Tutorial
Mining and Model Understanding on Medical Data

Part 3: Learning from Cohorts

Myra Spiliopoulou

KDD 2019, Anchorage: Tutorials Day - Aug 4, 2019

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG
INF
FACULTY OF COMPUTER SCIENCE
KMD

Agenda of this Part

- Cohorts - Basics and examples
- Learning from Cohorts - the conventional way (including clinical studies and RCTs)
- (KDD close) Learning from cohorts with expert involvement

Have you ever heard that . . .

Physical activity can reduce mortality risk and relates to better health status.

Where does this piece of information come from?

1. "Compared with individuals reporting no leisure time physical activity, we observed a 20% lower mortality risk among those performing less than the recommended minimum of 7.5 metabolic-equivalent hours per week" [Arem et al., 2015]
2. "Physical activity was strongly related to a better health status, with active people presenting the higher scores on health, followed by moderately actives." [Caballero et al., 2017]
3. . . .

Are these individuals representative? Of which population?

Two example studies referring to physical activity and health

1. “We pooled data from 6 studies in the National Cancer Institute Cohort Consortium (baseline 1992-2003). Population-based prospective cohorts in the United States and Europe with self-reported physical activity were analyzed in 2014. A total of 661 137 men and women (median age, 62 years; range, 21-98 years) and 116 686 deaths were included.” [Arem et al., 2015]
2. “From a initial sample of 12,099 participants at ELSA baseline [1], a total of 175 subjects (1.45%) were excluded from the analysis since they were unable to be interviewed through poor health, or through physical or cognitive disability. Moreover, 18 subjects (0.15%) were also excluded since their information was missing for at least the 25% of the self-reported health questions and measured tests at ELSA baseline.” [Caballero et al., 2017]

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

Cohorts

Examples

- **ELSA:** “English Longitudinal Study of Ageing” [Steptoe et al., 2012] “The ELSA study began in 2002 and is a biannual, longitudinal and nationally representative survey that focuses on adults aged 50 and over.” [Caballero et al., 2017]
- **MAS:** “The Sydney Memory and Ageing Study” [Sachdev et al., 2010] MAS “was initiated in 2005 to examine the clinical characteristics and prevalence of mild cognitive impairment (MCI) and related syndromes, and to determine the rate of change in cognitive function over time.” [Sachdev et al., 2010]
- **SHIP:** “Study of Health in Pomerania” [Völzke et al., 2010] “After German reunification in 1990, there was a lack of scientifically valid data from East Germany to explain the regional differences in life expectancy and, consequently, a need for population-based research in northeast Germany.” [Völzke et al., 2010]

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."

Learning from Cohorts - the traditional way

- Studies on population-based cohorts deliver insights on disease prevalence, risk factors, protective factors in the populations under observation
- Clinical studies also use cohorts $\longrightarrow$

Have you heard that physical exercise might do as a treatment for depression **?**

Where does this piece of information come from?

1. “Exercise is moderately more effective than a control intervention for reducing symptoms of depression, but analysis of methodologicallyrobust trials only shows a smaller effect in favour of exercise. When compared to psychological or pharmacological therapies, exercise appears to be no more effective, though this conclusion is based on a few small trials.” [Cooney et al., 2013] [2]
2. “Exercise and ICBT were more effective than TAU (Treatment as Usual) by a general medical practitioner, and both represent promising non-stigmatising treatment alternatives for patients with mild to moderate depression.” [Hallgren et al., 2015] (example study that uses an RCT)
3. “Previous meta-analyses may have underestimated the benefits of exercise due to publication bias. Our data strongly support the claim that exercise is an evidence-based treatment for depression.” [Schuch et al., 2016] [3]
4. “There are a number of theories as to why exercise may be an efficacious treatment or augmentation strategy for depression. Depression is associated with low physical activity levels and may have both physiological and psychological antidepressant effects.” [Koppelmans and Weisenbach, 2019]

[2] The measure they use is Standardized Mean Difference, usually shortened to SMD.

