Knowledge Discovery from Cohorts, Electronic Health Records and further Patient-related data

Myra Spiliopoulou

KDD 2018, London, August 19

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

INF FACULTY OF COMPUTER SCIENCE

KMD

Cohorts [Glenn, 2005]

The term “cohort”

Quoting [Glenn, 2005], page 2: “The term *cohort* originally referred to a group of warriors or soldiers, and the term is now sometimes used in a very general sense to refer to a number of individuals who have some characteristic in common.”

The term “cohort” in “cohort analysis”

Quoting [Glenn, 2005], page 2: “Here and in other literature on cohort analysis, however, the term is used in a more restricted sense to refer to those individuals (human or otherwise) who experienced a particular event during a specified period of time. The kind of cohort most often studied by social scientists is the human *birth cohort*, that is, those persons born during a given year, decade, or other period of time.”

Cohort Analysis [Glenn, 2005]

The term “cohort analysis”

Quoting [Glenn, 2005], page 3: “The term *cohort analysis* is usually reserved for studies in which two or more cohorts are compared with regard to at least one dependent variable measured at two or more points in time.”

Purposes of Cohort Analysis [Glenn, 2005], pages 1-2

- “Assessing the effects of aging”
- “Understand[ing] the sources and nature of social, cultural and political change.”

Counter-examples – [Glenn, 2005], page 3

- *Cross-sectional study*: Comparison of different groups of individuals with respect to some characteristic/variable – such a study “is conducted with data collected at one point in time, or, more accurately, within a short period of time.”
- *Panel study:* Comparison of the the attitudes of a group of individuals at two distinct timepoints – such a study “measures the characteristics of the same individuals at more than one point in time.”

Subtypes of Mild Cognitive Impairment [Guan et al., 2017]

Mild Cognitive Impairment (MCI) comes in different forms: **amnestic MCI** and **non-amnestic MCI** are considered different in etiology (what causes them) and in outcome (symptoms and evolution)

Goal: Classification of MCI subtypes for early, targeted intervention

Source: 184 participants from MAS - 42 aMCI, 27 naMCI, 115 cognitive normal

Sydney Memory and Aging Study - MAS [Sachdev et al., 2010]

Longitudinal study of community-dwelling persons, aged 70-90, recruited from two regions of Sydney

Several exclusion criteria at baseline, including diagnosized dementia, some psychotic disorders, multiple sclerosis etc.

Subtypes of MCI - Data

[Guan et al., 2017]

Inclusion criteria

- MRI scans from baseline and from 2-year followup (wave-2)
- Diagnosis at wave-2 as eiher cognitive normal (CN) or MCI

Data: features derived from MRI

Figure removed

Subtypes of MCI - Workflow [Guan et al., 2017]

Figure removed

Findings: There are MRI features (from baseline, wave-2 or longitudinal) which exhibit differences among pairs of classes - according to the used significance tests.

Subtypes of MCI - Over/Undersampling [Guan et al., 2017]

Quoting from 4th and 5th page:

> *"K-means clustering (Macqueen, 1967) algorithm was used for oversampling, where new synthetic data were generated by clustering the minority class data. Briefly, Ns samples were clustered into Ns/3 clusters, and Ns/3 centroids were generated. Then these centroids and the original samples were combined for the next iteration of oversampling. The oversampling procedure was repeated until the size of minority class was 2/3 the size of the majority class. K-Medoids clustering (Hastie et al., 2001) algorithm was used for undersampling, where actual data points from the majority class were chosen as the cluster centers. The final training set was a combination of the oversampled minority class data and the undersampled majority class data. While resampling the training set, the test set remained the same. The training set was resampled 3 times to reduce the bias due to random data generation. Then the feature selection method was applied on those resampled training sets, thus producing 3 learning models. These models were combined using majority voting, where the final label of an instance was decided based on the majority votes received from all the models."*

Examples and characteristics

Figure removed

Example: Sydney Memory and Ageing Study [Sachdev et al., 2010]

Protocol includes questionnaires, medical tests and (for some participants) MRI.

Some variables are recorded for only a small subset of the cohort.

Example: English Longitudinal Study of Ageing (ELSA) [Steptoe et al., 2012]

Figure removed

Protocol includes several questionnaires, physical examinations and performance (mostly in waves 2, 4, 6), blood tests and DNA recordings (waves 2, 4, 6)

Across time, there have been changes in the protocol, implying gaps in the recordings.

