Knowledge Discovery from Complex High Dimensional Data

Sangkyun Lee[1(✉)] and Andreas Holzinger[2,3]

[1] Artificial Intelligence Unit LS8, Computer Science Department,
Technische Universität Dortmund, Dortmund, Germany
sangkyun.lee@tu-dortmund.de

[2] Research Unit HCI-KDD, Institute for Medical Informatics,
Statistics and Documentation, Medical University Graz, Graz, Austria
a.holzinger@hci-kdd.org

[3] Institute for Information Systems and Computer Media,
Graz University of Technology, Graz, Austria

Abstract. Modern data analysis is confronted by increasing dimensionality of problems, mainly contributed by higher resolutions available for data acquisition and by our use of larger models with more degrees of freedom to investigate complex systems deeper. High dimensionality constitutes one aspect of "big data", which brings us not only computational but also statistical and perceptional challenges. Most data analysis problems are solved using techniques of optimization, where large-scale optimization requires faster algorithms and implementations. Computed solutions must be evaluated for statistical quality, since otherwise false discoveries can be made. Recent papers suggest to control and modify algorithms themselves for better statistical properties. Finally, human perception puts an inherent limit on our understanding to three dimensional spaces, making it almost impossible to grasp complex phenomena. For aid, we use dimensionality reduction or other techniques, but these usually do not capture relations between interesting objects. Here graph-based knowledge representation has lots of potential, for instance to create perceivable and interactive representations and to perform new types of analysis based on graph theory and network topology. In this article, we show glimpses of new developments in these aspects.

1 Introduction

Thanks to modern sensing technology, we witness rapid increase in data dimensions in numerous domains, for example high-resolution images, large-scale social networks, high-throughput genetic profiles, just to name a few. In most cases, the number of measured entities (features) grows in a much faster rate than the number of observations: pictures taken with smart phones have few million pixels, whereas we may have only few hundreds or thousands of photos.

Our main interest is such "high dimensional" data: to be more specific, a data set is high dimensional when the number of features (p) is larger than

S. Michaelis et al. (Eds.): Morik Festschrift, LNAI 9580, pp. 148–167, 2016.
DOI: 10.1007/978-3-319-41706-6_7

the number of observations (n) by a few magnitude. A good example is gene expression study data.

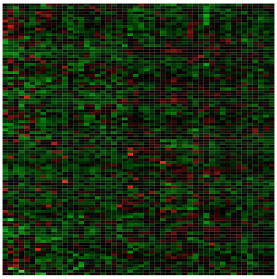


Fig. 1. Gene expression measurement samples of 100 genes (rows) from 50 breast cancer subjects (columns). GEO accession no. GSE11121. (Color figure online)

Figure 1 shows a part of breast cancer data from the Gene Expression Omnibus[1], which contains expression values of $p = 22k$ transcripts measured by the Affymetrix GeneChip Human Genome U133A microarrays. Typically, the number of observations is much smaller in this type of data, due to the cost involved to handle human subjects in a limited time. In the figure, the color represents high (green/bright) or low (red/dark) values of expression, and a primary task using the color intensity values is to identify genes that have different expression patterns in different groups of subjects. Genes with differential expression are then further investigated by wet experiments to identify their roles in biochemical pathways, their relations to other genes, and so forth.

A surprising misconception about high dimensionality is that data analysis would produce better outcome with higher dimensional data, because of increased amount of available information. In a way, this makes sense, for instance we can see objects more clearly in high-resolution digital photographs. In data science, an increased number of input features may allow for building more accurate predictors. However, realizing such predictors comes with extra cost in several aspects.

First, high dimensionality brings computational challenges to data analysis. Obviously, extra memory space will be needed, but also efficient computation algorithms will be required to obtain the best hypothesis for explaining data. The task of finding such a hypothesis is typically described as an optimization problem, where a parametrized function is fitted to data minimizing the mismatch between predictions and observed responses of interest (e.g., categories of objects, severity levels of a disease, etc.)

Secondly, an important task of identifying a (possibly small) subset of features contributing to prediction becomes statistically more challenging as dimension grows. Simply speaking, the reason is that performing multiple hypothesis tests to distinguish important features takes more statistical power, in other words, requires larger sample sizes. There have been quite a few literature on the conditions when we can identify relevant features: later we will discuss some of the recent results on lasso-type regression.

Third, due to limitations in human perception, understanding structures in high dimensional spaces is inherently difficult for us. In particular for interdisciplinary research, the outcome of data analysis would have to be shaped in a form easily perceivable by domain experts who may not be computer scientists. Graph-based representations of data space and analysis outcomes have lots of potential for this purpose: we will demonstrate some examples in biomedical data analysis.

[1] Gene Expression Omnibus http://www.ncbi.nlm.nih.gov/geo/.

2 Sparse Variable Selection and Estimation

There have been a lot of improvements in convex optimization, in particular for dealing with composite objective functions which are interesting for extracting understandable structures from high-dimensional data.

We consider a standard setting for data analysis: a set of m training data points $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ are given, where $\mathbf{x}_i \in \mathcal{X}$ is an input point and $y_i \in \mathcal{Y}$ is a response of interest. Typically, $\mathbf{x}_i$ is a vector and $y_i \in \{-1, +1\}$ for binary classification and $y_i \in \mathbb{R}$ for regression tasks, but both $\mathbf{x}_i$ and y_i can be more structured objects such as strings [51] or trees [38]. A goal of data analysis is to find a function $h_{\mathbf{w}}(\mathbf{x})$ parametrized by a vector $\mathbf{w} \in \mathbb{R}^n$, which best reflects the data in terms of a certain error measure between responses and predictions made by $h_{\mathbf{w}}(\mathbf{x})$. Finding the best parameters vector $\mathbf{w}$ can be formulated as follows,

$$\mathbf{w}^* = \operatorname*{arg\,min}_{w \in \mathbb{R}^n} \; \frac{1}{m} \sum_{i=1}^m \ell(y_i, h_{\mathbf{w}}(\mathbf{x}_i)) + \Psi(\mathbf{w}) = f(\mathbf{w}) + \Psi(\mathbf{w}). \tag{1}$$

Here, $\ell(y_i, h_{\mathbf{w}}(\mathbf{x}_i)) : \mathbb{R}^n \to \mathbb{R}$ is a *loss* function between a prediction $h_{\mathbf{w}}(\mathbf{x}_i)$ and an observed response y_i, which is convex in terms of $\mathbf{w}$. A function $f(\mathbf{w}) : \mathbb{R}^n \to \operatorname{dom} f$ is convex if for all $\mathbf{w}, \mathbf{v} \in \operatorname{dom} f$, the following holds for some $\alpha \geq 0$,

$$f((1-\lambda)\mathbf{w} + \lambda\mathbf{v}) \leq (1-\lambda) f(\mathbf{w}) + \lambda f(\mathbf{v}) - \frac{\alpha}{2}\lambda(1-\lambda)\|\mathbf{w} - \mathbf{v}\|^2.$$

If there exists $\alpha > 0$, f is called α-strongly convex. The second part $\Psi(\mathbf{w}) : \mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ in the objective is a *regularizer*, which is a proper ($\Psi(\mathbf{w}) \equiv +\infty$ is not true) convex function used to control certain statistical properties of the estimation process. Ψ also can be the indicator function of a convex set $\mathcal{W}$, i.e., $\Psi(\mathbf{w}) = 0$ if $\mathbf{w} \in \mathcal{W}$ and $\Psi(\mathbf{w}) = +\infty$ otherwise.

