Informatics for Health: Connected Citizen-Led Wellness and Population Health
R. Randell et al. (Eds.)

doi:10.3233/978-1-61499-753-5-101

Exploratory Clustering for Patient Subpopulation Discovery

Dragan GAMBERGER[a,1], Bernard ŽENKO[b], Nada LAVRAČ[b,c], and for the Alzheimer's Disease Neuroimaging Initiative[2]
[a] *Rudjer Bošković Institute, Croatia*
[b] *Jožef Stefan Institute, Slovenia*
[c] *University of Nova Gorica, Slovenia*

Abstract. Exploratory Clustering is a novel general purpose clustering tool which is especially appropriate for medical domains in which we need to identify subpopulations that are similar in two different data layers. The tool implements the multi-layer clustering algorithm in a framework that enables iterative experiments by the user in his search for relevant patient subpopulations. A unique property of the tool is integration of clustering and feature selection algorithms. Differences in values of most relevant attributes are used to demonstrate decisive properties of constructed clusters. Usefulness of the tool is illustrated on a task of discovering groups of patients with similar cognitive impairment.

Keywords. Data clustering, biomarkers, Alzheimer's disease

1. Introduction

In this work we present a novel publicly available web application for data clustering, which is useful for detection of relevant subpopulations that are similar in two different data layers at the same time. A typical application domain is medicine where, for example the first layer comprises biological or genetic data while the second layer comprises clinical data. Detection of subpopulations homogeneous in these two layers is relevant for understanding relations between biological and clinical variables and for biomarker identification. If the objective of data analysis is medical prognosis, then the first layer can consist of baseline patient information while the second layer can contain corresponding longitudinal data. A nice property of this approach is that if the resulting clusters are homogeneous at the same time in different data layers, the quality of clustering increases (e.g., in multi-view clustering [1] and redescription mining [2]).

It is known that objective evaluation of the quality of clustering is practically impossible [3]. For the same data different solutions are possible and selection of the optimal one depends on human understanding of the data analysis problem, meaning that

[1] Corresponding author: Dragan Gamberger, Rudjer Bošković Institute, Bijenička 54, 10 000 Zagreb, Croatia; E-mail: dragan.gamberger@irb.hr.

[2] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

human expert knowledge is essential for high quality clustering. Our goal when developing the Exploratory Clustering tool was to design an extremely simple tool that medical researchers will be able to use by themselves. This should make it easier to generate medically and scientifically relevant data analysis results.

Section 2 presents the basic concepts underlying the implemented tool, Section 3 describes the data upload page, while Section 4 presents and discusses the results for a small set of patients with cognitive impairment extracted from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [4].

2. System Description

Exploratory Clustering is a web application, therefore the user does not need to download and install any software. Instead he uses the web browser to upload data to the computing server, to interactively guide the analysis process and to get the results of the analysis.

The analysis with the Exploratory Clustering tool is an iterative process. The user uploads data and receives a result that is optimal according to the implemented clustering algorithm. In the next step the user can ask for the refinements of the current solution. The refinements can go in two directions. Either the user can ask for modification of the current solution by increasing or decreasing the size of the constructed clusters or he can ask for a new clustering solution from a different subset of input data. The process can be iterated many times, enabling the user to employ his expert preferences in order to select the optimal clustering result from a large set of potentially good solutions. It must be noted that the user selects the direction in which the refinements should be executed, while the clustering algorithm determines how each refinement is actually implemented. This ensures that results of all iterations reflect relations existing in the data and present potentially good solutions.

Exploratory Clustering combines clustering and feature selection algorithms. Integration of feature selection into the clustering process is important because it enables detection and elimination of irrelevant variables, making it possible to cluster also high dimensional data where instances are described by many variables (attributes). Additionally, this approach enables detection of variables that are most responsible for the current clustering result. By showing these variables to the user and especially by computing and presenting their average values (or mode values for categorical variables) for each cluster, the user can better understand the meaning of the constructed clusters and significance of differences among them. Specifically for exploratory clustering this information is of ultimate relevance for the user because it is the basis for selecting the optimal solution.

