CMIM-2: AN ENHANCED CONDITIONAL MUTUAL INFORMATION MAXIMIZATION CRITERION FOR FEATURE SELECTION

Jorge R. Vergara[1], Pablo A. Estévez[1]

[1]Department of Electrical Engineering
Universidad de Chile
(jorgever, pestevez)@ing.uchile.cl

Abstract

A new greedy feature selection criterion is proposed as an enhancement of the conditional mutual information maximization criterion (CMIM). The new criterion, called CMIM-2, allows detecting relevant features that are complementary in the class prediction better than the original criterion. In addition, we present a methodology to approximate the conditional mutual information to spaces of three variables, avoiding its estimation in high-dimensional spaces. Experimental results for artificial and UCI benchmark datasets show that the proposed criterion outperforms the original CMIM criterion.

Key words: Feature selection, conditional mutual information, information theory, relevance, redundancy.

1 Introduction

Feature selection plays an important role in the improvement of accuracy, efficiency and scalability of classifiers [10, 17]. Typically, supervised learning in classification is described as finding the relationship between a set of input features $F = \bigcup_{i=1}^{m} f_i$ and a vector class C. The relevant features are often unknown a priori in real problems. Usually among the features introduced to represent certain domain there are irrelevant and relevant features, and many of these relevant features are redundant to the vector class [14, 26]. On the other hand, the high dimensionality of data can cause the problem known as curse of dimensionality [11]. It has been demonstrated empirically that reducing the number of redundant and/or irrelevant features dramatically increases the computational speed of the classifiers and even their performance. In addition, it contributes to have a better understanding of the data and the classification models [10, 14]. The goal of feature selection is to find a subset of

relevant but non-redundant variables [26]. This can be formalized as the selection of the minimum subset S with features from the original set F, such that $P(C|S)$ is as close as possible to $P(C|F)$, where $P(C|S)$ and $P(C|F)$ are approximations of the probability distribution function of the class given the training set [14, 26]. The minimum subset S is called optimal subset.

In practice, the exhaustive search for an optimal subset of features of cardinality $|S| = d$ requires the evaluation of 2^d possible subsets, increasing the number of candidate subsets exponentially with the number of features. This combinatorial problem is known to be NP-hard [4]. In order to avoid evaluating all subsets candidates, many feature selection algorithms try to approximate the optimal subset of features. These methods fall into three categories [22]: Filters, Wrappers and Embedded. Filters evaluate the relevance of the features based on the intrinsic properties of the data, being independent of the learning process. Wrappers and embedded methods are dependent on the learning process and evaluate the relevance of the features according to the accuracy obtained by the classifiers. Wrappers, unlike embedded methods, define a classifier to find the optimal subset, so that various subsets of features are generated and evaluated. When searching for the optimal subset in the space of all subsets of features, the search method is wrapped around the model of classification. On the other hand, the search for an optimal subset of features in embedded methods is built into the classifier, so the search method is embedded in the classification model. Generally, the learning dependent methods yield subsets of features of better quality than the filter methods. However, the computational cost associated with learning and working in high dimensional spaces leaves the filter methods as a good option for preprocessing data.

In this paper, we propose a filter method that uses mutual information as a criterion for finding the optimal subset of features. Mutual information (MI) has been widely applied in feature selection methods [2, 3, 7, 8, 14, 19, 26]. The more important properties of MI are: (i) its ability to quantify nonlinear dependencies between features, and (ii) its invariance under transformations of the space [15]. Battiti [2] proposed a heuristic approach to find the optimal subset of features, an algorithm called Mutual Information Feature Selection (MIFS). It is based on using MI to rank the relevance of the features with respect to C and also on estimating the redundancy of the candidate feature to be selected with respect to the previously selected variables. Several variants of the MIFS algorithm have been proposed, that are more efficient in the management of the relevant features [7, 16, 19, 26]. However, all these methods work under the assumption that the relevance (irrelevance) of a feature is associated with the degree of its dependence with the class vector C. But it can occur that some features acting independently do not provide any information about C, but grouped together they do [8, 18]. An algorithm that deals with this problem is the Conditional Mutual Information Maximization

(CMIM) proposed by Fleuret [8], which makes a tradeoff between the predictive power of the candidate feature (relevance to the class vector C) and its independence from all features previously selected. This means that a candidate feature f_i would be selected only if it provides high information about C and this information cannot be provided by any of the previously selected features. The latter is understood as the non-redundancy of f_i with respect to the previously selected features.

