Data Mining Cup 2009

Team FH Brandenburg I + II

Gliederung

Vorbereitung

Aufgabe

Vorgehensweise

Fazit

"A CURIOUS COLLECTION. " - Sidney Paget

Team FH Brandenburg I + II

Carsten Schwenke, 4. Semester Bachelor

Hannes Uhlmann, 4. Semester Bachelor

Christian Heinrich, 3. Semester Master

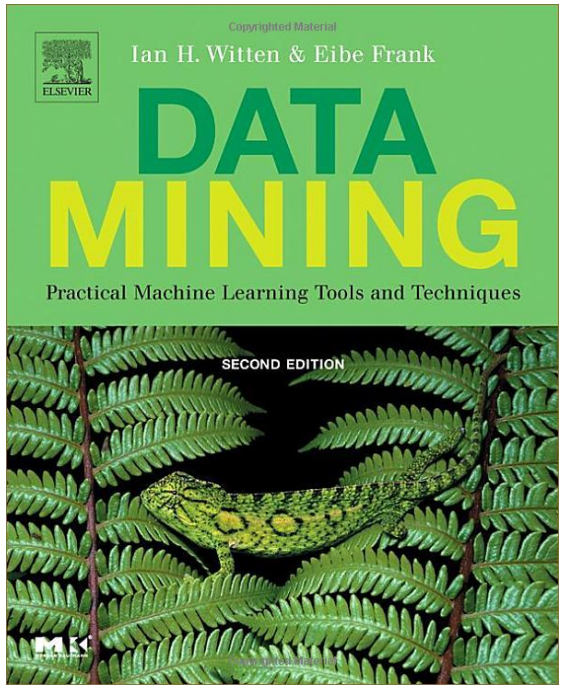
Christian Freye, 2. Semester Master

Andy Bertz, 2. Semester Master

Vorbereitung

Einarbeitung in das Thema Data-Mining, DMC und Rapidminer

- Seminare mit Ingo Boersch
- Selbststudium
- Rapidminer Tutorials
- Lernen von Siegern



Aufgabe

Vorhersage: Verkaufszahlen von 8 Buchtiteln an 2418 verschiedenen Standorten

Trainingsdaten: Simulierte Einkaufsdaten von 2394 Standorten

Merkmale: Verkaufzahlen von Büchern in 1856 Warengruppen

Vorgehensweise

Phasen Knowledge Discovery in Databases

Datenselektion / -extraktion
- Welche Daten notwendig und verfügbar?

Datenreinigung und Vorverarbeitung
- Fehlende Werte, Ausreißer, Inkonsistenzen

Datentransformation
- Format für DM (einzelne Tabelle), Aggregation, Aufteilung in Trainings- und Testdaten

Data Mining (10 .. 20% Zeitaufwand)
- Finden von Mustern

Interpretation und Evaluation
- Präsentation
- Bewertung mit Testdaten

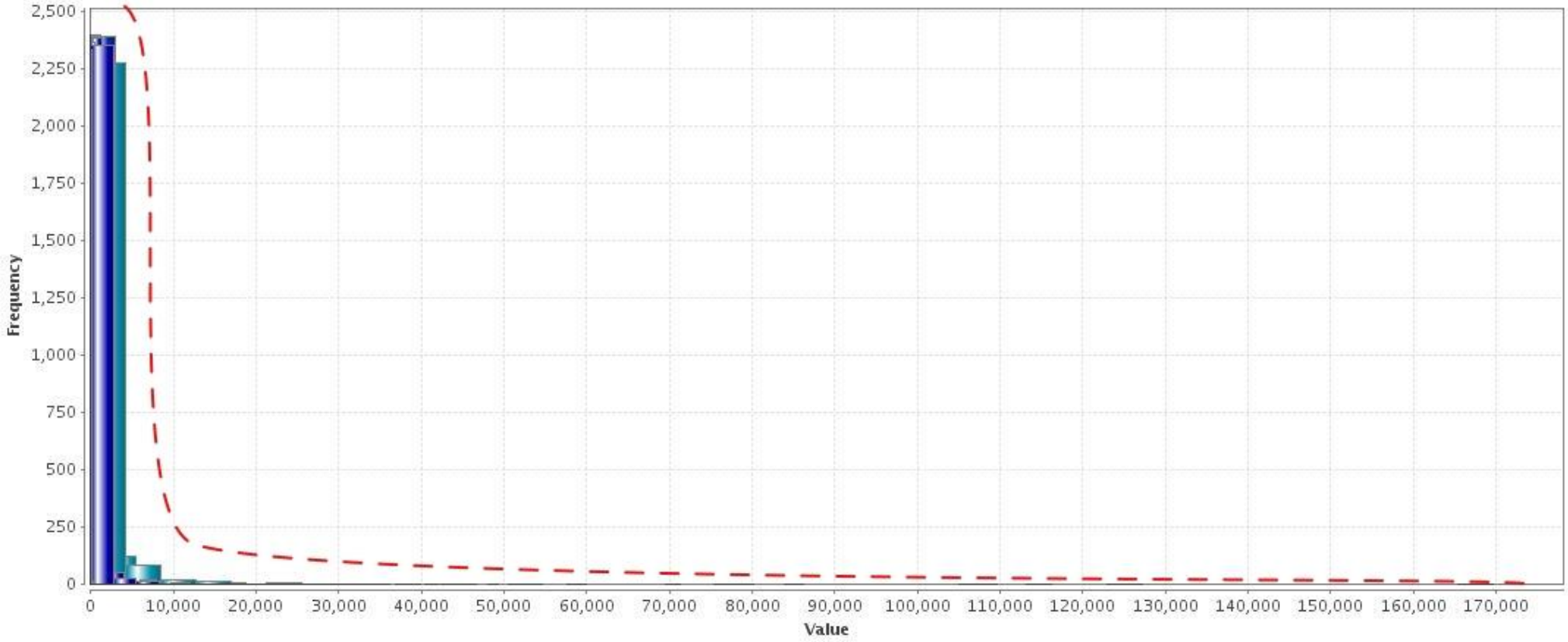
[Boersch09]

