Ontology-Guided Principal Component Analysis: Reaching the Limits of the Doctor-in-the-Loop

Sandra Wartner[1], Dominic Girardi[1(✉)], Manuela Wiesinger-Widi[1], Johannes Trenkler[2], Raimund Kleiser[2], and Andreas Holzinger[3]

[1] Research Unit Medical Informatics at RISC Software GmbH, Johannes Kepler University, Hagenberg and Linz, Austria
{sandra.wartner,dominic.girardi,manuela.wiesinger-widi}@risc-software.at
[2] Institute of Neuroradiology at Neuromed Campus of the Kepler University Klinikum, Linz, Austria
[3] Research Unit, HCI-KDD, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Graz, Austria

Abstract. Biomedical research requires deep domain expertise to perform analyses of complex data sets, assisted by mathematical expertise provided by data scientists who design and develop sophisticated methods and tools. Such methods and tools not only require preprocessing of the data, but most of all a meaningful input selection. Usually, data scientists do not have sufficient background knowledge about the origin of the data and the biomedical problems to be solved, consequently a doctor-in-the-loop can be of great help here. In this paper we revise the viability of integrating an analysis guided visualization component in an ontology-guided data infrastructure, exemplified by the principal component analysis. We evaluated this approach by examining the potential for intelligent support of medical experts on the case of cerebral aneurysms research.

Keywords: Principal component analysis · Ontology · Data mining · PCA · Data warehousing · Doctor-in-the-loop

1 Introduction

Medicine is constantly turning into a data intensive science and the quantity of available health data is enormously increasing - far beyond what a medical doctor can handle [4]. Within such large amounts of data, relevant *structural* and/or *temporal* patterns ("knowledge") are often hidden and not accessible to the medical doctors [14].

However, the real problem is not only in the large quantities of data (colloquially called: "big data"), but in "complex data". Medical doctors today are confronted with complex data sets in arbitrarily high dimensions, mostly heterogeneous, semi-structured, weakly-structured and often noisy [15] and of poor data quality. The handling and processing of this data is known to be a major technical obstacle for (bio-)medical research projects [2]. However, it is not only

M.E. Renda et al. (Eds.): ITBAM 2016, LNCS 9832, pp. 22–33, 2016.
DOI: 10.1007/978-3-319-43949-5_2

the data handling that contains major obstacles, also the application of advanced data analysis and visualization methods is often only understandable for data scientists. This situation will become even more dramatic in the future due to the ongoing trend towards personalized medicine with the goal of tailoring the treatment to the individual patient [12].

Interestingly, there is evidence that human experts sometimes still outperform sophisticated algorithms, e.g., in the instinctive, often almost instantaneous interpretation of complex patterns. A good example is diagnostic radiologic imaging, where a promising approach is to fill the semantic gap by integrating the physicians high-level expert knowledge into the retrieval process by acquiring his/her relevance judgments regarding a set of initial retrieval results [1].

Consequently, the integration of the knowledge of a domain expert may sometimes greatly enhance the knowledge discovery process pipeline. The combination of both human intelligence and machine intelligence, by putting a "human-in-the-loop" would enable what neither a human nor a computer could do on their own. This human-in-the-loop can be beneficial in solving computationally hard problems, where human expertise can help to reduce an exponential search space through heuristic selection of samples, and what would otherwise be an NP-hard problem, reduces greatly in complexity through the input and the assistance of a medical doctor into the analytics process [13]. This approach is supported by a synergistic combination of methodologies of two areas that offer ideal conditions towards unraveling such problems: Human-Computer Interaction (HCI) and Knowledge Discovery/Data Mining (KDD), with the goal of supporting human intelligence with machine intelligence to discover novel, previously unknown insights into data (HCI-KDD approach [11]).

From the theory of human problem solving it is known that, for example, medical doctors can often make diagnoses with great reliability – but without being able to explain their rules explicitly [6]. Here this approach could help to equip algorithms with such "instinctive" knowledge. The importance of this approach becomes clearly apparent when the use of automated solutions due to the incompleteness of ontologies is difficult [3].

