ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

HOMEWORK 04

Khai thác dữ liệu đồ thị

Sinh viên thực hiện: 21127229 - Dương Trường Bình

Giảng viên hướng dẫn: Lê Ngọc Thành
Lê Nhựt Nam

Lớp: 21KHDL

Mục lục

1 Problem 1

Câu hỏi: Trình bày điểm giống và khác nhau giữa bài toán dự đoán liên kết và bài toán suy luận liên kết bị thiếu?

Trả lời:

Điểm giống nhau:

1. **Mục tiêu chung:** Cả hai bài toán đều nhằm xác định các liên kết tiềm năng trong một mạng lưới hoặc đồ thị mà hiện tại chưa được quan sát thấy.
2. **Phương pháp tiếp cận:** Cả hai bài toán thường sử dụng các kỹ thuật phân tích tương tự để đánh giá khả năng tồn tại của một liên kết, như phân tích cấu trúc tô-pô của mạng, độ tương đồng giữa các nút, v.v.
3. **Ứng dụng:** Cả hai bài toán đều có ứng dụng rộng rãi trong nhiều lĩnh vực như mạng xã hội, sinh học, thương mại điện tử, hệ thống khuyến nghị, v.v.
4. **Thách thức:** Cả hai bài toán đều phải đối mặt với thách thức của việc xử lý dữ liệu lớn và phức tạp, cũng như đảm bảo độ chính xác của dự đoán hoặc suy luận.

Điểm khác nhau:

1. **Khung thời gian:**
   - *Dự đoán liên kết:* Tập trung vào việc dự đoán các liên kết mới có thể xuất hiện trong tương lai.
   - *Suy luận liên kết bị thiếu:* Tìm kiếm các liên kết đã tồn tại nhưng chưa được quan sát hoặc ghi nhận trong dữ liệu hiện tại.
2. **Bản chất của dữ liệu:**
   - *Dự đoán liên kết:* Thường làm việc với dữ liệu động, có tính thời gian.
   - *Suy luận liên kết bị thiếu:* Thường làm việc với dữ liệu tĩnh hoặc ảnh chụp của mạng tại một thời điểm cụ thể.
3. **Bản chất của liên kết:**
   - *Dự đoán liên kết:* Các liên kết được dự đoán thực sự chưa tồn tại tại thời điểm phân tích.

- *Suy luận liên kết bị thiếu:* Các liên kết được suy luận có thể đã tồn tại nhưng bị bỏ sót do hạn chế trong quá trình thu thập dữ liệu.

4. **Động lực:**
   - *Dự đoán liên kết:* Thường được sử dụng để hiểu về sự phát triển và động lực của mạng theo thời gian.
   - *Suy luận liên kết bị thiếu:* Thường được sử dụng để cải thiện tính đầy đủ và chính xác của dữ liệu mạng hiện có.
5. **Phương pháp đánh giá:**
   - *Dự đoán liên kết:* Đánh giá bằng cách so sánh các dự đoán với các liên kết thực tế xuất hiện trong tương lai.
   - *Suy luận liên kết bị thiếu:* Đánh giá bằng cách kiểm tra độ chính xác của các liên kết được suy luận so với dữ liệu thực tế đã biết nhưng bị ẩn đi, thường thông qua các thử nghiệm với dữ liệu được kiểm soát.
6. **Ứng dụng cụ thể:**
   - *Dự đoán liên kết:* Thường được sử dụng trong các hệ thống đề xuất, dự báo xu hướng, phân tích sự phát triển của mạng xã hội, v.v.
   - *Suy luận liên kết bị thiếu:* Thường được sử dụng trong việc hoàn thiện dữ liệu, phát hiện các mối quan hệ tiềm ẩn trong các bộ dữ liệu không đầy đủ, cải thiện chất lượng dữ liệu thu thập được.

2 Problem 2

Câu hỏi Cho ví dụ về bài toán dự đoán liên kết trong một miền tri thức cụ thể?

Trả lời:

Một ví dụ cụ thể về bài toán dự đoán liên kết trong lĩnh vực nghiên cứu khoa học là dự đoán sự hợp tác giữa các nhà khoa học trong tương lai. Chúng ta có thể mô tả bài toán này như sau:

- **Miền tri thức:** Mạng lưới cộng tác nghiên cứu khoa học
- **Mô tả đồ thị:**

  - Đỉnh (V): Mỗi đỉnh đại diện cho một nhà nghiên cứu.
  - Cạnh (E): Mỗi cạnh biểu thị sự cộng tác giữa hai nhà nghiên cứu, thể hiện qua việc họ đồng tác giả trong ít nhất một bài báo khoa học.
  - Thuộc tính thời gian: Mỗi cạnh có một dấu thời gian t(e) tương ứng với thời điểm bài báo được công bố.

