

Übungsaufgaben zur Lehrveranstaltung

DATA MINING

im Studiengang Informatik

modifizierte Version (einige Aufgaben entfernt und eigene Aufgaben hinzugefügt) der Aufgaben aus
Tan/Steinbach/Kumar: *Introduction to Data Mining*. Pearson Education, 2006.

Technische Universität Ilmenau
Fakultät für Informatik und Automatisierung
Fachgebiet Künstliche Intelligenz
apl. Prof. Dr.-Ing. habil. Rainer Knauf
Sommersemester 2012
Stand 09. Juli 2012

0 Einführung

1. Für folgende Aufgaben entscheide man, ob es sich um eine Data Mining Aufgabe handelt:
 - (a) Partitionierung von Kunden eines Unternehmens bzgl. Geschlecht
 - (b) Partitionierung von Kunden eines Unternehmens bzgl. Profitabilität
 - (c) Ermittlung des Umsatzes eines Unternehmens
 - (d) Sortierung einer Studentendatenbank nach Matrikel-Nummer
 - (e) Voraussage eines Würfelerggebnisses
 - (f) Vorhersage von Aktienkursen durch Analyse des Kursverlaufs der Vergangenheit
 - (g) Ableitung des Dauer - EKG zwecks Erkennung von Anomalien
 - (h) Aufzeichnung seismischer Wellen zur Erkennung der Erdbebenaktivität
 - (i) Ermittlung der Frequenz eines Tones.
2. Sie beraten ein Suchmaschinen-Betreiber. Erklären Sie, in ihm welcher Weise Data Mining Verfahren hilfreich sein können! Benennen Sie konkrete Beispiele dafür, wie die Techniken Clustering, Klassifikation, Assoziationsregel-Generierung und Anomalie-Erkennung angewandt werden können!
3. Welche der u.g. Daten bergen die Gefahr des Missbrauchs persönlicher Daten:
 - (a) Daten aus Volkszählungen der Jahre 1900 bis 1950
 - (b) IP-Adressen und Besuchszeiten von Web-Nutzern auf Ihrer Webseite
 - (c) Bilder erdumkreisender Satelliten
 - (d) Namen und Adressen von Telefonbuch-Einträgen
 - (e) im Web gesammelte Namen und e-Mail Adressen

1 Daten

1. Jedes der folgenden Datenobjekte beschreibt medizinische Informationen über einen Patienten in je einem 5-Tupel von Werten:

012	232	33.5	0	10.7
020	121	16.9	2	210.1
027	156	24.0	0	427.6
...

Ein Statistiker wirft einen Blick auf diese Daten und meint *“Feld 2 und 3 sind im Wesentlichen dasselbe.”* Wie kommt er darauf?

2. Man klassifiziere die Attribute in Tabelle 1 als binär, diskret oder kontinuierlich sowie als qualitativ (nominales oder ordinales Attribut) oder quantitativ (Intervall-Attribut oder rationales Attribut).
3. Der Marketing Direktor beschreibt Ihnen eine Idee, die Kundenzufriedenheit zu ermitteln, wie folgt:

“Es ist so einfach, dass ich gar nicht glauben mag, dass noch niemand vorher darauf gekommen ist. Ich registriere einfach die Anzahl von Kundenbeschwerden für jedes einzelne Produkt. In einem Data Mining Buch las ich, dass die Anzahl ein rationales Attribut ist. Aber als ich die Produkte mit meinem Kundenzufriedenheits-Maß bewertete und das Ergebnis meinem Chef präsentierte, sagte er, ich hätte etwas Offensichtliches übersehen und meine Bewertung ist nicht zu gebrauchen. Ich glaube, er war nur verwirrt, weil unser Bestseller-Produkt nach dieser Bewertung die schlechteste Kundenzufriedenheit aufweist.

Können Sie mir helfen, ihn vom Nutzen meiner Methode zu überzeugen?”

- (a) Wer hat Recht, der Marketing-Direktor oder der Chef?
Wenn Sie meinen, der Chef habe Recht, schlagen Sie ein besseres Maß zur Ermittlung der Kundenzufriedenheit vor!
 - (b) Hat der Marketing-Direktor den Attribut-Typ seines Zufriedenheitsmaßes richtig klassifiziert?
4. Einige Monate später erscheint o.g. Marketing Direktor erneut bei Ihnen und schlägt ein besseres Maß der Kundenzufriedenheit über ein Produkt oder eine Produktgruppe vor:

“Wenn wir neue Produkte entwickeln, schaffen wir typischerweise zunächst mehrere Varianten und evaluieren, welche am besten von den Kunden angenommen wird. Wir geben den Test-Kunden alle Produktvarianten und bitten Sie, diese entsprechend ihrer Präferenz zu ordnen. Allerdings sind unsere Test-Kunden sehr unentschlossen, insbesondere wenn es sich um mehr als zwei Varianten handelt. Deshalb testen wir mitunter ewig.

¹Eine ISBN besteht aus 13 Ziffern: Dem 3-stelligen Präfix, das die Buchbranche identifiziert (gegenwärtig 978), gefolgt von der 9-stelligen Kernnummer und der ermittelten Prüfziffer, die die Integrität der gesamten 13-stelligen ISBN bestätigt.

Attribut	diskret		kontinuierlich	qualitativ		quantitativ	
	binär			nominal	ordinal	Intervall	rational
Zeitangabe "Vormittag" und "Nachmittag"							
Helligkeit, gemessen durch Belichtungsmesser (Lichtstrom ϕ pro Flächeneinheit in lux)							
"gefühlte" Helligkeit, durch Menschen wahrgenommen							
Winkel, gemessen in Grad (zwischen 0 und 360)							
Bronze-, Silber und Goldmedaille als sportliches Ergebnis							
Höhe über dem Meeressgrund (NN)							
Anzahl der Patienten eines Krankenhauses							
ISBN Nummer eines Buches ¹							
Lichtdurchlässigkeit mit den Werten "undurchlässig", "halbdurchlässig" und "transparent"							
militärische Dienstgrade							
Weglänge vom Ortszentrum zum Uni-Campus							
Dichte eines Stoffes in g/cm^3							
Nummer einer Garderoben-Marke							
Kraftfahrzeug-Kennzeichen in Deutschland							

Table 1: Klassifizierung von Attributen

Ich schlug vor, diese Vergleiche immer nur paarweise durchzuführen und die Ergebnisse für das Ranking zu nutzen. Wenn man z.B. 3 Varianten hat, bittet man die Test-Kunden 1 mit 2, 2 mit 3 und 3 mit 1 zu vergleichen.

Auf diese Weise konnten wir die Testzeit auf ein Drittel senken. Allerdings klagen unsere Mitarbeiter darüber, dass man aus diesen Ergebnissen kein konsistentes Ranking ableiten kann.”

- (a) Ist dieses Verfahren geeignet, um aus den resultierenden Daten ein Ranking abzuleiten? Begründen Sie Ihre Ansicht!
 - (b) Kann man diesen Ansatz so verfeinern, dass er zum Ranking geeignet ist? Mit anderen Worten: Kann man aus dem Ergebnis des paarweise Vergleichens ein ordinales Maß ableiten?
 - (c) In der originalen Version des Rankings wurde für jede Variante ein durchschnittliches Ranking (ein durchschnittlicher Platz “auf der Hitliste” der Test-Kunden) als arithmetischer Mittelwert ermittelt. Halten Sie das für sinnvoll? Können Sie eine bessere Methode vorschlagen?
5. Können Sie sich eine Situation vorstellen, bei der eine Identifikations-Nummer für eine Vorhersage nützlich sein kann?
 6. Ein Bildungspsychologe möchte Testresultate mit einer Assoziations-Analyse überprüfen. Der Test besteht aus 100 Fragen mit je 4 möglichen Antworten.
 - (a) Wie würden Sie die Testdaten konvertieren, damit sie für eine solche Analyse geeignet sind?
 - (b) Welchen Attributstyp würden Sie dabei bekommen und wie viele Attribute würden dabei entstehen?
 7. Welche dieser beiden Messwert-Reihen hat wahrscheinlich eine höhere temporale Autokorrelation
 - (a) der tägliche Temperaturverlauf über viele Tage oder
 - (b) die täglich aufgezeichnete Regenmenge über viele Tage?

Begründen Sie Ihre Antwort!

8. Man diskutiere, warum die Attribute einer Dokument-Term-Matrix asymmetrisch diskret oder asymmetrisch kontinuierlich sind!
9. Für eine Herde asiatischer Elefanten werden folgende Attribute gemessen: Gewicht, Höhe, Stoßzahn-Länge, Rüssel-Länge, Ohr-Fläche. Welches Ähnlichkeitsmaß schlagen Sie vor, um Elefanten zu vergleichen oder zu clustern?
10. Diese Aufgabe dient dem Vergleich verschiedener Ähnlichkeitsmaße.
 - (a) Berechnen Sie
 - i. die Hamming Distance (L_1 Norm)
 - ii. den Jaccard-Koeffizienten
 für die beiden binären Vektoren $x = [0101010001]$ und $y = [0100011000]$.

