Advances in interactive Knowledge Discovery (15w2181 - July, 24–26, 2015)

Andreas Holzinger[1], Randy Goebel[2], Vasile Palade[3], Massimo Ferri[4], Katharina Morik[5], Sibylle Hess[6], Katharina Holzinger[7], Nitesh Chawla[8], Yuzuru Tanaka[9], Mateusz Juda[10], Mirko Cesarini[11], and Sou-Cheng Choi[12]

[1] Research Unit HCI-KDD, Inst. for Medical Informatics, Medical University Graz, and Center for Biomarker Research, Austria
[2] Centre for Machine Learning, University of Alberta, Canada
[3] Cogent Computing Applied Research Centre, Coventry University, UK
[4] Vision Mathematics Group, Department of Mathematics, University Bologna, Italy
[5] Artificial Intelligence Group, Technical University Dortmund, Germany
[6] SFB 876, Technical University Dortmund, Germany
[7] Faculty of Natural Sciences, Karl-Franzens University Graz, Austria
[8] Interdisciplinary Center for Network Science, University of Notre Dame, US
[9] Meme Media Laboratory, Hokkaido University, Sapporo, Japan
[10] Mrozek Group, Jagiellonian University, Krakow, Poland
[11] Department of Statistics, University of Milano Bicocca, Italy
[12] NORC University of Chicago, and Illinois Inst. of Technology, US

Abstract. A worldwide challenge in health systems is how to deal with big data. The trend towards personalized medicine has resulted in an explosion in the amount of complex data. The BIRS Workshop 15w2181 was dedicated to stimulate the cross-domain integration and appraisal of different fields towards tackling the grand challenge of discovering knowledge in complex data sets. The workshop brought together 10 international well-known researchers and two young female students, from Canada, Europe, Japan, and the United States, with diverse backgrounds, complementary competencies, but common interests and a shared vision: to make sense of big and complex data sets. The workshop particularly tried to contribute advancements in promising novel areas at the intersection of machine learning and topological data science.

1 Introduction and Motivation for Research

A huge problem in health systems worldwide is how to deal with the enormously increasing data sets. However, the volume of this data is only one aspect, the problem is that experts in the life sciences are confronted with very complex, heterogeneous, high-dimensional, vague, incomplete, noisy, and weakly structured data sets [1], [2], [3], and large amounts of unstructured information [4].

This so-called "big data" problem in the medical domain is driven by the trend towards precision P4-medicine (Predictive, Preventive, Participatory, Personalized medicine)[5], and has resulted in an explosion in the amount of gen-

erated data sets, in particular "-omics" data, for example from genomics, proteomics, metabolomics, but also epigenetics, transcriptomics, lipidomics, fluxomics, phenomics, microbiomics, etc. [6]. The trend is in moving from a reactive to a proactive medicine, closely related to systems approaches to disease, heavily data driven, hence machine learning approaches are indispensable [7].

The well-known challenges with such data include the complexity of feature dimensions (scaling and mapping problems), the heterogeneity of the data (problems of data integration, data fusion, data curation, data mapping), the change over time, and the typical (bio-)medical data problem: uncertainty of the data, bad quality, false, incomplete, missing data and the constant danger of modelling artifacts.

For the application of machine learning methods, the often mentioned problem of large amounts of data, is rather an advantage: Big data actually can provide benefits, because usually with only a few hundred training examples, there is the danger of random guessing; but using millions of training examples will raise the precision.

The issue of large data sets connects to the question: "What constitutes predictable structures in the world?" as something might be predictable – but not comprehensible [8]. Machine learning researchers study algorithms being capable of learning from data and because learning is an important aspect of intelligent behavior, thus machine learning has become a modern and central aspect of research in artificial intelligence. The most obvious example of learning occur in humans, so there is a natural bridge between research in machine learning and cognitive science, which is strongly related to HCI.

There is a paradigmatic shift from classical science, where the expert first states the question and then collect the data, to data science, where the expert first collects the data and then ask questions. A central challenge in this new approach is to ask relevant questions so to find relevant *structural* patterns and/or *temporal* patterns ("knowledge") in such data, because those are often hidden and not directly accessible to the expert [9].

