OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG
FACULTY OF COMPUTER SCIENCE

Knowledge Management & Discovery Lab

AI Research @ KMD

Learning algorithms

- Supervised and semi-supervised model learning on streams: adaption to data drift, dealing with evolving feature spaces
- Entity-centric learning: predictions on very few data, predictions on trajectories with gaps, learning on trajectories with systematically missing data (mainly epidemiology and mHealth)
- Semisupervision and active learning: label/information exploitation in evolving feature spaces
- Matching and comarison of models and patterns drawn from partially overlapping populations or samples

Interacting with the experts

- Experiments on acquiring new forms of information from an expert (mainly epidemiology and clinical research)
- Dynamics of expert-delivered knowledge
- Interpretability of models and patterns

Main ongoing projects @ KMD

- ImmunLearning (2019 - 2021) EFRE “Entwicklung eines Tests zur Diagnostik für Immunkompetenz bei Senior*innen mit Hilfe von Data-Mining-Methoden” (with Univ Medicine OVGU)
- CHRODIS+ (2017-2020) EU Joint Action on “Implementing good practices for chronic diseases”
- OSCAR (2017-2019): DFG project”Opinion Stream Classification with Ensembles and Active leaRners” (with Univ Hannover)

Further cooperations in medical research

- Epidemiological Research: Learning on high-dimensional longitudinal data (U Medicine Greifswald)
- Clinical Research: Modeling and predicting patient evolution on streams with gaps - clinical studies & mHealth (U Medicine Regensburg)

Subpopulation Discovery Framework

Motivation

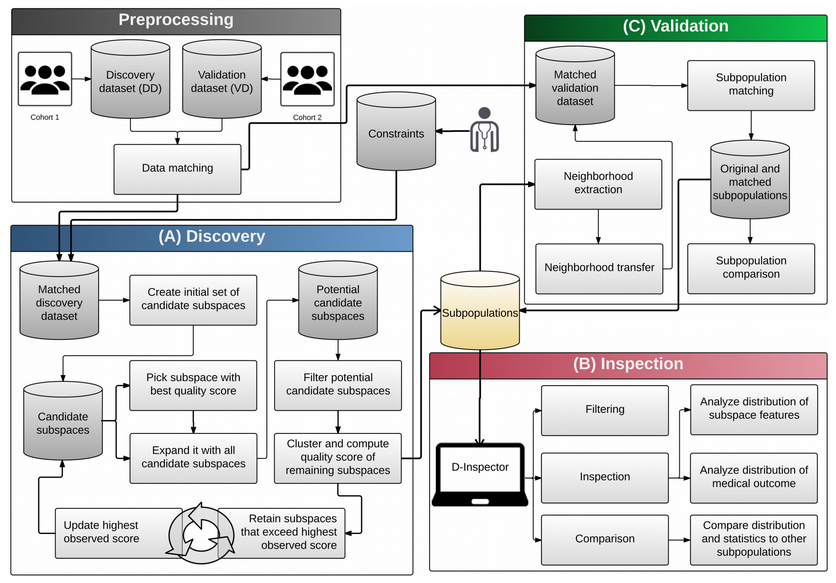
We propose an intelligent system that assists epidemiology experts in analysing the data of a population-based epidemiological study, in identifying relevant variables for an outcome and subpopulations with increased disease prevalence, and in validating the findings in an independent cohort.

Approach

Our system DIVA supports the Discovery, Inspection and VAlidation of subpopulations with increased prevalence of an outcome, without requiring parameter tuning. DIVA takes as input the cohort of an epidemiological study with all variables specified in the study’s protocol, as well as inputs from the expert on the similarity of a small number of cohort participants. DIVA uses semi-supervised subspace clustering and subspace construction to identify sets of variables – subspaces – that promote participant similarity with respect to the outcome and with respect to the expert inputs, and then discovers subpopulations with increased outcome prevalence in those subspaces. DIVA uses visual analytics techniques to assist the expert in juxtaposing, filtering and inspecting the characteristics of these subpopulations. DIVA aligns the cohort used for discovery to a second, independent cohort, and then checks whether the subpopulations found in the original cohort are also present in the second one.