The immediate integration of the domain expert into data exploration has already proved to be very effective, for example in knowledge discovery [9], or in subspace clustering [17], compelling the domain expert to face the major challenge of detecting mutual influences of variables. Having already an idea of those dependencies, the domain expert's goal, here: the medical doctor, is to confirm his suspicions; contrary to the data scientist, who has hardly any domain knowledge and therefore no insight in reasonable input for specific tools. Frequently, for many domain experts it is even not possible to have access to worthwhile, already long-time existing data analysis tools, including, e.g., the Principal Component Analysis (PCA), due to a lack of mathematical knowledge, on the one side, and missing computational knowledge, on the other side. Consequently, 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 enormous complexity of the medical research domain [5].

A survey from 2012 among hospitals from Germany, Switzerland, South Africa, Lithuania, and Albania [23] showed that only 29 % of the medical personnel of responders were familiar with practical applications 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 of data mining tools is the relatively high technical obstacle that often needs to be taken in order to apply 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 address this issue, 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 preprocessing, 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. In this context, medical experts are offered assistance in their research purposes.

In many cases, these domain experts are unfamiliar with the variety of mathematical methods and tools which greatly simplify data exploration. In order to overcome impediments concerning mathematical expertise or the selection and application of suitable methods, we propose ontology-guided implementations for domain-expert-driven data exploration. One of those major methods is Principal Component Analysis (hereinafter referred to as PCA, see Sect. 2), representing a powerful method for dimensionality reduction.

In order to assist domain experts data preprocessing is automated as far as possible using the user-defined domain-ontology to overcome technical obstacles already in advance. Thus, the domain expert is capable of performing the fundamental analysis on his own. By selecting data of interest and starting the calculations, PCA is run in the background and results in an inbuilt visualization for more convenient access of data information.

2 Principal Component Analysis

Principal Component Analysis (PCA) is a method for reducing the dimension of a data set such that the new set contains most of the information of the original set and can be interpreted more easily. PCA was first described by Pearson [25]

and since then has been reinvented in different fields such as Economic Sciences [16], Psychology [28,29], and Chemistry [20,27] under different names like Factor Analysis or Singular Value Decomposition. Also in other fields, including Geo Sciences and Social Sciences, PCA is an established method. For a good introduction to PCA see for example [18,26].

In the following paragraph we sketch the main idea of PCA. We are given a set of observations of m variables. PCA then computes the direction of maximal variance in these data in m-dimensional space. This direction forms a new variable (a linear combination of the original variables), the first principal component. This process is repeated with the remaining variance of the data until a specified number of principal components is reached or a specified percentage of the original variance is covered (explained variance of the system). Every succeeding principal component is orthogonal to the preceding ones and adds to the explained variance of the new system. There cannot be more principal components than original variables and if their number is equal then the explained variance is 100 %. Mathematically, PCA is a solution to the eigenvalue problem of the covariance resp. correlation matrix of the original variables where the eigenvectors form the principal components and the eigenvalues indicate the importance of the components (the higher the eigenvalue, the higher the explained variance of the component).

Of interest in interpreting the results of a PCA are the scores (projection of the original data points into the new vector space), loadings (eigenvectors multiplied by the square root of the corresponding eigenvalues, i.e., the loadings also include variance along the principal components), residuals and their respective plots. The score plot depicts the scores with respect to two selected principal components that form the axes of the plot. It is used to detect outliers and patterns in the data. The loadings plot depicts the original variables with respect to two selected principal components that form the axes of the plot. It is used to examine correlations between the original variables and to examine the extent to which the variables contribute to the different principal components. The biplot displays both scores and loadings simultaneously and allows to investigate the influence of the variables on the individual data points or groups of data points, respectively.