In the remainder of this paper, section 2 describes a background on information theory and its application to feature selection. Section 3 presents a description of the original CMIM algorithm and its limitations for feature selection. Section 4 describes the proposed algorithm based on an enhancement of conditional mutual information criterion. Section 5 presents the simulation results on artificial and benchmark datasets, demonstrating the effectiveness of the proposed method. Finally, the conclusions are drawn in Section 6.

2 Background on information theory

2.1 Mutual information

A machine learning can be considered as a system that reduces the uncertainty of a vector of classes (or outputs) C by extracting the information contained in the input set F. Feature selection aims at finding the minimum subset of features $S \subset F$ that yields the highest information about the output. The information theory of Shannon [23] provides an efficient way to quantify the amount of information among random variables through MI.

Definition 1. *Given the feature set* $\boldsymbol{F}$*, the mutual information [5] is defined as amount of reduced uncertainty of the output class* $\boldsymbol{C}$*,*

$$I(F;C) \quad = \quad \sum_{f=1}^{N_f} \sum_{c=1}^{N_c} P(f,c) \log \frac{P(f,c)}{P(f)P(c)}, \tag{1}$$

where $P(f); f = 1, \dots, N_f$ *is the probability of the different features of the set* F*,* $P(c); c = 1, \dots, N_c$ *is the probability of the output class* C *and* $P(f,c)$ *is the joint probability of* F *and* C*.*

In the context of information theory, MI allows us to quantify the reduction of uncertainty of the output class C if the input set F is known.

On the other hand, the feature selection problem is formally defined as follows [2]:

Definition 2. *Given an initial set* $\boldsymbol{F}$ *with* $\boldsymbol{m}$ *features and the output class* $\boldsymbol{C}$*, find the subset* $\boldsymbol{S} \subset \boldsymbol{F}$ *with* $\boldsymbol{d}$ *features that is optimal in the sense that* $\boldsymbol{I}(\boldsymbol{S};\boldsymbol{C})$ *is maximum among all subsets of cardinality* $\boldsymbol{d}$*.*

2.2 Type of variables in feature selection

Traditionally, information theory is used to quantify concepts of relevance and redundancy widely used in feature selection methods. Next, we formalize these concepts and describe the different types of interaction between variables.

Relevance

Given an input set F and output class C, the first step is to find which features have more information to describe C. The decision of which features should be chosen is usually associated to the degree of dependency of each single feature when used to describe C. However, it can occur that a group of features is more relevant than the same features acting independently. This implies that there are levels of relevance.

Kohavi and John [13] used a probabilistic approach to quantify the concept of relevance and characterized features as relevant or irrelevant. Later, Yu and Liu [26] added another level by distinguishing between strong relevance, which selects those feature of F that provide the highest information with respect to C and this information does not exist in other features; and the weak relevance, which selects those features that only give a certain level of knowledge (relative to C) and can be replaced by other features without loss of information. Using information theory, these levels of relevance are formalized by the following definition.

Definition 3. *Given a set of features* $\boldsymbol{F}$*:*

A feature $\boldsymbol{f_i}$ *is irrelevant to* $\boldsymbol{C}$ *iff:*

$$\forall\, S \subseteq F \backslash f_i : I(f_i; C|S) = 0. \tag{2}$$

A feature $\boldsymbol{f_i}$ *is strongly relevant to* $\boldsymbol{C}$ *iff:*

$$I(f_i; C|F \backslash f_i) > 0. \tag{3}$$

A feature $\boldsymbol{f_i}$ *is weakly relevant to* $\boldsymbol{C}$ *iff:*

$$I(f_i; C|F \backslash f_i) = 0 \quad \wedge \quad \exists\, S \subset F \backslash f_i : I(f_i; C|S) > 0. \tag{4}$$

Redundancy

The concept of redundancy is associated with the level of dependency among two or more features in F, which can be quantified by the common information shared by features. The redundancy has the following properties: it is non-linear, symmetric, non-negative and non-decreasing with the number of features. The latter property is justified by the fact that, unlike the relevance, the amount of redundancy can never decrease when other variables are added [18].

The condition of redundancy can be captured in terms of MI through the notion of Markov blanket [14, 26], which is formalized in the following definition.

