Data Mining

Martin Weis

Introduction
Data=information?
Value types

Visualisation

Software

Algorithms
Simple algorithms
k-NN

Processes
k-NN
Rule
Tree
Naive Bayes
Linear regression
Feature selection

Literature

References

Data Mining

An introduction

Martin Weis[1]

June 23, 2016

[1] Martin.Weis.newsadress@gmx.de

Overview

1. Introduction
   - Data=information?
   - Value types
2. Visualisation
3. Software
4. Algorithms
   - Simple algorithms
     - k-NN
5. Processes
   - k-NN
   - Rule
   - Tree
   - Naive Bayes
   - Linear regression
   - Feature selection
6. Literature

Definition of data mining

*Data mining, also called **knowledge discovery** in databases, in computer science, the process of discovering interesting and useful **patterns** and **relationships** in large volumes of **data**[2].*

[2] Source: Encyclopædia Britannica

Data mining application fields I

Who uses data mining?

- ware houses/shops: customer profitability
- insurances: risk assesment
- finance sector: prediction, scoring
- healthcare: risk assesment
- oil and gas industry: exploration
- social networks: personal profiles

Data mining application fields II

Real life applications you might know

- shops/ads: *you might also like...*
- mail: *X-Spam-Status: Yes, hits=6.1 required=5.0*
- social networks: *you also know these persons?*
- finance: *scoring*

Data vs. information

What is data? what is information?

- data is not information
- raw data (useless) → extract information automatically

Data recorded facts

Information patterns underlying the data

Data organisation

- data is usually organised in “tables”
- instances are described by a fixed set of *features* (*attributes*, *variables*)

Value types I

Values can have different types:

ordinal categorical, nominal, discrete values

numerical real, integer

Nominal

- values are distinct symbols (names)
- no relation between nominal values (no order, no distance)
- possible operation: test equality
- example: Herbicide names in a trial

Value types II

Ordinal

- quantities that have a natural ordering
- distances, if measurable, need not to be the same
- example: time as DAT (days after treatment)

Numerical

- Interval quantities are ordered and measured in fixed/equal units
- is zero defined?
- possible operations: all mathematical operations
- example: temperature

Visualisation Overview

Visualise your data!

- simple visualisation tools are useful
- explore the value ranges, outliers, distribution
- check against domain knowledge (plausibility)

Visualisation possiblilities

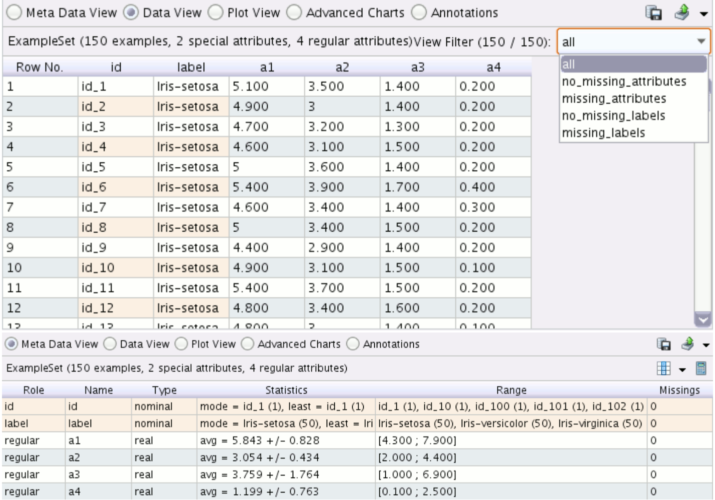
- histogram
- scattergram
- graph plots
- 2D/3D plotting

Visualisation: Example data set

Iris data

- *Iris flower data set* or *Fisher's Iris data set* (1936)
- multivariate data
- 50 samples of each of three species of Iris flowers (*Iris setosa*, *Iris virginica* and *Iris versicolor*)
- four features measured
  - length and width of sepal and petal [cm]

Visualisation examples – Tabular data and descriptive statistics I



○ Meta Data View ● Data View ○ Plot View ○ Advanced Charts ○ Annotations

ExampleSet (150 examples, 2 special attributes, 4 regular attributes)View Filter (150 / 150): all

- all
- no_missing_attributes
- missing_attributes
- no_missing_labels
- missing_labels

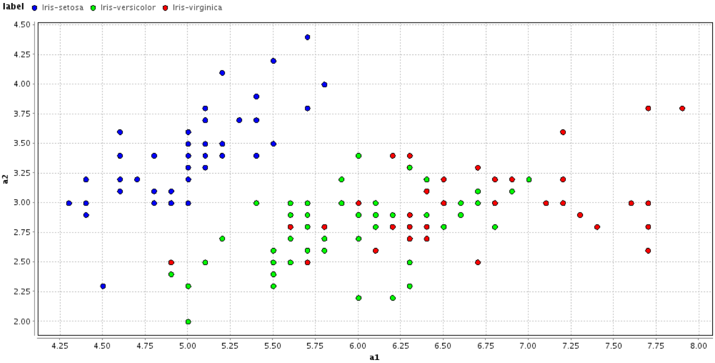
Row No.	id	label	a1	a2	a3	a4
1	id_1	Iris-setosa	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	4.900	3	1.400	0.200
3	id_3	Iris-setosa	4.700	3.200	1.300	0.200
4	id_4	Iris-setosa	4.600	3.100	1.500	0.200
5	id_5	Iris-setosa	5	3.600	1.400	0.200
6	id_6	Iris-setosa	5.400	3.900	1.700	0.400
7	id_7	Iris-setosa	4.600	3.400	1.400	0.300
8	id_8	Iris-setosa	5	3.400	1.500	0.200
9	id_9	Iris-setosa	4.400	2.900	1.400	0.200
10	id_10	Iris-setosa	4.900	3.100	1.500	0.100
11	id_11	Iris-setosa	5.400	3.700	1.500	0.200
12	id_12	Iris-setosa	4.800	3.400	1.600	0.200
13	id_13	Iris-setosa	4.800	3	1.400	0.100

● Meta Data View ○ Data View ○ Plot View ○ Advanced Charts ○ Annotations

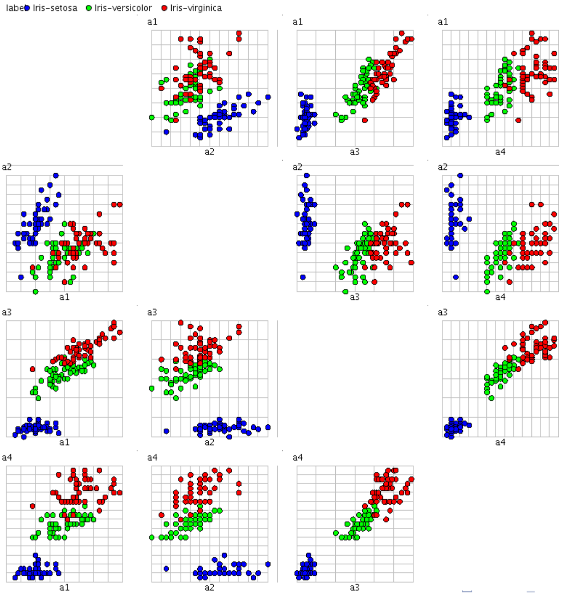
ExampleSet (150 examples, 2 special attributes, 4 regular attributes)

Role	Name	Type	Statistics	Range	Missings
id	id	nominal	mode = id_1 (1), least = id_1 (1)	id_1 (1), id_10 (1), id_100 (1), id_101 (1), id_102 (1)	0
label	label	nominal	mode = Iris-setosa (50), least = Iri	Iris-setosa (50), Iris-versicolor (50), Iris-virginica (50)	0
regular	a1	real	avg = 5.843 +/- 0.828	[4.300 ; 7.900]	0
regular	a2	real	avg = 3.054 +/- 0.434	[2.000 ; 4.400]	0
regular	a3	real	avg = 3.759 +/- 1.764	[1.000 ; 6.900]	0
regular	a4	real	avg = 1.199 +/- 0.763	[0.100 ; 2.500]	0