- (b) Welches der beiden o.g. Ähnlichkeitsmaße (Hamming oder Jaccard) kommt dem Simple Matching Coefficient näher und welches dem Kosinus Koeffizienten¹?
- (c) Die genetische Struktur einer Spezies sei repräsentiert als binärer Vektor, in welchem ein Attribut den Wert 1 hat, falls das Tier ein bestimmtes Gen besitzt und 0, falls dem nicht so ist.
Stellen Sie sich vor, Sie wollen ermitteln, wie ähnlich zwei Organismen verschiedener Spezies sind, indem Sie ermitteln, wie viele gleiche Gene beide haben. Welchen der beiden Ansätze würden Sie bevorzugen, Hamming oder Jaccard? Warum?
- (d) Wenn Sie zwei Organismen der selben Spezies, z.B. zwei Menschen, genetisch vergleichen wollen, würden Sie dann Hamming, Jaccard oder ein anderes Ähnlichkeitsmaß verwenden?²

11. Ermitteln Sie folgende Ähnlichkeits- bzw. Distanzmaße!

- (a) Kosinus Koeffizient, Korrelation, und Euklidische Distanz von $x = [1, 1, 1, 1]$ und $y = [2, 2, 2, 2]$.
- (b) Kosinus Koeffizient, Korrelation, Euklidische Distanz und Jaccard Koeffizient von $x = [0, 1, 0, 1]$ und $y = [1, 0, 1, 0]$.
- (c) Kosinus Koeffizient, Korrelation und Euklidische Distanz von $x = [0, -1, 0, 1]$ und $y = [1, 0, -1, 0]$.
- (d) Kosinus Koeffizient, Korrelation und Jaccard Koeffizient von $x = [1, 1, 0, 1, 0, 1]$ und $y = [1, 1, 1, 0, 0, 1]$.
- (e) Kosinus Koeffizient und Korrelation von $x = [2, -1, 0, 2, 0, -3]$ und $y = [-1, 1, -1, 0, 0, -1]$.

12. Diese Aufgabe dient dem Vergleich zwischen dem Kosinus Koeffizienten und der Korrelation.

- (a) Welchen Wertebereich hat der Kosinus Koeffizient?
- (b) Wenn zwei Datenobjekte einen Kosinus Koeffizienten von 1 aufweisen, sind sie dann identisch?
- (c) Gibt es einen Zusammenhang zwischen dem Kosinus Koeffizienten und Korrelation?³
- (d) Die Abbildung 1(a) zeigt den Zusammenhang zwischen dem Kosinus Koeffizienten und der Euklidischen Distanz für 100.000 zufällig generierte Punkte (Paare von Datenobjekten mit ausschließlich positiven Attribut-Werten), welche auf eine L_2 Norm (Länge bzw. Abstand zum Ursprung im n -dimensionalen Euklidischen Raum) von 1 normalisiert wurden.
Welche generelle Aussage läßt sich aus diesen Daten von 100.000 Punkten mit einer L_2 Norm von 1 ableiten?

¹Hamming ist ein Distanzmaß, während alle anderen in der Aufgabe erwähnten Maße tatsächlich Ähnlichkeitsmaße sind.

²Menschen haben 99.9 % aller Gene gemeinsam. Bei anderen Spezies verhält sich das ähnlich.

³Hinweis: Achten Sie auf statistische Größen wie Mittelwert und Standardabweichung für Fälle, in denen diese beiden Größen gleich bzw. verschieden sind.

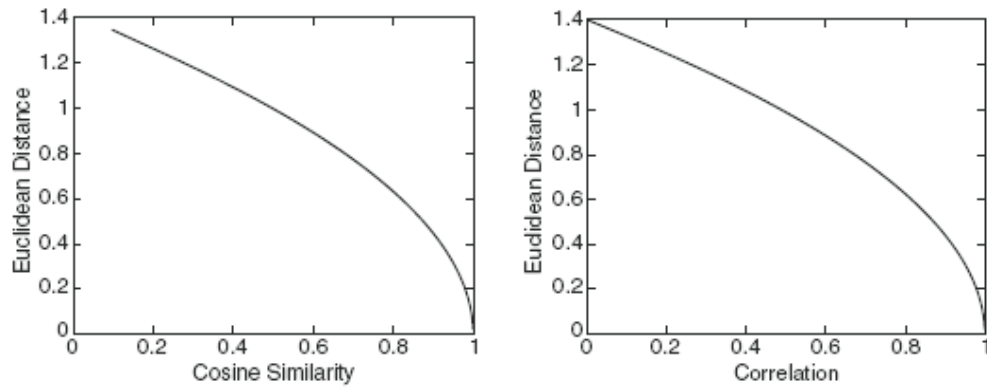


Figure 1: Euklidische Distanz vs. Kosinus Koeffizient und Korrelation

- (e) Die Abbildung 1 (b) zeigt den Zusammenhang zwischen der Korrelation und der Euklidischen Distanz für 100.000 zufällig generierte Punkte (Paare von Datenobjekten mit ausschließlich positiven Attribut-Werten), welche auf einen Mittelwert von 0 und eine Standardabweichung von 1 standardisiert wurden.

Welche generelle Aussage läßt sich aus diesen Daten von 100.000 Punkten mit einem Mittelwert von 0 und einer Standardabweichung von 1 ableiten?

- (f) Leiten Sie den mathematischen Zusammenhang zwischen Kosinus Koeffizient und Euklidischer Distanz für Daten mit einer Länge (nach L_2 Norm) von 1 ab!
- (g) Leiten Sie den mathematischen Zusammenhang zwischen Korrelation und Euklidischer Distanz für Daten ab, die durch
- Subtraktion des Mittelwertes und
 - Division durch die Standardabweichung
- standardisiert wurden!

13. Man zeige, dass das Distanzmaß für Datenobjekte A , und B

$$d(A, B) = |A - B| + |B - A|$$

die folgenden Eigenschaften aufweist:

- (a) d ist nicht negativ: $\forall X, Y : d(X, Y) \geq 0$
- (b) $\forall X, Y : d(X, Y) = 0$ gdw. $X = Y$
- (c) d ist symmetrisch: $\forall X, Y : d(X, Y) = d(Y, X)$
- (d) $\forall X, Y, Z : d(X, Z) \leq d(X, Y) + d(Y, Z)$
14. Wie würden Sie Zeitreihen eines Attributs von einem Wertebereich $[-1, 1]$ für folgende Anwendungsszenarien auf einen Wertebereich $[0, 1]$ abbilden?
- (a) Partitionieren (Clustern) zu Gruppen mit sehr ähnlichem Datenobjekten
- (b) Vorhersage des Betrags des Wertes (nicht der Richtung) zukünftiger Zeitreihen.
15. Schlagen Sie zwei Möglichkeiten vor, ein Ähnlichkeitsmaß mit einem Wertebereich $[0, 1]$ in ein Distanzmaß mit einem Wertebereich $[0, \infty]$ zu transformieren!

16. Distanzmaße sind gewöhnlich über Paaren von Objekten definiert.

- (a) Schlagen Sie zwei Möglichkeiten vor, Distanzmaße auf eine Menge von (mehr als zwei) Objekten zu definieren!
- (b) Wie könnte man die Distanz zwischen zwei Punktmengen im Euklidischen Raum definieren?

2 Klassifikation

2.1 Entscheidungsbäume

1. Entwickeln Sie einen vollständigen Entscheidungsbaum für die Paritätsfunktion (siehe Tabelle 2)! Lässt sich der Baum vereinfachen?

A	B	C	D	Klasse
F	F	F	F	T
F	F	F	T	F
F	F	T	F	F
F	F	T	T	T
F	T	F	F	F
F	T	F	T	T
F	T	T	F	T
F	T	T	T	F
T	F	F	F	F
T	F	F	T	T
T	F	T	F	T
T	F	T	T	F
T	T	F	F	T
T	T	F	T	F
T	T	T	F	F
T	T	T	T	T

Table 2: Paritätsfunktion

2. Tabelle 3 zeigt das Trainings-Set für eine binäre Klassifikation.

- (a) Ermitteln Sie den Gini-Index der Gesamtmenge der Trainingsdaten bzgl. der Klassenzugehörigkeit!
- (b) Ermitteln Sie den Gini-Index des Attributs *Kundennummer* bzgl. der Klassenzugehörigkeit für dessen verschiedene Ausprägungen!
- (c) Ermitteln Sie den Gini-Index des Attributs *Geschlecht* bzgl. der Klassenzugehörigkeit für dessen verschiedene Ausprägungen!
- (d) Ermitteln Sie den Gini-Index des nominalen Attributs *Autotyp* bzgl. der Klassenzugehörigkeit für dessen verschiedene Ausprägungen!
- (e) Ermitteln Sie den Gini-Index des Attributs *Konfektionsgröße* bzgl. der Klassenzugehörigkeit für die verschiedenen Ausprägungen!

Kundennummer	Geschlecht	Autotyp	Konfektionsgröße	Klasse
1	m	Van	S	C0
2	m	Sport	M	C0
3	m	Sport	M	C0
4	m	Sport	L	C0
5	m	Sport	XL	C0
6	m	Sport	XL	C0
7	f	Sport	S	C0
8	f	Sport	S	C0
9	f	Sport	M	C0
10	f	Luxus	L	C0
11	m	Van	L	C1
12	m	Van	XL	C1
13	m	Van	M	C1
14	m	Luxus	XL	C1
15	f	Luxus	S	C1
16	f	Luxus	S	C1
17	f	Luxus	M	C1
18	f	Luxus	M	C1
19	f	Luxus	M	C1
20	f	Luxus	L	C1

Table 3: Trainingsdaten für eine binäre Klassifikation

- (f) Welches der Attribute Geschlecht, Autotyp und Konfektionsgröße ist am besten zum Splitting geeignet?
- (g) Erklären Sie, warum man das Attribut Kundennummer trotz seines guten Gini-Index nicht im Entscheidungsbaum verwenden sollte!

