Khai khoáng dữ liệu là gì?

Trích xuất từ dữ liệu những thông tin hữu ích nhưng tiềm ẩn, chưa được biết.

Những gì khai khoáng dữ liệu có thể làm?

Mô tả (Description), Ước lượng (Estimation), Dự đoán (Prediction), Phân lớp (Classification), Gom nhóm (Clustering), Kết hợp (Association)

Mô tả chi tiết quá trình khám phá tri thức/phát hiện tri thức/khai phá dữ liệu?

Quy trình khai phá dữ liệu là một chuỗi lặp (iterative) và tương tác (interactive) gồm các bước (giai đoạn) bắt đầu với dữ liệu thô (raw data) và kết thúc với tri thức (knowledge of interest) đáp ứng được sự quan tâm của người sử dụng.

Bao gồm các bước cơ bản sau đây:

- Chọn lọc dữ liệu (Selection)
- Tiền xử lý dữ liệu (Preprocessing)
- Chuyển đổi dữ liệu (Transformation)
- Khai phá dữ liệu (Data mining)
- Đánh giá kết quả mẫu (Interpretation evaluation)

Tại sao phải tiền xử lý dữ liệu?

No quality data, no quality mining results

- Để đưa ra quyết định hiệu quả cần phải dựa trên dữ liệu chất lượng
- Kho dữ liệu cần phải được tích hợp bởi các dữ liệu chất lượng
- Trùng lắp hay thiếu dữ liệu sẽ dẫn đến việc thống kê sai hay hiểu nhầm đặc điểm của dữ liệu

Data preparation, cleaning, and transformation comprises the majority of the work in a data mining application (90%)

Data in the real world is dirty

- Không đầy đủ (incomplete) thiếu giá của thuộc tính, thiếu thông tin thuộc tính quan tâm, … Ví dụ: nghề nghiệp = “”
- Nhiễu (noisy) chứa sai sót hoặc ngoại lệ. Ví dụ: dung lượng = “-40”

- Không nhất quán (inconsistent) có sự sai biệt trong mã hoặc tên. Ví dụ: tuổi = 42, ngày sinh = “03/07/1997” hay đánh giá “1, 2, 3” và đánh giá “A, B, C”, …

Công việc/ nhiệm vụ của Tiền xử lý dữ liệu:

- Làm sạch (Data Cleaning): điền các giá trị còn thiếu, làm trơn các dữ liệu nhiễu (smooth noisy data), xác định hay loại bỏ các ngoại lệ, giải quyết dữ liệu không nhất quán.
- Tích hợp (Data Integration): tích hợp nhiều cơ sở dữ liệu, khối dữ liệu, tệp tin hoặc ghi chú
- Chuyển đổi (Data Transformation): Chuẩn hoá dữ liệu (scaling to a specific range), Kết hợp dữ liệu (aggregation)
- Giảm thiểu (Data reduction): Obtains reduced representation in volume but produces the same or similar analytical results; Data discretization: with particular importance, especially for numerical data; Data aggregation, dimensionally reduction, data compresstion, generalization

Data cleaning – làm sạch dữ liệu

- Fill in missing values – Bổ sung dữ liệu bị thiếu;
- Identify outliers and smooth out noisy data: nhận diện phần tử biên và giảm thiểu nhiễu;
- Correct inconsistent data – Xử lý dữ liệu không nhất quán;
- Fill in missing values – Xử lý dữ liệu bị thiếu;

Nguyên nhân gây thiếu dữ liệu:

- Sự cố thiết bị.
- Không tương thích với dữ liệu trước đó nen giá trị (mới) bị xoá đi.
- Dữ liệu không được nhập vào (lỗi người nhập liệu).
- Không lưu trữ lịch sử hay sự thay đổi của dữ liệu (thông tin thuyên chuyển của cán bộ trong 1 đơn vị).

