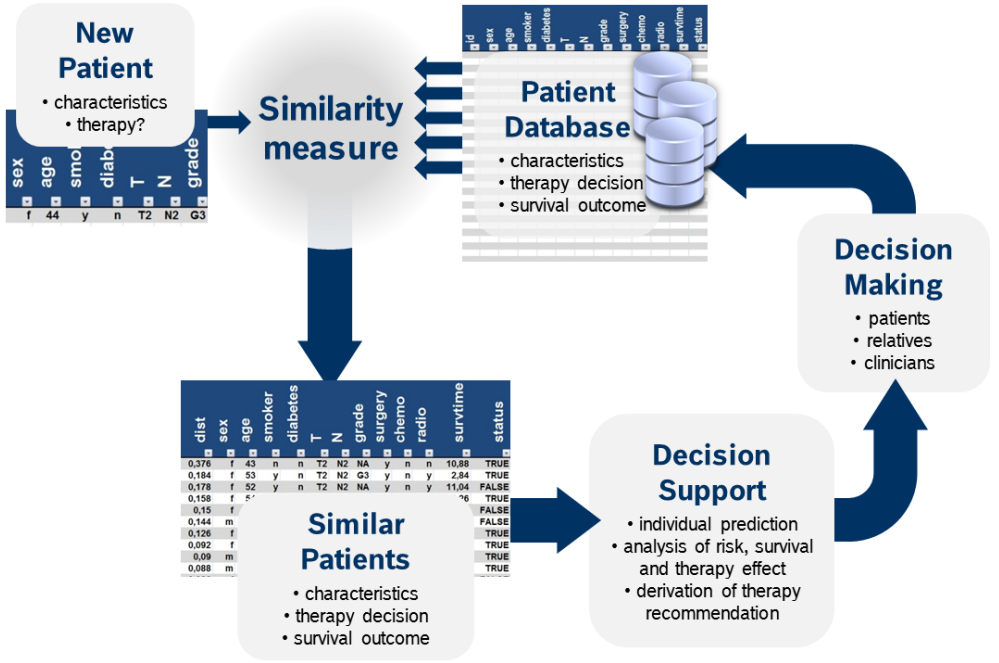
Online Clinical Data Mining

Patient Similarity Analytics

Objective: Individualized Therapy

Two possible approaches:

1. Rule Based Reasoning:
   - Get diagnose of a new case
   - Use knowledge to find therapy decision for this case
2. Case Based Reasoning:
   - Search for similar cases in a database
   - Use therapy which performed best on this similar cases



Patient Similarity Analytics

Objective: Define a similarity measure for num. and cat. features

Let $\mathbf{x}_i = \left(x_i^1, \ldots, x_i^p\right)^T$ and $\mathbf{x}_j = \left(x_j^1, \ldots, x_j^p\right)^T$ be the feature vectors of two patients i and j

A first rough idea may be to calculate the L1-distance of each feature and sum them up:

$$\| \boldsymbol{x}_i - \boldsymbol{x}_j \|_{L_1}^2 = \sum_{k=1}^{p} |x_i^k - x_j^k|$$

Problems:

1. Well defined for numerical features, but what's about categorical features, e.g. pathological N with N0,…, N4 or gender (male/female)
2. $|x_i^k - x_j^k|$ is not scale invariant, e.g. if you change from meters to centimetres, the contribution of this feature would explode by a factor of 100
3. All variables are treated the same way, e.g. lung cancer: smoker (no/yes) seems to be more important than the city you live

Patient Similarity Measures

Weighted Distance Measure

Two steps have to be done to overcome above problems:

1. Generalize the L1-distance on the feature scale, such that it can deal with different variable types
2. Each feature needs to be weighted suitable
    1. to rule out dependency on the scale
    2. account for varying importance across features
    3. impact size

This leads to following weighted distance measure:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{p} |\alpha\left(x_i^k, x_j^k\right) d(x_i^k, x_j^k)|$$

