Neurocomputing 358 (2019) 255–274

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Stopping rules for mutual information-based feature selection

Jan Mielniczuk[a,b,*], Paweł Teisseyre[a]

[a] Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, Warsaw, 01–248, Poland
[b] Faculty of Mathematics and Information Sciences, Warsaw University of Technology, Warsaw, Poland

ARTICLE INFO

Article history:
Received 21 November 2018
Revised 14 March 2019
Accepted 17 May 2019
Available online 22 May 2019

Communicated by Jiayu Zhou

Keywords:
Entropy
Mutual information
Interaction information
Feature selection
Multiple hypothesis testing
Stopping rules

ABSTRACT

In recent years feature selection methods based on mutual information have attracted a significant attention. Most of the proposed methods are based on sequential forward search which add at each step a feature which is most relevant in explaining a class variable when considered together with the already chosen features. Such procedures produce ranking of features ordered according to their relevance. However significant limitation of all existing methods is lack of stopping rules which separate relevant features placed on the top of the ranking list from irrelevant ones. Finding an appropriate stopping rule is particularly important in domains where one wants to precisely determine the set of features affecting the class variable and discard the irrelevant ones (e.g. in genome-wide association studies the goal is to precisely determine mutations in DNA affecting the disease). In this work we propose stopping rules which are based on distribution of approximation of conditional mutual information given that all relevant features have been already selected. We show that the distribution is approximately chi square with appropriate number of degrees of freedom provided features are discretized into moderate number of bins. The proposed stopping rules are based on quantiles of the distribution and related p-values which are compared with thresholds used in multiple hypothesis testing. Importantly the proposed methods do not require additional validation data and are independent from the classifier. The extensive simulation experiments indicate that the rules separate relevant features from the irrelevant ones. We show experimentally that Positive Selection Rate (fraction of features correctly selected as relevant with respect to all relevant features) approaches 1, when sample size increases. At the same time, False Discovery Rate (fraction of irrelevant features selected with respect to all selected features) is controlled. The experiments on 17 benchmark datasets indicate that the classification models, built on features selected by the proposed methods, in 13 cases achieve significantly higher accuracy than the models based on all available features.

1. Introduction

1.1. Motivation

In the paper we address the problem of selecting relevant features in classification problem using information theoretic methods. The last years have witnessed a rapid and substantial advancement of feature selection methods coping with high dimensionality of data, especially when the relevant features are believed to be sparsely scattered among all features (for a comprehensive review see [13]). The existing methods can be divided into two groups: model-based and model-free methods. Model-based methods assume a specific structure of data generation. The behaviour of the procedures when the assumptions are not met i.e. the assumed model is misspecified is scarcely understood. Here we focus on fully non-parametric and model-free approach based on mutual information (MI) which has several important advantages. First it avoids reliance on a particular model which allows to find all features associated with the class variable, not only those which are indicated by an employed model. MI-based methods, unlike some classical approaches (e.g. regularization techniques used in logistic regression such as LASSO), are able to detect both linear and nonlinear dependencies between features and class variable. Moreover, some advanced MI based criteria are able to discover interactions between features as well as to take redundancy between features into account (we refer to [6] for a recent approach to discovery of interactions). Finally information-theoretic approach is versatile as it can be used for both classification and regression tasks, i.e.

* Corresponding author at: Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, Warsaw, 01–248, Poland.
E-mail addresses: Jan.Mielniczuk@ipipan.waw.pl (J. Mielniczuk), Pawel.Teisseyre@ipipan.waw.pl (P. Teisseyre).

https://doi.org/10.1016/j.neucom.2019.05.048
0925-2312/

nominal and quantitative class variable as well as for any type of the features.

1.2. Related work

The aforementioned advantages led to development of feature selection methods based on information theoretic concepts. In recent years several such methods have been proposed, we refer to [3,34] and [18] for recent comprehensive reviews of MI-based methods. Brown et. al. [3] unified several of these techniques using one framework. Theoretical properties of MI-based methods are also discussed in the recent paper [20], see also [26] for theoretical evaluation of the methods. The idea of the approach in a nutshell, is to choose, among available features $X_1, \ldots, X_p$ a subset of their indices S (of a given size) such that the joint mutual information $I(X_S, Y)$ between X_S and class variable Y is maximal. Finding an optimal feature set is usually hard to accomplish because the search space grows exponentially with the number of features. Moreover, adequate estimation of mutual information in many dimensions requires prohibitive amount of data. As a result various greedy algorithms have been developed including forward selection, backward elimination and genetic algorithms applied to low-dimensional approximations of MI or Conditional Mutual Information (CMI). Nowadays sequential forward selection (SFS) is the most commonly adopted solution. SFS starts from an empty set of features and in each step add a candidate feature which is most relevant in explaining Y when it is considered together with the already chosen variables. Such procedure is described in detail in Section 3. The most of the existing MI-based methods use SFS scheme, e.g. MIFS [2], JMI [38], CMIM [11], MRMR [27], CIFE [19], CMICOT [32], among others.

What is missing in all such proposed methods, however, is a designed stopping rule which would with large probability separate relevant and irrelevant features, that is it would determine on a greedy path the last relevant variable beyond which no extra information is gained when the process is pursued. Without such stopping rule we obtain ranking of *all* features defined by the order in which they appear on the greedy path, such ranking under sparsity is clearly superfluous and may be outright misleading when noise variables are unnecessarily compared and ranked.

The relevance of this problem for selection methods in general and those based on MI in particular is widely recognized, in the later context we refer to the review article [34], which points out that finding stopping rule is 'one of the most-important open problems in the field of MI-based feature selection'. This is particularly important in domains where the main goal of the analysis is to determine precisely the set of features affecting the class variable and discard the irrelevant ones. For example in Genome-Wide Association Studies (GWAS) [37] the research is focused on finding mutations of genes influencing the disease. At the same time the number of false discoveries (i.e. spurious mutations selected as relevant) should be controlled. The methods which only produce the ranking of all available features are not sufficient in this case.

Although the problem of determining the stopping rule seems to be crucial in some domains, there is a lack of simple solutions in machine learning literature. The frequently used ad hoc method is based on validation data. It involves plotting classification error with respect to nested groups of features ranked according to some criterion. Then the stopping rule is determined as the point at which the curve flattens out. Despite its simplicity the method suffers from some drawbacks. First it requires additional validation data, which may be problematic in the domains where collecting data is costly (e.g. in GWAS the number of patients is usually limited). Secondly it involves using a classifier, whose choice may affect the curve. In addition focusing on a particular classifier is superfluous when feature selection itself is the main goal of the analysis.

1.3. Contribution

In order to to overcome drawbacks of the existing approaches we propose a method which neither depends on a classifier nor does require validation data. The method is based on distribution of approximation of Conditional Mutual Information *given that all relevant features have been already detected along the greedy path*. It is argued in Section 3.2 that such distribution may be approximated by chi square distribution with appropriate number of degrees of freedom. As at each step we choose a new candidate among several available ones, a modifications of usual thresholds equal to quantiles of reference distributions are needed in order to account for that. In this way we arrive at several proposals of stopping rules which are defined and discussed in Section 4. Fig. 7 in Section 6 illustrates that the threshold based on appropriately chosen quantile of chi-squared distribution allows to correctly separate groups of relevant and irrelevant features.

We also note that distribution of CMI can be approximated using permutation methods, namely at each stage of building a greedy path, values of all unselected variables are permuted and Monte Carlo distribution of CMI is obtained. This was tried for small selection problems, see [29], however the approach is extremely computationally demanding as the permutation of the candidate feature must be done in every step of SFS procedure. The method becomes infeasible for large number of features and that is why we have not considered it here.

Let us also discuss the generality of the proposed method. As mentioned before direct use of the multivariate conditional information is not possible due to difficulties in estimation of mutual information in high-dimensional feature space. As a result, the MI-based methods, proposed in the literature, use various approximations of CMI. In this paper we focus on the most natural, second-order approximation of CMI following from so-called Möbius representation, discussed in Section 2. The considered approximation results in so-called CIFE criterion [19]. It seems impossible to propose a universal rule that would be valid for all MI-based methods. This is due to the fact that the reference distribution of the score function given that all relevant features have been already selected may differ between various approximations of CMI. Despite this we believe that some concepts used in our approach can be transferred to other methods, which is discussed in Section 7.

Note that as the discussed methods consist in sequential addition of new features, candidate features are compared to those chosen already. In order to choose the most promising candidates we apply multiple testing approach (see e.g. [7]) in which the outcome of the comparison for a particular candidate is treated as a separate test value. Depending on the method used this results either in a choice of a one or a batch of features which are added to the set of the already chosen ones.

Finally we stress that there are various desirable properties of stopping rule which can be taken into account. The most stringent one is that it separates exactly relevant and redundant features in majority of cases (selection consistency). Another, more lenient one is that in a chosen subset of features contains all relevant features (screening property). It is known that some popular selectors such as LASSO [33] posses only screening property for all but very restrictive experimental designs. One may also want to control values of relative criteria such as Positive Selection Rate (PSR) and False Discovery Rate (FDR) defined in Section 6. We will address this issue when discussing properties of introduced stopping rules.

Our contribution to the subject is as follows.

1. We propose a novel approach to stopping rules for mutual information sequential feature selection which is based on multiple testing.
2. Distribution of approximation of the conditional mutual information is theoretically and numerically investigated.
3. We perform experiments on both artificial and real datasets which indicate that the proposed methods separate relevant features from the irrelevant ones with large probability and allow to build classification models with higher accuracy when compared with their counterparts based on all features.

The paper is structured as follows: in Section 2 we discuss information theoretic preliminaries and in Section 3 mutual information-based feature selection. In Section 4 we discuss approximate distribution of CMI and justify them theoretically and numerically. This leads to proposals of stopping rules in Section 5, which are investigated in Section 6 for both artificial and real data. In Section 7 we conclude the paper and discuss future research.

2. Preliminaries

2.1. Notation

We consider qualitative p features $X_1, \ldots, X_p$ and a qualitative class variable Y. Let X_S be a subset of features $X_1, \ldots, X_p$, indexed by set $S \subset \{1, \ldots, p\}$. We denote by $p(x_j) := P(X_j = x_j)$, $x_j \in \mathcal{X}_j$ a probability mass function corresponding to X_j, where $\mathcal{X}_j$ is a domain of X_j and $|\mathcal{X}_j|$ is its cardinality. The domain of a class variable is $\mathcal{Y}$. Joint probability will be denoted by $p(x_i, x_j) = P(X_i = x_i, X_j = x_j)$. Notation $\hat{p}(x_j)$ will be used for the sample estimate of $p(x_j)$.

2.2. Entropy and mutual information

Below we recall basic quantities considered in Information Theory. Entropy for discrete random variable X_j is defined as

$$H(X_j) = -\sum_{x_j} p(x_j) \log p(x_j). \tag{1}$$

Entropy quantifies the uncertainty of observing random values of X_j. If large mass of the distribution is concentrated on one particular value of X_j then its entropy is low. If all values are equally likely then $H(X_j)$ is maximal. The above definition can be naturally extended to the case of random vectors (i.e. when X_j is a multivariate random variable) by using multivariate mass function instead of univariate mass function. The conditional entropy of Y given X_j is

$$H(Y|X_j) = \sum_{x_j} p(x_j) H(Y|X_j = x_j). \tag{2}$$

The mutual information (MI) between X_j and Y is

$$I(Y, X_j) = H(Y) - H(Y|X_j) = \sum_{x_j, y} p(x_j, y) \log \frac{p(x_j, y)}{p(x_j)p(y)}. \tag{3}$$

This can be interpreted as the amount of uncertainty in Y which is removed when X_j is known which is consistent with the intuitive meaning of mutual information as the amount of information that one variable provides about another. The second equality in (3) indicates that it determines how similar the joint distribution is to the product of marginal distributions. It can be expressed as a Kullback–Leibler divergence between these two distributions. The mutual information is equal zero if and only if X_j and Y are independent. MI is always non-negative and its value corresponds to a strength but not to a direction of dependence i.e. mutual information does not distinguish between positive and negative association. MI is useful within the context of feature selection because it gives a way to quantify the relevance of a feature (or feature subset) with respect to the class variable. The conditional mutual information

$$\begin{aligned} I(Y, X_j|X_i) &= H(Y|X_i) - H(Y|X_i, X_j) \\ &= \sum_{x_i} p(x_i) \sum_{x_j, y} p(x_j, y|x_i) \log \frac{p(x_j, y|x_i)}{p(y|x_i)p(x_j|x_i)} \\ &= \sum_{x_i, x_j, y} p(x_i, x_j, y) \log \frac{p(x_i, x_j, y)p(x_i)}{p(y, x_i)p(x_j, x_i)}. \end{aligned} \tag{4}$$

measures the conditional dependence between X_j and Y given X_i. Note that the conditional mutual information is mutual information of X_j and Y given $X_i = x_i$ averaged over values of x_i. It is equal zero if and only if X_j and Y are conditionally independent given X_i. For more properties of the basic measures described above we refer to [4,39].

2.3. Interaction information

An important quantity, used in next sections, is interaction information (II) [21]. The 3-way interaction information is defined as

$$\begin{aligned} II(X_i, X_j, Y) = &- H(X_i) - H(X_j) - H(Y) + H(X_i, Y) \\ &+ H(X_j, Y) + H(X_i, X_j) - H(X_i, X_j, Y). \end{aligned} \tag{5}$$

It can be easily proved that II can be also written as

$$II(X_i, X_j, Y) = I((X_i, X_j), Y) - I(X_i, Y) - I(X_j, Y), \tag{6}$$

which is more intuitive form. Namely, it follows from (6) that II can be interpreted as a part of the mutual information of (X_i, X_j) and Y which is due solely to interaction between X_i and X_j in predicting Y i.e. the part of $I((X_i, X_j), Y)$ which remains after subtraction of amount of individual informations between Y and X_i and Y and X_j. In other words, II is obtained by removing the main effects from the term describing the overall dependence between Y and the pair (X_i, X_j). Interaction information can be also written as

$$II(X_i, X_j, Y) = I(Y, X_j|X_i) - I(Y, X_j), \tag{7}$$

which is consistent with the intuitive meaning of existence of interaction as a situation in which the effect of one variable on the class variable depends on the value of another variable. The properties of 3-way II, in the context of finding interactions between genes, are discussed in recent works see e.g. [23,24,36]. The 3-way II can be extended to the general case of m variables. The m-way interaction information [12,14] is

$$II(X_1, \ldots, X_m) = -\sum_{T \subseteq \{1, \ldots, m\}} (-1)^{m-|T|} H(X_T). \tag{8}$$

For $m = 2$, equality (8) reduces to mutual information, whereas for $m = 3$ it reduces to (5). The concept of m-way interaction information is useful in the context of feature selection. Namely it follows from so-called Möbius representation (cf e.g. [22]) that the joint mutual information between feature subset X_S and a class variable Y can be expressed as

$$I(X_S, Y) = \sum_{k=1}^{|S|} \sum_{\{t_1, \ldots, t_k\} \subseteq S} II(X_{t_1}, \ldots, X_{t_k}, Y). \tag{9}$$

Thus low-dimensional approximations of $I(X_S, Y)$ can be obtained by truncation of the number of terms in (9).

