УДК 0048:681.3

А.Г. КРИВОХАТА, О.В. КУДІН, А.О. ЛІСНЯК
Запорізький національний університет

ОГЛЯД МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ КЛАСИФІКАЦІЇ АКУСТИЧНИХ ДАНИХ

Останнім часом все більшого поширення набувають автоматичні системи розпізнавання звуку, зображень, відео. Такі системи знаходять різноманітні застосування на виробництві, інформаційній безпеці тощо. Серед методів, що застосовуються у таких задачах, слід виділити методи машинного навчання, як найбільш популярні та перспективні. У статті запропоновано огляд алгоритмів машинного навчання класифікації акустичних даних різного походження (природні звуки, мова, музика). Акцент робиться на глибинному машинному навчанні.

Ключові слова: акустичні дані, класифікація, машинний слух, машинне навчання.

А.Г. КРИВОХАТА, А.В. КУДИН, А.А. ЛИСНЯК
Запорожский национальный университет

ОБЗОР МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ АКУСТИЧЕСКИХ ДАННЫХ

В последнее время все большее распространение получают автоматические системы распознавания звука, изображений, видео. Такие системы находят различные применения в производстве, информационной безопасности и т.д. Среди методов, которые применяются в таких задачах, следует выделить методы машинного обучения, как наиболее популярные и перспективные. В статье предложен обзор алгоритмов машинного обучения классификации акустических данных разного происхождения (природные звуки, речь, музыка). Акцент делается на глубоком машинном обучении.

Ключевые слова: акустические данные, классификация, машинный слух, машинное обучение.

A.G. KRYVOKHATA, O.V. KUDIN, A.O. LISNYAK
Zaporizhzhya National University

A SURVEY OF MACHINE LEARNING METHODS FOR ACOUSTIC DATA CLASSIFICATION

The task of developing machines with sensing capabilities, such as vision and hearing is one of the challenging problems in engineering and computer science. Machine hearing is the field where different problems such as audio tagging have been formulated. Tag is a short name for a label applied to some audio by an automatic tagging algorithm. There are many applications for automated audio tagging systems, for example, music recommendation systems, estimating song similarity, etc. Many researchers exploit machine learning algorithms for developing such systems. However, there is no "one-fits-all" approach to solving the problem. Thus, an overview of state-of-the-art approaches and methods is relevant.

Basically, automated audio tagging systems can be decomposed into four parts: audio representation, features extraction, machine learning algorithm and accuracy estimation.

Audio representation stage means that raw signal is segmented into shorter signal chunks by some windowing process. In traditional methods, a common approach is to convert the original acoustic signal to frames of a certain length.

Receiving a compact representation of the acoustic characteristics of a signal is the aim of feature extraction stage. Zero-crossing rate, Spectrum shape, Short-Time Fourier Transform, Mel-frequency cepstral coefficients are widely used for feature extraction.

Machine learning methods like K-means, SVM, KNN, decision trees etc. are traditionally used in audio classification. In the last two decades, the deep learning based methods have also gained popularity for audio tagging, for example, convolutional neural networks or recurrent neural networks. Deep neural networks can operate on both raw acoustical signal and features extracted from the signal.

Accuracy estimation stage deploys quality assessment methods, for example, expert evaluation.

The purpose of this work is the analytical review of recent publications on the classification of acoustic data by means of machine learning with an emphasis on the application of deep machine learning methods.

We propose a possible direction for further development of machine hearing systems based on the analysis of publications and respective methods. This approach can use different types of ensemble learning methods with classifiers based on feature extraction and deep neural networks.

Keywords: acoustic data, classification, machine hearing, machine learning.

Постановка проблеми

Сучасний розвиток засобів телекомунікації та поширеність інструментів для редагування вмісту інтернет сайтів призводить до того, що в глобальній мережі Інтернет поряд з текстовою інформацією великого поширення набувають мультимедійні дані різного вмісту, зокрема, акустичні дані. Прикладом акустичних даних можуть бути музичні записи, записи лекцій, доповідей, записи звуків різного походження тощо. Для можливості пошуку серед таких даних, зазвичай, використовуються метадані, які описують в текстовому вигляді вміст відповідного аудіо файлу. Формування таких метаданих виконується вручну, що не завжди зручно при обробці великих об'ємів даних. Тому, актуальною задачею є розробка автоматизованих систем класифікації акустичних даних.

