Dominic Girardi & Andreas Holzinger 2018. Dimensionality Reduction for Exploratory Data Analysis in Daily Medical Research. In: Giabbanelli, Philippe J., Mago, Vijay K. & Papageorgiou, Elpiniki I. (eds.) Advanced Data Analytics in Health. Cham: Springer International, pp. 3-20, doi:10.1007/978-3-319-77911-9_1.

Dimensionality reduction for exploratory data analysis in daily medical research

Dominic Giradi[1] & Andreas Holzinger[2]

[1] RISC Software GmbH, Research Unit Medical Informatics, Linz University, Austria
dominic.girardi@risc.uni-linz.ac.at
[2] Holzinger Group, HCI-KDD, Institute for Medical Informatics/Statistics,
Medical University Graz, Austria andreas.holzinger@medunigraz.at

Abstract. In contrast to traditional, industrial applications such as market basket analysis, the process of knowledge discovery in medical research is mostly performed by the medical domain experts themselves. This is mostly due to the high complexity of the research domain, which requires deep domain knowledge. At the same time, these domain experts face major obstacles in handling and analyzing their high-dimensional, heterogeneous, and complex research data. In this paper, we present a generic, ontology-centered data infrastructure for scientific research which actively supports the medical domain experts in data acquisition, processing and exploration. We focus on the system's capabilities to automatically perform dimensionality reduction algorithms on arbitrary high-dimensional data sets and allow the domain experts to visually explore their high-dimensional data of interest, without needing expert IT or specialized database knowledge.

Keywords: Dimensionality Reduction, Visual Analytics, Knowledge Discovery

1 Introduction

The classic process of knowledge discovery in data (KDD) is a well known and meanwhile widely accepted process in the field of computer science and very important to create the *context* for developing methods and tools needed to cope with the ever growing flood of data [1] and the complexity of data. Particularly in the medical area we face not only increased volume and a diversity of highly complex, multi-dimensional and often weakly-structured and noisy data, but the pressing need for integrative analysis and modeling of data [2], [3], [4].

Fayyad et al. (1996) [1] define the process of knowledge discovery '*as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*'. This process, which usually contains steps such as understanding the problem and data, data preparation, and data mining, is mostly performed by so called data scientists [5]. In this view the (biomedical) domain expert is rather put into a supervising, consulting and customer only role. In the HCI-KDD approach [6] this is different, as the domain expert takes on an active role and the goal is to integrate the expert directly into the machinery loop [7]. There is a need for easy-to-use data exploration system, mainly driven by the fact, that the analysis is done by domain experts, and not computer scientist [8].

It is very important to note that there is a huge difference between work-flows in biomedical research and clinical research, hence the KDD process in daily clinical research differs significantly from standard research work-flows. The role of the domain expert turns from a passive external supervisor - or customer - to an active actor of the process, which is necessary due to the complexity of the research domain [9]. However, these domain experts are now confronted with large amounts of highly complex, high-dimensional, heterogeneous, semi-structured, weakly-structured research data [10] of often poor data quality. The handling and processing of this data is known to be a major technical obstacle for (bio-)medical research projects [11]. However, it is not only the data handling that contains major obstacles, also the application of advanced data analysis and visualization methods is often only understandable for data scientists or IT experts. A survey from 2012 among hospitals from Germany, Switzerland, South Africa, Lithuania, and Albania [12] showed that only 29% of the medical personnel of responders were familiar with a practical application of data mining. Although this survey might not be representative globally, it clearly shows the trend that medical research is still widely based on standard statistical methods. One reason for the rather low acceptance rate is the relatively high technical obstacle that needs to be taken in order to apply often complex algorithms combined with the limited knowledge about the algorithms themselves and their output. Especially in the field of exploratory data analysis deep domain knowledge of the human expert is a crucial success factor.

In order to overcome some of the mentioned obstacles and to help to contribute to a concerted effort in dealing with increasing volumes of data, we present a generic, ontology-based data infrastructure for scientific research that supports the research domain expert in the knowledge discovery process; beginning with the definition of the data model, data acquisition and integration and acquisition, to validation and exploration. The idea is to foster a paradigm shift that moves the domain expert from the edge of the process to the central actor. The system is based on a generic meta data-model and is able to store the actual domain ontology (formal description of the research domain data structures) as well as the corresponding structured research data. The whole infrastructure is implemented at a higher level of abstraction and derives its manifestation and behavior from the actual domain ontology at run-time. The elaborated structural meta-information is used to reach a sufficiently high degree of automation in data processing, particularly for exploratory data analysis. This allows the domain experts to visualize more easily their complex and high dimensional research data. In this paper we present two different ontology-guided visualization methods: a parallel coordinate visualization and non-linear mapping visualization.

