

КЛАССИФИКАЦИЯ РАМАНОВСКИХ СПЕКТРОВ СЫВОРОТКИ ПАЦИЕНТОВ С ИСПОЛЬЗОВАНИЕМ PLS-DA И АУГМЕНТАЦИИ MIXUP

А.И. Дьяченко¹, Ю.А. Христофорова¹

¹Самарский университет, г. Самара
e-mail: diychenkosmav@gmail.com

Аннотация

В работе рассматривается задача бинарной классификации рамановских спектров сыворотки пациентов с хронической сердечной недостаточностью и респираторными заболеваниями. Спектры были предварительно обработаны, включая удаление базовой линии методом асимметричных наименьших квадратов, сглаживание и нормализацию на стандартное отклонение. Для классификации использовался метод Partial Least Squares Discriminant Analysis. Исследовано влияние аугментации данных методом Mixup при различном увеличении объема выборки на значение площади под ROC-кривой.

Ключевые слова: рамановская спектроскопия, PLS-DA, Mixup, аугментация данных, ROC AUC, хроническая сердечная недостаточность, респираторные заболевания.

Введение

Рамановская спектроскопия широко применяется для анализа биологических образцов благодаря своей чувствительности к молекулярному составу тканей и биожидкостей [1]. В сочетании с методами машинного обучения она представляет собой эффективный инструмент для решения задач медицинской диагностики, в том числе при дифференциации заболеваний различной этиологии.

Одной из ключевых проблем при анализе медицинских спектральных данных является ограниченный объем выборок, что может приводить к нестабильности классификационных моделей и завышенной оценке качества. В связи с этим особый интерес представляет использование методов аугментации данных, позволяющих искусственно увеличивать объем обучающей выборки без проведения дополнительных измерений [2].

Целью данной работы является анализ эффективности классификации рамановских спектров сыворотки пациентов с хронической сердечной недостаточностью (ХСН) и респираторными заболеваниями с использованием метода PLS-DA, а также оценка влияния аугментации данных методом Mixup и размера выборки на значение метрики ROC AUC.

1. Материалы и методы

В исследовании использовался датасет рамановских спектров сыворотки пациентов, относящихся к двум нозологическим группам: хроническая сердечная недостаточность и респираторные заболевания. Классификация выполнялась в бинарной

постановке задачи. Размер реальной выборки варьировался от 50 до 216 спектров на класс, что позволило исследовать влияние объёма данных на качество классификации.

Все спектры подвергались предварительной обработке, включающей удаление базовой линии методом асимметричных наименьших квадратов (ALS), сглаживание спектров и нормализацию интенсивности на стандартное отклонение. Данные этапы обеспечивали снижение влияния шумов и фоновых искажений и повышали устойчивость последующей классификации.

В качестве классификатора использовался метод PLS-DA, эффективный при работе с высокоразмерными и коррелированными спектральными признаками [3]. Для повышения устойчивости модели и предотвращения переобучения выполнялся подбор оптимального числа латентных переменных на основе кросс-валидации. В качестве критерия оптимальности использовалась минимизация среднеквадратичной ошибки (RMSE).

В результате анализа установлено, что оптимальное число латентных переменных составляет 9, при котором достигается минимальное значение ошибки $RMSE = 0.2463 \pm \pm 0.1229$. Дальнейшее увеличение числа латентных переменных не приводило к улучшению качества модели и сопровождалось ростом вариативности результатов, что указывает на отсутствие дополнительной полезной информации в последующих компонентах.

Соответствующий график зависимости значения ошибки от количества латентных переменных представлен на рисунке 1.

