Controlling costs in feature selection: information theoretic approach

Paweł Teisseyre[1,2][0000−0002−4296−9819] and Tomasz Klonecki[1][0000−0002−0216−3685]

[1] Institute of Computer Science, Polish Academy of Sciences, Poland
[2] Faculty of Mathematics and Information Science, Warsaw University of Technology,
{Pawel.Teisseyre,Tomasz.Klonecki}@ipipan.waw.pl

Abstract. Feature selection in supervised classification is a crucial task in many biomedical applications. Most of the existing approaches assume that all features have the same cost. However, in many medical applications, this assumption may be inappropriate, as the acquisition of the value of some features can be costly. For example, in a medical diagnosis, each diagnostic value extracted by a clinical test is associated with its own cost. Costs can also refer to non-financial aspects, for example, the decision between an invasive exploratory surgery and a simple blood test. In such cases, the goal is to select a subset of features associated with the class variable (e.g., the occurrence of disease) within the assumed user-specified budget. We consider a general information theoretic framework that allows controlling the costs of features. The proposed criterion consists of two components: the first one describes the feature relevance and the second one is a penalty for its cost. We introduce a cost factor that controls the trade-off between these two components. We propose a procedure in which the optimal value of the cost factor is chosen in a data-driven way. The experiments on artificial and real medical datasets indicate that, when the budget is limited, the proposed approach is superior to existing traditional feature selection methods. The proposed framework has been implemented in an open source library[3].

Keywords: cost sensitive feature selection · information theory · mutual information

1 Introduction

Feature selection in supervised classification is a crucial task in many biomedical applications. Feature selection improves the comprehensibility of the considered model and allows to discover the relationship between features and the target variable. Most importantly, it helps to build models with better generalization and larger predictive power [7]. Last years have witnessed a rapid and substantial advancement of feature selection methods coping with the high dimensionality of data. However, most existing methods usually assume that all features have

[3] Python package: https://github.com/kaketo/bcselector

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_37

the same cost, which may be inappropriate as in some situations the acquisition of feature values is costly. For example, in a medical diagnosis, obtaining some information is inexpensive (e.g., gender or age of the patient), but each diagnostic value extracted by a clinical test is associated with its own cost. In general, feature costs may also correspond to non-financial factors, for example time or difficulty in obtaining administrative data (e.g., due to privacy reasons) [18]. Other examples of feature costs include risk associated with certain diagnostic examinations (such as general anesthesia [9], diagnostic X-rays [6]). Finally, the costs may correspond to a choice of diagnostic procedure, e.g., the decision between invasive exploratory surgery and a simple blood test. Ignoring the costs may lead to choosing features that yield a powerful model, but the model cannot be used in practice as high cost is incurred in the prediction [21]. In such cases, it may be better to have a model with an acceptable classification performance, but a much lower cost.

In this work, we focus on a model-free feature selection approach based on the information theory, which has several important advantages. First, it avoids reliance on a particular classification model which allows to find all features associated with the class variable, not only those which are indicated by the employed model. Information theoretic methods, unlike some classical approaches (e.g., logistic regression with lasso regularization), are able to detect both linear and non-linear dependencies between features and class variables. Moreover, some advanced criteria are able to discover interactions between features as well as to take the redundancy of features into account. The information theoretic approach is versatile as it can be used for both classification and regression tasks, i.e., nominal and quantitative class variables, as well as for any type of features. Finally, information theoretic filter methods are usually computationally much faster than their model-based counterparts (such as lasso or random forest variable importance measures). Methods from the latter group require fitting complex classification models, which may be challenging for datasets having a large number of features.

We propose a novel greedy feature selection method that takes into account information on feature costs. In each step of the proposed procedure, we select a feature that maximizes the proposed criterion. Our criterion consists of two components describing the feature relevance and its cost, respectively. The first term is an approximation of the conditional mutual information (CMI) between a candidate feature and a target variable under the condition of already selected features. The approximation of CMI is divided by a second term which is proportional to the cost of the candidate feature. Moreover, we introduce a cost factor that controls the trade-off between feature relevance (measured by CMI) and its cost. We argue that the cost factor plays an important role in cost sensitive feature selection, although it is neglected in most related methods. In particular, its choice should depend on the assumed budget. What distinguishes our idea from previous related methods is a data-driven method of choosing the optimal value of the cost factor.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_37

The paper is organized as follows. We discuss related work in Section 2 and introduce the basic concepts of the information theory in Section 3. In Section 4, we introduce the proposed method and discuss the results of the experiments in Section 5. Section 6 concludes the paper.