2 Related Research

The idea of using meta models to automatically create parts (data access layer, user interfaces) of data intensive software systems is a widely established method in model-driven engineering (MDE) [13], which is a promising approach to address platform complexity [14]. However, the MDE approach in general or concrete realizations such as the meta model-based approach for automatic user interface generation by da Cruz et al. (2010) [15] — just to provide an example — are used by software engineers to create source code or source code skeletons at development time. Our system derives the structure of the user interface from the meta model at run-time. There is no source code generation. From this perspective our system is related to the Margitte system by Renggli et al. (2007) [16]. Whilst the Margitte system is a general purpose framework, based on a self-describing meta-model, our system is based upon a meta-entity-relationship model - stored in a relational database - and clearly focused and specialized on scientific data acquisition and data processing considering the medical researcher as both a main user and administrator. There is a close relation to ontology-based systems: Zavaliy & Nikolski (2010) [17] describe the use of an ontology for data acquisition, motivated by the demand of adaptive data structures. They used an OWL (Web Ontology Language) [18] ontology to model their domain, which consists of four concepts (*Person*, *Hospital*, *Diagnosis* and *Medication*). However, there is no information given on user interface generation. Aside from this work, it was very hard to find any related work on absolutely generic ontology-based data acquisition systems. In most publications ontologies are used for information extraction from text [19], [20], [21], or to enrich the existing data with structural and semantic information, or to build a cross-institutional knowledge base. In [22] the authors describe the usage of ontologies for inter-hospital data extraction to improve patient care and safety by analyzing hospital activities, procedures, and policies. Here, the ontology is strongly connected to the project. In [23] e.g. the authors describe an ontology based system for extracting research relevant data out of heterogeneous hospital information systems. Here again, the purpose is to find a superior common data model to compare data of different medical institutions. The most comparable work was published in 2014 by Lozano et al. [24], who also present an ontology-based system, called OWLing. Their intention is comparable to the above-mentioned, but their implementation is completely based upon the web ontology-language OWL. The work of [25] is also based on OWL and uses inference and reasoning to support medical staff in pre-operative risk assessment. They use the ontology as knowledge base for decision support. Although its structure is adaptable, the absolute genericity is no objective of this project. Generally speaking, many of these projects use ontologies to store explicit domain knowledge and use this knowledge for inference, reasoning, and decision support — which is the fundamental idea behind ontologies: to make knowledge of the domain expert accessible to computer algorithms. Our approach aims in the opposite direction: we use ontologies to enable the domain expert to explore their complex, high-dimensional and voluminous research data on their own. The domain-ontology — holding structural and semantic information about the research data — allows the software to automatize many data processing operations and to provide guidance and assistance to the researcher. In this way, the technical obstacles which a non-IT user is confronted with, when he is working with complex data structures are reduced. The ontology-based approach allows domain-specific guidance and assistance on the one hand, while it guarantees domain-independence on the other hand. For applying the system in a completely different research domain only the domain-ontology needs to be (re-)defined, without changes to the software itself.

In this paper we concentrate on a particularly important aspect: Scientists working with large volumes of high-dimensional data are immediately confronted with the problem of dimensionality reduction: finding meaningful low-dimensional structures hidden in their high-dimensional observations - to solve this hard problem is a primary task for data analysts and designers of machine learning algorithms and pattern recognition systems [26].
Actually, we are challenged by this problem in everyday perception: extracting meaningful, manageable features out of the high-dimensional visual sensory input data [27].

The general problem of dimensionality reduction can be described as follows:

Given the p-dimensional random variable $\mathrm{x} = (x_1, \ldots, x_p)^T$, the aim is to find a representation of lower dimensions, $\mathrm{s} = (s_1, \ldots, s_k)^T$ with $k < p$, which (hopefully) preserves the "meaning" of the original data. However, meaning is an human construct, same as "interesting", and is influenced by tasks, personal preferences and past experiences and many other environmental factors [28]. This makes it essential to put the domain expert into the loop of the knowledge discovery process [6], and to be aware of the important aspect that *feature extraction and feature transformation* is key for understanding data. The major challenge in KDD applications is to extract a set of features, as small as possible, that accurately classifies the learning examples [29].
Assuming n data items, each represented by an p-dimensional random variable $\mathrm{x} = (x_1, \ldots, x_p)^T$, there are two types of feature transformation techniques: linear and non-linear.

In linear techniques, each of the $k < p$ components of the new transformed variable is a linear combination of the original variables:

$$s_i = w_{i,1}x_1 + \ldots w_{i,p}x_p, \quad \text{for} \quad i = 1, \ldots, k, \quad \text{or}$$

$$\mathrm{s} = \mathbf{W}\mathrm{x},$$

where $\mathbf{W}_{k \times p}$ is the linear transformation weight matrix.

Expressing the same relationship as

$$\mathrm{x} = \mathbf{A}\mathrm{s},$$

with $\mathbf{A}_{p \times k}$, we note that the new variables s are also called the hidden or the latent variables. In terms of an $n \times p$ feature-document matrix $\mathbf{X}$, we have

$$\mathrm{S}_{i,j} = w_{i,1}\mathrm{X}_{1,j} + \ldots w_{i,p}\mathrm{X}_{p,j}, \quad \text{for} \quad i = 1, \ldots, k \quad \text{and} \quad j = 1, \ldots, n$$

where j indicates the jth realization, or, equivalently,

$$\mathbf{S}_{k\times n} = \mathbf{W}_{k\times p}\mathbf{X}_{p\times n},$$

$$\mathbf{X}_{p\times n} = \mathbf{A}_{p\times k}\mathbf{S}_{k\times n}.$$

There are various methods how to attack this general problem, the two most known include: principal component analysis (PCA) and multidimensional scaling (MDS). PCA is the oldest and most known technique [30], and can be used in various ways to analyze the structure of a data set and to represent the data in lower dimension; for a classic application example see [31], and for details refer to [32]:

Tenenbaum et al. (2000) describe an approach for solving dimensionality reduction problems, which uses easily measured local metric information to learn the underlying global geometry of the data set, which is very different from classical techniques such as PCA and MDS. Their approach is capable of discovering the nonlinear degrees of freedom that underlie complex natural observations. In contrast to previous algorithms for nonlinear dimensionality reduction, the approach by Tenenbaum et al. efficiently computes a globally optimal solution, and moreover, for an important class of data manifolds, is guaranteed to converge asymptotically to the true structure [33].

3 Theoretical Background

3.1 Exploratory Visual Analytics

In contrast to statistical approaches aimed at testing specific hypotheses, Exploratory Data Analysis (EDA) is a quantitative tradition that seeks to help researchers understand data when little or no statistical hypotheses exist [34]. Often the exploratory analysis is rather based on visualization of the data than on descriptive statistic and other data mining algorithms. This graphical approach is often referred to as visual analytics. A scientific panel of the National Visualization and Analytics Center (a section of the Department of Homeland Security) defined visual analytics as the science of analytical reasoning facilitated by interactive visual interfaces [35]. Behind this rather technical definition Thomas and Cook provide a very practical view on the usage of visual analytics: End-users can apply visual analytics techniques to gain insight into large and complex data sets to identify the expected and to discover the unexpected [35]. [36] state, that one of the main purposes of such approaches is to gain insight into the data and to create new hypotheses. The central aspect of visual analytics is the integration of the human domain expert into the data analytics process . This allows to take advantage of his/her flexibility, creativity, and background knowledge and combine these skills with the increasing storage capacities and computational power of today's computers [37].

