**ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC SÀI GÒN KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KỲ**

**MÔN: CƠ SỞ DỮ LIỆU NÂNG CAO**

**ĐỀ TÀI: MÔ HÌNH CƠ SỞ DỮ LIỆU QUAN HỆ VỚI CÁC THUỘC TÍNH CÓ GIÁ TRỊ XÁC SUẤT CHO THÔNG TIN KHÔNG CHẮC CHẮN VÀ KHÔNG CHÍNH XÁC**

**Sinh viên thực hiện:** Lê Thị Lan Anh - 3122411006

Lê Thị Mỹ Hương – 3122411077

Đỗ Minh Quân - 3122411166

Trần Bùi Ty Ty - 3122411247

**Giảng viên phụ trách:** **PGS.TS.** Nguyễn Hòa

***Thành phố Hồ Chí Minh, năm 2024***

**LỜI CAM ĐOAN**

Nhóm chúng tôi xin cam đoan rằng:

Những kết quả nghiên cứu được trình bày trong tiểu luận cuối kỳ này là công trình nghiên cứu của chúng tôi dưới sự hướng dẫn của thầy ***PGS.TS*** *Nguyễn Hòa.*

Toàn bộ thông tin, dữ liệu và ý kiến từ nguồn khác được sử dụng trong đề tài này đã được trích dẫn và thể hiện đúng nguồn gốc của chúng. Chúng tôi đã tuân thủ các quy tắc và nguyên tắc về trích dẫn và tôn trọng bản quyền. Tất cả dữ liệu được sử dụng trong nghiên cứu đã được thu thập, xử lý và trình bày một cách trung thực và chính xác.

Chúng tôi không làm giả dữ liệu hoặc biểu đồ để làm cho kết quả trở nên sai lệch. Chúng tôi sử dụng tài liệu và nguồn tài nguyên của các tác giả khác một cách có trách nhiệm và chính trực.

Chúng tôi đã tuân thủ tất cả các quy định và yêu cầu của trường học và những nguyên tắc đạo đức trong việc sử dụng tài liệu từ nguồn khác. Tôi hiểu rằng vi phạm tính trung thực và đạo đức trong nghiên cứu có thể có hậu quả nghiêm trọng và có thể dẫn đến hình phạt từ trường học. Chúng tôi xem xét đây là một phần quan trọng của quá trình nghiên cứu và cam kết tuân thủ nghiêm ngặt các nguyên tắc này.

*Tháng 12 năm 2024*

***Trân trọng***

**LỜI CẢM ƠN**

Chúng tôi xin gửi lời biết ơn sâu sắc tới thầy Nguyễn Hòa về sự hướng dẫn, hỗ trợ và cơ hội mà thầy đã tạo ra cho chúng tôi trong quá trình thực hiện báo cáo về đề tài "Mô hình cơ sở dữ liệu quan hệ với các thuộc tính có giá trị xác suất cho thông tin không chắc chắn và không chính xác". Sự tận tâm của thầy đã giúp chúng tôi hiểu sâu hơn về chủ đề này, từ việc nắm bắt khái niệm cơ bản đến áp dụng kiến thức vào thực tế.

Chúng tôi rất trân trọng những kiến thức quý báu mà thầy đã chia sẻ và sự động viên không ngừng từ phía thầy. Quý thầy không chỉ là một người hướng dẫn mà còn là một nguồn động viên lớn lao đối với sự phát triển nghề nghiệp của chúng tôi.

Những kiến thức và kinh nghiệm mà chúng tôi học được từ thầy sẽ luôn là nền tảng vững chắc giúp chúng tôi tiến xa hơn trong sự nghiệp của mình. Một lần nữa, chúng tôi xin gửi lời biết ơn chân thành đến thầy Nguyễn Hòa vì sự hỗ trợ và sự cống hiến không ngừng.

*Tháng 12 năm 2024*

***Trân trọng***

**MỤC LỤC**

[DANH MỤC KÝ HIỆU 1](#_Toc184822579)

[DANH MỤC BẢNG 2](#_Toc184822580)

[LỜI MỞ ĐẦU 3](#_Toc184822581)

[CHƯƠNG 1. TỔNG QUAN VỀ MÔ HÌNH CÁC MÔ HÌNH CSDL 4](#_Toc184822582)

[**1.1. Mô hình cơ sở dữ liệu truyền thống – CRDB:** 4](#_Toc184822583)

[**1.2. Mô hình cơ sở dữ liệu quan hệ xác suất – PRDB:** 5](#_Toc184822584)

[**1.2.1. Loại đầu tiên C-1PRDB:** 5](#_Toc184822585)

[**1.2.2. Loại thứ hai C-2PRDB:** 5](#_Toc184822586)

[**1.2.3. So sánh C-1PRDB và C-2PRDB:** 7](#_Toc184822587)

[**1.2.4. Hạn chế của PRDB:** 7](#_Toc184822588)

[**1.3. Mô hình cơ sở dữ liệu xác suất với khoảng giá trị - IPRDB:** 8](#_Toc184822589)

[**1.3.1. Định nghĩa mô hình IPRDB:** 8](#_Toc184822590)

[**1.3.2. Cách xây dựng mô hình IPRDB:** 8](#_Toc184822591)

[CHƯƠNG 2. CƠ SỞ TOÁN HỌC ĐỂ PHÁT TRIỂN MÔ HÌNH IPRDB 10](#_Toc184822592)

[**2.1. Tổng quan giá trị xác suất** 10](#_Toc184822593)

[**2.2 Diễn giải xác suất của quan hệ nhị phân trên tập hợp** 10](#_Toc184822594)

[**2.3. Các chiến lược kết hợp xác suất** 12](#_Toc184822595)

[CHƯƠNG 3. MÔ HÌNH DỮ LIỆU IPRDB 16](#_Toc184822596)

[**3.1. Khái niệm về lược đồ IPRDB** 16](#_Toc184822597)

[**3.2. Phụ thuộc hàm IPRDB** 19](#_Toc184822598)

[**3.2. Khoá trong IPRDB** 21](#_Toc184822599)

[CHƯƠNG 4. ĐẠI SỐ QUAN HỆ VÀ TÍNH CHẤT CỦA IPRDB 22](#_Toc184822600)

[**4.1. Phép chọn – Selection** 22](#_Toc184822601)

[**4.2. Phép chiếu - Projection** 26](#_Toc184822602)

[**4.3. Phép tích – Descartes** 26](#_Toc184822603)

[**4.4. Phép kết hợp – Join** 27](#_Toc184822604)

[**4.5. Giao, hợp, hiệu - Intersection, Union and Difference** 29](#_Toc184822605)

[**4.6. Tính chất phép toán đại số** 34](#_Toc184822606)

[CHƯƠNG 5. MÔ HÌNH IPRDB TRONG BIỂU DIỄN VÀ XỬ LÝ THÔNG TIN 38](#_Toc184822607)

[**5.1. Mở rộng của IPRDB trong việc biểu diễn dữ liệu** 38](#_Toc184822608)

[**5.2. Hiệu quả của IPRDB trong xử lý dữ liệu** 39](#_Toc184822609)

[CHƯƠNG 6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 41](#_Toc184822610)

[**1. Kết luận** 41](#_Toc184822611)

[**2. Hướng phát triển** 41](#_Toc184822612)

[TÀI LIỆU THAM KHẢO 43](#_Toc184822613)

# **DANH MỤC KÝ HIỆU**

Các ký hiệu và quy ước chung sau đây được sử dụng trong bài báo cáo:

|  |  |
| --- | --- |
| **Ký hiệu** | **Ý nghĩa** |
|  | Quan hệ tập con cổ điển |
|  | Phép toán giao tập cổ điển |
|  | Phép toán hợp tập cổ điển |
|  | Quan hệ nhỏ hơn hoặc bằng trên trường các số thực/khoảng |
|  | Quan hệ lớn hơn hoặc bằng trên trường các số thực/khoảng |
|  | Phép toán hội xác suất của hai khoảng ứng với hai biến cố |
|  | Phép toán tuyển xác suất của hai khoảng ứng với hai biến cố |
|  | Phép toán trừ xác suất của hai khoảng ứng với hai biến cố |
| *Pr* | Hàm tính xác suất của một quan hệ/sự kiện |
| *prob* | Hàm tính xác suất của các quan hệ hai ngôi |
| *probS,I,o* | Hàm tính diễn dịch xác suất của các biểu thức |
| *min* | Hàm tính giá trị nhỏ nhất của một tập các số thực |
| *max* | Hàm tính giá trị lớn nhất của một tập các số thực |
|  | Tập tất cả các số tự nhiên |
| *℘* | Hàm ánh xạ của mỗi thuộc tính |
| θ | Một quan hệ nhị phân thuộc tập các ký hiệu: =, ≠, ≤, ≥, <, >, ⇒ |

# **DANH MỤC BẢNG**

[**CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI 4**](#_Toc183900380)

[Bảng 1.1 Ví dụ mô hình CSDL truyền thống 4](#_Toc184822683)

[Bảng 1.2 Ví dụ mô hình CSDL C-1PRDB 5](#_Toc184822684)

[Bảng 1.3 Ví dụ mô hình CSDL C-2PRDB 6](#_Toc184822685)

[Bảng 1.4 So sánh hai loại CSDL của PRDB 7](#_Toc184822686)

[Bảng 1.5 Ví dụ mô hình CSDL IPRDB 8](#_Toc184822687)

[CHƯƠNG 2.CƠ SỞ TOÁN HỌC ĐỂ PHÁT TRIỂN MÔ HÌNH IPRDB 10](#_Toc183900390)

[Bảng 2.1 Định nghĩa về chiến lược kết hợp xác suất 11](#_Toc183899760)

[Bảng 2.2 Lược đồ quan hệ PATIENT 13](#_Toc183899761)

[**CHƯƠNG 3. MÔ HÌNH DỮ LIỆU IPRDB** **16**](#_Toc184821724)

[Bảng 3.1 Quan hệ PATIENT 19](#_Toc183899768)

[Bảng 3.2 Quan hệ DOCTOR 20](#_Toc183899769)

[CHƯƠNG 4. ĐẠI SỐ QUAN HỆ VÀ TÍNH CHẤT CỦA IPRDB 22](#_Toc184821728)

[Bảng 4.1 Quan hệ σφ (PATIENT) 22](#_Toc183899781)

[Bảng 4.2 Quan hệ Π{P\_AGE, P\_DISEASE, D\_COST} ⊕in (PATIENT) 25](#_Toc183899782)

[Bảng 4.3 Quan hệ PATIENT1 28](#_Toc183899783)

[Bảng 4.4 Quan hệ PATIENT2 28](#_Toc183899784)

[Bảng 4.5 Quan hệ PATIENT3 = PATIENT1⋈⊗in PATIENT2 29](#_Toc183899785)

[Bảng 4.6 Quan hệ PATIENT4 30](#_Toc183899786)

[Bảng 4.7 Quan hệ PATIENT3 ∩⊗ PATIENT4 31](#_Toc183899787)

[Bảng 4.8 Quan hệ DIAGNOSE1 32](#_Toc183899788)

[Bảng 4.9 Quan hệ DIAGNOSE2 32](#_Toc183899789)

[Bảng 4.10 Quan hệ DIAGNOSE1 ∪⊕in DIAGNOSE2 33](#_Toc183899790)

[Bảng 4.11 Quan hệ DIAGNOSE1 ∪⊝in DIAGNOSE2 34](#_Toc183899791)

# **LỜI MỞ ĐẦU**

Trong thế giới hiện đại, dữ liệu đóng vai trò vô cùng quan trọng trong mọi lĩnh vực từ khoa học, kinh doanh đến công nghệ thông tin. Tuy nhiên, không phải lúc nào dữ liệu thu thập được cũng đảm bảo tính chính xác và chắc chắn. Sự không chắc chắn và không chính xác có thể xuất hiện do các yếu tố như lỗi đo lường, thông tin không đầy đủ hoặc sự không rõ ràng trong việc thu thập và xử lý thông tin.

Để giải quyết vấn đề này, mô hình cơ sở dữ liệu truyền thống cần được mở rộng để hỗ trợ việc quản lý và xử lý dữ liệu với các giá trị có xác suất, nhằm phản ánh tốt hơn bản chất thực tế của thông tin. Mô hình cơ sở dữ liệu quan hệ với các thuộc tính có giá trị xác suất chính là một hướng tiếp cận sáng tạo, cung cấp nền tảng lý thuyết và công cụ thực tiễn để lưu trữ và truy vấn dữ liệu không chắc chắn.

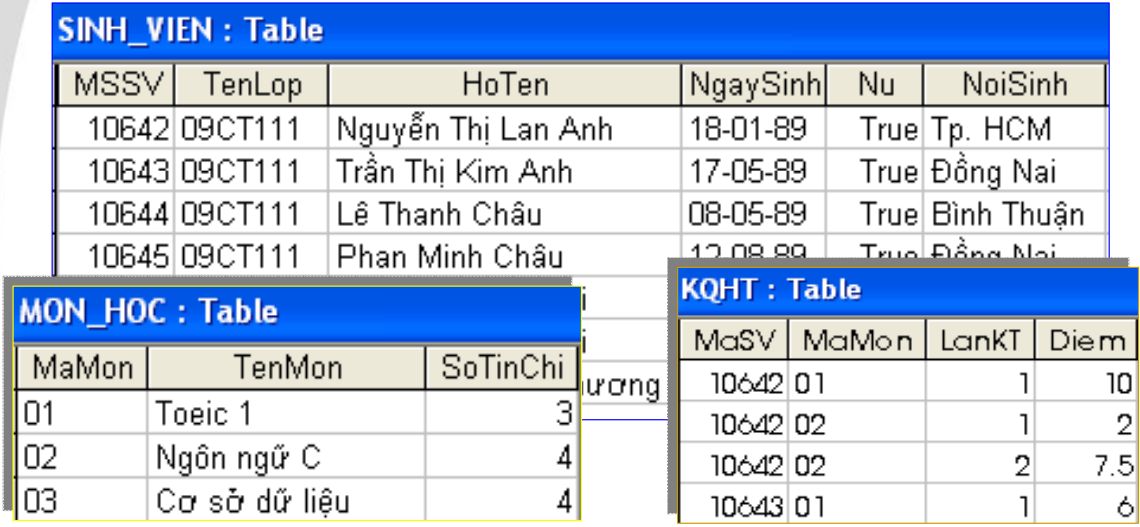
***"Mô hình cơ sở dữ liệu quan hệ với các thuộc tính có giá trị xác suất cho thông tin không chắc chắn và không chính xác"*** hướng đến việc nghiên cứu và phát triển một mô hình cơ sở dữ liệu đáp ứng được các yêu cầu này. Đề tài không chỉ phân tích những hạn chế của các mô hình truyền thống mà còn đề xuất các giải pháp mở rộng, bao gồm cấu trúc dữ liệu, toán tử và ngôn ngữ truy vấn phù hợp, từ đó giúp cải thiện khả năng xử lý dữ liệu không chắc chắn trong các hệ thống thông tin hiện đại.

