DOI: https://doi.org/10.15276/ict.01.2024.23
УДК 004.81; 167.7

Метод побудови ансамблевих класифікаторів для розпізнавання аудіо-даних різної природи

Андронаті Олександр Кирилович[1)]
Аспірант каф. Інформаційних систем;
ORCID: https://orcid.org/orcid.org/ 0009-0009-1794-5864; alex.andronati@gmail.com.
Scopus Author ID: 58677655800
Арсірій Олена Олександрівна[1)]
Д-р техніч. наук, професор, завідувачка каф. Інформаційних систем
ORCID: https://orcid.org/orcid.org/0000−0001−8130−9613; e.arsiriy@gmail.com. Scopus Author ID: 54419480900
Ніколенко Анатолій Олександрович[1)]
Канд. техніч. наук, доцент каф. Інформаційних систем
ORCID: https://orcid.org/0000-0002-9849-1797; anatolyn@ukr.net. Scopus Author ID: 57197491335
Кундієв Олег Ігорович[1)]
Аспірант каф. Інформаційних систем
ORCID: https://orcid.org/orcid.org/ 0009-0001-4499-1630; oleh.kundiiev@pm.me
[1)] Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

АНОТАЦІЯ

У роботі розроблено метод побудови ансамблевих класифікаторів для розпізнавання аудіо-даних різної природи. Метод є дата орієнтовним та потребує виконання наступних кроків. На першому кроці обираються вхідні датасети, які трансформуються та розподіляються на навчальні та тестові вибірки відповідно. Для задачі розпізнавання аудіо-емоцій обрано набори даних RAVDESS, розпізнавання жанру музики виконується на датасеті GTZAN. На другому кроці як елементи ансамблевих класифікаторів створено та досліджено сім наступних класифікаторів: K Nearest Neighbors, Support Vector Machine, Random Forest, XGBoost, Multilayer Perceptron, Convolutional Neural Network і Long Short-term Memory. В процесі навчання елементарних класифікаторів налаштування відпопідних гіперпараметрів виконувалось за допомогою підходу Grid Search. На третьому кроці елементарні класифікатори об'єднувались з використанням методу ансамблювання stacking з такими видами агрегування, як soft voting, hard voting, soft voting з використанням функції GOMPERTZ. Було перевірено всі можливі ансамблеві комбінації починаючі із трьох елементарних класифікаторів для розпізнавання аудіо-емоцій та музичних жанрів. Тому загальна кількість ансамблів, які досліджувались в роботі становила 297. Результати дослідження для проблеми класифікації аудіо-емоцій показали, що точність розпізнавання за метрикою Accuracy кращого ансамблевого класифікатора на 8.1% вища, ніж у кращого із елементарних класифікаторів у його складі, якій базується на MLP, а за метрикою F1 цей показник є вищим на 8 %. Для задачі розпізнавання музичних жанрів відповідні показники вище на 5.6 % та на 5.2 % відповідно.

Ключові слова: ансамблеві класифікатори; налаштування гіперпараметрів машинне навчання; глибоке навчання; аудіо-дані

Актуальність. Розпізнавання аудіо-даних в інтелектуальних системах – це процес аналізу аудіо-сигналів для розуміння їх змісту або характеристик. Це важливий компонент інтелектуальних систем, які можуть працювати з голосовими командами, розуміти мову або аналізувати аудіо-дані. Наприклад, у системах розпізнавання мови аудіо-дані використовуються для мовного перекладу, створення текстових транскрипцій, у системах біометричної ідентифікації аудіо-сигнали використовуються для аутентифікації користувача, у медичних діагностичних системах розпізнавання аудіо-сигналів є компонентом виявлення деяких серцевих захворювань і такого стану, як апное (зупинка дихання). У цьому дослідженні автори наводять приклади розпізнавання аудіо-даних для визначення емоційного стану людини, які базуються на попередніх роботах авторів [1-3] та в системах аналізу аудіо-контенту для визначення ключових характеристик музичних жанрів, які є важливими для музичних платформ і рекомендаційних систем. Побудова систем розпізнавання емоцій і аналізу аудіо-контенту вимагає створення та використання складних обчислювальних методів через неструктурованість і великий обсяг аудіо-даних. Відомо, що машинне навчання надає ефективні інструменти для роботи з такими даними. Різні алгоритми машинного навчання, такі як нейронні мережі, дерева рішень і метод опорних векторів, вже успішно застосовані в