Example: Study of Health in Pomerania (SHIP) [Völzke et al., 2011]

SHIP-Core

- SHIP-0: n=4338, 1997-2001
- SHIP-1: n=3300, 2002-2006
- SHIP-2: n=2333, 2008-2012
- SHIP-3: . . .

Figure removed

Figure removed

SHIP-TREND

- TREND-0: n=4420, 2008-2012
- TREND-1: . . .

Protocol includes questionnaires, medical tests and (from SHIP-2 on) MRI.

Across time, there have been changes in the protocol, implying gaps in the recordings.

The effects of time on a longitudinal population-based study

As time progresses

- the cohort shrinks
- new technologies become available ⇒ new diagnostic tests
- new scientific questions emerge ⇒ new assessments

⇒ the protocol of the study is modified

⇒ gaps in the recordings, systematically missing data – and labels

Exploiting systematically incomplete timestamped data

How to incorporate the unlabeled data into the learning process?

- **Key idea 1:** Exploit people similarity during learning
  $\Downarrow$
  Clustering-andThen-classification
- **Key idea 2:** Use similarity as a feature
  $\Downarrow$
  ClusterIDs as features
- **Key idea 3:** Model people similarity across the time axis

$$\Downarrow$$

- cohort member $:=$ vector of value-sequences [Hielscher et al., 2014]
- cohort member $:=$ member of an evolving cluster [Niemann et al., 2015]

Learning from incomplete value-sequences

[Hielscher et al., 2014]

Turning sequences of values into new features Figure removed

- Discretization: stepwise partitioning of the continuous range of values into segments, so that gain is maximized
- Within-feature density-based clustering of the value-sequences
- Deriving sequence-features to exploit the cross-wave similarity of participants for each feature

Learning from incomplete value-sequences

[Hielscher et al., 2014]

Most important sequence-features

stea_seq: most important sequence-feature for the female subpopulation

Figure removed

stea_seq: important sequence-feature for the male subpopulation

ggt_s_seq: important sequence-feature for the male subpopulation

Exploiting patient evolution for learning

[Niemann et al., 2015]

Figure removed

Learning from evolving clusters [Niemann et al., 2015]

Most important evolution features Figure removed

Constraint-based clustering & subspace clustering

Clustering with instance-based constraints

For a set of clusters ζ and two distinct instances x, y:

- A *Must-Link constraint* on x, y is satisfied by ζ if there is a $C \in \zeta$ so that $x, y \in C$.
- A *Cannot-Link constraint* on x, y is satisfied by ζ if there are $C_1, C_2 \in \zeta$ so that $x \in C_1, y \in C_2$ and $C_1 \cap C_2 = \emptyset$.

DRESS – Discovery of Relevant Example-constrained SubspaceS [Hielscher et al., 2016]

Given a dataset D and a set of ML and NL constraints,
find the ”best” subspace S of the feature space F:

- The clustering in S is of best quality.
- The clustering in S satisfies the constraints.

DRESS [Hielscher et al., 2016]

Quality of a subspace S

Quality wrt constraint satisfaction

$$q_{constraints}(S) = \frac{|ML(S)| + |NL(S)|}{|ML| + |NL|}$$

Cluster stretching with respect to constraints

$$q_{dist}(S) = \frac{\sum_{(x,y)\in NL} d_S(x,y)}{|NL|} - \frac{\sum_{(x,y)\in ML} d_S(x,y)}{|ML|}$$

Overall subspace quality

$$q(S) = q_{constraints}(S) \cdot q_{dist}(S)$$

DRESS workflow [Hielscher et al., 2016]

Figure removed

DRESS evaluation [Hielscher et al., 2016]

Alternatives for feature selection

- No feature selection: all features used for learning
- Correlation-based Feature Selection [Hall, 2000], using m% of the labeled instances
- DRESS, using n% of the labeled instances

