KHAI PHÁ DỮ LIỆU – CHƯƠNG 2
TIỀN XỬ LÝ DỮ LIỆU (DATA PREPROCESSING)

Tổng hợp kiến thức, câu hỏi tự luận và trắc nghiệm

1. GIỚI THIỆU VỀ TIỀN XỬ LÝ DỮ LIỆU

1.1. Tại sao cần tiền xử lý dữ liệu?

Dữ liệu thực tế thường **không hoàn hảo**: thiếu giá trị (NULL), không nhất quán (inconsistent), có nhiễu (noisy), lỗi định dạng. Tiền xử lý dữ liệu là các bước xử lý dữ liệu gốc để **nâng cao chất lượng dữ liệu**, từ đó nâng cao chất lượng kết quả khai phá.

1.2. Các tiêu chí chất lượng dữ liệu

- **Accuracy (Tính chính xác):** Giá trị thực/đúng được ghi lại.
- **Currency/Timeliness (Tính cập nhật):** Dữ liệu có sẵn và còn hiệu lực tại thời điểm cần dùng.
- **Completeness (Tính đầy đủ):** Tất cả giá trị cho mọi thuộc tính đều được ghi lại.
- **Consistency (Tính nhất quán):** Tất cả dữ liệu cùng loại được biểu diễn theo cùng một cách/định dạng.

1.3. Các kỹ thuật tiền xử lý chính

1. **Data Cleaning (Làm sạch):** Xử lý dữ liệu thiếu, loại bỏ nhiễu và sửa không nhất quán.
2. **Data Integration (Tích hợp):** Kết hợp dữ liệu từ nhiều nguồn thành Data Warehouse.
3. **Data Transformation (Biến đổi):** Chuẩn hóa, tổng hợp, tổng quát hóa.
4. **Data Reduction (Rút gọn):** Giảm kích thước dữ liệu nhưng giữ nguyên thông tin.
5. **Data Discretization (Rời rạc hóa):** Chuyển đổi thuộc tính liên tục thành khoảng rời rạc.

2. MÔ TẢ VÀ TÓM TẮT DỮ LIỆU

2.1. Độ đo xu hướng trung tâm

- **Mean (Trung bình):** $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- **Weighted arithmetic mean:** $\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$
- **Median (Trung vị):** Giá trị giữa trong dữ liệu đã sắp xếp.
- **Mode (Yếu vị):** Giá trị xuất hiện nhiều nhất.
- **Midrange (Trung điểm):** $\frac{\max + \min}{2}$

2.2. Độ đo phân tán

- **Quartiles (Tứ phân vị):** Q1 (25th percentile), Q2 = Median (50th), Q3 (75th).
- **IQR (Khoảng tứ phân vị):** IQR $= Q3 - Q1$.
- **Xác định ngoại lệ:** $\geq Q3 + 1.5 \times \text{IQR}$ hoặc $\leq Q1 - 1.5 \times \text{IQR}$.
- **Extreme outlier:** $\geq Q3 + 3 \times \text{IQR}$ hoặc $\leq Q1 - 3 \times \text{IQR}$.
- **Variance (Phương sai):** $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$
- **Standard deviation:** $\sigma = \sqrt{\sigma^2}$

Phân phối lệch: Nếu Mean $<$ Mode thì dữ liệu lệch âm (negatively skewed); ngược lại lệch dương.

3. LÀM SẠCH DỮ LIỆU

3.1. Xử lý dữ liệu thiếu

Nguyên nhân:

- Khách quan (dữ liệu không tồn tại, lỗi hệ thống).
- Chủ quan (lỗi con người).

Giải pháp:

- Không sử dụng bản ghi thiếu (ignore/delete).
- Cập nhật thủ công.
- Thay thế tự động: hằng số toàn cục, giá trị thường gặp (mode), giá trị trung bình (local/global mean), giá trị dự đoán (từ hồi quy/phân lớp).
- Phòng ngừa ngay từ thiết kế: ràng buộc toàn vẹn CSDL.

3.2. Phát hiện ngoại lệ và loại bỏ nhiễu

Dữ liệu bất thường:

- **Ngoại lệ (Outlier):** Đối tượng không tuân theo đặc tính chung của tập dữ liệu.
- **Nhiễu (Noise):** Các ngoại lệ bị loại bỏ/không được chấp nhận.

