MythMiner

Ein Empfehlungssystem für Fernsehprogramme auf Basis von RapidMiner

Balázs Bárány

Linuxwochen Wien, 7. 5. 2011

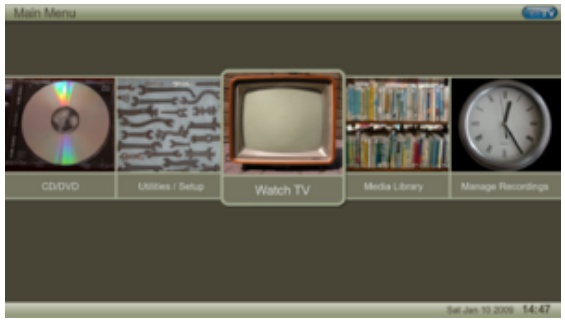
Inhalt

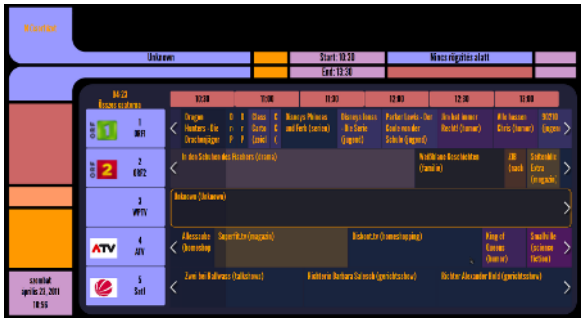
1. Voraussetzungen für MythMiner
   - MythTV
   - RapidMiner
2. Data Mining und Text Mining
   - Data Mining
   - Explorative Datenanalyse
   - Lernalgorithmen
   - Text Mining
3. MythMiner
   - Nutzung von MythMiner
   - Implementation von MythMiner

MythTV

Open Source Digital Video Recorder

- Komplettlösung für „Unterhaltungscomputer”
  - Fernsehen, Video, Audio, Bilder, Wetter, Nachrichten, ...
- Verteilte Architektur: Steuerung, Backend(s), Frontend(s)
- Web-Frontend für Programmierung, Verwaltung, Streaming
- Theme-Support





Installation und Nutzung von MythTV

Voraussetzungen

- Angepaßte Hardware: Wohnzimmerkompatibler PC, Fernsehkarte, Fernbedienung, HDMI
- Fertige Pakete in vielen Linux-Distributionen
- Eigene Distributionen: Mythbuntu, MythDora, KnoppMyth
- Quelle für Programminhalte (EPG) notwendig

RapidMiner

Open-Source-System für Data Mining

- Komplette Umgebung für Data Mining:
  - Grafische Oberfläche, visuelle Modellierung
  - „Entdeckung" der Daten = explorative Datenanalyse, Visualisierung
  - Datenintegration: Zusammenfassung und Umwandlung von Daten aus unterschiedlichen Quellen
  - Datenmodellierung für Vorhersagen, Maschinelles Lernen
  - Berichtsfunktionen
  - Plugins für diverse Aufgaben

RapidAnalytics

- Serverversion von RapidMiner
  - Community-Version eignet sich für periodische Job-Ausführung, Datenintegration, nicht interaktives Reporting usw.
  - Repository: Speicherbereich für gemeinsame Arbeit an RapidMiner-Projekten
  - Erweiterte Enterprise-Version mit interaktiven Berichts- und Analysefunktionen

Data Mining

Definition

„Unter Data Mining (der englischen Begriff bedeutet in etwa „Datenschätze heben") versteht man die systematische Anwendung von Methoden, die meist statistisch-mathematisch begründet sind, auf einen Datenbestand mit dem Ziel, neue Muster zu erkennen." (Wikipedia)

- Andere Bezeichnungen
  - Machine learning = Maschinelles Lernen
  - Knowledge Discovery in Databases = Entdeckung von Wissen in Datenbanken
  - generell: Predictive Analytics = Vorhersage-Analytik

Data Mining

Definition

„Unter Data Mining (der englischen Begriff bedeutet in etwa „Datenschätze heben") versteht man die systematische Anwendung von Methoden, die meist statistisch-mathematisch begründet sind, auf einen Datenbestand mit dem Ziel, neue Muster zu erkennen." (Wikipedia)

- Andere Bezeichnungen
  - Machine learning = Maschinelles Lernen
  - Knowledge Discovery in Databases = Entdeckung von Wissen in Datenbanken
  - generell: Predictive Analytics = Vorhersage-Analytik

