Using Cluster Ensembles to Identify Psychiatric Patient Subgroups

Vincent Menger[1,2(✉)], Marco Spruit[1], Wouter van der Klift[1], and Floor Scheepers[2]

[1] Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
{v.j.menger,m.r.spruit,w.vanderklift}@uu.nl
[2] Department of Psychiatry, University Medical Center Utrecht, Utrecht, The Netherlands
f.e.scheepers-2@umcutrecht.nl

Abstract. Identification of patient subgroups is an important process for supporting clinical care in many medical specialties. In psychiatry, patient stratification is mainly done using a psychiatric diagnosis following the Diagnostic and Statistical Manual of Mental Disorders (DSM). Diagnostic categories in the DSM are however heterogeneous, and many symptoms cut across several diagnoses, leading to criticism of this approach. Data-driven approaches using clustering algorithms have recently been proposed, but have suffered from subjectivity in choosing a number of clusters and a clustering algorithm. We therefore propose to apply cluster ensemble techniques to the problem of identifying subgroups of psychiatric patients, which have previously been shown to overcome drawbacks of individual clustering algorithms. We first introduce a process guide for modelling and evaluating cluster ensembles in the form of a Meta Algorithmic Model. Then, we apply cluster ensembles to a novel cross-diagnostic dataset from the Psychiatry Department of the University Medical Center Utrecht in the Netherlands. We finally describe the clusters that are identified, and their relations to several clinically relevant variables.

Keywords: Cluster ensembles · Mental healthcare · Psychiatry · Patient subgroups · Patient stratification · Applied data science

1 Introduction

Identification of patient subgroups is an important process that is able to guide clinical treatment in many medical specialties. In psychiatry, the main construct for stratifying patients is a psychiatric diagnosis, typically performed using the Diagnostic and Statistical Manual of Mental Disorders (DSM). This manual describes various high level disorders such as depressive disorders, anxiety disorders, and developmental disorders, with sub-types for each category. It defines

D. Riaño et al. (Eds.): AIME 2019, LNAI 11526, pp. 252–262, 2019.
https://doi.org/10.1007/978-3-030-21642-9_31

clear diagnostic criteria based on symptoms—a major depressive disorder for instance can only be diagnosed after eight symptoms have been assessed, including depressed mood, weight loss, fatigue, and inability to concentrate, and at least five were observed in a two-week period. While the DSM is by far the most widely adopted standard for diagnosis, in recent years its rigid approach has been subject to criticism. Research for instance shows that the DSM has little biological validity (i.e. lack of connection to biomarkers), that diagnostic categories are not specific (i.e. large heterogeneity exists within groups), and that symptoms often cut across diagnostic categories [4].

This critique on the DSM has seeded data-driven approaches that seek interesting subgroups using relevant datasets rather than using expert elicited criteria. For this purpose, various clustering algorithms that are able to discover latent subgroups have been applied to patient data. One major downside of a clustering approach however is the need to select an appropriate number of clusters and an appropriate clustering algorithm, which both have been shown to provide challenges for researchers [12]. The majority of studies rely on a single metric for choosing the right number of clusters, and subsequently apply a single clustering algorithm [14], while both choices can have significant impact on the results that are obtained. Consequently, as of yet no consensus exists on either the number or nature of psychiatric patient subgroups that can be derived in this data-driven way.

In this work we therefore propose to apply cluster ensembles, i.e. combinations of multiple clustering algorithms, to this problem. This enables identification of distinct subgroups that can directly inform treatment, while overcoming the downsides of individual clustering algorithms. Previous work has already shown that cluster ensembles often improve robustness, stability and accuracy over individual clustering algorithms, both in general and in the medical domain, yet this approach is still rare in mental healthcare research [8,22].

The contribution of this work is twofold. First, we present a process guide for modelling and evaluating cluster ensembles in the form of a Meta Algorithmic Model, as introduced in [20]. This guide aims to support researchers in applying cluster ensembles in their particular (medical) domain. Second, we apply a cluster ensemble approach to a novel cross-diagnostic dataset of 1,098 Youth Self Report (YSR) questionnaires of adolescents that were treated at the University Medical Center Utrecht in the Netherlands. Since these questionnaires were routinely captured during treatment, using them to identify patient subgroups, if present, can have direct applicability in the psychiatric practice [16,17]. After applying the cluster ensemble approach, we examine key characteristics of the clusters we obtained, and assess their relation to several clinically relevant variables including DSM diagnosis.

