К. Ж. Жаксылык

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ ЗВУКОВ ОКРУЖАЮЩЕЙ СРЕДЫ

Данная статья посвящена вопросу автоматической классификации звуков окружающей среды, которая является важной задачей в области обработки аудиоданных. В статье представлен обзор основных подходов к классификации звуков, включая методы машинного обучения, глубокого обучения и статистического анализа. Представлен анализ основных наборов данных (датасетов) используемых в данной области для тренировки алгоритмов машинного обучения.

Введение

Окружающий звук/аудио — это богатый источник информации, который можно использовать для определения контекста человека в повседневной жизни. Почти каждая деятельность воспроизводит некоторые звуковые образы, например, речь, ходьба, стирка или набор текста на компьютере. В большинстве мест также обычно есть определенный звуковой рисунок, например, рестораны, офисы или улицы города.

Автоматическая классификация звуков окружающей среды (АКЗОС, Environmental Sound Classification – ESC) – направление исследований в области обработки аудиосигналов, в котором ставится задача получения информации о звуковых событиях, происходящих в окружающей среде, и обнаружении определенных типов звуков, которые представляют интерес. Это область, которая находится в стадии активного развития, поскольку большинство традиционных исследований сосредоточено на речевых и музыкальных сигналах. Классификация звуков может быть использована для автоматического аудиомониторинга среды, что может быть полезно в системах безопасности (звуки дтп, взрывы), защиты окружающей среды (контроль за уровнем шума), управления умным домом (проникновение нарушителей), автономных автомобилях (гудок встречной машины) и пр.

Создание системы автоматической классификации звуков подразумевает решение нескольких важных промежуточных задач, ответы на которые необходимы для построения всей системы: сбор данных для обучения, выбор методов предварительной обработки, извлечения признаков, отбора признаков и выбора метода классификации данных (рис. 1).

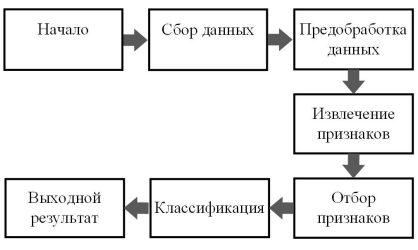

Рис. 1 – Основные этапы работы автоматической классификации звуков окружающей среды

Далее приведем обзор и сравнительный анализ актуальных подходов, характерных для каждого из этапов работы системы.

I. Этапы

Обзор открытых наборов данных

Первый этап в АКЗОС — сбор данных. В данной статье акцент будет придаваться этой части, т.к. выбор подходящего датасета для классификации звуков окружающей среды, является критически важным для достижения хорошей точности и эффективности работы модели. Датасет должен содержать достаточно разнообразные сигналы, чтобы покрыть все возможные звуки, которые могут встретиться в окружающей среде. Это позволит модели лучше обобщать и распознавать новые звуковые сигналы.

Для задач классификации звуков существует несколько открытых наборов данных, таких как UrbanSound8K, ESC-50, AudioSet и т.д.

Одним из самых часто применяемых для проведения исследований является набор данных «UrbanSound8K». Датасет UrbanSound8K, включает в себя 8732 маркированных звукозаписей из 10 классов, среди них представлены такие примеры звуков как: звуки транспортных средств, звуки речевой активности (в частности людей), звуки животных, звуки строительных работ, звуки выстрелов и звуки оповещения. Все файлы были записаны в реальных условиях на улицах и в помещениях различных городов по всему миру. Продолжительность каждой записи составляет менее 4 секунд. Данные были взяты из бесплатного репозитория звуков.

Датасет ESC-50 состоит из 2000 звуковых записей, продолжительность каждой составляет 5 секунд. Все звуки были записаны в формате WAV с частотой дискретизации 44,1 кГц и разрядностью 16 бит. Каждая звуковая запись сопровождается файлом JSON, который содержит метаданные, такие как название файла, номер класса и описание звука.