Recent research in data mining has concentrated on developing efficient machine learning algorithms for analyzing big data sets arising in many domains of real life, and especially those in the biomedical domain, such as for discovering patterns in biomedical data, or for modelling uncertainties in clinical decision making. With the increasing volume of biomedical and health care data becoming available everyday, structured learning and graphical models, such as probabilistic dependency networks, probabilistic decision trees, Bayesian networks and Markov Random Fields, are becoming popular tools in biomedical data mining research. Many problems in the area can be formulated as probabilistic inference problems. In addition, other machine learning approaches, including modern developments in clustering evaluation and alternative learning methods deriving from the area of computational intelligence area, especially deep learning, have found their application to data mining problems on Big Data.

2 Presentations

Andreas Holzinger: *Challenges of biomedicine, health and the life sciences and the chances of Interactive Machine Learning for Knowledge Discovery.* Andreas opened the workshop with an overview of the variations and complexity of data sets from biomedicine, health care and the life sciences and the problems and challenges biomedical researchers of today are faced, when trying to gain insight into their data to discover unknown unknowns. He emphasized that machine learning algorithms are indispensable, and a best practice today is demonstrated by autonomous vehicles ("Google car"). However, in complex domains such as biomedicine, where we deal with uncertain, probabilistic, noisy, incomplete and weakly structured data of high-dimensions the application of fully automatic machine learning algorithms endangers the modelling of artifacts. Consequently, Andreas accentuated the importance of supporting human intelligence with *interactive* machine learning by application of active learning algorithms, thereby putting the human-in-the-loop. The long-long term goal of the Holzinger group is to contribute towards cognitive computing systems, that learn and interact naturally with experts together to extend what neither a human nor a computer could do on its own.

Randy Goebel: *The role of logic and machine learning within a general theory of visualization.* Randy pointed out that the role of logic and machine learning in visualization is not familiar to many colleagues, but the idea of visual inference requires inductive transformations from base data to visual data. Randy emphasized that these transformations need to be constrained by inference principles, including the construction of layers of knowledge, which generally are difficult to construct by hand. The idea is to describe how logic, learning, and visualization are connected, in order to help enable humans to make better inferences from growing volumes of data in every area of application. This is highly relevant for the biomedical domain.

Vasile Palade: *Class Imbalance Learning.* Vasile demonstrated that class imbalance of data is commonly found in many data mining tasks and machine learning applications to real-world problems. When learning from imbalanced data, the performance measure used for model selection plays a vital role. The existing and popular performance measures used in class imbalance learning, such as the Gm and Fm, can still result in sub-optimal classification models. Vasile first presented a new performance measure, called the Adjusted Geometric-mean (AGm), which overcomes the problems of the existing performance measures when learning from imbalanced data. Support Vector Machines (SVMs) has become a very popular and effective machine learning technique, but which can still produce sub-optimal models when it comes to imbalanced data sets. Vasile presented then FSVM-CIL (Fuzzy SVM for Class Imbalance Learning), an effective method to train FSVMs with imbalanced data in the presence of outliers and noise in the data. Finally, Vasile discussed some efficient re-sampling methods for training SVMs with imbalance data in the context of applications.

Katharina Morik: *Big Data and Small Devices.* Katharina showed that big data are produced by various sources. Most often, they are distributedly stored

at computing farms or clouds. Analytics on the Hadoop Distributed File System (HDFS) then follows the MapReduce programming model (batch layer). It is complemented by the speed layer, which aggregates and integrates incoming data streams in real time. Katharina emphasized that when considering big data and small devices, obviously, we imagine the small devices being hosts of the speed layer, only. Analytics on the small devices is restricted by memory and computation resources. The interplay of streaming and batch analytics offers a multitude of configurations. The collaborative research center SFB 876 investigates data analytics for and on small devices regarding runtime, memory and energy consumption. Katharina investigated in her talk graphical models, which generate the probabilities for connected (sensor) nodes. Resource-restricted methods deliver insights fast enough for a more interactive analysis.

Sibylle Hess: *Investigation of Code Tables to compress and describe the underlying characteristics of binary databases.* Sibylle is a young Computer Science student and inspected the spectrum of methods (from frequent pattern mining to numerical optimization) to extract the pattern set that describes a binary database best. Invoking the Minimum Description Length (MDL) principle, this objective can be stated as: find the code table that compresses the database most. Sibylle pointed out that a particularly interesting interpretation of this task, relating it to biclustering, arises from the formulation as a matrix factorisation problem. Finally, Sibylle stressed that biclustering has a variety of applications in research fields such as collaborative filtering, gene expression analysis and text mining. She emphasized that the derived matrix factorisation analogy provides a new perspective on distinct data mining subfields (unifying biclustering and pattern mining concepts such as Krimp), initialising a cross-over of their applications and interpretations of derived models.