2 Background and Related Work

Clustering algorithms have previously been used in mental healthcare research for stratifying patients with a common psychiatric diagnosis, such as schizophrenia, depression, or autism [14]. The number of clusters ranges from two to seven,

typically selected based on a single measure such as Bayesian Information Criterion or Ward's method. Most researchers then apply one algorithm to their dataset, such as K-means Clustering, Hierarchical Clustering, or Latent Class Cluster Analysis. Clusters of various natures have been found, for instance based on differences in symptoms [5], treatment outcome and onset [3], and patient functioning [6]. A smaller number of studies focused on stratifying patients in a cross-diagnostic setting. A study by Olinio et al. for instance found six subtypes differing in presence of depression, anxiety, or a mixture of both [18], while Lewandowski et al. reported a neuropsychologically normal subtype, a globally impaired subtype, and two mixed cognitive profiles [13], and Kleinman et al. found a cluster with diminished sustained attention, inhibitory control and vigilance, and increased impulsiveness, and a second converse cluster [11]. So far, cluster ensembles have only been applied once in mental health research in a study by Shen et al. who used this technique to identify four subtypes of pervasive developmental disorders [19]. They reported differences in severity, in problems with language acquisition and impairment, and in aggressive behaviour.

To reduce variability in clustering outcomes, such as for example described above, cluster ensembles were proposed based on the principle that multiple weak partitions in combination can provide a more accurate and objective outcome than a single strongly optimized clustering [7]. This is analogous to ensemble learning techniques such as Boosting and Random Forests in the supervised domain. First, during the generation stage, a number of diverse partitions are obtained, ideally with strengths and weaknesses in different parts of the solution space [8]. This is for instance achieved by using multiple clustering algorithms and different algorithm parameters, by subsetting data, and by projecting data to subspaces [22]. The result of the generation stage on a dataset $X = \{x_i, ..., x_n\}$ with n observations is a partition set $P = \{p_1, ..., p_m\}$ of m partitions, where each $p_i = \{C_1^i, ..., C_k^i\}$ assigns every observation to a single cluster C_i out of k clusters. In the subsequent consensus stage, an optimal partitioning is obtained using partition set P. Various procedures have been proposed based on object co-occurrence in clusters, such as majority voting [23], or the graph-based Cluster-based Similarity Partitioning Algorithm (CSPA) [21]. Another type of approach finds the median partition in $p^* \in P$, for instance defined as the partition that maximizes similarity with all other partitions in P [24]. Cluster ensembles have recently successfully been applied in several biomedical domains [2].

3 Meta Algorithmic Model

To support researchers in applying cluster ensembles to their (medical) domain, we propose a Meta Algorithmic Model (MAM) of cluster ensemble modelling and evaluation (Fig. 1). Our MAM is an extension of the original work of Spruit and Jagesar [20], that was aimed at supervised learning tasks. In their words, MAMs are intended to provide "highly understandable and deterministic method fragments — i.e. activity recipes — to guide application domain experts without in-depth Machine Learning expertise". Method fragments are specified as a

combination of a Unified Modelling Language (UML) activity diagram showing processes, and a UML class diagram showing concepts.

The cluster ensemble modelling process, shown on the left of Fig. 1, starts with loading a prepared dataset. Then, in the generation stage multiple methods for introducing diversity in the cluster portfolio are used, including observation and feature sampling, choosing clustering algorithms and selecting a number of repetitions. After a number of clusters and a distance measure are selected, the cluster portfolio is created. In the subsequent consensus stage a consensus function should be selected, and weak partitions can be trimmed from the cluster portfolio. During the evaluation stage, internal index criteria (e.g. Carlinski-Harabasz, Silhouette) can be evaluated, and clusters can be visualized after applying a dimension reduction algorithm to the dataset. Cluster characteristics can be identified based on the cluster assignments of the dataset, and an external evaluation (e.g. using expert evaluation, or comparison to a reference class)