Results

We applied DIVA to the third wave (SHIP-2) of the SHIP-CORE cohort of the Study of Health in Pomerania for the liver disorder “hepatic steatosis”, and on the first wave (TREND-0) of the SHIP-TREND cohort of the same study for the thyroid gland disorder “goitre”. We found that most of the subpopulations extracted automatically, and subsequently ranked and filtered by the modules of DIVA, had significantly higher disease prevalence than the general population. We varied the amount of inputs needed from the expert to drive the subpopulation extraction process and found that a very small amount of information, namely the outcome of as few as 4 cohort participants, is sufficient for the identification of several relevant variables and subpopulations. We used a subset of TREND-0 for the validation on goitre and the complete TREND-0 for the validation on hepatic steatosis and found that the significant difference in prevalence for the identified subpopulation also holds in the validation data.



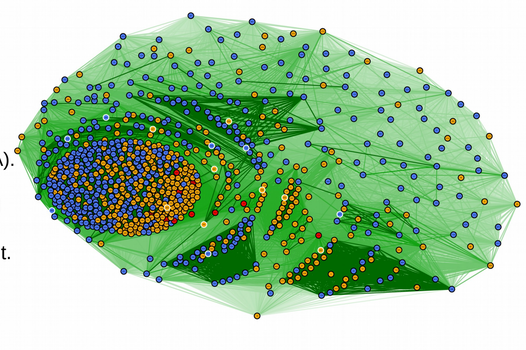
AI and mHealth apps for patient empowerment

Motivation

Chronic diseases cost EU economies around 115 billion euros a year. With the help of mHealth technology, patients can be empowered and supported in their self-management.

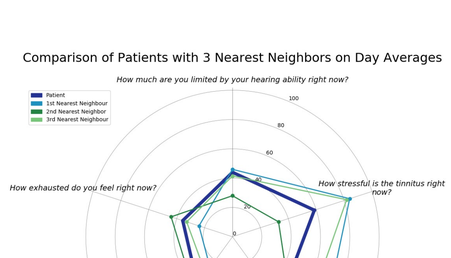
The source of data are Ecological Momentary Assessments (EMA). They…

- offer the opportunity to gain insight into diseases that traditional studies cannot offer.
- enable experts to support personalized approaches to treatment.
- are good applicable for highly variable and person-specific diseases such as tinnitus.



Approach

- quantify study records against key performance indicators such as "performance" and "compliance"
- comparing groups of patients
- identify groups of patient that can benefit most from mHealth technology
- find methods to integrate this technology into the daily lives of healthcare professionals
- find best practices that facilitate caregiver-patient interaction and reduce costs



Results

Preliminary analysis on field data includes:

- Clusters of EMA recordings, using DTW and Frechet distance to deal with different interaction intensities
- First supervised learning results on static and dynamic data to predict consistency of interaction

Exploiting Entity Information for Stream Classification

Motivation

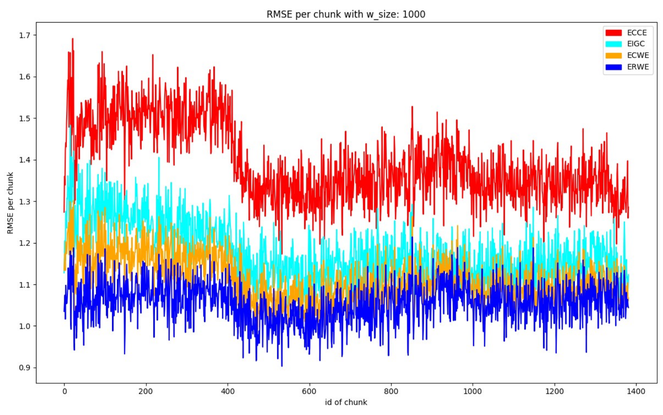
We predict the next rating of an Amazon product from the Tools and Home Improvement category based on the review text. We consider each product to be an entity which has instances (reviews with star rating) tied to it.

