Список літератури

1. *Берлекэмп Э.* Алгебраическая теория кодирования. – М.: Мир, 1971. – 450 с.
2. *Мак-Вильямс Ф., Слоэн Н.* Теория кодов, исправляющих ошибки. – М.: Связь, 1979. – 370 с.
3. *Shannon C.E.* A mathematical theory of communication // Bell Syst. Tech. J. 1948. – Vol. 27. – P. 379–423, 623–656.
4. *Шенон К.Е.* Работы по теории информации и кибернетики. – М.: ИЛ., 1963. – 829 с.
5. *Гитис Э.И.* Преобразователи информации для электронных цифровых вычислительных устройств. – М.: Энергия, 1970. – 400 с.
6. *Кодирование* информации (двоичные коды) / Н.Т. Березюк, А.Г. Андрущенко, С.С. Мошицкий и др. – Харьков: Вища шк., 1978. – 252 с.
7. *Хэмминг Р.В.* Теория кодирования и теория информации. – М.: Радио и связь, 1983. – 176 с.
8. *Блейхут Р.* Теория и практика кодов, контролирующих ошибки. – М.: Мир, 1986. – 576 с.
9. *Білецький А.Я., Білецький О.А.* Синтез кодів Грея // Вісн. НАУ. – 2002.– № 1. – С. 29–34.

Стаття надійшла до редакції 10.04.02.

УДК 007.5;681.3;681.518.3

Л.О. Жук, канд. техн. наук, проф.,
В.Ф. Сураєв, канд. техн. наук, доц.

СИСТЕМАТИЗАЦІЯ ТА ОЦІНКА МЕТОДІВ ОБРОБКИ ДАНИХ ДЛЯ ВИЗНАЧЕННЯ СТАНУ ТА СТРУКТУРИ СКЛАДНИХ ОБ'ЄКТІВ

Систематизовано методи обробки спектрометричних даних, обґрунтовано ефективні комп'ютерні процедури для швидкого експрес-аналізу стану чи структури складних об'єктів при проведенні масштабних експериментів, забезпеченні технологічних процесів чи наукових досліджень.

Обробка спектрометричних даних лежить в основі визначення стану та структури об'єктів різної природи: картографування поверхневого шару ґрунтів регіонів чи держав, виявлення корисних копалин, оцінка стану водоймищ, навколишнього середовища, посівів, аналіз паливно-мастильних матеріалів та ін.

Спектрограма поглинання та віддзеркалення (відбиття) для окремої складової об'єкта в чистому вигляді має своє унікальне відображення з певною кількістю піків (смуг) певної форми і висоти на певних довжинах хвиль λ, а спектрограма об'єкта з кількох компонент є суперпозицією складових та шуму. Обробка відповідних цифрових відліків має на меті виявлення піків та їх параметрів, а за цими даними можна робити висновки про наявність тих чи інших компонентів в об'єкті (якісний аналіз), а в деяких випадках і їх відносну кількість (кількісний аналіз).

В ідеальному випадку без шумів спектрограми можуть мати невиразну картину та навіть зміщені піки, як наслідок накладення складових. У реальному випадку шумовий фактор може бути дуже сильним, особливо при дистанційному одержанні спектрограм (із супутників чи інших літальних апаратів), і меншим для стаціонарних умов (наукові дослідження в лабораторії, технологічні процеси на підприємствах та ін.). Всі ці фактори потрібно враховувати при обробці даних.

Відомі методи обробки доцільно розбити на дві групи: розклад спектрограм на складові за допомогою моделювання та використання процедур та критеріїв теорії розпізнання образів.

Розглянемо першу групу. Загально прийнято окремі смуги моделювати кривими Гауса $\Psi_г(\lambda)$ або Лоренца $\Psi_л(\lambda)$:

$$\Psi_г(\lambda) = ye^{\frac{-5{,}545(\lambda-\lambda_0)^2}{2w^2}};$$

$$\Psi_л(\lambda) = \frac{y}{1 + \frac{4(\lambda - \lambda_0)^2}{w^2}},$$

де λ_0 – положення максимуму; y – значення максимуму; w – півширина (ширина смуги на рівні $y/2$).

Значно рідше смугу моделюють комбінацією кривих Гауса та Лоренца $\Psi_R(\lambda)$:

$$\Psi_R(\lambda) = q\Psi_Г(\lambda) + (1-q)\Psi_л(\lambda),$$

а q підбирають емпірично, причому $0 < q < 1$.

