Vorlesung Wissensentdeckung

Einführung

Katharina Morik, Claus Weihs

LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

5.4.2011

Gliederung

Problem

- Wir haben nur eine endliche Menge von Beispielen. Alle Funktionen, deren Werte durch die Beispiele verlaufen, haben einen kleinen Fehler.
- Wir wollen aber für alle Beobachtungen das richtige y voraussagen. Dann sind nicht mehr alle Funktionen, die auf die Beispiele gepasst haben, gut.
- Wir kennen nicht die wahre Verteilung der Beispiele.
- Wie beurteilen wir da die Qualität unseres Lernergebnisses?

Lern- und Testmenge

Wir teilen die Daten, die wir haben, auf:

Lernmenge: Einen Teil der Daten übergeben wir unserem Lernalgorithmus. Daraus lernt er seine Funktion $f(x) = \hat{y}$.

Testmenge: Bei den restlichen Daten vergleichen wir $\hat{y}$ mit y.

Aufteilung in Lern- und Testmenge

- Vielleicht haben wir zufällig aus lauter Ausnahmen gelernt und testen dann an den normalen Fällen. Um das zu vermeiden, verändern wir die Aufteilung mehrfach.

  leave-one-out: Der Algorithmus lernt aus $N-1$ Beispielen und testet auf dem ausgelassenen. Dies wird N mal gemacht, die Fehler addiert.

- Aus Zeitgründen wollen wir den Algorithmus nicht zu oft anwenden.

  Kreuzvalidierung: Die Lernmenge wird zufällig in n Mengen aufgeteilt. Der Algorithmus lernt aus $n-1$ Mengen und testet auf der ausgelassenen Menge. Dies wird n mal gemacht.

Kreuzvalidierung

- Man teile alle verfügbaren Beispiele in n Mengen auf. z.B. $n = 10$.
- Für i=1 bis i=n:
  - Wähle die i-te Menge als Testmenge,
  - die restlichen $n - 1$ Mengen als Lernmenge.
  - Messe die Qualität auf der Testmenge.
- Bilde das Mittel der gemessenen Qualität über allen n Lernläufen. Das Ergebnis gibt die Qualität des Lernergebnisses an.

Was wissen Sie jetzt?

- Sie haben Anwendungsbeispiele gesehen.
- Als Aufgaben der Modellbildung haben Sie Clustering, Klassifikation, Regression gesehen.
- Sie wissen, was die Kreuzvalidierung ist.

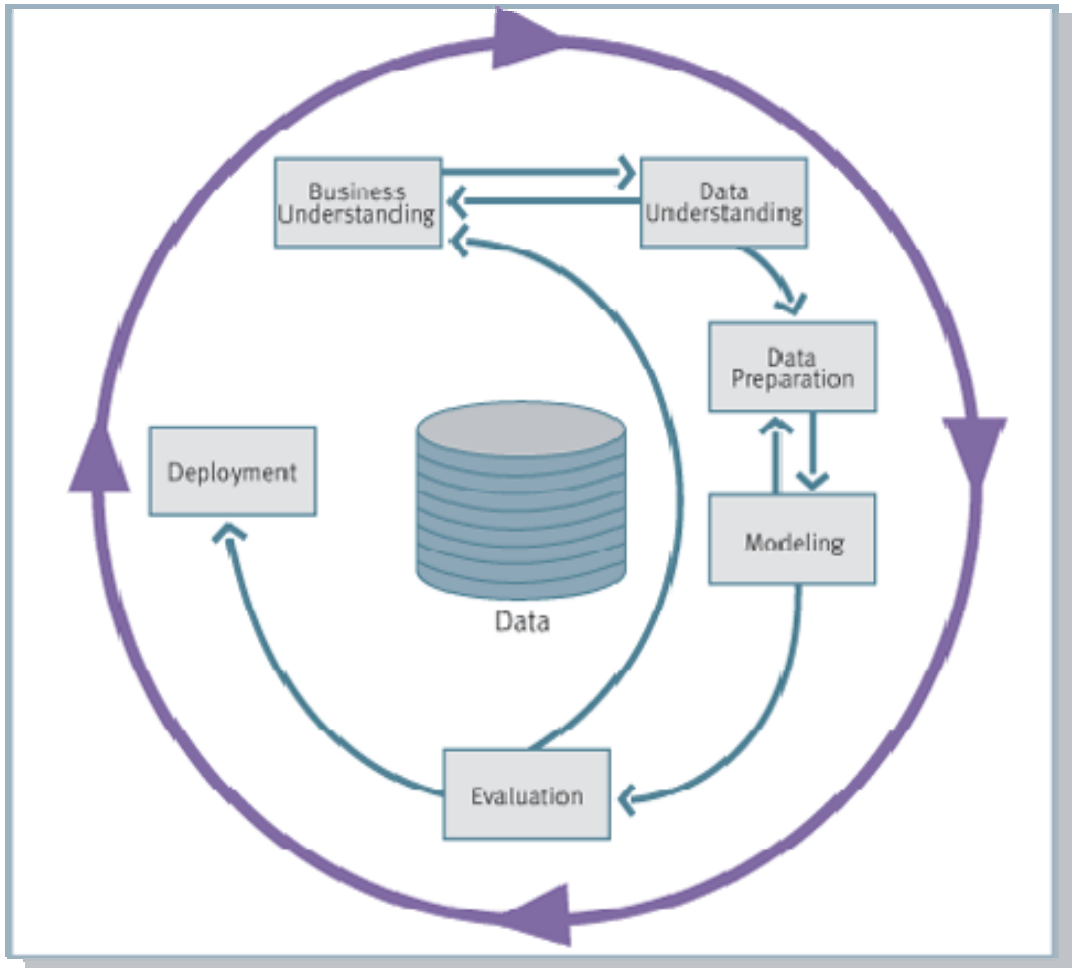
Was wissen Sie noch nicht?