Definition 4. *(Markov blanket) Given a feature* $\boldsymbol{f_i}$*, the subset* $\boldsymbol{M} \subset \boldsymbol{F}$ ($\boldsymbol{f_i} \notin \boldsymbol{M}$) *is a Markov blanket for* $\boldsymbol{f_i}$ *if:*

$$I(f_i; \{C, F \backslash \{M, f_i\}\}) | M) = 0, \tag{5}$$

which expresses that $\boldsymbol{M}$ *must contain not only the information of* $\boldsymbol{f_i}$ *about* $\boldsymbol{C}$*, but also the information from the remainder of features* $\boldsymbol{F} \backslash \{\boldsymbol{M}, \boldsymbol{f_i}\}$.

The concept of Markov blanket is according to Koller and Sahami [14] a stronger condition of conditional independence, so it is not possible to discriminate between redundant or irrelevant features [18].

3 Conditional mutual information maximization (CMIM) criterion and its limitations

One strategy to find the optimal subset $S \subset F$ would be to evaluate all possible subsets in F of cardinality d, however, this is impossible in practice due to the combinatorial explosion of possible solutions. To avoid an exhaustive search, the greedy selection strategy [2] begins with an empty set of selected features (S) and successively adds features one by one. The greedy selection algorithm delivers the most relevant features using the following procedure:

1. Initialization: Set $F \leftarrow$ 'initial set of m features', $S \leftarrow$ 'empty set'.
2. Computation of the MI between the output class and each feature: $\forall\, f_i \in F$, compute $I(f_i; C)$.
3. Selection of the first feature: Find the feature f_i that maximizes $I(f_i; C)$. Set $F \leftarrow F \setminus f_i$, $S \leftarrow f_i$.
4. Greedy selection: Repeat until $|S| = d$.
   (a) Computation of the MI between features: $\forall\, f_i \in F$, compute $I(\{f_i, S\}; C)$.
   (b) Selection of the next feature: Choose the feature $f_i \in F$ that maximizes $I(\{f_i, S\}; C)$ and set $F \leftarrow F \setminus f_i$, $S \leftarrow f_i$.
5. Output the set S containing the selected features.

The greedy selection algorithm finds the optimal subset of features S by selecting a new feature $f_i \in F \backslash S$ incrementally through maximization of

$$I(\{f_i, S\}; C). \tag{6}$$

Using the chain rule property of MI, the functional (6) can be rewritten as

$$I(\{f_i, S\}; C) = I(S; C) + I(f_i; C|S). \tag{7}$$

In (7), the second term on the right hand side, $I(f_i; C|S)$, measures the relevance of the candidate variable f_i to predict the output C under the influence of set S. The first term $I(S; C)$ is the information about class C given by the previously selected feature set S, but this information is common to all candidate variables and therefore it can be discarded. Thus, the greedy selection algorithm can be modified to find the subset S that maximizes $I(f_i; C|S)$, which represents a criterion of relevance [3]. In analytical terms, this approach yields the most relevant variables according to

$$REL = \arg \max_{f_i \in F \backslash S} I(f_i; C|S). \tag{8}$$

The greedy selection algorithm, using the criterion of relevance, avoids the evaluation of $\binom{m}{d}$ candidate subsets, but estimating entropy or mutual information in high-dimensional spaces is computationally intractable [8,19].

The conditional mutual information maximization algorithm (CMIM) [8, 25] approximates the relevance criterion (8), by considering the MI between the candidate variable f_i and the output class C given each one of the variables in the set S, separately. It allows preserving a certain tradeoff between the power prediction of f_i with respect to the output and the independence of the candidate feature with each one of the variables previously selected. CMIM considers that feature f_i is relevant only if it has information on C and this information is not contained by any of the variables already selected. Formally, the CMIM iterative scheme selection is expressed as follows,

$$\text{CMIM} = \begin{cases} \arg \max_{f_i \in F} \{I(f_i; C)\} & \text{for } S = \emptyset \\ \arg \max_{f_i \in F \backslash S} \left\{ \min_{f_j \in S} I(f_i; C|f_j) \right\} & \text{for } S \neq \emptyset. \end{cases} \tag{9}$$

The justification of using the minimum function in (9) is based on an approximation of the concept of Markov blanket. Fleuret [8] considers that the set M in equation (5) consists of a single feature in S. Therefore the feature f_i can be discarded of the selection process if there exists a feature $f_j \in S$ such that f_i and C are conditionally independent given f_j. Since MI is always positive we have

$$\min_{f_j \in S} I(f_i; C|f_j) = 0. \tag{10}$$

Moreover, the feature $f_i \in F \backslash S$ with the highest value of $I(f_i; C|f_j)$ is the most relevant one, which justifies the maximum function in (9).

