УДК 004.032.26(043.2)

ДОСЛІДЖЕННЯ АЛГОРИТМІВ ПРОВЕДЕННЯ КЛАСТЕРНОГО АНАЛІЗУ ДЛЯ ВИРІШЕННЯ ЗАДАЧ НЕРУЙНІВНОГО КОНТРОЛЮ

В.С. Єременко
Кандидат технічних наук, доцент, завідувач лабораторії
Науково-дослідна лабораторія систем неруйнівного контролю*
Контактний тел.: (044) 406-74-35, (067)209-07-69
E-mail: nau_307@ukr.net

А.В. Переїденко*
*Кафедра інформаційно-вимірювальних систем
Національний авіаційний університет
пр. Комарова, 1, корпус 11, ауд. 408, м. Київ, Україна, 03680
Контактний тел.: (044) 406-74-35, (093)711-10-70
E-mail: zoolkis@meta.ua

Дана загальна характеристика процедурі кластерного аналізу даних. Наведені результати дослідження різних функцій відстані як критерію про схожість об'єктів. Описано систему для проведення кластерного аналізу і дослідження достовірності кластеризації із застосуванням алгоритмів на основі описаних мір близькості. Систему реалізовано в середовищі LabVIEW 8.5

Ключові слова: кластерний аналіз, функція відстані, нейронна мережа

Дана общая характеристика процедуры кластерного анализа данных. Приведены результаты исследования разных функций расстояний как критерия о схожести объектов. Описана система для проведения кластерного анализа и исследования достоверности кластеризации с применением алгоритмов на основе описаных мер близости. Система реализована в среде NI LabVIEW 8.5

Ключевые слова: кластерный анализ, функция расстояния, нейронная сеть

This article is devoted to realization system of the cluster analysis without etalon samples. Main different vector space metrics were analyzed using the special control system. System was created with NI LabVIEW 8.5

Key words: vector space metric, cluster, cluster analysis, neural network

1. Вступ

Моніторинг технічного стану виробів із композиційних матеріалів ґрунтується на використанні діагностичної інформації, отриманої безпосередньо в процесі технічної діагностики та неруйнівного контролю. Процес моніторингу містить процедури отримання, перетворення та аналізу діагностичної інформації, а кінцевим етапом є прийняття рішення про технічний стан контрольованого об'єкту. Якість та ефективність розпізнавання безпосередньо впливає на достовірність діагностики в цілому. Тому розробка системи розпізнавання (класифікатора) стану виробів із композиційних матеріалів для своєчасного виявлення пошкоджень є важливою та актуальною задачею. В умовах наявності мінімальної кількості інформації про образи, що розпізнаються, та обмеженій кількості образів для навчання, а також враховуючи, що однією зі складних для виконання задач є виготовлення спеціальних еталонних зразків з різними типами дефектів притаманних контрольованому матеріалу, в роботі пропонується вирішити поставлену задачу дігностики технічного стану виробів із композитів на основі використання штучних нейронних мереж, які здатні проводити нелінійну кластеризацію, а також є гнучкими та здатними до розпізнавання за ознаками на основі сучасних методів обробки інформації.

2. Постановка задачі

Метою роботи є дослідження можливості контролю виробів із композитів, проведення якісного кластерного аналізу без попереднього навчання на еталонних зразках із застосуванням в якості апарату обробки експериментальних даних штучні нейронні мережі. В такому випадку відпадає необхідність мати еталонні об'єкти.

Особливість задачі кластеризації полягає в тому, що класи об'єктів спочатку не відомі. Результатом кластеризації є розбиття об'єктів на групи, що задовольняють деякому критерію оптимальності [1]. Цей крите-

рій може бути деяким функціоналом, що виражає рівні бажаності різних варіантів розбиття і об'єднання. На відміну від задач класифікації, кластерний аналіз не потребує апріорних припущень про набір даних, не накладає обмеження на представлення досліджуваних об'єктів, дозволяє аналізувати показники різних типів даних (інтервальні дані, частоти, бінарні дані). При цьому необхідно пам'ятати, що змінні повинні вимірюватися в конгруентних (порівнянних) шкалах.