Data Mining

Definition

„Unter Data Mining (der englischen Begriff bedeutet in etwa „Datenschätze heben") versteht man die systematische Anwendung von Methoden, die meist statistisch-mathematisch begründet sind, auf einen Datenbestand mit dem Ziel, neue Muster zu erkennen." (Wikipedia)

- Andere Bezeichnungen
  - Machine learning = Maschinelles Lernen
  - Knowledge Discovery in Databases = Entdeckung von Wissen in Datenbanken
  - generell: Predictive Analytics = Vorhersage-Analytik

Anwendungen für Data Mining

- Technik: Spamfilter, Vorhersage von Maschinenschäden, ...
- Wissenschaft: Textkategorisierung, Gensequenz-Analysen, Vulkanausbrüche und Erdbeben, ...
- Wirtschaft: Vorhersage von Kundenverhalten, Kündigerprävention, Marketingoptimierung, ...
- Privat

Anwendungen für Data Mining

- Technik: Spamfilter, Vorhersage von Maschinenschäden, ...
- Wissenschaft: Textkategorisierung, Gensequenz-Analysen, Vulkanausbrüche und Erdbeben, ...
- Wirtschaft: Vorhersage von Kundenverhalten, Kündigerprävention, Marketingoptimierung, ...
- Privat

Anwendungen für Data Mining

- Technik: Spamfilter, Vorhersage von Maschinenschäden, ...
- Wissenschaft: Textkategorisierung, Gensequenz-Analysen, Vulkanausbrüche und Erdbeben, ...
- Wirtschaft: Vorhersage von Kundenverhalten, Kündigerprävention, Marketingoptimierung, ...
- Privat

Anwendungen für Data Mining

- Technik: Spamfilter, Vorhersage von Maschinenschäden, ...
- Wissenschaft: Textkategorisierung, Gensequenz-Analysen, Vulkanausbrüche und Erdbeben, ...
- Wirtschaft: Vorhersage von Kundenverhalten, Kündigerprävention, Marketingoptimierung, ...
- Privat

Ein einfaches Beispiel

- Problem: Entscheidung über den Sommerurlaub
  - Variablen, Attribute: Eigenschaften, Merkmale
    - Hotel, Ort, Land, Meerblick ja/nein: Texte = *kategoriale Variablen* (Unterscheidung möglich)
    - Entfernung, Preis: Zahlen = *numerische Variablen* (Rechnen möglich)
    - Mischformen je nach Art der Modellierung: Anreise: Zeit, Kosten, Bequemlichkeit; „Sterne" des Hotels: anscheinend numerisch, aber auch kategorial interpretierbar
  - Quelle der Daten: Prospekte, Internet, Werbung in Medien, ...
  - Datenhaltung: Bei dieser Datenmenge eignet sich eine Tabellenkalkulation perfekt
    - Struktur ist wichtig: Urlaubsangebote in Zeilen, Attribute in Spalten

Ein einfaches Beispiel

- Problem: Entscheidung über den Sommerurlaub
  - Variablen, Attribute: Eigenschaften, Merkmale
    - Hotel, Ort, Land, Meerblick ja/nein: Texte = *kategoriale Variablen* (Unterscheidung möglich)
    - Entfernung, Preis: Zahlen = *numerische Variablen* (Rechnen möglich)
    - Mischformen je nach Art der Modellierung: Anreise: Zeit, Kosten, Bequemlichkeit; „Sterne" des Hotels: anscheinend numerisch, aber auch kategorial interpretierbar
  - Quelle der Daten: Prospekte, Internet, Werbung in Medien, ...
  - Datenhaltung: Bei dieser Datenmenge eignet sich eine Tabellenkalkulation perfekt
    - Struktur ist wichtig: Urlaubsangebote in Zeilen, Attribute in Spalten

Ein einfaches Beispiel

- Problem: Entscheidung über den Sommerurlaub
  - Variablen, Attribute: Eigenschaften, Merkmale
    - Hotel, Ort, Land, Meerblick ja/nein: Texte = *kategoriale Variablen* (Unterscheidung möglich)
    - Entfernung, Preis: Zahlen = *numerische Variablen* (Rechnen möglich)
    - Mischformen je nach Art der Modellierung: Anreise: Zeit, Kosten, Bequemlichkeit; „Sterne" des Hotels: anscheinend numerisch, aber auch kategorial interpretierbar
  - Quelle der Daten: Prospekte, Internet, Werbung in Medien, ...
  - Datenhaltung: Bei dieser Datenmenge eignet sich eine Tabellenkalkulation perfekt
    - Struktur ist wichtig: Urlaubsangebote in Zeilen, Attribute in Spalten