Fill in missing values – Xử lý dữ liệu bị thiếu

Cách xử lý dữ liệu bị thiếu

- Bỏ qua các bản ghi có dữ liệu bị thiếu
- Bổ sung dữ liệu bị thiếu bằng tay
- Bổ sung dữ liệu bị thiếu tự động:
    - Giá trị trung bình của thuộc tính, của thuộc tính cùng lớp
    - Giá trị hằng số nhất định
    - Giá trị có thể xảy ra nhất

Identify outliers and smooth out noisy data – Xử lý dữ liệu bị nhiễu

- Dữ liệu nhiễu là dữ liệu (đối tượng) không tuân theo đặc tính/ hành vi chung của tập dữ liệu.
- Giá trị không chính xác do:
    - Lỗi do thiết bị thu thập dữ liệu
    - Vấn đề nhập dữ liệu: người dùng hoặc máy có thể sai
    - Vấn đề truyền dữ liệu: sai từ thiết bị gửi/nhận/truyền
    - Hạn chế của công nghệ: ví dụ, phần mềm có thể xử lý không đúng
    - Thiết nhất quán khi đặt tên: cũng một tên song cách viết khác nhau

Identify outliers and smooth out noisy data – Xử lý dữ liệu bị nhiễu (giải pháp)

- Phân khoảng (Bining):
    - Sắp dữ liệu tăng và chia “đều” vào các thùng (bin).
    - Làm trơn: theo trung bình, theo trung vị,…
    - Hồi quy (Regression): Gắn dữ liệu với một hàm hồi quy (regression function), …
- Phân cụm (Clustering): Phát hiện liệu và loại bỏ các ngoại lai (sau khi đã xác định các cụm).
- Kết hợp giữa máy tính (phát hiện) và kiểm tra của con người (hiệu chỉnh).

Handle noisy data - Nắm bắt dữ liệu nhiễu

- Phương pháp đóng thùng (Bining):
    - Sắp xếp dữ liệu tăng và chia “đều” vào các thùng.

    - Làm trơn: theo trung bình, theo trung tuyến, theo biên…
- Phân cụm (Clustering):
    - Phát hiện và loại bỏ ngoại lai (outliers).
- Kết hợp kiểm tra máy tính và con người:
    - Phát hiện giá trị nghi ngờ để con người kiểm tra (chẳng hạn, đối phó với ngoại lai có thể).
- Hồi quy:
    - Làm trơn: ghép dữ liệu theo các hàm hồi quy.

Correct inconsistent data – Xử lý dữ liệu không nhất quán

- Dữ liệu được ghi nhận khác nhau, ví dụ: 9/3/2018 và 3/9/2018, …
- Nguyên nhân gây ra sự không nhất quán:
    - Sự không nhất quán trong các qui ước đặt tên hay mã dữ liệu.
    - Định dạng không nhất quán của các vùng nhập liệu.
    - Thiết bị ghi nhận dữ liệu.
- Giải pháp:
    - Tạo các ràng buộc khi nhập liệu.
    - Điều chỉnh dữ liệu không nhất quán bằng tay sau khi nhập liệu.
    - Viết các giải thuật điều chỉnh, chuyển đổi tự động.

Data integration – tích hợp dữ liệu

- Tích hợp dữ liệu (Data integration):
    - Kết hợp dữ liệu từ nhiều nguồn vào một kho dữ liệu thống nhất
- Tích hợp ở mức mô hình (Schema integration):
    - Tích hợp metadata từ các nguồn khác nhau
    - Ví dụ: A.cust-id ≡ B.customID
- Vấn đề xác định thực thể (để tránh dư thừa dữ liệu):
    - Cần xác định các thực thể (identities) trên thực tế từ nhiều nguồn dữ liệu
    - Ví dụ: Bill Clinton ≡ B. Clinton
- Phát hiện và xử lý các mâu thuẫn đối với giá trị dữ liệu:

    - Đối với cùng một thực thể trên thực tế, nhưng các giá trị thuộc tính từ nhiều nguồn khác nhau lại khác nhau. Các lý do thực thể:
        - Các cách biểu diễn khác nhau.
        - Mức đánh giá, độ đo (scales) khác nhau. Ví dụ: hệ đo lường mét với hệ đo lường của Anh.
- Dư thừa dữ liệu (redundant data) thường xuyên xảy ra, khi tích hợp dữ liệu từ nhiều nguồn (ví dụ: từ nhiều CSDL):
    - Định danh đối tượng: Cùng một thuộc tính (hay cùng một đối tượng) có thể mang các tên (định danh) khác nhau trong các CSDL khác nhau.
    - Dữ liệu suy ra được: Một thuộc tính trong một bảng có thể là một thuộc tính được suy ra (derived attribute) trong một bảng khác. Ví dụ: “Annual Revenue” và “Monthly Revenue”.
- Các thuộc tính dư thừa có thể được phát hiện bằng phân tích tương quan (Correlation analys): Pearson, Cosine, chi-square, …
- Yêu cầu chung đối với quá trình tích dữ liệu: Giảm thiểu (tránh được là tốt nhất) các dư thừa và các mâu thuẫn
    - Giúp cải thiện tốc độ của quá trình khai phá dữ liệu, và nâng cao chất lượng của các kết quả (tri thức) thu được.

Data transformation – chuyển đổi dữ liệu

- Biến đổi dữ liệu: quá trình biến đổi hay kết hợp dữ liệu vào những dạng thích hợp cho quá trình khai phá dữ liệu
- Làm trơn dữ liệu (smoothing): ước lượng Laplace
- Kết hợp dữ liệu (aggregation).
- Chuẩn hoá (normalization):
    - min-max normalization
        - Giá trị cũ: $v \in [minA, maxA]$
        - Giá trị mới: $v' \in [new_minA, new_maxA]$
            - Ví dụ: chuẩn hoá điểm số từ $0 - 4.0$ sang $0 - 10.0$
            - Đặc điểm của phép chuẩn hoá min-max?

$$v' = \frac{v - minA}{maxA - minA}(new_maxA - new_minA) + new_minA$$

- z-score normalization
  - Giá trị cũ: v tương ứng với mean $\overline{A}$ và standard deviation δ_A
  - Giá trị mới: $v' = \frac{v - \overline{A}}{\delta_A}$
    - Đặc điểm của chuẩn hoá z-score?

Data reduction – Thu giảm dữ liệu

- Kho dữ liệu chứa tới hàng TB:
  - Phân tích/khai phá dữ liệu phức mất thời gian rất dài khi chạy trên tập toàn bộ dữ liệu.
- Rút gọn/ thu giảm dữ liệu:
  - Có được trình bày gọn của tập dữ liệu mà nhỏ hơn nhiều về khối lượng mà sinh ra cùng (hoặc hầu như cùng) kết quả.
- Chiến lược rút gọn dữ liệu:
  - Rút gọn đặc trưng – loại bỏ thuộc tính không quan trọng.
  - Kết hợp khối dữ liệu.
  - Thu giảm chiều: PCA (phân tích thành phần chính).
- Rút gọn đặc trưng (như., lựa chọn tập con thuộc tính):
  - Lựa chọn tập nhỏ nhất các đặc trưng mà phân bố xác suất của các lớp khác nhau cho giá trị khi cho giá trị của các lớp này gần như phân bổ vốn có đã cho giá trị của các đặc trưng.
  - Rút gọn # của các mẫu trong tập mẫu dễ dàng hơn để hiểu dữ liệu.
- Kết hợp khối dữ liệu (data cube aggregation):
  - Dạng dữ liệu: additive, semi-additive (numerical).
  - Kết hợp dữ liệu bằng các hàm nhóm: average, min, max, sum, count, …
    - Dữ liệu ở các mức trừu tượng khác nhau.
    - Mức trừu tượng càng cao giúp thu giảm lượng dữ liệu càng nhiều.

Các tiêu chí lựa chọn giải thuật:

- Mô hình cần dễ hiểu hay không?
- Độ chính xác của mô hình
- Thời gian xây dựng mô hình
- Thời gian dự đoán
- Đặc tính của dữ liệu như kiểu dữ liệu, số lượng phần tử, số lượng chiều, …

Cây quyết định:

- **Kết quả sinh ra dễ diễn dịch** (if … then …).
- Khá đơn giản, nhanh, hiệu quả, được sử dụng nhiều.
- Trong nhiều năm qua, cây quyết định được bình chọn là giải thuật được sử dụng nhiều nhất và thành công nhất.
- **Làm việc đối với kiểu dữ liệu số và liệt kê.**
- Được ứng dụng thành công trong hầu hết các lĩnh vực về phân tích dữ liệu, phân loại văn bản, thư rác, phân loại gen, …
- **Cấu trúc cây:**
    - Nút trong được tích hợp với điều kiện để kiểm tra rẽ nhánh
    - Nút lá được gán nhãn tương ứng với lớp của dữ liệu
    - Một nhánh trình bày cho dữ liệu thỏa mãn điều kiện kiểm tra
    - **Dữ liệu mới đến** được phân loại bằng cách duyệt từ nút gốc của cây cho đến khi đụng đến nút lá, từ đó rút ra lớp của đối tượng cần xét.
- **Một số giải thuật để xây dựng cây quyết định:**
    - ID3 (Quinlan 79)
    - CART – Classification and Regression Tree (Brieman et al. 84)
    - Assistant (Cestnik et al. 87)
    - C4.5 (Quinlan 93)
    - See5 (Quanlan 97)
    - Orange (Demsar Zupan 98-03)
    - …

Máy học vector hỗ trợ - Support vector machines:

- Tìm siêu phẳng trong không gian N-dim để phân loại dữ liệu

- Ứng dụng:
    - Nhận dạng: tiếng nói, ảnh, chữ viết tay, …
    - Phân loại văn bản, khai mỏ dữ liệu văn bản.
    - Phân tích dữ liệu gen, nhận dạng bệnh, công nghệ bào chế thuốc.
    - Phân tích dữ liệu marketing
    - Phân tích cảm xúc qua khuôn mặt
    - …
- Ưu điểm:
    - Cho kết quả rất tốt trong thực tế, mô hình có độ chính xác cao.
    - Chịu đựng được nhiễu.
    - Hiệu quả khi xử lý dữ liệu có số lượng thuộc tính lớn.
    - Thành công trong nhiều ứng dụng.
- Khuyết điểm:
    - **Khó dịch kết quả.**
    - Quá trình học mô hình SVM tốn nhiều thời gian do độ phức tạp cao.
    - **Chỉ làm việc với dữ liệu số**
    - Tham số SVM và hàm nhân khó điều chỉnh.

Phương pháp k láng giềng – KNN

- Rất đơn giản, **không có quá trình học.**
- Khi phân loại mất nhiều thời gian, do quá trình tìm kiếm k dữ liệu lân cận. Sau đó phân loại dựa trên **majority vote** (hồi quy dựa trên giá trị trung bình)
- Kết quả phụ thuộc vào việc chọn khoảng cách sử dụng:
    - Khoảng cách Minkowski (căn bậc q), Manhattan (căn bậc 1), Euclid (căn bậc 2)
- **Có thể làm việc trên nhiều kiểu dữ liệu khác nhau** (Chú ý việc chuẩn hóa dữ liệu, giả sử các thuộc tính có độ quan trọng như nhau? Gán trọng số quan trọng cho mỗi thuộc tính?).
- Được ứng dụng thành công trong hầu hết các lĩnh vực tìm kiếm thông tin, nhận dạng, phân tích dữ liệu, …

Cải thiện độ chính xác với các phương pháp tập hợp mô hình:

- Xây dựng tập hợp các mô hình cơ sở dựa trên tập học.
- Kết hợp các mô hình khi phân loại cho độ chính xác cao.
- Dựa trên cơ sở:
    - Bias: thành phần lỗi độc lập với mẫu dữ liệu học.
    - Variance: thành phần lỗi do biến động liên quan đến sự ngẫu nhiên của tập học.
    - $\boldsymbol{Errors = bias^2 + varience}$
- Các giải thuật phổ biến:
    - **Bagging, Random forest – rừng ngẫu nhiên:** (averaging) xây dựng tập hợp các mô hình cơ sở **độc lập** nhau, kết hợp sự phân loại của các mô hình, giảm variance.
    - Boosting: xây dựng tập hợp các mô hình cơ sở **tuần tự** (tập trung lên các lỗi sinh ra từ các mô hình trước), AdaBoost và arcing, giảm bias.
- Áp dụng cho nhiều giải thuật cơ sở khác nhau như Cây quyết định, SVM, naïve Bayes, …
- Giải quyết các vấn đề về phân loại, hồi quy, gom nhóm, …
- **Cho kết quả tốt, tuy nhiên *không thể dịch/rất có có thể dịch* được kết quả sinh ra.** Ví dụ: một rừng bao gồm hàng trăm cây quyết định, …

