EuroVis Workshop on Visual Analytics (2017)
M. Sedlmair and C. Tominski (Editors)

Subpopulation Discovery and Validation in Epidemiological Data

S. Alemzadeh[1] , T. Hielscher[2], U. Niemann[1], L. Cibulski[1],T. Ittermann[3], H. Völzke[3], M. Spiliopoulou[2], B. Preim[1]

[1]Department of Simulation and Graphics, Otto-von-Guericke University Magdeburg, Germany
[2]Department of Technical and Business Information Systems, Otto-von-Guericke University Magdeburg, Germany
[3]University Medicine Greifswald, Germany

Abstract

Motivated by identifying subpopulations that share common characteristics (e.g. alcohol consumption) to explain risk factors of diseases in cohort study data, we used subspace clustering to discover such subpopulations. In this paper, we describe our interactive coordinated multiple view system Visual Analytics framework S-ADVIsED for SubpopulAtion Discovery and Validation In Epidemiological Data. S-ADVIsED enables epidemiologists to explore and validate findings derived from subspace clustering. We investigated the replication of a selected subpopulation in an independent population.

Categories and Subject Descriptors (according to ACM CCS): I.3.8 [Computer Graphics]: Applications—

1. Introduction

Epidemiologists investigate the factors which contribute to the outbreak of diseases. Thus, they identify risk factors related to life style, genetic predisposition, socio-demographic factors, and environmental factors as well as protective factors that reduce the likelihood of getting a disease [Woo13]. With the increasing amount of cohort study data, the traditional hypothesis-driven and statistics-focused approaches usually fail to identify subpopulations that have a risk for a specific disease which strongly deviates from the global mean [Obe04].

Subspace clustering and subgroup discovery are methods for the identification of subpopulations which share determinant factors. Identified patterns are expressed in the form of interpretable rules. For instance, a significant subpopulation could be phrased as "While in the study population only 18 % exhibit goiter, in the subpopulation described by $BMI > 30.5\,kg/m^2 \wedge TSH \leq 1.5\,mU/l$ it is 52 %." Each condition in the rule antecedent corresponds to an axis-parallel hyperplane in the attribute space. Thus, subgroup discovery algorithms return descriptions of subspaces that are limited to an axis-parallel, hyper-rectangular shape. Subspace clustering seeks for clusters in any subset of dimensions. Due to the complexity of the clusters' shapes, subspace clusters need to be transformed to hyper-rectangles such that they can be described as rules. Visual representations are essential to enable the user to explore subspace clustering results and to steer the process of transformation of subpopulations in detail and develop trust in the results. The latter is essential, since epidemiologists in general are skeptical to data mining results that may produce a very large amount of unreliable findings. In our collaboration with epidemiologists we noticed their need for replication and validation of data mining findings.

Our proposed S-ADVIsED framework combines visualization techniques and data mining concepts for discovery and validation of subspace clusters. S-ADVIsED allows to interactively explore subpopulations based on the user preferences. Our contributions in this paper include:

- Visual support for the identification of subpopulations in cohort study data
- Validating the findings in a second, independent cohort
- Exploration and comparison of subpopulations

2. Epidemiological Background

In this section, we describe cohort study data and terms used in this paper to address epidemiologists' requirements. Epidemiology research focused on the determinant and distributions in a specific population [Woo13].
Epidemiology data contain information acquired by interview (e.g. sociodemographic, medication use), physical examination (e.g. measuring blood pressure and BMI), laboratory tests (e.g. diabetes and TSH) and medical images (e.g. MR images) [PKH*16]. Combining these features leads to a heterogeneous, high-dimensional and large data set. Our analyses are based on the Study of Health in Pomerania (SHIP) [Völ12]. The study was performed in different waves, SHIP-0 (from 1997 to 2001), SHIP-1 (from 2002 to 2006) and SHIP-2 (from 2008 to 2012). Since this cohort gets older and smaller (e.g. due to persons moving to other regions), a new cohort the SHIP-TREND, was established in parallel with the SHIP-2 study. In this work, we focus on the fatty liver as a widespread disorder. Participants with a liver fat concentration of more than 10 % are considered as positive for fatty liver.