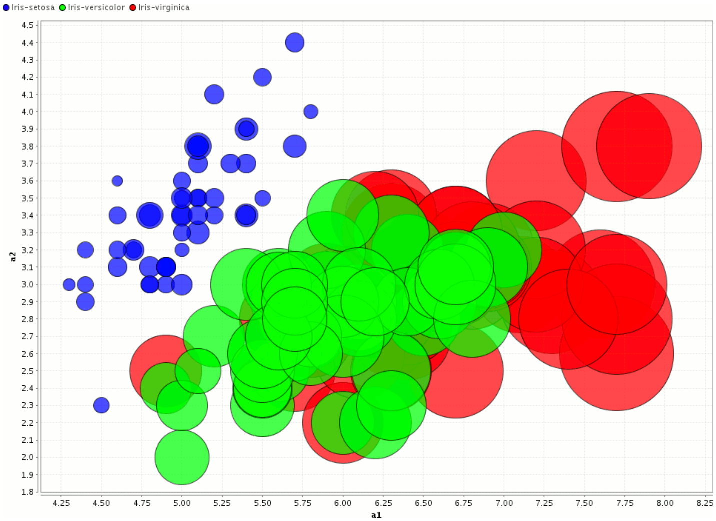
Visualisation examples – Scatter plot



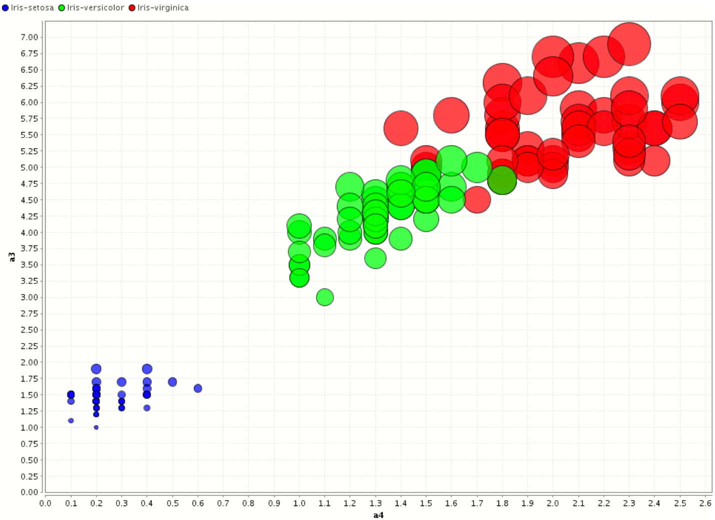
Visualisation examples – Scatter plot matrix