Ein einfaches Beispiel

- Problem: Entscheidung über den Sommerurlaub
  - Variablen, Attribute: Eigenschaften, Merkmale
    - Hotel, Ort, Land, Meerblick ja/nein: Texte = *kategoriale Variablen* (Unterscheidung möglich)
    - Entfernung, Preis: Zahlen = *numerische Variablen* (Rechnen möglich)
    - Mischformen je nach Art der Modellierung: Anreise: Zeit, Kosten, Bequemlichkeit; „Sterne" des Hotels: anscheinend numerisch, aber auch kategorial interpretierbar
  - Quelle der Daten: Prospekte, Internet, Werbung in Medien, ...
  - Datenhaltung: Bei dieser Datenmenge eignet sich eine Tabellenkalkulation perfekt
    - Struktur ist wichtig: Urlaubsangebote in Zeilen, Attribute in Spalten

Ein einfaches Beispiel

- Problem: Entscheidung über den Sommerurlaub
  - Variablen, Attribute: Eigenschaften, Merkmale
    - Hotel, Ort, Land, Meerblick ja/nein: Texte = *kategoriale Variablen* (Unterscheidung möglich)
    - Entfernung, Preis: Zahlen = *numerische Variablen* (Rechnen möglich)
    - Mischformen je nach Art der Modellierung: Anreise: Zeit, Kosten, Bequemlichkeit; „Sterne" des Hotels: anscheinend numerisch, aber auch kategorial interpretierbar
  - Quelle der Daten: Prospekte, Internet, Werbung in Medien, ...
  - Datenhaltung: Bei dieser Datenmenge eignet sich eine Tabellenkalkulation perfekt
    - Struktur ist wichtig: Urlaubsangebote in Zeilen, Attribute in Spalten

Beispiel-Tabelle

Strukturierte Erfassung der Angebote:

	A	B	C	D	E	F	G	H
1	**Bezeichnung**	**Ort**	**Land**	**Meerblick**	**Hotel**	**Tage**	**Preis**	**Entscheidung**
2	Niriides Beach	Kolymbia	Griechenland	ja	****	7	€ 371,00	interessant
3	Viva Alcudia Sun Village	Playa de Muro	Spanien	nein	***	7	€ 340,00	nicht interessant
4	La Moraleja	Cala San Vincente	Spanien	ja	*****	9	€ 895,00	zu teuer
5	Pulicinu	Baia Sardinia	Italien	nein	****	9	€ 685,00	nicht interessant
6	Colonna Beach	Porto Rotondo	Italien	ja	****	9	€ 720,00	interessant
7	Arbatasar	Arbatax	Italien	ja	****	9	€ 774,00	interessant
8	La Torre	Bari Sardo	Italien	ja	****	9	€ 978,00	zu teuer
9	L'Hostellerie du golf	Mandelieu	Frankreich	nein	***	9	€ 639,00	nicht interessant
10	Mediterranee Menton	Menton	Frankreich	nein	***	7	€ 653,00	nicht interessant

Welche Attribute könnten noch wichtig sein? Sind die Angebote vollständig vergleichbar?

Explorative Datenanalyse

„Entdecken" der Daten

- Kennzahlen der Verteilung der einzelnen Variablen
  - Kategorial: Liste der Werte, häufigster Wert, seltenster Wert
  - Numerisch: Minimum, Maximum, Mittelwert, Varianz, ...
  - Anzahl der fehlenden Daten
- Grafische Darstellung
  - Histogramme, Dichtekurven: Verteilung einer Variablen
  - Streudiagramme (Scatterplots): Zwei Variablen
  - Diagrammtypen für mehr Variablen
- Statistische Tests: Korrelation, Verteilungen, ...

