Analiza danych o wielkim wymiarze (Statistical inference for high-dimensional data)

Jacek Koronacki

Instytut Podstaw Informatyki PAN

Warszawa, 20 kwietnia i 16 maja 2016

Outline

- Introduction or setting the stage
- Multiple hypothesis testing
- Monte Carlo approaches to feature and model selection
- Regularization approaches to model selection for linear regression, and more
- Bayesian approaches to model selection
- Back to Monte Carlo - more on the ID part of the MCFS-ID algorithm
- In lieu of a conclusion - a word on Big Data Analytics from a statitical perspective

Acknowledgements

Whenever I refer to some contributions with my (usually) more or (sometimes) less modest input, it goes without saying that these contrubutions were made by a group of close collaborators. Of these collaborators, Michał Dramiński has been my closest one, and has greatly and solely developed our original ideas. The group, skillfully headed by our friend Jan Komorowski, includes Michał J. Dąbrowski, Klev Diamanti, Marcin Kierczak, Marcin Kruczyk and Susanne Bornelöv.

Introduction or setting the stage

A major challenge in the analysis of many biological data matrices is due to their sizes: relatively small number of records (samples), often of the order of tens, versus thousands of attributes or features for each record.

An obvious example, albeit rather classical today, are microarray gene expression experiments (here, the features are genes or, more precisely, their expression levels). Another, and a very specific one, is that of analyzing molecular interaction networks underlying HIV-1 resistance to reverse transcriptase inhibitors (here, the features are some physicochemical properties of amino acids). In Next Generation Sequencing / Genome-Wide Association Studies, while we have thousands observations, each consists of hundreds of thousands of features.

Introduction or setting the stage

By far, it is not only in Life Sciences, where problems of this type appear and have to be dealt with.

Indeed, in our own work, we met fascinating problems of commercial origin, including transactional data from a major multinational FMCG (fast-moving consumer goods) company and geological data from oil wells operated by a major American oil company.

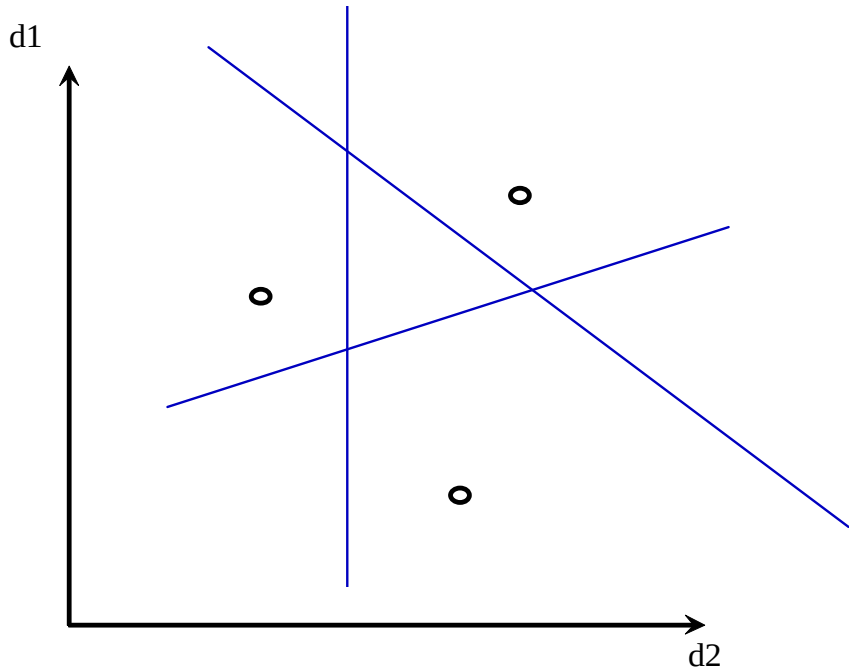
Introduction or setting the stage

Such tasks, regardless of whether the data are to explain a quantitative (as in regression) or categorical (as in classification) trait, are quite different from typical data mining problems, in which the number of features is much smaller than the number of samples.

Indeed, in a sense, these are ill-posed problems. It is immediately clear in the case of linear regression fitted by least-squares.

Introduction or setting the stage

For two-class classification, at least from the geometrical point of view, the task is trivial, since in a d-dimensional space, as many as $d + 1$ points can be divided into two arbitrary and disjoint subsets by some hyperplane, provided that these points do not lie in a proper subspace of the d-dimensional space.



Introduction or setting the stage

It is another matter that the hyperplane (or any other classification rule) found should have the generalization ability.

In any case, whether in classification or in regression, since it is rather a rule than an exception that most features in the data are not informative, it is of utmost importance to select the few ones that are informative and that may form the basis for class prediction or building a proper regression model.

That is, before building a classifier or a regression model, or while building any of them, we would like to find out which features are specifically linked to the problem at hand and should be included in the solution.

Introduction or setting the stage

Mathematically, properly formulated sparsity constraints should be included when seeking a solution. As we shall see, this requirement can be fulfilled by randomization or regularization.

Regarding classification one more important issue should be emphasized:

More often than not, rather than obtaining the best possible classifier, the Life Scientist needs to know which features contribute best to classifying observations (samples) into distinct classes and what are the interdependencies between the features which describe the observation.

Introduction or setting the stage

When dealing with multiple explanatory variables (features), one needs to address the problem of hypothesis testing, We therefore begin our exposition with a brief discussion of multiple hypothesis testing.

We then turn, and confine ourselves, to the area of supervised learning. Within the context of very high dimensional problems, in particular the *small n large p problems*, it is reasonable to divide the whole into three (more or less) separate families of approaches to such learning:

- Monte Carlo methods
- Regularization approaches (with a penalty for model complexity)
- Bayesian approaches.

Introduction or setting the stage

Important remark: It should be emphasized that these three families of approaches are not disjunctive but are partly overlapping. In particular, penalty for model complexity can be Bayesian (like Bayesian Information Criterion, BIC), what pertains to Bayesian regularization. Moreover, it is of utmost interest, and adds to their inherent beauty, that methods from different families share, or have similar, mathematical foundations.