3. Опис вирішення задачі

У роботі розглянута можливість застосування нейронних мереж Кохонена для вирішення задач кластеризації при проведенні неруйнівного контролю виробів з композиційних матеріалів.

Формально задача кластеризації формується наступним чином. Нехай X – множина об'єктів, Y – множина номерів (імен, міток) кластерів. Задано функцію відстані між об'єктами $\rho(x,x')$, також є кінцева навчальна вибірка об'єктів $X^m = \{x_1,\ldots,x_m\} \in X$. Потрібно розбити вибірку на непересічні підмножини (кластери) таким чином, щоб кожен кластер складався з об'єктів, близьких за метрикою ρ, а об'єкти різних кластерів істотно відрізнялися. При цьому кожному об'єкту $x_i \in X^m$ приписується номер кластера y_i. Алгоритм кластеризації – це функція $a: X \to Y$, яка будь-якому об'єкту $x \in X$ ставить у відповідність номер кластера $y \in Y$. Множина Y в деяких випадках відома заздалегідь, проте частіше ставиться задача визначити оптимальне число кластерів, з погляду одного або іншого критерію якості кластеризації.

Методи кластеризації розрізняються за правилами побудови кластерів. В якості таких правил виступають критерії, що використовуються при вирішенні питання про «схожість» об'єктів. Критерієм для визначення схожості і відмінності кластерів є відстань між векторами на діаграмі розсіювання (рис. 1).

Для обчислення відстані між об'єктами використовуються різні міри схожості (міри подібності), які також називаються метриками або функціями відстаней [2]. Способів визначення міри відстані між кластерами існує декілька.

У роботі досліджувалися деякі найбільш поширені способи визначення відстані.

1). Перший і найпоширеніший спосіб – обчислення евклідової відстані між двома векторами X і Y в n-вимірному просторі, коли відомі їх координати x_i і y_i, де $i = \overline{1,n}$. Воно просто є геометричною відстанню в багатовимірному просторі і обчислюється наступним чином:

$$\rho(X,Y) = \sqrt{\sum_i (x_i - y_i)^2}.$$

Слід зазначити, що евклідова відстань (та її квадрат) обчислюється за початковими, а не за стандартизованими даними. Це звичайний спосіб його обчислення, який має певні переваги (наприклад, відстань між двома об'єктами не змінюється при введенні в аналіз нового об'єкту, який може виявитися викидом). Проте, на відстані можуть сильно впливати відмінності між осями, по координатах яких обчислюються ці відстані.

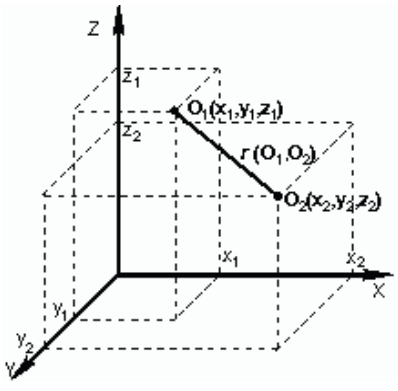

Рис. 1. Відстань між двома векторами в просторі

2). Іноді застосовується квадрат стандартної евклідової відстані, щоб додати великої ваги віддаленішим один від одного об'єктам. Ця відстань обчислюється за формулою:

$$\rho(X,Y) = \sum_i (x_i - y_i)^2.$$

3). Манхеттенська відстань (відстань міських кварталів) або так звана «хеммінгова» або «сіті-блок» відстань – ця відстань є просто середнім різниць по координатах. В більшості випадків ця міра відстані приводить до таких же результатів, як і для звичайної відстані Евкліда. Проте слід зазначити, що для цього заходу вплив окремих великих різниць (викидів) зменшується (оскільки вони не підносяться до квадрату). Манхеттенська відстань обчислюється за формулою:

$$\rho(X,Y) = \sum_i |x_i - y_i|.$$

4). У разі, коли необхідно визначити два об'єкти як "різні", якщо вони відрізняються по якомусь одному вимірюванню варто використовувати відстань Чебишева. Відстань Чебишева обчислюється за формулою:

$$\rho(X,Y) = \max(|x_i - y_i|).$$

5). Іноді для того, щоб суттєво збільшити або зменшити вагу, що відноситься до розмірності, для якої відповідні об'єкти сильно відрізняються використовується степенева відстань:

$$\rho(X,Y) = \sqrt[r]{\sum_i |x_i - y_i|^p},$$

де r і p – параметри, що визначаються дослідником. Параметр p відповідає за поступове зважування різниць за окремими координатами, параметр r відповідає за прогресивне зважування великих відстаней між об'єктами. Якщо обидва параметри – r і p, рівні двом, то ця відстань збігається з відстанню Евкліда.

6). У роботі було також розглянуто відстань на основі косинуса:

$$\rho(X,Y) = 1 - \frac{\left|\sum_{i=1}^{N} x_i \cdot y_i\right|}{\sqrt{\sum_{i=1}^{N} x_i^2} \cdot \sqrt{\sum_{i=1}^{N} y_i^2}}.$$

7). Відстань Камбера описується виразом:

$$\rho(X,Y)=\sum_i \frac{|x_i-y_i|}{x_i+y_i},\ x_i\neq -y_i\,.$$

8). Відстань Махаланобіса обчислюється таким чином:

$$\rho(X,Y)=(X-Y)^T\mathbf{C}^{-1}(X-Y),$$

де X, Y – вектори середніх значень змінних відповідно однієї та другої групи, C^{-1} – обернена коваріаційна групова матриця, ()T – оператор транспонування.

9). Відстань χ^2 визначається на основі таблиці зв'язаності, складеної з об'єктів X та Y, які частіше за все є векторами частот. Тут розглядаються очікувані значення елементів, що дорівнюють $E(x_i)=\frac{x_n\cdot(x_i+y_i)}{x_n+y_n}$ та $E(y_i)=\frac{y_n\cdot(x_i+y_i)}{x_n+y_n}$, де $x_n\neq -y_n$, а відстань χ^2 має вид кореня з відповідного показника:

$$\rho(X,Y)=\sqrt{\sum_{i=1}^{n}\frac{(x_i-E(x_i))^2}{E(x_i)}+\sum_{i=1}^{n}\frac{(y_i-E(y_i))^2}{E(y_i)}}\,.$$

Також в задачах кластеризації можуть бути використані і ряд інших мір близькості. Для визначення відстані між кластерами авторами були досліджені декілька правил (методів) об'єднання або зв'язку для двох кластерів.

1). Метод ближнього сусіда або одинарний зв'язок. В цьому випадку відстань між двома кластерами визначається відстанню між двома найбільш близькими об'єктами (найближчими сусідами) в різних кластерах. Цей метод дозволяє виділяти кластери як завгодно складної форми за умови, що різні частини таких кластерів сполучені ланцюгами близьких один до одного елементів. В результаті роботи цього методу кластери представляються довгими "ланцюгами" або "волокнистими" кластерами, "зчепленими разом" тільки окремими елементами, які випадково виявилися ближчими ніж інші один до одного.

2). Метод найбільш віддалених сусідів або повний зв'язок. При використанні даного методу відстані між кластерами визначаються найбільшою відстанню між будь-якими двома об'єктами в різних кластерах ("найбільш віддаленими сусідами"). Метод добре використовувати, коли об'єкти дійсно походять з різних "згущень". Якщо ж кластери мають в деякому роді подовжену форму або їх природний тип є "ланцюговим", то цей метод не слід використовувати.

