МЕТОД ПОНИЖЕНИЯ РАЗМЕРНОСТИ ПРОСТРАНСТВА ПРИЗНАКОВ НА ОСНОВЕ LASSO-РЕГРЕССИИ ДЛЯ ЗАДАЧИ 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: daniil.krasnoproshin@gmail.com, vashkevich@bsuir.by

В работе предложен метод понижения размерности пространства признаков, основанный на применении LASSO-регрессии, для повышения эффективности распознавания эмоций по речи. Предложенный метод позволяет автоматически отбирать наиболее значимые признаки для классификации эмоций, что способствует повышению точности и скорости распознавания. Предложенный метод может применятся с различными моделями классификации, в частности исследовалось применение метода с классификаторами на основе линейного дискриминантного анализа (ЛДА) и машин опорных векторов (МОВ). Предложенный метод тестировался с использованием набора данных RAVDESS. Показано, что в случае использования МОВ метод позволяет повысить метрику невзвешенной средней полноты с 45,6 до 47,3%, а в случае использования ЛДА с 46,0 до 48,4%. При этом число признаков сокращается приблизительно на 30%.

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 [1] и рекуррентных нейронных сетей [2], что позволяет более эффективно улавливать временные и частотные закономерности в речевых сигналах.

Тем не менее, нейросетевые подходы имеют недостатки, которые могут ограничивать их применимость. Сюда можно отнести высокую вычислительную сложность, а также необходимость обучения глубоких моделей на больших объемах данных. Кроме того, нейронные сети часто характеризуются низкой интерпретируемостью, что затрудняет понимание причинно-следственных связей между входными данными и прогнозами модели. Это усложняет анализ результатов и его понимание, что может быть нежелательным в некоторых приложениях, особенно связанных с медициной.

Таким образом, исследования с использованием более простых подходов сохраняют свою актуальность. Во-первых, они обладают высокой вычислительной эффективностью, что позволяет проводить анализ данных на обычных компьютерах. 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) или линейного дискриминантного анализа (ЛДА), 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. Разработка системы распознавания эмоций

На рис. 1 представлен процесс разработки системы распознавания эмоций. Согласно схеме на рис. 1 процесс разработки основан на использовании аннотированной речевой базы, в которой содержатся образцы речевых сигналов с указанием эмоций с которой они произнесены. Вначале выполняется предварительная обработка аудиосигналов, которая включает вычисление мел-частотных кепстральных коэффициентов (МЧКК), а также их первой и второй производных [1-2], а также ряда статистик, таких как межквартильный размах, коэффициент ассиметрии и эксцесс.

На следующем этапе выполняется отбор признаков, которые способствуют повышению производительности системы распознавания эмоций. Данный этап выполняется с учетом классификатора, который будет использоваться в дальнейшем в системе распознавания эмоций.

Рисунок 1 – Процесс разработки системы распознавания эмоций по речи

Отбор признаков позволяет снизить размерность признакового пространства и улучшить обобщающую способность модели, сохраняя при этом высокую точность классификации или даже повышая её. На заключительном этапе производится оценка производительности системы распознавания эмоций с использованием показателя UAR (англ. *unweighted average recall*), вычисляемом при помощи процедуры перекрестной проверки по 5 блокам [4].

2. Процедура отбора признаков

Для отбора признаков разработан ряд методов, таких как LASSO [5], Relief [4], mRMR и др. [6]. Однако, большинство из них разработано в контексте бинарной классификации. В данной работе разработан метод отбора признаков для задачи многоклассовой классификации, основанный на использовании LASSO-регрессии. Метод LASSO (англ. *least absolute shrinkage and selection operator*) представляет собой технику, позволяющую сократить число предикторов в задаче линейной регрессии [5]. В предыдущих работах [7, 8] LASSO показал свою эффективность в отборе признаков, применительно к задачам бинарной классификации речевых сигналов. В данной работе расширяется применимость метода LASSO для отбора признаков в задаче многоклассовой классификации. Предлагается следующий алгоритм отбора признаков:

