INF

FAKULTÄT FÜR
INFORMATIK

Seminar / Project topics Summer Term 2016

Tommy Hielscher

Medical Data Mining

Discovering knowledge from medical data

- Finding risk factors associated with specific diseases / disorders
- Identifying subpopulations with common characteristics
- Modeling the evolution of individuals and subpopulations over time
- Prediction of medical outcomes

Team Project: Feature generation from sequences of medical records

Consider a set of patients each described by a sequence of medical records.

e.g.: the Body-Mass-Index measured on three consecutive points in time

id	*som_bmi_s0*	*som_bmi_s1*	*som_bmi_s2*	*Class*
x_1	20.250	21.049	23.950	Neg.
x_2	25.930	28.727	31.869	Pos.
x_3	31.380	30.004	26.645	Neg.
...	...	...	...	...
x_n	26.400	27.912	30.613	Pos.

Team Project: Feature generation from sequences of medical records

What can we learn from this sequences about the target concept?

Considering the sequences associated with an object, what is the meaning of…

- *distinct values within a sequence?*
- *sequence statistics?*
- *the direction of the induced trajectory?*
- *the fluctuation of sequence values / number and kind of significant direction changes within the induced trajectory?*
- *…*

Team Project: Feature generation from sequences of medical records

What can we learn from this sequences about the target concept?

Considering the sequences associated with an object, what is the meaning of…

- *distinct values within a sequence?*
- *sequence statistics?*
- *the direction of the induced trajectory?*
- *the fluctuation of sequence values / **number and kind of significant direction changes within the induced trajectory?***
- *…*

Team Project: Feature generation from sequences of medical records

(1) Generate new object–features containing the implicit knowledge of the sequences about the target concept:

- Find and read literature on the topic (or related topics) of “feature generation from record sequences” and “time–series abstraction”

- Choose, make own suggestions and elaborate which features should be generated with respect to the ***number and kind of significant direction changes within the induced trajectory*** for classification purposes, when objects are described by short record–sequences

Team Project: Feature generation from sequences of medical records

(2) Implementation and evaluation of feature generation:

- Extend (Java) an operator to generate the features for a given data set
- Evaluate whether the generated features enhance classification performance in comparison to using only non–derived features

Software Project /Team Project / Individual Project: Implementation of (semi–)supervised subspace clustering algorithms

Consider epidemiological study data and the disorder fatty liver.

Medical researcher might be interested in the identification of subpopulations and the relevant features for these subpopulations that exhibit distinct distributions regarding fatty liver.

With (semi–)supervised subspace clustering, they might find…

- Groups of participants that have fatty liver and drink much alcohol
- Groups of participants that have fatty liver and a high BMI
- Females with fatty liver and moderate BMI after the onset of the menopause
- Young participants with low BMI that do not exhibit fatty liver
- …

Software Project /Team Project / Individual Project: Implementation of (semi–)supervised subspace clustering algorithms

For these groups different features and value–ranges of these features are differently associated with fatty liver.

The goal of this project is to implement (1) a subspace clustering algorithm in a given framework and (2) evaluate the findings of the algorithm on real epidemiological study data.

The project will be scaled according to team size.

(I.e. pre–selection of algorithm to implement, extent of the evaluation and literature review)