First ideas of introducing PCA in the medical field came up in the early 70 s, gradually increasing. From 2006 onwards, the annual increase of research results is still growing very fast, comprising already about 670 scientific results in 2015 on NCBI [22]. Currently, PCA establishes a satisfying solution in various medical sub domains for different purposes. The application field ranges from image processing, like image compression or recognition [21], to data representation, for facilitating analysis.

3 Ontology-Guided PCA

In this section, we briefly review the main integration actions of the PCA method (see Sect. 2) into the data infrastructure. Above all the term ontology has to be

clarified as there is a degree of uncertainty around the terminology, whereby for computer scientists an ontology is described as formal descriptions, properties and relationships between objects in the world [30].

3.1 Ontology-Guided Clinical Research Infrastructure

The theoretic base for the already mentioned ontology-based research infrastructure is a revision and adaption of the established process models for knowledge discovery. In the commonly known definitions of this process (see [19] for a good overview) the domain-expert is seen in a supervising, consulting and customer role. A person who is outside the process and assists in crucial aspects with domain knowledge and receives the results. All the other steps of the process are performed by so called data analysts, who are supported by the domain-experts in understanding the for the current research project relevant aspects of the research domain and interpreting the results. We revised these process models and proposed a new, domain-expert-centric process model for medical knowledge discovery [8]. Based upon this process model we developed a generic research infrastructure, which supports the domain expert throughout the whole process — from data model, over data acquisition and - integration, data processing, and quality-management to data exploration. The research infrastructure is domain independent and derives its current appearance and behavior from the user-defined domain ontology at run-time. The researcher is able to define what data structure he or she needs to answer the research questions. This definition — the domain ontology — then builds the base for the whole system. From a user's point of view, the infrastructure consists of three main modules:

1. The Management Tool: This Java rich-client application allows the user to defined and maintain the domain ontology. Furthermore, the whole data set of the system can be searched, filter, processed and analyzed in this application. The here presented work is integrated into this application.
2. The Data Integration Module: This module is a plug-in to an established open-source ETL (Extract-Transform-Load) Suite. It allows to access structured data from almost arbitrary sources and to properly integrate this data into the research infrastructure.
3. The Web Interface: If data needs to be acquired manually, the web interface offers domain-derived forms to view, enter, process the data via a web browsers. In the clinical context this is often necessary when information from semi- or unstructured documents (e.g. doctor's letters, care instructions, etc.) needs to be stored in a structured way.

For more detailed information on this infrastructure the reader is kindly referred to [7].

3.2 Background Processes

The execution of PCA requires structured processing of data. In our data infrastructure all of those preparatory steps are based on ontological meta-information and are automatically performed in the background. In this case

the domain expert neither has to be concerned about data types, data transformation, starting the corresponding algorithm nor about collecting and depicting results. Therefore, solely a few steps remain, explicitly data selection and parameter setting, in order to start PCA. After variable selection out of a (sub) set of data and adjusting parameter configuration PCA performs the projection into lower dimensional space. The result is visualized in interactive two dimensional charts (loading-, score- and biplot), offering manipulation of axes and therefore displaying different combinations of principal components. Backtracking to the pristine records can establish a better idea of relationships when selecting the scores. In a few steps data is ready for analysis.

3.3 Implementation

For the actual implementation, we used the WEKA library [10] for performing PCA. Therefore some partial integration has been necessary in order to acquire mathematical PCA output for visualization purposes.

Step 1. First, an ontology-guided transformation of the data into the WEKA **data** structure (*weka.core.Instances*) and converting our variables to WEKA conform **attributes** (*weka.core.Attribute*) has to be performed. An **evaluator** (*weka.attributeSelection.PrincipalComponents*, defining the evaluation method) has to be configured by setting the variance covered by the principal component as well as whether the correlation or the covariance matrix has to be used. All of those transformations are performed in an ontology-guided, hidden behavior from the researcher's perspective.