The tool is based on the multi-layer clustering methodology described in [5, 6]. We decided to use this methodology because it enables both single and two layer clustering and because it can work with correlated layers (e.g., in multi-view clustering correlations between views are not allowed [1]). The second property is important especially for medical applications. In contrast to most other clustering tools [7], the multi-layer clustering algorithm determines the number of clusters and their optimal size automatically, thus users do not have to adjust any parameters of the clustering algorithm. In the final result some or even many instances may remain unclustered. In this way the constructed clusters correspond to sets of similar instances, while other instances remain unclustered. In some cases unclustered instances may be interpreted as outliers.

3. Exploratory Clustering Web Application

Exploratory Clustering tool is available at `http://rr.irb.hr/exploC/`. Because of the limited space we are not able to include the screen-shot of the data upload page but the reader can check it on the web. The page also has the link to instructions for data preparation, which include two tutorials describing the tool and its application.

In its basic form the Exploratory Clustering can be used as a standard clustering tool for data sets with up to 1,000 instances and up to 1,000 attributes. In this case it is only necessary to specify a data file for layer 1. Optionally, the user can upload also a file with the names of attributes, a file with the names of instances, and a file with some known classification of examples. Upload of optional files does not affect the clustering result but it can increase the understandability of the results that are presented to the user.

For two-layer clustering the user has to prepare and upload also the data file for layer 2. If biological data are uploaded in layer 1 then layer 2 is for clinical data or if baseline data are in layer 1 then longitudinal data are in layer 2. The second layer can include also up to 1,000 attributes. Optionally, the user can upload also the names of attributes in the second data layer.

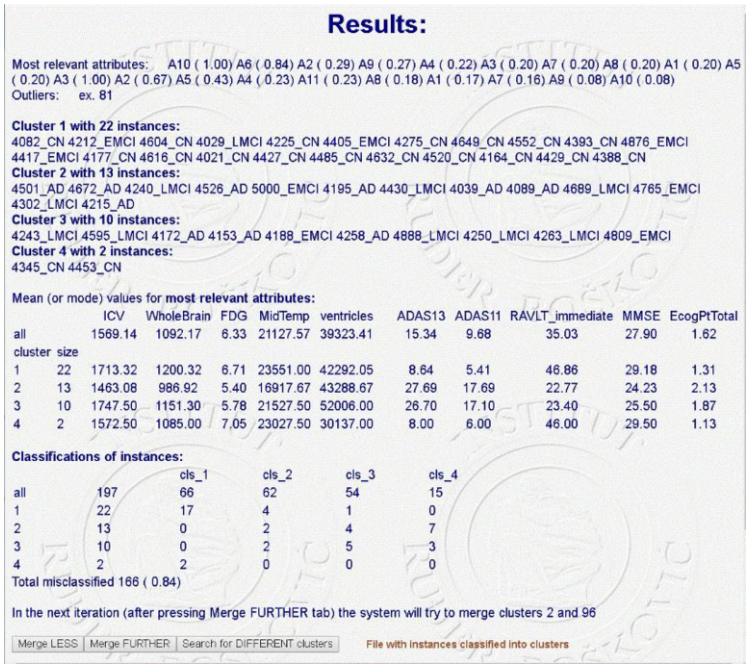
The user does not have to specify any parameters but can select increased reliability of the results. Increased reliability means execution of more iterations for computation of the similarity of instances [5, 6]. With this option the computation takes more time and its use is not recommended for data sets with more than 500 instances.

4. Illustrative Example

A data set of 197 male patients that have problems with dementia in used to illustrate the use of the tool. The data set is a subset of patients from the ADNI database [4] for which extensive clustering experiments have been performed and already reported in [5, 6]. In the first layer are 15 biological measurements like ABETA peptides, TAU and PTAU proteins, and MRI volumetric data together with 41 laboratory variables like number of red blood cells and total bilirubin values. In the second layer are 147 clinical variables like Alzheimer's Disease Assessment Scale (ADAS13) and Mini Mental State Examination (MMSE) score together with 40 symptoms like nausea and dizziness.

