УДК 681.3 http://doi.org/10.15407/pp2024.01.048

Д.В. Рагозін, А.Ю. Дорошенко

КЛАСИФІКАЦІЯ НИЗЬКОЧАСТОТНИХ СИГНАЛІВ ЗА ДОПОМОГОЮ МЕТОДІВ КЛАСТЕРИЗАЦІЇ

У статті розглянуто методи класифікації сигналів звукового та інфразвукового діапазону за допомогою алгоритмів кластеризації у випадках, коли присутня лише загальна апріорна інформація, наприклад, невідомі типи об'єктів, які генерують відповідні сигнали. Розглянуто підготовку даних звукового діапазону, особливості первинної обробки набору даних, параметри вибору алгоритму залежно від особливостей набору даних. Подано приклади кластеризації набору даних за допомогою алгоритму OPTICS, описано можливості каскадної обробки набору даних.

Ключові слова: класифікація даних, навчання без контролю, кластеризація даних, обробка звуку.

Вступ

Сучасні методи обробки сигналів часто зосереджені на застосуванні нейронних мереж. Однак нейромережеві методи мають значний недолік – велику енергоємність, що у випадку достатньо довгого аналізу є фактором, який обмежує застосування нейромереж. Також застосування певних типів нейромереж обмежено у разі необхідності попереднього навчання нейромережі. Тому для практичних застосувань цікаво проаналізувати можливості застосування методів, альтернативних нейромережам і оцінити їхню ефективність у вирішенні задач класифікації сигналів низьких частот. У статті ми розглянемо особливості задач класифікації сигналів, особливості обробки звукових сигналів, сучасні методи аналізу сигналів, практичне застосування методів і можливості подальшого їхнього розвитку.

Обробка низькочастотних сигналів

Зазначимо, що низькочастотні сигнали не обмежуються електромагнітними хвилями, і обробка низькочастотних електромагнітних хвиль є обмежено корисною, оскільки для практичної передачі інформації в абсолютній більшості випадків використовуються високі частоти. Звукові та інфразвукові хвилі несуть у собі значну кількість інформації, яка може бути проаналізована в автоматичному режимі і в ощадливому режимі споживання енергії. Останнє важливо у разі автономної роботи датчиків, які додатково можуть і не випромінювати електромагнітну енергію.

Важливо, що на відміну від «домену» електромагнітних хвиль, де маємо апріорну інформацію щодо спектру сигналу і певних сигнатур сигналу, апріорна інформація щодо звукового сигналу зазвичай відсутня. Під сигнатурою тут ми розуміємо типові співвідношення потужностей сигналу в піддіапазонах виділених частот і їхню типову зміну з плином часу. Це дуже грубе визначення, оскільки багато факторів впливає на конкретні вимоги до аналізу спектру. Детальніше з питаннями симуляції сигнатур можна ознайомитися в [1], де розглядається один із варіантів протоколу Wi-Fi, який використовується як база для великого сімейства протоколів передачі даних. Отож, під час аналізу сигналів електромагнітного спектру ми знаємо з чим стикаємось, можемо проаналізувати нижні рівні комунікацій у визначеннях моделі OSI [2] і певним чином скористатися отриманою інформацією. Наприклад, у практичній площині розпізнавання об'єктів або радіоелектронної боротьби. Описаний аналіз електромагнітного спектру стає можливим, оскільки використання діапазонів електромагнітного спектру формалізоване. Однак у випадку, коли об'єкт не випромінює електромагнітні хвилі (режим радіомовчання), необхідно застосовувати інші методи. У звуковому діапазоні така стандартизація неможлива, окрім випадків штучного обмеження потужності генерації звука під час роботи якогось обладнання. Тому після пе-

рвинного аналізу звукового спектру ми маємо скористатися одним із методів навчання без контролю (англійською мовою unsupervised learning), найбільш актуальні з яких розглянуто в [3]. Зокрема, специфічні для промислових завдань методи навчання без контролю розглянуті в [4].

Нашою метою є класифікація об'єктів, які створюють певну електромагнітну або звукову картину до певних типів за умови, що апріорно ми не знаємо, які у нас типи об'єктів, чи з'являються у нас нові об'єкти. Далі будемо також позначати таку класифікацію терміном «кластеризація».

Розглянемо спочатку первинну обробку звукового сигналу, потім вторинну. Тобто спочатку методи, які дозволяють редукувати розмірність даних з метою подальшої ефективної обробки і кластеризації редукованих даних.

