

Гавор М.І.

РОЗРОБКА СПОСОБУ КЛАСИФІКАЦІЇ ІНФОРМАЦІЙНИХ СИГНАЛІВ ШЛЯХОМ ПОБУДОВИ ОПТИМАЛЬНОГО ДЕРЕВА РІШЕНЬ

Метою статті було відобразити мій план по здійсненню та написанню магістерської дисертації, а також вказати напрям подальшого дослідження за даною темою.

В ході виконання статті було розглянуто один із ефективних методів автоматичного аналізу масиву інформаційних даних на основі побудови дерева рішень бінарних вейвлет-перетвореннях. Були вказані задачі які розв'язуються за допомогою дерева рішень, визначена структура дерева пакетного вейвлет-перетворення, а також розглянута побудова оптимальної структури бінарного дерева пакетного вейвлет-перетворення за ентропією Шенона.

Ключові слова: *дерево рішень, вейвлет, вейвлет-перетворення, інформаційний масив даних, вейвлет-коефіцієнти, вейвлет-декомпозиція, ентропія Шенона.*

The purpose of the article was to reflect my plan for the implementation and writing of a master's thesis, as well as to indicate the direction of further research on this topic.

In the course of the article, one of the effective methods of automatic analysis of an array of information data based on the construction of a decision tree of binary wavelet transforms was considered. The problems to be solved with the help of the decision tree were specified, the structure of the batch wavelet transform tree was determined, and the construction of the optimal structure of the binary tree of the batch wavelet transform according to Shannon entropy was considered.

Вступ

Метою роботи є дослідження ефективних методів автоматичного аналізу масиву інформаційних даних є побудова, так званого дерева рішень.

В даній роботі мною буде розглянуто класифікація інформаційних сигналів на основі оптимальної структури бінарного дерева вейвлет-перетворення.

Основна перевага даного способу полягає в більш точній і надійній класифікації для випадків наявності у інформаційних сигналах шумових компонент.

Актуальність теми

Як показує аналіз наукових публікацій, існуючі методи ідентифікації та класифікації інформаційних сигналів або забезпечують задовільну точність, за умови досить складних алгоритмів їх використання, або на рівні зі швидкодією та простотою алгоритмів навпаки не дозволяють досягти заданої точності. Тобто, вони можуть застосовуватись для вирішення тих чи інших вибіркових задач, але забезпечення оптимального співвідношення між точністю класифікації, швидкодією та простотою алгоритмів залишається досить складною проблемою.

Тому у даній роботі представлений метод класифікації інформаційних сигналів на основі використання бінарного дерева рішень пакетних вейвлет-перетворень.

Задачі які розв'язуються за допомогою дерева рішень

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

- опис даних – інформація зберігається у компактній формі дерева рішень, яке утримує точний опис об'єктів;
- класифікація – віднесення об'єктів до одного із раніше визначених класів, при цьому цільова змінна повинна мати дискретні значення;
- регресія – встановлення залежності цільової змінної, яка має неперервне значення, від вхідних змінних за допомогою дерева рішень.

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

- CART (Classification and Regression Tree) – це алгоритм побудови бінарного дерева рішень – дихотомічної класифікаційної моделі. Кожний вузол дерева при розбитті має тільки двох нащадків (листіків). За допомогою даного алгоритму вирішуються задачі класифікації та регресії;

- С 4.5 – алгоритм побудови дерева рішень кількість нащадків у вузлі якого при цьому необмежена. Використовується тільки для вирішення задач класифікації.

Структури дерева пакетного вейвлет-перетворення

При застосуванні пакетних вейвлетів вхідний сигнал описується трьома параметрами: позиції та масштабу, а також частотою. Фактично, пакетне вейвлет-перетворення являється адаптивним. Ця адаптація не потребує відомостей про статистичні властивості сигналів і дозволяє більш точно враховувати особливості аналізованого сигналу шляхом вибору відповідної оптимальної форми дерева розкладу. Це дозволяє забезпечити мінімальну кількість вейвлет-коефіцієнтів при їх класифікації.

Для дослідження планується розробка методу ідентифікації та класифікації ГЕН за допомогою побудови оптимального дерева. Як відомо, бінарні дерева доволі просто можуть бути представлені у вигляді списків або масивів. При цьому кожний елемент дерева має поле даних і два поля показників. Один показник зв'язує елемент з правим нащадком, а другий – з лівим (рис. 1).

