Dr. Andreas Hotho, Robert Jäschke
Fachgebiet Wissensverarbeitung

5.2.2009

12. Übung „Knowledge Discovery“

Wintersemester 2008/2009

1 Preprocessing und Datenbeschaffenheit

1. Diskutieren Sie das Verhältnis von Preprocessing und Datenverständnis.

   Das Verständnis der Daten ist ein wesentlicher Schritt für die korrekte Ableitung von Preprocessing-Schritten. Nur mit dem nötigen Verständnis der Daten ist der Analytiker in der Lage, die richtigen Schritte so abzuschätzen, dass die Daten für die Verarbeitung durch den Algorithmus in adäquater Form vorliegen. Mit Hilfe von statistischen Datencharakteristiken oder einer explorativen Analyse/Visualisierung kann man Preprocessingschritte wie Reduktion, Ableitung und Transformation für einen entsprechenden Datensatz bestimmen und deren Erfolg wiederum überprüfen. Eine klare Trennung von Datenverständnis und Preprocessing ist daher selten gegeben.

2. In der Statistik unterscheidet man die Attributausprägungen einer vorgegebenen Menge von Daten mittels Skalen mit unterschiedlichem Skalenniveau. Nennen Sie die drei wichtigsten Skalenniveaus und beschreiben Sie sie kurz; geben Sie jeweils ein Beispiel an. Was bedingt ein Skalenniveau bei der Untersuchung von Daten?

   Nominalskalierte Merkmale: Ausprägungen sind qualitativ, keine Ordnung möglich (rot, grün).

   Ordinalskalierte Merkmale: Ausprägungen können geordnet, aber Abstände nicht interpretiert werden (Tafelwein, Qualitätswein, prämierter Qualitätswein).

   Kardinalskalierte Merkmale: Ausprägungen sind Zahlen, Interpretation der Abstände möglich (metrisch).

   mögliche Auswirkungen:

   - Informationsgehalt der Daten
   - sinnvolle Anwendbarkeit von Rechenoperationen

3. Welches Ziel wird in Bezug auf die spätere Anwendung von Algorithmen mit dem Preprocessing der Daten im Data Mining verfolgt? Nennen Sie in diesem Zusammenhang zwei Beispiele, in denen Algorithmen bestimmte Preprocessingschritte erzwingen.

Die Daten müssen so vorverarbeitet werden, dass die in Ihnen enthaltenen Informationen bestmöglich dem anzuwendenden Verfahren zur Verfügung stehen.

SVM brauchen z. B. numerische Daten

Entscheidungsbäume funktionieren besser mit kategorischen Daten

4. Nennen Sie je fünf Preprocessingschritte zu Data Cleansing und Data Manipulation und beschreiben Sie drei davon genauer.

Data Cleansing

- consistency (Konsistenz)
- detail/aggregation level (Aggregationsniveau)
- pollution (Verunreinigung)
- relationship (Beziehungen)
- range (Definitionsbereich)
- defaults
- duplicate or redundant variables
- missing and empty values (fehlende Werte)

Data Manipulation

- reverse pivoting
- reducing dimensionality
- increasing dimensionality
- sparsity (schwach besetzte Werte)
- monotonicity (Monotonie der Daten)
- outliers (Ausreisser)
- numerating categorical values
- anachronisms
- relation between variable via pattern in the variable
- combinatorial explosion

Eine genauere Beschreibung der Schritte ist im Skript zu finden.

2 Vergleich verschiedener Verfahren

1. Überlegen Sie sich einen Datensatz, der sich gut für ein Clustering mit OPTICS eignet, aber schlecht for kMeans. Finden Sie auch ein Beispiel, bei dem die Umkehrung gilt?

2. Finden Sie einen Datensatz, der schlecht für eine Klassifikation mit einem Entscheidungsbaum geeignet ist, dafür aber gute Ergebnisse bei kNN bringt. Finden Sie auch ein Beispiel, bei dem die Umkehrung gilt?