Первинна обробка сигналу

Звісно, первинна обробка низьких частот проводиться шляхом частотного аналізу, але є декілька можливостей такого аналізу. Практичні частоти для аналізу звукової інформації починаються від 0.1 Гц і закінчуються на 30 кГц (ультразвук). Таким чином діапазон обробки складає до 18 октав (під октавою ми розуміємо діапазон частот від f до $2f$ Гц). Стандартизовано у промисловій обробці потужність хвиль вимірюється у частинах діапазону у 1/3 октави (піддіапазонах). Для стандартного звукового діапазону 20 Гц - 20 кГц (10 октав), потужність сигналу вимірюється у 30 піддіапазонах у третю частину октави. Зазначимо, що ця технологія вимірювання є практичним спадком часів, коли частотний аналіз виконувався за допомогою фільтрів смуг сигналу шириною у третю частину октави на стандартизованих частотах. Зараз обчислювальна потужність збільшилась і дозволяє проводити Фур'є аналіз спектру, як швидкий, так і звичайний, або їхню комбінацію. Кінцевою метою аналізу є отримання векторів $p_i = (p_0, p_1, \ldots, p_n)$, де p_i є потужністю хвиль для певного піддіапазону. Послідовність таких векторів далі може використовуватися для класифікації.

Порівняно з електромагнітними хвилями, звуковий діапазон є важчим для аналізу, оскільки велика кількість діапазонів може мати сигнатури незалежних сигналів.

Потрібно зробити ремарку, що з практичних міркувань необов'язково аналізувати весь діапазон від 0.1 Гц до 30 кГц. Звукові хвилі починаються з 5 Гц, вібрацію практичніше вимірювати до 50 Гц, у підсумку це залежить від типу вимірювань і встановлення датчиків. Далі дискретне перетворення Фур'є для певного сигналу дає нам лише певне уявлення про сигнал, його характерну потужність у діапазоні, оскільки дискретне перетворення Фур'є вимірюється для значень частоти, кратних частоті дискретизації. Тобто, якщо перетворення Фур'є дає нам амплітуду і фазу для частот $(f_0, f_1, f_2, \ldots, f_{n-1}, f_n)$, то різниця сусідніх частот $(f_i - f_{i-1})$ завжди однакова. Проте у природному середовищі практично ніколи не зустрічаються частоти, які точно співпадають з будь-яким f_i, тому на вихідному графіку амплітуд частот ми бачимо не пікові значення частоти, а дзвоноподібні хвилі (див. рис. 1). Для точного вимірювання частоти сигналу можна використати дробове перетворення Фур'є, яке може визначити амплітуду і фазу для будь-якої частоти. Але з міркувань обчислюваної потужності це неефективно і практично неможливо без попереднього аналізу сигналу. Тому піддіапазони в третину октави 1) добре усереднюють сигнал; 2) розносять гармоніки сигналу у різні піддіапазони. Далі на базі піддіапазонів можливо проводити вторинну обробку сигналу.

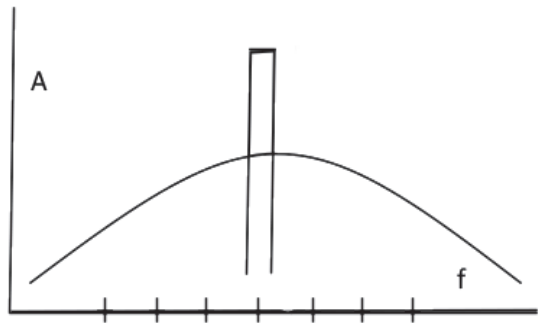

Рис. 1. Дзвоноподібний спектр сигналу у порівнянні з сигналом, частота якого співпадає з однією з частот сітки перетворення Фур'є.

Вторинна обробка сигналу

Для вторинної обробки сигналу з n піддіапазонами використовуємо вектор середніх значень амплітуд для кожного піддіапазону $a^n = (a_1, a_2, \ldots, a_n)$. Фазова інформація відкидається, бо для неї неможливо знайти «середнє» значення. Оскільки сигнал постійно змінюється, весь час спостережень за сигналом має бути розбитий на певні характерні інтервали залежно від часу протікання процесу, за яким ми спостерігаємо. Наприклад, у разі прихованого спостереження за рухом людей або техніки на дорозі можна спостерігати за сигналом і робити усереднення протягом інтервалів 5-7 секунд. У випадку великих приміщень інтервал може бути 15 секунд. Таким чином сигнал від спостережень є послідовність векторів $\{ a^n \}$, яку вже можна класифікувати.

Як було зазначено вище, у нас немає апріорної інформації про спектр сигналів, які спостерігаються, тому ми мусимо зробити певні припущення щодо вигляду класифікованого сигналу. На рис. 2 представлені приклади наборів даних $\{ a^2 \}$ (двовимірні) у поясненнях до пакету класифікації Scikit [5], заради ясності зроблені лише двовимірними. У більшості випадків аналізу розмірність сигналу перевищує 10.