2.4. Estimation of entropy-based terms

In previous sections we implicitly assumed that we have knowledge of the underlying distributions. In practice the true probabilities have to be estimated from data. The sample estimate of $p(x_j)$ is given as the frequency of occurrence of the event $X_j = x_j$ divided by the sample size, i.e. $\hat{p}(x_j) = \#\{i \in \{1,\ldots,n\} : X_j^{(i)} = x_j\}/n$, where $X_j^{(1)},\ldots,X_j^{(n)}$ are realizations of r.v. X_j. Analogously we estimate conditional and joint probabilities. The entropy is estimated as

$$\hat{H}(X_j) = -\sum_{x_j \in \mathcal{X}_j} \hat{p}(x_j)\log \hat{p}(x_j).$$

The problem of estimating mutual information and interaction information reduces to that of entropy estimation. The sample estimators of quantities described in previous sections will be denoted using 'hat' symbol. Entropy estimation for continuous data is highly non-trivial and we do not discuss it in the present work. For more information on entropy estimation procedures, we refer the reader to [25].

3. Mutual information-based feature selection

3.1. Sequential forward search (SFS)

In this work we focus on feature selection based on mutual information (MI). MI-based feature selection is concerned with identifying a feature subset of fixed size $1 \le k \le p$ that maximizes the joint mutual information with a class variable Y, i.e we look for

$$\arg\max_{S:|S|=k} I(X_S, Y),$$

where X_S denotes a subset of features $X_1,\ldots,X_p$, indexed by set $S \subset \{1,\ldots,p\}$. Finding an optimal feature set is usually infeasible because the search space grows exponentially with the number of features. As a result most employed algorithms are based on sequential suboptimal strategies and low-dimensional approximations of $I(X_S, Y)$. In particular, sequential forward selection (SFS) is the most commonly used solution. SFS algorithms start from an empty set of features and add, in each step, the feature that jointly, i.e. together with already selected features, has the maximal joint mutual information with the class. Formally, assume that S is a set of already chosen features, S^c is its complement and X_j, $j \in S^c$ is a candidate feature. In each step we add a feature whose inclusion gives the most significant improvement of the mutual information, i.e. we find

$$\arg\max_{j\in S^c}[I(X_{S\cup\{j\}}, Y) - I(X_S, Y)] = \arg\max_{j\in S^c} I(X_j, Y|X_S). \tag{10}$$

The equality in (10) follows from (6) and (7). Observe that (10) indicates that we select a feature that has the maximum association with the class given the already chosen features.

3.2. Approximations of conditional mutual information

Criterion (10) is appealing and attracted a significant attention. However in practice the estimation of joint mutual information (or conditional mutual information) is problematic even for small cardinality of S. This makes a direct application of (10) infeasible. A rich body of work in the MI-based feature selection literature approaches this difficulty by approximating the high-dimensional joint MI with a sum of low-dimensional MI terms. The natural way to approximate the conditional mutual information (CMI) is to use Möbius representation (9) which gives

$$\begin{aligned} &I(X_{S\cup\{j\}}, Y) - I(X_S, Y) \\ &\quad = I(X_j, Y|X_S) = \sum_{k=0}^{|S|}\sum_{\{i_1,\ldots,i_k\}\subseteq S} II(X_{i_1},\ldots,X_{i_k}, X_j, Y) \\ &\quad = I(X_j, Y) + \sum_{i\in S} II(X_i, X_j, Y) + \sum_{i_1,i_2\in S: i_1<i_2} II(X_{i_1}, X_{i_2}, X_j, Y) \\ &\qquad + \ldots + II(X_{i_1},\ldots,X_{i_{|S|}}, X_j, Y). \end{aligned} \tag{11}$$

The above formula allows to obtain various natural approximations of CMI. For example, consideration of only the first term of the sum in (11) leads to first-order approximation equal to $I(X_j, Y)$, which is a simple univariate filter, frequently used as a preprocessing step in high-dimensional data analysis. However this method suffers from many drawbacks; it does not take into account possible interactions between features and redundancy of some features. In particular, choosing top k features with respect to their MI with Y will result in possible inclusion of redundant features. In this paper we focus on second order approximation, which is a compromise between relatively accurate approximation of CMI and low computational cost. Note that it involves three-way interactions $II(X_i, X_j, Y)$, which account for both interactions and redundancy. The positive value of II indicates the existence of interaction, e.g. for $Y = XOR(X_i, X_j)$, being indicator function of the event $\{X_i \neq X_j\}$ we have $II(X_i, X_j, Y) = \log(2) > 0$, when all three variables are binary. On the other hand, the negative value of II indicates redundancy, e.g. for $Y = X_i = X_j$, we have $II(X_i, X_j, Y) = -\log(2) < 0$, when all three variables are binary. The score function for X_j is defined as a second order approximation of (11), i.e.

$$\begin{aligned} J(X_j, S) &= I(X_j, Y) + \sum_{i\in S} II(X_i, X_j, Y) \\ &= I(X_j, Y) + \sum_{i\in S}[I(Y, X_j|X_i) - I(Y, X_j)] \\ &= I(X_j, Y)(1 - |S|) + \sum_{i\in S} I(Y, X_j|X_i). \end{aligned} \tag{12}$$

The second equality in (12) follows from property (7). Note that $J(X_j, \emptyset) = I(X_j, Y)$.

In literature (12) is known as CIFE (Conditional Infomax Feature Extraction) [19] criterion. Analogous approximation was also considered for multi-label case, see [15]. Observe that in (12) we take into account not only relevance of the candidate feature, but also the possible interactions between the already selected features and the candidate feature. The empirical evaluation indicates that (12) is among the most successful MI-based methods, see [3] for extensive comparison of several MI-based feature selection methods. Some additional assumptions lead to other score functions. For example MIFS (Mutual Information Feature Selection) criterion [2]

$$I(X_j, Y) - \sum_{i\in S} I(X_i, X_j), \tag{13}$$

is a special case of (12). MIFS is obtained from (12) by additionally assuming that X_i and X_j, $j \in S$ are conditionally independent given Y. Let us also mention that more accurate approximations of CMI can be considered [35]. However using higher-order (> 3) approximations of CMI becomes problematic. First, the computational cost increases significantly. Let $|S^c|$ be a number of candidate features. For the first-order approximation of (11) one needs to calculate $|S^c|$ terms. For the second-order approximation, the complexity is $(|S| + 1)|S^c|$ terms; for the third-order term it is $(1 + |S| + \binom{|S|}{2})|S^c|$ terms, etc. Secondly, the estimation of multivariate entropy terms is difficult, particularly for small or moderate sample sizes. Note that the second order approximation of

(11) requires estimation of 3-dimensional probabilities in order to estimate entropy-based terms, the third order approximation requires estimation of 4-dimensional probabilities, etc. In general we have to estimate $(r+1)$-dimensional probabilities to compute terms in r-th order approximation of (11). When features are discretized into $|\mathcal{X}| = b$ bins, there are b^{r+1} possible combinations of feature values. This gives, on average, n/b^{r+1} observations for each combination, where n is a sample size. For example, assume that we consider 2-nd order approximation for moderate sample size $n = 1000$ and $b = 5$. Than we have, on average, $1000/125 = 8$ observations for each combination. For the 3-rd order approximation we have only $100/625 = 1.6$ observations for each combination, which makes efficient estimation infeasible. Due to above problems the most of the existing MI-based methods use only entropy terms of order up to 3.

3.3. Stopping rule

The Sequential Forward Search (SFS) procedures, described in Section 3.1 allow to find the ranking of features, starting from the most relevant one and ending with the least relevant one according to the chosen score function. However SFS procedures do not include stopping rule. This means that candidate features are added even if they are not relevant any more, i.e. they are conditionally independent from the class variable, given the already selected features, thus if $I(Y, X_j|X_S) = 0$ for any $j \in S^c$. As pointed out in Section 3.2, CMI cannot be directly used in SFS procedure. Instead we use approximation (12). Let S_k be a set of indices corresponding to features selected in k-th step of the SFS procedure, where $S_0 = \emptyset$. In k-th step we set $S_{k+1} = S_k \cup \{j_k\}$, such that $j_k = \arg\max_{j \in S_k^c} J(X_j, S_k)$. Thus stopping rule involves approximating

$$t := \arg\min_{1 \le k \le p} J(X_j, S_k) = 0, \forall j \in S_k^c. \tag{14}$$

Observe that $J(X_j, S_k)$ includes unknown terms and needs to be estimated from data as

$$\begin{aligned}\hat{J}(X_j, S_k) &= \hat{I}(X_j, Y) + \sum_{i \in S_k} \hat{II}(X_i, X_j, Y) \\ &= \hat{I}(X_j, Y) + \sum_{i \in S_k} [\hat{I}(Y, X_j|X_i) - \hat{I}(Y, X_j)] \\ &= \hat{I}(X_j, Y)(1 - |S_k|) + \sum_{i \in S_k} \hat{I}(Y, X_j|X_i). \end{aligned} \tag{15}$$

Thus t needs to be estimated from data as well. Observe that it may happen that $J(X_j, S_k) = 0$ whereas $\hat{J}(X_j, S_k) > 0$. The aim of the next sections is to approximate the distribution of $2n\hat{J}(X_j, S_k)$ under the null hypothesis that X_j are conditionally independent given X_{S_k} (denoted by $X_j \perp Y|X_{S_k}$) and propose stopping rules $\hat{t}$ based on the quantiles of this distribution.

4. Distributions of approximation for conditional mutual information

Approximate distribution of $\hat{J}(X_j, S)$ is based on the following Theorem, the proof of which is given below. For simplicity, we write in this Section S instead of S_k.

Theorem 1. *Let X, Y and Z be three qualitative variables having $|\mathcal{X}|$, $|\mathcal{Y}|$ and $|\mathcal{Z}|$ values, respectively. Assume that Y and Z are independent given X. Then*

$$2n\hat{I}(Y, Z|X) \approx \sum_{i=1}^{|\mathcal{X}|} W_i, \tag{16}$$

where W_i has asymptotically χ^2 distribution with $(|\mathcal{Y}| - 1)(|\mathcal{Z}| - 1)$ degrees of freedom and $\approx$ means that both sides differ by a term which is negligible in probability for n tending to infinity.

Proof. In order to ease notation we will use $\hat{p}_{ijk} = \hat{p}(X = x_i, Y = y_j, Z = z_k)$, $\hat{p}_{ij} = \hat{p}(X = x_i, Y = y_j)$ and so on. We write using (4)

$$\begin{aligned} 2n\hat{I}(Y, Z|X) &= 2n \sum_{i,j,k} \hat{p}_{ijk} \log \frac{\hat{p}_{ijk}\hat{p}_i}{\hat{p}_{ij}\hat{p}_{ik}} \\ &= 2n \sum_i \frac{1}{\hat{p}_i} \sum_{j,k} \hat{p}_{ijk}\hat{p}_i \log\left(1 + \frac{\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}}{\hat{p}_{ij}\hat{p}_{ik}}\right). \end{aligned} \tag{17}$$

Using the expansion $\log(1 + x) = x - x^2/2 + O(x^3)$ for small x we have that

$$\begin{aligned} \log\left(1 + \frac{\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}}{\hat{p}_{ij}\hat{p}_{ik}}\right) &= \frac{\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}}{\hat{p}_{ij}\hat{p}_{ik}} - \frac{1}{2}\frac{(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik})^2}{(\hat{p}_{ij}\hat{p}_{ik})^2} \\ &\quad + O\left(\frac{(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik})^3}{(\hat{p}_{ij}\hat{p}_{ik})^3}\right). \end{aligned} \tag{18}$$

Plugging the above expansion into (17) we see that the term pertaining to the last term in (18) is bounded for some $C > 0$ by

$$\begin{aligned} &C \times 2n \sum_i \frac{1}{\hat{p}_i} \sum_{j,k} \frac{(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik})^3}{(\hat{p}_{ij}\hat{p}_{ik})^3} \\ &\le C \times 2n \sum_i \frac{1}{\hat{p}_i} \times \frac{\max_{j,k} |\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}|}{\min_{j,k}(\hat{p}_{ij}\hat{p}_{ik})^2} \sum_{j,k} \frac{(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik})^2}{\hat{p}_{ij}\hat{p}_{ik}} \\ &= C \times 2 \sum_i \frac{1}{\hat{p}_i^2} \times \frac{\max_{j,k} |\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}|}{\min_{j,k}(\hat{p}_{ij}\hat{p}_{ik})^2} \\ &\quad \times n\hat{p}_i \sum_{j,k} \frac{(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik})^2}{\hat{p}_{ij}\hat{p}_{ik}}. \end{aligned} \tag{19}$$

Moreover, due to conditional independence and convergence $\hat{p}_{ij} \to p_{ij} > 0$, $\hat{p}_{ik} \to p_{ik} > 0$ we have that

$$|\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}| \to |p_{ijk}p_i - p_{ij}p_{ik}| = 0, \tag{20}$$

when $n \to \infty$. In view of (20) and (23) below, the last term in (19) is a sum of products of two terms such that the first terms converge to zero and the second terms have chi-squared distribution. Therefore, it follows from Slutsky's theorem (cf. Section 1.5.4 in [30]) that the bound in (19) converges in probability to 0, when $n \to \infty$. Thus we have that $2n\hat{I}(Y, Z|X)$ is approximately equal to

$$\begin{aligned} &2n \sum_i \frac{1}{\hat{p}_i} \sum_{j,k} [\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik} + \hat{p}_{ij}\hat{p}_{ik}] \\ &\quad \times \left[\frac{\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}}{\hat{p}_{ij}\hat{p}_{ik}} - \frac{1}{2}\frac{(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik})^2}{(\hat{p}_{ij}\hat{p}_{ik})^2}\right] \\ &= 2n \sum_i \frac{1}{\hat{p}_i} \sum_{j,k} \left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik} + \frac{(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik})^2}{\hat{p}_{ij}\hat{p}_{ik}}\right. \\ &\quad \left. - \frac{1}{2}\frac{(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik})^2}{\hat{p}_{ij}\hat{p}_{ik}} + \frac{1}{2}\frac{(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik})^3}{(\hat{p}_{ij}\hat{p}_{ik})^2}\right) \\ &\approx n \sum_i \frac{1}{\hat{p}_i} \sum_{j,k} \left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik} + \frac{(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik})^2}{\hat{p}_{ij}\hat{p}_{ik}}\right) \\ &= n \sum_i \frac{1}{\hat{p}_i} \sum_{j,k} \frac{(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik})^2}{\hat{p}_{ij}\hat{p}_{ik}}, \end{aligned} \tag{21}$$

where the approximation in (21) is obtained analogously to (19) and the last equality follows from noting that $\sum_{j,k} \hat{p}_{ijk}\hat{p}_i -$

