Analysis of Patient Groups and Immunization Results Based on Subspace Clustering

Michael Hund[1], Werner Sturm[2], Tobias Schreck[2], Torsten Ullrich[3], Daniel Keim[1], Ljiljana Majnaric[4], and Andreas Holzinger[5,6]

[1] Data Analysis and Visualization Group, University of Konstanz, Germany
[2] Institute for Computer Graphics and Knowledge Visualization, Graz University of Technology, Austria
[3] Fraunhofer Austria Reseach GmbH, Austria
[4] Faculty of Medicine, JJ Strossmayer University of Osijek, Croatia
[5] CBmed - Center for Biomarker Research in Medicine, Graz, Austria
[6] Research Unit HCI-KDD, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Austria

Abstract. Biomedical experts are increasingly confronted with what is often called *Big Data*, an important subclass of high-dimensional data. High-dimensional data analysis can be helpful in finding relationships between records and dimensions. However, due to data complexity, experts are decreasingly capable of dealing with increasingly complex data. Mapping higher dimensional data to a smaller number of relevant dimensions is a big challenge due to the *curse of dimensionality.* Irrelevant, redundant, and conflicting dimensions affect the effectiveness and efficiency of analysis. Furthermore, the possible mappings from high- to low-dimensional spaces are ambiguous. For example, the similarity between patients may change by considering different combinations of relevant dimensions (subspaces). We show the potential of subspace analysis for the interpretation of high-dimensional medical data. Specifically, we analyze relationships between patients, sets of patient attributes, and outcomes of a vaccination treatment by means of a subspace clustering approach. We present an analysis workflow and discuss future directions for high-dimensional (medical) data analysis and visual exploration.

Keywords: Knowledge Discovery and Exploration, Subspace Clustering, Subspace Analysis, Subspace Classification, Classification Explanation

1 Introduction

Today, experts in Life Sciences are not only confronted with large amount of data, but particularly with high-dimensional data e.g., by the trend towards personalized medicine [1]. A big challenge of biomedical informatics research is to gain knowledge from these complex high-dimensional data sets [2]. Within such data, relevant *structural* and/or *temporal* patterns ("knowledge") are often hidden and not accessible to the expert. While automatic data analysis can provide candidate patterns for user exploration, it is not always clear which analysis methods are suitable for a given problem. Often, methods which consider the full data space are applied. However, these may fail to deliver useful results due to the *curse of dimensionality* [3]. We present a case study on the applicability of full- and subspace-based analysis methods on a real-world immunization data set. We present a potentially effective analysis workflow, which can help to understand the relationship of clusters of patients in context of attribute similarities and outcomes of an immunization treatment. We also provide a discussion of limitations and possible extensions to subspace analysis in this domain.

2 Related Work

We briefly survey related work in the area of clustering including subspace methods and interactive data exploration.

Cluster Analysis. Cluster analysis is a widely known tool to reduce large data sets to a smaller number of clusters, which can be compared with each other and in relation to some target attribute of interest [4]. Traditional clustering approaches such as *k-means* or *hierarchical clustering* [4] take all dimensions into account. However, it has been shown that for many dimensions the so-called *curse of dimensionality* may prevent effective cluster analysis, as the similarity measure may become less discriminant [5, 3]. To this end, subspace cluster algorithms search for clusters not in the whole data space, but within different subsets of dimensions (called *subspaces*) in which discriminating clusters can be found [6].

Interactive Data Exploration. Data analysis algorithms typically require parameters to be set, and often, multiple solutions need to be considered before arriving at a satisfactory result. To this end, methods of interactive and visual exploration of the data and the analysis outputs can be very helpful. Specifically, many visualization techniques have been developed for exploration of high-dimensional data and clusterings. For example, Parallel Coordinate Plots [7] map high-dimensional data to Polylines, allowing the user to discern groups in data and potentially relevant relationships, effective for moderate numbers of dimensions. Another standard approach is to reduce data dimensionality and show relationships of data points by their positions in a data projection [8]. In [9], users could compare data clusterings with constituent data dimensions.