Due to the continuously increasing number of dimensions and heterogeneity of cohort study data, important associations might be

DOI: 10.2312/eurova.20171118

overlooked. Furthermore, the goal of epidemiologists is not only to assess the global effect of a determinant, but also to find groups of study participants which are similar w.r.t. common protective and risk factors.

3. Subspace Clustering

The efficiency of traditional clustering methods is hampered in high-dimensional feature space caused by an effect that is referred to as curse of dimensionality [PHL04]. Subspace clustering algorithms can overcome this issue by automatically discovering relevant subspaces and performing a clustering within these subspaces.

Constraint-based clustering. For the discovery of subspace clusters we use the constraint-based subspace clustering algorithm *DRESS* (*D*iscovery of *R*elevant *E*xample-constrained *S*ub*S*paces [HS*16]). The method incorporates expert knowledge on the similarity of study participants (must-link and not-link constraints) to find clusters in subspaces that satisfy these constraints. Constraints should reflect the similarity of participants w.r.t. a medical outcome. To find clusters, *DRESS* starts with a quality scoring of each subspace of cardinality one. Initially, these subspaces constitute the candidate set of subspaces. Subspace quality is scored by considering the distance between must-link and not-link constrained participants as well as the portion of satisfied constraints to all constraints in the respective subspace. To satisfy a must-link / not-link constraint, both constrained participants have to lie within the same / different cluster. From here, *DRESS* iteratively picks the best scored subspace S_{can} and merges it with all remaining subspaces in the candidate set. To reduce complexity only the resulting subspaces that satisfy a filter criterion are considered. For subspaces that satisfy this criterion, the full quality is calculated, which involves a density-based clustering with DBSCAN [EKSX96] where parameters are estimated [NHS*15]. When the quality of a subspace exceeds the highest yet observed quality q_{best}, *DRESS* retains it as a candidate subspace for further extension, updates q_{best} and stores all contained clusters. At the end of an iteration, S_{can} and all merge candidates that led to a new q_{best} are removed from the candidate set. *DRESS* terminates when the candidate set is empty and returns a ranking of subspaces and their associated clusters.

4. Related Work

We describe related works for both the analysis of cohort study data and the visualization of subspace clusters.

Visual Analytics on Cohort Study Data. Zhang et al. proposed an interactive visual analytics tool to analyze cohort populations [ZGP14]. Cohort Analysis via Visual Analytics (CAVA) comprises three main parts: The *Cohort*, *views* and *analytics* for analyzing cohorts via interaction with user. Krause et al. [KPS16] provided an interactive framework for Supporting Iterative Cohort Construction With Visual Temporal Queries (COQUITO). COQUITO enables medical researchers to explore the dataset by iterative queries via constraints. Klemm et al. [KOJL*14] presented a visual analytics system to identify subpopulations on the basis of data interactions using three global clustering on shape parameters characterizing the spinal canal to better understand backpain using SHIP data. Klemm et al. [KLG*16] enabled epidemiologists to enter regression formula and search for dimension combinations related to an outcome, e.g. increased breast density. With heat maps indicating strong correlations users are guided towards potentially relevant factors.

Visual Analysis of Subspace Clusters. Assent et al. [AKMS07], Tatu et al. [TZB*12] and Hund et al. in [HBS*16] presented visualization techniques to show the similarity between subspace clusters and to illustrate their distribution.

5. Visual Analytics Support for Epidemiological Analysis

Epidemiologists mostly rely on statistical methods and simple visualizations. Data mining methods are useful to them when they can trust their findings. A visual analytics system where data mining is not just a black box combined with explicit support for validation is essential to support epidemiologists.

Here, we propose S-ADVIsED as a web-based visual analytics framework using d3.js library [BOH11] that combines several visualization methods for discovery, validation and comparison of subpopulations. The screenshot of S-ADVIsED is shown in Fig. 1.

5.1. Requirements

Epidemiologists need to assess the quality of subspace clusters based on their intended measurements. To evaluate discovered subpopulations, we have to consider the following requirements for the visualization of subspace clusters [HFB*16]:

Dimensionality: Epidemiologists are more interested in low-dimensional subspace clusters to avoid overfitting. Knowledge derived from subspace clustering should be transferred to clinical practice, i.e. contribute to prevention, diagnosis and treatment of diseases.