- **Dữ liệu đầu vào:**
  - G[t_0, t'$_0$]: Đồ thị con chứa tất cả các cộng tác từ thời điểm t_0 đến t'$_0$ (ví dụ: từ năm 2010 đến 2020).
  - Thông tin bổ sung về các nhà nghiên cứu: lĩnh vực nghiên cứu, trường đại học, số lượng bài báo, chỉ số h-index, v.v.
- **Mục tiêu:** Dự đoán các cộng tác mới có khả năng xảy ra trong khoảng thời gian từ t_1 đến t'$_1$ (ví dụ: từ năm 2021 đến 2025).
- **Phương pháp tiếp cận:**
  1. Phân tích cấu trúc tô-pô của mạng lưới hiện tại:
     - Xác định các nhóm nghiên cứu (cộng đồng) hiện có.
     - Tính toán các độ đo trung tâm (centrality measures) cho mỗi nhà nghiên cứu.
  2. Áp dụng các độ đo tương đồng:
     - Common Neighbors: Số lượng cộng tác chung giữa hai nhà nghiên cứu.
     - Jaccard Coefficient: Tỷ lệ giữa số lượng cộng tác chung và tổng số cộng tác của hai nhà nghiên cứu.
     - Adamic/Adar: Xem xét tầm quan trọng của các cộng tác chung.
  3. Xem xét các yếu tố bổ sung:
     - Sự tương đồng trong lĩnh vực nghiên cứu.
     - Khoảng cách địa lý giữa các trường đại học.
     - Sự tương đồng trong mô hình xuất bản (ví dụ: các tạp chí ưa thích).
  4. Áp dụng các kỹ thuật học máy:
     - Sử dụng các thuật toán như Random Forest hoặc Gradient Boosting để kết hợp các đặc trưng và dự đoán xác suất hợp tác.

- **Kết quả và ứng dụng:**
  - Danh sách các cặp nhà nghiên cứu có khả năng cao sẽ cộng tác trong tương lai.
  - Có thể được sử dụng để:
    * Đề xuất cộng tác tiềm năng cho các nhà nghiên cứu.
    * Hỗ trợ các tổ chức tài trợ trong việc xác định các nhóm nghiên cứu tiềm năng.
    * Dự đoán xu hướng nghiên cứu mới nổi dựa trên các cộng tác được dự báo.
- **Thách thức:**
  - Xử lý dữ liệu lớn và phức tạp từ nhiều nguồn khác nhau.
  - Cân bằng giữa các yếu tố cấu trúc mạng và thông tin ngữ cảnh.
  - Đánh giá hiệu suất của mô hình dự đoán trong một lĩnh vực đang phát triển nhanh chóng.

3 Problem 3

Câu hỏi: Cho trước một đồ thị vô hướng, không trọng số $G = (V, E)$ thể hiện cấu trúc tô-pô của một mạng xã hội, trong đó mỗi cạnh $e = \langle u, v \rangle \in E$ thể hiện một tương tác giữa u và v xuất hiện ở thời gian $t(e)$. Với hai thời điểm t và $t' > t$, gọi $G[t, t']$ là đồ thị con của G bao gồm tất cả các cạnh trong khoảng thời gian từ t đến t'. Cho bốn thời điểm $t_0 < t'_0 \leq t_1 < t'_1$, ta có bài toán dự đoán liên kết:

- **Đầu vào:** $G[t_0, t'_0]$ (training interval)
- **Đầu ra:** Danh sách các cạnh không có trong $G[t_0, t'_0]$ mà được dự đoán xuất hiện trong $G[t_1, t'_1]$ (test interval)

Làm thế nào có thể đánh giá hiệu suất của phương pháp đề xuất cho việc giải quyết bài toán này? Nói cách khác, bạn hãy trình bày những metric có thể được sử dụng cho bài toán này.

