DOI: 10.35681/1560-9189.2022.24.2.275079

УДК 62:681.5:004

А. В. Волошко[1], Т. Е. Джеря[2]
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Проспект Перемоги, 37, 03056 Київ, Україна
[1]тел.: (050) 221-01-32, e-mail: avolosko820@gmail.com
[2]тел.: (066) 554-63-42, e-mail: tatyanakurus0202@gmail.com

Метод дерева рішень для ідентифікації і класифікації інформаційних сигналів

Останнім часом часто розглядається питання розвитку інтелектуальних активно-адаптивних електричних мереж в енергетиці. Інтелектуальні електричні мережі мають дуже багато різних аспектів. У статті розглянуто питання коректної класифікації інформаційних сигналів за допомогою методу дерева рішень. Це буде значно пришвидшувати процес віднесення графіка електричного навантаження до певного класу.

Ключові слова: *графік електричного навантаження, інформаційний сигнал, електроспоживання, дерево рішень, вейвлет-перетворення, вейвлет-базис.*

Вступ

Інформаційні потоки є важливим чинником для нормального функціонування сучасних електроенергетичних підприємств. Надмірний чи недостатній обсяг інформації призводить до необхідності створення ефективної системи управління цими потоками. За різними дослідженнями підприємства витрачають від 30 до 80–95 % часу на обробку інформації, тому що наявність актуальної і правдивої інформації, її своєчасне поширення вважається однією із найважливіших передумов забезпечення ефективного управління.

Завдяки оптимізації інформаційних потоків, чого можна досягти завдяки впровадженню розроблюваного методу, відбувається зменшення рівнів ієрархії, надлишку персоналу та спрощення структури організації. В умовах постійних змін і невизначеності організаційну структуру можливо пристосувати для функціонування та зробити її більш гнучкою. Це досягається шляхом збільшення швидкості та надійності передачі інформації, що в свою чергу може надати можливість дослідити та посилити взаємозв'язки між елементами системи, окремими підсистемами, суб'єктами управління по горизонталі та рівнями управління по вертикалі.

Аналізуючи надійність та ефективність функціонування електричних мереж на різноманітних часових і територіальних рівнях виникає необхідність у визначенні інтегральних (імовірнісних) характеристик параметрів режиму. Тобто спочатку потрібен достеменний і повний аналіз сигналів. Під «аналізом» сигналів мають на увазі не лише їхні чисто математичні функціональні перетворення, але й отримані на основі цих перетворень результати та висновки про специфічні особливості відповідних процесів і об'єктів.

Метою аналізу сигналів є:

— визначення та/або оцінка кількісних параметрів сигналів (енергія, середня потужність, середнє квадратичне значення та ін.);

— розкладання сигналів на елементарні складові для отримання всесторонньої, більш повної інформації про сигнал;

— порівняння ступенів «подібності» або «спорідненості» різних сигналів, у тому числі з певними кількісними оцінками [1].

Сигнали можуть бути об'єктами теоретичних досліджень і практичного аналізу лише тоді, коли визначений їхній математичний опис — математична модель. Математичний опис дозволяє абстрагуватися від фізичної природи сигналу та матеріальної форми його носія, здійснювати класифікацію сигналів, виконувати їхнє порівняння, встановлювати ступінь тотожності, моделювати системи обробки сигналів [2].

Метод ідентифікації і класифікації інформаційних сигналів на основі оптимальної структури бінарного дерева вейвлет-перетворення

Одним із ефективних методів автоматичного аналізу масиву інформаційних даних є побудова, так званого, дерева рішень. Дерево рішень — це графічний метод, що дозволяє пов'язати точки прийняття рішення, можливі стратегії і їхні наслідки з можливими факторами, умовами зовнішнього середовища, який застосовується в умовах ризику.

Прийняття рішень за рахунок даного методу умовно можна розділити на п'ять кроків. Перший крок — формулювання завдання: визначається можливість збору достовірної інформації для експерименту та реальних дій; визначаються події, котрі можуть виникнути за даних умов; установлення порядку виконання робіт (подій), у наслідках яких міститься корисна та доступна інформація. Другий крок — побудова «дерева рішень». Третій крок — оцінка стану середовища, тобто визначення ймовірності виникнення кожної події (за статистичними даними або експертним методом). Четвертий крок — установлення виграшів (чи програшів як виграшів зі знаком мінус) для кожної можливої комбінації альтернатив (дій) і станів середовища. П'ятий (останній) крок — вирішення задачі (прийняття рішень) [3].