Cluster Size: The number of participants in each subpopulation should be sufficiently large to support evidence of statistical significance.

In-depth Information: A clear and compact visualization showing the distributions of both involved and non-involved dimensions is essential.

Cluster Compactness: Participants who belong to one subspace cluster should be similar to each other with respect to their involved dimensions.

Object and Dimension Redundancy: It is necessary to indicate the overlap proportion of participants and dimensions in different subpopulations.

Comparison with Global Mean: It is crucial to compare different subpopulations with the whole population. As an example, epidemiologists are interested in investigating subpopulations that differ strongly from other subpopulations w.r.t. a specified attribute.

Dimension Variability: Subspace clustering algorithms minimize the sparsity of data by ignoring dimensions with higher variance. It might be interesting for epidemiologists to pursue the reason for incorporating a high variance dimension in a cluster.

5.2. Exploration and Pattern Discovery

The S-ADVIsED framework provides different well-known charts and an overall view of subspace clusters to fulfill the epidemiologists' requirements.

Global Overview. In the global view we have an overview of all subspace clusters and the main characteristics of subspace clusters: We illustrate subspace clusters by donut charts, since they have a simple representation and we are able to encode enough information in them to show different specifications of subpopulations (Fig. 2). In the encoding of subspace clusters, sectors

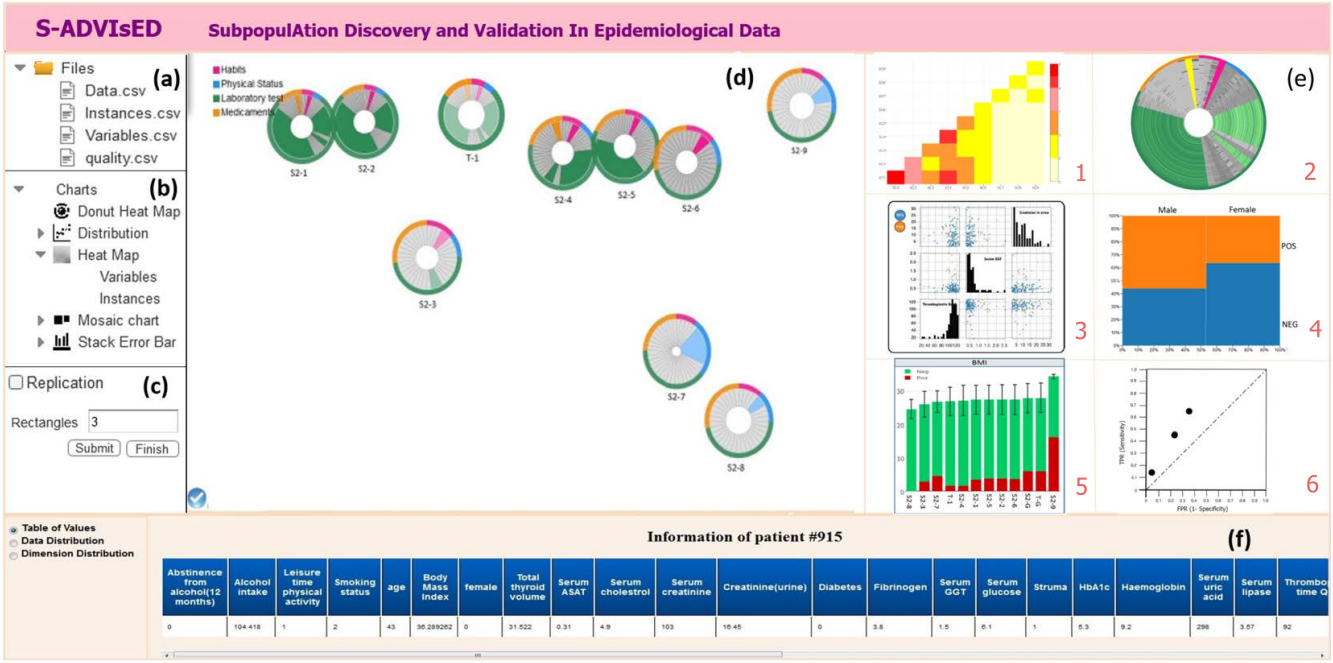

Figure 1: *User interface of S-ADVIsED: (a) tree view of input files, (b) charts panel juxtaposes pairwise dimension distributions of the selected subspace cluster by scatterplots (both numeric), mosaic charts (both categorical), and stacked errorbars (mixed numerical / categorical) (c) replication settings, (d) global view of subspace clusters, (e) in-depth analysis, (f) statistical information.*

