Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Peer Kröger
Anna Beer, Florian Richter

Knowledge Discovery in Databases

WS 2017/18

Übungsblatt 1: Aufgabenstellungen im Data Mining

Besprechung: 09. und 10.11.2017

Aufgabe 1-1 Data Mining Aufgaben

Welche Aufgaben für das Data Mining (Clustering, Outlier Detection, Klassifikation, etc.) verbergen sich hinter den folgenden Anwendungen? Ist die Aufgabe überwacht (supervised) oder nicht überwacht (unsupervised)?

(a) **Texterkennung/OCR**:
Beim Passieren der Brennerautobahn existiert seit einigen Jahren die Möglichkeit per E-Maut zu zahlen. Dabei wird bei Zahlungseingang das Nummernschild des Autos registriert. Beim Passieren der Mautstation fährt das Auto dann durch eine gesonderte Schranke die nur aufgeht, wenn das Nummernschild des Fahrzeugs als registriert erkannt wurde. Die Erkennung erfolgt dabei voll automatisch per digitaler Kamera.

(b) **Computer Aided Diagnosis:**
Patienten, die an Blutkrebs leiden, können in zwei Kategorien (ALL und AML) eingeteilt werden. Da sich die Therapien dieser beiden Arten teilweise sehr stark unterscheiden und sogar manchmal die Therapie für AML sehr schädlich für ALL-Patienten sein kann (und umgekehrt), versucht man, neue Patienten anhand von speziellen Daten (sog. Gen-Expressionsdaten) zu unterscheiden. Dazu werden die Daten der neuen Patienten mit den Daten der Patienten, deren Blutkrebstyp bereits bekannt ist, verglichen.

(c) **Cheat Detection**
Der Betreiber eines Multiplayer-Online-Spiels will sein System gegen verschiedene Verstöße der Benutzerrichtlinien abdecken. Dazu gehören die Verwendung von Bot-Programmen, das Manipulieren von Zeitstempeln im Kommunikation Protokoll und die Vorhersage verwendeter Zufallszahlen. Zur Erkennung von verdächtigem Verhalten wird Data Mining auf den erhältlichen Benutzerdaten verwendet.

(d) **Mensch und Maschine**
Moderne WWW-Suchmaschinen beantworten Benutzeranfragen, die aus nur einem oder wenigen Suchtermen bestehen. In der Regel liefert eine Anfrage dabei eine sehr große Ergebnismenge, die mit Hilfe eines Ranking Algorithmus nach Relevanz sortiert wird. Durch diese Sortierung kann der User dann selber entscheiden, wieviele Links er besuchen will. Die Problematik hierbei ist zum einem den Inhalt einer Ergebnisseite richtig zu erkennen. Zum anderen besteht die Notwendigkeit, dass wirklich hilfreiche Seiten höher gerankt werden als weniger hilfreiche Seiten, auch wenn beide inhaltlich zum Suchbegriff passen. Wortmehrdeutigkeiten stellen dabei ebenfalls ein großes Problem dar. Zum Beispiel kann sich die Suche nach dem Begriff "Golf" auf das Auto, den Sport oder den geographischen Begriff beziehen. Data Mining Techniken werden hier eingesetzt, um das Ranking zu optimieren und mögliche Ergebnis-Mengen nach dem jeweiligen Begriffskontext zu gruppieren.

(e) **Recommendation Systems**
Ein Online-Kaufhaus möchte für registrierte Kunden Artikel bestimmen, die dem Kunden beim Einloggen unaufgefordert angeboten werden. Dabei kann man auf die bereits gekauften Artikel des Kunden

zurückgreifen, um so die Interessengebiete des Kunden besser vorhersagen zu können. Zum Beispiel bietet es sich an, jemandem, der das Buch “Herr der Ringe” gekauft hat, auch die DVDs der Verfilmung anzubieten. Eine weitere ähnliche Aufgabe ist die Bestimmung von geeigneten Kombiangeboten zu einem bereits ausgewählten Artikel.