3. Tabelle 4 zeigt das Trainings-Set für eine binäre Klassifikation.

lfd. Nr.	a_1	a_2	a_3	Klasse
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

Table 4: Trainingsdaten für eine binäre Klassifikation

- (a) Wie hoch ist die Entropie der Trainingsdaten bzgl. Klassenzugehörigkeit?
- (b) Wie hoch ist der Informationsgewinn von a_1 und a_2 bzgl. dieser Trainingsdaten?

- (c) a_3 ist ein kontinuierliches Attribut. Ermitteln Sie den Informationsgewinn für jeden sinnvollen Split-Wert.
- (d) Welches der Attribute a_1 , a_2 und a_3 ist das beste bzgl. des Informationsgewinns?
- (e) Welches der Attribute a_1 , a_2 ist besser bzgl. des Klassifikationsfehlers?
- (f) Welches der Attribute a_1 , a_2 ist besser bzgl. des Gini-Index?
4. Tabelle 5 zeigt die Trainingsdaten für eine binäre Klassifikation.

A	T	T	T	T	T	F	F	F	T	T
B	F	T	T	F	T	F	F	F	T	F
Klasse	+	+	+	-	+	-	-	-	-	-

Table 5: Trainingsdaten für eine binäre Klassifikation

- (a) Berechnen Sie den Informationsgewinn beim Splitten bzgl. der Attribute A und B auf der Basis der Entropie! Welches Attribut erzielt den höheren Informationsgewinn?
- (b) Berechnen Sie den Informationsgewinn beim Splitten bzgl. der Attribute A und B auf der Basis des Gini-Index! Welches Attribut erzielt den höheren Informationsgewinn?
- (c) Bild 2 zeigt, dass der Gini-Index und die Entropie im Intervall $[0, 0.5]$ beide monoton steigen und im Intervall $[0.5, 1]$ beide monoton fallen. Ist es unter diesen Umständen möglich, dass der Informationsgewinn auf der Basis des Gini-Index ein anderes Attribut favorisiert als der auf der Basis der Entropie?

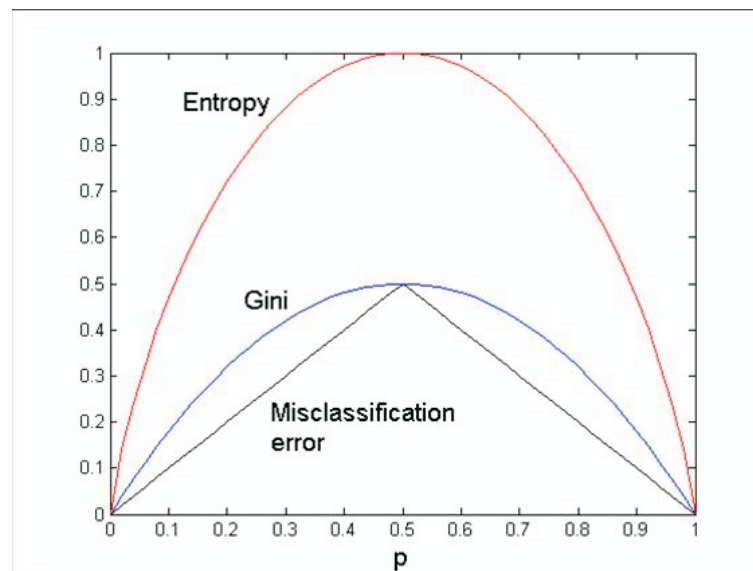


Figure 2: Gini-Index, Entropie und Klassifikationsfehler

5. Tabelle 6 zeigt die Klassenverteilung für Trainingsdaten einer binären Klassifikation.

X	Y	Z	Anzahl der Datensätze für Klasse +	Anzahl der Datensätze für Klasse -
0	0	0	5	40
0	0	1	0	15
0	1	0	10	5
0	1	1	45	0
1	0	0	10	5
1	0	1	25	0
1	1	0	5	20
1	1	1	0	15

Table 6: Trainingsdaten für eine binäre Klassifikation

- (a) Entwickeln Sie einen Entscheidungsbaum auf der Basis maximalen Informationsgewinns bzgl. des Klassifikationsfehlers! Wie hoch ist der Klassifikationsfehler des gesamten Entscheidungsbaumes?
- (b) Wiederholen Sie die Prozedur der Baumentwicklung, aber nehmen Sie X als Splitting-Attribut der 1. Ebene! Wie hoch ist dann der Klassifikationsfehler des gesamten Entscheidungsbaumes?
- (c) Vergleichen und diskutieren Sie die Ergebnisse der beiden Teilaufgaben!
6. Die Tabelle 7 fasst Trainings-Daten einer binären Klassifikation mit 3 binären Attributen zusammen.

A	B	C	Anzahl der Datensätze für Klasse	
			+	-
F	F	F	0	25
F	F	T	0	5
F	T	F	25	0
F	T	T	0	20
T	F	F	0	0
T	F	T	20	0
T	T	F	0	0
T	T	T	5	0

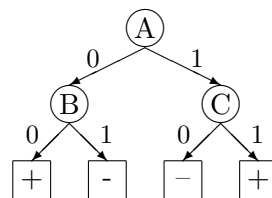
Table 7: Trainingsdaten für eine binäre Klassifikation

- (a) Ermitteln Sie für jedes Attribut den Klassifikationsfehler und den Informationsgewinn bezüglich des Klassifikationsfehlers, wenn man es zum Splitting auf der ersten Ebene heranzöge!
Welches Attribut ist demnach zu bevorzugen?
- (b) Setzen Sie die Baumkonstruktion für die zweite Ebene, d.h. für die dabei entstehenden Unterbäume des Wurzelknotens fort!
- (c) Wie viele Trainings-Datensätze werden durch den so entstandenen Entscheidungsbaum falsch klassifiziert?
- (d) Wiederholen Sie die Baumkonstruktion, wählen Sie dabei jedoch C als Splitting-Attribut auf der ersten Ebene! Wie hoch ist dann der Klassifikationsfehler?

- (e) Was kann man aus diesen Ergebnissen über die “gierige” (greedy) Heuristiken (mit frühzeitigem maximalen Informationsgewinn) ableiten?
7. Anhand der Trainingsdaten aus Tabelle 8 wurde ein Entscheidungsbaum gemäß Abbildung 3 entwickelt.

lfd. Nr.	A	B	C	Klasse
1	0	0	0	+
2	0	0	1	+
3	0	1	0	+
4	0	1	1	-
5	1	0	0	+
6	1	0	0	+
7	1	1	0	-
8	1	0	1	+
9	1	1	0	-
10	1	1	0	-

Table 8: Trainingsdaten für eine binäre Klassifikation

Figure 3: Aus Tabelle 8 abgeleiteter Entscheidungsbaum T

- (a) Ermitteln Sie den Generalisierungsfehler auf der Basis des “optimistic approach”, d.h. der Trainingsdaten!
- (b) Ermitteln Sie den Generalisierungsfehler auf der Basis des “pessimistic approach”! Verwenden Sie dabei einen Malus $\Omega(t_i) = 0.5$ für jeden Blattknoten t_i !
- (c) Ermitteln Sie den Generalisierungsfehler auf der Basis der Validierungsmenge in Tabelle 9!

lfd. Nr.	A	B	C	Klasse
11	0	0	0	+
12	0	1	1	+
13	1	1	0	+
14	1	0	1	-
15	1	0	0	+

Table 9: Validierungsdaten für den Entscheidungsbaum T (Abb. 3)

8. Zur Behandlung einer schweren Krankheit werden typischerweise die Therapien a , b , und c verordnet, welche nur sequentiell angewandt werden können. Zur Klasse P

gehören diejenigen Therapiefolgen, welche zur Heilung eines Patienten führten; zur Klasse N gehören diejenigen Therapiefolgen, die keinen Heilerfolg erbrachten.

Entwerfen Sie einen Entscheidungsbaum zur Beurteilung der Heilungschancen von Therapiefolgen für $P = \{bc, ac\}$ und $N = \{cb, acb\}$!

2.2 Regelbasierte Klassifikatoren

1. Gegeben seien folgende Attribute und Wertebereiche zur Bewertung von Gebrauchtwagen:

- *Klimaanlage* : $\{instand, defekt\}$
- *Motor* : $\{kraftvoll, durchzugsschwach\}$
- *Fahrleistung* : $\{hoch, normal, gering\}$
- *Rost* : $\{ja, nein\}$

Aus einem Datensatz wurde folgender regelbasierte Klassifikator erzeugt:

Fahrleistung = *hoch* \rightarrow *Wert* = *gering*

Fahrleistung = *gering* \rightarrow *Wert* = *hoch*

Klimaanlage = *instand* \wedge *Motor* = *kraftvoll* \rightarrow *Wert* = *hoch*

Klimaanlage = *instand* \wedge *Motor* = *durchzugsschwach* \rightarrow *Wert* = *gering*

Klimaanlage = *defekt* \rightarrow *Wert* = *gering*

- (a) Sind die Regeln gegenseitig ausschließend?
 - (b) Ist die Regelmenge vollständig?
 - (c) Ist es nötig, die Regeln zu ordnen?
 - (d) Ist die Festlegung eine default Klasse nötig?
2. Für ein top-down Regelgenerierungs-Verfahren soll der Wert einer Konjunktion B von

- $r_1 : A \rightarrow C$

zu

- $r_2 : A \wedge B \rightarrow C$

ermittelt werden.