Один з напрямків виявлення складових смуг – цілеспрямований перебір параметрів піків при вибраній формі. Критерій оцінки – сума квадратів відхилень між реальною і змодельованою спектрограмами.

Основними недоліками є:

– неоднозначність рішення, і тільки знаючи не менш, ніж третину параметрів складових, можна отримати правильну відповідь [1];

– необхідні апріорні знання про точну кількість складових, а відстань між піками повинна бути не менше півсуми півширин відповідних смуг.

Розщеплення спектрального контуру $f(\lambda)$ на окремі смуги виконують також за допомогою перетворення Фур'є [2]:

$$g(p) = \int_{-\infty}^{\infty} f(\lambda)\cos p\lambda d\lambda,$$

де $g(p)$ – Фур'є-образ.

При цьому роблять припущення, що складові описуються кривими Гауса, а їх півширини однакові. Для визначення кількості смуг m, їх розташування та інтенсивностей використовують спеціальні рівняння та формули.

Перевагою цього підходу можна вважати стійкість процедури до шумової складової. Суттєвий недолік – вимога рівності півширин всіх смуг, що на практиці майже не зустрічається, звідси – вузьке застосування або ж грубість моделі та низька якість результатів.

Для одержання інформації про складові $f(\lambda)$ може бути використаний метод обчислення вищих моментів спектральних смуг. При цьому допускається, що форма смуг відома, наприклад, Гауса чи Лоренца. Вищі моменти характеризують асиметрію, гостровершинність та ін. Точність не висока, оскільки як шуми діють також найближчі складові спектра.

Усі розглянуті методи базувалися на тому, що форма модельних смуг відома. Метод Аленцева-Фока [3] це обмеження знімає.

Якщо $f(\lambda)$ має m смуг невідомої форми $\Psi_1(\lambda), \Psi_2(\lambda), \ldots, \Psi_i(\lambda), \ldots, \Psi_m(\lambda)$, то знімають N спектрограм $f_1(\lambda), f_2(\lambda), \ldots, f_j(\lambda), \ldots, f_N(\lambda)$ із $N > m$ експериментів, причому в кожному з них вкладення складових $\Psi_i(\lambda)$ змінюється, а форма ні. Тоді кожна експериментальна крива $f_j(\lambda)$ моделюється так:

$$f_j(\lambda) = \sum_{i=1}^{m} a_{ji}\Psi_i(\lambda),$$

де a_{ij} – постійні коефіцієнти.

Якщо при одержанні спектрограм забезпечено певний масштаб, а складові зовсім не перекриваються, то за допомогою розроблених у роботі [3] певних процедур можна вирахувати a_{ij}, а далі послідовно невідомі складові $\Psi_i(\lambda)$.

Принциповий недолік – проведення N експериментів із зміною накладення кожної смуги, що навіть в лабораторних умовах проблематично або неможливо. Вимога, щоб смуги зовсім не перекривалися, ще більш звужує коло застосувань. Використання даних, одержаних не лабораторним шляхом (наприклад, з літального апарата) взагалі виключається.

Усі розглянуті методи обробки потребують деяких апріорних знань про складові (кількість смуг, їх форма, розташування). Для визначення відповідних параметрів у кожному окремому випадку можуть бути різні підходи, що базуються на попередніх знаннях про об'єкти, їх стан та умови одержання $f(\lambda)$.

Незалежно від цього слід виділити дві процедури попередньої обробки даних, що дозволяють підвищити розрізнення складного контуру $f(\lambda)$ та виділити його інформативні параметри.

Суть першої полягає у звуженні смуг за рахунок перетворень Фур'є штучним розширенням спектра $S(\omega)$ кривої $f(\lambda)$, що досягається діленням $S(\omega)$ на спектр $S_0(\omega)$, вибраної функції $\Psi_0(\lambda)$. При цьому аналізується перетворена функція $f_{\text{пер}}(\lambda)$:

$$f_{\text{пер}}(\lambda) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-j\omega\lambda}\frac{S(\omega)}{S_0(\omega)}d\omega .$$

Найкращі результати одержують, коли форма смуг відома, а їх півширини приблизно рівні [4]. При цьому звуження смуг досягається не більше, ніж у 1,5 рази.

