Assignment 02

Team 06

Qinyang Wu, st174540@stud.Uni-Stuttgart.de, 3519174
Huicheng Qian, st169665@stud.uni-stuttgart.de, 3443114
Kuang-Yu Li, st169971@stud.uni-stuttgart.de, 3440829

Task 1:

a) The **support** of an itemset can be defined as a fraction of transactions that contain an itemset.
s({B, C, F}) = (σ({B,C,F})) / (|T|) = 0,3
Thus, the support of itemset {B, C, F} is 0,3.

b) The **confidence** of an association rule such as {X} → {Y} can be defined as measuring how often items in itemset {Y} appear in transactions that contain itemset {X}.
c({B, C} → {F}) = (σ({B,C,F})) / (σ({B,C})) = 0,3 / 0,3 = 1
Thus, the confidence of rule {B, C} → {F} is 1.

c) Use the **Apriori algorithm** to determine all frequent itemsets when minsup = 0,25:

i) when k = 1, then the length of candidate itemset is 1,

candidate itemset	support	Frequent? (“Yes” when ≥ 0,25)
{A}	0,5	Yes
{B}	0,4	Yes
{C}	0,5	Yes
{D}	0,4	Yes
{E}	0,2	No
{F}	0,7	Yes
{G}	0,2	No
{H}	0,1	No

Then, {A}, {B}, {C}, {D}, {F} are frequent itemsets.

ii) when k = 2,

candidate itemset	support	Frequent? (“Yes” when ≥ 0,25)
{A, B}	0,1	No
{A, C}	0,1	No
{A, D}	0,2	No
{A, F}	0,3	Yes
{B, C}	0,3	Yes
{B, D}	0,2	No
{B, F}	0,3	Yes
{C, D}	0.1	No
{C, F}	0.4	Yes
{D, F}	0,1	No

Then, {A, F}, {B, C}, {B, F}, {C, F} are frequent itemsets.

iii) when k = 3,

candidate itemset	support	Frequent? (“Yes” when ≥ 0,25)
{B, C, F}	0,3	Yes

Then, only {B, C, F} is the frequent itemset since {A, B, C}, {A, B, F}, {A, C, F} are no candidates as not all their subsets are frequent (apriori property). And the Apriori algorithm stops here because there is only one frequent itemset.

Thus, {A}, {B}, {C}, {D}, {F}, {A, F}, {B, C}, {B, F}, {C, F}, {B, C, F} are all the frequent itemsets based on transactions T.

Task 2:

a) To improve cost efficiency, the variety of drugs used in the MSHA should be reduced. To achieve this goal, it is necessary to identify typical combinations of drugs that doctors prescribe or that are used in certain departments.

Technique: Association Rule Discovery, Descriptive
Reason: In order to identify typical combinations of drugs, we need to analyze the frequency of use of different drugs and the correlation between different drugs according to the drug's current use situation. The goal is descriptive and Association Rule can be applied.
Attributes:

- drug: drugs MSHA used
- department: departments in MSHA
- doctor: doctor in which department
- Min support: the lowest expected value of frequent used drugs combination, which is defined by ourselves
- Min confidence: the lowest confidence for association rule, which is defined by ourselves

Patterns:
The a typical pattern could be a set of association rules for each department

{Drug A, Drug B} -> {Drug C} with confidence 0.98
{Drug A, Drug C} -> {Drug E} with confidence 0.97
…
{Drug E, Drug F} -> {Drug G} with confidence 0.1

Then we can ranked the rules according the confidence and try eliminate the used of drugs that appears in lower rank of the rule
Examples:
If department = "department of anesthesiology" and minsupport = 0,5" and assume the attributes "drug" as a set of transactions D1 = { Drug A, Drug B, Drug C, Drug D}

Database D1

ID	items
1	Drug A, Drug D
2	Drug A, Drug B, Drug D
3	Drug A, Drug B, Drug C
4	Drug B, Drug C
5	Drug A, Drug C, Drug D

⇓

candidate itemset	support	Frequent? (“Yes” when ≥ 0,5)
{Drug A}	0,8	Yes
{Drug B}	0,6	Yes
{Drug C}	0,6	Yes
{Drug D}	0,6	Yes

⇓

candidate itemset	support	Frequent? (“Yes” when ≥ 0,5)
{Drug A, Drug B}	0,4	No
{Drug A, Drug C}	0,4	No
{Drug A, Drug D}	0,6	Yes
{Drug B, Drug C}	0,4	No
{Drug B, Drug D}	0,2	No
{Drug C, Drug D}	0,2	No

Apriori algorithm stops here.
Thus, in drug combinations, only {Drug A, Drug D} is frequent which means the identify typical combinations of drugs in the department of anesthesiology is the combination of Drug A and Drug D.

If a patient comes to the department of anesthesiology, or asks the doctor who works in the department of anesthesiology, the doctor will recommend the drug combination of Drug A and Drug D.

b) The MSHA management identified unusual high costs for drugs in certain departments. The goal is to do some root cause analysis.

