Information-Theoretic Feature Selection Using High-Order Interactions

Mateusz Pawluk[1], Paweł Teisseyre[2], and Jan Mielniczuk[1,2]

[1] Warsaw University of Technology, Faculty of Mathematics and Information Science, Warsaw, Poland
m.pawluk@mini.pw.edu.pl
[2] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
teisseyrep@ipipan.waw.pl miel@ipipan.waw.pl

Abstract. Feature selection is one of the major challenges in machine learning. In this paper, we focus on mutual information based methods, which attracted a significant attention in recent years. A clear limitation of the most existing methods is that they usually take into account only low-order interactions between features (up to 3rd order). We propose a novel criterion which takes into account both 3-way and 4-way interactions and can be possibly extended to the case of higher order terms. The basic component of our criterion is interaction information which is a measure of interaction strength derived from information theory. We show that our method is able to find interactions which remain undetected when using standard methods. We prove some theoretical properties of the introduced criterion and interaction information.

1 Introduction

Feature selection is one of the major problems in machine learning [1]. Feature selection is a crucial step for several reasons. First it improves the understandability of the considered model and allows to discover the relationship between features and the class (target) variable. Secondly, it helps to build models with better generalization and larger predictive power [2]. Finally, it allows to reduce the computational cost of fitting the model.

In this paper, we focus on mutual information (MI) based feature selection. This approach has several important advantages. First MI, unlike some classical measures (e.g. Pearson correlation), is able to capture both linear and non-linear dependencies among random variables. Secondly MI based criteria do not depend on any particular model which allows to find all features associated with the class variable, not only those which can be captured by an employed model. This is particularly important in the domains where feature selection itself is the main goal of the analysis, e.g. in human genetics where finding mutations of genes influencing the disease is a crucial problem. Moreover, some advanced MI based criteria are able to discover interactions between features as well as to take redundancy between features into account. Finally information-theoretic approach can be used for both classification and regression tasks, i.e. nominal

and quantitative class variable as well as for any type of the features. In this work we focus on classification problem, but the method can be easily extended to regression.

In recent years many algorithms based on mutual information have been proposed. A clear limitation of the existing methods is that they usually take into account only low-order interactions (up to 3rd order). This can be a serious drawback when some complex dependencies exist in our data. For example recent studies in genetics indicate that high-order interactions between genes may contribute to many complex traits [3] and it is crucial to identify them in order to efficiently predict the trait. Taylor et. al [3] give two examples of high-order interactions: one example of three-locus interactions that influence body weight in a cross of two chicken lines and another that showed a pair of genetic interactions involving five or more loci that determine colony morphology in a cross of two yeast strains. We propose a novel criterion called Interaction Information Feature Selection (IIFS) that takes into account both 3-way and 4-way interactions and can be possibly extended to the case of higher order terms. The basic component of our method is interaction information, which is a non-parametric measure of interaction strength derived from information theory. Our method is a generalization of Conditional Infomax Feature Extraction (CIFE) criterion [4] whose limitation is that it only considers 3-way interaction terms. We show that our method is able to find interactions which remain undetected when using standard methods. We also prove some theoretical properties of 4-way interaction information and of the novel criterion. Moreover we experiment with two different methods of multivariate entropy estimation: plug-in estimator based on data discretization and knn-based Kozachenko-Leonenko estimator [5].

The paper is structured as follows. In Section 2 we recall the definition of interaction information and prove some new theoretical properties of 4-way interaction information. In Section 3 we define the problem and review the existing methods. In Section 4 we present our method and discuss its theoretical properties, Section 5 contains the results of numerical experiments.

2 Interaction Information

First we define basic quantities used in Information Theory. We consider the discrete class variable Y and features $X_1, \ldots, X_p$, which can be either continuous or discrete. For sake of simplicity we write definitions only for discrete variables. We first recall the definition of the entropy for discrete class variable:

$$H(Y) = -\sum_{y} P(Y = y) \log P(Y = y). \tag{1}$$

Entropy quantifies the uncertainty of observing random values of Y. If large mass of the distribution is concentrated on one particular value of Y then the entropy is low. If all values are equally likely then $H(Y)$ is maximal. Let $S = (X_1, \ldots, X_m)$ be a subset of the original feature set of size $m = 1, \ldots, p$. The entropy of S is defined analogously to (1), with a difference that multivariate probability is

used instead of univariate probability. The conditional entropy of S given class variable Y can be written as