The latter approaches have in common to consider all input data dimensions at once. In other work, visualizations to explore clusters in subspaces, by a combination of heatmap and glyph representations in the so-called ClustNails approach [10] were proposed. The system was applicable to any subspace clustering approach. In [11], 2D projections of the data in alternative subspaces were applied, to identify complementary, orthogonal or redundant subspaces; again, the approach was applicable to different subspace selection methods. Another system to rely on subspace cluster comparison is VISA [12], which implement a simple glyph alternative to represent and compare subspace clusters. In [13], visual comparison of data groups across dimensions using linked views in an encompassing system was presented.

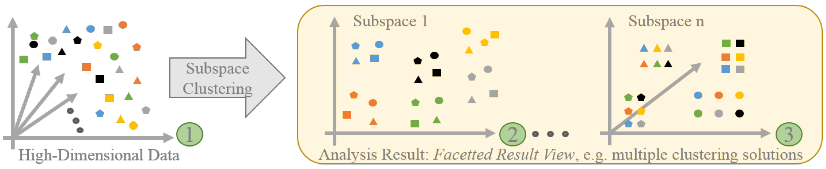

Subspace Clustering
High-Dimensional Data
1
Subspace 1
Subspace n
2
3
Analysis Result: *Facetted Result View*, e.g. multiple clustering solutions

Fig. 1. Subspace clustering: algorithms compute multiple, alternative solution in different subspaces, i.e. clustering by color (subspace 1) or by shape (subspace 2).

3 Data Analysis with Subspace Clustering

As discussed in Section 2, subspace clustering can be a promising tool for analysis of high-dimensional data with respect to multiple different groups in data and their relationships to dimensions. Its main idea is illustrated in Fig. 1. The goal is to understand data in terms of (a) groups of similar records (clusters), and (b) the underlying dimensions (subspaces). As an outcome of subspace cluster analysis applied on a high-dimensional data ①, depending on the parametrization and/or subspace cluster method, clusterings in multiple *different subspaces* may be found, constituting different notions of similarity; e.g. grouping according to color ② or shape ③. Each subspace cluster may give rise to a different interpretation. Depending on the underlying algorithm, we can state that all cluster members are similar to each other w.r.t. the dimensions of the subspace. The main difference to *feature selection* [14] is that subspace analysis aims for different patterns in different subspaces while feature selection typically determine a single subspace to optimize a quality criterion such as the classification error.

For our experiments in Section 4.3, we rely on a subspace clustering approach called *Proclus* (Projected Clustering) [15]. Proclus is similar to *k-means* [4] as it generates, by an iterative process, a *partition* of the data. Each data point can belong to one cluster, and each cluster is represented by a prototype point (medoid). Proclus needs two parameters: the number of clusters C and the average dimensionality per subspace D. The subspace computation starts by a random initialization of medoids. In a refinement step, for each of the C medoids a well-fitting subspace of average dimensionality D is found. This is achieved by finding dimensions that show a low variance of the distances between the respective medoid and its cluster members. The resulting subspace contains dimensions in which the values of the cluster members are similar. While other subspace clustering methods are available [6], we chose Proclus for its simplicity, efficiency, and robustness to noise, using the *OpenSubspace Framework* [16] implementation.

4 Use Case: Explanations for Vaccination Outcomes

We study the potential of a subspace clustering-based analysis on a real-world medical analysis problem. We introduce a relevant dataset from a clinical research, describe our analysis goals, present results of initial experiments, and interpret them from the domain perspective.

4.1 Considered Data Set and Analysis Goal

Data Set Used. The examined data set is based on a real patient data which describes volunteers vaccinated against influenza. Patients were selected to represents a high-risk population for influenza complications. All subjects were suffering of multiple (age-related) chronic medical conditions which interfere with the immune system. The investigated group of subjects consists of 35 male and 58 female persons aged between 50 and 89 years. The data set contains 61 dimensions describing clinical parameters such as sex, age, anthropometric measures, hematological, and biochemical tests. In addition to that, dimensions containing diagnosis results of common chronic diseases are included. Finally, a single target attribute representing the positive or negative vaccination outcome (36 positive, 57 negative) is included. Further details about the dataset and the underlying influenza vaccination can be found in [17].

