Tomasz Klonecki

Cost-constrained feature selection using information theory

Rozprawa doktorska

Promotor: dr hab. Paweł Teisseyre

Instytut Podstaw Informatyki
Polskiej Akademii Nauk
Warszawa, 2025

Acknowledgements

Firstly, I would like to express my deepest gratitude to my supervisor dr hab. Paweł Teisseyre, for his guidance, encouragement, and patience throughout the years of my doctoral journey. His invaluable advice and support have shaped not only this dissertation but also my approach to research in general.

I am also grateful to prof. dr hab. Jan Mielniczuk for his insightful comments and for always challenging me to think critically and ambitiously.

I would also like to express my heartfelt gratitude to my grandfather prof. dr hab. Witold Klonecki, whose mathematical riddles and puzzles during my childhood first sparked my fascination with problem-solving. His curiosity and joy in exploring ideas planted the seed that eventually grew into my academic path.

Finally, I wish to thank my family for their unconditional love and support. To my parents, for always encouraging me to pursue my passions and to my girlfriend Marta for her love, understanding, and constant belief in me — this work would not have been possible without you.

Contents

Streszczenie

Klasyczne metody selekcji cech skupiają się na maksymalizacji mocy predykcyjnej modelu, pomijając fakt, że pozyskanie wartości zmiennych objaśniających (cech) może być związane z kosztami finansowymi, ryzykiem lub obciążeniem czasowym. W praktyce prowadzi to do powstawania modeli które, choć dokładne, są trudne do wykorzystania w warunkach ograniczonych zasobów (np. finansowych). Problem ten jest szczególnie ważny w medycynie, gdzie zmienne objaśniające odpowiadają wynikom badań diagnostycznych, które mogą być związane ze znacznymi kosztami lub ryzykiem wystąpienia niepożądanych działań u pacjentów. Odpowiedzią na te problemy są metody selekcji cech uwzględniające ich koszty. Metody te pozwalają pogodzić dwa zadania: maksymalizację mocy predykcyjnej modelu oraz minimalizację kosztów pozyskiwania danych.

W pracy opracowano trzy podejścia. Po pierwsze, zaproponowano sekwencyjne metody selekcji oparte na teorii informacji, które wprowadzają, zależne od kosztów zmiennych, kary do kryteriów bazujących na informacji wzajemnej. Opracowane metody bazują na bardzo ogólnej mierze informatywności, opisującej istotność danej zmiennej w kontekście zmiennych włączonych już do modelu. Zaproponowana miara bazuje na dolnym ograniczeniu informacji wzajemnej i może być stosowana w ogólnej sytuacji wielowymiarowej zmiennej odpowiedzi. To obejmuje klasyczny przypadek klasyfikacji binarnej i wieloklasowej, jak również przypadek klasyfikacji wielo-etykietowej. Metoda pozwala uwzględniać interakcje wyższych rzędów, obejmujące różne zmienne objaśniane, ale również interakcje obejmujące różne zmienne odpowiedzi. Po drugie, sformalizowano problem grupowej selekcji cech, w którym koszty przypisane są do zbiorów zmiennych, takich jak grupy parametrów powiązanych z pojedynczym badaniem klinicznym. Po trzecie, rozszerzono podejścia modelowe oparte na minimalizacji ryzyka empirycznego na przypadek metod penalizowanych, w których kary są zależne od kosztów zmiennych.

Skuteczność zaproponowanych metod została zbadana w eksperymentach, przeprowadzonych na danych symulacyjnych oraz na danych rzeczywistych, w tym na bazie danych medycznych MIMIC. Wyniki potwierdzają, że podejścia kosztowe przewyższają klasyczne algorytmy w sytuacji ograniczonego budżetu, zapewniając porównywalną lub wyższą dokładność przy znacznie niższych kosztach.

Rozprawa wnosi istotny wkład w teorię i praktykę uczenia maszynowego uwzględniającego koszty, dostarczając narzędzi przydatnych w medycynie, finansach i innych dziedzinach, gdzie decyzje muszą łączyć skuteczność predykcyjną z ograniczeniami zasobów.

Abstract

Classical feature selection methods focus on maximizing the predictive power of a model, overlooking the fact that obtaining the values of explanatory variables (features) may be associated with financial costs, risks, or time burdens. In practice, this leads to the development of models that, although accurate, are difficult to apply under resource constraints (e.g., financial). This issue is particularly important in medicine, where explanatory variables correspond to the results of diagnostic tests, which can involve significant costs or risks of adverse effects for patients. A response to these problems is the development of cost-constrained feature selection methods. These methods aim to reconcile two objectives: maximizing the predictive power of the model and minimizing the cost of data acquisition.

In this work, three approaches have been developed. First, sequential selection methods based on information theory are proposed, introducing cost-dependent penalties into criteria based on mutual information. The developed methods rely on a very general informativeness measure that describes the relevance of a given variable in the context of variables already included in the model. The proposed measure is based on the lower bound of mutual information and can be applied in the general case of a multivariate response variable. This includes the classical cases of binary and multiclass classification, as well as multilabel classification. The method allows for the inclusion of higher-order interactions involving different features, as well as interactions involving different response variables. Second, the problem of group feature selection is formalized, in which costs are assigned to sets of variables, such as groups of parameters associated with a single clinical test. Third, model-based approaches based on empirical risk minimization are extended to penalized methods, in which penalties are cost-dependent.

The effectiveness of the proposed methods was evaluated in experiments conducted on both simulated and real-world data, including the MIMIC medical database. The results confirm that cost-constrained approaches outperform classical algorithms under budget constraints, providing comparable or higher accuracy at significantly lower costs.

The dissertation makes a significant contribution to the theory and practice of cost-aware machine learning by providing tools that are useful in medicine, finance, and other fields where decisions must combine predictive effectiveness with resource limitations.

Notation

Notation	Description
$T > 0$	Maximum allowable budget
n	Number of observations in training data
n_t	Number of observations in test data
p	Number of features
q	Number of labels (target variables)
$\mathcal{F} = \{1, \ldots, p\}$	Set of indices of the features
$\mathcal{L} = \{1, \ldots, q\}$	Set of indices of the labels
$x = (x_1, \ldots, x_p)$	Feature vector
$y = (y_1, \ldots, y_q)$	Label vector
$x^i = (x_1^i, \ldots, x_p^i)$	Feature vector of i-th observation
$y^i = (y_1^i, \ldots, y_q^i)$	Label vector of i-th observation
$X = (X_1, \ldots, X_p)$	Random variable corresponding to x
$Y = (Y_1, \ldots, Y_q)$	Random variable corresponding to y
X_S	Subvector of X corresponding to subset of indices $S \subseteq \mathcal{F}$
Y_A	Subvector of Y corresponding to subset of indices $A \subseteq \mathcal{L}$
$c_1, \ldots, c_p$	Costs of the features $X_1, \ldots, X_p$
$c(S) = \sum_{j \in S} c_j$	Cost associated with subset of features S
$\mathcal{F} = G_1 \cup G_2 \cup \ldots \cup G_K$	Groups of the features
$c(G_1), \ldots, c(G_K)$	Costs of the groups
$I(Y, X)$	Mutual information between Y and X
$I(Y, X \mid Z)$	Conditional mutual information between Y and X given Z
$II(Y, X, Z)$	Interaction information among Y, X and Z
$l(y^i, \hat{y}^i)$	Loss function measuring the prediction quality of $\hat{y}^i$

Chapter 1

Introduction

1.1 Motivation and description of the problem

Nowadays, machine learning models are used in many fields related to both science and business [125]. Supervised learning algorithms are used in medicine [18], economics [103], finance [3], banking [71], insurance [116] and the broadly understood entertainment industry [2] among others. Most artificial intelligence (AI) tools are based on machine learning models [125]. The effective performance of these models depends to a large extent on the quality of the training data, in particular on the input data. Obtaining relevant information, that can be useful in making decisions by the model, is often associated with costs [A1, 14, 57, 165].

For example, in medicine, models used to predict the occurrence of a disease are based on the results of diagnostic tests [5, 108] or information about genetic mutations [126]. Obtaining such information is associated with costs and these costs are incurred for each patient [111]. Importantly, in medical applications, costs can vary widely, ranging from minimal expenses for basic measurements like blood pressure to substantial costs associated with invasive diagnostic procedures such as endoscopy or biopsy [146]. Additionally, feature costs may encompass non-monetary considerations like the risks associated with certain diagnostic tests [51] or the time required to obtain and process feature values [143]. Another example of feature cost is the data acquisition time, reflecting the duration required to collect or measure specific features, particularly significant in real-time or latency-sensitive scenarios. Another example involves data storage costs, which depend on the volume, frequency, and retention period of collected information, and can become substantial with high-dimensional or continuously streamed data. Additionally, there are monetary expenses linked to obtaining data from external providers, which often charge fees per request or subscription.

Moreover, costs can be associated with entire groups of features, which is a common scenario across various research domains [A3, 106]. In medical diagnostics, such groups often correspond to sets of parameters obtained from a single diagnostic test. By covering the cost of the test, one gains access to all features within that group. For example, a complete blood count (CBC) provides multiple parameters related to blood cell characteristics. Feature groups may also include various statistical measures, such as the mean, median, or standard deviation

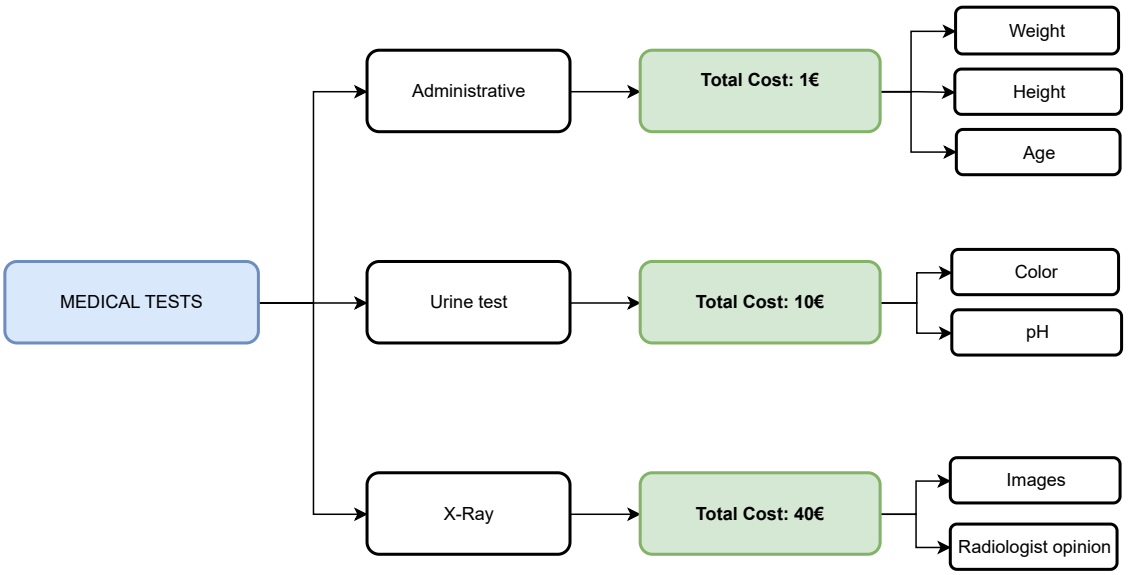


FIGURE 1.1: Cost associated with medical diagnostic tests.

of a single medical parameter monitored over time, such as blood pressure during a patient's hospital stay [168]. Importantly, the cost associated with acquiring different feature groups can vary significantly. For instance, collecting administrative data is typically much less expensive than conducting advanced diagnostic procedures.

Figure 1.1 illustrates a hierarchical grouping of medical examinations into three example categories: Administrative, Urine test, and X-Ray. Each category contains specific diagnostic components and an aggregated cost. This structure highlights the aggregation of individual diagnostic elements into broader procedural categories for cost estimation. The costs of features are associated with the entire group of features, which are obtained from a single diagnostic test. Of course, in simple feature selection problems, we can treat each feature as a separate group. However, in many practical applications, it is more convenient to consider groups of features, which are obtained from a single diagnostic test. The cost values presented in the diagram are based on the data from laboratories and medical centers in Warsaw, Poland. They were converted from PLN to EUR. Administrative costs were assigned a value of 1 EUR, because we assumed it as the minimal cost needed to process the data.

Most popular machine learning algorithms ignore information about the costs associated with obtaining variable values. This can lead to training a model with high predictive power, but the operation of such a model can be problematic from an economic point of view. Making a prediction can generate very high costs [59, 63, 169]. The area of machine learning in which feature costs are taken into account in the model building process is called *cost-sensitive machine learning* or *cost-constrained machine learning* [133]. A natural approach to reducing the cost of making predictions is to use feature selection methods. Feature selection can be performed

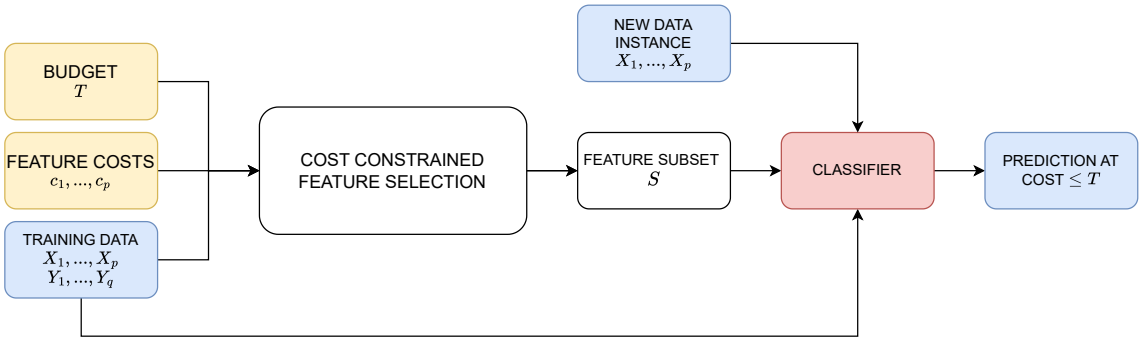


FIGURE 1.2: Flowchart of the model-free, cost-constrained feature selection.

before training the final classification or regression model, as a pre-processing step, in which case we talk about model-free feature selection methods [A2, 14, 57, 90]. Another solution is to take costs into account during the learning of the final model, in which case we talk about model-based methods [31, 165].

The problem of feature selection is a classic problem of machine learning [19, 33, 66, 67, 101]. Since in many practical applications we deal with a very high dimension of the feature vector (e.g. a large number of variables characterizing digital images, texts, patients), feature selection becomes a key task. First, eliminating irrelevant features allows building a model with greater predictive power, because we avoid the effect of overfitting [34, 55, 155]. Second, reducing the number of features allows reducing the computational cost associated with fitting the appropriate model [21, 50, 87]. Third, feature selection methods are important for explaining and discovering the structure of the dependencies between features and labels [13, 74, 124].

However, classical feature selection methods do not take into account information about their costs. This can lead to selecting a subset of features that are useful for prediction, but which generate too high a cost. Instead, very often, it is preferable to employ a feature subset with satisfactory performance that aligns with the specified budget.

In this thesis, we investigate feature selection methods that take into account information about costs. The idea of these methods is to find the most informative subset of features, the cost of which does not exceed the budget given by the user. Of course, the informativeness of a set of features can be measured in many different ways [83, 147]. Figure 1.2 shows a flowchart of model-free cost-constrained feature selection procedure.

In this thesis, we mainly focus on model-free methods based on information theory concepts, such as Mutual Information (MI), Conditional Mutual Information (CMI) and Interaction Information (II) [17]. First of all, model-free methods are usually computationally fast compared to model-based methods. Second, they allow freedom in choosing the final classification model. Third, information-theoretic methods enable the detection of nonlinear dependencies in data,

interactions between variables, and redundancies [17]. In the thesis, for the sake of completeness, we also consider model-based methods based on penalized empirical risk minimization, which are discussed in Chapter 7.

The analysis of cost-constrained methods raises new challenges and interesting research questions. A significant challenge is to find a compromise between the importance of a given feature and its cost [14]. The optimal strategy depends on several factors, such as the size of the budget or the number of relevant variables. For example, in a situation where our budget is very large and the number of relevant features is small, taking into account cost information may not be necessary, and cost-constrained methods will perform very similarly to traditional methods that ignore costs. On the other hand, in the case of a small available budget and a large number of relevant features needed for prediction, the above two groups of methods may perform completely differently. Cost-constrained methods will identify features with slightly lower predictive power but with significantly lower costs. The second challenge is to conduct simulation experiments. Currently, there is a lack of publicly available datasets with costs assigned to individual features. An exception is the medical dataset MIMIC [61] considered in our work, in which the costs of variables were assigned by experts [139]. Thus the need arises to develop procedures in which costs are artificially generated. One of the contributions of this thesis is the proposal of a framework that uses so-called proxy variables. Proxy features are new variables generated from original features. In experiments, we can control the cost and informativeness of proxy variables with respect to the informativeness and costs of the original variables. In this way, we can simulate situations favorable and unfavorable for cost-constrained methods. For example, in a situation where proxy variables are only slightly less informative with respect to the original variables, and the cost of proxy variables is much smaller, cost-constrained methods will tend to choose proxy features and consequently achieve better results when the budget is limited. The third challenge is the situation when costs are assigned to the entire group of features and not individual features. In such a problem, we may be interested in selecting entire groups or features from individual groups. This problem raises new difficulties, such as describing the informativeness of the entire group of features or limiting too many redundant features from a given group. Additionally, detecting interactions between features belonging to one group or features from different groups is an interesting task.

Finally, we mention that in this thesis we focus on classification problems. Importantly, most of the developed feature selection criteria have been designed to work for the case of traditional single-label classification as well as for the multi-label classification situation [12, 32, 91, 162]. Multi-label classification involves predicting multiple binary variables simultaneously. This is an important task in medical applications, where our goal is to predict the occurrence of multiple diseases that may be present in a patient at the same time [43, 112, 137, 168]. Of course, the multi-label classification situation raises interesting challenges for feature selection methods,

such as taking into account interactions between different target variables. We review multi-label classification methods in Chapter 2.

1.2 Contributions of this thesis

1. We propose a novel cost-constrained sequential feature selection method, described in Chapter 5. The method is based on the greedy selection of the most useful features, taking into account the cost information. The novelty of the method consists in two components. First, we use a new score that describes the informativeness of the added feature in the context of the features selected in the previous steps. This score is based on the lower bound of the mutual information. The introduced criterion is a generalization of the popular JMI criterion [8], for the case of multiple labels and the case involving higher-order interactions. The second component is the penalty for the cost of adding a feature. Furthermore, we analyze the impact of the cost factor on the selection of features and propose a method for selecting the optimal value of the cost factor.

2. We formalize and describe the group cost-constrained feature selection (GCC-FS) problem, in which costs are assigned to groups of variables, not to individual variables.

3. We propose three approaches to address GCC-FS problem. The first one is based on sequential selection of individual features, the second one is based on selection of groups of features and the third is a two-step procedure based on the use of so-called shadow features. The advantage of the last method is that it avoids the need to select the cost-factor parameter and allows to reduce the number of redundant variables.

4. Conducting simulation experiments, in which the proposed methods are compared with existing algorithms. The experiments were performed on real and artificial data, using the costs of features assigned by experts and different strategies for generating artificial costs.

1.3 Publications

The results in this thesis were first described in the following publications:

[A1] Paweł Teisseyre, Tomasz Klonecki, Controlling costs in feature selection: information theoretic approach, *Proceedings of the International Conference on Computational Science ICCS*, 2021.

[A2] Tomasz Klonecki, Paweł Teisseyre, Jaesung Lee, Cost-constrained feature selection in multi-label classification using information theoretic approach, *Pattern Recognition*, 2023.

[A3] Tomasz Klonecki, Paweł Teisseyre, Jaesung Lee, Cost-constrained group feature selection using information theory, *Proceedings of the International Conference on Modeling Decisions for Artificial Intelligence MDAI*, 2023.

[A4] Tomasz Klonecki, Paweł Teisseyre, Feature selection under budget constraint in medical applications: analysis of penalized empirical risk minimization methods, *Applied Intelligence*, 2023.

[A5] Tomasz Klonecki, Paweł Teisseyre, Jaesung Lee, Cost-constrained multi-label group feature selection using shadow features, *Proceedings of the 6-th Polish Conference on Artificial Intelligence Conference PP-RAI*, 2025.

Moreover, the proposed methods were used in the analysis of real data, in the context of selecting features useful for the classification of bipolar disorder episodes. These analyses were described in the following paper.

[A6] Olga Kaminska, Tomasz Klonecki, Katarzyna Kaczmarek-Majer, Feature Selection in Bipolar Disorder Episode Classification Using Cost-Constrained Methods, *Proceedings of the Workshop Explainable Artificial Intelligence and Process Mining Applications for Healthcare*, 2024.

1.4 Structure of this thesis

This dissertation is organized into several chapters, each addressing a key aspect of cost-constrained feature selection in supervised learning.

Chapter 2 introduces the fundamental problem of feature selection, outlining its significance in supervised learning tasks. It provides a comprehensive overview of existing feature selection algorithms, distinguishing between model-free and model-based approaches. Additionally, this chapter surveys the literature on feature-cost-aware models, with a particular focus on recent advances in cost-constrained feature selection methods and tree-based algorithms. Finally, we also discuss the multi-label classification problem, which is considered in the following chapters, in the context of cost-sensitive feature selection.

Chapter 3 presents the essential background on information-theoretic concepts that underpin many of the methods discussed in this thesis. Topics such as entropy, mutual information, and their relevance to feature selection are covered to equip the reader with the necessary theoretical tools.

Chapter 4 reviews traditional feature selection methods based on information theory. The chapter formulates the feature selection problem using mutual information and discusses its computational challenges. We provide an overview of popular algorithms for both single-label and multi-label classification. These methods form the theoretical basis for the cost-aware techniques proposed in subsequent chapters.

In Chapter 5, we introduce our proposed cost-constrained sequential feature selection (CSSF) method. This chapter details the algorithmic framework, discusses its theoretical properties, and analyzes its advantages over existing methods in terms of both accuracy and computational efficiency.

Chapter 6 addresses the challenge of group cost-constrained feature selection, where features are organized into groups with associated costs. We propose two novel approaches for group-based selection and analyze their effectiveness in various scenarios.

Chapter 7 explores model-based strategies for cost-constrained feature selection. This includes a discussion of penalized empirical risk minimization techniques and their application to cost-constrained settings, highlighting both theoretical and practical considerations.

Chapter 8 summarizes the main contributions of this dissertation, highlighting sequential, group-based, and penalized approaches to cost-constrained feature selection. It emphasizes their practical significance in domains such as healthcare, acknowledges limitations related to cost annotations, and outlines future directions including complex cost structures, deep learning models and validation on large-scale datasets.

Moreover, chapters 5, 6 and 7 present the results of extensive simulation studies and experiments conducted on real-world datasets. The performance of proposed methods is evaluated and compared against existing approaches, demonstrating their practical utility.

1.5 Notation

In this dissertation, we will use the following notation. The random variables $X = (X_1, \ldots, X_p)$ and $Y = (Y_1, \ldots, Y_q)$ represent the feature vector and the response (target) variable vector, respectively. A feature vector is sometimes called a vector of predictors or attributes or simply a vector of variables. It is useful to operate on index sets, so we introduce the notation $\mathcal{F} = \{1, \ldots, p\}$ for the set of feature variable indices, and $\mathcal{L} = \{1, \ldots, q\}$ for the set of response variable indices, also called labels. Often, by $S \subseteq \mathcal{F}$, we will denote a subset of indices from $\mathcal{F}$, and by X_S, we will denote the subvector of X corresponding to S. In the case of single-label classification, $q = 1$, while in the case of multi-label classification, Y consists of $q > 1$ binary variables. The coordinates of the Y vector may indicate, for example, the occurrence of particular diseases in medical applications. The training set containing n observations is denoted by $\{(x^i, y^i) : i = 1, \ldots, n\}$, where $x^i = (x^i_1, \ldots, x^i_p)$ is the feature vector for the i-th observation, being a realization of the random variable X. Similarly, $y^i = (y^i_1, \ldots, y^i_q)$ is the label vector for the i-th observation, being a realization of the random variable Y. The used notation is summarized in the Table at the beginning of the dissertation.

Chapter 2

Related topics

2.1 Feature selection in supervised learning

Feature selection is an important preprocessing step that involves identifying and selecting a subset of relevant features. By discerning the most informative features, we aim to enhance model performance, reduce computational complexity, and mitigate the effects of overfitting [21, 50, 134]. Feature selection techniques can be broadly categorized into two main groups: model-free methods and model-based methods [83, 86, 130, 145].

Among model-free methods, the simplest are filter approaches. Filter methods [76] assess the relevance of features independently of the learning algorithm used. These techniques typically involve evaluating each feature based on statistical measures, such as correlation, mutual information, or significance tests [9, 30, 80, 110, 124]. Features are ranked or scored according to their individual importance, and a subset of the top-ranked features is selected for further analysis. Filter methods are computationally efficient and can handle high-dimensional datasets effectively. However, the methods that are used in filters might not analyze interactions and dependencies between features. More advanced model-free methods take into account interactions between variables and redundancies. In this group, the most popular methods are those based on concepts taken from information theory [17]. A detailed review of methods based on information theory can be found in Chapter 4.

Model-based methods require training a learning model to select an optimal set of features. This group of methods is often further divided into two subgroups: wrappers and embedded methods [62, 70]. Wrapper methods incorporate the learning algorithm directly into the feature selection process [22, 26, 153]. These techniques utilize a specific model (e.g., decision trees) to evaluate the performance of different feature subsets. The feature selection process becomes an iterative search problem, where various combinations of features are evaluated based on their predictive performance. Wrapper methods often employ strategies like forward selection or backward elimination to identify the optimal subset of features. While wrapper methods can potentially yield better feature subsets tailored to the chosen learning algorithm, they are computationally intensive and may be prone to overfitting, especially with limited data. Embedded

methods integrate feature selection into the model training process itself. These techniques incorporate feature selection as an inherent component of the learning algorithm, ensuring that the selected features are optimized for the model's performance metric. Common examples of embedded methods include regularization techniques like Lasso [42, 153] or its variants [160] and decision tree-based algorithms [65] with built-in feature importance measures. In particular, importance measures based on tree ensemble models have gained much popularity. Examples include importance measures based on Random Forest [16], on the XGBoost algorithm [23], as well as the Boruta algorithms [75] or MCFS [36, 37]. Embedded methods may be limited by the specific characteristics and assumptions of the chosen learning algorithm.

Moreover, in many real-world applications, features naturally form groups based on domain-specific structures or semantic relationships. For example, in genomics, genes can be grouped by biological pathways, while in image analysis, pixels can be clustered by spatial proximity. Traditional feature selection methods, which treat features independently, may overlook such underlying structures, potentially leading to suboptimal performance or interpretability. The concept of selecting entire groups of features has received increasing attention in the literature. Among the most widely studied methodologies is the group lasso and its various extensions [122, 158], which promote group-wise sparsity in linear models. Alternative approaches based on information-theoretic criteria have also been proposed, such as those in [82], which exploit structural dependencies among features. However, these methods typically disregard the costs associated with acquiring feature groups.

2.2 Feature costs in machine learning models

Two principal types of costs are typically distinguished in machine learning: misclassification cost and feature cost. In the case of misclassification cost, it is assumed that different types of prediction error incur unequal penalties, and the learning objective is to minimize the expected total cost. This problem has garnered considerable attention within the machine learning community; a complete survey can be found in [39, 81, 114].

On the other hand, we have feature costs, which refer to the cost associated with acquiring or measuring input features. Feature selection methods that account for such costs are referred to as cost-constrained feature selection methods. These methods aim to identify a subset of features that minimizes acquisition cost while maintaining satisfactory predictive performance. A comprehensive recent review of cost-constrained feature selection algorithms across various applications is given in article [7]. Like traditional selection methods, cost-constrained methods can be roughly divided into two groups: model-free methods and model-based methods.

When it comes to model-free methods, it is worth mentioning the prior work by Bolón-Canedo et al. [14], who extended traditional filter methods such as Correlation-based Feature Selection (CFS) and Minimum Redundancy Maximum Relevance (MRMR) by incorporating a

cost term into their evaluation criteria, thus accounting for feature acquisition costs. The methods described in this thesis can be considered as an extension of this initial approach. In [90], the novel cost-constrained method, called CFSM, was proposed for multi-label classification. In the CFSM, the relevance score for the candidate feature, based on mutual information, is multiplied by the factor depending on the cost of the candidate feature. The method uses neighborhood granularity to transform traditional logical labels into label distribution forms. The details of the method are described in Section 5.4.2. Importantly, the above methods [14, 90] ignore conditional dependencies between the candidate feature and class variable (or class variables), given the already selected features. Hence, the methods cannot handle interactions involving pairs of features and class variable.

Among model-based methods, a number of cost-constrained feature selection strategies have been proposed in the context of regression and classification. For example, ParLiR (Parsimonious Linear Regression), introduced in [49], is an adaptation of the least-angle regression algorithm that integrates feature costs into the selection process. At each iteration, ParLiR computes cost-adjusted correlation scores for all candidate features, where each score is defined as the correlation coefficient between the feature and the current residual vector, penalized by a term proportional to the cost of the feature. As a result, features with higher costs receive lower selection scores. Cost-constrained modification of the popular AIC criterion [1] has been proposed in [57]. The idea is that the increase in the AIC criterion after adding a candidate feature is divided by the cost of the candidate variable. This makes it difficult to select a variable with a very high cost.

An important group of methods is based on the use of tree models. In the classification setting, the authors of [165] proposed a cost-aware feature selection method based on random forests, wherein feature selection during the construction of base decision trees is guided by probabilities inversely proportional to feature costs. Davis et al. [31] presented a cost-constrained adaptation of the decision tree algorithm, introducing a novel feature selection criterion that seeks to maximize information gain while minimizing feature cost. Ji and Carin [59] approached the cost-constrained classification problem through the lens of a partially observable Markov decision process, enabling sequential decision-making under cost constraints. Min et al. [98] redefined the information gain metric to incorporate feature cost, and [97] employed a backtracking algorithm for feature selection under test-time budget constraints.

Another important group are methods that take into account variable costs by using appropriate penalties in the empirical risk minimization scheme. For example, a cost-constrained adaptation of the lasso for logistic regression was introduced by [15], who assigned different penalty factors to distinct data modalities, such as clinical features, gene expression, methylation, and copy number variations. Teisseyre et al. [139] proposed a cost-constrained adaptive lasso combined with classifier chains for multi-label classification. Their approach leverages

the notion that features selected in earlier stages of the chain are more likely to be relevant later, reducing penalties for previously selected features. Modifications to other penalties (such as MCP or SCAD) that take costs into account are considered in Chapter 7. Besides lasso-type penalties, support vector machines have also been adapted to cost-constrained feature selection using $\ell_1 - norm$ formulations [60].

Recently, an intriguing method called *cheap knockoffs* was presented by [156]. It relies on artificially generated knockoff features, mirroring the correlation structure of original features [4, 20]. This approach requires more expensive features to compete against a greater number of knockoffs, thereby introducing stricter selection criteria for costlier variables. Specifically, each feature competes against multiple knockoffs, with costlier features facing more competitors. A feature is considered significant only if it outperforms all its knockoff counterparts, thus ensuring expensive features undergo stricter selection criteria. Additionally, the authors derived an upper bound for the weighted false discovery proportion, representing the fraction of total feature costs expended on irrelevant features.

Grouping features naturally aligns with the concept of feature acquisition costs, as groups often represent jointly measurable features with shared costs. This connection opens promising opportunities for cost-constrained modeling. An important extension of feature selection arises when acquisition costs are associated not with individual features, but with entire groups. In many practical scenarios, groups of features correspond to measurements obtained jointly through a single procedure. For instance, a medical test may provide multiple biomarkers simultaneously, but its cost is incurred only once when the test is performed. In such settings, assigning costs at the group level reflects the true economic or operational constraints more accurately than assigning costs to individual features. This formulation introduces a combinatorial challenge: selecting a subset of groups that yields high predictive performance while minimizing total cost. Most research in cost-constrained feature selection has focused on individual features, and only a limited number of studies have explicitly addressed the case of group-wise cost modeling. An exception is the article [106], where group costs are explicitly incorporated in the context of segmenting backscatter images for product analysis. In their approach, the informativeness of feature groups is assessed using the performance of a classifier on a validation set. While conceptually straightforward, this method entails repeated model training and evaluation, leading to high computational overhead, particularly for large-scale datasets.

2.3 Multi-label classification

The feature selection methods considered in this thesis were designed to work in a very general case, encompassing not only the classical classification problem with a single target variable, but also the case of multi-label classification (MLC) in which many binary target variables are

considered simultaneously [12, 47, 91, 162]. For example, we may be interested in predicting the occurrence of multiple diseases in a single patient based on some patient characteristics such as genetic features, diagnostic test results, age, etc. Co-occurrence of two or more diseases in one patient is referred to as multimorbidity. Multimorbidity is associated with significant reductions in functional status, quality of life and increased risk of death [43, 48] and therefore its prediction is an important task. Applying multi-label methods for multimorbidity prediction has been investigated in several works [168], [139], [112]. Recently, multi-criteria decision-making approach has been explored for cost-constrained feature selection in multi-label settings [100]. Moreover, the problem of multi-morbidity taking into account costs of the individual features is considered in [A2].

The main difference between multiclass and multi-label classification is that in the former case, an observation can be assigned only one label from a set of labels, while in the case of MLC, an object can be assigned multiple labels simultaneously. In the context of disease prediction, multi-class classification we want to predict one disease (out of many) that the patient suffers from, whereas the MLC corresponds to the case when patient can suffer from multiple diseases at the same time.

An interesting task in multi-label classification is modeling the conditional dependencies between labels. In recent years, various classification methods have been proposed, which can be divided into two important groups. The first group is the problem transformation methods. An example is the BR (Binary Relevance) approach [102, 141], in which we build independent models for each of the labels. An improvement of the BR method is the CC (classifier chain) method [32, 88, 118, 119, 120, 136], in which a chain of models is created, in such a way that each successive model predicting the label Y_k uses as input variables all the features X and the labels $Y_1, \ldots, Y_{k-1}$. This group also includes methods based on the Label Powerset (LP) power transformation [141]. The second important group consists of algorithm adaptation methods, which modify existing algorithms for the single-label problem to the MLC situation. Examples include ML-KNN method [161], which is an adaptation of KNN for MLC, multi-label Random Forest [72], multi-label SVM (MLTSVM) [24], multi-label Naive Bayes Classifier [163] or multi-label deep learning neural networks [117]. An exhaustive empirical comparison of many algorithms can be found in [12, 91].

