

第 1 章

粗糙集的基本概念

- 1 前言
- 2 粗糙集的產生
- 3 粗糙集的基本假設—知識
- 4 粗糙集的特性
- 5 粗糙集、模糊集、實證理論與灰色理論的異同性
- 6 粗糙集的研究內容
- 7 粗糙集的未來發展

1.1 前言

在自然科學，社會科學和工程技術的很多領域中，都不同程度地涉及到對不確定性（uncertainty）問題和對不完備（imperfect）資訊的處理。從實際系統中所得到的數據往往包含著雜訊，不夠精確甚至不完整，如果採用純數學上的假設以來消除或迴避這種不確定性，效果往往不甚理想。反之如果對這些資訊進行合適地處理，會有助於相關實際系統問題的解決。

正因為如此，多年來研究人員一直在努力尋找能處理不確定性和不完整（備）性的有效途徑，首先發現機率統計方法和模糊集是其中的兩種方法，並且已經廣泛的應用於一些實際領域。但是這些方法有時候需要一些數據的附加資訊或知識，例如統計機率分布及模糊歸屬函數等，不過這些資訊並不是很容易的可以得到，因此產生了許多衍生的研究方式，例如灰色理論、可拓理論及粗糙集（rough set），本書則鎖定在粗糙集的研究上。

例 1.1 例舉各種不確定性

- 解：**
- (1) 隨機性：隨機現象的不確定性，這是已經有很久歷史的經典概念。
 - (2) 模糊性：模糊概念下的不確定性，通常是由經驗法則歸屬函數發生的不確定性（1968 年由 Zadeh 提出）。

(3) 粗糙性：資訊系統中知識和概念的不確定性。 

此外，還有 1948 年仙農（C. E. Shannon）以起源於經典熱力學，借用資訊熵（entropy）的概念以度量系統的隨機程度，從資訊理論觀點奠定了這理論基礎，並且成功實現了隨機性的數理觀念。現在熵廣泛用於不確定性度量，而仙農熵則被稱為系統的熵。

1.2 粗糙集的產生

Z. Pawlak（波蘭；1982 年）針對 1904 年謂詞邏輯創始者弗里格（G. Frege；德國，1848~1925）的邊界線區構想提出了粗糙集的概念，並出版了第一本關於粗糙集的書。對粗糙集而言，Z. Pawlak 將有關想法確認的個體都歸屬在邊界線區域內，而這種邊界線區域則被定義為上近似集和下近似集的差集。由於粗糙集可以用明確的數學公式加以描述，所以模糊集元素的數目是可以被計算的，亦即在 0 和 1 之間的模糊度是可以計算而求得。粗糙集主要特點恰好能處理不分明問題的常規性，並且以不完整的資訊或知識去處理一些不明確現象的能力。

以數學的方式而言，粗糙集則是利用以後會更詳細說明的下近似集和上近似集的方式，在不需要任何先驗假設或額外的相關數據資訊之下，依據所觀察及度量到某些不精確的結果而進行數據分類的能力。而所謂的分類就是推理、學習和決策中的關鍵問題。二十多年來，粗糙集經過許多電腦學家、數學家不懈地研究，已經在理論上日漸趨

於完善。特別是由於上一世紀八十年代末和九十年代初在知識發現等領域得到了相當成功的應用，而越來越受到國際上的廣泛關注。粗糙集理論模型應用層面廣泛，涵蓋醫學工程、製程管理、財務工程等，目前主要大量應用於企業破產預警、資料庫行銷與金融投資預測三大領域。主要的原因是：粗糙集理論中內的特徵主要是可以藉由歷史資料庫（知識庫），挖掘資訊中隱藏的模型藉以預測未來。

而這三大領域的歷史資料庫皆由多種屬性資訊表建立，並且擁有相似的預測特徵。本書將有系統地介紹智能處理研究領域近年來迅速發展的粗糙集理論。該理論不僅能研究精確知識的表達、學習及歸納等方法，也具有大量數據資訊發現，推理知識和分辨系統的特徵、過程及對象等功能。粗糙集理論的應用通常是結合統計方法、模糊集、神經網絡及灰色理論，用以進行推理學習、處理不完整數據和解決不精確性等等問題。經由與其他理論進行整合應用，對於粗糙集理論的屬性、不可辨識與結果檢測進行探討與改良，使得粗糙集理論更具深度與廣度。

由於粗糙集是一種刻劃不確定性和不完整性的數學工具，能有效地分析不精確，不完整（incomplete）和不一致（inconsistent）等各種不完備的資訊，還可以對數據進行分析和推理，從其中發現隱含的知識和潛在的規律。經由歸納，可以得知粗糙集能夠有效地處理下列問題：

- (1) 不確定或不精確知識的表達
- (2) 經驗學習並從經驗中獲取知識
- (3) 不一致資訊的分析
- (4) 根據不確定及不完整的知識而進行推理

- (5)在保留資訊的前提下進行數據約簡
- (6)加以數據分類，分解近似建模型
- (7)識別並評估數據間的相互依賴關係
- (8)系統內各個因子之分類及重要性分析