Задачі, які розв'язуються за допомогою дерева рішень можуть бути об'єднані в основні три класи:

1) опис даних: застосування «дерева рішень» дозволяє зберігати інформацію про вибірку даних у компактній і зручній для обробки формі, що містить у собі точні описи об'єктів;

2) класифікація: застосування «дерева рішень» дозволяє справитися із завданнями класифікації, тобто відношення об'єктів до одного із описаних класів;

3) регресія: якщо змінна має недостовірні значення, то застосування «дерева рішень» дозволяє визначити залежність цієї цільової змінної від незалежних (вхідних) змінних.

Із великої кількості алгоритмів, які реалізують дерева рішень, найбільш поширеними є наступні два:

1) CART (Classification and Regression Tree) — це алгоритм побудови бінарного дерева рішень, тобто дихотомічної класифікаційної моделі. Кожен окремий вузол дерева при розбитті має тільки двох нащадків (два листки). За допомогою даного алгоритму вирішуються задачі класифікації і регресії;

2) C 4.5 — алгоритм побудови дерева рішень, кількість нащадків у вузлі якого при цьому необмежена. Використовується тільки для вирішення задач класифікації.

Алгоритм та опис методу дерева рішень

Процес побудови дерев рішень полягає в послідовному, рекурсивному розбитті навчальної великої кількості на підмножини із застосуванням вирішальних правил у вузлах. Процес розбиття триває до тих пір, поки усі вузли в кінці усіх гілок не будуть оголошені листям. Оголошення вузла листом може статися природним чином (коли він міститиме єдиний об'єкт, або об'єкти тільки одного класу), або після досягнення деякої умови зупинки, що задається користувачем (наприклад, мінімальне допустиме число прикладів у вузлі або максимальна глибина дерева).

У результаті деяких робіт за алгоритмом Малла послідовна двополосна фільтрація вхідного сигналу проводиться тільки для низькочастотної області, виходячи із припущення, що дана область містить основну частину інформації. У результаті отримується «однобоке» дерево рішень, що для багатьох випадків обробки інформації є припустимим. Але попередні дослідження свідчать, що використання їхніх результатів за алгоритмом Малла не забезпечують вирішення питань класифікації для випадків наявності в інформаційних сигналах шумових компонент.

У зв'язку з цим надалі розглядається питання застосування пакетних вейвлетів для операцій послідовного частотного перетворення як низькочастотних, так і високочастотних коефіцієнтів з метою одержання збалансованого дерева вейвлет-перетворення — бінарного дерева [4].

При застосуванні пакетних вейвлетів вхідний сигнал описується трьома параметрами: позиції та масштабу (як для звичайного вейвлет-перетворення), а також частотою. Як відомо, таке пакетне вейвлет-перетворення є адаптивним. Ця адаптація не потребує навчання або відомостей про статистичні властивості сигналів і дозволяє більш точно враховувати особливості сигналу, що аналізується, шляхом вибору відповідної оптимальної форми дерева розкладу. В процесі дослідження це дозволило забезпечити мінімальну кількість вейвлет-коефіцієнтів при заданій точності стискання/відновлення сигналів, їхньої класифікації [5].

Дослідження проводилося в плані розробки методу класифікації інформаційних сигналів за допомогою побудови оптимального дерева тому що, по-перше,

бінарні дерева доволі просто можуть бути представлені у вигляді списків, або масивів. При цьому кожний елемент дерева має поле даних і два поля покажчиків. Один покажчик зв'язує елемент з правим нащадком, а другий — з лівим (рис. 1).