Multiple hypothesis testing

Univariate approach based on multiple hypothesis testing: while disregarding interactions between features, it is statistically sound and all to well illustrates the intricacy of the problem:

Assume a two-class classification case. For each k-th feature we are interested in testing the null hypothesis H_{0k} of no relationship between the decision attribute (class) and the feature against the alternative that such a relationship does exist.

For each k-th feature, $k = 1, \ldots, d$, a natural test statistic is a t-statistic

$$\frac{\bar{x}_{1k} - \bar{x}_{2k}}{s_{1k} + s_{2k}}$$

although examined without assuming normal distribution of the feature.

A real catch is that we have to perform not one but d such tests!

Multiple hypothesis testing

The battery of tests should have a fixed level of the probability of type one error, e.g.,

$$\mathrm{FWER} \equiv \text{family-wise error rate} = P(FP \geqslant 1) \leqslant \alpha$$

where FP stands for the number of false positives (i.e., type I errors)

or

$$\mathrm{FDR} \equiv \text{false discovery rate} = E(FP/(FP + TP)) \leqslant \alpha$$

as well as a reasonable power of the whole procedure, e.g.,

$$P(TP \geqslant 1)$$

where TP stands for the number of true positives.

Multiple hypothesis testing

Bonferroni's (1936) classical procedure, under which any null hypothesis is rejected at level α/d, controls the FWER,

$$\mathrm{FWER} \equiv \text{family-wise error rate} = P(FP \geqslant 1) \leqslant \alpha,$$

for arbitrary test statistics joint null distributions; that is,

$$P(FP \geqslant 1) \leqslant \sum_{i \in \mathcal{H}_0} P_{H_{0i}}(\text{i-th test rejects}) \leqslant \frac{h}{d}\alpha \leqslant \alpha,$$

where $\mathcal{H}_0$ runs over the indices corresponding to true null hypotheses and $h = |\mathcal{H}_0|$.

(Under independence of test statistics and complete null hypothesis,

$$\mathrm{FWER} = 1 - (1 - \alpha/d)^d;$$

the FWER is smaller, if they are positively dependent.)

Multiple hypothesis testing

Note that under the Bonferroni procedure any null hypothesis is rejected regardless of the values of test statistics for other hypotheses.

A more sophisticated procedure of Benjamini and Hochberg (1995; see the next slide) controls the FDR,

$$\mathrm{FDR} \equiv \text{false discovery rate} = E(FP/(FP+TP)) \leqslant \alpha,$$

for independent test statistics (or, more generally, for positively regression dependent test statistics).

Multiple hypothesis testing

The Banjamini and Hochberg procedure:

1. Let

$$p_{(1)} \leqslant p_{(2)} \leqslant \cdots \leqslant p_{(d)}$$

denote the observed ordered p-values

2.

$$L = \max\{j : \ p_{(j)} < \alpha \cdot \frac{j}{d}\}$$

3. Reject all hypotheses H_{0j}, such that $p_{(j)} \leqslant p_{(L)}$.

Thus, the p-values must be obtained, but this can be done by a simple resampling procedure.

For this section see Dudoit and van der Laan (2008).

MC approaches: Model selection for linear regression - Random Subspace Method (RSM)

Mielniczuk and Teisseyre (2011) and (2013): Let $T_{i,m}$ be a t-statistic for i-th predictor in a linear regression model m with $|m|$ predictors. We have:

$$\frac{T_{i,m}^2}{n-|m|} = \frac{\mathrm{RSS}_{m-\{i\}} - \mathrm{RSS}_m}{\mathrm{RSS}_m}$$

It follows that the value of $T_{i,m}^2$ can serve as a measure of, simulatneously, the importance of the i-th predictor in model m and the quality of this very model.

MC approaches: Model selection for linear regression - Random Subspace Method (RSM)

In the RSM, a random subset m of features (predictors), of size $|m|$ smaller than the number of all features d and a number of observations n, is chosen. The model is fitted in the reduced feature space by OLS. Each of the selected features is assigned a weight describing its relevance in the considered submodel.

Random selection of features is repeated many times, corresponding submodels are fitted and the final weights (scores) of all d features are computed on the basis of all submodels.

The final model can then be constructed based on predetermined number of the most significant predictors or using a selection method applied to the nested list of models given by the ordering of predictors.

MC approaches: MCFS-ID Algorithm of Draminski et al.: the Monte Carlo Feature Selection (or MCFS) part

In what follows we begin with a brief description of an effective method for ranking features according to their importance for classification regardless of a classifier to be later used. Our procedure is conceptually very simple, albeit computer-intensive.

We consider a particular feature to be important, or informative, if it is likely to take part in the process of classifying samples into classes "more often than not".