Thông qua nghiên cứu này, chúng tôi hy vọng sẽ góp phần vào sự phát triển của khoa học dữ liệu, đồng thời mở ra những hướng đi mới trong việc ứng dụng cơ sở dữ liệu vào các lĩnh vực yêu cầu xử lý thông tin phức tạp và không chắc chắn.

# **CHƯƠNG 1. TỔNG QUAN VỀ MÔ HÌNH CÁC MÔ HÌNH CSDL**

## **1.1. Mô hình cơ sở dữ liệu truyền thống – CRDB:**

CRDB là mô hình cơ sở dữ liệu cơ bản nhất, được thiết kế để lưu trữ và truy vấn dữ liệu có cấu trúc rõ ràng. Có lợi trong việc mô hình hóa, thiết kế và triển khai các hệ thống quy mô lớn.



Hình 1.1 Ví dụ mô hình cơ sở dữ liệu truyền thống

Tuy nhiên, trong CRDB, các giá trị thường được định nghĩa bằng cách sử dụng các kiểu dữ liệu cố định (như số nguyên, số thực, hoặc chuỗi ký tự). Vì vậy, CRDB không hỗ trợ biểu diễn trực tiếp giá trị như "~500" hoặc "[0.5, 0.7]", không hỗ trợ logic mờ hoặc tính toán xác suất, gây hạn chế trong các ứng dụng cần phân tích hoặc suy luận dựa trên dữ liệu không chắc chắn. CRDB thay thế giá trị mặc định bằng giá trị NULL hoặc loại bỏ các bản ghi không đầy đủ gây khó khăn trong việc diễn giải và xử lý, dẫn đến kết quả không trực quan hoặc không chính xác. Các ràng buộc toàn vẹn dữ liệu (như khóa ngoại, duy nhất) trong CRDB yêu cầu dữ liệu phải chính xác gây khó khăn khi cần biểu diễn dữ liệu có sai số hoặc mơ hồ.

|  |  |  |  |
| --- | --- | --- | --- |
| **P\_ID** | **D\_ID** | **P\_DISEASE** | **D\_COST** |
| P226 | D014 | Lung Cancer | $30 |
| P226 | D014 | Tuberculosis | $35 |
| P228 | D016 | Lung Cancer | $30 |
| P262 | D022 | Dyspepsia | $5 |

Bảng 1.1 Ví dụ mô hình CSDL truyền thống

* Một bệnh chỉ có thể được ghi nhận là tồn tại hoặc không tồn tại.
* Không có biểu diễn nào cho mức độ không chắc chắn về bệnh. Vì vậy, chỉ có một hàng cho mỗi kết hợp giữa P\_ID (ID bệnh nhân) và P\_DISEASE (Bệnh). Chẳng hạn, P226, họ được chẩn đoán mắc Lung Cancer và Tuberculosis. Tuy nhiên, không có thông tin về xác suất hoặc mức độ tin cậy của các chẩn đoán này, tất cả đều được coi là đúng với mức tin cậy 100%.

## **1.2. Mô hình cơ sở dữ liệu quan hệ xác suất – PRDB:**

PRDB là một mô hình được mở rộng từ CRDB thiết kế để xử lý dữ liệu không chắc chắn bằng cách gắn xác suất cho các giá trị hoặc các mối quan hệ trong cơ sở dữ liệu. PRDB có 2 loại:

### **1.2.1. Loại đầu tiên C-1PRDB:**

Mỗi dòng dữ liệu được gán một xác suất trong khoảng [0, 1] để biểu diễn mức độ không chắc chắn. Các phép toán đại số của C-1PRDB được định nghĩa bằng cách mở rộng các phép toán đại số của CRDB thông qua việc tính toán và kết hợp xác suất của các bộ dữ liệu trong các quan hệ C-1PRDB. Tuy nhiên vì không thể biểu diễn các mối quan hệ phụ thuộc phức tạp giữa các giá trị hoặc thực thể. Điều này hạn chế ứng dụng trong các tình huống mà dữ liệu có sự tương quan cao.

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Tên** | **Vị trí** | **Xác suất** |
| 1 | An | Kỹ sư | 0.9 |
| 2 | Bình | Quản lý | 0.8 |

Bảng 1.2 Ví dụ mô hình CSDL C-1PRDB

### **1.2.2. Loại thứ hai C-2PRDB:**

Mỗi giá trị của thuộc tính được gán xác suất trong khoảng [0, 1] để biểu diễn mức độ không chắc chắn của giá trị đó. Các phép toán đại số của C-2PRDB được định nghĩa bằng cách mở rộng các phép toán của CRDB và sử dụng các phép toán trên xác suất hoặc phân phối xác suất để tính toán và kết hợp xác suất của các giá trị thuộc tính.

|  |  |  |
| --- | --- | --- |
| **ID** | **Tên** | **Kinh nghiệm quản lý** |
| 1 | An | Có |
| 2 | Bình | Không |

|  |  |
| --- | --- |
| **ID dự án** | **Lĩnh vực** |
| P1 | Công nghệ |
| P2 | Y tế |

|  |  |  |
| --- | --- | --- |
| **ID nhân viên** | **ID dự án** | **Xác suất làm Quản lý dự án** |
| 1 | P1 | 0.9 |
| 1 | P2 | 0.6 |
| 2 | P1 | 0.3 |
| 2 | P2 | 0.4 |

Bảng 1.3 Ví dụ mô hình CSDL C-2PRDB

### **1.2.3. So sánh C-1PRDB và C-2PRDB:**

|  |  |  |
| --- | --- | --- |
| **Đặc điểm** | **C-1PRDB** | **C-2PRDB** |
| **Xác suất gắn với** | Giá trị đơn lẻ | Giá trị và mối quan hệ phụ thuộc |
| **Độc lập dữ liệu** | Giả định dữ liệu độc lập | Xử lý cả các phụ thuộc giữa giá trị |
| **Phức tạp tính toán** | Đơn giản, nhanh chóng | Phức tạp, yêu cầu tài nguyên lớn |
| **Ứng dụng** | Thích hợp cho dữ liệu đơn giản | Phù hợp với dữ liệu phức tạp, có phụ thuộc |
| **Khả năng biểu diễn** | Hạn chế với các quan hệ phức tạp | Biểu diễn được các mối quan hệ phức tạp |

Bảng 1.4 So sánh hai loại CSDL của PRDB

### **1.2.4. Hạn chế của PRDB:**

PRDB giả định rằng xác suất cho mỗi sự kiện là một giá trị cụ thể, tuy nhiên, trong thực tế, xác suất thường không thể xác định chính xác mà chỉ nằm trong một khoảng giá trị. Việc sử dụng PRDB dễ dẫn đến sai lệch nếu dữ liệu đầu vào không đủ chính xác, gây khó khăn trong việc đảm bảo tính toàn vẹn khi dữ liệu xác suất không chính xác hoặc không đầy đủ và khó biểu diễn các mối quan hệ xác suất phức tạp khi có sự chồng chéo hoặc không chắc chắn giữa các giá trị.

***Ví dụ những vấn đề trong thực tế mà PRDB không thể giải quyết:***

* Dự báo thời tiết: "Xác suất mưa nằm trong khoảng [60%, 80%]".
* Chẩn đoán bệnh: "Bệnh nhân có khả năng mắc bệnh trong khoảng [40%, 60%]".
* Ước tính doanh thu trong tương lai với xác suất nằm trong khoảng [30%, 60%].

## **1.3. Mô hình cơ sở dữ liệu xác suất với khoảng giá trị - IPRDB:**

### **1.3.1. Định nghĩa mô hình IPRDB:**

Do những hạn chế từ CRDB và PRDB, mô hình IPRDB được đề xuất. Mô hình này sử dụng các khoảng xác suất thay vì các giá trị xác suất đơn lẻ để biểu diễn thông tin không chắc chắn, giúp tăng cường tính linh hoạt và khả năng xử lý thông tin mơ hồ.

|  |  |  |  |
| --- | --- | --- | --- |
| **P\_ID** | **D\_ID** | **P\_DISEASE** | **D\_COST** |
| P226 | D014 | {(lung cancer, [0.3, 0.6]), (tuberculosis, [0.4, 0.7])} | {($30, [0.3, 0.4]), ($35, [0.6, 0.7])} |
| P228 | D016 | {(lung cancer, [1, 1])} | {($30, [1, 1])} |
| P255 | D020 | {(hepatitis, [0.58, 0.96]), (pancreatitis, [0.2, 0.7]), (cholecystitis, [0.2, 0.6])} | {($7, [0.2, 0.4]), ($8, [0.76, 1])} |

Bảng 1.5 Ví dụ mô hình CSDL IPRDB

* Mỗi bệnh của bệnh nhân được biểu diễn kèm theo khoảng xác suất thay vì chỉ một giá trị xác suất cụ thể. Điều này cho biết mức độ khả năng bệnh nhân có thể mắc bệnh nằm trong khoảng nào.
* Ví dụ, P226 có khả năng mắc ung thư phổi với xác suất nằm trong khoảng [0.3, 0.6] nên mức độ tin cậy là từ 30% đến 60%. Cũng có khả năng bệnh nhân này mắc lao phổi với xác suất nằm trong khoảng [0.4, 0.7] (mức độ tin cậy 40% đến 70%).
* Đối với bệnh nhân P228, mức độ chắc chắn về chẩn đoán ung thư phổi là hoàn toàn, với xác suất [1, 1] (100%).

### **1.3.2. Cách xây dựng mô hình IPRDB:**

Để xây dựng IPRDB có 3 phương pháp chính:

1. Đề xuất các giá trị xác suất của kiểu dữ liệu để biểu diễn các thuộc tính có giá trị không chắc chắn và không chính xác
2. Sử dụng các phép toán trên các khoảng xác suất và diễn giải xác suất của quan hệ nhị phân trên các tập hợp để tính toán mức độ không chắc chắn của các phụ thuộc hàm, khóa và quan hệ trên xác miền giá trị của các thuộc tính.
3. Định nghĩa các chiến lược kết hợp các gái trị xác suất để phát triển các phép toán đại số quan hệ mới.

Mô hình IPRDB mang đến một giải pháp toàn diện và linh hoạt cho việc biểu diễn và xử lý dữ liệu không chắc chắn. Việc sử dụng khoảng xác suất giúp phản ánh rõ hơn các mức độ tin cậy của thông tin và mở rộng khả năng ứng dụng trong thực tế. Đồng thời, các chiến lược kết hợp và phép toán xác suất được phát triển trong mô hình này cho phép khai thác dữ liệu không chắc chắn hiệu quả hơn, đặc biệt trong các lĩnh vực yêu cầu xử lý thông tin phức tạp như y tế, tài chính, và khoa học dữ liệu.

# **CHƯƠNG 2. CƠ SỞ TOÁN HỌC ĐỂ PHÁT TRIỂN MÔ HÌNH IPRDB**

## **2.1. Tổng quan giá trị xác suất**

**Định nghĩa 2.1:**

Giả sử τ là một kiểu dữ liệu và D là miền của τ. Một giá trị xác suất trên miền của τ là một tập hợp hữu hạn các cặp {(v₁, [l₁, u₁]), ..., (vₘ, [lₘ, uₘ])}, trong đó:

vᵢ ∈ D và 0 ≤ lᵢ ≤ uᵢ ≤ 1, với mỗi i = 1, 2, ..., m

Giá trị xác suất pv = {(v₁, [l₁, u₁]), ..., (vₘ, [lₘ, uₘ])} biểu thị rằng giá trị của pv là một phần tử trong tập {v₁, ..., vₘ} và xác suất để giá trị của pv là vᵢ nằm trong khoảng [lᵢ, uᵢ].

Do đó, một giá trị xác suất vừa thể hiện sự không chắc chắn của giá trị, vừa thể hiện sự mơ hồ trong xác suất của giá trị đó. Một giá trị xác suất pv = {(v₁, [l₁, u₁]), ..., (vₘ, [lₘ, uₘ])} tương ứng với một hàm phân phối xác suất p trên tập V = {v₁, ..., vₘ}, sao cho p(vᵢ) ∈ [lᵢ, uᵢ] và ∑ p(vᵢ) ≤ 1**.**

Giả sử bệnh của một bệnh nhân được chẩn đoán là viêm gan (hepatitis) với xác suất nằm trong khoảng từ 0.5 đến 0.7 hoặc viêm túi mật (cholecystitis) với xác suất nằm trong khoảng từ 0.3 đến 0.5. Thông tin này có thể được biểu diễn dưới dạng giá trị xác suất:

*pv = {(hepatitis, [0.5, 0.7]), (cholecystitis, [0.3, 0.5])}*

Lưu ý: Một giá trị xác suất có thể được ký hiệu bởi pv = {(v, I) | v ∈ D, I = [l, u] ⊆ [0, 1]}.