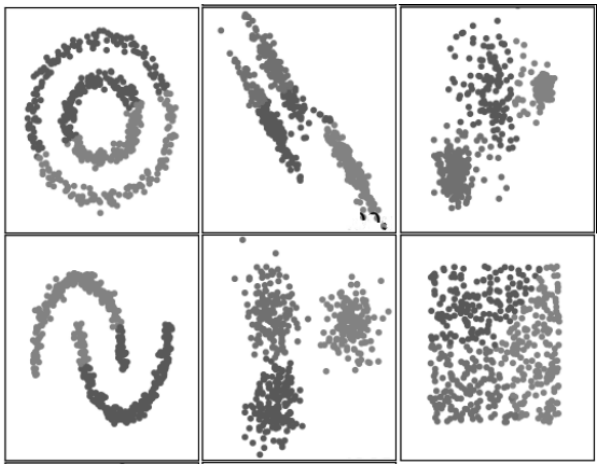

Рис 2. Різні типи наборів даних для кластеризації (приклад з бібліотеки SciKit).

Згідно експериментальних даних, різні об’єкти (легкові автівки, вантажні, важка бронетехніка, пішоходи, тварини середнього розміру, промислове обладнання, побутові прибори) дають (інфра)звукову картину певного сорту, яка локалізована в певних частотах, тобто не дає шумоподібних сигналів або сигналів у широкій смузі частот. Так автор спостерігав сигнал від геофона, спричинений рухом танка Т-64 на швидкості 20 км/год на відстані у 50 м, і це була синусоїда однієї частоти і практично однакової амплітуди, яка на графіках з рис. 2 виглядатиме як крапки, сконцентровані практично в одному місці. У підсумку зазначимо, що отримані набори даних виглядатимуть як правий у верхньому ряду та як середній у нижньому ряду на рис 2, що дозволяє оцінити, які властивості алгоритму класифікації мають значення для подальшого аналізу, а які – ні. Зазначимо, що у первинному сигналі досить багато шуму і набір даних переважно має сигнатури шумів, і тут ми розглядаємо набір даних, вже очищений від шуму, методи «очистки» від шуму розглядаються нижче.

Отож, ми вважаємо, що кількість типів об’єктів, які генерують звукові хвилі, обмежена (автівки, вантажівки, тяжка бронетехніка, пішоходи, велосипедисти, мотоцикли, тварини), також можуть бути інші події (далекі або близькі різкі звуки), які на рис. 2 мали б вигляд хаотично розташованих крапок. Останні мають класифікуватися окремо. Звісно, нейромережа могла б акуратніше класифікувати події, але не завжди є можливість залучити канал зв’язку або обчислювальні потужності чи коректно розробити нейромережу для всіх можливих випадків.

На рис. 3 показана типова картина звукових шумів з одного з випадків розглянутих у [6]. Якщо придивлятися до лівої частини рисунку, ми побачимо на темному тлі більш вертикальні світлі смуги, які позначають відносно більшу спектральну щільність енергії сигналу на певних частотах, а відносно перебігу часу (вертикальна вісь) ми бачимо, що частотна картина зберігається у часі. Далі робиться підсумок енергії по піддіапазонах, отримується вектор амплітуд a^n, як описано вище, і результат передається до класифікатора. Виходячи з наявних і описаних вище обмежень, ми класифікуватимемо звуки за допомогою методів кластеризації, а також опишемо особливості цих методів і далі наведемо приклади кластеризацій.

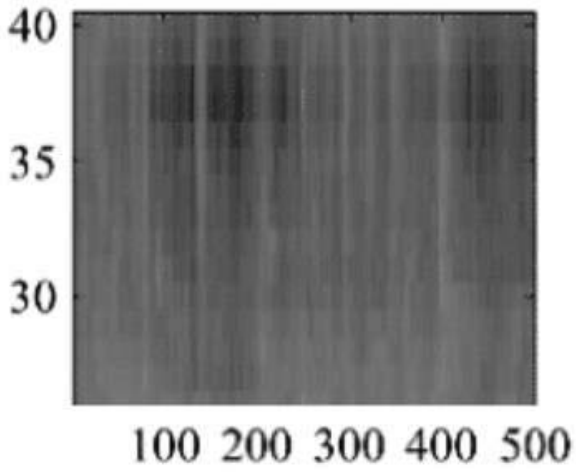

Рис 3. Приклад картини шумів у випадку декількох груп людей, які одночасно розмовляють (горизонтальна вісь – частота, вертикальна – час).