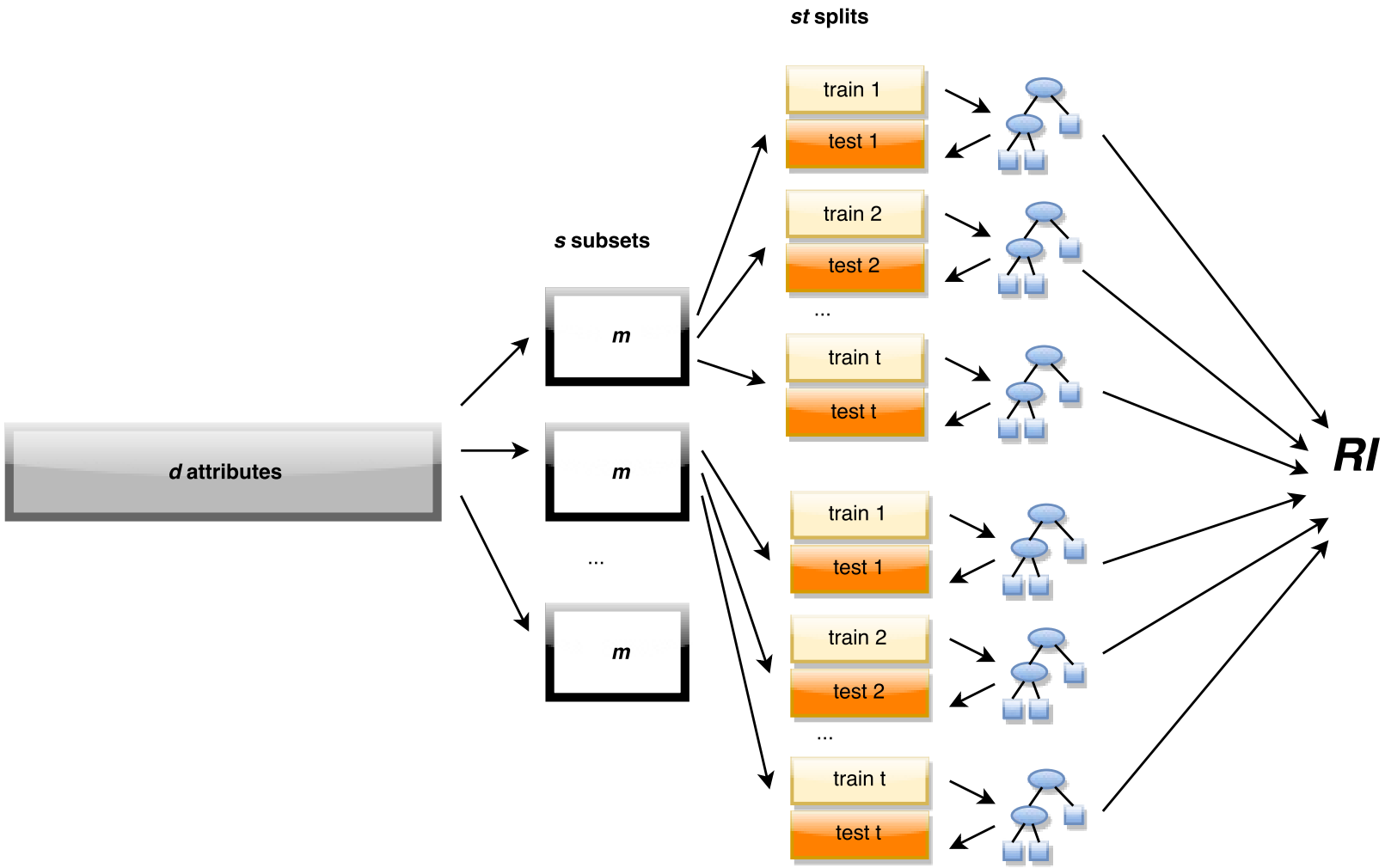
This "readiness" of a feature to take part in the classification process, termed relative importance of a feature, is measured via intensive use of classification trees. When assessing relative importance of a feature, the aforementioned "readiness" of the feature to appear in a given tree is suitably moderated by the (weighted) accuracy this tree.

MC approaches: MCFS-ID Algorithm: the MCFS part

In the main step of the procedure, we estimate relative importance of features by constructing thousands of trees for randomly selected subsets of features.

More precisely, out of all d features, s subsets of m features are selected, m being fixed and $m << d$, and for each subset of features, t trees are constructed and their performance is assessed. Each of the t trees in the inner loop is trained and evaluated on a different, randomly selected training and test sets which come from a split of the full set of training data into two subsets: each time, out of all n samples, about 66% of samples are drawn at random for training (in such a way as to preserve proportions of classes from the full set of training data) and the remaining samples are used for testing.

MC approaches: MCFS-ID Algorithm: the MCFS part



MC approaches: Interdependency Discovery, i.e., the ID part of the MCFS-ID Algorithm

In the MCFS part of the algorithm, a cutoff between informative and non-informative features is provided. From now on, our interest is confined to the set of informative features.

This approach to interdependency discovery is significantly different from known approaches which consist in finding correlations between features or finding groups of features that behave similarly in some sense across samples (e.g., as in finding co-regulated features).

The focus is on identifying features that "cooperate" in determining that a sample belongs to a particular class. A directed graph of such "cooperating" features is constructed.

MC approaches: Interdependency Discovery, i.e., the ID part of the MCFS-ID Algorithm

For an exposition of the MCFS-ID algorithm in its full-flegded versions, see Draminski et al. (2008), (2010), (2016a) and (2016b)

Regarding the ID part of the algorithm, see also next to the last section of this presentation.

Regularization approaches: Model selection for linear regression - ℓ_1 regularization

The Lasso (Least Absolute Shrinkage and Selection Operator):

As usual, we are given n observations, each with d explanatory variables (predictors), $(x_{i1}, x_{i2}, \ldots, x_{id})$, and one response variable, y_i,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_d x_{i,d} + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where ε_i are i.i.d. random errors with mean 0 and unknown variance σ^2, and $\beta_0, \ldots, \beta_d$ are unknown parameters.

Minimize

$$\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\}$$

subject to

$$\sum_{j=1}^{p} |\beta_j| \leqslant t.$$

Regularization approaches: ℓ_1 regularization

The Lasso, in contrast to ridge regression (i.e., ℓ_2 regularization), eliminates for small t some variables from the model. It can thus be used as a feature selection method, although one should be aware that the method is likely to include too many (and incorrectly ordered) variables.

For exhaustive account of the Lasso and related approaches see Bühlmann and van de Geer (2011) and Hastie, Tibshirani and Wainwright (2015). For an important extension of the idea see Pokarowski and Mielniczuk (2015), where a three-stage algorithm for selecting a regression model is proposed, with LASSO used in the 1st stage for screening of predictors (features); the proposed algorithms are given in the next 2 slides. See also Bogdan et al. (2015), where the regularizer is a sorted ℓ_1 norm.