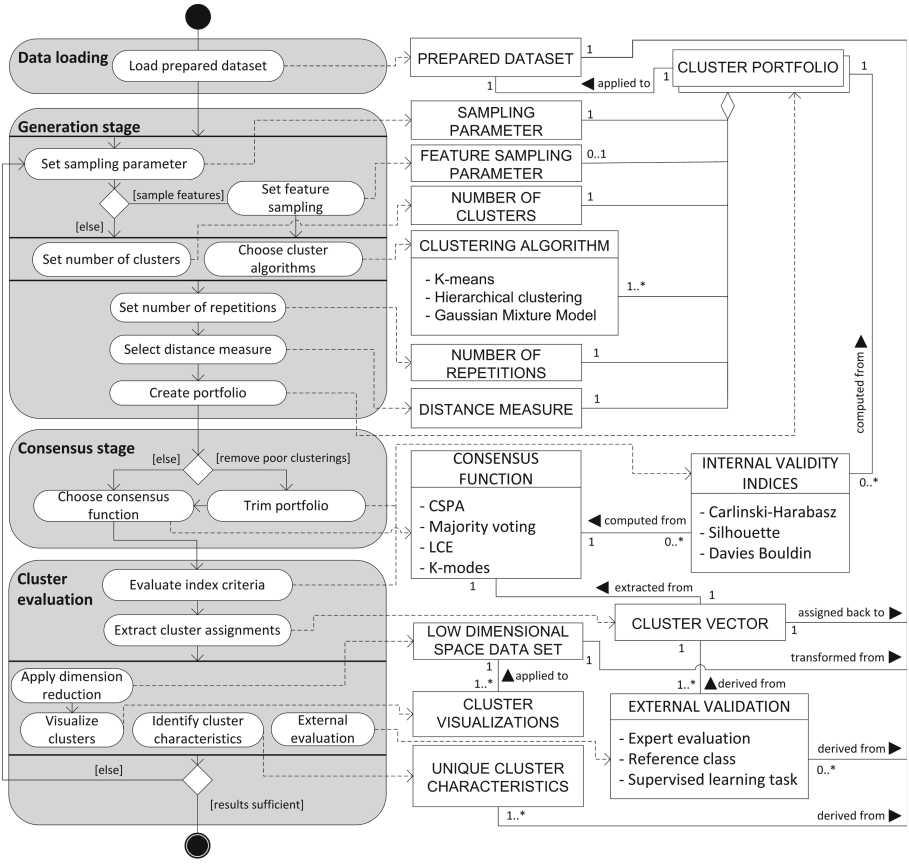

Fig. 1. Method fragment of the Meta Algorithmic Model for cluster ensemble modelling and evaluation.

can finally be performed. The class diagram on the right of Fig. 1 shows which concepts need to be instantiated in relation to each process step.

4 Applying Cluster Ensembles

4.1 Dataset

We applied the cluster ensemble modelling approach in Fig. 1 to a novel cross-diagnostic dataset of adolescent patients who were treated at the Psychiatry Department of the University Medical Center Utrecht in the Netherlands. The dataset consisted of Youth Self Report (YSR) questionnaires, a standardized checklist aimed at adolescents. It consists of 112 items in the form 'I am/have/feel *symptom/behaviour*', which a respondent can indicate as 'not true', 'somewhat or sometimes true', and 'very true or often true'. The YSR defines eight outcome scales by summing responses of specific item subsets: (1) Anxious depressed, (2) Withdrawn depressed, (3) Somatic complaints, (4) Social problems, (5) Thought problems, (6) Attention problems, (7) Rule breaking behaviour, and (8) Aggressive behaviour. We dismissed 50 reports with more than five percent out of 112 items missing, and for the remaining YSRs imputed missing values with the median score of that item. If multiple reports of a patient were present ($n = 175$), we used only the first report, under the assumption that treatment effect is smallest at this point. Our final dataset consists of 1,098 YSRs. The mean age of respondents was 14.7 years ($SD = 2.2$), and 44.5% of respondents were female.