Датасет ESC-10 содержит звукозаписи из 10 категорий звуков природной среды. Он состоит из 400 звуковых файлов в формате WAV, длительностью от 1 до 5 секунд. Каждый файл был проаннотирован вручную и отнесен к одной из

10 категорий звуков. ESC-10 является подмножеством ESC-50.

Предварительная обработка

Предварительная обработка необходима для удаления шума и усиления аудиосигнала. Аудиосигналы должны быть предварительно обработаны, чтобы их можно было использовать для выделения признаков или классификации. На этапе предобработки часто применяются методы снижения размерности (Dimensionality Reduction), такие как метод главных компонент (Principal Component Analysis), линейный дискриминантный анализ (Liner Discriminant Analysis), автоэнкодер. Эти методы предварительной обработки необходимы для уменьшения размера произвольно длинных спектрограмм. Спектрограммы сглаживаются и удаляются шумы. Звуковые сигналы усиливаются с использованием набора перцептивных фильтров (Perceptual filterbank) и метода на основе подпространства (Subspace–based methods) [12].

Обзор методов извлечения и классификации признаков

Признаки для звуковой классификации в основном характеризуются тремя категориями: (рис. 2)

- кепстральные признаки;
- временные признаки;
- спектральные признаки.

Рис. 2 – Категории классификации признаков ESC

Задача идентификации и классификации звуков окружающей среды является довольно сложной, поскольку данные типы звуков менее структурированы и полны шумов, чем другие типы аудиосигналов. Чтобы обеспечить эффективность работы классификатора, важно найти отличительное и информативное звуковое представление в качестве набора признаков и применить к нему задачу классификации с надежным алгоритмом и моделью.

Для облегчения выполнения классификации, решающее значение имеет отбор аудио признаков, передаваемых в классификаторы. Для этого используются либо непосредственно отсчёты речевого сигнала и их производные (временные признаки), либо спектральные характеристики такие как коэффициенты оконного преобразования Фурье (STFT) и Mel–спектрограммы (спектральные признаки) (Mel–spectrum), либо же Mel–частотные кепстральные коэффициенты (MFCC) (кепстральные признаки). По сравнению с формой звукового сигнала, частотно–временные представления сохраняют больше информации и имеют меньшие размеры.

Несмотря на то, что большая по объему и сложная модель может классифицировать признаки непосредственно из необработанного сигнала, это может привести к большим вычеслительным затратам [1]. В исследовании Muhammad Huzaifah “Comparison of time–frequency representations for environmental sound classification using convolutional neural networks” было проведено сравнение различных частотно–временных представлений, где в качестве классификатора использовались сверточные нейронные сети (CNNs) [2]. Результаты показали, что Mel–спектрограммы превзошли Mel–частотные кепстральные коэффициенты (MFCC) и показали хорошие характеристики для разных наборов данных. MFCC применяет дискретное косинусное преобразование (DCT) к Mel – спектрограммам, эта операция декоррелирует спектральные энергии и теряет локальный паттерн в частотно–временном представлении [3].

Основные типы классификаторов машинного обучения и глубоких нейронных сетей, используемые для решения задачи автоматической классификации звуков (рис. 3.)

