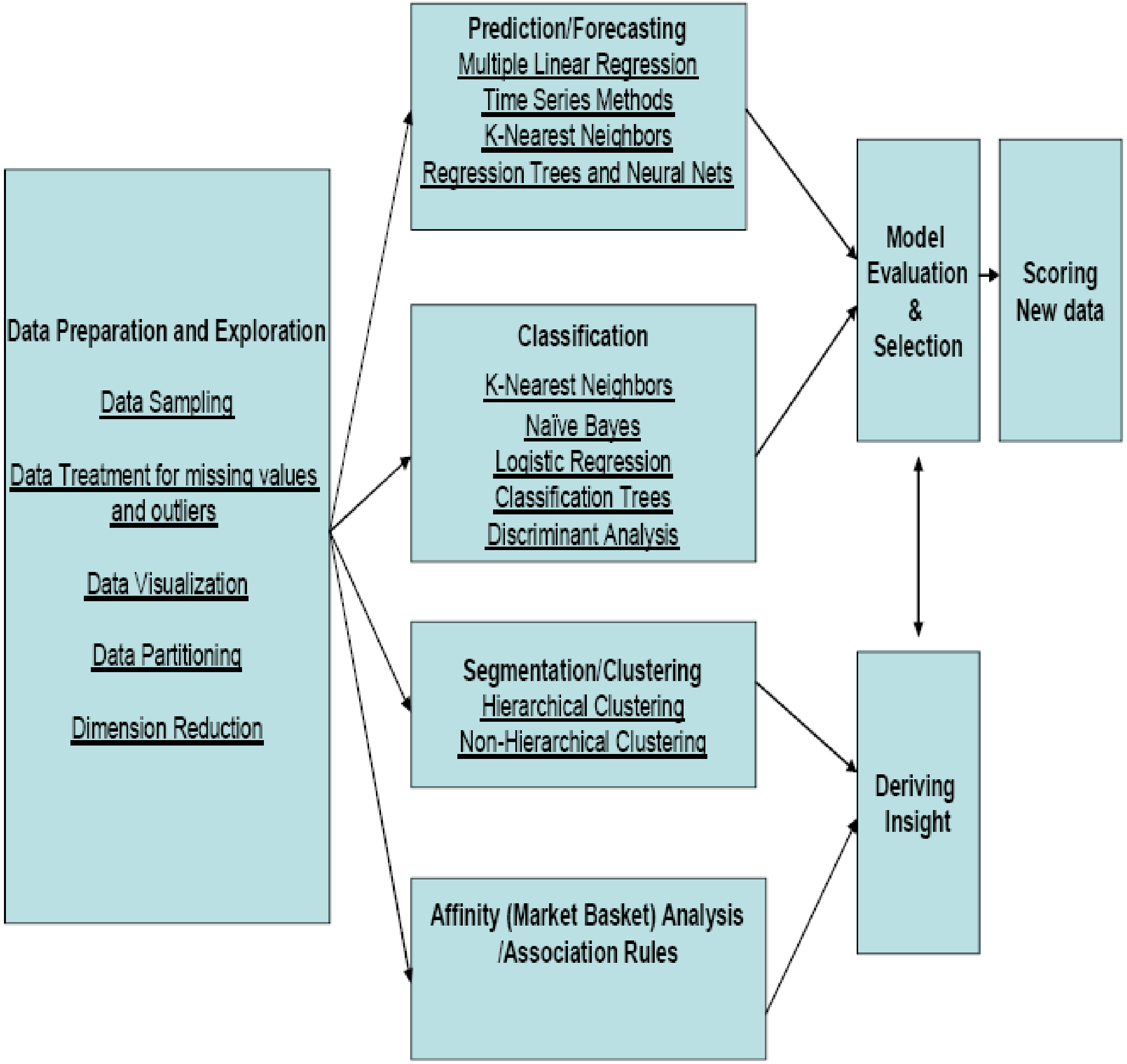
  STATISTICA Data Miner , Insightful Miner , SPSS-Clementine

- **Scenario #3.** If the company wants to do data mining modeling with lower-level business analysts, then the best tool will have a relatively high degree of automation:

  KXEN , Insightful Miner

- **Scenario #4.** If the company has it own in-house analytical tools that require some enhancement to provide data mining capability, then the best data mining tool will be one that is easily embedded into their existing systems: KXEN

