**Исследование возможностей классификации эмоций в невербальном речевом поведении на различных наборах данных**

**Аннотация**

**Ключевые слова:** эмоциональные вычисления, распознавание эмоций, визуализация многомерных данных

**1. Введение**

Эмоции играют одну из ключевых ролей в человеческих отношениях, мотивациях и поступках. Способность распознавания эмоций во взаимодействии человек-компьютер – важный шаг в сторону улучшения цифрового окружения. Область знаний, отвечающая за изучение и разработку систем и устройств, которые могут распознавать, интерпретировать и моделировать человеческие эмоции и состояния называется Эмоциональные Вычисления и является одной из активно развивающихся междисциплинарных областей, охватывающей информатику, психологию и когнитивистику.

В настоящий момент не существует единого подхода к определению и измерению эмоции. В когнитивистике существуют модели, позволяющие произвести некоторую формализацию эмоциональных данных[1]. Две наиболее распространенные модели – дискретная, предполагающая существование нескольких «базовых» эмоциональных состояний[2], и непрерывная, в которой эмоция рассматривается в качестве «базиса» в некоем эмоциональном пространстве[3]. Тем не менее, отсутствие единого подхода делает задачу распознавания эмоций трудно формализуемой. С такими задачами лучше других справляются алгоритмы машинного обучения. В случае дискретной модели эмоций задача сводится к классификации, в случае непрерывной модели – к регрессии. В данной работе рассматривается дискретная модель и, соответственно, задача классификации.

Задача распознавания эмоций нацелена на разработку методов точной классификации эмоциональных состояний человека с использованием одной или нескольких модальностей: визуальную (мимика, жестикуляция, поза) и слуховую (речь). Речевой канал, в свою очередь, содержит в себе как семантическую, так и акустическую информацию. Анализу роли последней в задаче классификации эмоций и посвящено предлагаемое исследование.

Однако, далеко не вся информация является классифицируемой, поэтому первым шагом в данном исследовании является визуализация имеющихся данных с целью оценить их распределение [Нужно что-то наплести про необходимость визуализации многомерных данных и тд, но в голову вообще ничего не лезет, не знаю как это более грамотно приплести сюда, первое предложение этого абзаца звучит очень притянуто за уши]. Таким образом, настоящее исследование ставит перед собой следующие задачи: попытка классификации эмоциональной окраски речи по акустическим параметрам звукового сигнала, визуализация распределения данных в многомерном признаковом пространстве.

Статья организована следующим образом: обзор литературы и методов решения задачи приведен в разделе 2. В разделе 3, мы описываем выбранные нами подходы к визуализации и классификации. В разделе 4 представлены описания и результаты экспериментов, их интерпретация и обсуждение – в разделе 5. В разделе 6 мы подводим итоги работы и намечаем вектор дальнейшего исследования.

**2. Обзор литературы**

Решение любой задачи машинного обучения подразумевает определение некоторого набора признаков, извлекаемого из объектов классификации. В данном случае объектом классификации является звуковая запись человеческой речи. Наиболее распространенным является эвристический подход: извлечение множества параметров из акустического сигнала с последующей конкатенацией их в признаковый вектор. Наиболее часто авторы используют просодические характеристики (тон, энергия, амплитуда) [4], однако одних лишь просодических признаков зачастую бывает недостаточно. К более сложным признакам относятся частоты основных формант [5], мел-кепстральные коэффициенты (MFCC) [6], коэффициенты линейного предсказания (LPC)[7]. Одним из state-of-the-art методов является построение спектрограмм, что сводит задачу к классификации изображений [8]. Распространенной практикой является применение статистических функций (среднее, минимум, максимум) и оценка значений производных.

Результатом работы авторов [9][10] стал инструмент OpenSMILE, позволяющий извлекать широкий спектр параметров звукового сигнала, а также применять различные функции к этим параметрам. Благодаря этому, появились наборы параметров, которые могут быть легко извлечены с помощью этого инструмента. К этим наборам относятся как очень большие, как, например, стандартные наборы признаков для конференций INTERSPEECH, содержащие более 5000 элементов [11], так и попытки унифицировать признаковое пространство для задач аффективных вычислений [12].

Выбор классификатора – также важный этап в решении задачи. Среди исследований наиболее популярны такие классификаторы, как машины опорных векторов [13], алгоритм k ближайших соседей [14], классификация с использованием скрытых марковских моделей [15] или нейронных сетей [16].