Phương pháp phát hiện ngoại lệ:

- **Statistical distribution-based:** Dựa trên phân phối thống kê (IQR rule).
- **Distance-based:** Dựa trên khoảng cách tới các điểm lân cận.
- **Density-based:** Vùng thưa thớt được coi là ngoại lệ.
- **Deviation-based:** So sánh với đặc tính tổng quát của nhóm.

Phương pháp loại bỏ nhiễu (Noise Removal):

- **Binning:** Sắp xếp dữ liệu vào các bin (bucket), sau đó làm mịn bằng: giá trị trung bình bin (bin means), trung vị bin (bin median), hoặc biên bin (bin boundaries).
- **Regression:** Dùng đường hồi quy để dự đoán và thay thế giá trị nhiễu.
- **Cluster analysis:** Nhóm dữ liệu và xác định điểm nhiễu nằm ngoài cụm.

3.3. Xử lý không nhất quán dữ liệu

Nguyên nhân: Cách đặt tên/mã hóa không nhất quán, định dạng khác nhau, lỗi hệ thống/con người.

Ví dụ: “2004/12/25” vs “25/12/2004”; vi phạm ràng buộc khóa ngoại.

Giải pháp: Dùng metadata để sửa, áp dụng ràng buộc dữ liệu, sửa thủ công hoặc tự động.

4. TÍCH HỢP DỮ LIỆU

4.1. Các vấn đề tích hợp

- **Entity identification (Nhận dạng thực thể):** Cùng một thực thể nhưng có tên khác nhau ở các nguồn (ví dụ: *cust_id* và *cust_No*; "Male" và "Nam").
- **Schema integration:** Gộp lược đồ từ nhiều nguồn.
- **Redundancy (Dư thừa):** Thuộc tính A có thể suy diễn từ B. Phát hiện bằng phân tích tương quan.
- **Data value conflicts (Xung đột giá trị):** Cùng đối tượng nhưng giá trị khác nhau do biểu diễn/đơn vị khác nhau (ví dụ: GPA [0,4] vs [0,10]; "yes" vs "1").

4.2. Phân tích tương quan để phát hiện dư thừa

Thuộc tính số – Hệ số tương quan Pearson:

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{(N-1)\sigma_A \sigma_B}$$

- $r_{A,B} > 0$: A và B tương quan thuận (có thể xóa một).
- $r_{A,B} = 0$: A và B độc lập.
- $r_{A,B} < 0$: A và B tương quan nghịch.

Thuộc tính danh mục – Kiểm định Chi-square (χ^2):

$$\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}, \quad e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{N}$$

- Bậc tự do (DoF) = $(r-1)(c-1)$.
- Nếu $\chi^2_{\text{tính toán}} \geq \chi^2_{\text{bảng}}$: bác bỏ giả thuyết độc lập $\Rightarrow$ A và B tương quan.

Ví dụ: Nghiên cứu 1500 người về giới tính và sở thích đọc. $\chi^2_{\text{tính}} = 507.93 \gg \chi^2_{\text{bảng}} = 10.828$ (DoF=1, α=0.001) $\Rightarrow$ giới tính và sở thích đọc tương quan.

5. BIẾN ĐỔI DỮ LIỆU

5.1. Các kỹ thuật biến đổi

1. **Smoothing (Làm mịn):** Loại bỏ nhiễu (binning, regression, clustering).

2. **Aggregation (Tổng hợp):** Tóm tắt dữ liệu chi tiết thành dữ liệu tổng hợp (min, max, sum, avg). Hỗ trợ data reduction và phân tích đa mức.

3. **Generalization (Tổng quát hóa):** Thay thế dữ liệu chi tiết bằng khái niệm cấp cao hơn theo phân cấp khái niệm (ví dụ: điểm số cụ thể $\rightarrow$ GPA $\rightarrow$ xếp loại: Giỏi/Khá/Trung bình).

4. **Normalization (Chuẩn hóa):** Biến đổi giá trị về miền giá trị chuẩn.

5. **Attribute/Feature construction:** Tạo thuộc tính mới từ thuộc tính hiện có.

5.2. Các phương pháp chuẩn hóa

(a) Min-max normalization:

$$v' = \frac{v - \min_A}{\max_A - \min_A} \times (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

Biến đổi $v \in [\min_A, \max_A]$ về $[\text{new\_min}_A, \text{new\_max}_A]$.

(b) Z-score normalization:

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

Chuẩn hóa theo trung bình $\bar{A}$ và độ lệch chuẩn σ_A.

(c) Normalization by decimal scaling:

$$v' = \frac{v}{10^j}$$

trong đó j là số nguyên nhỏ nhất sao cho $\max(|v'|) < 1$.

