

Odkrywanie współzależnych cech w danych silnie wielowymiarowych

Jacek Koronacki

Praca wspólna z Michałem Dramińskim

Poznań, 22 kwietnia 2016

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

Let us 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".

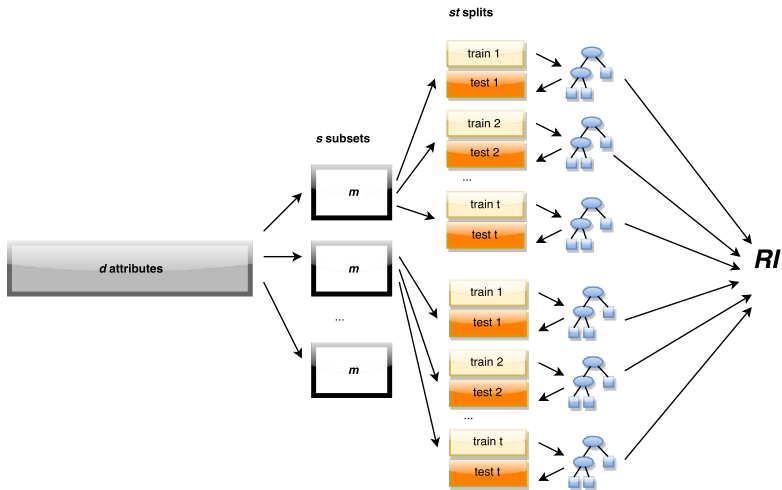
This "readiness" of a feature to take part in the classification process, termed relative importance of a feature, is measured via intensive use of classification trees. When assessing relative importance of a feature, the aforementioned "readiness" of the feature to appear in a given tree is suitably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

MCFS-ID Algorithm: the MCFS part, contd.

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 $m \ll d$, 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.

MCFS-ID Algorithm: the MCFS part, contd.



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.

Our 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.

The ID part of MCFS-ID Algorithm, contd.

To be more specific: 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 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* \rightarrow *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 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, $\tau = 1, \dots, s \cdot t$, 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

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

where $\text{GR}(n_k(\tau))$ 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 MCFS-ID Algorithm, contd.

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

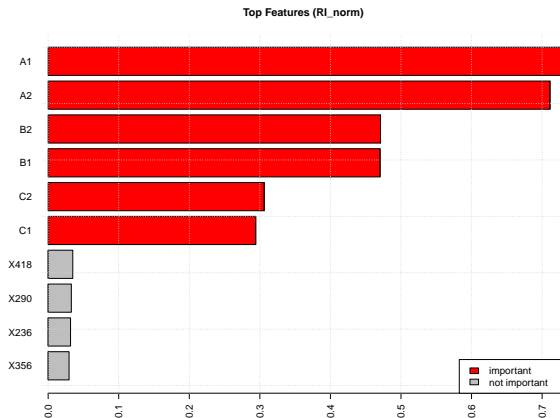
The ID part of MCFS-ID Algorithm, contd.

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 MCFS-ID Algorithm - a toy example

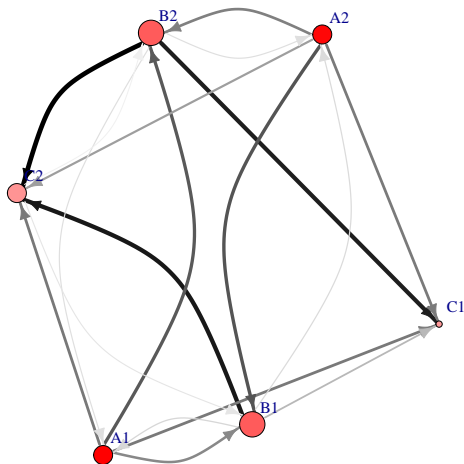
Consider objects from 3 classes, A , B and C , that contain 40, 20 and 10 objects, respectively (70 objects altogether). For each object, create 6 binary features ($A1$, $A2$, $B1$, $B2$, $C1$ and $C2$) that are 'ideally' or 'almost ideally' correlated with *class* feature. If an object's '*class*' equals ' 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 MCFS-ID Algorithm - a toy example



Rysunek: Top features selected by MCFS-ID.

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

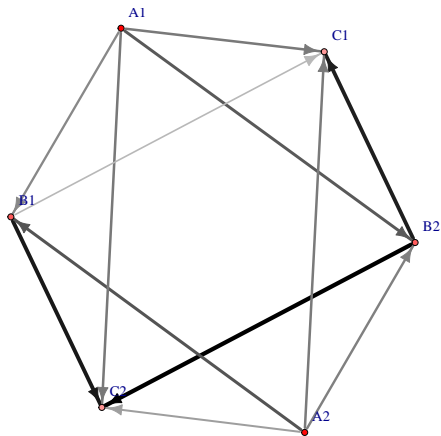


Rysunek: ID-Graph for artificial data.

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



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

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. (2016) for our proposal.

In lieu of a conclusion let state only 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, its explanatory power has been verified on a number of molecular and medical examples.

Acknowledgements

We thank our close collaborators, Jan Komorowski, Michal J. Dabrowski, Klev Diamanti, Marcin Kierczak, Marcin Kruczyk and Susanne Bornelöv, who have built on the MCFS-ID, most notably by providing a host of new insights and results within the area of bioinformatics.

- 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.
- Damiński, M., Koronacki, J., Komorowski, J.: A study on Monte Carlo Gene Screening. In: Intelligent Information Processing and Web Mining. 2005; Springer, 349-356.
- Damiński M, Rada Iglesias A, Enroth S, Wadelius C, Koronacki J, Komorowski J.: Monte Carlo feature selection for supervised classification. Bioinformatics. 2008; 24, 110-117.
- Damiński 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.
- Damiński 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. 2016; Springer, 285-304.

Please consult the given references for a vast literature on the topic discussed.