Katharina Holzinger: *Darwin, Lamarck, Baldwin, Mendel: What can we learn from them?*. Katharina is a young student of Natural Sciences and discussed the potential of evolutionary algorithms, inspired by biological mechanisms observed in nature, such as selection and genetic changes, to find the best solution for a given optimisation problem. Contrary to Darwin, and according to Lamarck and Baldwin, organisms in natural systems learn to adapt over their lifetime and allow to adjust over generations. Whereas earlier research was rather reserved, more recent research underpinned by the work of Lamarck and Baldwin, finds that these theories have much potential, particularly in upcoming fields such as epigenetics. Katharina emphasized particularly the integration of the Theories of Gregor Mendel, which could be helpful for knowledge discovery.

Nitesh Chawla: *Big Data and Small Data for Personalized and Population Health care.* Nitesh showed that proactive personalized medicine can bring fundamental changes in health care. He asked the question: "Can we then take a data-driven approach to discover nuggets of knowledge and insight from the big data in health care for patient-centered outcomes and personalized health care?" and Nitesh asked if we may answer the question: "What are my disease risks and how to best manage it?". Particularly he pointed out the importance of the question: "How to scale this at the population level?". Nitesh discussed some work

of his group that takes the data and networks driven thinking to personalized health care and patient-centered outcomes. He demonstrated the effectiveness of population health data to drive personalized disease management and wellness strategies, and in effect impacting population health. Nitesh also shared various pilots under-way that take the algorithms and tools on a "road-show".

Yuzuru Tanaka: *Exploratory Visual Analytics for the Discovery of Complex Analysis Scenarios for Big Data.* Yuzuru emphasized that the data-centric approach is increasing its significance in varieties of scientific research areas and large-scale social cyber-physical systems. Yuzuru showed examples from disparate areas: Biomedical research and urban-scale winter road management. Through his involvement in three major projects on these subjects, Yuzuru recognized a big gap between the state-of-the-art big data core technologies and both the data-centric research for the analysis of clinico-genomic trial data and the big data approach to the optimization of social system services. During the last couple of decades, the enabling core technologies for big data analysis have made remarkable advances in both analysis and management technologies. Yuzuru pointed out that we still lack methodologies to find out the best analysis scenario for finding out such solutions as personalized medicines or optimized snow removal plans from a given clinico-genomic trial data set or from given traffic and weather data sets. He also proposed exploratory visual analytics to support analysts to find out complex analysis scenarios, and the coordinated multiple views and analyses framework as its application framework.

Mateusz Juda: *Homology of big data - algorithms and applications.* Mateusz started with demonstrating homology as a well known and powerful tool in pure mathematics and he stressed that for many years it was impossible to use this tool in applied science because of data size and cubical algorithms for computing homology. He explained that new preprocessing methods give us now a possibility to apply homology for real data. Discrete Morse theory is an example of such a tool, which simplifies data without changing its topological information. Mateusz introduced discrete Morse theory and its application to homology computations and he showed how to construct a discrete vector field (Morse matching) using parallel and distributed algorithms. Mateusz also showed an application of this tool to knots detection and classification in a biological context.

Massimo Ferri: *Persistent topology for natural shape analysis and image retrieval.* Massimo emphasized that data are more and more often of "natural" origin (pictures or 3D meshes representing living beings, faces, handwritten words, hand-drawn sketches etc.). He pointed out that classical mathematical techniques do not fit well the task of analyzing, comparing, classifying, retrieving such data. On the contrary topology (and in particular algebraic topology) is, by its very nature, the part of mathematics which formalizes qualitative aspects of objects; therefore topological data processing and topological data mining well integrate with more classical mathematical tools. Massimo then concentrated on persistent homology, which combines geometry and algebraic topology in the study of pairs (X,f) where X is an object (typically a topological space) and f is a continuous function defined on X (typically with real values). One ap-

plication is the extraction of topological features of an object out of a cloud of sample points. Another class of applications uses f as a formalization of a classification criterion; in this case various functions can give different criteria, cooperating in a complex classifier. Massimo explained that persistent homology is studied by several teams throughout the world both from the theoretical and computational viewpoints and has already given rise to several applications: dermatological diagnosis, evolution of hurricanes, signature recognition, gesture recognition; retrieval of trademarks, 3D meshes, hand-drawn sketches etc.