$$H(S|Y) = \sum_y P(Y = y)H(S|Y = y). \tag{2}$$

The joint mutual information between S and class variable Y is

$$I(S, Y) = H(S) - H(S|Y). \tag{3}$$

This can be interpreted as the amount of uncertainty in S which is removed when Y is known which is consistent with the intuitive meaning of mutual information as the amount of information that one variable provides about another. Moreover the conditional mutual information between S and Y given variable Z is defined as

$$I(S, Y|Z) = H(S|Z) - H(S|Y, Z). \tag{4}$$

We recall a definition of m-way interaction information (II) [6, 7]

$$II(S) = II(X_1, \ldots, X_m) = -\sum_{T \subseteq S} (-1)^{|S|-|T|} H(T), \tag{5}$$

which generalizes the 3-way interaction information proposed in [8]. For $m = 2$, interaction information reduces to mutual information. The definition of interaction information is identical to that of multivariate mutual information $I(S)$ [8] except for a change in sign in the case of an odd number of variables, i.e. $II(S) = (-1)^{|S|} I(S)$. II can be understood as the amount of information common to all variables (or set of variables), but that is not present in any subset of these variables. Interestingly, m-way interaction information can be also defined using recursive formula

$$II(X_1, \ldots, X_m) = II(X_1, \ldots, X_{m-1}|X_m) - II(X_1, \ldots, X_{m-1}), \tag{6}$$

where $II(X_1, \ldots, X_{m-1}|X_m) = \sum_x P(X_m = x) II(X_1, \ldots, X_{m-1}|X_m = x)$. The next formula (also known as Möbius representation) [9–11] shows the relationship between II and joint mutual information $I(S, Y)$ which will be useful in the context of the proposed feature selection method

$$I(S, Y) = I((X_1, \ldots, X_m), Y) = \sum_{k=1}^{m} \sum_{T \subseteq S : |T|=k} II(T \cup Y). \tag{7}$$

To better grasp the concept of II, let us discuss in more detail 3-way and 4-way interactions. It follows from Möbius representation (7) that

$$II(X_1, X_2, Y) = I((X_1, X_2), Y) - I(X_1, Y) - I(X_2, Y), \tag{8}$$

which indicates that interaction information can be interpreted as a part of the mutual information of (X_1, X_2) and Y which is due solely to interaction between

X_1 and X_2 in predicting Y i.e. the part of $I((X_1, X_2), Y)$ which remains after subtraction of individual informations between Y and X_1 and Y and X_2. In other words, II is obtained by removing the main effects from the term describing the overall dependence between Y and the pair (X_1, X_2). Here let us mention that 3-way interaction information is a commonly used measure for detecting interactions between genes in genome-wide case- control studies [12, 13]. For 4-way interaction we have from (7) and (8) that

$$\begin{aligned} II(X_1, X_2, X_3, Y) = & I((X_1, X_2, X_3), Y) \\ & - I((X_1, X_2), Y) - I((X_1, X_3), Y) - I((X_2, X_3), Y) \\ & + I(X_1, Y) + I(X_2, Y) + I(X_3, Y). \end{aligned} \tag{9}$$

Observe that both terms $I((X_1, X_2), Y)$ and $I((X_1, X_3), Y)$ in (9) contain $I(X_1, Y)$ as summands (cf (8)) and as a result $I(X_1, Y)$ is subtracted twice. To account for it we add $I(X_1, Y)$ in the last line of (9). The remaining pairs are treated analogously. The simplest examples of 3-way and 4-way interactions are XOR problems. In XOR $Y = 1$ when the number of input variables taking value 1 is odd. It is easy to check that input binary variables are mutually independent and marginally independent from a class variable. For 3-dimensional case we have $I(X_1, Y) = I(X_2, Y) = 0$ and $II(X_1, X_2, Y) = I((X_1, X_2), Y) = H(Y) - H(Y|X_1, X_2) = H(Y) = \log(2)$. For 4-dimensional case all terms, except the first one, are zero. i.e. $II(X_1, X_2, X_3, Y) = I((X_1, X_2, X_3), Y) = H(Y) - H(Y|X_1, X_2, X_3) = H(Y) = \log(2)$.

Some properties of 4-way Interaction Information which has not been discussed in the literature are discussed below. For the sake of clarity we assume that all variables are discrete and let $p_{ijkl} = P(X_1 = x_i, X_2 = x_j, X_3 = x_k, Y = y_l)$, where P denotes the distribution of (X_1, X_2, X_3, Y). Moreover, $KL(P||Q)$ stands for Kullback-Leibler divergence between P and Q, defined as $KL(P||Q) = \sum_{i,j,k} p_{ijk} \log(p_{ijk}/q_{ijk})$.