Screening-Selection (SS) procedure of Pokarowski and Mielniczuk

A version of SOS (JMLR (2015)) with 'O' (for 'ordering') removed

Algorithm 1 SS

Input: y, X and λ
Screening (Lasso)
$\widehat{\beta} \equiv \widehat{\beta}(\lambda) = \text{argmin}_\gamma \{\|y - X\gamma\|^2 + 2\lambda|\gamma|_1\}$;
order nonzero coefficients:
$|\widehat{\beta}_{j_1}| \geqslant |\widehat{\beta}_{j_2}| \geqslant \ldots \geqslant |\widehat{\beta}_{j_s}|$, where $s = |\text{supp}\widehat{\beta}|$;
set $\mathcal{J} = \{\{j_1\}, \{j_1, j_2\}, \ldots, \{j_1, \ldots, j_s\}\}$;
Selection (GIC)
$\widehat{T} = \text{argmin}_{J \in \mathcal{J}} \{SSE_J + \lambda^2|J|\}$
Output: $\widehat{\beta}^{SS} = (X_{\widehat{T}}^T X_{\widehat{T}})^{-1} X_{\widehat{T}}^T y$

SOSnet algorithm of Pokarowski and Mielniczuk

- Use Lasso with $\lambda_{i=0,1,\ldots,m}$ to choose set of predictors I_i;
- Fit linear model $y \sim x_{I_i}, i = 0, 1, \ldots, m$;
- Order predictors according to (t-statistics)^2;
- Construct $\mathcal{M} = \cup$ nested models ;
- Use GIC on $\mathcal{M}$ to choose a final model.

Regularization approaches: Support Vector Machines - ℓ_2 regularization. And more

We skip an exposition of SVMs. Regarding their use for Big Data Analytics, we refer to Tan et al. (2014) and to Priyadarshini and Agarwal (2015).

There are more statistical approaches to dealing with high-dimensional data than those already hinted to and the Bayesian ones. See Bühlmann and van de Geer (2011) for an approach which stems from undirected graphical modeling and is based on inferring zero partial correlations for variable selection (the so-called PC-simple algorithm).

A still another and promising approach, which builds on ranking the marginal correlations and is referred to as sure independence screening, has been introduced by Fan and Lv (2008); see also Fan and Song (2010).

Broman and Speed (2002): Let

$$y_i = \mu + \sum_{j=1}^{d} \beta_j x_{ij} + \varepsilon_i,$$

where $x_{ij} = 1$ or $x_{ij} = 0$ and the ε_i are i.i.d. and normally distributed, $N(0, \sigma^2)$ (in fact, x_{ij} represents genotype at marker j for individual i). The task is to select a model for which Schwarz's Bayesian Information Criterion (BIC) assumes the minimal value;

$$BIC = n \cdot \log RSS(\beta) + \frac{1}{2} k \log n,$$

where k is the number of parameters β_j in the model. It was observed by Broman i Speed that the BIC tends to overestimate the number of parameters in the model. Accordingly, they proposed the 1st modification of the BIC.

Model selection for linear regression - Bayesian approaches

The Bayesian model selection advocates choosing the model M that maximizes posterior probability of the model given the data, this probability being proportional to

$$L(y|M)\pi(M),$$

where $\pi(M)$ is a prior probability for model M (Schwartz assumed noninformative uniform prior π), and

$$L(y|M) = \int L(y|M,\beta)f(\beta|M)d\beta,$$

$f(\beta|M)$ being some prior distribution on the vector of model parameters; for a wide class of these distributions one gets

$$\log L(y|M) = \log L(y|\beta) - \frac{1}{2}(k+2)\log n.$$

For the family of normal linear regression models, maximization of this last expression is equivalent to minimization of the BIC.

Model selection for linear regression - Bayesian approaches

Bogdan et al. (2004) introduced another modification of BIC (mBIC), assuming binomial prior distribution, $\mathrm{Bin}(d, c/d)$, with some fixed c, for the model size. See Bogdan et al. (2011) for later developments and Frommlet et al. (2012) for application of their approach to Genome-Wide Association Studies.

It is easy to extend the outlined approach to include regression models with interactions. It is also possible to extend it to include generalized linear models (possibly with constraints on the model's parameters).

The outlined approach is by far not the only one possible among this strand of Bayesian approaches; e.g., a similar approach is that based on the extended BIC, and a completely different approach, which bears some relationship with support vector machines, is that of relevance vector machines. (See, e.g., Chen and Chen (2008) and (2011), and Tipping (2001), Fletcher (2010) and Saarela et al. (2010).)

Nonparametric Bayesian approaches

Let Y be a response and $X = (X^{(1)}, \ldots, X^{(p)} \in R^p$ be explanatory variables. Assume

$$Y = f(X) + \varepsilon,$$

with ε normally distributed, $N(0, \sigma^2)$.

Usually, a Gaussian Process (GP) prior for f is assumed to have zero mean and square exponential covariance function (kernel function) $\exp(-\|x - x'\|^2/c)$. Such processes are smooth in a well-known sense. Other kernels can be used, and another smoothness conditions on f can be imposed.

It should be emphasized that the above mentioned use of a kernel function casts the whole approach into the area of ML with kernels (kernel machines). Indeed, some far reaching similarities (and differences) with ridge regression, SVMs, as well as with spline models are obvious and deserve separate analysis.

Nonparametric Bayesian approaches, contd.

An excellent exposition of Gaussian processes for ML is given in Rasmussen and Williams (2006); another excellent, albeit short, introduction to GPs in ML can be found in Bishop (2006). In neither of these expositions problems pertaining to dealing with Big Data are addressed, although Rasmussen and Williams (2006) has a chapter on Approximation Methods for Large Datasets.

Nonparametric Bayesian approaches, contd.

However interesting GPs for ML are, within the context of Big Data Analytics, special emphasis has to be placed on variable selection and/or variable projections. Loosely speaking, such mechanisms can be included into the nonparametric Bayesian approach by adding more randomness into the process, i.e., introducing suitable hyperparameters. See Tokdar (2011) for variable selection and linear projection proposals which have been shown to give consistent (in probability, and at a known rate) estimators of an unknown f; e.g., for f depending on $d < p$ variables, the rate of convergence is

$$n^{-\frac{\alpha}{2\alpha+d}} (\log n)^k$$

for any $k > p + 1$.