- Es gibt viele verschiedene Modellklassen. Damit werden die Lernaufgaben spezialisiert.
- Es gibt unterschiedliche Qualitätsfunktionen. Damit werden die Lernaufgaben als Optimierungsaufgaben definiert.
- Es gibt auch noch mehr Aufgaben: Finden häufiger Mengen!
- Der Gesamtablauf des Data Mining hat eine feste Struktur, die mehr enthält als nur den Lernschritt.
- Das Ziehen von Stichproben und wie diese verwendet werden können, lernen Sie kennen.
- Die Dimensionsreduktion ist ein wichtiger Vorverarbeitungsschritt.

CRISP-DM: CRoss Industry Standard Process for Data Mining (http://www.crisp-dm.org)

- Zusammenarbeit von NCR, SPSS und DaimlerChrysler
- NCR: Mehrwert für Data Warehouse Kunden
- SPSS: Konzept für Data Mining Produkt 'Clementine'
- DaimlerChrysler: Praktische Erfahrung
- KEINE theoretische, akademische Entwicklung,
- SONDERN Entwicklung aus praktischer Erfahrung an realen Problemen.

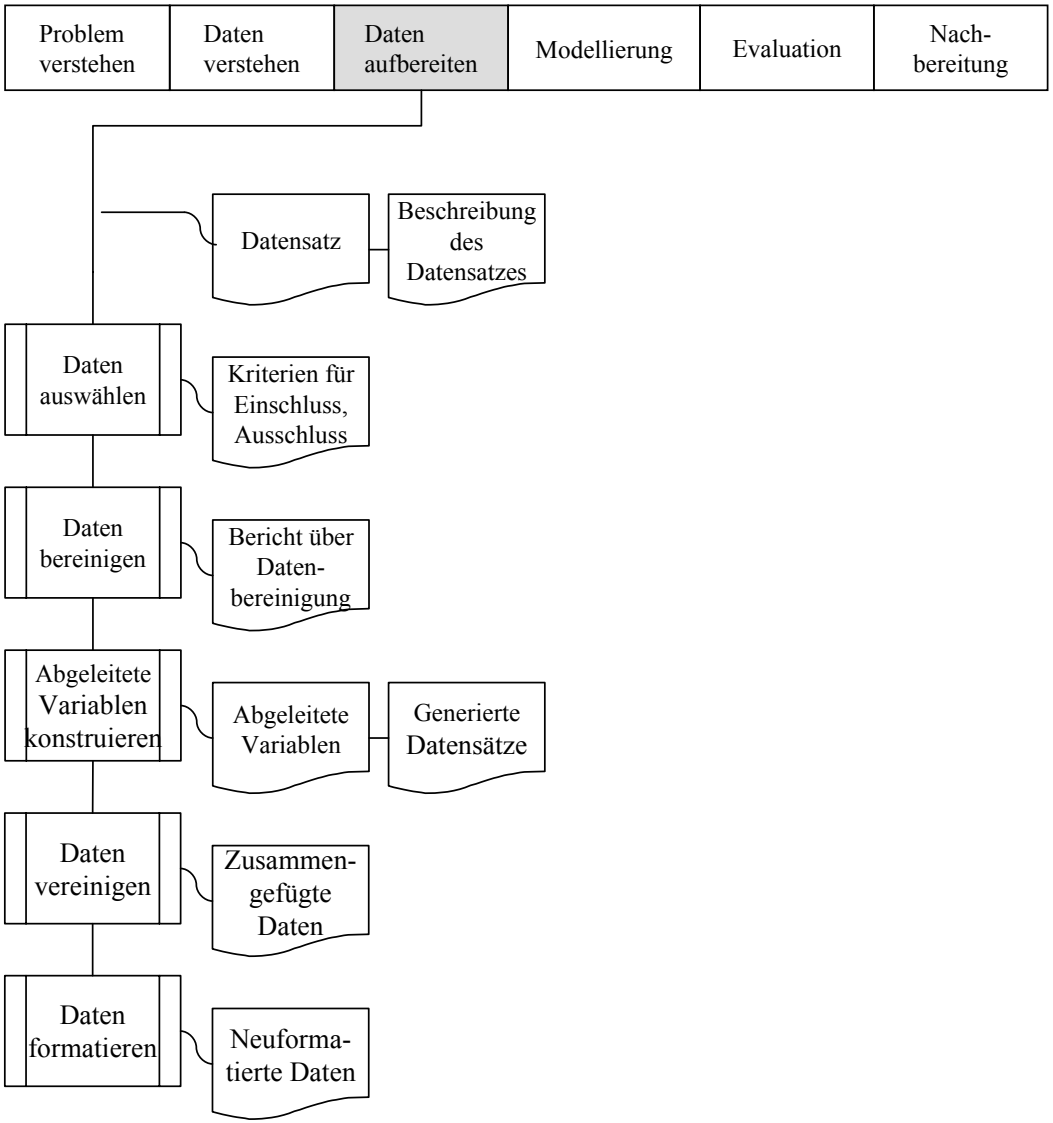
Übersicht


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

CRISP: Schritte

- **Problem verstehen**: Analyseziele, Situationsbewertung, Datenanalyseziele, Projektplan
- **Daten verstehen**: Sammeln, beschreiben, untersuchen, Qualität von Rohdaten
- **Daten aufbereiten**: Ein- und Ausschluss, Bereinigung, Transformation von Variablen
- **Modellierung**: Methoden- und Testdesignwahl, Schätzung, Modellqualität
- **Evaluierung**: Modell akzeptieren, Prozess überprüfen, nächste Schritte
- **Nachbereitung**: Anwendungs- und Wartungsplan, Präsentation, Bericht