[3] Publication bias: [Ioannidis et al., 2014]

Randomized Controlled Trials (RCTs)

- ⋆ The Cochrane Review [Cooney et al., 2013] draws conclusions from 35 RCTs that compared exercise to either no treatment (some RCTs) or to a control intervention (the other RCTs).
- ⋆ The meta-review of [Schuch et al., 2016] draws conclusions from studies on RCTs, after controlling for publication bias.

RCTs – very informally

A set of study participants is selected (following inclusion and exclusion criteria, which are published when reporting on the RCT).

The participants are placed **randomly** in two (equisized) groups: those that are exposed to the `Treatment` and those that serve as `Controls`.

An outcome is specified – to be measured at the end of the exposure.

The participants do not know to which group they belong. Ideally, the treating physician(s) do not know either; for some types of RCTs, this is compulsory.

Core idea: If the two groups are identical in all but the treatment (vs control), then any difference in the outcome (resp. the likelihood of occurence of the outcome's values) must be due to the treatment.

Example of an RCT: the workflow [Hallgren et al., 2015]

Figure 1 from [Hallgren et al., 2015]:
It shows the flow-chart of the reported RCT, across 4 phases (top-down):

- “Enrolment” (includes “Baseline assessment”)
- “Allocation” of the participants to the three interventions at random
- “Intervention” (namely: physical exercise, internet-based Cognitive Behavioural Therapy, Treatment as Usual – each with a duration of 12 weeks)
- “Post-treatment assessment” that took place after 3 months and included again the “Baseline assessment”, as well as a an “exit survey”

Learning from cohorts the traditional way - Any KDD?

The traditional approach of choice encompasses:

- Descriptive statistics
- MICE to deal with missing values
- PCA, ICA to reduce the feature space
- ANOVA, ANCOVA, Linear regression, multi-level models
- Correction for multiple testing (usually: Bonferroni correction)

KDD methods may be used in the workflow-of-learning:

At the beginning:

- AL to acquire labels – rare!
- NNs (on images/signals) to acquire features – frequent

Next to descriptive statistics:

- K-means, SOMs etc, to identify clusters of participants
- to select features interactively

Next to statistical methods:

SVMs, RFs etc, to learn on high-dimensional data and identify predictive fetures – rare!

NEXT: (KDD-close) Learning from Cohorts with Expert Involvement

Placing the expert in the loop

If the expert is willing to teach a learner:

- ActL: If users are repeatedly asked for labels, they may find this annoying or even "lose track of what they were teaching" [a]
- ReinfL: Users prefer to give positive rather than negative rewards [b]

[a] [Amershi et al., 2014], quoting from cite (Cakmak et al, 11).
[b] [Amershi et al., 2014], citing (Thomas & Braezal, 08) and (Knox & Stone, 12).

The expert also wants to do more: [Amershi et al., 2014]

- provide features, weights, changes in weights etc [a]
- experiment with different model inputs
- query the learner about its decisions [b].

[a] [Amershi et al., 2014], citing the findings of an experiment by (Stumpf et al, 07) on 500 different inputs for the improvement of a classifier.
[b] [Amershi et al., 2014], citing the findings on an interactive prototype by (Kulesza et al, 11).

What does the expert want to say – and what not?

- The expert is not necessarily willing to provide labels.
- The expert is willing to provide inputs for a planned study, provided that the impact on the outcome is transparent. Such as:
  - Queries to build a cohort from hospital data cf. CAVA [Zhang et al., 2014]
  - Labels, if there are none cf. CAESAR-ALE [Nissim et al., 2016]
  - Constraints to reduce the feature space DRESS [Hielscher et al., 2018]
  - Time to inspect the outputs cf. DIVA [Hielscher et al., 2018]

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

- **Goal:** get new insights about a population of patients
- **Parties involved:** team of physicians + team of technologists
- **Data:** EHR of hospital patients (timeseries of patient recordings)

Conventional workflow – from [Zhang et al., 2014] with extensions

At the beginning, there is a question/observation – a concrete phenomenon that must be explained (cf. use cases in [Zhang et al., 2014]).