2 Related work

Feature selection methods based on the information theory have attracted significant attention in recent years. Various criteria have been proposed ranging from a simple MIM filter [11] (involving the computation of the mutual information between a class variable and a candidate feature) to more powerful methods like CIFE [12], JMI [22] or IIFS [15] that take into account high-order interactions between features as well as possible redundancies between features. We refer to review articles [19] and [3]. In the latter one, the authors analysed dozens of feature selection methods both theoretically and experimentally. Most of the information theoretic methods only produce a ranking of the features and do not select a subset of relevant ones (see however [13]).

In the machine learning literature, there are some attempts to include cost information in the feature selection. The task is challenging as it is necessary to find a trade-off between the feature relevance and its cost. The method most related to our approach was proposed by [2] in which the popular information theoretic filter mRMR was modified by adding a penalty for the feature cost to the term describing the feature relevance. In our contribution, we propose a more general framework in which feature relevance can be measured by any approximation of conditional mutual information. Moreover, the method proposed in [2] lacks the choice of cost factor parameter; our method aims to fill this gap. Another related method has been described in the recent paper [8]. In this approach, the feature relevance is measured using an increase of the Akaike Information Criterion (AIC). The feature relevance term is simply divided by the cost. Unlike in our method, there is no cost factor. Importantly, the method is based on a parametric model whose quality is measured using the AIC, whereas in our approach we consider a more flexible, model-free criterion which is able to detect non-linear dependencies among variables. There are also some attempts to modify existing classification methods in which the feature selection is embedded in the base learner. For example, [23] proposed a random forest-based feature selection algorithm that incorporates the feature cost into the base decision tree construction process. In particular, when constructing a base tree, a feature is randomly selected with a probability inversely proportional to its associated cost. Although the method is appealing, it is not clear how to control the trade-off between the feature relevance and its cost and how to optimize the prediction performance within the assumed budget. Davis et al. [5] present a cost sensitive modification of the ID3 decision tree algorithm. They propose a new cost sensitive feature selection criterion that maximizes the information gain while minimizing the cost. The modification of the lasso method for logistic

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_37

regression was considered in [17]. The authors introduced a regularization term which depends on the feature costs.

3 Background

In this section, we review the basic concepts used in the information theory: the mutual information and the conditional mutual information [4], which are necessary to introduce a general framework of feature selection. First, we discuss some notations. We consider a target class variable Y and a vector of features $\mathbf{X} = (X_1, \ldots, X_p)$, where p is the number of all considered features. In addition, we denote by $\mathbf{X}_S$ a subvector of $\mathbf{X}$ corresponding to a subset of some features $S \subseteq \{1, \ldots, p\}$.

3.1 Mutual Information

The mutual information (MI) is the basic measure of dependence between two variables. MI between the class variable Y and a candidate feature X_k is defined as

$$I(Y, X_k) = H(Y) - H(Y|X_k),$$

where $H(Y)$ is the entropy of the class variable and $H(Y|X_k)$ is the conditional entropy. MI is a popular non-negative measure of association and equals 0 only if Y and X_k are independent. The MI can be also interpreted as the amount of uncertainty in the class variable which is removed by knowing the other variable X_k. In this context, it is often called information gain. In the context of feature selection, the MI is used to assess the individual relevance of the feature X_k, i.e., it measures marginal dependence between Y and X_k. Estimation of the MI is a challenging problem, especially in the case of continuous features [14]. In our experiments, we discretize all continuous features and use a plug-in estimator of the entropy in which the probabilities are estimated by fractions.

3.2 Conditional mutual information and its approximations

The conditional mutual information (CMI) is a crucial concept in the feature selection [3]. Most feature selection methods based on the information theory are forward sequential procedures that start from an empty set of features and, in each step, add a new feature from a set of candidate features. The CMI is used to measure how the candidate feature is associated with the class variable conditioned on the already selected features. The CMI is defined as

$$I(Y, X_k|\mathbf{X}_S) = H(Y|\mathbf{X}_S) - H(Y|X_k, \mathbf{X}_S),$$

where Y is a class variable, X_k is a candidate feature and $\mathbf{X}_S$ is a vector of features corresponding to already selected features. Importantly, it may happen that the candidate feature X_k is associated with the class variable, i.e., $I(Y, X_k) > 0$ but it is redundant when considering together with features S.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_37