6. RÚT GỌN DỮ LIỆU

6.1. Các chiến lược rút gọn

1. **Data cube aggregation:** Tổng hợp dữ liệu theo nhiều mức độ chi tiết khác nhau (ví dụ: tổng doanh thu theo tuần/tháng/quý).

2. **Attribute subset selection:** Loại bỏ thuộc tính dư thừa hoặc không liên quan, giữ nguyên phân phối xác suất. Đây là bài toán tối ưu – áp dụng heuristics.

3. **Dimensionality reduction:**

    - **PCA (Principal Component Analysis):** Chiếu dữ liệu lên không gian chiều thấp hơn.

- **Wavelet transforms:** Biến đổi wavelet để nén dữ liệu.
- **Correlation analysis:** Loại bỏ thuộc tính có tương quan cao.

4. **Numerosity reduction:** Giảm số lượng bản ghi/đối tượng.
    - **Parametric:** Lưu mô hình thay vì lưu dữ liệu thô (ví dụ: mô hình hồi quy).
    - **Nonparametric:** Histogram, Clustering, Sampling.

6.2. Các phương pháp lấy mẫu (Sampling)

- **SRSWOR** (Simple Random Sample Without Replacement): Lấy mẫu ngẫu nhiên không hoàn lại.
- **SRSWR** (Simple Random Sample With Replacement): Lấy mẫu ngẫu nhiên có hoàn lại.
- **Cluster sample:** Chọn ngẫu nhiên một số cụm, lấy toàn bộ đối tượng trong cụm đó.
- **Stratified sample:** Phân tầng dữ liệu, lấy mẫu theo tỷ lệ từng tầng.

7. RỜI RẠC HÓA DỮ LIỆU VÀ PHÂN CẤP KHÁI NIỆM

7.1. Rời rạc hóa (Data Discretization)

Mục đích: Giảm số giá trị của thuộc tính liên tục bằng cách chia miền giá trị thành các khoảng (interval) và gán nhãn cho mỗi khoảng.

Các phương pháp rời rạc hóa thuộc tính số:

- **Binning:** Chia thành các bin đều hoặc theo tần suất.
- **Histogram analysis:** Phân tích phân phối để xác định khoảng.
- **Chi-square merging (χ^2):** Gộp các khoảng lân cận có phân phối lớp tương tự.
- **Cluster analysis:** Phân cụm dữ liệu 1D, mỗi cụm là một khoảng.
- **Entropy-based discretization:** Tối thiểu hóa entropy khi chia khoảng.
- **Intuitive partitioning:** Chia theo quy tắc tự nhiên (ví dụ: chia theo bội số của 10).

7.2. Xây dựng phân cấp khái niệm (Conceptual Hierarchy)

Mục đích: Hỗ trợ khai phá ở nhiều mức trừu tượng.
Có thể xây dựng phân cấp cho:

- **Dữ liệu danh mục/rời rạc:** Mô tả tường minh theo nhóm hoặc theo quan hệ ngữ nghĩa.
- **Dữ liệu số:** Sử dụng rời rạc hóa để tạo phân cấp (ví dụ: điểm số $\rightarrow$ GPA $\rightarrow$ xếp loại).

8. TÓM TẮT

- Dữ liệu thực tế: thiếu, nhiễu, không nhất quán – tiền xử lý là bắt buộc.
- **Data Cleaning:** Xử lý missing data, loại bỏ nhiễu (binning, regression, clustering), sửa không nhất quán.
- **Data Integration:** Nhận dạng thực thể, phát hiện dư thừa (Pearson, χ^2), giải quyết xung đột giá trị.
- **Data Transformation:** Làm mịn, tổng hợp, tổng quát hóa, chuẩn hóa (min-max, z-score, decimal).
- **Data Reduction:** Data cube, attribute subset selection, PCA, sampling.
- **Data Discretization:** Chuyển liên tục $\rightarrow$ khoảng; phân cấp khái niệm hỗ trợ khai phá đa mức.

9. CÂU HỎI TỰ LUẬN

Câu 1. Tại sao tiền xử lý dữ liệu là bước quan trọng không thể thiếu trong quy trình KDD? Trình bày 4 tiêu chí chất lượng dữ liệu và giải thích tầm quan trọng của từng tiêu chí.

Câu 2. So sánh Mean, Median, Mode và Midrange. Trong tình huống nào thì Median là độ đo xu hướng trung tâm tốt hơn Mean? Cho ví dụ.