Наиболее прогрессивными и эффективными на данный момент являются так называемые end-to-end подходы, которые работают непосредственно с дискретизированным аудиосигналом в формате WAV и в которых предобработка, извлечение релевантных признаков и классификация объединены в единый «черный ящик». Для таких подходов обычно используются сверточные глубокие нейронные сети[17], трансферное обучение[18], сети долгой краткосрочной памяти (LSTM)[19].

Для визуализации многомерных данных наиболее часто используются алгоритмы снижения размерности, среди которых анализ главных компонент [20], линейный дискриминантный анализ [21], алгоритм t-distributed stochastic neighbor embedding (t-SNE)[22].

**3. Методология**

3.1. Извлекаемые признаки

Нами выбран набор признаков Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), поскольку данный набор - попытка стандартизации параметров, используемых в эмоциональных вычислениях[12]. eGeMAPS содержит 88 параметров звукового сигнала, среди которых: статистические функции от основной частоты и амплитуды, мел-кепстральные коэффициенты (MFCC) 1-4, приближенная оценка количества слов в секунду. Эти параметры наиболее полно отражают основные акустические признаки эмоций. Также положительной особенностью данного набора параметров является независимость размерности итогового признакового вектора от длительности звукового сигнала. Признаки извлекаются с помощью инструмента OpenSMILE.

3.2. Алгоритм классификации

В качестве классификатора был выбран алгоритм k ближайших соседей, который определяет класс объекта путем определения класса большинства объектов из числа k ближайших соседей объекта . Для оценки расстояния между объектами в kNN-алгоритме могут использоваться различные меры расстояния, такие как евклидова мера, манхэттенская мера, косинусная мера и др [23].

3.3. Предобработка данных

Выбранный алгоритм классификации работает с евклидовыми расстояниями между векторами признаков, поэтому требует симметричного и унимодального распределения данных. В связи с этим, все векторы признаков были подвергнуты нормализации по двум методам с целью сравнения влияния на качество классификации: *i)* Нормализация по методу мин-макса: линейное преобразование данных в диапазоне от 0 до 1; *ii)* Нормализация по Z-показателю: масштабирование данных на основе среднего значения и стандартного отклонения.

3.4. Визуализация многомерных данных

Чтобы получить представление о распределении данных в признаковом пространстве, использован алгоритм уменьшения размерности t-SNE. Он является одним из самых распространенных методов визуализации многомерных данных, который позволяет группировать их в пространствах малой размерности и тем самым получать визуализации, по которым можно делать выводы о возможности и целесообразности классификации представленных данных.

**4. Эксперименты**

4.1. Использованные наборы данных

Имеющиеся в нашем распоряжении наборы данных см. в Таблице 1. Здесь и далее приняты следующие сокращения для базовых эмоций: ANG - Anger (гнев); HAP - Happiness (радость); SAD - Sadness (печаль); NEU - Neutral (нейтральная эмоция); DIS - Disgust (отвращение); FEA - Fear (страх); BOR - Boredom (скука); SUR - Surprise (удивление); EXC - Excitement (возбуждение); FRU - Frustration (негодование); CAL - Calm (спокойствие).

|  |  |  |  |
| --- | --- | --- | --- |
| Название | Базовые эмоции | N записей | Язык |
| IEmoCap[24] | ANG, HAP, SAD, NEU, EXC, FRU | 7304 | Англ. |
| CREMA-D[25] | ANG, HAP, SAD, NEU, DIS, FEA | 7442 | Англ. |
| Emo-DB[26] | ANG, HAP, SAD, NEU, DIS, FEA, BOR | 535 | Нем. |
| RAVDESS[27] | ANG, HAP, SAD, NEU, DIS, FEA, SUR, CAL | 1440 | Англ. |
| SAVEE[28] | ANG, HAP, SAD, NEU, DIS, FEA, SUR | 480 | Англ. |
| TESS[29] | ANG, HAP, SAD, NEU, DIS, FEA, SUR | 2798 | Англ. |

Таблица 1. Наборы данных, подлежащие исследованию.