Brief Overview of the Data Mining Process



Data Preparation and Exploration

Data Sampling

Data Treatment for missing values and outliers

Data Visualization

Data Partitioning

Dimension Reduction

Prediction/Forecasting
Multiple Linear Regression
Time Series Methods
K-Nearest Neighbors
Regression Trees and Neural Nets

Classification
K-Nearest Neighbors
Naïve Bayes
Logistic Regression
Classification Trees
Discriminant Analysis

Segmentation/Clustering
Hierarchical Clustering
Non-Hierarchical Clustering

Affinity (Market Basket) Analysis /Association Rules

Model Evaluation & Selection

Scoring New data

Deriving Insight

Data Mining Tools for Analytics Applications

- **Data Preparation and Exploration**

  - involves sampling of the data

  - cleaning the data (handling missing values and data entry errors)

  - visualizing the data

- **Classification**

  - the data is separated into classes. E.g.: a buyer or a non-buyer.

  - develop rules using a data set where the classification is known and then apply these rules to another data set where the classification is unknown.

- **Prediction**

  - similar to classification except that instead of predicting a binary (either/or type) classes, the goal is to predict the numerical value of the variable (for e.g. the probability or amount of purchase).

- **Segmentation/Clustering**

  - similar records in the data are grouped based on clustering algorithms that minimize within-cluster variability while maximizing between-cluster variability in the data. Business applications include customer segmentation.

- **Affinity (Market Basket)Analysis**

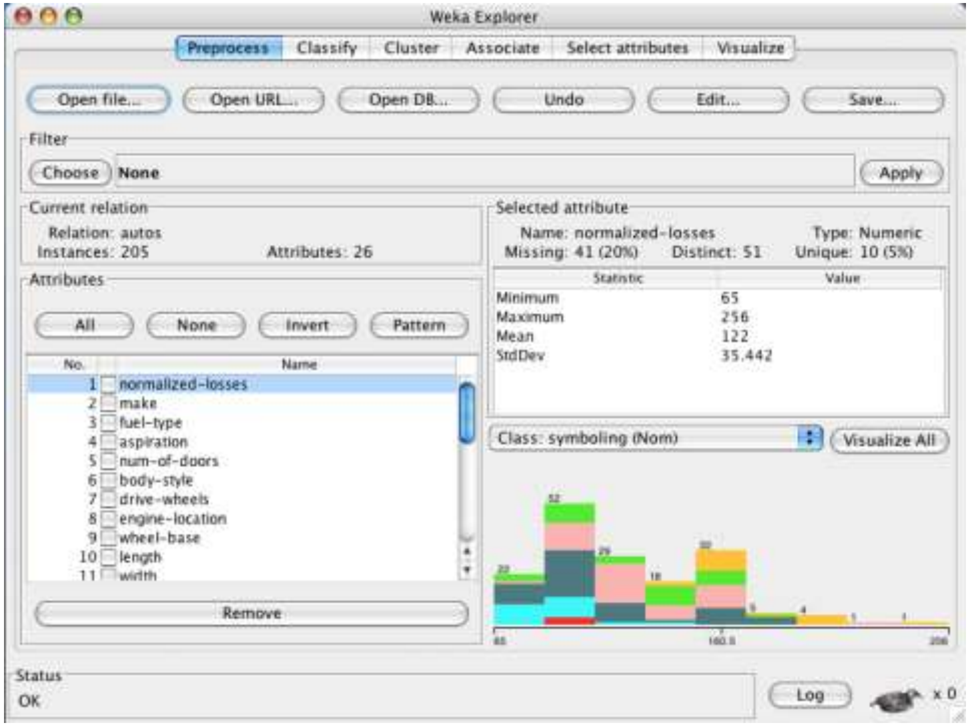
  - used in the context of large customer databases to help identify associations between purchased items.

University of Waikato (NZ): WEKA

- Waikato Environment for Knowledge Analysis (WEKA)
- Developed in Java by the Department of Computer Science
- Command line or GUI
- Lots of algorithms or add your own
- Free

Explorer: pre-processing the data

- Data can be imported from a file in various formats: ARFF, CSV, C4.5, binary
- Data can also be read from a URL or from an SQL database (using JDBC)
- Pre-processing tools in WEKA are called “filters”
- WEKA contains filters for:
  - Discretization, normalization, resampling, attribute selection, transforming and combining attributes, …