Vorgehensweise | Datenanalyse

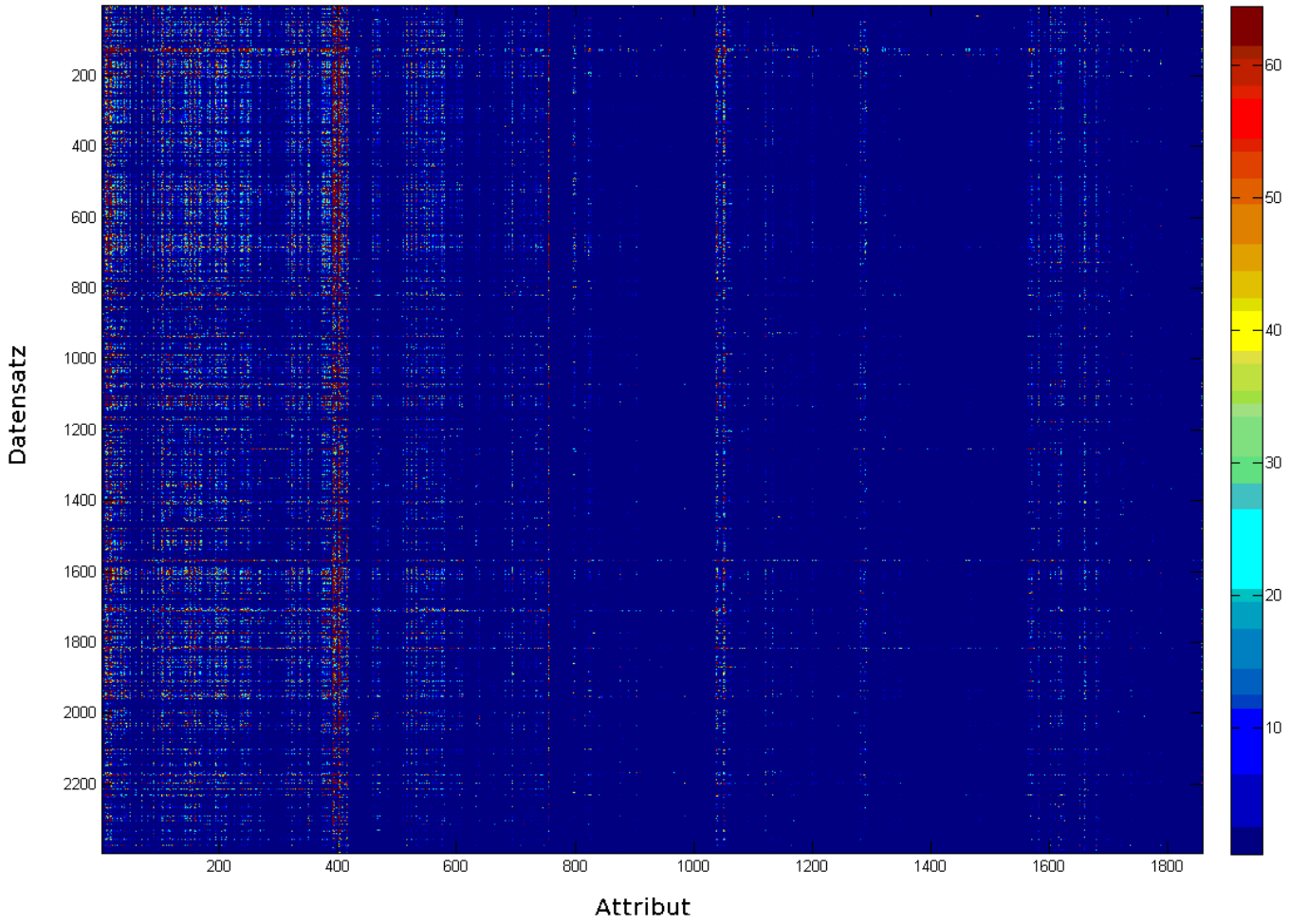
ID	WG 00000	WG 23000	WG 23110	WG 23120	...	WG 12600	WG 12700	WG 12900	T1	...	T8
3377	0	0	10	0	...	0	0	0	0	...	0
4355	0	0	0	0	...	0	0	0	0	...	0
1119	0	0	7	0	...	0	0	0	0	...	0
4429	0	0	0	0	...	0	0	0	0	...	0
2428	3	0	3233	0	...	0	0	0	0	...	0
1866	6	8	1451	4	...	0	0	0	1	...	7
2458	0	6	347	0	...	0	0	0	0	...	0
4531	7	12	1302	10	...	0	0	1	1	...	3
1884	0	0	339	0	...	0	0	0	0	...	0
5524	0	0	474	0	...	0	0	6	0	...	4
3647	0	3	372	0	...	0	0	19	1	...	4
...	...	...	...	...	...	...	...	...	...	...	...