$\hat{p}_{ij}\hat{p}_{ik}=0$. It is easy to see that the obtained expression equals

$$n\sum_i \hat{p}_i \sum_{j,k}\frac{(\hat{p}_{ijk}/\hat{p}_i-\hat{p}_{ij}\hat{p}_{ik}/\hat{p}_i^2)^2}{\hat{p}_{ij}\hat{p}_{ik}/\hat{p}_i^2}$$

$$=\sum_i n_i \sum_{j,k}\frac{(\hat{p}_{ijk}/\hat{p}_i-\hat{p}_{ij}\hat{p}_{ik}/\hat{p}_i^2)^2}{\hat{p}_{ij}\hat{p}_{ik}/\hat{p}_i^2}=:\sum_{i=1}^{|\mathcal{X}|}W_i, \tag{22}$$

where $n_i=n\hat{p}_i$. Observe that W_i is exactly chi square statistics for testing independence of Y and Z on the strata $X=x_i$, which under assumed independence of Y and Z given X has asymptotic χ^2 distribution with $(|\mathcal{Y}|-1)(|\mathcal{Z}|-1)$ degrees of freedom

$$W_i=n_i\sum_{j,k}\frac{(\hat{p}_{ijk}\hat{p}_i-\hat{p}_{ij}\hat{p}_{ik})^2}{\hat{p}_{ij}\hat{p}_{ik}}\approx\chi^2_{(|\mathcal{Y}|-1)(|\mathcal{Z}|-1)} \tag{23}$$

due to Fisher (c.f. [10]), see also [1], section 3.2.1 and [31], Theorem 6.9). In the last referenced monograph it is shown that the number of degrees of freedom of the limiting χ^2 distribution equals $p-s-1$, where $p=|\mathcal{Y}||\mathcal{Z}|$ and $s=(\mathcal{Y}|-1)+(|\mathcal{Z}|-1)$, as under hypothesis of conditional independence $(|\mathcal{Y}|-1)+(|\mathcal{Z}|-1)$ marginal conditional probabilities of Y and Z given $X=x_i$ are needed to determine the conditional distribution of (Y, Z) given $X=x_i$. Thus we have that the number of degrees of freedom equals $|\mathcal{Y}||\mathcal{Z}|-(|\mathcal{Y}|-1)-(|\mathcal{Z}|-1)-1=(|\mathcal{Y}|-1)(|\mathcal{Z}|-1)$. □

It is also conjectured that under assumed conditional independence of Y and Z given X variables W_is are asymptotically independent and the sum $\sum_{i=1}^{|\mathcal{X}|}W_i$ has chi square distribution with $|\mathcal{X}|(|\mathcal{Y}|-1)(|\mathcal{Z}|-1)$ degrees of freedom approximately. The result is true in particular case when $|\mathcal{Y}|=|\mathcal{Z}|=2$ and $\mathcal{X}$ is arbitrary (cf. [28]).

We apply this reasoning to the summands $\hat{I}(X_j,Y|X_i)$ of $\hat{J}(X_j,S)$ in (15). Then each $\hat{I}(X_j,Y|X_i)$ is approximately chi square distributed with $|\mathcal{X}_i|(|\mathcal{X}_j|-1)(|\mathcal{Y}|-1)$ degrees of freedom. In view of (15), the above result and the fact that $2n\hat{I}(X_j,Y)$ has approximately χ^2 distribution with $(|\mathcal{X}_j|-1)((|\mathcal{Y}|-1)$ degrees of freedom we approximate $2n\hat{J}(X_j,S)$ with chi square distribution with

$$d=d(j,|S|)=(|\mathcal{X}_j|-1)(|\mathcal{Y}|-1)\sum_{i=1}^{|S|}|\mathcal{X}_i| +(1-|S|)(|\mathcal{X}_j|-1)(|\mathcal{Y}|-1) \tag{24}$$

degrees of freedom. Note that in view of representation (16) expected value of $2n\hat{J}(X_j,S)$ is approximately $d(j,|S|)$ (recall that the expected value of chi square distribution χ^2_d with d degrees of freedom is d).

Further justification of the number of degrees of freedom in (24) follows from ([12]) and the first representation in (15). Namely, Han [12] proved that provided that X_i, X_j and Y are jointly independent then $2n\hat{I}(X_i,X_j,Y)$ is approximately chi-square with $(|\mathcal{X}_i|-1)(|\mathcal{X}_j|-1)(|\mathcal{Y}|-1)$ degrees of freedom. Summing up degrees of freedom for the summands in (15) we obtain

$$(\mathcal{X}_j|-1)(|\mathcal{Y}|-1)\sum_{i=1}^{|S|}(|\mathcal{X}_i|-1)+(|\mathcal{X}_j|-1)(|\mathcal{Y}|-1)=d(j,|S|).$$

We give below a convincing numerical evidence that for a qualitative features with small number of values or coarse discretization (small number of discretization bins) this distribution closely approximates distribution of $2n\hat{J}(X_j,S)$ when X_j is a randomly chosen candidate such that $X_j\perp Y|X_S$. In order to check this, we considered two artificial datasets M1 and M2 for which Y depends on 2 and 4 relevant features, respectively and all features have $|\mathcal{X}|=2,5$ or 10 possible values (the description of the datasets is given in Section 6.1).

For each run of simulated data from these models greedy search of S containing prescribed number of variables ($|S|=5,10,15$ and 20) was performed and only the runs for which S contained all relevant variables were retained. Thus in the case of M2 and $|S|=20$ set S contains 4 relevant variables and 16 spurious ones; the choice of the latter depending on the run. For a chosen set S, distribution of $2n\hat{J}(X_j,S)$ over all potential $p-|S|$ with $p=2000$ candidates in S^c is calculated and compared with chi-square distribution with d degrees of freedom by means of quantile plot. In order to account for variability due to particular run, 200 such runs were performed and for every chosen $0<\alpha<1$, 10th and 90th percentile of empirical quantiles $\hat{q}_{1,\alpha},\dots\hat{q}_{200,\alpha}$ of $\hat{J}$ were plotted against quantile of chi-square distribution with d dfs. Envelopes of quantile plots obtained in this way are shown in Figs. 1–6 for datasets M1 and M2, various levels of discretization and values of $|S|$.

It is seen that for coarse discretization when $|\mathcal{X}|\le 5$ the agreement between empirical distribution of $2n\hat{J}(X_j,S)$ and chi-square distribution is very high with slight curvature in the plot occurring for large $|S|$. Note small variability of the empirical distribution between runs indicated by the narrowness of the envelope bands with only a slight increase of variability in the right tail. For finer discretizations ($|\mathcal{X}|\ge 10$) the approximation deteriorates. The plots indicate that upper quantiles of empirical distribution are smaller than chi-square distribution which leads to larger number of undetected relevant variables and smaller PSR. Thus it seems reasonable to approximate distribution of $2n\hat{J}(X_j,S)$ when continuous variables are discretized using up to 5 levels. We also note that there are many important studies such as GWAS for which all features are qualitative and have a small number of categories (in the case of GWAS there are three categories corresponding to three possible genotypes at each locus).

We remark also that the structure of (15) suggests possible different approximation of distribution of CMI being distribution of two weighted independent chi squares, the first with the weight 1 and degrees of freedom equal to $(|\mathcal{X}_j|-1)(|\mathcal{Y}|-1)\sum_{i=1}^{|S|}|\mathcal{X}_i|$ and the second one with the weight $1-|S|$ and $(\mathcal{X}_j|-1)(\mathcal{Y}|-1)$ degrees of freedom. This have been tried but no significant improvement in approximating distribution of $\hat{J}(X_j,S)$ has been obtained.

5. Stopping rules

5.1. Proposed methods

In order to simplify the exposition we assume now that all features X_i have the same number of values equal to $|\mathcal{X}|$. Then number of degrees of freedom in (24) then equals

$$d(S)=|S||\mathcal{X}|(|\mathcal{X}|-1)(|\mathcal{Y}|-1)+(1-|S|)(|\mathcal{X}|-1)(|(|\mathcal{Y}|-1).$$

We will introduce and investigate stopping rules which are based on our main finding that distribution of $2n\hat{J}(X_j,S)$ is closely approximated by $\chi^2_{d(S)}$. We will employ multiple hypotheses testing approach, which accounts for the fact that at each stage we choose one or more candidates among many using multiple testing. The first approach uses Bonferroni correction [8], p. 74 to determine stopping point, the second one is a group of methods which allow multiple features added at each step while relaxing Bonferroni-based threshold at the same time.

Let $S_0=\emptyset$ and

$$j_k=\text{argmax}_{j\in S_k^c}2n\hat{J}(X_j,S_k) \tag{25}$$

be the index of candidate feature with the maximal score (note that for $k=0$, $j_0=\text{argmax}_{j\in S_k^c}2n\hat{I}(X_j,Y)$). Then $S_{k+1}=S_k\cup\{j_k\}$ and

$$\hat{t}_{\text{chi-Bonf}}=\text{argmin}_{k=1,\dots,p}\{2n\hat{J}(X_{j_k},S_k)\le\chi^2_{1-\alpha_k,d(S_k)}\}, \tag{26}$$

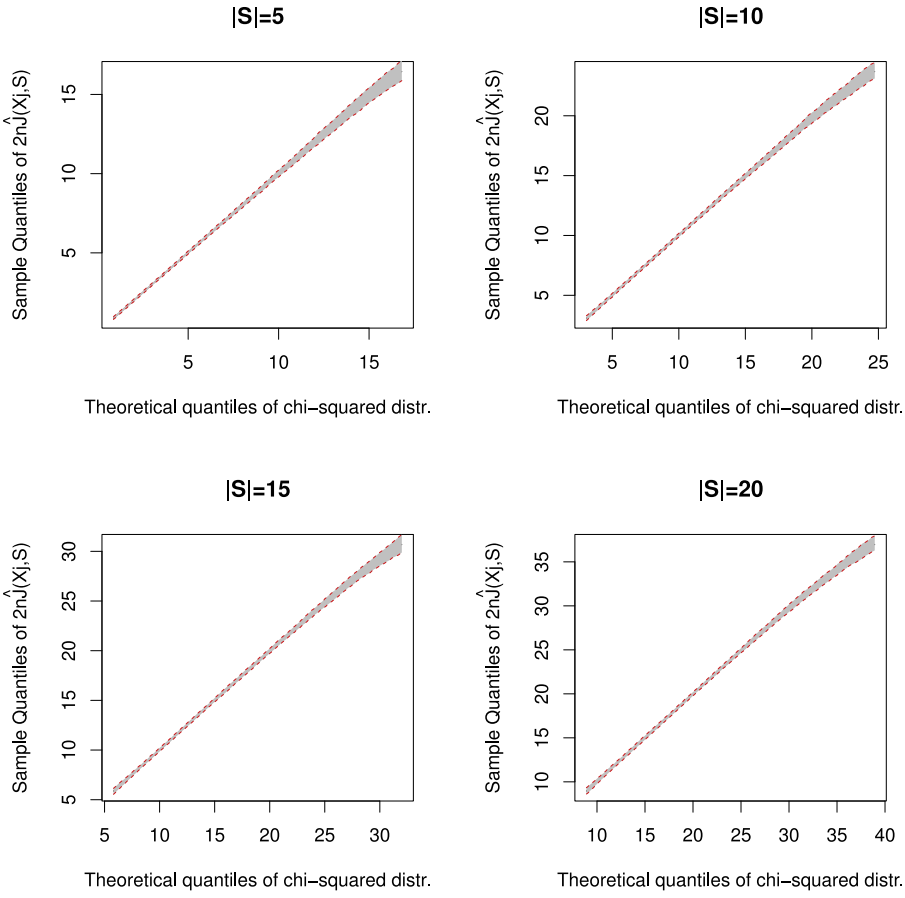

Fig. 1. Envelopes of quantile plots for simulated dataset M1 and for discretization with $|\mathcal{X}| = 2$ bins.

where $\alpha_k = \alpha/(p - |S_k|)$. Note that as the number of candidates at step k equals $p - |S_k|$, magnitude of Bonferroni correction changes at each step to account for that. Note also that for this method $|S_k| = k$, in contrast to the second method when multiple features can be added at each step. The SFS procedure with stopping rule chi-Bonf is described by Algorithm 1.

Algorithm 1: Sequential forward search (SFS) with stopping rule chi-Bonf.

Input : Training data of size n containing features $X_1, \ldots, X_p$ and class variable Y.

$S_0 = \emptyset$

for $k = 0, 1, \ldots$ **do**

  $\alpha_k = \alpha/(p - |S_k|)\, j_k = \arg\max_{j \in S_k^c} 2n\hat{I}(X_j, S_k)$

  if $2n\hat{I}(X_{j_k}, S_k) \leq \chi^2_{1-\alpha_k, d(S_k)}$ **then**

    stop

  else

    $S_{k+1} \leftarrow S_k \cup \{j_k\}$

Output : Relevant features S_k.

Fig. 7 visualizes the method; we show here how $\max_{j \in S_k^c} 2n\hat{I}(X_j, S_k)$ depends on k for one simulation trial in the case of four artificial datasets (see Section 6.1 for detailed description of the simulated datasets). Number of relevant features (i.e. those influencing class variable) varies from 2 to 12. Red line corresponds to the threshold $\chi^2_{1-\alpha_k, d(S_k)}$. Features above the red line are selected as relevant whereas features below the red line are recognized as irrelevant. Observe that all relevant features are correctly recognized as relevant for all considered datasets. In the case of dataset M3 (bottom-left figure) one irrelevant feature is incorrectly detected as relevant. Let us also discuss how the choice of parameter α affects the performance. Parameter α corresponds to the significance level of the test of conditional independence of X_j and Y, given X_{S_k}. If the value of α is small then it is more difficult to recognize candidate feature X_j as relevant (i.e. reject the null hypothesis). For large α, X_j is more likely to be recognized as relevant. In most experiments we set $\alpha = 0.05$, which is a default value in hypothesis testing. The influence of α on the performance is discussed in Section 6.1, see Fig. 12.