Theorem 1. *We have (i) $II(X_1, X_2, X_3, Y) = KL(P||P_K)$, where P_K corresponds to mass function p^K defined as*

$$p^K_{ijkl} = \frac{\prod_{S:|S|=3} p_S \prod_{S:|S|=1} p_S}{\prod_{S:|S|=2} p_S} = \frac{p_{ijk} p_{ijl} p_{jkl} p_{ikl} p_i p_j p_k p_l}{p_{ij} p_{ik} p_{il} p_{jk} p_{jl} p_{kl}}. \tag{10}$$

(ii) If $X_1 \perp X_2 | W$, where W is any subset (including $\emptyset$) of $\{X_3, Y\}$ then $II(X_1, X_2, X_3, Y) = 0$.
(iii) Let $\eta = \sum_{i,j,k,l} p^K_{ijkl}$. If $\eta \leq 1$ and $II(X_1, X_2, X_3, Y) = 0$ then $P = P_K$.

Proof. (i) follows from (5) and definition of Kullback-Leibler divergence. (ii) is a consequence of (10) and assumptions. In order to prove (iii) note that $KL(P||Q) = 0$ implies $P = Q$ not only in the case when Q is probability distribution but also in the case when total mass of Q does not exceed 1. This yields the result when applied to $Q = P_K$.

Observe that P_K is not necessarily probability distribution. Condition $\eta \leq 1$ is sufficient condition which ensures that $P = P_K$ when $II = 0$. P_K is generalization of Kirkwood approximation [14] to four-dimensional case.

3 Problem formulation and previous work

In this work we focus on feature selection based on mutual information (MI). MI-based feature selection is concerned with identifying a fixed-size subset $S \subset \{1, \ldots, p\}$ of the original feature set that maximizes the joint mutual information between S and class variable Y. Finding an optimal feature set is usually unfeasible because the search space grows exponentially with the number of features. As a result various greedy algorithms have been developed including forward selection, backward elimination and genetic algorithms. Today sequential forward selection is the most commonly adopted solution. Forward selection algorithms start from an empty set of features and add, in each step, the feature that jointly, i.e. together with already selected features, achieves the maximum joint mutual information with the class. Formally, assume that S is a set of already chosen features, S^c is its complement and $X_k \in S^c$ is a candidate feature. The score for feature X_k is

$$J(X_k) = I(S \cup X_k, Y) - I(S, Y). \tag{11}$$

Obviously the second term in (11) does not depend on X_k and it can be omitted, however it is more convenient to use this form. In each step we add a feature that maximizes $J(X_k)$. Criterion (11) is equivalent to

$$J(X_k) = I(X_k, Y|S), \tag{12}$$

see [15] for the proof. Observe that (12) indicates that we select a feature that achieves the maximum association with the class given the already chosen features. Criterion (11) (or equivalently (12)) is appealing and attracted a significant attention. However in practice the estimation of joint mutual information is problematic even for small set S. This makes a direct application of (11) infeasible. A rich body of work in the MI-based feature selection literature approaches this difficulty by approximating the high-dimensional joint MI with low-dimensional MI terms. These approximations may by accurate provided some additional conditions on data distribution are satisfied. A comprehensive review of the existing methods can be found in [15], here we review some representative methods. One of the most popular methods is Mutual Information Feature Selection (MIFS) proposed in [16]

$$J_{\mathrm{MIFS}}(X_k) = I(X_k, Y) - \sum_{j \in S} I(X_j, X_k). \tag{13}$$

This includes the $I(X_k, Y)$ term to ensure feature relevance, but introduces a penalty to enforce low correlations with features already selected in S. The similar idea is used in Minimum-Redundancy Maximum-Relevance (MRMR) criterion [17]

$$J_{\mathrm{MRMR}}(X_k) = I(X_k, Y) - \frac{1}{|S|} \sum_{j \in S} I(X_j, X_k). \tag{14}$$

with the difference that the second term is averaged over features in S. Both MIFS and MRMR criteria focus on reducing redundancy, however they do not

take into account interactions between features. Brown et. al. [15] have shown that if the selected features from S are independent and class-conditionally independent given any unselected feature X_k then (11) reduces to so-called CIFE criterion [4]

$$J_{\mathrm{CIFE}}(X_k) = I(X_k, Y) + \sum_{j \in S} [I(X_j, X_k|Y) - I(X_j, X_k)]. \tag{15}$$

