2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

An Information Theoretic Feature Selection Framework Based on Integer Programming

Siqi Nie
Department of Electrical, Computer and Systems Engineering
Rensselaer Polytechnie Institute
Troy, NY 12180
Email: nies@rpi.edu

Tian Gao
Department of Electrical, Computer and Systems Engineering
Rensselaer Polytechnie Institute
Troy, NY 12180
Email: gaot@rpi.edu

Qiang Ji
Department of Electrical, Computer and Systems Engineering
Rensselaer Polytechnie Institute
Troy, NY 12180
Email: qji@ecse.rpi.edu

***Abstract*—We propose a general framework for information theoretic feature selection based on the integer programming. Filter feature selection methods usually rely on a greedy forward or backward selection heuristic to find a satisfactory set of features, as the exact search is a combinatorial problem. We formulate the existing filter information theoretic criteria into an integer programming problem, and by using objective functions, we can represent many different existing scoring criteria. The integer programming framework can be solved efficiently by the existing solvers. We demonstrate the superior performance of the integer programming formulation over its corresponding criterion empirically.**

I. Introduction

The high dimensionality of data has been a significant issue in many machine learning problems. With the fast developing techniques to sense and record the environment around us, some practical applications may have thousands of features. Employing all the feature may not only lead to massive computation burden, but also cause the issue of over-fitting to unimportant aspects of the data, because many of the features may be totally irrelevant, or redundant to the task. Therefore, it is essential to automatically identify a subset of all the features that are most relevant to the task, which is the problem of feature selection. In many fields, feature selection has become a necessity [6].

This work tackles the feature selection problem using a filter approach, which is independent of the classifiers. For filter methods, during the past two decades, researchers have proposed different kinds of criteria for feature selection. One basic idea is that selected features should have high correlation with the task, and given the selected features, the discarded features should have small influence on the task. The criteria can be formulated as score functions. But since the combination of subsets of all the features are exponential with the number of features, existing works commonly employ a simple greedy approach to use an approximated score and select a subset of features iteratively. Due to the nature of greedy approaches, the final solution may not be optimal. For example, in an iterative process, the subset of feature is built from an empty set, by adding features one by one only considering existing set of selected features. When the procedure terminates, one intuitive way of evaluating the selected features is according to the classification accuracy using these features. Besides classification performance, the score of the selected features is an alternative criterion. But with the completion of the selected feature set, the score of each feature is changed due to conditioning on a larger set, and hence there is no guarantee that the selected features still have large scores, or the features discarded are still irrelevant to the target variable. In other words, an approach that can find all the useful features simultaneously and discard the irrelevant features with theoretical guarantee is still missing in the literature.

In this work, we fill in the blank by proposing a integer programming method based on some commonly used criteria, that can automatically select all the features simultaneously. To demonstrate the effectiveness, we employ two commonly used criteria: Conditional Infomax Feature Extraction (CIFE) [5], and joint mutual information (JMI) [20], and formulate these criteria into an Integer Programming framework. With well-developed techniques for solving integer program, our method can identify a better subset of features with different sizes efficiently. The framework can be extended to other filter feature selection criteria as well.

The paper is structured as follows. Section 2 presents an overview of the related work. In Section 3 we briefly introduce the background for feature selection. Section 4 introduces the integer programming approach based on two different criteria, CIFE and JMI. Experimental evaluation is given in Section 5. The paper is concluded in Section 6.

II. Related Work

Feature selection methods can be roughly divided into three different directions: wrapper methods, embedded methods, and filter methods. The wrapper methods [10, 19] use the accuracy from a specific classifier as the criterion to choose features. Wrapper methods could find very good features for the specific classifier but are prone to overfit, are classifier-dependent, and very costly in computation, compared to the other two approaches. Embedded methods [4, 7, 16, 18, 22] relax the wrapper methods assumption and instead exploit a classifier specific criterion during the classifier construction process. These guided methods are less expensive in computation, less likely to overfit, but are still classifier-dependent. In contrast,

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

filter methods [8, 11, 12, 21] select relevant features by using only the statistics of the data to develop some score criteria. Least amount of assumptions are made. Filter methods are classifier-independent, and very fast in computation. Extensive studies on various data sets [1] show the competitive performance of filtered methods.