## **2.2 Diễn giải xác suất của quan hệ nhị phân trên tập hợp**

Để tính mức độ không chắc chắn của các mối quan hệ trên các giá trị thuộc tính trong IPRDB, ta sử dụng cách giải thích xác suất của quan hệ nhị phân trên các tập hợp[2] như dưới đây:

|  |  |
| --- | --- |
| **Chiến lược** | **Phép toán** |
| **Bỏ qua (Ignorance)** | ([l1, u1] ⊗ig [l2, u2]) = [max (0, l1 + l2 - 1), min (u1, u2)]  ([l1, u1] ⨁ig [l2, u2]) = [max (l1, l2), min (1, u1 + u2)]  ([l1, u1] **⊖**ig [l2, u2]) = [max (0, l1 - u2), min (u1,1- l2)] |
| **Độc lập (Independence)** | ([l1, u1] ⊗in [l2, u2]) = [(l1. l2, u1. u2)]  ([l1, u1] ⨁in [l2, u2]) = [l1 +l2 – (l1.l2), u1 + u2 – (u1. u2)]  ([l1, u1] **⊖**in [l2, u2]) = [l1. (1 - u2), u1. (1- l2)] |
| **Tương quan thuận (Positive correlation)** | ([l1, u1] ⊗pc [l2, u2]) = [min (l1, l2), min (u1, u2)]  ([l1, u1] ⨁ pc [l2, u2]) = [max (l1, l2), max (u1, u2)]  ([l1, u1] **⊖** pc [l2, u2]) = [max (0, l1 - u2), max (0, u1- l2)] |
| **Loại trừ nhau (Mutual Exclusion)** | ([l1, u1] ⊗me [l2, u2]) = [0, 0]  ([l1, u1] ⨁ me [l2, u2]) = [min (1, l1+l2), min (1, u1+u2)]  ([l1, u1] **⊖** me [l2, u2]) = [l1, min (u1, 1- l2)] |

Bảng 2.1 Định nghĩa về chiến lược kết hợp xác suất

**Định nghĩa 2.2:**

Khái niệm "Probabilistic interpretation" (diễn giải xác suất) của các mối quan hệ nhị phân giữa các tập hợp A và B, với các giá trị xác suất nằm trong khoảng [0,1]. Dưới đây là giải thích chi tiết:

Pr(AθB): Đây là xác suất cho mối quan hệ nhị phân giữa A và B, ký hiệu AθB, với θ là một quan hệ nhị phân thuộc tập các ký hiệu: =, ≠, ≤, ≥, <, >, ⇒.

Pr(AθB) được xác định bởi hai cách khác nhau:

***Cách 1:*** Pr(AθB) = p(uθv∣ u∈A, v∈B)

***Cách 2:*** Pr(A⇒B) = p(u∈B∣ u∈A):

**Ví dụ 2.2:** Một số diễn giải xác suất cụ thể về các quan hệ tập hợp dựa trên miền các số tự nhiên:

* Pr({3,4} = {4,5}) = 0.25: Vì các trường hợp có thể xảy ra gồm {3,4}, {3,5}, {4,4}, {4,5}. Nghĩa là chỉ có một phần tư khả năng rằng các phần tử của hai tập này bằng nhau theo quan hệ nhị phân.
* Pr ({3,4} ⇒ {4,5}) = 0.5: Đây là xác suất có điều kiện rằng nếu phần tử nằm trong tập {3,4}, thì nó cũng nằm trong tập {4,5}, với giá trị xác suất là 0.5. Vì 3 không thuộc tập {4,5}, 4 thuộc tập {4,5}. Nghĩa là chỉ có một phần hai khả năng thỏa mãn điều kiện.
* Pr (3=5) = 0.0: Đây là xác suất rằng 3=5, được tính bằng 0 vì điều này là không thể xảy ra.
* Pr (5⇒ {5,6}) =1.0: Đây là xác suất rằng nếu phần tử là 5, thì nó cũng nằm trong tập {5,6}, và điều này chắc chắn xảy ra với xác suất là 1.0.

## **2.3. Các chiến lược kết hợp xác suất**

Xác suất của một sự kiện có thể không được xác định hoặc tính toán chính xác. Khi đó, một khoảng xác suất có thể được sử dụng thay vì một giá trị. Giả sử hai sự kiện e​1 và e2​ có xác suất nằm trong các khoảng: [l1​, u1​] cho sự kiện e1 và [l2, u2​] cho sự kiện e2. Khi đó, khoảng xác suất của các chiến lược

* Tổ hợp (Conjunction) e1 ∧ e2,
* Rời rạc (Disjunction) e1 ∨ e2 và
* Hiệu (Difference) e1 \ e2

Có thể được tính toán bằng các chiến lược thay thế. Ta áp dụng các chiến lược conjunction, disjunction và difference như đã được trình bày trong Bảng 6. Trong bảng các toán tử ⊗, ⊕, và ⊖ lần lượt biểu diễn các phép giao, hợp và hiệu.

Ký hiệu [l1, u1] ⊆ [l2, u2] được sử dụng để biểu thị rằng l2 ≤ l1 và u1 ≤ u2. Ngoài ra, một giá trị xác suất duy nhất p có thể được xem như một khoảng [p, p] và phép toán p. [l, u] được tính dạng [p. l, p. u].

**2.4. Các chiến lược tổ hợp, rời rạc, và hiệu của các giá trị xác suất**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **P\_ID** | **P\_NAME** | **P\_AGE** | **P\_DISEASE** | **D\_COST** |
| P104 | John | {(65, [1, 1])} | {(lung cancer, [0.5, 0.5]),  (tuberculosis, [0.5, 0.5])} | {($30, [0.3, 0.6]),  ($35, [0.4, 0.7])} |
| P218 | Paul | {(43,[0.5,0.5]),  (44,[0.5, 0.5])} | {(hepatitis, [0.3, 0.5]),  (cirrhosis, [0.5, 0.7])} | {($6, [0.4, 0.6]),  ($7, [0.4, 0.6])} |
| P325 | Helen | {(36, [1, 1])} | {(duodenitis, [0.5, 0.5]),  (gastritis, [0.5, 0.5])} | {($8, [0.5, 0.5]),  ($9, [0.5, 0.5])} |
| P412 | Anne | {(15, [1, 1])} | {(bronchitis, [1, 1])} | {($7, [1, 1)} |
| P426 | George | {(36, [1, 1])} | {(duodenitis, [0.4, 0.5]),  (gastritis, [0.5, 0.6])} | {($8, [0.3, 0.5]),  ($9, [0.5, 0.7])} |

Bảng 2.2 Lược đồ quan hệ PATIENT

Để xây dựng các phép toán đại số quan hệ xác suất trong IPRDB, chẳng hạn như phép chiếu, phép nối, phép giao, phép hợp và phép hiệu, đề xuất các phép toán hội (conjunction), tuyển (disjunction) và hiệu (difference) của các giá trị xác suất làm cơ sởđể kết hợp xác suất của các giá trị không chắc chắn và không chính xác của các thuộc tính trong các quan hệ kết quả của các phép toán đại số này. Sự kết hợp của các giá trị xác suất được định nghĩa như sau.

**Định nghĩa 2.3**: Conjunction (Tổ hợp)

Giả sử pv1 và pv2 là hai giá trị xác suất và ⊗ là một chiến lược hội xác suất (probabilistic conjunction strategy). Phép hội của pv1 và pv2 dưới chiến lược ⊗, ký hiệu là pv1 ⊗ pv2: pv = {(v, I1 ⊗ I2) | (v, I1) ∈ pv1, (v, I2) ∈ pv2.

Kết quả của phép tổ hợp này là một giá trị xác suất pw được định nghĩa bằng pw= {(v, I1⊗I2)}, với v là các giá trị có trong pv1 hoặc pv2, và I1 ⊗ I2 là khoảng xác suất của chúng sau khi áp dụng phép hội (tổ hợp).

**Ví dụ 2.3:** Nếu pv1 = {(hepatitis, [0.4, 0.6]), (cholecystitis, [0.4, 0.6])} và pv2= {(hepatitis, [1.0, 1.0])}, phép toán hội xác suất pv1 ⊗in pv2 dưới chiến lược độc lập sẽ cho ra giá trị xác suất pw = {(hepatitis, [0.4, 0.6])}.

**Định nghĩa 2.4:** Disjunction (Rời rạc)

Cho pv1​ và pv2 là hai giá trị xác suất, ⊕ là chiến lược tuyển xác suất (probabilistic disjunction strategy). Phép tuyển của pv1​ và pv2​ dưới chiến lược ⊕, ký hiệu là pv1⊕ pv2, là giá trị xác suất pv được định nghĩa bởi:

pv = {(v, I1​) ∣ (v, I1​) ∈ pv1​ và ¬∃I2, (v, I2) ∈ pv2} ∪ {(v, I2​)∣(v, I2​)∈pv2​ và ¬∃(v,I1​)∈pv1​} ∪ {(v,I1​⊕I2​)∣(v,I1​)∈pv1​ và (v,I2​)∈pv2​}.

Kết quả của phép này là một giá trị xác suất pw được định nghĩa bằng pw= {(v, I1)} nếu v nằm trong pv1 nhưng không nằm trong pv2, và ngược lại. Nếu v có trong cả hai, thì khoảng xác suất là tổ hợp các giá trị từ pv1 và pv2.

**Ví dụ 2.4:** Giả sử pv1 = {(hepatitis, [0.2, 0.5]), (cholecystitis, [0.3, 0.6])} và pv2 = {(hepatitis, [0.3, 0.5]), (pancreatitis, [0.2, 0.6])} là các giá trị xác suất, khi đó pv1​⊕in​ pv2​ dưới chiến lược rời rạc xác suất là giá trị xác suất được tính như sau:

Ta có: phân bố xác suất

* pv1 = {(hepatitis, [0.2, 0.5]), (cholecystitis, [0.3, 0.6])}
* pv2 = {(hepatitis, [0.3, 0.5]), (pancreatitis, [0.2, 0.6])}

Giữ nguyên giá trị chồng lấp. Hợp giá trị “hepatitis” bằng chiến lược tuyển độc lập:

* Hepatitis = ([0.2, 0.5] ⨁in [0.3, 0.5]) = [0.2+0.3 – (0.2.0.3), 0.3+ 0.5– (0.3. 0.5)] = [0.44, 0.75]

Kết quả cuối cùng của phân bố xác suất sau khi hợp pv1 và pv2 là:

* pv = {(cholecystitis, [0.3, 0.6]), (pancreatitis, [0.2, 0.6]), (hepatitis, [0.44, 0.75])}.

**Định nghĩa 2.5:** Difference (Phép hiệu xác suất)

Giả sử pv1​ và pv2​ là hai giá trị xác suất và ⊖ là chiến lược hiệu xác suất (probabilistic difference strategy). Hiệu của pv1​ và pv2​ dưới ⊖, được ký hiệu là pv1⊖pv2​, là giá trị xác suất pv được định nghĩa bởi:

pv = {(v, I1​) ∣ (v, I1​) ∈ pv1​ và ∃ (v, I2​) ∈ pv2​} ∪ {(v, I1​⊖I2​) ∣ (v, I1​) ∈ pv1​ và (v, I2​) ∈ pv2​}. Trong đó, ⊖ biểu thị sự hiệu của khoảng xác suất liên quan đến hai giác trị xác suất pv1 và pv2.

**Ví dụ 2.5:** Xét bệnh nhân John với thuộc tính D COST (Chi phí điều trị hằng ngày). Trong thuộc tính D COST (Chi phí điều trị) của bệnh nhân John, hai giá trị $30 [0.3,0.6] và $35 [0.4,0.7] có các khoảng xác suất.

Áp dụng chiến lược hiệu bỏ qua:

([l1, u1] **⊖**ig [l2, u2])

= [max (0, l1 - u2), min (u1,1- l2)]

= ([0.3,0.6] **⊖**ig [0.4,0.7])

= [max (0, 0.3 – 0.7), min (0.6,1- 0.4)]

= [0,0.6].

# **CHƯƠNG 3. MÔ HÌNH DỮ LIỆU IPRDB**

Mô hình dữ liệu IPRDB là một cấu trúc với các khái niệm cơ bản như lược đồ, quan hệ xác suất và cơ sở dữ liệu để thể hiện dữ liệu và mối quan hệ giữa chúng.

## **3.1. Khái niệm về lược đồ IPRDB**

Một lược đồ IPRDB bao gồm một tập hợp các thuộc tính quan hệ được liên kết tương ứng với các miền xác định các giá trị xác suất của các thuộc tính đó. Sơ đồ IPRDB là phần mở rộng của sơ đồ CRDB với các thuộc tính có giá trị không chắc chắn và không chính xác. Lược đồ IPRDB được định nghĩa như sau:

**Định nghĩa 3.1:**

Lược đồ IPRDB là một cặp *R = (****U****, ℘),* trong đó :

1. ***U***= *{A*1*, A*2*, . . . , Ak}:* là tâp hợp các thuộc tính đôi một khác nhau.
2. ℘ là một hàm ánh xạ mỗi thuộc tính A ∈ ***U*** đến tập hợp tất cả các giá trị xác suất trên miền của A.

Để đơn giản hoá, chúng ta có thể sử dụng ký hiệu *R = (****U****, ℘)* và *R* để biểu thị *R = (****U****, ℘).* Miền giá trị của A được ký hiệu là dom(A).

**Ví dụ 3.1:** Giả sử ta có một lược đồ IPRDB cho thông tin về bệnh nhân tại một bệnh viện, với các thuộc tính như mã bệnh nhân, tên bệnh nhân, tuổi, bệnh, và chi phí điều trị hàng ngày. Lược đồ IPRDB này được biểu diễn dưới dạng một cặp *R = (****U****, ℘),* trong đó: ***U*** là tập hợp các thuộc tính *P\_ID, P\_NAME, P\_AGE, P\_DISEASE,* và *D\_COST* lần lượt mô tả thông tin về mã định danh, tên, tuổi, bệnh, và chi phí điều trị hàng ngày của từng bệnh nhân. ℘ một hàm ánh xạ mỗi thuộc tính A ∈ ***U*** đến một tập hợp các giá trị xác suất trên miền của thuộc tính đó. Ví dụ như thuộc tính P\_NAME có các bệnh nhân John, Paul, Helen, Anne, Geogre thì ta có ℘(P\_NAME) là {{(John, [1,1])}, {(Paul, [1,1])}, {(Helen, [1,1])}, {(Anne, [1,1])}, {(George, [1,1])}}

Một quan hệ IPRDB hoặc một quan hệ xác suất của IPRDB là một thể hiện của lược đồ IPRDB, trong đó mỗi thuộc tính quan hệ được liên kết với một giá trị xác suất để biểu diễn một giá trị không chắc chắn và không chính xác mà thuộc tính đó có thể nhận như định nghĩa dưới.