For cluster ensemble modelling, we used the eight outcome scales of the YSR as input data. Since these scales have a non-arbitrary zero value (i.e. absence of any symptoms), we chose to analyse them as ratio scales, using Euclidean distance, implicitly assuming equidistant item scores. Since the outcome scales are a sum of individual items measured on an ordinal scale, they could also be regarded as ordinal scales themselves. However, this distinction is often relatively unimportant in practice, especially when performing clustering [9]. Analysing these data as ratio scales furthermore allows a larger variety of clustering algorithms to be applied to this dataset, most likely improving clustering outcomes.

4.2 Cluster Ensemble Modelling

One risk of performing cluster analysis is obtaining clusters, while no natural grouping exists in a dataset. For this purpose, we computed the Hopkins statistic as a measure of clustering tendency [1]. This statistic is computed from a dataset X with n observations by creating a sample $Y \subseteq X$, and a set of uniform randomly sampled points U, with U and Y both of size $m \ll n$. Then, let q_i be the distance of $u_i \in U$ to its nearest neighbour in X, and let p_i be the distance of $y_i \in Y$ to its nearest neighbour in X, according to some distance measure d. The Hopkins statistic is finally given by:

$$H = \frac{\sum_{i=1}^{m} q_i}{\sum_{i=1}^{m} p_i + \sum_{i=1}^{m} q_i} \tag{1}$$

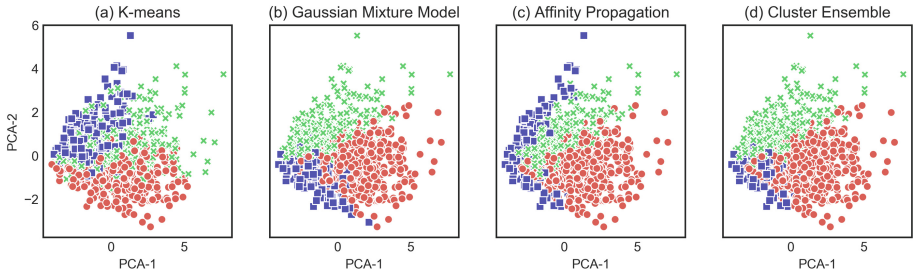

Fig. 2. Partitions of the dataset after applying Principal Component Analysis, both based on single algorithm (a–c), and combined in Cluster Ensemble (d).

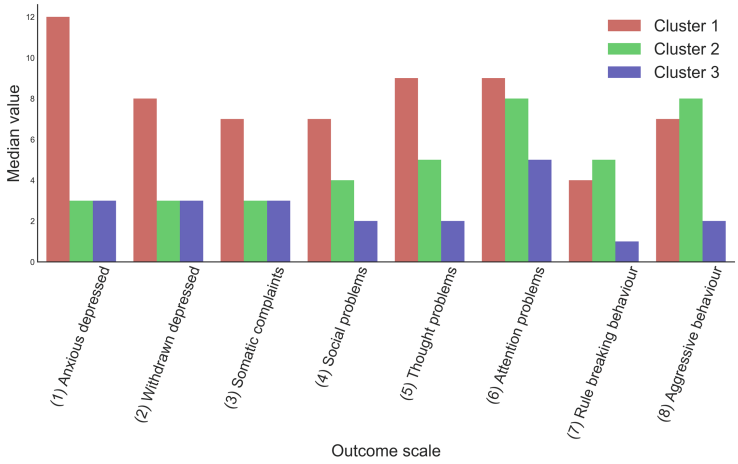

Fig. 3. Median YSR outcome scale value for each of the three identified clusters.

The Hopkins statistic ranges from 0 (uniformly distributed data), to 0.5 (randomly distributed data) to 1 (highly clusterable data). Computing this statistic for our dataset using Euclidean distance obtains $H = 0.71$. No definitive cut-off for cluster tendency has been established, but a value between 0.5 and 1 is regarded as indicative for high likelihood of significant clusters.

Next, an appropriate number of clusters k should be selected. Rather than rely on a single measure for determining this number, we used the R package `NbClust`, which computes 26 internal validity indices for several values of k, and proposes an optimal number of clusters based on a majority vote. We computed the validity indices in combination with both K-means clustering and hierarchical clustering, and set the number of clusters between two and seven. The majority vote shows that the optimal number of clusters $k = 3$ for our dataset, which we will use in all following steps.