3.2 Parallel Coordinates

Parallel Coordinates are well known method for loss-free dimensionality reduction. Instead of arranging the axis of a coordinate system orthogonally — which is limited to two respectively three axis (two dimensional computer displays in a three dimensional world) — they are arranged in a parallel way. A point in a two dimensional orthogonal coordinate system is represented by a line in a multi-dimensional parallel coordinate system. In the context of parallel coordinates, this is known as the point-line duality. Parallel coordinates as a mean of computer-based data visualization of higher dimensions were introduced by Alfred Inselberg [38] from the 1980ies up to now. Although parallel coordinates are known for some decades they are yet barely used in biomedical research [39]. On the other hand, their is an increasing number of publications on this raising from 14 in 1991 to 543 in 2011 on Google scholar [40], indicating the rising interest and relevance of this method.

3.3 Non-linear Mapping

Non-linear mapping is an application of multidimensional scaling (MDS). Multidimensional scaling is a method that represents measurements of similarity (or dissimilarity) among pairs of objects as distances between points of low-dimensional space [41]. Non-linear mapping extracts these measurements of dissimilarity from a data set in a high dimensional input space and subsequently uses MDS to scale them to a low-dimensional (usually two-dimensional) output space. In this way the high-dimensional topology of the input space is projected to the low-dimensional output space. A well known algorithm in this field is the Sammon's mapping algorithm by [42]. Sammon's mapping tries to reduce the error between the distance matrix in the high dimensional input space and the distance matrix of the low-dimensional output space, whereas the error E is defined as:

$$E = \frac{1}{\sum\limits_{i<j}[d^*_{ij}]} \sum_{i<j}^{N} \frac{[d^*_{ij}-d_{ij}]^2}{d^*_{ij}}$$

The term d^*_{ij} refers to the distance of two data points i and j in the high-dimensional input space and d_{ij} refers to the distance of i and j in the low-dimensional output space. The initial position of the data points in the output space, thus the initial values of d_{ij} are chosen randomly. It is now up to the implementation how to minimize the error value. Sammon used a deepest descend method in his paper, but there are also other ways, like the Kohonen heuristic [43, p. 35].

4 Method

4.1 An Ontology-Centered Infrastructure

The basic assumption behind the ontology-centered data infrastructure is the modified role of the domain expert in the knowledge discovery process. In common definitions this role is characterized as customer-like, consulting role [5], the main part of the process is performed by a so called data analyst. In every day medical research it is often the domain expert himself who takes the leading role in this process and is then confronted with technical barriers regarding handling, processing, and analyzing the complex research data [11]. This is known to be a major pitfall to (bio-)medical research projects [44].

Based on this assumption we developed a data infrastructure for scientific research that actively supports the domain expert in tasks that usually require IT knowledge or support, such as: structured data acquisition and integration, querying data sets of interest by non-trivial search conditions, data aggregation, feature generation for subsequent data analysis, data pre-processing, and the application of advanced data visualization methods. It is based upon a generic meta data-model and is able to store the current domain ontology (formal description of the actual research domain) as well as the corresponding research data. The whole infrastructure is implemented at a higher level of abstraction and derives its manifestation and behavior from the actual domain ontology at run-time. Just by modeling the domain ontology, the whole system, including electronic data interfaces, web portal, search forms, data tables, etc. is customized for the actual research project. The central domain ontology can be changed and adapted at any time, whereas the system prevents changes that would cause data loss or inconsistencies.

The infrastructure consists of three main modules:

1. Management Tool: The Management Tool is the main point of interaction for the research project leader. It allows the modeling and maintenance of the current domain ontology, as well as data processing, data validation, and exploratory data analysis.
2. Data Interface: The data interface is a plug-in to the well established open source ETL (Extraction-Transform-Load) suite Kettle, by Pentaho. Kettle allows the integration of numerous data sources and enables the user to graphically model his ETL process. For the final step, the data integration into our system and ontology-based data-sink interface was implemented.
3. Web Interface: The web interface is an automatically created web portal, which allows the users — depending on their permissions – to enter, view and edit existing data records. It is usually used to manually complement the electronically imported data with information that was not available electronically (e.g. hand-written care instructions, fever graphs, etc.) or in an un- or semi-structured way (e.g. doctors letters, image data, etc.).

An essential module of the Management Tool is an ontology-guided expression engine. It interweaves grammatical and structural meta information and allows the users to graphically model expression on their data that can be used for feature generation, the definition of complex search queries or data validity rules.

For further details on the ontology-guided expression engine the reader is kindly referred to [45], and for further information on the system itself to [46] and [47].

4.2 Ontology-Supported Data Exploration

It is the underlying paradigm of the whole infrastructure to move the researching domain expert in the central role of the knowledge discovery process. While the system supports the user in necessary tasks like data integration and processing, the real benefit of the paradigm shift occurs in the step of visual data exploration. Here the elaborate domain knowledge of the expert together with the general capability of the human mind to identify patterns in visualizations and the computational power of novel algorithms and systems can be combined, as [39] state. The aim is to support the expert end users in learning to interactively analyze information properties thus enabling them to visualize the relevant parts of their data [48].