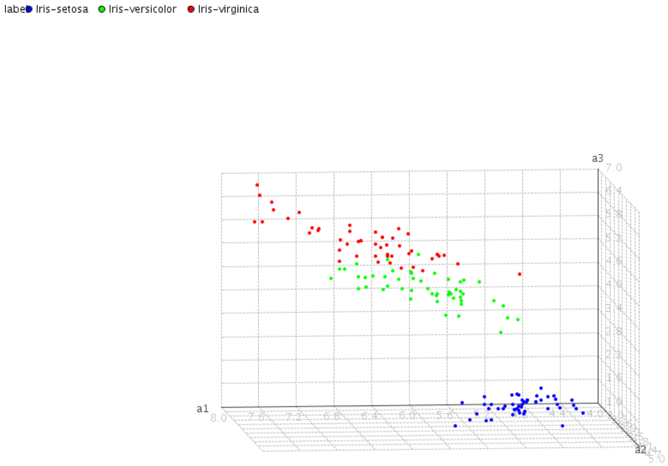
Visualisation examples – Bubble I



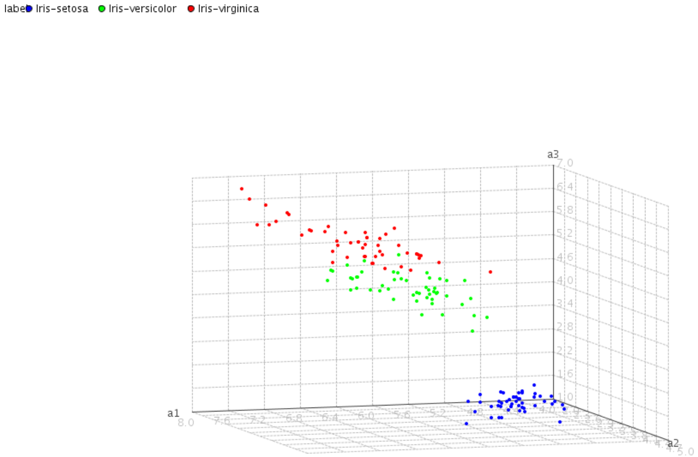
Visualisation examples – Bubble II



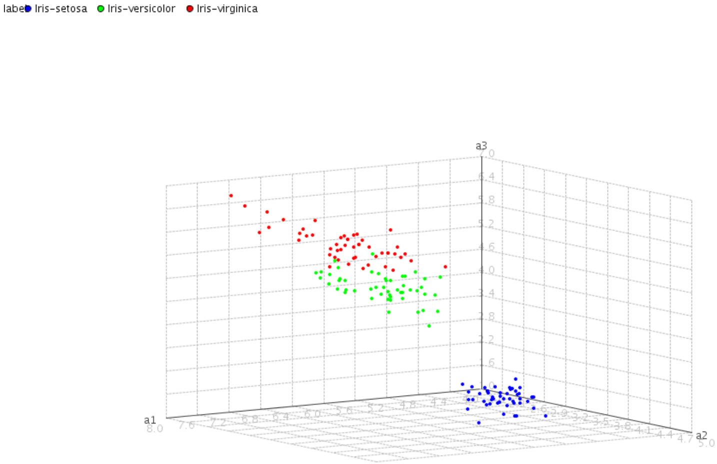
Visualisation examples – 3D plot I



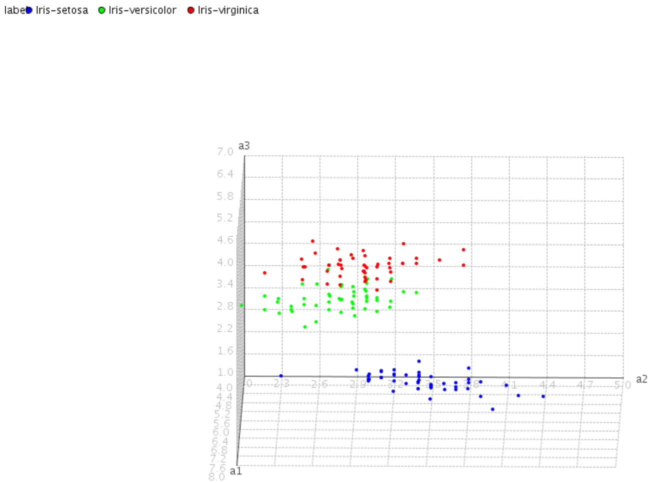
Visualisation examples – 3D plot II



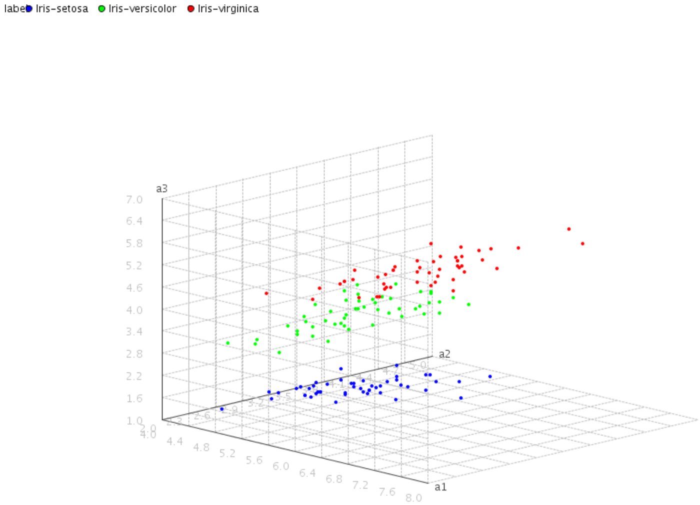
Visualisation examples – 3D plot III



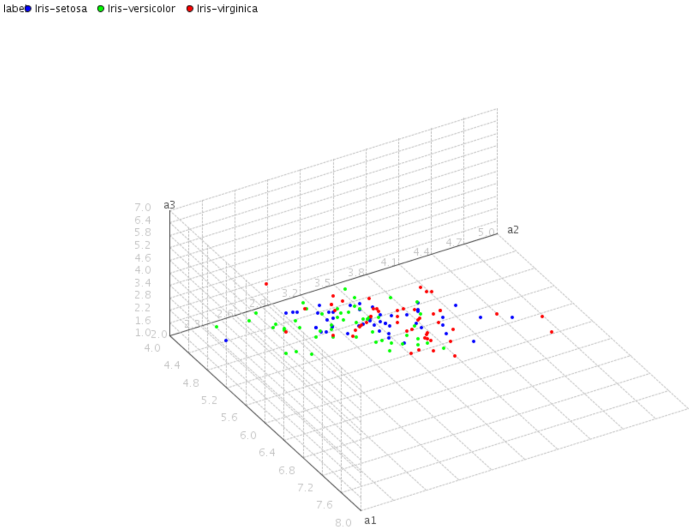
Visualisation examples – 3D plot IV



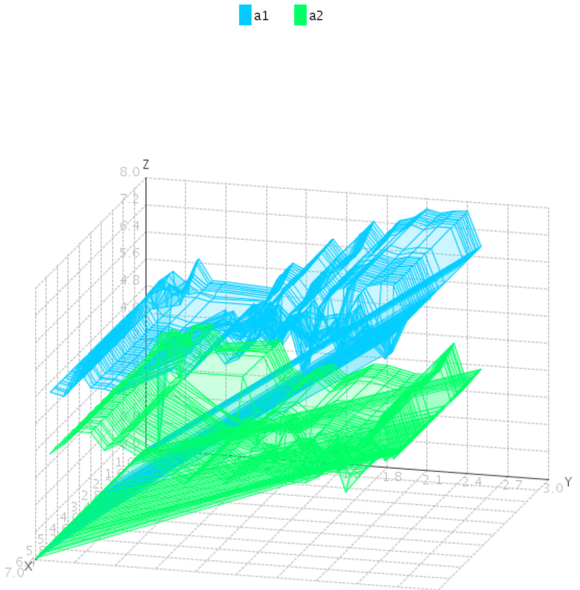
Visualisation examples – 3D plot V



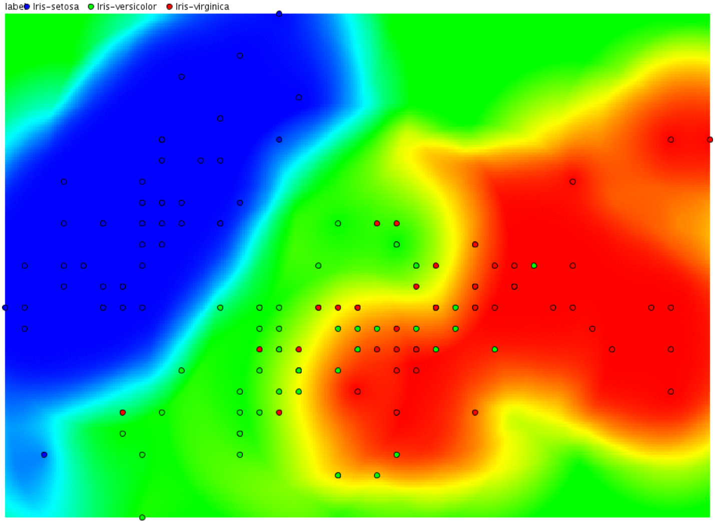
Visualisation examples – 3D plot VI



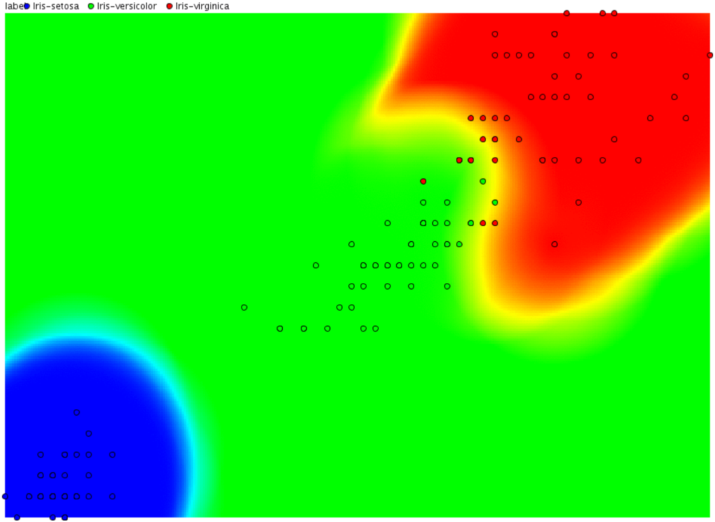
Visualisation examples – Surface 3D



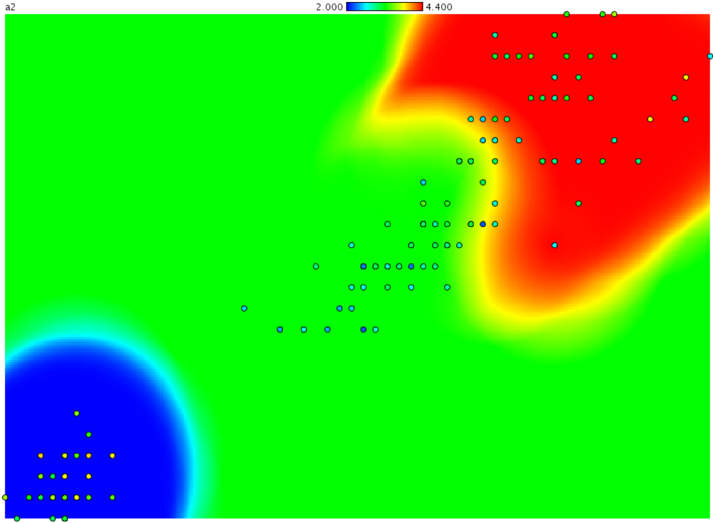
Visualisation examples – Density plots I



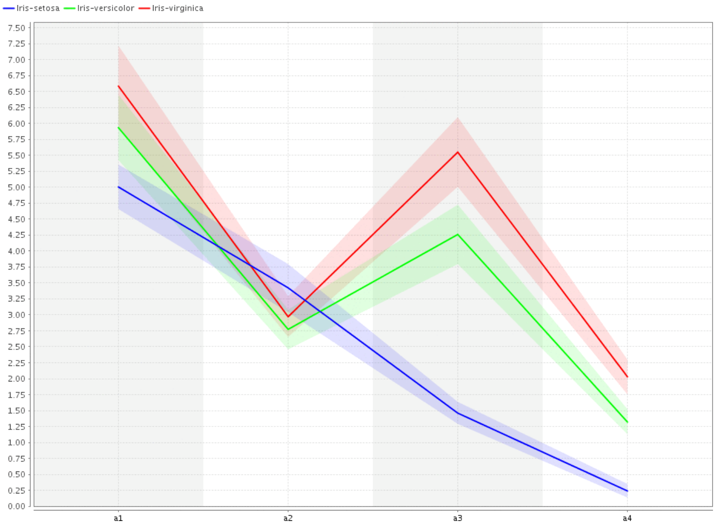
Visualisation examples – Density plots II



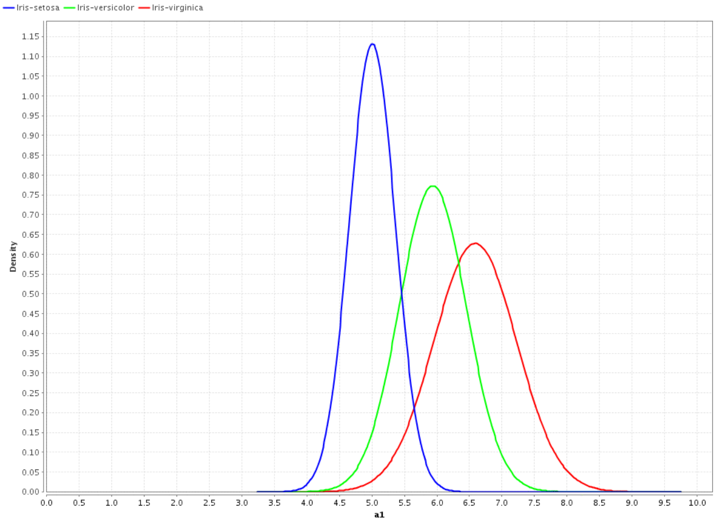
Visualisation examples – Density plots III



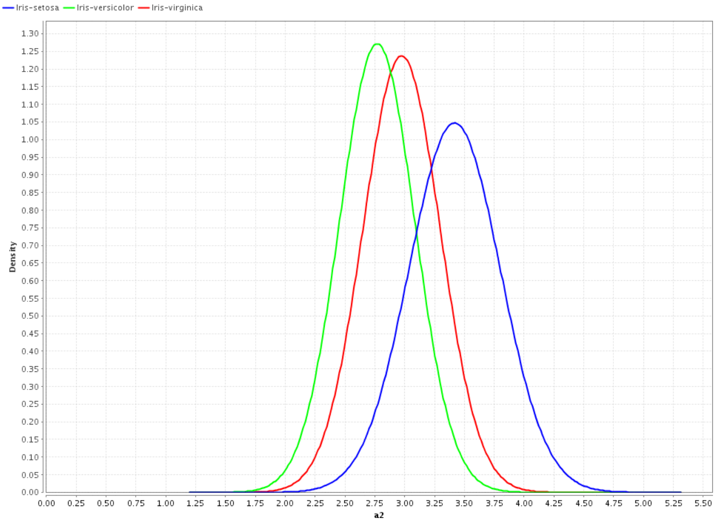
Visualisation examples – Deviation plot



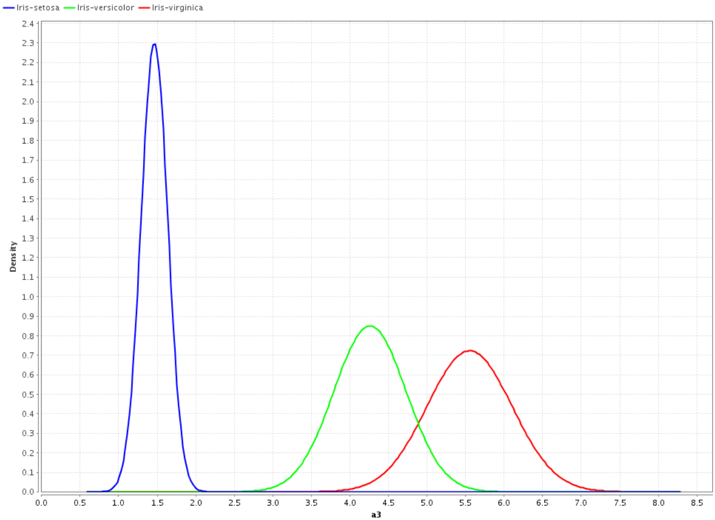
Visualisation examples – Distribution plots I