In the multi-label classification problem, feature selection methods were also considered [64, 132]. It is worth noting that variable selection for MLC involves some specific difficulties that do not occur in the case of traditional binary or multi-class classification [138]. The first challenge is the dimensionality of the label vector, which makes calculating even simple measures (such as mutual information) that describe the relationship between the selected feature and the label vector difficult. It is necessary to estimate the multidimensional probability distribution, which becomes more difficult the larger the number of labels considered. Secondly, in

MLC, the analysis of conditional relationships between features and labels is crucial. The studied feature may be independent of the selected label, but conditionally dependent if we consider another or other labels. A task directly related to the analysis of conditional dependencies is the problem of detecting interactions that may involve many variables as well as many labels. All this means that certain selection methods developed for binary classification often cannot be trivially transferred to the case of MLC. The problem of variable selection for MLC becomes interesting and demanding from a scientific point of view.

Among the selection methods for multi-label classification, we can distinguish model-free methods, based mainly on information theory. A more detailed review of these methods can be found in Sections 4.3 and 4.4. The second important group are model-based methods based on regularization. For example, parCC (parsimonious classifier chains) method [137] is based on CC and combining $\ell_{2,1}$ regularization to select features shared across the models and ℓ_1 regularization to select relevant labels in each model in the chain. Adaptive classifier chain (ACC) [139] is yet another modification of the CC method in which penalty factors are changing during fitting the consecutive models in the chain. Finally, elastic net regularization for feature selection combined with CC and BR approaches was used in [135].

Chapter 3

Information theory

Feature selection methods require the definition of measures of dependence between variables. First, our goal is to use general measures that enable the detection of various types of dependence between variables, including nonlinear dependencies. Furthermore, we are interested not only in analyzing the dependence between two variables but also in detecting interactions involving two or more variables. The feature selection methods discussed in the following chapters are sequential in nature; that is, at each step, we select the most informative feature in the context of the features selected in the previous steps. Therefore, we need measures of dependence that allow us to describe the usefulness of a given feature in predicting the target variable, but in combination with the features previously selected. Finally, in sequential methods, it is also important to consider the redundancy of a given feature. A variable may be correlated with the target variable but at the same time it may be useless in the context of the variables already selected.

Given these requirements, it is natural to use methods based on information theory. Developed in the late 1940s [128], Shannon's information theory revolutionized the understanding of communication systems and laid the foundation for modern digital communication, cryptography, data compression, and, most importantly in the context of this thesis, describing the strength of dependencies and interactions between variables.

In this chapter, we define the key concepts of information theory that underlie the feature selection methods discussed in subsequent chapters. In particular, we review fundamental quantities such as entropy, mutual information, conditional mutual information, and interaction information [29]. We present their basic properties and briefly discuss methods for estimating these measures from data.

3.1 Entropy and conditional entropy

One of the most important impacts of Shannon information theory is the notion of entropy, which states that we can establish a measure of uncertainty for any probability distribution.

While our focus in this study is on discrete variables, it is worth noting that analogous definitions can be extended to quantitative variables. This serves as our fundamental definition, upon which all other quantities explored in this thesis are based.

Definition 3.1 (Entropy). Let X be random variable taking values in $\mathcal{X} = \{1, \ldots, K\}$. Moreover, $P(X = x)$ denotes the probability mass function, i.e., probability that X takes value x. The entropy of random variable X is defined as

$$H(X) := -\sum_{x \in \mathcal{X}} P(X = x) \log P(X = x)$$

As entropy increases, so does uncertainty. Consider a random variable X that takes value x_0 with a probability of 1. In this case, the entropy of the variable is 0, indicating that there is no uncertainty in the value of the random variable. Entropy is traditionally calculated using a logarithm base of 2, which means that entropy is measured in bits. For example, a fair coin toss has an entropy of 1 bit. Another common solution is to use the natural logarithm, which we use in this thesis. The choice of logarithm type does not affect the basic properties of entropy. Entropy takes on a maximum value for a uniform distribution, i.e. when all categories have the same probability $P(X = 1) = P(X = 2) = \ldots = P(X = K) = 1/K$. In this case, $H(X) = \log(K)$.

Definition 3.2 (Conditional Entropy). Let X be a random variable taking values in $\mathcal{X} = \{1, \ldots, K\}$ and let Y be random variable taking values in $\mathcal{Y} = \{1, \ldots, L\}$. Moreover, $P(X = x|Y = y)$ denotes the conditional probability that X takes value x under the condition that Y takes value y. Conditional entropy of random variable X given Y is defined as

$$H(X|Y) := -\sum_{y \in \mathcal{Y}} P(Y = y) \sum_{x \in \mathcal{X}} P(X = x|Y = y) \log P(X = x|Y = y).$$

Conditional entropy describes the uncertainty associated with variable X, for $Y = y$, averaged over all possible values of variable Y.

The **chain rule for entropy** states that the joint entropy of p random variables can be decomposed as [29]:

$$H(X_1, \ldots, X_p) = \sum_{j=1}^{p} H(X_j|X_{j-1}, \ldots, X_1) \tag{3.1}$$

Therefore, the conditional entropy for $p = 2$ can be equivalently expressed as:

$$H(X|Y) = H(X, Y) - H(Y) \tag{3.2}$$

3.2 Mutual Information and conditional mutual information

In this section, mutual information and conditional mutual information are introduced. The mutual information $I(Y, X)$ defines the strength of the relationship between the variable X and the target variable Y. Mutual information takes non-negative values, large positive values indicate a strong dependence between variables. It is equal to zero if and only if X and Y are independent. This important property distinguishes mutual information from popular measures like Pearson's linear correlation, which only measures the strength of linear dependence between variables and can be zero even when the variables are dependent. The conditional mutual information $I(X, Y \mid Z)$ measures the conditional dependence between X and Y, given an additional variable Z. It takes non-negative values, large positive values indicate a strong conditional dependence between variables. It is equal to zero if and only if X and Y are conditionally independent given Z.

The theoretical properties of mutual information and conditional mutual information are described in detail in the book [29].

Definition 3.3 (Mutual Information). Let X be random variable taking values in $\mathcal{X} = \{1, \ldots, K\}$ and let Y be random variable taking values in $\mathcal{Y} = \{1, \ldots, L\}$. Moreover, $P(X = x, Y = y)$ denotes the joint probability mass function, i.e., probability that X takes value x and Y takes value y. Finally, $P(X = x)$ is the marginal probability mass function, i.e., probability that X takes value x. The mutual information between the random variables X and Y is defined as

$$I(X, Y) := \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(X = x, Y = y) \log \frac{P(X = x, Y = y)}{P(X = x) P(Y = y)}.$$

Mutual information measures the amount of information that one random variable has about another random variable. It quantifies the relationship between two variables. The higher the mutual information, the stronger the relationship between variables.

Mutual information can be equivalently expressed as:

$$I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X), \tag{3.3}$$

where $H(X)$ and $H(Y)$ are the marginal entropies, whereas $H(X|Y)$ and $H(Y|X)$ are the conditional entropies. In view of (3.3), mutual information describes how much uncertainty of Y is removed when we know X. Below we recall some further basic properties of mutual information, proofs of which can be found in the literature [29]. The following identity

$$I(X, Y) = H(X) + H(Y) - H(X, Y)$$

shows that mutual information is the reduction in joint uncertainty when both variables are observed together. Mutual information can also be interpreted as the Kullback–Leibler (KL) divergence [29] between the joint distribution $P(X,Y)$ and the product of its marginals $P(X)P(Y)$:

$$I(X,Y) = D_{\mathrm{KL}}\left[P(X,Y) \,\|\, P(X)P(Y)\right],$$

where the Kullback–Leibler (KL) divergence between probability distributions P and Q is defined as

$$D_{\mathrm{KL}}\left[P \,\|\, Q\right] = \sum_{x\in\mathcal{X}} P(X=x) \log \frac{P(X=x)}{Q(X=x)}.$$

This formulation emphasizes that mutual information measures the discrepancy between the actual joint distribution and what the distribution would be if X and Y were independent. If (X,Y) follows a bivariate normal distribution with correlation coefficient ρ, then the mutual information is given by:

$$I(X,Y) = -\frac{1}{2}\log(1-\rho^2).$$

The above closed-form expression shows that mutual information increases as the linear dependence between X and Y increases in magnitude.

Definition 3.4 (Conditional Mutual Information). Let X be random variable taking values in $\mathcal{X} = \{1,\ldots,K\}$ and let Y be random variable taking values in $\mathcal{Y} = \{1,\ldots,L\}$. Moreover, let Z be random variable taking values in $\mathcal{Z} = \{1,\ldots,M\}$. Let $P(Z=z)$ be the marginal probability mass function, i.e., probability that Z takes value z. The conditional information between X and Y, given Z is defined as

$$I(X,Y|Z) := \sum_{z\in\mathcal{Z}} P(Z=z) \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P(X=x,Y=y|Z=z) \log \frac{P(X=x,Y=y|Z=z)}{P(X=x|Z=z)P(Y=y|Z=z)}. \quad (3.4)$$

Conditional mutual information measures the conditional dependence between random variables. Note that in the above definition, we calculate the mutual information between variables X and Y, on a fixed layer $Z=z$, and then average the calculated value over all possible values of Z.

Below, we provide an important connection between mutual information and conditional mutual information, which will be used later in the dissertation. The following property [29] (called chain rule for mutual information)

$$I(Y,X|Z) = I(Y,(X,Z)) - I(Y,Z) \quad (3.5)$$

indicates that conditional mutual information between Y and X conditional Z can be written as

the increase in mutual information covering Y and Z related to the addition of the variable X. Finally, we mention that conditional mutual information can be expressed as the averaged (over Z) Kullback–Leibler (KL) divergence between the joint conditional distribution $P(X, Y|Z)$ and the product of its conditional marginals $P(X|Z)P(Y|Z)$:

$$I(X, Y|Z) = E_Z D_{\mathrm{KL}} \left[P(X, Y|Z) \,\|\, P(X|Z)P(Y|Z) \right]$$

The above formula indicates that the conditional mutual information measures the discrepancy between the actual joint conditional distribution and what the distribution would be if X and Y were conditionally independent given Z.

3.3 Interactions and redundancy

Features can be related to labels directly or through non-trivial interactions. One of the advantages of feature selection methods based on information theory is the ability to quantify interactions and redundancies. A useful measure used in information theory to describe interactions is Interaction Information [52, 58, 93].

Definition 3.5 (Interaction Information). Interaction Information (II) among variables $Z_1, \ldots, Z_K$ is defined as

$$II(Z_1, \ldots, Z_K) = -\sum_{S \subseteq \{1,\ldots,K\}} (-1)^{K-|S|} H(Z_S),$$

where Z_S is a subset of $Z = (Z_1, \ldots, Z_K)$ associated with the set S.

For $K = 1$, $II(Z_1) = -H(Z_1)$ equals the negative entropy. For $K = 2$ it simplifies to mutual information $II(Z_1, Z_2) = I(Z_1, Z_2)$ (second-order interaction). An interesting case is the third-order interaction for $K = 3$. The third-order Interaction Information can be written using MI and CMI:

$$II(Z_1, Z_2, Z_3) = I(Z_1, Z_2|Z_3) - I(Z_1, Z_2) \tag{3.6}$$

or equivalently

$$II(Z_1, Z_2, Z_3) = I((Z_1, Z_2), Z_3) - I(Z_1, Z_3) - I(Z_2, Z_3). \tag{3.7}$$

In the above formula, $I((Z_1, Z_2), Z_3)$ describes the dependence between the two-dimensional vector (Z_1, Z_2) and the variable Z_3. It is important that $I((Z_1, Z_2), Z_3) = 0$ implies $I(Z_1, Z_3) = 0$ and $I(Z_2, Z_3) = 0$, whereas the opposite implication is not true. The

third-order Interaction Information is symmetric with respect to all three variables. *Equation* (3.6) indicates that II measures the influence of variable Z_3 on the information between Z_1 and Z_2. On the other hand, *Equation* (3.7) suggests that II can be interpreted as the part of $I((Z_1, Z_2), Z_3)$ remaining after removing the main effects associated with Z_1 and Z_2. This corresponds to the intuitive understanding of interaction as a situation where two variables influence a third in a non-additive manner.

Interestingly, the third-order II (unlike MI) can take positive values and negative values. The positive value of II denotes synergistic interaction, whereas the negative value denotes redundancy. A classic example of synergistic interaction is XOR, i.e., $Z_3 = XOR(Z_1, Z_2) = I(Z_1 \neq Z_2)$,where $Z_1 \in \{0, 1\}$ and $Z_2 \in \{0, 1\}$ are independent and $P(Z_1 = 1) = 0.5$ and $P(Z_2 = 1) = 0.5$. In XOR example

$$II(Z_1, Z_2, Z_3) = I(Z_1, Z_2|Z_3) - I(Z_1, Z_2) = I(Z_1, Z_2|Z_3) > 0,$$

where the second equality follows from the independence of Z_1 and Z_2.

On the other hand, an example of redundancy is when $Z_1 = f(Z_3)$, and $Z_1 = Z_2$, therefore,

$$II(Z_1, Z_2, Z_3) = I(Z_1, Z_2|Z_3) - I(Z_1, Z_2) = -I(Z_1, Z_2) < 0.$$

In the context of the feature selection problem, two types of third-order interactions are of particular interest to us. Let us imagine that Y is a class variable, X_k is a candidate feature, and X_j is a variable selected as relevant in the previous steps. Positive interaction information

$$II(Y, X_k, X_j) = I(Y, X_k|X_j) - I(Y, X_k)$$

describes how including variable X_j strengthens the dependency between Y and X_k. The second interesting case concerns the multi-label situation, where we can consider a pair of binary class variables Y_1, Y_2, referring to the occurrence of two labels, and as before, a candidate variable X_k. The positive II among the three variables

$$II(Y_1, Y_2, X_k) = I(Y_1, X_k|Y_2) - I(Y_1, X_k)$$

indicates that including the second label Y_2 enhances the dependence between Y_1 and X_k.

Detection of interactions is of great practical importance, for example in genetics, interactions between genes are sought for disease prognosis [27, 28, 41]. There is an extensive literature on this topic; the presence of significant interactions between genes has been described in the context of prognosis of diseases such as breast cancer [121], ischemic heart disease [104] or Alzheimer's disease [167].

3.4 Estimating entropy-based terms

Below we briefly discuss the problem of estimating information-theoretic measures [107]. The most popular approach is to use plug-in estimators, which involve replacing the probabilities in the definition of entropy with their empirical counterparts. This method can be applied to discrete variables, whereas for continuous variables, discretization is necessary. Various discretization techniques are described e.g. in [150]. To demonstrate the idea, consider two discrete, univariate random variables, $X \in \{1, \ldots, K\}$ and $Y \in \{1, \ldots, L\}$. The estimators of the probability mass functions, based on sample $\{(x^i, y^i) : i = 1, \ldots, n\}$ of size n, are defined as

$$\widehat{P}(X = x) = \frac{|\{i : x^i = x\}|}{n}, \quad \widehat{P}(X = x, Y = y) = \frac{|\{i : x^i = x, y^i = y\}|}{n},$$

and

$$\widehat{P}(Y = y|X = x) = \frac{\widehat{P}(X = x, Y = y)}{\widehat{P}(X = x)},$$

where $|A|$ denotes cardinality of set A. To avoid zero probabilities caused by the fact that there are no observations belonging to a certain category in the observed data, so-called Laplace smoothing [11] is often used. This involves adding 1 to each count. The estimator after applying Laplace smoothing has the form:

$$\widehat{P}(X = x) = \frac{|\{i : x^i = x\}| + 1}{n + K}.$$

Based on the above estimators of the probability mass function, we define plug-in estimators of entropy, mutual information, and other quantities. For example, the estimators of entropy and conditional entropy have the form

$$\widehat{H}(X) = -\sum_{x \in \{1,\ldots,K\}} \widehat{P}(X = x) \log \widehat{P}(X = x),$$

and

$$\widehat{H}(X|Y) = -\sum_{y \in \{1,\ldots,L\}} \widehat{P}(Y = y) \sum_{x \in \{1,\ldots,K\}} \widehat{P}(X = x|Y = y) \log \widehat{P}(X = x|Y = y),$$

whereas the estimator of the mutual information can be written as

$$\widehat{I}(Y, X) = \widehat{H}(X) - \widehat{H}(X|Y).$$

The above estimation idea can naturally be applied to the case of multidimensional variables X, Y. The estimation task becomes particularly challenging when we are interested in higher-order terms. Important quantities, such as $I(X_S, Y)$ or $I(X, Y \mid X_S)$, with X_S being a sub-vector of X

corresponding to subset of indices $S \subseteq \mathcal{F}$, require the estimation of multidimensional probability distributions. For example, in the case of $I(X, Y)$, it is necessary to estimate the probability $P(X_1, \ldots, X_p, Y)$, which is challenging for larger p. One approach, used by many other authors [17], is to replace these measures with lower-order terms, i.e., terms that depend on probabilities involving at most a few variables. This can be done using various techniques, such as lower bounds on mutual information. Another possibility is to apply mutual information estimation methods that utilize variational inference, such as the MINE (Mutual Information Neural Estimation) algorithm [8]. This method is based on the so-called Donsker-Varadhan representation [35] for mutual information and the use of neural networks to optimize the appropriate risk function. Specifically, the Donsker-Varadhan representation allows us to express mutual information as

$$I(X, Y) = \sup_{T:\Omega \to R} [E_{P(X,Y)} T(X, Y) - \log E_{P(X)P(Y)} e^{T(X,Y)}],$$

where (X, Y) take values in some set Ω. Based on the above formula we have the following lower bound

$$I(X, Y) \geq \sup_{\theta \in \Theta} [E_{P(X,Y)} T_\theta(X, Y) - \log E_{P(X)P(Y)} e^{T_\theta(X,Y)}],$$

where $T_\theta(X, Y)$, $\theta \in \Theta$ is a family of some parametric functions. The idea in the MINE algorithm is to take neural networks as a parametric family and treat the above lower bound as an objective function that we maximize when learning the network. The downside of the second approach is the significant computational cost, which makes it less useful in the context of fast, model-free selection methods.

In the feature selection methods, discussed in this dissertation, we focus on the first approach, that is, the one based on lower-order terms. The details are given in Chapter 5. In addition, in our experiments, continuous variables are discretized. In this way, we avoid the difficult problem of estimating mutual information and conditional mutual information for continuous variables. In the experiments, we use plug-in estimators of mutual information and conditional mutual information described above.

Finally, we make an important remark. In the following sections, when defining various feature selection criteria, for the sake of notation simplicity we will operate on theoretical quantities such as $I(Y, X_k)$ or $I(Y, X_k|X_S)$. In practice, of course, theoretical quantities are replaced by estimators, in the above case $\widehat{I}(Y, X_k)$ and $\widehat{I}(Y, X_k|X_S)$. This convention is used in most works devoted to feature selection using information-theoretic methods.

Chapter 4

Traditional information-theoretic feature selection methods

In this chapter, we discuss feature selection methods grounded in information theory. The feature selection methods based on information theory can be divided into two groups, and the division is related to the task we intend to solve. We will also delve into different issues that arise in the task of feature selection, such as the estimation of information-theoretic measures and combinatorial problems related to the large number of analyzed subsets of variables.

4.1 All relevant feature selection vs. minimal optimal subset selection

In the existing literature on feature selection, two tasks are distinguished: All Relevant Feature Selection (ARFS) and Minimal Optimal Subset Selection (MOSS) [105, 157]. In this section, we will formally define both tasks and explain the differences between them.

The all-relevant feature selection strategy [75, 99, 138] involves identifying all features that are related to the labels. This task is particularly important in biomedical applications, where the main goal is to uncover the structure of dependencies that describe the relationship between the target variable and features. For example, in GWAS (Genome-Wide Association Studies) [144], researchers search for associations between the occurrence of a disease or a certain phenotype and mutations occurring in genes. In this problem, it is crucial to identify all genes associated with the disease. In the case of two genes linked to the disease (e.g., cancer) and strongly correlated with each other, the ARFS strategy will allow both genes to be chosen. Choosing only one of the above two features carries the risk of omitting a mutation that is causally linked to the disease, while it may be "replaced" by a mutation whose correlation with the disease is only spurious [105, 131]. Formally, we use the following definition.

Definition 4.1 (Set of all relevant variables). The set of all relevant variables S^*_{ARFS} consists of variables X_k such that there exists a subset $S \subseteq \mathcal{F} \setminus \{k\}$ such that

$$I(Y, X_k|X_S) > 0.$$

According to the above definition, a variable X_k can be relevant even if it is independent of the label vector, i.e., $I(Y, X_k) = 0$. Such a situation occurs, when X_k is conditionally dependent on Y, given a certain nonempty subset of other variables. This shows that using simple filters based on, for example, mutual information between the variable and the label vector is insufficient to detect all relevant features. Among the relevant variables, we can distinguish the strongly relevant and weakly relevant variables [70]. A variable is strongly relevant if $I(Y, X_k|X_{\mathcal{F}\setminus\{k\}}) > 0$. Strongly relevant variables are essential, containing information that cannot be obtained from other variables. A variable is called weakly relevant if $I(Y, X_k|X_{\mathcal{F}\setminus\{k\}}) = 0$ and there exists a subset $S \subset \mathcal{F} \setminus \{k\}$ such that $I(Y, X_k|X_S) > 0$. Weakly relevant variables also contain information about Y, but this information can also be obtained from other variables. The set of irrelevant variables is defined as follows:

Definition 4.2 (Set of all irrelevant variables). The set of all irrelevant variables consists of variables X_k such that for all subsets $S \subseteq \mathcal{F}$, we have $I(X_k, Y|X_S) = 0$.

Obviously, the set of all irrelevant variables is the complement of the set of all relevant variables. Lemma 4.3 below indicates that the irrelevancy of the feature X_k is related to its independence from the vector of labels Y and the remaining features $X_{\mathcal{F}\setminus\{k\}}$. The Lemma is stated and proved in [138], we show the proof below for completeness.

Lemma 4.3. The following implications hold.

1. Assume that $I(X_k, (Y, X_{\mathcal{F}\setminus\{k\}})) = 0$. Then, X_k is irrelevant.
2. Assume that X_k is irrelevant and $I(X_k, X_{\mathcal{F}\setminus\{k\}}) = 0$. Then, $I(X_k, (Y, X_{\mathcal{F}\setminus\{k\}})) = 0$.

Proof. Observe that $I(X_k, (Y, X_{\mathcal{F}\setminus\{k\}})) = 0$ is equivalent to

$$\underbrace{I(X_k, X_{\mathcal{F}\setminus\{k\}})}_{\text{Term 1}} + \underbrace{I(X_k, Y|X_{\mathcal{F}\setminus\{k\}})}_{\text{Term 2}} = 0, \tag{4.1}$$

which follows from the chain rule for MI (3.5). To prove 1), note that $X_k \perp (Y, X_{\mathcal{F}\setminus\{k\}})$ implies that $X_k \perp (Y, X_S)$ for any $S \subseteq \mathcal{F} \setminus \{k\}$, which implies that $I(X_k, (Y, X_S)) = 0$ and hence, in view of Eq. (4.1), $I(X_k, X_S) + I(X_k, Y|X_S) = 0$. This means that $I(X_k, Y|X_S) = 0$ for any S. Next, to show 2), we have to prove (4.1). Note that in 2), we assumed that $I(X_k, X_{\mathcal{F}\setminus\{k\}}) = 0$

and thus the first term in Eq. (4.1) is equal to 0. The second term in Eq. (4.1) is equal to 0 as X_k is irrelevant. □

The ARFS task involves detecting all relevant variables S^*_{ARFS} or (equivalently) eliminating all irrelevant variables according to Definition 4.2.

A representative example of the ARFS method is the one proposed in [99]. The authors calculate the CMI between the label and candidate feature, given a subset of the remaining features. The final score is obtained by taking the maximum score over the conditioning sets. To account for the different numbers of categories of the features, instead of the CMI, the corresponding p-values were considered. Among the model-based methods, it is worth mentioning the Boruta [75] and MCFS [37] algorithms, which are both based on decision trees. The Boruta algorithm is based on Random Forest feature importance measure [16, 56] and additionally it uses a testing procedure that allows the rejection of irrelevant features. MCFS uses more sophisticated, relevancy measures based on a large collection of decision trees fitted on subset of features and subsets of observations [36, 37]. Its major advantage is that the predictive power of each tree in the ensemble is considered in the measure definition and the interactions between the features can be detected.

The minimal optimal subset selection (MOSS) strategy focuses on identifying the minimal set of features that allows for accurate prediction of the label vector. Usually, this problem is described using the concept of a Markov boundary.

Definition 4.4 (Markov Boundary). The Markov boundary for a label vector Y is defined as the minimal set $S^*_{MOSS} \subseteq \mathcal{F}$ such that $I(Y, X_{\mathcal{F}\setminus S^*_{MOSS}} | X_{S^*_{MOSS}}) = 0$ and for any $S \subset S^*_{MOSS}$ we have $I(Y, X_{\mathcal{F}\setminus S} | X_S) > 0$.

The MOSS approach is used when the goal of analysis is to maximize the predictive power of the model. In this task, the elimination of redundant variables plays an important role, as including too many such variables in the model can reduce its predictive power [54]. The vast majority of existing selection methods are associated with the MOSS task.

The following simple example, discussed in [138], shows that the ARFS and MOSS tasks are not equivalent and moreover that S^*_{MOSS} does not have to be uniquely defined. Consider variables X_1, X_2, X_3 independently generated from any distribution, $X_4 = X_3$, and the remaining variables $X_5, \ldots, X_p$ are generated independently of X_1, X_2, X_3 and each other. We consider a conditional random field (CRF) model in which there are two labels $Y = (Y_1, Y_2)$, and the posterior distribution has the form:

$$P(Y_1 = y_1, Y_2 = y_2 | X_1 = x_1, \ldots, X_p = x_p) \propto \exp(y_1 x_1 + y_2 x_2 + x_3 y_1 y_2). \tag{4.2}$$

The symbol $\propto$ indicates equality up to a normalizing constant. In the above model, variables X_1 and X_2 affect the labels Y_1 and Y_2, while variable X_3 (or X_4) is associated with the interaction

between Y_1 and Y_2. The set of all relevant features $S^*_{ARFS} = \{1, 2, 3, 4\}$ contains 4 variables, while the minimal optimal subset $S^*_{MOSS} = \{1, 2, 3\}$ or $S^*_{MOSS} = \{1, 2, 4\}$ contains 3 variables and is not uniquely determined.

In summary, the choice between the MOSS and ARFS strategies depends on the goal of data modeling. The MOSS strategy is related to building a predictive model, while the ARFS strategy is related to explaining and discovering the full structure of dependencies in the data. The differences between these two approaches are described in the paper [129].

In the context of the problem raised in this thesis, the MOSS approach is much more natural. The ARFS approach may lead to the accumulation of costs related to strongly correlated and, consequently, redundant variables. Therefore, in the following parts we focus on the MOSS approach or its variants.

4.2 Problem statement

Let us recall that we have dataset with p features $X = (X_1, \ldots, X_p)$ and the target variable Y. The target variable may be multidimensional $Y = (Y_1, \ldots, Y_q)$, which is the case in multi-label classification. It is natural to write feature selection problem using mutual information as follows:

$$S^* = \arg\max_{S:|S|\leq K} I(Y, X_S), \tag{4.3}$$

where K is the maximum allowed number of features and X_S is the subvector of X corresponding to the features of the set S. Moreover, S^* is a novel set of predictors that optimize the outcome of mutual information between a set of predictors and the target variable.

In practice, the value of K can be pre-determined by the user or, more commonly, chosen by employing a classification model and selecting such a value of K for which the considered classification quality measure (e.g. accuracy) calculated on the validation set reaches its optimal value. Alternatively, stopping methods based on hypothesis testing are used [96].

The above approach, described by (4.3), produces two primary challenges, initially computing $I(Y, X_S)$, which might be challenging for a substantial number of predictors in S, and secondly, checking all combinations of predictors constitutes an NP-hard problem.

To simplify this task and address the complexity associated with verifying all combinations of predictors, most authors employ a greedy, iterative approach that incrementally adds one variable at a time. The single step of the procedure can be formulated as follows:

$$k_{opt} = \arg\max_k [I(Y, X_{S\cup\{k\}}) - I(Y, X_S)] = \arg\max_k I(Y, X_k|X_S), \tag{4.4}$$

where X_k is a new variable obtained in each step, S is already selected set of features and Y is the target variable. The second equality in (4.4) follows from the chain rule for mutual information (3.5). In every step of this iterative algorithm, we add one variable X_k to the

set S, until it reaches assumed number of features. The above score is usually referred to as $J_{cmi}(Y, X_k|X_S) = I(Y, X_k|X_S)$. Still, one part remains problematic, namely calculating the above terms $I(Y, X_S \cup X_k)$, $I(Y, X_S)$ or $I(Y, X_k|X_S)$ is challenging for a large number of predictors. We would need a proper approximation to solve this. In the next few sections, we will introduce various approaches to estimate it and obtain satisfying results for feature selection.

4.3 Sequential selection methods for single-label classification

Feature selection is a fundamental step in machine learning that improves model performance by reducing dimensionality and eliminating irrelevant or redundant features. In single-label classification, where each instance is assigned a single class label, selecting the most informative features is crucial for building efficient and accurate models. Among various feature selection approaches, sequential selection methods are widely used due to their simplicity and effectiveness in constructing an optimal feature subset. This section presents a detailed discussion of sequential feature selection algorithms. A comprehensive comparison can also be found in the review paper [17].

One approach to address the aforementioned challenges is an intuitive method [80] known as **Mutual Information Maximization (MIM)**. The core concept involves approximating the J_{cmi} criterion (4.4) by simply computing the mutual information for each variable individually:

$$J_{mim}(Y, X_k) = I(Y, X_k). \tag{4.5}$$

In applying this method for feature selection, features are ranked based on their J_{mim} scores, and the top K features are selected.

However, this type of approximation of the J_{cmi} may lack accuracy if interactions between variables exist. Furthermore, the MIM algorithm overlooks dependencies between variables, potentially selecting redundant variables. As an example, consider a dataset containing 3 features (X_1, X_2, X_3) and one target variable Y. The linear correlation coefficients between features and the target variable are presented in *Table* 4.1. The MIM algorithm would prioritize features

Table 4.1: Correlation Matrix for MIM example dataset.

	X_1	X_2	X_3	Y
X_1	1.00	0.98	0.37	0.95
X_2	0.98	1.00	0.36	0.94
X_3	0.37	0.36	1.00	0.39
Y	0.95	0.94	0.39	1.00

in the order X_1, X_2, and X_3. However, a notable issue arises: X_1 and X_2 exhibit a strong linear

correlation with each other, having a Pearson correlation coefficient of 0.98, indicating redundancy in the selected features concerning their information about the target variable Y.

In another method, called **Max-Relevance Min-Redundancy (MRMR)** introduced in [110], we observe the same term $I(X_k, Y)$ as in MIM, which describes the relevance of features. However, it also introduces a penalty term to promote low correlations between selected features. The score in MRMR algorithm is defined as follows:

$$J_{mrmr}(Y, X_k|S) = I(X_k, Y) - \frac{1}{|S|}\sum_{i\in S} I(X_k, X_i). \tag{4.6}$$

As an example, let us assume that we have a dataset with 3 features (X_1, X_2, X_3) and one target variable Y. Mutual information between variables is specified in *Table* 4.2. According to *Equation* (4.6), first iteration of the algorithm would be the same as in MIM, because S is an empty set, therefore MRMR selects X_1 first, as a result of maximizing mutual information between features and target variable. Next feature according to MIM would be X_2 with 1.28 score of mutual information with Y. The problem is that X_2 is strongly dependent with X_1, whereas X_3 is not. In *Table* 4.3, the MIM and MRMR scores are specified for X_2 and X_3, after selecting X_1 as first feature. The MRMR, chooses X_3 in the second step, whereas the MIM chooses X_2. So,

Table 4.2: Mutual Information Matrix for MRMR example dataset.

	X_1	X_2	X_3	Y
X_1	2.25	1.61	0.76	1.60
X_2	1.61	2.25	0.83	1.28
X_3	0.76	0.83	2.61	0.78
Y	1.60	1.28	0.78	2.17

Table 4.3: The MIM and MRMR scores in the second step.

	MIM	MRMR
X_2	1.28	-0.33
X_3	0.78	0.02

in contrast to MIM, the MRMR algorithm handles the redundancy problem within the data.

The next method **Mutual Information Feature Selection (MIFS)**, proposed in [6] extends MRMR algorithm with replacing fixed parameter based on the number of selected features to arbitrary value $\beta \geq 0$. The MIFS is defined as follows:

$$J_{mifs}(Y, X_k|S) = I(X_k, Y) - \beta\sum_{i\in S} I(X_k, X_i).$$

Parameter β can be tuned empirically, higher the value, stronger the emphasis on reducing inter-feature dependencies. It is worth to mention that using $\beta = 0$, would make MIFS equivalent to MIM, selecting features without taking into account dependencies among them.

The **Joint Mutual Information (JMI)** is somehow similar to *MRMR*, but it has one more component - conditional mutual information. Therefore it deals with class relevance-redundancy problem. JMI, proposed in [154] and later in [95], is defined as:

$$J_{jmi}(Y, X_k|S) = \sum_{i \in S} I((X_k, X_i), Y) = \sum_{i \in S} [I(X_i, Y) + I(X_k, Y|X_i)]. \tag{4.7}$$

Using the identity: $I(A, B|C) - I(A, B) = I(A, C|B) - I(A, C)$, we can rewrite J_{jmi} as:

$$J_{jmi}(Y, X_k|S) = \sum_{i \in S} [I(X_i, Y) + I(X_k, Y) - I(X_i, X_k) + I(X_k, X_i|Y)]. \tag{4.8}$$

Note that the term $\sum_{i \in S} I(X_i, Y)$ in *Equation* (4.8) does not depend on X_k argument, and therefore it can be skipped. Thus, the J_{jmi} criterion reduces to:

$$\begin{aligned} J_{jmi}(Y, X_k|S) &= \sum_{i \in S} [I(X_k, Y) - I(X_k, X_i) + I(X_k, X_i|Y)] \\ &= |S| \cdot I(X_k, Y) - \sum_{i \in S} [I(X_k, X_i) - I(X_k, X_i|Y)] \\ &\propto I(X_k, Y) - \frac{1}{|S|} \sum_{i \in S} [I(X_k, X_i) - I(X_k, X_i|Y)] \\ &= I(X_k, Y) + \frac{1}{|S|} \sum_{i \in S} II(Y, X_k, X_i), \end{aligned}$$

where the symbol $\propto$ denotes proportionality (equality up to a constant; this does not affect the choice of the candidate variable). In the final form, J_{jmi} is easy to interpret, because we have 2 main components. The first term $I(X_k, Y)$ describes the dependence between X_k and Y. The second part $\frac{1}{|S|} \sum_{i \in S} [I(X_i, X_k) - I(X_k, X_i|Y)]$ is related to synergistic interactions among features. In addition, the second term is also related to redundancy of the candidate feature X_k. The main difference between JMI and MRMR is the presence of the component $I(X_k, X_i|Y)$.