Прикладом автоматизованих систем обробки мультимедійних даних можуть бути рекомендаційні системи, які пропонують користувачам певний контент в залежності від даних, вказаних в профілі користувача та історії попередніх запитів. Також, актуальним напрямом в останні роки є машинний слух [14]. Однією з задач цього напряму є розробка ефективних методів класифікації звуків різного походження, наприклад, мови, музики, природних звуків тощо. При цьому, найбільш досліджуваними є саме задачі аналізу музики та мови [1]. Іншою задачею, яка досить часто розглядається авторами, є виявлення звукових подій. Ця задача спрямована на обробку неперервного акустичного сигналу та перетворення його в символічні описи відповідних звукових подій, присутніх на слухової сцені [13].

В загальному вигляді, процес аналізу цифрових акустичних даних зазвичай складається з декількох етапів. На початковому етапі виконується попередня обробка неперервного акустичного сигналу з метою представлення його у дискретному цифровому вигляді. При цьому, зазвичай, використовується ряд стандартних підходів [5, 16]. Далі виділяються ознаки акустичного сигналу, суттєві для розв'язання поставленої задачі аналізу. Серед найбільш широко вживаних ознак використовуються коефіцієнти перетворення Фур'є та автокореляції, мел-кепстральні коефіцієнти, хромограми. Після цього отримані ознаки використовуються як вхідні параметри математичної моделі (наприклад, алгоритму класифікації, кластеризації або нейронної мережі). На заключному етапі виконується верифікація отриманих результатів та впровадження розробленої системи акустичного аналізу.

На даний момент тематиці розробки систем машинного слуху присвячена велика кількість публікацій і продовжують з'являтися нові. Для того, щоб мати змогу виділити загальні тенденції серед безлічі робіт, необхідними є оглядові публікації, які містять аналіз спільних рис та відмінностей у підходах, які використовують різні автори.

Аналіз останніх досліджень і публікацій

Серед великої кількості оглядових робіт можна виділити декілька, які є найбільш загальними. Так, в оглядових статтях [1, 3, 19] наводиться опис компонент системи автоматичної класифікації звуків, яка містить модулі попередньої обробки, екстракції ознак, алгоритм навчання та модуль обчислень.

В [1] детально розглянуто підходи до виділення ознак сигналу. Наводяться критерії, за якими можна класифікувати мову, музику та природні звуки. Виділяються методи, засновані на фізичних властивостях сигналів та особливостях людського сприйняття звуків. Частіше за все використовуються методи виділення ознак, які представляють акустичний сигнал у таких областях: часовій, частотній, кепстральній та вейвлет.

Огляди [3, 19] містять аналіз загальних підходів та публікацій з автоматичної класифікації музичних записів за жанрами. Пропонується множина найбільш інформативних міток, які можуть використовуватись як класи при навчанні класифікаторів. Розглядаються найбільш вживані джерела розмічених акустичних даних, які можуть використовуватися у системах навчання з вчителем. Зазвичай, це відкриті музичні бази в мережі Інтернет, розмічені користувачами записи, наприклад, у соціальних мережах та дані, які згенеровано спеціально для розв'язання задач машинного слуху. В роботі [19] окремо розглядається питання оцінки ефективності систем класифікації музичних файлів за жанрами.

Попри наявність доволі докладних оглядів з даної тематики, більшість з них акцентуються на класичному підході до аналізу даних засобами машинного навчання, який складається з етапів: виділення ознак, навчання системи, верифікація та впровадження системи. Недостатня увага приділяється методам глибинного машинного навчання, які включають нейронні мережі з великим числом прихованих шарів.

Мета дослідження

Метою даної роботи є аналітичний огляд останніх публікацій з класифікації акустичних даних засобами машинного навчання з акцентом на застосування методів глибинного машинного навчання.

Викладення основного матеріалу дослідження

Для класифікації даних, які описані своїми ознаками, можуть використовуватися як статистичні методи (класифікатор Баєса [7], дискримінантний аналіз [4, 10], EM алгоритм тощо), так і методи, які ґрунтуються на мірах схожості та відмінності (метод k-середніх [5, 11, 17], метод опорних векторів [5, 10, 15], метод k найближчих сусідів [5, 6] тощо).

У роботі [7] розв'язується задача класифікації довільних аудіо даних. Пропонується автоматична система, яка розподіляє вхідні аудіо дані на сім класів, зокрема, різні типи музики, мова одного чи декількох

чоловік, звуки зовнішнього середовища. Виконано порівняння декількох ознак класифікації, показано, що найліпший результат досягається при використанні мел-частотних кепстральних коефіцієнтів та коефіцієнтів кодування з лінійним предиктором [16]. Процес аналізу аудіо даних складається з таких етапів: вилучення ознак, ідентифікація пауз, сегментація на фрагменти та класифікація фрагментів. Для класифікації використовується класифікатор Баєса.