For application of the cluster ensemble to our dataset, the R package `DiceR` was used, which implements various cluster ensemble techniques. In order to

find an appropriate subset of algorithms, we applied each of the twelve implemented algorithms with their standard settings to the dataset. We then selected three algorithms that obtain different partitions of the dataset, based on two-dimensional Principal Component Analysis (PCA) plots. These are the K-means algorithm (Fig. 2a), which minimizes the within-cluster sum of squares using an iterative approach, a Gaussian Mixture Model (Fig. 2b), which models the dataset with a mixture of multi-dimensional Gaussian probability distributions, and the Affinity Propagation algorithm (Fig. 2c), which approaches a dataset as a network in which data points communicate with all other points.

To obtain a diverse cluster portfolio, we used five reruns for each of the three clustering algorithms with a random subset of 80% of all data. The number of clusters k is fixed to three, as determined previously. We trimmed the cluster portfolio using a Rank Aggregation method: all partitions were ranked based on several internal validity indices, and the 75% highest partitions were retained. We finally used the Cluster-based Similarity Partitioning Algorithm (CSPA) to obtain a single clustering based on the cluster portfolio (Fig. 2d). All analysis code is made publicly available on GitHub[1].

5 Cluster Evaluation

Applying the cluster ensemble method to our dataset results in three clusters, which contain respectively 55.5%, 32.1%, and 12.5% of observations (Fig. 2d). The ensemble clustering shows strongest similarity with the Gaussian Mixture model, with some differences in the two smallest clusters, and greater differences with the K-means and Affinity Propagation partitions. To assess statistical significance of the three clusters found by the cluster ensemble approach, we used the sigClust method [10] which tests against a null hypothesis of all data being from a single Gaussian distribution. This results in $p = 0.01$ when applied to our dataset, indicating presence of significant clusters at the $\alpha = 0.05$ level.

Figure 3 shows the median value of the eight YSR scales over the three clusters, where distinctions among the three clusters can be observed. Cluster 1, the largest cluster, has the highest overall scores, especially in the two depressed scales (1–2). Values of other scales are among the highest as well in Cluster 1, with Rule Breaking Behaviour being the lowest item. Clusters 2 and 3 on the other hand generally have lower scores, with equal median outcomes on the Anxious depressed, Withdrawn depressed, and Somatic problems scales (1–3). For the other five scales, Cluster 2 shows higher outcomes. For the Rule Breaking Behaviour and Aggressive Behaviour scales (7–8), Cluster 2 shows higher median values than Cluster 1 as well.

To identify clusters' distinguishing characteristics, we integrated clinical notes from the EHR, i.e. pieces of text written by caregivers about treatment, that were de-identified using the DEDUCE method [15], in the two weeks surrounding YSR response. We extracted the 1000 most frequent terms from these texts, and computed the Spearman correlation coefficient for each term and

[1] http://www.github.com/vmenger/cluster-ensembles.

each of the three clusters vs the other two clusters. A psychiatrist then selected three informative terms among those with the highest positive correlation coefficients. For Cluster 1, the selected terms are *depressive*, *dejected*, and *suicidal*, which is in line with high scores in the two depressed scales. For Cluster 2, the terms *behavioural problems*, *adhd* (attention deficit hyperactivity disorder), and *distracted* are identified, which is in line with high scores on the Attention Problems and Aggressive Behaviour scales. For Cluster 3, these terms are *speech*, *verbal*, and *individual*. Based on these terms and Fig. 3, we describe Cluster 1 as 'depressive symptoms', and Cluster 2 as 'behavioural problems'. A comprehensive description of Cluster 3 is less evident, we therefore describe it as 'low severity'.

Table 1 shows the three clusters versus the main DSM diagnosis, which had been made definitive within 12 weeks of YSR response for a subset of 665 patients. The most common diagnosis for the three clusters respectively are Anxiety Disorder, Attention Deficit Disorder, and Pervasive Developmental Disorder (PDD). Diagnoses are typically present in several clusters, although they are usually most prominent in one single cluster, with the exception of PDDs.