1) Задача классификации на K-классов $\{C_1, C_2, \dots, C_K\}$ заменяется на K задач бинарной классификации, по схеме «один против всех». Это значит, что в начале все объекты из C_1 заносятся в один класс и им присваивается метка “1”, а все остальные объекты $\{C_2, \dots, C_K\}$ заносятся в другой класс и им присваивается метка “–1”. Так получается формулируется первая задача бинарной классификации. Затем процедура повторяется, только в отдельный класс помещаются объекты из C_2 и т.д. Применительно к рассматриваемой в работе задаче данный этап означает, что будут получены классификаторы, которые хорошо отделяют одну эмоцию (например, «удивление») от всех остальных.

2) Для решения задач бинарной классификации применяется метод отбора признаков на основе LASSO [8]. В результате получается K подмножеств признаков из их исходного полного набора.

3) Для каждого из K наборов признаков, полученных на шаге 2, применяется метод пошагового исключения переменных (англ. *BSS – backward-stepwise selection*) [5]. Этот этап позволяет выбрать оптимальное подмножество наиболее значимых признаков, снижая размерность данных и устраняя избыточность.

4) Полученные на шаге 3 K наборов признаков объединяются для формирования итогового набора признаков.

Предложенный алгоритм отбора признаков может помочь не только обеспечивает точность в решении многоклассовой задачи, но и поспособствовать повышению интерпретируемости модели, делая ее результаты более понятными.

3. Применение LASSO для отбора признаков в задаче бинарной классификации.

В данном разделе более подробно рассмотрен второй этап предлагаемой процедуры отбора признаков. Метод LASSO [5], используемый в данной работе, основан на решении задачи линейной регрессии:

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \quad (1)$$

где N – число примеров в обучающем наборе, y_i – метка класса i-го образца, x_{ij} – j-й признак i-го образца, β_j – j-й коэффициент линейной модели, λ – параметр регуляризации. Решение (1) при достаточно большом λ приводит к тому, что часть коэффициентов (предикторов) β становятся в точности нулевыми. Поэтому решая (1) для ряда возрастающих значений параметра регуляризации λ и фиксируя порядок, в котором модель «отбрасывает» признаки можно ранжировать их по значимости (первыми отбрасываются наименее значимые признаки).

На рис. 2 показано, как в процессе увеличения параметра регуляризации λ уменьшаются веса предикторов β_j. В приведенном случае $y_i = 1$ для векторов признаков, соответствующих речевым сигналам, содержащим эмоцию «счастье», и $y_i = -1$ для всех остальных эмоций.

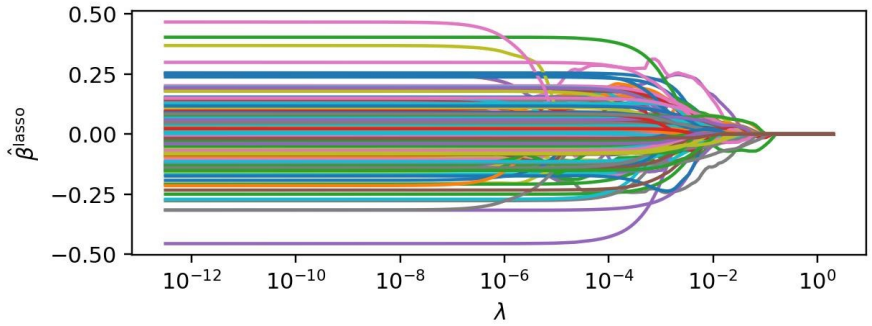

Рисунок 2 – Иллюстрация LASSO-регрессии

После того, как признаки ранжированы по значимости, выполнялась оценка полноты классификации при использовании возрастающего набора признаков. Т.е. требовалось узнать: какая будет точность классификации если использовать только один первый (самый важный) признак, первые два признака, первые три признака и т.д. После этого определялось оптимальное число признаков, которое обеспечивало наибольшую точность классификации. На рис. 3 показано, что для эмоции «счастье» максимальное значение точности (UAR) достигается при использовании первых 120 признаков из отобранных на предыдущем этапе.

