

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 ее 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 Постановка задачи

2.1 Применение квадратичной оптимизации для задачи отбора признаков

В [4] предлагается подход с применением квадратичной оптимизации для задачи выбора признаков. Основная идея предлагаемого подхода заключается в минимизации количества схожих признаков и максимизации количества релевантных признаков. Пусть \mathcal{J} — множество признаков в рассматриваемой модели, и $|\mathcal{J}| = n$. Зададим функционал $Q(\mathbf{a}) = \mathbf{a}^T \mathbf{Q} \mathbf{a} - \mathbf{b}^T \mathbf{a}$, где $\mathbf{a} \in \mathbb{R}^n$ и $\mathbf{Q} \in \mathbb{R}^{n \times n}$ — матрица схожести признаков, а $\mathbf{b} \in \mathbb{R}^n$ — вектор

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

релевантности признаков с целевым вектором. Матрицу \mathbf{Q} и вектор \mathbf{b} будем представлять как функции Sim и Rel соответственно, где $\text{Sim}: \mathcal{J} \times \mathcal{J} \rightarrow [0, 1]$, $\text{Rel}: \mathcal{J} \rightarrow [0, 1]$. Таким образом, необходимо решить задачу оптимизации:

$$\mathbf{a}^* = \arg \min_{\mathbf{a} \in \mathbb{B}^n} Q(\mathbf{a}).$$

Важно отметить, что задача целочисленного квадратичного программирования, сформулированная выше, является **NP**-полной, так как поиск минимума функции \mathbf{Q} ведется по вершинам булева куба $\mathbb{B}^n = \{0, 1\}^n$. Поэтому, чтобы можно было применять различные методы выпуклой оптимизации будем искать минимум функции по выпуклой оболочке булева куба $\text{Conv}(\mathbb{B}^n) = [0, 1]^n$.

Тогда получаем следующую задачу выпуклой оптимизации:

$$\begin{cases} \mathbf{z}^* = \arg \min_{\mathbf{z} \in [0, 1]^n} \mathbf{z}^T \mathbf{Q} \mathbf{z} - \mathbf{b}^T \mathbf{z} \\ \|\mathbf{z}\|_1 \leq 1 \end{cases}$$

Здесь, \mathbf{Q} и \mathbf{b} по-прежнему являются матрицей схожести признаков и вектором релевантности признаков соответственно. В данной работе функции Sym и Rel (или, другими словами, матрица \mathbf{Q} и вектор \mathbf{b} в обозначениях выше) задаются заранее до применения метода на основе сходств между признаками в датасете.

Далее положим, τ - пороговое значение для отбора признаков в данном методе, т.е. $z_i^* > \tau \Leftrightarrow a_i^* = 1 \Leftrightarrow j \in \mathcal{A}$, где $\mathcal{A} \subset \mathcal{J}$ - множество отобранных методом признаков.

Далее предлагается рассмотреть возможные применения байесовского подхода к данному методу квадратичного программирования.

2.2 Данные для экспериментов

В качестве данных для экспериментов по проверке предложенных подходов мы используем синтетические наборы данных из работы [3], в которых рассматриваются различные типы зависимости признаков между собой и с целевым вектором. Кроме того, будут проведены эксперименты и на собственных генерированных синтетических данных.

3 Базовый эксперимент

3.1 Цель

Как сказано в [4] метод квадратичного программирования улучшает качество отбора признаков для многих типов выборок. Однако, не во всех случаях этот метод дает наилучшие результаты. Цель базового эксперимента заключается в поиске и рассмотрении выборок с мультиколлинеарными признаками, на которых методу квадратичного программирования не удастся провести качественный отбор признаков. Далее планируется исследовать, в чем сходство выборок, на которых метод квадратичного программирования дает неоптимальные результаты, чтобы учесть полученные закономерности при разработке нового метода отбора признаков на основе байесовского подхода.

3.2 Описание данных

В качестве базового эксперимента рассмотрим синтетические данные и применим на них метод квадратичного программирования QBFS, описанный выше, для поиска значения \mathbf{a}^* , при котором значение функционала $Q(\mathbf{a})$ принимает наименьшее значение.

3.3 Результаты эксперимента

Результаты эксперимента и базовый код будут загружены немного позже. Надеюсь на понимание.

4 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 находятся в документе `authors-guide.pdf`. Работу над статьёй удобно начинать с правки \TeX -файла данного документа.

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++.

4.1 Название параграфа

Разделы и параграфы, за исключением списков литературы, нумеруются.

5 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