Although CMIM avoids redundancy, selects the relevant variables and avoids the multidimensional calculation of MI, its ability to identify and select variables interacting as groups with the output [12, 18] can be degraded when selecting the minimum value of conditional MI. To illustrate this limitation of CMIM we introduce the following example.

Example 1. Let x_1, x_2, x_3 and x_4 be four binary random variables, related by the xor ($\oplus$) function, $C = x_2 \oplus x_4$, as shown in Table 1. As can be seen from Table 1, none of the variables acting alone gives information of the class C, i.e., the relevance for each variable is null, $I(x_1; C) = I(x_2; C) = I(x_3; C) = I(x_4; C) = 0$. On the other hand, the pair of variables $\{x_2, x_4\}$ has the highest relevance, i.e., $I(\{x_2, x_4\}; C) = I(x_2; C|x_4) = I(x_4; C|x_2) = H(C) > 0$, where $H(C)$ is the entropy of C.

Table 1. XOR problem plus two irrelevant variables.

x_1	x_2	x_3	x_4	$C = x_2 \oplus x_4$
0	1	1	1	0
0	1	1	0	1
0	0	1	1	1
0	0	1	0	0

Since all variables have null relevance, the first feature to be selected using CMIM depends only on the order they are entered, i.e., the feature selected by CMIM is x_1. For the selection of the second variable is necessary to determine $I(x_i; C|x_1), i = \{2,3,4\}$ of the remaining candidate variables. However, x_1 is independent of the other variables and class C, therefore $I(x_i; C \mid x_1) = 0, i = \{2,3,4\}$. Considering again the order in which variables are entered, CMIM will select variable x_2.

After selecting the variable x_2, the next feature to be selected should be x_4 since $I(x_4; C|x_2)$ is greater than $I(x_3; C|x_2)$. However, CMIM computes the minimum between $I(x_4; C|x_2)$ and $I(x_4; C|x_1)$, which is zero. Thus, the variable x_4 is discarded as a relevant variable, and the variable x_3 is erroneously chosen as the third feature (order in which variables are entered). As the original CMIM criterion prioritizes those variables that give the minimum conditional MI, it will not find a solution to the XOR problem. In general, in problems where the variables are highly complementary (or dependent) to predict C, the CMIM algorithm will fail to find that dependence among the variables.

The new criterion proposed below changes the minimum function to the average function. Because $I(x_i; C \mid x_1) = 0, i = \{2,3,4\}$, the maximum function will be applied to $I(x_4; C|x_2)$ and $I(x_3; C|x_2)$. Since the latter term is zero, the variable x_4 will be correctly selected.

4 Enhanced conditional mutual information maximization criterion (CMIM-2)

The proposed feature selection criterion is an improvement of the CMIM criterion. It maintains the advantages of the original criterion, but it solves the problem of variables that are relevant in pairs.

Considering that we want to avoid the calculation of conditional MI in high dimensional spaces, we approach functional (8) by using arithmetical averages of the conditional MI. For this, let us define $S = \bigcup_{j=1}^{d} f_j$ where $|S| = d$, and $G_j = S \backslash f_j$, and using the chain rule repeatedly for MI, the following expression is obtained as

$$\begin{array}{rcl} I(f_i; C|S) & = & I(f_i; C|f_1) + [I(G_1; C|\{f_1, f_i\}) - I(G_1; C|f_1)] \\ I(f_i; C|S) & = & I(f_i; C|f_2) + [I(G_2; C|\{f_2, f_i\}) - I(G_2; C|f_2)] \\ \vdots & = & \vdots \quad\quad\quad \vdots \quad\quad\quad\quad \vdots \\ I(f_i; C|S) & = & I(f_i; C|f_d) + [I(G_d; C|\{f_d, f_i\}) - I(G_d; C|f_d)] \\ \hline d\, I(f_i; C|S) & = & \sum_{f_j \in S} \left[I(f_i; C|f_j) + \left[I(G_j; C|\{f_j, f_i\}) - I(G_j; C|f_j) \right] \right]. \end{array} \quad (11)$$

