INSTANCE SELECTION ALGORITHMS FOR PREDICTIVE MODELLING IN TELEHEALTH APPLICATIONS

Fabian WIESMÜLLER[a,b,c,], Dieter HAYN [a,c], Florian HOFFMANN[d], Sten HANKE[b], Peter KASTNER[e], Markus FALGENHAUER[a] and **Günter SCHREIER[a]**

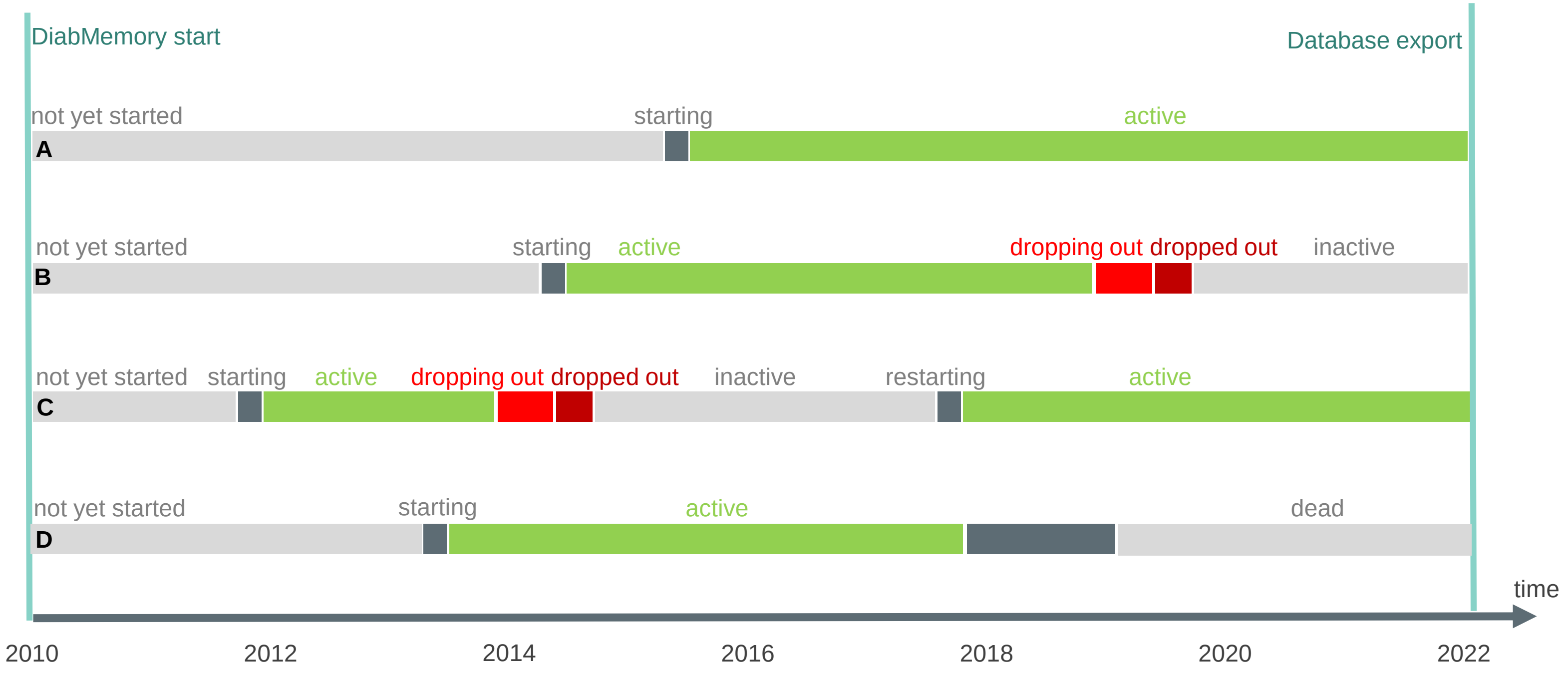
[a] *AIT Austrian Institute of Technology, Graz, Austria*
[b] *FH Joanneum, Graz, Austria*
[c] *Ludwig Boltzmann Institute for Digital Health and Prevention, Salzburg, Austria*
[d] *Versicherungsanstalt öffentlich Bediensteter, Eisenbahnen und Bergbau, Vienna, Austria*
[e] *telbiomed Medizintechnik und IT Service GmbH*

INTRODUCTION

- Increasing use of telehealth and telemedicine solutions
  - DiabMemory was developed in 2010
  - Provide care for diabetes patients

INTRODUCTION

- Adherence is an important aspect in DiabMemory
- Ongoing project works on predicting dropout patients
- A simplified version of this model achieved an AUROC of 0.70



Schematic representation of the monitoring period of four DiabMemory patients

INTRODUCTION

- Potential improvements for this prediction were identified
  - Different machine learning algorithms
  - Feature engineering and selection
  - Instance Selection (IS)