Методи класифікації за допомогою кластеризації

Для подальших експериментів було вирішено використовувати наявну алгоритмічну базу бібліотеки Scikit[5], яка на даний момент має найбільший алгоритмічний доробок методів класифікації.

Одним із найперших і найпростіших є метод K-means [7], який розподіляє точки набору даних на визначену кількість кластерів. Але у більшості випадків метод є дуже простим, має мінімальну параметризацію, тому досить швидко були розроблені покращені методи кластеризації. Розглянемо спочатку можливості параметризації таких методів.

Кількість кластерів. У нашому випадку досить часто виявляється, що ми не знаємо кількість кластерів, тобто кількість типів об'єктів, що будуть розрізнені в наборі даних. Кількість кластерів задається або певним числом, що має наслідок подальшої кластеризації об'єктів на більшу кількість типів, або автоматично. Автоматична кластеризація вимагає більшу обчислювальну потужність, оскільки дуже часто алгоритм вибудовує навколо набору даних потужні допоміжні структури даних. Кількість кластерів може задаватися неявно. Наприклад, мінімальна дистанція між точками одного кластеру (найпростіше), мінімальна кількість точок у кластері, або мінімально відносна щільність даних у кластері відносно загальної щільності даних. Окремо може задаватися метрика аномалій, «зайвих» точок, які не потрапляють у кластери, а вважаються або за визначенням нечастими аномаліями, або похибками вимірів.

Масштабованість. Основними її параметрами є кількість точок у наборі даних та кількість кластерів. Це впливає на швидкодію алгоритму та обсяг пам'яті й може бути фактором, який лімітує використання методу.

Типове використання. Є кластеризація загального призначення, є кластеризації для невеликої (десяток) або великої (сотні) кількості кластерів. Вирізняються методи, які краще працюють у випадку багатовимірних даних, оскільки використовують структури даних для акселерації доступу до масиву точок, як-от KD-дерев [8]. Можливі обмеження на зв'язки між кластерами, різні умови для визначення аномалій, введення ієрархії кластерів (що можливо за використання тих же KD-дерев [8]), задання різної щільності кластерів, індуктивне формування кластерів.

Метрика визначення кластерів. Звичайна евклідова відстань між точками, використання графів сусідів, лімітована сусідами метрика відстані між точками, відстань типу Махаланобіс.

Перед тим, як перейти до найбільш цікавих методів, підсумуємо й зробимо зауваження щодо параметризації. Звісно, методів менше, ніж усіх комбінацій параметрів кластеризації, оскільки виникнення нових методів стимулювалося необхідністю оптимізувати наявні методи для великих наборів даних. Великі обсяги даних вимагають швидкої індексації наявних даних, тому певні методи виникли як результат інтеграції простих методів кластеризації та програмних структур, що є акселераторами доступу до даних, таких як, KD-дерево [8]. Останнє вирішує важливу задачу пошуку сусідів точки у багатовимірному просторі з логарифмічною складністю, а без наявності KD-дерева пошук сусідів стає дуже дорогим за ресурсами завданням. У підсумку – для вибору ефективного алгоритму кластеризації необхідно мати апріорну інформацію щодо вхідних даних. Для звукового ді-

апазону вистачає тижня запису даних і подальшого візуального аналізу.

Окремо зазначимо, що певна кількість алгоритмів має імплементації лише в рамках бібліотек мовою Python, що може викликати погіршення швидкодії від 3 до 10 разів. Є імплементації багатьох алгоритмів від приватних осіб, але, зважаючи на складність внутрішніх структур даних, не можна стверджувати, що такі алгоритми відтестовані. Це накладає обмеження на обсяги даних, які можуть бути оброблені в реальному часі. Зупинимося на декількох алгоритмах.

Найбільш простим алгоритмом є K-means [7]. Оскільки історично він був розроблений раніше за інших, він є найпростішим і має найбільше недоліків. Основним практичним недоліком є те, шо неможливо відразу за апріорними даними встановити необхідну кількість кластерів. У випадку, коли кількості кластерів замало, декілька типів об'єктів можуть бути віднесені до одного кластеру – тобто ми їх не зможемо відрізнити. А якщо кластерів забагато, об'єкти одного типу будуть віднесені до різних кластерів, що спотворює аналіз. Це показано на рис. 7. Можна проаналізувати відстань між отриманими кластерами і, якщо відстань між геометричними центрами деяких кластерів менша за середнє значення, – зменшити кількість кластерів і перезапустити кластеризацію. Але такий метод погано працює у зворотньому напрямку – коли кластерів замало. Ускладнена й евристика, яка дозволяє зрозуміти, чи то дійсно мала геометрична відстань між класами, чи то така особливість набору даних. Ці обмеження привели до розробки більш досконалих методів. Наприклад, mean-shift [9], де параметризувалася необхідна щільність кластерів та мінімальний об'єм простору для формування кластеру. Кластеризація уточнюється ітеративно, водночас змінюється геометричний центр кластеру, тому цей метод ускладнений для паралелизації на відміну від K-means. Взагалі «покращені» методи відрізняються підвищеним споживанням пам'яті та обчислювальних ресурсів.