In terms of the the presented system, the exploratory data analysis takes place in the management tool. Here the user can query the data sets of interest and move them to a special area of the system, called the Exploration perspective. In contrast to the Data perspective, the data can not be deleted or manipulated — it is safe. It only can be removed from the current set of interest. This set of interest contains the selected records and by default all attributes these records have. The user can now easily remove unneeded attributes and create new ones by using the system's expression engine. The expression engine allows the user to aggregate data from all over the domain ontology data structures with one data set of interest. For each data set a standard report including descriptive statistics parameters and interactive histogram display can be shown. In here, data cleaning can be performed. Once the data set is cleaned and checked by the system's data validity engine, a number of data visualizations can be created automatically, ranging from simple scatter plots or histograms to parallel coordinates and non-linear mapping.

Ontology-Guided Visual Clustering It is an often re-occurring requirement in medical research to find groups of similar elements, e.g. patients with similar symptoms or anamnesis. This process is often referred to as clustering or unsupervised learning. Cluster analysis is defined as the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity [49].

Cluster algorithms try to find groups of similar records and group them into meaningful clusters. The cluster membership of each data record is usually marked with a cluster number or cluster label. Without any visual check the result of the clustering is very hard to interpret. It provides no information on the shape of each cluster and no information of the topology among the clusters. Although cluster analysis is an established state-of-the-art methods, its direct benefit for the domain expert is very limited.

The most significant difference between supervised and unsupervised learning is the absence of a target value. Supervised learning algorithms try to minimize the error between a target value and a calculated value. Since the error is known, results of supervised learning algorithms can be evaluated. So it is possible to determine which algorithm yielded the best model. For clustering algorithms there is no error, no difference between calculated and desired values. So it is not possible to determine which is the best clustering. Moreover the quality of the result also depends on the user of the clustering, even if this understanding of quality stands in a contradiction to a calculated quality value [50].

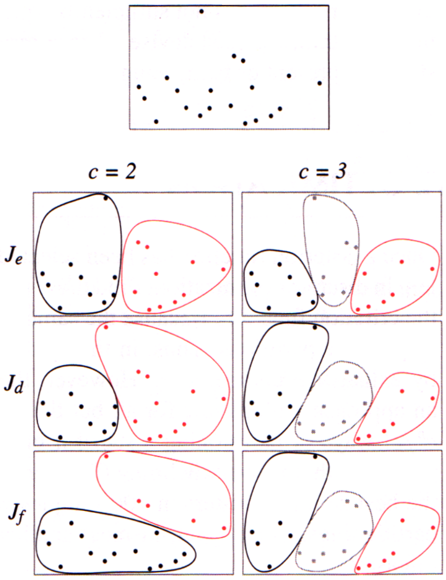


Fig. 1. One data set different criteria, different results (Duda et al., 2001).

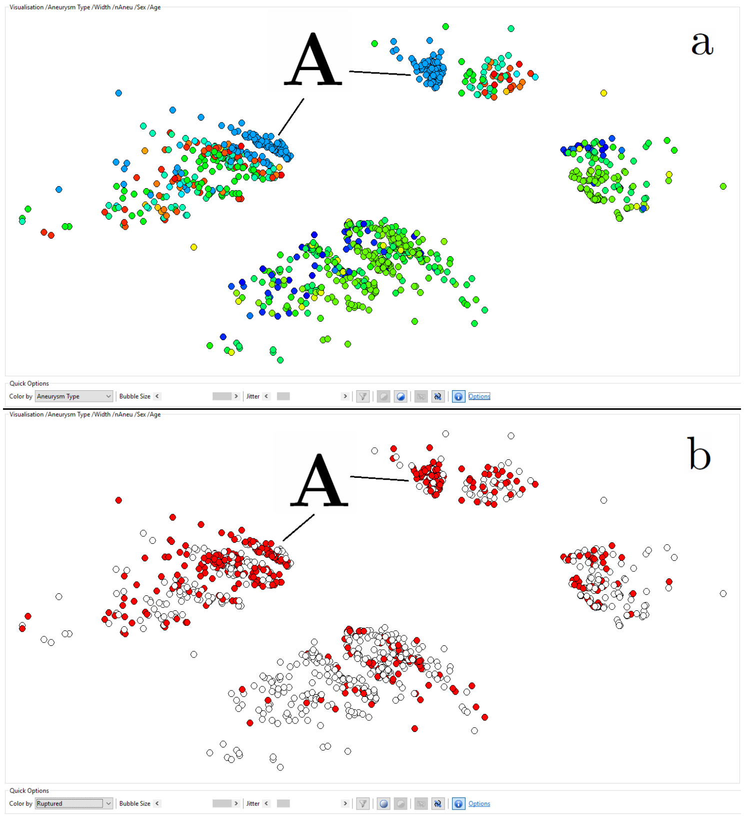
While there is no gold standard to compare to, there a number of quality criteria for unsupervised learning algorithms. The sum-of-squared-errors criteria [51] is one of them, which is minimized by some clustering algorithms (e.g. the k-means algorithm). However, Figure 1 shows up the limitations of those criteria. The data set itself doesn't show any obvious clustering. Three different cluster criteria were optimized for $c = 2$ (the number of clusters) and $c = 3$. All of these clusterings seem reasonable, there is no strong argument to favor one of them. While J_e (sum-of-squared-error criterion) tends to create two clusters of about equal size (for $c = 2$). J_d (determinant criterion [51] page 545) models two clusters of different sizes. Similar phenomena can be observed for $c = 3$ and the third criterion J_f (trace criterion [51] page 545). Although, there is no clustering with the data, the application of cluster algorithms would yield a number of clusters.

Furthermore, the result of some clustering algorithms may depend on a random initial state like k-means or the the EM algorithm [52]. Some of the algorithms need the number of clusters as an input parameter. Although there are strategies to overcome these shortcomings, like the consensus clustering meta heuristic [53], the suitability for daily use of standard unsupervised learning algorithms is very limited.

