

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ОБРАБОТКИ И КЛАССИФИКАЦИИ СПЕКТРАЛЬНЫХ ДАННЫХ НОВООБРАЗОВАНИЙ КОЖИ

Е.О. Сальникова¹, Ю.А. Христофорова¹

¹Самарский университет, г. Самара
e-mail: fume890@outlook.com

Аннотация

В работе рассматриваются два метода предварительной обработки спектров рамановского рассеяния кожи (с удалением базовой линии и без удаления) и две модели для их классификации на основе линейных дискриминантных методов, а также оценивается влияние, которое выбранные методы предобработки и машинного обучения оказывают на качество классификации. Результаты свидетельствуют о том, что выбранные методы предобработки значимо не влияют на качество классификации, но значимое улучшение достигается для модели на основе дискриминантного анализа частных наименьших квадратов.

Ключевые слова: метод ANOVA, линейный дискриминантный анализ (LDA), метод PLS-DA, кросс-валидация (K-fold), метод ALS, *ex vivo* спектры рамановского рассеяния кожи, рак кожи.

Введение

Рак кожи относится к числу наиболее распространённых онкологических заболеваний в мире и представляет собой значимую медико-социальную проблему. В структуре заболеваемости доминируют немеланомные формы, такие как базальноклеточный и плоскоклеточный рак кожи, тогда как наибольшую клиническую опасность представляет меланома, отличающаяся высокой степенью агрессивности и выраженным метастатическим потенциалом. На сегодняшний день инвазивная биопсия с последующим гистологическим исследованием остаётся «золотым стандартом» диагностики [1]. Однако, несмотря на высокую диагностическую точность, данный метод связан с рядом рисков, включая рубцевание тканей, повреждение нервных структур и возможную диссеминацию опухолевых клеток. Дополнительной проблемой является низкая специфичность отбора новообразований для проведения биопсии, что приводит к значительному числу необоснованных хирургических вмешательств, поскольку большинство удаляемых образований оказываются доброкачественными [2]. В связи с этим актуальной задачей является разработка эффективных неинвазивных методов предварительной диагностики рака кожи.

В рамках настоящей работы решается задача построения бинарных классификаторов на основе методов машинного обучения для анализа рамановских спектров кожи с целью предварительного определения рака.

1. Материалы и методы

Исходные данные – это *ex vivo* спектры кожи пациентов с гистологически подтвержденными злокачественными и доброкачественными опухолями. Спектры регистрировались с помощью портативного устройства с длиной волны зондирующего излучения 785 нм как в зоне новообразования, так и на здоровой коже. Измерения производились с временем экспозиции равным 3-м секундам и пятикратным усреднением при температуре детектора -15°C . Таким образом был составлен датасет, включающий в себя 377 спектров. Спектры регистрировались в диапазоне от 100 до 2300 cm^{-1} . Анализ спектров производился в наиболее информативном диапазоне от 265 до 2000 cm^{-1} .

Для оценки влияния методов предварительной обработки на результат классификации предобработка была исполнена в двух вариантах для одних и тех же спектров. В обоих вариантах выполнялось сглаживание фильтром Савицкого–Голея (ширина окна 15, порядок полинома 2) и нормировка по среднеквадратичному отклонению, однако во втором варианте дополнительно к этим методам использовалось вычитание базовой линии методом асимметричных наименьших квадратов (ALS) с параметрами $p = 0,01$, $\lambda = 10^5$, $N = 10$ [3]. Визуализация спектров после каждого из двух вариантов обработки приведена на рисунке 1 (а,б).