(f) **News Aggregation**
Eine Nachrichtenseite sammelt automatisch Meldungen aus verschiedenen Nachrichtenquellen um den Nutzer zu informieren. Da es aber häufig passiert, dass unterschiedliche Seiten über das selbe Thema berichten, sollen die Meldungen gruppiert werden. Die Überlappungen passieren dabei auf unterschiedlichen Ebenen: Zum einen gibt es natürlich breite Kategorien wie Sport und Politik, und Unterkategorien wie Fußball. Aber auch zu einem einzelnen Fußballspiel gibt es meist mehrere unterschiedliche Beiträge auf unterschiedlichen Seiten. Manche der Beiträge sind (weitgehend) identisch zu Agenturmeldungen, andere individuelle eigene Beiträge.

(g) **Extraktion von Daten / Web Scraping**:
Aus einer bekannten Filmdatenbank sollen eine Liste von Filmen und eine Liste von Schauspielern extrahiert werden (Lizenzprobleme seien für diese Aufgabe ignoriert).

(h) **Identifikation der wichtigsten Zulieferer**:
Ein großer Onlinehändler möchte wissen, welche Lieferanten für ihn am wichtigsten sind, d.h. den größten Umsatz beisteuern. Zu diesen könnten dann engere Beziehungen geknüpft werden, eine Übernahme der Firma erfolgen, oder ein neues Logistikzentrum nahe des Standortes des Lieferanten entstehen, um die Lieferzeiten zu verkürzen.

(i) **Bildsegmentation in medizinischen Bilddaten**:
Segmentation nennt man den Process des Unterteilens eines Bildes in verschiedene Teile. In der medizinischen Bildbearbeitung bedeuten diese Segmente meist verschiedene Zelltypen, Organe oder Pathologien, oder andere biologisch relevanten Strukturen. Medizinische Bildverarbeitung wird erschwert durch schlechte Bildqualität, niedrigen Kontrast und Rauschen oder andere Bildunklarheiten. Auch wenn es für Bilder schon viele Methoden gibt, werden und müssen diese meist noch für die Verwendung im medizinischen Bereich angepasst werden.

Begriffe in diesem Themenbereich sind unter anderem:

(i) Atlas-Based Segmentation:
Ein Experte bewertet einige Beispielbilder, anhand derer dann durch Extrapolation eine Aussage über das neue Bild gemacht wird. Dabei wird von den Trainingsdaten abstrahiert und daraus ein Modell entwickelt.

(ii) Shape-Based Segmentation:
Bei dieser Methode werden meist parametrisierte Modelle von Formen verwendet, die sich auf besondere Strukturen und Verläufe beziehen. Dabei wird die Form verändert, um dem neuen Bild zu gleichen

(iii) Interactive Segmentation:
Ein Arzt gibt während der Operation Informationen, wie zum Beispiel die Region oder die Grenze zu einem Segment. Ein Algorithmus kann dann die Zwischenergebnise verfeinern und somit genauer die Ausmaße eines Zelltypus definieren.

Aufgabe 1-2 Skalen-Niveaus von Merkmalen

Entscheiden Sie für jedes Merkmal des folgenden Datensatzes, ob es sich um ordinale, nominale oder metrische Merkmale handelt.