- (a) Im RIPPER Algorithmus entscheidet FOIL's information gain über die beste hinzuzufügende Konjunktion.

Bei der Ermittlung des Wertes von B decke

- r_1 350 positive und 150 negative DO ab und
- r_2 300 positive und 50 negative DO ab.

Wie hoch ist FOIL's information gain von r_2 bezüglich r_1 ?

- (b) Eine vorherige Version von RIPPER, genannt IREP, favorisiert diejenige Regelerweiterung, welche des Wert von

$$v_{IREP} = \frac{p + (N - n)}{P + N}$$

über einem Validierungs-Datensatz um den größten Wert erhöht bzw. belässt es es bei r_1 , falls keine Verfeinerungen diesen Wert erhöht. Hierbei ist

- P die Anzahl der positiven DO im Validierungs-Datensatz,
- N die Anzahl der negativen DO im Validierungs-Datensatz,
- p die Anzahl der positiven DO, die von der Regel abgedeckt werden und
- n die Anzahl der negativen DO, die von der Regel abgedeckt werden.

Ein Validierungs-Datensatz habe 500 positive und 500 negative DO.

- r_1 decke 200 positive und 50 negative DO ab.
- r_2 decke 100 positive und 5 negative DO ab.

Welcher der beiden Regeln würde IREP preferieren?

- (c) Ermitteln Sie v_{RIPPER} für r_1 und r_2 beim post-Pruning! Würde RIPPER beim post-Pruning die Konjunktion B als Verfeinerung akzeptieren?
3. C4.5 extrahiert Regeln indirekt aus einem Entscheidungsbaum. RIPPER hingegen extrahiert die Regeln direkt aus den Daten. Diskutieren Sie die Stärken und Schwächen beider Methoden!
4. Gegeben sei ein Datensatz mit 100 positiven und 400 negativen DO sowie drei Kandidaten für eine zu generierende Regel:

- $r_1: A \rightarrow +$ (deckt 4 positive und 1 negatives DO ab)
- $r_2: B \rightarrow +$ (deckt 30 positive und 10 negative DO ab)
- $r_3: C \rightarrow +$ (deckt 100 positive und 90 negative DO ab)

Ermitteln Sie die beste und die schlechteste Regel bzgl. der Metriken

- (a) Accuracy
- (b) FOIL's Informationsgewinn
- (c) Wahrscheinlichkeitsverhältnis
- (d) Laplace Metrik
- (e) m-estimate Metrik
5. Bild 4 zeigt die Abdeckung positiver und negative DO eines Datensatzes mit 29 positiven und 21 negativen DO durch drei Regeln r_1 , r_2 und r_3 . Ermitteln Sie beste und schlechteste Regel bzgl. der Metriken
- (a) Wahrscheinlichkeitsverhältnis
- (b) Laplace Metrik
- (c) m-estimate Metrik ($p_+ = \frac{29}{29+21} = 0.58$)

Ermitteln Sie die bessere der beiden Regeln r_2 und r_3 auf der Basis der Accuracy, nach dem r_1 bereits in die Regelbasis aufgenommen wurde und

- (d) kein von r_1 abgedecktes DO aus dem Datensatz entfernt wurde.
- (e) nur die positiven von r_1 abgedeckten DO aus dem Datensatz wurden.
- (f) alle von r_1 abgedeckten DO aus dem Datensatz wurden.

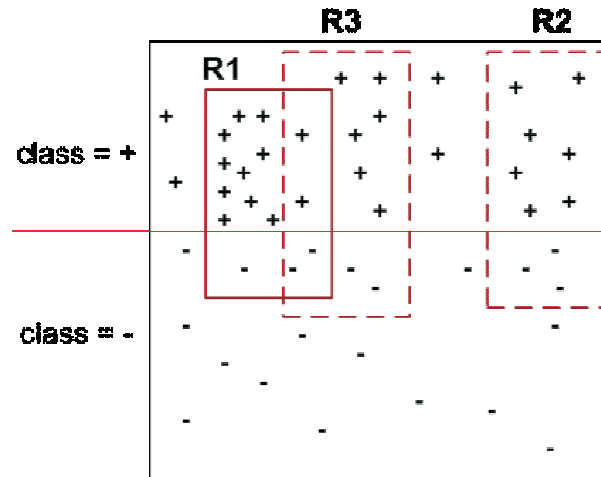


Figure 4: Datensatz und Abdeckung 3er Regeln

2.3 kNN Klassifikation

1. Jemand löst das Problem des Findens von k nächstgelegenen Nachbarn in einem Attribut-Raum durch den folgenden Algorithmus:

Algorithm **kNN** for finding K nearest neighbors.

- 1: **for** $i = 1$ to number of data objects **do**
- 2: Find the distances of the i -th object to all other objects.
- 3: Sort these distances in decreasing order.
(Keep track of which object is associated with each distance.)
- 4: **return** objects associated with the first K distances of the sorted list
- 5: **end for**

- (a) Beschreiben Sie eventuelle Probleme mit diesem Algorithmus, falls es mehrfache Objekte (verschiedene Objekte mit identischen Attribut-Werten) gibt! Sie dürfen dabei davon ausgehen, dass die distance - Funktion nur für identische Objekte den Wert Null liefert.
 - (b) Wie würden Sie o.g. Problem lösen?
2. Man betrachte den 1-dimensionalen Datensatz in Tabelle 10.

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	-	-	+	+	+	-	-	+	-	-

Table 10: Datensatz für Aufgabe 2

- (a) Klassifizieren Sie den Punkt $x' = 5.0$ anhand seiner 1, 3, 5 und 9 nächsten Nachbarn unter Anwendung des Mehrheitsprinzips!
- (b) Wiederholen Sie o.g. Klassifizierung unter Anwendung des gewichteten Abstandsmaßes

$$y' = \operatorname{argmax}_v \sum_{[x_i, y_i] \in kNN(x')} \frac{1}{d(x', x_i)^2} * \operatorname{ord}(v = y_i)$$

$$\text{mit } \text{ord}(\text{prop}) = \begin{cases} 1 & , \text{ falls } \text{prop} = \text{true} \\ 0 & , \text{ falls } \text{prop} = \text{false} \end{cases} .$$

3. Der kNN-Algorithmus kann erweitert werden, um auch nominale Attribute behandeln zu können. Eine Variante des Algorithmus' PEBLS (Parallel Exemplar-Based Learning System) ermittelt die Distanz bei nominalen Attributen durch eine so genannte Modified Value Difference Metric (MVDM). Für ein Paar nominaler Attribut-Werte v_1 und v_2 ist die MVDM-Distanz wie folgt definiert:

$$d(v_1, v_2) = \sum_{i=1}^k \left| \frac{n_{i1}}{n_1} - \frac{n_{i2}}{n_2} \right|$$

wobei

- n_{ij} die Anzahl von DO aus Klasse i mit dem Attribut-Wert v_j ist und
- n_j Gesamtzahl von DO mit dem Attribut-Wert v_j ist.

<i>home_owner</i>	<i>marital_status</i>	<i>annual_income</i>	<i>defaulted_borrower</i>
<i>yes</i>	<i>single</i>	124	<i>no</i>
<i>no</i>	<i>married</i>	100	<i>no</i>
<i>no</i>	<i>single</i>	70	<i>no</i>
<i>yes</i>	<i>married</i>	120	<i>no</i>
<i>no</i>	<i>divorced</i>	95	<i>yes</i>
<i>no</i>	<i>married</i>	60	<i>no</i>
<i>yes</i>	<i>divorced</i>	220	<i>no</i>
<i>no</i>	<i>single</i>	85	<i>yes</i>
<i>no</i>	<i>married</i>	75	<i>no</i>
<i>no</i>	<i>single</i>	90	<i>yes</i>

Table 11: Data Set for exercise 3

Die Tabelle 11 zeigt Trainingsdaten zur Beurteilung der Kreditwürdigkeit von Kunden einer Bank. Ermitteln Sie die MVDM-Distanz aller Paare von Werten für die nominalen Attribute *home_owner* und *marital_status*!

2.4 Bayes'sche Klassifikation

1. Der Anteil der Raucher unter nicht graduierten Studierenden betrage 15 %; der Anteil der Raucher unter graduierten Studierenden betrage 23 %. $\frac{1}{5}$ aller Studierenden seien graduiert, alle anderen nicht graduiert.
 - (a) Wie hoch ist die Wahrscheinlichkeit, dass ein Student, welcher raucht, graduiert ist?
 - (b) Ist ein zufällig ausgewählter Student wahrscheinlicher graduiert oder nicht graduiert?
 - (c) Ist ein zufällig ausgewählter Student, welcher Raucher ist, wahrscheinlicher graduiert oder nicht graduiert?

- (d) 30 % aller graduierten Studierenden leben im Wohnheim; bei den nicht graduierten Studierenden beträgt dieser Anteil nur 10 %.