1.3 粗糙集的基本假設—知識

「知識」這個概念在不同的範疇內有多種不同的含義。在粗糙集理論中「知識」被認為是一種分類能力，人們的行為是基於分辨對象為現實的或抽象的能力。例如在遠古時代，人們為了生存必須能分辨出什麼可以食用，什麼不可以食用；醫生診斷病患時，必須辨別出患者得的是哪一種病。這些根據事物的特徵差別將其分門別類的能力均可以看作是某種「知識」。由此可以得知知識在人工智慧中是一個非常重要的概念，而在粗糙集中，知識是對定義為全體討論領域的一個劃分，也是一種對對象進行分類的能力。

換言之，假設全集合為論域（universe；又稱域或對象） U ，（ $U \neq \phi$ ），其中 $\forall x \in U$ 稱為 U 中的一個概念（concept）， U 中的一個概念集合 $U = \{x_1, x_2, x_3, \dots, x_n\}$ 稱為關於 U 中的知識。此時 $x_i \subseteq U$ ， $x_i \neq \phi$ ， $x_i \cap x_j = \phi$ ， $i \neq j$ ， $i, j = 1, 2, 3, \dots, n$ ， $\bigcup_{i=1}^n x_i = U$ ，其中空集合本身也是一個概念集合。

例 1.2 集合：按照同一目的或特點，將所有的研究對象（元素）組合在一起。

表 1-1 · 集合論的代表性符號一覽表

符 號	U	Ω	ϕ	ω	A	$A^C = 1 - A$
集合論	論域	全集合	空集合	元集合	子集合	補集合
符 號	$A = B$		$A \subset B$		$A \cup B$	$A \cap B$ 或 AB
集合論	A 與 B 相等		A 是 B 的子集合		A 與 B 的聯集	A 與 B 的交集
符 號	$A - B$		$AB = \phi$		$A \cup B = \Omega$	
集合論	A 與 B 的差集		A 與 B 不相同		A 與 B 的構成的全集	

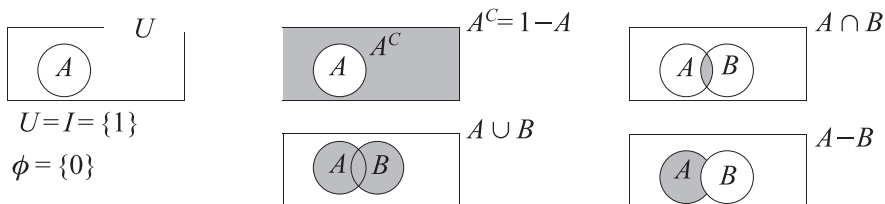


圖 1-1 · 集合的圖示法

(1) 集合表示方法

①列舉法： $U = \{x_1, x_2, x_3, \dots, x_n\}$

$A = \{a, b, \dots, n\}$ ， $B = \{1, 2, 3, \dots, m\}$ ， $C = \{\text{陳，林，李，蔡，史}\}$

②描述法

$A = \{x \mid \varphi(x)\}$ ： A 為滿足 $\varphi(x)$ 的一切 x 組成的集合

$B = \{x \mid x \in N, \text{ 並且 } 0 \leq x \leq 2\}$ ， $N = \{0, 1, 2, \dots, n, \dots\}$ ： $0 \sim 2$ 中最小元素的集合

(2) 空集合： $A = \{0\} = \{\} = \phi$

如果 $A = \{a, b, c, d\} \rightarrow \text{card}(A) = |A| = 4$ 表示該集合元素的數目

$a \in A$ ：元素 a 屬於 A ， $f \notin A$ ：元素 f 屬於 A 。

如果 $A = \{a, b, c, d\}$ ， $B = \{b, d\}$ 則 $A \subset B$ 。

例 1.3 給定人力資源的集合 $U = \{x_1, x_2, x_3, \dots, x_8\}$ ，並且假設有不同學歷（學士，碩士，博士）、學術領域（工程，商學，醫學）及性別（男，女）。這例子的目的是為說明「知識（分類能力獲得的概念或範疇的集合）」，「知識庫」和「近似空間」。

