Эффективное понижения размерности пространства признаков на основе LASSO-регрессии в задачах   
распознавания эмоций по речи

маг. Краснопрошин Д.В., доц. Вашкевич М.И.

Белорусский государственный университет информатики и радиоэлектроники  
ул. П. Бровки, 6, БГУИР, каф. ЭВС, 220013, Минск, Беларусь,   
e-mail: vashkevich@bsuir.by

В докладе показаны различные техники слияния медицинских изображений в области коэффициентов дискретного вейвлет-преобразования (ДВП), полученных как из известных функций, так и на основе целочисленного полифазного представления банка фильтров. Также предложен вариант слияния, использующий представление НЧ-компонент ДВП в области модуля максимума вейвлет-преобразования. Проведено попарное сравнение методов по метрике взаимной информации с применением статистических тестов. Результаты показали, что использование полифазного банка фильтров, а также представление НЧ-компонент ДВП в области модуля максимума вейвлет-преобразования позволяет получить лучше результаты слияния, чем использование двумерного ДВП с функциями Хаара.

# Введение.

Распознавание эмоций в речи представляет собой важный аспект, оказывающий влияние на технологии искусственного интеллекта (ИИ).

Одним из перспективных направлений в распознавании эмоций в речи является использование глубокого обучения для извлечения высокоуровневых признаков из аудиоданных. Многие исследования фокусируются на применении сверточных нейронных сетей и рекуррентных нейронных сетей, что позволяет более эффективно улавливать временные и частотные закономерности в речевых сигналах.

*– TODO: Отметить недостатки нейронных сетей и сказать, почему актуальны исследования с более простыми подходами. Отметить 1) высокую вычислительную эффективность; 2) интерпретируемость.*

Тем не менее, нейросетевые технологии имеют свои недостатки, которые могут ограничивать их применимость в некоторых сценариях исследований. Одним из основных недостатков нейронных сетей является высокая вычислительная сложность, особенно при обучении глубоких моделей на больших объемах данных. Это может создавать ограничения в доступности ресурсов для исследователей с ограниченным бюджетом или доступом к высокопроизводительным вычислительным системам.

Кроме того, нейронные сети часто характеризуются низкой интерпретируемостью, что затрудняет понимание причинно-следственных связей между входными данными и выходными прогнозами модели. Это усложняет анализ результатов и eго понимание, почему модель принимает определенные решения, что может быть нежелательным в некоторых приложениях, особенно в медицинской диагностике или правовых областях.

Исследования с использованием более простых подходов остаются актуальными и важными в свете этих недостатков нейронных сетей. Во-первых, они обладают высокой вычислительной эффективностью, что позволяет проводить анализ данных на обычных компьютерах или локальных вычислительных кластерах без необходимости доступа к дорогостоящим вычислительным ресурсам. Во-вторых, они обеспечивают более высокую интерпретируемость результатов, что позволяет исследователям лучше понимать, какие признаки или характеристики влияют на конечный результат классификации.

В контексте использования статистических моделей, таких как машины опорных векторов (МОВ), актуальной является задача понижения размерности признакового пространства. Это обусловлено не только стремлением к оптимизации вычислительной сложности алгоритма распознавания, но и стремлением к улучшению интерпретируемости модели. Сокращение размерности характеристического вектора способствует тому, что в результате остаются только признаки имеющие ключевое значение для процесса принятия решения. Все это приводит к интерпретируемости модели, делая ее результаты более понятными и прозрачными для конечного пользователя.

# 1. Общая архитектура системы распознавания

– краткое описание

– рисунок/схема

1) **Ввод данных:**

В системе предусмотрен ввод аудиозаписей, содержащих речь, как основного источника информации для анализа эмоциональной окраски. Для каждой аудиозаписи проводится предварительная обработка, включая извлечение характеристик а именно мел-частотных кепстральных коэффициентов, их первой и второй производных, а также ряда статистических показателей.

2) **Отбор признаков:**

Применяется простой статистический метод, а именно LASSO для выбора наиболее информативных характеристик. Это позволяет снизить размерность признакового пространства и улучшить обобщающую способность модели, сохраняя при этом высокую точность классификации.