Bootstrap AGGregatING – BAGGING và Random forests:

- Từ tập học LS có N phần tử.
- Xây dựng tập hợp **T mô hình cơ sở độc lập nhau**.
- Mô hình thứ i được xây dựng trên tập mẫu bootstrap.
    - Tại nút trong chọn ngẫu nhiên n' thuộc tính (n' <= n) và tính toán phân hoạch tốt nhất dựa trên n' thuộc tính này.
    - Cây được xây dựng đến độ sâu tối ta không cắt nhánh.
- Một bootstrap: lấy mẫu N phần tử có hoàn loại từ tập LS
- Khi phân loại: sử dụng **majority vote.**
- Hồi quy: tính giá trị trung bình của dự đoán của các mô hình.

Boosting:

- Từ tập học LS có N phần tử.
- Xây dựng tập hợp **T mô hình cơ sở tuần tự.**
- Mô hình thứ i được xây dựng trên tập mẫu lấy từ LS, tập trung vào các phần tử bị phân loại sai bởi mô hình thứ i-1 trước đó
- Khi phân loại: sử dụng **majority vote có trọng số.**
- Hồi quy: tính giá trị trung bình của dự đoán các mô hình **có sử dụng trọng số.**

Gom nhóm – Clustering: mô hình gom cụm dữ liệu (không có nhãn) sao cho các dữ liệu cùng nhóm có các tính chất tương tự nhau và dữ liệu của 2 nhóm khác nhau sẽ có các tính chất khác nhau:

  - Phương pháp học không giám sát.
  - Dữ liệu thường không có nhiều thông tin sẵn có như lớp (nhãn).
- Hierarchical clustering: Xây dựng một cây phân cấp dựa trên sự phân loại theo cấp bậc từ một tập hợp các dữ liệu. Dựa trên điểm cắt ở đâu mà ta thu được các cụm tương ứng
- K-Means: giải thuật đơn giản, cho kết quả dễ hiểu, cần cho tham số K cluster, kết quả phụ thuộc vào việc khởi động K tâm center của K cluster: có thể khắc phục bằng cách khởi động lại nhiều lần; khả nặng chịu nhiễu không tốt (bởi các phần tử ngoại lai): có thể khắc phục bằng K-Medoids, không sử dụng giá trị trung bình nhưng sử dụng phần tử ngay giữa.
- Mean Shift
- DBSCAN (Density Based Clustering)
- Agglomerative.

Ví dụ**:**

- Có thể giúp các nhà tiếp thị khám phá các nhóm khách hàng riêng biệt. Và họ có thể đặc trưng nhóm khách hàng của họ dựa trên các lịch sử mua hàng.
- Trong lĩnh vực sinh học, clustering được sử dụng để phân loại thực vật và động vật, phân loại gen có chức năng tương tự
- Clustering cũng giúp trong việc phân loại tài liệu trên web để phát hiện thông tin.

Gom nhóm:

- thường dựa trên cơ sở khoảng cách.
- nên chuẩn hóa dữ liệu.
- khoảng cách được tính theo từng kiểu của dữ liệu:
    - Kiểu số.
    - Kiểu nhị phân.
    - Kiểu rời rạc (nominal type).

Hồi quy – Regression:

- Một phương học tập có giám sát, dùng để dự đoán nhãn có giá trị liên tục.
- Trong khi sử dụng hồi quy tuyến tính, mục tiêu của chúng ta là để làm sao một đường thẳng có thể tạo được sự phân bố gần nhất với hầu hết các điểm. Do đó làm giảm khoảng cách (sai số) của các điểm dữ liệu cho đến đường đó.
- Hạn chế: Nhạy cảm với nhiễu (sensitive to noise).