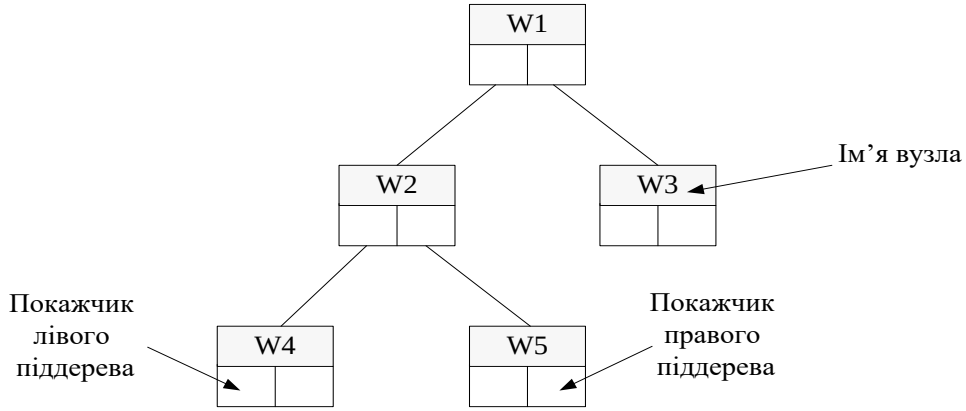

Рис. 1. Представлення бінарного дерева у вигляді спискової структури

Дослідження показали, що представлення бінарного дерева у вигляді масиву є найбільш ефективним. А саме, бінарне дерево завжди має строго означене число вершин на кожному рівні, які нумеруються зліва направо послідовно за рівнями та використовуються як індекси в одномірних масивах (рис. 2).

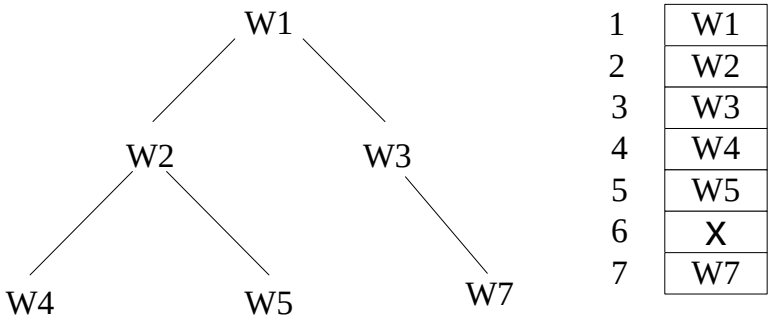

Рис. 2. Представлення бінарного дерева у вигляді масиву

Також, у зв'язку з тим, що в процесі обробки число рівнів дерева суттєво не змінюється, такий спосіб представлення бінарного дерева є значно більш економічним, на відміну від стекової структури. При цьому, адреса будь-якої вершини в одномірному масиві обчислюється, як

$$agp = (2k - 1) + (i - 1),$$

де k — номер рівня вершини; i — номер на рівні k в повному бінарному дереві.

Аналіз досліджень свідчить, що головним недоліком даного способу представлення бінарного дерева є те, що структура даних є статичною. Розмір масиву обирається, виходячи із максимально можливої кількості рівнів бінарного дерева. Тобто чим менш повним є дерево, тим менш раціонально використовується об'єм пам'яті.

Побудова оптимальної структури бінарного дерева пакетного вейвлет-перетворення за ентропією Шенона

При проведенні зворотної вейвлет-декомпозиції (реконструкція інформаційного сигналу) за допомогою побудови оптимального дерева декомпозиції незначні інформаційно-надлишкові чи непотрібні деталі сигналів виключаються. При цьому мірою оптимальності є кількість апроксимуючих і деталізуючих коефіцієнтів для реконструкції сигналу в межах заданої похибки.

Слід звертати особливу увагу на вибір критерію атрибута та відсікання гілок при побудові дерева рішень. Вибраний атрибут має розділяти множину так, щоб у підсумку отримані підмножники складалися із об'єктів, які відносяться до одного класу або ж є до нього максимально наближеними.

Як функцію вартості інформативності набору вейвлет-коефіцієнтів використано ентропію Шенона. Дана функція буде великою, якщо вейвлет-коефіцієнти приблизно однієї величини, і невеликою, якщо всі вейвлет-коефіцієнти, окрім декількох, близькі до нуля. Під ентропією розуміємо величину

$$E = \exp\left(-\sum\nolimits_{K=1}^{N} P_K \log\left(P_K\right)\right),$$

де x — вхідний сигнал.

Побудову оптимального дерева рішень (у вузлах дерева пакетного вейвлет-перетворення представлені значення ентропії) проведено шляхом визначення ентропії вузлів і його апроксимуючих і деталізуючих коефіцієнтів — нащадків. Якщо ентропія вузла є більшою за ентропію нащадків, подальша декомпозиція в даному вузлі закінчується, і дерево «обрізається». Алгоритм декомпозиції рекурсивно продовжується до досягнення глибини декомпозиції. На рис. 3 та рис. 4 представлено повне бінарне дерево графіків електричного навантаження (ГЕН) і оптимальне дерево декомпозиції ГЕН відповідно.

