УДК 004.932

АНСАМБЛЕВИЙ КЛАСИФІКАТОР ЕМОЦІЙ НА ОСНОВІ АУДІО-ФАЙЛІВ

Андронаті Олександр Кирилович
к.т.н., доцент каф. ІС Ніколенко Анатолій Олександрович
Національний університет «Одеська політехніка», УКРАЇНА

АНОТАЦІЯ. Розглянуто питання розробки та дослідження ансамблевого класифікатору емоцій на основі аудіофайлів. Доведено, що використання ансамблевого класифікатору для проблеми класифікації емоцій на основі аудіо-файлів призводить до значного підвищення якості класифікації.

Вступ. Дослідження в галузі штучного інтелекту та машинного навчання, пов'язані з автоматизованим розпізнаванням емоцій, спрямовані на отримання інформації про психічний стан людини, що є важливим для формування моделі соціальної взаємодії. Існує безліч інтелектуальних систем і технологій, заснованих на розпізнаванні емоцій мови, як от: системи людино-машинної взаємодії, онлайн-навчання, call-центрів, емоційного маркетингу. При розробці сучасних систем та технологій для класифікації емоційної мови використовуються алгоритми машинного навчання та нейронні мережі, але всі вони мають свої переваги та недоліки [1,2,3]. Отже вибір одного з них пов’язаний з деякими обмеженнями, специфічними для кожного алгоритму. Тому розробка ансамблю класифікаторів аудіо-даних для підвищення точності розпізнавання емоцій мови є актуальним науково-практичним завданням.

Метою роботи є розробка ансамблевого класифікатору для підвищення якості класифікації емоцій на основі аудіо-файлів.

Порівняння з існуючими аналогами. Для побудови класифікаторів аудіо-файлів з метою розпізнавання емоцій мови відомі рішення з використанням нейронних мереж різної архітектури або інших методів машинного навчання. В останні роки запропоновано використовувати наступні алгоритми класифікації: Support Vector Machine (SVM), K Nearest Neighbors (KNN), XGBoost, Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), LSTM, Random Forest [2,3]. Серед зазначених вище класифікаторів всі, окрім CNN та LSTM потребують представлення входу у вигляді вектору спектральних ознак. CNN використовує вхідні данні у вигляді 2-вимірного спектру аудіо-даних. LSTM (довга короткочасна пам'ять) використовується для аналізу аудіо-сигналу як часового ряду. Щодо точності класифікації, то в [2] система розпізнавання емоційної мови з використанням LSTM показує середнє значення точності класифікації 70.3% для розпізнавання 7 емоцій, а в [1] система побудована на MLP показує значення точності класифікації 75.1%, але для розпізнавання тільки 5 емоцій. Отже використання одного алгоритму для задачі розпізнавання емоційної мови не призводить до достатньо якісних результатів, що робить створення ансамблю для цієї задачі актуальним.

Основна частина роботи. Створення ансамблевого класифікатору емоцій на основі аудіо-файлів передбачає виконання наступних кроків:

1. Вибір даних для дослідження.
2. Витяг спектральних характеристик з аудіо-файлів та попередня обробка даних.
3. Налаштування гіперпараметрів алгоритмів, які будуть складовими частинами ансамблевого класифікатору, навчання кожного окремого алгоритму.
4. Побудова структури ансамблевого класифікатора та його навчання.
5. Порівняння різних ансамблевих класифікаторів з використанням різних метрик для мультикласової класифікації та вибір найкращого ансамблю.

В якості аудіо-даних обрано датасет RAVDESS, який містить 1440 файлів в форматі wav, кожний є записом голосу 24 професійних акторів. Кожний файл відповідає одній з 7 емоцій, таких як спокій, радість, сум, гнів, наляканість, здивованість та огида. Після проведення відповідних кроків попередньої обробки даних набір було розділено на тренувальну та тестову вибірку у співвідношенні 3/1, тобто 1080 файлів у тренувальній вибірці, 360 – в тестовій.

Після спектрального перетворення для кожного файлу розраховано такі характеристики: значення спектрального центроїду, спектрального спаду, спектральної пологості; 6 оцінок спектральних контрастів; значення частоти переходів через нуль; RMS; значення 20 мел-

частотних кепстральних коефіцієнтів (MFCC). У результаті отримуємо вхідний вектор ознак розмірністю 32 для подальшого використання в KNN, SVM, Random Forest, XGBoost та MLP класифікаторах. Для побудови входу LSTM-класифікатора 12 числових значень MFCC-коефіцієнтів для кожного файлу представляються у вигляді часового ряду розміром 125 відліків. Глибокі CNN-класифікатори на вхід отримують зображення розміром (128×128), які складаються із строк значень аудіо-файлу після спектрального перетворення.

Для розробки моделей окремих класифікаторів за алгоритмами машинного навчання та нейромережевих класифікаторів були налаштовані гіперпараметри за методикою Grid Search Cross Validation (GSCV). На першому кроці GSCV визначається набір значень для кожного з гіперпараметрів. На другому кроці GSCV фактично перебирає всі можливі комбінації значень гіперпараметрів та перевіряє точність класифікації за допомогою кросс-валідації.

Таким чином, в результаті було отримано 7 навчених за визначеними гіперпараметрами класифікаторів, які на тестовому наборі отримали найкраще передбачення для кожного класу (табл. 1). В термінології ансамблевих класифікаторів кожне таке передбачення має назву – vote.