The simplest example is the situation when S contains a copy of X_k. Another interesting situation is the case when $I(Y, X_k) = 0$ and $I(Y, X_k|\mathbf{X}_S) > 0$, i.e., there is no marginal effect of X_k, but the interaction between X_k and features from the set S exists. Estimation of the CMI is a very challenging problem, even for a moderate size of conditioning the set S and it becomes practically infeasible for the larger S. To overcome this problem, various approximations of CMI have been proposed, resulting in different feature selection criteria. We refer to [3] and [10] which clarify when various feature selection criteria can be indeed seen as approximations of the CMI. In the following, we briefly review the most popular ones. The simplest approximation is known as MIM (mutual information maximization) criterion defined simply as $I_{mim}(Y, X_k|\mathbf{X}_S) = I(Y, X_k)$, which totally ignores the conditioning set. The other popular method is MIFS (Mutual Information Feature Selection) proposed in [1]

$$I_{mifs}(Y, X_k|\mathbf{X}_S) = I(X_k, Y) - \sum_{j \in S} I(X_j, X_k), \tag{1}$$

in which the first term $I(X_k, Y)$ describes the feature relevance and the second is penalty enforcing low correlations with features already selected in S. Brown et. al. [3] have shown that if the selected features from S are independent and class-conditionally independent given any unselected feature X_k then CMI reduces to so-called CIFE criterion [12]

$$I_{cife}(Y, X_k|\mathbf{X}_S) = I(X_k, Y) + \sum_{j \in S} [I(X_j, X_k|Y) - I(X_j, X_k)]. \tag{2}$$

Note that CIFE criterion is much more powerful than MIM and MIFS as it takes into account possible interactions of order 2 between candidate feature X_k and features selected in the previous steps. There are also criteria that take into account higher-order interactions, see e.g., [15] and [20].

4 Controlling costs in feature selection

4.1 Problem statement

In this section, we describe an information theoretic framework for feature selection. It has to be recalled that Y is a class variable which is predicted using features $X_1, \ldots, X_p$. We assume that there are costs $c_1, \ldots, c_p \in (0, 1]$ associated with features $X_1, \ldots, X_p$. It can be denoted that by $C(S) = \sum_{j \in S} c_j$ a cost associated with a feature subset $S \subseteq \{1, \ldots, p\}$. The total cost is $TC = \sum_{j=1}^{p} c_j$. In addition, the assumption that we have a total admissible budget $B \leq TC$, can be made. The goal is to find a subset of features that allows to predict the class variable accurately within the assumed total budget B. The budget is a user-based parameter that can be manipulated according to current needs. Within an information theoretic framework, the problem can be stated as

$$S_{\text{opt}} = \arg \max_{S: C(S) \leq B} I(Y, \mathbf{X}_S), \tag{3}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_37

i.e., we aim to select a feature subset S_{opt} that maximizes joint mutual information between the class variable Y and the vector $\mathbf{X}_S$ within the assumed budget.

4.2 Greedy forward selection

Note that the number of possible subsets in the problem (3) may grow exponentially, which means that it is possible to solve it only for a small or moderate number of features. Moreover, the estimation of $I(Y, \mathbf{X}_S)$ is a challenging problem when S is large. In this work, we consider a sequential forward search which starts from an empty set of features and in each step adds a feature from a set of candidate features. We first describe the algorithm which finds the optimal feature subset within the assumed budget B, for the fixed value of a cost factor r (see Algorithm 1 for a detailed description). The core element of our algorithm is a cost sensitive criterion of adding a candidate feature. In the i-th step, from a set of candidate features $F_i(r)$ (see Algorithm 1), we select the feature with index $k_i(r)$ such that

$$k_i(r) = \arg \max_{k \in F_i(r)} \frac{I_{\text{approx}}(Y, X_k | \mathbf{X}_{S_{i-1}(r)})}{c_k^r}, \tag{4}$$

where I_{approx} is one of the approximations of CMI (see Section 3.2 for examples), $S_{i-1}(r)$ is a set of features selected in the previous step. Note that criterion (4) can be written in the alternative form $\arg\max_{k \in F_i(r)} [\log I_{\text{approx}}(Y, X_k | \mathbf{X}_{S_{i-1}(r)}) - r \log c_k]$. The first term corresponds to the relevancy of the candidate feature, whereas the second term is a penalty for its cost. We aim to select a candidate feature that maximizes the conditional mutual information with the class variable given already selected features, but at the same time we try to minimize the cost.