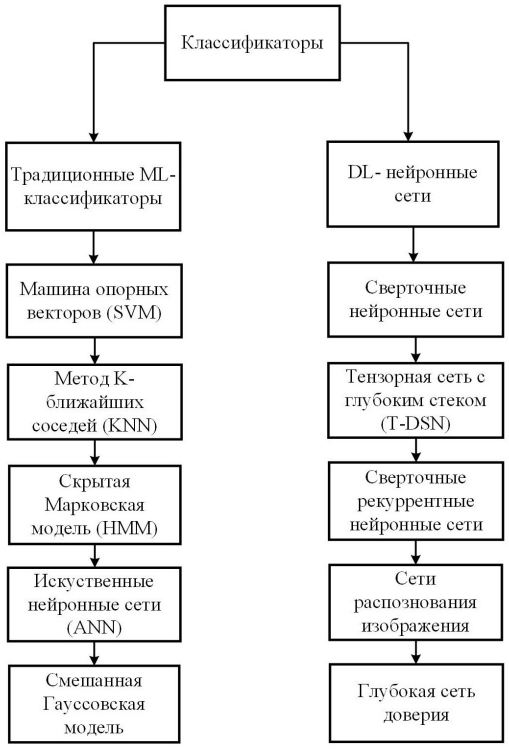

Рис. 3 – Классификаторы ML и DL

В последние годы во многих работах доказано, что модели на основе глубоких нейронных сетей более перспективны, чем традиционные классификаторы, при решении сложных задач классификации. Машины опорных векторов (SVM), модели Гауссовской смеси (GMM) и кластеризация k–средних широко используются в традиционных алгоритмах машинного обучения для задач классификации. Однако в приложениях для классификации звука данные классификаторы уязвимы к наличию шумов и чувствительны к временной динамике звука, что приводит к недостаточной надежности [4–7]. Наиболее популярной и относительно простой моделью глубокого обучения для задач классификации является сверточные нейронные сети (CNNs), которые обычно используются в приложениях компьютерного зрения и классификации изображений. Эта модель является многообещающей для задачи классификации звуков окружающей среды, поскольку звук можно интерпретировать как двумерное частотно-временное представление, в котором можно рассмотреть локализованные спектральные паттерны. В исследовании Karol J. Piczak "Environmental sound classification with convolutional neural networks" классификация звуков окружающей среды была самой первой работой, в которой оценивалась производительность задачи классификации городских звуков с использованием CNNs [8]. Его модель состоит из двух сверточных слоев с максимальным объединением, за которыми следуют два полносвязных слоя. Спектрограмма Log–Mel и ее дельта–информация использовались в качестве признаков аудиопредставления. Эксперимент был основан на трех общедоступных наборах данных:

- ESC–50;
- ESC–10;
- UrbanSound8K [9,10].

Для каждого набора данных была получена точность 81%, 65% и 73% соответственно. Zhichao Zhang и др. в исследовании "Deep convolutional neural network with mixup for environmental sound classification" предложили архитектуру CNNs с восемью сверточными слоями, за каждыми двумя сверточными слоями следовал слой максимального объединения [11]. Производительность этой предложенной CNN сравнивалась с VGG [12]. Результаты показали, что предлагаемая CNN работает лучше, чем VGG, оцененная по трем наборам данных ESC–50, ESC–10 и UrbanSound8K с использованием спектрограммы в качестве входных данных. Точность предложенной CNN составила 77%, 89% и 75% соответственно. В "Classifying environmental sounds using image recognition networks" исследователи Venkatesh Boddapati и др. применили AlexNet и GoogLeNet к спектрограммам звука и оценили наборы данных: ESC–50, ESC–10, UrbanSound8K [13–15].

Наилучшую точность дал GoogLeNet, которая составила 73%, 91% и 93% соответственно для каждого набора данных. Для тех же настроек GoogLeNet достигла более высокой точности классификации, чем AlexNet. Причина этого в том, что GoogLeNet значительно глубже и имеет гораздо больше слоев, чем AlexNet. В некоторых работах непосредственно использовались необработанные формы сигналов во временной области в качестве входных данных для модели классификации. В исследовании "Very deep convolutional neural networks for raw waveforms" автора Wei Dai и др. впервые использовали CNNs для классификации необработанных сигналов звуков окружающей среды [1]. Модель состоит из 34 слоев, где сверточные операции представляли собой 1–D свертки. Результат на UrbanSound8K составил 72%. Sajjad Abdoli и др. в своем исследовании "End–to–end environmental sound classification using a 1D convolutional neural network" разделили сигнал на перекрывающиеся кадры с помощью скользящего окна и использовали 1–D CNN, которая напрямую изучила признаки сигналов [16] . Точность модели на UrbanSound8K составила 89%, что является конкурентоспособным по сравнению с результатами других методов, использующих представления спектрограмм и 2–D CNN.