Note that the first term on the right hand size of equation (11) represents the MI between f_i and C given f_j, where f_j is the jth variable belonging to the subset S. The second term represents the remaining conditional MI present in high-dimensional spaces. Restricting the working space to variables $\{f_i, f_j\}$ and class C, i.e., avoiding the estimation of MI in spaces of more than three variables; the equation (11) can be simplified in a first order approximation to:

$$I(f_i; C|S) \approx \frac{1}{d} \sum_{f_j \in S} I(f_i; C|f_j). \quad (12)$$

The new proposed criterion for feature selection is defined as:

$$\text{CMIM-2} = \begin{cases} \arg\max\limits_{f_i \in F} \{I(f_i; C)\} & \text{for } S = \emptyset \\ \arg\max\limits_{f_i \in F \setminus S} \dfrac{1}{d} \sum\limits_{f_j \in S} I\left(f_i; C \middle| f_j\right) & \text{for } S \neq \emptyset. \end{cases} \tag{13}$$

The computational cost of CMIM-2 algorithm depends on the way of estimating $I(f_i; C|f_j)$, which in the our case is $\mathrm{O}(L \times M \times N \log N)$, where L is the number of classes of vector C, M is the number of features in set F and N is the number of data samples available. The estimation of MI is performed through contingency tables for discrete features and the Fraser's algorithm [9] for continuous features. Fraser's algorithm estimates MI by using adaptive histograms. The method chosen for sorting samples is heapsort [20], which has a complexity of $\mathrm{O}(N \log N)$ in the worst case.

5 Experiments

The proposed criterion was tested on three feature selection experiments. The first experiment uses the artificial data set MONK-1 [24], in order to show the importance of using the average conditional MI instead of the minimum. The second experiment involves two datasets described in Table 2, which are commonly used to compare feature selection techniques. The third experiment partially replicates one of the experiments described by Fleuret [8] on the Thrombin dataset. In this section the performance of the proposed criterion (CMIM-2) is compared with the original CMIM method and the criterion of selection based on the highest mutual information between a single feature and the vector class (RANK) [6].

5.1 Experiment 1: MONK-1

MONK-1 is one out of three problems generated from the MONK's problem [24], which describes the artificial domain of a robot using six attributes. Each problem is generated according to the classification task that must perform the robot, where the outputs are obtained as logical operations of the variables. For the MONK-1 problem, the output is obtained as:

Table 2. Datasets used in experiment 2. The column n contains the number of samples, column m contains the number of features, column c contains the number of classes, and column type contains the kind of features on each dataset: D: discrete, C: continuous.

Dataset	n	m	c	type
Arrhythmia	452	279	16	D,C
Spambase	4601	57	2	D,C

Table 3. Mutual information $I(x_i; C)$ between each variable and the vector class for the MONK-1 problem.

$C \backslash x_i$	x_1	x_2	x_3	x_4	x_5	x_6
C	0.0685	0.0073	0.0030	0.0182	**0.2987**	0.0031

$$C = (x_1 \equiv x_2) \vee (x_5 \equiv \text{'red'}), \tag{14}$$

where ≡ and ∨ stand for 'identical to' and OR function, respectively.

As can be seen from (14), the variables that provide information of class C are: $\{x_1, x_2, x_5\}$. Table 3 shows the relevant information of each variable with respect to the class ($I(x_i; C)$, $i = 1, \ldots, 6$) and Table 4 shows the conditional MI ($I(x_i; C|x_j)$) for all pair of variables.

As shown in Table 3, feature x_5 has the highest information regarding the output. This result was expected since x_5 does not interact with any other variable except the variable C. For selecting the second feature, the conditional mutual information $I\ (x_i; C|x_5)$ is estimated for each of the candidate features $i = (1,2,3,4,6)$. The results can be seen in the column x_5 of Table 4, where feature x_1 has the highest value. The variables selected so far are common to both CMIM and CMIM-2 criteria.

The third feature to be selected must be x_2, thereby completing the triplet of relevant variables for the MONK-1 problem. But as can be seen in the column x_1 of Table 4, the conditional information $I(x_2; C|x_1)$ is maximum. CMIM will ignore the maximum value because its goal is to find those features that have the greatest independence with respect to the variables previously selected (minimum in (9)), but also the greatest MI with respect to class C (maximum in (9)), i.e., $\max_i(\min(I(x_i; C|x_1), I(x_i; C|x_5)),\ i = \{2,3,4,6\}$. Thus, the feature selected by CMIM is x_4.

Table 4. Conditional mutual information $I(x_i; C|x_j)$ for each pair of variables of the MONK-1 problem.