Algorithm 1: Finding the optimal subset for the fixed cost factor r

Input : Y,$\mathbf{X} = (X_1, \ldots, X_p)$,$r$,$B$
$S_0(r) = \emptyset$, $F_1(r) = \{1, \ldots, p\}$
$I_{\text{approx-cum}}(r) = 0$
for $i = 1, \ldots, p$ **do**
  $k_i(r) = \arg\max_{k \in F_i(r)} \frac{I_{\text{approx}}(Y, X_k | \mathbf{X}_{S_{i-1}(r)})}{c_k^r}$
  if $C(S_{i-1}(r) \cup k_i(r)) \leq B$ **then**
    $S_i(r) := S_{i-1}(r) \cup k_i(r)$
    $F_{i+1}(r) := F_i(r) \setminus k_i(r)$
    $I_{\text{approx-cum}}(r) = I_{\text{approx-cum}}(r) + I_{\text{approx}}(Y, X_{k_i(r)} | \mathbf{X}_{S_{i-1}(r)})$
    $S(r) = S_i(r)$
  else
    $S(r) = S_{i-1}(r)$
    break for loop
  end
end
Output : $S(r)$, $I_{\text{approx-cum}}(r)$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_37

The cost factor r controls the trade-off between the relevancy of the candidate feature and its cost. Indeed, for $r = 0$, the cost c_k is ignored, whereas for larger r, the cost term plays a more important role. An interesting question arises: how to choose the optimal value of the parameter? In related papers, it is often stated that cost factors should be specified by the user according to his needs, see e.g., [2]. However, in practice, it is not clear how to select the optimal r. We argue that the choice of r should depend on the assumed budget B. Indeed, when B is large, say it is close to a total cost TC, then there is no need to take costs into account, so r should be close to zero. On the other hand, if B is small, then we need to take more into account the costs in order to fit into assumed budget, so r should be large. In order to find the feature subset corresponding to the optimal r, we propose the following procedure, described by the Algorithm 2. We run Algorithm 1 for different values of r, ranging between 0 and certain value $r_{\max}$. For each r we calculate the cumulative increments of the CMI related to the added candidate features. Finally, we choose r_{opt} corresponding to the largest cumulative increment and the feature subset corresponding to r_{opt}. The value $r_{\max}$ is chosen in the following way. Let $I_{\max} := \max_k I(X_k, Y)$ and $I_{\min} := \min_k I(X_k, Y)$ be the maximal and the minimal MIs, respectively. Next, let $c_{(1)} \leq c_{(2)} \leq \ldots \leq c_{(p)}$ be the feature costs sorted in ascending order. For $r = r_{max}$, we should select the cheapest feature regardless of its relevance. In particular, we could potentially have $I_{\max}/(c_{(2)}^{r_{\max}}) \leq I_{\min}/(c_{(1)}^{r_{\max}})$, as the cheapest feature with cost $c_{(1)}$ should be selected regardless of the value of mutual information. Using the above equation, we define $r_{\max} := \log(I_{\max}/I_{\min})/\log(c_{(2)}/c_{(1)})$. The number of values in the grid $0, \ldots, r_{\max}$ depends on the user preferences. For a denser grid, the optimal value of r can be chosen more precisely, but at the same time the computational cost of the procedure increases.

Algorithm 2: Finding the optimal feature subset with cost factor optimization

Input : Y,$\mathbf{X}$,B
for $r = 0, \ldots, r_{\max}$ **do**
  Run Algorithm 1 to obtain $S(r)$ and $I_{\text{approx-cum}}(r)$
end
$r_{\text{opt}} := \arg\max_{r=0,\ldots,r_{\max}} I_{\text{approx-cum}}(r)$
Output : $S(r_{\text{opt}})$

5 Experiments

The main goal of the experiments was to compare the proposed cost sensitive feature selection procedure with traditional feature selection that ignores information about feature costs (we used a standard sequential forward search with CIFE criterion as a representative traditional method). Regarding the proposed cost sensitive approach, we used the greedy procedure described in Algorithm 1 in which the conditional mutual information was approximated with CIFE criterion (2). We used the logistic regression model to calculate the ROC AUC score for the selected set of features. Moreover, the cost factor r was selected

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_37

using the Algorithm 2. We performed experiments on both artificial and real medical datasets. The proposed framework has been implemented in a publicly available Python package https://github.com/kaketo/bcselector.

5.1 Artificial dataset

The advantage of using an artificially generated dataset is that we can easily control the relationship between the feature relevancy and its cost. Below we present a method of generating the artificial data. We consider p original features with a cost equal to 1. The additional features are obtained from the original features by adding noise. The cost of additional features is inversely proportional to the variance of the noise. The above framework mimics a real scenario. For example, in a medical diagnosis we can perform the expensive diagnostic test which yields the accurate value of the feature or alternatively we can choose the cheaper diagnostic test which gives an approximate value of the feature. As an example, one may consider the medical ultrasonography (USG): the 3D scans are more effective and precise than traditional 2D scans, but they are also more expensive; the 2D scan can be regarded as an approximation of the 3D scan.