Besides biological and clinical data we also upload attribute names for both layers, names of examples and classification of examples according to the medical diagnosis that is not used as input data for clustering. Names of examples are a combination of the patient's RID number and the medical diagnosis that can be CN (cognitive normal), EMCI (early mild cognitive impairment), LMCI (late mild cognitive impairment) or AD (Alzheimer's disease). Classification of instances is in four classes so that patients with diagnosis CN are in class 1 while AD patients are class 4.

Figure 1 illustrates clustering results obtained on the described data set. The central part of the report is the list of constructed clusters. Each cluster is represented by a list of included instances. In this case the solution consists of four clusters with a total of 47 instances. The result demonstrates a high non-homogeneity of input instances with 150 out of 197 instances remaining unclustered. If the user is not satisfied by such weak clustering result he can iteratively press the tab "Merge FURTHER" at the bottom of the web page. In this way he can get even a solution with all 197 instances in only 2 clusters.



Results:

Most relevant attributes: A10 (1.00) A6 (0.84) A2 (0.29) A9 (0.27) A4 (0.22) A3 (0.20) A7 (0.20) A8 (0.20) A1 (0.20) A5 (0.20) A3 (1.00) A2 (0.67) A5 (0.43) A4 (0.23) A11 (0.23) A8 (0.18) A1 (0.17) A7 (0.16) A9 (0.08) A10 (0.08)
Outliers: ex. 81

Cluster 1 with 22 instances:
4082_CN 4212_EMCI 4604_CN 4029_LMCI 4225_CN 4405_EMCI 4275_CN 4649_CN 4552_CN 4393_CN 4876_EMCI 4417_EMCI 4177_CN 4616_CN 4021_CN 4427_CN 4485_CN 4632_CN 4520_CN 4164_CN 4429_CN 4388_CN
Cluster 2 with 13 instances:
4501_AD 4672_AD 4240_LMCI 4526_AD 5000_EMCI 4195_AD 4430_LMCI 4039_AD 4089_AD 4689_LMCI 4765_EMCI 4302_LMCI 4215_AD
Cluster 3 with 10 instances:
4243_LMCI 4595_LMCI 4172_AD 4153_AD 4188_EMCI 4258_AD 4888_LMCI 4250_LMCI 4263_LMCI 4809_EMCI
Cluster 4 with 2 instances:
4345_CN 4453_CN

Mean (or mode) values for **most relevant attributes:**

		ICV	WholeBrain	FDG	MidTemp	ventricles	ADAS13	ADAS11	RAVLT_immediate	MMSE	EcogPtTotal
all		1569.14	1092.17	6.33	21127.57	39323.41	15.34	9.68	35.03	27.90	1.62
cluster	size										
1	22	1713.32	1200.32	6.71	23551.00	42292.05	8.64	5.41	46.86	29.18	1.31
2	13	1463.08	986.92	5.40	16917.67	43288.67	27.69	17.69	22.77	24.23	2.13
3	10	1747.50	1151.30	5.78	21527.50	52006.00	26.70	17.10	23.40	25.50	1.87
4	2	1572.50	1085.00	7.05	23027.50	30137.00	8.00	6.00	46.00	29.50	1.13

Classifications of instances:

		cls_1	cls_2	cls_3	cls_4
all	197	66	62	54	15
1	22	17	4	1	0
2	13	0	2	4	7
3	10	0	2	5	3
4	2	2	0	0	0

Total misclassified 166 (0.84)

In the next iteration (after pressing Merge FURTHER tab) the system will try to merge clusters 2 and 96

Merge LESS | Merge FURTHER | Search for DIFFERENT clusters File with instances classified into clusters

Figure 1. Clustering result for 197 male patients with cognitive problems.