$x_i \backslash x_j$	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0	0.5096	0.0718	0.0654	**0.0640**	0.0745
x_2	**0.4483**	0	0.0095	0.0080	0.0100	0.0242
x_3	0.0062	0.0052	0	0.0116	0.0119	0.0251
x_4	0.0151	0.0189	0.0268	0	0.0331	0.0333
x_5	0.2942	0.3014	0.3076	0.3136	0	0.2993
x_6	0.0091	0.0200	0.0252	0.0182	0.0038	0

In CMIM-2, the MI values of the candidate variable $x_i, (\, i = \{2,3,4,6\})$ with respect to class C conditional on each of the variables previously selected ($\{x_1, x_5\}$) are weighted, i.e., $0.5 \cdot (I(x_i; C|x_1) + I(x_i; C|x_5))$. Thus CMIM-2 can capture additional variable information and select feature x_2 correctly.

5.2 Experiment 2: Benchmark datasets

In order to measure the performance of the proposed criterion two datasets available at the UCI repository [1] are used. The basic information for each dataset is given in Table 2.

The proposed selection criterion (CMIM-2) was compared with the original CMIM method [8] and the ranking method based on MI (RANK) [6].

For external validation, two classifiers were used: k-nearest neighbor classifier (KNN) and Support Vector Machine (SVM) with Gaussian kernel. The evaluation of feature subsets delivered by the different methods was made as follows: (i) the top 30 most relevant features delivered by each selection criterion on each data set were drawn, (ii) 10 rounds of cross validation were performed on subsets containing 1, 3, 6, 9, 12, 15, 18, 21, 24, 27 and 30 features drawn from the ranking yielded by each feature selection criterion. The neighborhood parameter in KNN and the kernel size parameter in SVM were selected by optimizing the validation error over the 10 rounds of cross validation. The percentage of correct classifications, i.e., the classifier accuracy, is presented in Figure 1 for the Arrhythmia data set.

Figure 1 shows a significant increase in the classifier accuracy for CMIM-2 criterion with respect to CMIM and RANK criterion on the Arrhythmia dataset. This means that there are variables in the Arrhythmia data set that are highly complementary for the task of predicting the class outputs [18]. The CMIM results are comparable to RANK, because the minimum function in (9) is looking for variables that are independent of previously selected subset of variables S, without taking into account the interaction between the candidate feature and the subset of selected variables S.

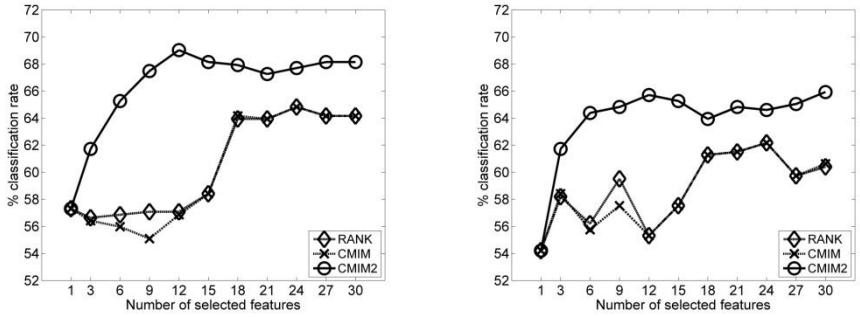

Figure 1. Average classification rate over 10 tests on the Arrhythmia dataset vs. number of variables selected by three feature selection criteria: RANK, CMIM and CMIM-2. (a) Using a Gaussian-SVM classifier. (b) Using a K-NN classifier.

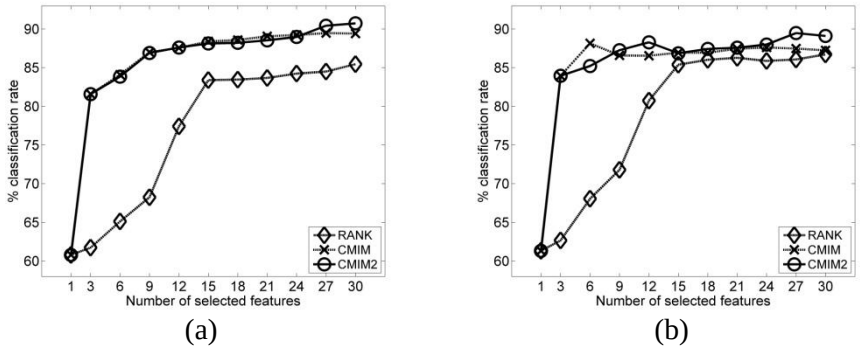

Figure 2. Average classification rate over 10 tests on the Spambase data set vs. number of variables selected by three feature selection criteria: RANK, CMIM and CMIM-2. (a) Using a Gaussian-SVM classifier. (b) Using a K-NN classifier.