Покращити кластеризацію намагалися за допомогою введення ієрархії кластерів [10], що може бути корисним, якщо ми знаємо, що набір даних має ієрархічні властивості – тобто різні типи об'єктів можуть генерувати дуже схожі спектри, але це працює лише у незначній кількості випадків.

Можна сказати, що метод DBSCAN [11] має революційні відмінності від [7],[9],[10] оскільки орієнтується на щільність точок набору даних у геометричному просторі і кластеризує дані за принципом знаходження областей простору з найбільшою щільністю даних. Тобто вимагає щоб для елемента даних на певній дистанції ε від нього було не менше s_{min} точок даних (*samples*). Таке визначення дозволяє мати кластери різної форми, не обов'язково опуклі. Якщо певні точки набору даних згідно з обмеженнями s_{min} і ε не можуть бути співставлені до певного кластеру, то вони вважаються outliers або називатимемо їх аномаліями. Зазвичай аномалії розташовані в областях простору, де немає великої щільності даних і з фізичної точки зору, – оскільки ми розглядаємо сигнали – звичайно є перешкодами. На рис. 4 – що взятий із коментарів до SciKit – аномалії зображено невеликими чорними колами, їхній алгоритм не зміг долучитися не до жодного з трьох кластерів.

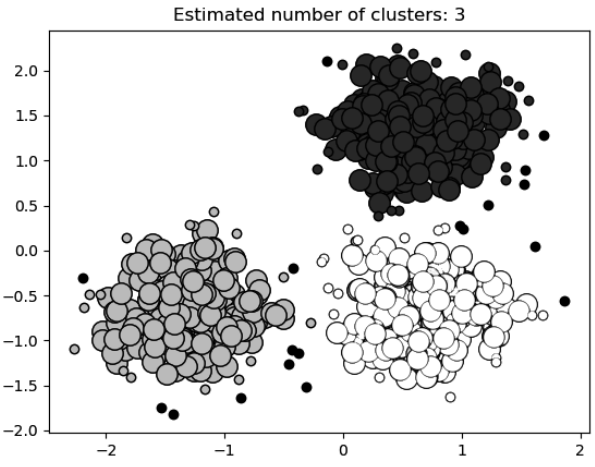

Рис. 4. Приклад кластеризації з аномаліями

Кількість аномалій регулюється вищевказаними параметрами s_{min} і ε, і кількість аномалій може бути регульована від 1% до 5% від загального обсягу даних. Така параметризація важлива у випадках, коли ми намагаємося класифікувати аномалії, і

подальша класифікація залежить від правильного визначення аномалій.

Із недоліків алгоритму відмітимо досить великий обсяг необхідної пам'яті. Розвиває ідеї DBSCAN алгоритм OPTICS [12], який дозволяє мати кластери з різними ε, і додає поняття досяжності, що фактично у порівнянні з DBSCAN дозволяє брати для кластерів не одне значення ε, а цілий інтервал значень $[\varepsilon_{min},\varepsilon_{max}]$. На додачу OPTICS оптимізований щодо обсягу використаної пам'яті.

Останнім із цікавих алгоритмів розглянемо BIRCH [13], який може оперувати даними, що не вміщуються в пам'яті. Кластеризація здійснюється на основі суб-кластерів, до яких намагаються віднести всі значення набору даних, а далі вже комбінувати су-кластери для досягнення необхідної кількості кластерів.

Методи [7],[8],[9],[10],[11],[12],[13] були розроблені для певних задач, для певних обмежень (серед яких і обсяг пам'яті, і швидкодія) і для певних апріорних знань про вхідні дані. Зазначимо, що серед усіх методів кластеризації неможливо виріznити найкращі і найгірші, оскільки всі вони орієнтовані на певні дані.

Якщо аналізувати наші апріорні дані – звуковий або електромагнітний спектр, то найбільш цікавим кандидатом виглядає алгоритм OPTICS [12], який ми далі використаємо для кластеризації.