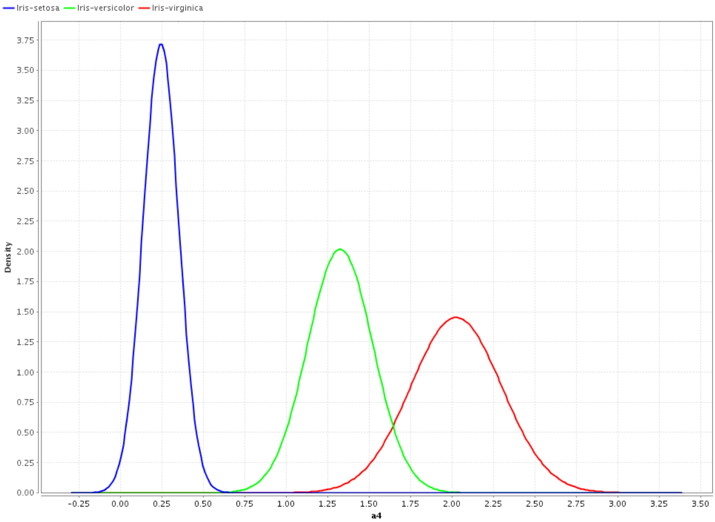
Visualisation examples – Distribution plots II



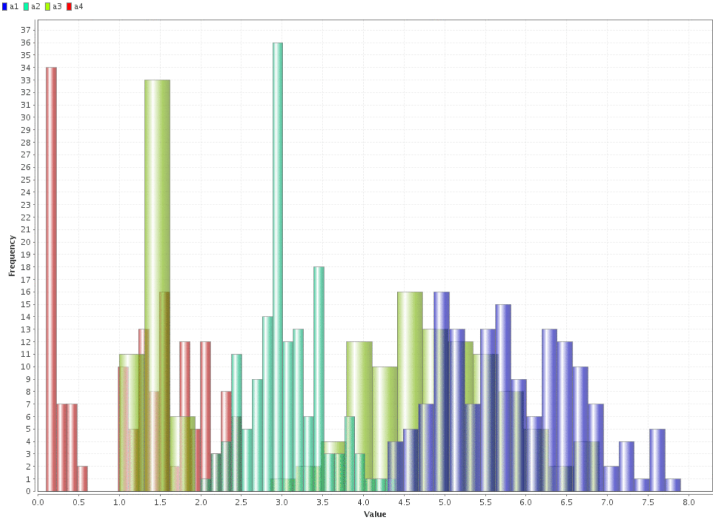
Visualisation examples – Distribution plots III



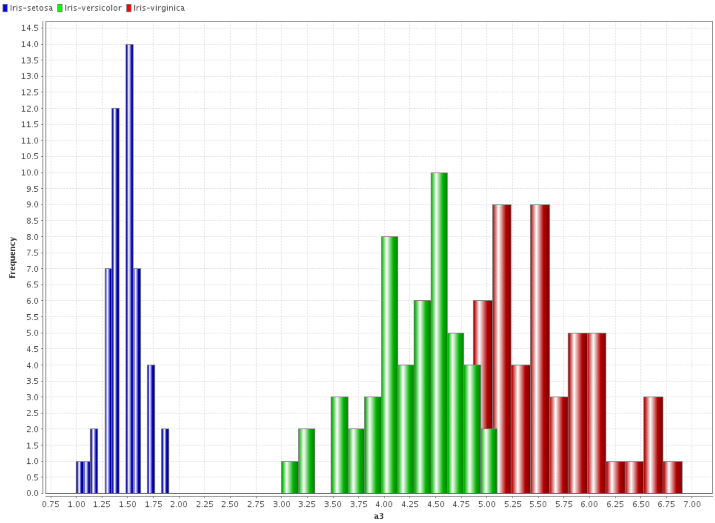
Visualisation examples – Distribution plots IV



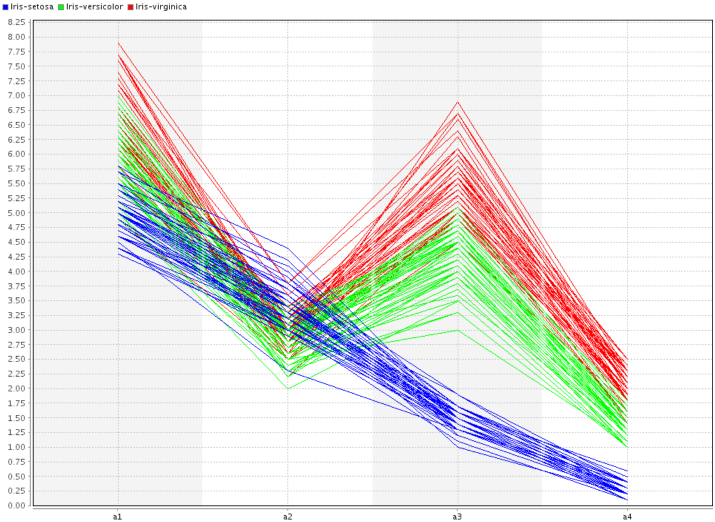
Visualisation examples – Histograms I



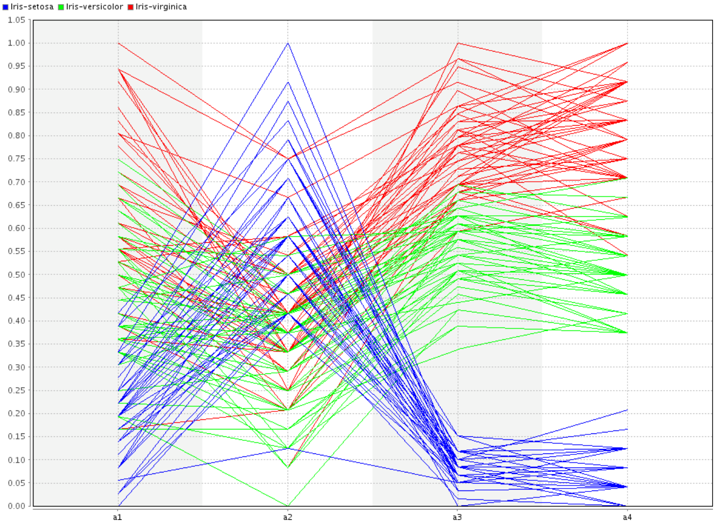
Visualisation examples – Histograms II



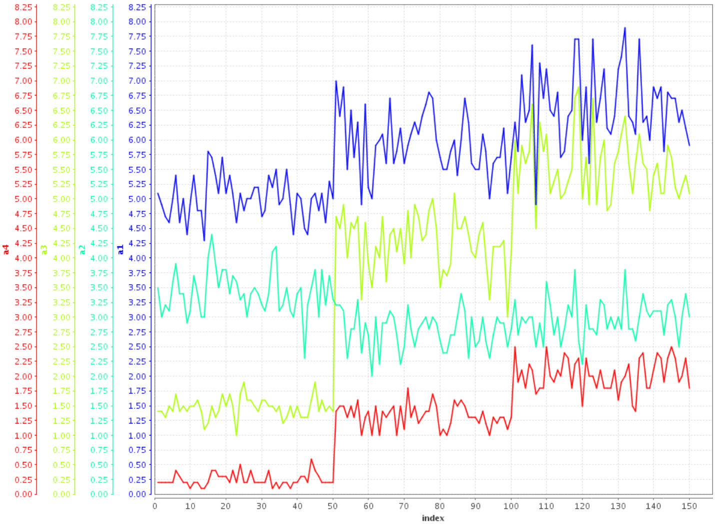
Visualisation examples – Parallel plot I