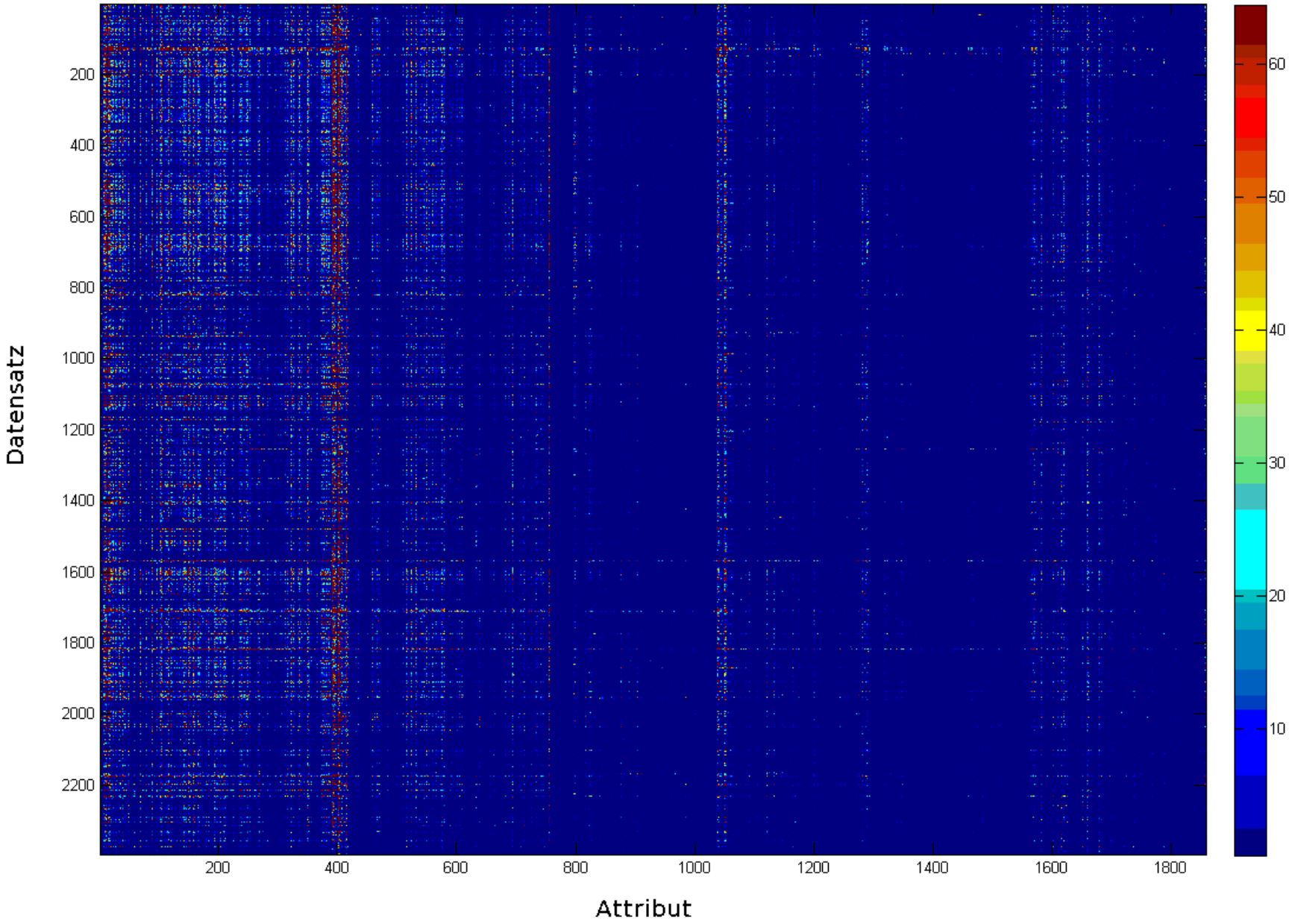
ID	WG 00000	WG 23000	WG 23110	WG 23120	...	WG 12600	WG 12700	WG 12900	T1	...	T8
3377	0	0	10	0	...	0	0	0	0	...	0
4355	0	0	0	0	...	0	0	0	0	...	0
1119	0	0	7	0	...	0	0	0	0	...	0
4429	0	0	0	0	...	0	0	0	0	...	0
2428	3	0	3233	0	...	0	0	0	0	...	0
1866	6	8	1451	4	...	0	0	0	1	...	7
2458	0	6	347	0	...	0	0	0	0	...	0
4531	7	12	1302	10	...	0	0	1	1	...	3
1884	0	0	339	0	...	0	0	0	0	...	0
5524	0	0	474	0	...	0	0	6	0	...	4
3647	0	3	372	0	...	0	0	19	1	...	4
...	...	...	...	...	...	...	...	...	...	...	...