The last considered criterion, proposed in [84], **Conditional Infomax Feature Extraction (CIFE)** maximizes the following score

$$J_{cife}(Y, X_k|S) = I(X_k, Y) - \beta \sum_{i \in S} [I(X_k, X_i) - I(X_k, X_i|Y)].$$

As we can see, J_{cife} is a generalization of J_{jmi}. Indeed, for $\beta = \frac{1}{|S|}$ we obtain that $J_{cife}(Y, X_k|S) = J_{jmi}(Y, X_k|S)$. We can control the tradeoff between the relevance term and the interaction term using parameter β. Often, the default value is $\beta = 1$ [17].

4.4 Sequential selection methods for multi-label classification

In modern machine learning applications, datasets often involve instances associated with multiple labels rather than a single label. This scenario, known as multi-label classification, arises in various domains, including text categorization, bioinformatics, image annotation, and medical diagnosis. Unlike traditional single-label classification, where each instance belongs to only one category, multi-label classification requires learning complex relationships between multiple labels. In multi-label classification we consider q binary target variables (labels): $Y_1, \ldots, Y_q$.

As the dimensionality of data increases, selecting the most relevant features becomes crucial for improving classification accuracy, reducing computational costs, and enhancing model interpretability. Feature selection is particularly challenging in multi-label settings due to label dependencies, where labels exhibit interdependencies that must be considered during selection.

First, the MIM criterion (4.5), used in single-label classification, can be naturally generalized into the **Multi-label Mutual Information maximization (MLMIM)**:

$$J_{mlmim}(Y, X_k) = \sum_{l=1}^{q} I(Y_l, X_k), \tag{4.9}$$

which can be interpreted as a simple approximation of $I(Y, X_k)$, which ignores interactions among labels. Of course, as with the MIM criterion, the above criterion also ignores interactions between variables and redundancies. The advantage of this approach is its low computational cost and applicability to a large number of labels.

The **approximate mutual information (AMI)** [77] is designed to capture the relationships between features and labels and to take into account the possible redundancy of the candidate feature X_k. The AMI criterion is defined as follows:

$$J_{ami}(Y, X_k|S) = \sum_{l=1}^{q} I(X_k, Y_l) - \sum_{i \in S} I(X_k, X_i) \tag{4.10}$$

Observe that for $q = 1$, AMI reduces to the MIFS criterion with $\beta = 1$. The first term in *Equation* (4.10) describes the dependencies between the candidate characteristic X_k and the labels, while the second can be treated as a penalty for the redundancy of X_k. The latter takes large values when X_k is correlated with already selected features.

The **Multivariate Mutual Information (MVML)** [78] allows detecting two types of interactions: feature-feature interaction and feature-label interaction. This approach enhances feature selection in high-dimensional multi-label datasets, leading to improved classification performance by retaining only the most discriminative and complementary features. The MVML is

defined as follows:

$$J_{mvml}(Y, X_k|S) = \sum_{l=1}^{q} I(X_k, Y_l) + \sum_{i\in S}\sum_{l=1}^{q} II(X_k, X_i, Y_l) + \sum_{l=1}^{q}\sum_{k=1}^{q} II(X_k, Y_l, Y_k)$$

The first term in MVML, as in the remaining methods, describes the dependencies between the candidate feature X_k and the labels. The second and third terms correspond to feature-feature and feature-label interactions, respectively. Importantly, MVML reduces to CIFE with $\beta = 1$, for $q = 1$.

The feature-feature interactions are also considered in **Max-dependency and min-redundancy (MDMR)** [85] method, in which the score has the following form:

$$J_{mdmr}(Y, X_k|S) = \sum_{l=1}^{q} I(X_k, Y_l) - \frac{1}{|S|}\sum_{i\in S} I(X_k, X_i) - \sum_{i\in S}\sum_{l=1}^{q} I(X_k, Y_l|X_i)$$

The feature-feature and feature-label interactions are taken into account in two related methods **Selected Terms of Feature Selection (STFS)** [46] defined as

$$J_{sfts}(Y, X_k|S) = \frac{1}{|\mathcal{L}|}\sum_{l=1}^{q} I(X_k, Y_l) + \sum_{l_i\in\mathcal{L}}\sum_{l_j\in\mathcal{L}-l_i} I(X_k, Y_{l_j}|Y_{l_i}) + \frac{1}{|\mathcal{L}||S|}\sum_{l\in\mathcal{L}}\sum_{i\in S} I(X_k, Y_l|X_i)$$

and **Double Conditional Relevance-Multi-label Feature Selection (DCR-MFS)** [164], which is defined as follows

$$J_{dcr-mfs}(Y, X_k|S) = \sum_{i\in S}\sum_{l=1}^{q} I(X_k, Y_l|X_i) + \sum_{l_i\in\mathcal{L}}\sum_{l_j\in\mathcal{L}-l_i} I(X_k, Y_{l_i}|Y_{l_j}) - \sum_{i\in S} I(X_k, X_i)$$

Finally, the **Scalable Criterion for a Large Label Set (SCLS)** [79] is a feature selection method specifically designed for multi-label classification with a large number of labels. The criterion efficiently evaluates candidate features by balancing relevance and redundancy while ensuring computational scalability.

$$J_{scls}(Y, X_k|S) = \sum_{l=1}^{q} I(X_k, Y_l) - \sum_{i\in S}\frac{I(X_k, X_i)}{H(X_k)}\sum_{l=1}^{q} I(X_k, Y_l)$$

The first term describes the dependencies between the candidate feature X_k and the labels, whereas the second term is related to the redundancy of the candidate feature.

Chapter 5

Cost-constrained feature selection

This chapter discusses the proposed feature selection method, which takes into account information about the costs of individual features. The method is based on the score function derived using the lower bound for mutual information. We consider a general situation covering cases of single and multi-label classification. This chapter is based on the results described in papers [A1, A2] coauthored by me.

5.1 Problem statement

The cost-constrained problem of feature selection can be stated using the information-theoretic framework:

$$S^* = \underset{S:c(S)<T}{\arg\max}\, I(Y, X_S), \tag{5.1}$$

where X_S denotes a subvector of X corresponding to the set $S \subseteq \mathcal{F}$, $T > 0$ represents a user-specified maximum admissible budget, and $c(S) = \sum_{j \in S} c_j$ indicates the cost associated with the subset S. The costs $c_1, \ldots, c_p \geq 0$ are assumed to be given. Moreover, we assume that the costs are normalized (i.e. $0 \leq c_i \leq 1$). If the original costs provided by the user are not in the range $[0, 1]$, we normalize them (i.e., $c_j \leftarrow c_j / \max_j c_j$). The budget range is $T \in [\min_{1 \leq j \leq p} c_j, c(\mathcal{F})]$. Setting $T = \min_{1 \leq j \leq p} c_j$ means that we can select only the cheapest feature, whereas $T = c(\mathcal{F})$ indicates that all features can be included in the model.

Solving problem (5.1) is challenging because the estimation of $I(Y, X_S)$ is demanding due to the dimensionality of X_S and also due to the dimensionality of Y in a multi-label situation when Y is a vector consisting of q coordinates. Many methods that approximate $I(Y, X_S)$ using low-dimensional terms have been proposed in recent years [73, 78, 149].

The proposed method [A2] is also based on this idea and uses the lower bound for $I(Y, X_S)$. In order to derive the formula for the lower bound, we first recall a simple characteristic of MI in Propositions 1 and 2.

Proposition 1. We let A and S be subsets of feature indices, such that $A \subseteq S \subseteq \mathcal{F}$ and X_A, X_S are subvectors of X, corresponding to A and S, respectively. Then, we have the following property:

$$I(X_S, Y) \geq I(X_A, Y), \tag{5.2}$$

where Y is a vector of labels.

Proof. We can write

$$\begin{aligned} &I(X_S, Y) - I(X_A, Y) \\ =&H(Y) - H(Y|X_S) - (H(Y) - H(Y|X_A)) = H(Y|X_A) - H(Y|X_S). \end{aligned}$$

Theorem 2.6.5 in [29] (conditioning reduces entropy; information cannot hurt) states that for $A \subseteq S$, we have $H(Y|X_S) \leq H(Y|X_A)$ as conditioning on more variables always reduces entropy. Thus, the inequality in (5.2) is true. □

Proposition 2. We let A be a subset of feature indices such that $A \subseteq \mathcal{F}$ and X_A is a subvector of X, corresponding to A. Moreover, we let B be a subset of indices, such that $B \subseteq \mathcal{L}$ and Y_B is a subvector of label vector Y, corresponding to B. Then, we have the following property:

$$I(X_A, Y) \geq I(X_A, Y_B), \tag{5.3}$$

Proof. The proof is analogous to the proof of Proposition 1. □

Based on Propositions 1 and 2, mutual information $I(Y, X_S)$ can be bounded from below by terms that are easier to estimate, leading to Theorem 1.

Theorem 1. We let $A \subseteq S$, where $|A| = a$, and $B \subseteq \mathcal{L}$, where $|B| = b$. Then we have

$$I(Y, X_S) \geq \frac{1}{\binom{|S|}{a}\binom{q}{b}} \sum_{\substack{A \subseteq S: \\ |A|=a}} \sum_{\substack{B \subseteq \mathcal{L}: \\ |B|=b}} I(Y_B, X_A). \tag{5.4}$$

Proof. By Proposition 1, $I(Y, X_A) \leq I(Y, X_S)$ for any subset $A \subseteq S$ because adding new variables to set A only increases the value of MI. Note that there are $\binom{|S|}{a}$ of such subsets, where a is the number of features in A. Averaging over all subsets A results in

$$I(Y, X_S) \geq \frac{1}{\binom{|S|}{a}} \sum_{\substack{A \subseteq S: \\ |A|=a}} I(Y, X_A), \tag{5.5}$$

as the average of the terms that are not greater than $I(Y, X_S)$ is also not greater than $I(Y, X_S)$.

Similarly, it follows from Proposition 2, that for fixed subset of features A, we have $I(Y_B, X_A) \leq I(Y, X_A)$, for any subset of labels $B \subseteq \mathcal{L}$, where $\mathcal{L} = \{1, \ldots, q\}$. Note that there are $\binom{q}{b}$ of such subsets, where b is a number of labels in B. Averaging over all $\binom{q}{b}$ subsets of labels B results in

$$I(Y, X_A) \geq \frac{1}{\binom{q}{b}} \sum_{\substack{B \subseteq \mathcal{L}: \\ |B|=b}} I(Y_B, X_A), \tag{5.6}$$

as the average of the terms that are not greater than $I(Y, X_A)$ is also not greater than $I(Y, X_A)$.

Combining the inequalities from (5.5) and (5.6) results in the lower bound of the joint MI:

$$\begin{aligned} I(Y, X_S) \geq & \frac{1}{\binom{|S|}{a}} \sum_{\substack{A \subseteq S: \\ |A|=a}} I(Y, X_A) \geq \frac{1}{\binom{|S|}{a}} \sum_{\substack{A \subseteq S: \\ |A|=a}} \frac{1}{\binom{q}{b}} \sum_{\substack{B \subseteq \mathcal{L}: \\ |B|=b}} I(Y_B, X_A) \\ = & \frac{1}{\binom{|S|}{a}} \frac{1}{\binom{q}{b}} \sum_{\substack{A \subseteq S: \\ |A|=a}} \sum_{\substack{B \subseteq \mathcal{L}: \\ |B|=b}} I(Y_B, X_A). \end{aligned}$$

□

Note that a and b can be treated as parameters controlling the number of features and labels included in the lower bound. Importantly, for larger values of a and b, the approximation is more accurate, but the estimation of MI terms becomes more challenging. Define:

$$I_{a,b}(Y, X_S) := \sum_{\substack{A \subseteq S: \\ |A|=a}} \sum_{\substack{B \subseteq \mathcal{L}: \\ |B|=b}} I(Y_B, X_A),$$

which is proportional to the lower bound of $I(Y, X_S)$, according to Theorem 1. The main idea of our approach is to use $I_{a,b}(Y, X_S)$ instead of $I(Y, X_S)$ in (5.1), which leads to a simpler optimization problem of the form:

$$S^* = \underset{S \subseteq \mathcal{F}: c(S) \leq T}{\arg\max} \; I_{a,b}(Y, X_S). \tag{5.7}$$

The dual optimization problem to (5.7) can be written as

$$S^* = \underset{S \subseteq \mathcal{F}}{\arg\max} \; \left[I_{a,b}(Y, X_S) - \lambda c(S) \right], \tag{5.8}$$

where parameter $\lambda > 0$ corresponds to budget T. Hereinafter, parameter λ is called cost-factor. It controls the cost penalty associated with the selected variables.

5.2 Cost-constrained sequential feature selection

Even though estimation of $I_{a,b}(Y, X_s)$ is easier than $I(Y, X_s)$ (provided that parameters a and b are small enough and/or the sample size is large), the problem (5.8) is still NP-hard and thus its direct solving is unfeasible because one needs to check 2^p feature subsets in the worst-case scenario. To remedy this, we use an iterative greedy sequential procedure, which starts from an empty set of features and in each step adds candidate feature k_{opt} that maximizes the increment of the objective function $I_{a,b}(Y, X_S)$. First, observe that the increment can be written as

$$J_{a,b}(Y, S, k) := I_{a,b}(Y, X_{S\cup\{k\}}) - I_{a,b}(Y, X_S) = \sum_{\substack{B\subseteq\mathcal{L}:\\ |B|=b}} \sum_{\substack{A\subseteq S:\\ |A|=a-1}} I(X_{A\cup\{k\}}, Y_B). \tag{5.9}$$

Then we select feature k_{opt} such that

$$\begin{aligned} k_{opt} =& \arg\max_{k\in\mathcal{F}\setminus S} \left[I_{a,b}(Y, X_{S\cup\{k\}}) - I_{a,b}(Y, X_S) - \lambda c(S\cup\{k\}) + \lambda c(S)\right] \\ =& \arg\max_{k\in\mathcal{F}\setminus S} \left[J_{a,b}(Y, S, k) - \lambda c_k\right]. \end{aligned} \tag{5.10}$$

In the following, we consider special cases of the general score function (5.10). Considering first the single label case $q = 1$ with $a = 2$ leads to

$$\begin{aligned} k_{opt} &= \arg\max_{k\in\mathcal{F}\setminus S} \left[\sum_{i\in S} I(Y, (X_k, X_i)) - \lambda c_k\right] \\ &= \arg\max_{k\in\mathcal{F}\setminus S} \left[\sum_{i\in S} I(Y, X_k|X_i) - \lambda c_k\right], \end{aligned} \tag{5.11}$$

where the last equality follows from (3.5), which states that

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

and the fact that $I(Y, X_i)$ does not depend on k. For $\lambda = 0$, criterion (5.11) reduces to JMI criterion [154]. Note that the above criterion considers interactions involving 3 variables: Y, X_k, X_i. Similarly, for $a = 3$ we obtain a score

$$k_{opt} = \arg\max_{k\in\mathcal{F}\setminus S} \left[\sum_{\{j_1,j_2\}\in S} I(Y, X_k|X_{j_1}, X_{j_2}) - \lambda c_k\right]$$

in which we condition on two variables, so we can include interactions involving 4 variables: Y, X_k, X_{j_1}, X_{j_2}.

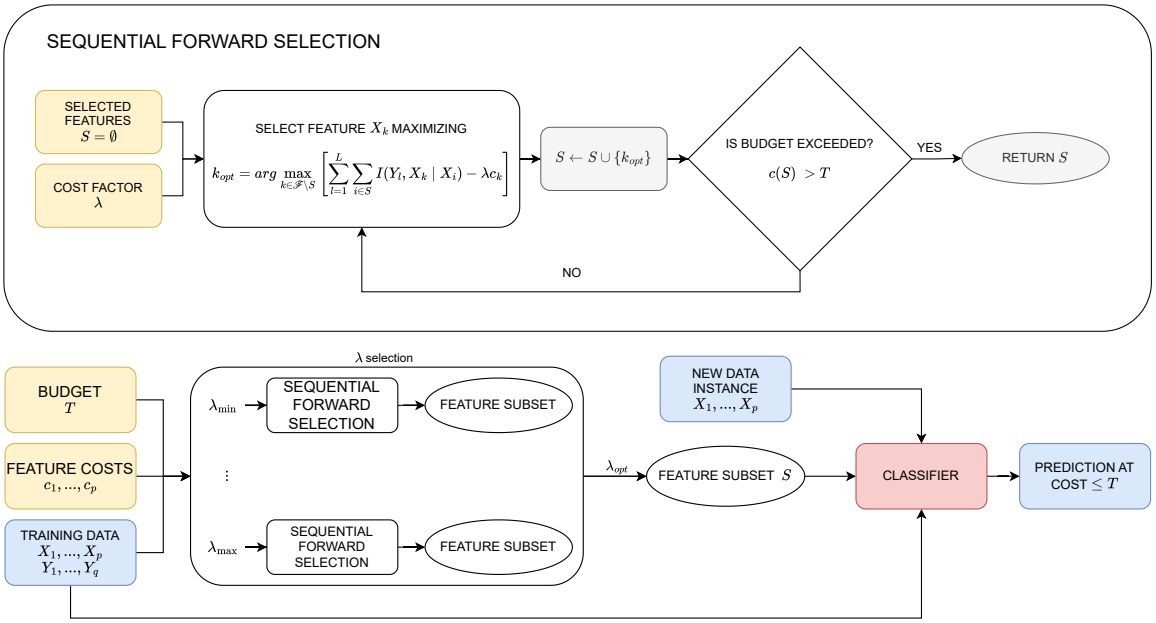


FIGURE 5.1: Flowchart of the cost-constrained sequential feature selection method with λ optimization. The costs of the features are denoted as $c_1, \ldots, c_p$.

We now consider the multi-label case $q > 1$. For $b = 1$ and $a = 2$, we obtain

$$k_{opt} = \arg\max_{k \in \mathcal{F} \setminus S} \left[\sum_{l=1}^{L} \sum_{i \in S} I(Y_l, X_k | X_i) - \lambda c_k \right], \tag{5.12}$$

which for $\lambda = 0$ is a simple generalization of the JMI criterion to the multi-label case. Such generalization has been considered in [127]. The second interesting case is obtained for $b = 2$ and $a = 2$

$$k_{opt} = \arg\max_{k \in \mathcal{F} \setminus S} \left[\sum_{\{l_1, l_2\} \in \mathcal{L}} \sum_{i \in S} I((Y_{l_1}, Y_{l_2}), X_k | X_i) - \lambda c_k \right],$$

The above criterion takes into account interactions of order 4, including variables $Y_{l_1}, Y_{l_2}, X_k, X_i$. As can be seen from the above examples, by choosing parameters a and b we can obtain different criteria. In general, the larger the values of a and b, the more flexible our criterion is, in the sense that we take into account higher-order interactions. On the other hand, for larger a and b, the estimation of MI-based terms becomes more demanding. Therefore, in practice, a compromise must be found between these issues. It depends on the amount of data and the difficulty of the variable selection problem, e.g. the presence of high-order interactions.

Figure 5.1 illustrates the overall flow of the proposed method. The top panel depicts the sequential forward selection process, where features that maximize the score function are added

one by one until the total cost exceeds the budget T. The bottom panel shows how this selection process is repeated for different values of the cost parameter λ. Among the resulting subsets, the one corresponding to the optimal λ is chosen. The method of selection is described in details in Section 5.3. The final subset can then be passed to any external classifier, which uses it to make predictions on new data, ensuring that the cost of prediction stays within the budget T. The proposed feature selection framework is model-agnostic and can be combined with any classification algorithm.

5.3 Choice of the cost factor

Cost factor $\lambda \geq 0$ in criterion (5.10) controls the trade-off between the relevance of the candidate feature X_k and its cost. The second term in (5.10) can be treated as a penalty for the feature costs. The cost is ignored for $\lambda = 0$, while the cost term plays a critical role for larger λ.

Determining the optimal value of λ is a challenging issue because it should depend on the budget T. When T is close to the total cost $\sum_{j=1}^{p} c_j$, there is no need to consider the costs, so λ can be close to zero. However, if T is small, we must consider the costs to fit a budget, so λ should be large. Moreover, the optimal λ depends on the number of relevant features needed to predict the target labels accurately. For example, when the number of relevant features is small, such that they fit the budget T, we can choose a small value of λ, even when T is small. A possible method of selecting the optimal λ is to use a validation set on which some evaluation measure (e.g., accuracy, F measure, or ROC AUC) is calculated for various feature subsets corresponding to the values of λ. Although this can be a straightforward solution, it creates three problematic issues. First, an external classifier is required, which can be computationally expensive to run for many values of λ. Secondly, additional validation dataset is needed, which can be problematic in some applications, where the total number of observations is limited. Finally, one has to choose a particular evaluation measure.

An alternative approach based on the grid search technique, which avoids such issues, has been proposed in papers [A1, A2]. In the method, we search lambdas distributed on the grid $\lambda \in [0, \lambda_{\text{max}}]$ and try to find the optimal lambda that optimizes a certain score. The first important issue is how to determine λ_{max}, which is defined as a value of λ for which only the cost determines the order of features in the criterion (5.10). The simplest approach works as follows. Consider the first step in the greedy procedure and two arbitrary features X_s and X_r, with costs c_s and c_r, such that $c_s > c_r$. We are interested in choosing λ_{max} for which the cheaper variable X_r will be chosen regardless of the value of the relevance term $I(Y, X_r)$, i.e., $I(X_s, Y) - \lambda_{\text{max}} c_s < I(X_r, Y) - \lambda_{\text{max}} c_r$. Therefore, λ_{max} should satisfy

$$\frac{I(X_s, Y) - I(X_r, Y)}{c_s - c_r} < \lambda_{\text{max}}.$$

We can determine the possible value of λ_{max} using the following simple upper bound

$$\frac{I(X_s,Y)-I(X_r,Y)}{c_s-c_r} \leq \frac{\max_s I(X_s,Y)}{c_s-c_r} \leq \frac{\max_s I(X_s,Y)}{\min\{|c_i-c_j| : i \neq j, c_i \neq c_j\}},$$

where the second inequality follows from $c_s - c_r \geq \min\{|c_i - c_j| : i \neq j, c_i \neq c_j\}$. This leads to

$$\lambda_{\text{max}} = \frac{\max_s I(X_s,Y)}{\min\{|c_i-c_j| : i \neq j, c_i \neq c_j\}},$$

which is easy to calculate and does not involve any additional computational cost, since we need to determine the values of $I(X_s, Y)$ anyway to perform the first step of the greedy procedure.

Now we can describe how we choose the optimal lambda value. We run the greedy procedure for a given $\lambda \in [0, \lambda_{\text{max}}]$. The algorithm selects subsequent variables, maximizing the (5.10) criterion at each step. We continue adding variables until the cost of the selected variables exceeds the budget T. Let us denote by $S(\lambda)$ the variables selected using such a procedure for a given λ. The optimal lambda is chosen as

$$\lambda_{\text{opt}} = \underset{\lambda \in [0,\lambda_{\text{max}}]}{\arg\max} \sum_{k \in S(\lambda)} I(X_k, Y).$$

In the multi-label case we use an analogous score

$$\lambda_{\text{opt}} = \underset{\lambda \in [0,\lambda_{\text{max}}]}{\arg\max} \sum_{k \in S(\lambda)} \sum_{l \in \mathcal{L}} I(X_k, Y_l).$$

So we choose such lambda that the cumulative sum of the values of the mutual information between Y and the selected variables is maximal. Of course, this method has its limitations. For example, it may turn out that for some selected variables the mutual information is close to zero, and these variables interact with previously selected variables, so that only conditional mutual information is nonzero. Such variables will be ignored in the above score function. Nevertheless, in practice the above criterion usually works effectively, and most importantly it is simple to calculate, and can be easily calculated for many lambda values.

We provide a simple example to illustrate the method for choosing the optimal value of the cost factor. We consider a simple dataset containing five binary features, $X_1, X_2, X_3, X_4, X_5 \in \{-1, 1\}$, and two target variables, $Y_1, Y_2 \in \{0, 1\}$. The target variables are derived from features X_1 and X_2 such that $P(Y_1 = 1|X_1) = \sigma(3X_1)$ and $P(Y_2 = 1|X_2) = \sigma(3X_2)$, where $\sigma(s) = (1+\exp(-s))^{-1}$ represents the sigmoid function. Additionally, X_3 and X_4 are slightly modified versions of X_1 and X_2, respectively, with 5% of their values randomly flipped, making them less informative than X_1 and X_2. The feature costs are defined as follows:

- The costs of X_1 and X_2 are $c_1 = c_2 = 1$.

TABLE 5.1: Illustrative example of choice factor optimization. Highlighted cells indicate selected feature subsets when the available budget is $T = 1$.

$T = 1$	Step	1	2	3	4	5
$\lambda = 0$	Selected features	$\{X_1\}$	$\{X_1, X_2\}$	$\{X_1, X_2, X_3\}$	$\{X_1, X_2, X_3, X_4\}$	$\{X_1, X_2, X_3, X_4, X_5\}$
	Total cost	**1**	2	2.5	3	3.25
	Cumulated Mutual Inf.	**0.489**	0.965	1.822	2.677	2.683
	Hamming loss	**0.312**	0.158	0.158	0.158	0.158
$\lambda = 1.952$ (λ_{max})	Selected features	$\{X_5\}$	$\{X_5, X_3\}$	$\{X_5, X_3, X_4\}$	$\{X_5, X_3, X_4, X_1\}$	$\{X_5, X_3, X_4, X_1, X_2\}$
	Total cost	0.25	**0.75**	1.25	2.25	3.25
	Cumulated Mutual Inf.	< 0.001	**0.371**	0.731	1.707	2.683
	Hamming loss	0.547	**0.310**	0.190	0.178	0.158
$\lambda = 0.253$ (λ_{opt})	Selected features	$\{X_3\}$	$\{X_3, X_4\}$	$\{X_3, X_4, X_2\}$	$\{X_3, X_4, X_2, X_1\}$	$\{X_3, X_4, X_2, X_1, X_5\}$
	Total cost	0.5	**1**	2	3	3.25
	Cumulated Mutual Inf.	0.370	**0.727**	1.703	2.677	2.683
	Hamming loss	0.332	**0.190**	0.178	0.158	0.158

- The costs of X_3 and X_4 are $c_3 = c_4 = 0.5$.
- The cost of X_5 is $c_5 = 0.25$.

Thus, features X_1 and X_2 are expensive and relevant. Features X_3 and X_4 are cheap and relevant, although they are less informative than X_1 and X_2. Finally, the last feature, X_5, is completely irrelevant and the cheapest. In Table 5.1, we present features selected in the consecutive steps of the method along with the total cost, accumulated mutual information value, and Hamming loss, corresponding to the given subset of features. To assess the quality of feature subsets generated by each method within the given budget, we employed binary relevance (BR) and the *k*-nearest neighbors (kNN) algorithm ($k = 10$) [162]. We chose BR-based kNN because of its simplicity and effectiveness, as BR transforms the problem using a one-versus-all approach to create a separate model for each label, avoiding the need for complex classification strategies. More sophisticated variants of BR-kNN, such as the kernel local label information method [44] and the kernel-based multilabel kNN [152], can be used as well. We measure the quality of the model with Hamming loss, defined as the fraction of labels that are incorrectly predicted out of the total number of labels. Values λ_{opt} and λ_{max} were selected using the procedure introduced in this chapter. We searched the grid of 100 λ values and selected the one that maximizes our metric. We report the results for λ values equal to 0, $\lambda_{max} = 1.952$, and $\lambda_{opt} = 0.253$. The highlighted cells indicate the selected feature subsets when the budget is $T = 1$. For $\lambda = 0$, as expected, we choose feature X_1, which is strongly correlated with one of the target variables and at the same time the most expensive. Unfortunately, in this case, we cannot add another feature because then we will exceed the assumed budget $T = 1$. Therefore, for the model based only on variable X_1 we get a non-satisfying Hamming loss value, equal to 0.312. For $\lambda = \lambda_{max}$, we first choose the cheapest but at the same time irrelevant variable X_5, and then the relevant variable X_3. The obtained model produces non-satisfying Hamming loss too, equal to 0.310 and in addition, it contains irrelevant variable X_5. For $\lambda = \lambda_{opt}$ we have a

model based on significant and cheap variables X_3, X_4. Moreover, we do not select the irrelevant variable X_5. Most importantly, for λ_{opt} we obtain the model with the highest predictive power. The Hamming loss is 0.190, which is much lower value than the values of the Hamming loss for $\lambda_{max} = 1.952$ and $\lambda = 0$. The above example shows that the proposed method allows to obtain a reasonable compromise between the informativeness of selected features and their cost.

5.4 Experiments

In this section, we present the experimental results of the methods discussed above. These are extended results compared to those described in papers [A1, A2]. In the case of selection methods that take into account cost information, the problem is the limited availability of datasets with assigned costs. Therefore, in most papers, strategies for generating artificial costs are considered. In this section, we describe 3 possible strategies called C1, C2 and C3. Our most important contribution is the proposal of the C1 scheme, which is based on the so-called proxy variables. Proxy variables are generated artificially based on the original variables. The introduction of proxy variables allows us to control the trade-off between the informativeness of variables and their cost. In Section 5.4.1, we describe the experimental framework, including proxy features and cost generation strategies, and in the following subsections, we present the results separately for single and multi-label classification.

5.4.1 Experimental framework

Unfortunately, datasets that include feature cost information are rare. One possible approach is to assign costs randomly [14, 165], though this does not accurately reflect real-world scenarios, where costs may exhibit correlations with feature relevance. An alternative strategy involves consulting domain experts who can estimate the costs as they would likely occur in practice. In cases where features represent the results of diagnostic tests, official price lists (often publicly available) can serve as a basis for assigning costs. This approach was applied to the MIMIC dataset [139]. However, assigning costs through expert consultation presents several challenges: it may be both expensive and time-consuming, and certain medical procedures can be difficult to price accurately. Another solution is to artificially generate costs using specific cost-setting strategies that aim to approximate real-world scenarios. This approach is straightforward to implement and allows the incorporation of information regarding feature relevance.

We begin by introducing the concept of proxy features. Alongside the original features $X_1, \ldots, X_p$, we consider proxy features $X_1^*, \ldots, X_p^*$ derived from the original features. Each proxy feature X_j^* is created by randomly permuting $\rho \cdot n$ values within the corresponding original feature X_j, where $\rho \in [0, 1]$ serves as a parameter, and n denotes the number of observations.

TABLE 5.2: Averaged values of estimated MI for different ρ and number of bins used for discretization.

Averaged Mutual Information	ρ	#bins= 2	#bins= 5	#bins= 10
$\widehat{I}(X_j, Y)$	0	0.448	0.493	0.522
$\widehat{I}(X_j^*, Y)$	0.05	0.392	0.421	0.450
$\widehat{I}(X_j^*, Y)$	0.1	0.345	0.365	0.386
$\widehat{I}(X_j^*, Y)$	0.3	0.197	0.203	0.215
$\widehat{I}(X_j^*, Y)$	0.5	0.096	0.099	0.108
$\widehat{I}(X_j^*, Y)$	0.9	0.005	0.006	0.009
$\widehat{I}(X_j^*, Y)$	1	0.001	0.002	0.004

Formally, let $X_j^{(1)}, \ldots, X_j^{(n)}$ be the values of feature X_j in the training data. In order to generate the values of the proxy feature $X_j^{*(1)}, \ldots, X_j^{*(n)}$, we randomly select a set $R \subset \{1, \ldots, n\}$ of size $\rho \cdot n$. Then, for $i \in R$ we set $X_j^{*(i)} = X_j^{(\sigma(i))}$ and for $i \notin R$ we set $X_j^{*(i)} = X_j^{(i)}$, where $\sigma : R \to R$ is some non-identity permutation of the set R. When $\rho = 1$, we permute all values, breaking the dependence between the proxy variable and the target variable Y. For $\rho = 0$, the proxy variable matches the original variable. In general, when $\rho \in (0, 1)$, the dependence between the proxy variable X_j^* and target variable Y is weaker than the dependence between the original variable X_j and Y. The proxy variables can be treated as noisy copies of the original variables. Permuting feature values is a widely recognized statistical method employed to emulate independence between two variables and construct independence tests [10]. To investigate the diminishing effect of shuffling on the relationship between X_j^* and a target variable Y, we present a straightforward computational illustration. We consider a scenario where $Y \in \{0, 1\}$, with equal probabilities $P(Y = 1) = P(Y = 0) = 0.5$, and X_j is generated using conditional distributions $X_j|Y = 0 \sim N(0, 1)$ and $X_j|Y = 1 \sim N(3, 1)$. In Table 5.2, we present the values of estimated mutual information $I(Y, X_j)$ alongside the estimated values of $I(Y, X_j^*)$, where X_j^* is created using varying ρ values and different discretization bin counts. The reported outcomes are the averages from 1000 iterations of Y and X_j, each of a size of $n = 1000$. As expected, the mutual information decreases with increasing ρ, irrespective of the number of bins employed in discretizing X_j^*.

Within the scope of the experimental framework under consideration, it is assumed that the costs associated with features are not constant. Consequently, the cost-constrained methods can produce divergent outcomes compared to conventional approaches. In light of the unavailability of datasets that encompass predefined costs, we consider three distinct strategies to synthetically generate the costs. By employing proxy features as described above, it becomes feasible to effectively control the relationship between feature relevance and associated costs. We present the following three strategies:

- **Strategy C1:** the original features $X_1, \ldots, X_j, \ldots, X_p$ are assigned cost $c_j = 1$, while the cost of the proxy features X_j^* is $c_j^* = \Psi \cdot c_j$, where $\Psi \in (0, 1)$ is another parameter that controls the relationship between the costs of the original and proxy feature. For example, when $\Psi = 0.5$, the cost of the proxy feature is two times less than the cost of the original feature. Generally, when ρ and Ψ are small, selecting the proxy feature instead of the original feature is preferable to reduce the cost. The framework above mimics a real scenario. For example, in medical diagnostics, the original feature may correspond to some expensive diagnostic test that always returns an error-free result, whereas the proxy feature may be a cheaper counterpart that returns incorrect values with some probability.

- **Strategy C2:** each feature X_j is assigned a specific cost $c_j = I(X_j, Y)$ and then the assigned costs are normalized $c_j \leftarrow c_j / \max_k c_k$.

- **Strategy C3:** costs are assigned randomly utilizing an uniform distribution $c_j \sim \mathcal{U}(0, 1)$ and then normalized $c_j \leftarrow c_j / \max_k c_k$.

In the case of C2 and C3, we eliminate the need to use proxy features, and costs are generated based on their relationship with the target variable (C2) or completely at random (C3). Note that scheme C1 is the most flexible due to the introduction of parameters ρ and Ψ and gives the greatest hope for the advantage of cost-constrained methods over traditional methods that ignore costs. This is because, with the appropriate selection of the values of parameters ρ and Ψ, proxy variables that are cheaper equivalents of the original variables can effectively replace them.