області аналізу аудіо-даних. Однак їх успіх обмежений складністю звукових сигналів і їхнього контексту, а також проблемою перенавчання [4]. Тому існує потреба в пошуку нових методів, які б забезпечили більшу ефективність і точність аналізу аудіо-даних.

Одним із потенційних рішень є використання ансамблевих класифікаторів, які поєднують декілька моделей для досягнення кращого результату, ніж окремі моделі. Ці ансамблі дозволяють поєднувати різні підходи та уникати обмежень окремих класифікаторів. Такий підхід відкриває нові перспективи у вивченні аудіо-даних і може забезпечити значне підвищення точності та ефективності їх аналізу.

Метою дослідження є підвищення точності розпізнавання аудіо-даних різної природи шляхом розробки методу їх ансамблевої класифікації. В якості прикладів аудіоданих для машинного навчання використовувались класифікації музичних жанрів та аудіо-емоцій.

Для досягнення мети дослідження вирішуються такі завдання:

1) Отримання спектральних характеристик із набору даних RAVDESS для розпізнавання аудіо-емоцій та набору даних GTZAN для розпізнавання музичних жанрів для створення вхідних табличних даних.

2) Навчання окремих класифікаторів, таких як KNN, SVM, Random Forest, XGBoost, MLP, CNN і LSTM з налаштуванням гіперпараметрів.

3) Використання різних типів прогнозної агрегації з різними комбінаціями базових класифікаторів для створення ансамблів.

4) Порівняння продуктивності ансамблів із застосуванням stacking-підходу та різних типів агрегації на двох наборах даних і вибір найкращих ансамблів на основі класифікаційних метрик.

Для побудови систем розпізнавання емоцій на основі машинного навчання використовуються такі відомі набори даних, як RAVDESS [5], CREMAD [6], SAVEE [7], а набір даних GTZAN [8] є одним із найвідоміших і широко використовуваних наборів даних у задачах класифікації музичних жанрів у системах аналізу аудіо-контенту для музичних платформ. Що стосується попередньої обробки даних, різні спектральні характеристики можуть використовуватися для різних завдань, наприклад, така характеристика, як спектральний контраст, може використовуватися для обох завдань, тоді як хроматичні характеристики можуть використовуватися лише для аналізу музичних жанрів.

Для розробки класифікаторів аудіо-даних добре відомі різні рішення, які використовують різні методи машинного та глибокого навчання. Нещодавно запропоновані алгоритми включають K Nearest Neighbors (KNN), Support Vector Machine (SVM) [9, 10], XGBoost, Multilayer Perceptron (MLP), Convolutional Neural Network (CNN) [11,12], Long Short-term Memory (LSTM) [13] і Random Forest [9]. Кожен із вищезазначених алгоритмів має свої унікальні переваги, однак використання їх окремо супроводжується певними обмеженнями, характерними для кожного з них, і призводить до недостатніх показників якості класифікації. Наприклад, у [12] показано, що використання згорткової нейронної мережі для класифікації аудіо-емоцій за допомогою набору даних RAVDESS призвело до значення accuracy = 72.7%. У статті [11] показано, що використання класифікатора KNN для розпізнавання музичних жанрів із використанням набору даних GTZAN дало значення accuracy = 70.9%.

Аналіз публікацій показує, що дослідження в галузі створення та використання ансамблевих класифікаторів зазнали значного прогресу за останні роки. Основні напрямки включають boosting, bagging, stacking та інші. Boosting, наприклад, заснований на побудові послідовного ряду слабких класифікаторів, які разом утворюють сильний класифікатор.