Vorgehensweise | Datenanalyse



Vorgehensweise | Datenanalyse




Datensatz
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
Attribut
200
400
600
800
1000
1200
1400
1600
1800
60
50
40
30
20
10

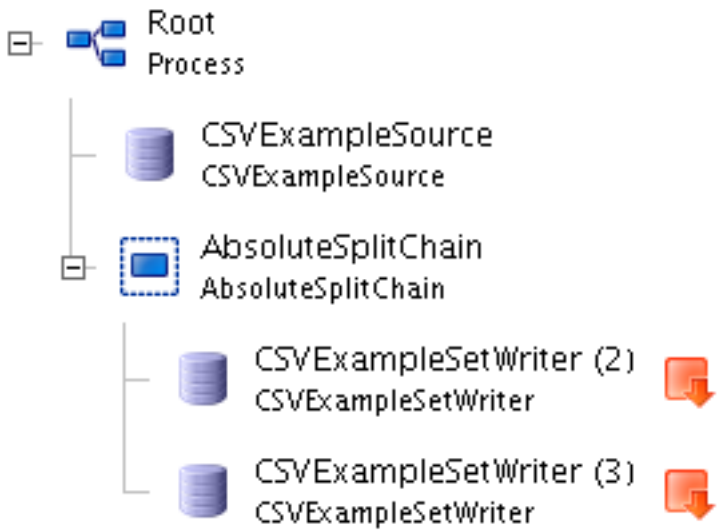
Vorgehensweise | Datenanalyse

Name	Value Type	Statistics	Range
T1	integer	avg = 0.255 +/- 1.676	[0.000 ; 51.000]
T2	integer	avg = 0.534 +/- 2.970	[0.000 ; 96.000]
T3	integer	avg = 0.195 +/- 1.449	[0.000 ; 34.000]
T4	integer	avg = 0.234 +/- 1.864	[0.000 ; 63.000]
T5	integer	avg = 1.950 +/- 19.377	[0.000 ; 853.000]
T6	integer	avg = 5.617 +/- 49.094	[0.000 ; 2,275.000]
T7	integer	avg = 0.631 +/- 10.139	[0.000 ; 438.000]
T8	integer	avg = 0.801 +/- 22.372	[0.000 ; 1,090.000]