stand for dimensions and their size depicts their variability based on the variance. The donut chart's radius size depicts the cluster size. A bigger radius means that it contains more objects. The colored sectors represent involved dimensions in subspace clustering results and the grays are non-involved ones. Linking and brushing techniques were implemented to show dimension overlaps. By clicking on each dimension the corresponding dimension in other subspace clusters will be highlighted. We provide a categorization of dimensions based on the suggestions of epidemiologists. Here, we propose four categories of dimensions: (1) *Habits* (e.g. smoking status), (2) *Physical condition* (e.g. BMI), (3) *Laboratory tests* (e.g. serum GGT concentration) and (4) *Medicaments* (e.g. amlidipine intake). We define the distance between the subspace clusters based on the shared dimensions and participants, as in [AKMS07]. To illustrate the distance (similarity) of subspace clusters a multi-dimensional scaling (MDS) is employed to project the clusters in 2D space. MDS is frequently used for evaluating clustering [AKMS07, HBS*16].

Charts and Validation. As illustrated in Fig. 1 (e), this part is used for analysis of subspace clusters and it is accessible via the chart panel. After finding any interesting pattern in one subspace cluster these findings should be validated through replication by expert users.

Charts. To show a compact overview of each subspace cluster, we provide a donut heat map (Fig. 1 (e)(2)). In the donut heat map, sectors stand for dimensions and rings represent individuals. Dimensions that do not contribute to one subspace cluster have a gray scale coloring, and involved dimensions are mapped to colors. Darker colors depict greater values; in contrast, brighter ones stands for smaller values. An optional sorting based on the variability is applied. Thus, participants with smaller values regarding the dimension with highest variance have a smaller radius. Well-known visualization techniques are used for the analysis of subspace clusters. Mosaic charts (Fig. 1 (e)(4)) are used to show the relationship of different nominal attributes. The user can dynamically select any two categorical dimensions (e.g. diabetes and fatty liver). Heat maps (Fig. 1 (e)(1)) enable the epidemiologist to identify the two clusters/ subpopulations that share the most dimensions or participants. The distribution of numeric dimensions is shown via a discretized scatterplot matrix (Fig. 1 (e)(3)) , equipped with histograms of each dimension in the main diagonal. Stacked errorbars

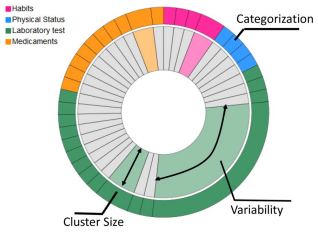

Figure 2: *Each subspace cluster is illustrated by a donut chart. Coloring is based on the categorization of dimensions and the radius of the donut depicts the cluster size.*

(Fig. 1 (e)(5)) compare different subspace clusters with the global mean based on any selected numeric dimension in the whole data set.

Validation. Subspace clustering algorithms usually produce lots of subspace clusters. Additionally, subspace clusters may have arbitrary shape and subpopulations need to be defined as intervals in the form of hyper-rectangles. One way to validate the subspace clusters is replication. This means, if a specified subspace cluster can be reproduced in an independent population, it is a relevant subpopulation. Subpopulations are considered similar if they differ similarly from the global mean regarding to a specific dimension. Therefore, S-ADVIsED lets epidemiologists adjust the shape of the selected cluster using a scatterplot matrix.

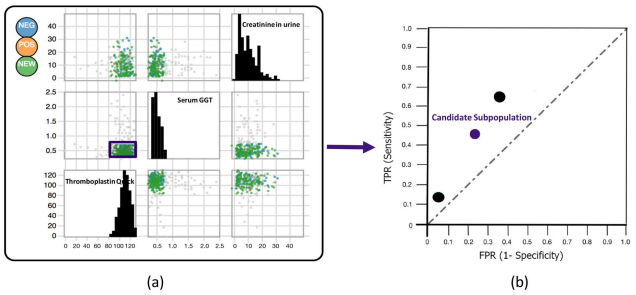

Figure 3: *Each submitted rectangle is a candidate to form a new subpopulation and transform the shape of the selected subspace cluster.*

6. Use Cases

In this section, we describe two uses cases. The first use case explains how analysts explore subspace clusters to select a subspace cluster for validation. The second use case demonstrates how the analyst checks reproducibility of selected subspace cluster. In all steps, expert users have an overall view of all subspace clusters (Fig. 1 (d)).