In order to overcome these drawbacks of classical cluster algorithms the decision was made to follow the visual analytics paradigm also in the task of finding clusters. Therefore, the potentially high dimensional research data needs to be mapped onto a two-dimensional display. Two well-known algorithms for this tasks are the Self-Organizing or Kohonen Map (SOM) [43] and the non-linear mapping algorithm Sammon's mapping [42]. Both algorithms try to minimize the error or mismatch between a topology in the n-dimensional source-space and the (mostly) two-dimensional target space. While the Kohonen Map yields a graph whose topology corresponds with the original topology, Sammon's mapping yields a cloud of dots, whereas each dot represents an input data record. In a theoretical evaluation [54] both algorithms were evaluated. Although the result of this evaluation preferred the SOM, practical tests with domain experts showed that the dot cloud was way easier to interpret than the network yielded by the SOM.

For the user, the non-linear mapping algorithm is hidden behind the notion 'Visual Clustering'. The only configuration, which is required by the user, is to select which attributes should be taken into account for the calculation of the distance or dissimilarity of two records. Then, the algorithm normalizes the data. Subsequently, a distance matrix is calculated, whereas for numerical variables an Euclidean distance is used and the Jaccard Metric [55] for categorical variables. Finally, the result is presented in a scatter plot. Via mouse wheel the user is able to change the variable that is used to color the dots. In this way, not only patterns in the topology of the data can be identified but also the correlation to other attributes according to the coloring. Within the plot, the user is able to select data sets of interest and directly access and process the underlying data records.

Figure 2 shows the visualization of a real-world medical data set of 1032 cerebral aneurysms. A cerebral aneurysm is the dilation, ballooning-out, or bulging of part of the wall of an artery in the brain [57]. The data was collected using the previously described ontology-centered infrastructure, at the Institute for Radiology at the Campus Neuromed of the Medical University Linz. The parameters for the visual clustering were chosen by the medical experts. A more detailed description of the data set and the experiments can be found in [56]. The visualization revealed the high dimensional structure of the data set showing clusters of ruptured aneurysms (colored in red in Figure 2 - section b). When coloring parameter was switched to the location of the aneurysm (section a in Figure 2), it was clearly visible that two of the previously seen clusters of ruptured aneurysms share the same location, which lead to the research-hypothesis that aneurysm in this location might be more prone to rupture. A hypothesis which turned out to be evident in medical literature [58]. Aside this confirmation and reproduction of already existing knowledge, the visualization yielded further interesting patterns which resulted in research hypotheses that are currently addressed by the medical researchers.



Visualisation /Aneurysm Type /Width /nAneu /Sex /Age

Quick Options

Color by: Aneurysm Type

Visualisation /Aneurysm Type /Width /nAneu /Sex /Age

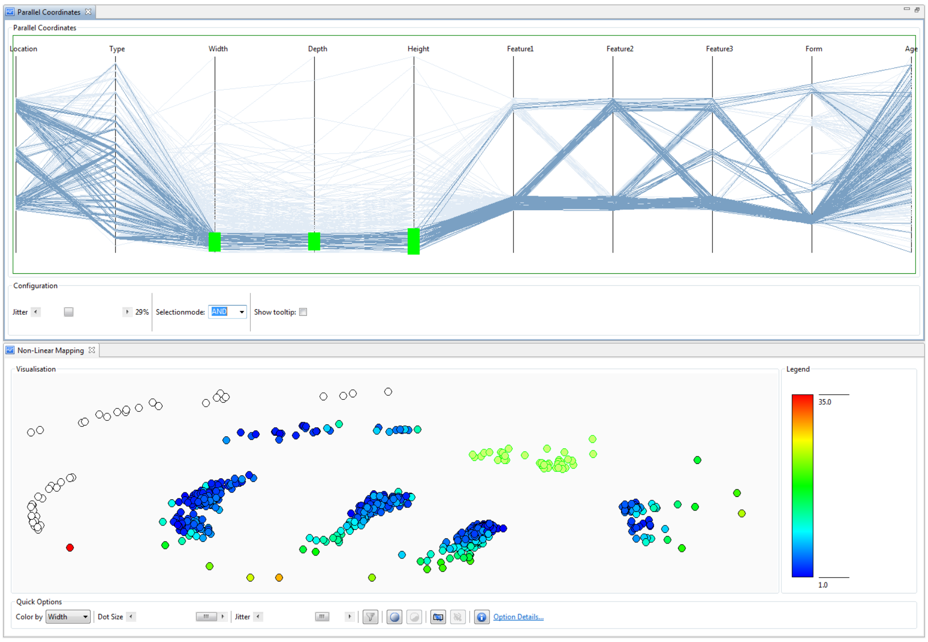
Quick Options

Color by: Ruptured

Fig. 2. An ontology-guided non-linear mapping of 1032 cerebral aneurysms with a distance calculation based on the following features: Aneurysm.Width, Aneurysm.Location, Patient.Number of Aneurysms, Patient.Age. a) The aneurysms are colored according to their location. b) The aneurysms are colored according to their rupture state: red are ruptured, white are none-ruptured [56]

Plots showing the same set of interest are linked with each other in a way that the selection on one plot is automatically highlighted on all other plots as well. Figure 3 shows two linked displays of a parallel coordinate system and a visual clustering. The selection in the parallel coordinate system is highlighted in the scatter plot of the clustering.

Ontology-Guided Parallel Coordinate View Comparable to the visual clustering, the user is able to visualize any data set of interest in form of a parallel coordinate system. For this visualization no further configuration is required. The system automatically normalizes numerical attributes and maps categorical attribute to numeric ranges. By hovering the mouse over the axes the current numeric or the underlying categorical value are shown as a tool tip. Categori-



Jitter 29%

Selectionmode: AND

Show tooltip: □

Color by Width

Jitter

Fig. 3. A simultaneous display of a parallel coordinate system with a scatter plot showing the result of a non-linear mapping. The selection in the parallel coordinate system is highlighted (in green) in the scatter plot. Both visualization show the same data set as seen in Figure 2

.

cal and Boolean values can cause horizontal lines to overlap, which leads to the effect that the user is not able to distinguish whether there is only one record with a certain configuration or many. Therefore, a jitter slider was integrated into the visualization that adds a random offset to values of Boolean and categorical attributes and let the corresponding lines drift apart. So it is perceptible if there are overlapping values. The selection within the parallel coordinate view is propagated throughout the system and highlighted in all other visualizations showing the same data set (see Figure 3). Furthermore, selected records can be opened and edited, which allows the quick changing from data exploration to data revision or editing.