Vorverarbeitung

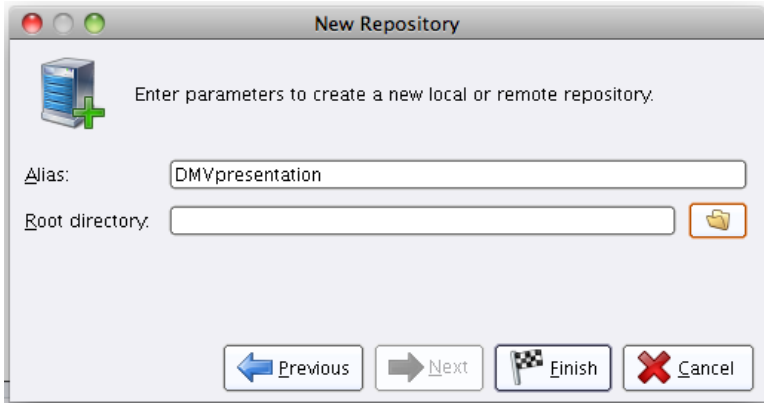


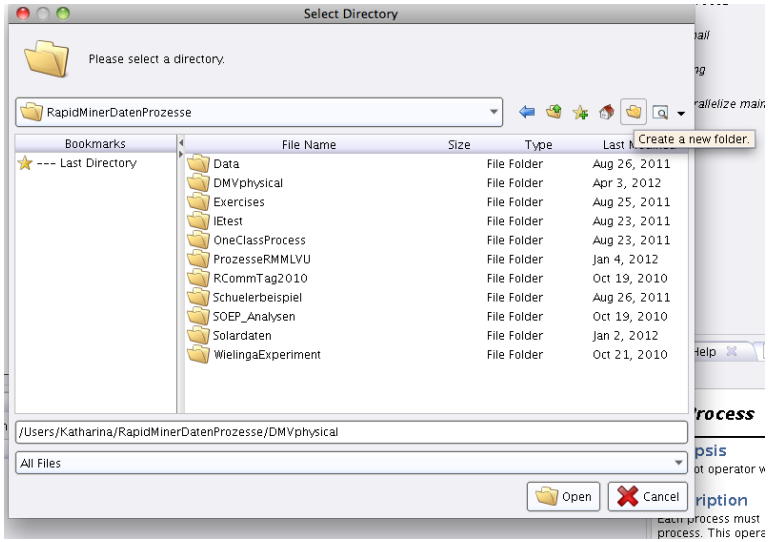
RapidMiner ist eine Umgebung, die den gesamten Prozess unterstützt.

- Operatoren für alle Verarbeitungsschritte,
- Prozess wird mit allen Parametern etc. dokumentiert –> Reproduzierbarkeit!
- Leicht zu erweitern durch eigene Operatoren, plug-ins.