Figure 2 shows the results for the Spambase data set, where very similar classification rates were obtained for CMIM and CMIM-2 methods. In our extensive simulations we have found that the CMIM-2 performance is superior or equal to CMIM's performance but never worse.

5.3 Experiment 3: Thrombin data set

The Thrombin data set was created for predicting molecular bioactivity for drug design. This database contains 1,909 samples and 139,351 features (active or inactive).

In all experiments 10 rounds of cross validation were performed to choose the best model [21], where each partition kept the proportion of positive and negative examples. Moreover, we conducted Student's paired two-tailed t-test in order to evaluate if there are statistical significant inferences between averages of cross validation error of CMIM-2 and RANK or CMIM.

Because the Thrombin data set is highly unbalanced (42 positive examples and 1,867 negative examples), the training and validation errors were measured by using the balanced error rate (BER), defined as follows:

$$\text{BER}=\frac{FP+FN}{2}, \tag{15}$$

where FP is the false positive rate and FN is the false negative rate.

The validation experiments were performed by choosing the top 5 and 10 relevant features delivered by the feature selection criterion. These features were entered as inputs to a SVM classifier with Gaussian kernel and a KNN classifier. The neighborhood parameter of KNN and kernel size of SVM were adjusted by minimizing BER. Due to the unbalanced data, other two classifiers were used: linear perceptron (PCT) and a Naïve Bayesian classifier (NB).

Tables 5 and 6 show the prediction error for different combinations of classifiers and feature selection methods using the top 5 and 10 features selected, respectively. In addition, the p-values show that the difference in average errors between CMIM-2 and CMIM, and CMIM-2 and RANK, are statistically significant.

Table 5. Average BER obtained with different combinations of feature selection methods and classifiers for the first 5 featured selected of the Thrombin dataset.

Classifier	**CMIM-2**	**CMIM**		**RANK**	
	BER	BER	p-value	BER	p-value
PCT	**15.17**	18.46	0.32	21.18	0.13
NB	**12.76**	14.22	0.62	18.81	0.05
SVM	**15.85**	26.61	0.03	24.74	0.02
KNN	**18.28**	25.36	0.08	29.41	0.02

Among all combinations, the CMIM-2-NB has the lower BER in validation, confirming the effectiveness of the proposed method. For the same classifier, it is found that CMIM-2 obtained in all cases lower BER values than CMIM and RANK, except for PCT with 10 features.

Note that the results delivered by CMIM-2 for each classifier do not necessarily match those published by Fleuret [8], since we use the whole set of 139,351 variables, while Fleuret used 2,500 randomly selected features only.

Table 6. Average BER obtained with different combinations of feature selection methods and classifiers for the first 10 featured selected of the Thrombin dataset.

Classifier	CMIM-2	CMIM		RANK	
	BER	BER	p-value	BER	p-value
PCT	20.91	26.29	0.18	**18.68**	0.49
NB	**15.21**	16.59	0.55	18.76	0.08
SVM	**17.16**	21.83	0.26	20.91	0.19
KNN	**17.10**	24.33	0.03	22.77	0.14

6 Conclusions

An enhanced conditional mutual information algorithm for feature selection has been proposed. The new algorithm, called CMIM-2, is able to detect pairs of relevant variables that act complementarily in predicting the class. Experimental results for artificial and UCI benchmark datasets show that the proposed algorithm outperforms the original CMIM algorithm.

An advantage of the proposed approach is that it is possible to improve the proposed criterion by considering a higher-order approximation of eq. (11). It would be of interest to study how it influences the selection of relevant variables that act complementarily in sets of three o more features. These ideas can be used to identify and formalize new levels of interaction among variables, beyond the traditional definitions of relevance and redundancy.

Acknowledgment

This work was funded by Conicyt-Chile under grant Fondecyt 1080643.

References

1. Asuncion A., Newman D., 2007. *UCI machine learning repository*. http://www.ics.uci.edu/~mlearn/MLRepository.html, University of California, Irvine, School of Information and Computer Sciences.
2. Battiti R., 1994, *Using mutual information for selecting features in supervised neural net learning*, IEEE Transactions on Neural Networks, Vol. 5, No 4, pp. 537–550.
3. Bell D. A., Wang H., 2000, *A formalism for relevance and its application in feature subset selection*, Machine Learning, Vol. 41, No. 2, pp. 175–195.
4. Blum A. L., Rivest R. L., 1992, *Training a 3-node neural networks is NP-complete*, IEEE Transactions on Neural Networks, Vol. 5, No. 1, pp. 117–127.

