Abstract

We address the challenging high-dimensional regression setting with correlated predictors, where the regression coefficients in a generalized linear model vary from sparse to dense. While some existing methods at hand can solve a sparse or a dense problem, there is a need for a method that can perform well without any knowledge about the true sparsity in feasible computing time since, in real applications, the true degree of sparsity is often unknown or unclear. With the help of stochastic dimension reduction tools – variable screening and random projection – we build an ensemble method to close this gap. We introduce a novel data-driven random projection method that leverages the underlying correlations among predictors, enhancing prediction performance over conventional random projections. This random projection is particularly useful in scenarios where the number of predictors far exceeds the number of observations. Central to our method is a ridge-type estimator for variable screening, allowing for effective dimensionality reduction while accounting for the response-predictor relationship. Our method builds an ensemble by repeatedly applying a probabilistic variable screening step followed by our proposed data-driven random projection and using the reduced set of predictors in a regular (generalized) linear model with a dimension smaller than the number of observations. Extensive simulations demonstrate that our method outperforms traditional random projection tools and classical sparse and dense methods across varying sparsity and covariance settings in variable ranking and prediction at a low computational cost. We validate the performance of the proposed framework through real data applications involving continuous, count, and binary responses, showing its advantages in both interpretability and prediction accuracy. An implementation of this approach is available in the **spar** package in `R`, which enables users to build predictive generalized linear models with high-dimensional predictors seamlessly.

Kurzfassung

Wir befassen uns mit der anspruchsvollen Problemstellung der hochdimensionalen Regression mit korrelierten Prädiktoren, bei der die Regressionskoeffizienten in einem verallgemeinerten linearen Modell von spärlich bis dicht variieren. Während einige der vorhandenen Methoden ein spärliches oder dichtes Problem lösen können, besteht ein Bedarf an einer Methode, die ohne Kenntnis der wahren Spärlichkeit mit vertretbarem Zeitaufwand eine gute Leistung erbringt, da in realen Anwendungen der wahre Grad der Spärlichkeit oft unbekannt oder unklar ist. Mit Hilfe von stochastischen Dimensionsreduktionsmethoden - Variablenscreening und Zufallsprojektion - entwickeln wir eine Ensemble-Methode, um diese Lücke zu schließen. Wir führen eine neuartige datengetriebene Zufallsprojektionsmethode ein, die die zugrundeliegenden Korrelationen zwischen den Prädiktoren nutzt und dadurch die Vorhersageleistung gegenüber herkömmlichen Zufallsprojektionen verbessert. Diese Zufallsprojektion ist insbesondere in Szenarien, in denen die Anzahl der Prädiktoren die Anzahl der Beobachtungen weit übersteigt, hilfreich. Im Mittelpunkt unserer Methode steht ein Ridge-Schätzer für das Variablenscreening, der eine effektive Dimensionsreduktion ermöglicht und gleichzeitig die Beziehung zwischen Zielvariable und Prädiktoren berücksichtigt. Unsere Methode baut ein Ensemble auf, indem sie wiederholt ein probabilistisches Variablenscreening durchführt, gefolgt von der von uns entwickelten datengesteuerten Zufallsprojektion und der Verwendung der reduzierten Prädiktoren in einem (verallgemeinerten) linearen Modell mit niedriger Dimension. Ausführliche Simulationen zeigen, dass unsere Methode herkömmliche Zufallsprojektionsmethoden und klassische spärliche und dichte Methoden bei unterschiedlichen Spärlichkeits- und Kovarianzkonfigurationen in der Variablenreihung und -vorhersage bei geringen Rechenkosten übertrifft. Wir validieren die Effektivität des vorgestellten Algorithmus anhand von realen Datenanwendungen, die kontinuierliche und binäre Zielvariablen sowie Zähldaten beinhalten, und heben seine Vorteile in Bezug auf Interpretierbarkeit und Vorhersagegenauigkeit hervor. Eine Implementierung dieses Ansatzes ist im Paket **spar** in `R` verfügbar, mit dem Benutzer nahtlos prädiktive verallgemeinerte lineare Modelle mit hochdimensionalen Prädiktoren erstellen können.