Задача распознавания эмоций по голосу осложнена отсутствием единой методики составления размеченных наборов данных. Находящиеся в открытом и полуоткрытом доступе датасеты значительно отличаются друг от друга качеством записи, количеством дикторов и представленными базовыми дискретными эмоциональными категориями. В связи с этим, приняты решения:

* 1. Для каждого набора данных провести многоклассовую классификацию. Сравнить результаты, сопоставить их с визуализациями.
  2. Для каждого набора данных провести бинарную классификацию негативной эмоциональной окраски по стратегии «один против остальных». Для этого добавить в каждый из наборов данных вторую разметку: NEG (Negative) – Негативная эмоция, REST – остальные эмоциональные состояния. Категории ANG, SAD, FRU, DIS, FEA отнести к негативным эмоциям, HAP, EXC, NEU, BOR, SUR, CAL – к остальным.; [Тут и вылезает моя изначальная формулировка: негативные – НЕнегативные. Поскольку категории Нейтральная, скука, спокойствие не являются ни негативными, ни позитивными. Относятся скорее к нейтральным. Я пробовал и такую, трехклассовую разметку. Но в таком случае, во первых, получается явный дизбаланс распределения сэмплов по категориям: примерно 2:1:1 негативная:нейтральная:позитивная, во-вторых, результаты у классификации получаются еще более плачевные чем для бинарной разметки: нейтральная категория практически не распознаваема, путается и с негативными, и с позитивными. Поэтому и было принято решение категории нейтральная и позитивная объединить в категорию НЕнегативная. В общем, нужно как-то вылезать, а как – не соображу. UPD: вроде, сообразил, см. в исправлениях в режиме рецензирования.] Проанализировать результаты бинарной классификации, сопоставить с визуализациями.
  3. Путем слияния всех наборов данных на английском языке, получить обобщенный набор данных English Assembly, содержащий 19462 образца и произвести попытку как мультиклассовой, так и бинарной классификации. При этом, для мультиклассовой классификации принято решение ограничиться образцами, помеченными шестью самыми распространенными базовыми эмоциями: ANG, HAP, DIS, FEA, NEU, SAD. В случае же бинарной классификации, используются все образцы.

4.2. Визуализация распределения данных в признаковом пространстве.

Мы используем библиотеку ScikitLearn [30], в которой имеется имплементация алгоритма t-SNE. Чтобы не перегружать визуализации и не усложнять работу алгоритма, количество подаваемых на вход алгоритма образцов данных было сокращено до случайной выборки, состоящей из 1000 образцов с сохранением пропорций распределения эмоциональных классов. Использованные нами параметры подобраны исходя из соображений, описанных в [31]: n\_components=2, n\_iter=5000; learning\_rate=0.1; early\_exaggeration=2. Параметр perplexity варьировался от 5 до 75, таким образом, было получено по 70 визуализаций для каждого набора данных, из которых было выбрано по одной наиболее удачной. Эти визуализации представлены на рисунках.

4.3. Классификация.

Для классификации также используется библиотека ScikitLearn, имеющая в своем составе алгоритм KNeighborsClassifier. Для экспериментов нами были использованы параметры этого алгоритма по умолчанию, кроме отвечающего за количество ближайших соседей k. Этот параметр варьировался от 2 до 75. Данные были предварительно разбиты на обучающую и тестовую выборку в пропорции 7:3, с сохранением пропорций распределения эмоциональных классов. В качестве метрики качества классификации выбрана β - средняя точность предсказания на тестовой выборке[Каким символом или сокращением обозвать?]. На рисунках изображены зависимости β от k для данных без предобработки, с мин-макс нормализацией и с Z-нормализацией. Лучшие результаты мультиклассовой и бинарной классификации представлены в Таблице 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Набор данных | Мультиклассовая | | Бинарная | |
| k | β | k | β |
| IEmoCap | 69 | 0,45 | 25 | 0,66 |
| Crema-D | 55 | 0,49 | 39 | 0,75 |
| Emo-DB | 13 | 0,72 | 27 | 0,86 |
| RAVDESS | 5 | 0,58 | 7 | 0,78 |
| SAVEE | 7 | 0,62 | 11 | 0,68 |
| TESS | 20 | 0,98 | 9 | 0,99 |
| English Assembly | 17 | 0,60 | 13 | 0,73 |

Таблица 2. Результаты классификации.

Также для облегчения интерпретации результата классификации и визуализаций распределения, на каждом наборе данных были построены матрицы ошибок лучшего классификатора. Они представлены на рисунках. [Как оформлять рисунки?]

