Clustering Socio-demographic and Medical Attribute Data in Cohort Studies

Paul Klemm[1], Lisa Frauenstein[1], David Perlich[1], Katrin Hegenscheid[2], Henry Völzke[2], Bernhard Preim[1]

[1]Otto-von-Guericke University Magdeburg, Germany
[2]Ernst-Moritz-Arndt University Greifswald, Germany
klemm@isg.cs.uni-magdeburg.de

Abstract. Longitudinal epidemiological studies like the Study of Health in Pomerania (SHIP) analyze a group of thousands of subjects (a cohort) by imposing a multitude of socio-demographic and biological factors. Epidemiological findings rest upon hypotheses which yield a selection of disease-specific cohort study parameters. They are then analyzed for significant interactions to identify risk factors. We propose an alternative approach by incorporating clustering algorithms with a Visual Analytics system to form subject groups which are the basis for an exploratory analysis of the underlying parameter interactions. We investigated three clustering techniques (k-Prototypes, DBSCAN and hierarchical clustering) for their suitability in these data sets. With our system, groups can be automatically determined to provide insights into this complex data.

1 Introduction

Epidemiological long-term cohort studies like the Study of Health in Pomerania [1] comprise a large range of sociodemographic and medical attributes of thousands of individuals to assess disease-specific risk factors. Knowledge about disease-influencing factors may affect its prevention, diagnosis and treatment. The resulting heterogeneous data space comprises several hundred variables gathered with different epidemiological instruments like interviews, clinical examinations and medical imaging. Risk factors are assessed in a sequential hypothesis driven way by domain experts. Hypotheses formulation follows observations in clinical routine. To validate the hypotheses, an attribute list is compiled and analyzed using regression analysis to check for statistical plausibility [2].

Visual Analytics methods integrate data analysis and visual exploration for analyzing huge data spaces to generate and validate hypothesis. An important technique for exploring heterogeneous data sets (data of different type) using Visual Analytics approaches is brushing over the subjects' attributes, which allows for subject grouping. We aim to enhance the exploratory data analysis approach by automatically generating subject groups using clustering algorithms. Our work is based on a data set that was compiled to analyze lower (*lumbar*) back pain. Privacy protection of the participants is ensured by anonymization of their personal data. Shape-based clustering of medical image data using spine detection models was carried out before [3]. Attributes of the lower spine canal, like curvature, torsion and vertebra positions, were extracted from the spine detection [4] to enhance the data set with shape-related parameters. Deterministic cluster results are a major requirement to ensure statistical resilience. Clustering subjects aims to reveal undiscovered correlations. Our main contributions are:

– Assessing three clustering methods (k-Prototypes, DBSCAN and hierarchical agglomerative clustering) for their suitability in cohort studies.
– Incorporating the clustering methods in a web-based Visual Analytics framework for browsing cohort study data.

2 Materials and Methods

In this section, we describe the epidemiological data set we used, followed by a brief overview of the incorporated clustering methods.

2.1 The Spine Data Set

A list of 77 attributes for 2333 subjects was compiled by domain experts at the University of Greifswald to analyze back pain. A finite element method was used to detect the lumbar spine in the MRI scans [4]. Curvature and torsion as well as the vertebra positions of the lumbar spine canal were extracted using the detection model. The data set is heterogeneous in terms of data types. A majority of 62 attributes are ordinal, i.e., results of multiple choice questionnaires related to lifestyle factors and medical background like back pain history. Body size measurements and parameters derived from the image data are covered from 17 scalar variables. A challenge when analyzing cohort studies are subjects with missing values for some attributes. Reasons for incomplete data range from medical/ethical to personal issues. The clustering workflow must account for missing data.