Generation of artificial data

1. Generate p independent random variables $X_1, \cdots, X_p \sim N(0,1)$ of size n. Let $x_i^{(j)}$ be the i-th value of j-th feature. We set $c_1 = c_2 = \cdots = c_p = 1$.
2. For each observation $i = 1, \ldots, n$, calculate the following term:

$$\sigma_i = \frac{e^{\sum_{j=1}^{p} x_i^{(j)}}}{1 + e^{\sum_{j=1}^{p} x_i^{(j)}}}.$$

3. Generate target variable $Y = \{y_1, \cdots, y_n\}$, where y_i is drawn from the Bernoulli distribution with the success probability σ_i.
4. Generate p noisy random features $e_1, \cdots, e_p$, where $e_j \sim N(0, \sigma)$.
5. Create additional p noisy features, defined as: $X'_j := X_j + e_j$. For each noisy feature we assign cost $c'_j = \frac{1}{\sigma+1}$.
6. Steps 4 - 5 are repeated for different values of σ and finally we obtain $(k+1) \times p$ features, where k is a number of repetitions of steps 4 - 5.

We present the illustrative example for $n = 10000$, $p = 4$ and $k = 4$. This setting yields 20 features in total (4 original and 16 noisy features). Noisy features were generated for four values of σ, randomly selected from $[1, 10]$. Figure 1 (top-left panel) shows the mutual information between considered features and the target variable. It is important to note that the mutual information for noisy features is always lower than for the original features. The left-bottom panel presents the costs of the considered features; note that noisy features have much lower costs than the original features. In the right panel we present the averaged results of 10 trials of feature selection performed for various fractions of the total cost. On OX axis, the budgets are described as a percent of the total cost. On OY axis,

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_37

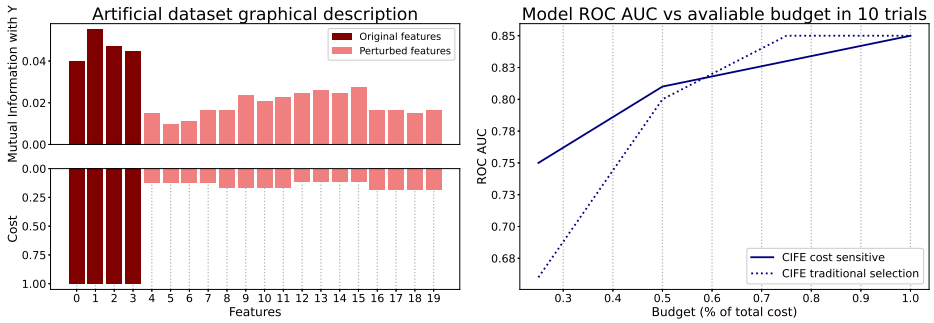

Fig. 1. Feature selection for artificial dataset.

we can see the ROC AUC score of the logistic regression model built on the selected features within the assumed budget. We can see that until 60% of the total cost, cost sensitive method performs better. This is due to the fact that,

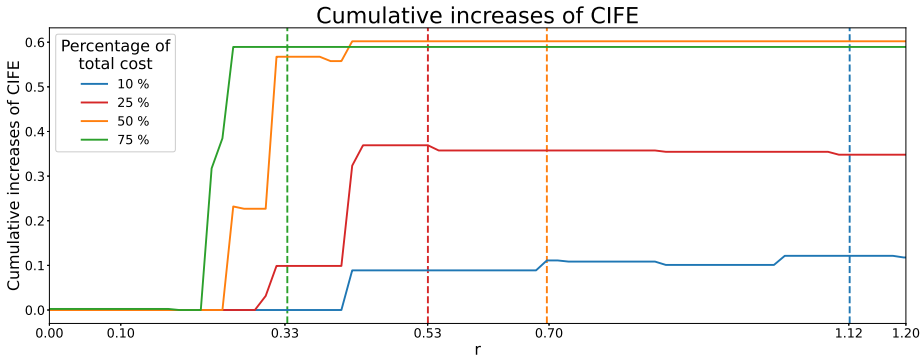

Fig. 2. Artificial dataset. Cumulative increases of CIFE for different values of r and different budgets. Vertical lines correspond to maximum of the curves.

in this case, traditional methods can only use a fraction of all original features (say 1 or 2 out of 4 original features) within the assumed budget, which deteriorates the predictive performance of the corresponding classification model. On the other hand, the cost sensitive method aims to replace the original features by their cheaper counterparts, which allows to achieve higher accuracy of the corresponding model. When the budget exceeds 60% of the total cost, the traditional feature selection method tends to perform better than the proposed method, which is associated with the fact that, in this case, traditional methods include all original features (i.e., those which constitute the minimal set allowing for accurate prediction of the target variable) which results in a large predictive power of the corresponding model. For a larger budget, cost sensitive methods

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_37

include both noisy features as well as the original ones. The noisy features become redundant when considering together with the original ones. This results in slightly lower prediction accuracy of the corresponding model. As expected, the cost sensitive methods are worth considering when the assumed budget is limited.