Impact of the feature selection on the performance of a classifier

Variant	Avg F-score	Avg AUC	Avg Accurracy	Avg Sensitivity	Avg Specificity	Avg #features
SHIP2·578						
Baseline on kNN	0.544	**0.862**	0.808	0.486	**0.911**	57
CFS(100%) on kNN	0.567	**0.862**	0.813	0.514	0.909	3.69
CFS(3.8%) on kNN	0.547	0.821	0.803	0.526	0.892	3.05
DRESS(3.8%) on kNN	**0.594**	0.859	**0.814**	**0.573**	0.890	17.35
Baseline on C4.5	0.612	0.724	*0.823*	0.591	**0.897**	/
C4.5 on 3.8% of the data	0.560	0.637	0.719	**0.718**	0.719	/
DRESS(3.8%) on C4.5	**0.620**	**0.819**	*0.822*	0.621	0.886	/

DRESS [Hielscher et al., 2016]

Subpopulations found with DRESS Figure removed

Closing on **Learning from cohorts of population-based studies**

- The best of all worlds on data quality and data transparency !
- Relatively few participants
- Very large feature space
- Systematically incomplete data due to protocol changes and due to participants exiting the cohort

Many ML solutions are around

- √ for learning
- √ for feature extraction Transparent workflows needed
- √ for the construction of dynamic features, given gaps
- √ for expert involvement see also next section
- √ for validation of the findings on other cohorts see next section

VISIT THE KMD LAB:

- http://www.kmd.ovgu.de/
- Faculty of Computer Science, Otto-von-Guericke-University Magdeburg
- Sendmail at: myra@ovgu.de

- Thank you!

Acknowledgements: German Research Foundation project OSCAR "Opinion Stream Classification with Ensembles and Active Learners"

Bibliography

[Glenn, 2005] Glenn, N. D. (2005).
Cohort Analysis.
Quantitative Applications in the Social Sciences. SAGE, 2nd edition.

[Guan et al., 2017] Guan, H., Liu, T., Jiang, J., Tao, D., Zhang, J., Niu, H., Zhu, W., Wang, Y., Cheng, J., Kochan, N. A., Brodaty, H., Sachdev, P., and Wen, W. (2017).
Classifying mci subtypes in community-dwelling elderly using cross-sectional and longitudinal mri-based biomarkers.
Frontiers in Aging Neuroscience, 9:309.

[Hall, 2000] Hall, M. A. (2000).
Correlation-based feature selection for discrete and numeric class machine learning.
In *Proc. of 17th Int. Conf. on Machine Learning*, pages 359–366, San Francisco, CA, USA. Morgan Kaufmann.

[Hielscher et al., 2014] Hielscher, T., Spiliopoulou, M., Völzke, H., and Kühn, J.-P. (2014).
Mining longitudinal epidemiological data to understand a reversible disorder.
In *Proc. of Symposium on Intelligent Data Analysis*, pages 120–130.

[Hielscher et al., 2016] Hielscher, T., Spiliopoulou, M., Völzke, H., and Kühn, J.-P. (2016).
Identifying relevant features for a multi-factorial disorder with constraint-based subspace clustering.
In *Proc. of IEEE Symposium on Computer-Based Medical Systems.*

Bibliography

[Niemann et al., 2015] Niemann, U., Hielscher, T., Spiliopoulou, M., Völzke, H., and Kühn, J. (2015).
Can We Classify the Participants of a Longitudinal Epidemiological Study from Their Previous Evolution?
In *IEEE Symposium on Computer-Based Medical Systems*, pages 121–126.

[Sachdev et al., 2010] Sachdev, P. S., Brodaty, H., Reppermund, S., Kochan, N. A., Trollor, J. N., Draper, B., Slavin, M. J., Crawford, J., Kang, K., Broe, G. A., et al. (2010).
The sydney memory and ageing study (mas): methodology and baseline medical and neuropsychiatric characteristics of an elderly epidemiological non-demented cohort of australians aged 70–90 years.
International Psychogeriatrics, 22(8):1248–1264.

[Steptoe et al., 2012] Steptoe, A., Breeze, E., Banks, J., and Nazroo, J. (2012).
Cohort profile: the english longitudinal study of ageing.
International journal of epidemiology, 42(6):1640–1648.

[Völzke et al., 2011] Völzke, H., Alte, D., Schmidt, C. O., Radke, D., Lorbeer, R., Friedrich, N., Aumann, N., Lau, K., Piontek, M., Born, G., et al. (2011).
Cohort profile: the Study of Health In Pomerania.
Int. J. of Epidemiology, 40(2):294–307.