In this paper, we focus on the filter feature selection methods using information theory and establish a better searching framework. Some popular information theoretic algorithms include Mutual Information Maximization (MIM) [14], Conditional Infomax Feature Extraction (CIFE) [5], Mutual Information Feature Selector (MIFS) [13], max-Relevance min-Redundancy (MRMR) [15], and Joint Mutual Information (JMI) [20]. All these methods propose different scoring criteria to select features but they are all fitted under one unified conditional likelihood maximization framework [3], giving an insight on the relationships among all these information theoretic criteria. These algorithms have two-fold approximation under the conditional likelihood framework. First, filter information theoretic algorithms use a simplified score of the conditional likelihood estimation. Existing criteria cannot select a set at the same time and instead elect to choose a single feature at each step. Secondly, the heuristic does not consider the relationship between a single queried feature and the entire selected set of features due to a large search space. Actually, existing criteria choose features only based on the current selected features through a forward greedy approach. These two approximations reduce the optimality and performance of the feature search.

Based on the conditional likelihood maximization framework, we propose to use a better search framework, i.e. the integer programming based approach, to select an optimal set of features, rather than the traditional forward sequential selection. Many embedded algorithms such as [2] uses the integer programming and often enforce the sparsity constraint. A zero-one integer programming [9] is proposed but it uses a Box classifier, which would make it classifier-dependent, and uses a linear combination of the class label and the features as the constraint, instead of a mutual information based criterion. To the best of our knowledge, integer programming has not been used in any filter information theoretic feature selection algorithm. We directly solve for a desired set of features and thus does not need a greedy select process that only choose one feature at a time. This reduces the first approximation of the existing methods, even though we still use the same approximated score. In addition, our method makes the decision to select each feature or not by considering the entire optimal features set, rather than just the current set of optimal features. This reduces the second approximation on the conditioned set mentioned above.

III. Background

In this section, we give a brief introduction of the information theoretic concepts, which will be used in the following sections.

In information theory, entropy measures the uncertainty of a random variable. The entropy, denoted $H(X)$, is defined as,

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x)\,, \tag{1}$$

where the lower case x denotes a possible value that the variable X can take from the value space $\mathcal{X}$. Following the rules of probability theory, conditional entropy of X given Y is defined as,

$$H(X|Y) = -\sum_{y\in\mathcal{Y}} p(y)\sum_{x\in\mathcal{X}} p(x|y)\log p(x|y)\,. \tag{2}$$

Conditional entropy measures the uncertainty remaining in X given that the value of another random variable Y is known.

Mutual Information between X and Y is the information shared by X and Y, which is defined as follows:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \sum_x\sum_y p(xy)\log\frac{p(xy)}{p(x)p(y)}\,, \end{aligned} \tag{3}$$

where $H(X)$ is the entropy, and $H(X|Y)$ is the conditional entropy of X given Y. Mutual information can be interpreted as the uncertainty in X which is removed by knowing Y. The mutual information can also be conditional.

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|YZ) \\ &= \sum_z p(z)\sum_x\sum_y p(xy|z)log\frac{p(xy|z)}{p(x|z)p(y|z)}\,. \end{aligned} \tag{4}$$

Conditional mutual information can be thought as the information still shared between X and Y after the value of a third variable Z is observed. The concept of conditional mutual information is the key for deriving different criteria for feature selection.

IV. Integer Program Formulation

In this section, we introduce an integer programming formulation for feature selection based on different feature selection criteria. First, a general objective function is proposed, which works together with any criteria function. Second, Integer Program (IP) formulation based on two commonly use criteria (CIFE and JMI) is given. We then generalize our framework to any criteria with linear terms, and give a performance guarantee in terms of scores.

A. Conditional Likelihood Maximization Framework

We first introduce some notations that will be used throughout this paper. An indicator vector $\mathbf{d} = \{d_i\}$ denotes the selected features,

$$d_i = \begin{cases} 1 & \text{feature } X_i \text{ is selected} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Notation $V = \{X_1, \ldots, X_n\}$ is the set of all feature variables, and T is the target variable (or class label). Notation U indicates the set of selected features, which is the full feature vector $\mathbf{x}$ projected onto the dimensions specified by $\mathbf{d}$. Notation $\tilde{U}$ denotes the set of unselected features.