Автоматичний аналіз акустичних даних знаходить застосування також і в комп'ютерній лінгвістиці. В статті [10] автори провадять лінгвістичне дослідження впливу наголосу на певні частини речення в англійській мові на сприйняття змісту. Описана в роботі методологія дослідження передбачає використання акустичного класифікатора (метод опорних векторів та лінійний дискримінантний аналіз). Для визначення ознак, що мають найбільший вплив на результат класифікації використовується метод Борута.

У роботі [11] вводиться поняття матриці часової залежності, яка містить дані про кількість повторень певних типів інтервалів у звукових даних. Кластеризація аудіо даних на основі отриманої матриці часової залежності виконується за допомогою метода k-середніх. Проілюстровано використання розроблених алгоритмів для кластеризації аудіо та відео документів.

В останні роки все більше робіт присвячено використанню нейронних мереж, як при вилученні ознак з даних, так і безпосередньо при класифікації.

Лінійна та нелінійна модель нейронів для виділення ознак з аудіо даних наводиться в роботі [2]. При цьому моделюється спектрально-часове поле сприйняття нейрону, а для навчання використовують дані, отримані безпосередньо з нейронів слухового апарату птахів виду зяблик-зебра. Досліджено вплив нелінійної функції активації нейронів на результат класифікації за шістьма класами. Як класифікатор використовувалися три методи: класифікатор Баєса, метод опорних векторів та метод гауссових сумішей. Показано, що використання нелінійної моделі нейронів дозволяє підвищити якість класифікації на 15%.

Метод деформаційного дискримінантного аналізу звукових даних розглядається в [4]. Застосування цього методу до попередньо обробленого звукового сигналу дозволяє отримати стійкий до шуму вектор ознак. Кожен шар деформаційного дискримінантного аналізу, в свою чергу, використовує метод головних компонент та згорткові нейронні мережі [22] для генерування вектору ознак з вхідного звукового файлу та класифікації. Як приклад описано архітектуру системи, що дозволяє знаходити звукові відрізки у потоковому аудіо.

У статті [6] запропоновано підхід до автоматичної класифікації музичних записів за жанрами. Особливістю підходу є те, що вихідні музичні файли розбиваються на три частини, для яких обчислюються власні 15-мірні вектори ознак. Після цього кожен з векторів передається на вхід класифікатору. Результат отримується шляхом голосування. На етапі визначення ознак використовується швидке перетворення Фур'є та обчислюються такі ознаки, як спектральний центроїд та інші. Використовуються два типи класифікаторів: метод k найближчих сусідів та багатошарова нейронна мережа з одним прихованим шаром.

Згорткові нейронні мережі для отримання стійких ознак з аудіо даних використовуються в роботах [12, 13]. Особливістю використання цього типу нейронних мереж є те, що на вхід можна подавати як вектори ознак, так і «сирі», тобто, попередньо не оброблені акустичні дані у цифровому форматі. В [12] додатково застосовано рекурентні нейронні мережі з вентильними рекурентними вузлами для класифікації акустичних даних. Проводяться обчислювальні експерименти на базі звуків оточуючого середовища. У роботах [13, 21] розв'язується задача виявлення специфічних звукових подій, наприклад, плачу дитини, звуку сигналізації або пострілів у публічних місцях тощо.

У статті [15] пропонується автоматична система екстракції ознак, в якій використовується генетичні алгоритми для отримання оптимальної, з точки зору певної задачі, комбінації операторів. Оператором можуть виступати різноманітні перетворення сигналу з метою отримання додаткових характеристик: алгоритми перетворення Фур'є, фільтри, алгоритми визначення інтервалів тощо. Функція пристосованості генетичного алгоритму обчислюється як міра якості класифікації аудіо даних, для класифікації використовується метод опорних векторів.

Робота [17] присвячена задачі класифікації музики та мовлення. Пропонується метод оптимізації числа ознак для задачі класифікації аудіо записів музики та мовлення у форматі MPEG. Для класифікації використовується метод k-середніх та нечітка нейронна мережа. Для виділення тих ознак, які найбільш впливають на результат класифікації застосовуються два підходи: аналіз головних компонент та генетичний алгоритм.