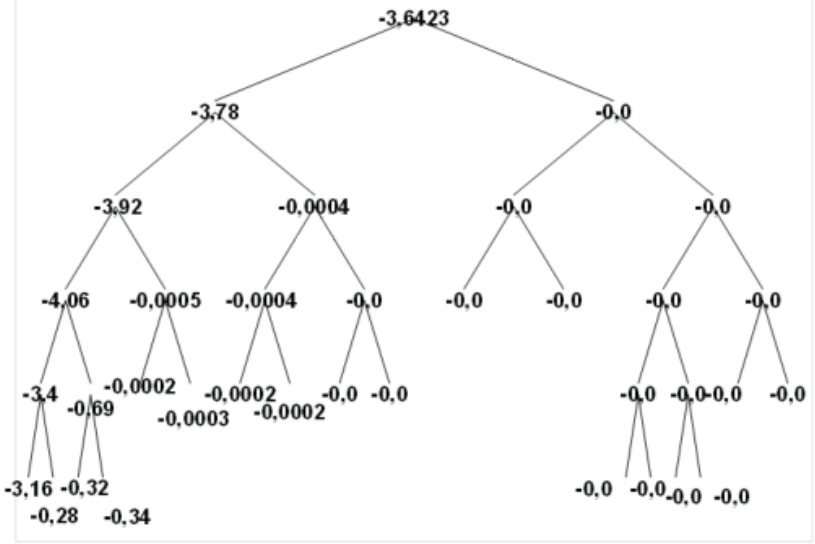

Рис. 3. Повне бінарне дерево пакетного вейвлет-перетворення

Як випливає з рис. 3, значення ентропії у вузлі [2 2], [3 1], [3 2], [3 3], [3 7], [4 2], [4 3], [4 4], [4 5], [4 6], [4 7], [4 14], [4 15], [5 2], [5 3] більше ніж у їхніх на-

щадків, Тобто у даних вузлах дерево декомпозиції повинно бути «обрізане», що і показано на рис. 4.

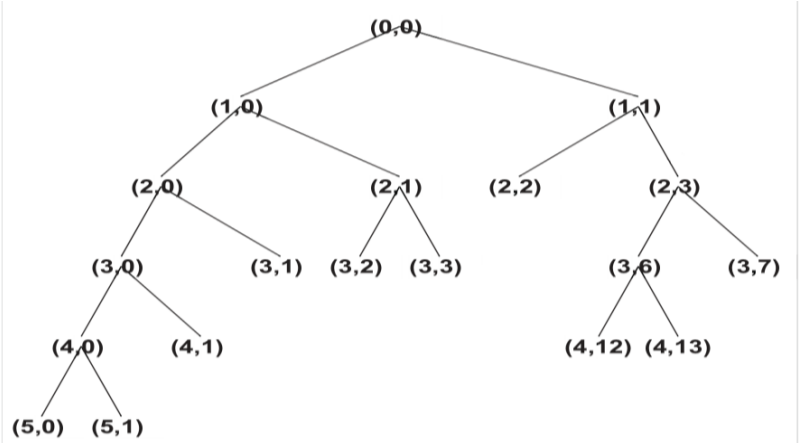

Рис. 4. Оптимальне дерево вейвлет-декомпозиції (у вузлах указані значення ентропії)

Для побудови вектора ознак класифікації графіків електричного навантаження на основі оптимального дерева декомпозиції було проаналізовано добові графіки електричного навантаження за період січень-грудень 2021 року (365 добових, погодинних ГЕН). За допомогою розробленого методу кластеризації з урахуванням адаптивного коефіцієнта варіації відносно всієї вибірки значень виділено дванадцять класів. Кожен клас містить від 10 до 15 вибіркових значень ГЕН, чотири класи при цьому було виділено як основні — вони містять у собі по 20–25 ГЕН. У межах одного класу структури оптимальних дерев декомпозиції є однаковими. Порівняння проводилося за допомогою послідовних алгоритмів та алгоритму Вітербі. Види оптимальних дерев трьох основних класів представлено на рис. 5.