Lernalgorithmen

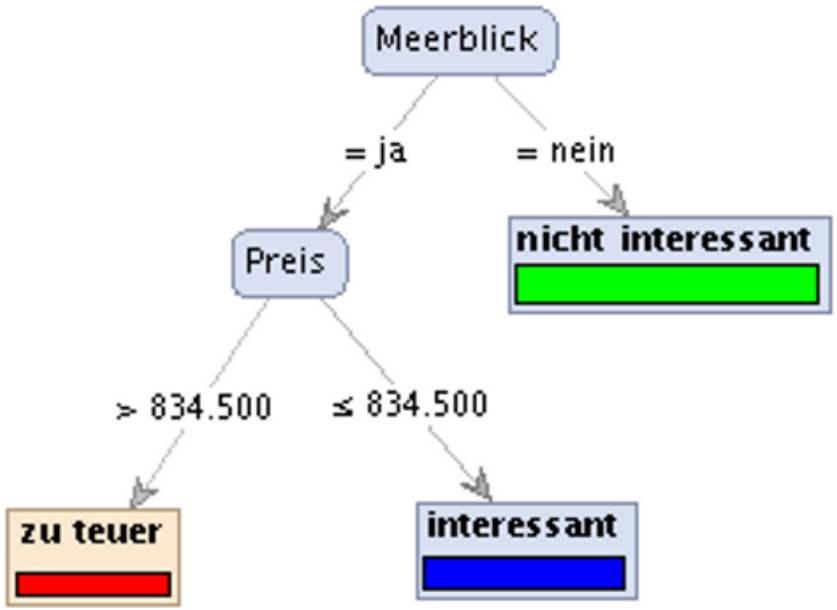
- Lernalgorithmen „bauen" Modelle aus den Daten
  - „informierte Verfahren" (supervised learning): Zielvariable bekannt
  - „uninformierte Verfahren" (unsupervised learning): keine bekannte Zielvariable
- Verschiedene Methoden, unterschiedlich anwendbar
  - Mathematisch: Regression, Support Vector Machines, ...
  - Entscheidungsbäume, Regelerkennung, ...
  - Neuronale Algorithmen, ...
  - Manche sowohl für kategoriale als auch für numerische Daten verwendbar, andere nur für die eine oder andere Klasse

Lernalgorithmen

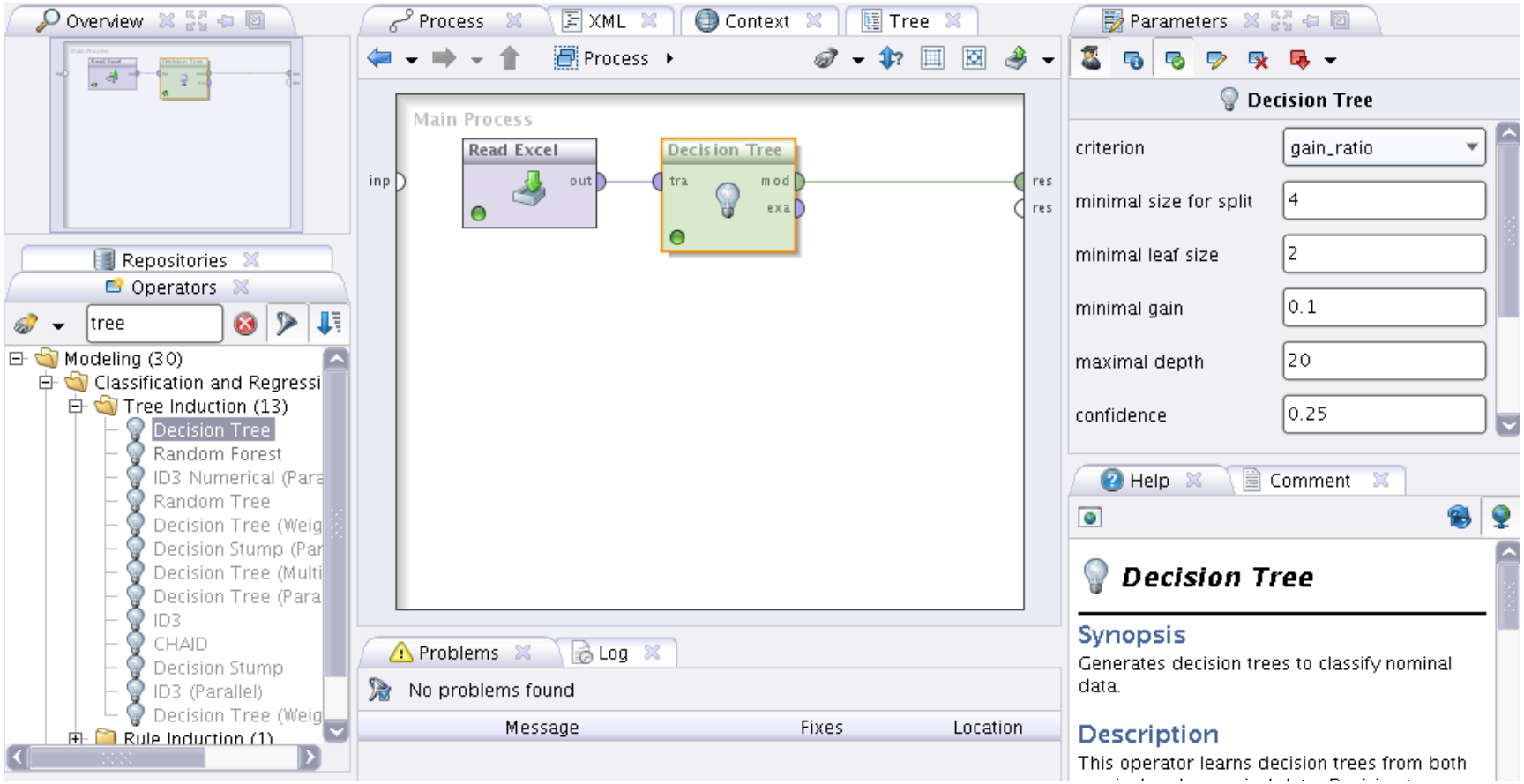
- Lernalgorithmen „bauen" Modelle aus den Daten
  - „informierte Verfahren" (supervised learning): Zielvariable bekannt
  - „uninformierte Verfahren" (unsupervised learning): keine bekannte Zielvariable
- Verschiedene Methoden, unterschiedlich anwendbar
  - Mathematisch: Regression, Support Vector Machines, ...
  - Entscheidungsbäume, Regelerkennung, ...
  - Neuronale Algorithmen, ...
  - Manche sowohl für kategoriale als auch für numerische Daten verwendbar, andere nur für die eine oder andere Klasse