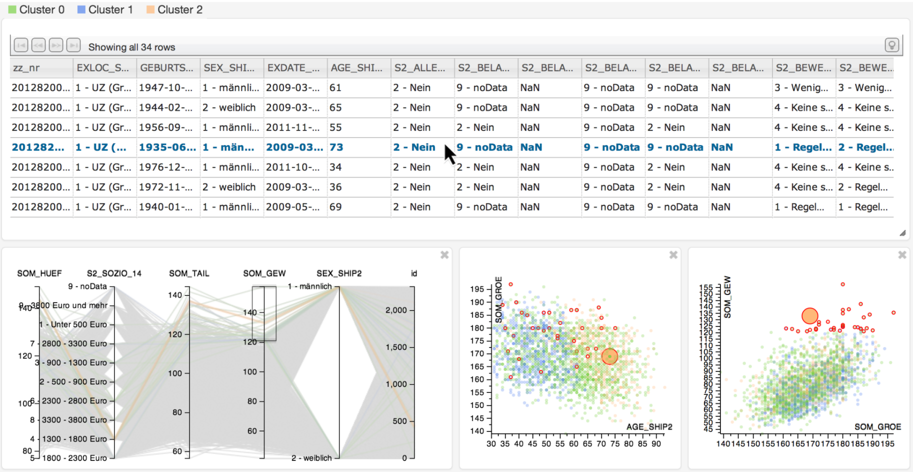
2.2 Clustering Workflow

The clustering methods are embedded into a in-house developed Visual Analytics system, which comprises different views for ordinal and metric variables (Fig. 1). To trigger clustering, the user selects either all parameters of a data set or a subset from a list. Due to missing values, the system immediately displays the number of subjects that are omitted in the clustering step given the current attribute selection. The selection of the clustering method and its parameters closes this process, which returns computed groups that are rendered as seen in Fig. 1.

2.3 Clustering Methods

Clustering methods divide the space spanned by data elements so that it maximizes the distance between groups and minimizes the within-groups variance.

Fig. 1. Embedding a clustering result within the Visual Analytics framework. A k-Prototypes clustering with $k = 3$ results in three color-coded clusters. All subjects with body weight above 120 *kg* are brushed using parallel coordinates and highlighted in the scatterplots with red circles. One subject of cluster two is selected in the list view, which increases its opacity in the parallel coordinates and its radius in the scatterplots.



🟩 Cluster 0 🟦 Cluster 1 🟧 Cluster 2

Showing all 34 rows

zz_nr	EXLOC_S...	GEBURTS...	SEX_SHI...	EXDATE_...	AGE_SHI...	S2_ALLE...	S2_BELA...	S2_BELA...	S2_BELA...	S2_BELA...	S2_BELA...	S2_BEWE...	S2_BEWE...
20128200...	1 - UZ (Gr...	1947-10-...	1 - männli...	2009-03-...	61	2 - Nein	9 - noData	NaN	9 - noData	9 - noData	NaN	3 - Wenig...	3 - Wenig...
20128200...	1 - UZ (Gr...	1944-02-...	2 - weiblich	2009-03-...	65	2 - Nein	9 - noData	NaN	9 - noData	9 - noData	NaN	4 - Keine s...	4 - Keine s..
20128200...	1 - UZ (Gr...	1956-09-...	1 - männli...	2011-11-...	55	2 - Nein	2 - Nein	NaN	9 - noData	2 - Nein	NaN	4 - Keine s...	4 - Keine s..
201282...	**1 - UZ (...**	**1935-06...**	**1 - män...**	**2009-03...**	**73**	**2 - Nein**	**9 - noData**	**NaN**	**9 - noData**	**9 - noData**	**NaN**	**1 - Regel...**	**2 - Regel..**
20128200...	1 - UZ (Gr...	1976-12-...	1 - männli...	2011-10-...	34	2 - Nein	2 - Nein	NaN	9 - noData	2 - Nein	NaN	4 - Keine s...	4 - Keine s..
20128200...	1 - UZ (Gr...	1972-11-...	2 - weiblich	2009-03-...	36	2 - Nein	2 - Nein	NaN	9 - noData	2 - Nein	NaN	4 - Keine s...	2 - Regel...
20128200...	1 - UZ (Gr...	1940-01-...	1 - männli...	2009-05-...	69	2 - Nein	9 - noData	NaN	9 - noData	9 - noData	NaN	1 - Regel...	1 - Regel...

Measurement of Distance. Clustering heterogeneous data attributes at the same time requires distance measurements that consider different data types [5]. We calculated the similarity between numerical attributes using the Euclidean distance. Ordinal attribute values are compared in a binary fashion, having distance *0* when they are identical and distance *1* otherwise. The factor γ can be used to weight elements [5]. We applied three different clustering techniques.

k-Means and k-Prototypes. Dividing the data into k clusters using randomly generated centroids, each data point is iteratively attached to its closest centroid. K-Prototypes [6] enhances k-Means to allow for the clustering of ordinal and scalar attributes using the previously described weighted distance. The random initialization of centroids renders the k-Prototypes clustering results non-deterministic. This is not suitable for epidemiological applications where reproducibility of all results is required [2]. Therefore, the initial centroid positions are computed by placing centroids near values that are close to each other.