Yang (2014) has noticed that Tokdar's proposal can be considered effective only if $d << p$.

Nonparametric Bayesian approaches, contd.

Yang (2014) has provided a general framework to assess the minimax risks for regression problems under ℓ_2 loss (see there for an excellent account of earlier, sometimes pioneering, results in the area). He has introduced a general class of Bayesian sieve estimators which, under certain (more or less restrictive) conditions, achieve the optimal minimax risk when f depends on $d << \min\{n, p\}$ variables or is a sum of finitely many, k, functions, each of which depends on $d_s << \min\{n, p\}$ variables.

He has shown also that a GP regression approach can lead to the minimax optimal adaptive rate in estimating f under some conditions when the function's domain lies on a Riemannian manifold.

See also Yang and Dunson (2014) and Yang and Tokdar (2015).

Back to Monte Carlo - more on the ID part of the MCFS-ID Algorithm

For a given training set of samples, an ensemble of decision trees has been constructed within the MCFS part of the algorithm. Each decision rule provided by each tree has the form of an "ordered conjunction" of conditions imposed on particular separate features. (Note that trees are "flexible" classifiers, where flexibility amounts to classifier's ability to produce rules as complex as is needed.)

Clearly then, each decision rule points to some interdependencies between the features appearing in the conditions. Indeed, the information included in such decision rules, when properly aggregated, reveals interdependencies (however complex they may prove) between features which are best "correlated" with or, as has been said, "cooperate" in determining, the samples' classes.

The ID part of the MCFS-ID Algorithm, contd.

To see how an ID-Graph is built, let us recall again that each node in each of the multitude of classification trees represents a feature on which a split is made. Now, for each node in each classification tree its all antecedent nodes can be taken into account along the path to which the node belongs.

For each pair [*antecedent node* $\longrightarrow$ *given node*] we add one directed edge to our ID-Graph from *antecedent node* to *given node*.

The edges are found along the paths in all the $s \cdot t$ MCFS trees. Clearly, the same edge can appear more than once even in a single tree.

The ID part of the MCFS-ID Algorithm, contd.

The strength of the interdependence between two nodes, actually two features, connected by a directed edge, termed ID weight of a given edge (ID weight for short), is defined in the following way:

For node $n_k(\tau)$ in the τ-th tree, $\tau = 1, \ldots, s \cdot t$, and its antecedent node $n_i(\tau)$, ID weight of the directed edge from $n_i(\tau)$ to $n_k(\tau)$, denoted $w[n_i(\tau) \rightarrow n_k(\tau)]$, is equal to

$$w[n_i(\tau) \rightarrow n_k(\tau)] = \mathrm{GR}(n_k(\tau)) \left(\frac{\text{no. in } n_k(\tau)}{\text{no. in } n_i(\tau)} \right), \qquad (1)$$

where $\mathrm{GR}(n_k(\tau))$ stands for gain ratio for node $n_k(\tau)$, $(\text{no. in } n_k(\tau))$ denotes the number of samples in node $n_k(\tau)$ and $(\text{no. in } n_i(\tau)$ denotes the number of samples in node $n_i(\tau)$.

The ID part of the MCFS-ID Algorithm, contd.

The final ID-Graph is based on the sums of all ID weights for each pair [*antecedent node* $\longrightarrow$ *given node*].

That is, for each directed edge found, its ID weights are summed over all occurrences of this edge in all paths of all MCFS classification trees.

For a given edge, it is this sum of ID weights which becomes the ID weight of this edge in the final ID-Graph.

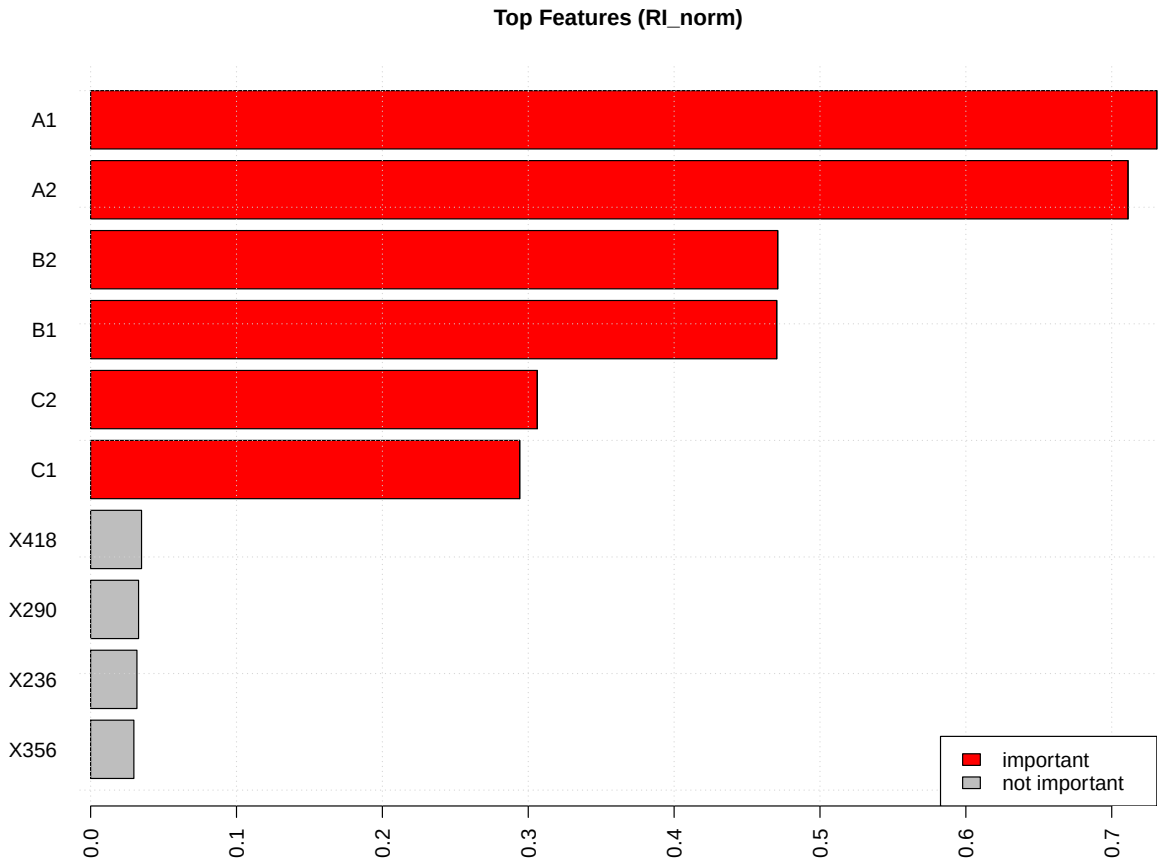
The ID part of the MCFS-ID Algorithm, contd.