In view of (8), the second term in (15) is equal $\sum_{j \in S} II(X_j, X_k, Y)$, so it is seen that CIFE is able to detect 3-way interactions. Yang and Moody [18] have proposed using Joint Mutual Information (JMI)

$$J_{\mathrm{JMI}}(X_k) = \sum_{j \in S} I((X_j, X_k), Y), \tag{16}$$

which is equal up to a constant to

$$J_{\mathrm{JMI}}(X_k) = |S| I(X_k, Y) + \sum_{j \in S} [I(X_j, X_k|Y) - I(X_j, X_k)]. \tag{17}$$

JMI is a similar to CIFE, with the difference that in JMI the marginal relevance term plays more important role than the overall interaction term.

4 Feature selection based on interaction information

In this Section we describe a proposed approach which can be seen as a generalization of CIFE. Our method considers not only 3-way interactions but also 4-way interactions.

4.1 Proposed criterion: IIFS

In our method we make use of Möbius representation. Recall that S is a set of already selected features of size m and X_k is a candidate feature. First observe that it follows from Möbius representation (7) that

$$J(X_k) = I(S \cup X_k, Y) - I(S, Y) = \sum_{k=0}^{m} \sum_{T \subset S : |T| = k} II(T \cup X_k \ \cup Y). \tag{18}$$

In the proposed method IIFS (Interaction Information Feature Selection) we define a score

$$J_{\mathrm{IIFS}}(X_k) = I(X_k, Y) + \sum_{j \in S} II(X_j, X_k, Y) + \sum_{i,j \in S : i < j} II(X_i, X_j, X_k, Y), \tag{19}$$

which is a third order approximation of (18). The first term in (19) takes into account marginal relevance of the candidate feature whereas the second and the third terms describe the 3 and 4-way interactions, respectively. Note that IIFS

can be seen as an extended version of CIFE which is a second order approximation of $J(X_k)$, namely

$$J_{\mathrm{IIFS}}(X_k) = J_{\mathrm{CIFE}}(X_k) + \sum_{i,j\in S:i<j} II(X_i, X_j, X_k, Y). \tag{20}$$

It is possible to consider higher order terms in (18), however it would increase the computational cost and make the estimation even more difficult. Below we state some properties of the introduced criteria.

Theorem 2. *The following properties hold.*
(i) Assume that $X_k \perp Y$. Then

$$J_{CIFE}(X_k) = \sum_{j\in S} I(X_k, Y|X_j). \tag{21}$$

(ii) Assume that $X_k \perp Y$ and $X_k \perp Y|X_j$ for any $X_j \in S$. Then

$$J_{IIFS}(X_k) = \sum_{i,j\in S:i<j} I(X_k, Y|X_i, X_j). \tag{22}$$

(iii) Assume that $X_i \perp X_j|X_k$ and $X_i \perp X_j|X_k, Y$, for some $X_i, X_j \in S$. Then $II(X_i, X_j, X_k, Y)$ does not depend on X_k.
(iv) If $|S| = 2$ then $\operatorname{argmax}_{X_k\in S^c} J_{IIFS}(X_k) = \operatorname{argmax}_{X_k\in S^c} J(X_k)$.

Proof. To prove (i) observe that property (6) implies

$$II(X_j, X_k, Y) = I(X_k, Y|X_j) - I(X_k, Y). \tag{23}$$

Under assumption $X_k \perp Y$ we have $I(X_k, Y) = 0$ which, together with (23) and (15) yields (21). Let us now prove (ii). It follows from (6) that

$$II(X_i, X_j, X_k, Y) = II(X_j, X_k, Y|X_i) - II(X_j, X_k, Y) \tag{24}$$

and

$$II(X_j, X_k, Y|X_i) = I(X_k, Y|X_j, X_i) - I(X_k, Y|X_i). \tag{25}$$

Under assumption (ii) we have that $I(X_k, Y) = 0$, $II(X_j, X_k, Y) = 0$ and $I(X_k, Y|X_i) = 0$ and thus $II(X_i, X_j, X_k, Y) = I(X_k, Y|X_j, X_i)$ which yields (22). Let us now prove (iii). Using (6) we can write

$$\begin{aligned} &II(X_i, X_j, X_k, Y) = II(X_i, X_j, Y|X_k) - II(X_i, X_j, Y) \\ &= I(X_i, X_j|X_k, Y) - I(X_i, X_j|X_k) - II(X_i, X_j, Y). \end{aligned} \tag{26}$$

Assumptions of (iii) implies that $I(X_i, X_j|X_k, Y) = I(X_i, X_j|X_k) = 0$, which yields the assertion in view of (26). Finally note that (iv) follows from the fact that for $|S| = 2$ equations (18) and (19) are equivalent. i.e. Möbius representation gives an exact value of $J(X_k)$.