Mirko Cesarini: *Data Quality in Schema free (big) data.* Mirko focused in his presentation on the challenges and open problems emerging when complex data sets are used to obtain insights about a population e.g., analysing job offers using data from web job boards, inspecting the job history of the working population (starting from administrative records), and analysing cellular network traffic. He pointed out that a huge set of weakly structured data can be derived from information sources containing a variety of data types. Mirko explained that in such a context, techniques ranging from formal methods to machine learning can identify and exploit information structures (both hidden and visible) to check data consistency, to ameliorate the data (e.g., fixing inconsistencies), and to create synthetic representations of the original data.

Sou-Cheng Choi: *Machine Learning for Machine Data in Computational Social Sciences.* In this last talk of the workshop, Sou-Cheng presented machine learning and high-accuracy prediction methods of rare events in semi-structured or unstructured log files produced at high velocity and high volume by NORC's computer-assisted telephone interviewing network. These machine log files are generated by their internal Voxco Servers for a telephone survey. Sou-Cheng and her colleagues adapt natural language processing (NLP) techniques and data-mining methods to train powerful learning and prediction models for error messages in the absence of source code, updated documentation, and relevant dictionaries. Such approaches can be useful for applications in other domains, e.g. the biomedical domain.

3 Challenges, Open Problems and Future Outlook

Big challenges in today's biomedical domain are in the development of new methods, algorithms and tools for the effective analysis and interpretation of complex biomedical data [10]. Within such data sets, relevant structural and/or temporal patterns ("knowledge") are often hidden, difficult to extract, thus not directly accessible to a biomedical expert. Consequently, a major challenge is in interactive Knowledge Discovery and Data Mining which relies heavily on machine learning approaches. However, many of the classical methods are based on the assumption that the data objects under consideration are represented in terms of feature vectors, or collections of attribute values; Bunke (2003) [11], for example, argued that graphs have a representational power that is significantly higher than the representational power of feature vectors. Moreover, graph-theory provides powerful tools to map data structures and to find novel connections between

data objects [12] and allow the application of statistical and machine learning techniques [13].
Methods from computational geometry and algebraic topology may also be of great help [14], and could be combined with machine learning approaches, e.g. evolutionary algorithms [15], [16]. Promising future research routes in this field are in interactive visual data mining together with graph-based data analysis [17], [18]. Another benefit of a graph-based data structure is in the applicability of methods from network topology and network analysis as well as data mining, e.g. small-world phenomenon [19] and cluster analysis [20], to mention only two.

3.1 Persistence

Similarity is a concept which sounds very natural to a human being, but is very difficult to formalize for use in a machine; it is anyway of paramount importance in data retrieval and data mining. The most widely accepted formalization is by defining a group of transformations (of the image plane, of the space where a 3D shape is embedded, of a parameter space etc.) such that two objects, which can be mapped into each other by a transformation of the group, are considered to be similar. The classical geometrical transformation groups adopted in Computer Vision and Pattern Recognition generally suffer from a rigidity, which can only be smoothened by a heavy use of statistics. Topological transformations (*homeomorphisms*), on the other hand, are by their very nature too "free". Another problem is that different observers might have different similarity concepts depending, e.g., on their specific tasks. Finally, "human" perception of similarity is obviously limited to 2D and 3D sensorial experience, whereas geometry and topology extend their reach to any dimension.

Applications suggested to adapt topology — and in particular its branch called *homology* — to take the observer's viewpoint into account and to restrict consequently the set of transformations. This started in the 90's with the work of P. Frosini in Bologna [21, 22], with what was then known as *size functions theory*. The main idea was to convey the observer's viewpoint into a function, called *measuring* (now *filtering*) function. The object is no more only a set — or, more precisely, a topological space — but a pair (X, f) of a space X and a function f defined on X and with real values. The filtration given by the sublevel sets of f then moderates the "freedom" of the topological setting; moreover, the possible use of different functions on the same space gives the method a powerful modularity. Modularity and freedom have proved to make this theory a winner in shape classification and retrieval, when it comes to shapes of a natural origin (tree leaves, leukocytes, handwritten letters, signatures, tumor cells, gestures, face profiles, echocardiographic sequences, cyclones, etc.).