З іншого боку, bagging складається з навчання кількох незалежних класифікаторів на різних підмножинах даних та об’єднання їхніх результатів. Stacking поєднує прогнози кількох базових моделей, використовуючи іншу модель, щоб дізнатися, які комбінації є найефективнішими.

Серед цих методів stacking є найпростішим у реалізації та ефективно використовується для створення ансамблевих класифікаторів із різноманітних складових [12]. Існують різні типи агрегації прогнозів для stacking: soft voting, hard voting, soft voting з використанням нечіткого ранжуванням Гомперца.

У методі hard voting остаточне рішення в ансамблі приймається на основі голосів кожної моделі класифікатора. Всі моделі роблять свої прогнози для кожного випадку, а остаточне рішення визначається більшістю голосів моделей. Hard voting використовується, коли класифікатори мають однакову вагу та можуть забезпечити адекватні прогнози щодо вхідних даних.

На Рис. 1 зображено один із класифікаторів hard voting, розроблений у роботі [1].

У soft voting кожен класифікатор в ансамблі призначає вагу кожному класу на основі ймовірностей. Замість прийняття простого двійкового рішення кожен класифікатор визначає ймовірності для кожного класу, а остаточне рішення визначається шляхом зіставлення ймовірностей з їх вагами. Цей метод дозволяє враховувати впевненість кожної моделі та приймати більш обґрунтовані рішення на основі цієї впевненості.

Soft voting з використанням нечіткого ранжуванням Гомперца є одним із варіантів агрегації для ансамблевого класифікатора, який поєднує в собі переваги м'якого голосування та методу нечіткого ранжирування Гомперца. Цей підхід використовується для врахування невизначеності та неоднорідності даних шляхом запровадження нечіткого ранжирування. Перш за все, soft voting визначає ймовірності для кожного класу для кожного класифікатора в ансамблі. Кожен класифікатор надає свої прогнози, а ймовірності для кожного класу обчислюються як середнє арифметичне ймовірностей, наданих кожним класифікатором. Потім використовується нечіткий метод ранжирування Гомперца для оцінки впевненості кожного класифікатора в його прогнозах і приведення цих значень впевненості в більш послідовну та стандартизовану форму. Метод Гомперца використовує функцію Гомперца для наближення нечіткості та невизначеності в даних, дозволяючи числові значення, які відображають рівень впевненості кожного класифікатора в його прогнозах.

У той же час функція Гомперца є окремим випадком узагальненої сигмоїдальної функції і має вигляд (1):

$$f(t) = a*\exp(-b*\exp(-ct)), \quad (1)$$

де a – верхня асимптота $a*exp(-b*exp(-inf))=a*exp(0)=a$; b, c – додатні числа (параметри росту); b задає зсув по x; c визначає масштабування в x; число Ейлера e = 2.71828...

Тобто функція Гомперца є сигмоїдальною функцією, яка описує зростання як найповільніше на початку та в кінці певного періоду часу. Крива наближається до правої або майбутньої асимптоти функції набагато повільніше, ніж до лівої або нижньої асимптоти. Функцію Гомперца можна використовувати для агрегування прогнозів від різних
класифікаторів завдяки її здатності моделювати зростання чи спад значень з часом. Це може бути особливо корисним, коли необхідно врахувати різну важливість або вагу прогнозів залежно від їх надійності чи ефективності.

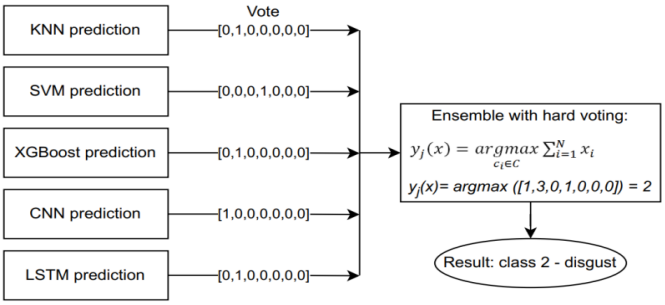

Рис. 1. **Hard voting ансамбль**

Так, soft voting з використанням нечіткого ранжування Гомперца дозволяє врахувати невизначеність і неоднорідність даних, що підвищує точність і стабільність ансамблевого класифікатора при класифікації аудіо-даних [15, 16, 17]. Таким чином, використання підходу stacking для побудови ансамблевих класифікаторів стає актуальним для підвищення точності класифікації аудіо-даних.