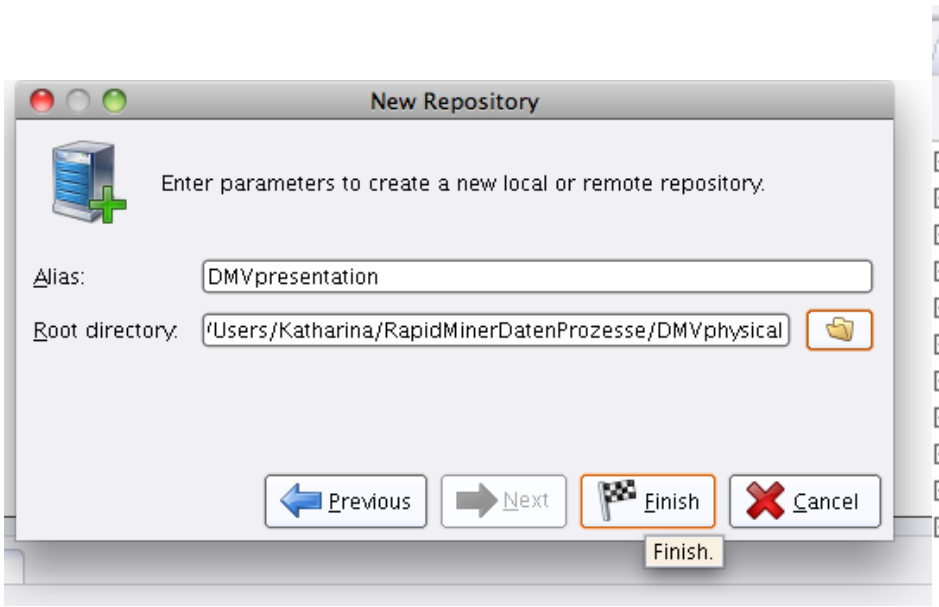
Prozesse und Daten speichern

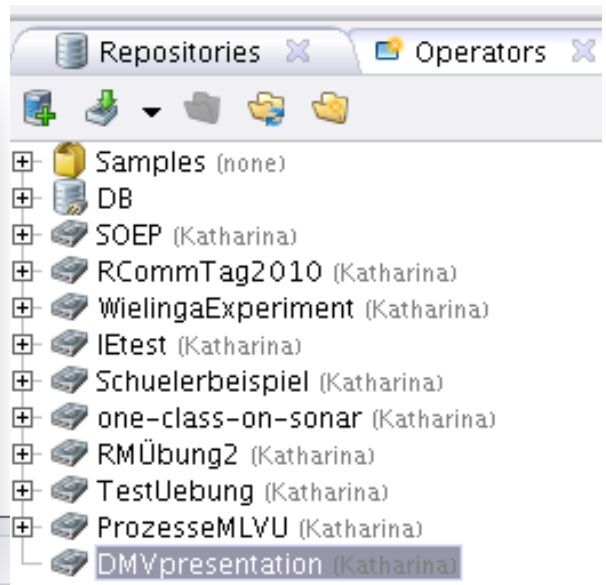
Für Ihre Daten und Prozesse müssen Sie einen Ort auf Ihrem Rechner vorsehen. Dieser Ort erhält einen symbolischen Namen als *Alias*und wird physikalisch bezeichnet durch ein *Root directory*.



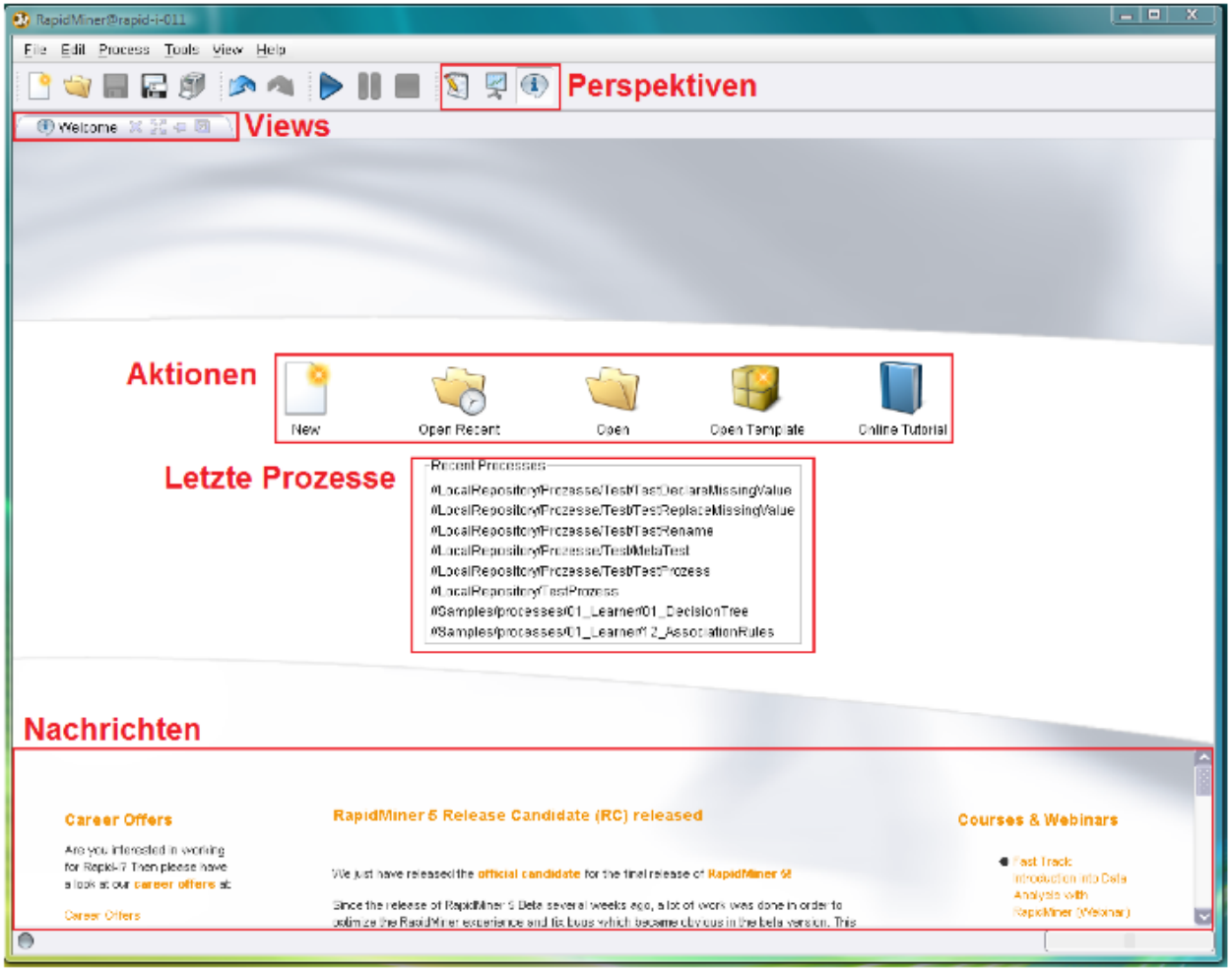


Repository


New Repository
Enter parameters to create a new local or remote repository.
Alias:
DMVpresentation
Root directory:
/Users/Katharina/RapidMinerDatenProzesse/DMVphysical
Previous
Next
Finish
Cancel
Finish.