Figure 2 visualizes the selection of the cost factor r described by the Algorithm 2, for one trail. Vertical dashed lines correspond to the optimal parameter values for different values of the budget.

5.2 MIMIC-II dataset

We performed an experiment on the publicly available medical database MIMIC-II [16] which provides various medical data about patients from the intensive care unit and their diseases. We randomly selected 6500 patients and chose hypertension disease as the target variable. We used 33 variables which refer to basic medical interviews and results of various medical tests. The costs of the features are provided by the experts and they are based on the prices of diagnostics tests in laboratories. We used the cost data described in [17]. Before running the algorithm, the original costs are normalized in such a way that $c_j \in (0, 1]$. It should be noted here that in most countries the relations between the prices of different diagnostic tests are similar.

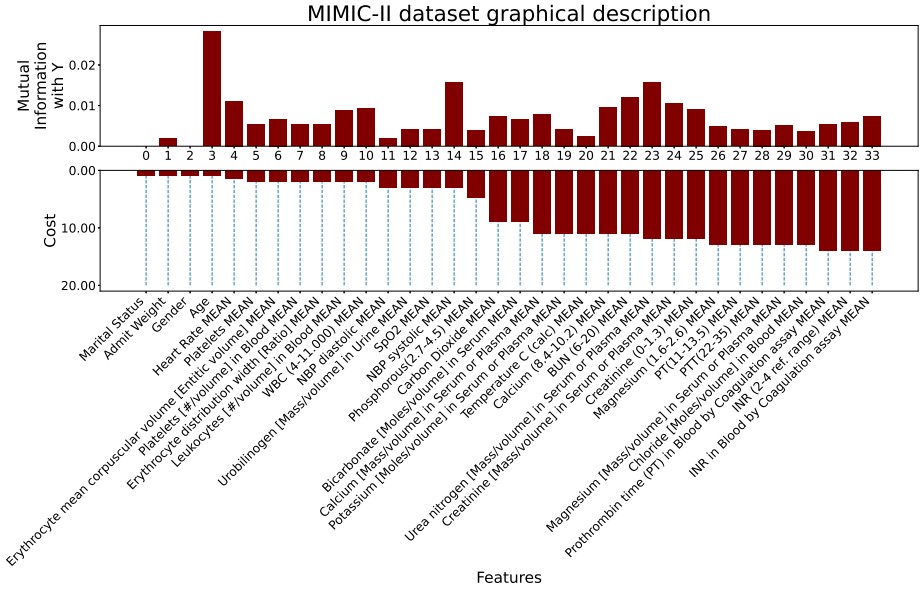

Fig. 3. MIMIC-II dataset. Basic characteristics of features.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_37

Figure 3 depicts the values of mutual information between considered features and the target variable as well as the costs of the features. Features are sorted according to the increasing cost. Values of the first four features (Marital status, Admit weight, Gender and Age), which are based on basic interviews with patients, are really cheap to collect. Note that the variable *Age* is highly correlated with the class variable although it has low cost. Therefore, we can expect that this feature will be selected as relevant by both traditional and cost sensitive methods. Values of the remaining features are possible to obtain using various medical tests. We can distinguish three groups of features: results of blood tests, blood pressure measurements and urine analysis. There are two conspicuous features: *NBP systolic* (number 14) and *urea nitrogen in serum or plasma* (number 23), both of them are moderately correlated with the target variable, but their cost is rather high.