Table 1. Main DSM diagnoses per cluster, if finalized within 12 weeks.

Disorder	Cluster 1	Cluster 2	Cluster 3	Total
Anxiety disorder	76 (19.2%)	7 (3.5%)	13 (18.3%)	14.4%
Developmental disorder				
Attention deficit disorder	39 (9.8%)	82 (41.4%)	10 (14.1%)	19.7%
Pervasive developmental disorder	103 (26.0%)	51 (25.8%)	24 (33.8%)	26.8%
Other	15 (3.8%)	15 (7.6%)	2 (2.8%)	4.8%
Eating disorder	10 (2.5%)	2 (1.0%)	3 (4.2%)	2.3%
Mood disorder	65 (16.4%)	13 (6.6%)	7 (9.9%)	12.8%
Psychotic disorder	24 (6.1%)	7 (3.5%)	4 (5.6%)	5.3%
Personality disorder	27 (6.8%)	1 (0.5%)	0 (0.0%)	4.2%
Other	37 (9.3%)	20 (10.1%)	8 (11.3%)	9.8%
Total	**396 (100%)**	**198 (100%)**	**71 (100%)**	**100%**

We finally integrate several clinically relevant variables, including Global Assessment of Functioning (GAF) score at start and end of treatment, a seven-point burden of disease indicator, and length of treatment (Table 2). Although Cluster 1 has the highest overall YSR outcome scale scores, the GAF scores both at start and end of treatment are relatively low. The difference between these groups are assessed with a Kruskal-Wallis one-way analysis of variance test. Results show that significant differences in GAF score at start and end of treatment and in length of treatment exist at the $\alpha = 0.05$ level, but not in burden of disease. This indicates that clusters do not only differ in YSR outcome scales, but also in variables that are relevant in clinical practice.

Table 2. Average value of clinically relevant variables per cluster. P-value is assessed using a Kruskal-Wallis test, * indicates significance at the $\alpha = 0.05$ level. GAF = Global Assessment of Functioning.

Variable	Cluster 1	Cluster 2	Cluster 3	P-value
GAF at start of treatment	45.9	50.0	46.0	0.008*
GAF at end of treatment	53.6	56.7	51.5	0.012*
Burden of disease	4.5	4.5	4.8	0.550
Length of treatment (days)	132.6	175.3	160.7	0.003*

6 Discussion and Conclusion

In line with previous research, our results point out that different clustering algorithms indeed obtain different partitions. Cluster ensembles are a useful method to overcome such issues. By applying our proposed cluster ensemble approach to a dataset of YSR questionnaires, we obtained three distinct patient subgroups. Patients with the same DSM diagnosis are typically represented in multiple clusters, indicating that the three clusters are to some extent a novel stratification of adolescent patients. We furthermore identified significant differences in GAF both at start and end of treatment, and in length of treatment. Although absolute differences among clusters are modest, this shows that patient subgroups do not only differ in the YSR outcome scales.

The clustering outcomes of this study are limited by both the type of data and the specific patient population that reported it. The dataset includes eight outcome scales that are general, but may not capture all dimensions of patient well-being, and whether the clusters we obtained generalize to other populations should be the topic of further research. The main contribution of this research however lies in the cluster ensemble approach, and the process guide introduced with the Meta Algorithmic Model. Such cluster ensemble approaches are able to eliminate one source of variance in reported psychiatric patient subgroups, and can thereby in the future contribute to the identification of a more robust and objective stratification of psychiatric patients.

References

1. Banerjee, A., Dave, R.: Validating clusters using the Hopkins statistic. In: IEEE International Conference on Fuzzy Systems (2004)
2. Boongoen, T., Iam-On, N.: Cluster ensembles: a survey of approaches with recent extensions and applications. Comput. Sci. Rev. **28**, 1–25 (2018)
3. Cole, V.T., Apud, J.A., Weinberger, D.R., Dickinson, D.: Using latent class growth analysis to form trajectories of premorbid adjustment in schizophrenia. J. Abnorm. Psychol. **121**(2), 388–395 (2012)
4. Cuthbert, B.N., Insel, T.R.: Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC Med. **11**(1), 126 (2013)

