Data and Webmining WS 2018/2019
Übungsblatt 1

Aaron Winziers - 1176638; Michael Wolz - 1195270

13. November 2018

Aufgabe 1

a)

1. Verstehen der Anwendungsdomäne

   Identifikation der verfügbaren Daten

   Festlegung des KDD Ziels

2. Zieldatenfestlegung (Selektion)

   Festlegung der Datenbanken, Datensätze, Attribute die untersucht werden sollen.

3. Vorverarbeitung und Datenbereinigung

   Erkennung und Eliminierung von Datenfehlern (Ausreißern) und von fehlenden Einträgen

4. Datenreduktion und Projektion (Transformation)

   Identifikation der nützlichen Attribute für die KDD Aufgabe

   Reduktion der Dimension (Attribute)

   Berechnung abgeleiteter Attribute

   Reduktion der zu bearbeitenden Daten (Sampling)

5. Auswahl der Data Mining Aufgabenklasse

   um welche Art von Data Mining Aufgabe handelt es sich, z.B. Klassifikation, Regression, Assoziation, Clustering

6. Wahl des Data Mining Algorithmus

   für den gewählt Aufgabenklasse: bestimme einen geeigneten Algorithmus

   je nach Algorithmus: Bestimmung von Modellparametern

7. Data Mining durchführen

   Anwendung des Algorithmus auf den vorverarbeiteten Daten

8. Interpretation

   gefundene Muster werden interpretiert

   ggf. weitere Iteration und Wiederholung der Schritte 1-7

9. Konsolidierung des KDD Ergebnisses

   Präsentation der Ergebnisse und Dokumentation

b)

c)

Aufgabe 2

a)

Klassifikation

- Einordnung einer Beobachtung in eine von n Klassen
- Klassenbeschreibung = gelernte Regeln
- Regeln werden angeendet um neue Beobachtungen zu klassifizieren

Regression

- Lernen einer Funktion zur Abbildung einer Beobachtung auf einen Zahlenwert
- Approximation einer Datenmenge durch mathematische Funktion
- Ziel: möglichst gute Datenerfassung, d.h. möglichst geringe Fehler machen
- Durch Funktion können Vorhersagen getroffen werden

Clustering

- Einteilung von Datensätzen in Gruppen, z.B. disjunkte Gruppen oder hierarchische Gruppierungen
- Ermittlung von Klassen für unklassifizierte Datenpunkte
- Objekte möglichst Homogen
- Cluster untereinander möglichst heterogen

Abhängigkeitsanalyse

- Erkennen von gesetzmäßigen Abhängigkeiten in Daten, z.B. in Form von Regeln

b)

Synthetisches Lernen

- Bildung induktiver Schlüsse
- Wahrheitsgehalt der Schlussfolgerung nicht gesichert
- Neues Wissen wird geraten

Analytisches Lernen

- Bildung deduktiver Schlüsse
- wahrheitserhaltend: Schlussfolgerungen sind nachweislich korrekt
- Man lernt hier eigentlich kein wirklich neues Domänenwissen

Lernen durch Analogie

- Bildung analoger Schlüsse
- Wahrheitsgehalt der Schlussfolgerung nicht gesichert
- Basieren auf Ähnlichkeit

Aufgabe 3

a)

Klassifikator

- Ein Klassifikator für eine Menge M ist eine Abbildung $f : M \to I$, wobei I eine Menge ist, die Indexmenge genannt wird.
- Wenn $I = \{0, 1\}$, dann heißt

  $P = \{x \in M | f(x) = 1\}$ die Menge der *positiven* Elemente und

  $N = \{x \in M | f(x) = 0\}$ die Menge der *negativen* Elemente

Klassifikationsbeschreibung

- Unterscheide: Klassifikator und Klassifikatorbeschreibung
- Verschiedene Möglichkeiten der Beschreibung:

  Aufzählung aller Elemente (nur bei endlicher Grundmenge)

  Angabe einer prädikatenlogischen Formel

  Angabe eines C-Programms

  ...
- Beachte

  Eine Klassifikatorbeschreibung bestimmt einen eindeutigen Klassifikator

  Ein Klassifikator kann mehrere unterschiedliche Klassifikatorbeschreibungen besitzen

Konzeptbegriff

- Konzepte sind Klassifikatorbeschreibungen
- Definition: Ein Konzept ist ein einstelliges Prädikat über einer Grundmenge M.
- Schreibweise:

  Für $x \in M$,

  K(x) : x gehört zum Konzept K (x ist ein positives Beispiel)

  ¬K(x) : x gehört nicht zum Konzept K (x ist ein negatives Beispiel)

Aufgabe 4

b)

Sie verändern sich nicht da die Precision und recall unabhängig von der Menge der True-Negatives(78) sind und die Menge der False-Positives (4) sich nicht verändert

c)

Recall = 1 : Kommt in diesem Fall vor wenn alle relevante Dokumente tatsächlich als relevant klassifiziert werden.

Precision = 1 : Kommt in diesem Fall vor wenn alle als relevant klassifizierte Dokumente tatsächlich relevant sind.

Wenn recall = 1, ist das Konzept vollständig da die vollständige Menge der relevanten Dokumente als relevant klassifiziert werden.