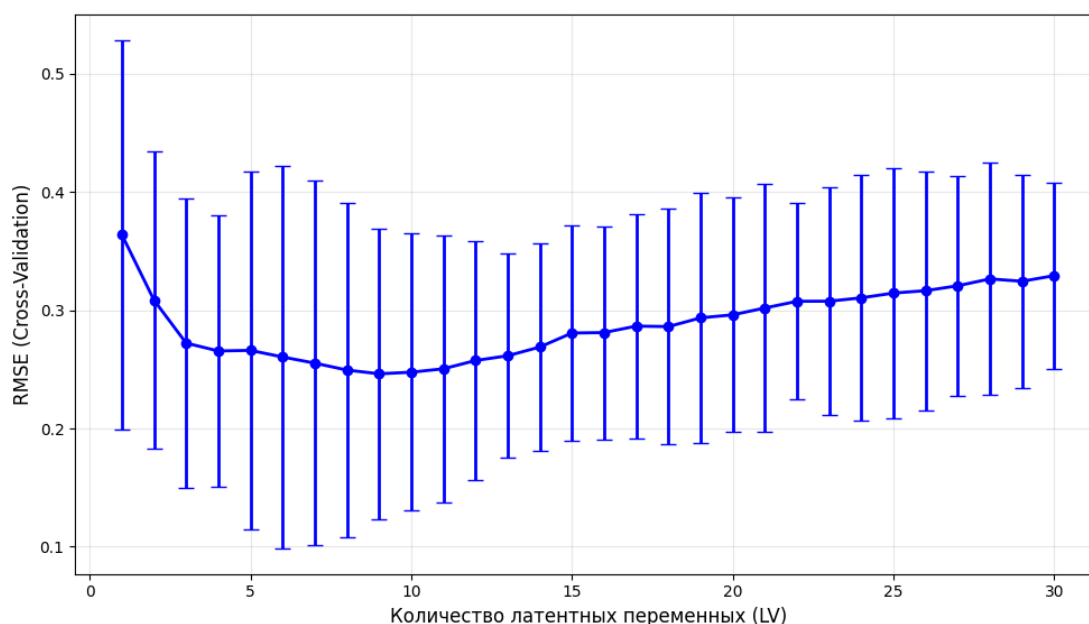


Рисунок 1 – Зависимость RMSE от числа латентных переменных модели PLS-DA

Для увеличения объёма обучающей выборки применялся метод аугментации Миксир, заключающийся в формировании новых спектров путём линейной комбинации исходных данных [4]. Рассматривались различные коэффициенты аугментации, соответствующие увеличению эффективного размера выборки до 432, 1080, 2160, 4320 и 10800 спектров.

Качество классификации оценивалось по метрике ROC AUC. Для каждого эксперимента вычислялись среднее значение ROC AUC и стандартное отклонение по результатам кросс-валидации.

2. Результаты

Для реальных данных без аугментации установлено, что увеличение размера выборки приводит к росту качества классификации. Так, при размере класса 50 среднее значение ROC AUC составило 0.955 ± 0.031 , тогда как при размере 216 спектров было достигнуто значение 0.986 ± 0.006 , являющееся наилучшим результатом для реальных данных.

Применение аугментации Mixup позволило дополнительно повысить устойчивость модели. Максимальное значение ROC AUC = 0.987 ± 0.007 было получено при увеличении эффективного размера выборки до 4320 спектров, что соответствует коэффициенту аугментации 20. При этом улучшение по сравнению с полной реальной выборкой составило около 0.06%.

На рисунке 2 представлена зависимость среднего значения ROC AUC от эффективного размера выборки. Видно, что использование аугментации позволяет достичь качества классификации, сопоставимого с результатами, полученными на реальных данных меньшего объёма, при сохранении низкого стандартного отклонения.