The theory was extended around year 2000 by H. Edelsbrunner (then at Duke University, now at the Inst. of Science and Technology of Austria) [23] and later developed also by G. Carlsson's team (in Stanford) [24] by what is called *persistent homology*; this theory is gaining much resonance in the applied mathematics community, and is the main stream of the ESF project ACAT (Applied and Computational Algebraic Topology) gathering 13 national teams.

A thorough survey on (1-dimensional) persistence is [25]. In the last few years the extension of the theory to filtering functions with multidimensional range is the object of a hard investigation [26–29].

Keystones of persistent homology are:

- *Persistence diagrams* For each nonnegative integer i, there is a persistence diagram consisting of a set of points in the plane, which condenses the essential information on the pair (X, f) (through the homology modules of degree i of the sublevel sets of the filtration given by f). More specifically, a persistence diagram carries the information on the *Persistent Betti Number* function: This associates, to each pair of real numbers $u \leq v$, the dimension of the image of the i-homology morphism induced by the inclusion of the sublevel set under u into the one under v. Size functions are the particular case of Persistent Betti Number functions for $i = 0$, which is actually the most important one for practical applications.
- *Natural pseudodistance* Given pairs (X, f) and (Y, g), there is a way to measure how much a homeomorphism from X to Y distorts the filtrations given by f and g. The minimum of such measures among all possible homeomorphims is a (pseudo)distance which formalizes the dissimilarity of the two pairs by a number.
- *Matching distance* Given the persistence diagrams (of the same degree i) of two pairs (X, f) and (Y, g), there is a well-defined distance between the diagrams which is an optimal lower bound for the natural pseudodistance of the pairs. Computation of this distance is essential for retrieval, but implies a very heavy computational burden, due to what is known as combinatorial explosion.

Challenges in Persistence Multidimensional ranges for filtering functions may be crucial for applications. In fact, cooperation of different viewpoints on the same problem (expressed by different filtering functions $f_1, \ldots, f_h$) has proved to be very successful, but that was done just by building different classifiers and mixing their outputs by fuzzy logic. Now it is possible to have just one classifier using a single multidimensional filtering function whose components are the old separate ones $f_1, \ldots, f_h$. There are examples showing that shape comparison by a multidimensional filtering function is finer than the use of the separate components [30, 31]. There are anyway theoretical and computational hurdles that the community is trying to overcome.

The matching distance is too heavy to compute in data mining. Preprocessing by a sloppier but much faster distance is necessary. This seems to be the case if persistence diagrams are transformed into complex polynomials and the distance is computed on their coefficients; preliminary results support this strategy [32].

Most important, the choice of invariance groups and an interactive selection of filtering functions (or components thereof) are a special benefit of persistence;

they promise to be of great advantage for relevance feedback in a human-in-the-loop scheme.

3.2 Evolutionary algorithms for big data processing

Evolutionary computing algorithms can be used for solving various optimization tasks as part of the solution for complex problems that involve big data and high dimensionality. Another recent employment of evolutionary algorithms [33], with very good potential application to big data processing, is to solve data sampling problems. Sampling is a basic and one of the most important tasks in data processing, statistics and machine learning, and acquiring good samples is not an easy task for an arbitrary probability distribution of data or when the data space is huge. Evolutionary Sampling proposed in [34] combines the popular rejection sampling method with other strategies within a probabilistic framework in order to obtain an optimal approximation of any pointwise computable density function by using finite samples, which is a fundamental problem in statistics and machine learning area. The paper also argues that many machine learning problems can be described as, or could be converted into corresponding, density function approximation problem problems, where the Evolutionary Sampling approach can be employed for training and as a machine learning method not as a mean for acquiring data only. Theoretical and experimental studies have demonstrated that the Evolutionary Sampling learning can be used to solve many practical application problems which can be expressed as density function approximation problems within a probabilistic framework.

Challenges and Future Work

- Investigating the role of evolutionary computing techniques when dealing with optimization and learning problems involving big data, such as for dynamic and very high dimensional problems, multi-objective big data analytics problems, or big data driven optimization of complex systems, is a very good and promising avenue for future research, which has not been properly tackled yet in the literature.
- Developing new nature inspired evolutionary algorithms. Investigating the performance (convergence and time complexity) of our previous swarm intelligent algorithms, such as the Random Drift Particle Swarm Optimization [35], and Quantum Particle Swarm Optimization variants [36], and apply them to big data and complex optimisation problems from bioinformatics and computational biology.
- Investigate the performance of the Evolutionary Sampling method in solving big data problems, especially from the biological and medical domain.