5.4.2 Methods

In the case of single-label classification, we compare the proposed cost-constrained method (5.11) against the following baselines: MIM [80], MRMR [110] and JMI [95, 154]. In the case of multi-label classification, we compare the proposed cost-constrained method (5.12) against the following baselines: AMI [77], MDMR [85], MVML [78], SCLS [79], STFS [46] and DCR-MFS [164]. Finally, we compare our method with the existing cost-constrained method CFSM [90]. The description of CFSM is provided below.

The **Cost-constrained Feature Selection on Multi-label data (CFSM)** integrates label enhancement through neighborhood granularity and incorporates cost constraints to optimize the trade-off between feature relevance and cost. Neighborhood granularity provides a mechanism to characterize the local structure of data by defining neighborhoods of instances within a similarity threshold. This allows the dependency between features and labels to be captured

through neighborhood mutual information, which is particularly effective for continuous attributes compared to traditional mutual information measures. The CFSM is defined as follows:

$$J_{cfsm}(Y, X_k, c_k|S) = \Big(\sum_{l=1}^{q} SIG_l \cdot I(X_k, Y_l)\Big) \cdot (1 - c_k),$$

where c_k is k-th feature cos. The SIG term in CFSM represents the significance of a label $l \in \mathcal{L}$ in the entire multi-label dataset. It quantifies how much each label contributes to describing the instances in the dataset after label enhancement. Higher values of SIG_l indicate that label l appears more frequently and is more influential in the dataset. For an enhanced dataset where instances have label distributions rather than just binary labels, the significance of a label is calculated as:

$$SIG_l = \frac{1}{n}\sum_{i=1}^{n} d_l^{x^i}$$

and $d_l^{x^i}$ represents the description degree of label l for instance x^i which is obtained through the label enhancement process. For a given instance x^i, its description degree for a positive label l is defined as:

$$d_l^{x^i} = \frac{\sum\limits_{x^j \in \delta(x^i)} \Big|\{x^j : y_l^j = 1\}\Big|}{\sum_{l=1}^{q} \sum\limits_{x^j \in \delta(x^i)} \Big|\{x^j : y_l^j = 1\}\Big|}$$

where the $|\cdot|$ represents the cardinality of the set, $\delta(x^i)$ is the neighborhood of x^i, determined by neighborhood granularity. The numerator counts how many neighboring instances also have label l and the denominator sums this value across all labels, ensuring that all description degrees sum to 1 for each instance.

To evaluate the quality of the feature subsets generated by each method in the single-label scenario, we employed a logistic regression model. To assess the quality of the feature subsets generated by each method in a multi-label scenario within the given budget, we employed the binary relevance (BR) method in conjunction with the *k*-nearest neighbors (*kNN*) algorithm ($k = 10$) [162]. While more advanced multilabel classifiers, such as the kernel-based local label information method [44] and the discernibility-based multilabel *kNN* approach [152], are available for evaluating the effectiveness of feature selection methods, we opted for the BR-based *kNN*. This choice was motivated by the simplicity and efficacy of BR, which transforms the multilabel problem into multiple binary classification tasks using a one-versus-all approach, thereby eliminating the need for complex classification strategies.

5.4.3 Evaluation measures

To evaluate the predictive power of selected feature subsets in a single-label scenario, we employed traditional scoring measures such as accuracy, precision, recall and F1.

To assess the predictive power of the selected feature subsets in a multi-label scenario, we employed three widely used example-based evaluation measures: the Hamming loss, the ranking loss and F1 score [91]. Although additional metrics, such as multilabel accuracy, were also considered in the experiments, the conclusions drawn were consistent with those based on the Hamming loss, ranking loss and F1. Therefore, we do not present the numerical results for these additional measures in this work.

We now recall the definitions of the considered metrics. Let $y = (y_1, \ldots, y_q)$ denote the vector of actual target variables and $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_q)$ denote the vector of predicted target variables. Hamming loss measures the frequency of misclassified example-label pairs, where either a label not associated with an example is predicted, or a label associated with an example is not predicted. A smaller Hamming loss value indicates better performance. The Hamming loss, defined as follows, represents the average number of incorrect predictions:

$$\text{Hamming loss}(y, \hat{y}) = \frac{1}{n_t q} \sum_{i=1}^{n_t} \sum_{j=1}^{q} \mathbb{1}(y_j^i \neq \hat{y}_j^i),$$

where $\mathbb{1}$ is the indicator function.

Next, we define the ranking loss. Let $f(x^i, y_j^i)$ be a real-valued function that quantifies the confidence that j-th label is a relevant (active) label for observation x^i (e.g., the confidence that $y_j^i = 1$). Let $R^i = \{j : y_j^i = 1\}$ be the set of active labels and $\bar{R}^i = \{j : y_j^i = 0\}$ be the set of nonactive labels, for instance i. Ranking loss measures the average proportion of label pairs that are incorrectly ordered. A lower ranking loss value indicates better performance. The ranking loss is then defined as:

$$\text{Ranking loss}(y, \hat{y}) = \frac{1}{n_t} \sum_{i=1}^{n_t} \sum_{y_j^i \in R^i, y_k^i \in \bar{R}^i} \frac{1}{|R^i| \cdot |\bar{R}^i|} \mathbb{1}(f(x^i, y_j^i) < f(x^i, y_k^i)),$$

where $\mathbb{1}$ is the indicator function.

The example-based F1 measure in a multilabel classification problem is the harmonic mean of example-based precision and recall, averaged over all instances in the test set. It is defined as:

$$F1 = \frac{1}{n_t} \sum_{i=1}^{n_t} \frac{2 \cdot Precision_i \cdot Recall_i}{Precision_i + Recall_i},$$

where the precision for the i-th observation is the proportion of correctly predicted labels (true positives) out of all predicted labels:

$$\text{Precision}_i = \frac{\sum_{j=1}^{q} \mathbb{1}(\hat{y}_j^i = 1 \wedge y_j^i = 1)}{\sum_{j=1}^{q} \mathbb{1}(\hat{y}_j^i = 1)}$$

and the recall for the i-th observation is the proportion of correctly predicted labels (true positives) out of all actual labels:

$$\text{Recall}_i = \frac{\sum_{j=1}^{q} \mathbb{1}(\hat{y}_j^i = 1 \wedge y_j^i = 1)}{\sum_{j=1}^{q} \mathbb{1}(y_j^i = 1)}$$

and $\mathbb{1}$ is the indicator function.

Finally, it is worth noting that each dataset used in the experiments was divided into training and testing subsets. The training set was utilized to perform feature selection and train the classification models, while the evaluation metrics were computed on the test set. In experiments, we split the data into train and test size with 80% and 20% respectively.

5.4.4 Single label case: results for artificial data

In the case of artificially generated datasets, the set of true relevant variables that determine the value of the target variable Y is known. Consequently, the efficiency of the feature selection techniques can be evaluated directly without using an external classifier and calculating its accuracy on the test data. Moreover, we can analyze the frequency with which genuinely relevant features are chosen. In our experiments, the artificial dataset is generated as described below. First, we generate a binary target variable $Y \in \{0, 1\}$ from the Bernoulli distribution with success probability π, where the class prior π is treated as a parameter. Then we generate feature vectors in each class from p-dimensional multivariate Gaussian distributions

$$X|Y = 0 \sim \mathcal{N}(0, I), \qquad X|Y = 1 \sim \mathcal{N}(\mu, \Sigma_{ij}),$$

where I is the identity matrix, $\Sigma_{ij} = a^{|i-j|}$, where $a \in [0, 1]$ is a parameter and finally

$$\mu = \{\underbrace{\alpha, \ldots, \alpha}_{\text{relevant}}, \underbrace{0, \ldots, 0}_{\text{irrelevant}}\},$$

where the number of irrelevant features is equal to the parameter p_{noise}. In total, we have 6 parameters (n, p p_{noise}, a, π and α) and their values can be controlled in the experiments. Since they mainly influence the difficulty of the classification task, rather than the performance of the feature selection methods, we present the results for one chosen setting of the parameters: $n = 1000$, $p = 10$, $p_{noise} = 5$, $\alpha = 1$ and $\pi = 0.5$. Other settings have also been analyzed.

We intend to answer two research questions.

- How do the considered methods perform for cost strategies C1-C3 for different budgets T?
- What is the impact of parameters Ψ, ρ and a?

Our results are presented in Tables 5.3 - 5.4 and extended results in Appendix A describe different parameters ρ and a. Each table shows results for the budget levels (1, 2, and 5) and various evaluation metrics such as Accuracy, F1 score, Precision, and Recall. For each combination of budget and metric, the mean performance of each method is presented along with its standard deviation.

The effectiveness of the proposed method is pronounced in strategy C1 (Tables 5.3, 5.4, A.1), whereas its advantages are less apparent in strategies C2 and C3 (Table A.2), indicating a context-dependent performance profile. The method demonstrates superior performance for small values of ρ and Ψ, which is consistent with theoretical expectations, as in this case proxy features are highly correlated with the target variable and at the same time they are much cheaper than their counterparts among original variables. Therefore, with a limited budget of $T = 1$, we can use more variables to effectively predict the target variable. Additionally, the parameter a exerts a more substantial influence on the classification performance of models incorporating the full set of features.

5.4.5 Single label case: results for real data

This Experimental Section contains experiments related to the selection of single-label cost-constrained features. Figure 5.2 contains two subplots (A and B), comparing the *ROC AUC* metric against the budget for two example datasets: *Banknote* (A) [89] and *Breast cancer* (B) [148]. It evaluates the performance of the proposed method against three established techniques: Joint Mutual Information (JMI), Mutual Information Maximization (MIM), and Minimum Redundancy Maximum Relevance (MRMR). The primary goal is to assess how well these methods perform in terms of the ROC AUC metric across varying computational budgets.

The proposed method appears to have high performance in both datasets, achieving the largest *ROC AUC* with a smaller budget. Error bars indicate variability in performance (standard deviation across multiple trials). The proposed method generally shows smaller or comparable uncertainty to other methods. In the Banknote dataset (A), the proposed method outperforms other methods across most budgets. In the Breast dataset (B), the proposed method achieves a competitive performance, closely matching MRMR as the budget increases. An important advantage of the proposed method is that it selects features already under a budget smaller than 1. In contrast, the competing methods tend to prioritize expensive features at the early stages, which makes it impossible to include any features when the budget is limited below this threshold.

TABLE 5.3: Results of artificial dataset feature selection for the cost generation method C1 and parameters $\pi = 0.5$ and $\alpha = 1$ and $\Psi = 0.1$

(A) Additional parameters: $\rho = 0.1$ and $a = 0.1$

Budget	**Metric**	Proposed	MIM	MRMR	JMI
1	Accuracy	**0.806 ± 0.024**	0.659 ± 0.028	0.659 ± 0.028	0.659 ± 0.028
2		**0.824 ± 0.030**	0.744 ± 0.022	0.754 ± 0.022	0.728 ± 0.031
5		0.824 ± 0.014	0.821 ± 0.027	**0.856 ± 0.029**	**0.856 ± 0.029**
1	F1 score	**0.811 ± 0.020**	0.667 ± 0.040	0.667 ± 0.040	0.667 ± 0.040
2		**0.829 ± 0.023**	0.746 ± 0.031	0.757 ± 0.028	0.731 ± 0.029
5		0.827 ± 0.015	0.825 ± 0.021	**0.860 ± 0.026**	**0.860 ± 0.026**
1	Precision	**0.806 ± 0.041**	0.663 ± 0.046	0.663 ± 0.046	0.663 ± 0.046
2		**0.823 ± 0.045**	0.753 ± 0.058	0.762 ± 0.047	0.739 ± 0.070
5		0.829 ± 0.029	0.825 ± 0.035	**0.854 ± 0.039**	**0.854 ± 0.039**
1	Recall	**0.818 ± 0.033**	0.673 ± 0.058	0.673 ± 0.058	0.673 ± 0.058
2		**0.837 ± 0.033**	0.743 ± 0.040	0.756 ± 0.053	0.729 ± 0.045
5		0.826 ± 0.024	0.827 ± 0.046	**0.867 ± 0.028**	**0.867 ± 0.028**

(B) Additional parameters: $\rho = 0.1$ and $a = 0.9$

Budget	**Metric**	Proposed	MIM	MRMR	JMI
1	Accuracy	**0.818 ± 0.025**	0.671 ± 0.037	0.671 ± 0.037	0.671 ± 0.037
2		**0.848 ± 0.006**	0.772 ± 0.042	0.716 ± 0.081	0.776 ± 0.029
5		0.886 ± 0.032	0.857 ± 0.018	0.881 ± 0.047	**0.902 ± 0.024**
1	F1 score	**0.831 ± 0.014**	0.683 ± 0.025	0.683 ± 0.025	0.683 ± 0.025
2		**0.860 ± 0.013**	0.774 ± 0.049	0.725 ± 0.076	0.786 ± 0.033
5		0.898 ± 0.029	0.866 ± 0.021	0.895 ± 0.040	**0.910 ± 0.025**
1	Precision	**0.795 ± 0.036**	0.675 ± 0.044	0.675 ± 0.044	0.675 ± 0.044
2		**0.807 ± 0.027**	0.778 ± 0.063	0.719 ± 0.069	0.764 ± 0.041
5		0.823 ± 0.047	0.831 ± 0.043	0.816 ± 0.067	**0.847 ± 0.045**
1	Recall	**0.873 ± 0.044**	0.696 ± 0.066	0.696 ± 0.066	0.696 ± 0.066
2		**0.922 ± 0.031**	0.772 ± 0.058	0.739 ± 0.117	0.810 ± 0.034
5		0.990 ± 0.009	0.906 ± 0.046	**0.994 ± 0.011**	0.984 ± 0.014

(C) Additional parameters: $\rho = 0.5$ and $a = 0.1$

Budget	**Metric**	Proposed	MIM	MRMR	JMI
1	Accuracy	**0.659 ± 0.028**	**0.659 ± 0.028**	**0.659 ± 0.028**	**0.659 ± 0.028**
2		0.728 ± 0.031	0.744 ± 0.022	**0.754 ± 0.022**	0.728 ± 0.031
5		0.823 ± 0.020	**0.856 ± 0.029**	**0.856 ± 0.029**	**0.856 ± 0.029**
1	F1 score	**0.667 ± 0.040**	**0.667 ± 0.040**	**0.667 ± 0.040**	**0.667 ± 0.040**
2		0.729 ± 0.032	0.746 ± 0.031	**0.757 ± 0.028**	0.731 ± 0.029
5		0.827 ± 0.013	**0.860 ± 0.026**	**0.860 ± 0.026**	**0.860 ± 0.026**
1	Precision	**0.663 ± 0.046**	**0.663 ± 0.046**	**0.663 ± 0.046**	**0.663 ± 0.046**
2		0.741 ± 0.067	0.753 ± 0.058	**0.762 ± 0.047**	0.739 ± 0.070
5		0.823 ± 0.022	**0.854 ± 0.039**	**0.854 ± 0.039**	**0.854 ± 0.039**
1	Recall	**0.673 ± 0.058**	**0.673 ± 0.058**	**0.673 ± 0.058**	**0.673 ± 0.058**
2		0.722 ± 0.044	0.743 ± 0.040	**0.756 ± 0.053**	0.729 ± 0.045
5		0.833 ± 0.035	**0.867 ± 0.028**	**0.867 ± 0.028**	**0.867 ± 0.028**

TABLE 5.4: Results of artificial dataset feature selection for the cost generation method C1 and parameters $\pi = 0.5$ and $\alpha = 1$ and $\Psi = 0.5$

(A) Additional parameters: $\rho = 0.1$ and $a = 0.1$

Budget	**Metric**	Proposed	MIM	MRMR	JMI
1	Accuracy	**0.702 ± 0.028**	0.659 ± 0.028	0.659 ± 0.028	0.659 ± 0.028
2		**0.784 ± 0.027**	0.744 ± 0.022	0.754 ± 0.022	0.728 ± 0.031
5		0.828 ± 0.034	0.821 ± 0.027	**0.856 ± 0.029**	**0.856 ± 0.029**
1	F1 score	**0.706 ± 0.025**	0.667 ± 0.040	0.667 ± 0.040	0.667 ± 0.040
2		**0.788 ± 0.018**	0.746 ± 0.031	0.757 ± 0.028	0.731 ± 0.029
5		0.831 ± 0.032	0.825 ± 0.021	**0.860 ± 0.026**	**0.860 ± 0.026**
1	Precision	**0.711 ± 0.027**	0.663 ± 0.046	0.663 ± 0.046	0.663 ± 0.046
2		**0.791 ± 0.043**	0.753 ± 0.058	0.762 ± 0.047	0.739 ± 0.070
5		0.835 ± 0.045	0.825 ± 0.035	**0.854 ± 0.039**	**0.854 ± 0.039**
1	Recall	**0.704 ± 0.059**	0.673 ± 0.058	0.673 ± 0.058	0.673 ± 0.058
2		**0.787 ± 0.030**	0.743 ± 0.040	0.756 ± 0.053	0.729 ± 0.045
5		0.827 ± 0.040	0.827 ± 0.046	**0.867 ± 0.028**	**0.867 ± 0.028**

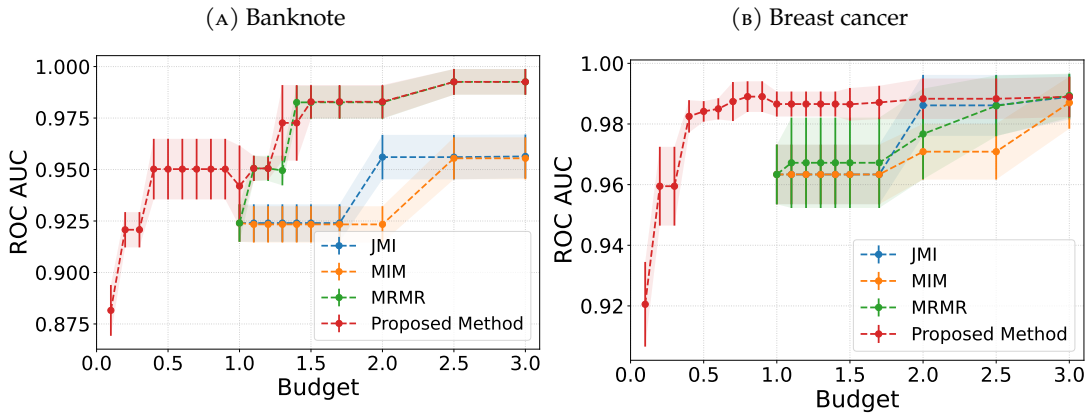
(B) Additional parameters: $\rho = 0.1$ and $a = 0.9$

Budget	**Metric**	Proposed	MIM	MRMR	JMI
1	Accuracy	**0.711 ± 0.056**	0.671 ± 0.037	0.671 ± 0.037	0.671 ± 0.037
2		**0.810 ± 0.032**	0.772 ± 0.042	0.684 ± 0.035	0.776 ± 0.029
5		0.847 ± 0.029	0.859 ± 0.019	0.869 ± 0.042	**0.902 ± 0.024**
1	F1 score	**0.718 ± 0.052**	0.683 ± 0.025	0.683 ± 0.025	0.683 ± 0.025
2		**0.823 ± 0.025**	0.774 ± 0.049	0.689 ± 0.038	0.786 ± 0.033
5		0.861 ± 0.027	0.867 ± 0.021	0.884 ± 0.034	**0.910 ± 0.025**
1	Precision	**0.718 ± 0.079**	0.675 ± 0.044	0.675 ± 0.044	0.675 ± 0.044
2		**0.785 ± 0.022**	0.778 ± 0.063	0.693 ± 0.062	0.764 ± 0.041
5		0.799 ± 0.050	0.834 ± 0.041	0.814 ± 0.065	**0.847 ± 0.045**
1	Recall	**0.722 ± 0.054**	0.696 ± 0.066	0.696 ± 0.066	0.696 ± 0.066
2		**0.868 ± 0.061**	0.772 ± 0.058	0.690 ± 0.061	0.810 ± 0.034
5		0.937 ± 0.022	0.906 ± 0.046	0.971 ± 0.011	**0.984 ± 0.014**

(C) Additional parameters: $\rho = 0.5$ and $a = 0.1$

Budget	**Metric**	Proposed	MIM	MRMR	JMI
1	Accuracy	**0.659 ± 0.028**	**0.659 ± 0.028**	**0.659 ± 0.028**	**0.659 ± 0.028**
2		0.728 ± 0.031	0.744 ± 0.022	**0.754 ± 0.022**	0.728 ± 0.031
5		**0.856 ± 0.029**	**0.856 ± 0.029**	**0.856 ± 0.029**	**0.856 ± 0.029**
1	F1 score	**0.667 ± 0.040**	**0.667 ± 0.040**	**0.667 ± 0.040**	**0.667 ± 0.040**
2		0.731 ± 0.029	0.746 ± 0.031	**0.757 ± 0.028**	0.731 ± 0.029
5		**0.860 ± 0.026**	**0.860 ± 0.026**	**0.860 ± 0.026**	**0.860 ± 0.026**
1	Precision	**0.663 ± 0.046**	**0.663 ± 0.046**	**0.663 ± 0.046**	**0.663 ± 0.046**
2		0.739 ± 0.070	0.753 ± 0.058	**0.762 ± 0.047**	0.739 ± 0.070
5		**0.854 ± 0.039**	**0.854 ± 0.039**	**0.854 ± 0.039**	**0.854 ± 0.039**
1	Recall	**0.673 ± 0.058**	**0.673 ± 0.058**	**0.673 ± 0.058**	**0.673 ± 0.058**
2		0.729 ± 0.045	0.743 ± 0.040	**0.756 ± 0.053**	0.729 ± 0.045
5		**0.867 ± 0.028**	**0.867 ± 0.028**	**0.867 ± 0.028**	**0.867 ± 0.028**

FIGURE 5.2: Single label case: Comparison of ROC AUC metric against budget.



5.4.6 Multi-label case: results for artificial data

For experiments made with artificial data in the multi-label classification problem, the set of relevant features is predefined, allowing the quality of feature selection methods to be evaluated directly without relying on an external classifier. This enables us to analyze the frequency with which truly relevant features are selected. Our goal was to generate a dataset that is challenging in terms of feature selection due to the presence of interactions between variables and conditional dependencies between labels. This latter element is the core of multi-label classification. The artificial dataset was created as follows. First, we generated binary features $X_1, \ldots, X_{50}$ independently using a Bernoulli distribution with a success probability $P(X_j = 1) = 0.5$. The cost of each feature was set to 1. Then, additional proxy features $X_1^*, \ldots, X_{50}^*$ were added as described in Section 5.4.1. We used parameters ($\Psi = 0.1$) and $\rho = 0.1$, resulting in a cost of 0.1 for each proxy feature. The target variables were generated sequentially using the chain rule. The first target variable Y_1 was generated using the posterior probability $P(Y_1 = 1|X_1 = x_1) = \sigma(x_1)$, where $\sigma(s) = \frac{\exp(s)}{1+\exp(s)}$ is a sigmoid function. Subsequent target variables were generated using the probabilities

$$P(Y_k = 1|X_{k-1} = x_{k-1}, X_k = x_k, Y_{k-1} = y_{k-1}) = \sigma(y_{k-1} + \text{XOR}(x_k, x_{k-1})),$$

for $k = 2, \ldots, q$, where $\text{XOR}(x_k, x_{k-1}) = \mathbb{I}(x_k \neq x_{k-1})$. In the experiments, we used $q = 5$ labels.

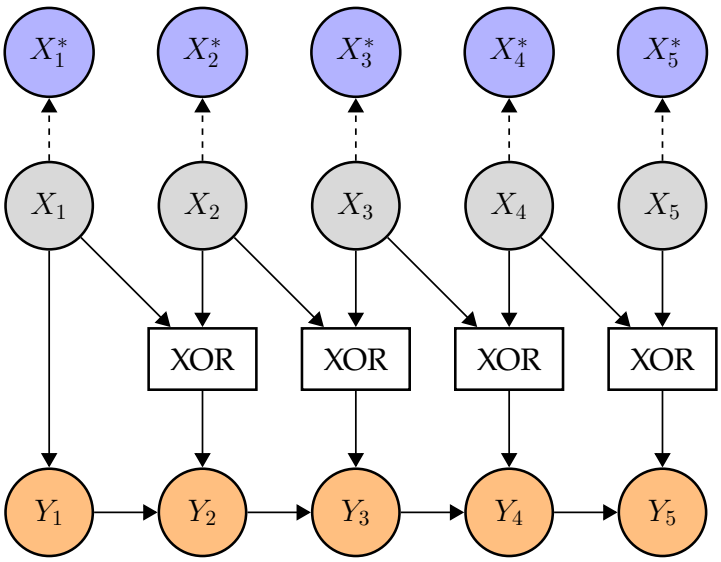
Figure 5.3 (top panel) illustrates the dependency structure of the artificial dataset, which forms a chain. Each target variable Y_k is generated based on the previous target variable Y_{k-1} and two features: X_{k-1} and X_k. Chain structures are commonly employed in multi-label classification [32, 119] due to their association with the factorization of the joint probability $P(Y_1, \ldots, Y_q|X) = P(Y_1|X) \prod_{l=2}^{q} P(Y_l|X, Y_1, \ldots, Y_{l-1})$. This approach allows modeling of the joint probability through conditional probabilities and describing the conditional dependencies between labels.

In this scenario, feature X_1 is marginally dependent on Y_1, while features $X_2, \ldots, X_5$ are marginally independent of the target variables (i.e., $I(Y_k, X_k) = 0$ for $k = 2, \ldots, 5$), but they exhibit conditional dependence on the target variables (i.e., $I(Y_k, X_k|X_{k-1}) > 0$ for $k = 2, \ldots, 5$). We observe interactions involving two features and one label, where interaction information is positive $II(Y_k, X_k, X_{k-1}) > 0$ for $k = 2, \ldots, 5$. This dataset poses a challenge for existing feature selection methods since identifying all relevant features is only possible when conditional dependencies are considered in the feature relevance measure. Features $X_6, \ldots, X_{50}$ are irrelevant and serve only as noisy features to make the feature selection problem more complex. Figure 5.3 (bottom panel) provides an example of the data generation process for the first two labels Y_1 and Y_2. Note that the probability of the first label $Y_1 = 1$ is higher when $X_1 = 1$. The second label is most likely to be equal to 1 when $Y_1 = 1$ and simultaneously $X_1 \neq X_2$ (as highlighted in the bolded rows).

The primary objective of the experiment conducted on the artificial dataset was to examine which types of features were selected by the considered methods. In the context of the artificial dataset, we can categorize the features into four distinct groups:

- **Relevant and expensive features:** Features X_1, X_2, X_3, X_4, and X_5, each with a cost of 1.
- **Relevant and cheap features:** Features $X_1^*, X_2^*, X_3^*, X_4^*$, and X_5^*, each with a cost of 0.1. These features are noisy variants of X_1, X_2, X_3, X_4, and X_5.
- **Irrelevant and expensive features:** Features $X_6, \ldots, X_{50}$, each with a cost of 1.
- **Irrelevant and cheap features:** Features $X_6^*, \ldots, X_{50}^*$, each with a cost of 0.1.

The total cost of selecting all the original relevant features $X_1, \ldots, X_5$ is 5. Therefore, if the available budget T is less than 5, it is not feasible to choose all the original relevant features. In such scenarios, the methods should prioritize selecting relevant inexpensive features that may perform slightly worse than the original ones but are significantly cheaper. Another intriguing question is whether the methods can identify features that influence the target variables solely through interactions. Figures 5.4 and 5.5 show the proportion of simulations (iterations in which the artificial dataset is generated) in which a particular feature was selected as relevant for budgets $T = 1$ and $T = 5$, respectively. The top five selected features are displayed. The proposed method initially selects all five relevant inexpensive features, demonstrating its ability to detect interactions and its preference for low-cost features. The CFSM method similarly favors cheaper features but fails to identify features that influence labels through interactions. As expected, traditional methods typically select costly features and, with a budget of $T = 1$, can only choose one relevant feature, X_1, which has a cost of 1. The MDMR, MVML, STFS, and DCR-MFS methods consider interactions and therefore detect interacting features like X_2 or X_3, but only when the available budget is sufficient (i.e., $T = 5$) to include them (Figure 5.5).



X_1	X_2	$XOR(X_1, X_2)$	$P(Y_1 = 1\|X_1) = \sigma(X_1)$	Y_1	$P(Y_2 = 1\|X_1, X_2, Y_1) = \sigma(Y_1 + XOR(X_1, X_2))$
0	0	0	0.5	0	0.50
0	0	0	0.5	1	0.73
0	1	1	0.5	0	0.73
0	**1**	**1**	**0.5**	**1**	**0.88**
1	0	1	0.73	0	0.73
1	**0**	**1**	**0.73**	**1**	**0.88**
1	1	0	0.73	0	0.50
1	1	0	0.73	1	0.73

FIGURE 5.3: Dependency structure for artificial dataset (top panel) and example of data generation mechanism for the first two labels Y_1 and Y_2. The bold rows correspond to the situation where the a posteriori probability of the second label is the highest.

5.4.7 Multi-label case: results for real data

We conducted experiments using publicly available datasets spanning various domains, including text mining, image recognition, and biology. All datasets were sourced from the widely used multilabel dataset repository, MULAN [142]. Table 5.5 provides an overview of key statistics for these datasets: n indicates the number of observations, p refers to the number of features, q represents the number of labels, LD denotes label density, and $Domain$ specifies the dataset domain. Certain datasets contained a large number of features and labels, making it impractical to conduct multiple experiments with varying parameter settings. To address this, we reduced the dimensionality of both the feature and label spaces. Specifically, we selected the 150 features with the highest entropy and limited the number of labels to a maximum of 50, prioritizing

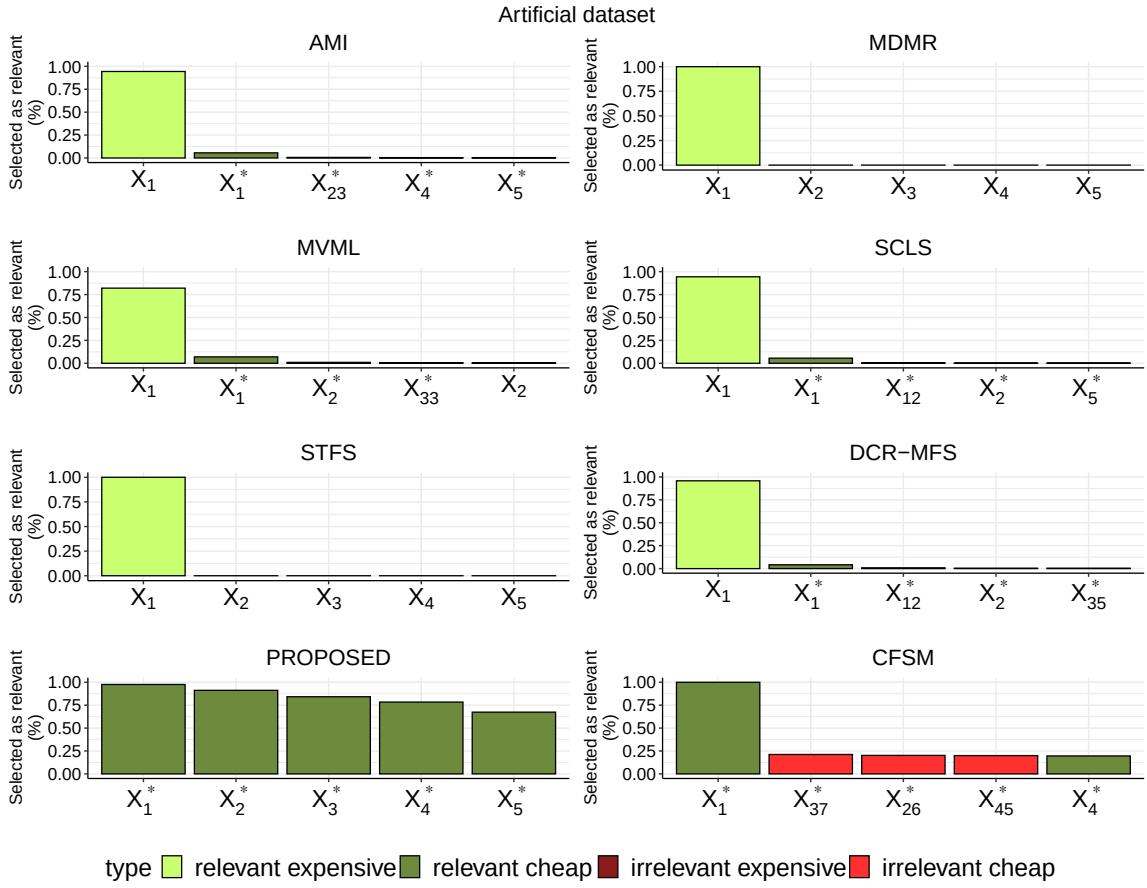


FIGURE 5.4: Fraction of features selected as relevant for the budget $T = 1$ within 1,000 experiments.

TABLE 5.5: Multi-label case: summary statistics of selected datasets.

Dataset	n	p	q	LD	Domain
Bibtex	7395	1836	159	0.015	Text
Bookmarks	87856	2,150	208	0.010	Text
Enron	1702	1,001	53	0.064	Text
Emotions	593	72	6	0.311	Audio
Genbase	662	1186	2	0.046	Biology
Medical	978	1449	45	0.028	Text
Scene	2047	294	6	0.179	Image
Yeast	2417	103	14	0.303	Biology

those with the highest prior probabilities. Additionally, we incorporated proxy features with costs of 0.1 alongside the original features, which had costs of 1, as detailed in **Strategy C1** in Section 5.4.1.