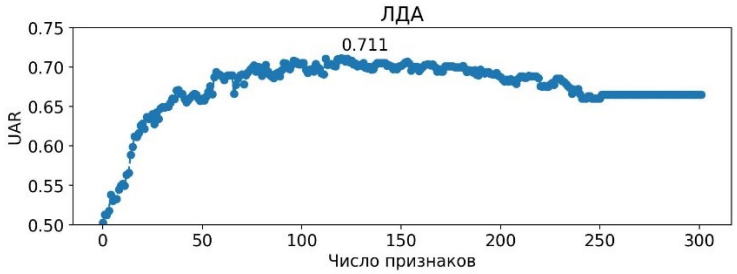

Рисунок 3 – Отбор подмножества признаков по результатам ранжирования с использованием LASSO-регрессии

Далее к полученному подмножеству признаков применялся метод пошагового исключения переменных [5].

4. Набор данных и извлечение признаков

Для проведения нашего исследования мы использовали набор данных Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [3]. который содержит записи от 24 актеров (12 мужчин и 12 женщин), включая 104 высказывания на каждого актера (60 речевых и 44 песенных). В контексте нашей работы, мы ограничились использованием речевых высказываний, что составило 1440 аудиофайлов в формате wav (16 бит, 48 кГц. RAVDESS содержит эмоциональные состояния: 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.

В данной работе речевые признаки рассчитывались на основании мел-частотных кепстральных коэффициентов (МЧКК) [2]. Расчет МЧКК относится к методам кратковременного анализа речевого сигнала, которые предполагают разбиение сигнала на фреймы (короткие сегменты).

В итоговый набор *исходных признаков* были включены среднее значение МЧКК (34 признака), среднеквадратичное отклонение МЧКК (34 признака), среднее от первой и второй производных от МЧКК (68 признаков), их среднеквадратическое отклонение (68), а также коэффициент асимметрии, эксцесс и межквантильный размах (по 34 признака для каждой характеристики соответственно). Таким образом, для каждого аудиофайла мы получаем 306-компонентный вектор надсегментных признаков МЧКК.

Для тестирования классификатора использовался метод перекрестной проверки по *k*-блокам (*k-fold cross-validation*)

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 [3]. 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.

6. Результаты экспериментов

Предложенный метод может применятся с различными моделями классификации, в частности исследовалось применение метода с классификаторами на основе линейного дискриминантного анализа (ЛДА) и машин опорных векторов (МОВ). Оценка производительности системы распознавания эмоций выполнялась с использованием невзвешенной средней полноты – UAR (*Unweighted Average Recall*). Результаты эксперимента приведены в таблице 1.

Таблица 1. Результирующий UAR для классификаторов на основе ЛДА и МОВ

Классификатор	Полный набора признаков (306)	Отобранный набор признаков
ЛДА	0,460	**0,484** (205 признаков)
МОВ (линейное ядро)	0.456	**0.473** (208 признаков)

После процедуры отбора признаков из изначального набора, состоящего из 306 характеристик, оставлено 205 признаков при повышении точности классификатора в случае с ЛДА и 208 признаков в случае с МОВ. 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, более высокая точности классификации 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.

На рис. 5 представлена матрица спутанности для лучшей модели. Анализ матрицы спутанности позволяет выявить важные закономерности в распознавании эмоций. Можно заметить, что наиболее часто неправильно классифицированной эмоцией является грусть (39%) и счастье (41%). Интересно, что «нейтральность» часто путается с «грустью» и «спокойствием», что позволяет предположить некоторое сходство их акустических характеристик. И наоборот, «гнев» имеет высокую точность распознавания (57%) и редко ошибочно классифицировалось как другая эмоция (за исключением отвращения), что указывает на отличительные особенности его акустического профиля. Эти результаты проливают свет на проблемы, с которыми сталкивается классификатор при различении тонких эмоциональных нюансов, и подчеркивают важность разработки функций и совершенствования моделей для улучшения эффективности распознавания эмоций.