5 Conclusion

In this paper we presented an ontology-based research infrastructure, grounded in the central paradigm to put the researching domain expert in a central position of the knowledge discovery process. After supporting the researcher in data integration and handling, we now address the field of data exploration. Here, we chose a primarily visual approach using two completely different means of dimensionality reduction: parallel coordinates on the one hand, and non-linear mapping on the other. Both are well-known means of visualization in the field of computer science. However, our experiences showed that they are practically not used for medical research. Consequently, to lower the technical barrier to use this type of visualization algorithms, we use the structural meta-information of the central ontology, to reduce the configuration and pre-processing requirements to a minimum. The main advantages of our approach can be summarized by:

– Domain Independence: The presented software is completely domain independent and can be adapted to any research domain by modeling the central domain ontology. The whole system adapts to this ontology at run-time.
– Assistance to Domain Experts: The structural and semantic information from the domain ontology is used to actively support the domain expert — who is usually no IT-expert — in technically challenging tasks such as data integration, data processing and data analytics.
– Fully Integrated Visual Analytics: The described data visualization algorithms are seamlessly integrated into the system and can be applied to any given data set of interest. The necessary data pre-processing and the launching of the algorithms is automatized based upon the structural information from the domain ontology. All visualization allow the interaction and a direct drill-down to the underlying data.

Practical experiences within the clinical context demonstrated that medical researchers are surprised at a first glance that the visual clustering often yields properly separable clusters. They were not aware of the existence of subgroups in their patient collective and the question of correlation with the clusters to other attributes immediately arose, a question that could often be answered just by coloring the dots in the cloud, according to the desired attribute. Generally, the dot cloud, yielded by the visual clustering, proved to be interpretable without further explanation. According to our experiences, the parallel coordinate visualization often confuses end users the first time they are confronted with it. As soon as the first selection across the axis is performed and subsequently moved along the axis, the fundamental principle of the visualization get clear and invites researchers to quickly check their hypothesis by setting and moving selection markers along the axis.

There are also risks and limitations in visual analytics generally and of course in our approach specifically. Similarly to any data-based method the quality of the visualization strongly depends on the quality of the underlying data. Nevertheless, we recommend to have the first look on the original raw data, because there might be something ”interesting” in it. Noisy or messy data yields meaningless or — even more dangerous — misleading visualizations and might lead to the danger of modelling artifacts [59]. To address this problem, the ontology based infrastructure supports a wide range of data quality and plausibility checks to increase data quality. Secondly, it is important for the users to understand that a visualization yields insights and hypotheses and not facts. Regarding the example which is shown in Figure 2, it is very important to understand that the clearly visible connection between the ruptured aneurysm and their location is just a hypothesis. There is no prove for a (statistically significant) correlation or even a causality. Both need to be examined using statistics and further medical research.

6 Future Research

There are a number of further research activities planned along the presented infrastructure. At first, we will support the visual approach by a machine learning module. Comparable to the consensus clustering algorithm [53], where clustering algorithms in different settings are executed and their result is consolidated, a supervised learning meta-algorithm will be developed. The base hypothesis behind this idea is the following: when different supervised learning algorithms in multiple configurations are able to predict a certain target value, by using the selected features, there is most probably a (non-linear) correlation between the features and the target values. If all or most algorithms fail, then we have to assume, that there is no correlation — at least within the given data set. Here again, the main objective is to enable the researching domain expert to use advanced machine learning algorithms, hence to combine the computational and algorithmic power of the computer with the intelligence and experience of the domain expert.

Clustering algorithms are based on distance or dissimilarity measures. Currently, for categorical attributes the Jaccard metric is used. Although it is able to cope with multi-selectable categorical attributes, it does not take into account a possible similarity of the category enumeration values. Given two patients with each two diagnoses: Patient 1 suffers from a flu and has a broken shinbone, while patient 2 has a broken fibular and a pneumonia. A distance measure that ignores the similarity of flu and pneumonia, and broken shinbone and broken fibular returns a big dissimilarity between patient 1 and 2. With the knowledge that a flu and a pneumonia are very similar diseases and a broken shinbone is almost the same as a broken fibular, the two patients would be considered very similar. It is now subject to our current research to integrate this knowledge into the distance measures for clustering following the ideas of [60], [61], or [62].

Currently, our system is designed to work with structured data. Unstructured data (e.g. free-text diagnoses, hand written care instructions, image data, etc.) can only be integrated into the system manually, using the automatically created web interface. For a more computer supported integration of electronically stored non-standardized (in medical jargon called: free text), an expansion and research in the field of information extraction is planned.

The system in its current state is suitable for storing highly structured data. It supports arbitrary complex and deep data structures. However, due to performance reason the number of attributes of each data class is limited. There is no fixed limit, however, keeping in mind performance and usability, it does not make sense to create more than 20 or 30 attribute to a class. This is an issue, when very high dimensional research data shall be integrated including genome expression data. Regarding the number of data sets the limiting factor is the performance of CPU respectively GPU of the user's machine. Especially the non-linear mapping has a very unfavorable run-time behavior. Experiments show acceptable performance up to several 10,000 of records with State-of-the-Art desktop PCs. Data sets with significantly higher number of data would require a preliminary sampling. For this kind of data special data storage and integration methods need to be found and future research will also focus on the integration of alternative visualization methods (e.g. [63], [64], consequently our work provides several additional and alternative future research possibilities. A very important future issue is to foster transparency, i.e. to explain why a decision had been made [65].