1. The (team of) physician(s) devise one or more hypotheses.
2. The physicians specify the cohort needed for the study of each hypothesis, possibly in interaction with a data analyst or DB expert.
3. The DB expert writes scripts to create the cohort and extract the data.
4. Data analysts build models according to the instructions of the physicians, e.g. on age and gender adjustment.
5. Physicians become a presentation/visualization of the model(s) and check whether their hypothesis is supported.
6. If necessary, GOTO 2.

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

Guidelines for a new workflow:

- Early cohort definition
- Flexible visualization
- Flexible analysis (picking analytics modules from a library)
- Cohort refinement and expansion
- Iterative analysis (building upon the results of previous iterations)

The elements of CAVA

[Zhang et al., 2014]

- **Cohorts:** a data construct

Dealing with the challenges of the feature space

Inner feature space: set of properties shared by all cohort members
Outer feature space: set of all properties of the cohort members

- **Views:** a library of visualization components for presentation and modification of cohorts
- **Analytics:** a library of software components for creation and modification of cohorts

CAVA operates on two databases:

Population database: contains all information about all individuals in the population; is expanded by information derived via analytics or views
Cohort database: contains the description of each cohort (as defined by the user) and the IDs of the cohort members

CAVA: components to begin with [Zhang et al., 2014]

A basic set of analytics components:

- *Batch analytics modules*, including
  - ”demographics module”
  - ”risk stratification module”
- *On-demand analytics modules*, including
  - ”patient similarity component” (published in AMIA 2010)
  - ”utilization analysis component” (published in AMIA 2012)
  - ”heart failure risk assessment component” (published in AMIA 2012)

and workflows for two example scenaria (in [Zhang et al., 2014]).

Interacting with the expert

The expert may be willing to provide:

- √ Queries to build a cohort from hospital data cf. CAVA [Zhang et al., 2014]
- → Labels, if there are none cf. CAESAR-ALE [Nissim et al., 2016]
- Constraints to reduce the feature space cf. DRESS [Hielscher et al., 2018]
- Time to inspect the outputs cf. DIVA [Hielscher et al., 2018]

AL for label acquisition [Nissim et al., 2017] and earlier works

Application area: Classification of condition severity (severe vs mild)
Input: SNOMED-CT and EHR

- CAESAR: Classification Approach for Extracting Severity Automatically from Electronic Health Records (earlier work)
  Labels are delivered by medical experts who inspect the conditions and decide between *severe* and *mild*.
- CAESAR-ALE: CAESAR with Active Learning Enhancement [Nissim et al., 2016]
  - **Goal:** reduce the labeling efforts
  - **Approach:** three active learning methods
  - **Findings:** labeling effort reduced, between 48% and 64% [4]
- CAESAR-ALE followup: exploit inputs from labelers who vary in their expertise [Nissim et al., 2017]

[4] NOTE: Reduced the effort for finding severe rather than mild conditions. They did not investigate the effort needed to find all conditions.

CAESAR-ALE procedure

[Nissim et al., 2016]

AL core for an SVM-classifier

Method that returns for each condition: (i) the confidence of the classifier to the label and (ii) the distance of the condition to the separating hyperplane.

AL methods

- SVM-Simple-Margin, proposed by Tong & Koller (JMLR 2000-2001)
- Method “Exploitation”: selects conditions which
  - are at the “severe” side
  - are far from the separating hyperplane
  - are far from each other
- Method “Combination_XA” that exchanges between the two other methods (one trial each) for a total of n trials

CAESAR-ALE to reduce labeling effort [Nissim et al., 2016]

Experiment 1 on 516 conditions (144 severe, 372 mild) in 10 randomly selected datasets for 10-fold cross validation

- Learning an initial classification model on 6 conditions
- Active learning on a pool of $n = 310$ conditions:

  AL method chooses $k = 5$ conditions at a time
  until the whole pool is processed

  and presents them to the expert
- Test on $m = 200$ conditions
- Estimate savings as

$$Reduction\_In\_Labeling\_Effort * Cost\_of\_Labeling(10529\ conditions) * 3$$

using 3 physician labelers (120 USD per hour)

Differences among labelers

[Nissim et al., 2016]

Experiment 2 with 3 physicians (have completed residency training) & 4 informatics experts (have at least a master degree)

- Learning an initial classification model on 6 conditions
- Active learning on a pool of $n = 100$ conditions:
  AL method chooses $k = 5$ conditions at a time
  until the whole pool is processed
  and presents them to the expert
- Test on $m = 410$ of the 516 conditions in the gold standard

Findings: The two CAESAR-ALE samplers

- find the severe conditions faster; after 62 trials, they have found more than 80 of the 144 severe conditions.
- lead to lower variance among the labelers than the SVM-based sampler

Interacting with the expert

The expert may be willing to provide:

- $\surd$ Queries to build a cohort from hospital data cf. CAVA [Zhang et al., 2014]
- $\surd$ Labels, if there are none cf. CAESAR-ALE [Nissim et al., 2016]
- $\rightarrow$ Constraints to reduce the feature space DRESS [Hielscher et al., 2018]
- Time to inspect the outputs cf. DIVA [Hielscher et al., 2018]
  especially for validation

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 Table showing the effect of DRESS-based feature selection on kNN and C4.5: comparison to the models learned over all features and to models learned with CFS for feature selection

DRESS [Hielscher et al., 2016]

Subpopulations found with DRESS Figure showing mosaic plots for four variables depicting subpopulations with prevalence differences

Interacting with the expert

The expert may be willing to provide:

- √ Queries to build a cohort from hospital data cf. CAVA [Zhang et al., 2014]
- √ Labels, if there are none cf. CAESAR-ALE [Nissim et al., 2016]
- √ Constraints to reduce the feature space DRESS [Hielscher et al., 2018]
- Time to inspect the outputs cf. DIVA for Inspection and Validation [Hielscher et al., 2018]

D-INSPECTOR as part of DIVA

[Hielscher et al., 2018]

Figure 2 from [Hielscher et al., 2018] removed

The validation issue

- $\surd$ Model validation
- ? Validation of the findings

on a dataset drawn independently from the same population

Constraint-based learning and Validation on Cohorts

[Hielscher et al., 2018]

The DIVA framework:

- Discovery: Given cohort dataset D and a set of ML/NL constraints,
  find groups of participants
  within subspaces
  which best describe the concept, as reflected in the constraints,
  where "best" refers to participant similarity/separation and constraint satisfaction.
- Inspection: Given these groups (subpopulations),
  provide ways to identify and analyze the most distinct ones w.r.t. to the medical outcome.
- Validation: Enable experts to investigate whether discovered supopulations are generalizable or not.

DIVA components [Hielscher et al., 2018]

Figure removed

Validation Procedure in DIVA [Hielscher et al., 2018]

Figure removed

Validation example using DIVA [Hielscher et al., 2018]

Learning and Validation with two cohorts:

SHIP-CORE (3rd wave) & SHIP-TREND (1st wave)

Figure removed

Validation example using DIVA [Hielscher et al., 2018]

Learning and Validation with two SHIP cohorts: Discovery of subpopulations with increased prevalence of hepatic steatosis

CORE (3rd wave) for learning & and TREND (1st wave) for validation

Figure removed

Takeaway messages from this part

- Cohort analysis adheres to well-established workflows.
- KDD methods find their way into these workflows, often in data preparation, sometimes in the learning tasks
- KDD methods have large potential for learning on large feature spaces,
- for cohort construction, refinement and validation,
- and for the exploitation of expert knowledge in the learning process

Visual analytics is essential for expert involvement.

Understanding the workflow and placing KDD methods *in it* is even more essential.