Ist ein zufällig ausgewählter Student, welcher raucht und im Wohnheim wohnt, wahrscheinlicher graduiert oder nicht graduiert? Es darf unterstellt werden, dass die Attribute “*im Wohnheim leben*” und “*rauchen*” unabhängig voneinander sind.

2. Man betrachte den Datensatz in Tabelle 12.

lfd. Nr.	A	B	C	$Klasse$
1	0	0	0	+
2	0	0	1	-
3	0	1	1	-
4	0	1	1	-
5	0	0	1	+
6	1	0	1	+
7	1	0	1	-
8	1	0	1	-
9	1	1	1	+
10	1	0	1	+

Table 12: Datensatz zu Aufgabe 2

- (a) Ermitteln Sie die geschätzten bedingten Wahrscheinlichkeiten $p(A|+)$, $p(B|+)$, $p(C|+)$, $p(A|-)$, $p(B|-)$ und $p(C|-)$ als relative Häufigkeit im Datensatz!
- (b) Ermitteln Sie die wahrscheinlichste Klassenzugehörigkeit des DO $A = 0$, $B = 1$ und $C = 0$ anhand der Bayes'schen Formel aus den in Aufgabe (a) ermittelten Werten!
- (c) Ermitteln Sie die bedingten Wahrscheinlichkeiten anhand des m-estimate Ansatzes mit $p' = 1/2$ and $m = 4$.
- (d) Wiederholen Sie Aufgabe (b) unter Nutzung der m-estimate Schätzungen aus Aufgabe (c) für die bedingten Wahrscheinlichkeiten.
- (e) Vergleichen Sie beider Methoden der Schätzung der bedingten Wahrscheinlichkeiten! Unter welchen Umständen ist welche besser?

3. Man betrachte den Datensatz in Tabelle 13.

- (a) Ermitteln Sie die geschätzten bedingten Wahrscheinlichkeiten $p(A = 1|+)$, $p(B = 1|+)$, $p(C = 1|+)$, $p(A = 1|-)$, $p(B = 1|-)$ und $p(C = 1|-)$ wie in Aufgabe 2 (a)!
- (b) Ermitteln Sie die wahrscheinlichste Klassenzugehörigkeit des DO ($A = 1, B = 1, C = 1$) anhand der Bayes'schen Formel aus den in Aufgabe (a) ermittelten Werten!
- (c) Vergleichen Sie $p(A = 1)$, $p(B = 1)$ und $p((A = 1) \wedge (B = 1))$. Welche Aussage über die Relation zwischen A und B kann man daraus ableiten?
- (d) Wiederholen Sie die Analyse wie in Aufgabe (c) für $p(A = 1)$, $p(B = 0)$ und $p((A = 1) \wedge (B = 0))$!
- (e) Vergleichen Sie $p((A = 1) \wedge (B = 1)|+)$ mit $p(A = 1|+)$ und $p(B = 1|+)$! Sind die Variablen innerhalb der Klasse voneinander unabhängig?

lfd. Nr.	A	B	C	Class
1	0	0	1	-
2	1	0	1	+
3	0	1	0	-
4	1	0	0	-
5	1	0	1	+
6	0	0	1	+
7	1	1	0	-
8	0	0	0	-
9	0	1	0	+
10	1	1	1	+

Table 13: Datensatz zu Aufgabe 3

4. Abbildung 5 zeigt das Bayes'sche Belief Network für den Datensatz aus Tabelle 14. Alle Attribute sind binär.

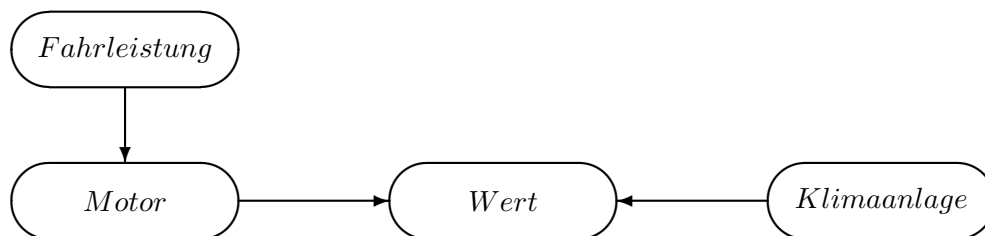


Figure 5: Bayes'sches Belief Network zu Aufgabe 4

<i>Fahrleistung</i>	<i>Motor</i>	<i>Klimaanlage</i>	Anzahl von DO mit	
			<i>Wert = hoch</i>	<i>Wert = gering</i>
<i>hoch</i>	<i>gut</i>	<i>instand</i>	3	4
<i>hoch</i>	<i>gut</i>	<i>defekt</i>	1	2
<i>hoch</i>	<i>schlecht</i>	<i>instand</i>	1	5
<i>hoch</i>	<i>schlecht</i>	<i>defekt</i>	0	4
<i>gering</i>	<i>gut</i>	<i>instand</i>	9	0
<i>gering</i>	<i>gut</i>	<i>defekt</i>	5	1
<i>gering</i>	<i>schlecht</i>	<i>instand</i>	1	2
<i>gering</i>	<i>schlecht</i>	<i>defekt</i>	0	2

Table 14: Datensatz zu Aufgabe 4

- (a) Entwickeln Sie die Wahrscheinlichkeitstabellen für jeden Knoten des Netzes!
 (b) Ermitteln Sie $p(\text{Motor} = \text{schlecht} \wedge \text{Klimaanlage} = \text{defekt})$ anhand dieses Netzes!
5. Man berechne anhand des in Abbildung 6 gezeigten Bayes'schen Netzes folgende Wahrscheinlichkeiten:
- (a) $p(B = \text{good} \wedge F = \text{empty} \wedge G = \text{empty} \wedge S = \text{yes})$
 (b) $p(B = \text{bad} \wedge F = \text{empty} \wedge G = \text{not_empty} \wedge S = \text{no})$

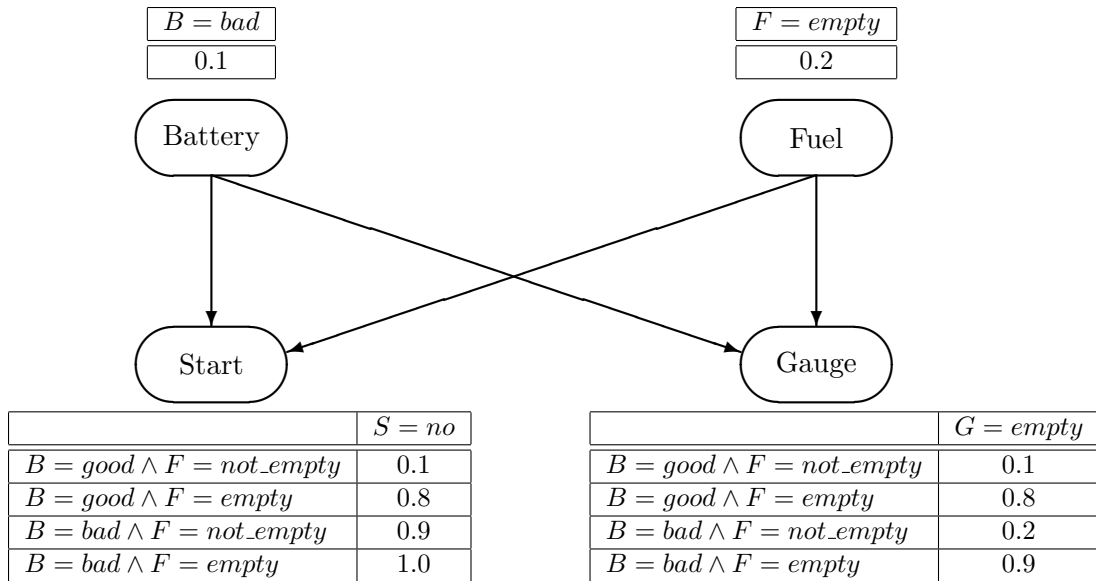


Figure 6: Bayes'sches Belief Network zu Aufgabe 5

- (c) Berechnen Sie die Wahrscheinlichkeit für $S = yes$ unter der Bedingung $B = bad$.
6. Evaluieren Sie die Performance 2er Modelle (Klassifikatoren) M_1 und M_2 . Der Trainings-Datensatz enthält 26 binäre Attribute A bis Z .

Nr.	tatsächl. Klasse	$p(+ A, B, \dots, Z, M_1)$	$p(+ A, B, \dots, Z, M_2)$
1	+	0.73	0.61
2	+	0.69	0.03
3	-	0.44	0.68
4	-	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09
7	-	0.08	0.38
8	-	0.15	0.05
9	+	0.45	0.01
10	-	0.35	0.04

Table 15: A posteriori Wahrscheinlichkeiten

Tabelle 15 enthält die a posteriori Wahrscheinlichkeiten für die (interessierende) positive Klasse, welche durch Anwendung des Modells auf einen Test-Datensatz ermittelt wurden. Da es sich um binäre Klassifikation handelt, ist $p(-) = 1 - p(+)$ und $p(-|A, \dots, Z) = 1 - p(+|A, \dots, Z)$. Das Ziel ist, positive DO zu erkennen.

- (a) Zeichnen Sie die ROC-Kurve für M_1 und M_2 in ein und dasselbe) Diagramm! Welches Modell halten Sie für besser? Begründen Sie Ihre Antwort!
- (b) Der sog. cutoff-Schwellwert t für das Modell M_1 sei 0.5, d.h. DO mit einer a posteriori-Wahrscheinlichkeit $p > 0.5$ werden als positive DO klassifiziert. Ermitteln Sie Precision, Recall und die F_1 -Metrik für M_1 mit $t = 0.5$!