Основний недолік – вимога рівності півширин смуг $f(\lambda)$. Крім того, процедура не формалізована, оскільки результат залежить також від вдалого вибору $\Psi_0(\lambda)$, хоча деякі рекомендації на цей рахунок є [4]. Ми також не одержуємо ніяких числових характеристик, а лише в деяких випадках розщеплюємо слабо перекриті смуги.

Процедура, що базується на багаторазовому диференціюванні $f(\lambda)$, як засіб попередньої обробки має незрівнянні переваги перед розглянутою: високу спроможність розрізнення; формалізованість та простоту; наочність результатів, що подані в графіках; інформативність (дозволяє виявити не тільки піки, але і їх параметри); автоматично виключає сталу складову; не потребує попередньої інформації про $f(\lambda)$.

За першою похідною уточнюють положення точок перегину кривої $f(\lambda)$, інші похідні парного порядку дають змогу виділити смуги, що перекриваються. При цьому кількість та положення піків майже не міняються. Похідні непарних порядків починаючи з третьої не використовуються через незручність інтерпретації результатів обробки.

Основний недолік – можливість появи несправжніх піків, як наслідок взаємодії похідних парних порядків, що при розшифруванні можуть сприйматися як дійсні. Для їх виключення розроблені ефективні способи [5], що враховують знаки похідних різних порядків. Крім того, доцільним є попередній аналіз одержаних даних.

Найбільші труднощі – зростання відносного рівня шумової складової при диференціюванні експериментальних даних, що обмежує порядок похідних та відповідно рівень спроможності розрізнення.

Ефективним засобом підвищення стійкості процедури є використання цифрових фільтрів на основі формул, одержаних при апроксимації ланок $f(\lambda)$ поліномами обраного степеня за методом найменших квадратів:

$$f^*(\lambda_i) = \sum_{k=-\frac{M-1}{2}}^{\frac{M-1}{2}} b_k f(\lambda_{i+k}), \qquad (1)$$

де $f^*(\lambda_i)$ – похідна заданого порядку в точці λ_i; M – кількість відліків на ланці апроксимації (M – непарне); b_k – постійні коефіцієнти.

Вибір параметрів фільтра (1) залежить від рівня шуму у вхідних даних та структури $f(\lambda)$, а методики обчислення b_k вже розроблені.

Якщо є можливість накопичення даних, то знижують шумову складову попереднім усередненням.

Проаналізовані методи обробки через моделювання ми віднесли до першої групи. За їх результатами проводять в основному якісний аналіз об'єктів, враховуючи присутність тих чи інших піків у спектральних кривих. У деяких обмежених випадках можливо проводити і кількісний аналіз, що потребує певної моделі суміші, знання спектральних функцій всіх компонентів, у тому числі і тих, що входять в суміш, але не є предметом визначення. Ставляться і деякі інші обмеження.

Отже, при використанні методів обробки першої групи рівень визначення стану чи структури об'єктів залежить виключно від рівня адекватності відповідних моделей. Відсутність моделі або її складність виключає або суттєво знижує можливість одержання відповіді. При цьому відповідні процедури слабо або зовсім не формалізуються, що суттєво знижує продуктивність обробки з використанням комп'ютерних технологій. Це обмежує їх застосування лабораторними дослідженнями чи стаціонарними технологічними процесами та робить проблематичним або неможливим використання при обробці великих масивів даних для проведення швидких експрес-аналізів у процесі масштабних експериментів.

Друга група методів обробки повністю виключає моделювання і базується на використанні процедур та критеріїв теорії розпізнання образів.

При цьому із спектрограми формується n-вимірний образ-вектор X структури чи стану об'єкта $X=(x_1,x_2,\ldots,x_n)$, де $x_1,x_2,\ldots,x_n$ – сукупність ознак.

Далі за допомогою вибраних критеріїв та вирішальних правил образ відноситься до певного класу об'єктів, на основі чого робиться якісний чи кількісний аналіз.

Основна проблема – вибір інформативних ознак при формуванні образу X у кожному окремому випадку, оскільки критерії та процедури розпізнання розроблені і відомі.

На перший погляд, як $x_1,x_2,\ldots,x_n$ потрібно брати всі відліки $f(\lambda)$, і чим їх більше, тим краще. Але досвід показав, що такий підхід хибний. Це можна пояснити «шумовою» дією неінформативних ознак, яких у цьому випадку значна більшість. Крім того, суттєво ускладнюється реалізація відповідних критеріїв.