**Định nghĩa 3.2: *U***= *{A*1*, A*2*, . . . , Ak}* là tâp hợp k các thuộc tính đôi một khác nhau. Một quan hệ IPRDB 𝑟trên lược đồ *R = (****U****, ℘),* là tập hợp hữu hạn các phần tử *{t1, t2,. . , tn},* trong đó *ti = (pʋi1, pʋi2, . . . , pʋik)* là một danh sách gồm k giá trị xác suất pʋij={(ʋij, [lij, uij ]) | *ʋ*ij ∈ dom(Aj), [lij , uij] ⊆ [0*,* 1]}, j = 1, 2, *. . .*, *k* sao cho *pʋij* ∈ *℘*(*Aj* ) với mọi i = 1, 2, . . . , n.Mỗi phần tử ti trong quan hệ 𝑟 trên R(U,℘) được gọi là một bộ (tuple) trên ***U.*** Giá trị xác suất pʋij biểu diễn giá trị không chắc chắn và không chính xác của thuộc tính***A***j trong bộ *ti.* Chúng ta viết *ti.A*j hoặc *ti*[***A****j*] để biểu thị pʋij và viết [ti] để thay thế cho (Vi1, Vi2, . . . , Vik) trong đó Vij = {ʋij | (ʋij, [lij, uij]) ∈ *pʋij* }. Với mỗi tập hợp các thuộc tính *H ⊆* *{A*1*, A*2*, . . . , Ak}* ký hiệu *ti*[H] biểu diễn phần còn lại của bộ ti sau khi loại bỏ các giá trị của những thuộc tính không thuộc *H.* Ngoài ra chúng ta chỉ quan tâm đến một quan hệ duy nhất trên một lược đồ, thì có thể tống nhất tên của quan hệ và tên lược đồ đó.

**Ví dụ 3.2:** Một cơ sở dữ liệu về bệnh nhân tại phòng khám của một bệnh viện, một quan hệ IPRDB đơn giản, có tên là PATIENT, dựa trên lược đồ IPRDB PATIENT({P ID, P NAME, P AGE, P DISEASE, D COST}, ℘) có thể được biểu diễn như ***Bảng 8***.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **P\_ID** | **P\_NAME** | **P\_AGE** | **P\_DISEASE** | **D\_COST** |
| P104 | John | {(65, [1, 1])} | {(lung cancer, [0.5, 0.5]),  (tuberculosis, [0.5, 0.5])} | {($30, [0.3, 0.6]),  ($35, [0.4, 0.7])} |
| P218 | Paul | {(43, [0.5, 0.5]),  (44, [0.5, 0.5])} | {(hepatitis, [0.3, 0.5]),  (cirrhosis, [0.5, 0.7])} | {($6, [0.4, 0.6]),  ($7, [0.4, 0.6])} |
| P325 | Helen | {(36, [1, 1])} | {(duodenitis, [0.5, 0.5]),  (gastritis, [0.5, 0.5])} | {($8, [0.5, 0.5]),  ($9, [0.5, 0.5])} |
| P412 | Anne | {(15, [1, 1])} | {(bronchitis, [1, 1])} | {($7, [1, 1)} |
| P426 | George | {(36, [1, 1])} | {(duodenitis, [0.4, 0.5]),  (gastritis, [0.5, 0.6])} | {($8, [0.3, 0.5]),  ($9, [0.5, 0.7])} |

Bảng 3.1 Quan hệ PATIENT

Trong quan hệ này, các thuộc tính *P\_ID, P\_NAME, P\_AGE, P\_DISEASE,* và *D\_COST* lần lượt mô tả thông tin về mã định danh, tên, tuổi, bệnh, và chi phí điều trị hàng ngày của từng bệnh nhân. Trong thực tế, khi chẩn đoán, các bác sĩ có thể không chắc chắn về bệnh của bệnh nhân. Ngoài ra, ngay cả khi biết bệnh, chi phí điều trị hàng ngày của bệnh nhân cũng không chắc chắn. Ví dụ, thông tin về bệnh nhân John cho biết John 65 tuổi, bệnh của John có thể là ung thư phổi hoặc lao với xác suất 0.5 cho mỗi bệnh. Chi phí điều trị hàng ngày của John có thể là 30$ với xác suất trong khoảng [0.3, 0.6] hoặc 35$ với xác suất trong khoảng [0.4, 0.7].

Lưu ý rằng, với mỗi thuộc tính ***A*** trong lược đồ **PATIENT**, *℘(*A) bao gồm tất cả các giá trị xác suất trên miền của ***A*** ***(Định nghĩa 3.1)*.** Ngoài ra, để đơn giản hóa, mỗi giá trị xác suất {(ʋ,[1, 1])} sẽ được biểu diễn dưới dạng một giá trị đơn ***ʋ*** (như các giá trị xác suất của thuộc tính P\_ID). Vì nếu một thuộc tính nhận giá trị xác suất như vậy, thì nó chỉ nhận giá trị ʋ với xác suất 1.0 **(*Định nghĩa 2.1*)**. Nói cách khác, thuộc tính chắc chắn nhận giá trị ʋ.

Cơ sở dữ liệu quan hệ IPRDB là một phần mở rộng của CRDB, với các lược đồ và quan hệ IPRDB như được định nghĩa dưới đây.

**Định nghĩa 3.3**: Cơ sở dữ liệu quan hệ IPRDB trên một tập hợp các thuộc tính là một tập hợp các mối quan hệ IPRDB tương quan với tập hợp các lược đồ IPRDB của chúng.

**Ví dụ 3.3:** Giả sử chúng ta có một cơ sở dữ liệu IPRDB về bệnh nhân với hai lược đồ IPRDB lần lượt là lược đồ IPRDB PATIENT (*Bảng 8*) và lượt đồ IPRDB DOCTOR (*Bảng 9*)

|  |  |  |  |
| --- | --- | --- | --- |
| **D\_ID** | **D\_NAME** | **D\_SPECIALITY** | **AVAILABILITY** |
| D101 | Dr.Smith | Cardiologist | {(Morning, [0.3, 0.6]),  (Afternoon, [0.4, 0.7])} |
| D102 | Dr.Lee | Pulmonologist | Full day, [1,1] |

Bảng 3.2 Quan hệ DOCTOR

Định nghĩa 3.3 cho biết rằng một cơ sở dữ liệu IPRDB là một tập hợp các quan hệ IPRDB, trong đó mỗi quan hệ được xây dựng dựa trên một lược đồ IPRDB. Ví dụ trên mô tả một cơ sở dữ liệu quan hệ IPRDB gồm hai bảng (PATIENT, DOCTOR), mỗi bảng đại diện cho một mối quan hệ IPRDB dựa trên lược đồ riêng của nó. Mỗi bảng có các thuộc tính với giá trị xác suất thể hiện tính không chắc chắn trong dữ liệu thực tế.

## **3.2. Phụ thuộc hàm IPRDB**

Các phụ thuộc chức năng (functional dependencies) đóng vai trò quan trọng trong CRDB (Cơ sở dữ liệu quan hệ truyền thống). Khái niệm phụ thuộc chức năng xác suất trong IPRDB được mở rộng từ CRDB với các thuộc tính có giá trị xác suất. Trước tiên, chúng ta định nghĩa phép đo xác suất để xác định mức độ bằng nhau của hai giá trị xác suất của cùng một thuộc tính đối với hai bộ dữ liệu khác nhau trong một quan hệ IPRDB.

**Định nghĩa 3.4:** Giả sử *R(****U****, ℘)* là một lược đồ IPRDB, 𝑟 là một quan hệ trên R và  t1 và t2 là  hai  bộ  trong  𝑟, *A*là thuộc tính của ***U*** và ⊕ là một chiến lược kết hợp xác suất. Khoảng xác suất để các giá trị của thuộc tính A của hai bộ t1 và t2  bằng nhau dưới ⊗, kí hiệu là p(t1.*A*  = t2.*A*), là ( ([*l1i, u1i*] ⊗ [*l2j, u2j*]) . *Pr*(*ʋ1i = ʋ2j*) ), trong đó *t1.A =* {(ʋ11, [*l11, u11*]), *. . . ,* (ʋ1m, [*l1m, u1m*])}, *t2.A =* {(ʋ21, [*l21, u21*]), *. . . ,* (ʋ2n, [*l2n, u2n*])} và ⊕ là toán tử chiến lược loại trừ nhau (mutual exclusion probabilistic disjunction operator).

**Ví dụ 3.4:** Giả sử lược đồ IPRDB *R(****U****, ℘)* có thuộc tính 𝐴, với quan hệ 𝑟 chứa hai bộ t1 và t2. Các giá trị xác suất của thuộc tính 𝐴 trong mỗi bộ được biểu diễn như sau: t1.A={(‘A’, [0.4, 0.6]), (‘B’, [0.3, 0.5])}, t2.A={(‘A’, [0.5, 0.7]), (‘C’, [0.2, 0.4])}

Ta có : p(t1.A = t2.A)

= ( ([*l1i, u1i*] ⊗ [*l2j, u2j*]) . *Pr*(*ʋ1i = ʋ2j*) )

= (([*l11, u11*] ⊗ [*l21, u21*]) . *Pr*(*ʋ11 = ʋ21*) ) ⊕ ( ([*l12, u12*] ⊗ [*l21, u21*]) . *Pr*(*ʋ12 = ʋ21*))

⊕ (([*l11, u11*] ⊗ [*l22, u22*]) . *Pr*(*ʋ11 = ʋ22*))

⊕ (([*l12, u12*] ⊗ [*l22, u22*]) . *Pr*(*ʋ12 = ʋ22*))

= ( [0.5, 0.6] . 0.1 ) ⊕ ( 0 . 0 ) ⊕ ( 0 . 0 ) ⊕ ([0.3, 0.4] . 0 )

= [0.5, 0.6]

Phụ thuộc hàm xác suất trong IPRDB là một phần mở rộng của phụ thuộc hàm trong CRDB, với các thuộc tính có giá trị không chắc chắn và không chính xác như sau.

**Định nghĩa 3.5:** Giả sử *R(****U****, ℘)* là một mô hình IPRDB, 𝑟 là bất kỳ quan hệ nào trên 𝑅, ⊗ là một chiến lược kết hợp xác suất (probabilistic conjunction strategy), và X, Y là hai tập con khác rỗng của ***U.*** Phụ thuộc hàm xác suất X trên Y được ký hiệu là X →⊗​ Y tồn tại khi và chỉ khi:

∀t1, t2 ∈ 𝑟 : ⊗A∈X p(t1.A =⊗ t2.A) ≤ ⊗A∈Y p(t1.A =⊗ t2.A).

Có thể thấy rằng định nghĩa này bao hàm định nghĩa của CRDB (Cơ sở dữ liệu quan hệ chắc chắn). Ngoài ra, dễ dàng nhận thấy rằng với mọi lược đồ IPRDB *R(****U****, ℘),* thì ***U*** →⊗Y ***​***với Y ⊆ ***U*** tồn tại dưới mọi chiến lược kết hợp xác suất.

**Ví dụ 3.5** : Trong mọi quan hệ 𝑟 trên lược đồ PATIENT với tập thuộc tính ***U*** = { P\_ID, P\_NAME, P\_AGE, P\_DISEASE, P\_COST } trong ví dụ 3.2, các giá trị của thuộc tính P\_ID, biểu diễn định danh của bệnh nhân, là duy nhất và đôi một khác nhau. Do đó, với hai bộ dữ liệu t1, t2 ∈ r và một thuộc tính A ∈ ***U,*** ta có: p(t1.A =⊗ t2.A)0 và p(t1.P\_ID =⊗ t2. P\_ID) = 0 (do P\_ID luôn duy nhất và khác nhau giữa các bệnh nhân). Từ đó suy ra được p(t1.P\_ID =⊗ t2. P\_ID) ⊗A∈Y p(t1.A =⊗ t2.A) với Y ⊆ ***U*** (*định nghĩa 3.5*). Vậy nên tồn tại một phụ thuộc hàm IPRDB P\_ID →⊗Y trên lược đồ PATIENT dưới mọi chiến lược xác xuất

Trong IPRDB, phụ thuộc hàm có yếu tố xác suất (probabilistic). Khi biết giá trị của X, ta chỉ có thể xác định giá trị của Y với một mức độ xác suất nhất định, có thể không phải là 1. Điều này có thể làm cho các tính chất phụ thuộc hàm trong IPRDB phức tạp hơn và ít xác định hơn so với CRDB. Phụ thuộc hàm có thể thay đổi tùy thuộc vào chiến lược kết hợp xác suất ⊗, và phụ thuộc vào mức độ xác suất mà chúng ta áp dụng Chính vì thế, điều này có thể tạo ra một mức độ không xác định giữa các mối quan hệ giữa các thuộc tính trong IPRDB, làm cho các tính chất của phụ thuộc hàm ít ổn định hơn. IPRDB yêu cầu tính toán các xác suất liên quan đến phụ thuộc hàm, điều này có thể tạo ra các biến động và yêu cầu thêm các chiến lược kết hợp xác suất trong các phép toán. Tóm lại, IPRDB mở rộng các tính chất của phụ thuộc hàm trong CRDB nhưng với sự bổ sung của yếu tố xác suất, điều này làm cho các tính chất trong IPRDB có thể linh hoạt hơn nhưng cũng phức tạp hơn so với trong CRDB.

## **3.2. Khoá trong IPRDB**

**Định nghĩa 3.6:** Giả sử *R(****U****, ℘)* là một mô hình IPRDB, 𝑟 là bất kỳ quan hệ nào trên 𝑅, ⊗ là một chiến lược kết hợp xác suất (probabilistic conjunction strategy). Một tập thuộc tính *K* ⊆ ***U*** là khóa của 𝑅 dưới ⊗ nếu giá trị của các thuộc tính trong 𝐾 là xác định và tồn tại một phụ thuộc hàm xác suất 𝐾→⊗ ***U*** sao cho không có bất kỳ tập con thực sự nào của 𝐾 thỏa mãn tính chất này.