Figures 5.6, 5.7 and 5.8 illustrate how the evaluation measures (Hamming loss, ranking loss, F1) vary across different budgets T for three selected datasets. As expected, we observe a significant advantage of cost-constrained methods (the proposed method and CFSM) over traditional

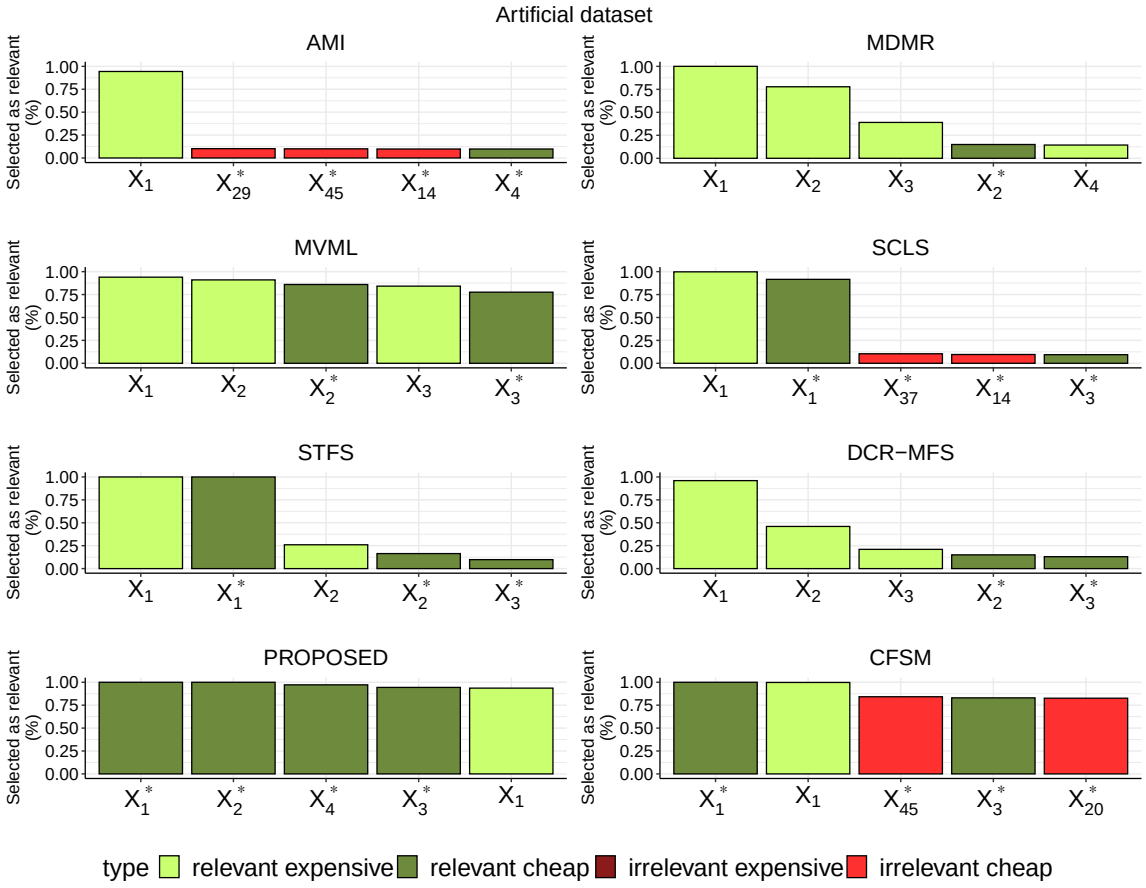
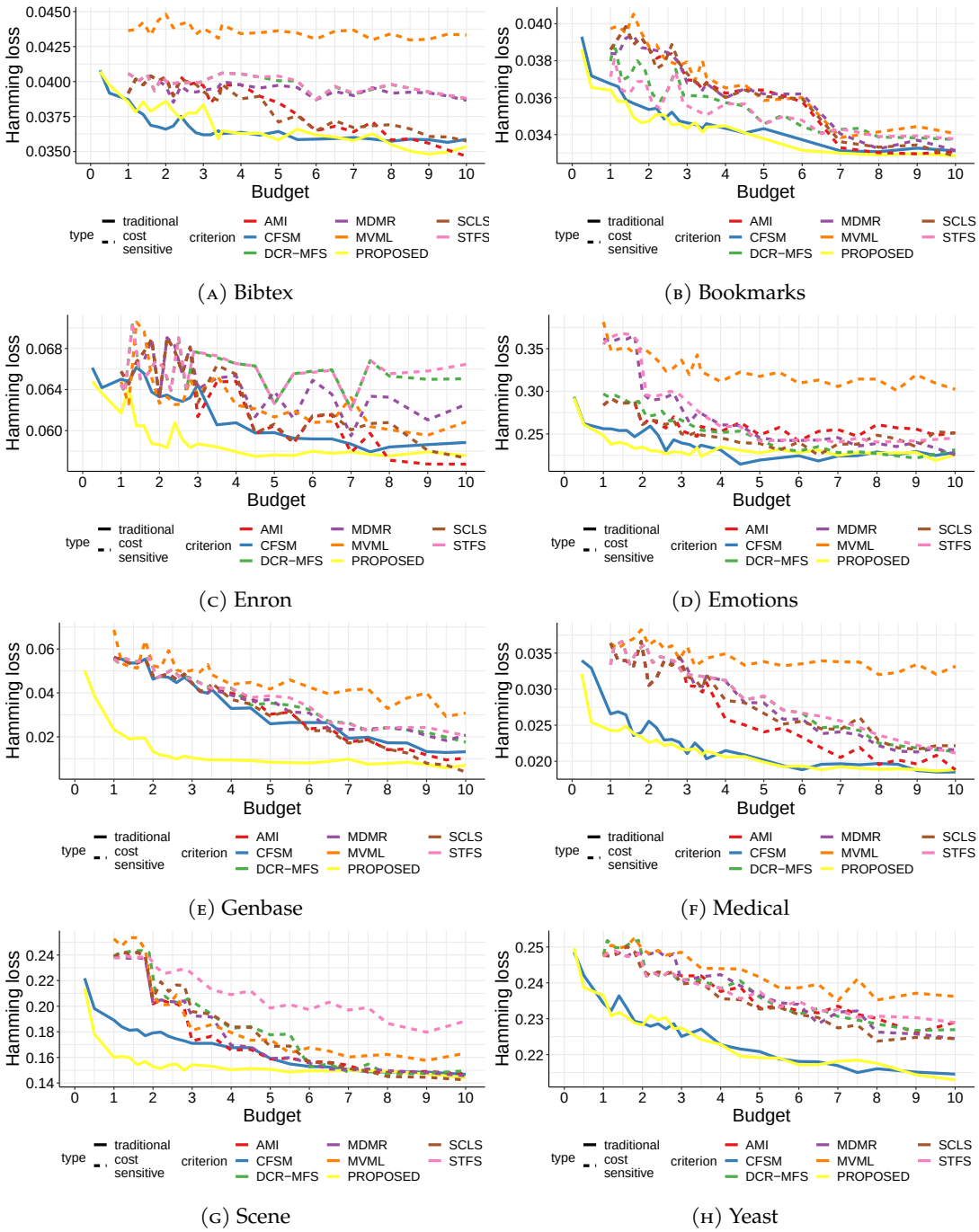


FIGURE 5.5: Fraction of features selected as relevant for the budget $T = 5$ within 1,000 experiments.

methods under low-budget conditions. For instance, in the *Scene* dataset, both losses of the proposed method are two times lower than those of traditional methods when the budget is $T = 1$. The proposed method consistently outperforms other traditional algorithms across all metrics for lower budgets for *Bibtex*, *Bookmarks*, *Emotions*, *Genbase*, *Medical*, *Yeast* datasets. On the other hand, the *Enron* dataset classification problem seems to be complicated for a lower number of features, therefore cost-constrained methods perform better for a slightly higher budget in this situation. Notably, the proposed method performs on par with or better than CFSM, consistent with previous conclusions, as the proposed method utilizes more versatile relevance measures that account for feature interactions. Additionally, unlike CFSM, the proposed method optimizes the penalty parameter λ, which may further enhance its performance. For larger budgets, the differences between the methods become less significant, as traditional and cost-constrained methods can select all relevant features, yielding a powerful classification model.



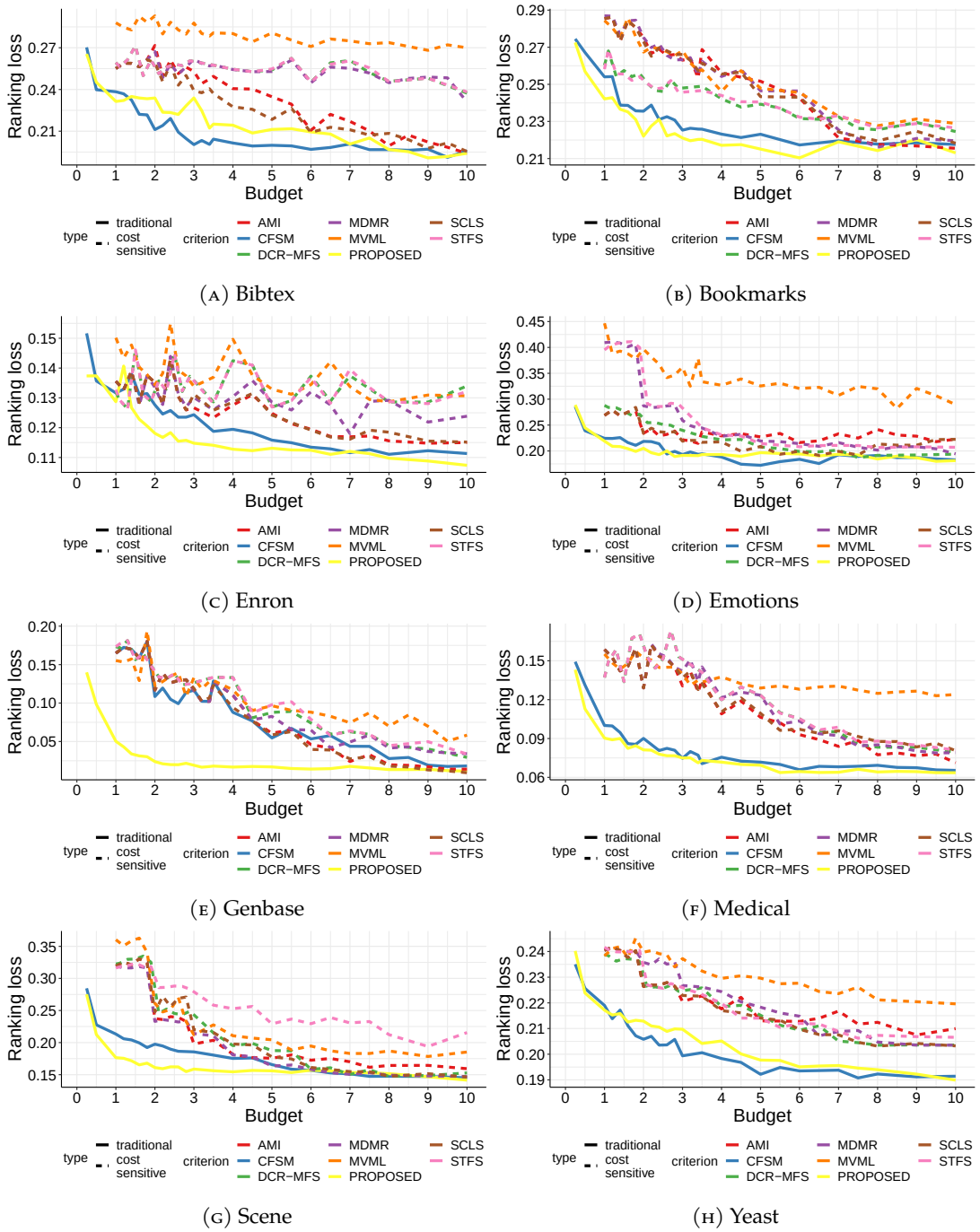
(A) Bibtex (B) Bookmarks

(C) Enron (D) Emotions

(E) Genbase (F) Medical

(G) Scene (H) Yeast

FIGURE 5.6: Hamming loss measure for budget $T \leq 10$ for selected datasets.



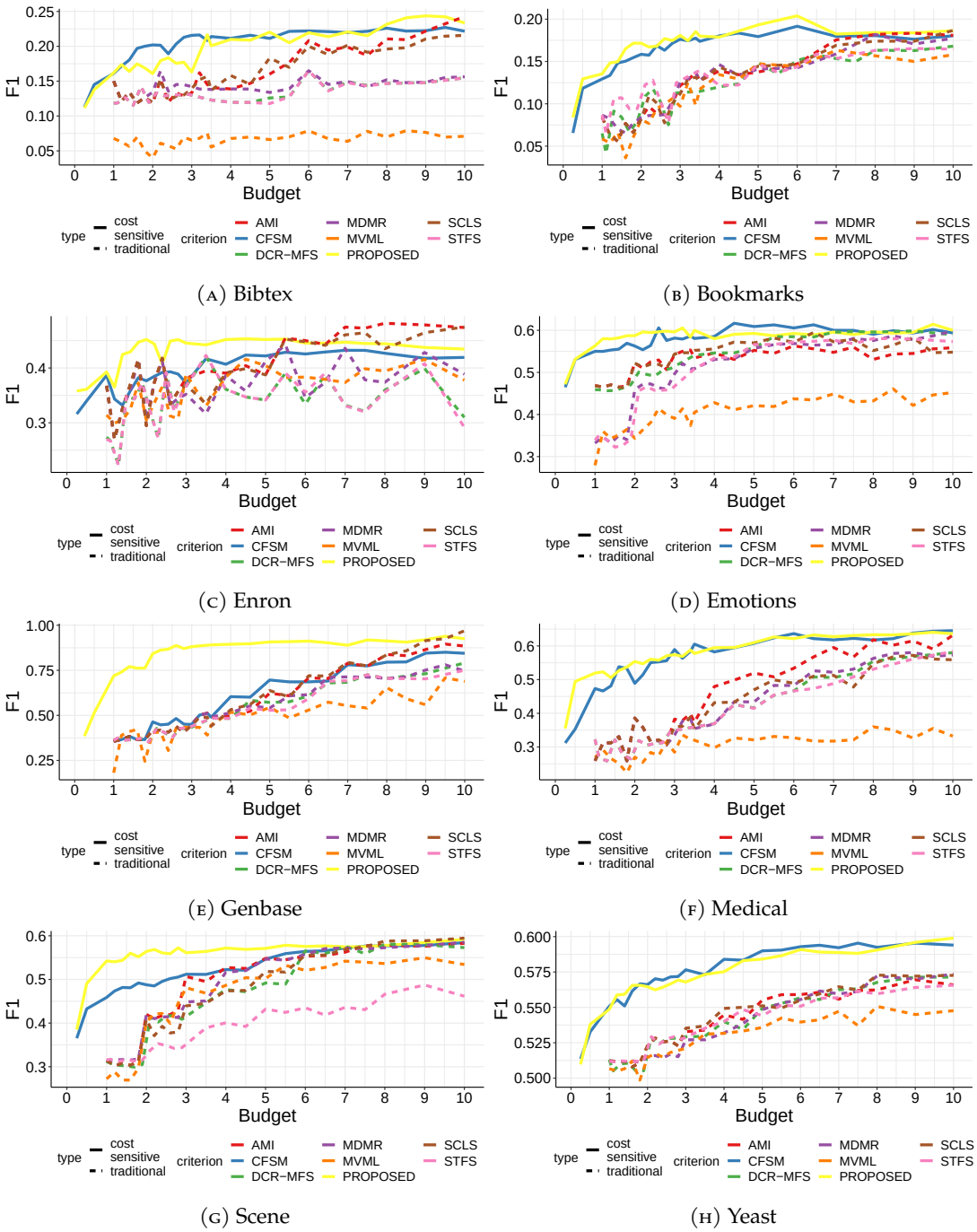
(A) Bibtex (B) Bookmarks

(C) Enron (D) Emotions

(E) Genbase (F) Medical

(G) Scene (H) Yeast

FIGURE 5.7: Ranking loss measure for budget $T \leq 10$ for selected datasets.



(A) Bibtex (B) Bookmarks

(C) Enron (D) Emotions

(E) Genbase (F) Medical

(G) Scene (H) Yeast

FIGURE 5.8: F1 measure for budget $T \leq 10$ for selected datasets.

5.4.8 Case study: Medical dataset MIMIC-II

We performed an experiment on the publicly available medical database MIMIC-II [123] which provides various medical data about 19773 patients from the intensive care unit and their 10 possible diseases (*copd, diabetes, fluid, hypertension, hypotension, kidney, lipoid, liver, thrombosis, thyroid*). The dataset includes 305 features, encompassing various types of medical data. The low-cost features include administrative information such as *age, weight, or marital status*. A second group consists of simple medical tests, typically collected during standard medical interviews (e.g., *heart rate*). More expensive features correspond to laboratory blood tests (e.g., *potassium or calcium levels*). Moreover, ICU patients were monitored periodically, resulting in repeated measurements for selected tests; for these, we computed simple summary statistics (mean, median, standard deviation). An important advantage of the MIMIC dataset is that feature costs were assigned by domain experts [139]. The cost values are derived from the prices of diagnostic tests obtained from Polish analytical laboratories, with the primary source being the official price list of ALAB laboratories. Although the costs are expressed in Polish currency, it is worth emphasizing that the absolute values are not critical, since the relative pricing structure of diagnostic tests tends to be similar across different countries.

In this chapter we selected hypertension disease as the target variable. Then we split the data into train and test size with 80% and 20% respectively. This dataset has been used in previous research; for more information on feature extraction and preprocessing, we refer to [168]. To streamline the process, we performed a preselection, identifying 30 features with the highest mutual information with the target. Additionally, to enhance the feature set without significantly increasing costs, we included 4 administrative features (*Marital status*, *Admit weight*, *Gender*, and *Age*), which are very inexpensive to utilize. By the conclusion of the experiment, the finalized feature set comprised a total of 34 features which refer to basic medical interviews and results of various medical tests.

Figure 5.9 depicts the values of mutual information between the considered features and the target variable as well as the costs of the features. Features are sorted according to 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 connected with the class variable although it has a low cost. Therefore, we can expect that this feature will be selected as relevant by both traditional and cost-constrained 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 interesting features: *Age* (number 3) and *NBP systolic* (number 14), which are highly correlated with the target variable, but their cost is relatively low. On the other hand, feature *urea nitrogen in serum or plasma* (number 23) is also highly correlated with the target variable, but at the same time, its cost is also high.

Figure 5.10 visualizes the results of feature selection for various budgets for traditional and

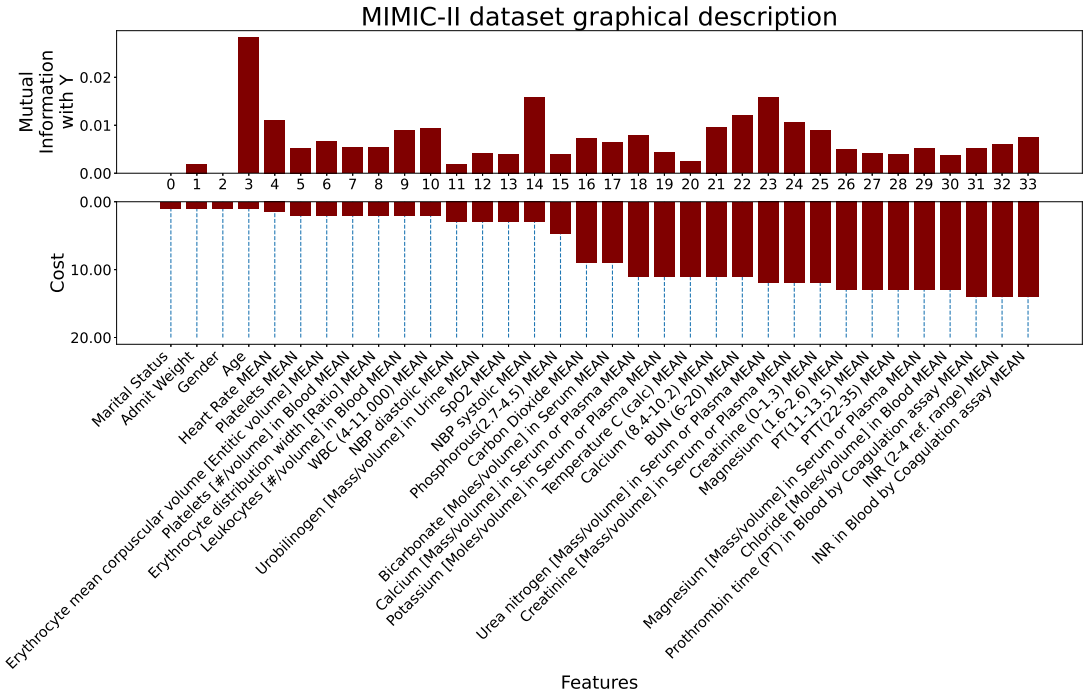


FIGURE 5.9: MIMIC-II dataset. Basic characteristics of features.

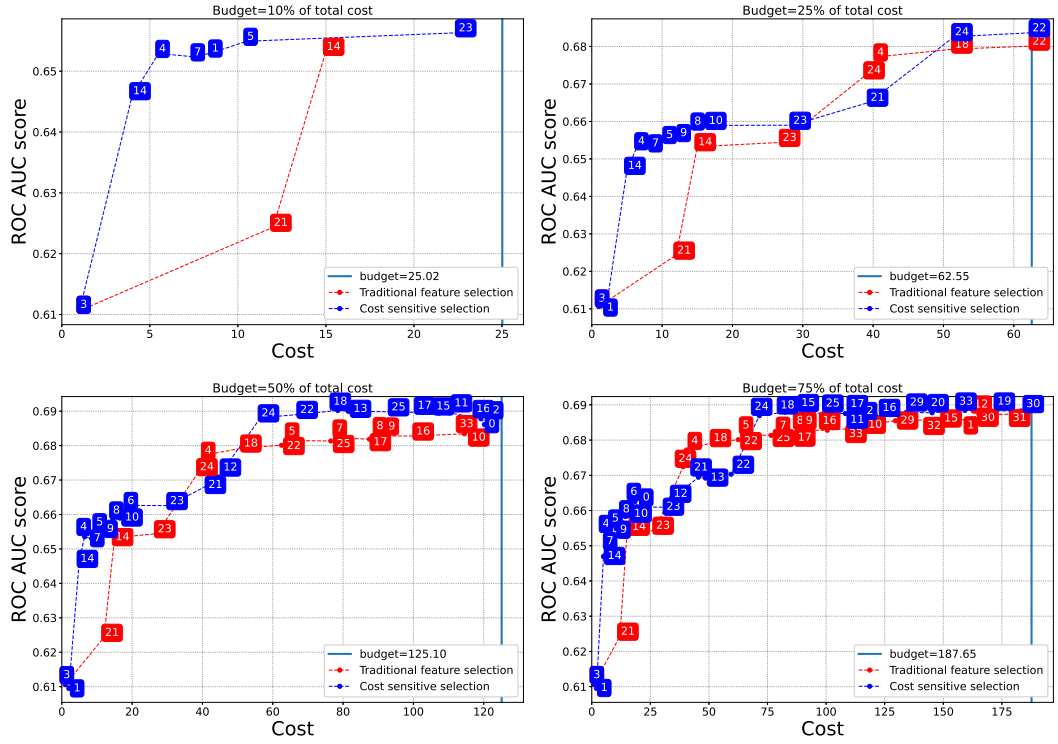


FIGURE 5.10: Feature selection for MIMIC-II dataset.

cost-constrained methods. The model which was used as a classifier was a simple logistic regression. The figure shows how the ROC AUC depends on the number of features used to train the model. The cost factor parameter λ_{opt} is calculated for each budget as in 5.3, 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 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 *calcium* (*8.4 - 10.2*) and the cost-constrained method selects *NBP systolic*, which is really cheap and has a high positive value of the mutual information with the target. In the next steps, the cost-constrained algorithm favors cheap features with moderate value of the mutual information, which explains why *mean heart rate* or *mean platelet volume in blood* are selected. The most important observation is that the cost-constrained feature selection method achieves a higher score when the budget is low. For higher budgets, the performance of both approaches is comparable. 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-constrained methods select features that serve as cheaper substitutes for the relevant expensive features.

Chapter 6

Group cost-constrained feature selection

In many real-world applications, especially in domains such as healthcare, finance, and bioinformatics, features are naturally organized into groups, and each group may be associated with distinct acquisition or computational costs. Unlike in the previous chapter, costs are assigned to the entire group of features, not to individual features. Incurring the cost of a group, we gain access to all features belonging to a given group. Importantly, we incur the same cost regardless of whether we use one variable from a given group or all variables from that group. For example, in medical applications, a group may include various parameters related to a general blood test. By paying the cost of a general blood test, we get access to various parameters such as Red Blood Cell Count (RBC), Hemoglobin (Hb), Hematocrit (Hct), among many others. These parameters form a group. Within this framework, many situations are possible. A group may contain only relevant or irrelevant variables or some relevant and some irrelevant variables. Furthermore, variables may interact within a group, but interactions between variables from different groups are also possible. It may happen that only one variable from the group is informative and the others are strongly correlated with it and therefore redundant. In such a situation, considering the entire group does not make sense, it is enough to use one variable from the group. The groups may also differ in size. Of course, in the situation where each group contains only one variable, the problem is reduced to that considered in the previous chapters.

6.1 Problem statement

We assume that the feature set $\mathcal{F}$ consists of K disjoint groups $G_1, \ldots, G_K$, where $\mathcal{F} = G_1 \cup \ldots \cup G_K$ and the intersection of any two groups is empty, i.e., $G_i \cap G_j = \emptyset$ for $i \neq j$. Each group G_i has an associated cost $c(G_i)$. When the value of one feature from a group is acquired, the values of all other features in the same group are obtained at no additional cost. The total cost of a subset of features $S \subseteq \mathcal{F}$ is given by:

$$c(S) = \sum_{k=1}^{K} c(G_k)\mathbb{1}(\exists j \in S : j \in G_k), \tag{6.1}$$

where $\mathbb{1}(A) = 1$, if event A occurs.

Within the information-theoretic framework, the grouped cost-constrained feature selection (GCC-FS) problem can be formulated as:

$$S_{opt} = \underset{S:c(S)\leq T}{\arg\max}\, I(Y, X_S), \tag{6.2}$$

where T, as previously, represents a user-defined maximum budget, and X_S denotes the feature vector corresponding to the subset $S \subseteq \mathcal{F}$. Intuitively, this problem seeks to identify a subset of features that maximizes the dependence on the label vector while ensuring that the total cost of the selected features does not exceed the specified budget T.

6.2 Sequential forward selection for GCC-FS

Since solving (6.2) would require an exhaustive search across all possible subsets of features (which is computationally infeasible due to the 2^p combinations), researchers typically utilize sequential forward selection methods. These methods assess the relevance of candidate features one by one, given the features already selected. In this section we propose two sequential methods that are able to select most important features, taking into account that the features are grouped. The methods are described in [A3].

The first method, called **Cost-Constrained Single Feature Selection (CC-SFS)**, is grounded in the sequential addition of individual features. Let S denote the set of features selected in the preceding steps. We define a contextual cost function for the candidate feature $k \in \mathcal{F} \setminus S$, taking into account what features were selected in the previous steps:

$$c(k, S) = \begin{cases} 0 \text{ if } k \in G \text{ and } \exists j \in S : j \in G \\ c(G) \text{ if } k \in G \text{ and } \nexists j \in S : j \in G. \end{cases}$$

If the candidate feature X_k belongs to a group G from which at least one feature has already been selected, then the cost associated with adding X_k is zero. Otherwise, the cost incurred corresponds to that of the group G. We initialize with an empty feature set, $S = \emptyset$, and iteratively add a candidate feature according to the update rule $S \leftarrow S \cup \{k_{opt}\}$, where

$$k_{opt} = \underset{k\in\mathcal{F}\setminus S}{\arg\max}\, \left[I(X_k, Y \mid X_S) - \lambda c(k, S)\right], \tag{6.3}$$

and $\lambda > 0$ is a cost-factor that balances the relevance of the feature against its associated cost. The selection process continues until the accumulated cost exceeds the predefined budget, that is, $c(S \cup \{k_{opt}\}) > T$, where $c(S)$ is defined in *Equation* (6.1). As with the methods described in the previous chapter, estimating $I(X_k, Y|X_S)$ is difficult due to the dimension of S. We therefore use the same approach as described in Chapter 5.2, which involves using a lower bound on the mutual information. This allows us to replace (6.3) by the following

$$k_{opt} = \arg\max_{k \in \mathcal{F} \setminus S} \left[J_{a,b}(Y, S, k) - \lambda c(k, S) \right],$$

where $J_{a,b}(Y, S, k)$ is defined in (5.9). In particular, for $q = 1$ and $a = 2$, we get

$$\begin{aligned} k_{opt} &= \arg\max_{k \in \mathcal{F} \setminus S} \left[\sum_{i \in S} I(Y, (X_k, X_i)) - \lambda c(k, S) \right] \\ &= \arg\max_{k \in \mathcal{F} \setminus S} \left[\sum_{i \in S} I(Y, X_k | X_i) - \lambda c(k, S) \right] \end{aligned} \tag{6.4}$$

The only difference between (6.4) and (5.11) lies in using $c(k, S)$ instead of c_k. The above criterion is used in the experiments. The **Cost-Constrained Group Feature Selection** (CC-GFS) is the second approach that considers the whole group structure of the features and operates by adding an entire group of features at each iteration, rather than individual features. The procedure begins with the empty set $S = \emptyset$, and at each step, a candidate group is selected and added according to the update rule $S \leftarrow S \cup \{G_{opt}\}$, where

$$G_{opt} = \arg\max_{G} \left[I(X_G, Y|X_S) - \lambda c(G) \right], \tag{6.5}$$

where $\lambda > 0$ is a parameter controlling the trade-off between the relevance of the group of features and its cost. The candidate groups are added until we exceed the budget, i.e., $c(S) > T$.

Selecting groups of features instead of individual features presents both advantages and limitations. The primary advantage lies in the ability to capture synergistic interactions among features within a group. For instance, consider a group $G = \{1, 2\}$ such that the target variable is defined as $Y = \text{XOR}(X_1, X_2)$. In this case, $I(X_1, Y) = I(X_2, Y) = 0$, meaning that neither X_1 nor X_2 would be identified as relevant by the Single Feature Selection method. However, the joint mutual information $I(X_G, Y) > 0$, and thus group G would be recognized as relevant by the Group Feature Selection method.

Despite its advantages, the group-based selection strategy also introduces potential drawbacks. For example, a group might contain only a single informative feature, with the remaining features being pure noise. Additionally, including an entire group may lead to the selection of redundant features, especially when strong correlations exist among features within the group. Moreover, some features in a candidate group may be highly correlated with those already

selected, which can further increase redundancy. The inclusion of excessive redundant or irrelevant features may negatively impact model performance.

To address these concerns, it is necessary to introduce an elimination step to prune redundant features from the set selected by the Group Feature Selection method. A natural criterion for elimination is to remove a feature $r \in S$ such that

$$I(X_r, Y \mid X_{S\setminus\{r\}}) = 0, \tag{6.6}$$

i.e., the feature provides no additional information given the remaining selected features.

However, evaluating condition (6.6) is often challenging. First, the true value of the conditional mutual information is unknown and must be estimated, which becomes increasingly difficult as the size of S grows. Second, verifying (6.6) typically involves statistical hypothesis testing, which can be unreliable in settings with small sample sizes or large conditioning sets.

As a practical alternative, condition (6.6) can be approximated by a computationally simpler criterion. One such heuristic is to eliminate feature r if it is strongly associated with any other feature X_j in the set $S \setminus \{r\}$. Specifically, feature r is removed when

$$\frac{I(X_r, X_j)}{H(X_r)} \geq 1 - \varepsilon, \tag{6.7}$$

for some $j \in S \setminus \{r\}$, where $\varepsilon > 0$ is a parameter, which defines a threshold $\tau = 1 - \varepsilon$. The following Lemma provides a formal justification for using (6.7) as a surrogate for the more stringent condition in (6.6).

Lemma 1. Let $r \in S$. Assume that $I(X_r, X_j)/H(X_r) \geq 1 - \varepsilon$ for $j \in S \setminus \{r\}$, $H(X_r) > 0$ and an arbitrarily small $\varepsilon > 0$. Then $I(X_r, Y|X_{S\setminus\{r\}}) \leq \varepsilon H(X_r)$.

Proof. Using the assumption we have

$$\frac{I(X_r, X_j)}{H(X_r)} = \frac{H(X_r) - H(X_r|X_j)}{H(X_r)} = 1 - \frac{H(X_r|X_j)}{H(X_r)} \geq 1 - \varepsilon$$

and thus $H(X_r|X_j) < \varepsilon H(X_r)$. Using this and the fact that conditioning on a smaller subset of features increases the entropy, we have

$$\begin{aligned} I(X_r, Y|X_{S\setminus\{r\}}) &= H(X_r|X_{S\setminus\{r\}}) - H(X_r|X_{S\setminus\{r\}}, Y) \\ &\leq H(X_r|X_{S\setminus\{r\}}) \leq H(X_r|X_j) \leq \varepsilon H(X_r). \end{aligned} \tag{6.8}$$

□

It follows from Lemma that $\varepsilon \approx 0$ implies that $I(X_r, Y|X_{S\setminus\{r\}}) \approx 0$. Finally, we will discuss the practical use of the criterion (6.5). As in all previous methods, it is problematic to calculate

the candidate relevance term, which in this case is equal to $I(X_G, Y|X_S)$. In this method, this is a challenge both with respect to the dimension of S and with respect to the dimension of G. Our approach, similar to previous methods, relies on Theorem 1 and mutual information lower bound. We describe the details below. First note that criterion (6.5) can be written as

$$\begin{aligned} G_{opt} &= \arg\max_G \left[I(X_G, Y|X_S) - \lambda c(G)\right] \\ &= \arg\max_G \left[I(X_{S\cup G}, Y) - I(X_S, Y) - \lambda c(G)\right] \\ &= \arg\max_G \left[I(X_{S\cup G}, Y) - \lambda c(G)\right], \end{aligned}$$

where the last equality follows from the fact that $I(X_S, Y)$ does not depend on the candidate group G and thus can be omitted. It follows from Theorem 1 that the first term, can be bounded from below as follows

$$I(X_{S\cup G}, Y) \geq \frac{1}{\binom{|S\cup G|}{a}} \frac{1}{\binom{q}{b}} \sum_{A\subseteq S\cup G:|A|=a} \sum_{B\subseteq \mathcal{L}:|B|=b} I(X_A, Y_B),$$

where a and b are parameters. Our idea is to replace $I(X_{S\cup G})$ with a term that is proportional to the lower bound, which leads to

$$G_{opt} = \arg\max_G \left[I_{a,b}(X_{S\cup G}, Y) - \lambda c(G)\right],$$

where

$$I_{a,b}(X_{S\cup G}, Y) = \sum_{A\subseteq S\cup G:|A|=a} \sum_{B\subseteq \mathcal{L}:|B|=b} I(X_A, Y_B)$$

Consider the situation of single-label classification ($q = 1$) and $a = 2$. Then we obtain

$$\begin{aligned} &I_{a,b}(X_{S\cup G}, Y) = \\ &\sum_{i\in G, j\in S} I(Y, (X_i, X_j)) + \sum_{i,j\in G:i<j} I(Y, (X_i, X_j)) + \sum_{i,j\in S:i<j} I(Y, (X_i, X_j)). \end{aligned}$$

Note, however, that the last term does not depend on the candidate group G, so it can also be omitted. This leads to the final criterion

$$G_{opt} = \arg\max_G \left[\sum_{i\in G, j\in S} I(Y, (X_i, X_j)) + \sum_{i,j\in G:i<j} I(Y, (X_i, X_j)) - \lambda c(G)\right] \tag{6.9}$$

If all groups consist of only one variable, criterion (6.9) reduces to (5.11). Criterion (6.9) is used in the numerical experiments described below.