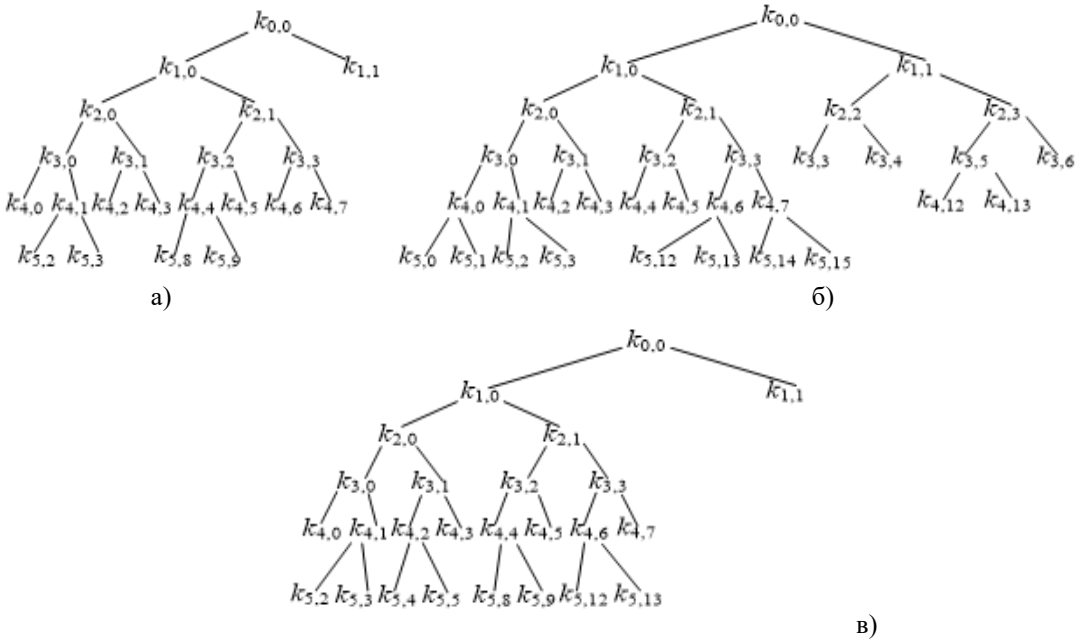

Рис. 5. Оптимальні дерева: а) вихідні; б) святкові; в) робочі дні

Більш ефективним алгоритмом віднесення ГЕН до відповідного класу виявився стековий алгоритм. Алгоритм віднесення ГЕН до одного із трьох класів представлено на рис. 6.

Згідно з рис. 6, на якому формується алгоритм віднесення ГЕН до одного із класів, порогові значення $\delta 1$, $\delta 2$, $\delta 3$ адаптуються до технологічного процесу та характеру виробництва на стадії попередньої статистичної обробки реальних графіків. Проведений порівняльний аналіз даного методу кластеризації з відомими показав наявність похибки 1-го роду в межах ~8 %, та 2-го роду в межах ~7,5 %. Економія пам'яті для зберігання значень ГЕН знаходиться в межах 35 %, швидкість класифікації і пошуку типових графіків у базі даних збільшується в 1,6 рази.