2.1 Sparsity-Inducing Regularization

An intriguing use of the convex minimization in (1) is to extract the most relevant features in data vectors $\mathbf{x}$ that contribute to minimizing the averaged loss. In particular, when a generalized linear model is considered so that $h_{\mathbf{w}}(\mathbf{x}) = f(\langle \mathbf{w}, \mathbf{x} \rangle)$ for a convex function f, where $\langle \cdot, \cdot \rangle$ is an inner product, we can set unimportant components of $\mathbf{w}$ to zero to turn-off their contribution to prediction. Such componentwise switching-off can be achieved by minimizing $\Psi(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$ at the same time, where $\lambda > 0$ is a tuning parameter. With least squares loss function, i.e., $\ell(y_i, h_{\mathbf{w}}(\mathbf{x}_i)) = (y_i - h_{\mathbf{w}}(\mathbf{x}_i))^2$, the problem (1) is called as the lasso problem [66].

Variants. The idea can be extended to incorporate a combination of ℓ_2 and ℓ_1 regularization, i.e., $\Psi(\mathbf{w}) = \lambda\{(1-\alpha)\|\mathbf{w}\|_2^2 + \alpha\|\mathbf{w}\|_1\}$ for some given $\lambda > 0$ and $\alpha \in [0, 1]$. This regularization is called the elastic net [80], which tends to select all correlated features together compared to the selection by lasso where some correlated features may not be selected. In addition, for $\alpha < 1$ the regularizer $\Psi(\mathbf{w})$ makes the objective strongly convex in $\mathbf{w}$, which can lead to better convergence rate e.g. in gradient descent algorithms.

When certain grouping of features is known a priori, then we can use $\Psi(\mathbf{w}) = \sum_{g \in G} \|\mathbf{w}_g\|_2$ for subvectors $\mathbf{w}_g$ of $\mathbf{w} \in \mathbb{R}^n$ corresponding to groups $g \subset \{1, 2, \dots, n\}$. This particular setting is useful when it is preferable to select groups rather than individual components. For instance, a group of binary variables may encode a single multinomial variable of interest. This setting within (1) is known as group-lasso [74]. When groups may overlap, a modified version in [36] is recommended to avoid turning-off all groups sharing a demoted variable. Interested readers can find more details in an introductory article [48].

2.2 Accelerated Proximal Gradient Descent Algorithm

When the convex functions ℓ is smooth (continuously differentiable) and Ψ is possibly nonsmooth but "simple" (the meaning will be clarified later), one of the best algorithm for solving the optimization problem (1) is the accelerated proximal gradient descent algorithm, also known as FISTA [7].

Similarly to the gradient descent, the proximal gradient descent algorithm considers a simple quadratic approximation of the smooth part ℓ in the objective, augmented with Ψ, that is,

$$f(\mathbf{w}) + \Psi(\mathbf{w}) \approx f(\mathbf{w}_k) + \langle \nabla f(\mathbf{w}_k), \mathbf{w} - \mathbf{w}_k \rangle + \frac{L}{2}\|\mathbf{w} - \mathbf{w}_k\|_2^2 + \Psi(\mathbf{w}), \qquad (2)$$

where $L > 0$ is the Lipschitz constant of the gradients ∇f,

$$\|\nabla f(\mathbf{w}) - \nabla f(\mathbf{v})\| \le L\|\mathbf{w} - \mathbf{v}\|_2^2, \quad \forall \mathbf{w}, \mathbf{v} \in \operatorname{dom} f.$$

Given these, the proximal gradient method chooses the next iterate as the minimizer of the right-hand side expression of (2),

$$\begin{aligned} \mathbf{w}_{k+1} &= \arg\min_{\mathbf{w}} \ \langle \nabla f(\mathbf{w}_k), \mathbf{w} - \mathbf{w}_k \rangle + \frac{L}{2}\|\mathbf{w} - \mathbf{w}_k\|^2 + \Psi(\mathbf{w}) \\ &= \arg\min_{\mathbf{w}} \ \frac{1}{2}\|\mathbf{w} - (\mathbf{w}_k - (1/L)\nabla f(\mathbf{w}_k))\|^2 + (1/L)\Psi(\mathbf{v}) \\ &= \operatorname{prox}_{(1/L)\Psi}(\mathbf{w}_k - (1/L)\nabla f(\mathbf{w}_k)). \end{aligned} \qquad (3)$$

Here, we have defined the *proximal operator* associated with a function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ of a given point $\mathbf{z} \in \mathbb{R}^n$ as

$$\operatorname{prox}_h(\mathbf{z}) := \arg\min_{\mathbf{w} \in \mathbb{R}^n} \ \frac{1}{2}\|\mathbf{w} - \mathbf{z}\|^2 + h(\mathbf{w}).$$

From this definition, we can interpret that the update in (3) computes the next iterate $\mathbf{w}_{k+1}$ as a point which is close to the given gradient descent point $\mathbf{z} = \mathbf{w}_k - (1/L)\nabla f(\mathbf{w}_k)$ and minimizes $h = (1/L)\Psi$ at the same time. We call h (or Ψ) is "simple" if the proximal operator can be computed efficiently.

This procedure can be *accelerated* using an ingenious technique due to Nesterov [59]. The modified version uses another sequence of vectors $\mathbf{v}_k$ composed as a particular linear combination of the two past iterates,

$$\mathbf{v}_{k+1} = \mathbf{w}_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{w}_k - \mathbf{w}_{k-1}), \quad t_{k+1} = \frac{1}{2}(1 + \sqrt{1 + 4t_k^2}).$$

Then, the next iterate $\mathbf{w}_{k+1}$ is computed based on $\mathbf{v}_k$, not $\mathbf{w}_k$,

$$\mathbf{w}_{k+1} = \text{prox}_{(1/L)\Psi}(\mathbf{v}_k - (1/L)\nabla f(\mathbf{v}_k))$$

This method generate iterates $\{\mathbf{w}_k\}$ converging to an optimal solution $\mathbf{w}^*$ with the a sublinear rate $\mathcal{O}(1/k^2)$ [7], that is,

$$[f(\mathbf{w}_k) + \Psi(\mathbf{w}_k)] - [f(\mathbf{w}^*) + \Psi(\mathbf{w}^*)] \leq \frac{2L\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{(k+1)^2}.$$

This achieves the best convergence rate as a first-order optimization method [59], and it becomes slower only by a constant factor if line-search is involved.

2.3 Consistency in Variable Selection

One of the important questions regarding the solution $\mathbf{w}^*$ of (1) with ℓ_1 regularization is that if the "true" set of important variables (often called as the support) will be identified. This type of discussion is based on a data generation model that an $m \times n$ training data matrix $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_m^T)$ and responses $\mathbf{y} \in \mathbb{R}^m$ are related by

$$\mathbf{y} = \mathbf{X}\mathbf{w}^\circ + \epsilon$$

where ϵ is a vector of m i.i.d. random variables with mean 0 and variance σ^2. Here, $\mathbf{w}^\circ$ defining the relation is the true weight vector we try to estimate, by a solution $\mathbf{w}^*$ of (1) with $\Psi(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$.