Patient Similarity Measures

Weighted Distance Based on Coefficients from Cox PH Model

- $\alpha\left(x_i^k, x_j^k\right)$ denotes individual weighting for feature k
- In the survival context, we define $\alpha\left(x_i^k, x_j^k\right)$ and $d\left(x_i^k, x_j^k\right)$ as (Klenk et al., 2010)
  - If feature k is numerical

$$d\left(x_i^k, x_j^k\right) = x_i^k - x_j^k$$

$$\alpha\left(x_i^k, x_j^k\right) = \frac{\hat{\beta}^k}{\sum \beta^k}$$

  - If feature k is categorical

$$d\left(x_i^k, x_j^k\right) = \begin{cases} 0 \text{ if } x_i^k = x_j^k \\ 1 \text{ if } x_i^k \neq x_j^k \end{cases}$$

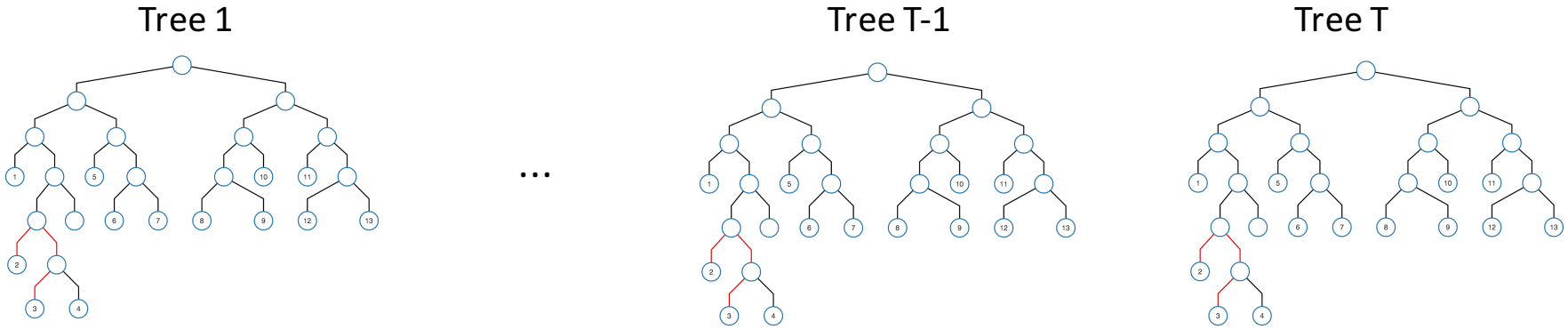
$$\alpha\left(x_i^k, x_j^k\right) = \frac{\hat{\beta}^{k,level_i} - \hat{\beta}^{k,level_j}}{\sum \beta^k}$$

This measure solves problems 1- 3.

Drawback: Patients with missing values are dropped

Patient Similarity Measures

Another Measure: Random Survival Forests Proximity Measure



- Patient i and j are more similar, if the fraction of trees in which patient i and j share the same terminal node is close to 1 (Breiman, 2002)