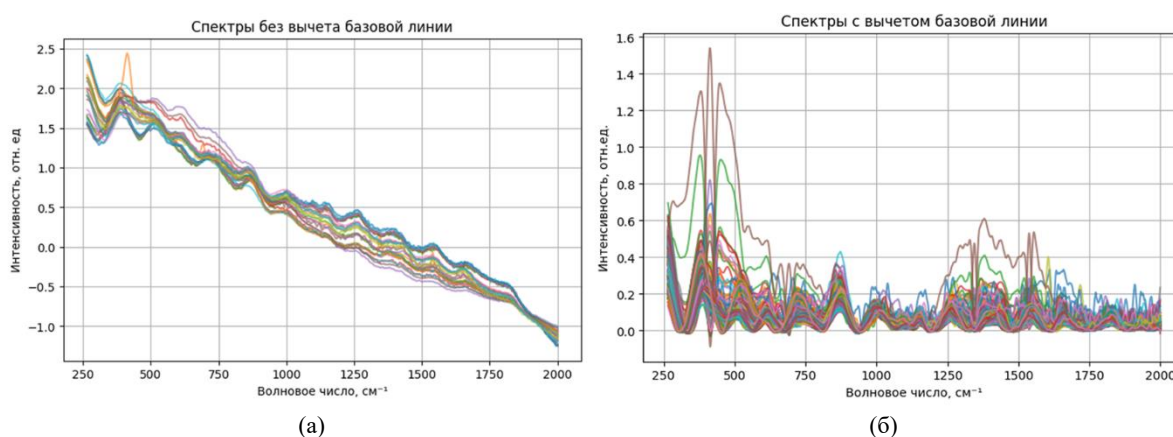


Рисунок 1 – Спектры рамановского рассеяния кожи после предварительной обработки без удаления базовой линии (а) и с удалением (б)

Для решения задачи классификации были реализованы две модели: первая основана на методе ANOVA с последующим линейным дискриминантным анализом (LDA), вторая – на методе PLS-DA [4]. Обе модели оценивались с использованием 10-кратной кросс-валидации (K-fold). Классификация выполнялась для двух классов: «норма» (183 спектра) и «патология» (194 спектра).

2. Результаты

В качестве метрики качества классификации использовалась площадь под ROC-кривой (ROC AUC), позволяющая количественно оценить разделяющую способность моделей. Результаты классификации обеих моделей для спектров, обработанных по первому алгоритму предобработки (без удаления базовой линии), представлены на рисунке 2, а для спектров с удаленной базовой линией – на рисунке 3.

Для спектров, не подвергавшихся коррекции базовой линии, установлено, что классификатор, основанный на комбинации ANOVA и LDA, демонстрирует относительно низкое качество классификации (ROC AUC = $0,536 \pm 0,073$). В то же время модель PLS-DA показывает более высокие значения данной метрики (ROC AUC = $0,698 \pm 0,114$), что свидетельствует о её лучшей способности к разделению классов в исходных данных.

При использовании предобработки с удалением базовой линии методом ALS наблюдается некоторое улучшение качества для модели ANOVA + LDA (ROC AUC = $0,563 \pm 0,055$). Однако для модели PLS-DA значения ROC AUC несколько снижаются (до $0,672 \pm 0,096$), оставаясь при этом выше, чем у альтернативной модели.

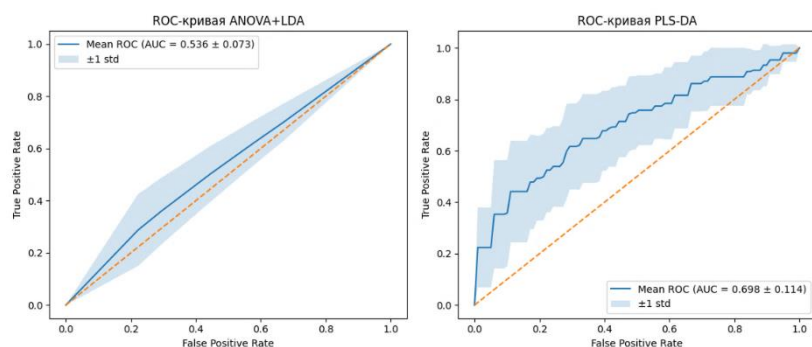


Рисунок 2 – ROC-кривые моделей на основе ANOVA с LDA и PLS-DA для спектров без вычета базовой линии