Таблиця 1 – Результати класифікації окремих алгоритмів

Алгоритм	Accuracy	F1-score
KNN	0.572	0.561
SVM	**0.692**	**0.687**
Random Forest	0.597	0.588
XGBoost	0.650	0.646
MLP	0.681	0.678
CNN	0.611	0.606
LSTM	0.522	0.522

Для раціонального вибору архітектури ансамблевого класифікатора було використано два способи отримання фінального передбачення із результатів окремих класифікаторів – hard voting та soft voting. Hard voting передбачає вибір прогнозу з найбільшою кількістю голосів, тоді як soft voting передбачає об’єднання ймовірностей кожного прогнозу в кожній моделі та вибір прогнозу з найвищою загальною ймовірністю. Таким чином з різною комбінацією алгоритмів-складових (від 3 до 7 в складі ансамблю) було побудовано 198 ансамблевих класифікаторів (по 99 з з використанням hard voting та soft voting), які були порівняні між собою з використанням таких метрик, як точність, F-1 міра, AUC.

Серед спроектованих ансамблевих класифікаторів найкращі значення класифікаційних метрик отримано ансамблевим класифікатором з KNN, SVM, CNN, LSTM з агрегацією soft voting. Результати наведено в матриці невідповідності (таблиця 2), де гнів – 0, спокій – 1, огида – 2, наляканість 3, радість – 4, нейтральна – 5, сум – 6, здивованість – 7.

Таблиця 2 – Матриця невідповідностей для ансамблю класифікаторів з KNN, SVM, CNN, LSTM з агрегуванням soft voting

		Передбачене значення							
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
справжнє значення	**0**	35	0	7	1	0	0	0	5
	1	0	44	1	0	0	2	1	0
	2	2	0	43	0	1	0	2	0
	3	0	1	6	32	1	0	6	2
	4	3	1	2	2	34	1	1	4
	5	0	8	0	0	1	14	1	0
	6	0	1	3	1	2	3	38	0
	7	0	2	4	3	1	0	0	38

З матриці невідповідності були отримані значення accuracy = 0.772 та F1 = 0.771. Отриманий результат пояснюється тим, що до ансамблю включені окремі класифікатори, які

відрізняються за алгоритмами навчання та представленням вхідних аудіо-даних, крім того агрегація soft voting є в цілому більш гнучкою, адже вона враховує ступінь впевненості рішення окремих класифікаторів, тобто значення ймовірності.

Наприкінці наведемо порівняння результатів розпізнавання аудіо-даних емоцій із тестової вибірки датасета RAVDESS кращого ансамблю з результатами окремих класифікаторів у його складі (таблиця 3). За результатами можна побачити, що точність ансамблевого класифікатора за метрикою Accuracy на 8% вище за точність кращого класифікатора у його складі (SVM), а за метрикою F1 таке підвищення дорівнює 8.4%.

Таблиця 3 – Порівняння результатів розпізнавання емоцій аудіо-даних

Назва	Значення	
	Accuracy	F1-score
Ансамбль	**0.772**	**0.771**
	Значення/ Різниця в метриках в %	
KNN	0.572/ **+20%**	0.561/ **+21%**
SVM	0.692/ **+8%**	0.687/ +**8.4%**
CNN	0.611/ +**16.1%**	0.606/ **+16.5%**
LSTM	0.522/ **+25%**	0.522/ **+24.9%**

Висновки. В роботі досліджується та обґрунтовується можливість підвищення точності розпізнавання емоцій людини за рахунок ансамблевої класифікації аудіо-даних. В якості набору даних обрано датасет RAVDESS. В якості алгоритмів-складових ансамблю побудовано 4 класифікатора за популярними алгоритмами машинного навчання KNN, SVM, Random Forest, XGBoost, та 3 нейромережеві моделі MLP, CNN та LSTM. Досліджено якість класифікації окремих алгоритмів. Показано, що SVM-класифікатор показує найкращу точність за метриками Accuracy = 0.692 та F1 = 0.687.

З використанням агрегацій hard voting та soft voting побудовано 198 ансамблевих класифікаторів з кількістю алгоритмів складових від 3 до 7. За проведеним дослідженням найкращі значення класифікаційних метрик отримано ансамблем класифікаторів KNN, SVM, CNN, LSTM з агрегацією soft voting. Показано що точність кращого ансамблевого класифікатора вище точності кращого алгоритму у його складі (SVM) за метрикою Accuracy на 8% а за метрикою F1 на 8.4%.

Отже використання ансамблевих класифікаторів для розпізнавання емоцій людини із аудіо-файлів є перспективним напрямом дослідження.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1.Petrosiuk, D. V., Arsirii, O. O., Babilunha, O. Ju. & Nikolenko, A. O. “Deep Learning Technology of Convolutional Neural Networks for Facial Expression Recognition”. Applied Aspects of Information Technology. Publ. Nauka i Tekhnika. Odessa: Ukraine. 2021; Vol.4 No.2: 192–201. DOI: https://doi.org/10.15276/aait.02.2021.6

2.L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. Ali Mahjoub, and C. Cleder, ‘Automatic Speech Emotion Recognition Using Machine Learning’, Social Media and Machine Learning. IntechOpen, Feb. 19, 2020. doi: 10.5772/intechopen.84856.

3.R.D.G. Ayon, M.S. Rabbi, U. Habiba, M. Hasana; Bangla Speech Emotion Detection using Machine Learning Ensemble Methods;, Advances in Science, Technology and Engineering Systems Journal, vol. 7, no. 6, pp. 70-76 (2022). https://dx.doi.org/10.25046/aj070608

ENSEMBLE CLASSIFIER OF EMOTIONS BASED ON AUDIO FILES

Oleksandr Andronati
Ph.D., associate professor, Department IS Anatolii Nikolenko
Odesa Polytechnic National University, UKRAINE

ANNOTATION. Development and research of ensemble classifier of emotions based on audio files is considered. It is proved that the use of ensemble classifier for the problem of classification of emotions based on audio files leads to a significant increase in the quality of classification.