Consistency results has been established first by Knight and Fu [41], for the cases where n and $\mathbf{w}^\circ$ are independent of m and some regularity conditions hold. In estimation consistency, they showed that $\mathbf{w}^* \to \mathbf{w}^\circ$ in probability as $m \to \infty$, and $\mathbf{w}^*$ is asymptotically normal when $\lambda = o(m)$. In variable selection consistency, they also showed that when $\lambda \propto \sqrt{m}$, the true set of important variables are identified in probability, that is,

$$\mathbb{P}(\{i : \mathbf{w}_i^* \neq 0\} = \{i : \mathbf{w}_i^\circ \neq 0\}) \to 1, \quad \text{as } m \to \infty.$$

In high dimensions, the growth of dimensions n is restricted in a way that $s\log(n) = o(m)$, where s is the sparsity of the true signal $\mathbf{w}^\circ$ [56,76]. In addition, other conditions are required for the design matrix $\mathbf{X}$, namely the *neighborhood stability conditions* [56] or the equivalent *irrepresentable conditions* [76,79] that are almost necessary and sufficient for lasso to identify the true support for the cases where n is fixed or n grows with m. Roughly speaking, these conditions state that the irrelevant covariates are orthogonal to relevant ones.

The conditions however may not be satisfied in practice, and finding weaker conditions is in active research, e.g. [37]. Also, more general notions of variable selection consistency have been discussed in other context, e.g. in stochastic online learning [49].

3 Sparse Graph Learning

From a sparse solution $\mathbf{w}^*$ of (1), we can find a set of relevant features, and also can prioritize them by the magnitude of the coefficient vector $\mathbf{w}^*$ for further investigation, e.g. bio-chemical studies of chosen genes to clarify their roles in a complex system. However, its outcome is essentially a ranked list of features which does not tell much about the relations of covariates: the latter type of information would be more helpful to understand the underlying system. In this view, we consider another learning model which produces a graph of features, where connections between nodes (features) represents a certain statistical dependency.

3.1 Gaussian Markov Random Field

The Gaussian Markov Random Field (GMRF) is a collection of n jointly Gaussian random variables represented as nodes in a graph $G = (V, E)$, with a set of n vertices V and a set of undirected edges E. In this model we consider random vectors $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ with a mean vector μ and a covariance matrix Σ, whose probability density is given as

$$p(\mathbf{x}) = (2\pi)^{-n/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right).$$

The edges represent conditional dependency structure: in GRMFs, the variables $\mathbf{x}_i$ and $\mathbf{x}_j$ associated with the nodes i and j are *conditionally independent* given all the other nodes [45] when there is no edge connecting the two nodes, or equivalently the corresponding entry in the precision matrix satisfies $\Sigma^{-1}_{ij} = 0$. That is,

$$\Sigma^{-1}_{ij} = 0 \quad \Leftrightarrow \quad \begin{aligned} &P(\mathbf{x}_i, \mathbf{x}_j | \{\mathbf{x}_k\}_{k \in \{1,2,\ldots,n\} \setminus \{i,j\}}) \\ &= P(\mathbf{x}_i | \{\mathbf{x}_k\}_{k \in \{1,2,\ldots,n\} \setminus \{i,j\}}) P(\mathbf{x}_j | \{\mathbf{x}_k\}_{k \in \{1,2,\ldots,n\} \setminus \{i,j\}}). \end{aligned}$$

This also implies that we can consider the precision matrix Σ^{-1} as a weighted adjacency matrix for an undirected graph representing a GMRF.

3.2 Sparse Precision Matrix Estimation

Assuming that $\mu = \mathbf{0}$ without loss of generality (i.e. subtract the mean from data points), the likelihood function to describe the chance to observe a collection of m i.i.d. samples $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ from $\mathcal{N}(\mathbf{0}, \Sigma^{-1})$ becomes

$$L(\Sigma^{-1}, \mathcal{D}) = \prod_{i=1}^{m} p(\mathbf{x}_i) \ \sim \ \prod_{i=1}^{m} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}\mathbf{x}_i^T \Sigma^{-1} \mathbf{x}_i \right).$$

Therefore the log likelihood function (omitting constant terms and scaling by $2/m$) becomes,

$$LL(\varSigma^{-1}, \mathcal{D}) = \log\det(\varSigma^{-1}) - \mathrm{tr}\,(S\varSigma^{-1}).$$

Here we have defined $S := \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^T$ as the sample covariance matrix.

Minimizing the negative log likelihood plus a sparsity inducing norm on the prediction matrix $\varTheta = \varSigma^{-1}$ can be stated as

$$\min_{\varTheta \in \mathbb{R}^{n\times n}} \quad -LL(\varTheta, \mathcal{D}) + \lambda\|\varTheta\|_1, \quad \text{subject to } \ \varTheta \succ 0,\ \varTheta^T = \varTheta. \tag{4}$$

The ℓ_1 norm of $\varTheta$ here is defined elementwise, that is, $\|\varTheta\|_1 := \sum_{i=1}^{n}\sum_{j=1}^{n} |\varTheta_{ij}|$.

The sparse precision matrix estimation in (4) is a convex optimization problem proposed by Yuan and Lin [75]. Due to its special structure maximizing the determinant of a matrix, they applied an interior point algorithm [68], which may not be suitable for high dimensions n due to the complexity $\mathcal{O}(n^6 \log(1/\epsilon))$ to obtain an ϵ-suboptimal solution. A more efficient block coordinate descent algorithm has been suggested by Banerjee et al. [3], to solve the dual problem of (4). Each subproblem of this block coordinate descent formulation can be cast as a lasso problem in forms of (1), and this fact has been used by Friedman, Hastie, and Tibshirani to build the graphical lasso algorithm [24]. However, each subproblem of these solvers still involves quite large $(n-1)\times(n-1)$ matrices, resulting in $\mathcal{O}(sn^4)$ complexity for s sweeps of all variables. Many research articles have contributed more efficient optimization algorithms (for a brief survey, see [47]).

3.3 Graph Selection Consistency

Regarding the statistical quality of the solution $\varTheta^*$ of (4), we can ask similar questions to those in Sect. 2.3, that if the solution identifies the true graphical structure, or equivalently the true set of edges or the nonzero patterns in the true model $\varTheta^\circ$. In other word, we check if following property holds:

$$P\left(\left\{(i,j) : \varTheta_{ij}^* \neq 0\right\} = \left\{(i,j) : \varTheta_{ij}^\circ \neq 0\right\}\right) \to 1 \ \text{ as } \ m \to \infty.$$

The sparse graph learning problem (4) has a very similar structure to the sparse variable selection problem (1), and they share very similar consistency results, e.g. [75]. Algorithms using random sampling have been recently proposed, such as bolasso [2] and stability selection [57], which require weaker conditions to achieve variable selection consistency.