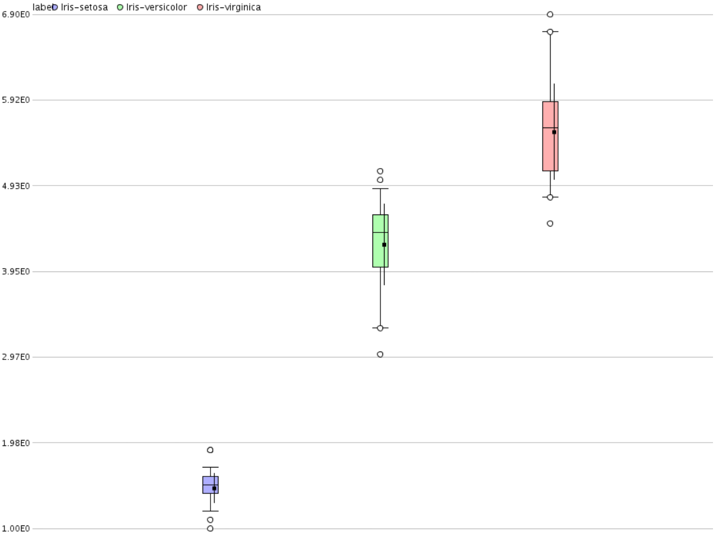
Visualisation examples – Parallel plot II



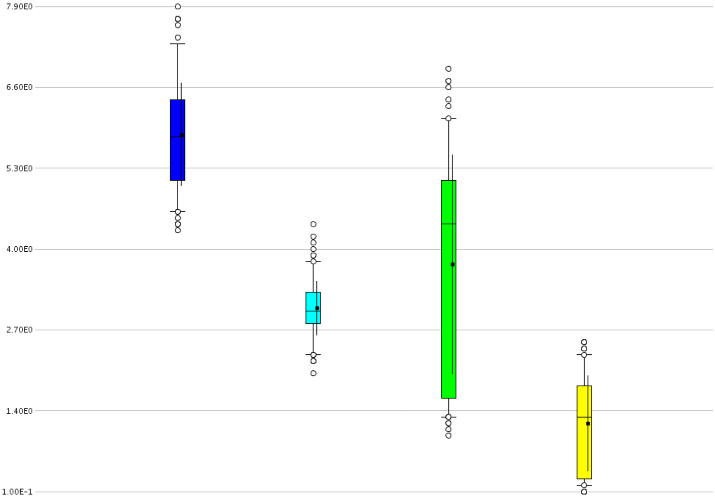
Visualisation examples – Series



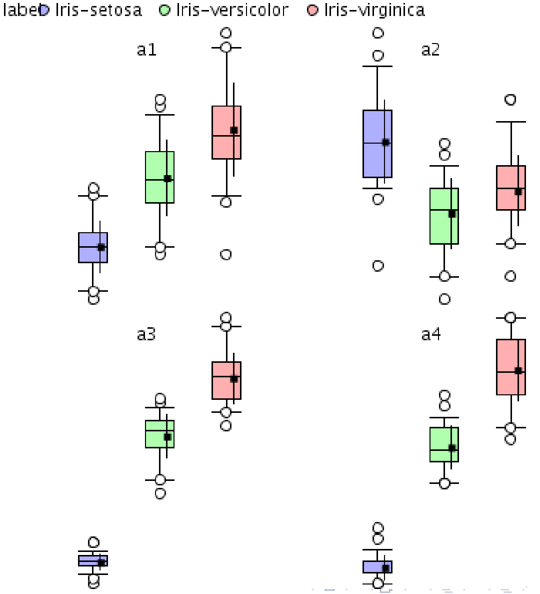
Visualisation examples – Quartiles I



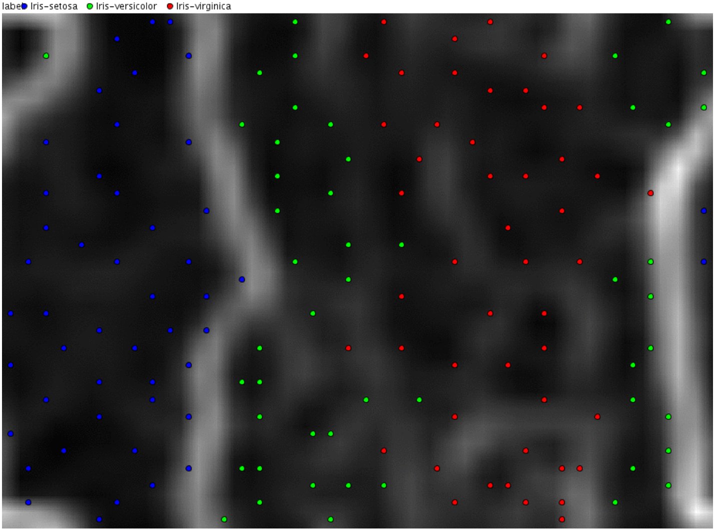
Visualisation examples – Quartiles II



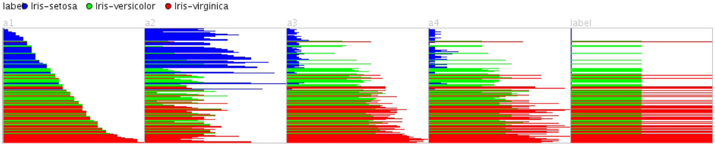
Visualisation examples – Quartiles III



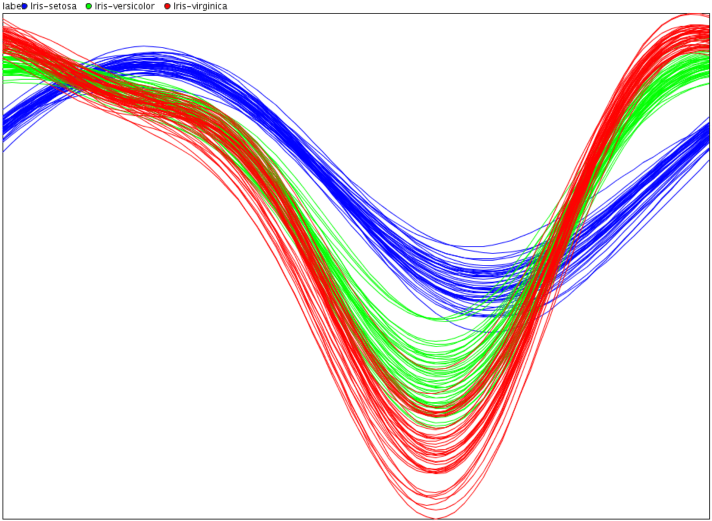
Visualisation examples – SOM – self-organising maps



Visualisation examples – Survey



Visualisation examples – Andrews curves



Software for data mining I

Statistical software

implements most of the statistical algorithms

- R http://www.r-project.org/, esp. packages rattle, RWeka and the CRAN Task View: Machine Learning & Statistical Learning

Software for data mining II

Data mining software

targeting the workflow and process setup

- RapidMiner: former YALE http://rapid-i.com/ Mierswa et al., 2006
- KNIME: Konstanz Information Miner http://knime.org/, runs in Eclipse IDE.
- WEKA: www.cs.waikato.ac.nz/ml/weka/ Witten and Frank, 2005; Hall et al., 2009

comparison of data mining software [german]: Lanig et al., 2010

Algorithms: overview

Main functions

- Classification: assign a class to a sample
- Clustering: identify groups of items (no a priori class information)
- Association: identify relationships
- Forecasting: estimates future values

Algorithms: what is a learner?

Machine learning algorithms

- branch of artificial intelligence
- evolve behaviors based on empirical data
- **learner**
  - takes examples
  - captures characteristics of interest of their unknown underlying probability distribution
- recognize complex patterns
- learner must generalize from the given examples

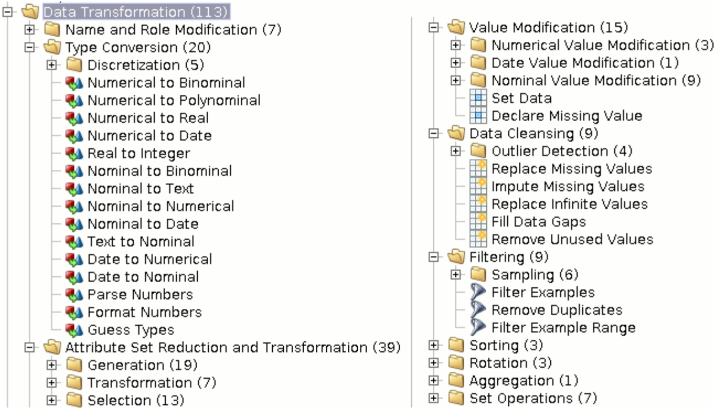
Source: WP:Machine_learning

Machine learning algorithms I

List of widely used ML algorithms

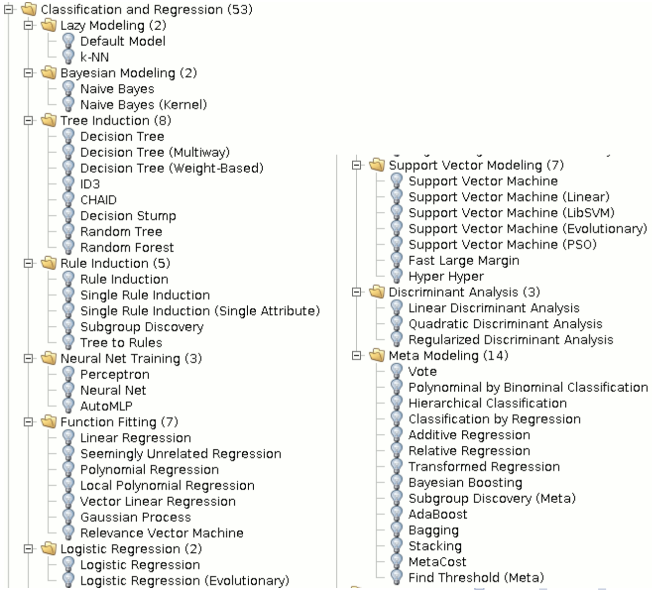
- decision-tree
- rule-based learning
- artificial neural network (ANN)
- bayesian learning
- support vector machines
- instance-based learning
- ensemble methods
- genetic algorithms
- graph-based learning
- regression