Таким чином виникає задача вибору оптимального набору ознак, коли збережено всі інформативні ознаки при мінімальному їх числі n. Найчастіше його знаходять за допомогою покрокової процедури: спочатку вибирають найкращу ознаку, потім другу і т. д. Таким способом можна одержати оптимальний набір, коли ознаки незалежні одна від одної, або коли залежність лінійна, тоді потрібно зробити відомі математичні перетворення, що знімають залежність. Якщо зв'язок нелінійний або невідомий, то проблему вибору оптимального набору ознак вирішити неможливо.

У більшості випадків оптимальний набір ознак формують за значеннями $f(\lambda)$ на певних довжинах хвиль, але використовують й інші величини (коефіцієнти кольору, параметри піків). Як критерії ефективності альтернативних наборів доцільно використовувати показники, обчислені на навчальній виборці:

– ентропію H навчальної вибірки:

$$H=-\frac{1}{L}\sum_{k=1}^{L}\sum_{j=1}^{Q}P_{kj}\log P_{kj};$$

– середню кількість інформації $\bar{I}$ в множині реалізацій наборів ознак відносно множини класів розпізнання:

$$\bar{I}=\frac{1}{L}\sum_{k=1}^{L}(H-\sum_{j=1}^{Q}P_{kj}\log P_{kj});$$

– значення середньої ймовірності $\overline{P}$ помилки розпізнання:

$$\overline{P} = \frac{1}{L}\sum_{k=1}^{L}(1 - \max_{j=1,\ldots,Q} P_{kj}),$$

де L – кількість об'єктів вибірки; Q – число класів; P_{kj} – імовірність віднесення k-го образу до j-го класу.

Про ефективність підходу на основі розпізнання образів свідчить такий приклад. За спектрограмами, одержаними з літака, проведено картографування та класифікацію поверхневого шару ґрунтів України та Молдови на три класи, а після спеціальної обробки спектрограм – на дев'ять класів [6]. Набір ознак включав кілька значень $f(\lambda)$ на певних довжинах хвиль.

Аналогічні методики мали успіх при визначенні стану сільськогосподарських рослин та дослідженні Землі з космосу.

Як набір ознак можна використовувати параметри піків (їх положення, півширини, інтенсивності або інше в одному вектор-образі), але це потребує таких процедур обробки, які пов'язані з моделюванням, або ж урахування специфіки досліджуваних об'єктів чи їх стану, що пов'язано з усіма розглянутими недоліками.

Авторами запропонована така процедура вибору оптимального набору ознак:

– експериментальна крива $f(\lambda)$ диференціюється парну кількість раз за допомогою цифрових фільтрів (1);

– як ознаки використовуються інтенсивності максимумів та мінімумів на диференційованій кривій;

– одержані дані нормуються за максимальним значенням одного з піків, що зменшує кількість ознак на одиницю, за рахунок перерозподілу інформації між ними (один стає лінійно залежним).

Отже, чисто формальними діями виходять або відразу на оптимальний набір, або ж на мінімальну кількість ознак, серед яких можуть бути неінформативні, що пов'язані з фоном. При цьому оптимальний набір одержують за рахунок описаної покрокової процедури, а нормування проводять до значення одного з неінформативних, що збільшує оптимальний набір на 1.

Якщо є можливість окремо зняти фонову спектрограму і працювати з різницею між реальною та фоновою кривими, то відразу отримуємо оптимальний набір.

Основне обмеження – стійкість процедури цифрової фільтрації і відповідно значень ознак. Диференціювання та нормування зменшують дію дрейфових складових та зміни коефіцієнта передачі при різних умовах одержання спектрограм. Усередненням, якщо є можливість, ослаблюють складову випадкового шуму у вхідних даних. Особливу увагу слід зосередити на виборі параметрів цифрового фільтру. Критерій – максимальна похибка, яка не повинна перевищувати кількох відсотків.

Запропонована авторами процедура, що тепер широко використовується, була перевірена на прикладі об'єктів, що є сумішшю частинок фотосистем 1 та 2 з хлоропластів вищих рослин [7]. До навчальної вибірки входило п'ять класів об'єктів з вмістом фотосистеми 1 в 0, 25, 50, 75 та 100 %. При цьому всі образи навчальної вибірки класифікувалися вірно, а поза навчальною вибіркою з імовірністю не гірше 0,96, тобто, кількісний аналіз проводився на досить високому рівні.