У [18] розглядається задача класифікації музичних записів та співу птахів. Для класифікації використовується комбінація методу кластеризації за допомогою самоорганізаційних карт Кохонена та класифікатор на основі нейронних мереж векторного квантування.

В [20] розглядається випадок задачі класифікації аудіо даних, при якому дані для навчання включають фоновий шум, а дані тестового набору записано без шуму. В цьому випадку розподіли ознак відповідних даних для тренування та тестування можуть бути схожі, але зміщенні один відносно другого. Розроблено модифікацію логістичної регресії для нівелювання зміщень при навчанні та тестуванні класифікатора.

Окремою задачею є підготовка акустичних даних різного походження та позначення їх відповідними мітками, що необхідно для реалізації методів машинного навчання з вчителем. У роботі [9] визначається онтологія, що формалізує набір аудіо даних, який може використовуватися для навчання систем машинного слуху. Онтологія, яка вводиться в статті, визначає систему можливих категорій звуків для розпізнавання. Пов'язаний з цією роботою ресурс research.google.com/audioset містить набір розмічених вручну звукових роликів з YouTube (понад 2 мільйони файлів). Кожен з десяти секундних сегментів може відноситься до одного чи декількох класів онтології. Також відомими платформами з даними для систем машинного слуху є freesound.org, DCASE (dcase.community) та, певною мірою, kaggle.com.

Висновки

З аналізу літературних джерел можна зробити висновок, що задача класифікації акустичних даних і, загалом, розробки систем машинного слуху є досить актуальною. Опубліковані на даний момент наукові статті можна умовно розділити на три категорії.

До першої відносять роботи, в яких виконується попередня обробка сигналу з метою сегментації та вилучення ознак, далі навчається класифікатор на вхід якого подаються вектори ознак. В цих роботах, зазвичай, застосовується перетворення Фур'є, обчислення мел-частотних кепстральних коефіцієнтів та інших частотних або спектральних характеристик сигналу. З класифікаторів частіше застосовуються метод опірних векторів, k-найближчих сусідів, дерева прийняття рішень, метод k-середніх, нейронні мережі та інші. Може також застосовуватися ансамбль декількох класифікаторів, в такому випадку клас-переможець визначається шляхом голосування.

До другої категорії можна віднести публікації, в яких автори намагаються автоматизувати процес побудови оптимального набору ознак для застосування класифікаторів. Серед підходів, які застосовуються для такої автоматизації, можна виділити генетичні алгоритми та нейронні мережі. Класифікатори використовуються ті ж, що і в публікаціях попередньої категорії.

В публікаціях третьої категорії застосовуються підходи глибинних нейронних мереж. Часто це згорткові нейронні мережі, на вхід яких можуть подаватися як дані без попередньої обробки, так і набори ознак акустичних даних. Ефективність такого підходу пояснюється багатошаровою архітектурою згорткових нейронних мереж. Передбачається наявність декількох типів шарів: шари згортки, в яких виділяються певного виду ознаки, агрегувальні шари, в яких відбувається зменшення розмірності та декілька повністю зв'язних шарів, в яких виконується класифікація. До недоліків такого підходу можна віднести складність налаштування нейронних мереж зі складною архітектурою та вимогливість до обчислювальних ресурсів. Реалізація глибинних нейронних мереж, зазвичай, потребує системи паралельних та розподілених обчислень, залучення графічних процесорів GPU.

На основі наведеного в роботі аналізу публікацій та методів, що в них застосовуються, можна запропонувати можливий варіант подальшого розвитку систем машинного слуху. Такий підхід може використовувати різні види ансамблевого навчання із застосуванням класифікаторів на основі ознак та глибинних нейронних мереж. Таким чином, різні класифікатори, на вхід яких подаються різні вектори ознак або дані без попередньої обробки, можуть бути відносно ефективними на різних даних, але об'єднуватися в один ефективний класифікатор-ансамбль. Перевагою такого підходу може бути його адаптивність з точки зору вимогливості до обчислювальних ресурсів, оскільки, за необхідністю, можна коректувати кількість класифікаторів, які беруть участь в аналізі.