*DBSCAN. D*ensity-*B*ased *S*patial *C*lustering of *A*pplications with *N*oise computes clusters based on object density. Elements are density-connected when they are reachable by a chain of dense objects. Density-connected elements form a cluster. Outliers are objects that are not associated to a cluster via density. DBSCAN is steered by parameters, which define the distance between neighbors (ϵ) and the number of neighbors that a "dense" element must comprise ($minPts$). The method is independent of a predefined cluster number and accounts for outliers.

Hierarchical Agglomerative Clustering. The stepwise aggregation of the closest elements into a cluster yields a dendrogram whose levels represent clusters. By varying the minimum similarity, the desired number of clusters is obtained. The method is known to be outlier-prone.

3 Results

The difficulty of comparing cluster results in this application domain is twofold. First, we cannot measure the accuracy of a result due to missing ground truth. Second, the presented clustering methods have different parameters, which have a strong impact on their results. We tried to minimize the difference in the results by focussing on the same numerical and categorical parameters.

Table 1. Dice's coefficients for clustering results of k-Prototypes and DBSCAN.

Cluster Number	Algorithms	Dice's Coefficient
2	$k - Prototypes/DBSCAN$ ($\epsilon = 1.3$)	0.634
	$k - Prototypes/DBSCAN$ ($\epsilon = 1.4$)	0.655
	$k - Prototypes/DBSCAN$ ($\epsilon = 1.5$)	0.657
3	$k - Prototypes/DBSCAN$ ($\epsilon = 0.9$)	0.720
	$k - Prototypes/DBSCAN$ ($\epsilon = 1.1$)	0.644
	$k - Prototypes/DBSCAN$ ($\epsilon = 1.2$)	0.646
6	$k - Prototypes/DBSCAN$ ($\epsilon = 1.0$)	0.406

K-Prototypes was tested in a range of two to ten clusters. The cluster sizes range from 94 to 487 subjects. No overly large or small clusters are computed.

DBSCAN's parameter *minPts* equals the minimum cluster size. Since epidemiologists are interested in larger groups of subjects, this value needs to be fairly high. Ester and colleagues argue that the impact of *minPts* is little above a certain threshold [7]. We set this value empirically to 50, which produces size-balanced clusters. Parameter ϵ defines the size of an object's neighborhood. Set low, ϵ leads to many small outlier clusters, which we want to avoid–an ϵ value between 0.6 and 0.8 classifies 1602 subjects as outliers and is therefore not reasonable. Parameter ϵ set to 0.9 to 1.2 results in balanced clusters.

Hierarchical Agglomerative Clustering creates very unbalanced trees for our data. Many clusters only contain one element. Complete-Linkage produced the best results in terms of cluster size, but still yields one large cluster containing almost all subjects. Hence, this method was discarded for use on our data.

3.1 Comparison using Dice's Coefficient

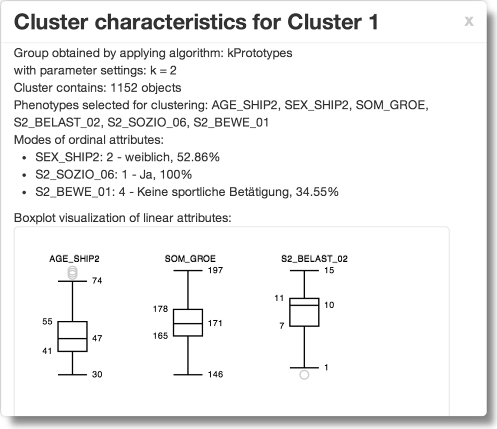
We used Dice's coefficient to compare the clustering results under use of different parameters [8]. It is defined as $\frac{2(A \bigcap B)}{|A|+|B|}$, where A and B are the clusters to compare and $A \bigcap B$ is the amount of elements in A and B. Dice's coefficient is 0 for disjunct and 1 for identical clusters. Since the hierarchical agglomerative clustering results are not plausible, we only compared k-Prototypes and DBSCAN. The results for clusters with size 2, 3 and 6 for DBSCAN with corresponding k-Prototypes results can be found in Table 1. While Dice's coefficient for 2 to 3 clusters is close to 0.65, it is only at 0.4 for 6 clusters. Cluster results are similar while there is a decreasing similarity for an increasing cluster number. This reflects the missing ground truth problem–these results are only an expression of similarity, not plausibility. The latter can only be determined in the context of epidemiological reasoning whether the groups represent meaningful correlations.