3). Метод K-середніх. В загальному випадку цей метод визначає рівно K різних кластерів, розташованих на можливо великих відстанях один від одного. Програма починає з K випадково вибраних кластерів, а потім змінює приналежність об'єктів до них, щоб мінімізувати відмінність всередині кластерів і максимізувати відмінність між кластерами. У кластеризації за методом K-середніх програма переміщує об'єкти з одних кластерів в інші, щоб отримати найбільш значущий результат при проведенні дисперсійного аналізу.

Для вирішення задачі кластеризації при проведенні неруйнівного контролю виробів з композиційних матеріалів була реалізована нейронна мережа Кохонена. Нейронні мережі Кохонена – це клас нейронних мереж, основним елементом яких є шар Кохонена. Шар Кохонена складається з адаптивних лінійних суматорів (лінійних формальних нейронів). Як правило, вихідні сигнали шару Кохонена обробляються за правилом «переможець отримує все»: найбільший сигнал перетворюється на одиничний, останні – в нуль [3].

В результаті роботи були досліджені описані вище міри близькості. В якості експериментальних даних для дослідження були використані дані, отримані при проведенні контролю зразків композиційних матеріалів методом низькошвидкісного удару [4]. Досліджуємий зразок мав п'ять характерних зон – без дефектну і чотири зони з різним ступенем пошкодженості (дефекту). Інформативними параметрами для аналізу були амплітуда та довжина імпульсу прийнятого сигналу. Для порівняння мір близькості центри кластерів скупчення точок у двовимірному просторі для кожної із зон досліджуємого зразка (рис. 2) були знайдені за допомогою штучної нейронної мережі (НМ) Кохонена, а також як арифметичні центри скупчення векторів.

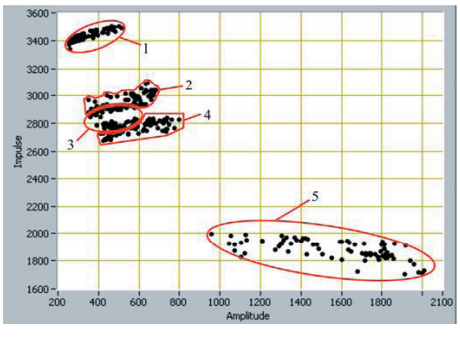

Рис. 2. Розміщення векторів з різних ділянок зразка композиту
1-4 — ділянки з різним ступенем дефекту,
5 — бездефектна ділянка

На рис. 3 зображено інтерфейс системи для дослідження різних мір близькості. Систему було розроблено з використанням пакету NI LabVIEW 8.5.

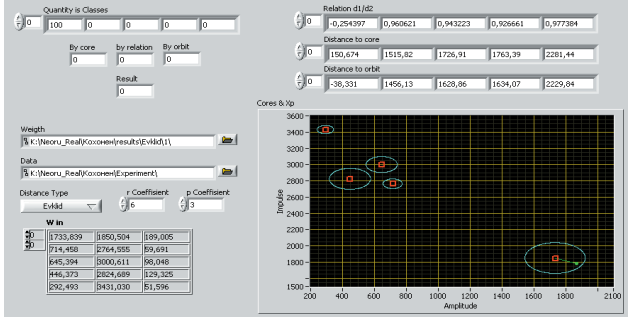

Рис. 3. Інтерфейс системи для дослідження мір близькості

Результати дослідження описаних мір близькості можна представити у вигляді табл. 1 і 2. У відповідні таблиці занесено достовірність приєднання вектору до певного кластеру (ділянки). Достовірність з якою об'єкт або вектор відноситься до певного визначеного кластеру залежить від методу за яким було знайдено центри скупчення векторів, що належать до кожної із зон експериментального зразка композиційного мате-

ріалу. В таблицях приведено результати застосування алгоритмів кластеризації з використанням тільки тих мір близькості, достовірність яких склала більше 80%. Результати, які було отримано із застосуванням арифметичних центрів і центрів, знайдених шляхом проведення кластеризації НМ Кохонена мають певні відмінності.