Let us briefly comment the above statements. Items (i) and (ii) of Theorem 2 indicate that under additional assumptions CIFE and IIFS reduce to simpler and more intuitive forms. Using the forms given in (i) and (ii) one may easily give an example showing the advantage of IIFS over CIFE. Indeed, under assumption (ii) we have $J_{\mathrm{CIFE}}(X_k) = 0$ and we may conclude that $J_{\mathrm{IIFS}}(X_k) > 0$ if there exists a pair $X_i, X_j \in S$ such that $I(X_k, Y|X_i, X_j) > 0$. In this case IIFS recognizes X_k as a relevant whereas CIFE treats X_k as a spurious feature. In addition [15] has showed that if assumptions of (iii) hold for any $k \in S^c$, maximization of $J_{\mathrm{CIFE}}(X_k)$ is equivalent to maximization of $J(X_k)$. In (iii) we confirm that indeed in this case the 4-way interaction term can be omitted.

5 Experiments

The aim of the experiments is to compare the performance of the proposed method IIFS with other popular methods discussed in Section 3: MIFS, MRMR, JMI and CIFE.

5.1 Artificial data

The main advantage of the experiments on artificial data is that we can directly investigate which method is able to detect the particular types of interactions. We consider two simulation models, including 3-way and 4-way interactions, respectively. To make a task more challenging we assume in both cases that features are continuous. To assess the quality of the methods we introduce the following measure. Let t be a set of relevant features influencing Y and $j_1, j_2, \ldots, j_p$ be features sequentially selected by the given method. The selection rate (SR) is defined as

$$SR = \frac{|\{j_1, \ldots, j_{|t|}\} \cap t|}{|t|}, \tag{27}$$

i.e. SR is a fraction of relevant features among first $|t|$ selected. For example if we have two relevant features X_1, X_2 then $t = \{1, 2\}$. When the method produces a list $\{1, 2, 5, \ldots\}$ then $SR = 1$. On the other hand if the method gives $\{1, 5, 2, \ldots\}$ then $SR = 0.5$, as one spurious feature X_5 is ranked higher than the relevant feature X_2. In the following we describe two simulation models.

Simulation model 1 (3-way interaction model). We consider 50 uniformly distributed features: $X_1 \sim U[0, 3]$, $X_j \sim U[0, 2]$, for $j = 2, \ldots, 50$. Only two first features X_1 and X_2 are relevant, i.e. class variable Y depends only on X_1 and X_2, the remaining features are spurious. Table 1 shows the joint distribution of X_1, X_2, Y. This model is an extension of 2-dimensional XOR; note that $Y = 1$ when $X_1 \in A, X_2 \in B$ or $X_1 \in B, X_2 \in A$. It is easy to verify that for this model we have: $I(X_1, Y) > 0$, $I(X_j, Y) = 0$, for $j = 2, \ldots, 50$ and $II(X_1, X_2, Y) > 0$, thus we have one main effect corresponding to X_1 and one 3-way interaction.

Simulation model 2 (4-way interaction model). We consider 50 uniformly distributed features: $X_1, X_2 \sim U[0, 3]$, $X_j \sim U[0, 2]$, for $j = 3, \ldots, 50$. Class

variable Y depends on X_1, X_2, X_3 whereas the remaining features are spurious. Table 2 shows the joint distribution of X_1, X_2, Y. This model is an extension of 3-dimensional XOR. It is easy to verify that for this model we have: $I(X_1, Y), I(X_2, Y) > 0$, $I(X_j, Y) = 0$, for $j = 3, \ldots, 50$ and $II(X_1, X_2, X_3, Y) > 0$, thus we have two main effects corresponding to X_1 and X_2 and moreover one 4-way interaction.

Table 1: Simulation model 1 (3-way interaction model). Notation: $A = [0, 1]$, $B = (1, 2]$, $C = (2, 3]$.