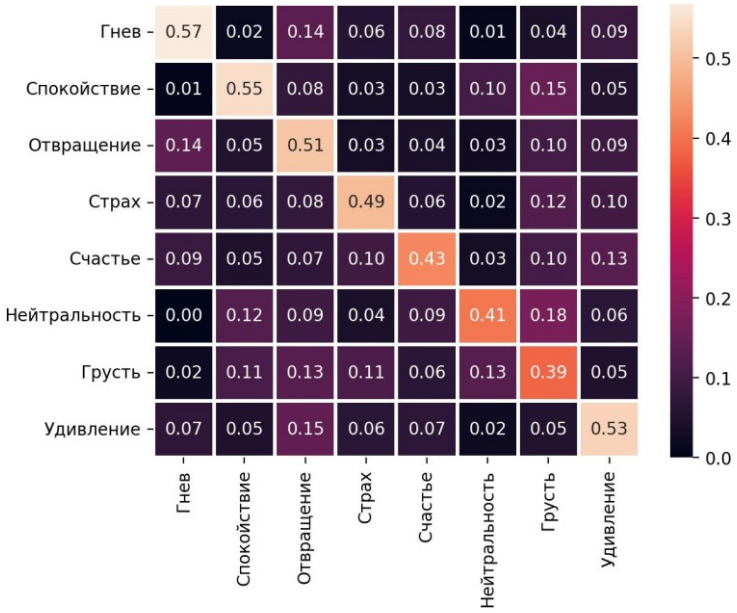

Рис. 5. Матрица спутывания (ЛДА)

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.

7. Заключение

Исследование, проведенное на основе анализа речевого набора 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. Issa D., M. Demirci F., Yazici A. Speech Emotion Recognition with Deep Convolutional Neural Networks. Biomedical Signal Processing and Control, vol. 59, 2020.
2. On C. K., Pandiyan P. M., Yaacob S., and Saudi A. Mel-Frequency Cepstral Coefficient Analysis in Speech Recognition. 2006 International Conference on Computing & Informatics, 2006, pp. 1–5.
3. Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning/ C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J.M. Montero, F. Fernández-Martínez // Sensors. – 2021. – vol. 21. – pp. 1 – 29.
4. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлевают знания из данных / пер. с англ. А.А. Слинкина. М.: ДМК Пресс, 2015. 400 с.
5. Джеймс Г. и др. Введение в статистическое обучение с примерами на языке R / Г. Джеймс, Д. Уиттон, Т. Хасти, Р. Тибширани //М.: ДМК Пресс, 2016. – 450 с.
6. Huang S. H. Supervised feature selection: A tutorial //Artif. Intell. Res. – 2015. – Т. 4. – №. 2. – С. 22-37.
7. Tsanas A. et al. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease //IEEE transactions on biomedical engineering. – 2012. – Т. 59. – №. 5. – P. 1264-1271.
8. Лихачёв Д. С. и др. Комбинированный метод отбора информативных признаков для выявления речевых патологий по голосу //Доклады БГУИР. – 2023. – Т. 21. – №. 4. – С. 110-117.

METHOD FOR REDUCING THE DIMENSIONALITY OF THE FEATURE SPACE BASED ON LASSO REGRESSION FOR SPEECH EMOTION RECOGNITION

Krasnoproshin D.V.,Vashkevich M.I.

Belarusian State University of Informatics and Radioelectronics
6, P. Brovki str., Computer Engineering Department, 220113, Minsk, Belarus,
e-mail: daniil.krasnoproshin@gmail.com, vashkevich@bsuir.by

A method for reducing the dimension of the feature space, based on the use of LASSO regression, to increase the effectiveness of speech emotion recognition is proposed. The proposed method automatically selects the most significant features for the classification of emotions, that increases the accuracy and speed of recognition. The proposed method can be used with various classification models, in particular, we test the method using classifiers based on linear discriminant analysis (LDA) and support vector machines (SVM). The proposed method was tested using the RAVDESS dataset. It is shown that in the case of SVM, the method allows to increase the metric of unweighted average recall (UAR) from 45.6 to 47.3%, and in the case of using LDA from 46.0 to 48.4%. At the same time, the number of features is reduced by about 30%.