3.4 Breast Cancer Gene Dependency Graphs

To demonstrate graph extraction using the Gaussian MRF, we used a genomic data set consisting of gene expression profiles of $n = 20492$ features (genes, more specifically, transcripts) from $m = 362$ breast cancer patients. The data set was created combining three gene expression data sets available from the Gene Expression Omnibus, with the accession IDs GSE1456, GSE7390, and GSE11121.[2]

Figure 2 shows the graph learned separately on subgroups of patients determined by their "grade" of cancer progression: 1 (almost normal), 2 (faster growth) and 3 (much faster growth). The parameter $\lambda = 1.6$ was chosen for all cases which produced small numbers of connected components. Only the connected components with at least two nodes are shown for compact visualization. The color of node represents the p-values of the likelihood ratio test, for the case of using each node (gene) as an univariate predictor for overall survival time under the Cox proportional hazard model [16]. Colors are assigned to five p-value intervals in $[10^{-5}, 1)$, equally sized in logarithmic scale, where darker colors indicate smaller p-values.

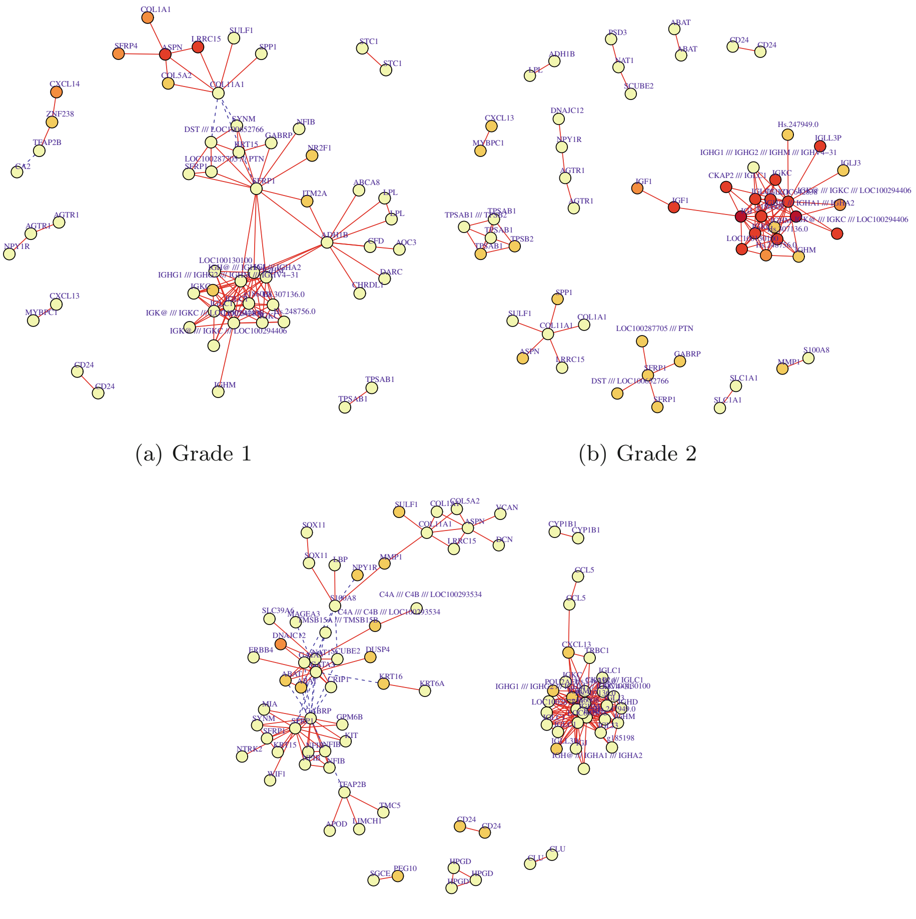
The visualization in Fig. 2 looks quite easy to comprehend even for no biology expert. For example, genes with many neighbors in the graphs (so called hub nodes) turned out to have important roles in breast cancer development, including ASPN [11], SFRP1 [40], and ADH1B [50], even though some (e.g. ADH1B) may not be interesting as univariate predictors considering their p-value.

4 Graph-Based Discovery in Medical Research

An ongoing trend in many scientific areas is the application of network analysis for knowledge discovery. The underlying methodology is the representation of the data by a graph representing a relational structure. Benefits can be created in a blend of different approaches and methods and a combination of disciplines including graph theory, machine learning, and statistical data analysis. This is particularly applicable in the biomedical domain: large-scale generation of various data sources (e.g. from genomics, proteomics, metabolomics, lipidomics, transcriptomics, epigenetics, microbiomics, fluxomics, phenomics, cytomics, connectomics, environomics, exposomics, exonomics, foodomics, toponomics, etc.) allows us to build networks that provide a new framework for understanding the molecular basis of physiological and pathological health states. Many widespread diseases, for example diabetes mellitus, [20], involve enormous interactions between thousands of genes. Although, modern high-throughput techniques allows the identification of such genes amongst the resulting omics data, a functional understanding is still the grand challenge. A major goal is to find diagnostic biomarkers or therapeutic candidate genes.

Network-based methods have been used for quite a while in biology to characterize genomic and genetic mechanisms. Diseases can be considered as abnormal

[2] The CEL files from the GEO were normalized and summarized for transcripts using the frozen RMA algorithm [55]. Then only the verified (grade A) genes were chosen for further analysis according to the NetAffx probeset annotation v33.1 of Affymetrix ($n = 20492$ afterwards). Also, microarrays with low quality according to the GNUSE [54] error scores > 1 were discarded ($m = 392$ afterwards).



(a) Grade 1

(b) Grade 2

(c) Grade 3

Fig. 2. Graphical representation of transcript relations corresponding to breast cancer subgroups. Node color represents the p-value of each node (genes) as univariate predictor of overall survival times (darker color = smaller p-value). Edge types represent correlation: solid = positive and dashed = negative. Node labels show the corresponding gene symbols. (Color figure online)

perturbations of critical cellular networks. The progress and intervention in complex diseases can be analyzed today using network theory. Once the system is represented by a network, methods of network analysis can be applied, not only to extract useful information regarding important system properties, but also to investigate its structure and function. Various statistical and machine learning methods have been developed for this purpose and have already been applied to networks [19]. The underlying structure of such networks are graphs. Graph theory [25] provides tools to map data structures and to find unknown connections between single data objects [21,65]. The inferred graphs can be further analyzed by using graph-theoretical, statistical and machine learning techniques [18].

A mapping of already existing and in medical practice approved *knowledge spaces* as a conceptual graph and the subsequent visual and graph-theoretical analysis may bring novel insights on hidden patterns in the data, which exactly is the goal of knowledge discovery [28]. Another benefit of the graph-based data structure is in the applicability of methods from network topology and network analysis and data mining, e.g. small-world phenomenon [4,39], and cluster analysis [42,72].

However, the biomedical domain is significantly different from other real world domains. Mostly, the processes are data-driven trial-and-error processes, used as help to extract patterns from large data sets by way of predefined models through an fully automated tool without human involvement [77]. Many machine learning researchers pay much attention to find algorithms, models and tools to support such fully automated approaches. The Google car is currently a best practice example [64], at the same time little attention is paid to include the human into this loop.

The reason for this huge difference is the high complexity of the biomedical research domain itself [14]. It is inevitable for the future biomedical domain expert to switch from the classical consumer-like role [44] to an active part in the knowledge discovery process [27,30]. However, this is not so easy, because it is well known that many biomedical research projects fail due to the technical barriers that arise to the domain experts in data integration, data handling, data processing, data visualization and analysis [1,34,43]. A survey from 2012 among hospitals from Germany, Switzerland, South Africa, Lithuania, and Albania [60] showed that only 29 % of the medical professionals were familiar with any practical application of data mining methods and tools. Although this survey might not be representative globally, it clearly shows the trend that medical research is still widely based on standard statistical methods.