Câu 3. Giải thích IQR (Interquartile Range) và quy tắc phát hiện ngoại lệ dựa trên IQR. Sự khác biệt giữa outlier và extreme outlier là gì?

Câu 4. Trình bày 3 nguyên nhân gây ra dữ liệu thiếu và 5 phương pháp xử lý dữ liệu thiếu. Phương pháp nào phù hợp nhất trong trường hợp dữ liệu thiếu hoàn toàn ngẫu nhiên?

Câu 5. Giải thích 3 phương pháp Binning (bin means, bin median, bin boundaries) để loại bỏ nhiễu. Minh họa bằng tập dữ liệu: {4, 8, 15, 21, 21, 24, 25, 28, 34} với 3 bin.

Câu 6. Trình bày các vấn đề trong tích hợp dữ liệu (Data Integration). Tại sao "redundancy" là vấn đề nghiêm trọng và cách phát hiện nó như thế nào?

Câu 7. Giải thích hệ số tương quan Pearson $r_{A,B}$. Khi $r_{A,B} = -0.9$, điều đó có nghĩa là gì đối với hai thuộc tính A và B? Khi nào thì nên xóa một trong hai thuộc tính?

Câu 8. Giải thích kiểm định Chi-square (χ^2) để phân tích tương quan giữa hai thuộc tính danh mục. Bậc tự do và mức ý nghĩa trong kiểm định χ^2 có vai trò gì?

Câu 9. So sánh 3 phương pháp chuẩn hóa: min-max, z-score và decimal scaling. Phương pháp nào phù hợp khi dữ liệu có phân phối Gaussian? Khi nào nên dùng decimal scaling?

Câu 10. Giải thích thuộc tính/tính năng xây dựng (attribute/feature construction) trong Data Transformation. Cho 2 ví dụ cụ thể về cách tạo thuộc tính mới từ thuộc tính hiện có.

Câu 11. Trình bày chiến lược **Attribute Subset Selection** trong rút gọn dữ liệu. Tại sao đây là bài toán tối ưu? Phương pháp heuristics nào được áp dụng?

Câu 12. So sánh các phương pháp lấy mẫu: SRSWOR, SRSWR, Cluster sample và Stratified sample. Phương pháp nào đảm bảo tính đại diện tốt nhất cho dữ liệu phân tầng?

Câu 13. Giải thích **PCA (Principal Component Analysis)** trong giảm chiều dữ liệu. Ý tưởng chính của PCA là gì? Khi nào nên sử dụng PCA trong tiền xử lý DM?

Câu 14. Trình bày phương pháp **Entropy-based discretization**. Tiêu chí nào được sử dụng để tìm điểm chia tối ưu? Tại sao entropy lại hữu ích cho rời rạc hóa?

Câu 15. Giải thích khái niệm **Conceptual Hierarchy (Phân cấp khái niệm)**. Phân cấp khái niệm hỗ trợ khai phá dữ liệu như thế nào? Cho ví dụ xây dựng phân cấp cho thuộc tính "địa điểm".

Câu 16. Phân biệt **lossless reduction** và **lossy reduction** trong rút gọn dữ liệu. Mỗi loại có ví dụ điển hình nào? Khi nào có thể chấp nhận lossy reduction?

Câu 17. Giải thích **phân phối lệch (skewed distribution)**. Data với Mean $<$ Median $<$ Mode có phân phối như thế nào? Điều này ảnh hưởng thế nào đến việc chọn độ đo xu hướng trung tâm?

Câu 18. Trình bày phương pháp **Numerosity Reduction** dạng tham số (parametric). Tại sao lưu mô hình thay vì lưu dữ liệu thô lại giúp rút gọn dữ liệu? Hạn chế của phương pháp này?

Câu 19. Giải thích vấn đề **"Data Value Conflicts"** trong tích hợp dữ liệu. Cho ít nhất 3 ví dụ cụ thể về xung đột giá trị và cách giải quyết.

Câu 20. Trình bày quy trình tiền xử lý hoàn chỉnh cho tập dữ liệu điểm thi gồm 100 sinh viên, có một số điểm NULL, một số điểm bất thường ($>$ 100), điểm từ các lớp khác nhau có thang điểm khác nhau ([0,10] và [0,100]).

10. CÂU HỎI TRẮC NGHIỆM

Câu 1. Tiêu chí chất lượng dữ liệu nào đảm bảo rằng tất cả giá trị cho mọi thuộc tính đều được ghi lại?