Technique: Clustering / Segmentation, Descriptive
Reason: In order to find the reason for unusual high costs for drugs, we need to group people to find out the characteristics of people in these certain departments. The people include both doctors and patients. The high cost for drugs could also result from certain procedures performed in that department. This is a descriptive goal. The segmentation could be done via clustering of doctors, patients and procedures. We can investigate based on the clusterings independently.

Attributes:

- Patients information: age, gender risk group, diagnosis...
- Doctors: type of specialty, year of experience...
- Procedures: treatments, type of treatments, duration...

Patterns:

The clustering of the patients, doctors, and procedures according to their attribute

Examples

The clustering of patients could show the root cause of high cost on drugs in certain departments. For example, the department with 80% of patients diagnosed of chronic disease, which can only be controlled by drug, explains the reason that the department spends most money on drugs

c) The MSHA offers various services to former patients as part of its aftercare program. Hence, the hospital would like to inform the patients about suitable services four weeks after they have been discharged. The goal is to provide such suggestions based on services that were offered to patients or services patients made use of in the past.

Technique: Classification, Predictive

Reason: Through the services that were offered to patients or services patients made use of in the past to predict the patient's status and provide services for them after four weeks. This is predictive. Since service attributes are categorical, we should use classification. Based on the past data, we can generate a decision tree for where an inner node is a single attribute of the patient, an edge in the categories of an attribute, and a leaf node represents a service that the patient received. The tree could be generated with recursive methods using information gain.

Attributes:

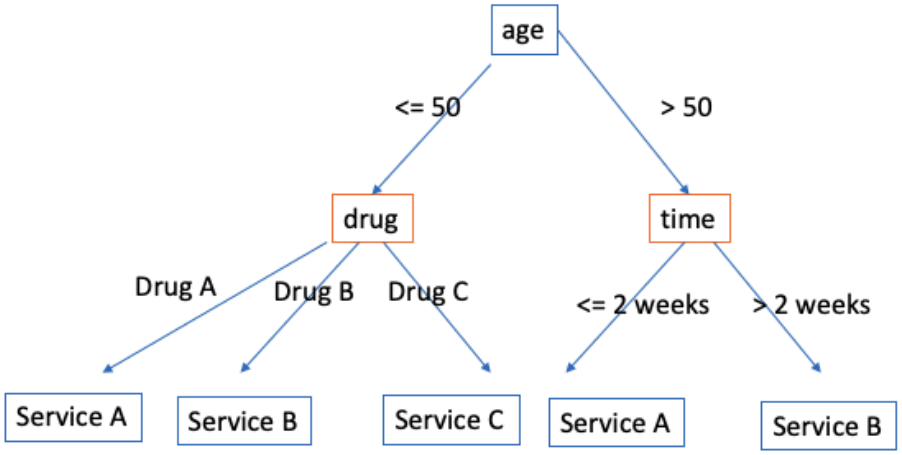
- age: age interval range of patient
- gender: gender of a patient
- disease: disease of a patient
- diagnosis: diagnosis from a doctor
- drug: drugs patient used
- time: the time interval range since patient discharged from the hospital
- service: services patients made use of in the past

Pattern:

The output pattern could be a classification model. For example a decision tree. The leaves of the tree are services and the internal nodes are attributes with split.

Example:

We generated a decision tree based on the given data and attributes as shown below.



Then we can use this decision tree model for providing suggestions.
i) If a patient with the following attributes:
age = 22, drug = drug A,
Then he or she should receive service A
ii) If a patient with following attributes:
age = 50, time = 30 days,
Then he or she should receive service B

d) For planning purposes, the health alliance needs a good estimate for the time patients will stay at a hospital. The goal is to provide such an estimate based on data about previous patients that already have been discharged.

Technique: Regression, Predictive
Reason: Through the analysis and calculation of the existing samples from previous patients that already have been discharged, obtain a continuous result including the patient's recovery and the prediction of discharge time. This requires a predictive technique since the estimated time is a numeric value. Therefore, the method should be a regression. In this case, we can use existing data to train a linear regression model.
Attributes

- Costs of patient's procedure
- Age of patient
- Frequency of patient visiting the hospital
- Number of drugs prescribed for the patients
- Number of department patient has visited
- Previous duration of stay of patient
  All the attributes should be numerical.

Pattern
By used all the data for linear regression model training, we will be able to get all the weightings for the attributes for the expected duration of patient
Duration in hours = w_{age} * (age) + w_{cost} * (cost) + w_{freq} * (freq) + … + w_n *(n^{th} attribute)
Example

if a patient with the following attributes:
age = 22, Costs = 10, Freq = 10, Drugs = 10, Prev-stay = 10
w_{age} = 0.1, w_{cost} =0.2, w_{freq} = 1, w_{drug} = 0.3, $w_{prev\text{-}stay}$ = 0.3
Estimated duration = 0.1 * 22 + 0.2 * 10 + 1 * 10 + 0.3 * 10 + 0.3 * 10 = 20.2 hours