To turn the life sciences into data intensive sciences [28], consequently, there is urgent need for usable and useful data exploration systems - which are in the direct work flow of the biomedical domain expert [81]. A possible solution to solve such problems is in a hybrid approach to put the human into the machine learning loop [22,63].

4.1 Medical Knowledge Space

This example shows the advantage of representing large data sets of medical information using graph-based data structures. Here, the graph is derived from a standard quick reference guide for emergency doctors and paramedics in the German speaking area; tested in the field, and constantly improved for 20 years: The handbook "Medikamente und Richtwerte in der Notfallmedizin" [58] (German for Drugs and Guideline Values in Emergency Medicine, currently available in the 11th edition accompanies every German-speaking emergency doctor as well as many paramedics and nurses. It has been sold 58,000 times in the

German-speaking area. The 92-pages handbook (size: 8 × 13 cm) contains a comprehensive list of emergency drugs and proper dosage information. Additionally, important information for many emergency situations is included.

The data includes more than 100 essential drugs for emergency medicine, together with instructions on application and dosage depending on the patient condition, complemented by additional guidelines, algorithms, calculations of medical scores, and unit conversion tables of common values. However, due to the traditional list-based interaction style, the interaction is limited to a certain extent. Collecting all relevant information may require multiple switches between pages and chapters, and knowledge about the entire content of the booklet. In consequence to the alphabetical listing of drugs by active agents, certain tasks, like finding all drugs with common indications, proved to be inefficient and time consuming.

Modeling relationships between drugs, patient conditions, guidelines, scores and medical algorithms as a graph (cf. Fig. 3) gives valuable insight into the structure of the data set. Each drug is associated with details about its active agent and pharmacological group; brand name, strengths, doses and routes of administration of different products; indications and contraindication, as well as additional remarks on application. Consequently, a single drug itself can be represented as connected concepts. Shared concepts create links between multiple drugs with medical relevance, and provide a basis for content-aware navigation.

The interconnection of two drugs, namely adrenaline and dobutamine, is shown in Fig. 4. The left-hand side illustrates the main three types of relations inducing medical relevance; shared indications, shared contra-indications and shared pharmacological groups. Different node colors are used to distinguish between types of nodes such as active agents, pharmacological groups, applications, dosages, indications and contra-indications. The right-hand side highlights the connection of adrenaline and dobutamine by a shared indication.

Links to and between clinical guidelines, tables and calculations of medical scores, algorithms and other medical documents, follow the same principle. On the contrast to a list-based interaction style, these connections can be used for identification and visualization of relevant medical documents, to reorganize the presentation of the medical content and to provide a fast and reliable contextual navigation.

The explosive growth of complexity of networks have overwhelmed conventional visualization methods and future research should focus on developing more robust and efficient temporally aware clustering algorithms for dynamic graphs, i.e. good clustering will produce layouts that meet general criteria, such as cluster colocation and short average edge length, as well as minimize node motion between time steps [52]. The use of multi-touch interfaces for graph visualization [32] extends graph manipulation capabilities of users and thereby can be used to solve some of the visualization challenges.

4.2 DrugBank

DrugBank is an comprehensive, open, online database that combines detailed drug data with drug target information, first released in 2006 [71]. The current

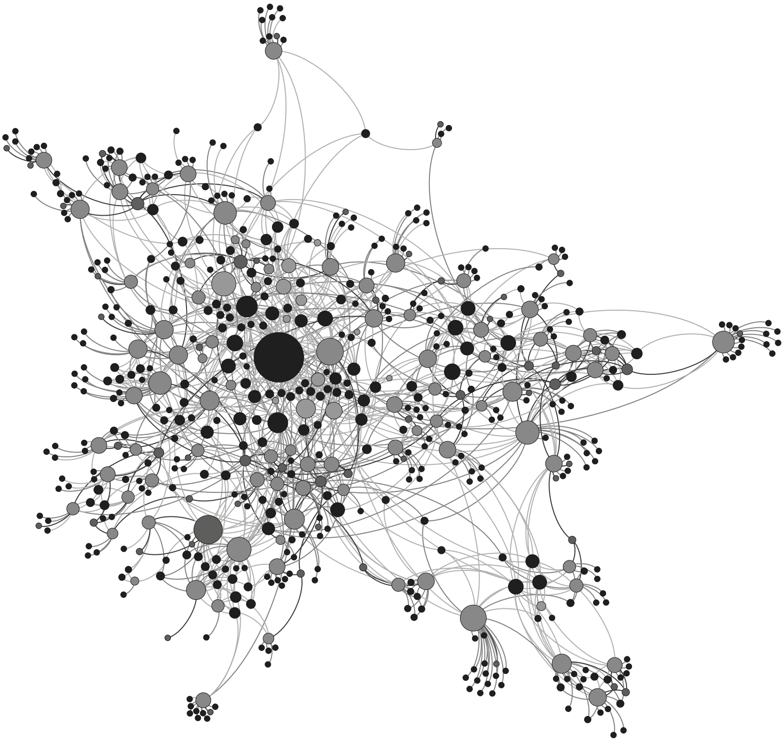


Fig. 3. Graph of the medical data set showing the relationship between drugs, guidelines, medical scores and algorithms.

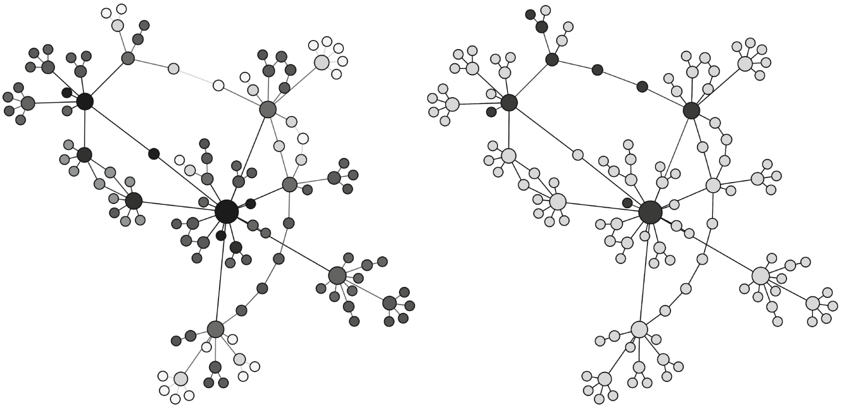


Fig. 4. Interconnection between two drugs, "Adrenaline" and "Dobutamine"; connections to and between clinical guidelines, tables and calculations of medical scores, algorithms and other medical documents, follow the same principle. (Color figure online)

version (DrugBank 4.2) includes 7759 drug entries, with each entry containing more than 200 data fields devoted to drug/chemical data, as well as drug target and protein data.

DrugBank includes drug descriptions, chemical structures and properties, food and drug interactions, mechanisms of action, patent and pricing data, nomenclature, synonyms, etc. Previous versions of DrugBank have been widely used to facilitate drug discovery and constant updates have it expanded to contain data on drug metabolism, absorption, distribution, metabolism, excretion and toxicity and other kinds of quantitative structure activity relationships information [46]. Users may query DrugBank in several different ways via the provided web interface, including simple text queries, chemical compounds queries and protein sequence searches. Alternatively the full database can be downloaded in XML format for further data processing and exploration.