	F_1	F_2	F_3	F_4	F_5	F_6
I_1	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
I_2	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}
I_3	X_{31}	X_{32}	X_{33}	X_{34}	X_{35}	X_{36}
I_4	X_{41}	X_{42}	X_{43}	X_{44}	X_{45}	X_{46}
I_5	X_{51}	X_{52}	X_{53}	X_{54}	X_{55}	X_{56}
I_6	X_{61}	X_{62}	X_{63}	X_{64}	X_{65}	X_{66}

INTRODUCTION

- Potential improvements for this prediction were identified
  - Different machine learning algorithms
  - Feature engineering and selection
  - Instance Selection (IS)

	F_1	F_2	F_3	F_4	F_5	F_6
I_1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
I_2	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}
I_3	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}
I_4	x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	x_{46}
I_5	x_{51}	x_{52}	x_{53}	x_{54}	x_{55}	x_{56}
I_6	x_{61}	x_{62}	x_{63}	x_{64}	x_{65}	x_{66}

INTRODUCTION

- Potential improvements for this prediction were identified
  - Different machine learning algorithms
  - Feature engineering and selection
  - Instance Selection (IS)

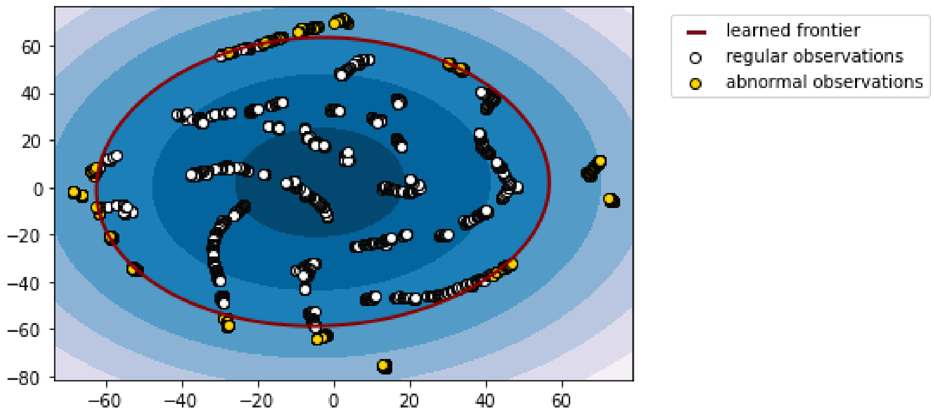
	F_1	F_2	F_3	F_4	F_5	F_6
I_1	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
I_2	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}
I_3	X_{31}	X_{32}	X_{33}	X_{34}	X_{35}	X_{36}
I_4	X_{41}	X_{42}	X_{43}	X_{44}	X_{45}	X_{46}
I_5	X_{51}	X_{52}	X_{53}	X_{54}	X_{55}	X_{56}
I_6	X_{61}	X_{62}	X_{63}	X_{64}	X_{65}	X_{66}

METHODS - DATASET

- Results in a feature matrix of
  - 4,200 days x 3,030 features per patient
    - Padding for patients with shorter monitoring periods
- After feature selection 4,200 days x 344 features
- Initially a large feature set with (partially) redundant data

METHODS - INSTANCE SELECTION

- Intrinsic Instance Selection
  - Random selection
  - Binning
  - Dropout Aligned
  - Sampling at dropout
- Supervised Instance Selection
  - Based on Nearest Neighbor algorithm
- Unsupervised Instance Selection
  - One Class Support Vector Machine (One Class SVM)
  - Mutual Information (MI)



One Class SVM applied to one patient

RESULTS

Algorithm	AUROC (in %)	Event per Non-Event	Number of Instances
Dropout Aligned	69.91	1.16	1,052
Random Selection	68.80	0.10	11,220
Binned	69.19	0.02	73,661
Sampling at Dropout	72.78	1.00	1,138
Classification	68.91	0.08	300,584
One Class SVM	75.88	0.45	21,559
MI	59.67	0.01	162,425

RESULTS

Algorithm	AUROC (in %)	Event per Non-Event	Number of Instances
Dropout Aligned	69.91	1.16	1,052
Random Selection	68.80	0.10	11,220
Binned	69.19	0.02	73,661
Sampling at Dropout	72.78	1.00	1,138
Classification	68.91	0.08	300,584
One Class SVM	75.88	0.45	21,559
MI	59.67	0.01	162,425

One Class SVM performed with the highest AUROC

DISCUSSION AND CONCLUSION

- One Class SVM outperformed the second-best approach by 3.1%

Algorithm	AUROC (in %)	Event per Non-Event	Number of Instances
Sampling at Dropout	72.78	1.00	1,138
One Class SVM	75.88	0.45	21,559

- One Class SVM needs approx. 19 times the data
- One Class SVM not transparent compared to e.g., Sampling at Dropout
- Keep the setting in mind
  - Diabetes vs heart failure
- Keep the ML algorithm in mind
  - Random forest vs neural networks
- Instance Selection improves performance and runtime

THANK YOU!