We investigate the hypothesis that knowledge of an entity's evolution can improve prediction results.

We compare a entity-ignorant model against entity-centric models and mixed models that combine the predictions of an entity-ignorant and entity-centric model.

Approach

The instances arrive in a stream and the entity-ignorant model (EIGC) is trained on each instance whereas the entity-centric models are only trained on the instances that belong to that particular entity. We have one ensemble that uses only the predictions from the entity-centric classifiers (ECCE), one that creates the mean between entity-centric and entity-ignorant prediction (ECWE) and one that puts weights on both predictions according to the past error (ERWE). As we have ordinal labels we use RMSE as a performance metric so lower values are better.



Results

Using only the entity-centric classifiers (ECCE) is worse than an entity-ignorant model (EIGC) but **combining the predictions leads to better result**s and the best result comes from the error-weighted ensemble (ERWE). The lower RMSE of the mixed models comes mostly from having fewer extreme errors.

Explainable Phenotyping on Clinical Data

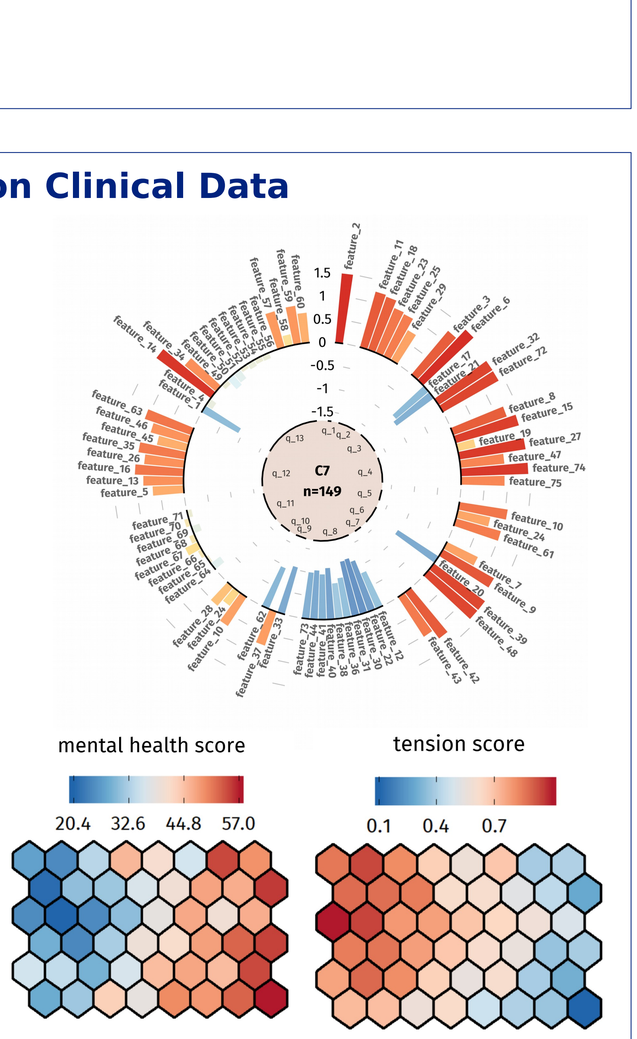
Motivation

The identification of different disease phenotypes can contribute to the determination of a suitable treatment pathway for patients.

Approach

We propose a combination of a bottom-level clustering using self-organizing maps and a top-level clustering using X-Means to deal with the following tasks:

- Which features are representative to a specific phenotype?
- How can we visualize high-dimensional clusters of patients involving dozens or hundreds of features?
- How can we juxtapose multiple phenotypes and highlight characteristics that differentiate themselves from the population average?



Results

Using tinnitus patient questionnaire data, we discovered multiple unique tinnitus phenotypes. The radial barchart visualization shows a cluster of *149 severely affected* tinnitus patients in 77 dimensions.