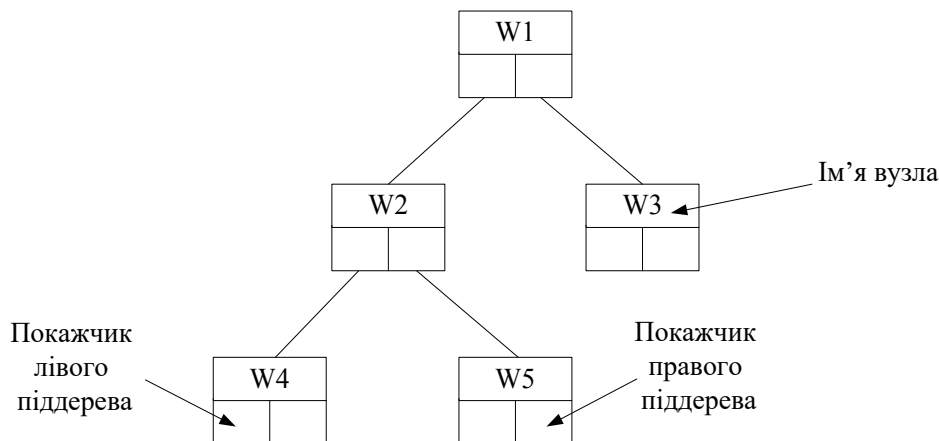


Рисунок 1– Представлення бінарного дерева у вигляді спискової структури

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 і використовуються в якості індексів у одновірних масивах.

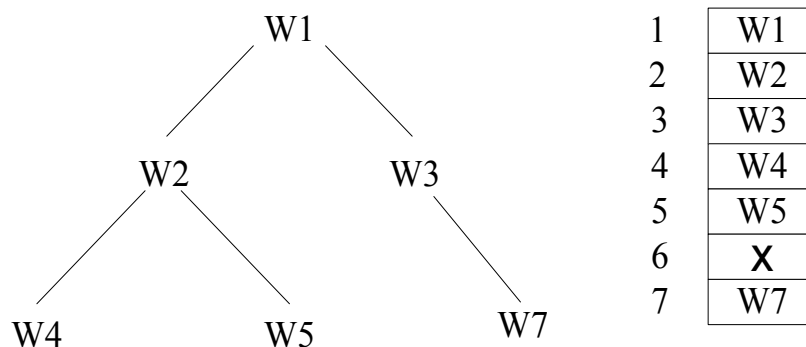


Рисунок 2 – Представлення бінарного дерева у вигляді масиву

Також, в 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.

Аналіз наукових досліджень свідчить, що головним недоліком даного способу представлення бінарного дерева являється 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

За допомогою побудови оптимального дерева при проведенні зворотної вейвлет-декомпозиції (реконструкція інформаційного сигналу) незначні, інформаційно-надлишкові, або зовсім непотрібні деталі сигналів виключаються. Мірою оптимальності при цьому являється кількість апроксимуючих і деталізуючих коефіцієнтів для реконструкції сигналу у межах заданої похибки.

При побудові дерева рішень особливу увагу слід приділяти питанням вибору критерію атрибута та відсікання віток. Обраний атрибут повинен розбивати множину так, щоб отримані в підсумку підмножини склались із об'єктів, які належать до одного класу, або є максимально наближеними до нього.

В якості функції вартості інформативності набору вейвлет-коефіцієнтів використовується ентропія Шенона. Дана функція буде великою, якщо вейвлет-коефіцієнти приблизно однієї величини, і невеликою, якщо всі вейвлет-коефіцієнти, окрім декількох, близькі до нуля. Під ентропією Шенона розуміється величина

$$E = \exp\left(-\sum_{k=1}^N P_k \log P_k\right),$$

де $P_k = |x_k|^2 / \|x\|^2$, x – вхідний сигнал, вейвлет-коефіцієнти якого $\{x_k\}_{1 \leq k \leq N}$ зустрічаються з ймовірністю $\{P_k\}_{1 \leq k \leq N}$.

Побудову оптимального дерева рішень 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.