	1	2	3	4	5	6
X_1	A	A	B	B	C	C
X_2	A	B	A	B	A	B
Y	0	1	1	0	0	0
$P(X_1, X_2, Y)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Table 2: Simulation model 2 (4-way interaction model). Notation: $A = [0, 1]$, $B = (1, 2]$, $C = (2, 3]$.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X_1	A	B	A	A	B	B	A	B	C	C	C	C	A	A	B	B
X_2	A	A	B	A	B	A	B	B	A	B	A	B	C	C	C	C
X_3	A	A	A	B	A	B	B	B	A	A	B	B	A	B	A	B
Y	0	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1
$P(X_1, X_2, X_3, Y)$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$

Table 3: Computational times.

Feature Selection	CIFE	JMI	MIFS	MRMR	IIFS
MADELON	16.312 secs	16.228 secs	16.089 secs	16.147 secs	1.109 mins
GISETTE	1.156 hours	1.153 hours	1.091 hours	1.124 hours	2.701 hours
MUSK	11.719 secs	11.048 secs	13.587 secs	14.217 secs	15.746 secs
BREAST	0.887 secs	0.425 secs	0.515 secs	0.499 secs	0.988 secs

Figure 1 shows how selection rate (SR) depends on sample size n. In the case of model 1 the methods which take into account 3-way interactions (JMI, CIFE, IIFS) produce the same rankings. They detect successfully both relevant features: X_1 and X_2. MIFS and MRMR are able to detect only one relevant feature. In the case of model 2, MIFS, MRMR, JMI and CIFE are able to detect only 2 relevant features X_1, X_2 but they fail to select feature X_3. Selection rate (SR) for MIFS, MRMR, JMI and CIFE converges to 2/3. As expected only IIFS chooses all 3 relevant features, which results in $SR = 1$ for sufficiently large sample size. The above experiment shows that there is no significant difference

between IIFS, JMI, CIFE when only 3-way interactions occur. In the case of 4-way interaction model, IIFS is significantly superior to other methods. Moreover we analyse how the method of entropy estimation influences the results. We used two methods: standard plug-in method based on data discretization with b bins (solid line) and knn-based Kozachenko-Leonenko estimator [5], with $k = 10$ (dashed line). For small $b = 2$ it is seen that knn-based method is superior to plug-in method. For $b = 5$, plug-in method works better than knn-based method in the case of model 1, whereas knn-based method is a winner for model 2.

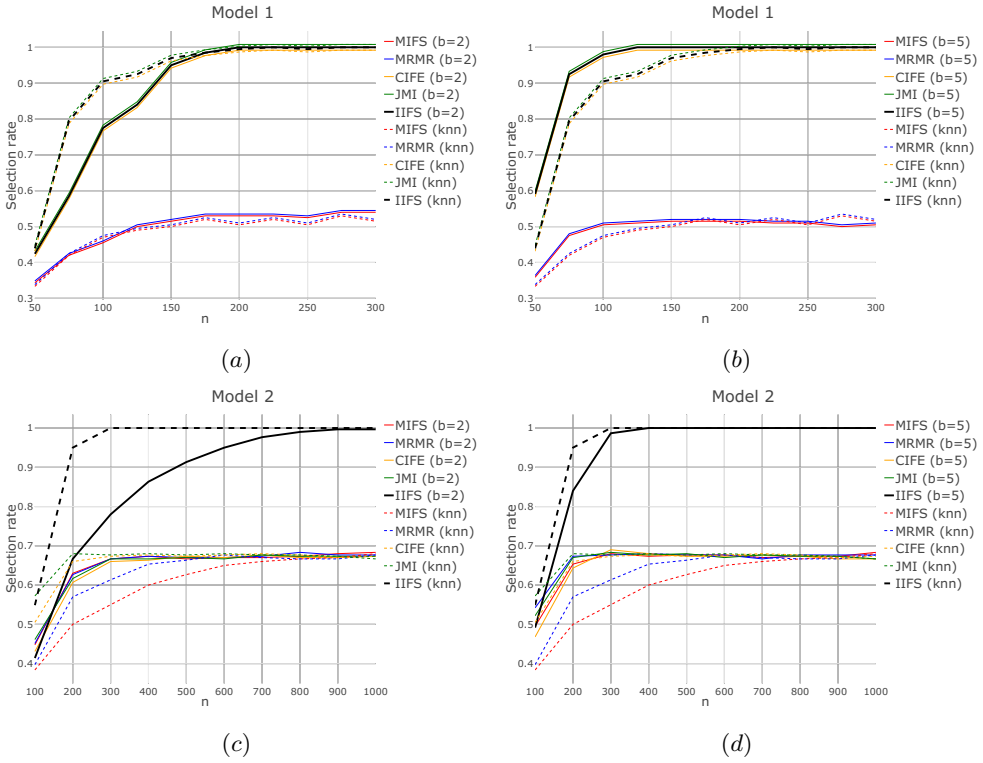

Fig. 1: Selection rate w.r.t. sample size n for simulation models 1 (a)-(b) and 2 (c)-(d). Parameter b corresponds to the number of bins in discretization, 'knn' in brackets corresponds to knn-based entropy estimation.