Exploration and Replication. To start, the expert user may be interested to see the similarity between subspace clusters separately with respect to the shared dimensions or participants. So, she can select the heat map from the charts panel (Fig. 1 (b)(1)). As next step, by zoom-in and tooltips (to see involved dimensions) she selects a subspace cluster. For example, we selected S2-1 with three numeric dimensions related to blood examinations (thromboplastin time Quick, serum GGT, creatinine in urine) and two nominal dimensions related to life style and medication use (smoking status, enalapril), whereas all participants are ex-smokers. Next, by selecting the donut heat map from the chart panel the user can get an overview of the selected subspace cluster in a compact view via in-depth analysis panel. The user can interactively click on a participant (ring) and see the table of values in the footer. Moreover, all subspace clusters that share this participant will be highlighted in the global view panel. For the next step, based on the type of information, the user can select a chart and determine its parameter flexibly, i.e. in (Fig. 1 (e)) part 3, the user selected a scatterplot matrix with selecting fatty liver from the sub-menu as discretization parameter. By activating the replication (Fig. 1 (c)), the validation phase based on the specified parameter for the number of selected ranges (rectangles) by the expert user will be started. The user can see combined participants of the selected subspace cluster with the SHIP-TREND population. A scatterplot matrix is provided, defined by the involved numeric dimensions. As shown in Fig. 3 (a), orange points are positive and blue ones are negative fatty liver participants from SHIP-2 data. The green points are participants from SHIP-TREND. For the next step, the user can define the desired ranges for dimensions by drawing a rectangle in one pair of dimensions. While the user is drawing and expanding a rectangle, she can see highlighted corresponding individuals who are located inside the rectangle w.r.t. other pairs of involved dimensions, see (Fig. 3 (a)). Next, the labels of SHIP-TREND participants are predicted based on the drawn rectangle. To predict labels, 1-nearest neighbor classification is applied. The user is enabled to draw multiple rectangles in different pairs of dimensions with distinct positions and diameter. Each drawn rectangle is a candidate to transform the selected subspace cluster and to form a new subpopulation with SHIP-TREND objects within specified intervals. As next step, the epidemiologist should define ranges in terms of boundaries and a distribution regarding the outcome of fatty liver. To do this, the tool displays Receiver Operating Characteristic (ROC) curves (Fig. 3 (b)(6)). A ROC curve shows the relationship between true positive rate (TPR or sensitivity) and false positive rate (FPR or 1-specificity). The TPR measures the fraction of correctly classified diseased (positive) study participants. At the final step, the new subpopulation is generated and integrated to the ovarall view. The selected subspace cluster will be transformed to the defined rectangular range.

Check reproducibility. In the following, the reproducibility of the selected subpopulation is investigated. Different measures are specified by epidemiologists to check the reproducibility of the intended subpopulation [Cib16]. One measure is distribution. By selecting the scatterplot matrix and mosaic chart, the analyst can check the distributions of both subpopulations regarding to the target dimension. Involved dimensions in both subpopulations must be the same. As we just consider involved dimensions, we have the same number of dimensions in both subpopulations. The analyst can see involved dimensions by the linking and brushing technique in the global view. The size of both subpopulations should be in the same range. It is achievable by comparing the radius size of subpopulations and for more detail by the bar charts in the footer. In our case, after the transformation phase S2-1 and T-1 have 95 and 104 participants, respectively. Subpopulations are considered as replicated if they deviate similarly from the global mean. Thus, the sorted stackbar chart shows the mean value of the whole population and subpopulation based on involved dimensions.

7. Conclusion & Future Work

We presented S-ADVIsED as a web-based visualization framework for the discovery of subpopulations in cohort study data. The design of the system was based on site visits at the epidemiology department and is largely based on ideas of epidemiologists, e.g. for transforming clustering results in subpopulations and validating such subpopulations. We intend to develop a method that maximizes the product of sensitivity and specificity delivering recommended hyper-rectangular approximation of a subpopulation which subsequently can be adjusted based on the epidemiologists' suggestions.