Analysis Perspectives. According to the domain expert (medical physician and researcher) who created the dataset, the reasons for a positive or negative vaccination outcome can neither be described by a single dimension, nor by a fixed combination of dimensions. Instead, a variety of different reasons may cause the positive or negative outcome. In the remainder of this paper, we analyze the above described dataset by means of a subspace clustering based method in order to discover multiple reasons for different outcomes. Our idea is to apply a subspace clustering algorithm to the dataset in order to find similarities of patients of the same outcome class. The subspace dimensions can be interpreted as possible explanations for an outcome.

Data Preprocessing. As shown above, the considered dataset is heterogeneous as it contains both numerical and nominal dimensions. Existing subspace clustering algorithms typically work on numerical data only. Also, for existing implementations there is no description how missing values are treated. As a consequence, we preprocessed the dataset in the following way: (1) We removed all patient records that have a missing value in any of its dimension. Afterwards the resulting dataset contains 29 patients with a positive and 42 patients with a negative outcome. (2) We transformed all nominal dimensions such as *sex*, *hypert*, or *statins* into a numerical representation. Due to the fact that all nominal dimension (except for diabetes mellitus (DM)) consist of only two different values (mainly *yes* and *no*), we converted the values to either 0 or 1. Finally, we normalized all dimensions linearly in the range $[0, 1]$. After this, all dimensions are numerical in the range of $[0, 1]$, enabling further analysis with equally weighted dimensions.

4.2 Experiments in Full-Space Analysis

In our initial experiments on the dataset, we found that a full-space analysis is not useful. We used data mining tools such as KNIME [18] to cluster patients into different groups, or applied different classification algorithms to correctly predict the vaccination outcome of a patient.

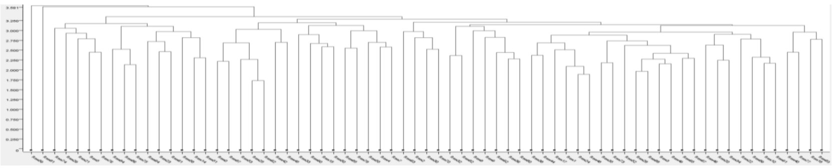


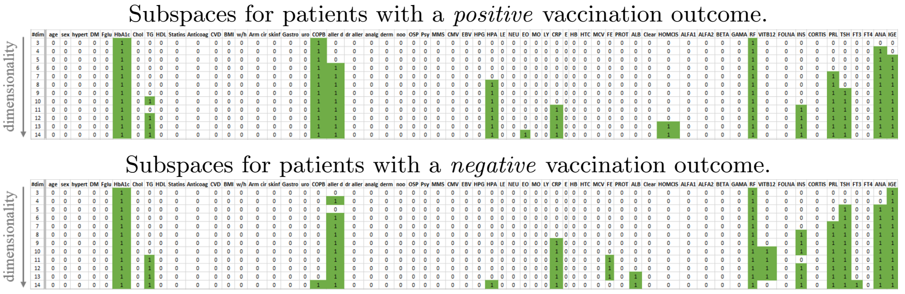
Fig. 2. Dendrogram illustrating the hierarchical clustering of our dataset (*Euclidean distance* and *average linkage* type). The x-axis represents the individual patients, while the y-axis indicate the dissimilarity between two patients or patients with clusters.

Full-Space Clustering: A hierarchical clustering was applied. The results are illustrated as a Dendrogram in Fig. 2. The x-axis is mapped to the individual patients, the y-axis represents the dissimilarity between two patients or a patient cluster. A large y-value corresponds to a high dissimilarity. From Fig. 2 we see that none of the patients are considered similar and, as a consequence, no useful grouping of patients can be identified. We assume the reasons as: (1) Patients are typically similar to each other only in a subset of dimensions, (2) a similarity in one dimension can be countered by a dissimilarity in another dimension, and (3) the *concentration effect* [5] affects the similarity computation in high-dimensional spaces.