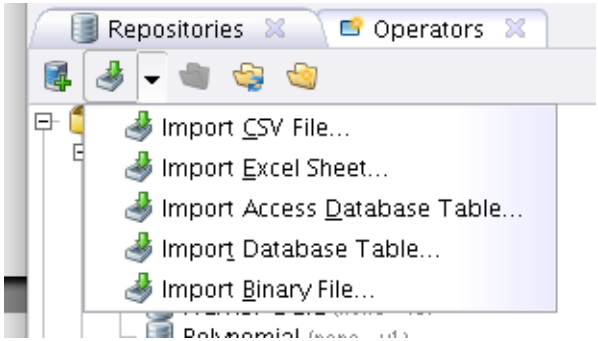

Repositories
Operators
Samples (none)
DB
SOEP (Katharina)
RCommTag2010 (Katharina)
WielingaExperiment (Katharina)
IEtest (Katharina)
Schuelerbeispiel (Katharina)
one-class-on-sonar (Katharina)
RMÜbung2 (Katharina)
TestUebung (Katharina)
ProzesseMLVU (Katharina)
DMVpresentation (Katharina)

RapidMiner Ansicht

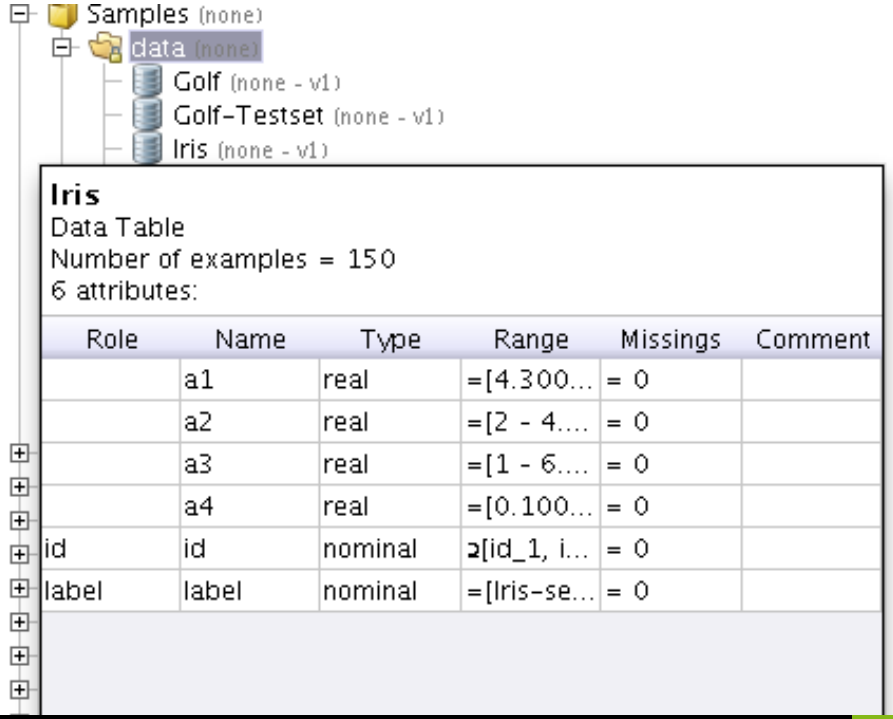


Daten einlesen

Einlesen ist für viele Formate vorbereitet.



Vorhandene Daten haben immer Metadaten dabei!

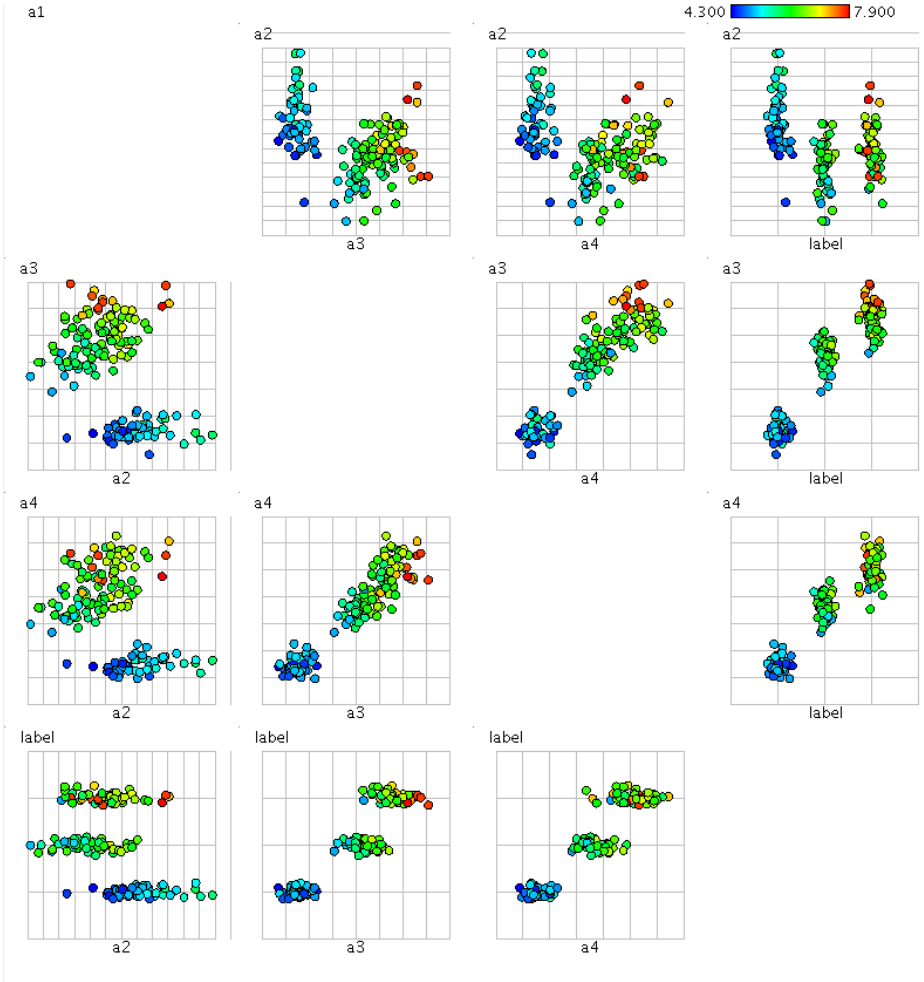


Iris
Data Table
Number of examples = 150
6 attributes:

Role	Name	Type	Range	Missings	Comment
	a1	real	=[4.300...	= 0	
	a2	real	=[2 - 4....	= 0	
	a3	real	=[1 - 6....	= 0	
	a4	real	=[0.100...	= 0	
id	id	nominal	ↄ[id_1, i...	= 0	
label	label	nominal	=[Iris-se...	= 0	

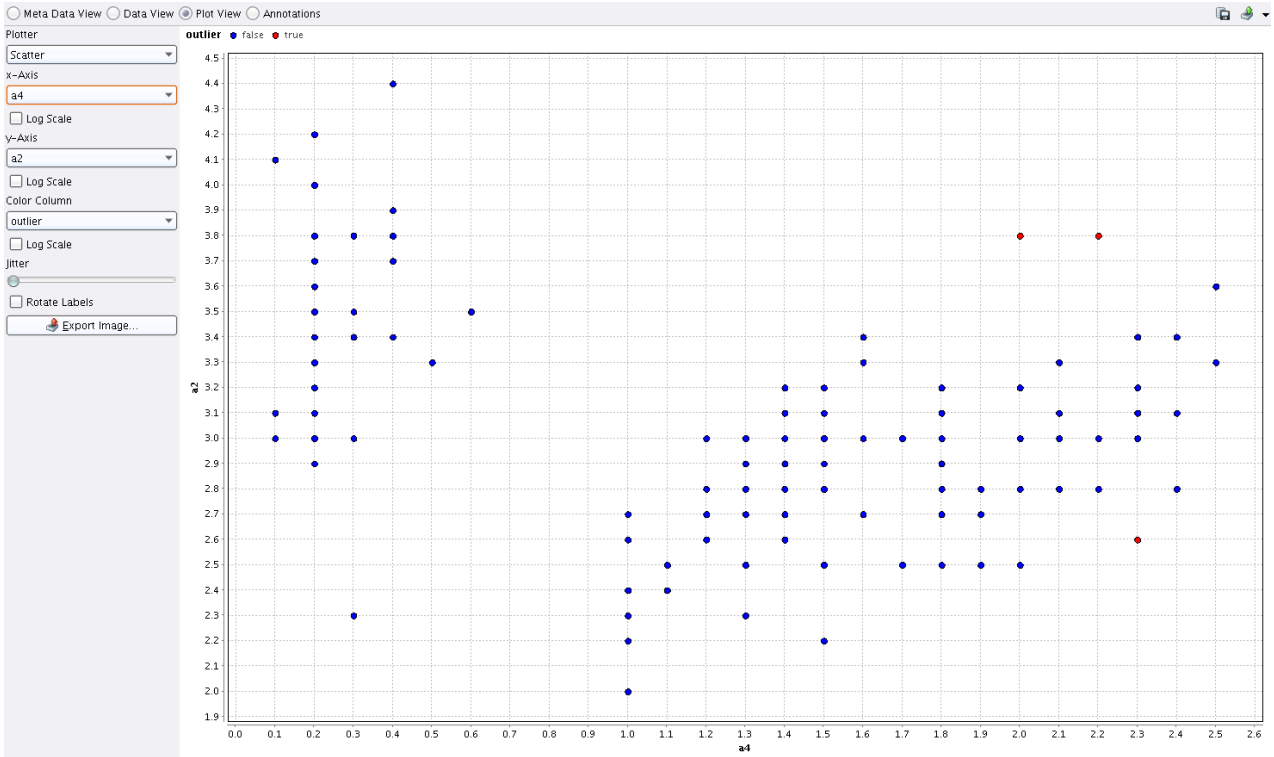
Daten ansehen

Retrieve, Prozess ablaufen, Ergebnisperspektive, Plot, z.B. Scatter Matrix (a1 als Farbe, Korrelation der anderen).