By checking the names of instances included into clusters we can conclude that constructed clusters are pretty consistent in respect of the diagnosis. In clusters 1 and 4 are mostly CN patients, in cluster 2 are mostly AD patients while in cluster 3 are mostly LMCI patients. Because we have prepared the file in which different diagnoses are coded with values 1-4 we have enabled generation of the classification report at the bottom of the web page. From this report it is easy to assess the consistency of clusters. For example, we see that in the largest cluster with 22 instances there are 17 CN patients, 4 EMCI patients and 1 LMCI patient. There are 166 misclassified instances, which corresponds to the sum of the number of unclustered instances and the number of minority class instances in all the clusters.

For expert evaluation the most interesting part of the report is the list of 5 most relevant attributes from each layer. For these 10 attributes the tool computes their average values for all 197 instances and then its average value for every constructed cluster. Large differences between reported values mean that the tool has been successful in detecting clusters that are substantially different. For example, for attribute ADAS13 the average value for all instances is 15.34 while average values for clusters 1 and 4 are about 8 and for clusters 2 and 3 the average values are about 27. But the data may reveal also some unexpected properties of constructed clusters. For example, for attribute ICV we have the average value for all instances 1,569, for cluster 2 with majority of AD patients we have substantially lower value 1,463 while for cluster 3 with majority of LMCI patients we have a substantially increased value equal to 1,747. In contrast, for attribute ventricles all clusters 1-3 have values higher than the average value for the complete population with highest value being 52,006 for cluster 3. This information can be very interesting for expert evaluation and for the user's decision if the constructed clusters are relevant.

5. Conclusion

To the best of our knowledge the Exploratory Clustering is the only clustering tool available as a web application, the tool that besides clusters of instances themselves also presents characterization of the constructed clusters, and the only tool that enables effective search for optimal solution over a set of different potentially good solutions. A simple user interface and parameter free clustering algorithm are additional advantages of the tool. Integration of feature selection into the clustering algorithm enables that in contrast to many other clustering algorithms that have a problem with the curse of dimensionality this tool can be used also for data sets with a large number of non-informative variables.

A serious drawback is time complexity of the tool, which is growing fast with the number of instances. An additional problem, especially when the data set has many variables, is that the refined solutions can only be slight modifications of the current solution and the user has to go through many iterations in order to get substantially novel clusters.

Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency core research programme *Knowledge Technologies* (P2-0103) and project *HinLife: Analysis of Heterogeneous Information Networks for Knowledge Discovery in Life Sciences* (J7-7303); the European Commission's support through *The Human Brain* project (FET Flagship grant FP7-ICT-604102), *MAESTRA* project (Gr. no. 612944), and *InnoMol* project (Gr. no. 316289); support of the Croatian Science Foundation project *Machine Learning Algorithms for Insightful Analysis of Complex Data Structures* (Gr. no. 9623).

References

[1] S. Sun, A survey of multi-view machine learning, *Neural Computing and Applications* **23** (2013), 2031-2038.
[2] L. Parada and N. Ramakrishnan, Redescription mining: structure theory and algorithms, *Proceedings of the association for the advancement of artificial intelligence* AAAI '05 (2005), 837–844.
[3] U. von Luxburg, R.C. Williamson and I. Guyon, Clustering: Science or art?, *In Guyon, I., Dror, G., Lemaire, V., Taylor, G. W., and Silver, D. L. (eds.), ICML Unsupervised and Transfer Learning* **27** (2012), 65–79.
[4] M.W. Weiner et al., The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception, *Alzheimer's Dementia* **8** (2012), S1-68.
[5] D. Gamberger, B. Ženko, A. Mitelpunkt and N. Lavrač, Homogeneous clusters of Alzheimer's disease patient population, *Biomedical Engineering Online* **15** (2016) S78.
[6] D. Gamberger, B. Ženko, A. Mitelpunkt, N. Shachar and N. Lavrač, Clusters of male and female Alzheimer's disease patients in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, *Brain Informatics* **3** (2016), 169-179.
[7] G. Gan, C. Ma and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, SIAM Philadelphia, 2007.