A. Accuracy.

B. Timeliness.

C. Completeness.

D. Consistency.

Câu 2. Với tập dữ liệu {25, 25, 40, 45, 50, 60, 60, 60, 65, 80, 85, 85}, Mode là:

A. 25.

B. 56.67.

C. 60.

D. 62.5.

Câu 3. IQR được tính bằng công thức:

A. $Q2 - Q1$.

B. $Q3 - Q1$.

C. $Q3 - Q2$.

D. $\max - \min$.

Câu 4. Một điểm dữ liệu được coi là **ngoại lệ (outlier)** theo quy tắc IQR nếu:

A. $x \geq Q3 + IQR$.

B. $x \geq Q3 + 1.5 \times IQR$ hoặc $x \leq Q1 - 1.5 \times IQR$.

C. $x \geq Q3 + 3 \times IQR$.

D. $x > \text{Mean} + 2\sigma$.

Câu 5. Phương pháp xử lý dữ liệu thiếu nào **không làm thay đổi** tập dữ liệu?

A. Thay bằng giá trị trung bình.

B. Thay bằng giá trị thường gặp (mode).

C. Bỏ qua (ignore) bản ghi thiếu.

D. Thay bằng giá trị dự đoán.

Câu 6. Trong phương pháp **Binning** để loại bỏ nhiễu, **Bin means** thực hiện:

A. Thay mỗi giá trị trong bin bằng giá trị biên gần nhất.

B. Thay mỗi giá trị trong bin bằng giá trị trung bình của bin.

C. Thay mỗi giá trị bằng trung vị của bin.

D. Xóa các giá trị ngoại lệ trong bin.

Câu 7. Khi hệ số tương quan Pearson $r_{A,B} = 0$, hai thuộc tính A và B:

A. Tương quan thuận hoàn toàn.

B. Tương quan nghịch hoàn toàn.

C. Độc lập với nhau.

D. Không thể kết luận.

Câu 8. Trong kiểm định Chi-square, bậc tự do (Degree of Freedom) được tính bằng:

A. $(r+1)(c+1)$.

B. $(r-1)(c-1)$.

C. $r \times c$.

D. $r+c-1$.

Câu 9. Min-max normalization biến đổi giá trị v từ miền $[\min_A, \max_A]$ về $[0,1]$ bằng công thức:

A. $v' = v/\max_A$

B. $v' = (v-\bar{A})/\sigma_A$

C. $v' = (v-\min_A)/(\max_A - \min_A)$

D. $v' = v/10^j$

Câu 10. Z-score normalization sử dụng:

A. Giá trị min và max.

B. Trung bình và độ lệch chuẩn.

C. Trung vị và IQR.

D. Chỉ giá trị max.

Câu 11. Phương pháp nào trong Data Transformation giúp tạo ra **thuộc tính mới** không có trong tập dữ liệu gốc?

A. Smoothing.

B. Aggregation.

C. Generalization.

D. Attribute/Feature construction.

Câu 12. Chiến lược rút gọn dữ liệu nào lưu **mô hình** thay vì dữ liệu thô?

A. Nonparametric numerosity reduction.

B. Parametric numerosity reduction.

C. Attribute subset selection.

D. Data cube aggregation.

Câu 13. **PCA (Principal Component Analysis)** thực hiện:

A. Loại bỏ các bản ghi dư thừa.

B. Chiếu dữ liệu lên không gian có số chiều nhỏ hơn, giữ lại phương sai lớn nhất.

C. Phân tầng dữ liệu để lấy mẫu đại diện.

D. Thay thế giá trị thiếu bằng giá trị trung bình.

Câu 14. Trong rút gọn dữ liệu, **Attribute Subset Selection** loại bỏ thuộc tính nào?

A. Chỉ thuộc tính số.

B. Thuộc tính dư thừa (redundant) hoặc không liên quan (irrelevant).

C. Thuộc tính có giá trị thiếu.

D. Chỉ thuộc tính danh mục.

Câu 15. Phương pháp lấy mẫu nào đảm bảo tính đại diện tốt nhất khi dữ liệu có các nhóm/tầng với tỷ lệ khác nhau?