In sum, an ID-Graph provides a general roadmap that not only shows all the most variable attributes that allow for efficient classification of the objects but, moreover, it points to possible interdependencies between the attributes and, in particular, to a hierarchy between pairs of attributes. High differentiation of the values of ID weights in the ID-Graph gives strong evidence that some interdependencies between some features are much stronger than others and that they create some patterns/paths calling for interpretation based on background knowledge.

The ID part of the MCFS-ID Algorithm - a toy example

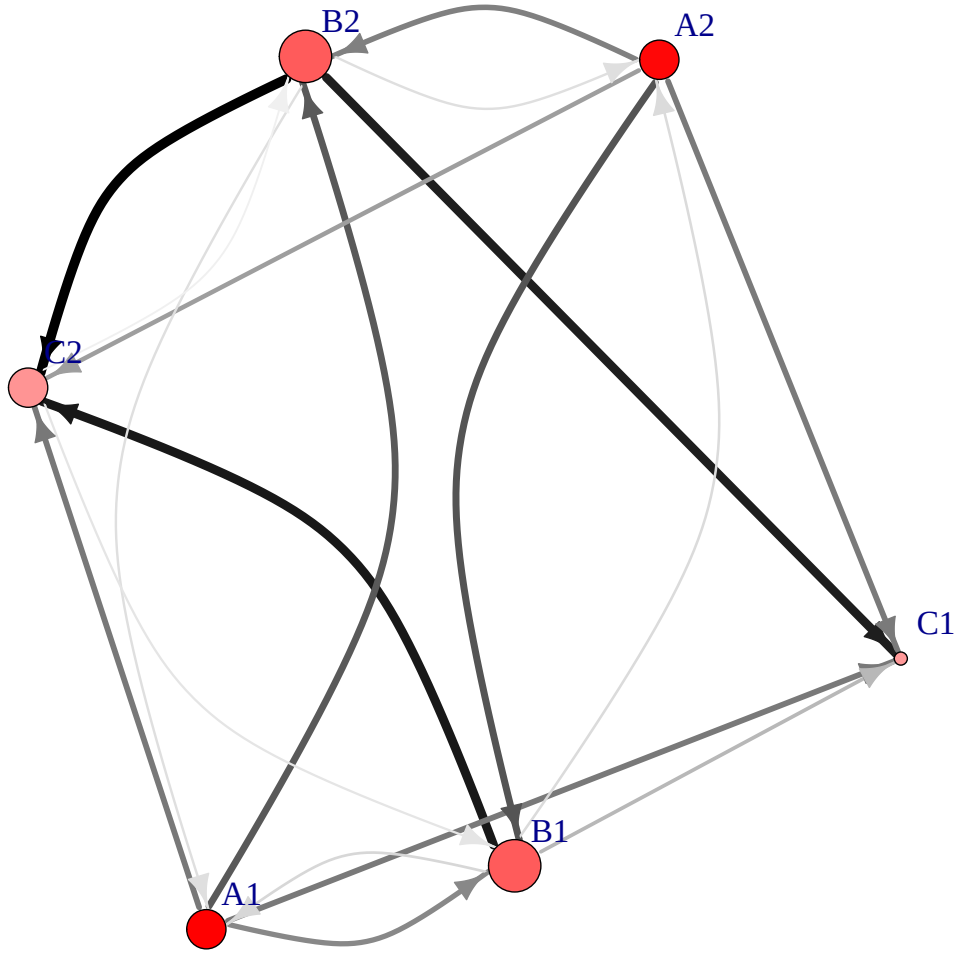
Consider objects from 3 classes, A, B and C, that contain 40, 20 and 10 objects, respectively (70 objects altogether). For each object, create 6 binary features ($A1$, $A2$, $B1$, $B2$, $C1$ and $C2$) that are 'ideally' or 'almost ideally' correlated with *class* feature. If an object's *'class'* equals *'A'*, then its features $A1$ and $A2$ are set to class value *'A'*; otherwise $A1 = A2 = 0$. If an object's *'class'* is *'B'* or *'C'*, we proceed analogously, but we introduce some random corruption to 2 observations from class *'B'* and to 4 observations from class *'C'*: in the former case, for each of the two observations and both attributes $B1/B2$, we randomly replace their value *'B'* by '0' and in the latter case, again for each of the four observations and both attributes $C1/C2$, we randomly replace their value *'C'* by '0'. The data also contains additional 500 random numerical features with uniformly [0,1] distributed values. Thus we end up with 6 nominal important features (3 pairs with different levels of importance for classification) and 500 randomly distributed.

The ID part of the MCFS-ID Algorithm - a toy example



Rysunek: Top features selected by MCFS-ID.