3) **Модель классификации:**

В качестве классификатора используется модель на основе линейного дискриминантного анализа.

Линейный дискриминантный анализ (ЛДА) — это статистический метод машинного обучения, который направлен на поиск линейных комбинаций признаков, наиболее эффективно разделяющих данные на различные классы. Принцип работы ЛДА заключается в поиске такого линейного преобразования признаков, которое максимизирует отношение межклассовой дисперсии к внутриклассовой дисперсии. Это достигается путем нахождения оптимальных весовых коэффициентов для линейной комбинации признаков, которые обеспечивают максимальное разделение между классами данных.

4) **Выходные данные:**

В результате обработки система выдает прогноз эмоциональной окраски для каждой аудиозаписи, отображая вероятности принадлежности к различным эмоциональным категориям.

# 2. Процедура отбора признаков

(САМОЕ ГЛАВНОЕ, ЧТО МЫ ХОТИМ ОПИСАТЬ – главный акцент!)

– Сказать, что мы решаем многоклассовую задачу;

– Сведенение K-классовой задачи к K задачам бинарной классификации;

– Описание Lasso для ранжирования признаков, отбор сокращенного числа признаков после ранжирования.

– Объединение признаков, полученных для K бинарных классификаторов.

В рамках данной работы особое внимание уделялось отбору признаков. По мнению авторов это может способствовать повышение эффективности и интерпретируемости модели распознавания эмоций в речевом сигнале. В данном контексте, решая многоклассовую задачу, авторы стремились найти оптимальный подход, который включает в себя следующие важные этапы:

1) Сведение к бинарным классификациям: многоклассовая задача эффективно разбивается на K отдельных бинарных задач классификации, где каждая эмоция рассматривается отдельно. Это позволяет более точно адаптировать модель к каждой эмоциональной категории.

2) Использование МЧКК для извлечения характеристик: МЧКК служит основой для извлечения признаков из речевых сигналов, предоставляя детальный аудио-анализ. Этот шаг формирует входные данные для дальнейшей обработки.

3) Применение метода Lasso для ранжирования признаков: процесс отбора признаков включает в себя использование метода Lasso. Этот механизм не только выделяет наиболее важные характеристики, но и присваивает им веса, что дополнительно улучшает качество модели.

Следует отметить, что **метод Lasso (Least Absolute Shrinkage and Selection Operator)** представляет собой технику регуляризации, которая используется для отбора признаков и снижения размерности данных. Он вводит штраф на сумму абсолютных значений коэффициентов регрессии, что позволяет занулить некоторые из них. В контексте ранжирования признаков, Lasso выделяет наиболее важные характеристики и придает им значимые веса. Алгоритм Lasso минимизирует следующую функцию стоимости:

|  |  |
| --- | --- |
|  | (1) |

где: – количество наблюдений,

p - количество признаков,

yi - целевая переменная для i-го наблюдения,

xij- значение j-го признака для i-го наблюдения,

βj- коэффициент регрессии для j-го признака,

λ - параметр регуляризации.

Параметр λ регулирует степень регуляризации. При увеличении λ Lasso уменьшает веса признаков, зануляя некоторые из них.

4) Отбор сокращенного числа признаков после ранжирования: после применения Lasso и ранжирования признаков по их важности, следующим шагом является отбор сокращенного числа признаков. Этот этап позволяет выбрать оптимальное подмножество наиболее значимых признаков, снижая размерность данных и устраняя избыточность.

Одним из подходов к отбору сокращенного числа признаков является установка порогового значения на веса признаков и отбор только тех, чьи веса превышают этот порог. Такой механизм позволяет создать более простую и интерпретируемую модель, которая все еще обладает высокой предсказательной силой.

5) Объединение признаков для K бинарных классификаторов: признаки, выделенные для каждой из K бинарных задач классификации, объединяются. Этот этап обеспечивает комплексный и гармоничный подход к распознаванию эмоций в речевом сигнале, где каждая эмоция получает оптимальный вес в итоговой модели.