Rapidminer Operators – Data transformation



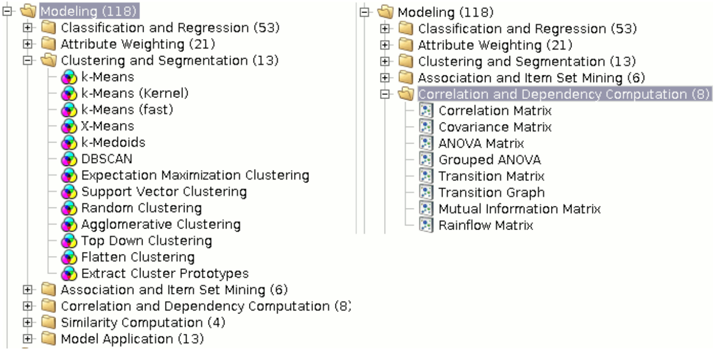
- Data Transformation (113)
  - Name and Role Modification (7)
  - Type Conversion (20)
    - Discretization (5)
    - Numerical to Binominal
    - Numerical to Polynominal
    - Numerical to Real
    - Numerical to Date
    - Real to Integer
    - Nominal to Binominal
    - Nominal to Text
    - Nominal to Numerical
    - Nominal to Date
    - Text to Nominal
    - Date to Numerical
    - Date to Nominal
    - Parse Numbers
    - Format Numbers
    - Guess Types
  - Attribute Set Reduction and Transformation (39)
    - Generation (19)
    - Transformation (7)
    - Selection (13)
  - Value Modification (15)
    - Numerical Value Modification (3)
    - Date Value Modification (1)
    - Nominal Value Modification (9)
    - Set Data
    - Declare Missing Value
  - Data Cleansing (9)
    - Outlier Detection (4)
    - Replace Missing Values
    - Impute Missing Values
    - Replace Infinite Values
    - Fill Data Gaps
    - Remove Unused Values
  - Filtering (9)
    - Sampling (6)
    - Filter Examples
    - Remove Duplicates
    - Filter Example Range
  - Sorting (3)
  - Rotation (3)
  - Aggregation (1)
  - Set Operations (7)

Rapidminer Operators – Classification and Regression



- Classification and Regression (53)
  - Lazy Modeling (2)
    - Default Model
    - k-NN
  - Bayesian Modeling (2)
    - Naive Bayes
    - Naive Bayes (Kernel)
  - Tree Induction (8)
    - Decision Tree
    - Decision Tree (Multiway)
    - Decision Tree (Weight-Based)
    - ID3
    - CHAID
    - Decision Stump
    - Random Tree
    - Random Forest
  - Rule Induction (5)
    - Rule Induction
    - Single Rule Induction
    - Single Rule Induction (Single Attribute)
    - Subgroup Discovery
    - Tree to Rules
  - Neural Net Training (3)
    - Perceptron
    - Neural Net
    - AutoMLP
  - Function Fitting (7)
    - Linear Regression
    - Seemingly Unrelated Regression
    - Polynomial Regression
    - Local Polynomial Regression
    - Vector Linear Regression
    - Gaussian Process
    - Relevance Vector Machine
  - Logistic Regression (2)
    - Logistic Regression
    - Logistic Regression (Evolutionary)
  - Support Vector Modeling (7)
    - Support Vector Machine
    - Support Vector Machine (Linear)
    - Support Vector Machine (LibSVM)
    - Support Vector Machine (Evolutionary)
    - Support Vector Machine (PSO)
    - Fast Large Margin
    - Hyper Hyper
  - Discriminant Analysis (3)
    - Linear Discriminant Analysis
    - Quadratic Discriminant Analysis
    - Regularized Discriminant Analysis
  - Meta Modeling (14)
    - Vote
    - Polynominal by Binominal Classification
    - Hierarchical Classification
    - Classification by Regression
    - Additive Regression
    - Relative Regression
    - Transformed Regression
    - Bayesian Boosting
    - Subgroup Discovery (Meta)
    - AdaBoost
    - Bagging
    - Stacking
    - MetaCost
    - Find Threshold (Meta)

Rapidminer Operators – Clustering and correlation



- Modeling (118)
  - Classification and Regression (53)
  - Attribute Weighting (21)
  - Clustering and Segmentation (13)
    - k-Means
    - k-Means (Kernel)
    - k-Means (fast)
    - X-Means
    - k-Medoids
    - DBSCAN
    - Expectation Maximization Clustering
    - Support Vector Clustering
    - Random Clustering
    - Agglomerative Clustering
    - Top Down Clustering
    - Flatten Clustering
    - Extract Cluster Prototypes
  - Association and Item Set Mining (6)
  - Correlation and Dependency Computation (8)
  - Similarity Computation (4)
  - Model Application (13)

- Modeling (118)
  - Classification and Regression (53)
  - Attribute Weighting (21)
  - Clustering and Segmentation (13)
  - Association and Item Set Mining (6)
  - Correlation and Dependency Computation (8)
    - Correlation Matrix
    - Covariance Matrix
    - ANOVA Matrix
    - Grouped ANOVA
    - Transition Matrix
    - Transition Graph
    - Mutual Information Matrix
    - Rainflow Matrix

Simple algorithms - try first

Simple structures

- one attribute does all
- all attributes equally involved
- weighted linear combination
- simple rules
- instance-based: prototypes

Example Operator: k-NN

Synopsis Classification with k-NN based on an explicit similarity measure.

Description A k nearest neighbor implementation.

Input training set: expects: ExampleSet, expects: ExampleSet

Output model: exampleSet:

Parameters

k The used number of nearest neighbors. Range: integer; 1-+?; default: 1

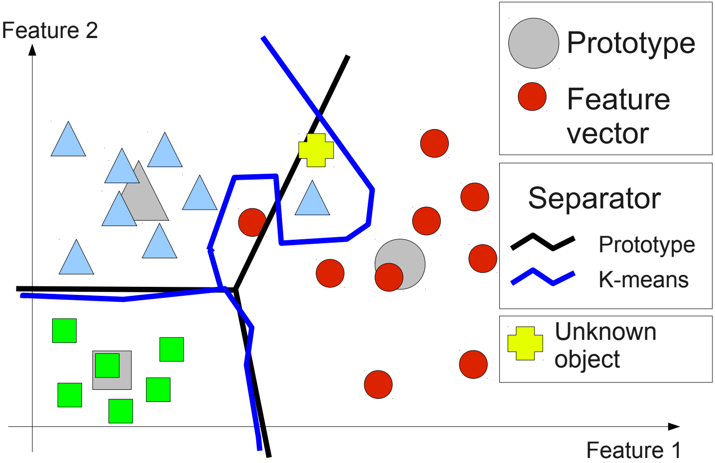
weighted vote Indicates if the votes should be weighted by similarity

measure types The measure type Range

mixed measure Select measure Range

⋮

k-NN/Prototype classification



Data mining process I

Preprocessing of data

- is the data set complete? missing values?
- normalisation
- type of data: change necessary?
- label data: appropriate?
- large data sets: sub-sampling, stratification
- separate training data, test data

Data mining process II

Selection of process

- data types can be handled by algorithms?
- which algorithms are suitable
- how can they be learned?
- start with simple algorithms, progress to more complex ones

Data mining process III

Attribute selection/weighting

- reduce the dimensionality of the feature space (curse of dimensionality)
- select only useful attributes
- attribute weighting
- forward/backward selection strategies
- selection strategy according to learner (e.g. use learner/cross-validation)

Learning procedure

- use training data to learn → generate model

Data mining process IV

Evaluation of model

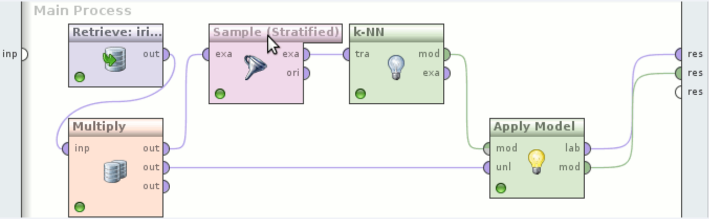
- cross-validation
- numerical measures describing the model fit
- contingency matrices, errors type I and II
- overfitting?