**Ví dụ 3.6:** Trong quan hệ PATIENT ở trên, nếu giả sử mỗi bệnh nhân có một mã định danh duy nhất tương ứng với giá trị của thuộc tính P\_ID, thì P\_ID là một khóa của lược đồ PATIENT dưới mọi chiến lược kết hợp xác suất**.**

# **CHƯƠNG 4. ĐẠI SỐ QUAN HỆ VÀ TÍNH CHẤT CỦA IPRDB**

## **4.1. Phép chọn – Selection**

**Định nghĩa 4.1:** Cho R là một lược đồ IPRDB, và X là tập hợp các biến bộ quan hệ. Khi đó, biểu thức lựa chọn được định nghĩa như sau:

* x.Aθc, với x∈X, A là thuộc tính trong R, θ là quan hệ nhị phân từ tập hợp {=, ≠, ≤, ≥, <, >, ⇒}, và c ∈dom(A).
* x.A1 θ x.A2, với x ∈ X, A1​ và A2 là hai thuộc tính trong R và ⊗ là một chiến lược kết hợp xác suất.
* α ⊗ β, với α và β là các biểu thức lựa chọn trên cùng một biến bộ quan hệ, và ⊗ là một chiến lược kết hợp xác suất.
* α ⊕ β, với α và β là các biểu thức lựa chọn trên cùng một biến bộ quan hệ, và ⊕ là một chiến lược rời rạc xác suất.

**Ví dụ 4.1:** Xét bảng dữ liệu **PATIENT** trong Ví dụ 3.1. Việc lựa chọn "tất cả bệnh nhân bị xơ gan và trả chi phí điều trị hàng ngày trên 5 USD" có thể được biểu diễn bằng biểu thức lựa chọn: x.P\_DISEASE=cirrhosis ⊕ x.D\_COST>5

Bây giờ, các điều kiện lựa chọn trong IPRDB được định nghĩa một cách chính thức dựa trên các biểu thức lựa chọn như sau.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **P\_ID** | **P\_NAME** | **P\_AGE** | **P\_DISEASE** | **D\_COST** |
| P218 | Paul | {(43, [0.5, 0.5]), (44,[0.5, 0.5])} | {(hepatitis, [0.3, 0.5]), (cirrhosishepatitis, [0.53,0.75])} | {($6, [0.4, 0.6]), ($7, [0.4, 0.6])} |

Bảng 4.1 Quan hệ σφ (PATIENT)

**Định nghĩa 4.2:** Cho R là một lược đồ IPRDB. Khi đó, các điều kiện lựa chọn được định nghĩa như sau:

* Nếu α là một biểu thức lựa chọn và [l, u] là một khoảng con của [0,1], thì (α, [l, u]) là một điều kiện lựa chọn.
* Nếu φ và ψ là các điều kiện lựa chọn trên cùng một biến bộ, thì ¬φ, (φ ∨ ψ), (φ∧ψ) là các điều kiện lựa chọn.

**Ví dụ 4.2:** Với bảng **PATIENT** trong Ví dụ 3.1, lựa chọn "tất cả bệnh nhân trên 40 tuổi với xác suất ít nhất 0.8, bị lao phổi (tuberculosis), và trả chi phí điều trị hàng ngày không vượt quá 30 USD với xác suất từ 0.5 đến 0.6" được biểu diễn bởi điều kiện lựa chọn:

(x.P\_AGE > 40, [0.8,1]) ∨ (x.P\_DISEASE = tuberculosis ∧ x.D\_COST ≤ 30, [0.5, 0.6])

**Định nghĩa 4.3:** Giả sử R là một lược đồ cơ sở dữ liệu xác suất quan hệ chưa xác định (IPRDB schema), r là một quan hệ trên R, x là một biến bộ (tuple variable), và t là một bộ trong r. Diễn giảixác suất của các biểu thức lựa chọn cho R, r, và t, ký hiệu là ProbR,r,t là ánh xạ một phần từ tập hợp tất cả các biểu thức lựa chọn đến tập hợp tất cả các khoảng đóng của [0,1], được định nghĩa theo các bước sau:

* ProbR,r,t​(x.Aθc) = ([li,ui].Pr(vi θ c))

Trong đó: t.A = { (v1​,[l1​,u1​]),…,(vk,[lk,uk​])} và ⨁ là toán tử loại trừ xác suất kết hợp.

* ProbR,r,t​(x.A1 ​θ x.A2​) = ( [l1​,u1​]⊗[l2​,u2​], Pr(v1n​θv2n​) )
* trong đó t.A1​={(v11​,[l1​,u1​]),…,(v1m​,[lm​,um​])}, t.A2​={(v21​,[l2​,u2​]),…,(v2n​,[ln​,un])}, Và ⨁ là toán tử loại trừ xác suất kết hợp.
* ProbR,r,t​(α ⊗ β) = ProbR,r,t(α) ⊗ ProbR,r,t(β).
* ProbR,r,t​(α ⨁ β) = ProbR,r,t(α) ⨁ ProbR,r,t(β).

**Ví dụ 4.3:** Giả sử R là lược đồ **PATIENT** và r là quan hệ **PATIENT**. Xét hàng thứ hai trong r, ký hiệu là t2​. Chúng ta có:

ProbR,r,t(x.P\_DISEASE=cirrhosis)

= [0.3,0.5] ⨯ Pr(hepatitis=cirrhosis) ⊗me [0.5,0.7] ⨯ Pr(cirrhosis=cirrhosis)

= [0.3,0.5]⨯0.0 ⊕me​ [0.5,0.7]⨯1.0

= [0,0] ⊕me⨯ [0.5,0.7]

= [0.5,0.7]

**Định nghĩa 4.4:** Cho R là một lược đồ trong IPRDB, và r là một quan hệ trên R. Việc thỏa mãn các điều kiện lựa chọn dưới ProbR,r,t được định nghĩa như sau:

* ProbR, r, t ⊨ (α)[l,u] nếu và chỉ nếu ProbR,r,t(α)⊆[l,u].

Điều này có nghĩa là xác suất của điều kiện α phải nằm trong khoảng xác suất [l, u]. Ví dụ: Nếu α là một điều kiện (biến hoặc một phép kiểm tra) và khoảng xác suất cho phép là từ 0.3 đến 0.7, thì xác suất của α phải nằm trong khoảng này để điều kiện được thỏa mãn.

* ProbR, r, t ⊨ ¬ φ nếu ProbR, r, t ⊭ φ.

Điều này có nghĩa là phủ định của một điều kiện φ thỏa mãn nếu như φ không thỏa mãn. Ví dụ: Nếu φ là một điều kiện có xác suất nằm ngoài khoảng yêu cầu, thì phủ định của φ sẽ thỏa mãn.

* ProbR, r, t ⊨ φ∧ω nếu ProbR, r, t ⊨ φ và ProbR, r, t ⊨ ω.

Điều này có nghĩa là phép và của hai điều kiện φ và ω thỏa mãn nếu cả hai điều kiện đều thỏa mãn. Ví dụ: Nếu φ và ω đều là những điều kiện với xác suất nằm trong khoảng yêu cầu, thì φ∧ω cũng sẽ thỏa mãn.

* ProbR, r, t ⊨ φ∨ω nếu ProbR, r, t⊨φ hoặc ProbR, r, t ⊨ ω.

Điều này có nghĩa là phép hoặc của hai điều kiện φ và ω thỏa mãn nếu ít nhất một trong hai điều kiện thỏa mãn. Ví dụ: Nếu φ hoặc ω (hoặc cả hai) có xác suất nằm trong khoảng yêu cầu, thì φ∨ω cũng sẽ thỏa mãn.

**Ví dụ:** Giả sử ta có điều kiện α là "Xác suất sinh viên A thi đỗ cao hơn 0.6". Nếu với dữ liệu xác suất ta biết rằng xác suất sinh viên A thi đỗ là 0.7, thì ta có thể viết: ProbR, r, t ⊨α để biểu thị rằng xác suất 0.7 thỏa mãn điều kiện α (vì 0.7 > 0.6).

**Định nghĩa 4.5:** Cho R là một lược đồ IPRDB, r là một quan hệ trên R và φ là một điều kiện lựa chọn trên một biến tuple trong r. Lựa chọn trên r cho φ, được biểu thị bằng σφ(r), là một quan hệ r\* trên R được chỉ định bởi r\* = {t∈r | ProbR, r, t ⊨ φ}.

**Ví dụ 4.4:** Giả sử r biểu thị quan hệ BỆNH NHÂN (PATIENT) trong Ví dụ 3.1, và R biểu thị lược đồ của nó. Truy vấn: “Xác định tất cả các bệnh nhân trên 40 tuổi với xác suất ít nhất là 0.9, bị xơ gan và trả chi phí điều trị hàng ngày không dưới 6 USD với xác suất nằm trong khoảng từ 0.3 đến 0.7” có thể được thực hiện bằng phép chọn σφ (PATIENT), trong đó:

**φ = (x.P\_AGE > 40)[0.9,1] ∧ (x.P\_DISEASE=cirrhosis ⊗ x.D\_COST ≥ 6) [0.3, 0.7]**

Chỉ có bộ thứ hai t2​ của quan hệ BỆNH NHÂN trong Ví dụ 3.1 thỏa mãn φ vì:

* **ProbR, r, t2 (x.P\_AGE>40)**

= [0.5,0.5] ⊗ Pr (43>40) ⊕me [0.5,0.5] ⊗ Pr (44>40)

= [0.5,0.5] ⊗ 1.0 ⊕me [0.5, 0.5] ⊗ 1.0

= [1.0, 1.0] ⊆ [0.9, 1].

* **ProbR, r, t2 (x.D\_COST ≥ 6)**

= [0.4,0.6] ⊗ Pr (6≥6) ⊕me​ [0.4,0.6] ⊗ Pr (7≥6)

= [0.4,0.6] ⊗ 1.0 ⊕me​[0.4,0.6] ⊗ 1.0

= [0.4,0.6] ⊕me ​[0.4,0.6]

= [0.8,1].

Từ kết quả của phép tính trong Ví dụ 4.3, ta có:

* **ProbR,r,t2 (x.P\_DISEASE = cirrhosis ⊗ x.D\_COST ≥ 6)**

= [0.5,0.7] ⊗ [0.8,1]

= [0.4,0.7] ⊆ [0.3,0.7].

Với các bộ khác, ta có:

* ProbR, r, t2(x.P\_DISEASE=cirrhosis ⊗ x.D\_COST ≥ 6) = [0,0] ⊄ [0.3,0.7], ∀i≠2

Do đó, kết quả của truy vấn được hiển thị trong Bảng 10

|  |  |  |
| --- | --- | --- |
| **P\_AGE** | **P\_DISEASE** | **D\_COST** |
| {(65, [1,1])} | {(lung cancer, [0.5, 0.5]),  (tuberculosis, [0.5, 0.5])} | {($30, [0.3, 0.6]),  ($35, [0.4, 0.7])} |
| {(43, [0.5, 0.5])},  {(44, [0.5, 0.5])} | {(hepatitis, [0.3, 0.5]),  (cirrhosis, [0.5, 0.7])} | {($6, [0.4, 0.6]),  ($7, [0.4, 0.6])} |
| {(15, [1,1])} | {(bronchitis, [1, 1])} | {($7, [1, 1)} |
| {(36, [1,1])} | {(duodenitis, [0.7, 0.75]),  (gastritis, [0.75, 0.8])} | {($8, [0.65, 0.75]),  ($9, [0.75, 0.85])} |

Bảng 4.2 Quan hệ Π {P\_AGE, P\_DISEASE, D\_COST} ⊕in (PATIENT)

## **4.2. Phép chiếu - Projection**

**Định nghĩa 4.6**:

Cho R(U, ℘) là một lược đồ IPRDB, r là một quan hệ trên R, H là một tập con của các thuộc tính của U, và ⊕ là một chiến lược phép rời rạc xác suất.

Phép chiếu của r trên H theo ⊕, ký hiệu là ΠH⨁(r), là quan hệ r\* trên lược đồ R\* được xác định như sau:

* R∗ = (H, ℘\*) và ℘\*(A)=℘(A), ∀*A ∈ H*.
* r∗= {t∗∣ t∗. A = u.A⊕…⊕w.A, ∀*A ∈ H*, ∃u,..., w∈r sao cho [u[H]]= … =[w[H]]}.

**Ví dụ 4.5:** Xét quan hệ PATIENT trên lược đồ PATIENT ({P\_ID, P\_NAME, P\_AGE, P\_DISEASE, D\_COST}, ℘) như trong Bảng 2. Phép chiếu của quan hệ này trên tập thuộc tính H= {P\_AGE, P\_DISEASE, D\_COST} theo chiến lược ⊕in​ là quan hệ ΠH⊕in(PATIENT) trên lược đồ R∗=({P\_AGE, P\_DISEASE, D\_COST}, ℘\*), được tính trong Bảng 4, với ℘\*(A) = ℘(A), ∀ A ∈ H. Trong quan hệ PATIENT, ta có t3[H] = t5[H], do đó hai bộ dữ liệu t3 và t5, được chiếu trên H và gộp lại thành bộ t4 theo chiến lược rời rạc xác suất ⨁in trong Bảng 4.

## **4.3. Phép tích – Descartes**

Đối với tích Descartes của hai quan hệ IPRDB như trong CRDB, chúng tôi giả định tập hợp các thuộc tính của các lược đồ của chúng là rời rạc, và mọi k-tuple t = (pv1, pv2, . . . , pvk) của các giá trị xác suất là một danh sách không có thứ tự. Tích Descartes của hai IPRDB mối quan hệ được mở rộng từ mối quan hệ của hai CRDB với các thuộc tính có giá trị không chắc chắn và không chính xác như sau.

**Định nghĩa 4.7:** Cho U1, U2 là hai tập thuộc tính không có phần tử chung nào, R1(U1,℘1), R2(U2,℘2) là hai lược đồ IPRDB, r1, r2 là hai quan hệ trên R1, R2 tương ứng. Tích Descartes của r1 và r2, được biểu thị bằng r1×r2 là mối quan hệ r trên R, được xác định bởi:

1. R =(U,℘), trong đó U =U1 ∪ U2

* ℘(A) = ℘1(A) nếu A ∈ U1
* ℘(A) = ℘2(A) nếu A ∈ U2.