The problem statement of feature selection under conditional likelihood maximization framework is to identify the minimum subset of features such that the conditional likelihood of the training labels is maximized. For i.i.d. data $\mathcal{D} = \{(\mathbf{x}^i, y^i); i = 1, \ldots, N\}$ the conditional likelihood of labels is,

$$\mathcal{L}(d|\mathcal{D}) = \prod_{i=1}^{N} p(y^i|U^i)\,. \tag{6}$$

The optima of the conditional likelihood coincide with that of the conditional mutual information [1]:

$$\arg\max_d \mathcal{L}(d|\mathcal{D}) = \arg\min_d I(\tilde{U}; t|U)\,. \tag{7}$$

Directly minimizing the conditional information above is intractable, due to the exponential number of possible sets U. To address this issue, researchers propose various filter criteria, to approximate the objective.

A filter criterion J is defined to measure the importance of a feature or feature subset. Given a criterion, current filter feature selection algorithms are typically a sequential search considering features one-by-one. However, one issue is that when selecting next feature, the conditional mutual information is based on the current selected features, so features are highly dependent on the previous selected features. On the contrary, in this work, we evaluate each feature based on the entire set (not current selected features), to keep or discard the feature. Since maximizing the conditional likelihood of labels is equivalent to minimizing the score of unselected features (Equation 7), we have the following objective function:

$$\min \sum_{k \in V \setminus U} J(x_k; t|U) + \lambda |U| \quad , \tag{8}$$

where $J(x_k; t|U)$ is some criterion of candidate feature x_k, given the selected features U (which also needs to be identified). $|U|$ denotes number of features selected. The objective function balances the score and size of the selected features using a parameter λ. It intends to minimize the score for unselected features, as well as the number of selected features at the same time. In general, a larger set of selected features will lead to small scores for the unselected features, and vice versa.

By making different assumptions, different criterion can be used for feature selection. In the following two subsections, we employ two criteria for feature selection: Conditional Infomax Feature Extraction (CIFE) and Joint Mutual Information (JMI). Such criteria have a linear formulation, which can serve as linear terms in the objective function of integer programming.

B. Conditional Infomax Feature Extraction

Conditional Mutual Information (CMI) criterion is defined as,

$$J_{\text{cmi}}(x_k) = I(x_k; t|U)\,. \tag{9}$$

It is a score function for a single feature, as the first-step approximation of CMI score of a feature set ($I(\tilde{U}; t|U)$).

Conditional Infomax Feature Extraction (CIFE) criterion is a second-step approximation of the CMI by making the following assumption,

Assumption 1: [3] For all unselected features x_k, assume the following,

$$\begin{aligned} p(\mathbf{x}|x_k) &= \prod_{j \in U} p(x_j|x_k) \\ p(\mathbf{x}|x_k t) &= \prod_{j \in U} p(x_j|x_k t) \end{aligned} \tag{10}$$

This states that the selected features X are independent and class-conditionally independent given the unselected feature x_k under consideration.

Under the assumption, the CMI criterion can be decomposed as

$$I'(x_k; t|U) = I(x_k; t) - \sum_{j \in U} I(x_j; x_k) + \sum_{j \in U} I(x_j; x_k|t)\,. \tag{11}$$

Here the prime $I'(x_k; t|U)$ is used to denote the approximation of the 'true' CMI $I(x_k; t|U)$. This approximated criterion is also known as CIFE, which we denote as $J_{\text{cife}}(x_k)$.

To consider the selected features jointly, we propose to identify all the selected features simultaneously. Under the criteria of CIFE, our general objective function (Equation 8) becomes,

$$\min \sum_{k \in V \setminus U} J_{\text{cife}}(x_k) + \lambda |U| \quad . \tag{12}$$