Для дослідження було використано два добре відомі набори даних: RAVDESS [5] і GTZAN [8]. Набір даних RAVDESS містить аудіозаписи мови, де актори висловлюють різні емоції. Він містить загалом 1440 файлів і 7 емоцій, таких як спокій, щастя, сум, гнів, страх, здивованість та огида. GTZAN – це набір даних, який містить аудіозаписи музики (1000 аудіо-файлів у форматі WAV), які представляють різні фрагменти музичних композицій. Кожен фрагмент має тривалість 30 секунд. Файли розділені на 10 різних музичних жанрів: блюз, класика, кантрі, диско, хіп-хоп, джаз, метал, поп, реггі, рок. Кожен музичний жанр містить 100 семплів у наборі даних, що забезпечує рівномірний розподіл. Окрім аудіо-файлів, набір даних також містить метадані про кожен семпл, наприклад назву файлу та музичний жанр.

Метою попередньої обробки аудіо-даних RAVDESS і GTZAN є отримання наступних спектральних характеристик [9]: спектральна пологість, спектральний центроїд, спектральний контраст, спектральний спад, швидкість переходу через нуль, мел-частотні кепстральні коефіцієнти, RMS. Для аналізу набору даних GTZAN були обрані додаткові хроматичні спектральні характеристики.

Етап попередньої обробки завершується розподілом отриманих даних на навчальну та тестову вибірки. У нас є 1080 екземплярів у тренувальному наборі та 360 у тестовому наборі для RAVDESS, 750 екземплярів у тренувальному наборі та 250 у тестовому наборі для GTZAN.

Для навчання окремих класифікаційних моделей машинного навчання (KNN, SVM, XGBoost, Random Forest) насамперед були налаштовані гіперпараметри за допомогою Grid Search. Ідея методу Grid Search полягає в тому, що він перебирає всі комбінації різних гіперпараметрів із попередньо визначеного списку значень, а потім використовує перехресну перевірку для визначення оптимального набору параметрів, який демонструє найкращу ефективність класифікації. Для навчання індивідуальних класифікаційних моделей глибокого навчання (MLP, CNN, LSTM) для кожного з алгоритмів нейронної мережі розроблено відповідну архітектуру, де експериментально встановлено оптимальну кількість прихованих шарів та інших гіперпараметрів.

Отже, ми отримали 7 окремих класифікаторів з оптимальними гіперпараметрами, які на тестовому наборі забезпечили найкращий прогноз для кожного класу.

З 7 індивідуальних класифікаторів із підходом stacking і трьома видами агрегації було складено наступну кількість класифікаторів (з урахуванням того, що мінімальна кількість класифікаторів для ансамблю становить 3) для кожної з двох проблем (2):

$$C_n^k = 3\,(C_7^3 + C_7^4 + C_7^5 + C_7^6 + C_7^7) = 297, \qquad (2)$$

де C_n^k – кількість комбінацій n, взятих k за раз, причому множення на 3 пояснюється існуванням трьох типів агрегації ансамблю.

Для вибору найкращого класифікатора використовувалися такі метрики, як accuracy і оцінка F1.

Для обох завдань результати класифікації за окремими класифікаторами на тестових вибірках дали недостатні результати. Для задачі класифікації емоцій найкращий результат показав багатошаровий персептрон з accuracy = 0.728, F1 = 0.726. Для завдання класифікації музичних жанрів найкращий результат також показав багатошаровий персептрон з accuracy = 0.74, оцінка F1 = 0.739. Точність класифікації окремих класифікаторів для обох завдань наведена в Таблиці 1.