Trả lời:

Để đánh giá hiệu suất của các phương pháp dự đoán liên kết trong bối cảnh này, chúng ta có thể sử dụng nhiều metric khác nhau. Dưới đây là các metric phổ biến và cách chúng được áp dụng:

1. Độ chính xác (Precision)

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Trong đó:

- TP (True Positive): Số liên kết được dự đoán đúng là xuất hiện trong $G[t_1, t_1']$.
- FP (False Positive): Số liên kết được dự đoán sai là xuất hiện trong $G[t_1, t_1']$.

Ý nghĩa: Tỷ lệ các liên kết được dự đoán đúng trong tổng số liên kết được dự đoán là sẽ xuất hiện.

2. Độ thu hồi (Recall)

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

Trong đó:

- FN (False Negative): Số liên kết thực sự xuất hiện trong $G[t_1, t_1']$ nhưng không được dự đoán.

Ý nghĩa: Tỷ lệ các liên kết thực sự xuất hiện được dự đoán đúng.

3. F1-score

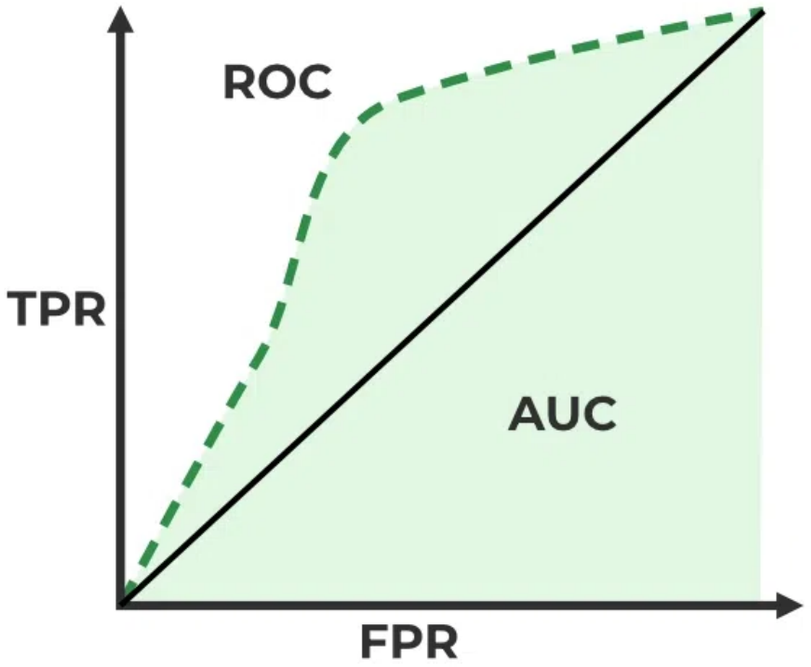
$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

Ý nghĩa: Trung bình điều hòa của Precision và Recall, cung cấp một số đo cân bằng giữa hai metric này.

4. AUC-ROC

AUC-ROC (Area Under the Curve - Receiver Operating Characteristic) đo diện tích dưới đường cong ROC, là một đồ thị biểu diễn tỷ lệ True Positive Rate (TPR) so với False Positive Rate (FPR) tại các ngưỡng khác nhau.

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN} \tag{4}$$



Ý nghĩa: Đánh giá khả năng phân biệt của mô hình giữa các liên kết sẽ xuất hiện và không xuất hiện. Giá trị AUC-ROC từ 0.5 (dự đoán ngẫu nhiên) đến 1 (dự đoán hoàn hảo).

5. AUC-PR (Area Under the Curve - Precision-Recall)

Tương tự như AUC-ROC, nhưng sử dụng đường cong Precision-Recall thay vì ROC.

Ý nghĩa: Đặc biệt hữu ích khi dữ liệu mất cân bằng, tức là số lượng liên kết thực tế ít hơn nhiều so với số lượng không liên kết.

6. Top-K Precision

$$Top - KPrecision = \frac{\text{Số liên kết đúng trong K dự đoán hàng đầu}}{K} \tag{5}$$

Ý nghĩa: Đánh giá độ chính xác của mô hình khi chỉ xem xét K dự đoán có điểm số cao nhất.

7. Mean Average Precision (MAP)

$$MAP = \frac{1}{|U|} \sum_{u \in U} \frac{1}{m_u} \sum_{k=1}^{m_u} \text{Precision}_u(@k) \tag{6}$$

Trong đó U là tập các nút, m_u là số liên kết thực của nút u, và $\text{Precision}_u(@k)$ là độ chính xác ở vị trí k trong danh sách dự đoán cho nút u.