Збір даних

Найпростішим варіантом збору даних є звичайний мікрофон і звичайний (офісний) комп'ютер. Цього обладнання вистачає для запису звукових сигналів у діапазоні 20Гц – 10кГц, також необхідний звуковий редактор для запису звуку. Дещо більший діапазон частот буде за умови використання напівпрофесійних або скерованих мікрофонів і підсилювача. У випадку вібраційних хвиль можна використати геофони (geophone), які продаються у маркетплейсах електроніки. Необхідним буде також підсилювач електричного сигналу, який може видати сигнал на лінійний вхід звукової карти.

Для підготовки набору даних для подальшої обробки із зроблених записів звукового діапазону можна використати Matlab або Python, оскільки ці середовища програмування мають найбагатші бібліотеки обробки сигналів. Сигнал ділиться на однакові інтервали (наприклад 10 секунд), довжина інтервалу може змінюватися відносно типу об'єкту спостереження. Для ділянок доріг може бути 10 с, для приміщень або складів – 15-20 с. Отримані інтервали сигналу обробляються за допомогою швидкого перетворення Фур'є (FFT), водночас нам достатньо амплітудно-частотної характеристики (АЧХ). Далі отримана АЧХ розділяється на піддіапазони, у кожному піддіапазоні амплітуди усереднюються до одного значення амплітуди. Для кожного інтервалу отримаємо вектор амплітуд $a^n = (a_1, a_2, \ldots, a_n)$, який буде елементом набору даних. Для інтервалу у 10 секунд щодоби матимемо 8640 векторів, і це число впливає на вибір метода кластеризації. Для аналізу добових даних бажано мати набори даних, що перекриваються у часі. У найпростішому випадку можна дробити добовий набір даних на множини A^D_A, A^D_P, де D – умовний номер доби, а A і P позначає першу і другу половини доби. Далі для кластеризації можна формувати набори (A^D_A, A^D_P), (A^D_P, A^{D+1}_A), (A^{D+1}_A, A^{D+1}_P), (A^{D+1}_P, A^{D+2}_A),… У разі наявності обчислювальних потужностей та пам'яті можна формувати набори (A^D, A^{D+1}), (A^{D+1}, A^{D+2}), (A^{D+2}, A^{D+3}), … або навіть довші набори. Далі перейдемо до дослідів із кластеризації.

Результати кластеризації

Для дослідів із кластеризації був обраний метод OPTICS [12], згідно з розмірковуваннями і обчисленнями обсягів набору даних, які подані у статті вище. Використовувалася бібліотека Scikit [5] для імплементацій методу OPTICS, довжина векторів набору була від 10 до 20. Нижче для ілюстративних випадків використана довжина вектору в 10, оскільки більша довжина вектору лише захаращує ілюстративні приклади. Горизонтальна вісь показує середню частоту піддіапазону, вертикальна вісь показує відносну потужність сигналу. Абсолютне зна-

чення потужності сигналу не має жодного значення для методів кластеризації. На рис. 5 показаний результат кластеризації для одного з вібраційних наборів даних.

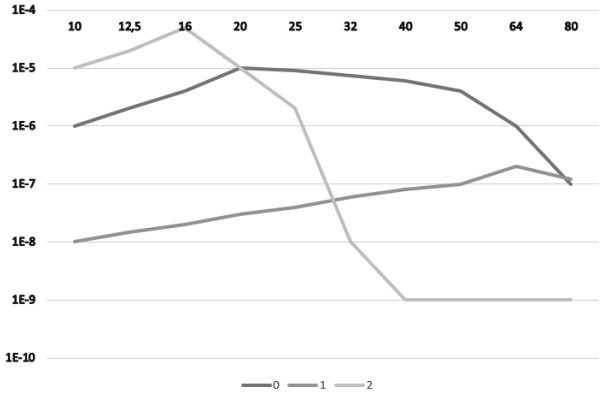

Рис. 5. Приклад коректної кластеризації на три кластери (OPTICS).

На рис. 5 ми бачимо що алгоритм визначив 3 кластери. У першому кластері (0) пік потужності зареєстрований на частоті 20 Гц і досить плавно падає до 50 Гц. Пік другого кластеру (1) припадає на 16 Гц і швидко спадає після 25 Гц. Третій кластер (2) представлений високочастотним шумом. Окремо зазначимо, і це важливо, що на частотну картину накладаються особливості мікрофона або поверхні, на якій стоїть геофон, у різних випадках вимірювань конкретні значення потужності можуть відрізнятися, але якісна картина співвідношення потужностей сигналу у кластерах залишається незмінною. На наступному рисунку 6 показаний ще один приклад роботи алгоритму. Перший кластер (0) має пік на низьких частотах і поступово спадає в області високих частот, другий кластер (1) має пік в області 40 Гц, третій кластер (2) має пік в області 16 Гц.