TABLE 6.1: Selected feature groups and their costs.

Group Name	Description	Cost	# Features
A	Administrative data from the interview (e.g., age, gender).	1.0	9
NBP	Non-invasive blood pressure	3.0	8
RL	RLL & RUL lung sounds frequency	9.0	4
UN	Urea nitrogen in serum or plasma	12.0	4
VR	Verbal response	2.0	3
HR	Heart rate	1.5	4
P	Platalets in blood	2.0	4

6.3 Experiments on sequential forward selection

The main objective of the experiments was to evaluate and compare the performance of the proposed cost-aware feature selection methods, CC-SFS (Cost Constrained Single Feature Selection) and CC-GFS (Cost Constrained Group Feature Selection), described in Section 6.2. As a baseline, we included a traditional feature selection approach that ignores group costs. Specifically, we used the JMI criterion [154] as a representative of such methods, which corresponds to applying CC-SFS with $\lambda = 0$.

Once again we conducted our experiments using the MIMIC dataset [123], for dataset description see Section 5.4.8. In this experiment we focused on predicting hypertension using $p = 305$ features, primarily derived from diagnostic test results. Crucially, the original features in the dataset have expert-assigned costs. In this work, we extend the original setup by grouping related features based on their source and assigning costs at the group level. Most groups contain four statistical measures (mean, median, standard deviation, and range) of a single medical parameter monitored over time; examples include creatinine or glucose levels in serum or plasma. Additionally, we define a group for administrative data, which includes basic demographic attributes such as age, gender, and marital status, typically collected during a patient interview. The full group cost assignments are available in the GitHub repository[1], and the selected groups along with their associated costs are listed in Table 6.1. Prior to running the algorithms, all costs are normalized to the range $[0, 1]$.

To evaluate the quality of the selected feature subsets, we used a logistic regression model. Feature selection and model training are performed on 80% of the data (training set), while the performance is assessed on the remaining 20% (test set) using the ROC AUC metric. The optimal value of λ is chosen through 5-fold cross-validation. Figure 6.1 shows the ROC AUC scores for models trained on features selected under different budget levels T. When the budget is limited to 10% of the total cost, the CC-GFS algorithm significantly outperforms the others, which can be attributed to its strategy of selecting entire feature groups in each iteration. As the budget

[1]MIMIC-II group costs: https://github.com/Kaketo/mimic-II-group-costs

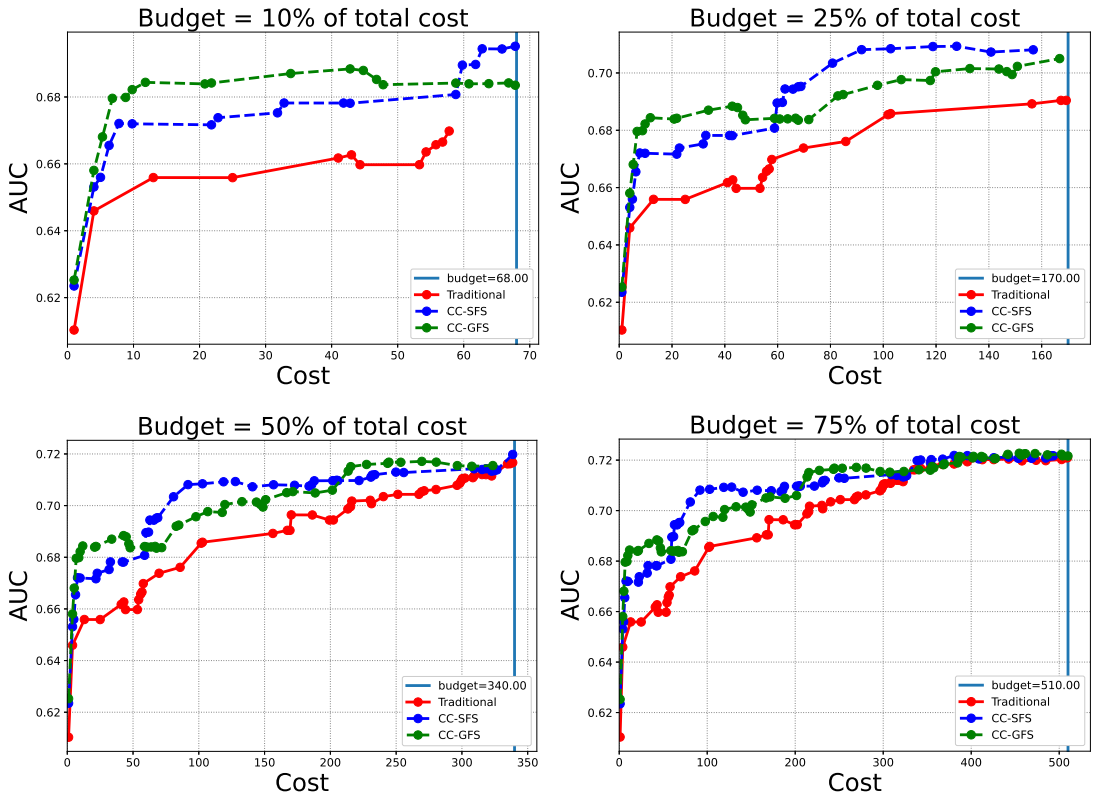


FIGURE 6.1: CC-SFS and CC-GFS for MIMIC-II dataset (hypertension).

increases to 25%, CC-SFS shows the best performance among all methods. At a 50% budget, all cost-constrained methods yield similar performance, though still outperforming the traditional method. For the highest considered budget (75%), all methods achieve similar AUC scores, indicating that most relevant features have already been selected by that point. Table 6.2 shows the features or groups chosen in the initial five steps of each algorithm. In the early stages, all methods consistently select features from the administrative and non-invasive blood pressure groups, which is expected given the known relationship between hypertension, age, and blood pressure. However, the selection patterns diverge in subsequent steps. The traditional method tends to favor features that are highly informative but also expensive, while CC-SFS focuses on features from already selected groups, thereby keeping the overall cost increase moderate. In contrast, CC-GFS continues to select entire groups in each iteration, which often results in the inclusion of multiple informative and typically low-cost features at once.

Experimental results on the large-scale MIMIC clinical dataset demonstrate that both methods significantly outperform traditional approach that ignores group structure and associated costs. The CC-GFS methods are particularly effective in low-budget scenarios, where their advantages are most evident. Among the two, CC-GFS typically selects a larger number of features and tends to achieve slightly better performance than CC-SFS when the available budget is low.

TABLE 6.2: Features selected in the first five steps.

	Step	1	2	3	4	5
Traditional	Num. of feat.	1	1	1	1	1
	Cost	1.0	3.0	9.0	12.0	0.0
	Group	A	NBP	RL	UN	NBP
CC-SFS	Num. of feat.	1	1	1	1	1
	Cost	1.0	0.0	0.0	3.0	0.0
	Group	A	A	A	NBP	NBP
CC-GFS	Num. of feat.	8	8	2	4	4
	Cost	1.0	3.0	2.0	1.5	2.0
	Group	A	NBP	VR	HR	P

6.4 Two-step feature selection for GCC-FS using shadow features

One of the difficulties associated with the methods CC-SFS and CC-GFS, described in the previous sections, is the need to choose the regularization parameter λ. Moreover, the CC-GFS method runs the risk of selecting too many redundant variables because it relies on adding an entire group of variables at each step. This can be problematic, especially when the groups contain many variables.

To overcome the above challenges, we introduce a novel yet conceptually simple and effective method based on using so-called shadow features. The algorithm as well as the experiments are described in [A5]. The method works in two main steps. First, we apply a standard feature selection approach - method described in *Equation* (6.3) with $\lambda = 0$ which produces an initial set of selected features, S_1. Among the remaining features, $\mathcal{F} \setminus S_1$, we can identify two subsets: A and C. Set $A = \{j_1, \ldots, j_a\}$ includes features with zero cost, as they come from groups that already have at least one selected feature. On the other hand, set C includes features from entirely unused groups. Since selecting any feature from C would exceed the total budget T, these features can no longer be considered. Because features in A come at no extra cost, they can still help improve the quality of the selected subset. The main idea of our method is to add a portion of features from A to the initial set S_1, aiming to enhance the final feature subset without exceeding the budget.

A straightforward approach is to add feature X_k, where $k \in A$, that maximizes the conditional relevance measure $I(X_k, Y \mid X_{S_1})$, and then iteratively continuing this process for the remaining features in A. However, a key challenge is to decide when to stop. The set A can be large, and including all its features can result in an overly complex model, increasing the risk of overfitting in downstream classification tasks. To address this issue, we propose using shadow features to define a stopping criterion. At the empirical level, shadow features are constructed by randomly permuting the values of the original features. Specifically, for each feature in A, we consider a corresponding shadow feature. The set of shadow features is denoted by

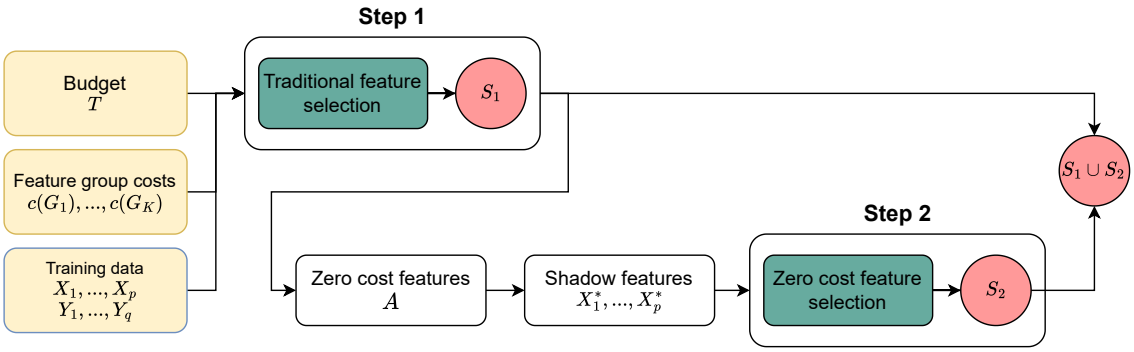


FIGURE 6.2: Flowchart of the Two-step feature selection for GCC-FS using shadow features

$X_{j_1}^*, \ldots, X_{j_a}^*$, where $a = |A|$. The shadow features are statistically independent of both the target variable Y and the original features $X_j, \ldots, X_p$, and therefore carry no predictive information about Y. The key advantage of shadow features is that their marginal distributions match those of the original features, i.e., $P(X_{j_k} = x) = P(X_{j_k}^* = x)$. This makes them good proxies for the original features in terms of distribution, while ensuring they are non-informative.

Additionally, we can formalize the fact that shadow features do not provide any useful information when combined with any selected subset of features, as shown in the following Lemma.

Lemma 6.1. Let X_k^* represent a shadow feature. Then, for any subset $S \subseteq \mathcal{F}$, we have $I(X_k^*, Y|X_S) = 0$.

Proof. Recall that $I(X_k^*, (Y, X_S)) = 0$ is equivalent to the following expression:

$$I(X_k^*, X_S) + I(X_k^*, Y|X_S) = 0, \tag{6.10}$$

which follows from the chain rule for mutual information [29]. By the definition of shadow features, we know that $I(X_k^*, (Y, X)) = 0$. Since independence $X_k^* \perp (Y, X)$ holds, it also implies $X_k^* \perp (Y, X_S)$ for any subset $S \subseteq \mathcal{F}$. Consequently, this leads to $I(X_k^*, (Y, X_S)) = 0$, which, according to (6.10), gives us $I(X_k^*, Y|X_S) = 0$. □

In the second stage of our method, we augment the initial feature set S_1 by adding a subset $S_2 \subseteq A$. We start with an empty set, i.e., $S_2 = \emptyset$, and iteratively add features according to the following rule: at each step, we select the feature

$$k_{\text{opt}} = \arg\max_{k \in A \setminus S_2} I(X_k, Y|X_{S_1 \cup S_2}).$$

Algorithm 1: Two-step GCC-FS using **shadow features**

Input : A label vector Y, features $X_1, \ldots, X_p$, budget T, group costs $c(G_1), \ldots, c(G_K)$
Output : The optimal set of features S_{opt}

```
# Step 1:
S_1 = ∅ #initialization
while c(S_1) ≤ T do
    k_opt = arg max_{k∈F\S_1} I(X_k, Y|X_{S_1}),
    S_1 ← S_1 ∪ {k_opt}.
end
# Step 2:
S_2 := ∅ #initialization
I*_max := 0 #initialization
A := {j_1, ..., j_a} # a set of zero-cost features
while A \ S_2 ≠ ∅ do
    k_opt = arg max_{k∈A\S_2} I(X_k, Y|X_{S_1∪S_2}).
    S_2 ← S_2 ∪ {k_opt}
    I*_max := max_{k∈A\S_2} I(X*_k, Y|X_{S_1∪S_2})
    if I*_max > I(X_{k_opt}, Y|X_{S_1∪S_2}) then
        break
    end
end
S_opt = S_1 ∪ S_2
return S_opt
```

and update the set as $S_2 \leftarrow S_2 \cup k_{\text{opt}}$. To decide when to stop adding features from A, we use shadow features as a reference. Specifically, we compute

$$I^*_{\text{max}} := \max_{k \in A \setminus S_2} I(X^*_k, Y | X_{S_1 \cup S_2}),$$

which is the maximum conditional mutual information among the shadow features. We stop the process when $I^*_{\text{max}} > I(X_{k_{\text{opt}}}, Y|X_{S_1 \cup S_2})$, meaning that the most informative shadow feature is more relevant (in terms of conditional mutual information) than the best remaining original feature. This suggests that further additions would likely be uninformative and possibly harmful due to overfitting. This stopping rule can be adjusted, for example, the algorithm could stop only if a certain percentage (e.g., 5%) of shadow features outperform the selected feature. The complete procedure is outlined in Algorithm 1. The computational complexity is driven by two main loops: one in Step 1 and the other in Step 2. In the worst case, Step 1 involves evaluating all features at each iteration, resulting in a complexity of $O(p^2)$, where p is the total number of features. Step 2 operates over the subset A of zero-cost features, and its complexity depends on the size of A. The diagram in Figure 6.2 shows the proposed algorithm in the graphical

form. Finally, we mention, that in practice, the conditional mutual information appearing in the above method is approximated using the methodology described for previous methods, which is based on the lower bound for the mutual information.

6.5 Experiments on two step feature selection

The main goal of the experiments was to assess the effectiveness of the method introduced in Section 6.4. For comparison, we employed methods based on Single Feature Selection (SFS) incorporating a cost penalty term λ from Section 6.2. Specifically, we examined scenarios with $\lambda = 0$, which represents traditional feature selection without any cost constraints, as well as scenarios with $\lambda = \lambda_{\max}$ and $\lambda = 0.5\lambda_{\max}$. The parameter $\lambda_{\max}$ was defined such that, in the initial selection step, the feature with the lowest cost would always be chosen ahead of higher-cost features, irrespective of their predictive value. Our intention was to demonstrate that prioritizing features purely based on minimal cost does not necessarily result in better model accuracy. The scenario using $0.5\lambda_{\max}$ represents an intermediate case, where the feature selection balances considerations of both cost and the correlation with the label vector.

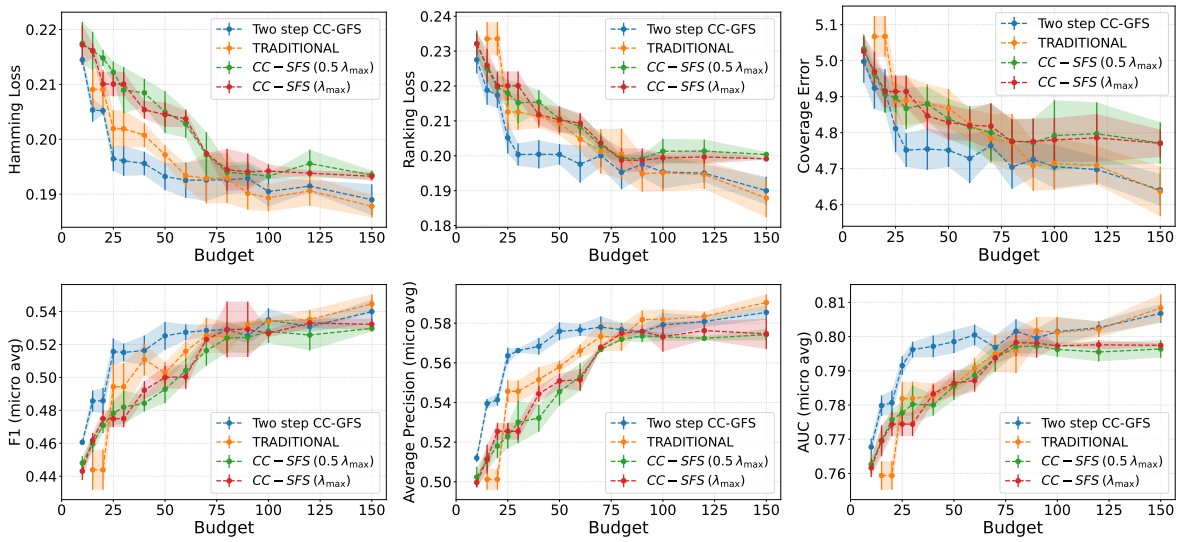


FIGURE 6.3: Performance of two-step CC-GFS using shadow features, CC-SFS and traditional feature selection method.

We conducted experiments using the MIMIC medical dataset [123], comprising data related to patients' medical conditions admitted to intensive care units (ICUs). Patients within the dataset were often diagnosed with multiple diseases, including hypertension (65% of patients), diabetes (31%), fluid imbalance (31%), lipid issues (30%), kidney disorders (29%), COPD (22%), thyroid problems (11%), hypotension (10%), liver conditions (6%), and thrombosis (5%). Unlike the experiments described in Section 6.3, we now consider the general case

of multi-label classification. Our goal is to predict the occurrence of the above 10 diseases in patients. Details about the features used and the assignment of costs to groups are described in the Section 6.3.

To evaluate the effectiveness of the selected feature subsets, we employed a multilabel k-Nearest Neighbours algorithm [161]. Feature selection and model training were both conducted using the training subset (80% of the data), while performance metrics were calculated using the validation set. This process of splitting data was repeated five times to estimate variability in the results. Figure 6.3 illustrates selected performance metrics for models trained on feature sets obtained under various budget constraints.

Formal definitions of these metrics can be found in Section 5.4.3; we also refer to [162]. For lower budgets (below 75), the proposed two-step algorithm significantly outperforms other methods across all metrics. However, when budgets increase to around 100, the performance of all methods becomes comparable, indicating that the most informative features have already been selected. The results suggest that the method based on shadow features (Two step CC-GFS) is particularly beneficial when predicting diseases based on inexpensive features.

Chapter 7

Model-based methods considering feature costs based on penalized empirical risk minimization

Feature costs can be incorporated into machine learning algorithms and feature selection algorithms in various ways. In this section, we discuss the embedded feature selection methods based on the general penalized empirical risk minimization (ERM) framework.

Currently, penalized empirical risk minimization (ERM) methods, employing penalties such as lasso [140], elastic net [54], or non-convex penalties like MCP and SCAD [159], hold a prominent position among feature selection techniques. They have become the benchmark in numerous medical applications, particularly when simultaneous prediction and feature selection are primary objectives. Penalized methods offer several significant advantages. Firstly, they are effective with high-dimensional datasets, unlike classical feature selection approaches based on information criteria [1] (e.g., AIC or BIC), which often fail when the number of features significantly exceeds the number of observations. Secondly, these methods are versatile, accommodating various empirical risk functions tied to different loss metrics, such as logistic, squared, or hinge loss. This flexibility enables their application across diverse supervised learning tasks, including regression, binary and multi-class classification, multi-label classification as well as survival analysis. Lastly, penalized ERM methods are categorized as embedded feature selection techniques, meaning they integrate feature selection directly with parameter estimation. This integrated approach is particularly advantageous when the goal is to construct a robust classification model without relying on external feature selection processes like filtering methods.

Penalized ERM methods have been effectively applied across numerous medical domains. Prominent examples include predicting SARS-CoV-2 pneumonia [151], breast cancer [69], real-time forecasting of endemic infectious diseases [25], cardiovascular disease prediction [68], identification of ischemic stroke [94], and forecasting visual field progression in glaucoma patients [45], among others. Additionally, several studies utilizing elastic net and non-convex

penalties have addressed tasks such as breast cancer survival prediction [113] and penile cancer detection [109].

The literature on cost-constrained variants of penalized ERM methods is much more limited. A cost-sensitive adaptation of the lasso for logistic regression was introduced by [15] and also used in [139] in the context of multi-label classification. Despite the widespread popularity of penalized ERM approaches, comprehensive analyses and comparisons of cost-aware penalized methods remain limited. In this chapter, we introduce modifications to non-convex penalties (such as MCP) that take into account feature costs and compare them with cost-sensitive lasso, cost-sensitive adaptive lasso and the method of cheap knockoffs [156].

7.1 Problem statement

The objective in this problem is to learn a supervised model that predicts the target variable for new instances using a subset of features whose total cost does not exceed a predefined budget T. In this chapter, we focus on the single label case, therefore $y^i \in \mathbb{R}$. Moreover, we focus on the linear predictor, that is, the prediction is a certain function of the linear combination of the features $(x^i)^T\beta$. The x^i is the feature vector and β is an arbitrary parameter vector. Additionally, we assume that the first coordinate of x^i is equal 1, which corresponds to the intercept. The quality of prediction $\hat{y}^i$ is measured using the loss function $l(y^i, (x^i)^T\beta)$. Commonly used loss functions include the squared loss, $l(y^i, (x^i)^T\beta) = (y^i - (x^i)^T\beta)^2$, which is typical in regression tasks, the logistic loss, $l(y^i, (x^i)^T\beta) = -[y^i \log(\sigma((x^i)^T\beta)) + (1 - y^i)\log(1 - \sigma((x^i)^T\beta))]$, where $\sigma(s) = (1 + \exp(-s))^{-1}$, corresponding to logistic regression, and the hinge loss, $l(y^i, (x^i)^T\beta) = \max\{0, 1 - y^i(x^i)^T\beta\}$, commonly used in support vector machines (SVM). A smaller loss value indicates that the predicted output is close to the actual target value. Within the Empirical Risk Minimization (ERM) framework for linear prediction [54], the goal is to minimize the empirical risk with respect to β:

$$\hat{R}(\beta) := \frac{1}{n}\sum_{i=1}^{n} l(y^i, (x^i)^T\beta),$$

which serves as an estimate of the theoretical risk defined by $R(\beta) = \mathbb{E}_{Y,X}[l(Y, X^T\beta)]$. Learning with a budget constraint T can be framed as a constrained optimization problem, where the goal is to solve

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min}\ \hat{R}(\beta) \quad \text{subject to} \quad \sum_{j=2}^{p} c_j \, \mathbb{1}\left[|\beta_j| \neq 0\right] \leq T, \tag{7.1}$$

where $\mathbb{1}(A)$ denotes the indicator function, i.e., $\mathbb{1}(A) = 1$ if event A occurs. Note that the summation is performed over the indices $2, \ldots, p$, because the first coordinate x^i is equal to one, which refers to the intercept, not the proper variable. The constraint in (7.1) ensures that the total cost of the selected features does not exceed the budget T. This formulation corresponds to identifying the best model according to empirical risk minimization under a predefined cost

limitation. When all feature costs are equal, i.e., $c_2 = \ldots = c_p$, the constrained optimization problem simplifies to the classical best subset selection problem [54]. In this setting, the objective is to find the optimal model using a limited number of features. The original formulation in (7.1) can be equivalently expressed in a penalized form:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \hat{R}(\beta) + \lambda \sum_{j=2}^{p} c_j \mathbb{1}\left[|\beta_j| \neq 0\right] \right\}, \tag{7.2}$$

where the regularization parameter $\lambda > 0$ controls the trade-off between prediction accuracy and total feature cost, and plays a role analogous to the budget T. More precisely, for any given budget T, there exists a corresponding value of $\lambda > 0$ such that the constrained problem and its penalized counterpart yield the same solution, and vice versa [53]. The second term in (7.2) serves as a cost-based penalty, with the parameter λ controlling the trade-off between model fit and total feature cost. Larger values of λ encourage the selection of less costly models. While this penalized formulation appears promising, it suffers from a major limitation: the optimization problem in (7.2) is non-convex and computationally intractable for even moderately sized feature sets, as it is known to be NP-hard [53]. This computational difficulty stems from the use of an ℓ_0-type penalty. In the following sections, we explore alternative penalty functions that offer more practical solutions.

7.2 Cost-sensitive lasso

A natural way to relax the optimization problem in (7.2) is to replace the computationally intractable ℓ_0-type penalty with an ℓ_1-type penalty. This leads to the cost-sensitive lasso formulation, where the objective becomes:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \hat{R}(\beta) + \lambda \sum_{j=2}^{p} c_j |\beta_j| \right\}. \tag{7.3}$$

More generally, one could consider an ℓ_q norm penalty with $q \geq 0$. However, lasso is particularly appealing because $q = 1$ is the smallest value that yields a convex constraint region, making the optimization problem convex and computationally tractable. In this sense, it provides the closest convex relaxation of the best subset selection problem. The cost-sensitive lasso has been applied in practice, for example in [15], where penalty weights c_j were assigned not to individual features, but to groups corresponding to different data modalities such as clinical, gene expression, methylation, and copy number variation.

Choosing the optimal regularization parameter λ is critical. In standard feature selection, this is typically done via cross-validation. However, in cost-constrained settings, the situation is more nuanced, as the total cost incurred by the selected feature subset depends directly on

the value of λ. A sufficiently large value of λ will shrink all coefficients to zero, resulting in zero total cost. This upper-bound value can be derived analytically. Conversely, a very small λ yields many non-zero coefficients and, therefore, a higher cost. In our experiments, we employ the following strategy. We define a decreasing sequence of regularization parameters $\lambda_1 > \lambda_2 > \ldots > \lambda_L$, and compute the corresponding costs $C(\lambda_1), \ldots, C(\lambda_L)$ for the selected features. We set $L = 100$ in our implementation. We then identify the largest index k such that the cost at λ_k satisfies $C(\lambda_k) \leq T$, while $C(\lambda_{k+1}) > T$. From the sequence $\lambda_1, \ldots, \lambda_k$, we select the value of λ that yields the best performance according to the evaluation metric used (e.g., AUC).

7.3 Cost-sensitive adaptive lasso

The core idea of the adaptive method is to adjust the penalty factors for individual features based on their estimated relevance. The approach consists of two main steps. In the first step, we fit a univariate model for each feature independently. Specifically, for each $j = 2, \ldots, p$, we solve:

$$\hat{\beta}_j^{(0)} = \arg\min_{\beta_j} \hat{R}_j(\beta_j), \quad \text{where } \hat{R}_j(\beta_j) = \frac{1}{n} \sum_{i=1}^{n} l(y^i, x_j^i \beta_j).$$

In the second step, we solve the following penalized optimization problem:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \hat{R}(\beta) + \lambda \sum_{j=2}^{p} \frac{c_j}{1 + |\hat{\beta}_j^{(0)}|} |\beta_j| \right\}.$$

This adaptive weighting scheme reduces the penalty for features that exhibit stronger univariate relationships with the target variable. In particular, if feature j is highly predictive, then $|\hat{\beta}_j^{(0)}|$ will be large, resulting in a smaller penalty term for that feature in the second step. While the initial step here uses a univariate approach to compute $\hat{\beta}_j^{(0)}$, alternative methods such as ordinary least squares (OLS) may be employed, especially in the context of regression [166].

7.4 Cost-sensitive non-convex penalties

We also investigate non-convex penalty functions, with particular focus on the Minimax Concave Penalty (MCP), a representative and highly promising method from this class [159]. In the cost-constrained formulation of MCP, the optimization problem is defined as:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \hat{R}(\beta) + \sum_{j=2}^{p} c_j \, P(\beta_j, \lambda, \gamma) \right\},$$

where the penalty function $P(\beta_j, \lambda, \gamma)$ is given by:

$$P(\beta_j, \lambda, \gamma) = \begin{cases} \lambda|\beta_j| - \dfrac{\beta_j^2}{2\gamma}, & \text{if } |\beta_j| \leq \gamma\lambda, \\ \dfrac{1}{2}\gamma\lambda^2, & \text{if } |\beta_j| > \gamma\lambda, \end{cases}$$

and $\gamma > 0$ is a tuning parameter that controls the concavity of the penalty.

Figure 7.1 compares three penalty functions: the ℓ_0-type penalty, the ℓ_1 lasso penalty, and the MCP. While the lasso penalty increasingly deviates from the ℓ_0 penalty as the magnitude of the coefficients grows, the MCP begins with the same rate of penalization as the lasso but gradually reduces it, eventually applying no additional penalty to large coefficients. This behavior makes MCP a closer and more flexible approximation to the ideal ℓ_0-type penalty.

Our experimental results confirm the effectiveness of MCP in the cost-constrained setting. We also consider an adaptive version of MCP, analogous to the adaptive lasso, where the second step employs MCP instead of lasso. Although other non-convex penalties, such as the Smoothly Clipped Absolute Deviation (SCAD) [40], can also be applied, our experiments indicate that MCP consistently outperforms SCAD by a small margin. Therefore, we focus on presenting results for MCP in this thesis.

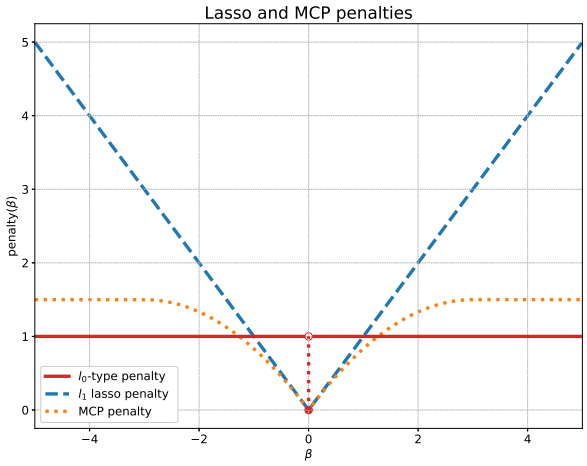


FIGURE 7.1: Lasso and MCP penalties for $\lambda = 1, \gamma = 3$

7.5 Experiments

The primary objective of the experiments is to investigate the differences between cost-constrained feature selection methods formulated within the penalized empirical risk

minimization framework and conventional feature selection approaches that disregard feature cost information. As the representative of the traditional approach, we adopt the lasso algorithm defined as

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \hat{R}(\beta) + \lambda \sum_{j=2}^{p} |\beta_j| \right\}.$$

In our experimental study, the standard lasso serves as the baseline. Additionally, we examine the performance of the *cheap knockoff* method [156], which is likewise built upon the lasso framework. To ensure a fair comparison, the experiments are restricted exclusively to methods grounded in penalized empirical risk minimization, deliberately excluding other categories of cost-constrained techniques such as filter-based methods discussed in previous chapters, which are independent of any specific classification model. In all evaluated approaches, the logistic loss function is employed. Several key research questions are addressed in the experiments: (i) what is the predictive performance of the analyzed methods under budget constraints, and which penalty configuration yields optimal results; (ii) to what extent the incorporation of cost information improves conventional feature selection; and (iii) what proportion of the total cost is expended on non-informative features.

7.5.1 Datasets

We evaluate the performance of cost-constrained feature selection methods on real-world medical datasets. The experimental study includes four distinct datasets. In the majority of cases, feature values correspond to the outcomes of medical diagnostic tests, while the binary target variables indicate the presence or absence of specific diseases. Table 7.1 provides a summary of the key characteristics of the analyzed datasets. Furthermore, in MIMIC dataset, feature costs have been assigned by domain experts. For the remaining datasets, artificial costs are generated according to the procedure described in Section 7.5.2, using **Strategy C2**. Recall, that in **Strategy C2**, costs are assigned proportionally to the value of mutual information between the class variable and individual features. As a result, features exhibiting stronger dependence on the target variable are associated with higher costs. This procedure is motivated by the observation that features containing valuable information (e.g., advanced diagnostic tests) are typically more expensive. At the same time, we acknowledge that certain easily accessible and inexpensive features, such as the patient's age, may also serve as important predictive factors. Nevertheless, this cost generation strategy appears more justified than, for instance, assigning costs at random.

The first real dataset is a clinical database MIMIC [123], discussed already in previous chapters (see detailed description in Section 5.4.8). In the experiments, described in this chapter, we focus on diabetes disease, which serves as a target variable. In the experiments, we selected a subset of 30 features exhibiting the highest value of the mutual information with the target

TABLE 7.1: Summary of real datasets.

Dataset	Costs	Original features	All features*	Observations	Positive observations
MIMIC (diabetes)	assigned by experts	30	90	19773	30.63%
Heart dataset	generated artificially	13	39	270	44.44%
Thyroid dataset	generated artificially	20	60	3772	6.12%
Alzheimer dataset	generated artificially	9	27	373	39.14%

* All features include original, proxy and noise features (see Section 7.5.2 for details).

variable. Figure 7.2a shows both the assigned feature costs and the mutual information values between the features and the considered target variable.

The second dataset is the widely used Cleveland heart disease dataset [38], which focuses on predicting the presence of heart disease. It contains 270 observations and 13 features, including basic patient information (e.g., age and sex), electrocardiographic results, and blood test outcomes. The third dataset concerns thyroid disease prediction [115]. It originates from the Garavan Institute in Sydney, Australia, and contains 3772 samples, with thyroid disease diagnosed in 6.12% of the cases. In addition to basic medical interview data (e.g., age, pregnancy), the dataset includes results from blood diagnostic tests specifically related to thyroid function (e.g., TSH, TT4). The final dataset employed in the experiments is obtained from the Open Access Series of Imaging Studies (OASIS) [92]. It contains magnetic resonance imaging (MRI) data related to Alzheimer's disease. The features include administrative variables (e.g., age, years of education) as well as MRI-based measurements. However, the number of features in this dataset is relatively small.

7.5.2 Experimental framework

Within the proposed framework, we construct two additional groups of features. The first group consists of proxy features, denoted as $X'_1, \ldots, X'_p$, which are derived from the original features $X_1, \ldots, X_p$ in the same way as they were created in Section 5.4.1. In addition to the p proxy features, we generate another set of p noise features, denoted by $X''_1, \ldots, X''_p$, which are independent of the target variable Y (obtained by applying $\rho = 1$). This procedure increases the complexity of the feature selection problem. Consequently, the final dataset contains $3p$ features, where p is the number of original features. When feature costs are not available, they are generated according to the following procedure. The costs of the original features are defined based on their relevance, quantified via mutual information, see **Strategy C2** in Section 5.4.1. The costs for the proxy and noise features are defined as $c'_j = c''_j = \Psi \cdot c_j$, where $\Psi \in (0, 1)$ is a parameter controlling the cost ratio between the original features and their corresponding proxy or noise features. For instance, if $\Psi = 0.5$, the proxy feature cost is set to half of the original feature cost. In general, lower values of ρ combined with lower values of Ψ increase

the benefit of replacing original features with their proxy counterparts. We investigate how the parameter ρ influences the performance of the examined methods. This framework is designed to approximate real-world scenarios. For example, in medical diagnostics, one may choose between performing an expensive diagnostic test (original feature), which provides an accurate measurement, or a cheaper alternative (proxy feature), which yields an approximate value. An illustrative case is medical ultrasonography: 3D scans offer higher precision and effectiveness compared to conventional 2D scans, but they also incur higher costs. In this context, a 2D scan may be regarded as a lower-cost approximation of a 3D scan.

