

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.