3.3 Graphical models for big data

Probabilistic and graphical models are popular tools to use when processing large scale data sets today, and Bayesian networks is undoubtedly the most common

approach within this family. However, inferring large Bayesian networks from data, such as inferring Genetic Regulatory Networks from genomic time series data, is a very challenging task in machine learning today. The current algorithms for inferring a Bayesian network from data, irrespective of the application area involved, work for small networks, of less than 100 nodes or so. But, the smallest networks we find in biology and medicine are at least an order of magnitude bigger; for example, the genetic regulatory network for the smallest organisms in nature and which are commonly studied by researchers in biology today has at least several thousands of nodes.

Challenges and Future Work

- Developing efficient algorithms for inferring large Bayesian networks from data, by extending the work in [37].
- Investigating the performance of the Dense Structural Expectation Maximisation algorithm proposed in [37] for large network inference problems, especially on networks generated from biology area, such as genetic regulatory networks, but not only.

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., Stocker, C., Dehmer, M.: Big complex biomedical data: Towards a taxonomy of data. In Obaidat, M.S., Filipe, J., eds.: Communications in Computer and Information Science CCIS 455. Springer, Berlin Heidelberg (2014) 3–18
3. Holzinger, A.: Biomedical Informatics: Discovering Knowledge in Big Data. Springer, New York (2014)
4. Holzinger, A., Stocker, C., Ofner, B., Prohaska, G., Brabenetz, A., Hofmann-Wellenhof, R.: Combining hci, natural language processing, and knowledge discovery - potential of ibm content analytics as an assistive technology in the biomedical domain. In: Springer Lecture Notes in Computer Science LNCS 7947. Springer, Heidelberg, Berlin, New York (2013) 13–24
5. Donsa, K., Spat, S., Beck, P., Pieber, T.R., Holzinger, A.: Towards personalization of diabetes therapy using computerized decision support and machine learning: some open problems and challenges. In Holzinger, A., Roecker, C., Ziefle, M., eds.: Smart Health, Lecture Notes in Computer Science LNCS 8700. Springer, Heidelberg, Berlin (2015) 235–260
6. Huppertz, B., Holzinger, A.: Biobanks a source of large biological data sets: 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. Springer, Berlin, Heidelberg (2014) 317–330
7. Holzinger, A.: Trends in interactive knowledge discovery for personalized medicine: Cognitive science meets machine learning. IEEE Intelligent Informatics Bulletin **15** (2014) 6–14
8. Jaekel, F., Schoelkopf, B., Wichmann, F.A.: Does cognitive science need kernels? Trends in cognitive sciences **13** (2009) 381–388