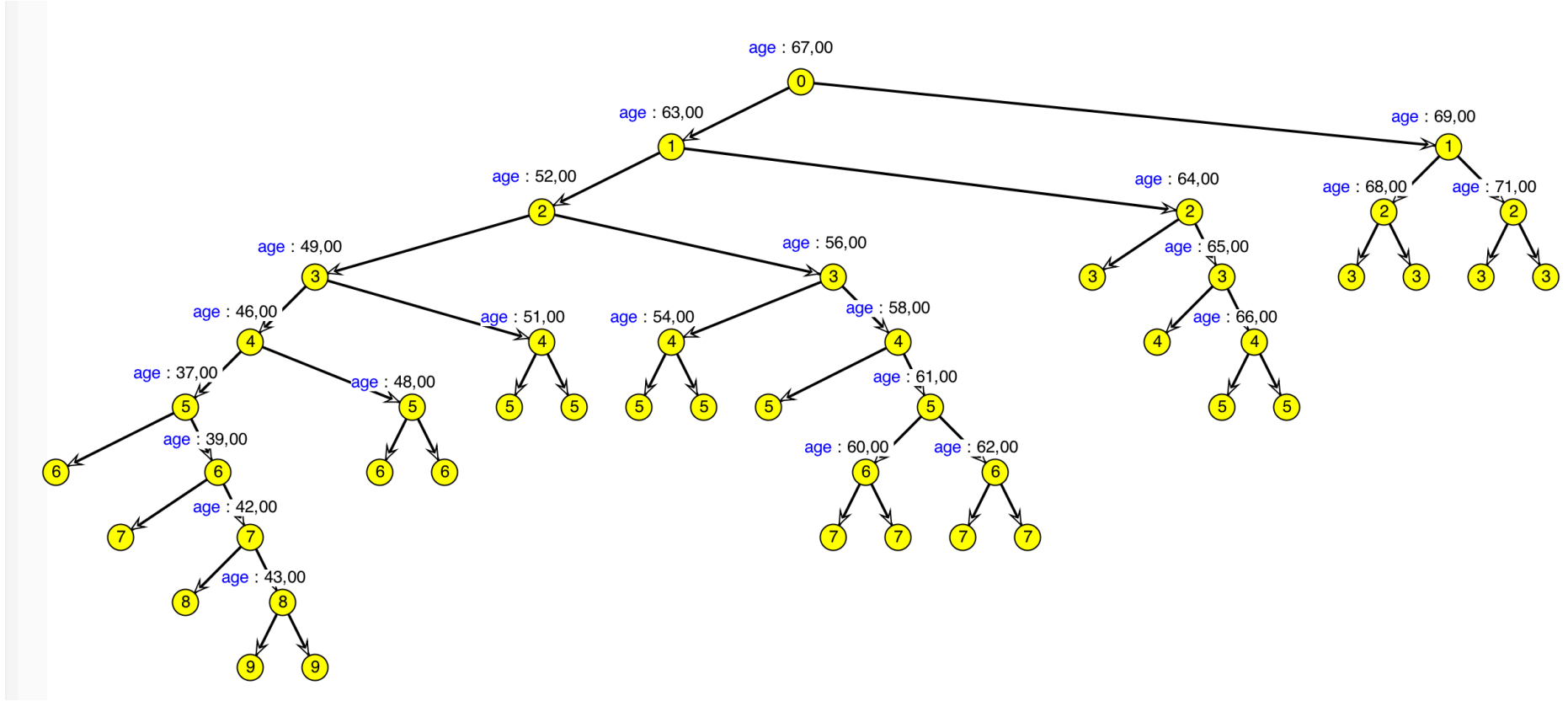
$$d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 = 1 - \frac{1}{M}\sum 1[x_i \; and \; x_j \; in \; same \; terminal \; node \; of \; tree \; t]$$

- $M(\leq T)$ is the number of trees that contains both patients
- RSF is accounting for the nature of the survival data in the splitting rule:
  1. Modified log-rank modeled after Gray‘s test
  2. Weighted log-rank score