- (c) Wiederholen Sie diese Analyse für M_2 mit demselben cutoff-Schwellwert $t = 0.5$! Vergleichen Sie die Resultate basierend auf den ermittelten F_1 -Metriken für beide Modelle! Welches Modell halten Sie für besser? Deckt sich diese Einschätzung mit der basierend auf der ROC-Kurve?
- (d) Wiederholen Sie die Analyse für M_1 mit $t = 0.1$! Welchen Schwellwert halten Sie für besser, $t = 0.5$ oder $t = 0.1$? Deckt sich Ihre Einschätzung mit der basierend auf der ROC-Kurve?

7. Gegeben sei folgender Trainingsdatensatz:

X	Y	Anzahl DO	
		+	-
0	0	0	100
1	0	0	0
2	0	0	100
0	1	10	100
1	1	10	0
2	1	10	100
0	2	0	100
1	2	0	0
2	2	0	100

- (a) Man konstruiere einen Entscheidungsbaum für u.g. Trainingsdaten. Das Auswahlkriterium für Split-Attribue sei der Klassifikationsfehler. Ist das Modell 100 % korrekt über den Trainingsdaten? Wie hoch ist sein Klassifikationsfehler?
- (b) Bestimmen Sie die Accuracy, Precision (F_0), Recall (F_∞), und die F_1 Metrik des Modells bzgl. der seltenen Klasse ”+”!
- (c) Man konstruiere einen neuen Entscheidungsbaum mit der folgenden Kostenfunktion : $C(i, j) = \begin{cases} 0, & \text{if } i = j \\ 1, & \text{if } i = +, j = - \\ \frac{\text{AnzahlDOderKlasse-}}{\text{AnzahlDOderKlasse+}}, & \text{if } i = -, j = + \end{cases}$
- (d) Bestimmen Sie die Accuracy, Precision (F_0), Recall (F_∞), und die F_1 Metrik des neuen Baumes bzgl. der seltenen Klasse ”+”!

3 Assoziations-Analyse

1. Man betrachte die Daten in Tabelle 16.

- (a) Ermitteln Sie den Support der Itemsets $\{e\}$, $\{b, d\}$, und $\{b, d, e\}$ unter Betrachtung der Transaktions-Nr. als je einen Warenkorb!
- (b) Ermitteln Sie die Confidence der Regeln $\{b, d\} \rightarrow \{e\}$ and $\{e\} \rightarrow \{b, d\}$ anhand des Ergebnisses aus Aufgabe (a)! Ist die Confidence eine symmetrische Metrik?
- (c) Wiederholen Sie (a) unter Betrachtung der Kunden-Nr. als je einen Warenkorb! Jedes Item sollte als binäre Variable geführt werden, welche den Wert 1 hat, falls der Kunde das Produkt gekauft hat und anderenfalls den Wert 0 hat.

Kunden-Nr. ID	Transaktions-Nr.	gekaufte Waren
1	0001	{a, d, e}
1	0024	{a, b, c, e}
2	0012	{a, b, d, e}
2	0031	{a, c, d, e}
3	0015	{b, c, e}
3	0022	{b, d, e}
4	0029	{c, d}
4	0040	{a, b, c}
5	0033	{a, d, e}
5	0038	{a, b, e}

Table 16: Warenkorb - Transaktionen zu Aufgabe 1

(d) Ermitteln Sie die Confidence der Regeln $\{b, d\} \rightarrow \{e\}$ and $\{e\} \rightarrow \{b, d\}$ anhand des Ergebnisses aus Aufgabe (c)!

(e) Seien

- s_1 und c_1 der Support bzw. die Confidence einer Regel r unter Betrachtung der Transaktions-Nr. als je einen Warenkorb und
- s_2 und c_2 der Support bzw. die Confidence einer Regel r unter Betrachtung der Kunden-Nr. als je einen Warenkorb.

Diskutieren Sie, ob es evtl. einen Zusammenhang zwischen s_1 und s_2 oder c_1 und c_2 !

2. (a) Wie hoch sind die Confidence-Werte von Regeln der Art $\emptyset \rightarrow A$ und $A \rightarrow \emptyset$?

(b) Inwieweit hat Confidence Monotonie-Eigenschaften?

Seien c_1 , c_2 , and c_3 die Confidence-Werte der Regeln

- $r_1 : \{p\} \rightarrow \{q\}$
- $r_2 : \{p\} \rightarrow \{q, r\}$
- $r_3 : \{p, r\} \rightarrow \{q\}$

c_1 , c_2 und c_3 mögen verschiedene Werte haben. Welche Relationen gibt es zwischen c_1 , c_2 und c_3 ? Welche Regel hat den geringsten Confidence-Wert?

(c) Wiederholen Sie die Analyse wie in (b) unter der Annahme, dass alle drei Regeln den gleichen Support haben! Welche Regel hat dann den höchsten Confidence-Wert?

(d) Ist Confidence transitiv?

Mögen die Confidence-Werte der Regeln $A \rightarrow B$ und $B \rightarrow C$ einen Schwellwert $minconf$ überschreiten. Ist es denkbar, dass $A \rightarrow C$ einen Confidence-Wert hat, der niedriger als $minconf$ ist?

3. Man betrachte die Warenkorb-Transaktionen in Tabelle 17.

(a) Wie viele Assoziations-Regeln (inklusive solcher mit einem Support von Null) kann man maximal aus diesen Daten Generieren?

(b) Wie groß kann ein häufiges Item-Set (unter der Bedingung, dass $minsup > 0$) maximal sein?

Transaktions-Nr.	gekaufte Waren
1	{ <i>Milch, Bier, Windeln</i> }
2	{ <i>Brot, Butter, Milch</i> }
3	{ <i>Milch, Windeln, Kekse</i> }
4	{ <i>Brot, Butter, Kekse</i> }
5	{ <i>Bier, Kekse, Windeln</i> }
6	{ <i>Milch, Windeln, Brot, Butter</i> }
7	{ <i>Brot, Butter, Windeln</i> }
8	{ <i>Bier, Windeln</i> }
9	{ <i>Milch, Windeln, Brot, Butter</i> }
10	{ <i>Bier, Kekse</i> }

Table 17: Warenkorb-Transaktionen zu Aufgabe 3

- (c) Wie viele Itemsets aus drei Items lassen sich aus diesen Daten bilden? Geben Sie eine allgemeine Formel zur Bildung von Itemsets der Größe k bei n ($n \geq k$) Items an!
- (d) Ermitteln Sie das Itemset (der Größe 2 oder größer) mit dem höchsten Support-Wert!
- (e) Ermitteln Sie ein Paar von Items a und b , für welches die Regeln $a \rightarrow b$ und $b \rightarrow a$ den selben Confidence-Wert haben!
4. Man betrachte die folgenden 3-Itemsets, welche aus den häufigen Items $F_1 = \{a, b, c, d, e\}$ entstanden sein mögen:

$$F_3 = \{\{a, b, c\}, \{a, b, d\}, \{a, b, e\}, \{a, c, d\}, \{a, c, e\}, \{b, c, d\}, \{b, c, e\}, \{c, d, e\}\}.$$

- (a) Geben Sie alle 4-Itemsets Kandidaten an, die durch die $F_{k-1} \times F_1$ Verschmelzung entstehen!
- (b) Geben Sie alle 4-Itemsets Kandidaten an, die durch den *a Priori* Algorithmus ($F_{k-1} \times F_{k-1}$ Verschmelzung) entstehen!
- (c) Geben Sie alle 4-Itemsets Kandidaten an, welche nach dem Pruning Schritt im *a Priori* Algorithmus übrig bleiben!
5. Der *a priori* Algorithmus verwendet eine “generate-and-test” Strategie zur Ermittlung häufiger Itemsets. Itemset-Kandidaten der Größe k werden durch Verschmelzung von Paaren häufiger Itemsets der Größe $k - 1$ generiert.

Ein solcher Kandidat wird in einem zusätzlichen Pruning - Schritt (außer dem danach vollzogenen Support - basierten Pruning) verworfen, falls sich einer seiner Untermengen als “nicht häufig” erweist.

Der *a priori* Algorithmus möge auf die in Tabelle 18 mit einem Mindest-Support von $minsup = 0.3$ angewandt werden.

- (a) Skizzieren Sie die Halbordnung \subseteq zwischen allen bis zu 4-elementigen Itemsets $\{a, b, c, d, e\}^*$ aus $\{a, b, c, d, e\}$ in Form eines Verbands! Markieren (oder färben) Sie die Knoten im Verband wie folgt:

Transaktions-Nr.	gekaufte Waren
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

Table 18: Warenkorb-Transaktionen zu Aufgabe 5

N : ..., falls das Itemset am a priori Algorithmus nicht als Kandidat-Itemset betrachtet wird.

Dies kann einer der folgenden Gründe haben:

- i. Es ist nicht generiert worden bei der Kandidaten-Generierung oder
- ii. es ist zwar generiert worden, ist aber beim Kandidaten-Pruning wieder entfernt wurde, weil es $k - 1$ - Teilmengen gab, welche nicht sich nicht als häufig erwiesen.

F : ..., falls das Itemset als häufig (frequent) identifiziert wird.