Beispiel-Entscheidungsbaum

Entscheidungsbaum aus den Urlaubsplanungs-Daten:



Erstellung des Entscheidungsbaumes in RapidMiner



Vorverarbeitung für Text Mining

- Die Daten stehen als Texte (Dokumente) zur Verfügung
- Diese müssen für die Lernverfahren in Attribute umgewandelt werden
  - Tokenisierung: Umwandlung in Wörter (seltener: Sätze oder andere Einheiten)
  - Säuberung: Stopwörter, Groß-/Kleinschreibung, Stammbildung (stemming)
  - N-Gramme bilden: aus Wörtern oder aus Buchstaben
  - Wert des Attributs: „Termfrequenz", „Termfrequenz und inverse Dokument-Frequenz" (TF/IDF)
- Das Ergebnis ist eine sehr „breite" Datenstruktur mit hunderten oder tausenden Attributen

Vorverarbeitung für Text Mining

- Die Daten stehen als Texte (Dokumente) zur Verfügung
- Diese müssen für die Lernverfahren in Attribute umgewandelt werden
  - Tokenisierung: Umwandlung in Wörter (seltener: Sätze oder andere Einheiten)
  - Säuberung: Stopwörter, Groß-/Kleinschreibung, Stammbildung (stemming)
  - N-Gramme bilden: aus Wörtern oder aus Buchstaben
  - Wert des Attributs: „Termfrequenz", „Termfrequenz und inverse Dokument-Frequenz" (TF/IDF)
- Das Ergebnis ist eine sehr „breite" Datenstruktur mit hunderten oder tausenden Attributen

Vorverarbeitung für Text Mining

- Die Daten stehen als Texte (Dokumente) zur Verfügung
- Diese müssen für die Lernverfahren in Attribute umgewandelt werden
  - Tokenisierung: Umwandlung in Wörter (seltener: Sätze oder andere Einheiten)
  - Säuberung: Stopwörter, Groß-/Kleinschreibung, Stammbildung (stemming)
  - N-Gramme bilden: aus Wörtern oder aus Buchstaben
  - Wert des Attributs: „Termfrequenz", „Termfrequenz und inverse Dokument-Frequenz" (TF/IDF)
- Das Ergebnis ist eine sehr „breite" Datenstruktur mit hunderten oder tausenden Attributen

Lernverfahren im Text Mining

- Verfahren muß für Datensätze mit sehr vielen Attributen geeignet sein
- Naive Bayes: in Spamfiltern beliebt
  - Sehr schnell
  - Bestehende Modelle mit neuen Daten aktualisierbar
- Support Vector Machine, neuronale Netze, ...
  - Probleme mit der Laufzeit bei großen Datenmengen
  - Nicht aktualisierbar, Modell muß neu aufgebaut werden, um neue Daten zu integrieren
  - Aber häufig bessere Genauigkeit