Ý nghĩa: Đánh giá hiệu suất của mô hình trên nhiều truy vấn khác nhau, có tính đến thứ tự của các dự đoán.

8. Normalized Discounted Cumulative Gain (NDCG)

NDCG đánh giá chất lượng của xếp hạng dự đoán, có tính đến vị trí của các liên kết đúng trong danh sách xếp hạng.

$$NDCG@k = \frac{DCG@k}{IDCG@k} \tag{7}$$

Trong đó DCG@k là Discounted Cumulative Gain tại vị trí k, và IDCG@k là DCG lý tưởng.
Ý nghĩa: Đánh giá chất lượng xếp hạng của mô hình, đặc biệt hữu ích khi quan tâm đến thứ tự của các dự đoán.

Lựa chọn metric phù hợp

Việc lựa chọn metric phù hợp phụ thuộc vào đặc điểm cụ thể của bài toán:

- Nếu dữ liệu mất cân bằng nghiêm trọng (số lượng liên kết mới xuất hiện ít hơn nhiều so với số lượng không liên kết), nên ưu tiên sử dụng AUC-PR và F1-score.
- Nếu quan tâm đến chất lượng xếp hạng của các dự đoán, nên sử dụng Top-K Precision hoặc MAP.
- Nếu cần đánh giá tổng quát khả năng phân biệt của mô hình, AUC-ROC là một lựa chọn tốt.
- Trong nhiều trường hợp, việc sử dụng kết hợp nhiều metric sẽ cung cấp cái nhìn toàn diện hơn về hiệu suất của mô hình.

4 Problem 4: Tiếp cận đỉnh dựa trên tương đồng tô-pô đồ thị (thiên hướng cục bộ)

Trong tiếp cận này, chúng ta sẽ cho một số độ đo:

- Common Neighbors
- Jaccard's Coefficient

- Adamic/Adar
- Preferential Attachment

Câu hỏi: Bạn chọn một trong bốn độ đo và trình bày về nó. Cho ví dụ minh họa.

Trả lời

Định nghĩa

Common Neighbors là một phương pháp đơn giản và hiệu quả để dự đoán liên kết trong đồ thị. Phương pháp này dựa trên nguyên tắc rằng hai đỉnh có nhiều láng giềng chung có khả năng cao sẽ kết nối với nhau trong tương lai. Đây là một trong những độ đo phổ biến được sử dụng để đánh giá mức độ tương đồng giữa hai đỉnh trong một đồ thị.

Trong các ứng dụng liên quan đến mạng xã hội, Common Neighbors giúp xác định sự liên quan giữa hai đối tượng dựa trên số lượng đỉnh chung mà cả hai cùng liên kết đến. Phương pháp này đặc biệt hữu ích trong việc gợi ý kết bạn, đề xuất sản phẩm, hoặc dự đoán các mối quan hệ tiềm năng trong mạng lưới.

Công thức và cách tính

Cho một đồ thị $G = (V, E)$ với tập hợp các đỉnh V và tập hợp các cạnh E. Độ đo Common Neighbors giữa hai đỉnh u và v được định nghĩa như sau:

$$CN(u, v) = |\Gamma(u) \cap \Gamma(v)|$$

Trong đó:

- $\Gamma(u)$ là tập hợp các láng giềng của đỉnh u
- $\Gamma(v)$ là tập hợp các láng giềng của đỉnh v
- $|\Gamma(u) \cap \Gamma(v)|$ là số lượng phần tử trong tập giao của $\Gamma(u)$ và $\Gamma(v)$

Nguyên lý hoạt động

Common Neighbors dựa trên hiệu ứng đóng tam giác (triangle closing effect) trong mạng xã hội. Hiệu ứng này cho rằng nếu hai người có nhiều bạn chung, họ có khả năng cao sẽ trở thành bạn của nhau. Phương pháp này giả định rằng nếu hai nút có nhiều láng giềng chung thì khả năng chúng sẽ kết nối với nhau là cao.