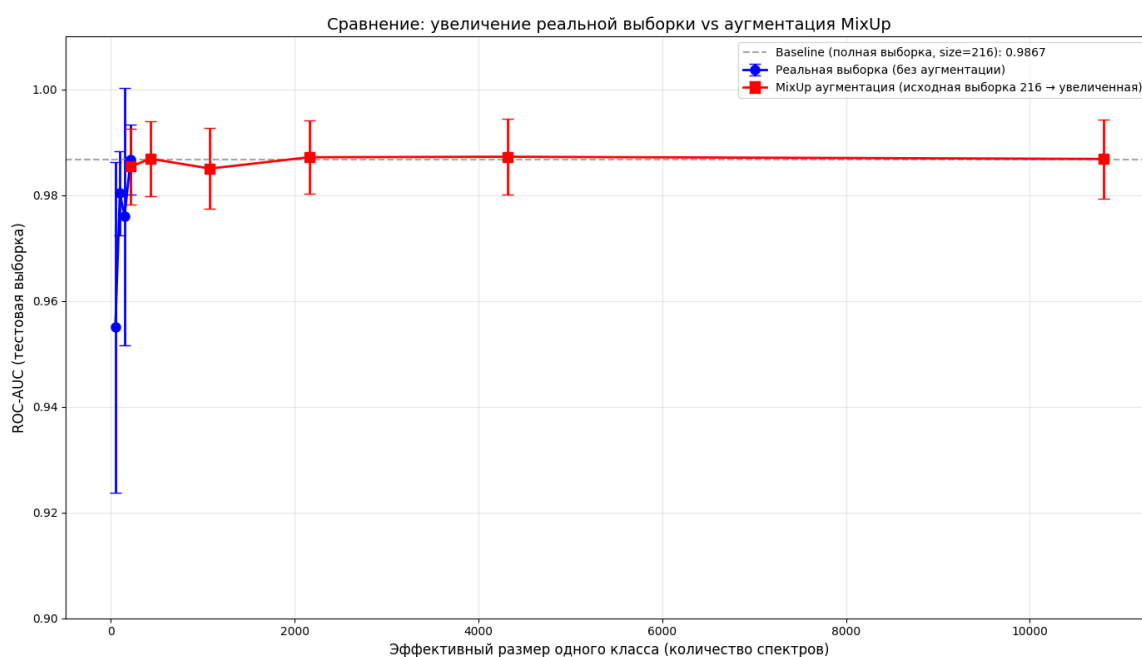


Рисунок 2 – Зависимость значения ROC AUC от эффективного размера выборки при использовании аугментации Mixup

Заключение

В работе проведён анализ классификации рамановских спектров сыворотки пациентов с хронической сердечной недостаточностью и респираторными заболеваниями с использованием метода PLS-DA. Показано, что увеличение объёма реальной выборки до 216 спектров способствует росту значения ROC AUC и снижению вариативности результатов.

Установлено, что применение аугментации данных методом Mixup позволяет достичь качества классификации, сопоставимого с результатами, полученными на полной реальной выборке. Это свидетельствует об эффективности данного подхода для анализа медицинских спектральных данных в условиях ограниченного объема экспериментальных измерений.

Благодарности

Работа выполнена при поддержке гранта РФФИ № 25-75-00146 (<https://rscf.ru/project/25-75-00146/>).

Список использованных источников

1. Popp, J. Modern Raman spectroscopy for biomedical applications: A variety of Raman spectroscopical techniques on the threshold of biomedical applications / J. Popp, C. Krafft, T. Mayerhöfer // *Optik & Photonik*. – 2011. – V. 6, № 4. – P. 24–28.
2. Shorten, C. A survey on image data augmentation for deep learning / C. Shorten, T.M. Khoshgoftaar // *Journal of Big Data*. – 2019. – V. 6.
3. Lee, L.C. Partial least squares-discriminant analysis for classification of high-dimensional data / L.C. Lee, C.Y. Liong, A.A. Jemain // *Analyst*. – 2018. – P. 3527–3536.
4. Zhang H., mixup: Beyond empirical risk minimization / H. Zhang [et al.] // *ICLR*. – 2018.

CLASSIFICATION OF SERUM RAMAN SPECTRA USING PLS-DA AND MIXUP DATA AUGMENTATION

A.I. Diychenko¹, Y.A. Khristoforova¹

*¹Samara University, Samara
e-mail: diychenkosmav@gmail.com*

Abstract

This paper addresses the binary classification of Raman spectra from patients with chronic heart failure and respiratory diseases. The spectra were preprocessed using asymmetric least squares baseline correction, smoothing, and normalization. A Partial Least Squares Discriminant Analysis classifier was applied with optimization of the number of latent variables based on cross-validation. The effect of Mixup data augmentation with different dataset expansion factors on Receiver Operating Characteristic – Area Under Curve performance was also investigated.

Keywords: Raman spectroscopy, PLS-DA, Mixup, data augmentation, ROC AUC, chronic heart failure, respiratory diseases.