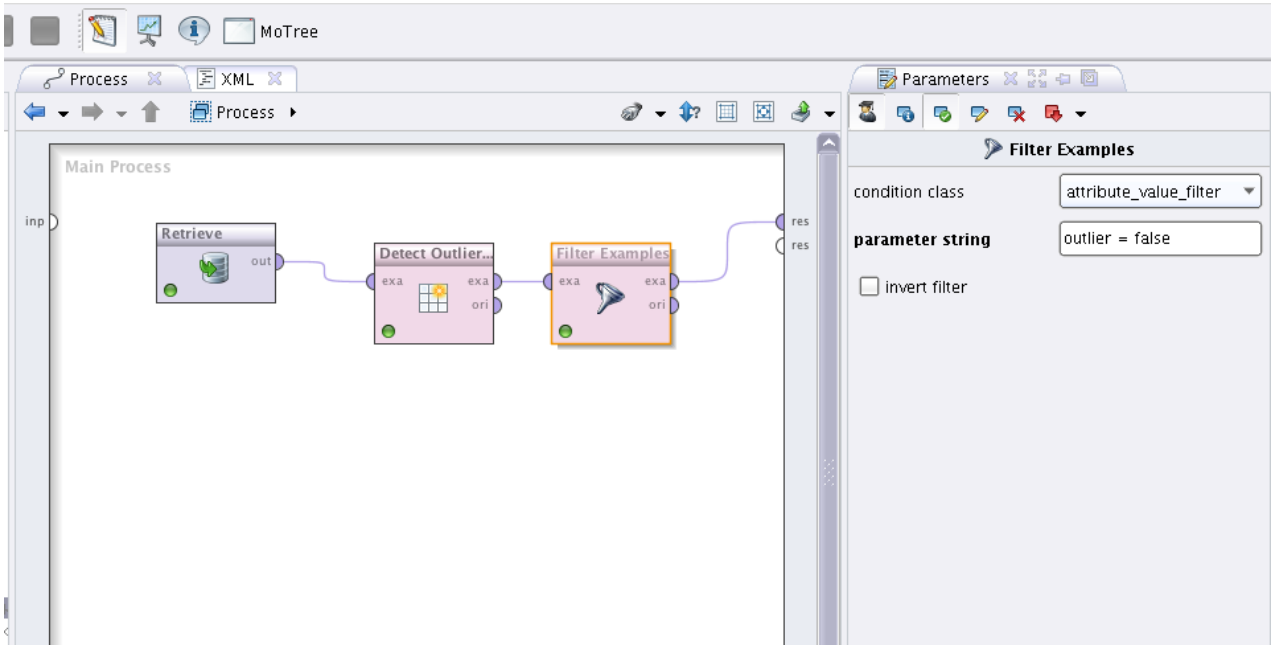
Ausreisser entdecken

RapidMiner hat unter Data Transformation/Data Cleansing Methoden zur Ausreissererkennung, hier Distanz-basiert.



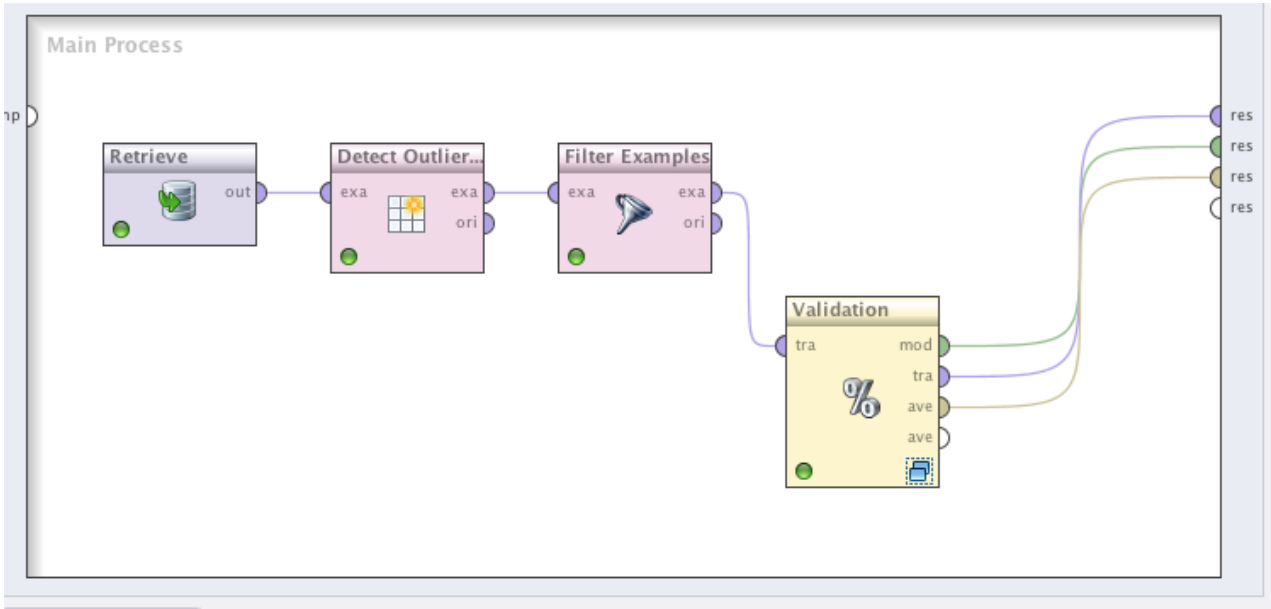
Ausreisser entfernen

Ausreisser sind durch *true* in der neuen Spalte *Outlier* gekennzeichnet.