References

1. Fayyad, U., Piatetsky-shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Magazine **17** (1996) 37–54
2. 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
3. Holzinger, A.: Introduction to machine learning and knowledge extraction (make). Machine Learning and Knowledge Extraction **1** (2017) 1–20
4. Holzinger, A., Malle, B., Kieseberg, P., Roth, P.M., Mller, H., Reihs, R., Zatloukal, K.: Machine learning and knowledge extraction in digital pathology needs an integrative approach. In: Springer Lecture Notes in Artificial Intelligence Volume LNAI 10344. Springer International, Cham (2017) 13–50
5. Kurgan, L.A., Musilek, P.: A survey of knowledge discovery and data mining process models. The Knowledge Engineering Review **21** (2006) 1–24
6. Holzinger, A. In: HumanComputer Interaction and Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? Springer, Heidelberg, Berlin, New York (2013) 319–328
7. Holzinger, A., Jurisica, I. In: Knowledge Discovery and Data Mining in Biomedical Informatics: The future is in Integrative, Interactive Machine Learning Solutions. Springer, Heidelberg, Berlin (2014) 1–18
8. Zudilova-Seinstra, E., Adriaansen, T.: Visualisation and interaction for scientific exploration and knowledge discovery. Knowledge and Information Systems **13** (2007) 115–117
9. Cios, K.J., William Moore, G.: Uniqueness of medical data mining. Artificial intelligence in medicine **26** (2002) 1–24
10. Holzinger, A., Stocker, C., Dehmer, M. In: Big Complex Biomedical Data: Towards a Taxonomy of Data. Springer, Berlin Heidelberg (2014) 3–18
11. Anderson, N.R., Lee, E.S., Brockenbrough, J.S., Minie, M.E., Fuller, S., Brinkley, J., Tarczy-Hornoch, P.: Issues in biomedical research data management and analysis: Needs and barriers. Journal of the American Medical Informatics Association **14** (2007) 478–488
12. Niakšu, O., Kurasova, O.: Data mining applications in healthcare: Research vs practice. Databases and Information Systems BalticDB&IS 2012 (2012) 58
13. Frankel, D.: Model driven architecture : applying MDA to enterprise computing. Wiley, New York (2003)
14. Schmidt, D.C.: Model-driven engineering. Computer **39** (2006) 25–31

15. Cruz, A.M.R., Faria, J.P.: A metamodel-based approach for automatic user interface generation. In Petriu, D., Rouquette, N., Haugen, A., eds.: Model Driven Engineering Languages and Systems. Volume 6394 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2010) 256–270
16. Renggli, L., Ducasse, S., Kuhn, A.: Magritte - a meta-driven approach to empower developers and end users. In Engels, G., Opdyke, B., Schmidt, D., Weil, F., eds.: Model Driven Engineering Languages and Systems. Volume 4735 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2007) 106–120
17. Zavaliy, T., Nikolski, I.: Ontology-based information system for collecting electronic medical records data. In: Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET), 2010 International Conference on. (2010) 125
18. McGuinness, D.L., van Harmelen, F.: Owl web ontology language overview: W3c recommendation (2004)
19. Tran, Q.D., Kameyama, W.: A proposal of ontology-based health care information extraction system: Vnhies. In: Research, Innovation and Vision for the Future, 2007 IEEE International Conference on. (2007) 1–7
20. Holzinger, A., Geierhofer, R., Modritscher, F., Tatzl, R.: Semantic information in medical information systems: Utilization of text mining techniques to analyze medical diagnoses. Journal of Universal Computer Science **14** (2008) 3781–3795
21. Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., Verspoor, K.: Biomedical text mining: State-of-the-art, open problems and future challenges. In Holzinger, A., Jurisica, I., eds.: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics, Lecture Notes in Computer Science LNCS 8401. Volume 8401. Springer, Berlin Heidelberg (2014) 271–300
22. Kataria, P., Juric, R., Paurobally, S., Madani, K.: Implementation of ontology for intelligent hospital wards. In: Hawaii International Conference on System Sciences, Proceedings of the 41st Annual, title=Implementation of Ontology for Intelligent Hospital Wards. (2008) 253
23. Kiong, Y.C., Palaniappan, S., Yahaya, N.A.: Health ontology system. In: Information Technology in Asia (CITA 11), 2011 7th International Conference on. (2011) 1–4
24. Lozano-Rubí, R., Pastor, X., Lozano, E.: Owling clinical data repositories with the ontology web language. JMIR Medical Informatics **2** (2014) e14
25. Bouamrane, M.M., Rector, A., Hurrell, M.: Using owl ontologies for adaptive patient information modelling and preoperative clinical decision support. Knowledge and information systems **29** (2011) 405–418
26. Kaski, S., Peltonen, J.: Dimensionality reduction for data visualization (applications corner). IEEE Signal Processing Magazine **28** (2011) 100–104
27. Holzinger, A.: Trends in interactive knowledge discovery for personalized medicine: Cognitive science meets machine learning. Intelligent Informatics Bulletin **15** (2014) 6–14
28. Beale, R.: Supporting serendipity: Using ambient intelligence to augment user exploration for data mining and web browsing. International Journal of Human-Computer Studies **65** (2007) 421–433
29. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces. Knowledge and information systems **12** (2007) 95–116
30. Pearson, K.: On lines and planes of closest fit to systems of points in space. Philosophical Magazine **2** (1901) 559–572