Таблиця 1

Точність застосування мір близькості із арифметичними центрами кластерів

Тип ділянки	Міри близькості				
	Чебишева	Махаланобіса	Евкліда	Степенева	Квадрат Евкліда
без дефекту	1,00	1,00	1,00	1,00	1,00
дефект 1	1,00	0,94	0,94	1,00	0,94
дефект 2	0,85	0,85	0,85	0,85	0,85
дефект 3	0,73	0,55	0,55	0,73	0,55
дефект 4	1,00	1,00	1,00	1,00	1,00
Загальна точність	0,92	0,87	0,87	0,92	0,87

На рис. 4 зображено достовірність віднесення об'єкту до кластеру із застосуванням різних мір близькості. Найкраща достовірність проведення кластерного аналізу досягається із застосуванням міри близькості Чебишева. Високу достовірність також можна отримати із застосуванням міри близькості Махаланобіса та Евкліда. Алгоритми кластеризації на основі інших мір близькості для вирішення поставлених задач показали достовірність нижче 80%, тому їх використання в даній ситуації вважається недоцільним.

Таблиця 2

Точність застосування мір близькості із центрами точок (кластерів), що знайдені шляхом застосування нейронної мережі Кохонена

Тип ділянки	Міри близькості		
	Чебишева	Махаланобіса	Евкліда
без дефекту	1,00	1,00	1,00
дефект 1	1,00	0,65	0,82
дефект 2	0,92	1,00	0,65
дефект 3	0,79	1,00	0,93
дефект 4	1,00	1,00	1,00
Загальна точність	0,94	0,93	0,88

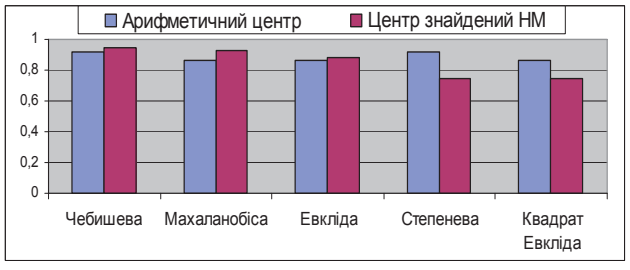

Рис. 4. Достовірність кластеризації із застосуванням різних мір близькості

4. Висновки

На основі отриманих результатів можна зазначити, що для вирішення задачі безеталонної дефектоскопії без попереднього навчання на еталонних зразках для знаходження відстані між вектором, що характеризує властивості об'єкта контролю, і центром відповідного кластеру найбільш доцільно використовувати міри близькості Чебишева, Махаланобіса або Евкліда. Застосування алгоритмів пошуку відстаней на основі цих мір близькості дозволяє отримати достовірність віднесення вектору до необхідного кластеру відповідно 94, 93 і 88 %. При використанні алгоритмів кластеризації на основі інших мір близькості, була отримана достовірність нижче 80%, тому для вирішення даних задач їх застосування є недоцільним.

Література

1. Дюран Б., Оделл П. Кластерный анализ. Пер. с англ. Е. З. Демиденко. Под ред. А. Я. Боярского. – М.: «Статистика», 1977. – 128 с.
2. Скворцов В.А. Примеры математических пространств. – М.: МЦНМО, 2002. – 24 с.
3. Хайкин Саймон. Нейронные сети: полный курс, 2-е издание.: Пер. с англ. – М.: Издательский дом «Вильямс», 2006. – 1104 с.
4. Еременко В.С., Мокийчук В.М., Овсянкин А.М. Обнаружение ударных повреждений сотовых панелей методом низкоскоростного удара // Техническая диагностика и неразрушающий контроль. – К., 2007.– №1. – с.24-27.