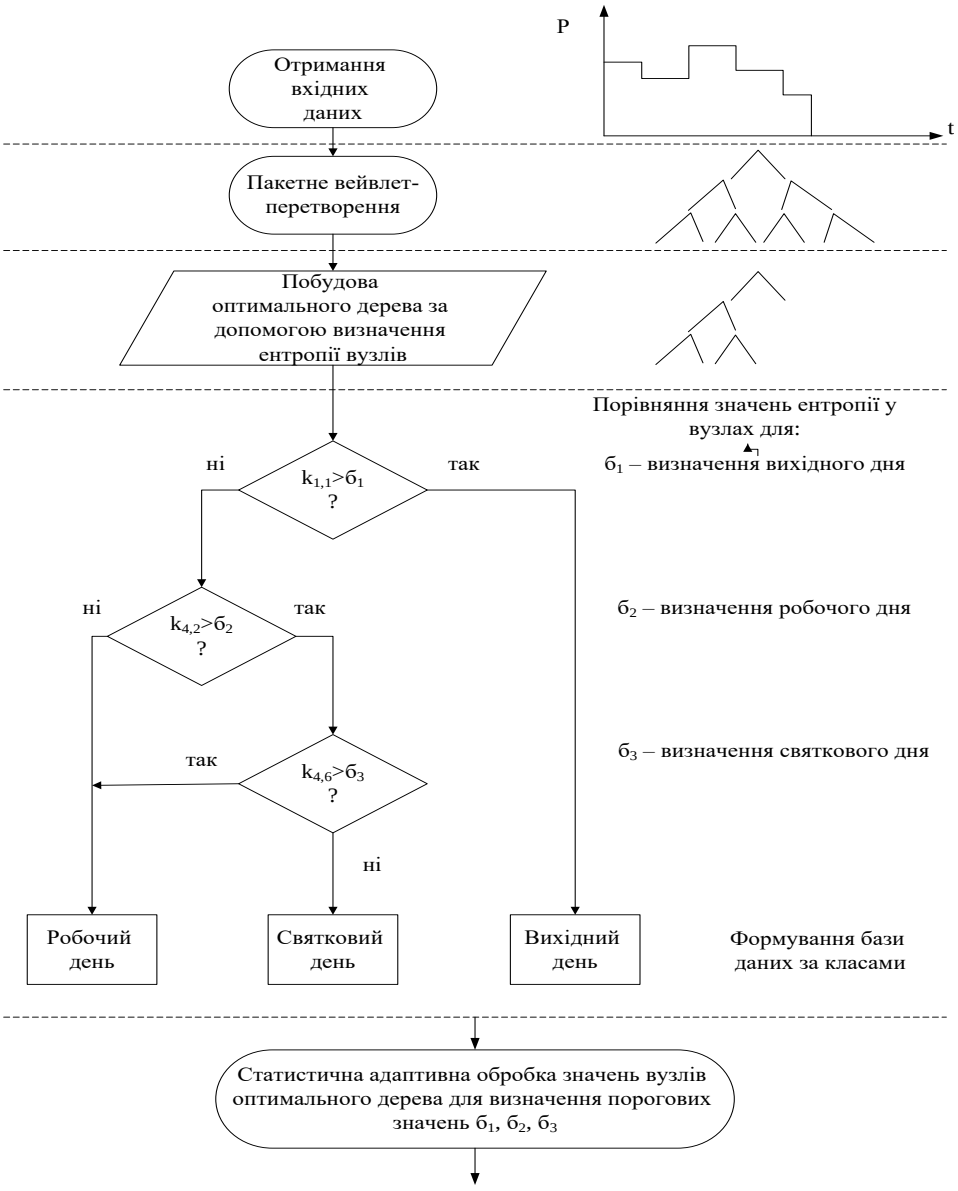

Рис. 6. Алгоритм віднесення ГЕН до одного із класів

Також у роботі проведено аналіз точності реконструкції ГЕН із оптимальних дерев бінарного дерева. На рис. 7 представлено значення абсолютних похибок реконструкції ГЕН із застосуванням різних базисів декомпозиції/відновлення.