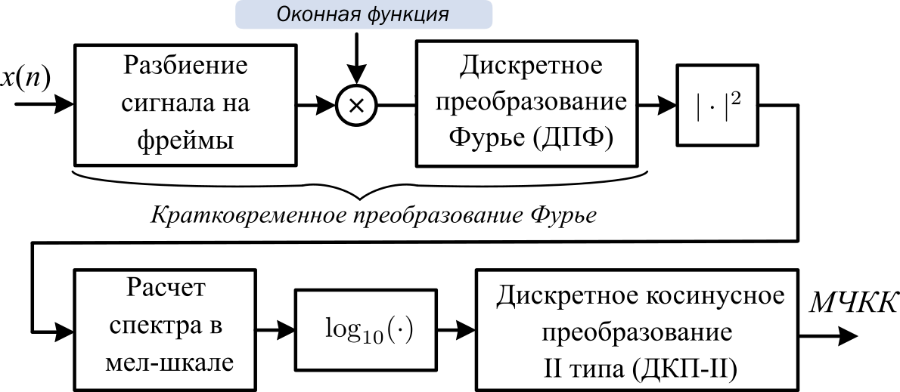
Согласно предположению авторов такая стратегия отбора признаков может помочь не только обеспечивает точность в решении многоклассовой задачи, но и активно поспособствовать интерпретируемости модели, делая ее результаты более понятными и прозрачными.

# 3. Извлечение признаков:

– КРАТКОЕ описание извлечения признаков, включая мел-кепстральные частотные коэффициенты и их производные.

- Внедрение статистических моментов (skewness, kurtosis, interquantile range) в матрицу признаков для более полного охвата характеристик аудиосигнала.

На начальном этапе в системе распознавания эмоций по речи выполняется предварительная обработка входных аудиоданных [1, 5]. В данной работе речевые признаки рассчитывались на основании МЧКК [8]. Расчет МЧКК относится к методам кратковременного анализа речевого сигнала, которые предполагают разбиение сигнала на фреймы (короткие сегменты). Считается, что в интервале от 10 до 30 мс голосовой сигнал можно считать стационарным. На рис. 1 представлена схема вычисления МЧКК.



В итоговый набор *исходных признаков* были включены среднее значение МЧКК ( признаков), среднеквадратичное отклонение МЧКК ( признаков), среднее от первой и второй производных от МЧКК (признаков), их среднеквадратическое отклонение (), а также коэффициент асимметрии, эксцесс и межквантильный размах (по признаков для каждой характеристики соответственно). Таким образом, для каждого аудиофайла мы получаем -компонентный вектор надсегментных признаков МЧКК. Использовагние таких статистических показателей как skewness, kurtosis и interquantile range, а также первой и второй произовных от МЧКК обусловлено стремеление наиболее полно охватить характеристики аудиосигнала.

Полученный базовый набор признаков далее подавался на вход системе, которая включала в себя метод Lasso для ранжирования признаков и отбора оптимального подмножества характеристик, как было сказано выше.

Для тестирования классификатора использовался метод перекрестной проверки по  блокам (*k-fold cross-validation*) [9], который заключается в следующем.

1) Разделить набор разбивается на *k* блоков;

2) В цикле для  выполняются следующие операции:

– блок  устанавливается в качестве тестового набора данных (*test data*)

– оставшиеся блоки принимаются как тренировочные данных (*train data*)

– выполняется обучение модели классификатора на тренировочных и оценивоется ее эффективность на тестовых данных

– сохранение результатов классификации для данных из тестового набора;

– сброс параметров модели до исходного состояния для следующей итерации.

3) Расчет оценки эффективности модели на основе сохраненных результатов классификации данных.

В данной работе данных были разбиты на блоки следующим образом (в скобках указаны номера актеров):

- блок 0: (2, 5, 14, 15, 16);

- блок 1: (3, 6, 7, 13, 18);

- блок 2: (10, 11, 12, 19, 20);

- блок 3: (8, 17, 21, 23, 24);

- блок 4: (1, 4, 9, 22).

Такой порядок разбиения был предложен в [12]. Выбранная стратегия заключается в том, что каждый блок должен содержать одинаковое количество случайно выбранных образцов для каждого класса. При этом должно выполняться условие, что каждый актер представлен либо обучающей, либо валидационной выборке, но не в обоих.