Step 2. The key component of the WEKA PCA is represented and performed by the **feature selection** (*weka.attributeSelection.AttributeSelection*), hence a **ranker** (*weka.attributeSelection.Ranker*, defining the search method) as well as the evaluator configured in step 1 have to be assigned. The specified ranker's task is to supervise whether the defined threshold (explained variance) or a specified number of components is reached, thus PCA has finished.

Step 3. After completion of the embedded WEKA PCA process, an internal PCA result class is prepared, carrying the most essential output including eigenvectors and eigenvalues, scores and loadings as well as the number of principal components and proposed features. Accordingly, the generated output is subject to back transformation in the prescribed ontology and is processed for being displayed in a scatter chart related visualization.

Step 4. In order to determine the quality of the result some key figures have to be determined. Therefore we take into account linearity of the input data, the size of the data set, the variance covered as well as tests on normal distribution. The outcome of the quality test is displayed within the visualization, supporting the researcher in evaluating the significance of the outcome. Since PCA is vulnerable to outliers, outlier detection is provided in the visualization, making it possible for the user to exclude these points and re-initiate a PCA. In particular, the rationale for various quality outcomes is the quality of the input data.

4 Results

We evaluated the viability of this approach to perform PCA within an ontology-guided data infrastructure for scientific research purposes on a data set of 1237 records, representing a cerebral aneurysm each. This vulnerability of a blood vessel is described as the dilation, ballooning-out, or bulging of part of the wall of an artery in the brain [24]. Those samples were taken from patients, registered by the Institute of Neuroradiology at Neuromed Campus of the Kepler University Klinikum. The aim of this collaboration was to collect and analyze the medical outcome data of their patients, who have cerebral aneurysms. The main research subject of the database is the clinical and morphological follow up of patients with cerebral aneurysms, which were treated either with an endovascular procedure, surgically or conservative [9].

We attempt to show the feasibility of the ontology-based research, done by the domain expert without assistance of a data scientist. In this context the domain experts are from the field of neurosurgery and neuro-radiology. The following parameters of the aneurysm were taken into account: the age of patient at diagnosis, number of aneurysms in total for this patient, the number of recorded clinical events (complications), the number of surgical treatments, the number of endovascular treatments, as well as the width, depth, height, neck and size of an aneurysm. It was not aim of this evaluation to discover new medical knowledge, but rather to verify the method by being able to demonstrate already known knowledge about the data.

The result in form of the loadings plot is shown in Fig. 1. It shows the first two principal components (PC) with the highest percentage of explained variance. The first PC is displayed on the x-axis and the second PC is shown on the y-axis. It is apparent from this plot that there is a strong relationship between the *Width*, *Depth* and *Height* of aneurysms, as they are located close to one another. From a medical point of view, this is obvious, since aneurysms are spheric in most of the cases. Another variable is in the surrounding of this variable cluster, namely the variable *Neck*. The neck describes the diameter of the opening of the aneurysm to the supplying blood vessel. Here again, a correlation is indicated by the nature of the matter. The bigger the aneurysm is in all its dimensions, the bigger the neck tends to be. On the other principal component, the opposing position of the number of endovascular treatments on the one hand and surgical treatments on the other side is appealing. This is given due to the fact, that most aneurysms are either treated the one or the other way. The position of the variable *Age of Patient at Diagnosis* very close to the center of the visualization indicates that there are hardly any correlations between this variable and the others and this variable has no influence on the shape of the data cloud.

While the previous observations were easily explained with already known facts, the opposing position of the width-depth-height-cluster on the left-hand-side of the first PC and the variable *Total Number of Aneurysms for Patient* on the right-hand-sind struck the attention of the medical researchers. Preceding visualization with other methods already indicated a (weak) reverse correlation between the total number of aneurysms a patient suffers from and the size of