2. r = {t | t.A = t1.A nếu A ∈ U1, t.A = t2.A nếu A ∈ U2, t1 ∈ r1, t2 ∈ r2}

## **4.4. Phép kết hợp – Join**

Phép kết hợp (Join) của hai quan hệ IPRDB là một phần mở rộng của phép nối tự nhiên của hai quan hệ CRDB với các giá trị xác suất như định nghĩa sau.

**Định nghĩa 4.8:** Giả sử U1 và U2​ là hai tập thuộc tính sao cho nếu chúng có cùng tên thuộc tính, thì những thuộc tính đó có cùng miền giá trị. Cho R1​(***U***1​,*℘*1​) và R2(***U***2, *℘*2) là hai lược đồ IPRDB, 𝑟1​ và 𝑟2​ là hai quan hệ trên R1​ và R2 tương ứng, và ⊗ là một chiến lược kết hợp xác suất . Phép kết hợp của 𝑟1 và 𝑟2 dưới ⊗ kí hiệu là 𝑟1 ⋈⊗ 𝑟2 là mối quan hệ r trên lược đồ R, được xác định bởi:

**1.** R =(U,℘), trong đó U =U1 ∪ U2 ,

* ℘(A)= ℘1(A) nếu A ∈ U1 - U2
* ℘(A) = ℘2(A) nếu A ∈ U2 - U1
* ℘1(A) ⊗ ℘2(A) nếu A ∩ U1 - U2

**2.** r = { t |

t.A = t1.A nếu A∈U1-U2,

t.A = t2.A nếu A∈U2 - U1,

t.A = t1.A⊗t2 nếu A∈ U1 ∩ U2, t1 ∈ r1, t2 ∈ r2}

**Ví dụ 4.8:** Cho hai quan hệ IPRDB PATIENT1 và PATIENT2 như trong Bảng 5 và 6, thì kết quả của phép nối chúng theo chiến lược kết hợp xác suất (probabilistic conjunction strategy) ⊗in (chiến lược độc lập) là mối quan hệ PATIENT1⋈⊗in PATIENT2 được tính như trong Bảng 7. Ở đây, các tên của mỗi mối quan hệ và lược đồ của nó là giống hệt nhau. Bộ ℘(A) cho mỗi thuộc tính A trong các lược đồ bao gồm giá trị xác suất trên dom(A).

|  |  |
| --- | --- |
| **P\_ID** | **P\_DISEASE** |
| P421 | { (bronchitis, [0.35, 0.45]), (bronchiectasis, [0.55, 0.65]) } |
| P829 | { (pancreatitis, [1, 1]) } |

Bảng 4.3 Quan hệ PATIENT1

|  |  |
| --- | --- |
| **P\_NAME** | **P\_DISEASE** |
| Peter | { (bronchiectasis, [1, 1]) } |
| Selena | { (pancreatitis, [0.4, 0.5]), (cirrhosis, [0.5, 0.6]) } |

Bảng 4.4 Quan hệ PATIENT2

***Chi tiết tính toán*:**

Ta có, trong bảng 5 thì t(P\_ID) = P421 có ʋ2  *=* bronchiectasis và t(P\_ID) = P829 có ʋ= pancreatitis lần lượt giống với ʋ = bronchiectasis của t(P\_NAME) = PETER và ʋ1 = pancreatitis của t(P\_NAME) = Selena trong bảng 6. Từ đó để tìm PATIENT1⋈⊗in PATIENT2  ta được:

Với t(P\_DISEASE) = bronchiectasis:

* [L1, U1] ⊗In [L2, U2] ≡ [L1 . L2, U1 . U2] *(biết (L,U) là khoảng xác suất)*

= (0.55, 0.65) ⊗in (1,1)

= [0.55 × 1, 0.65 × 1]

=  [0.55, 0.65]

Với t(P\_DISEASE) = pancreatitis:

* [L1, U1] ⊗In [L2, U2] ≡ [L1 . L2, U1 . U2] *(biết (L,U) là khoảng xác suất)*

= [1, 1] ⊗in [0.4, 0.5]

= [1 × 0.4, 1 × 0.5]

= [0.4, 0.5]

|  |  |  |
| --- | --- | --- |
| **P\_ID** | **P\_NAME** | **P\_DISEASE** |
| P421 | Peter | {(bronchiectasis, [0.55, 0.65]) } |
| P829 | Selena | {(pancreatitis, [0.4, 0.5])} |

*Kết quả thu được như sau :*

Bảng 4.5 Quan hệ PATIENT3 = PATIENT1⋈⊗in PATIENT2

## **4.5. Giao, hợp, hiệu - Intersection, Union and Difference**

Giao, hợp và hiệu của hai quan hệ IPRDB trên cùng một lược đồ là những quan hệ IPRDB trên lược đồ đó. Khi thực hiện các phép toán này, hai bộ dữ liệu có cùng khóa (tức là các bộ dữ liệu có giá trị khóa giống nhau) từ hai quan hệ sẽ được hợp nhất thành một bộ dữ liệu duy nhất trong quan hệ kết quả. Quá trình hợp nhất này sẽ được thực hiện thông qua một chiến lược kết hợp xác suất, điều này có nghĩa là các giá trị của các thuộc tính sẽ được kết hợp theo một phương thức xác suất, thay vì là các giá trị xác định như trong cơ sở dữ liệu quan hệ truyền thống (CRDB).

Trong cơ sở dữ liệu quan hệ truyền thống, các bộ dữ liệu có cùng khóa sẽ được coi là giống hệt nhau, và chúng có thể được kết hợp trực tiếp mà không cần phải quan tâm đến yếu tố xác suất. Tuy nhiên, trong IPRDB, vì các thuộc tính có thể có giá trị xác suất thay vì giá trị chắc chắn, việc kết hợp các bộ dữ liệu này trở nên phức tạp hơn. Việc áp dụng một chiến lược kết hợp xác suất cho phép phản ánh sự không chắc chắn trong các giá trị của các thuộc tính trong các quan hệ này.

Các phép toán giao, hợp và hiệu trong IPRDB có thể được coi là sự mở rộng của các phép toán tương ứng trong CRDB, với điểm khác biệt là chúng phải xử lý các giá trị thuộc tính có xác suất. Cụ thể, trong IPRDB, các phép toán này không chỉ đơn giản là việc kết hợp các bộ dữ liệu, mà còn phải tính đến sự xác suất trong các thuộc tính của chúng. Điều này cho phép mở rộng khả năng của cơ sở dữ liệu để xử lý các tình huống có sự không chắc chắn, ví dụ như trong các hệ thống mà dữ liệu bị nhiễu hoặc không hoàn chỉnh.

Vì vậy, giao, hợp và hiệu của hai quan hệ IPRDB là sự kết hợp giữa lý thuyết cơ sở dữ liệu quan hệ truyền thống và lý thuyết cơ sở dữ liệu xác suất, giúp mở rộng khả năng biểu diễn và thao tác với dữ liệu không chắc chắn. Các phép toán này sẽ được định nghĩa như sau.

**Định nghĩa 4.9:** Giả sử *R(****U****, ℘)* là một mô hình IPRDB, 𝑟1, 𝑟2 là hai quan hệ nào trên 𝑅, ⊗ là một chiến lược kết hợp xác suất (probabilistic conjunction strategy). Giao (**intersection**) của 𝑟1, 𝑟2  dưới ⊗, ký hiệu là 𝑟1 ⊗ 𝑟2, là một quan hệ IPRDB 𝑟trên R được định nghĩa như sau: {t | t.A = t1.A ⊗ t2.A, t1 𝑟1, t2 𝑟2, A ***U*** sao cho t1[K] = t2[K].

Lưu ý rằng giá trị của mỗi thuộc tính khóa là xác định theo định nghĩa 3.6. Do đó, ký hiệu t1[K] = t2[K] có thể được sử dụng trong định nghĩa 4.9. Hơn nữa, chúng ta có thể xác định duy nhất một bộ dữ liệu của một quan hệ dưới mỗi khóa của quan hệ đó. Vì vậy, quan hệ kết quả là duy nhất đối với tất cả các khóa.

**Ví dụ 4.5:** Có 2 quan hệ dữ liệu IPRDB PATIENT3 bảng 4.4 và PATIENT4 bảng 4.5. Áp dụng định nghĩa 4.5 ta có kết quả của phép toán PATIENT3 ⊗ PATIENT4  ở bảng 4.6.

|  |  |  |
| --- | --- | --- |
| **P\_ID** | **P\_NAME** | **P\_DISEASE** |
| P421 | Peter | {(bronchiectasis, [0.35, 0.6])} |
| P829 | Selena | {(pancreatitis, [0.42, 0.45])} |

Bảng 4.6 Quan hệ PATIENT4

*Tính toán chi tiết:*

**Bộ dữ liệu thứ nhất:**

* t1 = ( P421, Peter, {(bronchiectasis,[0.55,0.65])} ) (*bảng 12*)
* t2 = ( P421, Peter, {(bronchiectasis, [0.35,0.6])} ) (*bảng 13*)

P\_ID = P421, chúng ta kết hợp các giá trị xác xuất có ʋ giống nhau, trong trường hợp này là “bronchiectasis” bằng công thức [L1, U1] ⊗[L2, U2]

* [0.55,0.65] ⊗ [0.35,0.6] = [0.6, 0.65]
* Kết quả t = ( P421, Peter, { (bronchiectasis, [0.6, 0.65]) } )

**Bộ dữ liệu thứ hai:**

* t1 = ( P829, Selena, {(pancreatitis,[0.4,0.5])} ) (bảng 12)
* t2 = ( P829, Selena, {(pancreatitis,[0.42,0.45])} ) (bảng 13)

P\_ID = P829, chúng ta kết hợp các giá trị xác xuất có ʋ giống nhau, trong trường hợp này là “pancreatitis” bằng công thức [L1, U1] ⊗[L2, U2]

* [0.4,0.5] ⊗ [0.42,0.45] = [0.42,0.45]
* Kết quả t = ( P829, Selena, {(pancreatitis,[0.42,0.45])})

|  |  |  |
| --- | --- | --- |
| **P\_ID** | **P\_NAME** | **P\_DISEASE** |
| P421 | Peter | {(bronchiectasis, [0.6, 0.65])} |
| P829 | Selena | {(pancreatitis, [0.42, 0.45])} |

*Kết quả thu được như sau:*

Bảng 4.7 Quan hệ PATIENT3 ∩⊗ PATIENT4

**Định nghĩa 10:** Giả sử *R(****U****, ℘)* là một mô hình IPRDB, 𝑟1, 𝑟2 là hai quan hệ nào trên 𝑅, ⊕ là một chiến lược kết hợp xác suất (probabilistic conjunction strategy). Phép hợp (Union) của 𝑟1, 𝑟2  dưới ⊕, ký hiệu là 𝑟1 ⊕ 𝑟2, là một quan hệ IPRDB 𝑟trên R được định nghĩa như sau: 𝑟 = {t1 ∈ 𝑟 1|∀t2 ∈ 𝑟2, t1[K] ≠ t2[K]} ∪ {t2 ∈ 𝑟2 | ∀t1 ∈ 𝑟 1, t2[K] ≠ t1[K]} ∪ {t | t.A = t1.A ⊕ t2.A, t1 ∈ 𝑟1, t2 ∈ 𝑟2, A ∈ U sao cho t1[K] = t2[K]}

**Ví dụ 4.10:** Cho 2 quan hệ IPRDB DIAGNOSE1 và DIAGNOSE2 trên cùng một lược đồ DIAGNOSE( {P\_ID, D\_ID, P\_DISEASE, D\_COST}, ℘) như bảng 10 và 11, trong đó {P\_ID, D\_ID} là khóa của lược đồ và tập hợp ℘(A) của mỗi thuộc tính A trong DIAGNOSE bao gồm các giá trị xác suất trên dom(A). Phép hợp (union) của DIAGNOSE1 và DIAGNOSE2 dưới ⊗in là quan hệ DIAGNOSE1 ∪⊕in DIAGNOSE2 được tính như bảng 17.

|  |  |  |  |
| --- | --- | --- | --- |
| **P\_ID** | **D\_ID** | **P\_DISEASE** | **D\_COST** |
| P226 | D014 | {(lung cancer, [0.3, 0.6]),  (tuberculosis, [0.4, 0.7])} | {($30, [0.3, 0.4]),  ($35, [0.6, 0.7])} |
| P255 | D020 | {(hepatitis, [0.3, 0.8]),  (pancreatitis, [0.2, 0.7])} | {($8, [0.6, 1])} |

Bảng 4.8 Quan hệ DIAGNOSE1

|  |  |  |  |
| --- | --- | --- | --- |
| **P\_ID** | **D\_ID** | **P\_DISEASE** | **D\_COST** |
| P228 | D016 | {(lung cancer, [1, 1]) | {($30, [1, 1]) |
| P262 | D022 | {(dyspepsia, [1, 1])} | {($5, [1, 1])} |
| P255 | D020 | {(hepatitis, [0.4, 0.8]),  (cholecystitis, [0.2, 0.6])} | {($7, [0.2, 0.4]),  ($8, [0.4, 0.8])} |

Bảng 4.9 Quan hệ DIAGNOSE2

*Tính toán chi tiết:*

Các mã P\_ID khác nhau đều được ghi nhận lại

Ta có: t1  = ( P255, D020, {(hepatitis, [0.3, 0.8]), (pancreatitis, [0.2, 0.7])},{($8, [0.6, 1])} ) và t2 = ( P255, D020,{(hepatitis, [0.4, 0.8]), (cholecystitis, [0.2, 0.6]), {($7, [0.2, 0.4]), ($8, [0.4, 0.8])} ) có t1[K] = t2[K] đều bằng P255 và D020