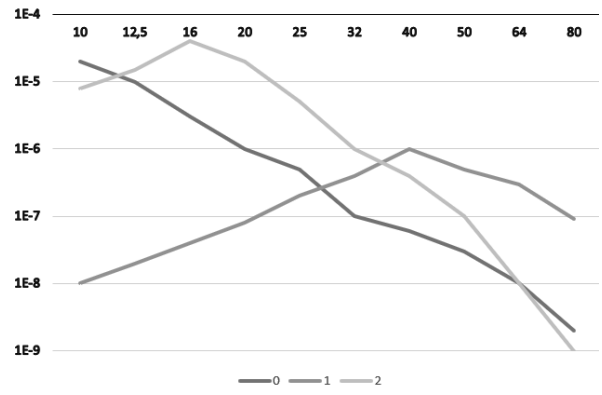

Рис. 6. Інший приклад коректної кластеризації для трьох кластерів.

На рис. 7 показаний один з перших дослідів роботи з простішими алгоритмами на базі K-means. На рисунку показаний приклад некоректної кластеризації.

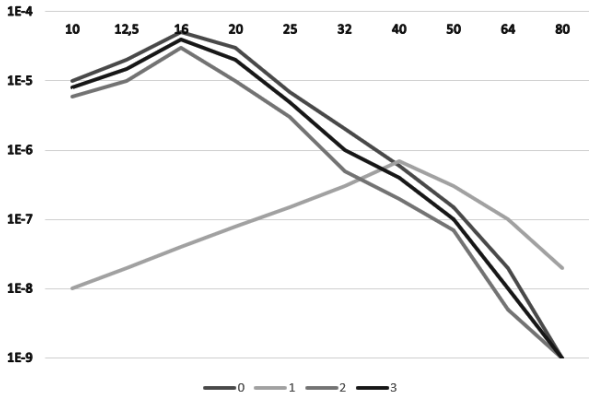

Рис. 7. Приклад некоректної кластеризації під час використання методу K-means.

Методу кластеризації, який орієнтований на наперед задане число кластерів, було задано перебільшену кількість кластерів – 4. Реальна кількість кластерів (визначена *a posteriori*) – 2. Таким чином алгоритм штучно «розтягнув» один із кластерів, у якого відносно більша кількість елементів, на три кластери (0,2,3). У разі, коли на цьому ж наборі даних алгоритмові пропонується розподілити дані на два кластери, лишаються кластери, наприклад, тільки 1 і 3.

Із практичного боку маємо ще декілька спостережень щодо підготовки даних до кластеризації. По-перше, не обов'язково мати діапазони в 1/3 октави, кластеризація добре працює і у випадку ½ октави. Більшість характерних шумів від об'єктів є синусоїдальними сигналами небагатьох частот, від 3 до 5, тому діапазон в 1/3 октави можна вважати оптимальним. Важка техніка може давати синусоїду лине однієї частоти, можливо тому, що сильні вібрації підвіски екранують інші частоти.

Маніпуляції з виділення піддіапазонів «низької», «середньої» та «високої» частоти шляхом простого ділення піддіапазонів на три групи, потім - з кластерізаціями у таких підгрупах, не дають нам переваг перед прямим використанням кластеризації. Такий висновок можна зробити як мінімум з тих засад, що доволі важко пояснити фізичний смисл такого вибору.

Алгоритми DBSCAN [11] та OPTICS [12], і не лише вони, дають нам також елементи набору даних, позначені як «без кластеру», або аномалії. Зазвичай це точки набору даних, які знаходяться далеко від геометричного центру кластеру, максимальна відстань залежить від описаних вище параметрів алгоритмів. Зважаючи на фізичний сенс набору даних, аномалії з більшими амплітудами на високих частотах можуть з великою вірогідністю вважатися шумом, а аномалії на низьких частотах можуть бути додатково розглянуті щодо їхніх джерел.

Каскадовані кластеризації

Для аналізу наборів даних необхідна їхня правильна підготовка, оскільки, залежно від фізичних особливостей середовища (вібрації дороги і вібрації цеха дуже різні), необхідно певним чином виділяти цікаві нам набори даних,. Сучасні алгоритми кластеризації, які можна знайти у [5], пропонують як кластеризацію, так і виділення аномалій. І дуже часто цікаві для аналізу сигнали на фоні первинного набору даних виглядають аномаліями. Тому важливе визначення каскадування кластеризацій, наприклад, на першому етапі виділення змістовних аномалій, а на другому етапі їхня кластеризація. На третьому етапі також можлива кластеризація аномалій, що залишилися після першої кластеризації. Конкретне каскадування дуже сильно залежить від об’єкта, аналіз якого ми проводимо, і конкретні поглиблені схеми каскадування повинні розглядатися виключно для своєї предметної області.