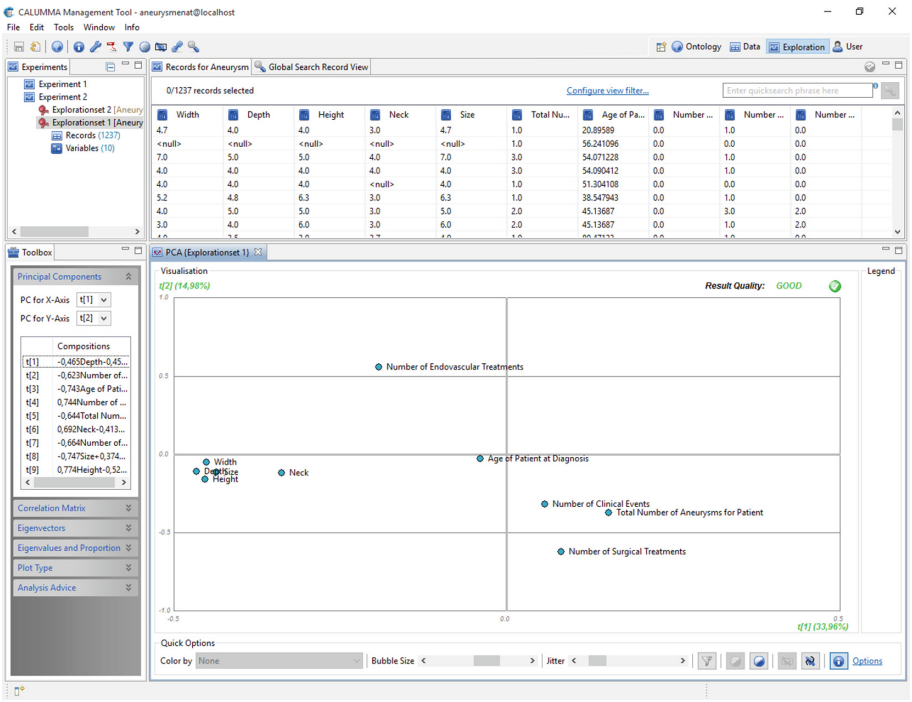

Fig. 1. A two dimensional loadings plot of the aneurysm data set, embedded in an ontology-guided data infrastructure. The x-axis represents the first principal component, expressing 33.96 % of the total variance. The second principal component conveys 14.98 %, depicted by the y-axis.

these aneurysms. All methods, including this PCA-run, showed evidence, that patients with numerous aneurysms tend to have smaller ones. This phenomenon will now be investigated. This is a very good example for what the ontology-guided approach for the doctor-in-the-loop is able to do. It allows the researching domain experts to explore their complex data and generate new hypotheses for subsequent research.

The automatic calculation of the relevant key figures indicate that this PCA-run yielded an acceptable and meaningful result. The covered percentage of the variance is acceptable and colored in green. The result of this key number calculation and interpretation is visualized in the upper right corner of the visualization and giving an indication to the researcher how reliable this output is.

5 Discussion and Conclusion

For considerably complex mathematical methods results cannot be interpreted unambiguously at first glance, contrary to a simple bar chart or box plot. However, this is all the more important to provide guidance throughout data process-

ing actions. The generated numerical output of the principal components method is conclusive for mathematical experts. By contrast, domain experts with basic mathematical and technical knowledge can neither see any immediate use regarding eigenvectors and eigenvalues, nor are they able to assess the significance of the output. This is precisely the point where assistance of an ontology-guided data infrastructure takes effect. Detecting relationships and correlations can be much more simplified by visualizing the results of the principal component method. Thus, Fig. 1 is quite revealing, as the visualized eigenvectors (loadings) substantially better illustrate strong relationships between the three variables (*Width*, *Depth*, *Height*) than a non-guided numerical output. As suspected, they are situated close together in the loadings plot, since aneurysms are rather circular in almost all cases.

This research sheds new light on the support for domain experts in mathematical and technical issues through smooth guidance of data exploration. The example of applying PCA pursues the objective to reveal previously unsuspected relationships when the number of input variables is supremely high. This way of ontology-guided data preprocessing is considered as an intermediate step in (medical) data analysis and requires extensive mathematical skill and knowledge. Only few systems are capable of intelligent assistance for guiding the medical domain expert through data analysis in an acquainted data infrastructure.