The ID part of the MCFS-ID Algorithm - a toy example

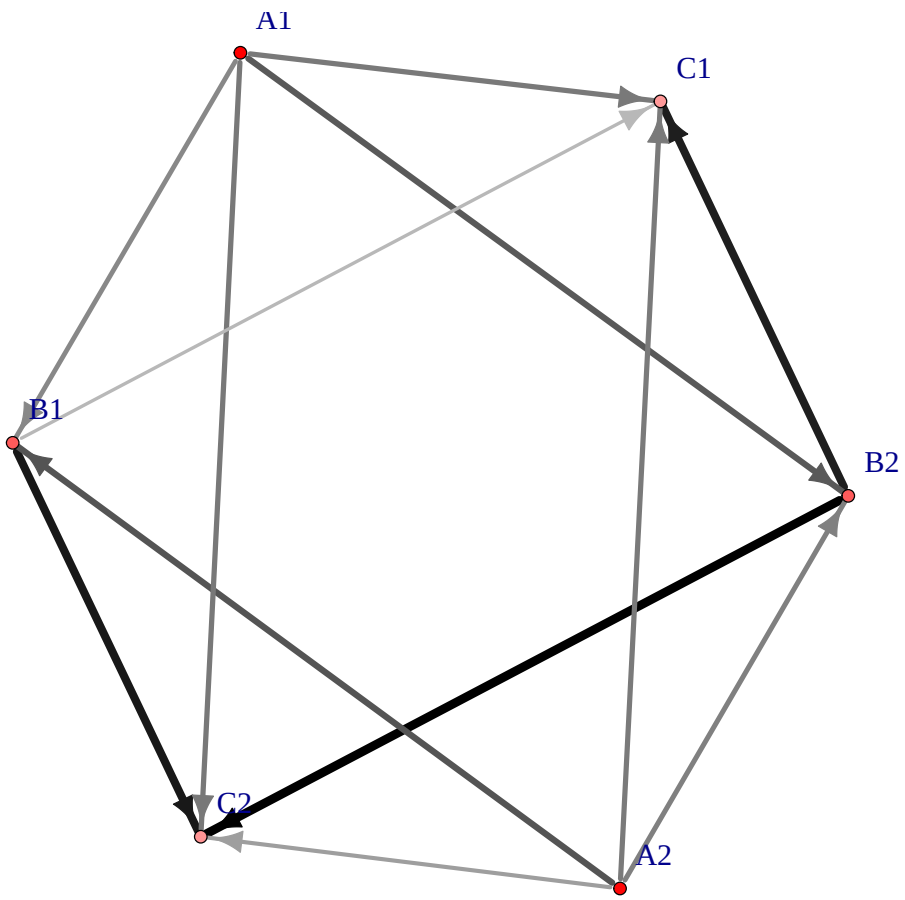


Rysunek: ID-Graph for artificial data.

The ID part of the MCFS-ID Algorithm - a toy example

In the ID-Graphs, as seen in the Figure, some additional information is conveyed with the help of suitable graphical means. The color intensity of a node is proportional to the corresponding feature's RI. The size of a node is proportional to the number of edges related to this node. The width and level of darkness of an edge is proportional to the ID weight of this edge. Since we would like to review only the strongest ID weights let us plot ID-Graph with only 12 top edges.

The ID part of the MCFS-ID Algorithm - a toy example



Rysunek: ID-Graph for artificial data, limited to top 6 features and top 12 ID weights.

The ID part of the MCFS-ID Algorithm, and more - Discovering interactions on a finer level

The ID-Graph does not tell the differences between the classes, i.e., it does tell what interdependencies make the samples belong to different classes but does not give rules which determine any given class. Accordingly and separately, a way to construct rule networks is also provided, where the networks are constructed from IF-THEN rules with one network per each decision class.

Please see Bornelöv, Marillet and Komorowski (2014) and Draminski et al. (2016a) for our proposal.

Concluding, let us add that while the current version of the MCFS-ID is a new one, it is already included in CRAN (The Comprehensive R Archive Network). Moreover, along with a module to discover rule networks their explanatory power has been verified on a number of molecular and medical examples.

In lieu of a conclusion - a word on Big Data Analytics from a statistical perspective

It seems now widely accepted that the term Big Data refers one to situations when data are characterized by at least three or four "Vs" (cf., e.g., chapter 1 in Japkowicz and Stefanowski (2016)):

- Volume - huge and, usually continuously increasing, size of the collected and analyzed data
- Velocity - high speed at which the data is generated and input into an analyzing system
- Variety - heterogenous and complex representations of the analyzed data
- Variability - changes in the structure of the data, as well as changes in how users want to interpret that data.

Clearly then, strictly speaking, Massive Data should not be confused with Big Data.

A word on Big Data Analytics from a statistical perspective, contd.

- Statistical approaches form an indispensable and crucial part of Machine Learning
- As of now, while statistical meta-analyses as well as, e.g., probabilistic methods of linking data from different sources are studied and developed, statistical approaches are best suited to deal only with Massive Data from a homogenous source
- As such, statistical approaches form an indispensable and crucial part of Big Data Analytics, however if used within homogenous settings
- The importance of statistical approaches follows from their explanatory power and methodological rigorousness

A word on Big Data Analytics from a statistical perspective, contd.

- A statistician is well aware that he/she can apply statistical techniques only when the data come from repetitions of some events. He/she is also well aware that the data at hand, when properly analyzed, can help answer only some specific questions, by far not any questions of interest. He/she is well equipped to examine data for possible biases or other faults.

- Methods of statisical learnig provide causal models when possible (feasible), and predictive algorithms (behavioral models) when deeper cognizance of the phenomenon under scrutiny is unavailable.

A word on Big Data Analytics from a statistical perspective, contd.

- Paradoxically, it was an extraordinary development of computer technologies what freed statisticians dealing with massive data from John Tukey's prison of Exploratory Data Analysis with its slogan "Let the data speak for themselves"
- In 1979, William Eddy, a not so famous as John Tukey but a more radical statistician proclaimed:

  "The data analytic method denies the existence of 'truth', the only knowledge is empirical.

  [...] If we can make without models, I think we should."
- Today, a nonmilitant statistician prefers to say:

  If we cannot make with models, we should make without them.

A word on Big Data Analytics from a statistical perspective, contd.