Ưu điểm

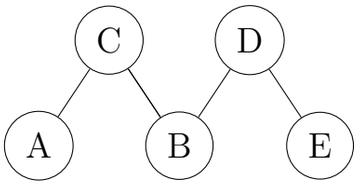
- Đơn giản, dễ hiểu và dễ triển khai
- Hiệu quả trong nhiều loại mạng xã hội và mạng sinh học
- Tính toán nhanh chóng, phù hợp cho các đồ thị lớn
- Không yêu cầu thông tin về thuộc tính của các đỉnh, chỉ cần cấu trúc topo của đồ thị

Nhược điểm

- Không xét đến cấu trúc toàn cục của đồ thị
- Có thể kém hiệu quả trong các mạng thưa (sparse networks)
- Không phân biệt được tầm quan trọng của các láng giềng chung
- Không tính đến sự khác biệt về số lượng láng giềng mà mỗi đỉnh có. Nếu một trong hai đỉnh có rất nhiều láng giềng, thì số lượng đỉnh chung có thể lớn ngay cả khi không có sự liên quan trực tiếp nào

Ví dụ minh họa

Xét đồ thị sau:



Ta sẽ tính Common Neighbors cho hai cặp đỉnh: (A,E) và (A,B)

1) Tính Common Neighbors cho cặp đỉnh (A,E):

- $\Gamma(A) = \{C\}$
- $\Gamma(E) = \{D\}$
- $CN(A, E) = |\Gamma(A) \cap \Gamma(E)| = |\{\}| = 0$

2) Tính Common Neighbors cho cặp đỉnh (A,B):

- $\Gamma(A) = \{C\}$
- $\Gamma(B) = \{C, D\}$

- $CN(A, B) = |\Gamma(A) \cap \Gamma(B)| = |\{C\}| = 1$

Kết luận: Trong ví dụ này, cặp đỉnh (A,B) có khả năng kết nối cao hơn cặp đỉnh (A,E) vì có nhiều láng giềng chung hơn. Cụ thể, A và B có 1 láng giềng chung là C, trong khi A và E không có láng giềng chung nào.

Kết luận

Common Neighbors là một phương pháp đơn giản nhưng hiệu quả trong việc dự đoán liên kết trong đồ thị. Mặc dù có một số hạn chế, phương pháp này vẫn được sử dụng rộng rãi trong nhiều ứng dụng thực tế, đặc biệt là trong các hệ thống gợi ý và phân tích mạng xã hội. Tuy nhiên, để có kết quả tốt hơn, người ta thường kết hợp Common Neighbors với các phương pháp khác hoặc sử dụng các biến thể cải tiến của nó.

5 Problem 5: Tiếp cận đường đi dựa trên tương đồng tô-pô đồ thị (thiên hướng toàn bộ)

Câu hỏi: Bạn chọn một trong hai độ đo (Hit hoặc Katz) và trình bày về nó. Cho ví dụ minh họa.

Trả lời:

Định nghĩa

Độ đo Katz là một phương pháp dự đoán liên kết dựa trên đường đi trong đồ thị, được đề xuất bởi Leo Katz vào năm 1953. Đây là một phương pháp tiếp cận toàn cục, xem xét không chỉ láng giềng trực tiếp mà còn cả cấu trúc tổng thể của đồ thị.

Công thức và cách tính

Cho một đồ thị $G = (V, E)$, độ đo Katz giữa hai đỉnh u và v được định nghĩa như sau:

$$Katz(u, v) = \sum_{l=1}^{\infty} \beta^l \cdot |paths_{u,v}^{<l>}|$$

Trong đó:

- $|paths_{u,v}^{<l>}|$ là số lượng đường đi độ dài l từ u đến v

- β là một hằng số giảm trừ (damping factor), thường chọn $0 < \beta < 1$
- l là độ dài của đường đi

Nguyên lý hoạt động

Độ đo Katz xem xét tất cả các đường đi có thể có giữa hai đỉnh, nhưng gán trọng số thấp hơn cho các đường đi dài hơn. Điều này dựa trên giả định rằng các kết nối gián tiếp (qua nhiều bước trung gian) ít quan trọng hơn các kết nối trực tiếp.