Application to new data

- use model for new data

Rapidminer: k-NN I

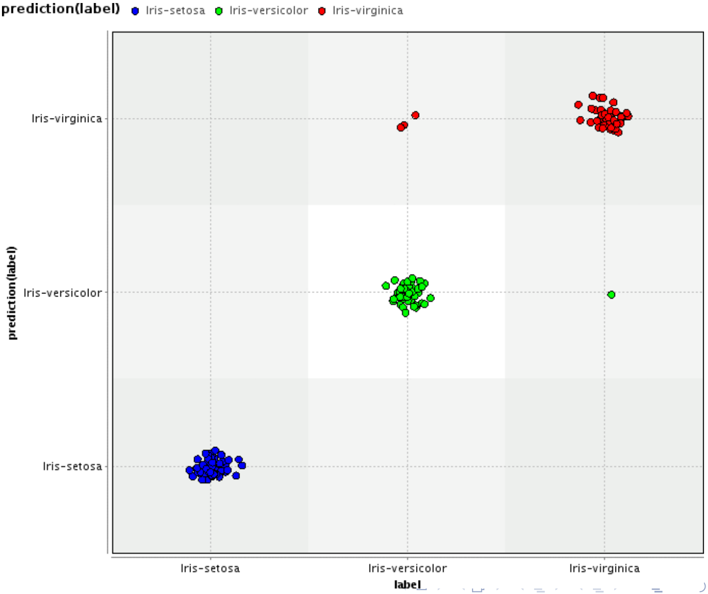
Rapidminer process for k-NN



Instance-based classification

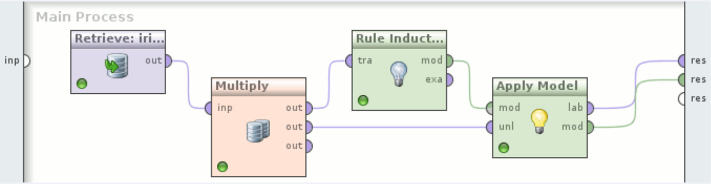
1-Nearest Neighbour model for classification. The model contains 15 examples with 4 dimensions of the following classes: Iris-setosa Iris-versicolor Iris-virginica

Rapidminer: k-NN II



Rule induction I

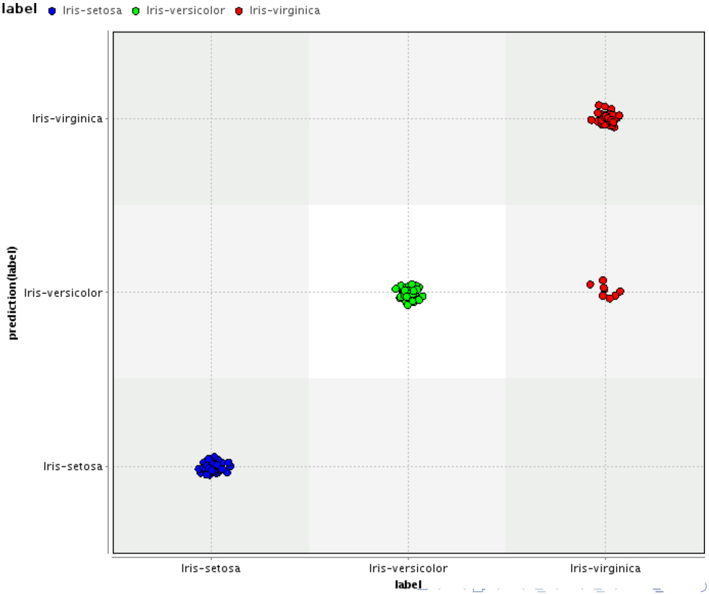
Rapidminer process for rule induction



Rules

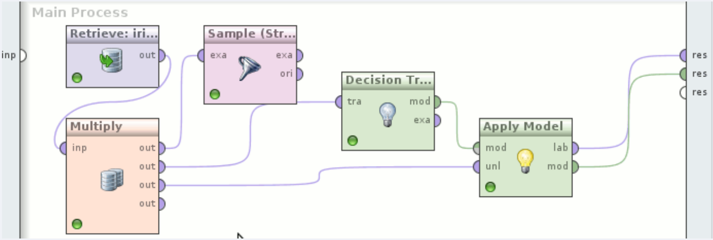
if a3 $\leq$ 2.450 then Iris-setosa (50/0/0)
if a3 $\leq$ 5.150 and a4 $\leq$ 1.850 then Iris-versicolor (0/50/8)
else Iris-virginica (0/0/37)
correct: 137 out of 145 training examples.