Список використаної літератури

1. Alias F. A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds / F. Alias, J.C. Socoro, X. Sevillano // J. Appl. Sci., 2016. – 6(5), 143. P. 1–44.
2. Bach J.-H. Automatic classification of audio data using nonlinear neural response models / J.-H. Bach, A.-F. Meyer, D. McElfresh, J. Anemüller // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 2012. – P. 357–360.
3. Bertin-Mahieux T. Automatic tagging of audio: the state-of-the-art. Machine audition: principles, algorithms and systems / T. Bertin-Mahieux, D. Eck, M. Mandel // IGI Global, 2011. – P. 334–352.
4. Burges C.J.S. Extracting noise-robust features from audio data / C.J.S. Burges, J.C. Platt, S. Jana // Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, FL, USA, 13–17 May 2002, 2002. – P. 1021–1024.
5. Camastra F. Machime learning for audio, image and video analysis / F. Camastra, A. Vinciarelli – London, Springer-Verlag, 2015. – 561 p.
6. Costa C.H.L. Automatic classification of audio data / C.H.L. Costa, Jr. J.D. Valle, A.L. Koerich // Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2004. – P. 562–567.
7. Dongge Li. Classification of general audio data for content-based retrieval / Li Dongge, Ishwar K. Sethi, Nevenka Dimitrova, Tom McGee // Pattern Recognition Letters, 2001. – Vol. 22(5). – P. 533–544.
8. Free sound General-Purpose Audio Tagging Challenge. URL: https://www.kaggle.com/c/freesound-audio-tagging/data (Дата звернення 06.06.2018)

9. Gemmeke J.F. Audio set: an ontology and human-labeled dataset for audio events / J.F. Gemmeke, D.P.W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal, M. Ritter // Proceedings of the Acoustics, Speech and Signal Processing International Conference, 2017.
10. Howel J. Acoustic classification of focus: on the web and in the lab / J. Howel, M. Rooth, M. Wagner // Laboratory Phonology: Journal of the Association for Laboratory Phonology, 2007. – 8(1): 16. – P. 1–41.
11. Ibrahim Z. Al A. Audio data analysis using parametric representation of temporal relations / Z. Al A. Ibrahim, I. Ferrane, P. Joly // IEEE International Conference on Information and Communication Technologies: from Theory to Applications (ICTTA), 2006.
12. Kong Q. Convolutional gated recurrent neural network incorporating spatial features for audio tagging / Q. Kong, Y. Xu, W. Wang, M.D. Plumbley // The 2017 International Joint Conference on Neural Networks (IJCNN 2017), Anchorage, Alaska, 2017.
13. Kong Q. A joint separation-classification model for sound event detection of weakly labelled data / Q. Kong, Y. Xu, W. Wang, M.D. Plumbley // ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 15 - 20 Apr 2018, Calgary, Canada, 2018.
14. Lyon R.F. Machine Hearing: An Emerging Field / R.F. Lyon // IEEE Signal Process. Mag, 2010. – Vol. 27. – P. 131–139.
15. Mierswa I. Learning feature extraction for learning from audio data / I. Mierswa, K. Morik // Technische Universität Dortmund. Technical Reports, 2004 – No. 55.
16. Oppenheim A.V. Discrete-Time Signal Processing. Third edition / A.V. Oppenheim // Pearson Education Limited, 2014 – 1055 p.
17. Rizzi A. Optimal short-time features for music/speech classification of compressed audio data / A. Rizzi, M. Buccino, M. Panella, A. Uncini // International Conference on Intelligent Agents. 28 Nov.-1 Dec. 2006. Sydney, NSW, Australia, 2006.
18. Stastný J., Skorpil V., Fejfar J. Audio data classification by means of new algorithms / J. Stastný, V. Skorpil, J. Fejfar // 36th International conference on Telecommunications and Signal Processing 2013, Rome, Italy, 2013. – P. 507–511.
19. Sturm B.L. A Survey of evaluation in music genre recognition / B.L. Sturm, A. Nürnberger, S. Stober, B. Larsen, M. Detyniecki (eds) // Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation. AMR 2012. Lecture Notes in Computer Science, 2014. – Vol 8382. – P. 29–66.
20. Wichern G. Automatic audio tagging using covariate shift adaptation / G. Wichern, M. Yamada, H. Thornburg, M. Sugiyama, A. Spanias // IEEE international conference Acoustics speech and signal processing (ICASSP), 14–19 Mar 2010, 2010.
21. Xu Y. Unsupervised feature learning based on deep models for environmental audio tagging / Y. Xu, Q. Huang, W. Wang, P. Foster, S. Sigtia, P.J.B. Jackson, M.D. Plumbley // IEEE/ACM transactions on audio, speech and language processing, 2017. – Vol 25., No. 6. – P. 1230–1241.
22. Zaccone G., Karim Md. R. Deep learning with TensorFlow / G. Zaccone, Md. R. Krim. – Packt Publishing, 2018. – 767 p.