Full-Space Classification: For the classification task, did not remove missing values but rather replace them by the average value of the dimension. We applied several classification algorithms to find useful predictors for the vaccination outcomes. Our experiments comprised e.g., *Decision-Trees*, *Bayes Classification*, and *Random Forest*. We split the dataset into a training set (80% of the records) and a validation set (20% of the records). For the validation, we measured the percentage of correctly classified patients after the model training. While the accuracy of the classification of the training dataset performs very well (approx. 84% for decision tree), the accuracy for the validation dataset dropped below 50% for some algorithm; which is worse than random classification. We assume the poor classification performance is caused by (1) the size of the training dataset which is too small, and (2) there are no global aspects allowing a classification into the two outcome classes. Instead, different combinations of features may be of relevance to predict the outcome properly.

4.3 Subspace Analysis: Initial Experiments and Results

To search for local explanations of the vaccination outcome, subspace analysis techniques are beneficial. In the following, we describe three different experiments that we conducted. The experiments apply the subspace clustering algorithm *Proclus* to different subsets of the data. We interpret the discovered subspaces as a mean to describe the similarity between a subset of patients and, as a consequence, as possible reasons for a vaccination outcome. Supplementary material of the experiments and the attribute description can be found on our website (`http://files.dbvis.de/bih2015/`).



Subspaces for patients with a *positive* vaccination outcome.

#dim	age	sex	hypert	DM	Fglu	HbA1c	Chol	TG	HDL	Statins	Anticoag	CVD	BMI	w/h	Arm cir	skinf	Gastro	uro	COPB	aller d	dr aller	analg	derm	noo	OSP	Psy	MMS	CMV	EBV	HPG	HPA	LE	NEU	EO	MO	LY	CRP	E	HB	HTC	MCV	FE	PROT	ALB	Clear	HOMCIS	ALFA1	ALFA2	BETA	GAMA	RF	VITB12	FOLNA	INS	CORTIS	PRL	TSH	FT3	FT4	ANA	IGE
3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0
5	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1
6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1
7	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1
8	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1
9	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	1	1
10	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	1	1
11	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	1	1
12	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	1	1
13	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	1	1	0	0	1	1
14	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	1	1	0	0	1	1

Subspaces for patients with a *negative* vaccination outcome.

#dim	age	sex	hypert	DM	Fglu	HbA1c	Chol	TG	HDL	Statins	Anticoag	CVD	BMI	w/h	Arm cir	skinf	Gastro	uro	COPB	aller d	dr aller	analg	derm	noo	OSP	Psy	MMS	CMV	EBV	HPG	HPA	LE	NEU	EO	MO	LY	CRP	E	HB	HTC	MCV	FE	PROT	ALB	Clear	HOMCIS	ALFA1	ALFA2	BETA	GAMA	RF	VITB12	FOLNA	INS	CORTIS	PRL	TSH	FT3	FT4	ANA	IGE
3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
5	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	1
6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	1
7	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	1	1
8	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	1	1
9	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	1	1
10	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	1	1	0	0	1	1
11	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	0	0	1	1
12	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	1	0	1	1	0	0	1	1
13	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	0	1	1	0	0	1	1
14	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	1	1	1	0	1	1

Fig. 3. Subspaces detected by experiment 2: subspace clustering (*Proclus*) applied separately to patients with a positive, or negative outcome. The columns represent the different dimensions (green indicate that dimension belongs to subspace). Each row represents a clustering result of different dimensionality.