Vorgehensweise | Vorverarbeitung

Aufteilen der Trainingsdaten für jede Klasse

Unterteilung in Trainingsmenge und Validierungsmenge (stratifiziert)

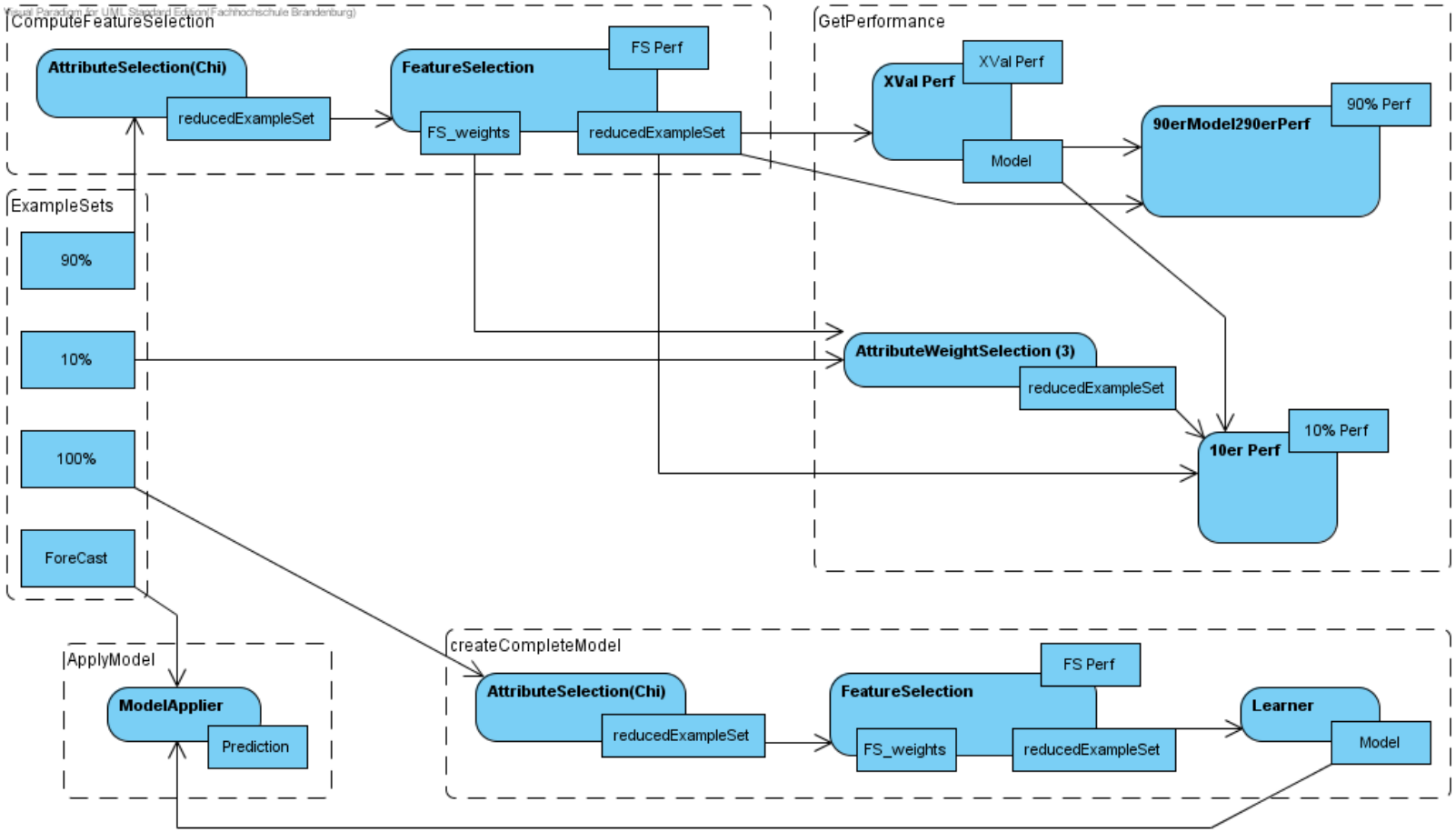


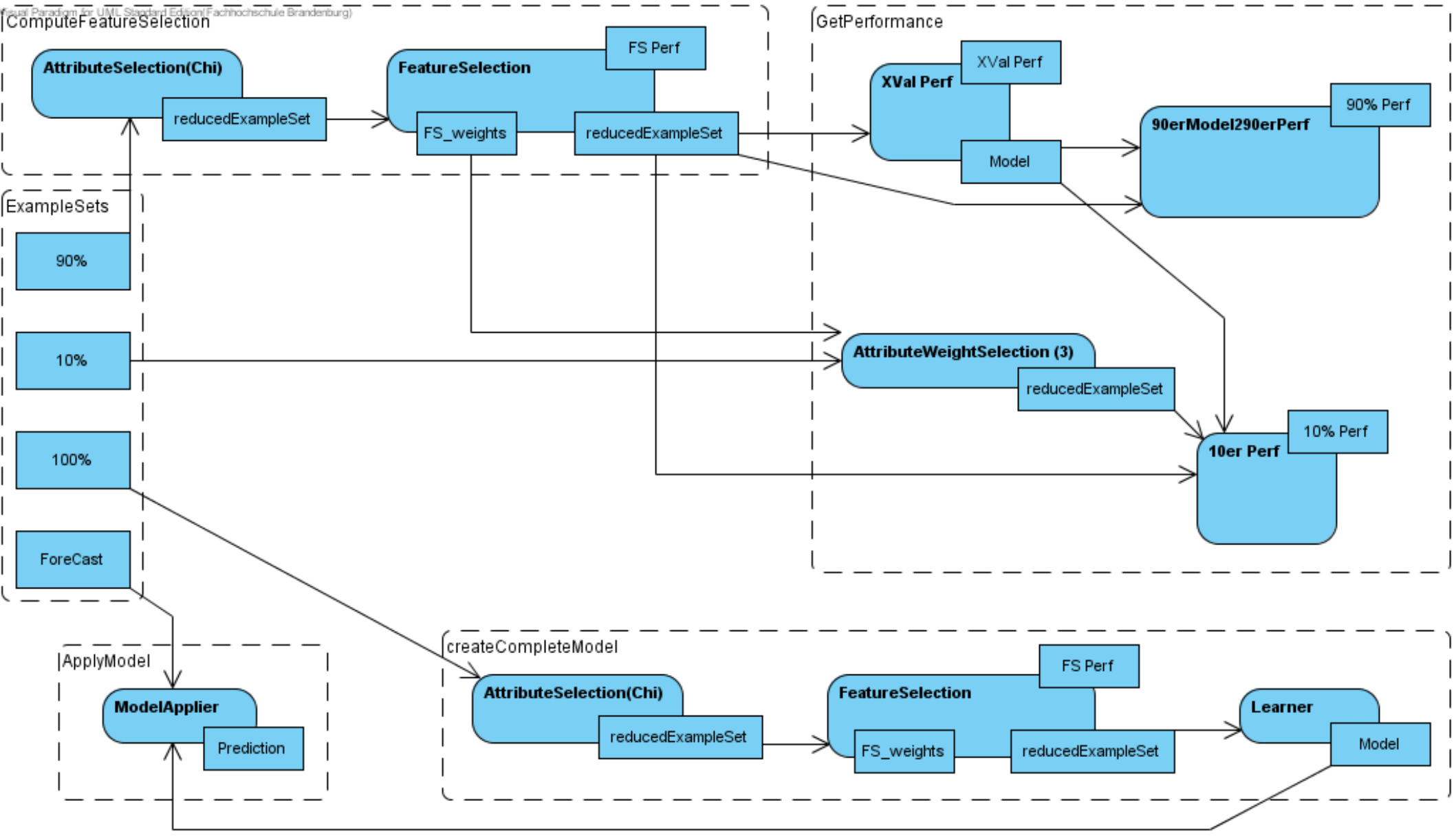
Vorgehensweise | Vorverarbeitung

Score	Zeitstempel	Eigenschaften / Änderungen
5.671	23.04.2009	• 5-fache Kreuzvalidierung mit DefaultLearner, der „0“ vorhersagt
5.158	24.04.2009	• 12-fache Kreuzvalidierung • DefaultLearner durch JMySVMLearner mit Standardeinstellungen ersetzt • Entfernen überflüssiger Attribute durch Operator RemoveUselessAttributes
5.064	25.04.2009	• 12-fache Kreuzvalidierung durch 5-fache Kreuzvalidierung ersetzt
4.182	30.04.2009	• Entfernen weiterer Attribute durch Feature Selection (Operator FeatureSelection) mit Vorwärtsselektion • 5-fache Kreuzvalidierung eingebettet in Feature Selection (wird deshalb in jedem Iterationsschritt der Feature Selection durchgeführt)