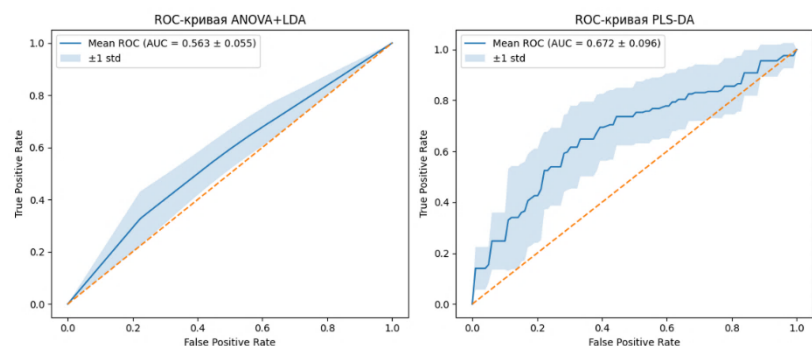


Рисунок 3 – ROC-кривые моделей на основе ANOVA с LDA и PLS-DA для спектров с вычетом базовой линии

Таким образом, вне зависимости от применения коррекции базовой линии, модель PLS-DA демонстрирует более высокие значения ROC AUC по сравнению с ANOVA + LDA.

Заключение

Результаты статистического анализа показали, что классификатор на основе PLS-DA обладает статистически значимым преимуществом по сравнению с подходом ANOVA + LDA. Это подтверждено с использованием парного t-теста при уровне значимости $p = 0,05$ как для спектров с коррекцией базовой линии, так и без неё.

Полученные данные позволяют сделать вывод о том, что метод PLS-DA является более эффективным инструментом для решения задачи классификации в рамках рассматриваемого эксперимента. Его преимущество, вероятно, связано с лучшей способностью учитывать скрытую структуру данных и взаимосвязи между признаками.

В то же время сравнительный анализ моделей одного типа, обученных на спектрах с различными вариантами предобработки (с удалением базовой линии и без него), не выявил статистически значимых различий. Это указывает на то, что в рамках данного исследования выбранные методы предобработки оказывают сопоставимое влияние на качество классификации.

В целом можно заключить, что выбор алгоритма классификации оказывает более существенное влияние на итоговый результат, чем применение рассматриваемого метода коррекции базовой линии на этапе предварительной обработки спектров.

Список использованных источников

1. Дерматоонкопатология: Иллюстрированное руководство для врачей / И. Н. Чупров, А. А. Сыдилов, Д. В. Заславский, Р. А. Насыров. – 2021. – 27 с.
2. English, D. R. Factors influencing the number needed to excise: excision rates of pigmented lesions by general practitioners / D. R. English // The Medical journal of Australia, 2004. – 18 p.
3. Peng, S. Asymmetric least squares for multiple spectra baseline correction / S. Peng // Analytica Chimica Acta. – 2010. – P. 64–66.
4. Lee, L. C. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data / L. C. Lee, C. Y. Liong, A. A. Jemain // Analyst. – 2018. – P. 3527.

COMPARATIVE ANALYSIS OF METHODS FOR PROCESSING AND CLASSIFICATION OF SPECTRAL DATA OF SKIN NEOPLASMS

E.O. Salnikova¹, Y.A. Khristoforova¹

¹*Samara University, Samara
e-mail: fume890@outlook.com*

Abstract

The study examines two preprocessing methods for Raman scattering spectra of the skin (with and without baseline correction) and two classification models based on linear discriminant approaches. It also evaluates the impact of the chosen preprocessing and machine learning methods on classification performance. The results indicate that the selected preprocessing methods do not significantly affect classification quality, while a substantial improvement is achieved with the model based on partial least squares discriminant analysis.

Keywords: ANOVA method, Linear Discriminant Analysis (LDA), PLS-DA method, K-fold cross-validation, ALS method, ex vivo Raman spectra of skin, skin cancer.