While the DrugBank database is a comprehensive resource for information on individual drugs, it does not provide an illustration of the overall structure of the data set. Representation as a graph can quickly and clearly create new insight into the DrugBank dataset, such as pattern in drug and food interactions, structures in drug and drug classification relations, or relations between drugs by common indications.

The DrugBank database contains 1191 distinct drug entries which list at least a single interaction with another drug. This allows us to define the node set representing these drugs, and the set as all edges between drugs, when an interaction between two drugs is listed. This construced graph contains 1213 nodes linked by 12088 edges, which reveals 22 nodes (e.g. “Sipuleucel-T”, “Pizotifen”, “Iodine”, etc.), listed as drug interaction without a corresponding drug entry in the DrugBank database. Figure 5 shows the visualization of the drug interaction graph, with the drug node size weighted by degree.

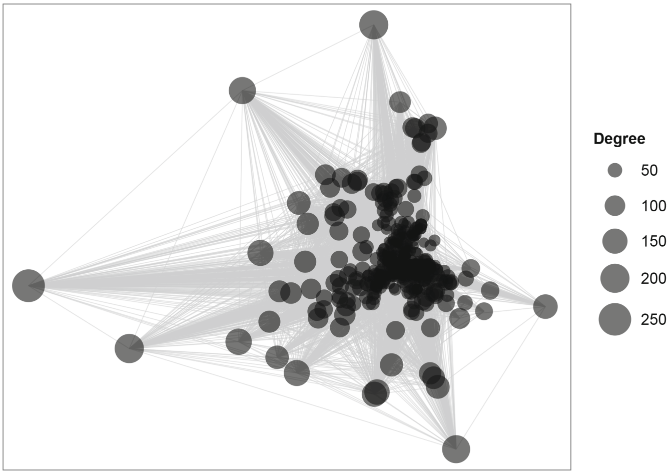


Fig. 5. Graph of drug interactions in the DrugBank database.

4.3 Biological Networks

Functions of life on a sub-cellular level rely on various complex interactions between different entities. Proteins, genes and metabolites interact to produce either healthy or diseased cellular processes. Our understanding of this network of interactions, and the interacting objects themselves, is continuously changing; and the graph structure itself is constantly changing and evolving as we age or as disease progresses.

Our methods for discovering new relationships and pathways change as well. A tool from Jurisica Group in Toronto may be of help here: NAViGaTOR 3 addresses such realities by having a very basic core rooted in graph theory, with the flexibility of a modular plugin architecture that provides data input and output, analysis, layout and visualization capabilities. NAViGaTOR 3 implements this architecture by following the OSGi standard[3]. Available API enables developers to expand standard distribution by integrating new features and extending the functionality of the program to suit their specific needs [61].

NAViGaTOR 3 was designed with the knowledge that a researcher may need to combine heterogeneous and distributed data sources. The standard distribution supports the loading, manipulation, and storage of multiple XML formats and tabular data. XML data is handled using a suite of file loaders, including XGMML, PSI-MI, SBML, KGML, and BioPAX, which store richly-annotated data and provide links to corresponding objects in the graph. Tabular data is stored using DEX (Martinez-Bazan, Gomez-Villamor and Escale-Claveras, 2011), a dedicated graph database from Sparsity Technologies[4].

Figure 6 shows an integrated graph by combining metabolic pathways, protein-protein interactions, and drug-target data. This metabolic data was collected in the Jurisca Lab, combining several steroid hormone metabolism pathways: androgen, glutathione, N-nitrosamine and benzo(a)pyrene pathway, the ornithine-spermine biosynthesis pathway, the retinol metabolism pathway and the TCA cycle aerobic respiration pathway. The figure highlights different pathways with different edge colors. The edge directionality highlights reactions and flow between the individual pathways. The data set was centred on steroid hormone metabolism and included data from hormone-related cancers [26]. The list of FDA-approved drugs used for breast, ovarian and prostate cancer was retrieved from the National Cancer Institute[5]. Afterwards the DrugBank[6] was searched for targets for each drug and those integrated in the graph structure.

5 Challenges and Future Research

A grand challenge is to discover relevant *structural* patterns and/or *temporal* patterns ("knowledge") in high dimensional data, which are often hidden and

[3] OSGi Standard http://www.osgi.org/Main/HomePage.

[4] DEX Graph Database http://www.sparsity-technologies.com/dex.

[5] National Cancer Institute, http://www.cancer.gov.

[6] DrugBank, http://www.drugbank.ca.

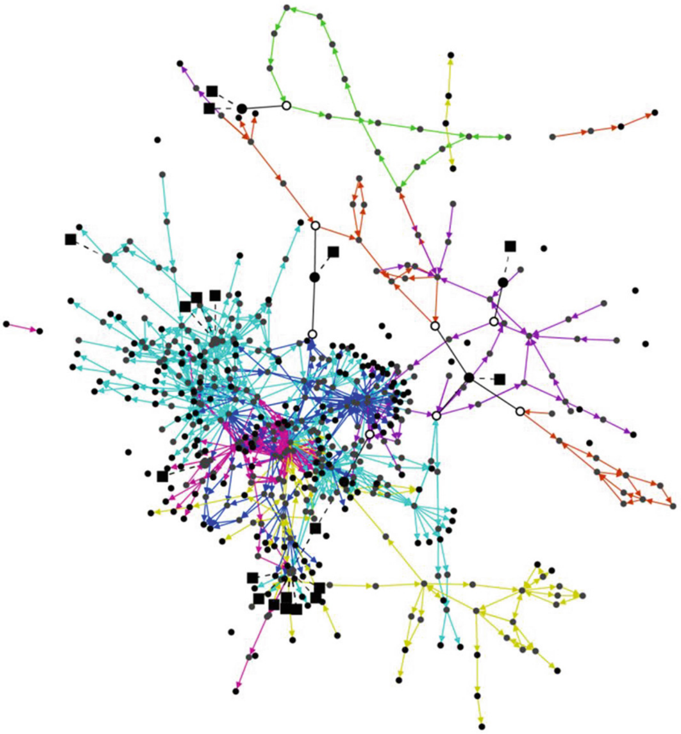


Fig. 6. Partially explored network: connecting drugs and metabolism. A network comprising metabolites, enzymes, and drugs of multiple pathways in the early stages of exploration. (Color figure online)

not accessible to the human expert but would be urgently needed for better decision support or for deeper investigation. Also, the fact that most data sets in the biomedical domain are weakly-structured or non-standardized add extra difficulties [28].

In medical research, these challenges are closely connected to the search for personalized medicine, which is a trend resulting in an explosion in data size (especially dimensionality): for instance "-omics" data, including data of genomics, proteomics, metabolomics, etc [35]. Whilst personalized medicine is the ultimate goal, stratified medicine has been the current approach, which aims to select the best therapy for groups of patients who share common biological characteristics. Here, machine learning approaches and optimization of knowledge discovery tools become imperative [53,61].

Optimization algorithms and techniques are now at the core of many data analysis problems. In high dimensional settings, statistical understanding of these algorithms is crucial not only to obtain quality solutions but also to invent new types of algorithms, as witnessed in recent literature [2,8,49,57]. Efficient and distributed algorithm implementations also become critical due to high computational demands. There are lots of active research in this regard based on optimization algorithms e.g. the ADMM [9] and block-coordinate descent methods [6,67].