Lernverfahren im Text Mining

- Verfahren muß für Datensätze mit sehr vielen Attributen geeignet sein
- Naive Bayes: in Spamfiltern beliebt
  - Sehr schnell
  - Bestehende Modelle mit neuen Daten aktualisierbar
- Support Vector Machine, neuronale Netze, ...
  - Probleme mit der Laufzeit bei großen Datenmengen
  - Nicht aktualisierbar, Modell muß neu aufgebaut werden, um neue Daten zu integrieren
  - Aber häufig bessere Genauigkeit

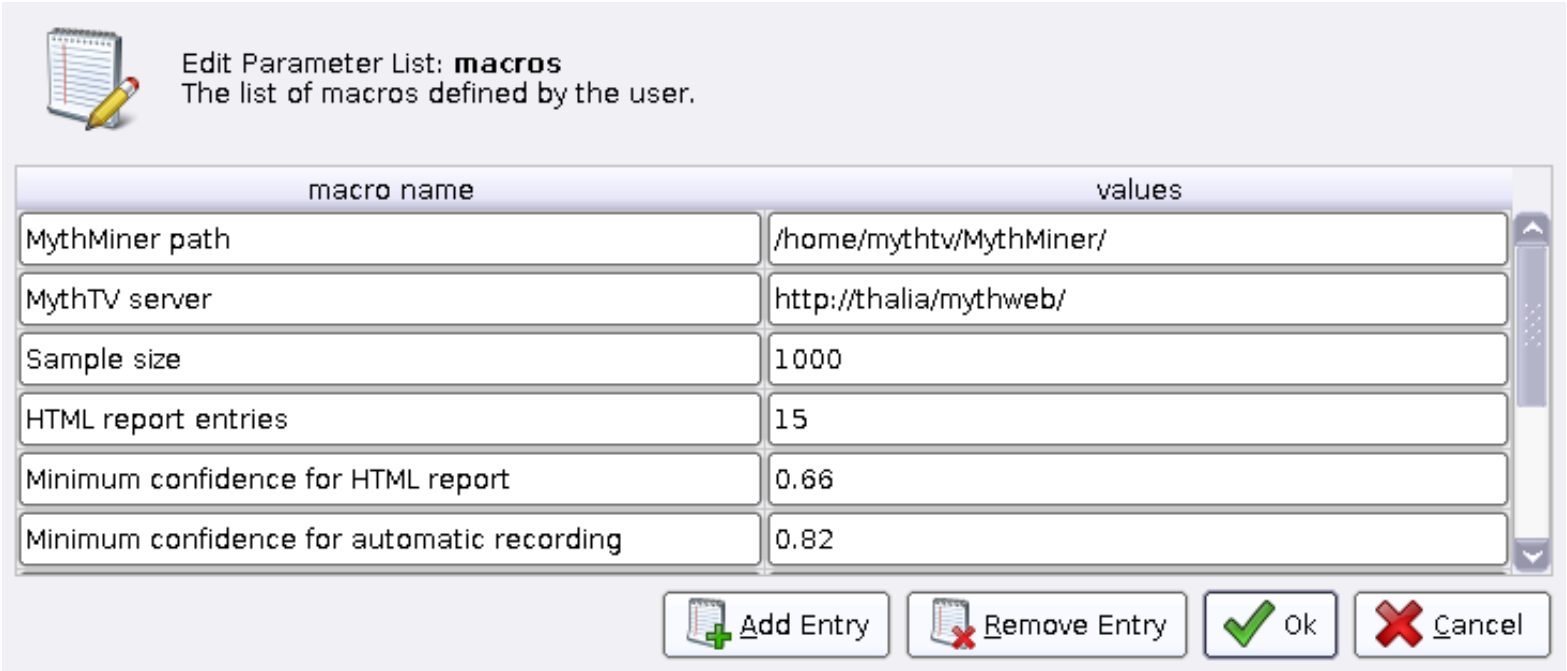
Lernverfahren im Text Mining

- Verfahren muß für Datensätze mit sehr vielen Attributen geeignet sein
- Naive Bayes: in Spamfiltern beliebt
  - Sehr schnell
  - Bestehende Modelle mit neuen Daten aktualisierbar
- Support Vector Machine, neuronale Netze, ...
  - Probleme mit der Laufzeit bei großen Datenmengen
  - Nicht aktualisierbar, Modell muß neu aufgebaut werden, um neue Daten zu integrieren
  - Aber häufig bessere Genauigkeit

MythMiner besorgen

- MythTV installieren und benutzen!
- RapidMiner installieren
- MythMiner herunterladen:
  `http://tud.at/programm/mythminer/`
  - entpacken
- In RapidMiner öffnen und konfigurieren
  - Datenbankverbindung zu MythTV anlegen
  - Schritt „Configure process" öffnen und anpassen