THANK YOU !

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

[Amershi et al., 2014] Amershi, S., Cakmak, M., Knox, W. B., and Kulesza, T. (2014).
Power to the people: The role of humans in interactive machine learning.
AI Magazine, 35(4):105–120.

[Arem et al., 2015] Arem, H., Moore, S. C., Patel, A., Hartge, P., De Gonzalez, A. B., Visvanathan, K., Campbell, P. T., Freedman, M., Weiderpass, E., Adami, H. O., et al. (2015).
Leisure time physical activity and mortality: a detailed pooled analysis of the dose-response relationship.
JAMA internal medicine, 175(6):959–967.

[Caballero et al., 2017] Caballero, F. F., Soulis, G., Engchuan, W., Sánchez-Niubó, A., Arndt, H., Ayuso-Mateos, J. L., Haro, J. M., Chatterji, S., and Panagiotakos, D. B. (2017).
Advanced analytical methodologies for measuring healthy ageing and its determinants, using factor analysis and machine learning techniques: the athlos project.
Scientific reports, 7:43955.

[Cooney et al., 2013] Cooney, G. M., Dwan, K., Greig, C. A., Lawlor, D. A., Rimer, J., Waugh, F. R., McMurdo, M., and Mead, G. E. (2013).
Exercise for depression.
Cochrane database of systematic reviews, (9).

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

Bibliography

[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.

[Hallgren et al., 2015] Hallgren, M., Kraepelien, M., Lindefors, N., Zeebari, Z., Kaldo, V., Forsell, Y., et al. (2015).
Physical exercise and internet-based cognitive–behavioural therapy in the treatment of depression: randomised controlled trial.
The British Journal of Psychiatry, 207(3):227–234.

[Hielscher et al., 2018] Hielscher, T., Niemann, U., Preim, B., Völzke, H., Ittermann, T., and Spiliopoulou, M. (2018).
A framework for expert-driven subpopulation discovery and evaluation using subspace clustering for epidemiological data.
Expert Systems with Applications, 113:147 – 160.

[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*.

[Ioannidis et al., 2014] Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., and David, S. P. (2014).
Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention.
Trends in cognitive sciences, 18(5):235–241.

Bibliography

[Koppelmans and Weisenbach, 2019] Koppelmans, V. and Weisenbach, S. L. (2019).
Mechanisms underlying exercise as a treatment for depression.
The American Journal of Geriatric Psychiatry, 27(6):617–618.

[Nissim et al., 2016] Nissim, N., Boland, M. R., Tatonetti, N. P., Elovici, Y., Hripcsak, G., Shahar, Y., and Moskovitch, R. (2016).
Improving condition severity classification with an efficient active learning based framework.
Journal of Biomedical Informatics, 61:44–54.

[Nissim et al., 2017] Nissim, N., Shahar, Y., Elovici, Y., Hripcsak, G., and Moskovitch, R. (2017).
Inter-labeler and intra-labeler variability of condition severity classification models using active and passive learning methods.
Artificial Intelligence in Medicine, 81:12 – 32.
Artificial Intelligence in Medicine AIME 2015.

[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.

Bibliography

[Schuch et al., 2016] Schuch, F. B., Vancampfort, D., Richards, J., Rosenbaum, S., Ward, P. B., and Stubbs, B. (2016).
Exercise as a treatment for depression: a meta-analysis adjusting for publication bias.
Journal of psychiatric research, 77:42–51.

[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., 2010] Völzke, H., Alte, D., Schmidt, C. O., Radke, D., Lorbeer, R., Friedrich, N., Aumann, N., Lau, K., Piontek, M., Born, G., et al. (2010).
Cohort profile: the study of health in pomerania.
International journal of epidemiology, 40(2):294–307.

[Zhang et al., 2014] Zhang, Z., Gotz, D., and Perer, A. (2014).
Iterative cohort analysis and exploration.
Information Visualization (Info Vis), pages 1–19.