5.2 Benchmark data

For more thorough assessment of developed criterion we used datasets from the NIPS Feature Selection Challenge [19] (MADELON and GISETTE) and UCI

repository [20] (BREAST and MUSK). NIPS datasets consist of training sets (2000 observations for MADELON and 6000 for GISETTE) and validation sets (600 observations for MADELON and 1000 for GISETTE), whereas for UCI datasets we used 10-fold cross-validation in order to calculate error rates. We carried out the same experiment as that described in [15], Section 6.1. In addition to methods considered in [15] we investigate the performance of the proposed method IIFS. Each criterion was used to generate a ranking for the top features. Then the original datasets were used to classify the validation data. As in [15] we used kNN method with $k = 3$ neighbours as a classifier. As an evaluation measure we considered Balanced Error Rate defined as

$$BER = 1 - 0.5 \cdot (\frac{TP}{TP + FN} + \frac{TN}{TN + FP}), \tag{28}$$

where TP, TN, FP, FN denote true positives, true negatives, false positives and false negatives, respectively. Results of our experiments are presented in Figure 2. We only present curves corresponding to plug-in estimator as knn-based entropy estimator worked much worse in this case possibly due to a prior discretization of the original data. For MADELON and MUSK datasets there is no significant improvement of IIFS compared to CIFE and JMI. So we may conclude that considering interactions of order higher than 3 does not improve the performance in this case. Note that for MADELON interactions play an important role; the methods which do not take into account interactions at all (MIFS and MRMR) fail. For GISETTE dataset the proposed criterion IIFS has the lowest error rate when the number of features varies between 20 and 100. For BREAST IIFS is also a winner. This suggests that taking into account high-order interactions helps in these cases. Interestingly, for GISETTE and BREAST, IIFS is significantly better than CIFE, which additionally indicates that including 4-way interaction term improves the performance. The computational times for IIFS are longer than for competitors (see Table 3) which is a price for taking into account high-order interactions. Note however that the times for IIFS, although longer than for CIFE, are of the same order.

6 Conclusions

In this paper we presented a novel feature selection method, named IIFS. Feature selection score in IIFS, based on interaction information, is derived from so-called Möbius representation of joint mutual information. Our method in an extension of CIFE criterion consisting in taking into account 4-way interaction terms. We discussed theoretical properties of 4-way interaction information (Theorem 1) as well as feature selection methods: CIFE and IIFS (Theorem 2). The numerical experiments show that there is no significant difference between IIFS, JMI and CIFE when only the interactions of order up to 3 are present. This means that estimation of absent 4-way interactions does not cause significant deterioration of IIFS performance. In the case of 4-way interactions IIFS is significantly superior to other methods. Future work will include the development of methods