For brevity, we introduce another parameter,

$$c_{j,k} = -I(x_j; x_k) + I(x_j; x_k|t)\,, \tag{13}$$

and denote the mutual information of x_k and t as $I_k = I(x_k; t)$. Notice that $|U| = \sum_{k=1}^{n} d_k$. With the indicator variables, CIFE score becomes:

$$J_{\text{cife}}(x_k) = I_k + \sum_{j=1}^{n} c_{j,k} d_j\,. \tag{14}$$

The range of $k \in V \setminus U$ in Equation 12 still involves the indicator variable. Notice that we do not compute score for a feature if it is already selected in the subset U. This issue is addressed by adding $(1 - d_k)$ as the multiplier for each score, as shown below,

$$(1 - d_k) J_{\text{cife}}(x_k) = (1 - d_k) \left(I_k + \sum_{j=1}^{n} c_{j,k} d_j \right) , \tag{15}$$

such that,

$$(1 - d_k) J_{\text{cife}}(x_k) = \begin{cases} 0 & \text{feature } X_k \text{ is selected} \\ J_{\text{cife}}(x_k) & \text{otherwise} \end{cases} . \tag{16}$$

By introducing $(1 - d_k)$, a quadratic term $d_k d_j$ also appears. To linearize the objective function, we employ an auxiliary optimization variable $y_{j,k} \in \{0, 1\}$ to represent $d_j d_k$, and

with some mathematical manipulation, we have the following constraints,

$$\frac{1}{2}(d_j + d_k) - \frac{1}{2} \leq y_{j,k} \leq \frac{1}{2}(d_j + d_k) \qquad (17)$$
$$y_{j,k} \in \{0, 1\}\,.$$

The functionality of the constraint is the same as $y_{j,k} = d_j d_k$, but in a linear form. Then,

$$(1 - d_k)J_{\text{cife}}(x_k) = I_k(1 - d_k) + \sum_{j=1}^{n} c_{j,k}d_j - \sum_{j=1}^{n} c_{j,k}y_{j,k}\,. \qquad (18)$$

To summarize, under CIFE criterion, the IP formulation of the problem is:

$$\begin{aligned} \min \quad & \sum_{k=1}^{n} ((1 - d_k)J_{\text{cife}}(x_k)) + \lambda \sum_{k=1}^{n} d_k \\ \text{s.t.} \quad & \\ & \tfrac{1}{2}(d_j + d_k) - \tfrac{1}{2} \leq y_{j,k} \leq \tfrac{1}{2}(d_j + d_k) \\ & d_k \in \{0, 1\} \\ & y_{j,k} \in \{0, 1\} \\ & \forall j, k \in \{1, 2, \ldots, n\}\,. \end{aligned}$$

Both the objective function and the constraints have a linear form, which makes the integer program easy to solve.

To obtain the formulation, I_k and $c_{j,k}$ need to be pre-computed in a pre-process procedure. We need to compute mutual information I_k n times. According to the definition of $c_{j,k}$ (Equation 13), we need to compute mutual information $I(x_j; x_k)$ and conditional mutual information $I(x_j; x_k|t)$. Due to the symmetry property of mutual information, the required computation of MI and CMI are both $n^2/2$.

C. Joint mutual information

Joint Mutual Information (JMI) [20] is an alternative criterion for feature selection.

$$J_{\text{jmi}}(x_k) = \sum_{j \in U} I(x_k x_j; t)\,. \qquad (19)$$

After some mathematical derivations, JMI can be re-written as:

$$J_{\text{jmi}}(x_k) = |U|I(x_k; t) - \sum_{j \in U} [I(x_j; x_k) - I(x_j; x_k|t)]\,. \qquad (20)$$

Compared with CMI, the difference of JMI is the term $|U|$.

Similar to the case of CIFE, we do not need the score of a feature that is already selected, and using the same approach to deal with the quadratic term, we have

$$(1 - d_k)J_{\text{jmi}}(x_k) = \sum_{j=1}^{n} (I_k + c_{j,k})(d_j - y_{j,k})\,. \qquad (21)$$

The formulation of feature selection using JMI as a criterion is

$$\begin{aligned} \min \quad & \sum_{k=1}^{n} (1 - d_k)J_{\text{jmi}}(x_k) + \lambda \sum_{k=1}^{n} d_k \\ \text{s.t.} \quad & \\ & \tfrac{1}{2}(d_j + d_k) - \tfrac{1}{2} \leq y_{j,k} \leq \tfrac{1}{2}(d_j + d_k) \\ & d_k \in \{0, 1\} \\ & y_{j,k} \in \{0, 1\} \\ & \forall j, k \in \{1, 2, \ldots, n\} \end{aligned}$$

Since JMI is the summation of mutual information $I(x_k x_j; t)$ (Equation 19), it is always positive. However, as an approximation to conditional mutual information, CIFE is likely to be negative considering its formulation (Equation 11). Generally speaking, CIFE for a feature is smaller than JMI for the same feature, which can also be thought as the effect of the term $|U|$ in JMI's formulation.

