

Байесовский выбор структур обобщенно-линейных моделей*

А. Д. Толмачев, А. А. Адуенко, В. В. Стрижов

tolmachev.a.d.@phystech.edu; aduenko1@gmail.com; strijov@ccas.ru

В данной работе исследуется проблема мультиколлинеарности и её влияние на эффективность методов выбора признаков. Рассматривается задача выбора признаков и различные подходы к её решению. Исследованы возможности применения байесовского подхода для метода отбора признаков на основе квадратичного программирования. В работе приводятся критерии сравнения различных способов отбора признаков, и проведено сравнение различных методов на тестовых выборках. Сделан вывод об эффективности рассматриваемых подходов на определенных типах данных.

Ключевые слова: *регрессионный анализ; мультиколлинеарность; байесовский подход; выбор признаков; квадратичное программирование*

1 Введение

Работа посвящена анализу применения байесовского подхода для методов отбора признаков и сравнительному анализу различных методов отбора признаков. Предполагается, что исследуемая выборка содержит значительное число мультиколлинеарных признаков. Мультиколлинеарность — это сильная корреляционная связь между отбираемыми для анализа признаками, совместно воздействующими на целевой вектор, которая затрудняет оценивание регрессионных параметров и выявление зависимости между признаками и целевым вектором. Проблема мультиколлинеарности и возможные способы её обнаружения и устранения описаны в [5, 6].

Задача выбора оптимального подмножества признаков является одной из основных задач предварительной обработки данных. Методы выбора признаков основаны на минимизации некоторого функционала, который отражает качество рассматриваемого подмножества признаков. В [1, 2] сделан обзор основных существующих методов отбора признаков.

В [4] предложен новый метод отбора признаков, использующий один из основных методов оптимизации, квадратичное программирование, для задачи отбора признаков. Цель данной работы состоит в анализе возможностей применения байесовского подхода для метода квадратичного программирования в задаче выбора признаков.

Важной частью этой работы является сравнение метода на основе байесовского подхода и других методов отбора признаков, описанных, например, в [3], на различных тестовых выборках.

2 Название раздела

Данный документ демонстрирует оформление статьи, подаваемой в электронную систему подачи статей <http://jmla.org/papers> для публикации в журнале «Машинное обучение и анализ данных». Более подробные инструкции по стилевому файлу `jmla.sty` и использованию издательской системы \LaTeX 2 ϵ находятся в документе `authors-guide.pdf`. Работу над статьёй удобно начинать с правки \TeX -файла данного документа.

*Работа выполнена в рамках курса «Моя первая научная статья», НИУ МФТИ, 2021

Обращаем внимание, что данный документ должен быть сохранен в кодировке UTF-8 without BOM. Для смены кодировки рекомендуется пользоваться текстовыми редакторами Sublime Text или Notepad++.

2.1 Название параграфа

Разделы и параграфы, за исключением списков литературы, нумеруются.

3 Заключение

Желательно, чтобы этот раздел был, причём он не должен дословно повторять аннотацию. Обычно здесь отмечают, каких результатов удалось добиться, какие проблемы остались открытыми.

Литература

- [1] Isabelle Guyon and André Elisseeff. An introduction of variable and feature selection. *J. Machine Learning Research Special Issue on Variable and Feature Selection*, 3:1157 – 1182, 01 2003.
- [2] Nuhu Ibrahim, H.A. Hamid, Shuzlina Rahman, and Simon Fong. Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. *Pertanika Journal of Science and Technology*, 26:329–340, 01 2018.
- [3] Aleksandr Katrutsa and Vadim Strijov. Stresstest procedure for feature selection algorithms. *Chemometrics and Intelligent Laboratory Systems*, 142, 02 2015.
- [4] Alexandr Katrutsa and Vadim V. Strijov. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. *Expert Syst. Appl*, 76:1–11, 2017.
- [5] Edward Leamer. Multicollinearity: A bayesian interpretation. *The Review of Economics and Statistics*, 55(3):371–80, 1973.
- [6] Ron Snee. Regression diagnostics: Identifying influential data and sources of collinearity, book review. *Journal of Quality Technology*, 01 1980.

Поступила в редакцию