In order to define group of methods which allow adding batches of features simultaneously let at step k denote by

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(p-|S_k|)}$$

ordered p-values corresponding to values of statistic $2n\hat{I}(X_j, S_k)$ for $j \in S_k^c$ with related features denoted as $X_{j_1}, \ldots X_{j_{p-|S_k|}}$. Thus X_{j_1} corresponds to $p_{(1)}$, X_{j_2} corresponds to $p_{(2)}$ and so on. P-values are calculated w.r.t. reference distribution $\chi^2_{d(S_k)}$. Moreover, define

$$k^* = \min\{j : p_{(j)} > \frac{\alpha}{p - |S_k| - j + 1}\} \tag{27}$$

and $J_k = \{j_1, \ldots, j_{k^*-1}\}$. Let $S_{k+1} = S_k \cup J_k$. Holm procedure [8], p. 79 defines a stopping rule as

$$\hat{t}_{\text{chi}-\text{Holm}} = \text{argmin}\{k = 1, 2, \ldots : J_k = \emptyset\}.$$

Bonferroni procedure is known to be conservative i.e. for a fixed number of tested hypotheses its family wise error rate is significantly smaller than α. To alleviate this drawback Holm's procedure

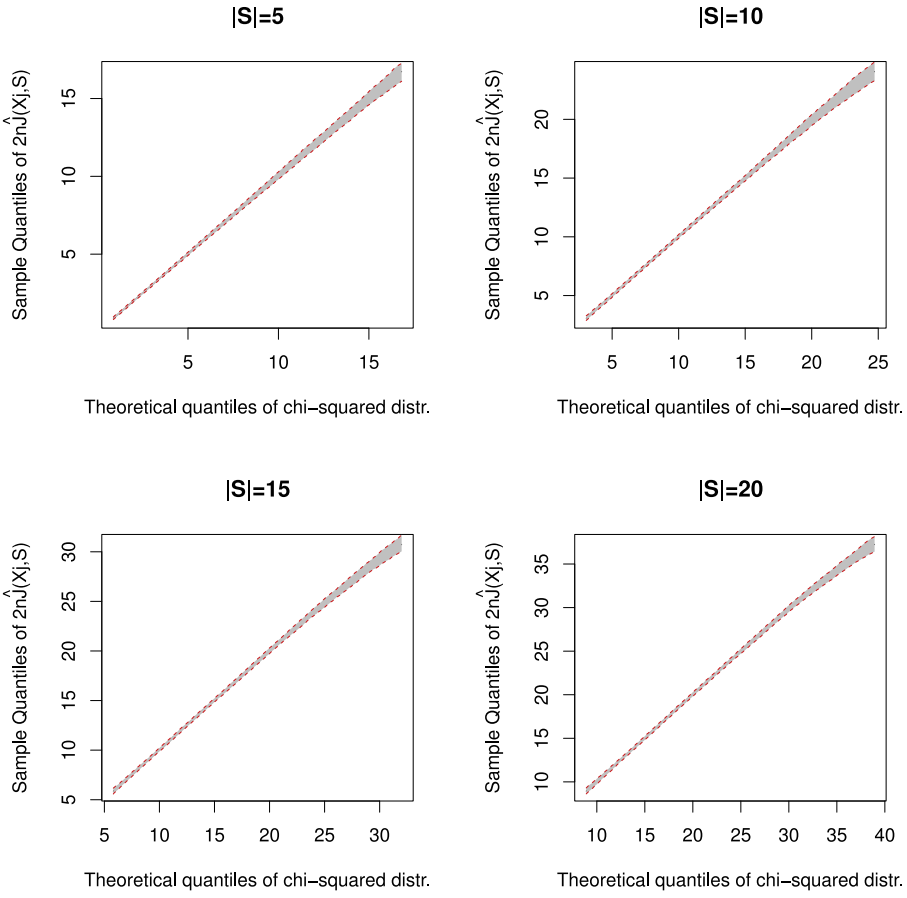

Fig. 2. Envelopes of quantile plots for simulated dataset M2 and for discretization with $|\mathcal{X}| = 2$ bins.

employs less restrictive thresholds on p-values. Note that for $p_{(j)}$ with $j = 1$ the threshold in (27) equals $\alpha/(p - |S_k|)$, that is it coincides with Bonferroni threshold. For $j > 1$ the threshold is larger and thus the condition is less restrictive. The particular form of the threshold in (27) ensures control of family wise error rate and at the same time it yields much more liberal procedure than Bonferroni procedure. We also note that the approach based on Holm's method also provides ranking of the candidates based on the currently calculated p-values (for the unsuccessful candidates at the stopping time the p-values calculated at this stage are compared in order to rank them). The batch SFS procedure (B-SFS) with stopping rule chi-Holm is described by Algorithm 2.

In order to define two remaining batch methods using the same ordering of p-values as before consider now ([8], p. 80)

$$k^* = \max\{j : p_{(j)} \leq \frac{j\alpha}{p - |S_k| + 1} =: \alpha_j\}, \tag{28}$$

$J_k = \{j_1, \ldots, j_{k^*}\}$, $S_{k+1} = S_k \cup J_k$ and
$\hat{t}_{\text{chi}-\text{BH}} = \text{argmin}_{k=1,\ldots,p} J_k = \emptyset$.

as before. This for fixed number of hypotheses is called Benjamini–Hochberg procedure. Note that now we consider downward crossings and not upward crossings and when one compares threshold curves for Holm and Benjamini–Hochberg procedures it turns out that the latter is more liberal. Finally, we define stopping time $\hat{t}_{\text{chi}-\text{BY}}$. It is based on Benjamini–Yakuteli proposal and it is defined analogously but with a lower threshold $\tilde{\alpha}_j = \alpha_j / c_{p-|S_k|}$, where $c_k = \sum_{i=1}^{k} 1/i$. This method controls False Discovery Rate (FDR) for a fixed number of hypotheses. Thus the proposed procedures use multiple testing approaches at each step of selection employing proposed reference distribution to calculate p-values. Note

Algorithm 2: Batch sequential forward search (B-SFS) with stopping rule chi-Holm.

Input : Training data of size n containing features $X_1, \ldots, X_p$ and class variable Y.

$S_0 = \emptyset$
for $k = 0, 1, \ldots$ **do**
  Calculate p-values of statistic $2n\hat{I}(X_j, S_k)$ for $j \in S_k^c$
  Order p-values $p_{(1)} \leq p_{(2)} \leq .. \leq p_{(p-|S_k|)}$($X_{j_1}, \ldots X_{j_{p-|S_k|}}$ are features corresponding to ordered p-values)
  $k^* = \min\{j : p_{(j)} > \frac{\alpha}{p-|S_k|-j+1}\}$
  $J_k = \{j_1, \ldots, j_{k^*-1}\}$
  if $J_k = \emptyset$ **then**
    stop
  else
    $S_{k+1} \leftarrow S_k \cup J_k$

Output : Relevant features S_k.

that the batch methods do not safeguard against jointly selecting strongly correlated variables, which act similarly. This behaviour will be observed when analysing their performance on real data sets.

5.2. Reference methods

As the methods against which our proposals will be compared we consider three rules. The first one disregards the fact that maximal value of $2n\hat{I}(X_j, S_k)$ over all $j \in S_k^c$ is calculated and compares

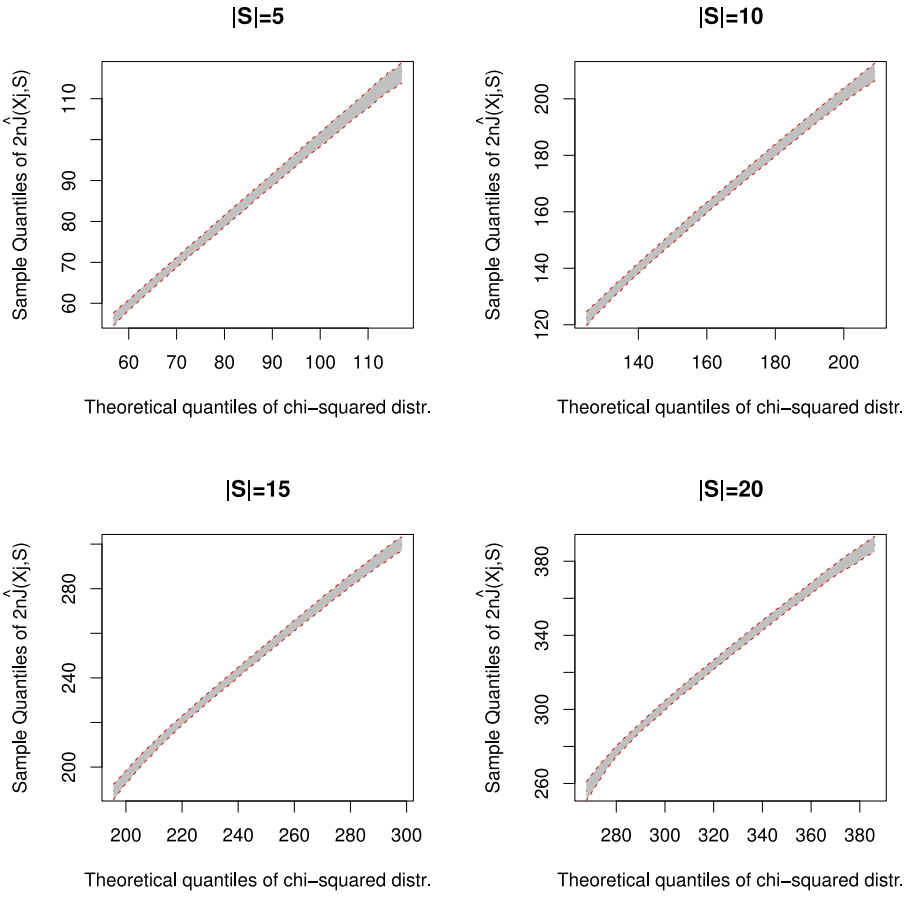

Fig. 3. Envelopes of quantile plots for simulated dataset M1 and for discretization with $|\mathcal{X}| = 5$ bins.

the maximal value with quantile of order $1-\alpha$ of $\chi^2_{d(S_k)}$ distribution. Thus it is defined as

$$\hat{t}_{chi} = \text{argmin}_{k=1,\ldots,p}\{2n\hat{J}(X_{j_k}, S_k) \leq \chi^2_{1-\alpha,d(S_k)}\}, \tag{29}$$

and $S_{k+1} = S_k \cup \{j(S_k)\}$, where j_k is defined as for Bonferroni rule. The two remaining rules correspond to AIC and BIC criteria when the number of parameters occurring in the penalty is defined as $d(S_k)$. Recall that under null hypothesis $d(S_k)$ is approximately equal the expected value of $2n\hat{J}(X_j, S_k)$. Namely let

$$\hat{t}_{AIC} = \text{argmin}_{k=1,\ldots,p}\{\max_{j\in S_k^c} 2n\hat{J}(X_j, S_k) \leq 2d(S_k)\} \tag{30}$$

$$= \text{argmin}_{k=1,\ldots,p}\{\max_{j\in S_k^c} 2n\hat{J}(X_j, S_k) - 2d(S_k) \leq 0\}. \tag{30}$$

Chosen index of the candidate at the step $k+1$ is defined as in (25) and $S_{k+1} = S_k \cup \{j_k\}$. Stopping rule $\hat{t}_{BIC}$ is defined analogously with the constant 2 in the definition of AIC criterion changed to $\log n$.

Some remarks are in order. Note that as our stopping rules use $\max_{j\in S_k^c} 2n\hat{J}(X_j, S_k)$ and we have established that distribution of $2n\hat{J}(X_j, S_k)$ for random X_j, $j \in S_k^c$ is close to $\chi^2_{d(S_k)}$, an obvious way to proceed would be to try to approximate distribution of the maximum by distribution of the maximum of corresponding chi-squares. As dependence structure of $J(X_j, S_k)$ for $j \in S_k^c$ is unknown one might check whether imposing assumption of their independence leads to stopping rule with interesting properties. However, under independence, the resulting stopping based on distribution of maximum of $p - |S_k|$ independent random variables having $\chi^2_{d(S_k)}$ distribution leads to Dunn-Šidák procedure ([8], p. 78). This in our experiments worked very similarly to chi-Bonf and thus we do not report corresponding results here.

6. Experiments

6.1. Consistency of feature selection methods

First we study empirically the consistency of the discussed methods, i.e. how precisely we can select the true relevant features (i.e. those influencing the class variable). We are interested in two issues: (1) how many (in relative sense) of the true relevant features are correctly selected as relevant and (2) how many of those irrelevant are incorrectly selected as relevant. Such study can be performed only on artificial datasets for which we exactly know which features are truly relevant, i.e. which features influence the class variable. To generate artificial data we use logistic regression model in which we can easily control the dependence between features and the class variable. Moreover, in this model we can define interaction terms in a straightforward way. We use the following notation. Let M, I be two vectors containing indices corresponding to main effect terms and interaction terms, respectively. Also we define logistic function $\sigma(s) = (1+\exp(-s))^{-1}$. The data generation scheme is as follows.

1. Generate independent features $X_1, \ldots, X_p$ from standard Gaussian distribution $N(0, 1)$.
2. Generate binary class variable $Y \in \{0, 1\}$ from Bernoulli distribution, with posterior probability

$$P(Y = 1|X_1, \ldots, X_p) = \sigma(X_M + f(X_M, X_I)),$$

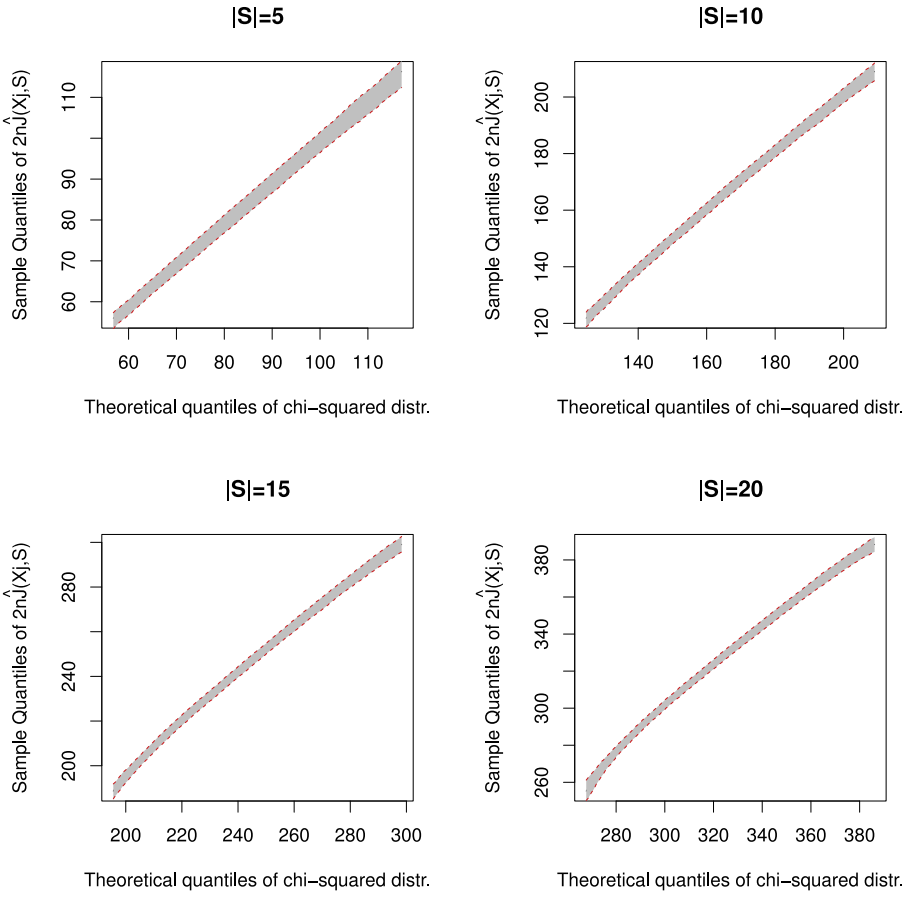

Fig. 4. Envelopes of quantile plots for simulated dataset M2 and for discretization with $|\mathcal{X}| = 5$ bins.

where f is a function which defines interactions between features corresponding to M and I.