Cách tính toán trên ma trận

Trong thực tế, công thức Katz có thể được tính toán hiệu quả bằng cách sử dụng đại số ma trận:

$$Katz = (I - \beta A)^{-1} - I$$

Trong đó:

- I là ma trận đơn vị
- A là ma trận kề của đồ thị
- β là hằng số giảm trừ, thường chọn $\beta < \frac{1}{\lambda_1}$ với λ_1 là giá trị riêng lớn nhất của A

Ưu điểm và nhược điểm

Ưu điểm:

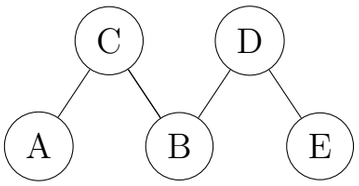
- Xem xét cấu trúc toàn cục của đồ thị, không chỉ giới hạn ở láng giềng trực tiếp
- Có thể phát hiện các mối quan hệ tiềm ẩn không rõ ràng từ cấu trúc cục bộ
- Hiệu quả trong việc xử lý các đồ thị lớn và phức tạp
- Có thể điều chỉnh độ quan trọng của các đường đi dài thông qua tham số β

Nhược điểm:

- Phức tạp hơn về mặt tính toán so với các phương pháp dựa trên láng giềng
- Việc chọn giá trị β phù hợp có thể ảnh hưởng đáng kể đến kết quả
- Có thể bị ảnh hưởng bởi các đỉnh có bậc cao (hub nodes) trong mạng

Ví dụ minh họa

Xét đồ thị đơn giản sau:



Ta sẽ tính độ đo Katz cho cặp đỉnh (A,E) và (A,B) với $\beta = 0.5$.

1) Cho cặp đỉnh (A,E):

- Đường đi độ dài 1: Không có
- Đường đi độ dài 2: Không có
- Đường đi độ dài 3: A-C-B-D-E (1 đường)
- Đường đi độ dài 4: A-C-B-C-D-E (1 đường)
- Đường đi độ dài 5: A-C-B-C-B-D-E (1 đường)

Tính toán:

$$\begin{aligned} Katz(A,E) &= 0.5^1 \cdot 0 + 0.5^2 \cdot 0 + 0.5^3 \cdot 1 + 0.5^4 \cdot 1 + 0.5^5 \cdot 1 + ... \\ &= 0.125 + 0.0625 + 0.03125 + ... \\ &\approx 0.21875 \end{aligned}$$

2) Cho cặp đỉnh (A,B):

- Đường đi độ dài 1: Không có
- Đường đi độ dài 2: A-C-B (1 đường)
- Đường đi độ dài 3: Không có
- Đường đi độ dài 4: A-C-B-C-B (1 đường)

Tính toán:

$$\begin{aligned} Katz(A,B) &= 0.5^1 \cdot 0 + 0.5^2 \cdot 1 + 0.5^3 \cdot 0 + 0.5^4 \cdot 1 + ... \\ &= 0.25 + 0.0625 + ... \\ &\approx 0.3125 \end{aligned}$$

Kết luận

Trong ví dụ này, ta thấy:

$$Katz(A, B) \approx 0.3125 > Katz(A, E) \approx 0.21875$$

Điều này cho thấy cặp đỉnh (A,B) có khả năng kết nối cao hơn cặp đỉnh (A,E). Kết quả này phù hợp với trực quan về cấu trúc đồ thị, vì A và B gần nhau hơn và có nhiều đường đi ngắn hơn giữa chúng so với A và E.

Độ đo Katz không chỉ xem xét đường đi ngắn nhất mà còn tính đến tất cả các đường đi có thể, giúp nó phản ánh tốt hơn mối quan hệ tổng thể giữa các đỉnh trong đồ thị. Điều này làm cho Katz trở thành một công cụ mạnh mẽ trong dự đoán liên kết, đặc biệt hữu ích trong các mạng phức tạp như mạng xã hội, mạng sinh học, hoặc mạng trích dẫn khoa học.

6 Problem 6: Nghiên cứu về nhân quả

Câu hỏi: Trình bày và so sánh các cấp độ: Associational (kết hợp), Interventional (can thiệp), Counterfactual (phản thực tế) trong suy luận nhân quả.

Trả lời:

6.1 Giới thiệu

Suy luận nhân quả là một lĩnh vực quan trọng trong thống kê và khoa học dữ liệu, nhằm xác định mối quan hệ nguyên nhân-kết quả giữa các biến. Judea Pearl, một nhà khoa học máy tính và thống kê học nổi tiếng, đã đề xuất ba cấp độ suy luận nhân quả: Associational, Interventional và Counterfactual. Mỗi cấp độ này cung cấp một mức độ hiểu biết sâu sắc hơn về mối quan hệ nhân quả.