31. Hoover, A., Jean-Baptiste, G., Jiang, X., Flynn, P.J., Bunke, H., Goldgof, D.B., Bowyer, K., Eggert, D.W., Fitzgibbon, A., Fisher, R.B.: An experimental comparison of range image segmentation algorithms. Pattern Analysis and Machine Intelligence, IEEE Transactions on **18** (1996) 673–689
32. Jackson, J.E.: A user's guide to principal components. Volume 587. John Wiley and Sons (2005)
33. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290** (2000) 2319–2323
34. Behrens, J.T., Yu, C.H. In: Exploratory Data Analysis. John Wiley & Sons, Inc. (2003)
35. Thomas, J., Cook, K.: A visual analytics agenda. Computer Graphics and Applications, IEEE **26** (2006) 10–13
36. Holzinger, A., Scherer, R., Seeber, M., Wagner, J., Müller-Putz, G.: Computational sensemaking on examples of knowledge discovery from neuroscience data: towards enhancing stroke rehabilitation. In: Information Technology in Bio-and Medical Informatics. Springer (2012) 166–168
37. Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H.: Visual analytics: Scope and challenges. Springer (2008)
38. Inselberg, A.: The plane with parallel coordinates. The Visual Computer **1** (1985) 69–91
39. Otasek, D., Pastrello, C., Holzinger, A., Jurisica, I.: Visual data mining: Effective exploration of the biological universe. In: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. Springer (2014) 19–33
40. Heinrich, J., Weiskopf, D.: State of the art of parallel coordinates. STAR Proceedings of Eurographics **2013** (2013) 95–116
41. Borg, I.: Modern multidimensional scaling : theory and applications. Springer, New York (1997)
42. Sammon, J.W.: A nonlinear mapping for data structure analysis. IEEE Transactions on Computers **18** (1969) 401–409
43. Kohonen, T.: Self-Organizing Maps - Third Edition. Springer (2001)
44. Franklin, J.D., Guidry, A., Brinkley, J.F.: A partnership approach for electronic data capture in small-scale clinical trials. Journal of Biomedical Informatics **44, Supplement 1** (2011) S103 – S108
45. Girardi, D., Küng, J., Giretzlehner, M.: A meta-model guided expression engine. In: Intelligent Information and Database Systems. Springer (2014) 1–10
46. Girardi, D., Arthofer, K., Giretzlehner, M.: An ontology-based data acquisition infrastructure. In: Proceedings of 4th International Conference on Knowledge Engineering and Ontology Development, Barcelona (2012) 155–160
47. Girardi, D., Dirnberger, J., Trenkler, J.: A meta model-based web framework for domain independent data acquisition. In: ICCGI 2013, The Eighth International Multi-Conference on Computing in the Global Information Technology. (2013) 133–138
48. Holzinger, A.: On knowledge discovery and interactive intelligent visualization of biomedical data-challenges in human-computer interaction & biomedical informatics. In: DATA. (2012)
49. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review. ACM Computing Surveys **31** (1999) 265–323
50. Elhawary, M., Nguyen, N., Smith, C., Caruana, R.: Meta clustering. Sixth IEEE International Conference on Data Mining **1** (2006) 107–118
51. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification - Second Edition. Wiley Interscience (2001)

52. Bradley, P.S., Fayyad, U.M.: Refining initial points for k-means clustering. Proceedings of the Fifteenth International Conference on Machine Learning (1998) 91–99
53. Monti, S., Tamayl, P., Mesirov, J., Golub, T.: Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning **52** (2003) 91–118
54. Girardi, D., Giretzlehner, M., Küng, J.: Using generic meta-data-models for clustering medical data. In: ITBAM, Vienna (2012) 40–53
55. Boriah, S., Chandola, V., Kumar, V.: Similarity measures for categorical data: A comparative evaluation. red **30** (2008) 3
56. Girardi, D., Küng, J., Kleiser, R., Sonnberger, M., Csillag, D., Trenkler, J., Holzinger, A.: Interactive knowledge discovery with the doctor-in-the-loop: a practical example of cerebral aneurysms research. Brain informatics **3** (2016) 133–143
57. NIH: Cerebral aneurysm information page (2010)
58. Bijlenga, P., Ebeling, C., Jaegersberg, M., Summers, P., Rogers, A., Waterworth, A., Iavindrasana, J., Macho, J., Pereira, V.M., Bukovics, P., et al.: Risk of rupture of small anterior communicating artery aneurysms is similar to posterior circulation aneurysms. Stroke **44** (2013) 3018–3026
59. Wartner, S., Girardi, D., Wiesinger-Widi, M., Trenkler, J., Kleiser, R., Holzinger, A.: Ontology-guided principal component analysis: Reaching the limits of the doctor-in-the-loop. In Renda, E.M., Bursa, M., Holzinger, A., Khuri, S., eds.: Information Technology in Bio- and Medical Informatics: 7th International Conference, ITBAM 2016, Porto, Portugal, September 5-8, 2016, Proceedings. Springer International Publishing, Cham (2016) 22–33
60. Hsu, C.C.: Generalizing self-organizing map for categorical data. Neural Networks, IEEE Transactions on **17** (2006) 294–304
61. Boutsinas, B., Papastergiou, T.: On clustering tree structured data with categorical nature. Pattern Recognition **41** (2008) 3613–3623
62. Gibert, K., Valls, A., Batet, M.: Introducing semantic variables in mixed distance measures: Impact on hierarchical clustering. Knowledge and Information Systems **40** (2014) 559–593
63. Lex, A., Streit, M., Kruijff, E., Schmalstieg, D.: Caleydo: Design and evaluation of a visual analysis framework for gene expression data in its biological context. In: Pacific Visualization Symposium (PacificVis), 2010 IEEE, IEEE (2010) 57–64
64. Mueller, H., Reihs, R., Zatloukal, K., Holzinger, A.: Analysis of biomedical data with multilevel glyphs. BMC Bioinformatics **15** (2014) S5
65. Holzinger, A., Plass, M., Holzinger, K., Crisan, G.C., Pintea, C.M., Palade, V.: A glass-box interactive machine learning approach for solving np-hard problems with the human-in-the-loop. arXiv:1708.01104 (2017)