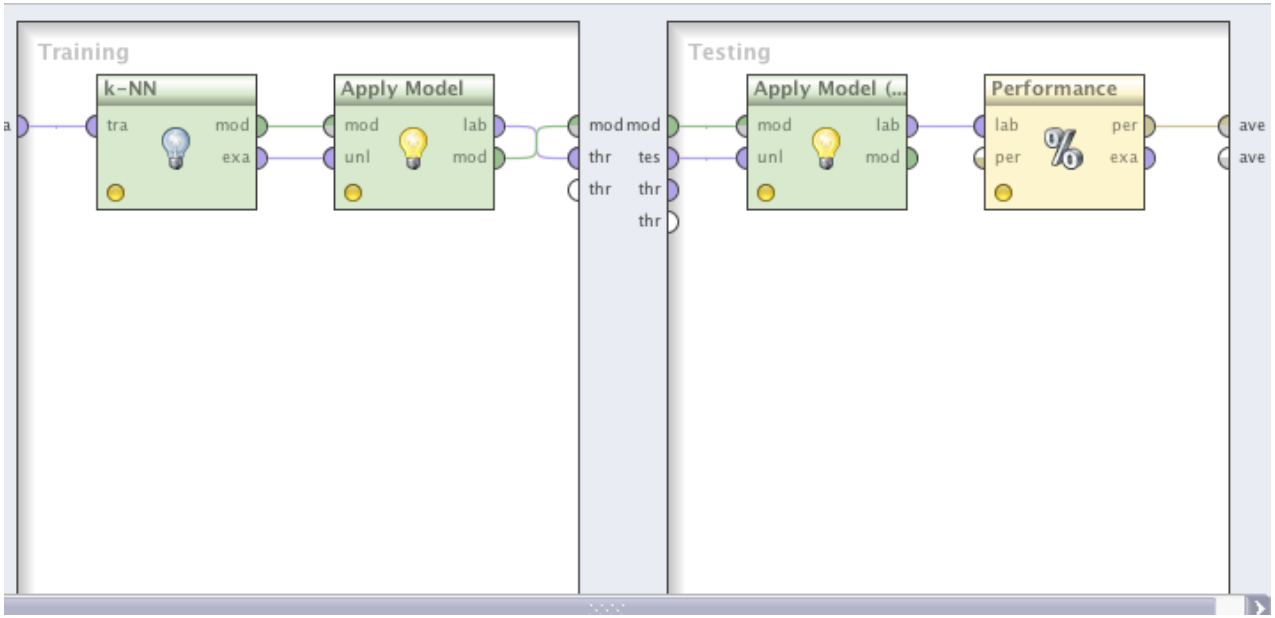
Filtern, so dass nur Beispiele mit *Outlier = false* übrig bleiben.

Kreuzvalidierung



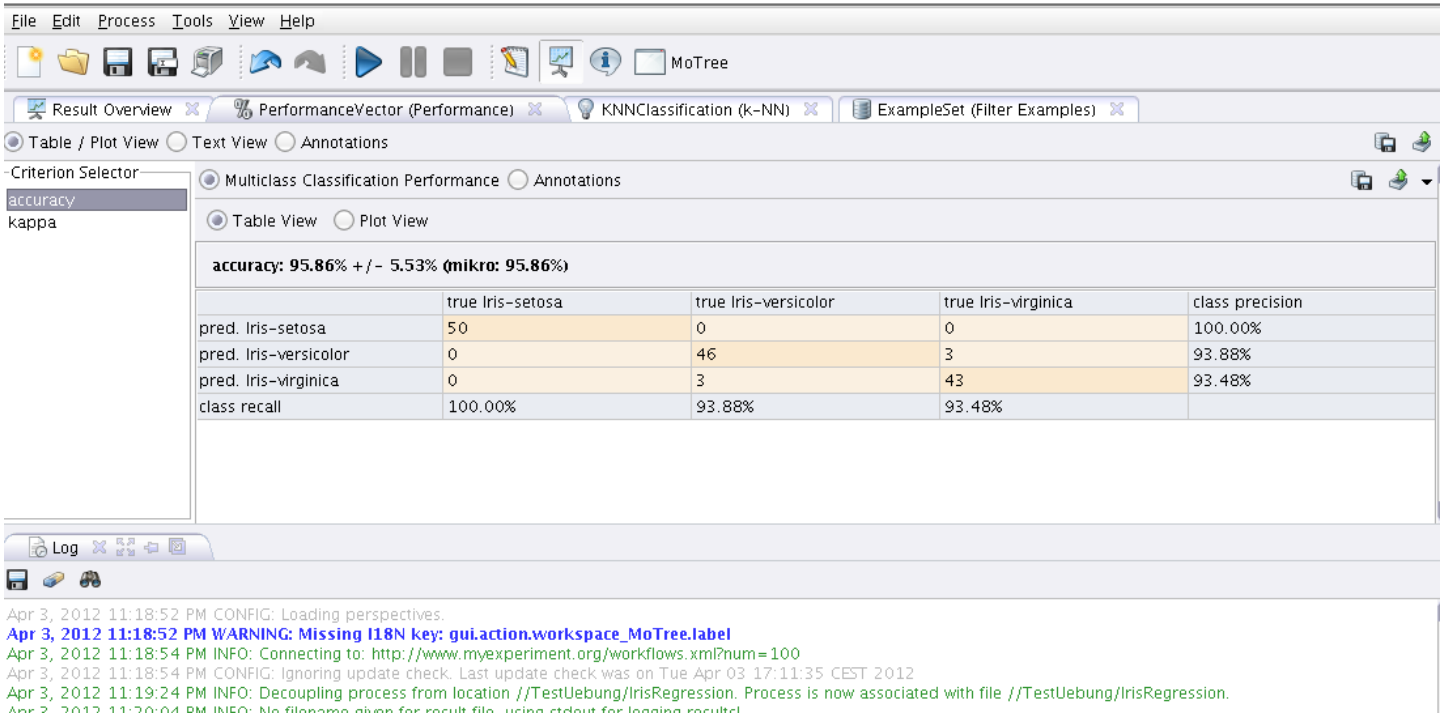
Modellierung

Innerhalb der Kreuzvalidierung wird das Modell gelernt, hier durch k-NN. Dann wird das gelernte Modell getestet und die Performanz als Durchschnitt ausgegeben.



Ergebnis

Resultatsperspektive



Was wissen Sie jetzt?

- Sie haben das CRISP kennengelernt, das den gesamten Ablauf der Wissensentdeckung beschreibt.
- Als Aufgaben der Modellbildung haben Sie Clustering, Klassifikation, Regression gesehen.
- Sie wissen, was die Kreuzvalidierung ist.
- Sie haben RapidMiner kennen gelernt:
  - Repository anlegen
  - Daten einlesen
  - Daten ansehen
  -