Bilevel optimization for feature selection in binary classification with support vector machines

Adejuyigbe O. Fajemisin, David Rey

SKEMA Business School, Université Côte d'Azur, Sophia Antipolis, France

{ade.fajemisin,david.rey}@skema.edu

Mots-clés : *bilevel optimization, machine learning, feature selection, branch-and-bound.*

1 Introduction

Feature selection is an important task undertaken during the development of predictive models. Excluding irrelevant features can have positive effects on prediction accuracy and model interpretability, especially when there is a large number of candidate features [2]. Numerous approaches exist for feature selection in the literature – shrinkage, subset selection and criterion based methods are examples of techniques for feature selection from a statistical point of view. Recently, optimization approaches for feature selection have received an increasing attention due to their potential improve subset selection approaches in the context of specific learning models. Some of the related optimization literature addresses feature selection by focusing on the training problem. In [4], a Mixed Integer Linear Programming (MILP) formulation is developed for an embedded feature selection problem in Support Vector Machines (SVM). A min-max approach to select features for nonlinear SVM classification has also been proposed in [3]. Other approaches have proposed to address feature selection at the validation level : in this case, the problem is to determine the optimal set of features based on hold-out data validation. In [1], the authors frame feature selection for SVMs as a bilevel optimization problem where the leader represents the validation step and the follower represents the training problem. They develop a genetic algorithm to solve this bilevel optimization problem. We build on this research stream and develop exact, bilevel optimization approaches for feature selection.

2 Problem formulation

We propose a bilevel optimization formulation for feature selection in binary classification tasks, where the classifier is an SVM with a linear kernel. Our formulation allows us to integrate both the training and testing phases of classification. In our bilevel model, the leader decides the features to be selected in order to maximize the classification accuracy, and the follower (i.e. the SVM) determines the optimal parameters of the separating hyperplane given the features selected by the leader.

Let Ω be the data consisting of $|\Omega| = m$ data points, where each data point consists of n features. Let $\Omega_T \subset \Omega$ be the training data and let $\Omega_V = \Omega \setminus \Omega_T$ be the validation data. For each data point $i \in \Omega$, we denote $\boldsymbol{x}_i \in \mathbb{R}^n$ its feature vector and $y_i \in \{-1, 1\}$ its class. Let $u_j \in \{0, 1\}$ be the decision variable representing whether feature $j = 1, \dots, n$ is selected (1) or not (0). Given the training set Ω_T, and the vector $\boldsymbol{u}$ determined by the leader, the parameterized follower problem $\mathsf{SVM}(\boldsymbol{u})$ is to determine the optimal parameters $(\boldsymbol{w}, b) \in \mathbb{R}^n \times \mathbb{R}$ of the separating hyperplane $\boldsymbol{w}^T \cdot \boldsymbol{x} + b$. Let ξ_i be the error associated with misclassifying data point $i \in \Omega_T$, and C is a regularization parameter used to penalize misclassified points.

Mathematically :

$$\mathsf{SVM}(\boldsymbol{u}) = \arg\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\sum_{j=1}^{n} w_j^2 + C\sum_{i\in\Omega_T}\xi_i, \tag{1a}$$

$$\text{s.t.} \quad y_i\left(\sum_{j=1}^{n} w_j x_{ij} + b\right) \geq 1-\xi_i, \qquad \forall i \in \Omega_T, \tag{1b}$$

$$\xi_i \geq 0, \qquad \forall i \in \Omega_T, \tag{1c}$$

$$w_j \in \mathbb{R}, \qquad \forall j = 1,\ldots,n, \tag{1d}$$

$$b \in \mathbb{R}, \tag{1e}$$

$$(1-u_j)\, w_j = 0, \qquad \forall j = 1,\ldots,n, \tag{1f}$$

Let $\boldsymbol{z} \in \{0,1\}^{m_V}$ be a decision variable such that $z_i = 1$ if validation data point $i \in \Omega_V$ is correctly classified and $z_i = 0$ otherwise. As the leader aims to maximize the accuracy, i.e. the number of correctly classified data points, the feature selection problem can thus be formulated as the bilevel optimization problem below :

$$\max_{\boldsymbol{u},\boldsymbol{z}} \quad \sum_{i\in\Omega_V} z_i, \tag{2a}$$

$$\text{s.t.} \quad z_i\left(y_i\left(\sum_{j=1}^{n} w_j^\star x_{ij} + b^\star\right) - 1\right) \geq 0, \qquad \forall i \in \Omega_V, \tag{2b}$$

$$z_i \in \{0,1\}, \qquad \forall i \in \Omega_V, \tag{2c}$$

$$u_j \in \{0,1\}, \qquad \forall j = 1,\ldots,n, \tag{2d}$$

$$(\boldsymbol{w}^\star, b^\star, \xi^\star) \in \mathsf{SVM}(\boldsymbol{u}). \tag{2e}$$

3 Solution approach and evaluation

Since the parameterized follower problem $\mathsf{SVM}(\boldsymbol{u})$ is convex, it can be reformulated using its first order optimality conditions. While this results in a single-level problem, the presence of Karush–Kuhn–Tucker (KKT) complementarity conditions results in a problem that is both non-convex and nonlinear. We develop customized branch-and-bound (B&B) algorithms to solve the resulting single-level, mathematical programming with complementarity constraints (MPCC) formulation. Our approach is validated using a variety of benchmark data sets and evaluation metrics, and is also compared to some classification methods.

Références

[1] Joseph Agor and Osman Y Özaltın. Feature selection for classification models via bilevel optimization. *Computers & Operations Research*, 106 :156–168, 2019.

[2] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[3] Asunción Jiménez-Cordero, Juan Miguel Morales, and Salvador Pineda. A novel embedded min-max approach for feature selection in nonlinear support vector machine classification. *European Journal of Operational Research*, 293(1) :24–35, 2021.

[4] Martine Labbé, Luisa I Martínez-Merino, and Antonio M Rodríguez-Chía. Mixed integer linear programming for feature selection in support vector machine. *Discrete Applied Mathematics*, 261 :276–304, 2019.