3.2 Visualization of Clustering Results

Enhancing the Visual Analytics framework by clustering capabilities for automatic grouping was a key motivation for this work. Each group is rendered using a different color and can therefore be differentiated in the linked plots (Fig. 1). We introduce an additional information window, which contains statistical information associated to each cluster (Fig. 2).

4 Discussion

We presented three methods for clustering epidemiological cohort study data to compute groups that capture data interactions. Linked to Visual Analytics systems, these methods provide an alternative way of gaining new insight into the complex interactions in these high-dimensional data sets. We found k-Prototypes and DBSCAN to be appropriate for our data. Hierarchical agglomerative clustering produced unbalanced cluster trees, yielding huge clusters containing almost all subjects and is therefore not suitable for our research. The clustering results are strongly dependent on the chosen variable types and the distance measure. Future extensions comprise better cluster group comparison to amplify hypothesis generation by highlighting influential parameters. Usability would benefit



Cluster characteristics for Cluster 1

Group obtained by applying algorithm: kPrototypes
with parameter settings: k = 2
Cluster contains: 1152 objects
Phenotypes selected for clustering: AGE_SHIP2, SEX_SHIP2, SOM_GROE, S2_BELAST_02, S2_SOZIO_06, S2_BEWE_01
Modes of ordinal attributes:

- SEX_SHIP2: 2 - weiblich, 52.86%
- S2_SOZIO_06: 1 - Ja, 100%
- S2_BEWE_01: 4 - Keine sportliche Betätigung, 34.55%

Boxplot visualization of linear attributes:

Fig. 2. Information window for a clustering resulting from the k-Prototypes algorithm. The clustering parameters yield a reproducible clustering result. The distribution of metric parameters in the cluster is displayed using box plots. The most frequent value of each ordinal parameter is displayed using percentage statements.

from automatic parameter designation using quality criteria. Missing data can be tackled with imputation [9]. For k-Prototypes, k could be derived by a knee function that plots the cluster number to a cluster quality measurement [10].

At the end, it falls to the user to validate the data for plausibility. A clustering-based automated grouping step can only highlight certain dependencies in the data set. It is no alternative to the classic epidemiological workflow, but rather an enhancement of the available tools, providing a different point of view. By combining both worlds, the huge cohort study data sets can be made tangible.

Acknowledgements: SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grant no. 03ZIK012), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania. This work was supported by the DFG Priority Program 1335: Scalable Visual Analytics.

References

1. Völzke H, Alte D, Schmidt C, et al. Cohort Profile: The Study of Health in Pomerania. International Journal of Epidemiology. 2011 Mar;40(2):294–307.
2. Thew S, Sutcliffe A, Procter R, de Bruijn O, McNaught J, Venters CC, et al. Requirements Engineering for e-Science: Experiences in Epidemiology. Software, IEEE. 2009;26(1):80–87.
3. Klemm P, Lawonn K, Rak M, Preim B, Tönnies KD, Hegenscheid K, et al. Visualization and Analysis of Lumbar Spine Canal Variability in Cohort Study Data. In: Proc. of International Workshop on Vision, Modeling and Visualization; 2013. p. 121–128.
4. Rak M, Engel K, Tönnies KD. Closed-Form Hierarchical Finite Element Models for Part-Based Object Detection. In: Proc. of International Workshop on Vision, Modeling and Visualization; 2013. p. 137–144.
5. Huang Z. Clustering large data sets with mixed numeric and categorical values. In: Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining; 1997. p. 21–34.
6. Huang Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery. 1998;2(3):283–304.
7. Ester M, Kriegel HP, Sander J, Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proc. of the Second International Conference on Knowledge Discovery and Data Mining.; 1996. p. 226–231.
8. Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology. 1945;26(3):297–302.
9. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle Introduction to Imputation of Missing Values. Journal of Clinical Epidemiology. 2006;59(10):1087 – 1091.
10. Salvador S, Chan P. Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms. In: Proc. of Tools with Artificial Intelligence; 2004. p. 576 – 584.