Таблиця 1. **Класифікаційна якість окремих класифікаторів**

Алгоритм	**Метрики для класифікації музики**		**Метрики для класифікації емоцій**	
	Accuracy	***F1-score***	***Accuracy***	***F1-score***
SVM	0.74	0.738	0.703	0.699
MLP	0.74	0.739	0.728	0.726
XGB	0.704	0.703	0.656	0.648
Random Forest	0.684	0.673	0.653	0.643
KNN	0.676	0.671	0.62	0.608
CNN	0.588	0.553	0.617	0.597
LSTM	0.448	0.442	0.597	0.592

Для завдання класифікації аудіо-емоцій серед 297 отриманих ансамблів за допомогою підходу stacking та трьох типів агрегації soft voting ансамбль, що складається з MLP, KNN, SVM, Random Forest, CNN і LSTM та використовує нечітке ранжування Гомперца, дав найкращий результат класифікаційних метрик. Були отримані accuracy = 0.808 і F1 = 0.806 (Таблиця 2).

Для завдання класифікації музичних жанрів серед 297 отриманих ансамблів за допомогою підходу stacking та трьох типів агрегації soft voting ансамбль, що складається з усіх семи розроблених індивідуальних класифікаторів та використовує нечітке ранжування Гомперца, дав найкращий результат класифікаційних метрик. Були отримані accuracy = 0.796 і F1 = 0.792 (Таблиця 2).

Отриманий результат пояснюється наступним: окремі компоненти в ансамблі відрізняються один від одного логікою роботи та представленням вхідних даних (табличні дані, зображення спектрограми, часові ряди), а також м’яке голосування є найбільш універсальним серед трьох типів агрегації, описаних вище, і враховує впевненість класифікаторів у своїх прогнозах.

Останнім етапом методу є порівняння якості класифікації найкращого з отриманих ансамблевих класифікаторів з класифікаторами, що були його складовими, на аудіо-даних із тестових наборів RAVDESS і GTZAN (Таблиця 2).

Було виявлено, що для задачі класифікації звукових емоцій точність ансамблю на 8.1 % вища, ніж у найкращого класифікатора у його складі (MLP), а F1 на 8 % вища. Для задачі розпізнавання музичного жанру відповідні числа становлять 5.6 % і 5.2 %.

Таблиця 2. **Показники точності розпізнавання кращих ансамблів**

Algorithm	**Метрики для класифікації музики**		**Метрики для класифікації емоцій**	
	Accuracy	***F1-score***	***Accuracy***	***F1-score***
Ensemble	0.796	0.792	0.808	0.806
	Різниця в метриках		***Різниця в метриках***	
KNN	18.9 %	19.8 %	12 %	12.1 %
SVM	10.6 %	10.8 %	5.6 %	5.4 %
CNN	19.2 %	21 %	20.8 %	23.9 %
LSTM	21.1 %	21.4 %	34.8 %	35 %
MLP	8.1 %	8 %	5.6 %	5.2 %
Random Forest	15.6 %	16.3 %	11.2 %	11.9 %

Представлена робота присвячена створенню методу, що дозволяє підвищити точність класифікації аудіо-даних. На відміну від традиційних методів, які включають трудомісткі підходи, такі як розширення вхідних наборів даних або проведення детального спектрального аналізу для збільшення простору функцій, ця робота досліджує потенціал підвищення точності розпізнавання аудіо-даних за допомогою використання ансамблевих класифікаторів. Набір даних RAVDESS був обраний для задачі розпізнавання мовних емоцій, а набір даних GTZAN – для задачі розпізнавання музичних жанрів.

Використовуючи сім різних класифікаторів (SVM, KNN, Random Forest, XGBoost, MLP, CNN, LSTM) як компоненти ансамблю, дослідження вивчає їх ефективність у розпізнаванні емоцій і музичних жанрів за допомогою таких показників, як точність і F1. Показано, що найкращий класифікатор для першої задачі – MLP показує точність = 0.74 і F1 = 0.739, найкращий класифікатор для другої задачі – MLP показує точність = 0.728 і F1 = 0.726.

