Aufgabe 1

1

Nach Fayyad (KDD) und bei CRISP-DM hat man beide Male das grundsätzliche Vorgehen, dass zuerst die Daten vorverarbeitet werden, dann das eigentliche Data-Mining stattfindet und zum Schluss die Ergebnisse beurteilt werden. CRISP-DM weißt dabei, aber untereinander, iterative Schritte auf, was dazu führen könnte, das der Prozess zwar länger dauert, dafür könnte aber ein besseres Ergebnis möglich sein. Ebenfalls beginnt CRISP-DM mit dem Business-Understanding-Schritt, welcher in der realen Anwendung vermutlich ein notwendiger, erster Schritt ist um zum Schluss nützliche Ergebnisse zu erhalten.

2

Bei der Subgruppenentdeckung geht es darum 'interessante' Gruppen gegeben einer Zielvariable zu finden. Weißt zum Beispiel eine gewissen Eigenschaft stärker darauf hin, dass die Zielvariable erfüllt ist, als es im Mittel der Fall ist.

3

Die Subgruppenentdeckung fällt in die Bereiche Modeling und Evaluation. Beim Modeling findet das wirklich Mining statt, welches hier das Herausfinden der Subgruppen ist. Bei der Evaluation werden diese dann bewertet. Die Subgruppenentdeckung kann zwar selbst ihre Ergebnisse nicht beurteilen, durch die gegebene Qualitätsfunktion ist allerdings eine gewisse Art von Evaluation bereits Teil ds Modelings.

4

Bei der Subgruppenentdeckung geht es nur darum, dass die Zielvariable erfüllt ist und die Fälle in denen diese nicht erfüllt sind, haben keinen starken Einfluss. Ebenso geht es dort auch nicht darum, ob die Subgruppe immer zusammen mit der Zielvariable auftaucht. Es geht viel mehr darum zu gucken, ob falls die Zielvariable bereits erfüllt ist, ob sich dann dort gewissen Gruppen finden lassen, bei denen die Zielvariable häufig zutrifft. Bei Assoziationsregeln geht es darum eine Implikation zu finden und damit mehr oder weniger in der Lage zu sein, gewisse Dinge vorher zu sagen. Bei den Subgruppen ist dies andersherum. Dort wird von der Zielvariable darauf geschlossen, welche Subgruppen diese erfüllen. Somit lässt sich mit Assoziationsregeln darauf schließen, was eintritt, gegeben der Vorbedingungen und bei Subgruppen wird darauf geschlossen, welche Vorbedingungen vermutlich notwendig sind, gegeben ein Ziel.

5

Häufig möchte man die top-k Subgruppen erhalten. Bei heuristischen Verfahren wird nicht garantiert, dass die Resultate auch wirklich die definitiv besten Gruppen sind. Bei vollständigen Verfahren allerdings schon. Dafür ist ein heuristisches Verfahren meist aber

deutlich schneller und man erhält mit diesen in den meisten Fällen immer noch gute Ergebnisse. Möchte man beispielsweise alle Ergebnisse erhalten oder unbedingt die korrekten Ergebnisse, so finden die heuristischen Verfahren keine Anwendung. Ist aber beispielsweise nur eine kleine Menge an Subgruppen oder ein schnelles Ergebnis erwünscht, so sollten heuristische Verfahren verwendet werden.

6

Um N zu bestimmen, wählt man einfach alle Zeilen aus und zählt diese (COUNT(*)). Um n zu bestimmen wählt man dann die Zeilen aus, in denen die Zielvariable den gewünschten Wert enthält und zählt diese `(SELECT COUNT(*) FROM table WHERE table.z=wert)`
Die Größe eine Subgruppe lässt sich analog dazu bestimmen `(SELECT COUNT(*) FROM table WHERE table.a=wert1 AND table.b=wert2 ...)`
Der Anteil an der Zielvariable wäre dann z. B.

```
SELECT (b.n_sum / a.n_sub) as t_p
FROM (
    SELECT COUNT(*) as n_sub
    FROM table WHERE table.a=wert1 AND table.b=wert2 ...
) as a, (
    SELECT COUNT(*) as n_sum
    FROM table WHERE table.a=wert1 AND table.b=wert2 ... AND table.z=wert
) as b
```

Eine binäre Variable in der Uni-DB wäre z. B. examine.Grade. Mit

```
SELECT COUNT(*) as N FROM examine
```

lässt sich N bestimmen.
Mit

```
SELECT COUNT(*) as n FROM examine WHERE examine.Grade=2
```

lässt sich n bestimmen.

Die Werte für die einzelnen Subgruppen wären dann z. B.:

```
SELECT (b.n_sum / a.n_sub) as t_p
FROM (
    SELECT COUNT(*) as n_sub
    FROM examine WHERE examine.StuNo<=28000
) as a, (
    SELECT COUNT(*) as n_sum
    FROM examine WHERE examine.StuNo<=28000 AND examine.Grade=2
) as b
```

Somit könnte man die Subgruppe StuNo<=28000 bestimmen.

Dieses Vorgehen kann man dann für alle vorher bestimmen Subgruppen durchführen.