Cluster Analyses of a Target Data Set from the IFCS Cluster Benchmark Data Repository: Introduction to the Special Issue

Iven Van Mechelen and Werner Vach

Abstract After a brief introduction to benchmarking in data analysis in general and in cluster analysis in particular, we describe the setup of the IFCS Cluster Benchmark Data Repository along with two challenges connected with it. The first of these challenges called for data sets to be contributed to the repository; the second one pertained to cluster analyses of the winning data set of the first challenge. Subsequently, we introduce the winning data set of the first challenge together with relevant meta-data. We conclude with a brief description of the organization of the present special issue, which comprises reports of analyses that have been submitted as contributions to the second challenge.

Iven Van Mechelen
KU Leuven, Tiensestraat 102 box 3713, B-3000 Leuven, Belgium,
✉ Iven.VanMechelen@kuleuven.be

Werner Vach
University of Freiburg, Stefan-Meier-Strasse 26, D-79104 Freiburg, Germany,
✉ wv@imbi.uni-freiburg.de

Archives of Data Science, Series B (Online First)
KIT Scientific Publishing
Vol. 1, No. 1, 2019

DOI 10.5445/KSP/1000085952/01
ISSN 2510-0564

1 Benchmarking in Data Analysis in General and in Cluster Analysis in Particular

In any data-analytic process many alternative methods are available for pre-processing the input of the process, for the actual data analysis (in the narrow sense), and for post-processing the output of the data analysis. This obviously gives rise to the question as to which methods are optimal (in which respect(s) and for which types of data). To answer this question, comparative evaluations of the methods in question are highly needed. Such comparative evaluations can be referred to as instances of benchmarking. In some cases, benchmarking can be based on mathematical-theoretical analyses. More often, however, it will be based on analyses of (empirical as well as simulated) data.

There is a fairly long benchmarking tradition in many subdomains of data analysis. Yet, the situation is rather different in the domain of unsupervised classification or cluster analysis where there is much less of such a tradition. This obviously considerably hampers a cumulative building of knowledge in the clustering domain. This lack of a benchmarking tradition goes with a dearth of both a conceptual and a data grasp.

Recently, a Task Force within the *International Federation of Classification Societies* (IFCS) has tried to address these gaps. To deal with the dearth of a conceptual grasp, a white paper has been written with an extensive discussion of the theoretical and conceptual underpinnings of data-based benchmarking in the field of cluster analysis, and of the practicalities of how to address benchmarking questions in clustering (along with foundational recommendations and guidelines) (Van Mechelen et al, 2018). To deal with the dearth of a data grasp, a blueprint of a cluster benchmark data repository has been devised, which will be explained in more detail below (Section 2). Subsequently, we will introduce two challenges connected with this repository (Section 3), along with the winning data set of the first challenge (Section 4). We will conclude with a brief description of the organization of the present special issue (Section 5).

2 The IFCS Cluster Benchmark Data Repository

Nowadays many subdomains of science face an increasing attention for all kinds of data sharing. There are many reasons for this, including not least concerns about reproducibility of research results. This is evidenced by the appearance of new data journals (e.g., *Scientific Data*, and *Open Health Data*) and of research data repositories (e.g., the *UCI Machine Learning Repository*), with associated surveys and registries.

In general, data sharing implies quite a number of major challenges, which may be easily underestimated in practice. Indeed, a sound communication of data should go far beyond a basic communication of numbers or data entries, in that it also requires a clear communication of the data structure, the nature and meaning of experimental units and variables involved in the data, information on what are admissible values for the variables and their meaning, the code(s) used to denote missing data entries, full information on possible required or recommended types of preprocessing, and so on. Within the context of cluster benchmarking, in addition to this, there is a clear need for extensive additional and nontrivial information on meaningful quality concerns and criteria for a clustering.