5. Cover T. M., Thomas J. A., 2006, *Elements of Information Theory*, 2nd ed. Wiley-Interscience.
6. Duch W., Winiarski T., Biesiada J., Kachel A., 2003, *Feature selection and ranking filter*, In International Conference Artificial Neural Networks (ICANN) and International Conference Neural Information Processing (ICONIP), pp. 251–254.
7. Estévez P. A., Tesmer M., Pérez C. A., Zurada J. M., 2009, *Normalized mutual information feature selection*, IEEE Transactions on Neural Networks, Vol. 20, No 2, pp. 189–201.
8. Fleuret F., Guyon I., 2004, *Fast binary feature selection with conditional mutual information*, Journal of Machine Learning Research, Vol. 5, pp. 1531–1555.
9. Fraser A. M., Swinney H. L., 1986, *Independent coordinates for strange attractors from mutual information*, Physical Review A, Vol. 33, No. 2, pp. 1134–1140.
10. Guyon I., Elisseeff A., 2003, *An introduction to variable and feature selection*, Journal of Machine Learning Research, Vol. 3, pp. 1157–1182.
11. Hastie T., Tibshiran R., Friedman J., 2001, *The Elements of Statistical Learning.* Springer.
12. Jakulin A., Bratko I., 2003, *Quantifying and visualizing attribute interactions*, ACM (Computing Research Repository) ,Vol. cs.AI/0308002, pp. –.
13. Kohavi R., John G. H., 1997, *Wrappers for feature subset selection*, Artificial Intelligence Vol. 97, No. 1-2, pp. 273 – 324.
14. Koller D., Sahami M., 1996, *Toward optimal feature selection*, Technical Report 1996-77, Stanford InfoLab.
15. Kullback S., 1997, *Information Theory and Statistics.* New York: Dover.
16. Kwak N., Choi C.-H., 2002, *Input feature selection for classification problems*, IEEE Transactions on Neural Networks, Vol. 13, No. 1, pp. 143–159.
17. Liu H., Dougherty E., Dy J., Torkkola K., Tuv E., Peng H., Ding C., Long F., Berens M., Parsons L., Zhao Z., Yu L., Forman G., 2005, *Evolving feature selection*, IEEE Intelligent Systems, Vol. 20, No. 6, pp. 64–76.
18. Meyer P., Schretter C., Bontempi G., 2008, *Information-theoretic feature selection in microarray data using variable complementarity*, IEEE Journal of Selected Topics in Signal Processing, Vol. 2, No. 3, pp. 261–274.
19. Peng H., Long F., Ding C., 2005, *Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp. 1226–1238.
20. Press W., Flannery B., Teukolsky S., Vetterling W., 1992, *Numerical Recipes in C, 2nd ed. Cambridge*, U.K.: Cambridge University Press.
21. Ripley B. D., 2008, *Pattern Recognition and Neural Networks.* Cambridge University Press.
22. Saeys Y., Inza I., Larranaga P., 2007, *A review of feature selection techniques in bioinformatics*, Bioinformatics, Vol. 23, No. 19, pp. 2507–2517.

23. Shannon C. E., 1948, *A mathematical theory of communication*, Bell System Technical Journal, Vol. 27, pp. 379–423, 625–56.
24. Thrun S., Bala J., Bloedorn E., Bratko I., Cestnik B., Cheng J., Jong K. D., Dzeroski S., Hamann R., Kaufman K., Keller S., Kononenko I., Kreuziger J., Michalski R., Mitchell T., Pachowicz P., Roger B., Vafaie H., de Velde W. V., Wenzel W., Wnek J., Zhang J., 1991, *The MONK's problems: A performance comparison of different learning algorithms*, Technical Report CMU-CS-91-197, Carnegie Mellon University, Computer Science Department, Pittsburgh, PA.
25. Wang G., Lochovsky F. H., 2004, *Feature selection with conditional mutual information MaxiMin in text categorization*, In proceedings of the thirteenth ACM international conference on information and knowledge management, New York, USA, pp. 342–349.
26. Yu L., Liu H., 2004, *Efficient feature selection via analysis of relevance and redundancy*, Journal Machine Learning Research, Vol. 5, pp. 1205–1224.