Filter
None

Current relation
Relation: autos
Instances: 205 Attributes: 26

Attributes

No.	Name
1	☐ normalized-losses
2	☐ make
3	☐ fuel-type
4	☐ aspiration
5	☐ num-of-doors
6	☐ body-style
7	☐ drive-wheels
8	☐ engine-location
9	☐ wheel-base
10	☐ length
11	☐ width

Selected attribute
Name: normalized-losses Type: Numeric
Missing: 41 (20%) Distinct: 51 Unique: 10 (5%)

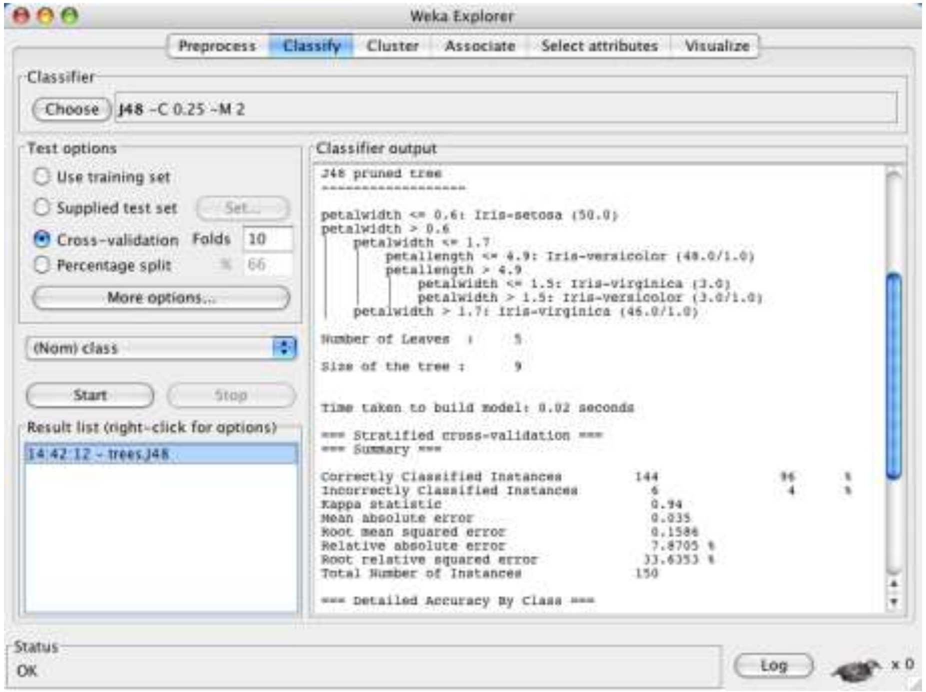
Statistic	Value
Minimum	65
Maximum	256
Mean	122
StdDev	35.442

Class: symboling (Nom)

Status
OK

Explorer: building “classifiers”

- Classifiers in WEKA are models for predicting nominal or numeric quantities
- Implemented learning schemes include:
  - Decision trees
  - support vector machines
  - multi-layer perceptrons
  - logistic regression
  - Bayes’ nets, …



Classifier: J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

Result list (right-click for options)

14:42:12 - trees.J48

Classifier output

```
J48 pruned tree
------------------

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5

Size of the tree :      9


Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.035
Root mean squared error                  0.1586
Relative absolute error                  7.8705 %
Root relative squared error             33.6353 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===
```

Status

OK

- “Meta”-classifiers include:
  - Bagging, boosting, stacking, error-correcting output codes, locally weighted learning, …

Explorer: clustering data

- WEKA contains "clusterers" for finding groups of similar instances in a dataset
- Implemented schemes are:
    - *k*-Means, EM, Cobweb, *X*-means, FarthestFirst
- Clusters can be visualized and compared to "true" clusters (if given)
- Evaluation based on loglikelihood if clustering scheme produces a probability distribution

Webography

- http://www.dataminingslab.com/pubs/kdd98_elder_abbott_nopics_bw.pdfhttp://www.cs.umd.edu/~waa/attack/frame.htm
- http://www.oracle.com/technology/products/bi/odm/pdf/odm_metaspectrum_1004.pdfhttp://www.cypherpunks.ca/bh2001/mgp00002.html
- http://www.icsti.org/documents/VTTDataMiningTools.pdf
- http://www.thearling.com/text/dmwhite/dmwhite.htm
- http://www.cs.waikato.ac.nz/ml/weka/
- http://www.information-management.com/specialreports/20040323/1000465-1.html
- http://databases.about.com/od/datamining/a/datamining.htm
- http://www.sqlserverdatamining.com/ssdm/

Thank you for your attention