In the above models features corresponding to set I interact with features corresponding to M. We assume that $|M| = |I| = m$. We consider the following functions.

1. Function $f_1(X_M, X_I) = \sum_{i=1}^{m} X_{M_i} X_{I_i}$. Datasets corresponding to f_1 are denoted as M1,...,M5.
2. Function $f_2(X_M, X_I) = \sum_{i=1}^{m} \max(X_{M_i}, X_{I_i})$. Datasets corresponding to f_2 are denoted as M1a,...,M5a.
3. Function $f_3(X_M, X_I) = \sum_{i=1}^{m} \min(X_{M_i}, X_{I_i})$. Datasets corresponding to f_3 are denoted as M1b,...,M5b.
4. Function $f_4(X_M, X_I) = \sum_{i=1}^{m} \mathbf{1}(X_{M_i} X_{I_i} < 0)$. Datasets corresponding to f_4 are denoted as M1c,...,M5c.
5. Function $f_5(X_M, X_I) = \sum_{i=1}^{m} \mathrm{sgn}(X_{M_i} X_{I_i})$. Datasets corresponding to f_5 are denoted as M1d,...,M5d.
6. Function $f_6(X_M, X_I) = \sum_{i=1}^{m} \mathbf{1}(X_{M_i} \geq X_{I_i})$. Datasets corresponding to f_6 are denoted as M1e,...,M5e.

For example, when $M = (1, 2)$, $I = (3, 4)$ and f_1 is chosen, the posterior probability is $P(Y = 1|X_1, \ldots, X_p) = \sigma(X_1 + X_2 + X_1X_3 + X_2X_4)$. In this case there are two main effect terms (X_1 and X_2) and two interaction terms. Features X_3, X_4 do not influence the class variable directly. Instead, they interact with features X_1 and X_2 in influencing Y. Since the results for different choices of f are similar, we only present the results for f_1 and f_2 in the paper, whereas the results for remaining functions are presented in Supplement. We denote by $T = \{M \cup I\}$ a set of all true relevant features, i.e. features affecting Y. In the above example $T = \{1, 2, 3, 4\}$. In experiments we consider five simulated datasets, described by Table 1. The number of relevant features varies from 2 to 30, e.g.

Table 1
Artificial datasets.

Dataset	Main effects	Interactions	Relevant features
M1, M1a	$M = (1)$	$I = (2)$	$T = \{1, 2\}$, $\lvert T\rvert = 2$
M2, M2a	$M = (1, 2)$	$I = (3, 4)$	$T = \{1, 2, 3, 4\}$, $\lvert T\rvert = 4$
M3, M3a	$M = (1, \ldots, 4)$	$I = (5, \ldots, 8)$	$T = \{1, \ldots, 8\}$, $\lvert T\rvert = 8$
M4, M4a	$M = (1, \ldots, 6)$	$I = (7, \ldots, 12)$	$T = \{1, \ldots, 12\}$, $\lvert T\rvert = 12$
M5, M5a	$M = (1, \ldots, 15)$	$I = (16, \ldots, 30)$	$T = \{1, \ldots, 30\}$, $\lvert T\rvert = 30$

in dataset M5 we allow for 15 main effects and 15 interactions. In all datasets, the number of main effect terms is equal to the number of interaction terms. In the case of artificial datasets, set T is known a priori (obviously this is not the case for real datasets). Thus, the quality of the considered feature selection methods can be assessed by comparing T with $\hat{T}$ being a set of features selected by the given method. As evaluation measures we use Positive Selection Rate (PSR) defined as

$$PSR(T, \hat{T}) = \frac{|T \cap \hat{T}|}{|T|},$$

which measures a fraction of correctly chosen relevant features with respect to all relevant features. Note that $PSR(T, \hat{T}) = 1$ indicates that all relevant features were selected. We also consider False Discovery Rate (FDR)

$$FDR(T, \hat{T}) = \frac{|\hat{T} \setminus T|}{|\hat{T}|},$$

which measures how many irrelevant features were selected with respect to all selected features. Similarly $FDR(T, \hat{T}) = 0$ indicates

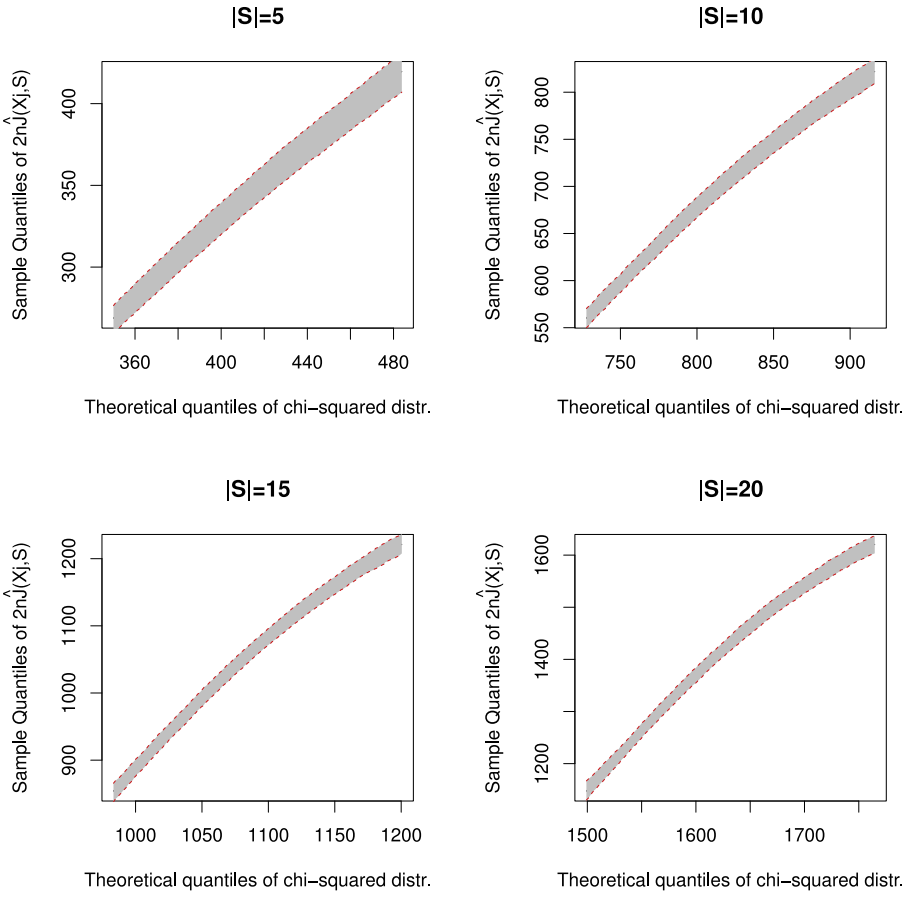

Fig. 5. Envelopes of quantile plots for simulated dataset M1 and for discretization with $|\mathcal{X}| = 10$ bins.

that no irrelevant features are included in the chosen set. Before running feature selection methods we discretize features $X_1, \ldots, X_p$ into $b = 2$ equal width bins partitioning the range of each feature.

Figs. 8 and 9 show how PSR and FDR depend on sample size n for fixed number of features $p = 100$, for datasets M1-M5. When the number of true relevant features is small (e.g. for dataset M1) then it is relatively easy to identify them correctly for sufficiently large sample size. Indeed, in the case of M1, for all considered methods, PSR ≈ 1, for $n \geq 500$. On the other hand, PSR is significantly smaller in the case of M5. The first proposed method chi-Bonf works very well in the case of datasets with small number of true relevant features (M1-M3). Its PSR is close to one and at the same time FDR is close to zero, for sufficiently large n. Its performance slightly deteriorates for datasets M4-M5 having many relevant features; in these cases the method selects too few relevant features. The other proposed methods (chi-Holm, chi-BH and chi-BY) are advantageous in the case of datasets with larger number of relevant features. Consider for example model M5. It is seen that for $n = 2000$, the methods proposed in Section 5 have FDR close to zero. At the same time chi-Holm, chi-BH and chi-BY have significantly larger PSR than chi-Bonf. Note that chi-BH has usually slightly larger FDR than chi-Bonf, chi-Holm and chi-BY. It is seen that both chi and AIC methods select too many irrelevant features, which results in significantly larger FDR than for other methods. The BIC method has significantly smaller PSR than other competitors for datasets M3-M5. The analogous results for datasets M1a-M5a are shown in Figs. 10 and 11. Conclusions for M1a-M5a are similar to those for M1-M5.

We also investigate how the choice of parameter α in Algorithm 1 affects PSR and FDR. As pointed out in Section 5.1, for larger α it is more likely to recognize canditate feature as relevant. This results in larger PSR, but at the same time larger FDR. Fig. 12 shows how PSR and FDR depend on α for chi-Bonf and $n = 500$. Both PSR and FDR grow with α. Note also that PSR depends strongly on the dataset.

6.2. Classification performance

In the case of real datasets the consistency of the feature selection methods cannot be assessed as the indices of the true

Table 2
Summary statistics of real datasets.

	n	p	p/n	classes
glass	214	9	0.04	6
segment	2310	19	0.01	7
wdbc	569	31	0.05	2
credit-a	690	38	0.06	2
spambase	4601	57	0.01	2
sonar	208	60	0.29	2
diabetes	768	8	0.01	2
heart-c	303	19	0.06	2
vote	435	32	0.07	2
waveform-5000	5000	40	0.01	3
Adult	32561	57	0.00	2
vehicle	846	18	0.02	4
ionosphere	351	34	0.10	2
credit-g	1000	48	0.05	2
Leukemia	72	3571	49.60	2
prostate	102	6033	59.15	2
madelon	2600	500	0.19	2

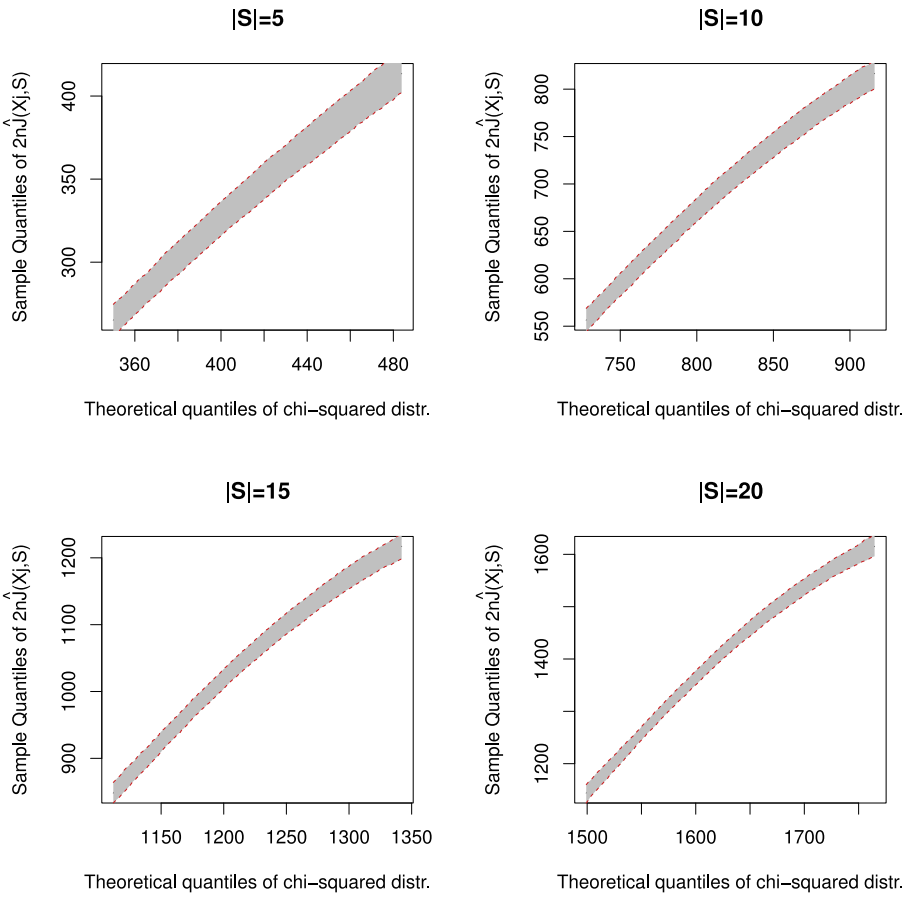

Fig. 6. Envelopes of quantile plots for simulated dataset M2 and for discretization with $|\mathcal{X}| = 10$ bins.

relevant features (i.e. those influencing the class variable) are not known. Instead we can assess the classification performance, e.g. the accuracy of the classifier built on the features selected by the given method. We consider 17 datasets from UCI machine learning repository [5] in addition to artificial data sets discussed before. These datasets are chosen to represent various characteristics. Most of them were already used in related studies on MI-based feature selection, see e.g. [3]. Table 2 shows basic statistics of the datasets: number of observations n, features p, ratio p/n and number of classes. The number of observations n varies from 72 to 32561, whereas number of features p ranges from 8 to 6033. The difficulty of the feature selection task can be measured using ratio p/n. The larger the ratio, the more challenging is the task. For the considered datasets the ratio ranges from 0.0018 to 59.1471. Most of the classification problems are binary (for 13 datasets) and 4 are multi-class problems. The quantitative features are discretized into 2 bins, whereas the discrete features are left intact. To make a feature selection task more challenging, for each dataset we add noisy features which are obtained by permuting the values of the original features. So for each dataset we have twice as many features as for original dataset. Obviously the noisy features are irrelevant in predicting the class variable. Classification performance is also assessed for artificial datasets described in Section 6.1. To estimate the classification accuracy we perform the following steps. First we split data into training set and validation set. Feature selection methods are launched on training set. Then we build a classifier on training set using selected features and finally we calculate the accuracy on validation set. The above steps are repeated for 50 random data splits. As a classifier we use a simple nearest neighbour classifier ($k = 10$) which avoids making any assumptions about the

Table 3
Accuracy (averaged over 50 data splits) for artificial datasets M1–M5. The winner method and the methods which are not significantly different from the winner (at a significance level 0.05) are in bold. The last row is a percentage of datasets for which the given method is a winner. The last two columns contain p-values of ANOVA F test for comparison between means 1–8 (the last but one column) and 1–7 (the last column).