The IFCS Cluster Benchmark Data Repository has been designed in such a way as to meet the different communication needs as outlined above. This repository, which is linked to the IFCS website at `http://ifcs.boku.ac.at/repository/`, will comprise data sets with and without given "true" clusterings. One of the distinctive features of the repository is that comprehensive meta-data will be supplied with each data set. This meta-data information, which providers of data will have to enter via a custom-made questionnaire, will comprise documentation on the specific nature of the clustering problem and on characteristics that useful clusters should fulfil (with scientific justification). In particular, the meta-data will cover the following topics:

- the subject matter background, data structure, admissible data values and their meaning, required or recommended types of data preprocessing, etc.
- a substantive justification of why a clustering of the data is needed (if available);

- whether an external variable is available that is to be used to judge the clustering result, and, if so, whether this relates to a known underlying true clustering or to some pragmatic aim (and, if so, to which one, and why);
- whether certain internal characteristics of the clustering are to be used to judge the clustering result; examples of such internal characteristics may refer to:
  - cluster membership (e.g., should all points be clustered and why? should clusters be allowed to overlap and why?);
  - within-cluster features (e.g., are there requirements on what should be the unifying ground for elements to belong to the same cluster, such as small within-cluster similarities or a common pattern of values on some cluster-defining variables, and, if so, why?);
  - between-cluster features (e.g., are there requirements on what should be the discriminating basis for elements to belong to different clusters, such as some form of separability, and, if so, why?);
  - aspects that go beyond the data set under study (e.g., is quality of inferences about some kind of population characteristics an issue and, if so, which one and why?).

3 Two Challenges Connected with the Repository

Two challenges have been organized in connection with the IFCS Cluster Benchmark Data Repository.

3.1 Challenge 1

The goal of the first challenge was for entrants to contribute data set(s) to the repository. Six submissions for this challenge were received, which were subsequently evaluated by the IFCS Task Force on Benchmarking. The evaluation criteria for the challenge included:

- technical correctness (which includes a correct specification of the numbers of objects and variables, of the data structure, and of the admissible data values);
- quality of the meta-data.

3.2 Challenge 2

The goal of the second challenge was for entrants to contribute cluster analyses of the winning data set of the first challenge. The required format for contributions was: a short report (with detailed justification of the performed analysis), a graphical representation of the analysis' result and its evaluation, and code of the analysis. Eight submissions with analyses of target data set were received. These were again evaluated by the IFCS Task Force on Benchmarking. The evaluation criteria for this challenge included:

- technical correctness and clarity of the report;
- linking choices in the analysis and the evaluation of its result to the meta-data, in addition to the quality of reflection about what constitutes a good clustering.

4 Target Data Set for the Second Challenge

In this section we will describe the winning data set of the first challenge, which also constituted the target data set for the second challenge.

- name of the data set: Baseline assessment and outcome measures of low back pain (LBP) patients;
- contributor: Werner Vach (University of Freiburg, Germany);
- papers to be cited when using this data set: Kongsted et al (2015); Nielsen et al (2016);
- subject matter background: data from a longitudinal study of adult LBP patients who consulted chiropractors;
- structure of the data: object by variable data;
- nature of the objects: 928 LBP patients;

- nature of the variables that were selected for the challenge (for a comprehensive list of all variable labels and descriptions, see Appendix):
  - 112 variables measured at baseline via self-report questionnaires and clinical examinations:
    - biographical: sex, age, educational level, etc.
    - pain history: duration of LBP, pain distribution, etc.
    - psychological: depression, recovery beliefs, avoidance, etc.
    - activity limitation: walk more slowly, cannot work, etc.
    - participation: stay home, decreased sexual activity, etc.
    - physical impairment: pain on flexion, pain on extension, etc.
  - 3 outcome variables measured via follow-up questionnaires at 2 weeks, 3 months, and 12 months after baseline:
    - global perceived improvement;
    - LBP intensity;
    - general health status measure for LBP (summary score of Roland-Morris Disability Questionnaire (Roland and Morris, 1983)).
- substantive justification of why a clustering is needed:
  - LBP is a highly heterogeneous condition.
  - There is a clear need for a better understanding of the mechanisms underlying this heterogeneity.
  - Such an understanding would benefit from knowledge of the prognosis of LBP.
  - There is an interest in a clinically useful grouping of patients based on their baseline characteristics only (between 3 and 12 groups).
  - It would be useful if patient groups could be characterized with reference to a few key variables (to reduce later data collection).
- quality concerns:
  - external criteria:
    - A clustering with a pragmatic aim will be looked for.
    - Longitudinal outcomes could be used as external variables to inform or validate the clustering.

- internal criteria:
  - cluster membership:
    - It is natural that some patients are on the border between different groups.
    - A small group of unclassifiable patients may be acceptable.
    - If clusters reflect different conceptual characteristics, it may be natural to allow for cluster overlap.
    - Clusters may vary in size; yet, a large number of small clusters would limit clinical acceptability.
  - within-cluster features:
    - unifying ground: There should be a sufficient degree of within-cluster similarity that would allow for a conceptual labelling of the clusters.