A. SRSWOR.

B. SRSWR.

C. Cluster sample.

D. Stratified sample.

Câu 16. Rời rạc hóa (Data Discretization) chuyển đổi:

A. Thuộc tính danh mục thành thuộc tính số.

B. Thuộc tính liên tục thành các khoảng (interval) rời rạc có nhãn.

C. Thuộc tính nhị phân thành đa giá trị.

D. Dữ liệu 2D thành dữ liệu 1D.

Câu 17. Phân cấp khái niệm (Conceptual Hierarchy) hỗ trợ DM bằng cách:

A. Giảm kích thước tập dữ liệu về 1 bản ghi.

B. Cho phép khai phá ở nhiều mức trừu tượng khác nhau.

C. Tự động phát hiện ngoại lệ.

D. Chuẩn hóa thuộc tính về khoảng [0,1].

Câu 18. **Entity identification** trong Data Integration giải quyết vấn đề:

A. Xung đột giá trị giữa các nguồn dữ liệu.

B. Hai tên khác nhau ở hai nguồn nhưng cùng chỉ một thực thể.

C. Dữ liệu số có đơn vị đo lường khác nhau.

D. Thiếu khóa chính trong bảng dữ liệu.

Câu 19. Phương pháp nào **không** phải là kỹ thuật loại bỏ nhiễu (noise removal)?

A. Binning.

B. Regression.

C. Cluster analysis.

D. Min-max normalization.

Câu 20. Với bộ dữ liệu 12 sinh viên: Q1 = 42.5, Q3 = 72.5. Giá trị nào sau đây là **ngoại lệ**?

A. 50.

B. 85.

C. 95.

D. 25.

Câu 21. Tiêu chí chất lượng dữ liệu **Consistency** vi phạm khi:

A. Một giá trị thuộc tính bị NULL.

B. Dữ liệu đã lỗi thời.

C. Cùng loại dữ liệu nhưng được biểu diễn theo nhiều định dạng khác nhau (ví dụ: ngày tháng).

D. Giá trị đo lường sai so với thực tế.

Câu 22. Trong **Decimal Scaling**, giá trị j được chọn là:

A. Giá trị lớn nhất trong tập dữ liệu.

B. Số nguyên nhỏ nhất sao cho $\max(|v'|) < 1$.

C. Số chữ số thập phân của giá trị trung bình.

D. Bậc của phương sai.

Câu 23. Generalization trong Data Transformation khác Aggregation ở chỗ:

A. Generalization chỉ áp dụng cho thuộc tính số.

B. Generalization thay thế dữ liệu chi tiết bằng khái niệm cấp cao hơn theo phân cấp khái niệm.

C. Aggregation tạo ra thuộc tính mới không có trong dữ liệu gốc.

D. Generalization chỉ áp dụng cho dữ liệu danh mục.

Câu 24. Phương pháp phát hiện ngoại lệ nào dựa trên việc so sánh một điểm với **mật độ** của vùng xung quanh nó?

A. Statistical distribution-based.

B. Distance-based.

C. Density-based.

D. Deviation-based.

Câu 25. **Data Cube Aggregation** trong rút gọn dữ liệu phù hợp với dữ liệu kiểu:

A. Chỉ dữ liệu nhị phân.

B. Additive và semi-additive (dữ liệu số có thể tổng hợp).

C. Dữ liệu văn bản không cấu trúc.

D. Dữ liệu ảnh và video.

Câu 26. Số phương pháp chuẩn hóa (normalization) được trình bày trong slide C2 là:

A. 2.

B. 3.

C. 4.

D. 5.

Câu 27. Nếu $r_{A,B} = 0.95$, nhà phân tích nên:

A. Giữ cả hai thuộc tính A và B vì chúng tương quan nghịch.

B. Xem xét xóa một trong hai thuộc tính vì chúng có tương quan thuận rất cao (dư thừa).

C. Không làm gì vì $r_{A,B} < 1$.

D. Nhân đôi thuộc tính để tăng thông tin.

Câu 28. Trong phương pháp Binning, **Bin boundaries** thực hiện:

A. Thay mỗi giá trị bằng giá trị trung bình của bin.

B. Thay mỗi giá trị bằng giá trị biên gần nhất (min hoặc max của bin).

C. Xóa toàn bộ giá trị trong bin.

D. Thay bằng giá trị trung vị của bin.

Câu 29. Trong Data Integration, **chi-square test** được dùng để:

A. Chuẩn hóa dữ liệu về khoảng [0,1].

B. Phân tích tương quan giữa hai thuộc tính danh mục.

C. Phát hiện ngoại lệ trong phân phối chuẩn.

D. Rời rạc hóa thuộc tính liên tục.

Câu 30. Mục tiêu chính của **Data Reduction** trong tiền xử lý dữ liệu là:

A. Tăng kích thước tập dữ liệu để có nhiều mẫu huấn luyện hơn.

B. Biến đổi tập dữ liệu thành tập nhỏ hơn trong khi giữ nguyên tính đầy đủ thông tin.

C. Chỉ giảm số thuộc tính (chiều) mà không giảm số bản ghi.

D. Loại bỏ tất cả các ngoại lệ khỏi tập dữ liệu.

Câu 31. Phân phối lệch âm (negatively skewed) có đặc điểm:

A. Mean $>$ Median $>$ Mode.

B. Mean $<$ Median $<$ Mode.

C. Mean $=$ Median $=$ Mode.

D. Mode $<$ Mean $<$ Median.

Câu 32. Giá trị e_{ij} trong công thức Chi-square được tính bằng:

A. Tần suất quan sát thực tế.

B. Tần suất kỳ vọng nếu hai thuộc tính độc lập: $\text{count}(A = a_i) \times \text{count}(B = b_j)/N$.

C. Trung bình của o_{ij} trên toàn bảng.

D. Phương sai của phân phối.

ĐÁP ÁN

Câu hỏi tự luận – Hướng dẫn trả lời

Câu	Nội dung cần trình bày
1	Dữ liệu thực thường thiếu, nhiễu, không nhất quán. 4 tiêu chí: Accuracy (giá trị đúng), Timeliness (cập nhật), Completeness (đầy đủ), Consistency (nhất quán). Chất lượng dữ liệu ảnh hưởng trực tiếp đến chất lượng mô hình DM.
2	Mean: trung bình tất cả giá trị; Median: giá trị giữa; Mode: giá trị phổ biến nhất; Midrange: (max+min)/2. Median tốt hơn khi dữ liệu lệch mạnh hoặc có ngoại lệ (ví dụ: thu nhập) vì không bị ảnh hưởng bởi giá trị cực đoan.
3	IQR = Q3 - Q1. Outlier: $\geq Q3 + 1.5 \times IQR$ hoặc $\leq Q1 - 1.5 \times IQR$. Extreme: $\geq Q3 + 3 \times IQR$. Extreme outlier xa hơn, ít khả năng là dữ liệu hợp lệ hơn outlier thông thường.
4	Nguyên nhân: không tồn tại, lỗi hệ thống, lỗi con người. 5 giải pháp: ignore, cập nhật thủ công, hằng số toàn cục, giá trị mode/mean, giá trị dự đoán. Khi missing hoàn toàn ngẫu nhiên: dùng giá trị mean toàn cục (global average) là hợp lý nhất.
5	Sắp xếp: {4, 8, 15, 21, 21, 24, 25, 28, 34}. 3 bin = {4,8,15}, {21,21,24}, {25,28,34}. Bin means: {9,9,9}, {22,22,22}, {29,29,29}. Bin median: {8,8,8}, {21,21,21}, {28,28,28}. Bin boundaries: {4,4,15}, {21,21,24}, {25,25,34}.
6	Vấn đề: entity identification, schema integration, redundancy (A suy ra từ B), data value conflicts. Redundancy nghiêm trọng vì làm lệch kết quả khai phá, tốn bộ nhớ và thời gian. Phát hiện bằng Pearson (số) hoặc chi-square (danh mục).
7	$r_{A,B} = -0.9$: tương quan nghịch rất mạnh (khi A tăng thì B giảm và ngược lại). Xem xét xóa một thuộc tính khi $\|r\|$ cao (ví dụ > 0.8) vì chúng mang thông tin gần như giống nhau. Không nên xóa cả hai vì sẽ mất thông tin.
8	$\chi^2 = \sum(o_{ij} - e_{ij})^2 / e_{ij}$. DoF = $(r-1)(c-1)$: số chiều của phân phối chi-square. Mức ý nghĩa α: xác suất bác bỏ giả thuyết đúng. So sánh $\chi^2_{\text{tính}}$ với $\chi^2_{\text{bảng}}$ tại DoF và α đã cho.
9	Min-max: bảo tồn quan hệ giữa các giá trị, nhạy với ngoại lệ. Z-score: phù hợp khi dữ liệu Gaussian, không biết min/max trước. Decimal: đơn giản nhưng không bảo tồn phân phối. Gaussian $\rightarrow$ dùng z-score. Khi muốn giá trị $\|v'\| < 1$ dễ dàng $\rightarrow$ decimal.