Patient Similarity Measures

Another Measure: Random Survival Forests Proximity Measure

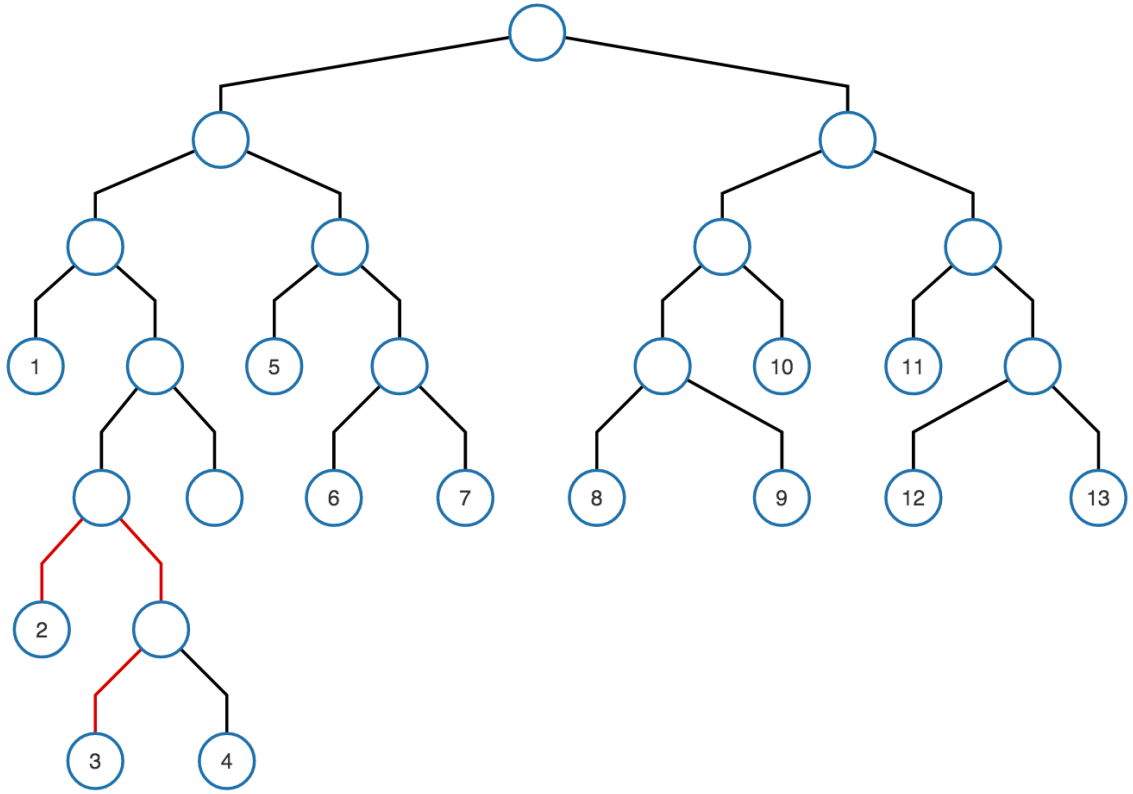
- Decision is binary and „close“ patients are counted as „far“ patients



Patient Similarity Measures

Random Survival Forests: A Modified Proximity Measure

Example Tree:



Distance between patient i und j:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 = 1 - \frac{1}{M}\sum_{t=1}^{M} 1/e^{w g_{ijt}}$$

$\boldsymbol{g_{ijt}} =$ number of edges between end nodes of patient i and j in tree t

$\boldsymbol{w} =$ weighting; with $\boldsymbol{w} = 0$ we get standard proximity measure

Example (red path): $\boldsymbol{g_{ijt}} = 3$

Outlook: Patient Similarity Measures & Matching

Validation Study

- Actually we prepare a validation study on simulated and real data to examine the effect of:
  1. exclusion of variables that have a significant effect on the survival
  2. inclusion of nonsense variables
  3. accuracy of the prediction for similar cases
  4. compare our matching with propensity score methods (Austin, 2012) and targeted learning (van der Laan & Rose)

Outlook: OCDM-Software

Actual Projects

- Cluster Analysis (Status: 80%)
- Matching:
  - Usage of the introduced distance methods (Cox Beta, Modified Proximity Measure) for matching (Status: 80%)
  - Propensity Score Methods for Survival Data (Status: 0%)
- Pathway analysis (e.g. therapy line)
  - Data integration (?)

Online Clinical Data Mining

Gefördert durch:

Robert Bosch Stiftung

Entwickelt von:

Universität Stuttgart in Kooperation mit der TTI GmbH – TGU MUON-STAT

Eingesetzt in:

RBK
Robert-Bosch-Krankenhaus
Klinik Schillerhöhe

OSP
Onkologischer Schwerpunkt Stuttgart

SLK-Kliniken