Obs.	Geschlecht	Grösse (cm)	Gewicht (kg)	Haarfarbe	Blutgruppe	Brille	Rauchen	Wohnlage
67	Frau	175	60	dunkelbl./braun	A	nein	gelegentlich	ruhig
68	Frau	176	52	hellblond	AB	ja	gelegentlich	ruhig
69	Frau	176	63	schwarz	A	ja	selten	sehr ruhig
70	Frau	179	65	dunkelbl./braun	0	ja	nie	ruhig
71	Frau	180	62	dunkelbl./braun	B	ja	nie	ruhig
72	Frau	180	70	dunkelbl./braun	A	ja	nie	ruhig
73	Frau	185	72	dunkelbl./braun	B	nein	nie	sehr ruhig
74	Frau	195	62	rot	0	ja	sehr viel	sehr ruhig
75	Frau	203	62	rot	AB	ja	sehr viel	sehr lärmig
76	Mann	165	53	dunkelbl./braun	A	nein	selten	ruhig
77	Mann	169	63	dunkelbl./braun	B	ja	selten	ruhig
78	Mann	169	72	dunkelbl./braun	A	nein	nie	ruhig
79	Mann	170	61	dunkelbl./braun	A	nein	nie	sehr ruhig
80	Mann	171	71	dunkelbl./braun	A	nein	viel	lärmig
81	Mann	173	61	schwarz	A	ja	nie	sehr ruhig
82	Mann	173	63	rot	A	nein	selten	lärmig
83	Mann	173	67	dunkelbl./braun	B	ja	nie	ruhig
84	Mann	175	68	dunkelbl./braun	.	nein	nie	ruhig
85	Mann	175	71	dunkelbl./braun	AB	nein	viel	ruhig
86	Mann	176	60	dunkelbl./braun	A	nein	selten	ruhig
87	Mann	177	64	dunkelbl./braun	AB	nein	nie	sehr lärmig

Aufgabe 1-3 Distanzmaße

Distanzmaße können wir nach ihren Eigenschaften in folgende Kategorien einteilen:

$d : S \times S \to \mathbb{R}_0^+$ $x, y, z \in S$:	reflexiv reflexive $x = y \Rightarrow d(x, y) = 0$	symmetrisch symmetric $d(x, y) = d(y, x)$	strikt strict $d(x, y) = 0 \Rightarrow x = y$	Dreiecksungleichung Triangle inequality $d(x, z) \leq d(x, y) + d(y, z)$
Unähnlichkeitsfunktion Dissimilarity function	$\times$			
(Symmetrische) Prämetrik (Symmetric) Pre-metric	$\times$	$\times$		
Semimetrik, Ultrametrik Semi-metric, Ultra-metric	$\times$	$\times$	$\times$	
Pseudometrik Pseudo-metric	$\times$	$\times$		$\times$
Metrik Metric	$\times$	$\times$	$\times$	$\times$

D.h., wenn ein Distanzmaß $d : S \times S \to \mathbb{R}_0^+$ für alle $x, y, z \in S$: reflexiv, symmetrisch und strikt ist sowie die Dreiecks-Ungleichung erfüllt, ist es eine Metrik. Wie Sie sehen, muß z.B. eine Prämetrik nicht *strikt* reflexiv sein. Machen Sie sich den Unterschied zwischen Reflexivität und Striktheit klar!

Anmerkung: Die Namen Distanzfunktion, Semi- und Pseudo-Metrik werden in der Literatur nicht einheitlich definiert.

Entscheiden Sie für die folgenden Funktionen $d(\mathbb{R}^n, \mathbb{R}^n)$ jeweils, ob es sich um ein Distanzmaß handelt, und wenn ja, in welche Kategorie es fällt.

(a) $d(x, y) = \sum_{i=1}^{n} (x_i - y_i)$

(b) $d(x, y) = \sum_{i=1}^{n} (x_i - y_i)^2$

(c) $d(x, y) = \sqrt{\sum_{i=1}^{n-1} (x_i - y_i)^2}$

(d) $d(x, y) = \sum_{i=1}^{n} \begin{cases} 1 & \text{falls} \quad x_i = y_i \\ 0 & \text{falls} \quad x_i \neq y_i \end{cases}$

(e) $d(x, y) = \sum_{i=1}^{n} \begin{cases} 1 & \text{falls} \quad x_i \neq y_i \\ 0 & \text{falls} \quad x_i = y_i \end{cases}$