Câu	Nội dung cần trình bày
10	Feature construction: tạo thuộc tính mới hữu ích hơn. VD1: từ (ngày_bán - ngày_sản_xuất) tạo thuộc tính "tuổi_sản_phẩm". VD2: từ (chiều_cao, cân_nặng) tạo "BMI = cân_nặng/chiều_cao²".
11	Attribute subset selection: loại bỏ thuộc tính dư thừa/không liên quan. Bài toán tối ưu vì: 2^n tập con có thể (với n thuộc tính). Heuristics: (1) Forward selection: bắt đầu từ $\emptyset$, thêm dần thuộc tính tốt nhất; (2) Backward elimination: bắt đầu từ tất cả, xóa dần thuộc tính tệ nhất.
12	SRSWOR: không hoàn lại, mỗi phần tử xuất hiện tối đa 1 lần. SRSWR: có hoàn lại, phần tử có thể lặp. Cluster: chọn cụm ngẫu nhiên, lấy toàn bộ cụm. Stratified: lấy theo tỷ lệ từng tầng $\rightarrow$ đại diện tốt nhất khi dữ liệu không đồng đều giữa các nhóm.
13	PCA tìm các hướng (principal components) có phương sai lớn nhất, chiếu dữ liệu lên không gian đó. Giữ k components đầu tiên thay vì n chiều gốc. Dùng khi: nhiều thuộc tính tương quan cao, cần giảm chiều trước khi áp dụng thuật toán DM.
14	Entropy-based: tại mỗi điểm chia, tính entropy của các lớp trong mỗi khoảng. Chọn điểm chia tối thiểu hóa entropy tổng có trọng số. Entropy hữu ích vì đo mức độ hỗn loạn/thuần của lớp trong mỗi khoảng.
15	Conceptual hierarchy: tổ chức theo cấp bậc trừu tượng. Hỗ trợ DM đa mức: khai phá ở mức tổng quát trước, sau đó drill-down. Ví dụ địa điểm: "Bình Thạnh" $\rightarrow$ "TP.HCM" $\rightarrow$ "Việt Nam" $\rightarrow$ "Châu Á".
16	Lossless: không mất thông tin – ví dụ: attribute subset selection (giữ nguyên phân phối). Lossy: mất một phần thông tin – ví dụ: Sampling, Binning. Chấp nhận lossy khi: mất mát thông tin nhỏ, lợi ích rút gọn lớn hơn chi phí mất mát.
17	Lệch âm (negatively skewed): đuôi dài bên trái; Mode $>$ Median $>$ Mean. Nên dùng Median vì không bị kéo lệch bởi các giá trị cực đoan nhỏ. Dữ liệu thu nhập thường có phân phối lệch dương (positively skewed).
18	Parametric: lưu tham số mô hình (ví dụ: θ_0, θ_1 của hồi quy tuyến tính) thay vì toàn bộ dữ liệu. Hạn chế: chỉ phù hợp khi dữ liệu tuân theo mô hình đã chọn; mất thông tin chi tiết ngoài mô hình.
19	Xung đột biểu diễn: "2004/12/25" vs "25/12/2004" $\rightarrow$ chuẩn hóa định dạng. Xung đột đơn vị: GPA [0,4] vs [0,10] $\rightarrow$ chuyển đổi về cùng thang. Xung đột mã hóa: "yes/no" vs "1/0" $\rightarrow$ thống nhất mã hóa. Giải quyết: dùng metadata và áp dụng quy tắc chuyển đổi nhất quán.

Câu	Nội dung cần trình bày
20	(1) Data Cleaning: xử lý NULL bằng mean/mode; xác định điểm > 100 là outlier → xóa hoặc thay bằng max hợp lệ; (2) Data Integration: xác định cùng sinh viên, cùng môn học từ các lớp khác nhau; (3) Data Transformation: chuẩn hóa thang điểm về [0,10] bằng min-max normalization (chia [0,100] cho 10); (4) Kiểm tra tính nhất quán sau xử lý.

Câu hỏi trắc nghiệm – Đáp án

Câu	ĐA	Câu	ĐA	Câu	ĐA	Câu	ĐA
1	C	11	D	21	C	31	B
2	C	12	B	22	B	32	B
3	B	13	B	23	B	33	C
4	B	14	B	24	C	34	B
5	C	15	D	25	B	35	B
6	B	16	B	26	B	36	B
7	C	17	B	27	B	37	B
8	B	18	B	28	C	38	B
9	C	19	D	29	B	39	B
10	B	20	C	30	A	40	B