За допомогою підходу stacking з трьома різними типами агрегації було створено 297 ансамблів із 3-7 компонентами для кожної з двох проблем класифікації. Було виявлено, що soft voting ансамбль, що складається з MLP, KNN, SVM, Random Forest, CNN і LSTM та використовує нечітке ранжування Гомперца, дав найкращий результат класифікаційних метрик для проблеми розпізнавання звукових емоцій; soft voting ансамбль, що включає класифікатори MLP, KNN, SVM, RF, CNN, LSTM, XGB та використовує нечітке ранжування Гомперца, дав найкращий результат класифікаційних метрик для проблеми розпізнавання музичних жанрів.

Примітно, що ансамблевий класифікатор досяг 8 % підвищення accuracy та 8.4 % покращення оцінки F1 порівняно з найкращим індивідуальним класифікатором (MLP) для проблеми класифікації музики. Ця результуюча точність на 8.7 % вища ніж отримана в [8].

Відповідні цифри для точності та оцінки F1 для проблеми класифікації емоцій становили 5.6 % та 5.4 %. Крім того, отримана точність і F1-оцінка на 3.6 % і 3.5 % вищі, ніж отримані в [1], а точність на 8.1 % вища, ніж отримані в [9].

Отже, використання ансамблевих класифікаторів для розпізнавання аудіо-даних емоцій і жанрів музики є перспективним напрямком досліджень.

СПИСОК ЛІТЕРАТУРИ

1. Andronati O., Antoshchuk S., Nikolenko A., Arsirii O., Babilunha O.б Petrosiuk D. “Ensemble classifiers of audio data for speech emotions recognition”. *13th International Conference on Advanced Computer Information Technologies (ACIT*). Wrocław, Poland. 2023. p. 623–626. DOI: https://doi.org/10.1109/ACIT58437.2023.10275604.

2. Arsirii O. O., Petrosiuk D. V. “Pseudo-labeling of transfer learning convolutional neural network data for human facial emotion recognition“. *Herald of Advanced Information Technology.* 2023; 6 (3): 203−214. DOI: https://doi.org/10.15276/hait.06.2023.13.

3. Arsirii O. O., Petrosiuk D. V. “An adaptive convolutional neural network model for human facial expression recognition”. *Herald of Advanced Information Technology.* 2023; 6 (2): 128–138. DOI: https://doi.org/10.15276/hait.06.2023.8.

4. Gourisaria M. K., Agrawal R., Sahni M. et al. “Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques”. *Discov Internet Things.* 2024; 4 (1). DOI: https://doi.org/10.1007/s43926-023-00049-y.

5. Livingstone S. R., Russo F. A. “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*. 2018; 13(5): p.e0196391. DOI: https://doi.org/10.1371/journal.pone.0196391.

6. “CREMA-D”. – Available from: https://paperswithcode.com/dataset/crema-d.

7. “SAVEE”. – Available from: https://paperswithcode.com/dataset/savee.

8. “GTZAN”. – Available from: https://paperswithcode.com/dataset/gtzan.

9. Ayon R. D. G., Rabbi M. S., Habiba U., Hasana M. “Bangla speech emotion detection using machine ensemble methods”. *Advances Learning in Science, Technology and Engineering Systems Journal.* 2022; 7 (6): 70–76. DOI: https://dx.doi.org/10.25046/aj070608.

10. Garg S., Varshney A. “Music genre classification”. *International Journal of Advanced Research in Computer and Communication Engineering*. 2022. DOI: https://doi.org/10.17148/IJARCCE.2022.11551.

11. Surabhi V, Saurabh M. “Speech emotion recognition: A review”. *International Research Journal of Engineering and Technology (IRJET)*. 2016; 03: 313–316.

12. Singh V., Prasad S. “Speech emotion recognition system using gender dependent convolution neural network”. *Procedia Computer Science*. 2023; 218: 2533–2540. DOI: https://doi.org/10.1016/j.procs.2023.01.227.