Graph-based approaches introduced above are closely related to the graph-based data mining and topological data mining, which are amongst the most challenging topics [31–33,62]. Graph-based data mining was pioneerined about two decades ago [15,17,73], and based upon active research subjects including subgraph categories, isomorphism, invariance, measures, and solution methods [70]. It also can involve content-rich information, e.g. relationship among biological concepts, genes, proteins and drugs, such as in [13] or network medicine [5].

A closely related method is topological data mining, which focuses more on topological spaces (or manifolds) equipped with measures defined for data elements. The two most popular topological techniques in the study of data are *homology* and *persistence*. The connectivity of a space is determined by its cycles of different dimensions. These cycles are organized into groups, called homology groups. Given a reasonably explicit description of a space, the homology groups can be computed with linear algebra. Homology groups have a relatively strong discriminative power and a clear meaning, while having low computational cost. In the study of persistent homology the invariants are in the form of persistence diagrams or barcodes [23]. For interested readers, we suggest papers about point cloud from vector space models [69], and persistent homology [10,12,78].

The grand vision for the future is to effectively support human learning with machine learning. The HCI-KDD network of excellence[7] is an initiative proactively supporting this vision, bringing together people with diverse background but with a shared goal of finding solutions for dealing with big and complex data sets. We believe such an endeavor is necessary to deal with the complex and interdisciplinary nature of the problem. A recent outcome of the network can be found here [29]. This shows that diverse techniques and new ideas need to be integrated for successful knowledge discovery with big and complex real data. Still, there are many emergent challenges and open problems, which we believe deserve further research.

References

1. 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. Inform. Assoc. **14**(4), 478–488 (2007)
2. Bach, F.R.: Bolasso: Model consistent Lasso estimation through the bootstrap. In: 25th International Conference on Machine Learning, pp. 33–40 (2008)
3. Banerjee, O., Ghaoui, L.E., d'Aspremont, A.: Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. J. Am. Med. Inform. Assoc. **9**, 485–516 (2008)
4. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)

[7] HCI-KDD Network: www.hci-kdd.org.

5. Barabási, A., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Science **12**(1), 56–68 (2011)
6. Beck, A., Tetruashvili, L.: On the convergence of block coordinate descent type methods. Science **23**(4), 2037–2060 (2013)
7. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. Science **2**(1), 183–202 (2009)
8. Bogdan, M., van den Berg, E., Sabatti, C., Su, W., Candes, E.J.: SLOPE - adaptive variable selection via convex optimization. (2014). arXiv:1407.3824
9. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Science **3**(1), 1–122 (2011)
10. Bubenik, P., Kim, P.T.: A statistical approach to persistent homology. Science **9**(2), 337–362 (2007)
11. Castellana, B., Escuin, D., Peiró, G., Garcia-Valdecasas, B., Vázquez, T., Pons, C., Pérez-Olabarria, M., Barnadas, A., Lerma, E.: ASPN and GJB2 are implicated in the mechanisms of invasion of ductal breast carcinomas. Science **3**, 175–183 (2012)
12. Cerri, A., Fabio, B.D., Ferri, M., Frosini, P., Landi, C.: Betti numbers in multidimensional persistent homology are stable functions. Science **36**(12), 1543–1557 (2013)
13. Chen, H., Sharp, B.M.: Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics **5**(1), 147 (2004)
14. Cios, K.J., Moore, G.W.: Uniqueness of medical data mining. BMC Bioinformatics **26**(1), 1–24 (2002)
15. Cook, D.J., Holder, L.B.: Graph-based data mining. BMC Bioinformatics **15**(2), 32–41 (2000)
16. Cox, D.R., Oakes, D.: Analysis of Survival Data. Monographs on Statistics & Applied Probability. Chapman & Hall/CRC, London (1984)
17. Dehaspe, L., Toivonen, H.: Discovery of frequent DATALOG patterns. BMC Bioinformatics **3**(1), 7–36 (1999)
18. Iordache, O.: Methods. In: Iordache, O. (ed.) Polystochastic Models for Complexity. UCS, vol. 4, pp. 17–61. Springer, Heidelberg (2010)
19. Dehmer, M., Basak, S.C.: Statistical and Machine Learning Approaches for Network Analysis. Wiley, Hoboken (2012)
20. 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., Röcker, C., Ziefle, M. (eds.) Smart Health. LNCS, vol. 8700, pp. 237–260. Springer, Heidelberg (2015)
21. Dorogovtsev, S., Mendes, J.: Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, Oxford (2003)
22. Duerr-Specht, M., Goebel, R., Holzinger, A.: Medicine and health care as a data problem: will computers become better medical doctors? In: Holzinger, A., Röcker, C., Ziefle, M. (eds.) Smart Health. LNCS, vol. 8700, pp. 21–39. Springer, Heidelberg (2015)
23. Epstein, C., Carlsson, G., Edelsbrunner, H.: Topological data analysis. BMC Bioinformatics **27**(12), 120201 (2011)
24. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical Lasso. BMC Bioinformatics **9**(3), 432–441 (2008)
25. Golumbic, M.C.: Algorithmic Graph Theory and Perfect Graphs. Elsevier, Amsterdam (2004)
26. Henderson, B.E., Feigelson, H.S.: Hormonal carcinogenesis. Carcinogenesis **21**(3), 427–433 (2000)

27. 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)
28. 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), I1 (2014)
29. Holzinger, A., Jurisica, I. (eds.): Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges, vol. 8401. Springer, Heidelberg (2014)
30. 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. LNCS, vol. 8401, pp. 1–18. Springer, Heidelberg (2014)
31. Holzinger, A., Malle, B., Giuliani, N.: On graph extraction from image data. In: Ślęzak, D., Tan, A.-H., Peters, J.F., Schwabe, L. (eds.) BIH 2014. LNCS, vol. 8609, pp. 552–563. Springer, Heidelberg (2014)
32. Holzinger, A., Ofner, B., Dehmer, M.: Multi-touch graph-based interaction for knowledge discovery on mobile devices: state-of-the-art and future challenges. In: Holzinger, A., Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. LNCS, vol. 8401, pp. 241–254. Springer, Heidelberg (2014)
33. Holzinger, A., Ofner, B., Stocker, C., Calero Valdez, A., Schaar, A.K., Ziefle, M., Dehmer, M.: On graph entropy measures for knowledge discovery from publication network data. In: Cuzzocrea, A., Kittl, C., Simos, D.E., Weippl, E., Xu, L. (eds.) CD-ARES 2013. LNCS, vol. 8127, pp. 354–362. Springer, Heidelberg (2013)
34. 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, pp. 3–18. Springer, Heidelberg (2014)
35. 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. LNCS, vol. 8401, pp. 317–330. Springer, Heidelberg (2014)
36. Jacob, L., Obozinski, G., Vert, J.P.: Group Lasso with overlap and graph Lasso. In: Proceedings of the 26th International Conference on Machine Learning (ICML), pp. 433–440 (2009)
37. Javanmard, A., Montanari, A.: Model selection for high-dimensional regression under the generalized irrepresentability condition. BMC Bioinformatics **26**, 3012–3020 (2013)
38. Joachims, T., Finley, T., Yu, C.N.: Cutting-plane training of structural SVMs. BMC Bioinformatics **77**(1), 27–59 (2009)
39. Kleinberg, J.: Navigation in a small world. Nature **406**(6798), 845–845 (2000)
40. Klopocki, E., Kristiansen, G., Wild, P.J., Klaman, I., Castanos-Velez, E., Singer, G., Stöhr, R., Simon, R., Sauter, G., Leibiger, H., Essers, L., Weber, B., Hermann, K., Rosenthal, A., Hartmann, A., Dahl, E.: Loss of SFRP1 is associated with breast cancer progression and poor prognosis in early stage tumors. Nature **25**(3), 641–649 (2004)
41. Knight, K., Fu, W.: Asymptotics for Lasso-type estimators. Ann. Stat. **28**(5), 1356–1378 (2000)
42. Koontz, W., Narendra, P., Fukunaga, K.: A graph-theoretic approach to nonparametric cluster analysis. Nature **100**(9), 936–944 (1976)