Rule induction II



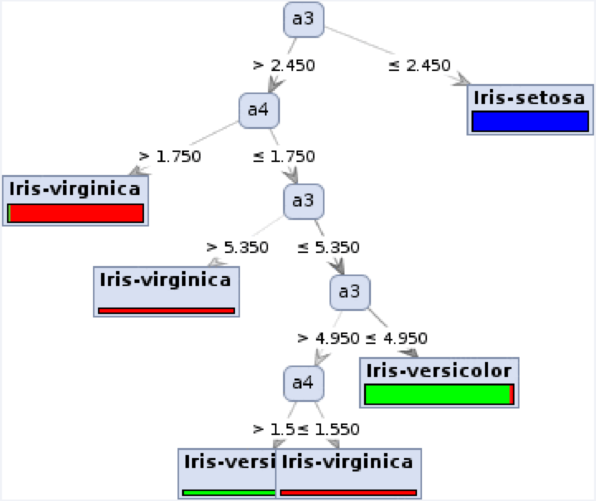
Decision Tree I

Rapidminer process for tree induction

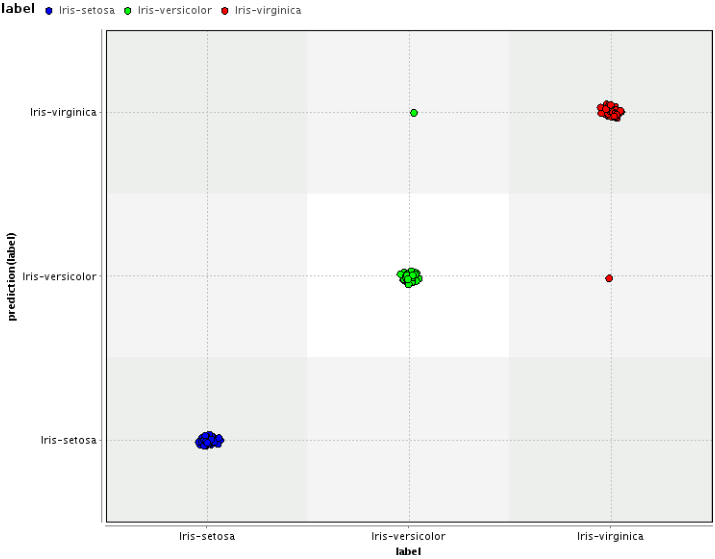


Decision Tree II

Tree

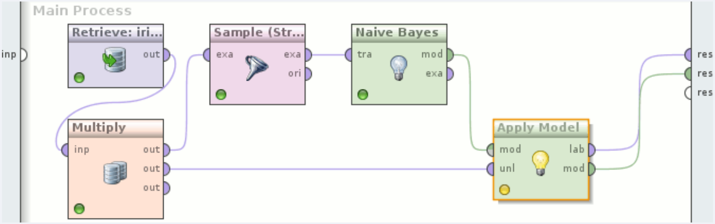


Decision Tree III



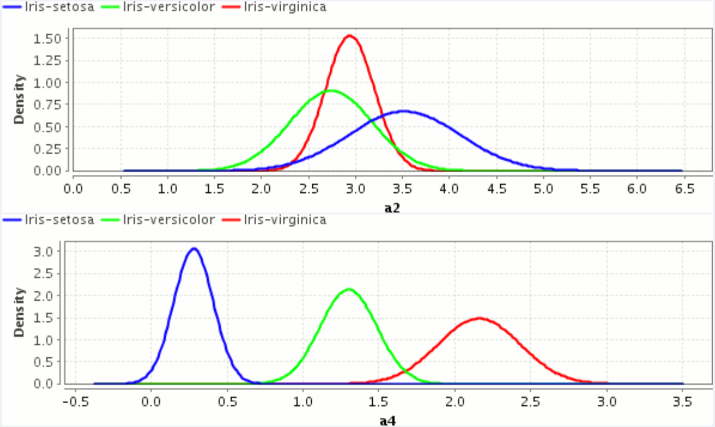
Rapidminer: naive Bayes I

Rapidminer process for naive Bayes



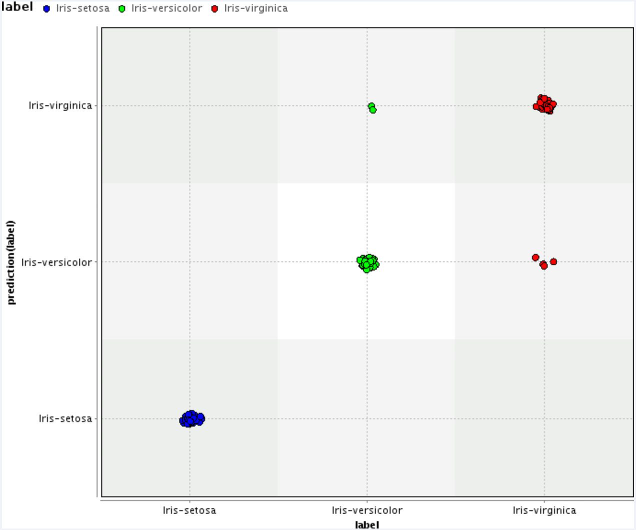
Rapidminer: naive Bayes II

Model

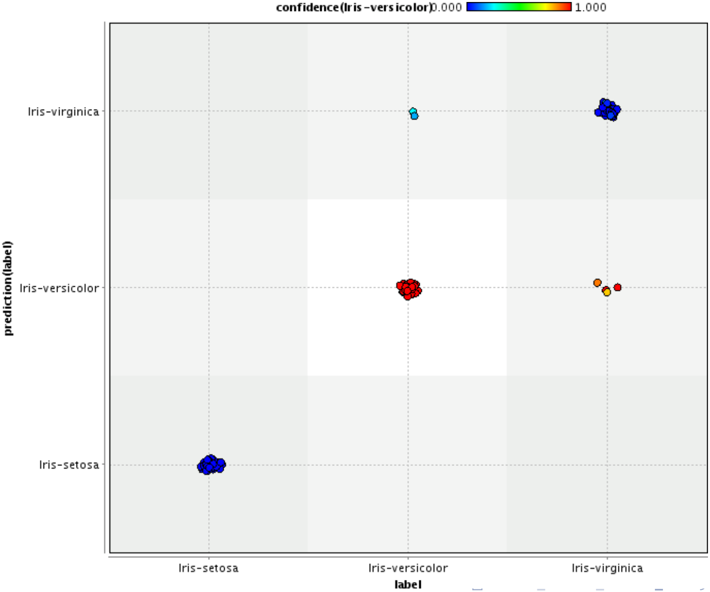


Rapidminer: naive Bayes III

Prediction



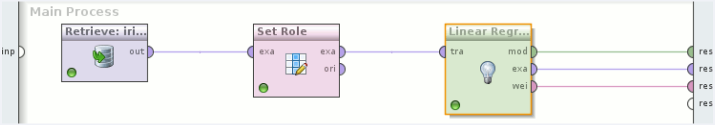
Rapidminer: naive Bayes (confidence)



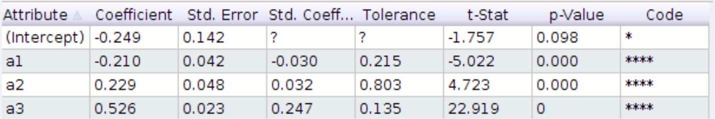
Linear regression I

Rapidminer process for linear regression

Change roles: set label to ‘labelunused’ and attribute a4 (numerical) as label (target)



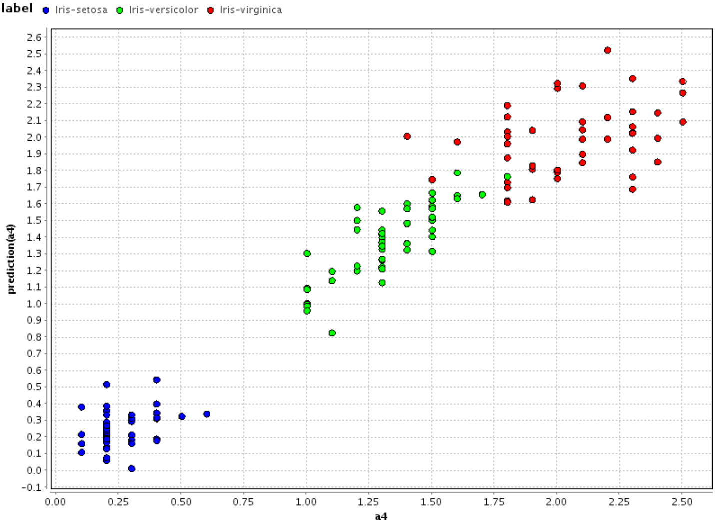
Result



Attribute	Coefficient	Std. Error	Std. Coeff...	Tolerance	t-Stat	p-Value	Code
(Intercept)	-0.249	0.142	?	?	-1.757	0.098	*
a1	-0.210	0.042	-0.030	0.215	-5.022	0.000	****
a2	0.229	0.048	0.032	0.803	4.723	0.000	****
a3	0.526	0.023	0.247	0.135	22.919	0	****

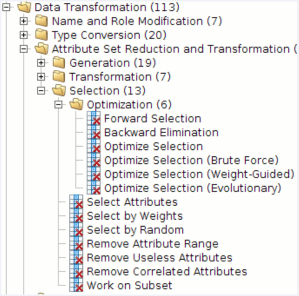
$$a_4 = -0.210a_1 + 0.229a_2 + 0.526a_3 - 0.249$$

Linear regression II



Feature (attribute) selection I

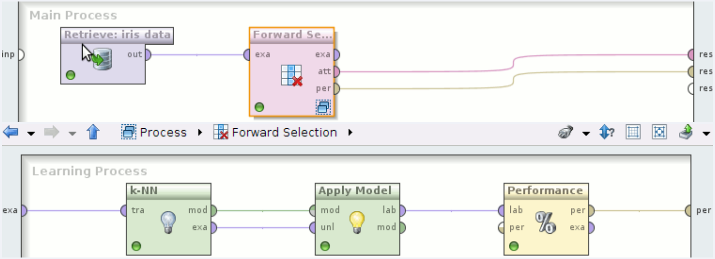
Feature selection algorithms



- Data Transformation (113)
  - Name and Role Modification (7)
  - Type Conversion (20)
  - Attribute Set Reduction and Transformation (
    - Generation (19)
    - Transformation (7)
    - Selection (13)
      - Optimization (6)
        - Forward Selection
        - Backward Elimination
        - Optimize Selection
        - Optimize Selection (Brute Force)
        - Optimize Selection (Weight-Guided)
        - Optimize Selection (Evolutionary)
      - Select Attributes
      - Select by Weights
      - Select by Random
      - Remove Attribute Range
      - Remove Useless Attributes
      - Remove Correlated Attributes
      - Work on Subset

Feature (attribute) selection II

Process for featureselection



Feature (attribute) selection III

Confusion matrix (Performance)

accuracy: 98.00%

True:	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa:	50	0	0
Iris-versicolor:	0	49	2
Iris-virginica:	0	1	48

Attributte weights

a1, a2	0
a3, a4	1

Data mining literature

Books

- Petersohn, 2005: *Data Mining: Verfahren, Prozesse, Anwendungsarchitektur*
- Witten and Frank, 2005: *Data Mining: Practical machine learning tools and techniques*

Articles

- Hall et al., 2009: “The WEKA data mining software: an update”
- Mierswa et al., 2006: “YALE: Rapid Prototyping for Complex Data Mining Tasks”
- Lanig et al., 2010: *Evaluation von Data Mining Werkzeugen*

Bibliography I

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten (2009). "The WEKA data mining software: an update". In: *SIGKDD Explor. Newsl.* 11 (1 Nov. 2009), pp. 10–18. ISSN: 1931-0145. DOI: 10.1145/1656274.1656278.

Lanig, S., M. Lemcke, and P. Mayer (2010). *Evaluation von Data Mining Werkzeugen*. ger. Tech. rep. Holzgartenstr. 16, 70174 Stuttgart: Institut für Visualisierung und Interaktive Systeme, 2010.

Bibliography II

Mierswa, I., M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler (2006). "YALE: Rapid Prototyping for Complex Data Mining Tasks". In: *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*. Ed. by L. Ungar, M. Craven, D. Gunopulos, and T. Eliassi-Rad. Philadelphia, PA, USA: ACM, Aug. 2006, pp. 935–940. ISBN: 1-59593-339-5. DOI: http://doi.acm.org/10.1145/1150402.1150531.

Petersohn, H. (2005). *Data Mining: Verfahren, Prozesse, Anwendungsarchitektur*. 1st ed. Oldenbourg Wissenschaftsverlag, 2005, p. 342. ISBN: 3486577158.

Witten, I. H. and E. Frank (2005). *Data Mining: Practical machine learning tools and techniques*. 2nd ed. San Francisco: Morgan Kaufmann, 2005.