Dataset	(1) chi2 (Bonf)	(2) AIC	(3) BIC	(4) chi2	(5) chi2 (Holm)	(6) chi2 (BH)	(7) chi2 (BY)	(8) full	pv (1–8)	pv (1–7)
M1	**0.644**	0.607	**0.648**	0.606	**0.643**	**0.642**	**0.642**	0.541	0.000	0.000
M2	**0.698**	0.674	**0.700**	0.665	**0.702**	**0.702**	**0.700**	0.554	0.000	0.000
M3	**0.694**	**0.696**	**0.673**	**0.683**	**0.702**	**0.708**	**0.686**	0.568	0.000	0.099
M4	0.649	0.663	0.617	0.659	0.663	**0.672**	0.656	0.568	0.000	0.000
M5	0.574	**0.605**	0.561	**0.608**	0.597	**0.609**	0.595	0.583	0.000	0.000
Winner	60%	40%	60%	40%	60%	100%	60%	0%		

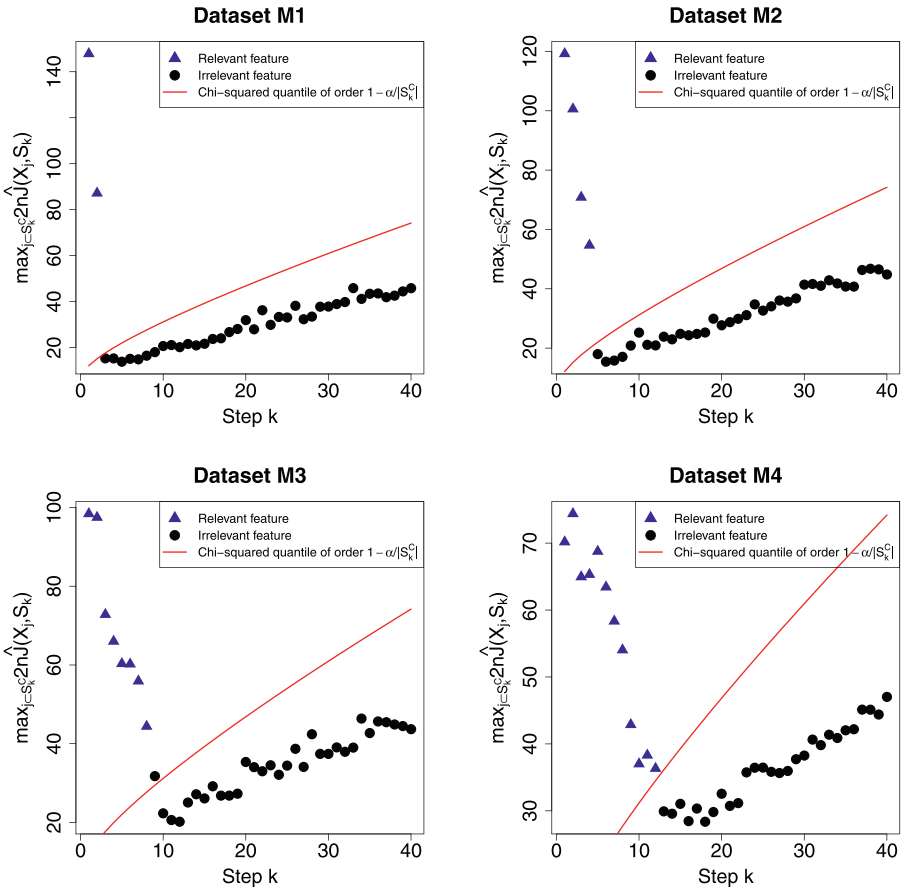

Fig. 7. Maximal improvement of the criterion function $\max_{j\in S_k^c} 2n\hat{J}(X_j, S_k)$ with respect to $k = 1, 2, \ldots$. Red line corresponds to the threshold based on quantile of chi-squared distribution with Bonferroni correction. Features above the red line are selected as relevant whereas features below the red line are recognized are irrelevant. Triangles are true relevant features, dots are true irrelevant features.

data as well as the need for parameter tuning. For this reason NN classifier was used by several authors to compare the classification performance of MI-based feature selection methods [3].

Table 3 shows accuracy (averaged over 50 data splits) for artificial datasets M1–M5. Column (8) contains the results for the classifier based on all features. The last row is a percentage of datasets for which the given method is a winner. The last two columns contain p-values of ANOVA F test for comparison between means of the methods 1–8 (the penultimate column) and methods 1–7 (the last column). In the case of significant dependence (p-value smaller than 0.05) we performed post-hoc tests (we used Tukey test, see e.g. [9]) to check whether the differences between the two particular methods were significant. The winner method and the methods which are not significantly different from the winner (at a significance level 0.05) are listed in bold. Observe that the proposed method chi-BH is a winner for all considered datasets. Methods chi and AIC work worse than the proposed methods which is due to their large FDR. Note also that the classifier based on all available features has significantly smaller accuracy than the winner method. This obviously underlines the need of constructing appropriate stopping rules in such cases.

Table 4 shows number of selected features (averaged over 50 data splits) for artificial datasets M1-M5. The last row is an averaged (over all datasets) fraction of selected features among all possible features. Note that BIC and chi-Bonf select few features, on average 3.2% and 4.2% of all available features, respectively. On the

Table 4
Number of selected features (averaged over 50 data splits) for artificial datasets M1–M5. The winner method (selecting the smallest number of features) and the methods which are not significantly different from the winner (at a significance level 0.05) are in bold. The last row is an averaged (over all datasets) fraction of selected features among all possible features. The last two columns contain p-values of ANOVA F test for comparison between means 1–8 (the last but one column) and 1–7 (the last column).

Dataset	(1) chi2 (Bonf)	(2) AIC	(3) BIC	(4) chi2	(5) chi2 (Holm)	(6) chi2 (BH)	(7) chi2 (BY)	(8) full	pv (1–8)	pv (1–7)
M1	**1.76**	7.82	1.94	8.72	1.78	1.80	**1.76**	100.00	0.00	0.00
M2	**3.46**	8.76	3.54	10.68	3.54	3.70	3.48	100.00	0.00	0.00
M3	5.44	9.68	**4.04**	12.78	5.82	6.32	5.30	100.00	0.00	0.00
M4	5.76	11.70	**3.70**	15.64	6.70	7.88	6.24	100.00	0.00	0.00
M5	4.80	12.14	**2.84**	18.10	7.16	9.20	7.12	100.00	0.00	0.00
% features	4.2%	10.0%	3.2%	13.2%	5.0%	5.8%	4.8%	100.0%		

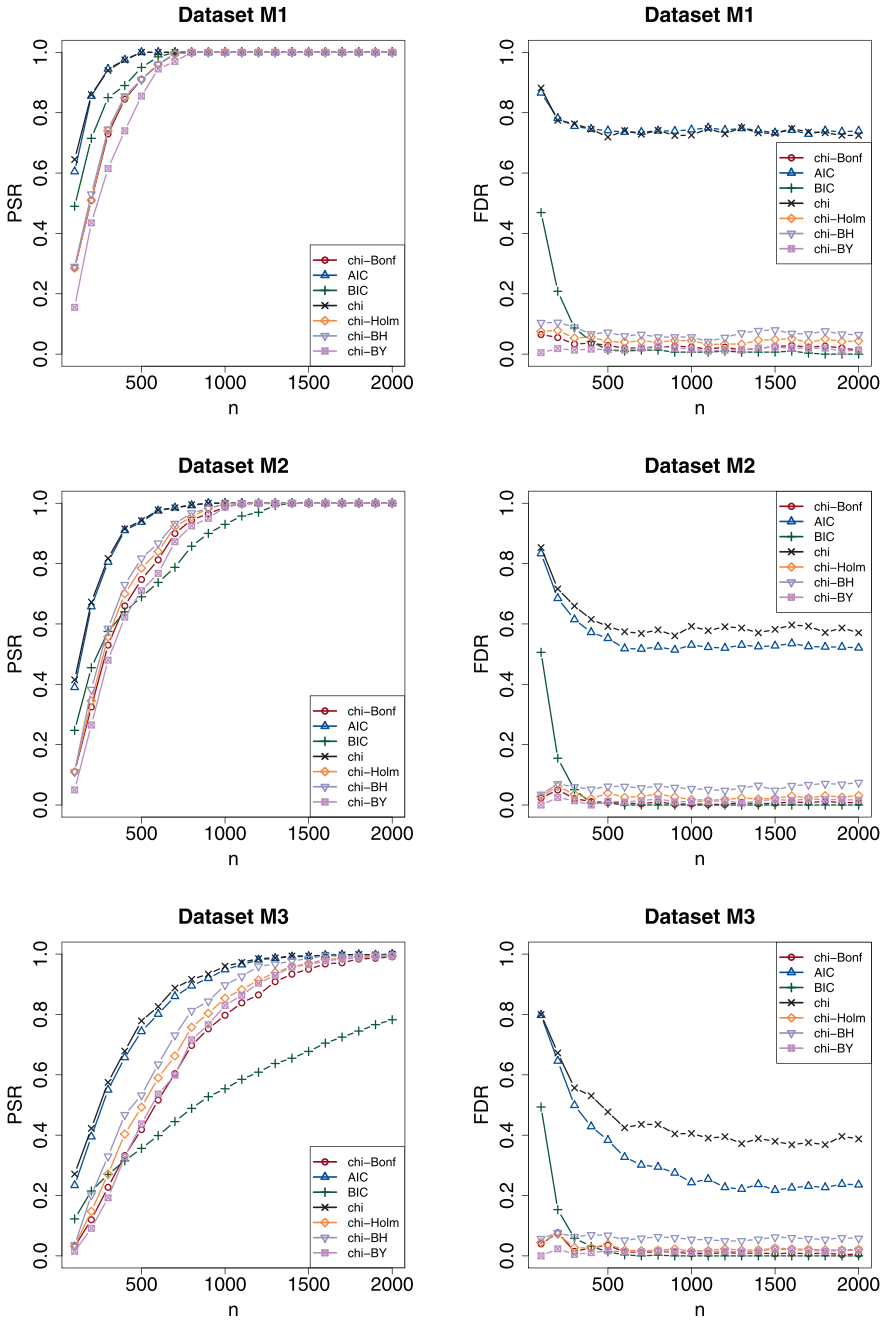

Fig. 8. PSR and FDR for simulation models M1–M3.

Table 5
Accuracy (averaged over 50 data splits) for artificial datasets M1a–M5a. The winner method and the methods which are not significantly different from the winner (at a significance level 0.05) are in bold. The last row is a percentage of datasets for which the given method is a winner. The last two columns contain *p*-values of ANOVA F test for comparison between means 1–8 (the last but one column) and 1–7 (the last column).

Dataset	(1) chi2 (Bonf)	(2) AIC	(3) BIC	(4) chi2	(5) chi2 (Holm)	(6) chi2 (BH)	(7) chi2 (BY)	(8) full	pv (1–8)	pv (1–7)
M1a	**0.705**	0.678	**0.705**	0.673	**0.704**	0.703	**0.704**	0.595	0.000	0.000
M2a	**0.768**	0.745	**0.767**	0.738	**0.769**	**0.770**	**0.764**	0.665	0.000	0.001
M3a	**0.798**	**0.798**	**0.794**	**0.790**	**0.804**	**0.805**	**0.803**	0.731	0.000	0.553
M4a	**0.814**	**0.823**	**0.807**	**0.815**	**0.816**	**0.826**	**0.816**	0.773	0.000	0.442
M5a	0.877	**0.889**	0.884	**0.890**	0.880	**0.884**	0.871	**0.890**	0.000	0.001
Winner	80%	60%	80%	60%	80%	80%	80%	20%		

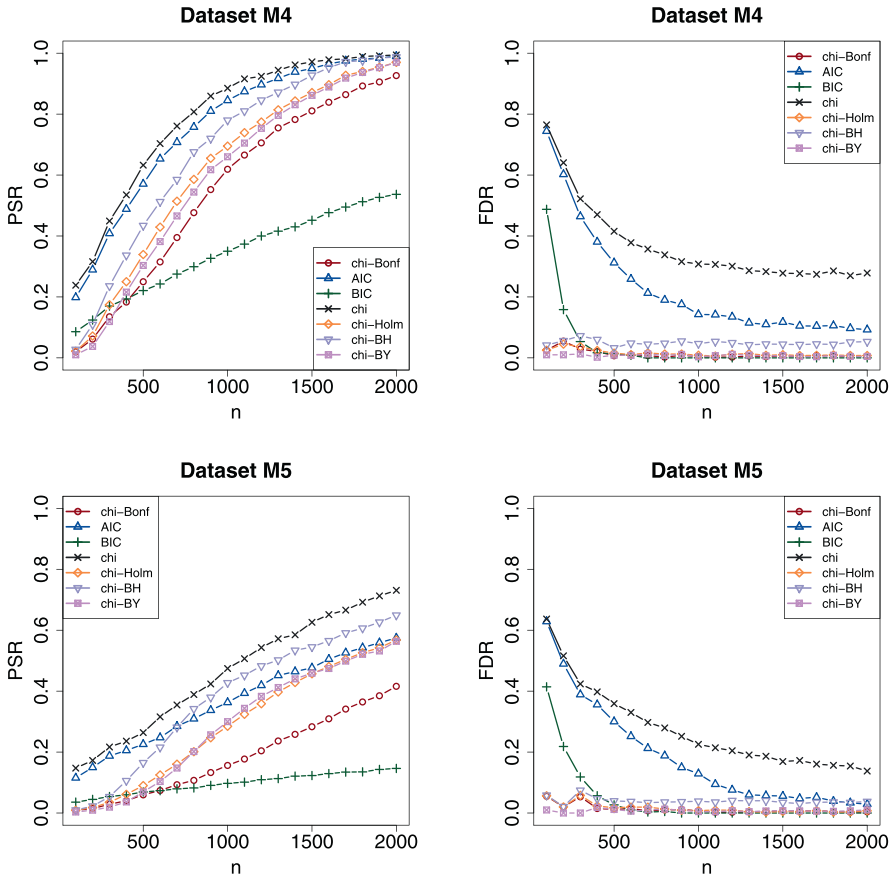

Fig. 9. PSR and FDR for simulation models M4–M5.

other hand AIC and chi select much more features, 10% and 13.2%, respectively. The chi-BH method, which is a winner w.r.t. accuracy, chooses moderate number of features, namely 5.8% of all possible features.

Tables 5 and 6 show accuracy and number of selected features for datasets M1a–M5a. The conclusions are similar as for M1-M5. First, chi and AIC work worse than other methods which is due to their large FDR. Secondly, the classifier based on all available features has usually significantly smaller accuracy than the winner method. Regarding number of features, BIC selects the least features (on average 3.3%), whereas chi selects the largest number of features (on average 14%).

Table 7 shows accuracy (averaged over 50 data splits) for real datasets. The proposed methods achieve the highest accuracy; chi-Holm and chi-BY are the winners for around 75% of datasets, whereas chi-BH is a winner for 65% of datasets. The classifier based on all available features has significantly smaller accuracy than the winner method in most cases. Table 8 shows number of selected features (averaged over 50 data splits) for real datasets. The chi-Bonf method selects the smallest number of features, on average 6% of all possible features, whereas chi-BH selects the largest number of features.