43. Kumpulainen, S., Jarvelin, K.: Barriers to task-based information access in molecular medicine. Nature **63**(1), 86–97 (2012)
44. Kurgan, L.A., Musilek, P.: A survey of knowledge discovery and data mining process models. Nature **21**(01), 1–24 (2006)
45. Lauritzen, S.L.: Graphical Models. Oxford University Press, Oxford (1996)
46. Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y.F., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjee, S., Dame, Z.T., Han, B.S., Zhou, Y., Wishart, D.S.: Drugbank 4.0: shedding new light on drug metabolism. Nature **42**(D1), D1091–D1097 (2014)
47. Lee, S.: Sparse inverse covariance estimation for graph representation of feature structure. In: Holzinger, A., Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. LNCS, vol. 8401, pp. 227–240. Springer, Heidelberg (2014)
48. Lee, S.: Signature selection for grouped features with a case study on exon microarrays. In: Stańczyk, U., Jain, L.C. (eds.) Feature Selection for Data and Pattern Classification, pp. 329–349. Springer, Heidelberg (2015)
49. Lee, S., Wright, S.J.: Manifold identification in dual averaging methods for regularized stochastic online learning. Nature **13**, 1705–1744 (2012)
50. Lilla, C., Koehler, T., Kropp, S., Wang-Gohrke, S., Chang-Claude, J.: Alcohol dehydrogenase 1B (ADH1B) genotype, alcohol consumption and breast cancer risk by age 50 years in a german case-control study. Nature **92**(11), 2039–2041 (2005)
51. Lodhi, H., Saunders, C., Shawe-Taylor, J., Watkins, N.C.C.: Text classification using string kernels. Nature **2**, 419–444 (2002)
52. Ma, K.L., Muelder, C.W.: Large-scale graph visualization and analytics. Nature **46**(7), 39–46 (2013)
53. Mattmann, C.A.: Computing: a vision for data science. Nature **493**(7433), 473–475 (2013)
54. McCall, M., Murakami, P., Lukk, M., Huber, W., Irizarry, R.: Assessing affymetrix genechip microarray quality. BMC Bioinformatics **12**(1), 137 (2011)
55. McCall, M.N., Bolstad, B.M., Irizarry, R.A.: Frozen robust multiarray analysis (fRMA). BMC Bioinformatics **11**(2), 242–253 (2010)
56. Meinshausen, N., Bühlmann, P.: High-dimensional graphs and variable selection with the Lasso. BMC Bioinformatics **34**, 1436–1462 (2006)
57. Meinshausen, N., Bühlmann, P.: Stability selection. BMC Bioinformatics **72**(4), 417–473 (2010)
58. Müller, R.: Medikamente und Richtwerte in der Notfallmedizin, 11th edn. Ralf Müller Verlag, Graz (2012)
59. Nesterov, Y.E.: A method of solving a convex programming problem with convergence rate $o(1/k^2)$. Soviet Math. Dokl. **27**(2), 372–376 (1983)
60. Niakšu, O., Kurasova, O.: Data mining applications in healthcare: research vs practice. In: Databases and Information Systems Baltic DB & IS 2012, p. 58 (2012)
61. 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. LNCS, vol. 8401, pp. 19–33. Springer, Heidelberg (2014)
62. Preuß, M., Dehmer, M., Pickl, S., Holzinger, A.: On terrain coverage optimization by using a network approach for universal graph-based data mining and knowledge discovery. In: Ślęzak, D., Tan, A.-H., Peters, J.F., Schwabe, L. (eds.) BIH 2014. LNCS, vol. 8609, pp. 564–573. Springer, Heidelberg (2014)

63. Schoenauer, M., Akrour, R., Sebag, M., Souplet, J.C.: Programming by feedback. In: Proceedings of the 31st International Conference on Machine Learning (ICML 2014), pp. 1503–1511 (2014)
64. Spinrad, N.: Google car takes the test. Nature **514**(7523), 528–528 (2014)
65. Strogatz, S.: Exploring complex networks. Nature **410**(6825), 268–276 (2001)
66. Tibshirani, R.: Regression shrinkage and selection via the Lasso. Nature **58**, 267–288 (1996)
67. Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. Nature **109**(3), 475–494 (2001)
68. Vandenberghe, L., Boyd, S., Wu, S.P.: Determinant maximization with linear matrix inequality constraints. Nature **19**(2), 499–533 (1998)
69. Wagner, H., Dłotko, P., Mrozek, M.: Computational topology in text mining. In: Ferri, M., Frosini, P., Landi, C., Cerri, A., Di Fabio, B. (eds.) CTIC 2012. LNCS, vol. 7309, pp. 68–78. Springer, Heidelberg (2012)
70. Washio, T., Motoda, H.: State of the art of graph-based data mining. Nature **5**(1), 59 (2003)
71. Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., Woolsey, J.: Drugbank: a comprehensive resource for in silico drug discovery and exploration. Nature **34**, D668–D672 (2006)
72. Wittkop, T., Emig, D., Truss, A., Albrecht, M., Boecker, S., Baumbach, J.: Comprehensive cluster analysis with transitivity clustering. Nature **6**(3), 285–295 (2011)
73. Yoshida, K., Motoda, H., Indurkhya, N.: Graph-based induction as a unified learning framework. Nature **4**(3), 297–316 (1994)
74. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Nature **68**, 49–67 (2006)
75. Yuan, M., Lin, Y.: Model selection and estimation in the Gaussian graphical model. Biometrika **94**(1), 19–35 (2007)
76. Zhao, P., Yu, B.: On model selection consistency of Lasso. Biometrika **7**, 2541–2563 (2006)
77. Zhengxiang, Z., Jifa, G., Wenxin, Y., Xingsen, L.: Toward domain-driven data mining. In: International Symposium on Intelligent Information Technology Application Workshops, pp. 44–48 (2008)
78. Zhu, X.: Persistent homology: an introduction and a new text representation for natural language processing. In: IJCAI, IJCAI/AAAI (2013)
79. Zou, H.: The adaptive Lasso and its Oracle properties. Biometrika **101**(476), 1418–1429 (2006)
80. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Biometrika **67**, 301–320 (2005)
81. Zudilova-Seinstra, E., Adriaansen, T.: Visualisation and interaction for scientific exploration and knowledge discovery. Biometrika **13**(2), 115–117 (2007)