D. Generalization

During the past several decades, many criteria have been proposed for feature selection by researchers. One group of them share the same property: linear combinations of Shannon Information terms, in the following form,

$$J = I(x_k; t) - \beta \sum_{j \in U} I(x_j; x_k) + \gamma \sum_{j \in U} I(x_j; x_k|t)\,. \qquad (22)$$

In the last two sections, the criteria we discussed can both be fitted into Equation 22. In the CIFE case, $\beta = \gamma = 1$; in the JMI case, $\beta = \gamma = 1/|U|$.

Besides those, many other criteria can be also derived from Equation 22. For example, Mutual Information Feature Selection criterion (MIFS) [13] with $\gamma = 0$ and $\beta \in [0, 1]$,

$$J_{\text{mifs}}(x_k) = I(x_k; t) - \beta \sum_{j \in U} I(x_j; x_k)\,. \qquad (23)$$

MIM [14] simply sets $\gamma = 0$ and $\beta = 0$.

Max-Relevance min-Redundancy criterion (MRMR) [15] is more complicated with the size of U involved, but similar to JMI,

$$J_{\text{mrmr}}(x_k) = I(x_k; t) - \frac{1}{|U|} \sum_{j \in U} I(x_j; x_k)\,. \qquad (24)$$

Such linear formulations can be easily transformed into the objective function of the integer programming framework. In general, the proposed IP based method works with any criterion that has a form of linear combination of Shannon Information terms.

E. Performance Guarantee

In this section we compare a subset of selected feature U returned by our algorithm, with another subset U' selected by other algorithm, but based on the same criterion. From our general objective function (Equation 8), as long as the two subsets have the same size ($|U| = |U'|$), we have

$$\sum_{k \in V \setminus U'} J(x_k; t|U') \geq \sum_{k \in V \setminus U} J(x_k; t|U)\,. \qquad (25)$$

Our algorithm gives a lower bound of the score of all the unselected features. Since the score is an approximation of the conditional mutual information of unselected features, as in Equation 7, by optimizing

$$\sum_{k \in V \setminus U} J(x_k; t|U)\,. \qquad (26)$$

We optimize the conditional likelihood $\mathcal{L}(d|\mathcal{D})$. In other words, based on the same criterion, the proposed algorithm can find the optimal solution under the Conditional Likelihood Maximization framework.

TABLE I
DATA SETS USED IN THE EXPERIMENTS

data	feature	example	class	ratio
breast	30	569	2	57
congress	16	435	2	72
heart	13	270	2	34
ionosphere	34	351	2	35
krvskp	36	3196	2	799
lungcancer	56	32	3	4
parkinsons	22	195	2	20
sonar	60	208	2	21
soybeansmall	35	47	4	6
spect	22	267	2	67
splice	60	3175	3	265
waveform	40	5000	3	333
wine	13	178	3	12

TABLE II
PERFORMANCE OF IPCIFE, SCORE OF UNSELECTED FEATURES.

data	CIFE	IPCIFE
breast	-1.633	-9.959
congress	-1.688	-5.652
heart	0.971	0.832
ionosphere	13.734	3.446
krvskp	0.463	0.410
lungcancer	30.579	-5.609
parkinsons	1.686	-1.419
sonar	11.253	6.988
soybeansmall	-7.899	-20.285
spect	0.370	0.052
splice	3.477	3.298
waveform	0.665	-1.008
wine	1.251	0.426
average	4.095	-2.191