Score	Zeitstempel	Eigenschaften / Änderungen
3.407	30.04.2009	• JMySVMLearner durch NearestNeighbors-Operator ersetzt
3.277	30.04.2009	• NearestNeighbors-Operator durch AttributeBasedVote-Operator ersetzt
3.047	30.04.2009	• AttributeBasedVote-Operator durch JMySVMLearner-Operator mit veränderten Parametern „convergence_epsilon“ und „L_pos“ ersetzt
2.814	01.05.2009	• RemoveUselessAttributes ersetzt durch AttributeWeightSelection-Operator, der überflüssiger Attribute mit Hilfe von Gewichten entfernt • Gewichte werden mit einem Chi-Quadrat-Test (Operator ChiSquaredWeighting) auf den Trainingsdaten erstellt und danach gespeichert • Diskretisieren der Daten für den ChiSquaredWeighting-Operator, weil dieser nur mit nominalen Attributen funktioniert (Operator AbsoluteDiscretization) • Entfernen weiterer Attribute durch Feature Selection (Operator FeatureSelection) mit Vorwärtsselektion • JMySVMLearner durch LinearRegression-Operator ersetzt

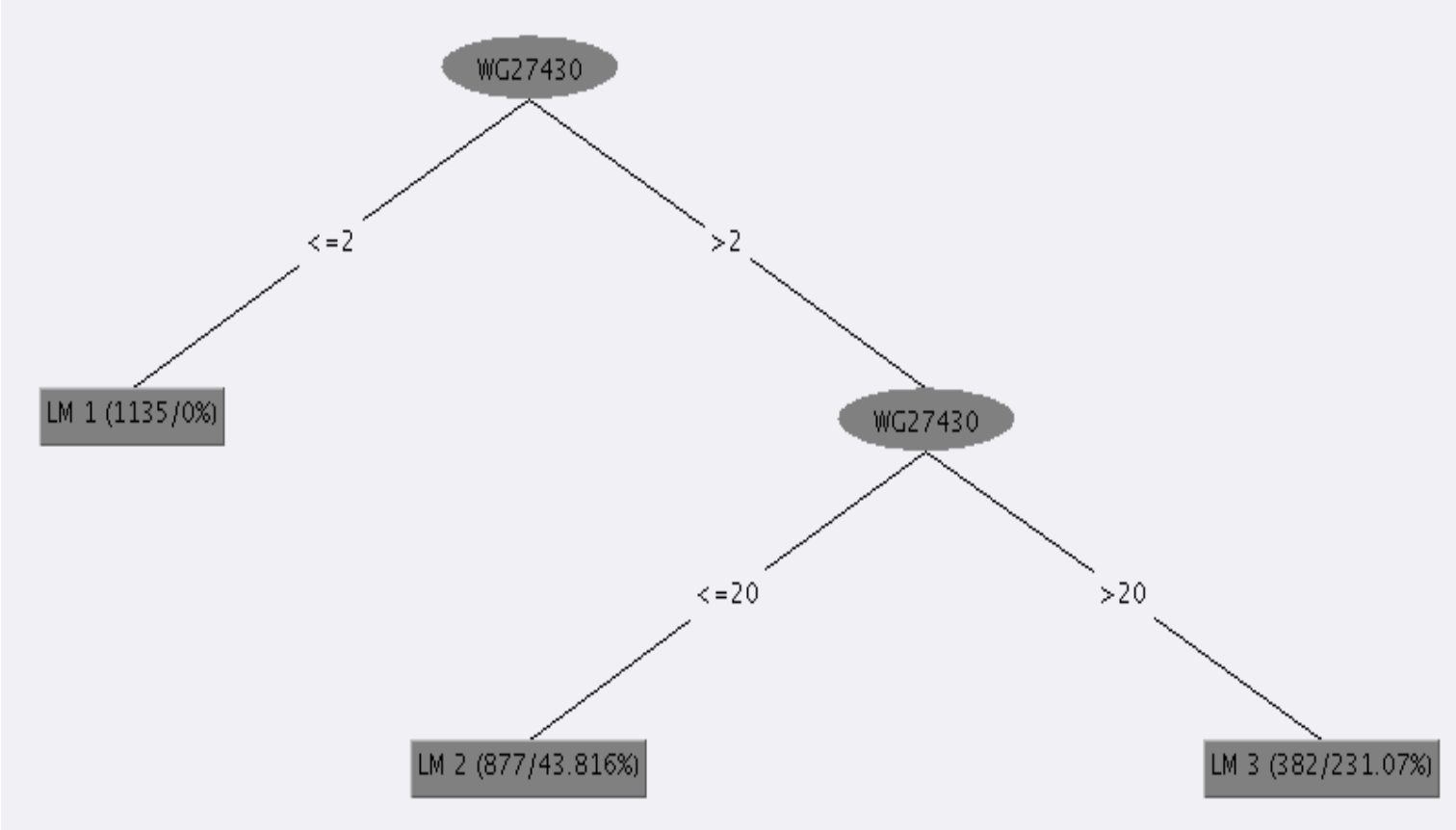
Vorgehensweise | Vorhersage




ComputeFeatureSelection
AttributeSelection(Chi)
reducedExampleSet
FeatureSelection
FS Perf
FS_weights
reducedExampleSet
GetPerformance
XVal Perf
XVal Perf
Model
90erModel290erPerf
90% Perf
ExampleSets
90%
10%
100%
ForeCast
AttributeWeightSelection (3)
reducedExampleSet
10er Perf
10% Perf
ApplyModel
ModelApplier
Prediction
createCompleteModel
AttributeSelection(Chi)
reducedExampleSet
FeatureSelection
FS Perf
FS_weights
reducedExampleSet
Learner
Model

Vorgehensweise | Vorhersage

Was ist ein Modellbaum?



Fazit

+ Lösung eingereicht

+ Erfahrungen von RWTH Aachen

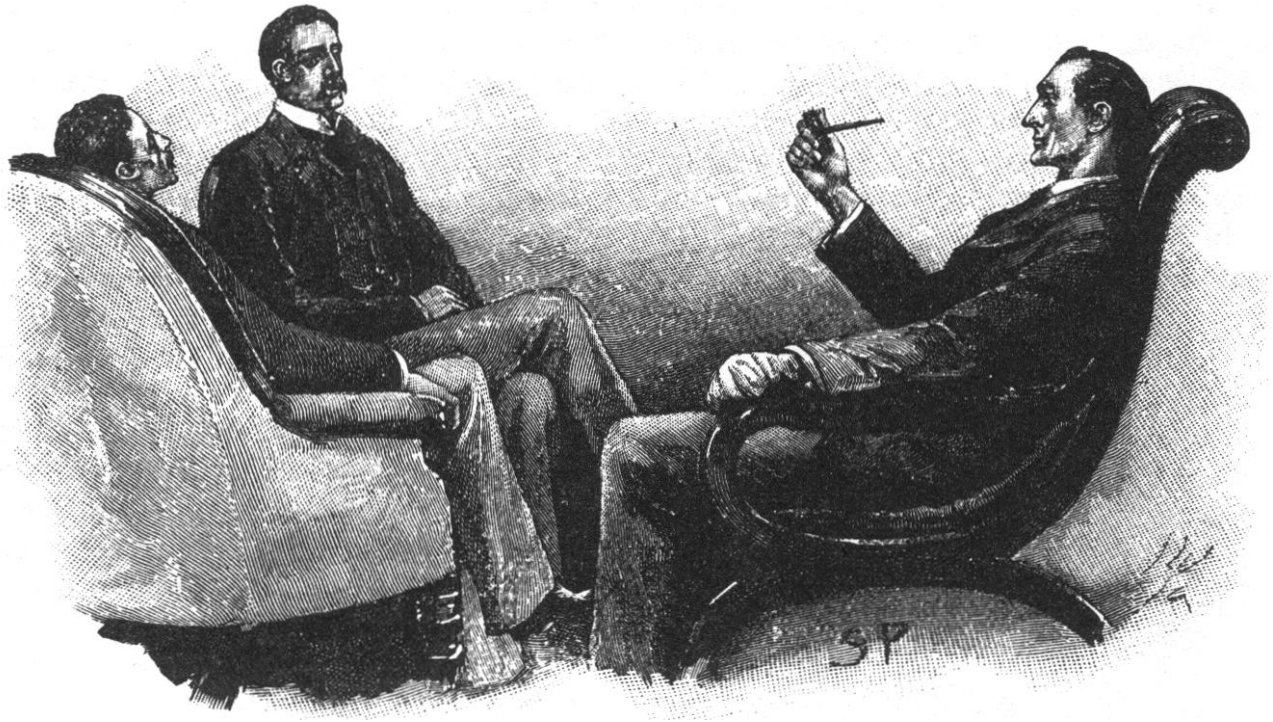
+ jetzt auch eigene DMC-Erfahrung für FHB

+/- Ansatz Rapidminer

- keine DMC Erfahrung

- Zeitaufteilung bei 8 Klassen

Vielen Dank!



"IS THERE ANOTHER POINT WHICH I CAN MAKE CLEAR? " - Sidney Paget