7.5.3 Evaluation measures

In classification problems, one of the most commonly used evaluation metrics is the AUC score, which assesses the quality of predictions. AUC, defined as the area under the receiver operating characteristic (ROC) curve, quantifies the classifier's ability to discriminate between classes. In contrast to many traditional metrics (such as accuracy, precision, or recall), AUC is independent of the threshold for posterior probabilities, making it a more general and robust performance measure. Beyond evaluating the predictive performance, it is also important to account for the costs associated with selecting irrelevant features. To this end, we propose a novel metric referred to as the Cost-Sensitive False Discovery Rate (CSFDR). The CSFDR quantifies the proportion of the total cost of selected features that is attributed to irrelevant variables. Formally, it is defined as follows. Let S denote the set of selected features, and let $C(S) = \sum_{j \in S} c_j$ represent the total cost of the features included in S. Let N denote the set of indexes corresponding to noise features $X_1'', \ldots, X_p''$ present in the dataset. The cost-sensitive false discovery rate (CSFDR) is defined as

$$\text{CSFDR}(S) = \frac{C(S \cap N)}{C(S)}. \tag{7.4}$$

A low value of CSFDR indicates that only a small portion of the total cost was allocated to noisy features. Conversely, when $\text{CSFDR}(S) \approx 1$, it suggests that the majority of the cost was spent on irrelevant variables. This measure can be directly computed, since the set N is known in advance within the experimental framework considered. The set N contains only irrelevant features, i.e., those that do not influence the target variable. However, for real datasets, the set of irrelevant features is not known a priori, and some of them may still be present among the original features $X_1, \ldots, X_p$. Consequently, in such scenarios, N represents only a subset of irrelevant features, and the CSFDR provides a lower bound on the true proportion of costs attributed to irrelevant variables. Nevertheless, even for real datasets, CSFDR remains a useful metric as it allows one to assess the portion of cost spent on artificially generated noise features.

7.5.4 Results

We report the results for the setting with $\Psi = 0.1$ and $\rho = 0.1$, which corresponds to the scenario where the costs associated with proxy and noisy features are substantially lower than those of the original features. In addition, proxy features exhibit high correlation with the original ones. This configuration is particularly interesting in our study, as it allows us to assess the potential advantages of cost-constrained methods under favorable conditions. The only parameter that differs in experiments is the number of original features p, because it is directly connected to the dataset itself. The results are shown in Figures 7.2, 7.3 and Tables 7.2, 7.3, 7.4, 7.5. We present the values, averaged over 100 simulations. We note substantial differences between the traditional method and cost-constrained approaches when the available budget is limited, specifically when it does not exceed $20\% - 40\%$ of the total cost. This finding is encouraging, as the superior performance of cost-constrained methods in low-budget regimes is a crucial advantage from the perspective of their practical application.

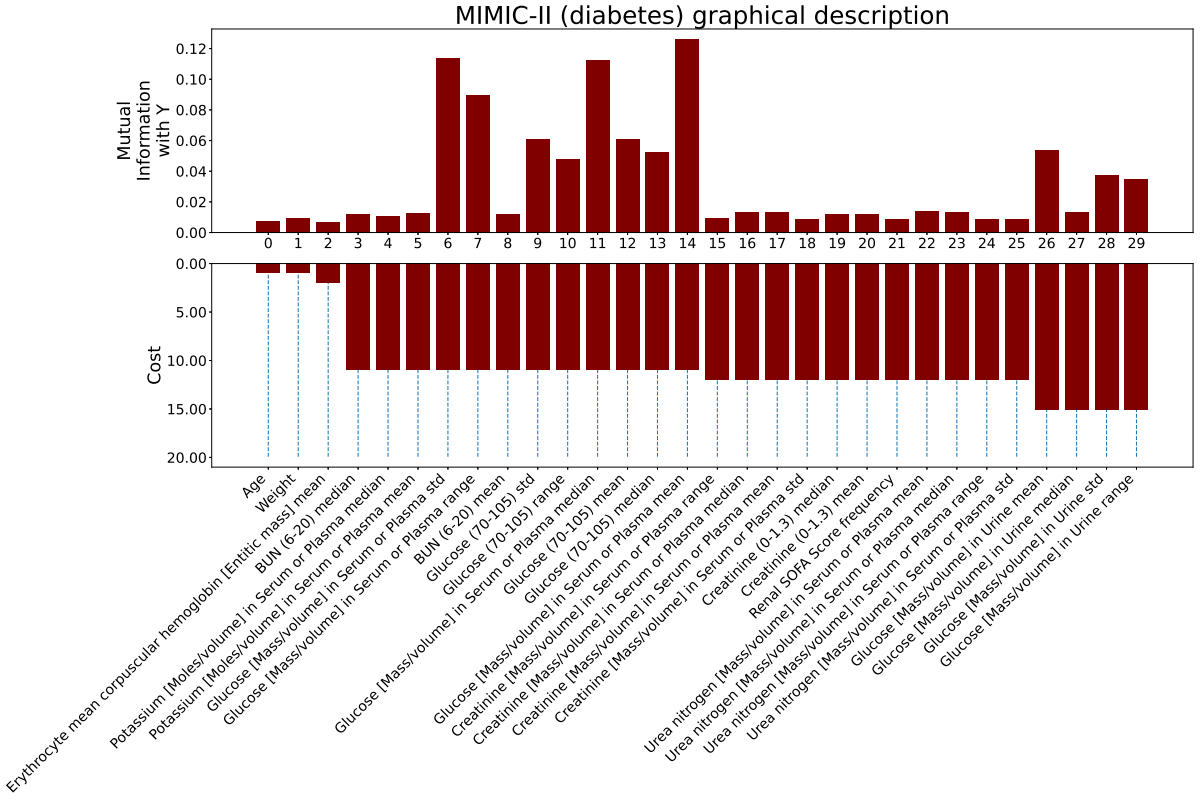
In the MIMIC dataset (Figure 7.2), the most notable difference is observed when the budget is approximately 10% of the total cost, the AUC achieved by the traditional lasso is around 0.8, whereas cost-constrained methods yield AUC values in the range of $0.8 - 0.9$. Furthermore, the cheap knock-off method demonstrates slightly inferior performance compared to the other cost-constrained methods under very limited budgets. However, as the budget increases, it tends to slightly outperform the remaining methods across most datasets. Notably, for larger values of budget, the differences among the methods become less pronounced.

In most datasets, the AUC corresponding to cost-sensitive lasso, adaptive lasso, MCP, and adaptive MCP reach their plateau earlier than those of the other methods. For the Alzheimer dataset (Figure 7.3c), both the traditional lasso and the cheap knock-off method reach the plateau relatively late, only when the budget reaches 25% of the total cost. Across the majority of experiments, no substantial differences in AUC are observed between the cost-sensitive lasso and adaptive lasso, nor between the cost-sensitive MCP and its adaptive variant. The most evident performance differences are observed for the heart dataset; see Figure 7.3a and Table 7.3.

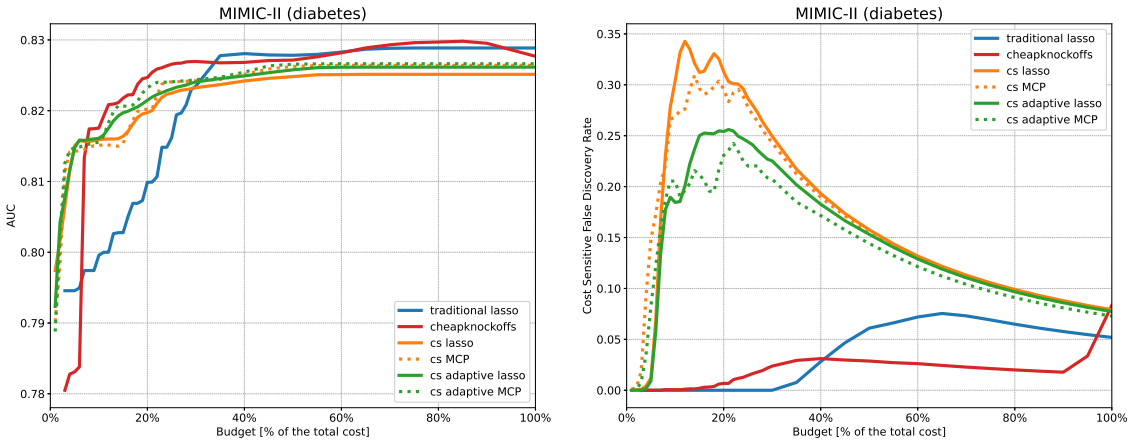
The analysis of CSFDR conducted on real datasets yields the conclusion that the CSFDR remains close to zero for both the traditional method and the cheap knock-off approach, indicating that these methods select only a small number of irrelevant features. However, this is accompanied by a substantially lower AUC, especially under limited budget conditions. Among the remaining cost-constrained methods, cost-sensitive MCP and its adaptive variant achieve notably lower CSFDR values compared to cost-sensitive lasso and adaptive lasso. For example, in the heart dataset (Figure 7.3), at $T = 0.2$, the CSFDR for cost-sensitive MCP is approximately three times lower than that obtained for cost-sensitive lasso. An interesting observation arises for the MIMIC diabetes dataset, where significant differences in CSFDR are observed both between lasso and adaptive lasso, and between MCP and adaptive MCP. Overall, the experimental

results suggest that cost-sensitive MCP may be recommended as the preferred method, as it offers a favorable trade-off between AUC maximization for small budgets and CSFDR minimization. Simultaneously, no consistent advantage of adaptive MCP over its standard counterpart is observed.

To gain a more detailed understanding of the behavior of the analyzed methods, we examine the types of features they tend to select. Figure 7.4 presents the selected features, categorized into three groups (original, proxy, and noisy features), for the MIMIC diabetes dataset with a budget of $T = 10\%$ of the total cost. Both the costs of the selected features and their mutual information with the target variable are reported. The traditional lasso selects only two original features, both of which are associated with high acquisition costs. The feature subset obtained by the cheap knock-off method includes three original features along with two inexpensive proxy features. In contrast, the feature subsets selected by the remaining cost-constrained methods differ substantially, containing a larger number of proxy features that compensate for the limited inclusion of original features. These proxy features are considerably cheaper than the original ones, while still exhibiting high correlation with the target variable. Moreover, it can be observed that the remaining cost-constrained methods also select some noisy features. In particular, cost-sensitive lasso selects 12 noisy features, whereas MCP includes only 7, which is consistent with the previously observed lower CSFDR values for MCP in comparison to cost-sensitive lasso.

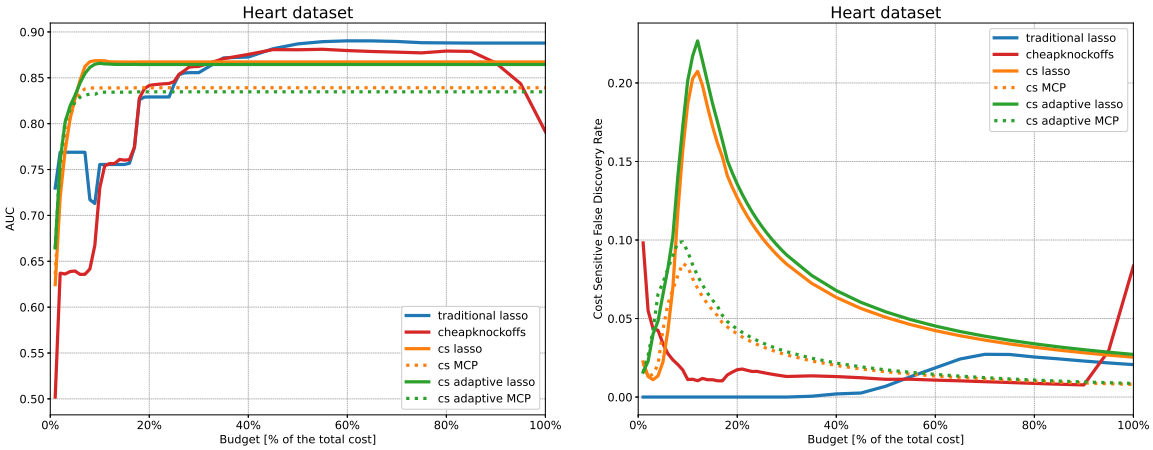


(A) Graphical description of features cost and relevance.

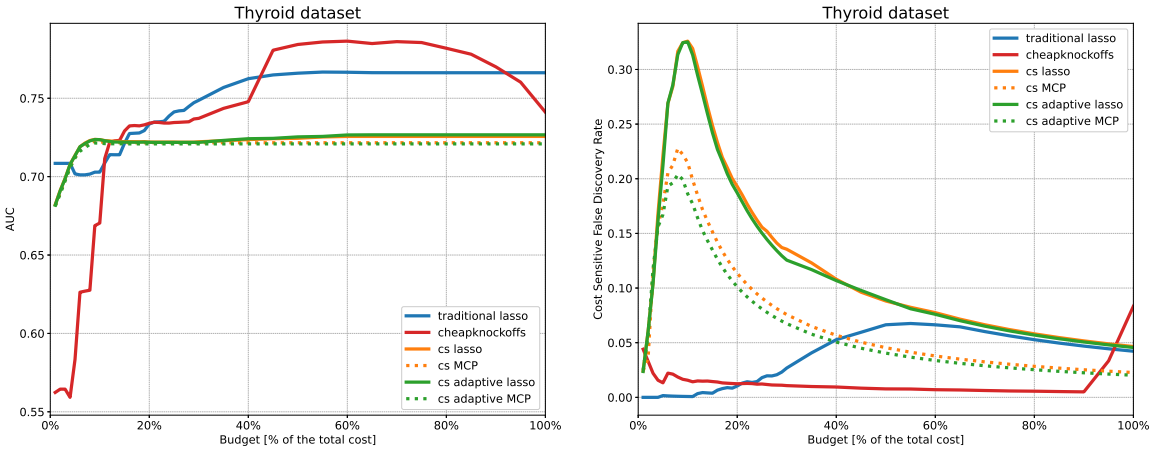


(B) AUC and CSFDR results ($p = 30$)

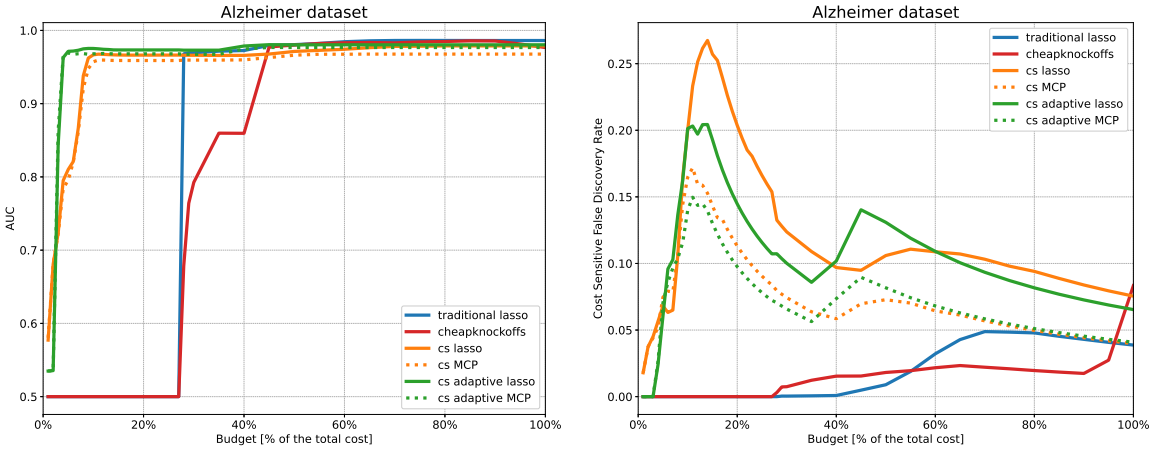
FIGURE 7.2: Experiment results for MIMIC-II (diabetes)



(A) AUC and CSFDR results for *heart* dataset ($p = 13$)



(B) AUC and CSFDR results for *thyroid* dataset ($p = 20$)



(C) AUC and CSFDR results for *alzheimer* dataset ($p = 9$)

FIGURE 7.3: Experiment results for *Heart*, *Thyroid* and *Alzheimer* datasets

TABLE 7.2: Mean and standard deviation of AUC for *MIMIC-II* (*diabetes*) dataset ($p = 30$).

Budget [%]	traditional lasso	cheapknockoffs	cs lasso	cs MCP	cs adaptive lasso	cs adaptive MCP
5	0.795 (±0.004)	0.783 (±0.009)	**0.815 (±0.004)**	0.814 (±0.003)	0.815 (±0.004)	0.815 (±0.003)
10	0.800 (±0.004)	**0.818 (±0.004)**	0.816 (±0.003)	0.815 (±0.003)	0.816 (±0.003)	0.816 (±0.004)
15	0.803 (±0.005)	**0.822 (±0.004)**	0.816 (±0.003)	0.815 (±0.004)	0.820 (±0.003)	0.821 (±0.003)
20	0.810 (±0.005)	**0.825 (±0.004)**	0.820 (±0.003)	0.820 (±0.003)	0.821 (±0.003)	0.823 (±0.004)
30	0.823 (±0.005)	**0.827 (±0.003)**	0.823 (±0.003)	0.824 (±0.003)	0.824 (±0.003)	0.824 (±0.003)
50	**0.828 (±0.003)**	0.827 (±0.003)	0.825 (±0.003)	0.826 (±0.003)	0.826 (±0.003)	0.827 (±0.003)

TABLE 7.3: Mean and standard deviation of AUC for *heart* dataset ($p = 13$).

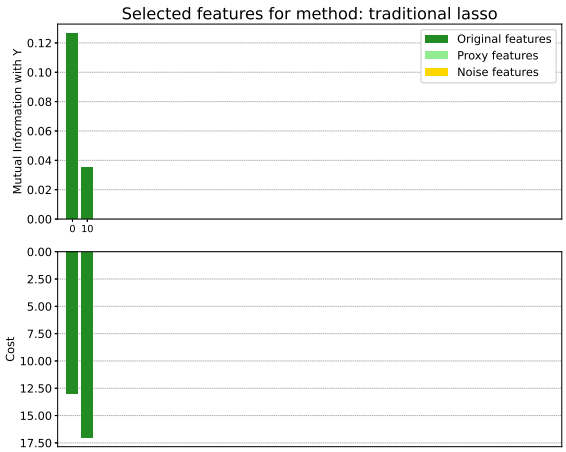
Budget [%]	traditional lasso	cheapknockoffs	cs lasso	cs MCP	cs adaptive lasso	cs adaptive MCP
5	0.769 (±0.123)	0.639 (±0.148)	0.827 (±0.047)	0.825 (±0.047)	**0.831 (±0.038)**	0.823 (±0.041)
10	0.755 (±0.045)	0.730 (±0.135)	**0.869 (±0.029)**	0.839 (±0.040)	0.866 (±0.031)	0.834 (±0.042)
15	0.755 (±0.045)	0.760 (±0.128)	**0.867 (±0.030)**	0.839 (±0.040)	0.865 (±0.032)	0.834 (±0.042)
20	0.829 (±0.037)	0.842 (±0.032)	**0.867 (±0.030)**	0.839 (±0.040)	0.865 (±0.031)	0.835 (±0.042)
30	0.856 (±0.028)	0.862 (±0.029)	**0.867 (±0.030)**	0.839 (±0.040)	0.865 (±0.031)	0.835 (±0.041)
50	**0.887 (±0.023)**	0.881 (±0.025)	0.867 (±0.030)	0.839 (±0.040)	0.865 (±0.031)	0.835 (±0.042)

TABLE 7.4: Mean and standard deviation of AUC for *thyroid* dataset ($p = 20$)

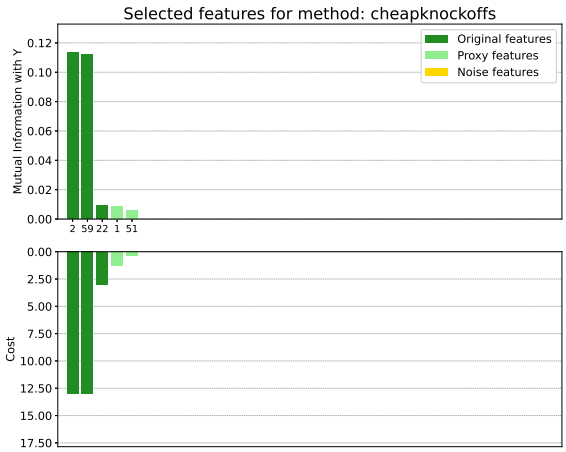
Budget [%]	traditional lasso	cheapknockoffs	cs lasso	cs MCP	cs adaptive lasso	cs adaptive MCP
5	0.702 (±0.016)	0.583 (±0.102)	**0.714 (±0.024)**	0.712 (±0.026)	0.713 (±0.023)	0.712 (±0.027)
10	0.703 (±0.017)	0.670 (±0.081)	0.724 (±0.021)	0.722 (±0.022)	**0.724 (±0.021)**	0.721 (±0.026)
15	0.721 (±0.019)	**0.729 (±0.019)**	0.722 (±0.022)	0.722 (±0.022)	0.722 (±0.022)	0.721 (±0.026)
20	0.734 (±0.017)	**0.734 (±0.018)**	0.722 (±0.022)	0.722 (±0.022)	0.722 (±0.022)	0.721 (±0.026)
30	**0.749 (±0.020)**	0.737 (±0.021)	0.722 (±0.022)	0.722 (±0.022)	0.722 (±0.022)	0.721 (±0.026)
50	0.766 (±0.017)	**0.784 (±0.019)**	0.724 (±0.021)	0.722 (±0.022)	0.725 (±0.020)	0.721 (±0.026)

TABLE 7.5: Mean and standard deviation of AUC for *alzheimer* dataset ($p = 9$)

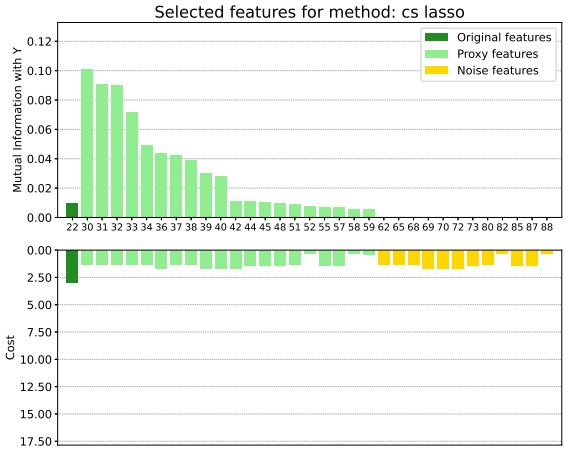
Budget [%]	traditional lasso	cheapknockoffs	cs lasso	cs MCP	cs adaptive lasso	cs adaptive MCP
5	0.500 (±0.000)	0.500 (±0.000)	0.810 (±0.051)	0.796 (±0.060)	**0.971 (±0.011)**	0.969 (±0.015)
10	0.500 (±0.000)	0.500 (±0.000)	0.967 (±0.018)	0.957 (±0.025)	**0.975 (±0.012)**	0.968 (±0.017)
15	0.500 (±0.000)	0.500 (±0.000)	0.966 (±0.016)	0.959 (±0.021)	**0.973 (±0.012)**	0.968 (±0.015)
20	0.500 (±0.000)	0.500 (±0.000)	0.966 (±0.016)	0.959 (±0.021)	**0.973 (±0.012)**	0.968 (±0.015)
30	0.969 (±0.006)	0.793 (±0.230)	0.966 (±0.016)	0.959 (±0.020)	**0.973 (±0.013)**	0.967 (±0.016)
50	**0.981 (±0.006)**	0.980 (±0.008)	0.971 (±0.012)	0.966 (±0.016)	0.980 (±0.008)	0.977 (±0.013)



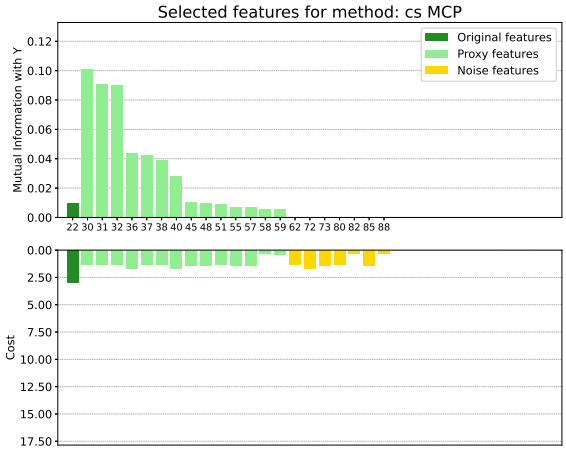
(A) Traditional lasso



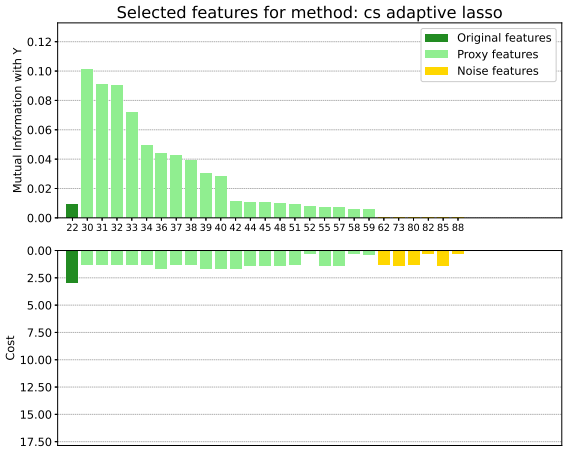
(B) Cheapknockoff



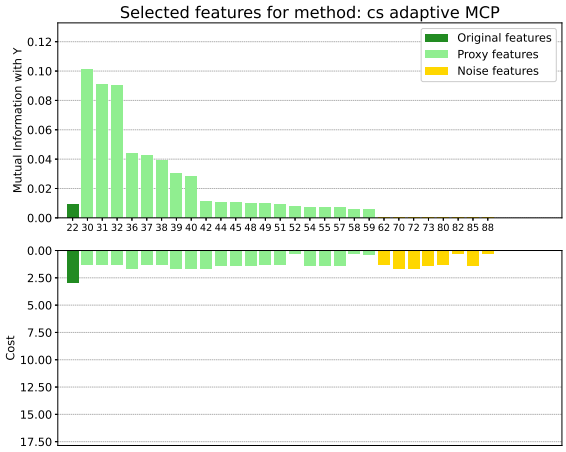
(C) Cost-sensitive lasso



(D) Cost-sensitive MCP



(E) Cost-sensitive adaptive lasso



(F) Cost-sensitive adaptive MCP

FIGURE 7.4: Cost and relevance of the selected features for MIMIC-II dataset, when the budget T is equal to 10% of the total cost.

Chapter 8

Conclusions

This dissertation investigated the problem of feature selection under explicit budget constraints. Traditional methods of feature selection aim to maximize predictive accuracy but ignore the costs associated with acquiring features or groups of features. Such costs, however, are critical in many real-world applications, particularly in healthcare, where diagnostic tests may vary considerably in price, invasiveness, and time requirements. To address this gap, the thesis developed and analyzed three complementary frameworks. The first focused on sequential information-theoretic selection with explicit cost penalties. The second introduced group cost-constrained selection that accounts for domain-driven structures. The third explored model-based penalized approaches that integrate costs directly into empirical risk minimization. Together, these approaches form a unified view of cost-aware feature selection, combining theoretical innovation with practical relevance.

8.1 Key Findings

First contribution of this dissertation was the development of cost-constrained sequential information-theoretic feature selection methods. These methods extend classical mutual-information based criteria by explicitly incorporating the costs of acquiring new features at each step of the selection process. The approach combines a measure of the informativeness of an added variable in the context of previously selected variables with a cost penalty that discourages the inclusion of expensive variables unless they provide substantial additional predictive value. A natural measure that quantifies this informativeness is the conditional mutual information, which is, however, difficult to estimate. A significant achievement, therefore, is the development of a method that allows replacing the difficult-to-estimate conditional mutual information with terms involving a smaller number of variables, which are much easier to estimate. The approach is based on the use of a lower bound on mutual information, leading to a very general informativeness measure. By appropriately selecting parameters, one can account for higher-order interactions involving both features and labels in multi-label classification. In cases with limited data, the parameters can be chosen such that the criterion excludes interaction terms. The experiments demonstrated that sequential

cost-aware methods consistently outperformed traditional information-theoretic algorithms when budgets were tight. In particular, the methods achieved higher predictive performance under low-budget regimes by prioritizing inexpensive but informative variables. By leveraging conditional dependencies and higher-order interactions, the sequential approach was able to identify compact subsets of features that balanced predictive strength with affordability. Another important finding was the effectiveness of the proposed procedure for optimizing the cost factor parameter. Rather than relying on manual tuning, the method employed a systematic search that adapted to the characteristics of the dataset and the available budget. This flexibility made the sequential algorithms more robust and easier to apply across diverse scenarios. Overall, the cost-constrained sequential information-theoretic framework illustrates how classical filter-style feature selection can be adapted to meet modern requirements of efficiency and feasibility. By combining information-theoretic measures with explicit cost modeling, the proposed methods offer a principled way to construct accurate models that operate within strict budget constraints.

Second contribution of this thesis was the introduction and analysis of group cost-constrained feature selection (GCC-FS), a framework in which costs are assigned to entire groups of features rather than to individual variables. This setting naturally reflects many practical scenarios, especially in medicine, where a single diagnostic test produces multiple related variables but incurs only one cost. For instance, a complete blood count provides a panel of correlated measurements at a fixed price, making group-based modeling more realistic than treating each measurement independently. The experimental results demonstrated the effectiveness of the proposed group selection algorithms in several respects. They consistently identified feature groups that provided strong predictive performance while respecting strict budget constraints. In both synthetic and real datasets, the methods achieved a superior balance between informativeness and cost compared with algorithms that operate only at the individual feature level. Moreover, the proposed two-step procedure based on shadow features was shown to be particularly effective at reducing redundancy. By introducing shadow features that competed with original variables, the method avoided the over-selection of highly correlated measurements within a group and promoted more diverse and informative subsets. This approach also eliminated the need to fine-tune a cost factor parameter, thereby simplifying application in practice. The advantages of GCC-FS were especially visible in the experiments conducted on the MIMIC-II dataset. In this case, traditional feature selection tended to focus on expensive tests, leading to high accuracy but at a prohibitive cost. Overall, the group selection methods developed in this thesis highlight the importance of moving beyond feature-level cost modeling. By treating related measurements as coherent units, GCC-FS reflects the realities of data acquisition processes, reduces redundancy, and improves the interpretability of the selected subsets, thereby enhancing both the effectiveness and the usability of machine learning models in cost-sensitive environments.

The third major contribution of this dissertation was the exploration of model-based penalized methods for cost-constrained feature selection. Unlike model-free approaches, which rely on filter-style information-theoretic criteria, these methods integrate cost-awareness directly into the learning objective through modified regularization terms. This strategy allows the model to simultaneously optimize predictive accuracy and acquisition cost, thereby providing a unified framework for cost-sensitive learning. The central idea was to adapt well-established penalization techniques such as lasso, adaptive lasso, and non-convex penalties including MCP and SCAD. By introducing cost-adjusted penalty factors, the algorithms discouraged the inclusion of expensive features unless they offered significant gains in predictive performance. The experimental results demonstrated that penalized cost-constrained models performed competitively or superiorly compared to their traditional counterparts. Across a variety of real-world datasets, the cost-aware versions achieved better classification metric values under budget restrictions while selecting substantially less costly feature subsets.

8.2 Practical Significance

The practical implications of this work are particularly relevant in domains where predictive performance must be balanced against resource expenditure. In healthcare, the proposed methods offer the potential to reduce reliance on expensive or invasive diagnostic tests while preserving diagnostic accuracy, thereby improving both patient outcomes and the efficiency of healthcare systems. More broadly, the principles developed here can be applied in finance, industry, and other domains where feature acquisition costs play a significant role in decision-making. The ability to explicitly manage the trade-off between informativeness and cost offers a way to design machine learning systems that are not only accurate but also economically feasible and socially responsible.

8.3 Limitations

Despite the advances presented in this thesis, certain limitations remain. A key challenge was the lack of publicly available datasets with reliable cost annotations, which necessitated the use of artificially generated proxy costs in many experiments. While this approach provided useful insights, validation on larger datasets with expert-assigned costs would further strengthen the conclusions.

Information-theoretic feature selection methods allow for the consideration of interactions between variables. At the theoretical level, the proposed methods support the use of interactions of arbitrarily high order. However, in practice, we focus on interactions involving at most three variables (feature-feature-label) or (feature-label-label). Accounting for higher-order interactions could lead to the selection of a variable set for which the model achieves even greater

predictive power. Unfortunately, incorporating such higher-order interactions requires the estimation of high-dimensional probability distributions, which becomes a very challenging task when the dataset is limited in size. This issue thus represents a limitation in applying the considered methods in situations where effective prediction relies on very high-order interactions.

8.4 Future Work

This work suggests several directions for future research.

First, in this thesis, we considered the costs assigned to individual variables or the costs assigned to entire groups of variables. In the future, it would be worth exploring scenarios involving more complex relationships between the costs of individual variables and the costs of groups. For example, one could consider a situation where the cost of one group depends on whether we have paid for variables from another group. It is also possible to explore a scenario where the cost of acquiring variables from several groups simultaneously is lower than the cost of acquiring each group separately.

Secondly, we focused on model-free section methods or simple linear models based on the penalized empirical risk minimization scheme. A promising extension is the exploration of cost-constrained deep learning architectures, allowing budget-aware modeling of high-dimensional and unstructured data.

Finally, validation on larger clinical datasets, in collaboration with practitioners, will be essential to demonstrate the real-world utility of cost-constrained feature selection.

8.5 Final Remarks

This dissertation has demonstrated that explicitly incorporating feature costs into feature selection leads to models that are both accurate and resource-efficient. By merging information-theoretic criteria, group cost modeling, and penalized optimization, the work enriches the theoretical foundations of cost-sensitive machine learning and provides practical tools for applications where resources are limited. The contributions of this thesis not only advance the methodological landscape of machine learning but also open pathways for impactful deployment in real-world domains where efficiency and interpretability are of paramount importance.