**Набор данных RAVDESS.**

Для проведения нашего исследования мы использовали набор данных Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [10]. RAVDESS предоставляет уникальную возможность исследования эмоциональных аспектов речи, содержа записи от 24 актеров (12 мужчин и 12 женщин), включая 104 разнообразных высказывания на каждого актера (60 речевых и 44 песенных). В контексте нашей работы, мы ограничились использованием речевых высказываний из RAVDESS, что составило 1440 аудиофайлов в формате wav (16 бит, 48 кГц) с учетом двойной записи каждой вокализации. Этот набор данных обеспечивает широкий спектр эмоциональных состояний, включая нейтральность, спокойствие, счастье, грусть, гнев, страх, удивление и отвращение. Важно подчеркнуть, что эмоциональные состояния были представлены на двух уровнях громкости, что улучшает обучение моделей в условиях повседневного разнообразия эмоций в реальных сценариях общения.

– Обоснование выбора данной базы данных для нашего исследования.

Данный датасет обладает контролируемыми условиями записи, разнообразием дикторов и эмоций, а также высоким качеством данных, что делает его надежным инструментом для анализа эмоций в речи.

6. Результаты экспериментов

– Представление результатов процесса отбора признаков и их влияния на качество классификации эмоций в речи.

– Подробное рассмотрение ключевых изменений в матрице признаков после применения методов отбора.

– Детальное сопоставление размеров исходной и отобранной матриц признаков.

– Обсуждение результатов в свете сохранения высокой точности классификации при снижении размерности.

– Сопоставление использования различных методов классификации (ЛДА, МОВ, деревья?)

На первом этапе эксперимента мы провели процедуру переопределения и нормализации данных. Далее, для каждой эмоции мы использовали метод LASSO для ранжирования признаков в порядке их важности для классификации. Для этого мы обучили модель LASSO на различных значениях параметра регуляризации и выбрали оптимальные признаки, которые наилучшим образом коррелируют с каждой из рассматриваемых эмоций.

Дополнительно стоит отметить, что в процессе упорядочивания и оптимизации количества признаков большая часть удаленных характеристик связана с изначально извлеченными мел-кепстральными коэффициентами. В то же время, первая и вторая производные, среднеквадратическое отклонение, а также коэффициенты асимметрии, эксцесса и межквантильный размах, вычисленные на основе исходных мел-кепстральных коэффициентов, почти не были затронуты в процессе отбора признаков. Это свидетельствует о высокой информативности указанных характеристик и их важной роли в распознавании эмоциональных состояний на основе аудиоданных.

Затем мы применили метод анализа дискриминантных признаков (LDA) для оценки классификационной эффективности на основе отобранных признаков. Для каждой эмоции мы последовательно добавляли признаки в модель LDA и оценивали её производительность с помощью метрики UAR (Unweighted Average Recall) на кросс-валидации по оставшимся наблюдениям.

Наконец, мы провели анализ ошибок классификации с помощью матрицы спутанности и оценили общую производительность модели на основе метрики UAR.

В итоге, после процедуры отбора признаков из изначального набора, состоящего из 306 характеристик, оставлено 205 признаков при относительном сохранении точности классификатора. Этот результат демонстрирует значительное уменьшение размерности признакового пространства, что является важным шагом в оптимизации аналитических процессов. Одновременно удалось сохранить высокий уровень классификационной точности, что подчеркивает эффективность примененной методологии отбора признаков.

Уменьшение размерности признакового пространства приводит к снижению вычислительной нагрузки и улучшению обобщающей способности классификатора. Это позволяет эффективнее обрабатывать и анализировать данные, что особенно важно в контексте ресурсозатратных задач, таких как анализ эмоций на основе звуковых данных. Более того, сохранение высокой классификационной точности при сокращении числа признаков подчеркивает информативность отобранных характеристик и их значимость для распознавания эмоциональных состояний.

Таким образом, данное исследование демонстрирует успешное сбалансированное сочетание между снижением размерности данных и сохранением качества классификации, что является важным вкладом в развитие методов обработки и анализа данных в области распознавания эмоций.

Этот эксперимент позволяет нам понять, какие аспекты речи наиболее информативны для распознавания различных эмоциональных состояний, что имеет важное значение для развития систем распознавания эмоций на основе звука.

