Praktikum Data Warehousing & Mining SS 2010

Zwischenergebnisse Data-Mining Cup

Gruppe 2: Patricia, Muhannad, Hong, Marusa, Jingyu, Dominik

Agenda

- Data Preprocessing
  - Irrelevante Attribute
  - Abgeleitete Attribute
  - Data Cleaning
- Gelernte Modelle & Evaluation
  - Entscheidungsbäume
  - Neuronale Netze

Dominik:

Data Preprocessing

Gruppe 2:
Patricia, Muhannad, Hong, Marusa, Jingyu, Dominik

Irrelevante Attribute

- = Attribute, die zum Lernen nicht verwendet werden sollen

- customernumber (trotzdem wichtig, da Primärschlüssel)
- Points (da fast immer 0)
- Delivtype (korreliert zu paymenttype)
- model (Korrelation mit entry von 0,98)
- Domain
- Title
- Alle Datumsfelder (s. Abgeleitete Attribute)

Abgeleitete Attribute

- = Attribute, die zusätzlich generiert wurden und aus anderen abgeleitet wurden.

- Delay of delivery (datedeliveryreal – deliverydatepromised)
- Deliverytime (datedeliveryreal – date)
- Differenz date – datecreated (fast immer ist 0)
- Unterschied invoicepostcode delivpostcode (selten befüllt)
- Digital, no digital oder vermischt(Abstrahieren von w1...w10)
- Monat eines Datums

c) Data Cleaning

- = Ersetzen von Ausprägungen der Attribute, um die Algorithmen besser verwenden zu können

- Jahr 4700 (deliverydatepromised): Im abgeleiteten Attribut delay of delivery berücksichtigt
- Datedeliveryreal bei ebooks ersetzen mit Bestelldatum (Auswirkungen auf delay of delivery)

Patricia:

Entscheidungsbäume

Was haben wir gemacht?

- Parametern ändern:
  - verschiedene Werte Kostenmatrix wählen ¿Welches Verhältnis?¿Wie gross mussen die Kosten sein?
  - Reduktionsgrad (verschiedene Werte: 25%,30%,50%,70%)
  - Minimum Anzahl Verzweigung (Reduktionsgrad höher ist es besser ein grosses Minimum)

  Man erhält verschiedene Ergebnisse und die bedeutsamkeit der Variablen sind verschieden.

- Unterschiedene Modelle zu kombinieren => Konfidenz
  - Dort wo h1 nicht sicher ist (Konfidenz < 0,5) kann modell h2 sicherer sein

Modelle h1 & h2

- Modell h1:

  - Reduktionsgrad: 30%
  - Punktzahl: 19923 (Test:61% Train:83%)
  - Aber Optionen :
    - Verstärkung anwenden
    - Punktzahl: 22006 (Test:66%Train:91,51%)
    - Rechenaufwand grösser.!

Kostenmatrix:

	0	1
0	0	0.5
1	2.5	0

- Modell h2:

  - Reduktionsgrad: 30%
  - Punktzahl: 20100 (Test: 67,76% Train: 87%)
  - Mit Verstärkerung : Punktzahl: 22525

Kostenmatrix :

	0	1
0	0	3
1	10	0

Gruppe 2:
Patricia, Muhannad, Hong, Marusa, Jingyu, Dominik

- Modell h1 + h2:
  - Verschiedene Bedeutsamkeit der Variablen
  - Übereinstimmen nicht in 18,85% der Fälle

  - Fusion: Senden wir voucher wenn maximum der Konfidenz beide Modelle vorhersagt, dass der Kunde mit der Wahrscheinlichkeit >0.6 nicht kaufen wird.

  - Punktzahl: 21152

Muhannad:

Neuronale Netze

Schneller Algorithmus

- Wählt eine geeignete Topologie für das Netz aus
  - Faustregel (rule of thumb)
  - Eigenschaften der Daten
- Eigenschaften: kleine verdeckte Schichten
  - Schneller zu trainieren
  - Besser zu verallgemeinern
- Modus
  - Einfach
  - Experte: Anzahl der verdeckten Schichten/Einheiten
    - Höhere Trainingzeit
    - Bessere Qualität

Schneller Algorithmus (forts.)