Áp dụng định nghĩa 4.10 ta có: t.A = t1.A ⊕in t2.A

Ta có: t = (P255, D020, {(hepatitis, [0.4,0.8]) ⊕in (hepatitis, [0.3,0.8]), (cholecystitis, [0.2, 0.6]), (pancreatitis, [0.2, 0.7])}, {($7, [0.2, 0.4]), {($8, [0.6, 1]) ⊕in ($8, [0.4, 0.8])} )

* Với A = P\_DISEASE ta có: (hepatitis, [0.4, 0.8]) ⊕in (hepatitis, [0.3, 0.8])

= [0.4, 0.8] ⊕in [0.3, 0.8] ( ⊕in = [L1 + L2 – (L1 ⨯ L2), U1 + U2 – (U1 ⨯ U2)])

= [ (0.4 + 0.3 – (0.4 ⨯ 0.3) ), (0.8 + 0.8 – (0.8 ⨯ 0.8)) ] = [0.58, 0.96]

* Với A = D\_COST ta có: {($8, [0.6, 1]) ⊕in ($8, [0.4, 0.8])}

= [0.6, 1] ⊕in [0.4, 0.8] ( ⊕in = [L1 + L2 – (L1 ⨯ L2), U1 + U2 – (U1 ⨯ U2)])

= [0.6 + 0.4 – (0.6 ⨯ 0.4), 1 + 0.8 – (1 ⨯ 0.8)] = [0.76, 1]

*Kết quả thu được:*

|  |  |  |  |
| --- | --- | --- | --- |
| **P\_ID** | **D\_ID** | **P\_DISEASE** | **D\_COST** |
| P228 | D016 | {(lung cancer, [1, 1]) | {($30, [1, 1]) |
| P226 | D014 | {(lung cancer, [0.3, 0.6]),  (tuberculosis, [0.4, 0.7])} | {($30, [0.3, 0.4]),  ($35, [0.6, 0.7])} |
| P262 | D022 | {(dyspepsia, [1, 1])} | {($5, [1, 1])} |
| P255 | D020 | {(hepatitis, [0.58, 0.96]),  (pancreatitis, [0.2, 0.7]),  ((cholecystitis, [0.2, 0.6])} | {($7, [0.2, 0.4]),  ($8, [0,76, 1])} |

Bảng 4.10 Quan hệ DIAGNOSE1 ∪⊕in DIAGNOSE2

**Định nghĩa 11:** Giả sử R(***U***, ℘) là lược đồ IPRDB, 𝑟 1 và 𝑟 2 là hai quan hệ trên R, K là giá trị của R và ⊝ là một chiến lược kết hợp xác suất, được ký hiệu 𝑟 1 ∪⊝ 𝑟 2 , quan hệ 𝑟 IPRDB qua R được định nghĩa bởi :

𝑟 = { t1 ∈ 𝑟 1 | ∀t2 ∈ 𝑟 2, t1[K] ≠ t2[K] } ∪ {t | t.A = t1.A ⊝ t2.A, t1 ∈ 𝑟 1, t2 ∈ 𝑟 2, A ∈ U sao cho t1[K] = t2[K]}

Lưu ý rằng quan hệ kết quả trong các định nghĩa 4.10 và 4.11 không phụ thuộc vào việc chọn khóa của lược đồ của nó.

**Ví dụ 4.11:** Cho 2 quan hệ IPRDB DIAGNOSE1 và DIAGNOSE2 trên cùng một lược đồ DIAGNOSE( {P\_ID, D\_ID, P\_DISEASE, D\_COST}, ℘) như bảng 10 và 11, trong đó {P\_ID, D\_ID} là khóa của lược đồ và tập hợp ℘(A) của mỗi thuộc tính A trong DIAGNOSE bao gồm các giá trị xác suất trên dom(A). ). Phép trừ của DIAGNOSE1 và DIAGNOSE2 dưới ⊝in là quan hệ DIAGNOSE1 ∪⊝in DIAGNOSE2 được tính như bảng 13 như sau.

***Tính toán chi tiết:***

* Xét t1 ∈ 𝑟1 | ∀t2 ∈ 𝑟2, t1[K] ≠ t2[K] có:

t1 = (P226, D014, {(lung cancer, [0.3, 0.6]), (tuberculosis, [0.4, 0.7])}, {($30, [0.3, 0.4]), ($35, [0.6, 0.7])}) không nằm trong DIAGNOSE2 nên vẫn giữ nguyên bộ dữ liệu này

* Xét {t | t.A = t1.A ⊝ t2.A, t1 ∈ 𝑟1, t2 ∈ 𝑟2, A ∈ U sao cho t1[K] = t2[K]}

t = (P255, D020, {(hepatitis, [0.4, 0.8]) ⊝in (hepatitis, [0.3, 0.8]), pancreatitis, [0.2, 0.7])} , {($7, [0.2, 0.4]), ($8, [0.4, 0.8])} )

Ta có: t = (P255, D020, {(hepatitis, [0.4, 0.8]) ⊝in (hepatitis, [0.3, 0.8]), (cholecystitis, [0.2, 0.6]), (pancreatitis, [0.2, 0.7])} , {($7, [0.2, 0.4]), {($8, [0.6, 1]) ⊝in ($8, [0.4, 0.8])} )

* Với A = P\_DISEASE ta có: (hepatitis, [0.4, 0.8]) ⊝in (hepatitis, [0.3, 0.8])

= [0.4, 0.8] ⊝in [0.3, 0.8] ( ⊝in = [L1 ⨯ (1 – U2), U1 ⨯ (1 - L2)])

= [0.4 ⨯ (1 - 0.8), 0.8 ⨯ (1 - 0.3)]

= [0.08, 0.56]

* Với A = D\_COST ta có: {($8, [0.6, 1]) ⊝in ($8, [0.4, 0.8])}

= [0.6, 1] ⊝in [0.4, 0.8] ( ⊝in = [L1 ⨯ (1 – U2), U1 ⨯ (1 - L2)])

= [0.6 ⨯ (1 - 0.8), 1 ⨯ (1 - 0.4)]

= [0.12 , 0.6]

*Kết quả thu được :*

|  |  |  |  |
| --- | --- | --- | --- |
| **P\_ID** | **D\_ID** | **P\_DISEASE** | **D\_COST** |
| P226 | D014 | {(lung cancer, [0.3, 0.6]),  (tuberculosis, [0.4, 0.7])} | {($30, [0.3, 0.4]),  ($35, [0.6, 0.7])} |
| P255 | D020 | {(hepatitis, [0.08, 0.56]),  (pancreatitis, [0.2, 0.7])} | {($8, [0.12 , 0.6])} |

Bảng 4.11 Quan hệ DIAGNOSE1 ∪⊝in DIAGNOSE2

## **4.6. Tính chất phép toán đại số**

Các tính chất cơ bản của các phép toán đại số trong IPRDB là sự mở rộng của các tính chất trong CRDB với các giá trị xác suất. Những tính chất này khẳng định rằng mô hình IPRDB là chặt chẽ và nhất quán.

**Mệnh đề 4.1:** Giả sử 𝑟 là một quan hệ trên lược đồ R trong IPRDB, là hai điều kiện chọn lọc. Khi đó, thực hiện phép chọn với điều kiện :

(1)

**Chứng minh:**  Giả sử s = . Theo định nghĩa 4.4 và 4.5, ta có:

(Do s = ta thay s bằng kèm điều kiện )

(Định nghĩa 4.4 - điều kiện 3: ProbR,r,t ⊨ ∧ khi và chỉ khi ProbR,r,t ⊨ φ và ProbR,r,t ⊨ .))

Do đó, phương trình đã được chứng minh. Phương trình được đồng định nghĩa, vì . Chính vì thế mệnh đề 4.1 đã được chứng minh và chứng minh rằng thứ tự thực hiện hai phép chọn không quan trọng vì kết quả là duy nhất.

**Mệnh đề 4.2.** Giả sử R là một lược đồ IPRDB, r là quan hệ trên R, ⊕ là Chiến lược phân tách xác suất (probabilistic disjunction strategy), A và B là hai tập con của R, A⊆B. Khi đó, ∏A⊕ ( ∏B⊕ (*r*) ) = ∏A⊕ (*r*) (2)

**Chứng minh:** Vì A⊆ B, nên A ∩ B =A và (2) là các mối quan hệ trên cùng một sơ đồ. Từ định nghĩa 4.6, dễ thấy ∏A⊕( ∏B⊕ (*r*)) = ∏A∩B⊕ (*r*) = ∏A⊕ (*r*) theo chiến lược phân tách xác suất. Do đó phương trình (2) đã được chứng minh.

**Mệnh đề 4.3.** Giả sử R1, R2 và R3 là các lược đồ IPRDB sao cho nếu chúng có cùng tên thuộc tính, thì các thuộc tính đó có cùng miền giá trị, r1, r2 và r3 là quan hệ so với R1, R2 và R3, tương ứng, ⊗ là một chiến lược kết hợp xác suất. Khi đó,

r1 ⋈⊗ r2 = r2 ⋈⊗ r1 (3)

(r1 ⋈⊗ r2) ⋈⊗ r3 = r1 ⋈⊗ (r2 ⋈⊗ r3)  (4)

Các phương trình (3) và (4) nói rằng sự tham gia của các quan hệ IPRDB là giao hoán và liên kết.

**Chứng minh:** Dễ thấy rằng r1⋈⊗ r2, r2⋈⊗ r1 là hai mối quan hệ trên cùng một lược đồ. Theo định nghĩa 2.3, sự kết hợp của các giá trị xác suất là giao hoán (do tính giao hoán của các chiến lược giao nhau xác suất). Vì vậy, theo định nghĩa 4.8, nó theo sau r1⋈⊗ r2 = r2⋈⊗ r1 (3) đã được chứng minh.

Theo Định nghĩa 4.8, kết quả của hai vế trong phương trình (4) là các quan hệ trên cùng một lược đồ. Hơn nữa, theo Định nghĩa 2.3, phép liên kết (conjunction) của các giá trị xác suất là kết hợp (associative). Dựa trên Định nghĩa 4.8 và tính kết hợp của phép nối tự nhiên (natural join) trong cơ sở dữ liệu quan hệ thông thường, suy ra rằng phép nối của các quan hệ trong IPRDB cũng có tính chất kết hợp. Do đó, điều này dẫn đến kết quả rằng (r1 ⋈⊗ r2) ⋈⊗ r3 = r1 ⋈⊗ (r2 ⋈⊗ r3)  (4)

Vì tích Descartes (Định nghĩa 4.7) là một trường hợp đặc biệt của phép nối, nên nó dẫn đến kết quả trực tiếp của Mệnh đề 4.3 dưới sau.

**Hệ quả 4.1:** Giả sử R1, R2 và R3 là các lược đồ IPRDB sao cho chúng không có các thuộc tính có tên giống nhau, r1, r2 và r3 là các quan hệ trên R1, R2 và R3 tương ứng. Khi đó,

r1⨯ r2 = r2⨯ r1 *(5)*

(r1⨯ r2)⨯ r3 = r1⨯ (r2⨯ r3)*(6)*

**Mệnh đề 4.4** Cho R là lược đồ IPRDB, r1, r2 và r3 là quan hệ trên R. Hãy là một chiến lược kết hợp ⊗ (hoặc ⊕ phân tách xác suất). Khi đó,

r1 ∩⊗ r2 = r2 ∩⊗ r1 *(7)*

(r1 ∩⊗ r2) ∩⊗ r3 = r1 ∩⊗ (r2 ∩⊗ r3)  *(8)*

r1 ∪⊕ r2 = r2 ∪⊕ r1 *(9)*

(r1 ∪⊕ r2) ∪⊕ r3 = r1 ∪⊕ (r2 ∪⊕ r3) *(10)*

Các phương trình (7), (8), (9), và (10) nói rằng phép giao và phép hợp của các quan hệ trong IPRDB là giao hoán và kết hợp.

**Chứng minh:** Từ tính giao hoán và kết hợp của các chiến lược kết hợp xác suất, ta có thể suy ra rằng phép kết hợp các giá trị xác suất cũng có tính giao hoán và kết hợp (Định nghĩa 2.3). Vì vậy, phép giao của các quan hệ IPRDB r1, r2 và r3 dưới chiến lược kết hợp xác suất ⊗ và mỗi khóa được chọn cũng có tính giao hoán và kết hợp. Từ đó, theo Định nghĩa 4.9, ta có các phương trình (7) và (8).

Từ tính giao hoán và kết hợp của các chiến lược phân ly xác suất, ta có thể suy ra rằng phép phân ly các giá trị xác suất cũng có tính giao hoán và kết hợp (Định nghĩa 2.4). Vì vậy, phép hợp của các quan hệ IPRDB r1, r2 và r3 dưới chiến lược phân ly xác suất ⊕ và mỗi khóa được chọn cũng có tính giao hoán và kết hợp. Từ đó, theo Định nghĩa 4.10, ta có các phương trình (9) và (10).

Các mệnh đề và chứng minh trong IPRDB chứng minh rằng các phép toán đại số cơ bản, bao gồm phép chọn, phép chiếu, phép giao, phép hợp, và phép kết, đều có tính chất giao hoán và kết hợp. Điều này có ý nghĩa quan trọng vì nó cho phép thực hiện các phép toán trên cơ sở dữ liệu quan hệ mà không làm thay đổi kết quả cuối cùng, bất kể thứ tự hoặc nhóm các phép toán được áp dụng.

Cụ thể, các mệnh đề cho thấy rằng trong IPRDB, các phép toán này không chỉ giữ được tính chất chuẩn của các phép toán trong cơ sở dữ liệu quan hệ thông thường mà còn hỗ trợ các chiến lược xác suất, như chiến lược giao hoán xác suất và chiến lược phân ly xác suất. Điều này cho phép xử lý các dữ liệu không chắc chắn và chứa thông tin xác suất một cách hiệu quả mà không ảnh hưởng đến tính chính xác của kết quả.

Lợi ích chính của các tính chất này là chúng giúp đảm bảo tính nhất quán và chặt chẽ của mô hình IPRDB khi thực hiện các phép toán, đặc biệt khi làm việc với dữ liệu không hoàn chỉnh hoặc không chắc chắn. Các phép toán có thể được áp dụng linh hoạt mà không cần lo lắng về thứ tự thực hiện, đồng thời giúp tối ưu hóa việc truy vấn và xử lý dữ liệu trong các hệ thống cơ sở dữ liệu phức tạp. Điều này rất hữu ích trong các ứng dụng yêu cầu xử lý dữ liệu xác suất hoặc dữ liệu không chắc chắn, như trong các hệ thống thông tin quyết định, phân tích dữ liệu lớn hoặc các mô hình học máy.