6.2 Các cấp độ suy luận nhân quả

6.2.1 Associational (Kết hợp)

Định nghĩa: Cấp độ này liên quan đến việc quan sát và mô tả mối quan hệ giữa các biến trong dữ liệu mà không can thiệp vào hệ thống.

Đặc điểm chính:

- Dựa trên dữ liệu quan sát (observational data).
- Sử dụng các phương pháp thống kê truyền thống như hồi quy, phân tích tương quan.
- Có thể phát hiện mối liên hệ giữa các biến nhưng không thể xác định quan hệ nhân quả.

Ví dụ: Quan sát mối quan hệ giữa việc hút thuốc và nguy cơ mắc ung thư phổi trong dân số.

Công cụ toán học: $P(y|x)$ - Xác suất của y khi quan sát thấy x.

6.2.2 Interventional (Can thiệp)

Định nghĩa: Cấp độ này liên quan đến việc can thiệp vào hệ thống bằng cách thay đổi giá trị của một biến và quan sát tác động lên các biến khác.

Đặc điểm chính:

- Đòi hỏi khả năng can thiệp vào hệ thống (thường thông qua thí nghiệm).
- Có thể xác định được hướng của mối quan hệ nhân quả.
- Loại bỏ được một số yếu tố gây nhiễu (confounding factors).

Ví dụ: Thực hiện một thử nghiệm ngẫu nhiên có đối chứng để đánh giá tác động của một loại thuốc mới đối với bệnh nhân.

Công cụ toán học: $P(y|do(x))$ - Xác suất của y khi can thiệp và đặt x một cách cưỡng bức.

6.2.3 Counterfactual (Phản thực tế)

Định nghĩa: Cấp độ này liên quan đến việc suy luận về những gì có thể xảy ra trong các tình huống giả định không thực sự xảy ra.

Đặc điểm chính:

- Đòi hỏi mô hình hóa chi tiết về cơ chế nhân quả.
- Cho phép suy luận về các tình huống không thể quan sát được trong thực tế.
- Cung cấp hiểu biết sâu sắc nhất về mối quan hệ nhân quả.

Ví dụ: Suy luận về việc liệu một bệnh nhân cụ thể có thể đã sống sót nếu họ đã được điều trị bằng một phương pháp khác.

Công cụ toán học: $P(y_x|x', y')$ - Xác suất của y nếu x đã được thực hiện, trong một thế giới mà chúng ta đã quan sát thấy x' và y'.

6.3 So sánh giữa các cấp độ

Tiêu chí	**Associational**	**Interventional**	**Counterfactual**
Độ phức tạp	Thấp	Trung bình	Cao
Yêu cầu dữ liệu	Dữ liệu quan sát	Dữ liệu thử nghiệm	Mô hình nhân quả chi tiết
Khả năng suy luận nhân quả	Hạn chế	Tốt	Rất tốt
Khả năng áp dụng thực tế	Cao	Trung bình	Thấp
Độ tin cậy của kết luận	Thấp	Cao	Rất cao (nhưng phụ thuộc vào giả định)

Bảng 1: So sánh giữa ba cấp độ suy luận nhân quả

6.4 Kết luận

Ba cấp độ suy luận nhân quả - Associational, Interventional, và Counterfactual - cung cấp một khuôn khổ toàn diện để hiểu và phân tích mối quan hệ nhân quả. Mỗi cấp độ đòi hỏi các phương pháp và giả định khác nhau, và cung cấp các mức độ hiểu biết khác nhau về mối quan hệ nhân quả.

Associational là cấp độ cơ bản nhất, dễ thực hiện nhưng có khả năng suy luận nhân quả hạn chế. Interventional cung cấp bằng chứng mạnh mẽ hơn về quan hệ nhân quả nhưng đòi hỏi khả năng can thiệp vào hệ thống. Counterfactual là cấp độ cao nhất, cho phép suy luận về các tình huống giả định, nhưng đòi hỏi mô hình hóa chi tiết và các giả định mạnh.

Trong thực tế, việc kết hợp cả ba cấp độ này có thể cung cấp hiểu biết toàn diện nhất về mối quan hệ nhân quả trong một hệ thống phức tạp.

Tài liệu tham khảo

[1] Network Science Book. *Network Science*. Available online: http://networksciencebook.com. (Accessed: August 2024).

[2] Slide của thầy Lê Ngọc Thành, *Topic 09 - Link prediction*, Trường Đại học Khoa học Tự nhiên, 2024