I : ..., falls das Itemset wegen nicht hinreichendem Support als nicht häufig (infrequent) identifiziert wird.

(b) Wie hoch ist der Anteil häufiger Itemsets im Verband?

(c) Wie hoch ist die Pruning-Rate des *a priori* Algorithmus bei diesem Datensatz?

Die Pruning-Rate ist definiert als Anteil der Itemsets, welche nicht als Kandidat betrachtet werden. Hierbei spielt es keine Rolle, ob sie (1) nicht als Kandidat nominiert wurden oder (2) beim Kandidaten-Pruning (der Teilmengen-Überprüfung) "durchgefallen" sind.

(d) Wie hoch ist die Rate so genannter "falscher Alarme", d.h. der Anteil der Kandidaten im Verband, welche mangels hinreichenden Supports als nicht häufig eingestuft wurden?

6. Zur effektiven Ermittlung des Support-Wertes eines Kandidaten-Itemsets nutzt der *a priori* Algorithmus einen Hash-Baum. Man betrachte den Hash-Baum für 3-Itemsets in Abbildung 7.

(a) Welche der Blattknoten des Hash-Baumes würden zur Suche von Kandidaten-Itemsets für die Transaktion $t = \{1, 3, 4, 5, 8\}$ aufgesucht werden?

(b) Welche der in den oben ermittelten Blattknoten befindlichen Kandidaten-Itemsets sind tatsächlich in t enthalten und würden zur Erhöhung des Support beitragen?

7. Man betrachte die folgenden 3-Itemsets:

$\{1, 2, 3\}$, $\{1, 2, 6\}$, $\{1, 3, 4\}$, $\{2, 3, 4\}$, $\{2, 4, 5\}$, $\{3, 4, 6\}$, $\{4, 5, 6\}$

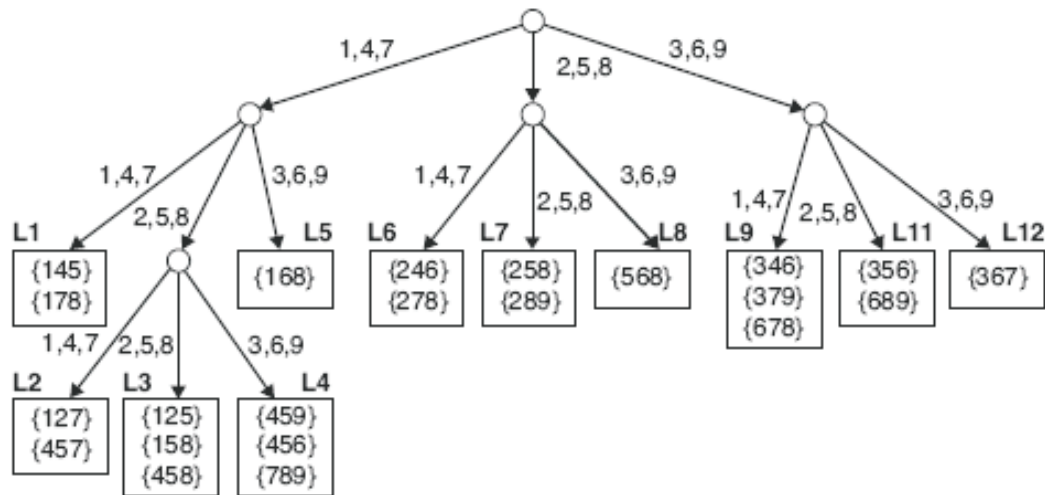


Figure 7: Hash-Baum zu Aufgabe 6

- (a) Konstruieren Sie einen binären Hash-Baum für die o.g. Kandidaten - k -Itemsets! Die Hash-Funktion soll alle ungeraden Items dem linken Unterbaum und alle geraden items dem rechten Unterbaum zuordnen! Ein Kandidaten 3-Itemset wird durch Berechnung der Hash-Funktion für das jeweils nächste Item in den Baum aufgenommen. Ist dabei ein Blatt erreicht, wird das Itemset nach folgenden Regeln in den Baum aufgenommen:
- **Regel 1:** Wenn die Tiefe des Blattknotens gleich k ist (die Wurzel habe die Tiefe 0), wird das Itemset unabhängig von der Anzahl der bereits im Blattknoten enthaltenen Itemsets in diesen Blattknoten aufgenommen.
 - **Regel 2:** Wenn die Tiefe des Blattknotens kleiner als k ist, wird das Itemset in den Blattknoten aufgenommen, falls die Anzahl der dort bereits abgelegten Itemsets kleiner als $maxsize$ ist. Für diesen Baum sein $maxsize = 2$.
 - **Regel 3:** Wenn die Tiefe des Blattknotens kleiner als k ist und die Anzahl der dort bereits abgelegten Itemsets bereits $maxsize$ ist, wird das Blatt durch einen Unterbaum ersetzt. Sowohl vorher im Blatt befindlichen Itemsets als auch das neue Itemset werden entsprechend ihrer Hash-Funktion in den linken oder rechten Unterbaum eingefügt.
- (b) Wie viele Blattknoten und wieviele nicht-Blattknoten hat der so entstehende Hash-Baum?
- (c) Gegeben sei eine Transaktion $t = \{1, 2, 3, 5, 6\}$. Welche Blattknoten des Hash-Baumes werden von dieser Transaktion aufgesucht? Welches sind die Kandidat - 3-Itemsets, die im Hash-Baum gefunden werden?
8. Man markiere die Knoten des in Abbildung 8 gezeigten Itemset - Verbands unter Betrachtung der Transaktionen in Tabelle 19
- mit M , (maximal) falls es sich um ein maximales häufiges Itemset handelt,
 - mit C , (closed) falls es sich um ein geschlossenes häufiges Itemset handelt,

TID	Waren
1	{a, b, d, e}
2	{b, c, d}
3	{a, b, d, e}
4	{a, c, d, e}
5	{b, c, d, e}
6	{b, d, e}
7	{c, d}
8	{a, b, c}
9	{a, d, e}
10	{b, d}

Table 19: Transaktionsdaten zu Aufgabe 8

- mit F , (frequent) falls es sich um ein häufiges Itemset handelt, welches weder maximal noch geschlossen ist und
- mit I , (infrequent) falls es sich nicht um ein häufiges Itemset handelt.

Der mindest-Support sei $minsup = 0.3$

9. Gegeben seien die Transaktionen in Tabelle 20.

- Zeichnen Sie den Verband der aus dem häufigen Itemset $\{a, c, e, g\}$ bildbaren Assoziationsregeln! Verbinden Sie die Knoten im Verband nur dann durch eine gerichtete Kante, wenn die beiden Startknoten der hineinführenden Kanten beim redundanzfreien Verschmelzen diejenigen Regeln repräsentieren, mit denen die Verschmelzung vorgenommen wird!
- Markieren Sie jeden Knoten des Verbands mit der Konfidenz der jeweiligen Regel!
- Die Mindestkonfidenz sein $minconf = 0.75$. Markieren Sie jeden Knoten
 - mit C (confident), falls dieser Knoten bei der redundanzfreien Auswahl von Regelpaaren durch deren Verschmelzung gebildet wird und für weitere Verschmelzungen der nächsten Ebene herangezogen werden kann.
 - mit I (inconfident), falls dieser Knoten bei der redundanzfreien Auswahl von Regelpaaren durch deren Verschmelzung gebildet wird, aber mangels Konfidenz nicht für weitere Verschmelzungen der nächsten Ebene herangezogen wird.
 - mit N (not formed), falls dieser Knoten bei der redundanzfreien Auswahl von Regelpaaren nicht gebildet wird.

10. Zur Evaluierung von Assoziations-Patterns wird häufig der Support und die Konfidenz verwendet, um uninteressante Regeln zu prunen.

- Geben Sie die Kontingenz-Tabelle für nachfolgende Regeln unter Verwendung der Transaktionsdaten in Tabelle 21 an:
 - $b \rightarrow c$
 - $a \rightarrow d$
 - $b \rightarrow d$

TID	Waren
1	{a, b, c, e, g}
2	{a, b, e, g}
3	{a, c, e, f, g}
4	{c, d, e, g}
5	{a, c, d, e, g}
6	{c, e, f, g}
7	{a, b, c, d, e, g}
8	{a, b, c, e, f}
9	{a, c, d, e, f, g}
10	{a, b, c, f, g}
11	{a, c, d, e, g}
12	{a, b, e, f}
13	{a, b, c, e, f, g}
14	{c, d, f, g}
15	{a, c, d, e, f, g}
16	{b, c, d, e, f}
17	{a, b, c, e, g}
18	{b, d, e, g}
19	{a, c, d, e, g}
20	{b, c, e, f}

Table 20: Transaktionsdaten zur Aufgabe 9

TID	Waren
1	{a, b, d, e}
2	{b, c, d}
3	{a, b, d, e}
4	{a, c, d, e}
5	{b, c, d, e}
6	{b, d, e}
7	{c, d}
8	{a, b, c}
9	{a, d, e}
10	{b, d}