In order to gain a deeper insight into performance of the methods we visually inspect the stopping points for selected datasets. Figs. 13 and 14 show how accuracy on validation set depends on the number of selected features for selected datasets and for one particular data split. For better clarity we only analyse the proposed methods: chi-Bonf, chi-Holm and chi-BH. It follows

Table 6
Number of selected features (averaged over 50 data splits) for artificial datasets M1a–M5a. The winner method (selecting the smallest number of features) and the methods which are not significantly different from the winner (at a significance level 0.05) are in bold. The last row is an averaged (over all datasets) fraction of selected features among all possible features. The last two columns contain p-values of ANOVA F test for comparison between means 1–8 (the last but one column) and 1–7 (the last column).

Dataset	(1) chi2 (Bonf)	(2) AIC	(3) BIC	(4) chi2	(5) chi2 (Holm)	(6) chi2 (BH)	(7) chi2 (BY)	(8) full	pv (1–8)	pv (1–7)
M1a	1.70	8.70	1.84	10.74	1.74	1.90	**1.58**	100.00	0.00	0.00
M2a	3.24	8.72	**3.14**	11.48	3.28	3.48	**3.12**	100.00	0.00	0.00
M3a	5.22	10.24	**4.08**	13.64	5.96	6.80	5.84	100.00	0.00	0.00
M4a	6.18	11.62	**4.84**	16.60	7.00	8.76	7.08	100.00	0.00	0.00
M5a	4.74	11.96	**2.80**	17.40	6.86	10.42	7.16	100.00	0.00	0.00
% features	4.2%	10.2%	3.3%	14.0%	5.0%	6.3%	5.0%	100.0%		

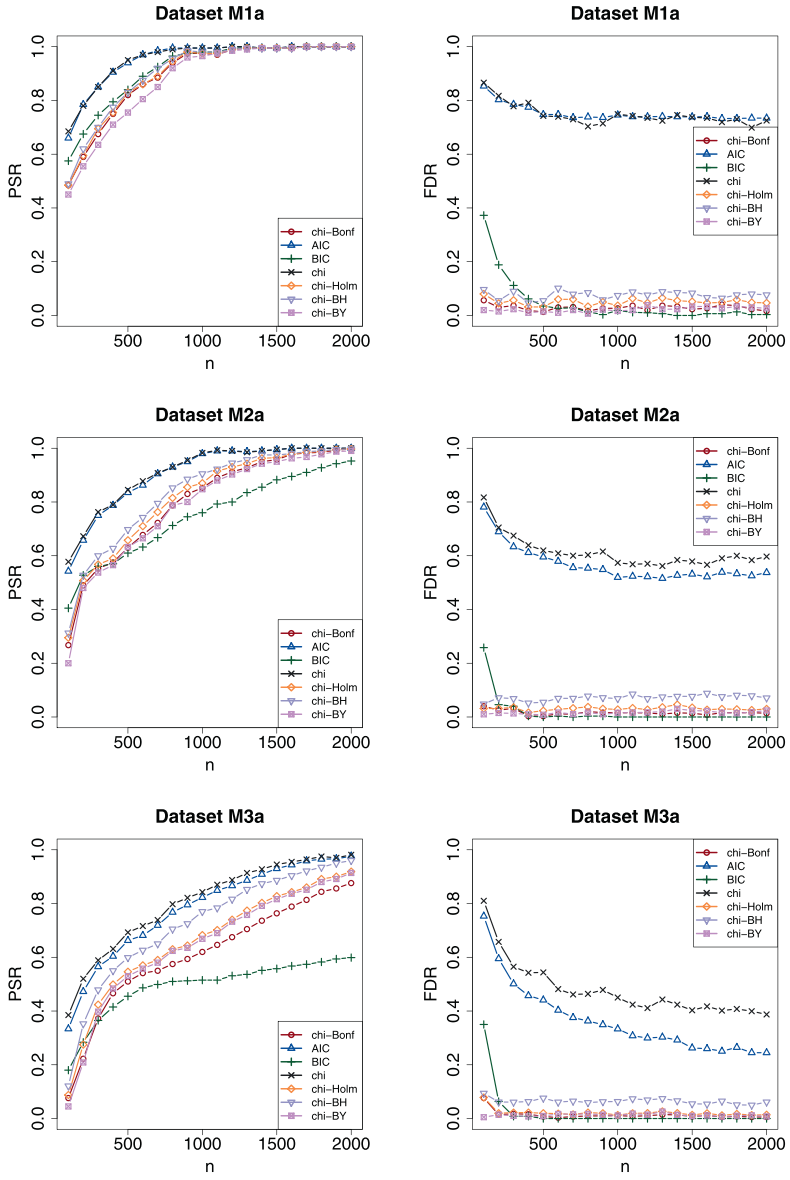

Fig. 10. PSR and FDR for simulation models M1a–M3a.

from Figs. 13 and 14 that all three methods work as expected for datasets M1–M4 and `madelon` i.e. they roughly find a point corresponding to maximal accuracy. In the case of M5, `spambase` and `credit-g`, the chi-Bonf method stops too early. It is also seen that chi-BH usually selects more features than the remaining two methods. Note that if there is a clear maximum of the curve, all the methods work similarly (see e.g. the curves for datasets M1 and M2 on Fig. 13). On the other hand, if the maximum is not pronounced the stopping points indicated by the methods differ (see e.g. the curve for `spambase` on Fig. 14).

7. Conclusions

In this paper we discussed a problem of finding optimal stopping rule for mutual information-based sequential forward selection methods. Such a rule allows to separate true relevant features from irrelevant features which is crucial in many domains. The proposed methods are based on the distribution of approximation of the conditional mutual information given that all relevant features have been already selected. We show that the distribution is approximately chi square with appropriate number of degrees of freedom. The choice of the reference distribution is justified theoretically (Theorem 1) and empirically. It turns out that for coarse discretization the agreement between empirical distribution of the approximation of conditional mutual information and reference chi-square distribution is very high. For finer discretization the approximation deteriorates. To validate the quality of the methods we performed experiments on both artificial and real datasets. The main conclusions from the experiments are as follows. First, the proposed batch methods chi-Holm, chi-BH and chi-BY have desirable consistency properties, i.e. they achieve large PSR and at the same time their FDR is close to zero. In this respect they are clearly superior to other methods. Secondly, the proposed batch methods chi-Holm, chi-BH and chi-BY have larger accuracy than other competitors for both artificial and real datasets. The limitation of batch methods is simultaneous inclusion of groups of correlated features which may result in large number of redundant features among selected ones for some real datasets. The chi-Bonf method usually

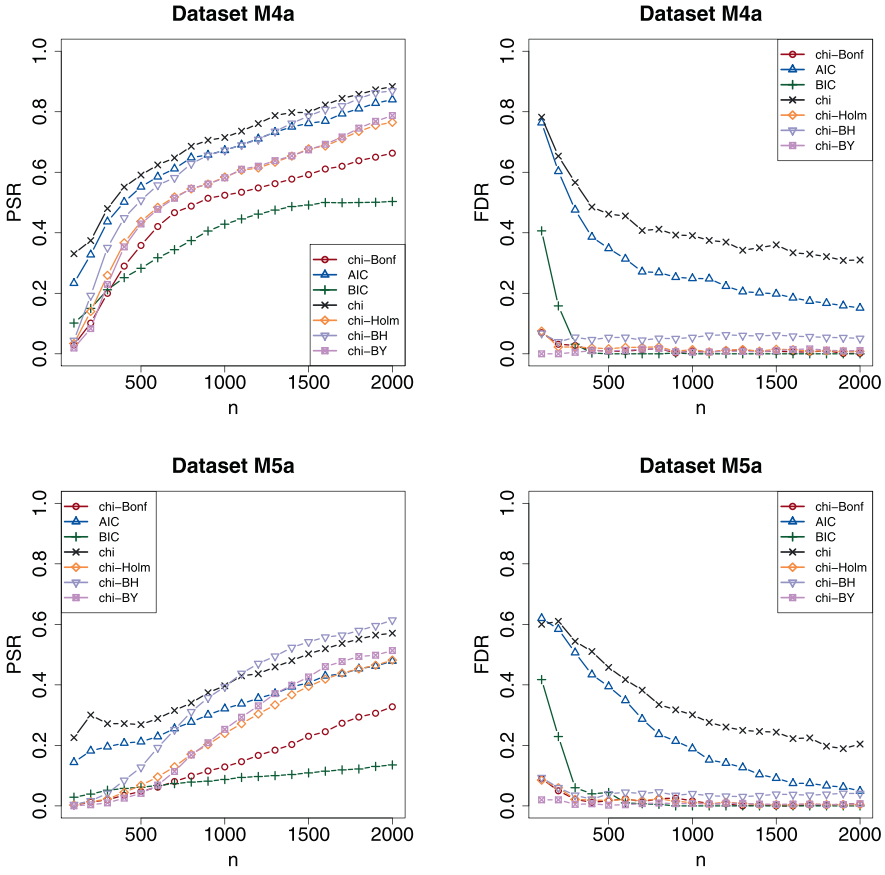

Fig. 11. PSR and FDR for simulation models M4a–M5a.

Table 7
Accuracy (averaged over 50 data splits) for real datasets. The winner method and the methods which are not significantly different from the winner (at a significance level 0.05) are in bold. The last row is a percentage of datasets for which the given method is a winner. The last two columns contain *p*-values of ANOVA F test for comparison between means 1–8 (the last but one column) and 1–7 (the last column).

Dataset	(1) chi2 (Bonf)	(2) AIC	(3) BIC	(4) chi2	(5) chi2 (Holm)	(6) chi2 (BH)	(7) chi2 (BY)	(8) full	pv (1–8)	pv (1–7)
glass	0.451	0.480	0.409	0.497	**0.525**	**0.521**	**0.519**	0.432	0.000	0.000
segment	0.806	0.807	0.829	0.801	**0.849**	0.840	**0.852**	0.672	0.000	0.017
wdbc	0.858	0.802	0.854	0.817	**0.901**	**0.888**	**0.897**	0.602	0.000	0.000
credit-a	**0.828**	0.779	**0.827**	0.778	0.804	0.767	0.805	0.634	0.000	0.000
spambase	0.699	0.766	0.683	0.765	0.769	**0.794**	0.770	0.687	0.000	0.000
sonar	0.590	**0.612**	**0.633**	**0.625**	0.602	0.608	0.564	0.610	0.000	0.000
diabetes	**0.703**	**0.727**	**0.716**	**0.725**	**0.713**	**0.716**	**0.711**	0.683	0.000	0.238
heart-c	**0.706**	**0.671**	**0.710**	**0.685**	**0.688**	**0.673**	**0.680**	0.582	0.000	0.146
vote	**0.918**	**0.904**	**0.920**	**0.909**	**0.901**	**0.897**	**0.899**	**0.891**	0.624	0.717
waveform	0.764	0.762	0.760	0.755	**0.835**	0.834	**0.835**	0.783	0.000	0.000
Adult	0.825	0.827	0.809	0.821	**0.839**	0.834	**0.838**	0.757	0.000	0.000
vehicle	**0.564**	**0.555**	0.544	0.488	0.532	0.530	0.533	0.425	0.000	0.004
ionosphere	**0.803**	**0.783**	**0.813**	**0.787**	**0.785**	**0.776**	**0.773**	0.723	0.000	0.295
credit-g	**0.655**	**0.673**	**0.659**	**0.672**	**0.660**	**0.664**	**0.662**	**0.665**	0.137	0.090
Leukemia	**0.713**	**0.743**	**0.762**	**0.734**	**0.771**	**0.803**	**0.757**	**0.805**	0.706	0.794
prostate	**0.729**	**0.683**	**0.720**	**0.694**	**0.674**	**0.679**	**0.653**	**0.626**	0.462	0.756
madelon	**0.749**	**0.792**	**0.788**	**0.755**	**0.751**	**0.756**	**0.750**	0.620	0.000	0.809
Winner	58.8%	58.8%	58.8%	52.9%	76.5%	64.7%	76.5%	23.5%		

stops too early, i.e. it selects too few features, whereas AIC and chi methods usually select too many irrelevant features. Finally, BIC method has significantly smaller PSR than other considered methods, i.e. it selects too few features. In view of the above findings we may recommend chi-Holm and chi-BH methods as the most promising ones among the methods proposed.

Let us also discuss some directions for future research. In this work we proposed a stopping rule for one particular criterion

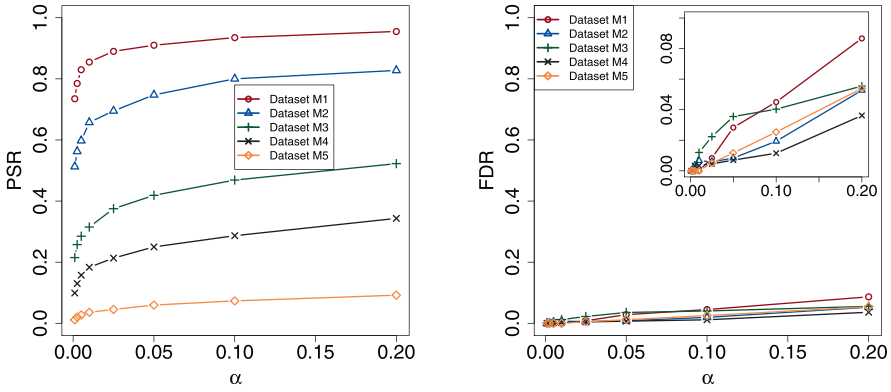

Fig. 12. PSR and FDR with respect to to α for chi-Bonf method and $n = 500$.

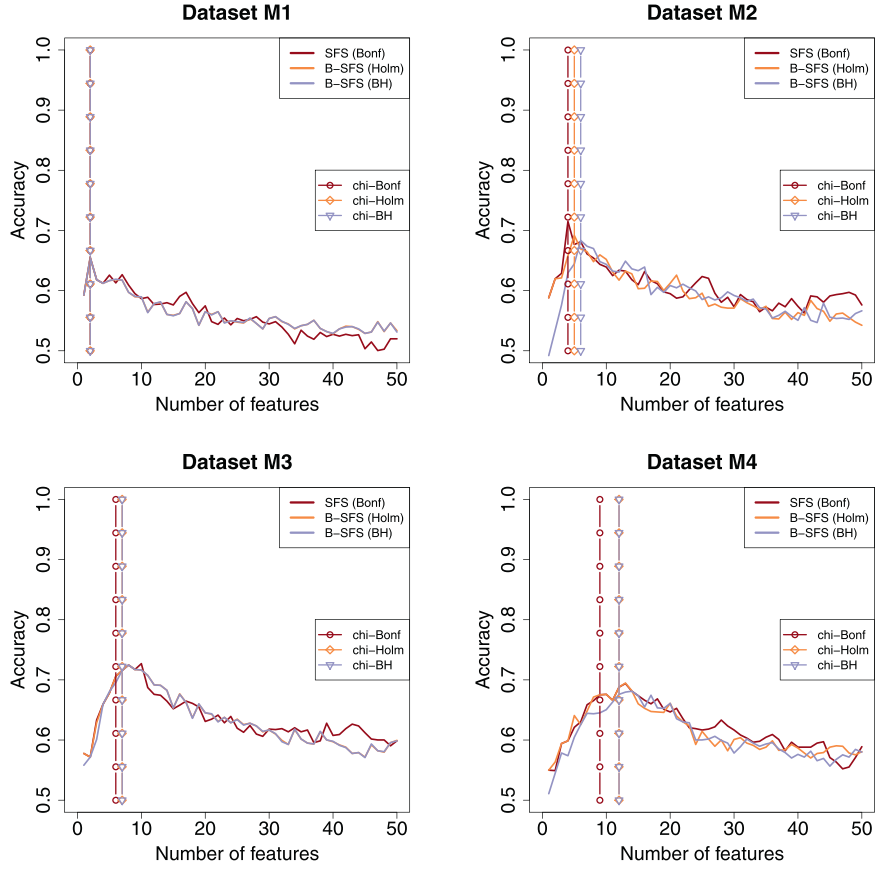

Fig. 13. Accuracy with respect to the number of features.