TABLE III
PERFORMANCE OF IPJMI, SCORE OF UNSELECTED FEATURES

data	CIFE	IPJMI
breast	1.616	1.376
congress	0.607	0.533
heart	0.574	0.539
ionosphere	3.985	3.120
krvskp	0.272	0.264
lungcancer	13.532	7.210
parkinsons	1.308	0.810
sonar	7.372	6.259
soybeansmall	3.283	2.933
spect	0.495	0.430
splice	2.148	2.148
waveform	2.106	2.071
wine	2.013	2.013
average	3.024	2.285

V. EXPERIMENTS

In this section we empirically evaluate the integer programming based method for feature selection. The comparison is between the IP method and greedy methods using the same criteria, in terms of score of unselected features and classification accuracy. We choose 13 data sets from UCI machine learning repository, as detailed in Table I. These data sets have a wide variety of example-feature ratios, which is to measure the difficulty of feature selection. If the data set has N data points, m features and c classes, the ratio for is N/mc. The smaller ratio a data set has, the harder it is for feature selection.

The parameter λ in the objective function controls the weight for the size of the selected features. By tuning the value of λ, our algorithm can return different numbers of selected features. In the experiment, we set $\lambda = 1$. If our IP algorithm returns K features, we select the top K features from CIFE and JMI algorithm, since both algorithm has a rank table of all features.

To empirically demonstrate the bound of the unselected scores, Table II and III shows the score of the unselected features using IP and greedy approaches. It can be seen that based on the same criterion, IP algorithm is never worse than the greedy approaches. This empirically proves our theory in Section IV-E.

For classification, to make fewer assumptions about data and avoid parameter tuning, we use a simple nearest neighbor classifier (3-NN). We also implement two other popular feature selection method: Incremental Association Markov Blanket (IAMB) [17], and max-Relevance min-Redundancy (MRMR) [15]. Table IV demonstrates the performance of integer programming using CIFE criterion (IPCIFE), compared with different methods. In most data sets, using the feature selected by IPCIFE, the recognition accuracy is higher. In terms of average error rate, the IPCIFE outperforms the greedy approach using CIFE criterion, and the other methods. Table V shows the similar result of IP method using JMI criterion (IPJMI). The proposed method does not necessarily have better performance in terms of classification accuracy on every data set, because the process we employ to select features is classifier-independent. Filter methods for feature selection do not guarantee the classification performance, but the average performance of our algorithm is comparable, and sometimes even higher than state-of-the-art methods. Specifically, based on the same criteria (CIFE or JMI), the IP method outperforms the greedy methods, which proves that jointly identifying the features can improve the selected features.

Our IP formulation, by considering the desired set of features at the same time and the relationships among them, is more complex than the greedy algorithms. All the information theoretic greedy heuristic that fits the conditional likelihood maximization framework is $O(N^2)$, where N is the number of features. Generally solving IP problem is NP-hard, but IP solvers hope and generally can successfully and efficiently solve the problem. What's more, if we set a time limit for the solver, it can return a currently best solution.

In practice, we use CPLEX as the solver. Typically the running time is less than 2 minutes. One worst case is for the lungcancer dataset, which takes 16 minutes and returns an approximated result with a relative gap 0.037%. It takes 3684882 MIP simplex iterations, 68983 branch-and-bound nodes, 3510 cover cuts, 7 Gomory cuts and 16 zero-half cuts.

TABLE IV
PERFORMANCE OF IPCIFE COMPARED WITH DIFFERENT METHOD IN TERMS OF CLASSIFICATION ERROR RATE (%)

data	iamb	mrmr	jmi	cife	ipcife
breast	0.130	0.095	0.074	0.196	0.060
congress	0.073	0.087	0.073	0.064	0.078
heart	0.185	0.207	0.207	0.259	0.230
ionosphere	0.097	0.131	0.136	0.148	0.119
krvskp	0.056	0.050	0.056	0.056	0.056
lungcancer	0.500	0.625	0.563	0.563	0.500
parkinsons	0.214	0.204	0.235	0.214	0.224
sonar	0.567	0.519	0.529	0.529	0.567
soybeansmall	0.000	0.250	0.083	0.458	0.042
spect	0.231	0.224	0.246	0.239	0.164
splice	0.130	0.211	0.130	0.152	0.130
waveform	0.239	0.290	0.243	0.280	0.250
wine	0.124	0.067	0.135	0.135	0.056
average	0.231	0.192	0.208	0.253	0.190