- Flooded by Big Data, some researchers claim essentially the same what radical proponents of EDA claimed decades ago. They say that, e.g., given Big Data, we can abandon causal explanations, since it suffices to know correlations which enable one to predict; cf. discussions of this issue in chapters 1 and 2 in Japkowicz and Stefanowski (2016).
- Even if any pretext can serve the purpose of regressing to foolishness, it is better to stay wise and try to understand, not only to predict.
- Happily, the earlier discussed methods of statistical learning are used and developed to advantage, and widely, within the Big Data settings.

Selective bibliography:

- Bishop C.: Pattern Recognition and Machine Learning. 2006; Springer.
- Bogdan M., Ghosh J.K., Doerge R.W.: Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting Quantitatve Trait Loci. Genetics. 2004; 167, 989-999.
- Bogdan, M., Chakrabarti, A., Frommlet, F., Ghosh, J.K.: Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. Annals of Statistics. 2011; 39(3), 1551-1579.
- Bogdan M., van den Berg E., Sabatti C., Su W., Candes E.J.: SLOPE—ADAPTIVE VARIABLE SELECTION VIA CONVEX OPTIMIZATION. The Annals of Applied Statistics. 2015; 9(3), 1103–1140.
- Bornelöv S., Marillet S., Komorowski J.: Ciruvis: a web-based tool for rule networks and interaction detection using rule-based classifiers. BMC Bioinformatics. 2014; 15:139.
- Broman K.W, Speed T.P, A model selection approach for the identification of quantitative trait loci in experimental crosses. J. Royal Statist. Soc. 2002; B 64, 641-656.
- Bühlmann P., van de Geer S., *Statistics for High-Dimensional Data*, Springer, 2011.
- Chen J., Chen Z., Extended Bayesian information criterion for model selection with large model space. 2008; Biometrika, 94, 759-771.
- Chen J., Chen Z., Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. 2011; arXiv:1107.2502
- Draminski M, Rada Iglesias A, Enroth S, Wadelius C, Koronacki J, Komorowski J.: Monte Carlo feature selection for supervised classification. Bioinformatics. 2008; 24, 110-117.
- Draminski M., Kierczak M., Koronacki J., Komorowski J.: Monte Carlo Feature Selection and Interdependency Discovery in Supervised Classification. In: Koronacki J., et al. (eds): Advances in Machine Learning II. 2010; Springer series: Studies in Computational Intelligence, Vol. 263, 371-385.
- Draminski M., Dąbrowski M.J., Diamanti K., Koronacki J., Komorowski J.: Discovering networks of interdependent features in high-dimensional problems. In: : N.Japkowicz and J.Stefanowski (eds.), Big Data Analysis: New Algorithms for a New Society. 2016a; Springer, 285-304.
- Draminski M., Koronacki J.: rmcfs: An R Package for Monte Carlo Feature Selection and Interdependency Discovery. 2016b; submitted.
- Dudoit S., van der Laan M. J., *Multiple Testing Procedures with Applications to Genomics*, Springer, 2008.
- Fan J., Lv J., Sure independence screening for ultra-high dimensional feature space. J. Royal Statist. Soc. 2008; B 70 (5), 849-911.
- Fan. J., Song R., Sure independence screening for generalized linear models with np-dimensionality. Ann. Statist. 2010; 38(6), 3567–3604.

Selective bibliography, contd.:

- Fletcher T., Relevance vector machines explained. 2010; Tech. report, www.cs.ucl.ac.uk/staff/T.Fletcher
- Frommlet F., Ruhaltinger F., Twaróg P., Bogdan M.: Modified versions of the Bayesian Information Criterion for genome-wide association studies Computational Statist. and Data Anal. 2012; 56(5), 1038-1051.
- Hastie T., Tibshirani R., Wainwright M.: Statistical Learning with Sparsity: The Lasso and Generalizations, CRC 2015.
- N.Japkowicz N., Stefanowski J. (eds.): Big Data Analysis: New Algorithms for a New Society. 2016b; Springer.
- Mielniczuk J., Teisseyre P.: Using Random Subset Method for prediction and variable importance assessment in linear regression. Computational Statist. and Data Anal. 2012; in press.
- Mielniczuk J., Teisseyre P.: Selection and Prediction for Linear Models using Random Subspace Methods. Proceedings of the Conference Information Technologies: Research and their Interdisciplinary Applications, Institute of Computer Science. 2013; 103-121.
- Pokarowski P., Mielniczuk J.: Combined ℓ_1 and Greedy ℓ_0 Penalized Least Squares for Linear Model Selection. J. Machine Learning Reserach. 2015; 16, 961-992.
- Priyadarshini A., Agarwal S.: A Map Reduce based Support Vector Machine for Big Data Classification. International Journal of Database Theory and Application. 2015; 8(5), 77-98.
- Rasmussen C.E., Williams C.K.I.: Gaussian Processes for Machine Learning. 2006; MIT Press.
- Saarela M., Elomaa T., Ruohonen K., An Analysis of Relevance Vector Machine Regression. In: Advances in Machine Learning, vol. 2; Springer, 2010.
- Tan M., Tsnag I.W., Wang L.: Towards Ultrahigh Dimensional Feature Selection for Big Data. J. Machine Learning Reserach. 2014; 15, 1371-1429.
- Tipping M.E.: Sparse Bayesian learning and the relevance vector machine. 2001; Journal of Machine Learning Research 1, 211-244.
- Tokdar S.T: Dimension adaptability of Gaussian process models with variable selection and projection. 2011; Tech. Rep., Duke Iniversity, arXiv:1112.0716v1.
- Yang Y.: Nonparametric Bayes for Big Data. PhD Thesis. 2014; Duke Unoversity.

Selective bibliography, contd.:

- Yang Y., Dunson D.B.: Bayesian Manifold Regression. 2014; Annals of Statistics, to appear.
- Yang Y, Tokdar S.T.: MINIMAX-OPTIMAL NONPARAMETRIC REGRESSION IN HIGH DIMENSIONS. 2015; Annals of Statistics 43(2), 652–674.