**5. Интерпретация и обсуждение результатов.**

По результатам классификации видно, что в подавляющем большинстве случаев, у алгоритма k ближайших соседей в случае мультиклассовой классификации не возникает проблем с выявлением гнева, также достаточно уверенно распознается печаль и нейтральная эмоциональная окраска. Это видно также на визуализациях t-SNE, где образцы этих эмоций занимают строго определенную область пространства. В случае бинарной классификации «Положительная эмоциональная окраска» - «Отрицательная эмоциональная окраска» мы можем наблюдать явный перевес качества классификации в сторону определения истинно «негативного» класса и большой ложноположительный показатель, что свидетельствует о некоторой степени переобученности модели, устранив которую, можно получить лучшие результаты.

На наборе данных TESS получены выделяющиеся результаты как классификации, так и визуализации, однако, это связано с особенностями этого набора данных – чрезмерное модулирование актерами эмоций, произнесение одной и той же фразы, всего два актера. В общем случае, лучшие результаты получены для тех наборов данных, в которых: а) Разделение и модуляция эмоциональных переживаний сильнее выражена (более яркая игра актеров); б) Произносятся одинаковые и короткие фразы; в) Меньшее количество актеров.

На некоторых наборах данных (SAVEE, IEmoCap) T-SNE выявляет кластеры по отличному от эмоциональной окраски признаку. Это могут быть кластеры для говорящих разного пола, или для разных актеров. Применение разделения по полам или использование идентификации говорящего может улучшить качество классификации.

[Здесь что-то нужно еще написать об успешности-неуспешности объединения датасетов в один большой. Результаты то в принципе норм, но какие выводы сделать из этого, не лезет в голову.]

**6. Выводы и дальнейшие планы.**

По итогу работы можно сделать вывод, что возможность и результаты классификации эмоций по речи методом k ближайших соседей во многом зависят от особенностей обучающего набора данных, а также от метода предобработки. В наши дальнейшие планы входит расширение списка используемых наборов данных (в том числе, на русском языке), методов классификации и предобработки. Также планируется применять рассмотренные методы в задачах мультимодального распознавания эмоций.

**7. Литература.**

[1]Cornelius, R.R., 1996. The Science of Emotion. Research and Tradition in the Psychology of Emotion. Prentice-Hall, Upper Saddle River, NJ

[2]Ortony, Andrew & Turner, Terence. (1990). What's Basic About Basic Emotions?. Psychological review. 97. 315-31. 10.1037/0033-295X.97.3.315.

[3]Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161–1178. <https://doi.org/10.1037/h0077714>

[4]Y. Yeqing and T. Tao, “An new speech recognition method based on prosodic analysis and svm in zhuang language,” Proc. 2011 International Conference on Mechatronic Science, Electric Engineering and Computer, 2011

[5]Khulage, A.A.. (2012). Extraction of pitch, duration and formant frequencies for emotion recognition system. IET Conference Publications. 2012. 7-9. 10.1049/cp.2012.2482.

[6]M. S. Likitha, S. R. R. Gupta, K. Hasitha and A. U. Raju, "Speech based human emotion recognition using MFCC," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, 2017, pp. 2257-2260, doi: 10.1109/WiSPNET.2017.8300161.

[7] A. Razak, R. Komiya, and M. Abidin, “Comparison between fuzzy and nn method for speech emotion recognition.” Proc. Information Technology and Applications, ICITA 2005.

[8] Niu, Yafeng & Zou, Dongsheng & Niu, Yadong & He, Zhongshi & Tan, Hua. (2017). A breakthrough in Speech emotion recognition using Deep Retinal Convolution Neural Networks.

[9] Eyben, Florian & Wöllmer, Martin & Schuller, Björn. (2010). openSMILE -- The Munich Versatile and Fast Open-Source Audio Feature Extractor. MM'10 - Proceedings of the ACM Multimedia 2010 International Conference. 1459-1462. 10.1145/1873951.1874246.

[10] Florian Eyben, Felix Weninger, Florian Gross, Björn Schuller: “Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor”, In Proc. ACM Multimedia (MM), Barcelona, Spain, ACM, ISBN 978-1-4503-2404-5, pp. 835-838, October 2013. doi:10.1145/2502081.2502224

[11] Schuller, Björn & Steidl, Stefan & Batliner, Anton & Vinciarelli, Alessandro & Scherer, Klaus & Ringeval, Fabien & Chetouani, Mohamed & Weninger, Felix & Eyben, Florian & Marchi, Erik & Mortillaro, Marcello & Salamin, Hugues & Polychroniou, Anna & Valente, Fabio & Kim, Samuel. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 148-152.