MythMiner-Konfiguration



Ergebnis der Ausführung

- HTML-Datei im ausgewählten Verzeichnis abgelegt
- Optional: Umwandlung mit dem mitgelieferten Shellskript: URL klickbar, schönere Tabelle
- Optional: tägliche E-Mail
- Nicht immer optimale Ergebnisse
  - Ein-zwei Arztserien ziehen die anderen hinein
  - Serien mit allgemeinem Inhalt verwässern das Ergebnis
- Aber an den meisten Tagen was Interessantes dabei

Ergebnis der Ausführung

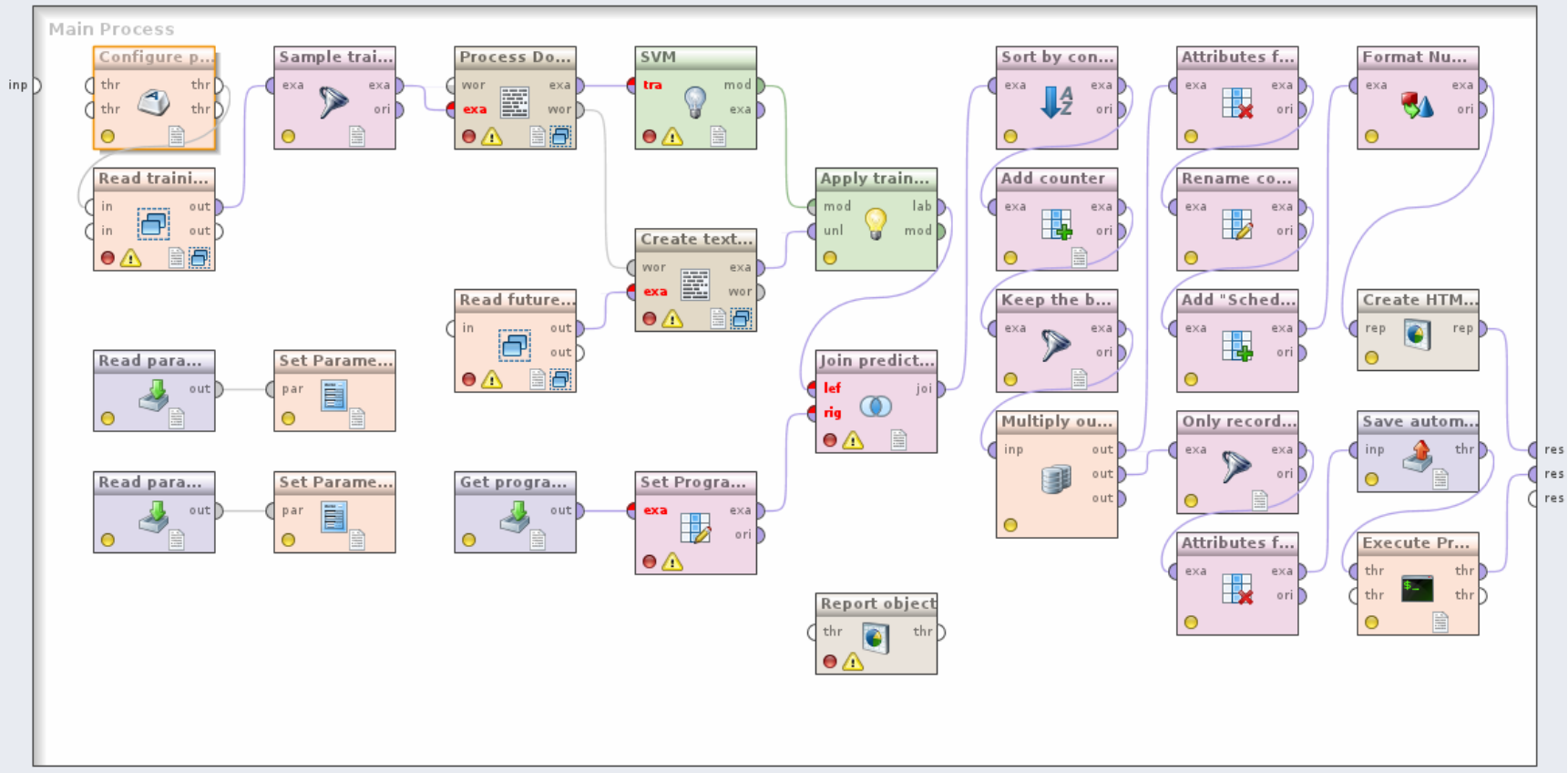
- HTML-Datei im ausgewählten Verzeichnis abgelegt
- Optional: Umwandlung mit dem mitgelieferten Shellskript: URL klickbar, schönere Tabelle
- Optional: tägliche E-Mail
- Nicht immer optimale Ergebnisse
  - Ein-zwei Arztserien ziehen die anderen hinein
  - Serien mit allgemeinem Inhalt verwässern das Ergebnis
- Aber an den meisten Tagen was Interessantes dabei