Quick and effortless access to different statistical and mathematical methods and tools often represents the fundamental challenge for medical experts due to the lack of comprehensive technological knowledge. Initially, it becomes necessary to give domain experts an understanding of the variety of available methods. Even if an appropriate method has been found, the major obstacle is the related realization, provided that the researcher is aware of mathematical science.

Not all results of a PCA-run are equally good and meaningful. It is very dangerous to use the PCA without further exploration of some statistical key numbers. The visualization tries to bring the result of the automatic calculation of these key features to the user. The percentage of the covered variance is colored, in a range from green (acceptable) over orange to red (unacceptable) (see Fig. 1). These two and the other key numbers are summarized in the info field *Result Quality* in the upper right corner of the visualization. There is of course no clear cut between the qualities of PCA results, but the sum of key numbers and their coloring provides guidance to the user to interpret the results. For very inexperienced users it provides a first barrier to use PCA results without any control of the key figures. This aspect clearly distinguishes the PCA from our preceding attempts (e.g., [9]) to make complex data mining and data visualization algorithms accessible to researching domain experts.

We realized that not all data mining and data visualization algorithms are meant to be used by non-data-scientists. We consequently try to push the technical barrier towards complex methods and algorithms in order to enable the biomedical domain experts to take advantage of them. Thus far, the results of these methods (non-linear mapping, parallel coordinates, etc.) were easy to interpret with limited danger to mis-interpretation. Here, in the case of PCA even

very promising looking visualization might be completely worthless, and without checking the corresponding key-figures an interpretation is not possible. Only an intelligent research-platform designed for domain-expert-driven knowledge discovery can help by automatically calculating these key-figures and bringing them to a prominent position in the user interface.

6 Open Challenges and Future Work

Since the feasibility for PCA only subsists as numerical attributes are used, a small part of variables can be taken into consideration. Further work is required to establish the viability of extensive and automated ontology-based data preprocessing. Thus, especially in the medical domain, information is stored in categorical or boolean attributes. Extending the data infrastructure will lead to a PCA for categorical variables (multiple correspondence analysis, factor analysis for mixed data).

References

1. Akgul, C.B., Rubin, D.L., Napel, S., Beaulieu, C.F., Greenspan, H., Acar, B.: Content-based image retrieval in radiology: Current status and future directions. J. Digit. Imaging **24**(2), 208–222 (2011)
2. 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. J. Am. Med. Inf. Assoc. **14**(4), 478–488 (2007)
3. Atzmüller, M., Baumeister, J., Puppe, F.: Introspective subgroup analysis for interactive knowledge refinement. In: Sutcliffe, G., Goebel, R. (eds.) FLAIRS Nineteenth International Florida Artificial Intelligence Research Society Conference, pp. 402–407. AAAI Press, Menlo Park (2006)
4. Buchan, I.E., Winn, J.M., Bishop, C.M.: A unified modeling approach to data-intensive healthcare. In: Hey, T., Tansley, S., Tolle, K. (eds.) The fourth paradigm: Data-Intensive Scientific Discovery, pp. 91–98. Microsoft Research, Redmond (2009)
5. Cios, K.J., William Moore, G.: Uniqueness of medical data mining. Artif. Intell. Med. **26**(1), 1–24 (2002)
6. Gigerenzer, G., Gaissmaier, W.: Heuristic decision making. Ann. Rev. Psychol. **62**, 451–482 (2011)
7. Girardi, D., Dirnberger, J., Giretzlehner, M.: An ontology-based clinical data warehouse for scientific research. Saf. Health **1**(1), 1–9 (2015)
8. Girardi, D., Kueng, J., Holzinger, A.: A domain-expert centered process model for knowledge discovery in medical research: putting the expert-in-the-loop. In: Guo, Y., Friston, K., Aldo, F., Hill, S., Peng, H. (eds.) BIH 2015. LNCS, vol. 9250, pp. 389–398. Springer, Heidelberg (2015)
9. 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 Inf., 1–11 (2016). (Online First Articles)

10. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
11. Holzinger, A.: Human-computer interaction and knowledge discovery (HCI-KDD): what is the benefit of bringing those two fields to work together? In: Cuzzocrea, A., Kittl, C., Simos, D.E., Weippl, E., Xu, L. (eds.) CD-ARES 2013. LNCS, vol. 8127, pp. 319–328. Springer, Heidelberg (2013)
12. Holzinger, A.: Trends in interactive knowledge discovery for personalized medicine: Cognitive science meets machine learning. IEEE Intell. Inf. Bull. **15**(1), 6–14 (2014)
13. Holzinger, A.: Interactive machine learning for health informatics: When do we need the human-in-the-loop? Springer Brain Inform. (BRIN) **3**, 1–13 (2016). http://dx.doi.org/10.1007/s40708-016-0042-6
14. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge discovery and interactive data mining in bioinformatics - state-of-the-art, future challenges and research directions. BMC Bioinform. **15**(S6), I1 (2014)
15. Holzinger, Andreas, Stocker, Christof, Dehmer, Matthias: Big complex biomedical data: towards a taxonomy of data. In: Obaidat, Mohammad S., Filipe, Joaquim (eds.) ICETE 2012. CCIS, vol. 455, pp. 3–18. Springer, Heidelberg (2014)
16. Hotelling, H.: Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. **24**, 417–441 (1933)
17. Hund, M., Bhm, 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: Making sense of complex datasets with the doctor-in-the-loop. Brain Inf. **3**, 1–15 (2016)
18. Kessler, W.: Multivariate Datenanalyse: für die Pharma-Bio- und Prozessanalytik. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim (2007)
19. Kurgan, L.A., Musilek, P.: A survey of knowledge discovery and data mining process models. Knowl. Eng. Rev. **21**(01), 1–24 (2006)
20. Malinowski, E.: A thesis in two parts: application of factor analysis to chemical problems. Stevens Inst. Technol. **2**(1–2), 54–94 (1961)
21. Nandi, D., Ashour, A.S., Samanta, S., Chakraborty, S., Salem, M.A., Dey, N.: Principal component analysis in medical image processing: a study. Int. J. Image Min. **1**(1), 65–86 (2015)
22. National Center for Biotechnology Information: Mesh search for principalcomponent analysis and medicine (2016). http://www.ncbi.nlm.nih.gov/
23. Niakšu, O., Kurasova, O.: Data mining applications in healthcare: research vs practice. Databases Inf. Syst. BalticDB&IS **2012**, 58 (2012)
24. NIH: Cerebral Aneurysm Information Page (April 2010). http://www.ninds.nih.gov/disorders/cerebral_aneurysm/cerebral_aneurysm.htm
25. Pearson, K.: On lines and planes of closest fit to systems of points in space. Philos. Mag. **2**, 559–572 (1901)
26. Rencher, A.: Methods of Multivariate Analysis. Wiley Series in Probability and Statistics. Wiley, Chichester (2002)
27. Sharaf, M., Illman, D., Kowalski, B.: Chemometrics. Wiley, New York (1986)
28. Thurstone, L.: Multiple-factor Analysis: A Development and Expansion of The Vectors of Mind. The university of Chicago committee on publications in biology and medicine. University of Chicago Press, New York (1947)
29. Thurstone, L., Thurston, T.: Factorial Studies of Intelligence. Psychometrika monograph suplements. The University of Chicago press, Chicago (1941)

30. Wang, B.B., Mckay, R.I., Abbass, H.A., Barlow, M.: A comparative study for domain ontology guided feature extraction. In: Proceedings of the 26th Australasian Computer Science Conference vol. 16, pp. 69–78. Australian Computer Society, Inc. (2003)