Using Entity Neighbourhood to Improve Entity-Centric Prediction

Motivation

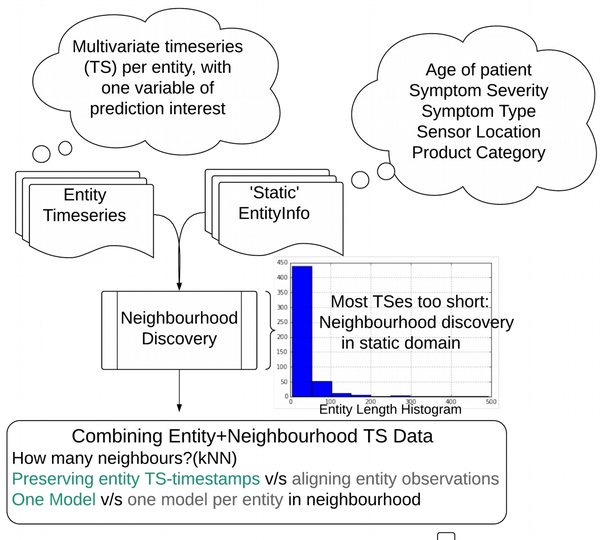
Most data streams have entity-level generating processes. However, stream mining methods do not consider the effect of the data-generating entity on data distribution.

Approach

Methods and workflow to answer the following questions:

1) *Can information from the 'neighbourhood' of an entity be used to improve the predictions of its future labels?*

2) *Can 'similarity' be learned in another space?:*
Many real-world datasets with entity-level time-series have severe sparsity / too little temporal information for meaningful computation of neighbourhoods. Can the computation of the neighbourhood be done over the

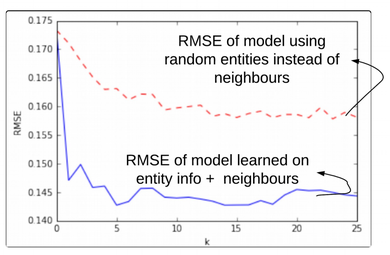


Results

The proposed method using kNN is compared against the results from k randomly selected entites from the dataset.

Entity neighbourhoods can be 'transferred' from another domain.

Across all datasets, for small values of 'k', kRE improves entity-level predictions.



Publications

1) C. Beyer, V. Unnikrishnan, U. Niemann, P. Matuszyk, E. Ntoutsi, and M. Spiliopoulou. Exploiting entity information for stream classification over a stream of reviews. In Proceedings of the Symposium on Applied Computing. ACM, 2019.

2) Tommy Hielscher, Uli Niemann, Bernhard Preim, Henry Völzke, Till Ittermann, and Myra Spiliopoulou. A Framework for Expert-Driven Subpopulation Discovery and Evaluation Using Subspace Clustering for Epidemiological Data. Expert Systems with Applications, (113):147 - 160, 2018.

3) V. Unnikrishnan, C. Beyer, P. Matuszyk, U. Niemann, R. Pryss, W. Schlee, E. Ntoutsi, and M. Spiliopoulou. Entity-level stream classification: exploiting entity similarity to label the future observations referring to an entity. In International Journal of Data Science and Analytics. 2019:1-5.

M.Sc. Christian Beyer
Tel.: +49 391 67 51798
christian.beyer@ovgu.de

Dr.-Ing. Tommy Hielscher
Tel.: +49 391 67 54905
tommy.hielscher@ovgu.de

M.Sc. Uli Niemann
Tel.: +49 391 67 54905
uli.niemann@ovgu.de

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

M.Sc. Miro Schleicher
Tel.: +49 391 67 52333
miro.schleicher@ovgu.de

M.Sc. Vishnu Unnikrishnan
Tel.: +49 391 67 51798
vishnu.unnikrishnan@ovgu.de

Prof. Dr. Myra Spiliopoulou
G29-135
Tel.: +49 391 67 58967
myra@ovgu.de
http://kmd.ovgu.de/