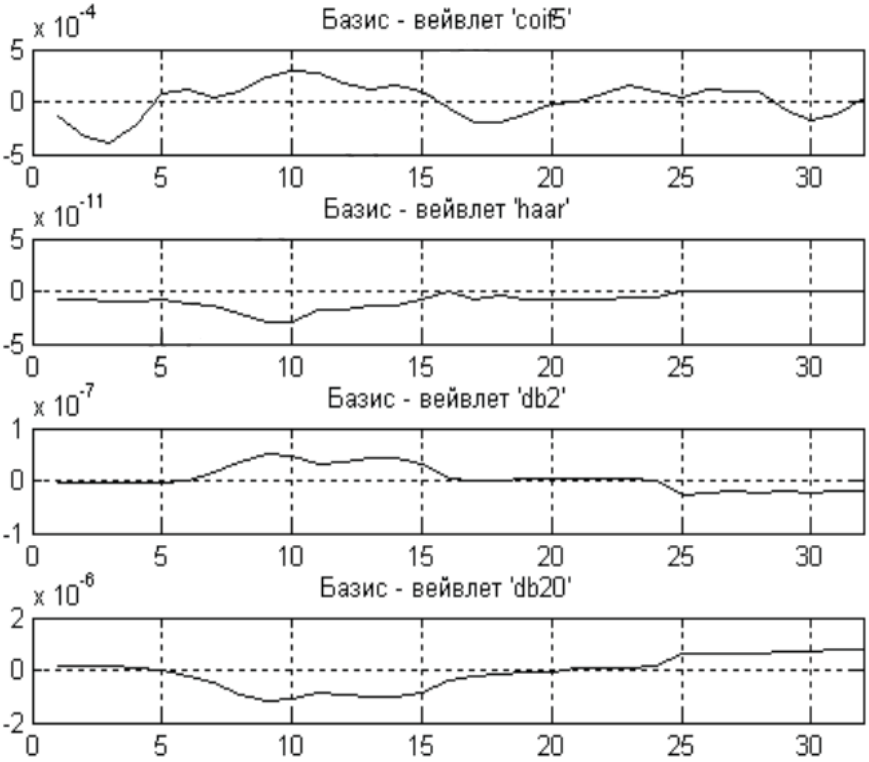

Рис. 7. Похибка відновлення ГЕН із застосуванням різноманітних вейвлет-базисів

Як випливає з даних на рис. 7, побудова оптимального дерева вейвлет-перетворення з урахуванням функції вартості на основі ентропії («обрізання» гілок повного бінарного дерева) забезпечує досить незначну похибку при його реконструкції (значення похибки значно менше похибки квантування). При цьому, виходячи з властивостей ГЕН, найбільш прийнятним базисом вейвлет-декомпозиції є вейвлет Хаара — похибка складає $\pm 5*10^{-11}$. Це пояснюється тим, що вейвлет-пакети, що створені за допомогою вейвлетів Хаара — розривні функції, а фільтр Хаара ($h[n]$) добре локалізований у часі, але не за частотою.

Висновки

Перевагою методу дерева рішень серед інших є інтуїтивність. Класифікаційна модель, представлена у вигляді дерева рішень, є інтуїтивною та спрощує розуміння розв'язуваної задачі. Результат роботи алгоритмів конструювання дерев рішень, на відміну, наприклад, від нейронних мереж, що представляють собою «чорні ящики», легко інтерпретується користувачем. Ця властивість дерев рішень не тільки важлива при віднесенні до певного класу нового об'єкта, але й корисна при інтерпретації моделі класифікації у цілому. Метод дерева рішень дoз-

воляє зрозуміти та пояснити, чому конкретний об’єкт відноситься до того або іншого класу.

Багато статистичних методів є параметричними, і користувач повинен заздалегідь володіти певною інформацією, наприклад, знати вид моделі, мати гіпотезу про вид залежності між змінними, припускати, який вид розподілу мають дані. Методи дерева рішень, на відміну від таких методів, будують непараметричні моделі. Отже, методи дерева рішень здатні розв’язувати такі задачі Data Mining, в яких відсутня апріорна інформація про вид залежності між досліджуваними даними.

Аналізуючи отримані результати, можна зробити висновки, що застосування оптимального дерева рішень як вектора ознак класифікації забезпечує суттєву економію пам’яті для зберігання інформаційних даних і збільшує швидкість доступу до баз даних, однак класифікація сигналів здійснюється зі значними похибками. Тому в подальшому основна увага приділяється підвищенню точності проведення класифікації при застосуванні оптимального дерева рішень як вектора ознак.

При проведенні дослідження програмного забезпечення енергетичних підприємств України було встановлено, що більшість систем управління виробництвом не відповідає сучасним вимогам, або потребує модернізації, також підприємства не можуть самостійно вирішувати задачі функціонування і організації якісного процесу передачі даних. Тому можна зробити висновок, що інтеграція більш швидкого та надійного методу класифікації інформаційних сигналів буде привабливою та доцільною для покращення роботи інформаційних систем підприємств компаній.

1. Жураковський Ю.П., Полторак В.П. Теорія інформації та кодування: підручник. Київ: ВШ, 2001. 255 с.

2. Лайонс Р. Цифровая обработка сигналов: Второе издание. Москва: ООО «БиномПресс», 2006. 656 с.

3. Дерево ухвалення рішень. URL: https://uk.wikipedia.org/wiki/Дерево_ухвалення_рішень

4. Волошко А.В., Лутчин Т.М., Терещенко Д.Ю. Метод класифікації інформаційних сигналів за допомогою побудови оптимального дерева вейвлет-перетворення. *Енергетика, екологія. Людина*. 2012. С. 81–85.

5. Гайдур Г.І., Прилєпов Є.В., Попов М.І. Принципи побудови дерева рішень на основі класифікаційного алгоритму С4.5. *Наукові записки Українського науково-дослідного інституту зв'язку*. 2018. № 1. С. 60–64.

Надійшла до редакції 30.11.2022