7. Заключение

– Краткое подведение итогов и ключевых выводов, основанных на анализе результатов исследования.

– Подчеркивание значимости эффективного отбора признаков для оптимизации методов распознавания эмоций в речи.

– Обсуждение потенциальных направлений для будущих исследований, включая расширение набора данных и оптимизацию методов отбора признаков.

Исследование, проведенное на основе анализа Ryerson Audio-Visual Database of Emotional Speech (RAVDES), позволило получить ценные результаты, подтверждающие эффективность методов распознавания эмоций в речи. Основываясь на анализе данных и результатов экспериментов, мы пришли к выводу, что эффективный отбор признаков играет ключевую роль в оптимизации методов распознавания эмоций в речи. Удалось продемонстрировать, что сохранение высокой классификационной точности при сокращении числа признаков позволяет значительно снизить вычислительную нагрузку и улучшить обобщающую способность классификатора.

Подчеркивается важность дальнейших исследований в этой области. В частности, существует потребность в расширении набора данных, что позволит провести более обширные и всесторонние исследования. Также важным направлением является оптимизация методов отбора признаков с целью улучшения производительности классификации и обеспечения более точного распознавания эмоций в речи.

В целом, результаты данного исследования представляют собой значимый вклад в развитие методов распознавания эмоций в речи и указывают на перспективы дальнейших исследований в этой области.

**Литература**

1. Хлесткин А. Ю. Методы формирования мультимодального изображения //Научные труды SWorld. – 2014. – Т. 4. – №. 4. – С. 14-18.
2. Li S. et al. Pixel-level image fusion: A survey of the state of the art // Information Fusion. – 2017. – vol. 33. – P. 100-112.
3. De Silva D. et al. Wavelet based edge feature enhancement for convolutional neural networks //Eleventh International Conference on Machine Vision (ICMV 2018). – SPIE, 2019. – P. 751-760.
4. Structurally orthogonal finite precision FPGA implementation of block-lifting-based quaternionic paraunitary filter banks for L2L image coding / N.A. Petrovsky, E.V. Rybenkov, A.A. Petrovsky // Digital Signal Processing (DSP’2017): Proc. 22nd Int. Conf., London, UK, Aug. 23-25. – 2017 – 5 P.
5. Qu G., Zhang D., Yan P. Medical image fusion by wavelet transform modulus maxima //Optics Express. – 2001. – Т. 9. – №. 4. – С. 184-190.
6. Дворкович, В.П. Новый подход к использованию вейвлет-фильтров при обработке изображений / В.П. Дворкович, А.В. Гильманшин // Цифровая обработка сигналов – 2008 – № 1 – С. 37–42.
7. Парфенюк М., Петровский А.А. Параунитарные банки фильтров на основе алгебры кватернионов: теория и применение // Цифровая обработка сигналов – 2008 – № 1 – C. 22-36.
8. Qu G., Zhang D., Yan P. Information measure for performance of image fusion //Electronics letters. – 2002. – vol. 38. – №. 7. – P. 313–315.
9. The whole brain atlas [Электронный ресурс]. – 2023. – Режим доступа: <http://www.med.harvard.edu/aanlib/home.html>

FUSION OF MEDICAL IMAGES BASED ON THE   
DISCRETE WAVELET TRANSFORM

Petrovsky N.A., Kuis I.S. Vashkevich M.I.

Belarusian State University of Informatics and Radioelectronics

6, P. Brovki str., Computer Engineering Department, 220113, Minsk, Belarus,  
e-mail: {nick.petrovsky, vashkevich[}@bsuir.by](mailto:}@bsuir.by)

The report shows various techniques for fusion of medical images in the domain of discrete wavelet transform (DWT) coefficients obtained both from Haar functions and based on an integer polyphase representation of the filter bank. A fusion method is also proposed that uses the representation of the low-frequency components of the DWT in the domain of the wavelet modulus maximum. A pairwise comparison of methods based on the metric of mutual information with the use of statistical tests was carried out. The results showed that the use of a polyphase filter bank decomposition, as well as the representation of the low-frequency DWT components in the wavelet modulus maximum domain, allows to obtain better fusion results than using a two-dimensional DWT with Haar functions.