[12]F. Eyben et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," in IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190-202, 1 April-June 2016, doi: 10.1109/TAFFC.2015.2457417

[13] Chavhan, Yashpalsing & Dhore, Manikrao & Pallavi, Yesaware. (2010). Speech Emotion Recognition Using Support Vector Machines. International Journal of Computer Applications. 1. 10.1007/978-3-642-21402-8\_35.

[14] Meftah I.T., Le Thanh N., Ben Amar C. (2012) Emotion Recognition Using KNN Classification for User Modeling and Sharing of Affect States. In: Huang T., Zeng Z., Li C., Leung C.S. (eds) Neural Information Processing. ICONIP 2012. Lecture Notes in Computer Science, vol 7663. Springer, Berlin, Heidelberg

[15] Nwe, Tin & Foo, S.W. & De Silva, Liyanage. (2003). Speech Emotion Recognition Using Hidden Markov Models. Speech Communication. 41. 603-623. 10.1016/S0167-6393(03)00099-2.

[16] Roy T., Marwala T., Chakraverty S. (2020) Speech Emotion Recognition Using Neural Network and Wavelet Features. In: Chakraverty S., Biswas P. (eds) Recent Trends in Wave Mechanics and Vibrations. Lecture Notes in Mechanical Engineering. Springer, Singapore

[17] Trigeorgis, George & Ringeval, Fabien & Brueckner, Raymond & Marchi, Erik & Nicolaou, Mihalis & Schuller, Björn & Zafeiriou, Stefanos. (2016). Adieu Features? End-to-end Speech Emotion Recognition using a Deep Convolutional Recurrent Network. 10.13140/RG.2.1.3842.7283.

[18] ElShaer, Mohamed & Wisdom, Scott & Mishra, Taniya. (2019). Transfer Learning From Sound Representations For Anger Detection in Speech.

[19] Gideon, John & Khorram, Soheil & Aldeneh, Zakaria & Dimitriadis, Dimitrios & Mower Provost, Emily. (2017). Progressive Neural Networks for Transfer Learning in Emotion Recognition. 1098-1102. 10.21437/Interspeech.2017-1637.

[20] Ivosev, G., Burton, L. and Bonner, R., 2008. Dimensionality reduction and visualization in principal component analysis. *Analytical chemistry*, *80*(13), pp.4933-4944.

[21] Wang, H., Ding, C. and Huang, H., 2010, September. Multi-label linear discriminant analysis. In *European conference on computer vision* (pp. 126-139). Springer, Berlin, Heidelberg.

[22] L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.

[23] Воронцов, К.В., 2011. Математические методы обучения по прецедентам (теория обучения машин). *Москва*, pp.119-121.

[24] Busso, Carlos & Bulut, Murtaza & Lee, Chi-Chun & Kazemzadeh, Abe & Mower Provost, Emily & Kim, Samuel & Chang, Jeannette & Lee, Sungbok & Narayanan, Shrikanth. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation. 42. 335-359. 10.1007/s10579-008-9076-6.

[25] "H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, “CREMA-D: Crowd-sourced emotional multimodal actors dataset,” IEEE Transactions on Affective Computing, 2014."

[26] "F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of German emotional speech,” in 9th European Conference on Speech Communication and Technology (Interspeech’2005 - Eurospeech), Lisbon, Portugal, September 2005, pp. 1517–1520."

[27] S. R. Livingstone и F. A. Russo, «The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English», PLOS ONE, т. 13, вып. 5, с. e0196391, май 2018.

[28] Jackson, Philip & ul haq, Sana. (2011). Surrey Audio-Visual Expressed Emotion (SAVEE) database.

[29] M. K. Pichora-Fuller и K. Dupuis, «Toronto emotional speech set (TESS)». Scholars Portal Dataverse, 2020

[30] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, pp.2825-2830.

[31] Poličar P.G., Stražar M., Zupan B. (2019) Embedding to Reference t-SNE Space Addresses Batch Effects in Single-Cell Classification. In: Kralj Novak P., Šmuc T., Džeroski S. (eds) Discovery Science. DS 2019. Lecture Notes in Computer Science, vol 11828. Springer, Cham