Appendix A

Additional figures and tables

TABLE A.1: Results of artificial dataset feature selection for the cost generation method C_1 and parameters $\pi = 0.5$ and $\alpha = 1$ and $\rho = 0.5$ and $a = 0.9$

(A) Additional parameters: $\Psi = 0.1$

Budget	**Metric**	Proposed	MIM	MRMR	JMI
1	Accuracy	**0.671 ± 0.037**	**0.671 ± 0.037**	**0.671 ± 0.037**	**0.671 ± 0.037**
2		0.739 ± 0.067	0.772 ± 0.042	0.692 ± 0.026	**0.776 ± 0.029**
5		0.811 ± 0.029	0.883 ± 0.027	0.811 ± 0.026	**0.902 ± 0.024**
1	F1 score	**0.683 ± 0.025**	**0.683 ± 0.025**	**0.683 ± 0.025**	**0.683 ± 0.025**
2		0.744 ± 0.071	0.774 ± 0.049	0.701 ± 0.039	**0.786 ± 0.033**
5		0.827 ± 0.030	0.892 ± 0.031	0.834 ± 0.020	**0.910 ± 0.025**
1	Precision	**0.675 ± 0.044**	**0.675 ± 0.044**	**0.675 ± 0.044**	**0.675 ± 0.044**
2		0.740 ± 0.074	**0.778 ± 0.063**	0.691 ± 0.054	0.764 ± 0.041
5		0.773 ± 0.052	0.840 ± 0.052	0.754 ± 0.039	**0.847 ± 0.045**
1	Recall	**0.696 ± 0.066**	**0.696 ± 0.066**	**0.696 ± 0.066**	**0.696 ± 0.066**
2		0.752 ± 0.085	0.772 ± 0.058	0.714 ± 0.046	**0.810 ± 0.034**
5		0.893 ± 0.037	0.953 ± 0.035	0.935 ± 0.013	**0.984 ± 0.014**

(B) Additional parameters: $\Psi = 0.5$

Budget	**Metric**	Proposed	MIM	MRMR	JMI
1	Accuracy	**0.671 ± 0.037**	**0.671 ± 0.037**	**0.671 ± 0.037**	**0.671 ± 0.037**
2		0.745 ± 0.064	0.772 ± 0.042	0.689 ± 0.022	**0.776 ± 0.029**
5		0.864 ± 0.037	0.883 ± 0.027	0.796 ± 0.030	**0.902 ± 0.024**
1	F1 score	**0.683 ± 0.025**	**0.683 ± 0.025**	**0.683 ± 0.025**	**0.683 ± 0.025**
2		0.748 ± 0.074	0.774 ± 0.049	0.699 ± 0.036	**0.786 ± 0.033**
5		0.874 ± 0.038	0.892 ± 0.031	0.812 ± 0.030	**0.910 ± 0.025**
1	Precision	**0.675 ± 0.044**	**0.675 ± 0.044**	**0.675 ± 0.044**	**0.675 ± 0.044**
2		0.746 ± 0.075	**0.778 ± 0.063**	0.689 ± 0.052	0.764 ± 0.041
5		0.823 ± 0.050	0.840 ± 0.052	0.765 ± 0.063	**0.847 ± 0.045**
1	Recall	**0.696 ± 0.066**	**0.696 ± 0.066**	**0.696 ± 0.066**	**0.696 ± 0.066**
2		0.753 ± 0.088	0.772 ± 0.058	0.711 ± 0.042	**0.810 ± 0.034**
5		0.933 ± 0.047	0.953 ± 0.035	0.869 ± 0.019	**0.984 ± 0.014**

TABLE A.2: Results of artificial dataset feature selection for parameters $\pi = 0.5$ and $\alpha = 1$ and $\rho = 0.1$ and $a = 0.1$

(A) Additional parameters: cost generation method C_2

Budget	**Metric**	Proposed	MIM	MRMR	JMI
1	Accuracy	**0.670 ± 0.028**	**0.670 ± 0.028**	**0.670 ± 0.028**	**0.670 ± 0.028**
2		0.750 ± 0.033	0.750 ± 0.033	**0.770 ± 0.022**	0.732 ± 0.046
5		**0.868 ± 0.015**	0.838 ± 0.018	0.857 ± 0.021	**0.868 ± 0.015**
1	F1 score	**0.681 ± 0.022**	**0.681 ± 0.022**	**0.681 ± 0.022**	**0.681 ± 0.022**
2		0.755 ± 0.039	0.755 ± 0.039	**0.776 ± 0.009**	0.735 ± 0.045
5		**0.872 ± 0.022**	0.842 ± 0.007	0.860 ± 0.028	**0.872 ± 0.022**
1	Precision	**0.671 ± 0.010**	**0.671 ± 0.010**	**0.671 ± 0.010**	**0.671 ± 0.010**
2		0.756 ± 0.082	0.756 ± 0.082	**0.774 ± 0.059**	0.744 ± 0.092
5		**0.862 ± 0.044**	0.842 ± 0.046	0.852 ± 0.046	**0.862 ± 0.044**
1	Recall	**0.694 ± 0.051**	**0.694 ± 0.051**	**0.694 ± 0.051**	**0.694 ± 0.051**
2		0.758 ± 0.024	0.758 ± 0.024	**0.783 ± 0.044**	0.730 ± 0.026
5		**0.883 ± 0.029**	0.845 ± 0.049	0.868 ± 0.020	**0.883 ± 0.029**

(B) Additional parameters: cost generation method C_3

Budget	**Metric**	Proposed	MIM	MRMR	JMI
1	Accuracy	0.653 ± 0.029	**0.694 ± 0.054**	0.666 ± 0.052	0.653 ± 0.029
2		0.807 ± 0.061	0.790 ± 0.032	0.808 ± 0.058	**0.810 ± 0.056**
5		**0.851 ± 0.025**	0.848 ± 0.022	0.842 ± 0.029	**0.851 ± 0.025**
1	F1 score	0.656 ± 0.025	**0.696 ± 0.071**	0.670 ± 0.054	0.656 ± 0.025
2		**0.813 ± 0.056**	0.794 ± 0.039	0.812 ± 0.057	0.812 ± 0.056
5		**0.851 ± 0.029**	0.848 ± 0.027	0.842 ± 0.034	**0.851 ± 0.029**
1	Precision	0.667 ± 0.066	**0.699 ± 0.075**	0.676 ± 0.078	0.667 ± 0.066
2		0.806 ± 0.077	0.791 ± 0.052	0.813 ± 0.070	**0.817 ± 0.067**
5		**0.862 ± 0.041**	0.859 ± 0.036	0.855 ± 0.050	**0.862 ± 0.041**
1	Recall	0.649 ± 0.023	**0.697 ± 0.094**	0.668 ± 0.053	0.649 ± 0.023
2		**0.822 ± 0.039**	0.799 ± 0.056	0.812 ± 0.050	0.809 ± 0.055
5		**0.841 ± 0.030**	0.837 ± 0.031	0.832 ± 0.038	**0.841 ± 0.030**

Bibliography

[A1] P. Teisseyre and T. Klonecki. "Controlling Costs in Feature Selection: Information Theoretic Approach". In: *Proceedings of the International Conference on Computational Science*. ICCS. 2021, pp. 483–496.

[A2] T. Klonecki, P. Teisseyre, and J. Lee. "Cost-constrained feature selection in multilabel classification using an information-theoretic approach". In: *Pattern Recognition* 141 (2023), pp. 1–18.

[A3] T. Klonecki, P. Teisseyre, and J. Lee. "Cost-constrained Group Feature Selection Using Information Theory". In: *Modeling Decisions for Artificial Intelligence*. MDAI. 2023, pp. 121–132.

[A4] T. Klonecki and P. Teisseyre. "Feature selection under budget constraint in medical applications: analysis of penalized empirical risk minimization methods". In: *Applied Intelligence* 53 (2023), pp. 29943–29973.

[A5] T. Klonecki, P. Teisseyre, and J. Lee. "Cost-constrained multi-label group feature selection using shadow features". In: *Proceedings of the 6-th Polish Conference on Artificial Intelligence Conference PP-RAI*. PP-RAI. 2025.

[A6] O. Kaminska, T. Klonecki, and K. Kaczmarek-Majer. "Feature Selection in Bipolar Disorder Episode Classification Using Cost-Constrained Methods". In: *Explainable Artificial Intelligence and Process Mining Applications for Healthcare*. 2024, pp. 36–40.

[1] H. Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*. 1998, pp. 199–213.

[2] N. Anantrasirichai and D. Bull. "Artificial intelligence in the creative industries: a review". In: *Artificial intelligence review* 55 (2022), pp. 589–656.

[3] A. Bahrammirzaee. "A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems". In: *Neural Computing and Applications* 19 (2010), pp. 1165–1195.

[4] R. Barber and E. Candès. "Controlling the false discovery rate via knockoffs". In: *Ann. Statist.* 43 (2015), pp. 2055–2085.

[5] G. Battineni et al. "Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis". In: *Journal of Personalized Medicine* 10 (2020), p. 21.

[6] R. Battiti. "Using Mutual Information for Selecting Features in Supervised Neural Net Learning". In: *Trans. Neur. Netw.* 5 (1994), pp. 537–550.

[7] S. Beiranvand, M. B. Dowlatshahi, and A. Hashemi. "A review on cost-based feature selection algorithms in the various applications of machine learning". In: *Journal of Mahani Mathematical Research* (2025).

[8] M. I. Belghazi et al. "MINE: Mutual Information Neural Estimation". 2018.

[9] M. Bennasar, Y. Hicks, and R. Setchi. "Feature selection using joint mutual information maximisation". In: *Expert systems with applications* 42 (2015), pp. 8520–8532.

[10] T. B. Berrett et al. "The Conditional Permutation Test for Independence While Controlling for Confounders". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82 (2019), pp. 175–197.

[11] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[12] J. Bogatinovski et al. "Comprehensive comparative study of multi-label classification methods". In: *Expert Systems with Applications* 203 (2022), pp. 1–18.

[13] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos. "Feature selection for high-dimensional data". In: *Progress in Artificial Intelligence* 4 (2015), pp. 65–75.

[14] V. Bolón-Canedo et al. "A framework for cost-based feature selection". In: *Pattern Recognition* 47 (2014), pp. 2481–2489.

[15] A-L. Boulesteix et al. "IPF-LASSO: Integrative ℓ_1-Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data". In: *Comput. Math. Methods Medicine* 2017 (2017), pp. 1–14.

[16] L. Breiman. "Random Forests". In: *Machine Learning* 45 (2001), pp. 5–32.

[17] G. Brown et al. "Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection". In: *Journal of Machine Learning Research* 13 (2012), pp. 27–66.

[18] U. Butt et al. "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications". In: *Journal of Healthcare Engineering* 2021 (2021), pp. 1–17.

[19] J. Cai et al. "Feature selection in machine learning: A new perspective". In: *Neurocomputing* 300 (2018), pp. 70–79.

[20] E. Candès et al. "Panning for Gold: Model-free Knockoffs for High-dimensional Controlled Variable Selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80 (2016).

[21] G. Chandrashekar and F. Sahin. "A survey on feature selection methods". In: *Computers and Electrical Engineering* 40 (2014), pp. 16–28.

[22] J. Chen et al. "Kernel feature selection via conditional covariance minimization". In: *Advances in Neural Information Processing Systems*. 2017.

[23] T. Chen and C. Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. 2016, pp. 785–794.

[24] W-J. Chen et al. "MLTSVM: A novel twin support vector machine to multi-label learning". In: *Pattern Recognition* 52 (2016), pp. 61–74.

[25] Y. Chen et al. "The utility of LASSO-based models for real time forecasts of endemic infectious diseases: A cross country comparison". In: *Journal of Biomedical Informatics* 81 (2018), pp. 16–30.

[26] H. Climente-González and J. C. Valderrama-Zurián. "Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data". In: *Bioinformatics* 35 (2019), pp. 427–435.

[27] H. J. Cordell. "Detecting gene-gene interactions that underlie human diseases". In: *Nature Review Genetics* 10 (2009), pp. 392–404.

[28] H. J. Cordell. "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans". In: *Human Molecular Genetics* 11 (2002), pp. 2463–2468.

[29] T. M. Cover and J. A. Thomas. *Elements of Information Theory* (*Wiley Series in Telecommunications and Signal Processing*). Wiley-Interscience, 2006.

[30] M. Dash and H. Liu. "Feature selection for classification". In: *Intelligent Data Analysis* 1 (1997), pp. 131–156.

[31] J. V. Davis et al. "Cost-sensitive decision tree learning for forensic classification". In: *European Conference on Machine Learning*. 2006, pp. 622–629.

[32] K. Dembczyński et al. "On label dependence and loss minimization in multi-label classification". In: *Machine Learning* 88 (2012), pp. 5–45.

[33] P. Dhal and C. Azad. "A comprehensive survey on feature selection in the various fields of machine learning". In: *Applied Intelligence* 52 (2022), pp. 11701–11726.

[34] T. G. Dietterich. "Overfitting and Undercomputing in Machine Learning". In: *ACM Computing Surveys* (*CSUR*) 27 (1995), pp. 326–327.

[35] M. D. Donsker and S. R. S. Varadhan. "Asymptotic evaluation of certain markov process expectations for large time, I". In: *Communications on Pure and Applied Mathematics* 28 (1975), pp. 1–47.

[36] M. Dramiński and J. Koronacki. "rmcfs: An R Package for Monte Carlo Feature Selection and Interdependency Discovery". In: *Journal of Statistical Software* 85 (2018), pp. 1–28.

[37] M. Dramiński et al. "Monte Carlo feature selection for supervised classification". In: *Bioinformatics* 24 (2008), pp. 110–117.

[38] D. Dua and C. Graff. *Heart Disease*. UCI Machine Learning Repository. 2017.

[39] C. Elkan. "The foundations of cost-sensitive learning". In: *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'01. 2001, pp. 973–978.

[40] J. Fan and R. Li. "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties". In: *Journal of the American Statistical Association* 96 (2001), pp. 1348–1360.

[41] R. Fan et al. "Entropy-based information gain approaches to detect and to characterize gene-gene and gene-environment interactions/correlations of complex diseases". In: *Genetic Epidemiology* 35 (2011), pp. 706–721.

[42] V. Fonti and E. Belitser. "Feature selection using lasso". In: *VU Amsterdam research paper in business analytics* 30 (2017), pp. 1–25.

[43] M. Fortin et al. "Multimorbidity and quality of life in primary care: a systematic review". In: *Health and Quality of Life Outcomes* 2 (2004), pp. 1–12.

[44] X. Fu, D. Li, and Y. Zhai. "Multi-label learning with kernel local label information". In: *Expert Systems with Applications* 207 (2022), p. 118027.

[45] Y. Fujino et al. "Applying "Lasso" Regression to Predict Future Visual Field Progression in Glaucoma Patients". In: *Investigative Ophthalmology & Visual Science* 56 (2015), pp. 2334–2339.

[46] W. Gao et al. "A unified low-order information-theoretic feature selection framework for multi-label learning". In: *Pattern Recognition* 134 (2023), p. 109111.

[47] E. Gibaja and S. Ventura. "A Tutorial on Multilabel Learning". In: *ACM Computing Surveys* 47 (2015), pp. 1–38.

[48] R. Gijsen et al. "Causes and consequences of comorbidity: A review". In: *Journal of Clinical Epidemiology* 54 (2001), pp. 661–674.

[49] R. Goetschalckx, K. Driessens, and S. Sanner. "Cost-sensitive parsimonious linear regression". In: *2008 Eighth IEEE International Conference on Data Mining*. 2008, pp. 809–814.

[50] I. Guyon and A. Elisseeff. "An introduction to variable and feature selection". In: *Journal of Machine Learning Research* 3 (2003), pp. 1157–1182.

[51] E. J. Hall and D. J. Brenner. "Cancer risks from diagnostic radiology". In: *The British Journal of Radiology* 81 (2008), pp. 362–378.

[52] T. S. Han. "Multiple mutual informations and multiple interactions in frequency data". In: *Information and Control* 46 (1980), pp. 26–45.

[53] T. Hastie. *Statistical learning with sparsity : the lasso and generalizations*. CRC Press, 2015.

[54] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2009.

[55] D. M. Hawkins. "The problem of overfitting". In: *Journal of Chemical Information and Computer Sciences* 44 (2004), pp. 1–12.

[56] H. Ishwaran. "The effect of splitting on random forests". In: *Machine Learning* 99 (2015), pp. 75–118.

[57] R. Jagdhuber et al. "Cost-constrained feature selection in binary classification: adaptations for greedy forward selection and genetic algorithms". In: *BMC bioinformatics* 21 (2020), p. 26.

[58] A. Jakulin and I. Bratko. "Quantifying and Visualizing Attribute Interactions: An Approach Based on Entropy". 2004.

[59] S. Ji and L. Carin. "Cost-sensitive feature acquisition and classification". In: *Pattern Recognition* 40 (2007), pp. 1474–1485.

[60] Y. Jing-Rung, L. Chun-Yu, and L. Donald. "Cost-effective and accuracy-oriented l1-norm support vector machine for enhanced feature selection". In: *Measurement* 253 (2025), p. 117506.

[61] A. E. W. Johnson et al. "MIMIC-III, a freely accessible critical care database". In: *Scientific Data* 3 (2016), pp. 1–9.

[62] A. Jović, K. Brkić, and N. Bogunović. "A review of feature selection methods with applications". In: *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 2015, pp. 1203–1208.

[63] D. Justus et al. "Predicting the computational cost of deep learning models". In: *2018 IEEE International Conference on Big Data (Big Data)*. 2018, pp. 2061–2068.

[64] S. Kashef, H. Nezamabadi-Pour, and B. Nikpour. "Multilabel feature selection: A comprehensive review and guiding experiments". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8 (2018), pp. 1–29.

[65] J. Kazemitabar et al. "Variable importance using decision trees". In: *Advances in neural information processing systems* (2017), p. 30.

[66] U. M. Khaire and R. Dhanalakshmi. "Stability of feature selection algorithm: A review". In: *Journal of King Saud University-Computer and Information Sciences* 34 (2022), pp. 1060–1073.

[67] S. Khalid, T. Khalil, and S. Nasreen. "A survey of feature selection and feature extraction techniques in machine learning". In: *2014 Science and Information Conference*. 2014, pp. 372–378.

[68] C. Khanji et al. "Lasso Regression for the Prediction of Intermediate Outcomes Related to Cardiovascular Disease Prevention Using the TRANSIT Quality Indicators". In: *Medical Care* 57 (2018), p. 1.

[69] S.M. Kim et al. "Logistic LASSO regression for the diagnosis of breast cancer using clinical demographic data and the BI-RADS lexicon for ultrasonography". In: *Ultrasonography* 37 (2018), pp. 36–42.

[70] R. Kohavi and G. H. John. "Wrappers for feature subset selection". In: *Artificial Intell.* 97 (1997), pp. 273–324.

[71] F. Königstorfer and S. Thalmann. "Applications of Artificial Intelligence in commercial banks–A research agenda for behavioral finance". In: *Journal of behavioral and experimental finance* 27 (2020), p. 100352.

[72] S. Kouchaki et al. "Multi-Label Random Forest Model for Tuberculosis Drug Resistance Classification and Mutation Ranking". In: *Frontiers in Microbiology* 11 (2020).

[73] M. Kubkowski, J. Mielniczuk, and P. Teisseyre. "How to Gain on Power: Novel Conditional Independence Tests Based on Short Expansion of Conditional Mutual Information". In: *Journal of Machine Learning Research* 22 (2021), pp. 1–57.

[74] M. Kuhn and K. Johnson. *Applied Predictive Modeling*. Springer, 2013.

[75] M. B. Kursa and W. R. Rudnicki. "Feature Selection with the Boruta Package". In: *Journal of Statistical Software* 36 (2010), pp. 1–13.

[76] C. Lazar et al. "A survey on filter techniques for feature selection in gene expression microarray analysis". In: *IEEE/ACM transactions on computational biology and bioinformatics* 9 (2012), pp. 1106–1119.

[77] J. Lee and D-W. Kim. "Approximating mutual information for multi-label feature selection". In: *Electronics Letters* 48 (2012), pp. 929–930.

[78] J. Lee and D-W. Kim. "Feature selection for multi-label classification using multivariate mutual information". In: *Pattern Recognition Letters* 34 (2013), pp. 349–357.

[79] J. Lee and D.-W. Kim. "SCLS: Multi-label feature selection based on scalable criterion for large label set". In: *Pattern Recognition* 66 (2017), pp. 342–352.

[80] D. D. Lewis. "Feature Selection and Feature Extraction for Text Categorization". In: *Proceedings of the Workshop on Speech and Natural Language*. HLT '91. 1992, pp. 212–217.

[81] C-L. Li and H-T. Lin. "Condensed filter tree for cost-sensitive multi-label classification". In: *International conference on machine learning*. 2014, pp. 423–431.

[82] H. Li et al. "Group feature selection with streaming features". In: *2013 IEEE 13th International Conference on Data Mining*. 2013, pp. 1109–1114.

[83] J. Li et al. "Feature selection: A data perspective". In: *ACM Computing Surveys (CSUR)* 50 (2017), p. 94.

[84] D. Lin and X. Tang. "Conditional Infomax Learning: An Integrated Framework for Feature Extraction and Fusion". In: *Computer Vision – ECCV 2006*. 2006, pp. 68–82.

[85] Y. Lin et al. "Multi-label feature selection based on max-dependency and min-redundancy". In: *Neurocomputing* 168 (2015), pp. 92–103.

[86] H. Liu and H. Motoda. *Computational Methods of Feature Selection*. CRC Press, 2007.

[87] H. Liu and L. Yu. "Toward integrating feature selection algorithms for classification and clustering". In: *IEEE Transactions on Knowledge and Data Engineering* 17 (2005), pp. 491–502.

[88] W. Liu and I. W. Tsang. "On the Optimality of Classifier Chain for Multi-label Classification". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems*. NIPS'15. 2015, pp. 712–720.

[89] V. Lohweg. *Banknote Authentication*. UCI Machine Learning Repository. 2012.

[90] X. Long et al. "Cost-sensitive feature selection on multi-label data via neighborhood granularity and label enhancement". In: *Applied Intelligence* 51 (2021), pp. 2210–2232.

[91] G. Madjarov et al. "An extensive experimental comparison of methods for multi-label learning". In: *Pattern Recognition* 45 (2012), pp. 3084–3104.

[92] D. Marcus et al. "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults". In: *Journal of cognitive neuroscience* 19 (2007), pp. 1498–507.

[93] W. J. McGill. "Multivariate information transmission". In: *Psychometrika* 19 (1954), pp. 97–116.

[94] Z. Meng et al. "Development and Validation of a LASSO Prediction Model for Better Identification of Ischemic Stroke: A Case-Control Study in China". In: *Frontiers in Aging Neuroscience* 13 (2021).

[95] P. E. Meyer and G. Bontempi. "On the Use of Variable Complementarity for Feature Selection in Cancer Classification". In: *Proceedings of the 2006 International Conference on Applications of Evolutionary Computing*. EuroGP'06. 2006, pp. 91–102.

[96] J. Mielniczuk and P. Teisseyre. "Stopping rules for mutual information-based feature selection". In: *Neurocomputing* 358 (2019), pp. 255–274.

[97] F. Min, Q. Hu, and W. Zhu. "Feature selection with test cost constraint". In: *International Journal of Approximate Reasoning* 55 (2014), pp. 167–179.

[98] F. Min et al. "Test-cost-sensitive attribute reduction". In: *Information Sciences* 181 (2011), pp. 4719–4732.

[99] K. Mnich and W. R. Rudnicki. "All-relevant feature selection using multidimensional filters with exhaustive search". In: *Information Sciences* 524 (2020), pp. 277–297.

[100] S.S. Mohanrasu et al. "Cost-sensitive feature selection for multi-label classification: multi-criteria decision-making approach". In: *Applied Computing and Informatics* (2025).

[101] L.C. Molina, L. Belanche, and A. Nebot. "Feature selection algorithms: a survey and experimental evaluation". In: *2002 IEEE International Conference on Data Mining, 2002. Proceedings.* 2002, pp. 306–313.

[102] E. Montañes et al. "Dependent binary relevance models for multi-label classification". In: *Pattern Recognition* 47.3 (2014), pp. 1494–1508.

[103] S. Mullainathan and J. Spiess. "Machine learning: an applied econometric approach". In: *Journal of Economic Perspectives* 31 (2017), pp. 87–106.

[104] M. R. Nelson et al. "A Combinatorial Partitioning Method to Identify Multilocus Genotypic Partitions That Predict Quantitative Trait Variation". In: *Genome Research* 11 (2001), pp. 458–470.

[105] R. Nilsson et al. "Consistent Feature Selection for Pattern Recognition in Polynomial Time". In: *Journal of Machine Learning Research* 8 (2007), pp. 589–612.

[106] P. Paclík et al. "On Feature Selection with Measurement Cost and Grouped Features". In: *Proceedings of Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition*. 2002, pp. 461–469.

[107] L. Paninski. "Estimation of Entropy and Mutual Information". In: *Neural Computation* 15 (2003), pp. 1191–1253.

[108] D. J. Park et al. "Development of machine learning model for diagnostic disease prediction based on laboratory tests". In: *Scientific reports* 11 (2021), p. 7567.

[109] M. Pavlou et al. "Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events". In: *Statistics in medicine* 35 (2015).

[110] H. Peng, F. Long, and C. Ding. "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (2005), pp. 1226–1238.

[111] K. A. Phillips et al. "The economic value of personalized medicine tests: what we know and what we need to know". In: *Genetics in Medicine* 16 (2014), pp. 251–257.

[112] P. D Polessa et al. "Comparing machine learning algorithms for multimorbidity prediction: An example from the Elsa-Brasil study". In: *PLoS One* 17 (2022), pp. 1–14.

[113] C. Porzelius, M. Schumacher, and H. Binder. "Sparse Regression Techniques in Low-dimensional Survival Data Settings". In: *Statistics and Computing* 20 (2010), pp. 151–163.

[114] S. Puthiya Parambath, N. Usunier, and Y. Grandvalet. "Optimizing F-measures by cost-sensitive classification". In: *Advances in neural information processing systems*. 2014.

[115] R. Quinlan. *Thyroid Disease*. UCI Machine Learning Repository. 1987.

[116] S. Rawat et al. "Application of machine learning and data visualization techniques for decision support in the insurance sector". In: *International Journal of Information Management Data Insights* 1 (2021), p. 100012.

[117] J. Read and F. Perez-Cruz. "Deep Learning for Multi-label Classification". 2014.

[118] J. Read et al. "Classifier chains for multi-label classification". In: *ECML/PKDD*. 2009, pp. 254–269.

[119] J. Read et al. "Classifier chains for multi-label classification". In: *Machine Learning* 85 (2011), pp. 333–359.

[120] J. Read et al. "Classifier Chains: A Review and Perspectives". In: *J. Artif. Int. Res.* 70 (2021), pp. 683–718.

[121] M. D. Ritchie et al. "Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer". In: *The American Journal of Human Genetics* 69 (2001), pp. 138–147.

[122] V. Roth and B. Fischer. "The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms". In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 848–855.

[123] M. Saeed et al. "Multiparameter intelligent monitoring in intensive care II: A public-access intensive care unit database". In: *Critical Care Medicine* 39 (2011), pp. 952–960.

[124] Y. Saeys, I. Inza, and P. Larrañaga. "A review of feature selection techniques in bioinformatics". In: *Bioinformatics* 23 (2007), pp. 2507–2517.

[125] I. Sarker. "Machine Learning: Algorithms, Real-World Applications and Research Directions". In: *SN Computer Science* 2 (2021).

[126] S. J. Schrodi et al. "Genetic-based prediction of disease traits: prediction is very difficult, especially about the future". In: *Frontiers in Genetics* 5 (2014), p. 162.

[127] K. Sechidis, N. Nikolaou, and G. Brown. "Information Theoretic Feature Selection in Multi-label Data through Composite Likelihood". In: *Structural, Syntactic, and Statistical Pattern Recognition*. 2014, pp. 143–152.

[128] C. E. Shannon. "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27 (1948), pp. 379–423.

[129] G. Shmueli. "To Explain or To Predict?" In: *Statistical Science* 25 (2010), pp. 289–310.

[130] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad. "A review of unsupervised feature selection methods". In: *Artificial Intelligence Review* 53 (2020), pp. 907–948.

[131] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. 2nd. MIT press, 2000.

[132] N. Spolaôr et al. "A Comparison of Multi-label Feature Selection Methods using the Problem Transformation Approach". In: *Electronic Notes in Theoretical Computer Science* 292 (2013), pp. 135–151.

[133] M. Tan. "Cost-sensitive learning of classification knowledge and its applications in robotics". In: *Machine Learning* 13 (1993), pp. 7–33.

[134] J. Tang, S. Alelyani, and H. Liu. *Feature selection for classification: A review*. CRC Press, 2014, pp. 37–64.

[135] P. Teisseyre. "CCnet: Joint multi-label classification and feature selection using classifier chains and elastic net regularization". In: *Neurocomputing* 235 (2017), pp. 98–111.

[136] P. Teisseyre. "Classifier chains for positive unlabelled multi-label learning". In: *Knowledge-Based Systems* 213 (2021), pp. 1–16.

[137] P. Teisseyre. "Learning classifier chains using matrix regularization: application to multimorbidity prediction". In: *Proceedings of the European Conference on Artificial Intelligence*. 2020.

[138] P. Teisseyre and J. Lee. "Multilabel all-relevant feature selection using lower bounds of conditional mutual information". In: *Expert Systems with Applications* 216 (2023), p. 119436.

[139] P. Teisseyre, D. Zufferey, and M. Słomka. "Cost-sensitive classifier chains: Selecting low-cost features in multi-label classification". In: *Pattern Recognition* 86 (2019), pp. 290–319.

[140] R. Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society (Series B)* 58 (1996), pp. 267–288.

[141] G. Tsoumakas and I. Katakis. "Multi-label classification: An overview". In: *International Journal of Data Warehousing and Mining* (2007), pp. 1–13.

[142] G. Tsoumakas et al. "MULAN: A Java Library for Multi-Label Learning". In: *Journal of Machine Learning Research* 12 (2011), pp. 2411–2414.

[143] P. D. Turney. "Types of Cost in Inductive Concept Learning". In: *Proceedings of the 17th International Conference on Machine Learning*. ICML'02. 2002, pp. 1–7.

[144] E. Uffelmann et al. "Genome-wide association studies". In: *Nature Reviews Methods Primers* 1 (2021), p. 59.

[145] R. J. Urbanowicz et al. "Relief-based feature selection: Introduction and review". In: *Journal of Biomedical Informatics* 85 (2018), pp. 189–203.

[146] A. Van den Bruel et al. "The evaluation of diagnostic tests: evidence on technical and diagnostic accuracy, impact on patient outcome and cost-effectiveness is needed". In: *Journal of Clinical Epidemiology* 60 (2007), pp. 1111–1116.

[147] J. R. Vergara and P. A. Estévez. "A review of feature selection methods based on mutual information". In: *Neural Computing and Applications* 24 (2014), pp. 175–186.

[148] W. Wolberg W. Street and O. Mangasarian. *Breast Cancer Wisconsin (Diagnostic)*. UCI Machine Learning Repository. 1993.

[149] S. Wangduk, D-W. Kim, and J. Lee. "Generalized Information-Theoretic Criterion for Multi-Label Feature Selection". In: *IEEE Access* 7 (2019), pp. 122854–122863.

[150] I. H. Witten et al. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2016.

[151] G. Wu et al. "A prediction model of outcome of SARS-CoV-2 pneumonia based on laboratory findings". In: *Scientific Reports* 10 (2020), pp. 1–9.

[152] W. Xin et al. "A weighted ML-KNN based on discernibility of attributes to heterogeneous sample pairs". In: *Information Processing & Management* 59 (2022), p. 103053.

[153] M. Yamada et al. "High-dimensional feature selection by feature-wise kernelized lasso". In: *Neural Computation* 26 (2014), pp. 185–207.

[154] H. H. Yang and J. Moody. "Data Visualization and Feature Selection: New Algorithms for Nongaussian Data". In: *Proceedings of the 12th International Conference on Neural Information Processing Systems*. NIPS'99. 1999, pp. 687–693.

[155] X. Ying. "An overview of overfitting and its solutions". In: *Journal of Physics: Conference Series*. 2019, p. 022022.

[156] G. Yu, D. Witten, and J. Bien. "Controlling Costs: Feature Selection on a Budget". In: *Stat* 11 (2022), pp. 1–22.

[157] L. Yu and H. Liu. "Efficient Feature Selection via Analysis of Relevance and Redundancy". In: *J. Mach. Learn. Res.* 5 (2004), pp. 1205–1224.

[158] M. Yuan and Y. Lin. "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68 (2006), pp. 49–67.

[159] C. Zhang. "Nearly unbiased variable selection under minimax concave penalty". In: *The Annals of Statistics* 38 (2010), pp. 894–942.

[160] H. Zhang et al. "Feature selection for neural networks using group lasso regularization". In: *IEEE Transactions on Neural Networks and Learning Systems* 30 (2019), pp. 123–134.

[161] M-L. Zhang and Z-H. Zhou. "ML-KNN: A lazy learning approach to multi-label learning". In: *Pattern Recognition* 40 (2007), pp. 2038–2048.

[162] M. Zhang and Z. Zhou. "A Review on Multi-Label Learning Algorithms". In: *IEEE Transactions on Knowledge and Data Engineering* 26 (2013), pp. 1819–1837.

[163] M.-L. Zhang, J.-M. Peña, and V. Robles. "Feature selection for multi-label naive Bayes classification". In: *Information Sciences* 179 (2009), pp. 3218–3229.

[164] P. Zhang and W. Gao. "Feature Relevance Term Variation for Multi-Label Feature Selection". In: *Applied Intelligence* 51 (2021), pp. 5095–5110.

[165] Q. Zhou, H. Zhou, and T. Li. "Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features". In: *Knowledge-Based Systems* 95 (2016), pp. 1–11.

[166] H. Zou. "The adaptive lasso and its oracle properties". In: *Journal of the American Statistical Association* 101 (2006), pp. 1418–1429.

[167] G. S. Zubenko, H. B. Hughes, and J. S. Stiffler. "D10S1423 identifies a susceptibility locus for Alzheimer's disease in a prospective, longitudinal, double-blind study of asymptomatic individuals". In: *Molecular Psychiatry* 6 (2001), pp. 413–419.

[168] D. Zufferey et al. "Performance comparison of multi-label learning algorithms on clinical data for chronic diseases". In: *Computers in Biology and Medicine* 65 (2015), pp. 34–43.

[169] Y. Zuo et al. "Performance and Cost Assessment of Machine Learning Interatomic Potentials". In: *The Journal of Physical Chemistry A* 124 (2020), pp. 731–745.