Experiment 1. In the first experiment, we apply Proclus to the preprocessed dataset and aim for subspace clusters that contain mainly patients of a single outcome. The dimensions of these clusters and the respective values of the cluster members are a means to describe the specific outcome. For this experiment, we vary Proclus' parameters (#clusters: 2-8; avg. #dimensions: 3-14). We evaluate each cluster with the *Entropy score* [16] which measures the purity of a cluster w.r.t. a specified class label. The supplementary material on our website provides an overview of the results. We can see that almost none of the detected clusters contain patients of only one specific class, but rather a mixture of both classes without a significant majority of a positive or negative outcome. We believe that this result is caused by (1) the computation strategy of Proclus which aims for large clusters, and (2) the dataset contains dimensions in which many patients are similar to each other - independent of their class label. These dimensions dominate the detected clusters and prevent Procus from finding clusters relevant for the description of the vaccination outcome (c.f. experiment 2).

Experiment 2. To find descriptive clusters for each vaccination outcome, we split the dataset into subsets according to the outcome class. Further analysis is applied to the individual subsets. In the first part, we configure *Proclus* to detect subspaces containing a *single cluster*. The dimensions of the subspace indicate global similarities of a class. For each subset the average number of dimensions varies between 3 and 14. The results can be found in Fig. 3. The different dimensions are indicated as columns while each row represents a subspace cluster with a different dimensionality. The cells of a row are marked with a green background, if the subspace contains the dimension. E.g., the first subspace for a positive outcome contains the dimensions: *HbA1c*, *COPB*, and *RF*.

Proclus determines the dimensions of a subspace cluster by ordering all dimensions by the variance of its cluster members, and selecting the dimensions with a minimum variance (c.f. Section 3). Therefore, subspaces with a larger dimensionality may include dimensions in which its cluster members are less similar. As all records belong to the same cluster, dimensions in lower-dimensional subspaces are more descriptive for an outcome class (w.r.t. global outcome similarity). Consequently, the height of the green bars in Fig. 3 illustrates the importance of a dimension for an outcome class. Except for *HPA* and *PRL*, the globally descriptive dimension are identical for both outcomes. This result is in-line with the detected subspaces of the first experiment (see supplementary material), i.e. the following set of dimensions is discriminative for all patients from a global perspective: *HbA1c*, *COPB*, *aller d*, *HPA*, *CRP*, *RF*, *INS*, *PRL*, *TSH*, *ANA*, and *IGE*. Most patients in our dataset are similar in these dimensions, however, we do not get much knowledge about the patients w.r.t. the vaccination outcome. This observation is confirmed by the second part of experiment 2. In addition to the first part, we also varied the #clusters between 2-4. The complete result can be found in the supplementary material. In summary, we can see that even for results with 4 clusters, the majority of dimensions is from the given set above. From the second experiment, we can conclude that subspace clustering helps to find dimensions in which patients of a specific class are similar to each other, hence these dimensions may be an indicator for the reason of the classification. However, experiment 2 shows, that dimensions in which most patients are similar to each other, highly influence the clustering results. As a consequence the subspaces for both outcome classes are similar to each other.

Experiment 3. In our last experiment, we concentrate on *more local patterns*. From the previous experiments, we know that all patients, and in particular all patients of one outcome class are similar to each other in the dimensions described above. To find more local patterns, we remove these dimensions from both subsets and re-apply Proclus. Afterwards, a heuristic is used to search for a result in which all patients are assigned to any subspace cluster, the cluster sizes are similar, and the number of dimensions is rather high. For both outcome classes, we selected a result with four different clusters and an average number of dimensions of 14. A table of the subspaces and the assigned cluster members for each outcome class can be found in the supplementary material. From these results, we can make the following observations: (1) The patients belonging to a subspace cluster are similar in all of the dimensions of the subspace; (2) For one outcome class, we found subspaces that differ significantly in their dimensions; (3) The relevant subspaces for a positive and a negative outcome class are different. We provided our results to the domain-expert who created the dataset. The expert liked the result very much and provided some insights into our findings:

Positive Vaccination Outcome : One subspace shows a group of patients that is homogeneous in all dimensions of the subspace. Based on the following dimension and its values, we can state that the group of patients is rather healthy, i.e. a positive vaccination outcome: The patients do not have hypertension, CVD, neoplasm, (attribute noo), psychiatric disorders and do not have adverse reaction to drugs (attribute dr aller). Furthermore, the patients do not use any of the following medications: statins, anticoagulants, or analgesics which results in preserved renal function (dimension CLEAR).