13. Kerkeni L., Serrestou Y., Mbarki M., Raoof K., Ali Mahjoub M., and Cleder C. “Automatic speech emotion recognition using machine learning”. *Social Media and Machine Learning. IntechOpen*. 2020. DOI: https://doi.org/10.5772/intechopen.84856.

14. Mohan M., Dhanalakshmi P., Kumar R. S. “Speech emotion classification using ensemble models with MFCC”. *Procedia Computer Science*. 2023; 218: 1857–1868. DOI: https://doi.org/10.1016/j.procs.2023.01.163.

15. Dhara T., Singh P. K. & Mahmud M. “A fuzzy ensemble-based deep learning model for EEG-based emotion recognition”. *Cogn Comput.* 2023. DOI: https://doi.org/10.1007/s12559-023-10171-2.

16. Sahoo K. K., Dutta I., Ijaz M. F., Woźniak M., Singh P. K. “TLEFuzzyNet: Fuzzy rank-based ensemble of transfer learning models for emotion recognition from human speeches”. *IEEE Access*. 2021; 9: 166518–166530. DOI: https://doi.org/10.1109/ACCESS.2021.3135658.

17. Kundu R., Basak H., Singh P. K. et al. “Fuzzy rank-based fusion of CNN models using Gompertz function for screening COVID-19 CT-scans”. *Scientific Reports*. 2021; 11 (1): 14133. DOI: https://doi.org/10.1038/s41598-021-93658-y.

DOI: https://doi.org/10.15276/ict.01.2024.23
UDC 004.81;167.7

A method of developing ensemble classifiers for recognizing audio data of various nature

Oleksandr K. Andronati [1)]
Postgraduate student, Department of Information Systems
ORCID: https://orcid.org/orcid.org/0009-0009-1794-5864; alex.andronati@gmail.com. Scopus Author ID 58677655800
Olena O. Arsirii[1)]
Dr. Sc., Professor, head of Department ofInformation Systems
ORCID: https://orcid.org/orcid.org/0000−0001−8130−9613; e.arsiriy@gmail.com. Scopus Author ID: 54419480900 **Anatoly O. Nikolenko**[1)]
PhD, Associate Professor, Department of Information Systems
ORCID: https://orcid.org/0000-0002-9849-1797; anatolyn@ukr.net. Scopus Author ID: 57197491335
Kunditv I. Oleg[1)]
Graduate student, Department of Information Systems
ORCID: https://orcid.org/orcid.org/ 0009-0001-4499-1630; oleh.kundiiev@pm.me
[1)] Odesa Polytechnic National University, 1, Shevchenko Ave. Odesa, 65044, Ukraine

ABSTRACT

The work developed a method of building ensemble classifiers for recognizing audio data of various nature. The method is a tentative date and requires the following steps. In the first step, the input datasets are selected, which are transformed and divided into training and test samples, respectively. RAVDESS datasets are chosen for the task of audio emotion recognition, music genre recognition is performed on the GTZAN dataset. In the second step, the following seven classifiers were created and investigated as elements of ensemble classifiers: K Nearest Neighbors, Support Vector Machine, Random Forest, XGBoost, Multilayer Perceptron, Convolutional Neural Network and Long Short-term Memory. In the process of training elementary classifiers, the corresponding hyperparameters were adjusted using the Grid Search approach. In the third step, elementary classifiers were combined using the stacking ensemble method with such types of aggregation as soft voting, hard voting, soft voting using the GOMPERTZ function. All possible ensemble combinations starting with three elementary classifiers for recognizing audio emotions and music genres were tested. Therefore, the total number of ensembles studied in the work was 297. The research results for the problem of audio emotion classification showed that the accuracy of recognition according to the Accuracy metric of the best ensemble classifier is 8.1% higher than that of the best elementary classifier in its composition, which is based on on MLP, and according to the F1 metric, this indicator is 8% higher. For the task of recognizing music genres, the corresponding indicators are higher by 5.6 % and 5.2 %, respectively.

Keywords: Ensemble classifiers; hyperparameter tuning; machine learning; deep learning; audio data