Beispiel für Ergebnis-Mail

Program list

Program info	Category	Title	Description	Scheduled for recording	Confidence
SW/BW, 08:30 - 09:15	geschichte	Rom - Marmor, Macht und Märtyrer: Vom Hüttendorf zur Metropole	Die erste Folge "Vom Hüttendorf zur Metropole" untersucht den Gründungsmythos der Stadt am Tiber, die Sage von Romulus, Remus und der Wölfin, und sie beleuchtet die Rolle der Nachbarvölker, den Einfluss der Etrusker und Griechen auf die frühe Siedlung. Sie fragt nach der politischen Organisation und den sozialen Problemen der römischen Republik. Wie funktionierte dieses antike Gemeinwesen? Was waren die entscheidenden Gründe, die Roms Größe und Macht bestimmten. Wie muss man sich die städtische Infrastruktur vorstellen?	yes	83.0 %
SW/BW, 09:15 - 10:00	geschichte	Rom - Marmor, Macht und Märtyrer: Der Aufstieg zur Kaiserstadt	Die zweite Folge spannt den Bogen vom Charisma Cäsars über die Expansion des römischen Weltreichs bis zur Friedensherrschaft des Kaiser Augustus. Er war es, der die Stadt aus Ziegeln zu einer Weltstadt aus Marmor machte. Von Rom aus erstreckte sich strahlenförmig ein gut ausgebautes Straßennetz bis in die letzten Winkel des Reiches. Dies erlaubte rasche Truppenverschiebungen, förderte die Handelsströme und den kulturellen Austausch zwischen den Provinzen und der		76.2 %

Überblick


Main Process
inp
Configure p...
Sample trai...
Process Do...
SVM
Read traini...
Create text...
Read future...
Read para...
Set Parame...
Read para...
Set Parame...
Get progra...
Set Progra...
Apply train...
Join predict...
Report object
Sort by con...
Add counter
Keep the b...
Multiply ou...
Attributes f...
Rename co...
Add "Sched...
Only record...
Attributes f...
Format Nu...
Create HTM...
Save autom...
Execute Pr...
res
res
res

Ablauf - Trainingsdaten und Modell

1. Konfiguration und Parameter fürs Lernverfahren einlesen (*Configure process; Read parameters / Set parameters*)
2. Trainingsdaten aus der Datenbank lesen: Unterscheidung zwischen aufgenommen und nicht aufgenommen (Unterprozess *Read training data*)
3. Stichprobe aus Trainingsdaten ziehen (*Sample training data*)
4. Datenstruktur aus Trainingsdaten bauen (Unterprozess *Process documents from data*)
   - Alles in Kleinbuchstaben, Tokenisierung, je nach Parametern Säuberungsverfahren anwenden
5. Aus den Trainingsdaten ein SVM-Modell bauen lassen (*SVM*)

Ablauf - Neue Daten und Modellanwendung

1. Morgige Daten aus der Datenbank lesen (Unterprozess *Read future data*)
2. Die morgigen Daten mit der Wortliste aus den Trainingsdaten verknüpfen (*Create text attributes from future data*). Das ist notwendig, weil das Modell auf dieser Wortliste basiert
3. Das SVM-Modell anwenden (*Apply trained model*)
4. Zusätzlich benötigte Attribute der morgigen Programme dazubinden (*Join predictions with data*)

Ablauf - Berichtserstellung

1. Morgige Programme nach dem Grad der Sicherheit (confidence) sortieren und nummerieren (*Sort by confidence, Add counter*)
2. Sendungen mit der minimalen konfigurierten Konfidenz behalten (*Keep the best results*)
3. Attribute formatieren: schönere Feldnamen für die Tabelle, Konfidenz als Prozent formatieren, ...
4. HTML-Seite mit der Tabelle speichern
5. Bei Sendungen mit genug Konfidenz die Daten in die Datenbank einfügen
6. MythTV-Reschedule auslösen (optional, aber empfehlenswert)

Schluß

- Fragen?