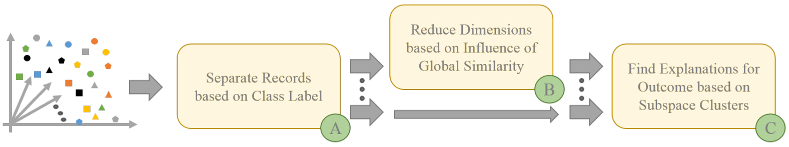

Separate Records based on Class Label
A
Reduce Dimensions based on Influence of Global Similarity
B
Find Explanations for Outcome based on Subspace Clusters
C

Fig. 4. Our proposed workflow to discover relations between patients, relevant dimensions and different class labels (here vaccination outcomes).

Negative Vaccination Outcome : One subspace show a clear reason for a negative vaccination outcome: although not having *DM*, adverse reactions on drugs (*dr aller*), not having increased *Fglu* values, not having anaemia (*E*, *HB*), patients can have negative vaccination outcoume due to impairment in some relevant pathophysiologic mechanisms, including slightly increased *MCV* (a sign of decreased *VITB12* and/or *FOLNA*), and decreased cortisol (*CORTIS*). Further examples for both outcome classes are described in the supplementary material.

4.4 Proposed Subspace Analysis Workflow

Based on our findings in the experiments described above, we propose a subspace clustering-based workflow (c.f. Fig. 4) to find relations between data records, dimensions, and associated class labels. The workflow consists of the main steps Ⓐ and Ⓒ as well as an optional step Ⓑ improving local similarity aspects.

The first step of the workflow is to separate all data records based on their class label Ⓐ. The subsequent steps are applied to each record subset individually. The optional step Ⓑ is in-line with the findings of the second experiment. In many datasets, there are dimensions that highly influence the detection of subspace clusters. On the one hand, these dimensions are interesting as they show the global similarity between data records. On the other hand, such dimensions can distort the results, e.g. a dataset with non-relevant dimensions in which all records are similar. Subspace clustering consider these dimensions as a relevant and add them to most clusters. In such a case, step Ⓑ can be applied to remove such dimensions. In Ⓒ, a subspace clustering is applied to the remaining dimensions to finally determine the similarities between records, dimensions, class labels.

5 Discussion

The explorative analysis of patient treatment data is a challenging task. As our experiments show, subspace clustering can be a valuable tool to discover relevant groups of patients w.r.t. different medical subspaces and their relationship to the treatment (here: vaccination outcome). As a key finding of our experiments, an analysis in the full attribute space may not be the best choice, but subspace methods can be an interesting tool, especially if used in an appropriate analysis workflow. We proposed one workflow, considered as promising starting point.

We also identify a number of extension possibilities to our approach. For one, we may need heuristic criteria which could select, from a large number of parameters (e.g., input dimensions, number of clusters, distance thresholds etc.) a small number of results which are not redundant but can be meaningfully interpreted. To this end, we need a formalization how to measure what alternative or complementary means in terms of dimensions, cluster size, and attribute subsets. We need to include additional medical background into such a specification. Visual interfaces may be particularly beneficial to this end. A key issue in visualization is how to effectively map patient records, cluster, and attribute properties to visual displays. Regarding data size, scalability of the cluster analysis may become an issue, which could be addressed by efficient implementations.

We considered Proclus which considers all dimensions of a subspace as equally important for the subspace. However, there may also exist non-linear relationships between attributes which might be relevant. Alternative analysis tools like non-linear multivariate regression could be considered to optimize attribute selection. Also on the preprocessing side, how to appropriately treat categorical and binary attributes in the analysis is a problem. We here chose standard approaches, but the expert may be needed to specify how to treat such attributes.

While often, analysis is handled by ad-hoc approaches, it would be desirable to have a software framework to allow a flexible, interactive specification of analysis workflows, to easily apply and re-use proven workflows. We imagine a workflow editor which could support the analysis process in a scalable way, and at the same time, allow experts to document which and why analysis steps were taken.