Висновки

У статті розглянуто можливості, найбільш цікаві алгоритми кластеризації і приклади роботи з набором даних для аналізу частотної характеристики сигналів звукового діапазону. Зокрема, розглянуто «живі» приклади аналізу сигналів вибраними методами кластеризації для звукового і вібраційного діапазонів з поясненням результатів. У сучасному світі результати таких аналізів використовуються для спостереження за користуванням об’єктами і для аналізу об’єктів для провадження певних видів промислової діяльності, для яких необхідно регулювати режими шуму і вібрацій.

References

1. F. Meneghello, N. Dal Fabbro, D. Garlisi, I. Tinnirello, M. Rossi. A CSI Dataset for Wireless Human Sensing on 80 MHz Wi-Fi Channels. IEEE Communications Magazine, 2023. https://doi.org/10.48550/arXiv.2305.03170
2. ISO/IEC 7498-1:1994 Information technology — Open Systems Interconnection — Basic Reference Model: The Basic Model. June 1999.
3. Liu, Y., Li, J. (2023). A Survey of Spectrum Sensing Algorithms Based on Machine Learning. // In Proc. Xiong, N., Li, M., Li, K., Xiao, Z., Liao, L., Wang, L. (eds) Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery. ICNC-FSKD 2022. Lecture Notes on Data Engineering and Communications Technologies, vol 153. Springer, Cham. https://doi.org/10.1007/978-3-031-20738-9_97
4. M. Salehi, H. Mirzaei, D. Hendrycks, Y. Li, M. H. Rohban, and M. Sabokrou, “A unified survey on anomaly, novelty, open-set, and out of-distribution detection: Solutions and future challenges,” arXiv preprint arXiv:2110.14051, 2021. https://arxiv.org/pdf/2110.14051.pdf
5. J. Hao, T. Ho Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. // Journal of Educational and Behavioral Statistics. Vol 44, Feb 2019. Doi: 10.3102/1076998619832248.
6. N. Alamdari, N. Kehtarnavaz. A Real-Time Smartphone App for Unsupervised Noise Classification in Realistic Audio Environments. // In Proc. IEEE Intl. Conf. on Consumer Electronics (ICCE), Jan 2019. 1-5. Doi: 10.1109/ICCE.2019.8662052.
7. D. Sculley. Web-scale k-means clustering. // In Proc. of 19th Intl. Conf. on World Wide Web, Apr. 2010. pp. 1177-1178. Doi: 10.1145/1772690.1772862.
8. J.L. Bentley. Multidimensional binary search trees used for associative searching. // Communications of the ACM. (1975) 18 (9): pp. 509–517. doi:10.1145/361002.361007.
9. D. Comaniciu, P. Meer. Mean shift: a robust approach toward feature space analysis // In Proc. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 603-619, May 2002, doi: 10.1109/34.1000236.

10. J.H. Ward, Jr. Hierarchical Grouping to Optimize an Objective Function, // In Journal of the American Statistical Association, 1963, vol 58, pp. 236–244.
11. M. Ester, H. P. Kriegel, J. Sander, X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise // In Proc. of the 2nd Inter. Conf. on Knowledge Discovery and Data Mining, 1996, pp. 226–231
12. M. Ankerst, M.M. Breunig, H.P. Kriegel, J. Sander. OPTICS: ordering points to identify the clustering structure. // In ACM Sigmod Record, 1999, vol. 28, No. 2, pp. 49-60.
13. T. Zhang, R. Ramakrishnan, M. Livny. BIRCH: An efficient data clustering method for large databases. // In ACM Sigmod Record, 1996, vol. 25, issue 2, pp. 103-114.

Одержано: 07.02.2024

Про авторів:

Рагозін Дмитро Васильович,
старший науковий співробітник.
Кількість публікацій в українських виданнях – більше 10.
Кількість зарубіжних публікацій – більше 5.
https://orcid.org/0000-0002-8445-9921

Дорошенко Анатолій Юхимович,
доктор фізико-математичних наук,
професор, завідувач відділу теорії комп'ютерних обчислень,
професор кафедри інформаційних систем та технологій Національного технічного університету України «КПІ імені Ігоря Сікорського».
Кількість наукових публікацій в українських виданнях – понад 200.
Кількість наукових публікацій в зарубіжних виданнях – понад 90.
Індекс Хірша – 7.
http://orcid.org/0000-0002-8435-1451

Місце роботи авторів:

Інститут програмних систем
НАН України,
03187, м. Київ-187,
проспект Академіка Глушкова, 40.
Тел.: (044) 526 3559.
E-mail: dmytro.rahozin@gmail.com,
dmytro.rahozin@ukr.net