# **CHƯƠNG 5. MÔ HÌNH IPRDB TRONG BIỂU DIỄN VÀ XỬ LÝ THÔNG TIN**

IPRDB là một phần mở rộng của CRDB và các mô hình PRDB lớp thứ hai với các giá trị xác suất (tức là các khoảng xác suất cho các giá trị). Ngoài ra, IPRDB còn có khả năng xử lý dữ liệu hiệu quả hơn các mô hình PRDB lớp thứ hai. Dưới đây là phần thảo luận chi tiết hơn về các kết quả đạt được.

## **5.1. Mở rộng của IPRDB trong việc biểu diễn dữ liệu**

Có hai lớp chính trong mô hình PRDB:

* **Lớp thứ nhất (C-1PRDB):** biểu diễn quan hệ xác suất dưới dạng tập hợp các bộ dữ liệu (tuple), mỗi bộ dữ liệu là một giá trị xác suất thuộc [0, 1]. Các phép toán đại số của C-1PRDB được định nghĩa bằng cách mở rộng các phép toán đại số của CRDB thông qua việc tính toán và kết hợp xác suất của các bộ dữ liệu trong các quan hệ C-1PRDB.
* **Lớp thứ hai (C-2PRDB):** biểu diễn quan hệ xác suất dưới dạng tập hợp các bộ dữ liệu (tuple). Mỗi thuộc tính trong bộ dữ liệu được liên kết với một giá trị xác suất duy nhất (v, p), biểu thị rằng thuộc tính có thể nhận giá trị v với xác suất p. EC-2PRDB mô hình mở rộng của C-2PRDB, trong đó mỗi thuộc tính được liên kết với một phân phối xác suất (v1, p1), ..., (vm, pm), biểu thị rằng thuộc tính có thể nhận một trong các giá trị vi với xác suất pi. Các phép toán đại số của C-2PRDB và EC-2PRDB được định nghĩa bằng cách mở rộng các phép toán của CRDB và sử dụng các phép toán trên xác suất hoặc phân phối xác suất để tính toán và kết hợp xác suất của các giá trị thuộc tính.

A diagram of a method

Description automatically generated

Hình 5.1 Minh họa sự mở rộng của IPRDB so với các mô hình CRDB, C-2PRDB, và EC-2PRDB

Mô hình IPRDB thuộc lớp C-2PRDB. Mỗi thuộc tính trong IPRDB được liên kết với một giá trị xác suất dưới dạng khoảng (v1, [l1, u1]), ..., (vm, [lm, um]), biểu thị rằng thuộc tính có thể nhận giá trị vi với xác suất thuộc [li, ui]. Các phép toán đại số của IPRDB được định nghĩa bằng cách mở rộng các phép toán của CRDB, sử dụng các giải thích xác suất cho các quan hệ nhị phân trên các tập hợp, và chiến lược kết hợp các khoảng xác suất. Bằng cách kết hợp các khoảng xác suất với các giá trị thuộc tính, IPRDB cho phép biểu diễn cả sự bất định của giá trị thuộc tính và sự không chính xác của xác suất cho các giá trị đó. Ngược lại, các mô hình như CRDB, C-2PRDB, và EC-2PRDB chỉ biểu diễn được sự bất định mà không thể biểu diễn sự không chính xác của xác suất (vì trong (v1, p1), ..., (vm, pm), xác suất pi là một số cụ thể).

## **5.2. Hiệu quả của IPRDB trong xử lý dữ liệu**

Do giá trị thuộc tính trong các quan hệ IPRDB là giá trị xác suất, việc tính toán và xử lý mô hình dữ liệu IPRDB hiệu quả hơn so với các mô hình dữ liệu C-2PRDB nơi mà giá trị thuộc tính là các cặp hàm phân phối xác suất của một tập giá trị.

Độ phức tạp tính toán của các phép toán đại số trong IPRDB là đa thức (polynomial) theo kích thước của các quan hệ xác suất, và nó hiệu quả tương đương với độ phức tạp tính toán của các phép toán đại số trong CRDB và EC-2PRDB.

* **Lựa chọn (selection):** Độ phức tạp O(n) với n bộ dữ liệu.
* **Chiếu (projection):** Độ phức tạp O(n) nhờ việc kết hợp xác suất các bộ trùng lặp.
* **Tích Descartes, hợp, giao, hiệu:** Độ phức tạp O(nm) với n và m là số bộ dữ liệu của hai quan hệ.

Như vậy, hiệu năng của mô hình IPRDB trong việc tính toán và xử lý thông tin bất định và không chính xác là tốt và có thể áp dụng trong thực tiễn.

# **CHƯƠNG 6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

## **1. Kết luận**

Đề tài này giới thiệu mô hình cơ sở dữ liệu quan hệ xác suất mới, gọi là IPRDB, mở rộng từ mô hình CRDB bằng cách sử dụng các thuộc tính có giá trị xác suất dạng khoảng để biểu diễn thông tin không chắc chắn và không chính xác.

**Nội dung chính:**

1. **Đặc điểm mô hình IPRDB:**

* Quan hệ được định nghĩa dưới dạng tập hợp các bộ, trong đó mỗi thuộc tính liên kết với một giá trị xác suất dạng khoảng.
* Giá trị xác suất này biểu diễn sự không chắc chắn và không chính xác của giá trị mà các thuộc tính có thể nhận.

1. **Mở rộng từ mô hình CRDB:**

* Các khái niệm cơ bản như lược đồ quan hệ, phụ thuộc hàm xác suất, khóa và các phép toán đại số quan hệ đã được mở rộng dựa trên mô hình CRDB.
* Sử dụng các chiến lược tổ hợp xác suất và các phép giao, hợp, hiệu của giá trị xác suất.

1. **Tính nhất quán và khả thi:**

* Các thuộc tính cơ bản và phép toán đại số quan hệ xác suất trong IPRDB được chứng minh có tính chất nhất quán và logic.
* IPRDB được đánh giá là một mô hình hiệu quả trong việc xử lý dữ liệu không chắc chắn và không chính xác.

## **2. Hướng phát triển**

Việc nghiên cứu và phát triển hệ thống quản lý cơ sở dữ liệu IPRDB cùng ngôn ngữ truy vấn đang được triển khai với mục tiêu đưa mô hình này vào ứng dụng thực tiễn. IPRDB là một mô hình cơ sở dữ liệu mạnh mẽ, được thiết kế để xử lý thông tin không chắc chắn và không chính xác, dựa trên các khái niệm và phép toán mở rộng từ mô hình CRDB. Mặc dù đã chứng minh được tiềm năng lớn, mô hình này vẫn còn nhiều cơ hội để cải thiện và mở rộng khả năng đáp ứng yêu cầu thực tiễn như:

1. **Nâng cao các chiến lược tổ hợp xác suất:** Phát triển thêm các chiến lược kết hợp xác suất nhằm cải thiện độ chính xác và hiệu suất khi xử lý dữ liệu phức tạp.
2. **Tích hợp công nghệ mới:** Áp dụng IPRDB vào các hệ thống trí tuệ nhân tạo (AI) và học máy (Machine Learning) để tối ưu hóa việc ra quyết định trong môi trường dữ liệu không chắc chắn.
3. **Hỗ trợ hiệu quả hơn cho truy vấn phức tạp:** Phát triển các thuật toán tối ưu hóa truy vấn để cải thiện tốc độ và hiệu quả khi làm việc với dữ liệu lớn.
4. **Ứng dụng thực tiễn:** Nghiên cứu các ứng dụng cụ thể của IPRDB trong các lĩnh vực như y tế (chẩn đoán bệnh), tài chính (phân tích rủi ro), và viễn thông (phân tích dữ liệu cảm biến).
5. **Mở rộng các khái niệm toán học:** Xây dựng thêm các mô hình toán học hỗ trợ, như xác suất đa chiều, để cải thiện khả năng biểu diễn dữ liệu phức tạp.

*Với những tiềm năng này, IPRDB hứa hẹn sẽ trở thành một mô hình cơ sở dữ liệu mạnh mẽ, đóng góp đáng kể vào sự phát triển của ngành khoa học dữ liệu và công nghệ thông tin.*

# **TÀI LIỆU THAM KHẢO**

[1] E.F. Codd, A relational model of data for large shared data banks, Communications of the ACM, vol.13, no.6, pp.377-387, 1970.

[2] G. Ozsoyoglu, Z. M. Ozsoyoglu, and V. Matos, Extending relational algebra and relational cal culus with set-valued attributes and aggregate functions, ACM Transactions on Database Sys tems, vol.12, no.4, pp.566-592, 1987.

[3] A. Silberschatz, H.F. Korth, and S. Sudar shan, Database system concepts, Seventh Edi tion, McGraw-Hill, 2019.

[4] D. Dey and S. Sarkar, A probabilistic rela tional model and algebra, ACM Transactions on Database Systems, vol.21, no.3, pp.339-369, 1996.

[5] D. Barbara, H. Garcia-Molina, and D. Porter, The management of probabilistic data, IEEE Transactions on Knowledge and Data Engineer ing, vol.4, no.5, pp.487-502, 1992.

[6] J. Bernad, C. Bobed, and E. Mena, Uncertain probabilistic range queries on multidimensional data, Information Sciences, vol. 537, pp.334 367, 2020.

[7] A. Ali, S. Talpur, and S. Narejo, Detecting faulty sensors by analyzing the uncertain data using probabilistic databases, Proceed ings of 3rd International Conference on Comput ing, Mathematics and Engineering Technologies, Sukkur, Pakistan, pp.143-150, 2020.

[8] V.V. Kheradkar and S. K. Shirgave, Query processing over relational cross model in uncer tain and probabilistic databases, Proceedings of 3Th International Conference on Arti cial Intelligence and Smart Energy, Coimbatore, India, pp.763-769, 2023.

[9] N. Fuhr and T. Rolleke, A probabilistic rela tional algebra for the integration of information retrieval and database systems, ACM Transactions on Information Systems, vol.15, no.1, pp.32-66, 1997.

[10] S. Zhang and C. Zhang, A probabilistic data model and its semantics, Journal of Research and Practice in Information Technology, vol.35, no.4, pp.237-256, 2003.

[11] Z. Ma and L. Yan, Advances in probabilis tic databases for uncertain information manage ment, Springer-Verlag Berlin Heidelberg, 2013.

[12] Y. Li, J. Chen, and L. Feng, Dealing with uncertainty: A survey of theories and practices, IEEE Transactions on Knowledge and Data En gineering, vol. 25, no.11, pp.2463-2482, 2013.

[13] I.I. Ceylan, A. Darwiche, and G.V.D. Broeck, Open-world probabilistic databases: Seman tics, algorithms, complexity, Journal of Ar ti cial Intelligence, vol.295, no.11, pp.103474 103513, 2021.

[14] H. Debbi, Explaining query answers in proba bilistic databases, International Journal of In teractive Multimedia and Arti cial Intelligence, vol.8, no.4, pp.140-152, 2023.

[15] L.V.S. Lakshmanan, N. Leone, R. Ross, and V.S. Subrahmanian, Probview: A exible prob abilistic database system, ACM Transactions on Database Systems, vol.22, no.3, pp.419-469, 1997.

[16] W. Zhao, A. Dekhtyar, and J. Goldsmith, Databases for interval probabilities, Interna tional Journal of Intelligent Systems, vol.19, no.9, pp.789-815, 2004.

[17] R. Ross and V.S. Subrahmanian, Aggregate operators in probabilistic databases, Journal of the ACM, vol.52, no.1, pp.54-101, 2005.

[18] D. Dey and S. Sarkar, Generalized normal forms for probabilistic relational data, IEEE Transactions on Knowledge and Data Engineer ing, vol.14, no.3, pp.485-497, 1992.

[19] T. Eiter, T. Lukasiewicz, and M. Walter, A data model and algebra for probabilistic complex val ues, Annals of Mathematics and Arti cial Intelligence, vol.33, pp.205-252, 2001.

[20] S.K. Lee, An extended relational database model for uncertain and imprecise information, Proceedings of 18th Conference on Very Large Data Bases, Vancouver, Canada, pp.211-220, 1992.

[21] H. Nguyen, A probabilistic relational database model and algebra, Journal of Computer Sci ence and Cybernetics, vol.31, no.4, pp.305-321, 2015.

[22] H. Nguyen, T.N. Nguyen, and T.T.N. Tran, A probabilistic relational database model with uncertain multivalued attributes, ICIC Express Letters, vol. 16, no.3, pp.241-248, 2022.

[23] H. Nguyen, Extending probabilistic relational database model with uncertain multivalued at tributes, International Journal of Innovative Computing, Information and Control, vol.18, no.5, pp.14771492, 2022.

[24] V. Biazzo, R. Giugno, T. Lukasiewicz, and V. S. Subrahmanian, Temporal probabilistic ob ject bases, IEEE Transactions on Knowledge and Data Engineering, vol.15, no.4, pp. 921-939, 2003.

[25] H. Nguyen, Extending relational database model for uncertain information, Journal of Computer Science and Cybernetics, vol.35, no.4, pp.355-372, 2019

[26] T. Friedman and G. Broeck, Symbolic querying of vector spaces: probabilistic databases meet re lational embeddings, Proceedings of 36th Con ference on Uncertainty in Arti cial Intelligence, Toronto, Canada, vol.124, pp.1268-1277, 2020.

[27] A. Gilad, A. Imber, and B. Kimelfeld, The consistency of probabilistic databases within dependent cells, Proceedings of 26th Interna tional Conference on Database Theory, Ioannina, Greece, pp. 22:1-22:19, 2023.

[28] T.V. Bremen and K.S. Meel, Probabilistic query evaluation: The combined FPRAS land scape, Proceedings of 42th ACM SIGMOD SIGACT-SIGAI Symposium on Principles of Database Systems, New York, USA, pp 339347, 2023.