II. Заключение

В статье рассмотрены основные подходы и методы к решению задачи классификации звуков окружающей среды. Каждый метод имеет свои преимущества и недостатки, и выбор метода зависит от конкретной задачи и доступных ресурсов. Исходя из вышеперечисленного анализа, наивысшая точность на данный достигается с использованием сверточных нейронных сетей(CNNs). В будущем для классификации звуков окружающей среды необходимо изучить более обширные наборы данных, т.к. нет более крупного набора эталонных данных, кроме общедоступного UrbanSound8K. В целом, автоматическая классификация звуков окружающей среды является активно развивающейся областью исследований, и ее применение может быть полезным во многих сферах человеческой деятельности.

Список литературы

1. Wei Dai [et al.] Very deep convolutional neural networks for raw waveforms / Wei Dai // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2017/ – ICASSP, 2017. – pp. 421–425.
2. Muhammad H. Comparison of time–frequency representations for environmental sound classification using convolutional neural networks // In arXiv preprint arXiv: 1706.07156 (2017).

3. Ossama Abdel–Hamid [et al.] Convolutional neural networks for speech recognition : In IEEE/ACM Transactions on audio, speech, and language processing, 22 Oct, 2014. – pp. 1533–1545.
4. Antonio J. A tool for urban soundscape evaluation applying support vector machines for developing a soundscape classification model / J. Antonio, R. Diego, R. Angel // Science of the Total Environment – №482, 2014, – pp. 440–451.
5. Michael J B. Machine learning in acoustics: Theory and applications / J B. Michael [et al.] // The Journal of the Acoustical Society of America 146.5, 2019, – pp. 3590–3628.
6. Jia–Ching W. Environmental sound classification using hybrid SVM/KNN classifier and MPEG–7 audio low–level descriptor / W. Jia–Ching [et ai.] // The 2006 IEEE international joint conference on neural network proceedings, 2006. – pp. 1731– 1735.
7. Salamon J. Unsupervised feature learning for urban sound classification / J. Salamon, J. P. Bello // 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015. – pp. 171–175.
8. Piczak K. J. Environmental sound classification with convolutional neural networks / K. J. Piczak // 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), 2015. – pp. 1–6.
9. Piczak K. J. ESC: Dataset for Environmental Sound Classification [Electronic resource]. – Mode of access: https://doi.org/10.7910/DVN/YDEPUT.
10. URBANSOUND8K DATASET [Electronic resource] – Mode of access: https://urbansounddataset.weebly.com/urbansound8k.html.
11. Zhang Z. Deep convolutional neural network with mixup for environmental sound classification/ Z. Zhang [et al.]// Chinese Conference on Pattern Recognition and Computer Vision (PRCV), – Springer, 2018. – pp. 356–367.
12. Simonyan K., Zisserman A. Very deep convolutional networks for large–scale image recognition // arXiv preprint arXiv: 1409.1556 (2014).
13. Boddapati V. Classifying environmental sounds using image recognition networks/V. Boddapati // Procedia computer science. – 2017. – Vol. 112, №C. – pp. 2048–2056.
14. Krizhevsky A. Imagenet classification with deep convolutional neural networks /. A. Krizhevsky, I. Sutskever, G. E Hinton // Advances in neural information processing systems, 2012. – pp. 1097–1105.
15. . Szegedy C. Going deeper with convolutions / C. Szegedy [et al.] // Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. – pp. 1–9.
16. Abdoli S. “End-to-end environmental sound classification using a 1D convolutional neural network” / S. Abdoli, P. Cardinal, A. L. Koerich // Expert Systems with Applications. – 2019. – Vol. 136. – pp. 252–263.

Жаксылык Куаныш, магистрант кафедры информационных технологий автоматизированных систем Белорусского государственного университета информатики и радиоэлекстроники, kuanysh.zhk@gmail.com

Научный руководитель: Захарьев Вадим Анатольевич, доцент кафедры систем управления Белорусского государственного университета информатики и радиоэлектроники, кандидат технических наук, доцент, zahariev@bsuir.by