5. Dollfus, S., Everitt, B., Ribeyre, J.M., Assouly-Besse, F., Sharp, C., Petit, M.: Identifying subtypes of schizophrenia by cluster analyses. Schizophr. Bull. **22**(3), 545–555 (1996)
6. Fountain, C., Winter, A.S., Bearman, P.S.: Six developmental trajectories characterize children with autism. Pediatrics **129**(5), e1112–e1120 (2012)
7. Fred, A.: Finding consistent clusters in data partitions. In: Kittler, J., Roli, F. (eds.) MCS 2001. LNCS, vol. 2096, pp. 309–318. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-48219-9_31
8. Ghaemi, R., Sulaiman, N., Ibrahim, M., Mustapha, N.: A survey: clustering ensembles techniques. Int. J. Comput. Inf. Eng. **3**(2), 365–374 (2009)
9. Harpe, S.E.: How to analyze likert and other rating scale data. Curr. Pharm. Teach. Learn. **7**(6), 836–850 (2015)
10. Huang, H., Liu, Y., Yuan, M., Marron, J.S.: Statistical significance of clustering using soft thresholding. J. Comput. Graph. Statist. **24**(4), 975–993 (2015)
11. Kleinman, A., et al.: Attention-based classification pattern, a research domain criteria framework. Aust. N. Z. J. Psychiatry **49**(3), 255–265 (2014)
12. Kuncheva, L., Hadjitodorov, S.: Using diversity in cluster ensembles. In: IEEE International Conference on Systems, Man and Cybernetics (2004)
13. Lewandowski, K.E., Sperry, S.H., Cohen, B.M., Ongur, D.: Cognitive variability in psychotic disorders: a cross-diagnostic cluster analysis. Psychol. Med. **44**(15), 3239–3248 (2014)
14. Marquand, A.F., Wolfers, T., Mennes, M., Buitelaar, J., Beckmann, C.F.: Beyond lumping and splitting: a review of computational approaches for stratifying psychiatric disorders. Biol. Psychiatry Cogn. Neurosci. Neuroimaging **1**(5), 433–447 (2016)
15. Menger, V., Scheepers, F., van Wijk, L.M., Spruit, M.: DEDUCE: a pattern matching method for automatic de-identification of Dutch medical text. Telematics Inform. **35**(4), 727–736 (2018)
16. Menger, V., Spruit, M., de Bruin, J., Kelder, T., Scheepers, F.: Supporting reuse of EHR data in healthcare organizations: the CARED research infrastructure framework. In: Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (2019)
17. Menger, V., Spruit, M., Hagoort, K., Scheepers, F.: Transitioning to a data driven mental health practice: collaborative expert sessions for knowledge and hypothesis finding. Comput. Math. Methods Med. **2016**, 1–11 (2016)
18. Olino, T.M., Klein, D.N., Lewinsohn, P.M., Rohde, P., Seeley, J.R.: Latent trajectory classes of depressive and anxiety disorders from adolescence to adulthood: descriptions of classes and associations with risk factors. Compr. Psychiatry **51**(3), 224–235 (2010)
19. Shen, J., Lee, P.H., Holden, J., Shatkay, H.: Using cluster ensemble and validation to identify subtypes of pervasive developmental disorders. In: AMIA Annual Symposium Proceedings, vol. 2007, pp. 666–670 (2007)
20. Spruit, M., Jagesar, R.: Power to the people! - Meta-algorithmic modelling in applied data science. In: Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (2016)
21. Strehl, A., Ghosh, J.: Cluster ensembles – a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. **3**, 583–617 (2003)
22. Topchy, A., Jain, A.K., Punch, W.: Combining multiple weak clusterings. In: Proceedings of the Third IEEE International Conference on Data Mining, p. 331 (2003)

23. Topchy, A.P., Law, M.H.C., Jain, A.K., Fred, A.L.: Analysis of consensus partition in cluster ensemble. In: Proceedings of the Fourth IEEE International Conference on Data Mining, ICDM 2004, pp. 225–232 (2004)
24. Vega-Pons, S., Ruiz-Shulcloper, J.: A survey of clustering ensemble algorithms. Int. J. Pattern Recognit Artif Intell. **25**(03), 337–372 (2011)