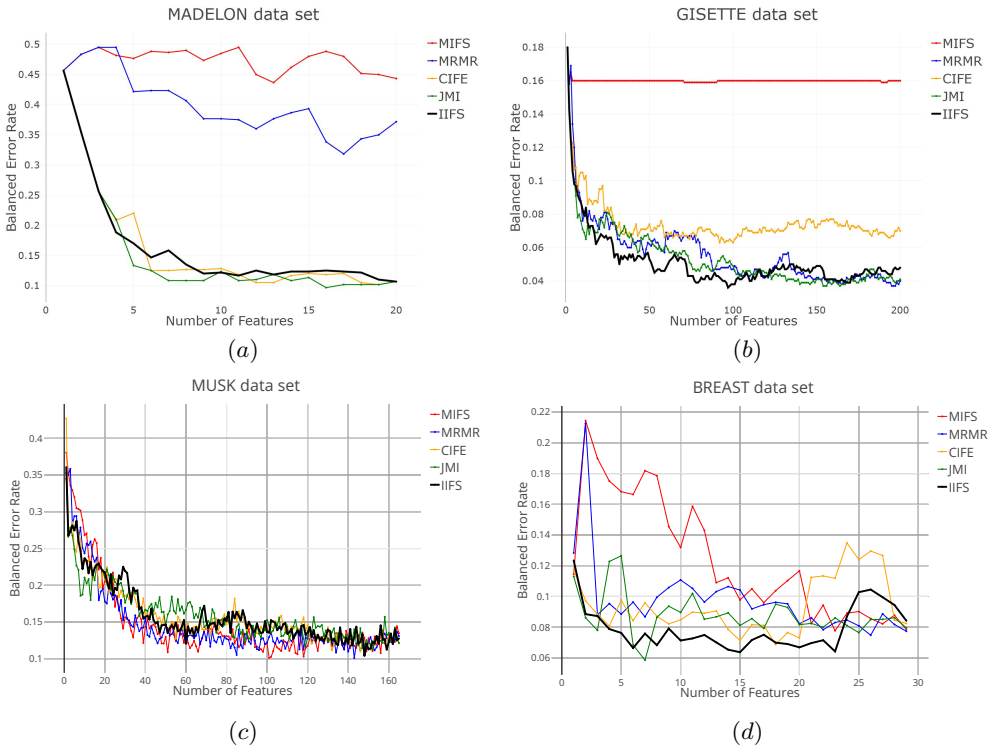

Fig. 2: Validation error curves for MADELON (a), GISETTE (b), MUSK (c) and BREAST (d) datasets.

considering high-order interactions as well as the comparison of IIFS with such methods, for example with a novel method proposed in [21].

References

1. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research **3** (2003) 1157–1182
2. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer (2009)
3. Taylor, M.B., Ehrenreich, I.M.: Higher-order genetic interactions and their contribution to complex traits. Trends in Genetics **31**(1) (2015) 34–40
4. Lin, D., Tang, X.: Conditional infomax learning: An integrated framework for feature extraction and fusion. In: Proceedings of the 9th European Conference on Computer Vision - Volume Part I. ECCV'06 (2006) 68–82
5. Kozachenko, L., Leonenko, N.: Sample estimate of the entropy of a random vector. Problemy Peredachi Informatsii **23**(2) (1987) 9–16
6. Jakulin, A., Bratko, I.: Quantifying and visualizing attribute interactions: An approach based on entropy. manuscript (2004)

7. Han, T.S.: Multiple mutual informations and multiple interactions in frequency data. Information and Control **46**(1) (1980) 26 – 45
8. McGill, W.J.: Multivariate information transmission. Psychometrika **19**(2) (1954) 97–116
9. Kojadinovic, I.: Relevance measures for subset variable selection in regression problems based on k-additive mutual information. Computational Statistics and Data Analysis **49**(4) (2005) 1205–1227
10. Meyer, P., Schretter, C., Bontempi, G.: Information-theoretic feature selection in microarray data using variable complementarity. IEEE Journal of Selected Topics in Signal Processing **2**(3) (2008) 261–274
11. Vergara, J.R., Estévez, P.A.: A review of feature selection methods based on mutual information. Neural Computing and Applications **24**(1) (2014) 175–186
12. Moore, J., Gilbert, J., Tsai, C., Chiang, F., Holden, T., Barney, N., White, B.: A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J. Theor. Biol. **241**(2) (2006) 256 – 261
13. Mielniczuk, J., Teisseyre, P.: A deeper look at two concepts of measuring gene-gene interactions: logistic regression and interaction information revisited. Genetic Epidemiology **42**(2) (2018) 187–200
14. Matsuda, H.: Physical nature of higher-order mutual information: intrinsic correlations and frustration. Physical Review E **62**(3 A) (2000) 3096–3102
15. Brown, G., Pocock, A., Zhao, M.J., Luján, M.: Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. Journal of Machine Learning Research **13**(1) (2012) 27–66
16. Battiti, R.: Using Mutual Information for Selecting Features in Supervised Neural-Net Learning. Ieee Transactions on Neural Networks **5**(4) (1994) 537–550
17. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Trans. on Pattern Analysis and Machine Intelligence **27**(8) (2005) 1226–1238
18. Yang, H.H., Moody, J.: Data visualization and feature selection: new algorithms for nongaussian data. Advances in Neural Information Processing Systems **12** (1999) 687–693
19. Guyon, I.: Design of experiments for the NIPS 2003 variable selection benchmark (2003)
20. Dheeru, D., Karra Taniskidou, E.: UCI machine learning repository (2017)
21. Shishkin, A., Bezzubtseva, A., Drutsa, A.: Efficient High-Order Interaction-Aware Feature Selection Based on Conditional Mutual Information. In: Advances in Neural Information Processing Systems. NIPS (2016) 1–9