References

[AKMS07] ASSENT I., KRIEGER R., MÜLLER E., SEIDL T.: Visa: Visual subspace clustering analysis. *SIGKDD Explor. Newsl. 9*, 2 (Dec. 2007), 5–12. 2, 3

[BOH11] BOSTOCK M., OGIEVETSKY V., HEER J.: D^3 data-driven documents. *IEEE transactions on visualization and computer graphics 17*, 12 (2011), 2301–2309. 2

[Cib16] CIBULSKI L.: Visual analytics support for analysis of cohort study data: Requirements and concepts. *Project report, Otto-Von-Guericke University Magdeburg* (2016). 4

[EKSX96] ESTER M., KRIEGEL H.-P., SANDER J., XU X.: A density-based algorithm for discovering clusters in large spatial databases with noise. AAAI Press, pp. 226–231. 2

[HBS*16] HUND M., BÖHM D., STURM W., SEDLMAIR M., SCHRECK T., ULLRICH T., KEIM D. A., MAJNARIC L., HOLZINGER A.: Visual analytics for concept exploration in subspaces of patient groups. *Brain Informatics* (2016), 1–15. 2, 3

[HFB*16] HUND M., FÄRBER I., BEHRISCH M., TATU A., SCHRECK T., KEIM D. A., SEIDL T.: Visual quality assessment of subspace clusterings. *Workshop on Interactive Data Exploration and Analytics* (2016). 2

[HS*16] HIELSCHER T., SPILIOPOULOU M., ET AL.: Identifying relevant features for a multi-factorial disorder with constraint-based subspace clustering. In *Computer-Based Medical Systems (CBMS), IEEE 29th International Symposium on* (2016), pp. 207–212. 2

[KLG*16] KLEMM P., LAWONN K., GLASSER S., NIEMANN U., HEGENSCHEID K., VÖLZKE H., PREIM B.: 3d regression heat map analysis of population study data. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 81–90. 2

[KOJL*14] KLEMM P., OELTZE-JAFRA S., LAWONN K., HEGENSCHEID K., VÖLZKE H., PREIM B.: Interactive visual analysis of image-centric cohort study data. *IEEE transactions on visualization and computer graphics 20*, 12 (2014), 1673–1682. 2

[KPS16] KRAUSE J., PERER A., STAVROPOULOS H.: Supporting iterative cohort construction with visual temporal queries. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 91–100. 2

[NHS*15] NIEMANN U., HIELSCHER T., SPILIOPOULOU M., VÖLZKE H., KÜHN J. P.: Can we classify the participants of a longitudinal epidemiological study from their previous evolution? In *Proc. of IEEE Symposium on Computer-Based Medical Systems* (2015), pp. 121–126. 2

[Obe04] OBENSHAIN M. K.: Application of data mining techniques to healthcare data. *Infection Control & Hospital Epidemiology 25*, 08 (2004), 690–695. 1

[PHL04] PARSONS L., HAQUE E., LIU H.: Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter 6*, 1 (2004), 90–105. 2

[PKH*16] PREIM B., KLEMM P., HAUSER H., HEGENSCHEID K., OELTZE S., TOENNIES K., VÖLZKE H.: Visual analytics of image-centric cohort studies in epidemiology. In *Visualization in Medicine and Life Sciences III*. Springer, 2016, pp. 221–248. 1

[TZB*12] TATU A., ZHANG L., BERTINI E., SCHRECK T., KEIM D., BREMM S., VON LANDESBERGER T.: Clustnails: Visual analysis of subspace clusters. *Tsinghua Science and Technology 17*, 4 (2012), 419–428. 2

[Völ12] VÖLZKE H.: Study of health in pomerania (ship). *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz 55*, 6-7 (2012), 790–794. 1

[Woo13] WOODWARD M.: *Epidemiology: study design and data analysis*. CRC press, 2013. 1

[ZGP14] ZHANG Z., GOTZ D., PERER A.: Iterative cohort analysis and exploration. *Information Visualization* (2014), 289–307. 2