解：(1) 按顏色、形狀及體積分類：

①學歷

學士： x_1, x_3, x_7

碩士： x_2, x_4

博士： x_5, x_6, x_8

②學術領域

工程： x_1, x_5

商學： x_2, x_6

醫學： x_3, x_4, x_7, x_8

③性別

男： x_2, x_7, x_8

女： x_1, x_3, x_4, x_5, x_6

換言之，定義三個新的屬性：學歷 R_1 、學術領域 R_2 及性別 R_3 ，根據這些屬性的新定義，我們可以得到下面的三種分類：

$$U|R_1 = \{\{x_1, x_3, x_7\} \{x_2, x_4\} \{x_5, x_6, x_8\}\}$$

$$U|R_2 = \{\{x_1, x_5\} \{x_2, x_6\} \{x_3, x_4, x_7, x_8\}\}$$

$$U|R_3 = \{\{x_2, x_7, x_8\} \{x_1, x_3, x_4, x_5, x_6\}\}$$

這樣，在我們的方法中，可以把知識的概念或範疇加以整理，獲得分類的能力。那麼， $U = \{x_1, x_2, x_3, \dots, x_n\}$ 中的任意概念族稱為關於 U 的抽象知識，簡稱知識。例如，子集合 $\{x_1, x_3, x_7\}$ 就是 $U = \{x_1, x_2, x_3, \dots, x_8\}$ 中按學歷分類的「學士程度的」知識， $\{x_1, x_5\}$ 就是 $U = \{x_1, x_2, x_3, \dots, x_8\}$ 中按學術領域分類的「工程的」知識等等。通常，我們不只是處理只有一個單獨的分類，而是處理 $U = \{x_1, x_2, x_3, \dots, x_n\}$ 上的一些分類族。一個 U 上的分類族則定義為一個 U 上的知識庫，它構成了一個特定論域的分類，如果我們定義這知識庫為 K ，則

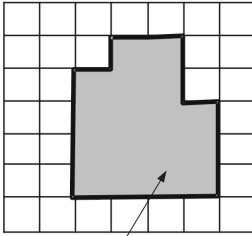
$$K = (U, R) = (U, R_1, R_2, R_3) \quad (1-1)$$

如果 R 是 U 上的劃分 $R = \{X_1, X_2, \dots, X_n\}$ 表達的等價關係，因此 (U, R) 又稱為近似空間，而 R 是 U 上等價關係的一個族集合。知識庫 K 的等價關係表也可以表示如表 1-2 所示。

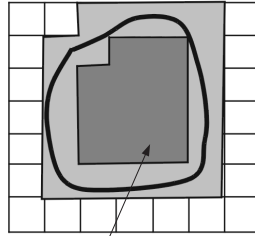
表 1-2 • $K = (U, R) = (U, R_1, R_2, R_3)$

U/R	R_1	R_2	R_3
x_1			
x_x			
$\vdots$			
x_n			

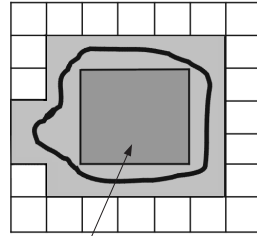
例 1.4 粗糙集「概念近似」的想法



確定的：
 X 與基本要素 $[x]_R$ 符合



非確定的：
 X 與基本要素 $[x]_R$ 不符合得很好



不能定義的：
 X 與基本要素 $[x]_R$ 不能符合

1.4 粗糙集的特性

1. 粗糙集是一種軟計算方法軟計算（soft computing）

此概念是由模糊集創始人 L. A. Zadeh 所提出的。軟計算中的主要工具包括粗糙集，模糊邏輯，神經網路，機率推理，信度網路（belief networks），遺傳算法（GA）與混沌（chaos）理論等。而傳統的計算方法則稱為硬計算（hard computing），是使用精確、固定和不變的算法表達和解決問題。其中軟計算是利用所允許的不精確性，不確定性和部分真實性以得到易於處理和成本較低的解決方案，達到與現實系統最佳的協調。

2. 粗糙集方法相當簡單實用並具有以下幾個特點

(1) 它能處理各種數據，包括不完整的數據以及擁有眾多變量（parameters）的數據。

- (2) 它能處理數據的不精確性和模稜兩可性（ambiguity），包括確定性和非確定性兩種情況。
 - (3) 它能求得知識的最小表達和知識的各種不同顆粒（granularity）層次。
 - (4) 它能從數據中找出概念，簡單並且易於操作的模式（pattern）。
 - (5) 它能產生精確而又易於檢查和證實的規則，特別適於智能控制中規則（rule）的自動生成。
3. 粗糙集的應用

粗糙集是一門實用性很強的學科，從誕生到現在雖然只有二十幾年的時間，但已經在不少領域取得了豐碩的成果。例如在近似推理，數字邏輯分析和約簡，建立預測模型、決策支援、控制算法獲取、機器學習算法和模式識別等等。粗糙集已同時被廣泛地應用於人工智慧、醫療數據分析自動控制、模式識別、語言識別以及各種智能資訊處理等領域，並且獲得了豐碩的成果。以下簡單介紹粗糙集應用的幾個主要領域。

(1) 從數據庫中將「知識」發現

現代社會中隨著資訊產業的迅速發展，大量來自金融，醫療及科學研究等不同領域的資訊被儲存在數據庫中，這些浩瀚的數據間隱含著許多有價值但鮮為人知的相關性。例如股票的價格和一些經濟指數有什麼關係，手術前病患的病理指標可能與手術是否成功存在某種關聯等等。日本在這一方面也有長足的進展，特別是感性工學方面的應用。

由於數據庫的過於龐大，人工處理這些數據幾乎是不可能的，