9. Otasek, D., Pastrello, C., Holzinger, A., Jurisica, I.: Visual data mining: Effective exploration of the biological universe. In Holzinger, A., Jurisica, I., eds.: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. Lecture Notes in Computer Science LNCS 8401. Springer, Heidelberg, Berlin (2014) 1934
10. 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(Suppl 6):S1** (2014)
11. Bunke, H.: Graph-based tools for data mining and machine learning. In Perner, P., Rosenfeld, A., eds.: Machine Learning and Data Mining in Pattern Recognition, Proceedings. Volume 2734 of Lecture Notes in Artificial Intelligence. Springer-Verlag Berlin, (Berlin) 7–19
12. Strogatz, S.: Exploring complex networks. Nature **410** (2001) 268–276
13. Dehmer, M., Emmert-Streib, F., Mehler, A.: Towards an Information Theory of Complex Networks: Statistical Methods and Applications. Birkhaeuser Boston (2011)
14. Holzinger, A.: On topological data mining. In Holzinger, A., Jurisica, I., eds.: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. Lecture Notes in Computer Science LNCS 8401. Springer, Heidelberg, Berlin (2014) 331–356
15. Holzinger, K., Palade, V., Rabadan, R., Holzinger, A.: Darwin or lamarck? future challenges in evolutionary algorithms for knowledge discovery and data mining. In Holzinger, A., Jurisica, I., eds.: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. Lecture Notes in Computer Science LNCS 8401. Springer, Heidelberg, Berlin (2014) 35–56
16. Holzinger, A., Blanchard, D., Bloice, M., Holzinger, K., Palade, V., Rabadan, R.: Darwin, Lamarck, or Baldwin: Applying evolutionary algorithms to machine learning techniques. In: The 2014 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2014), IEEE (2014) 449–453
17. Holzinger, A., Jurisica, I.: Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. In Holzinger, A., Jurisica, I., eds.: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. Lecture Notes in Computer Science LNCS 8401. Springer, Heidelberg, Berlin (2014) 1–18
18. Otasek, D., Pastrello, C., Holzinger, A., Jurisica, I.: Visual data mining: Effective exploration of the biological universe. In Holzinger, A., Jurisica, I., eds.: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. Lecture Notes in Computer Science LNCS 8401. Springer, Heidelberg, Berlin (2014) 19–34
19. Albert, R., Barabsi, A.: Statistical mechanics of complex networks. Reviews of Modern Physics **74** (2002) 47–97
20. Makrogiannis, S., Economou, G., Fotopoulos, S., Bourbakis, N.G.: Segmentation of color images using multiscale clustering and graph theoretic region synthesis. IEEE Transactions on Systems Man and Cybernetics Part A: Systems and Humans **35** (2005) 224–238
21. Frosini, P.: Measuring shapes by size functions. In: Intelligent Robots and Computer Vision X: Algorithms and Techniques, International Society for Optics and Photonics (1992) 122–133
22. Verri, A., Uras, C., Frosini, P., Ferri, M.: On the use of size functions for shape analysis. Biological cybernetics **70** (1993) 99–107

23. Edelsbrunner, H., Letscher, D., Zomorodian, A.: Topological persistence and simplification. (2000) 454–463
24. Carlsson, G., Zomorodian, A., Collins, A., Guibas, L.J.: Persistence barcodes for shapes. International Journal of Shape Modeling **11** (2005) 149–187
25. Edelsbrunner, H., Harer, J.: Persistent homology-a survey. Contemporary mathematics **453** (2008) 257–282
26. Frosini, P., Mulazzani, M.: Size homotopy groups for computation of natural size distances. Bulletin of the Belgian Mathematical Society Simon Stevin **6** (1999) 455–464
27. Carlsson, G., Zomorodian, A.: The theory of multidimensional persistence. Discrete & Computational Geometry **42** (2009) 71–93
28. Biasotti, S., Cerri, A., Frosini, P., Giorgi, D., Landi, C.: Multidimensional size functions for shape comparison. Journal of Mathematical Imaging and Vision **32** (2008) 161–179
29. Cerri, A., Di Fabio, B., Ferri, M., Frosini, P., Landi, C.: Betti numbers in multidimensional persistent homology are stable functions. Mathematical Methods in the Applied Sciences **36** (2013) 1543–1557
30. Cagliari, F., Di Fabio, B., Ferri, M.: One-dimensional reduction of multidimensional persistent homology. Proceedings of the American Mathematical Society **138** (2010) 3003–3017
31. Adcock, A., Rubin, D., Carlsson, G.: Classification of hepatic lesions using the matching metric. Computer vision and image understanding **121** (2014) 36–42
32. Di Fabio, B., Ferri, M.: Comparing persistence diagrams through complex vectors. (2015)
33. Xie, Z., Sun, J., Palade, V., Wang, S., Liu, Y.: Evolutionary sampling: A novel way of machine learning within a probabilistic framework. Information Sciences **299** (2015) 262–282
34. Jun, S., Palade, V., Xiao-Jun, W., Wei, F., Zhenyu, W.: Solving the power economic dispatch problem with generator constraints by random drift particle swarm optimization. Industrial Informatics, IEEE Transactions on **10** (2014) 222–232
35. Jun, S., Palade, V., Xiaojun, W., Wei, F.: Multiple sequence alignment with hiddenmarkov models learned by random driftparticle swarm optimization. Computational Biology and Bioinformatics, IEEE/ACM Transactions on **11** (2014) 243–257
36. Sun, J., Fang, W., Palade, V., Wu, X., Xu, W.: Quantum-behaved particle swarm optimization with gaussian distributed local attractor point. Applied Mathematics and Computation **218** (2011) 3763–3775
37. Fogelberg, C., Palade, V.: Dense structural expectation maximisation with parallelisation for efficient large-network structural inference. International Journal on Artificial Intelligence Tools **22** (2013)