Table 21: Transaktionsdaten zur Aufgabe 10

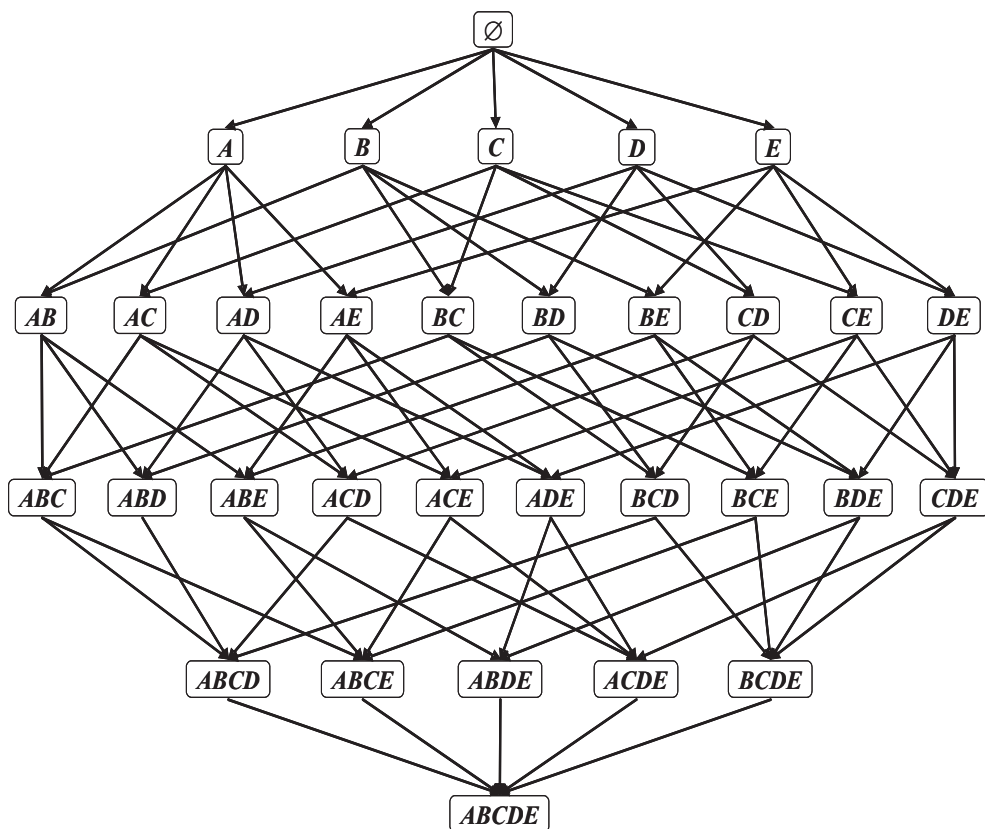


Figure 8: Itemset Verband zu Aufgabe 8

- iv. $e \rightarrow c$
 - v. $c \rightarrow a$
- (b) Verwenden Sie die Kontingents-Tabellen aus Aufgabe (a), um die folgenden “Interessantheits-Maße” für diese Regeln zu ermitteln! Stellen Sie ein Ranking der Regeln bezüglich dieser Maße auf!
- i. Support
 - ii. Konfidenz
 - iii. Interest Faktor
 - iv. Korrelation (ϕ - Koeffizient)
 - v. Interest Support (IS)
 - vi. Odds Ratio (α - Koeffizient)
- (c) Ermitteln Sie die Korrelation zwischen dem Ranking bezüglich der Konfidenz und dem der anderen fünf Maßen!
- Welches dieser Maße korreliert am besten mit den Konfidenz?
11. Gegeben seien Transaktionsdaten von 100 Transaktionen mit 20 Items. Der Support für Item a sei 0.25, der Support für Item b sei 0.9 und der Support für das Itemset $\{a, b\}$ sei 0.2. Der Midnestsupport sei $minsup = 0.1$ und die Mindestkonfidenz sei $minconf = 0.6$.

		A		
		1	0	
$C = 0$	B	1	0	15
		0	15	30
$C = 1$	B	1	5	0
		0	0	15

Table 22: Kontingenz-Tabelle zu Aufgabe 12

- (a) Ermitteln Sie die Konfidenz der Assoziations-Regel $\{a\} \rightarrow \{b\}$! Ist diese Regel interessant bezüglich der geforderten Mindestkonfidenz?
- (b) Ermitteln Sie das Lift-Maß des Assoziations-Patterns $\{a, b\}$. In welcher Weise sind a und b voneinander abhängig?
- (c) Welche Schlussfolgerung sollte man aus den Ergebnissen der Aufgaben (a) und (b) ziehen?
12. Tabelle 22 ist eine $2 \times 2 \times 2$ Kontingenz-Tabelle für die binären Variablen A und B bei verschiedenen Werten einer Steuer-Variablen C .
- (a) Ermitteln Sie den φ -Koeffizienten von A und B für $C = 0$, $C = 1$, und $C = 0$ oder 1 .
- (b) Was sollte man aus diesem Beispiel schlussfolgern?
13. Man betrachte das Verhältnis von Käufern von Fernsehern der Sorte High Definition Television (HDTV) und Fitness-Geräten (FG) in den Tabellen 23 und 24.

HDTV Käufer	FG Käufer		
	ja	nein	
ja	99	81	180
nein	54	66	120
	153	147	300

Table 23: Verhältnis von Käufern von HDTV und FG zu Aufgabe 13

Kunden-gruppe	HDTV Käufer	FG Käufer		total
		ja	nein	
Studenten	ja	1	9	10
	nein	4	30	34
Berufstätige	ja	98	72	170
	nein	50	36	86

Table 24: Verhältnis von Käufern von HDTV und FG zu Aufgabe 13

- (a) Ermitteln Sie die Odds Ratio Maße (die α Koeffizienten) für beide Tabellen.
- (b) Ermitteln Sie die Korrelationen (die φ Koeffizienten) für beide Tabellen.
- (c) Ermitteln Sie die Lift - Maße für beide Tabellen.

Beschreiben Sie für jedes der o.g. Maße, wie sich die Assoziationen durch die gemeinsame Betrachtung aller Kunden gegenüber den separaten Betrachtung von Studenten und Berufstätigen qualitativ ändert.

4 Cluster-Analyse

1. Gegeben sei ein Datensatz mit m Datenobjekten (DO) und K Clustern.
 - Die Hälfte der DO und Cluster befinde sich in einer “dichteren Region”.
 - Die (andere) Hälfte der DO und Cluster befinde sich in einer “weniger dichten” Region.
 - Bei Regionen sind wohl separiert voneinander.

Wie sollten die Centroiden angeordnet sein, um K Cluster mit minimalem “sum of the squared error” (SSE) zu finden:

- (a) Es sollten in jeder der Regionen gleich viele Centroiden angeordnet werden.
 - (b) Es sollten mehr Centroiden in der weniger dichten Region angeordnet werden.
 - (c) Es sollten mehr Centroiden in der dichteren Region angeordnet werden.
2. Man betrachte den Centroiden eines Clusters mit DO eines binären Transaktions-Datensatzes. Ein DO repräsentiert eine Transaktion.
 - (a) Welchen minimal- und Maximalwert kann eine Komponente (Attribut) des Centroiden haben?
 - (b) Was ist die Interpretation der Komponenten des Centroiden?
 - (c) Welche der Komponenten charakterisieren die DO (Transaktionen) am besten?
 3. Man konstruiere ein Beispiel mit drei natürlichen Clustern, die durch K-means (fast immer) gefunden werden, aber durch biscting K-means nicht.
 4. Der totale SSE (sum of squared error) ist die Summer der SSEs über alle Attribute.
 - (a) Was bedeutet es, wenn der SSE für ein Attribut in allen Clustern klein ist?
 - (b) Was bedeutet es, wenn der SSE für ein Attribut in einem Cluster klein ist, in allen anderen Clustern jedoch nicht?
 - (c) Was bedeutet es, wenn der SSE für ein Attribut in allen Clustern groß ist?
 - (d) Was bedeutet es, wenn der SSE für ein Attribut in einem Cluster groß ist, in allen anderen jedoch nicht?
 - (e) Wie kann man diese Information über die Attribute nutzen, um das Clustering zu verbessern?

5. Eine Abart des sogenannten “Leader” Algorithmus sei wie folgt definiert:

Jedes Cluster sei durch ein DO (seinen Leader) definiert. Jedes andere DO wird dem Cluster des nächstgelegenen Leaders zugeordnet, es sei denn, der Abstand zu selbigen überschreitet einen (nutzerdefinierten) Schwellwert. In diesem Fall wird ein solches DO zum Leader eines neuen Clusters erklärt.

Man diskutiere Vor- und Nachteile dieses Algorithmus gegenüber K-means!

6. Eine Voronoi-Diagramm für K Punkte (Centroiden) in einer Ebene ist eine Partitionierung aller Punkte dieser Ebene in K Regionen, bei welcher jeder Punkt der Ebene derjenigen Region zugeordnet ist, zu dessen Centroiden er den geringsten Abstand hat (siehe Abbildung 9).

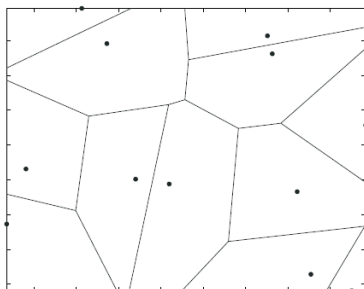


Figure 9: Voronoi Diagram zu Aufgabe 6

- (a) In welchem Zusammenhang stehen derartige Voronoi-Diagramme und K-means Cluster?
- (b) Was kann man aus den Voronoi-Diagrammen bzgl. der Form der entstehenden K-means Cluster ableiten?