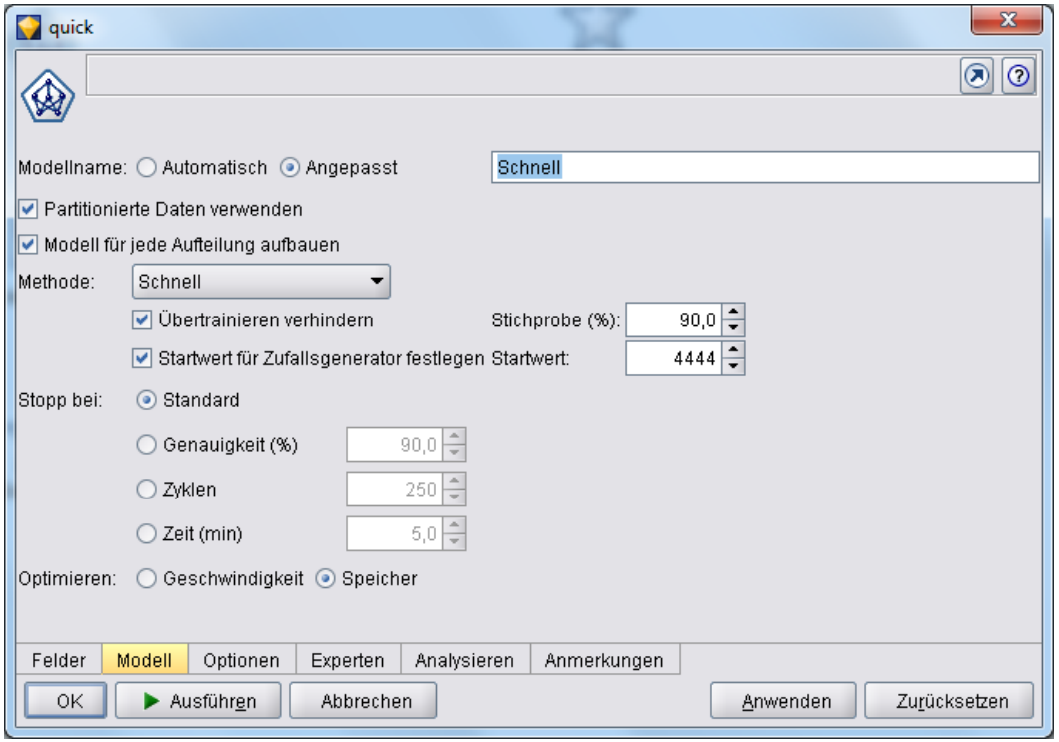
- Einfacher Modus:
  - Stichprobe: 50%, 75%, 90%, 100%
  - Punktzahl: 9310.5
- Experten Modus
  - Stichprobe: 50%, 75%, 90%, 100%
  - Anzahl der Schichten: 2, dann 3
  - Punktzahl: 9310.5 (gleich)
  - Grund: alle Modelle bis hier → 0
- Anderee Algorithmus: Dynamisch (nächste Folie)

Dynamischer Algorithmus

- Anfangstopologie erstellen
- Im Laufe des Training abwandeln:
  - Hinzufügen/Entfernen von verdeckten Einheiten
- Ergebnisse
  - Stichprobe: 50% (45 Minuten)
  - Stichprobe: 75% (circa: 2 Stunden)
  - Punktzahl: 9310.5 (keine Verbesserung!)
- Voraussichtlicher Grund: Daten-Vorverarbeitung
  - Verschiede Daten-Vorverarbeitung: NADA!
- Geringe Verbesserung (nächste Folie)

Startwert

- Startwert: 4444 (zufällig)
- Stichprobe: 90%
- Punktanzahl: 9590! (Geringer Fortschritt)



Gruppe 2:
Patricia, Muhannad, Hong, Marusa, Jingyu, Dominik

Vielen Dank für Euer Interesse!

Gruppe 2:
Patricia, Muhannad, Hong, Marusa, Jingyu, Dominik