(known as CIFE criterion), being a second order approximation of the conditional mutual information. We focused on this criterion as it seems to be very natural approximation, following from the Möbius representation (9). Future work should include designing stopping rules for other methods. Note however that it is rather impossible to propose a universal rule that would be valid for all MI-based methods. This is due to the fact the the distribution of the score function given that all relevant features have been already selected may be quite different for different methods. Despite this we believe that a general methodology used in our approach can be applied to other methods. For example consider JMI method [38] in which the score for the candidate X_j is $J(X_j, S) = |S|I(X_j, Y) + \sum_{i\in S}[I(X_j, X_i|Y) - I(X_j, X_i)]$. The number of degrees of freedom d can be calculated in analogous way as for CIFE using our Theorem 1. Then the quantile of the chi square distribution with d degrees of freedom could be used to define stopping rule for SFS with JMI criterion. We can proceed similarly for other popular methods, e.g. MIFS or MRMR (these two however do not take into account interactions between features as CIFE or JMI). Using stopping rules for other MI-based methods requires separate experiments and is left for future research.

The second important issue for future research is choosing more accurate reference distribution for finer discretization. The proposed chi square distribution is close to the empirical distribution for up to 5 bins but the approximation deteriorates for 10 and more bins. Moreover it would be desirable to improve batch methods in order to prevent adding the groups of strongly correlated features simultaneously in a single step of the greedy

Table 8
Number of selected features (averaged over 50 data splits) for real datasets. The winner method (selecting the smallest number of features) and the methods which are not significantly different from the winner (at a significance level 0.05) are in bold. The last row is an averaged (over all datasets) fraction of selected features among all possible features. The last two columns contain *p*-values of ANOVA F test for comparison between means 1–8 (the last but one column) and 1–7 (the last column).

Dataset	(1) chi2 (Bonf)	(2) AIC	(3) BIC	(4) chi2	(5) chi2 (Holm)	(6) chi2 (BH)	(7) chi2 (BY)	(8) full	pv (1–8)	pv (1–7)
glass	1.44	1.86	**1.00**	2.24	3.28	3.68	3.22	18.00	0.00	0.00
segment	5.52	5.26	**3.80**	6.10	14.12	15.86	14.64	38.00	0.00	0.00
wdbc	**1.98**	6.54	2.16	5.84	17.44	21.42	18.26	62.00	0.00	0.00
credit-a	**1.74**	7.82	2.16	7.90	5.36	8.74	5.64	76.00	0.00	0.00
spambase	2.84	5.86	**2.30**	5.78	7.90	16.34	9.08	114.00	0.00	0.00
sonar	4.38	25.00	4.84	49.62	6.10	9.52	**3.22**	120.00	0.00	0.00
diabetes	**1.76**	3.90	1.84	3.02	2.68	3.38	2.58	16.00	0.00	0.00
heart-c	**1.94**	5.88	2.38	5.58	7.82	10.54	8.16	38.00	0.00	0.00
vote	**1.96**	6.40	2.16	6.16	29.82	34.72	30.72	64.00	0.00	0.00
waveform	5.40	6.82	**4.86**	10.56	20.24	21.50	19.94	80.00	0.00	0.00
Adult	20.72	23.14	**11.00**	30.38	41.64	46.68	42.82	114.00	0.00	0.00
vehicle	6.92	7.50	**2.52**	11.02	13.88	15.80	13.98	36.00	0.00	0.00
ionosphere	**1.86**	18.64	2.28	15.12	19.94	24.54	19.24	68.00	0.00	0.00
credit-g	**1.58**	9.66	2.10	14.68	3.50	6.42	3.20	96.00	0.00	0.00
Leukemia	**0.55**	25.75	14.55	25.75	45.85	129.80	65.95	7142.00	0.00	0.00
prostate	**0.60**	70.35	43.70	50.80	554.20	757.65	403.05	12066.00	0.00	0.00
madelon	19.05	33.25	**13.30**	100.00	24.05	37.70	24.55	1000.00	0.00	0.00
% features	6.1%	12.5%	4.7%	15.1%	19.3%	23.5%	19.4%	100.0%		

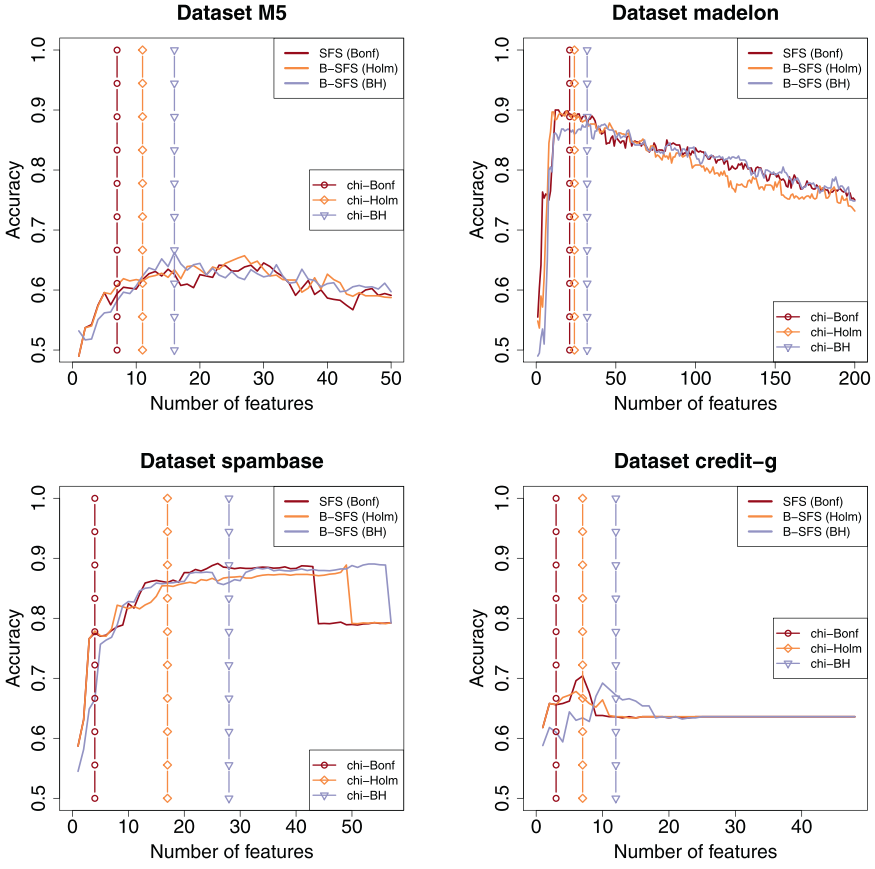

Fig. 14. Accuracy with respect to the number of features.

procedure. Finally it would be useful to propose stopping rules for more complex classification problems, e.g. multi-label classification, for which various MI-based methods have been developed [15–17] but stopping rules are missing.

Conflict of interest

None.

Acknowledgements

Remarks of two referees which led to substantial improvement of the exposition of the paper are gratefully acknowledged.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neucom.2019.05.048.

References

[1] A. Agresti, Categorical Data Analysis, Wiley, 2002.
[2] R. Battiti, Using mutual information for selecting features in supervised neural-net learning, IEEE Trans. Neural Netw. 5 (4) (1994) 537–550.
[3] G. Brown, A. Pocock, M.J. Zhao, M. Luján, Conditional likelihood maximisation: a unifying framework for information theoretic feature selection, J. Mach. Learn. Res. 13 (1) (2012) 27–66.
[4] T.M. Cover, J.A. Thomas, Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing), Wiley-Interscience, 2006.
[5] D. Dheeru, E. Karra Taniskidou, UCI machine learning repository, 2017, URL http://archive.ics.uci.edu/ml.
[6] M. Draminski, J. Koronacki, rmcfs: An R package for monte carlo feature selection and interdependency discovery, J. Stat. Softw. 85 (12) (2018) 1–28.
[7] S. Dudoit, M.J. van der Laan, Multiple Testing Procedures with Applications to Genomics, Springer, 2009.
[8] S. Dudoit, J. Shaffer, J. Boldrick, Multiple hypothesis testing in microarray experiments, Stat. Sci. 18 (2003) 71–103.
[9] J. Faraway, Linear Models with R, Chapman, 2014.
[10] R. Fisher, On the interpretation of chi square from contingency tables and calculation of *p*, J. R. Stat. Soc. 85 (1922) 87–94.
[11] F. Fleuret, Fast binary feature selection with conditional mutual information, J. Mach. Learn. Res. 5 (2004) 1531–1555.
[12] T.S. Han, Multiple mutual informations and multiple interactions in frequency data, Inf. Control 46 (1) (1980) 26–45.
[13] T. Hastie, R. Tibshirani, M. Wainwright, Statistical Learning with Sparsity: the Lasso and Generalizations, Springer, 2015. URL https://web.stanford.edu/~hastie/StatLearnSparsity/
[14] A. Jakulin, I. Bratko, Quantifying and visualizing attribute interactions: an approach based on entropy, 2004. URL https://arxiv.org/pdf/cs/0308002.pdf.
[15] J. Lee, D.-W. Kim, Feature selection for multi-label classification using multivariate mutual information, Pattern Recognit. Lett. 34 (3) (2013) 349–357.
[16] J. Lee, D.-W. Kim, Memetic feature selection algorithm for multi-label classification, Inf. Sci. 293 (2015a) 80–96.
[17] J. Lee, D.W. Kim, Mutual information-based multi-label feature selection using interaction information, Expert Syst. Appl. 42 (4) (2015b) 2013–2025.
[18] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, Feature selection: a data perspective, J. Mach. Learn. Res. (2016) 1–73.
[19] D. Lin, X. Tang, Conditional infomax learning: an integrated framework for feature extraction and fusion, in: Proceedings of the Ninth European Conference on Computer Vision–Volume Part I, in: ECCV, 2006, pp. 68–82.
[20] F. Macedo, M. Rosário Oliveira, A. Pacheco, R. Valadas, Theoretical foundations of forward feature selection methods based on mutual information, Neurocomputing 325 (2019) 67–89.
[21] W.J. McGill, Multivariate information transmission, Psychometrika 19 (2) (1954) 97–116.
[22] P. Meyer, C. Schretter, G. Bontempi, Information-theoretic feature selection in microarray data using variable complementarity, IEEE J. Sel. Top. Signal Process. 2 (3) (2008) 261–274.
[23] J. Mielniczuk, M. Rdzanowski, Use of information measures and their approximations to detect predictive gene-gene interaction, Entropy 19 (2017) 1–23.
[24] J. Mielniczuk, P. Teisseyre, A deeper look at two concepts of measuring gene–gene interactions: logistic regression and interaction information revisited, Genet. Epidemiol. 42 (2) (2018) 187–200.
[25] L. Paninski, Estimation of entropy and mutual information, Neural Comput. 15 (6) (2003) 1191–1253.
[26] C. Pascoal, M. Rosário Oliveira, A. Pacheco, R. Valadas, Theoretical evaluation of feature selection methods based on mutual information, Neurocomputing 226 (2017) 168–181.
[27] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[28] P. Peng, L. M., D. Aston, Likelihood ratio tests with three-way tables, J. Am. Stat. Assoc. 105 (490) (2010) 740–749.
[29] L. Schiatti, L. Faes, J. Tessadori, G. Barresi, L. Mattos, Mutual information-based feature selection for low-cost BCIs based on motor imagery, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2016, pp. 2772–2775. 2016-October
[30] R.J. Serfling, Approximation Theorems of Mathematical Statistics, Wiley, New York, 1980.
[31] J. Shao, Mathematical Statistics, Springer, New York, 2003.
[32] A. Shishkin, A. Bezzubtseva, A. Drutsa, Efficient high-order interaction-aware feature selection based on conditional mutual information, in: Proceedings of the Advances in Neural Information Processing Systems, in: NIPS, 2016, pp. 1–9.
[33] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. B 58 (1996) 267–288.
[34] J.R. Vergara, P.A. Estévez, A review of feature selection methods based on mutual information, Neural Comput. Appl. 24 (1) (2014) 175–186.
[35] N.X. Vinh, S. Zhou, J. Chan, J. Bailey, Can high-order dependencies improve mutual information based feature selection? Pattern Recognit. 53 (2016) 46–58.
[36] X. Wan, C. Yang, Q. Yang, T. Xue, X. Fan, N. Tang, W. Yu, Boost: a fast approach to detecting gene-gene interactions in genome-wide case-control studies, Am. J. Hum. Genet. 87 (3) (2010) 325–340.
[37] M.H. Wang, H.J. Cordell, K. Van Steen, Statistical Methods for Genome-wide Association Studies, Semin. Cancer Biol. 55 (2018) 53–60.
[38] H.H. Yang, J. Moody, Data visualization and feature selection: new algorithms for nongaussian data, Adv. Neural Inf. Process. Syst. 12 (1999) 687–693.
[39] R.W. Yeung, A First Course in Information Theory, Kluwer, 2002.

Jan Mielniczuk is full professor and deputy director for research at the Institute of Computer Science, Polish Academy of Science and professor at the Faculty of Mathematics and Information Science of Warsaw University of Technology. He received Ph.D. (1985) degree from the Warsaw University and habilitation (1996) degree from the Institute of Mathematics, Polish Academy of Sciences. In 2009 he received a title of Professor. His main research contributions concern computational statistics and data mining, in particular time series modelling and prediction, inference for high dimensional and misspecified data, model selection, computer intensive methods, asymptotic analysis and quantification of dependence. He is an author and coauthor of two books and over seventy articles. He is elected member of Committee of Mathematics of Polish Academy of Sciences.

Paweł Teisseyre received the Ph.D. degree from Institute of Computer Science, Polish Academy of Sciences, in 2013. He is currently an assistant professor at the institute. He also cooperates with Agency for Health Technology Assessment and Tariff System, Warsaw, Poland, in the field of medical data analysis. His research interests include feature selection in high-dimensional classification and regression problems, multi-label learning and medical data analysis.