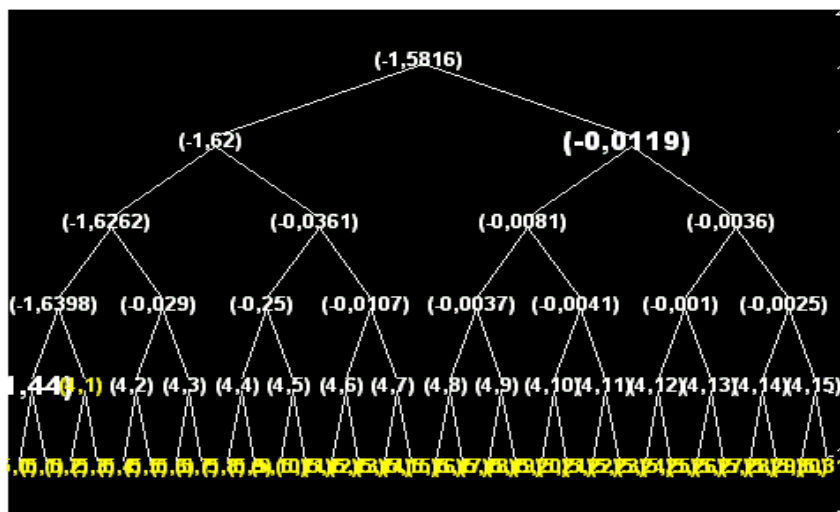


Рисунок 3 – Повне бінарне дерево пакетного вейвлет-перетворення

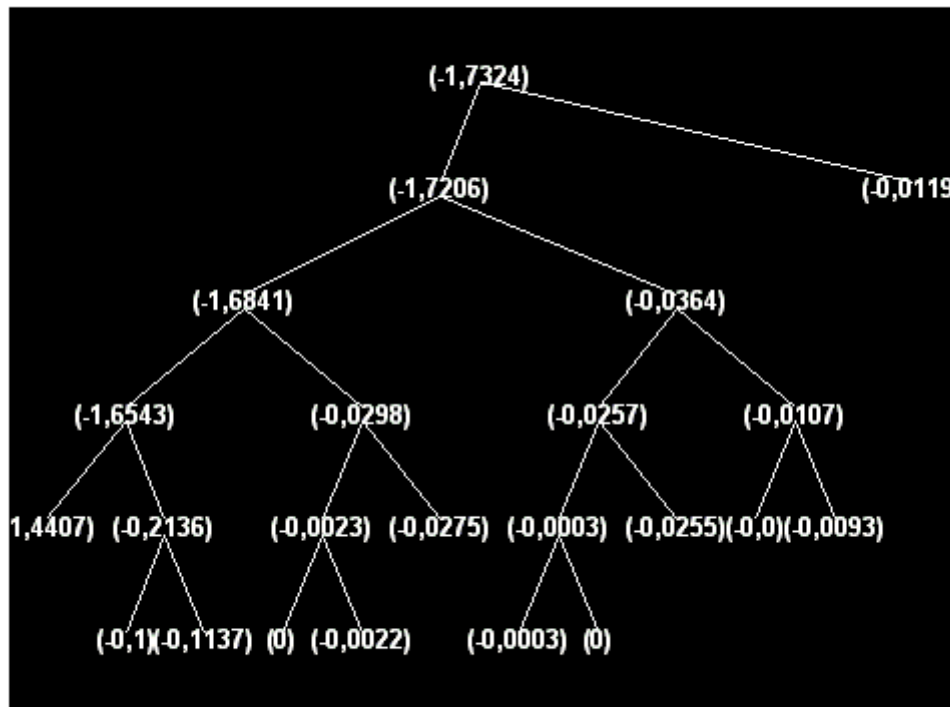


Рисунок 4 – Оптимальне дерево вейвлет-декомпозиції (у вузлах представлені значення ентропії)

Висновок

Підсумовуючи, необхідно відмітити, що застосування в якості вектору ознак класифікації оптимального дерева рішень зазвичай забезпечує суттєву економію пам'яті для зберігання інформаційних даних та збільшує швидкість доступу до баз даних, однак класифікація сигналів в певних випадках може мати значні похибки, тому необхідно проводити аналіз точності, для усіх можливих оптимальних дерев, і в таких випадках потрібно приділяти увагу підвищенню точності проведення класифікації при застосуванні у якості вектору ознак оптимального дерева рішень.

Перелік використаних джерел

1. Орлов А.И. Статистические методы и анализ данных// Социология. – 1991.- № 2. - С. 28 – 49.
2. А.В. Гармаш Лекция 13. Методы классификации и идентификации. Распознавание образов. - <http://chemstat.com.ru/node/13/>
3. Дж. Вэн Райзин. Классификация и кластер. Перевод с английского П.П.Кольцова под редакцией Ю.И.Журавлева// Издательство. – М.:Мир, 1980. – 389 с.
4. Daniuseviciute L. Wavelet-based Entropy Analysis of Electromyography during 100 Jumps/L. Daniuseviciute, K. Pukenas, A. Skurvydas // Electronics and Electrical Engineering. – 2010. – № 8 (104). – Р. 93 – 96.
5. Шеннон К. Работы по теории информации и кибернетике / Изд-во иностр. Литературы. – Москва. – 1963. – 832 с.
6. Волошко А.В. Метод класифікації інформаційних сигналів за допомогою побудови оптимального дерева вейвлет-перетворення/А.В. Волошко, Т.М. Лутчин, Д.Ю. Терещенко// Енергетика, екологія. Людина. - 2012. С. 81 – 85.