TABLE V
PERFORMANCE OF IPJMI COMPARED WITH DIFFERENT METHOD IN TERMS OF CLASSIFICATION ERROR RATE (%)

data	iamb	mrmr	jmi	cife	ipjmi
breast	0.130	0.095	0.074	0.196	0.060
congress	0.073	0.087	0.073	0.064	0.087
heart	0.185	0.207	0.207	0.259	0.222
ionosphere	0.097	0.131	0.136	0.148	0.119
krvskp	0.056	0.050	0.056	0.056	0.056
lungcancer	0.500	0.625	0.563	0.563	0.500
parkinsons	0.214	0.204	0.235	0.214	0.204
sonar	0.567	0.519	0.529	0.529	0.567
soybeansmall	0.000	0.250	0.208	0.458	0.208
spect	0.231	0.224	0.246	0.239	0.164
splice	0.130	0.211	0.130	0.152	0.130
waveform	0.282	0.290	0.294	0.288	0.340
wine	0.056	0.067	0.067	0.090	0.067
average	0.231	0.190	0.217	0.251	0.210

VI. CONCLUSION

In this work, we propose an integer programming formulation for feature selection, under the framework of conditional likelihood maximization. Unlike typical greedy approaches, features are selected jointly. Compared with the greedy approach based on the same criteria, the proposed framework has a lower bound of the score of the unselected features, which guarantees the performance in terms of information scores. The IP problem can be solved efficiently in most cases, and the effectiveness of classification performance is empirically demonstrated.

ACKNOWLEDGMENT

The work described in this paper is supported in part by a grant from the Office of Navy Research (ONR) under the grant number N00014-12-1-0868.

REFERENCES

[1] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. A review of feature selection methods on synthetic data. *Knowledge and information systems*, 34(3):483–519, 2013.

[2] Paul S Bradley and Olvi L Mangasarian. Feature selection via concave minimization and support vector machines. In *ICML*, volume 98, pages 82–90, 1998.

[3] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research*, 13:27–66, 2012.

[4] Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3, 2006.

[5] Ali El Akadi, Abdeljalil El Ouardighi, and Driss Aboutajdine. A powerful feature selection approach based on mutual information. *International Journal of Computer Science and Network Security*, 8(4):116, 2008.

[6] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[7] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[8] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[9] Manabu Ichino and Jack Sklansky. Optimum feature selection by zero-one integer programming. *Systems, Man and Cybernetics, IEEE Transactions on*, (5):737–746, 1984.

[10] Iñaki Inza, Pedro Larrañaga, Ramón Etxeberria, and Basilio Sierra. Feature subset selection by bayesian network-based optimization. *Artificial intelligence*, 123(1):157–184, 2000.

[11] Kenji Kira and Larry A Rendell. A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning*, pages 249–256. Morgan Kaufmann Publishers Inc., 1992.

[12] Igor Kononenko. Estimating attributes: analysis and extensions of relief. In *Machine Learning: ECML-94*, pages 171–182. Springer, 1994.

[13] Nojun Kwak and Chong-Ho Choi. Input feature selection for classification problems. *Neural Networks, IEEE Transactions on*, 13(1):143–159, 2002.

[14] David D. Lewis. Feature selection and feature extraction for text categorization. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, pages 212–217, 1992.

[15] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

[16] Alexander Statnikov, Constantin F Aliferis, Ioannis Tsamardinos, Douglas Hardin, and Shawn Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–643, 2005.

[17] Ioannis Tsamardinos, Constantin F Aliferis, Alexander R Statnikov, and Er Statnikov. Algorithms for large scale markov blanket discovery. In *FLAIRS Conference*, volume 2003, pages 376–381, 2003.

[18] Jason Weston, André Elisseeff, Bernhard Schölkopf, and Mike Tipping. Use of the zero norm with linear models and kernel methods. *The Journal of Machine Learning Research*, 3:1439–1461, 2003.

[19] Ian H Witten, Eibe Frank, and Mark A Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann,San Francisco, 2005.

[20] Howard Hua Yang and John E Moody. Data visualization and feature selection: New algorithms for nongaussian data. In *NIPS*, pages 687–702. Citeseer, 1999.

[21] Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224, 2004.

[22] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.