Дослідження похибок цифрової фільтрації в цьому конкретному випадку (з інших може бути не так) показало, що найкращою є формула (1) з п'яти точок, отримана при апроксимації ланок $f(\lambda)$ параболами:

$$f''(\lambda_i) = \frac{2f(\lambda_{i-2}) - f(\lambda_{i-1}) - 2f(\lambda_i) - f(\lambda_{i+1}) + 2f(\lambda_{i+2})}{7},$$

де $f''(\lambda_i)$ – друга похідна.

Фільтри четвертої похідної давали недопустиму похибку (більше 10 %).

Як критерій розпізнання $F(X,\overline{X}_j)$ в цьому випадку найкращим виявився баєсовський, що досить легко реалізується, оскільки оптимальний набір ознак n склав п'ять значень:

$$F(X,\overline{X}_j)=ln|C_j|+(X-\overline{X}_j)'C_j^{-1}(X-\overline{X}_j),$$

де $\overline{X}_j$ та C_j – відповідно n-вимірний вектор математичного сподівання та коваріаційна матриця ознак навчальних образів j-го класу.

Класифікація – за мінімумом критерію $F(X,\overline{X}_j)$.

Отже, показані можливості різних методів обробки спектрометричних даних та перспективність процедур з використанням критеріїв теорії розпізнання образів, на основі формалізованих правил дають змогу швидко обробляти величезні масиви експериментальних даних на комп'ютерах з метою якісного і навіть кількісного експрес-аналізу складу чи стану об'єктів. При цьому найбільш формалізованою та простою є методика вибору оптимального набору ознак за значеннями парних похідних в точках перегину.

Список літератури

1. *Антипова-Коротаева И.Н., Казанова Н.Н.* Математическое разложение сложных спектральных контуров на компоненты с частично известными параметрами // Журн. прикладной спектроскопии. – 1972. – Т.16. № 5. – С. 855–858.

2. *Гречушников В.Н., Калинкина И.Н., Старостина Л.С.* Разложение перекрытых спектральных линий методом Фур'е // Журн. прикладной спектроскопии. – 1975. – Т.23. №6. – С.1059–1066.

3. *Фок М.В.* Разделение сложных спектров на индивидуальные полосы при помощи обобщенного метода Аленцева // Тр. ФИАН. – 1972. – Т.59. – С. 3–10.

4. *Allen L.C., Gladney H.M.* Resolution enhancement for spectra of chemical and physical interest // The jornal of chemical physics. – 1964. – Vol. 40. № 11. – P. 3125–3143.

5. *Marrey J.R.* On determining spectral peak positions from composite spectra with a digital computer // Analytical Chemistry. – 1968. – Vol. 40. – P. 905–914.

6. *Кондратьев К.Я., Васильев О.В., Федченко П.П.* Опыт распознавания почв по их спектрам отражения // Почвоведение. – 1983. – № 4. – С. 5–17.

7. *Жук Л.А., Кочубей С.М., Сураев В.Ф.* Автоматизация экспресс-анализа количественного состава многокомпонентных объектов // Механизация и автоматизация управления. – 1989. – № 2. – С. 24–27.

Стаття надійшла до редакції 30.03.02.

УДК 004.032

Ю.М. Мінаєв, д-р техн. наук, проф.,
Бенамур Лієс,
М.М. Гузій, канд. техн. наук, доц.,
В.В. Давиденко, асп.

ІДЕНТИФІКАЦІЯ АТАК НА КОМП'ЮТЕРНІ МЕРЕЖІ НА ПІДСТАВІ НЕЙРОМЕРЕЖНИХ ТЕХНОЛОГІЙ У СИСТЕМІ МОДЕЛЮВАННЯ MATLAB/SIMULINK

Розглянуто методику побудови нейромережі для ідентифікації атак на комп'ютерні мережі за допомогою засобів пакету математичного моделювання MatLab/Neural Network/ Simulink. Наведено приклад, що ілюструє ефективність нейромережної технології.

Проблема виявлення атак на комп'ютерні мережі (КМ) є домінуючою в теорії та практиці захисту інформації. Сучасні системи виявлення атак IDS (Intrusion Detecting System) працюють на двох рівнях залежно від того, до якої інформації існує доступ. Загальним у цих підходах є пошук відповідних ознак (сигнатур), комбінації цих ознак (шаблонів), які вказують на ворожі дії або на їх підозру. Якщо пошук цих сигнатур та шаблонів виконується на рівні мережного