6 Conclusion and Future Outlook

The life sciences, biomedicine and health care are turning into a data intensive science, where we face not only increased volumes and a diversity of highly complex, multi-dimensional and often weakly-structured and noisy data, but also the growing need for integrative analysis and modeling [1]. Considering that analysis in the full attribute (feature) space may not be effective, we here explored subspace cluster analysis to study the relationship between patient data and immunization treatment outcome on a specific research data set. We found that a segmentation of the patients for treatment outcome followed by subspace clustering allowed to identify relevant patient groups and respective medical attributes, which can be a basis to generalize medical knowledge. Our proposed workflow is only a first step, and we identified a number of interesting challenges and extensions for future work in the area. The grand vision for the future is to effectively support human learning with machine learning - visualization is close to the end-user, hence indispensable within this approach.

Acknowledgment

We thank Ines Färber of RWTH Aachen for fruitful discussions on subspace clustering algorithms.

References

1. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge discovery and interactive data mining in bioinformatics - state-of-the-art, future challenges and research directions. BMC Bioinformatics **15** (2014) I1
2. Holzinger, A.: Biomedical Informatics: Discovering Knowledge in Big Data. Springer, New York (2014)
3. Hinneburg, A., Aggarwal, C.C., Keim, D.A.: What is the nearest neighbor in high dimensional spaces? In: Proc. Int. Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc. (2000) 506–515
4. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. 3rd edn. (Morgan Kaufmann Publishers Inc.)
5. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is "nearest neighbor" meaningful? In: Proc. Int. Conference on Database Theory. (1999) 217–235
6. Kriegel, H.P., Kröger, P., Zimek, A.: Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Transactions on Knowledge Discovery from Data (TKDD) **3** (2009) 1–58
7. Fua, Y.H., Ward, M., Rundensteiner, E.: Hierarchical parallel coordinates for exploration of large data sets. In: Proc. Conference on Visualization, IEEE CS Press (1999) 43–50
8. Buja, A., Swayne, D.F., Littman, M.L., Dean, N., Hofmann, H., Chen, L.: Data visualization with multidimensional scaling. Journal of Computational and Graphical Statistics **17** (2008) 444–472
9. Seo, J., Shneiderman, B.: Interactively exploring hierarchical clustering results. Computer **35** (2002) 80–86
10. 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** (2012) 419–428
11. Tatu, A., Maaß, F., Färber, I., Bertini, E., Schreck, T., Seidl, T., Keim, D.: Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In: Proc. IEEE Conf. Visual Analytics Science and Technology. (2012) 63–72
12. Assent, I., Krieger, R., Müller, E., Seidl, T.: Visa: visual subspace clustering analysis. SIGKDD Explor. Newsl. **9** (2007) 5–12
13. Turkay, C., Lex, A., Streit, M., Pfister, H., Hauser, H.: Characterizing cancer subtypes using dual analysis in caleydo StratomeX. IEEE Computer Graphics and Applications **34** (2014) 38–47
14. Liu, H., Motoda, H.: Computational Methods of Feature Selection. Chapman & Hall/CRC (2007)
15. Aggarwal, C., Procopiuc, C., Wolf, J., Yu, P., Park, J.: Fast algorithms for projected clustering. In: Proc. ACM Int. Conf. on Management of Data. (1999) 61–72
16. Müller, E., Günnemann, S., Assent, I., Seidl, T.: Evaluating clustering in subspace projections of high dimensional data. **2** (2009) 1270–1281
17. Trtica-Majnaric, L., Zekic-Susac, M., Sarlija, N., Vitale, B.: Prediction of influenza vaccination outcome by neural networks and logistic regression. Journal of biomedical informatics **43** (2010) 774–781
18. Berthold, M., Cebron, N., Dill, F., Gabriel, T., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B.: KNIME: The Konstanz Information Miner. In: Studies in Classification, Data Analysis, and Knowledge Organization, Springer (2007)