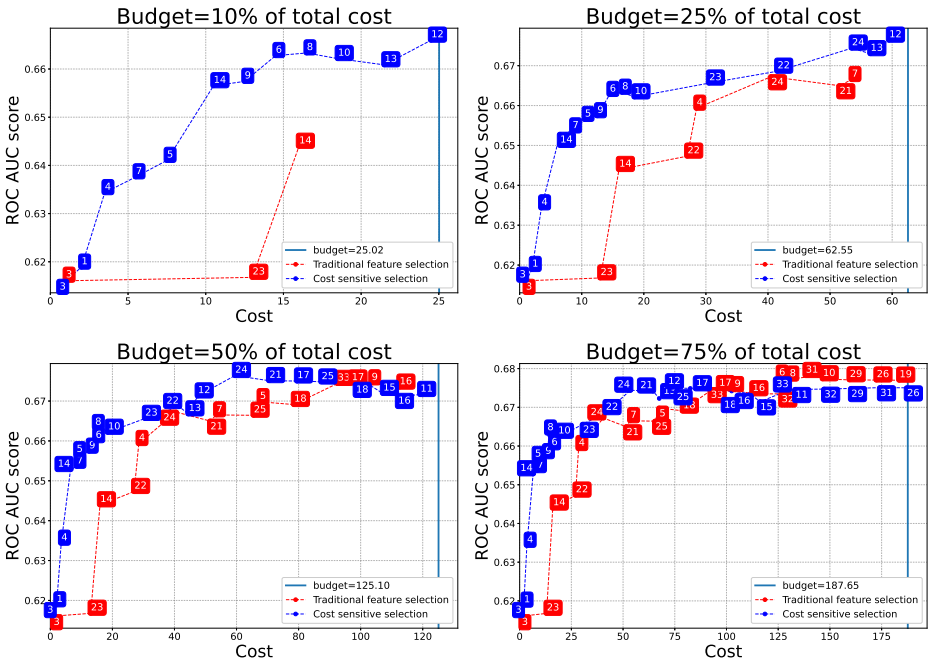

Fig. 4. Feature selection for MIMIC-II dataset.

Figure 4 visualizes the results of feature selection for various budgets for traditional and cost sensitive methods. The figure shows how the ROC AUC depends on the number of features used to train the model. The parameter r is calculated for each budget, therefore the sets of selected features may be different for different values of budget B. Observe that the variable *age* is selected as the most relevant feature in all cases. This can be easily explained as *age* has small

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_37

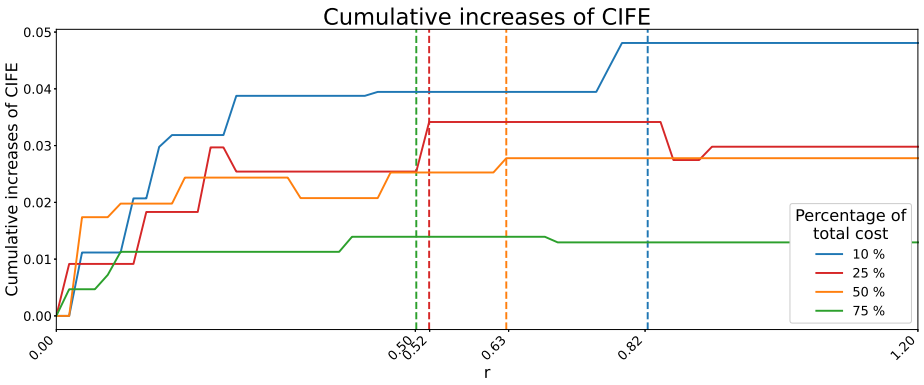

Fig. 5. MIMIC-II dataset. Cumulative increases of CIFE for different values and r and different budgets. Vertical lines correspond to maximum of the curves.

cost and high mutual information with the target variable. The first discrepancy between the methods can be seen in the second step, where the traditional method selects expensive *urea nitrogen in serum* and the cost sensitive method selects *weight* which is really cheap and has a positive value of the MI. In the next steps, the cost sensitive algorithm favors cheap features with moderate value of the MI, which explains why *mean heart rate* or *mean platelet volume in blood* are selected. The most important observation is that the cost sensitive feature selection method achieves higher accuracy when the budget is low. For higher budgets, the traditional methods tend to perform better (see left-bottom and right-bottom panels in the Figure 4). Thus, we observe the similar situation as for the artificial dataset. For a larger budget, traditional methods can include all relevant features, which results in a large predictive power of the model. For a limited budget, cost sensitive methods select features that serve as cheaper substitutes of the relevant expensive features.

Figure 5 visualizes the Algorithm 2. We can observe how the cumulative increments of the approximation of the CMI (CIFE approximation in this case) depend on r for different budgets. The vertical lines correspond to maximum of the curves. As expected, we obtain larger values of r_{opt} for smaller budgets, which is in line with the discussion in Section 4.2. When the budget is large, one should rather focus on the relevancy of the candidate features and not their cost. This explains why r_{opt} is smaller in this case.