5 Organization of the Special Issue

This special issue has been organized as follows: This introductory paper will be followed by six papers each of which will present a cluster analysis of the target data set that was described in the previous section (two further contributors did not submit a paper for this special issue). The special issue will conclude with a discussion paper in which the clusterings that result from the different analyses will be compared, using visualization methods, outcome means and confidence intervals, and cluster validation indices.

Acknowledgements The work on this paper has been supported in part by the Interuniversity Attraction Poles program of the Belgian government (grant IAP P7/06 to Iven Van Mechelen), and by the Research Fund of KU Leuven (grant GOA/2015/03).

References

Kongsted A, Kent P, Hestbaek L, Vach W (2015) Patients with low back pain had distinct clinical course patterns that were typically neither complete recovery nor constant pain: A latent class analysis of longitudinal data. The Spine Journal 15(5):885–895, DOI 10.1016/j.spinee.2015.02.012

Nielsen AM, Vach W, Kent P, Hestbaek L, Kongsted A (2016) Using existing questionnaires in latent class analysis: Should we use summary scores or single items as input? A methodological study using a cohort of patients with low back pain. Clinical Epidemiology 8:73–89, DOI 10.2147/CLEP.S103330

Roland MO, Morris RW (1983) A study of the natural history of back pain. Part 1: Development of a reliable and sensitive measure of disability in low back pain. Spine 8:141–144

Van Mechelen I, Boulesteix AL, Dangl R, Dean N, Guyon I, Hennig C, Leisch F, Steinley D (2018) Benchmarking in cluster analysis: A white paper. Manuscript submitted for publication, URL https://arxiv.org/abs/1809.10496

Appendix

List of variables in the Target Data Set for the second challenge:

Number	Label	Description(*)
1	id	Patient identifier
2	gen12m	Global perceived improvement 12 months after baseline consultation
3	vasl12m	LBP intensity 12 months after baseline consultation
4	rmprop12m	Roland-Morris summary score 12 months after baseline consultation
5	gen3m	Global perceived improvement 3 months after baseline consultation
6	vasl3m	LBP intensity 3 months after baseline consultation
7	rmprop3m	Roland-Morris summary score 3 months after baseline consultation
8	gen2w	Global perceived improvement 2 weeks after baseline consultation
9	vasl2w	LBP intensity 2 weeks after baseline consultation
10	rmprop2w	Roland-Morris summary score 2 weeks after baseline consultation
11	Bsex0	Sex

(*) Explanation of abbreviations:

AROM	Active Range Of Motion
LBP	Low back pain
SI-joint	Sacroiliac joint tests

Number	Label	Description(*)
12	Age	Age
13	Budd0	Highest educational level
14	Barb0	Work situation
15	Bfor0	Health insurance
16	Bfbe0	Physical work load
17	Bhoej0	Height
18	Bryg0	Smoking status
19	Dlva0	Duration of LBP
20	Dlsy0	Days with sick leave last month
21	Vasl0	LBP intensity
22	Vasb0	Leg pain intensity
23	Tlep0	Previous LBP episodes
24	Tlda0	More than 30 days of LBP last year
25	Okon0	Able to decrease pain
26	Okom0	Negative recovery belief
27	Oens0	Feel socially isolated
28	Obeh0	Treatment not essential
29	Start10	Pain has spread down leg(s)
30	Start20	Shoulder/neck pain
31	Start30	Have only walked short distances
32	Start40	Dressed more slowly last two weeks
33	Start50	Not safe to be physically active
34	Start60	Worrying thoughts a lot of the time
35	Start70	Terrible back pain, will never get better
36	Start80	Not enjoyed things used to enjoy
37	Start90	Bothersomeness of back pain last 2 weeks
38	Htil0	Self-rated general health
39	Rm10	Stay home most of the time
40	Rm20	Change position frequently
41	Rm30	Walk more slowly
42	Rm40	Not doing usual jobs around the house
43	Rm50	Use handrail to get upstairs
44	Rm60	Hold on to something to get out of an easy chair

(*) Explanation of abbreviations:

AROM	Active Range Of Motion
LBP	Low back pain
SI-joint	Sacroiliac joint tests

Number	Label	Description(*)
45	Rm70	Get dressed more slowly
46	Rm80	Only stand for short periods of time
47	Rm90	Try not to bend or kneel down
48	Rm100	Difficult to get out of a chair
49	Rm110	Back / leg painful almost all the time
50	Rm120	Difficult to turn over in bed
51	Rm130	Trouble putting on socks
52	Rm140	Only walk short distances
53	Rm150	Sleep less well
54	Rm160	Avoid heavy jobs around the house
55	Rm170	More irritable with people than usual
56	Rm180	Go upstairs more slowly
57	Rm190	Stay in bed most of the time
58	Rm200	Decreased sexual activity
59	Rm210	Rubbing / holding areas that hurt / are uncomfortable
60	Rm220	Do less daily work around the house
61	Rm230	Often express concern
62	Fabq10	Pain caused by physical activity
63	Fabq20	Physical activity makes worse
64	Fabq30	Physical activity might harm back
65	Fabq40	Should not do physical activity which (might) make pain worse
66	Fabq50	Cannot do physical activities which (might) make worse
67	Fabq60	Pain caused by work / accident at work
68	Fabq70	Work aggravated pain
69	Fabq80	Claim for compensation
70	Fabq90	Work is too heavy
71	Fabq100	Work makes / would make pain worse
72	Fabq110	Work might harm back
73	Fabq120	Should not do normal work with present pain
74	Fabq130	Cannot work with present pain

(*) Explanation of abbreviations:

AROM	Active Range Of Motion
LBP	Low back pain
SI-joint	Sacroiliac joint tests

Number	Label	Description(*)
75	Fabq140	Cannot work till pain is treated
76	Mdi1	Felt low in spirits / sad
77	mdi2	Lost interest in daily activities
78	Mdi3	Felt lacking in energy and strength
79	Mdi4	Felt less self-confident
80	Mdi5	Had a bad conscience
81	Mdi7	Have had difficulty in concentrating
82	Mdi8	Felt very restless / subdued / slowed down
83	Mdi9	Had trouble sleeping at night
84	Mdi10	Have suffered from reduced / increased appetite
85	Rmprop	Roland-Morris summary score
86	facetextrot	Pain on extension / rotation
87	facetsit	Best posture to sit
88	facetwalk	Best activity is not to walk
89	Paraspin_debut	Non-paraspinal pain onset
90	musclepalp	Painful muscle palpation
91	triggerpoint	Trigger points
92	nootherdisease	No other chronic disease
93	heartdisease	Heart disease
94	asthma	Asthma / allergy
95	psychdisease	Psychological disease
96	musculoskeldiseas	Musculoskeletal disease
97	otherchronicdisea	Other chronic disease
98	musclegroup_palp	Painful muscle group(s)
99	Pain_dis	Pain distribution
100	Domin_bp	LBP not dominating
101	Romflex	Pain on flexion (AROM)
102	Romext	Pain on extension (AROM)
103	Romsideglr	Pain on sideglide, right (AROM)
104	Romsidegll	Pain on sideglide, left (AROM)
105	Romrotr	Pain on right rotation (AROM)
106	Romrotl	Pain on left rotation (AROM)

(*) Explanation of abbreviations:

AROM	Active Range Of Motion
LBP	Low back pain
SI-joint	Sacroiliac joint tests

Number	Label	Description(*)
107	Mdtreduce	Reducible disc (diagnosis)
108	Mdtpartlyreduce	Partly reducibel disc (diagnosis)
109	Mdtnonreduce	Non-reducible disc (diagnosis)
110	Mdtdysfunc	Dysfunction syndrome (diagnosis)
111	Herndiscr	Indication of herniated disc, right
112	Herndiscl	Indication of herniated disc, left
113	Affstrenght	Affected muscular strength
114	Affsens	Affected sensibility
115	Affdtr	Affected deep tendon reflexes
116	sisep_comb	SI-joint: Separation test
117	siP4_comb	SI-joint: Thigh thrust
118	sigaens_comb	SI-joint: Gaenslens
119	sicompres_comb	SI-joint: Compression test
120	sithrust_comb	SI-joint: Sacral thrust
121	bmi	Body Mass Index
122	Start_risk	High-risk group (Keele STarT Back Screening Tool)

(*) Explanation of abbreviations:

AROM	Active Range Of Motion
LBP	Low back pain
SI-joint	Sacroiliac joint tests