6 Conclusions

In this paper, we proposed an information theoretic framework for cost sensitive feature selection. We developed a generic algorithm which allows to use various approximations of the conditional mutual information to assess the relevance of the candidate feature. Moreover, we use the penalty for the cost of the candidate feature. The strength of the penalty is controlled by the cost factor r.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_37

Importantly, we proposed a method of choosing the optimal value of r. The experiments on artificial and real datasets indicate that the proposed cost sensitive method allows to select features that yield a more accurate classification model when restrictions on the budget are imposed. The proposed method can be especially recommended when the assumed budget is low. There are many interesting issues left for future research. In this work, we assumed that each feature has equal extraction cost. However, in many medical applications, features are extracted in groups rather than individually, that is, the feature extraction cost is common for a whole group of features and one pays to simultaneously select all features belonging to such group instead of a single feature at a time. It would be interesting to adapt our method to such a case. Another interesting problem is to consider many target variables (e.g., many diseases) simultaneously, which in the machine learning community is known as a multilabel classification problem. In such cases, it is challenging to approximate the conditional mutual information as instead of a single variable Y we consider a multivariate variable $\mathbf{Y} = (Y_1, \ldots, Y_K)$.

References

1. R Battiti. Using Mutual Information for Selecting Features in Supervised Neural-Net Learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, 1994.
2. V. Bolón-Canedo, I. Porto-Díaz, N. Sánchez-Maroño, and A. Alonso-Betanzos. A framework for cost-based feature selection. *Pattern Recognition*, 47(7):2481–2489", 2014.
3. G. Brown, A. Pocock, M. J. Zhao, and M. Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13(1):27–66, 2012.
4. T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.
5. J. V. Davis, J. Ha, C. J. Rossbach, H. E. Ramadan, and E. Witchel. Cost-sensitive decision tree learning for forensic classification. In *Proceedings of the 17th European Conference on Machine Learning*, ECML'06, pages 622–629. Springer-Verlag, 2006.
6. E. J. Hall and D. J. Brenner. Cancer risks from diagnostic radiology. *The British Journal of Radiology*, 81(965):362–378, 2008.
7. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2009.
8. R. Jagdhuber, M. Lang, A. Stenzl, J. Neuhaus, and J. Rahnenfuhrer. Cost-constrained feature selection in binary classification: adaptations for greedy forward selection and genetic algorithms. *BMC Bioinformatics*, 21(2):307–333, 2020.
9. R. S. Lagasse. Anesthesia safety: Model or myth?: A review of the published literature and analysis of current original data. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 97(6):1609–1617, 2002.
10. M. Lazecka and J. Mielniczuk. Analysis of information-based nonparametric variable selection criteria. *Entropy*, 22(9):974, 2020.
11. D. D. Lewis. Feature selection and feature extraction for text categorization. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, page 212–217. Association for Computational Linguistics, 1992.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_37

12. D. Lin and X. Tang. 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, pages 68–82, 2006.
13. J. Mielniczuk and P. Teisseyre. Stopping rules for mutual information-based feature selection. *Neurocomputing*, 358:255–271, 2019.
14. L. Paninski. Estimation of entropy and mutual information. *Neural Comput.*, 15(6):1191–1253, 2003.
15. M. Pawluk, P. Teisseyre, and J. Mielniczuk. Information-theoretic feature selection using high-order interactions. In *Machine Learning, Optimization, and Data Science*, pages 51–63. Springer International Publishing, 2019.
16. M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L. W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark. Multiparameter intelligent monitoring in intensive care II: A public-access intensive care unit database. *Critical Care Medicine*, 39(5):952–960, 2011.
17. P. Teisseyre, D. Zufferey, and M. Słomka. Cost-sensitive classifier chains: Selecting low-cost features in multi-label classification. *Pattern Recognition*, 86:290–319, 2019.
18. P. D. Turney. Types of cost in inductive concept learning. In *Proceedings of the 17th International Conference on Machine Learning*, ICML'02, pages 1–7, 2002.
19. J. R. Vergara and P. A. Estévez. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24(1):175–186, 2014.
20. N. Vinh, S. Zhou, J. Chan, and J. Bailey. Can high-order dependencies improve mutual information based feature selection? *Pattern Recognition*, 53:46–58, 2016.
21. Z. E. Xu, M. J. Kusner, K. Q. Weinberger, M. Chen, and O. Chapelle. Classifier cascades and trees for minimizing feature evaluation cost. *Journal of Machine Learning Research*, 15(1):2113–2144, 2014.
22. H. H. Yang and J. Moody. Data visualization and feature selection: new algorithms for nongaussian data. *Advances in Neural Information Processing Systems*, 12:687–693, 1999.
23. Q. Zhou, H. Zhou, and T. Li. Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features. *Knowledge-Based Systems*, 95:1 – 11, 2016.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_37