Phân loại giải thuật học máy:

Supervised learning

- SVM
- Decision Trees
- Naive Bayes
- Linear Regression
- Ensemble methods

Unsupervised learning - Clustering

- K-means
- Affinity Propagation
- Mean Shift
- Spectral clustering
- Hierarchical clustering
- DBSCAN
- OPTICS

- BIRCH
- Clustering performance evaluation

Đánh giá hiệu quả của giải thuật học:

- Accuracy là tỉ lệ giữa số điểm được phân loại đúng và tổng số điểm. Accuracy chỉ phù hợp với các bài toán mà kích thước các lớp dữ liệu là tương đối như nhau. Hay trong tất cả các dự đoán của chúng ta, tỷ lệ dự đoán đúng là bao nhiêu?
    - $\text{accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$
- Confusion matrix giúp có cái nhìn rõ hơn về việc các điểm dữ liệu được phân loại đúng/sai như thế nào.
    - True Positive (TP): số lượng điểm của lớp positive được phân loại đúng là positive.
    - True Negative (TN): số lượng điểm của lớp negative được phân loại đúng là negative.
    - False Positive (FP): số lượng điểm của lớp negative bị phân loại nhầm thành positive.
    - False Negative (FN): số lượng điểm của lớp positiv bị phân loại nhầm thành negative
    - Với một cách xác định một lớp là positive, Precision được định nghĩa là tỉ lệ số điểm true positive trong số những điểm được phân loại là positive (TP + FP). Nói cách khác, có bao nhiêu dự đoán “positive” là thật sự “true” trong thực tế?
    - Recall được định nghĩa là tỉ lệ số điểm true positive trong số những điểm thực sự là positive (TP + FN). Nói cách khác, có bao nhiêu dự đoán “positive” đúng là do mô hình của chúng ta đưa ra?
    - Precision cao đồng nghĩa với việc độ chính xác của các điểm tìm được là cao. Recall cao đồng nghĩa với việc True Positive Rate cao, tức tỉ lệ bỏ sót các điểm thực sự positive là thấp.
    - $Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN}$

- $F1 - score = 2\frac{Precision\ x\ Recall}{Precision + Recall}$
- Trung bình điều hòa giữa Precision và Recall. Đây là chỉ số thay thế lý tưởng cho accuracy khi mô hình có tỷ lệ mất cân bằng mẫu cao.

- **Mean Absolute Error – MAE**: là một phương pháp đo lường sự khác biệt giữa hai biến liên tục. Giả sử X và Y là hai biến liên tục thể hiện kết quả dự đoán mô hình và kết quả thực tế. Chúng ta độ đo MAE:
  - $\mathrm{MAE} = \frac{\sum_{i=1}^{n} |\mathrm{Yi} - \mathrm{Xi}|}{n}$
- **Mean Squared Error – MSE**: là trung bình của bình phương sai số, tức là sự khác biệt giữa các giá trị được mô hình dự đoán và giá trị thực. MSE là một hàm rủi ro, tương ứng với giá trị kỳ vọng của sự mất mát sai số bình phương hoặc mất mát bậc hai.
  - $\mathrm{MSE} = \frac{\sum_{i=1}^{n} (\mathrm{Yi} - \mathrm{Xi})\text{^}2}{n}$

Hãy cho biết các mô hình biểu diễn văn bản: Vector Space Model, Bag Of Word (BOW), Graph-Based Model, TFIDF, One-hot-vector, Ma trận đồng xuất hiện.

Các chữ V nào nói lên tính chất quan trọng của Big Data? Volume (dung lượng), Velocity (tốc độ), Variety (tính đa dạng), Veracity (sự chính xác của dữ liệu), Value (Giá trị).

Overfitting (quá khớp) trong machine learning, tức là dữ liệu quá khớp với dữ liệu hay mô hình "học vẹt" tức là chỉ nhớ dữ liệu chứ không có khả năng tổng quát hóa để dự đoán các dữ liệu chưa được quan sát trên thực tế sử dụng. Nó có thể rất đúng trên dữ liệu quan sát có sẵn (100%) nhưng lại sai rất lớn trên các dữ liệu mới được đưa vào mô hình mà không nằm trong tập dữ liệu quan sát đã có.

Underfitting (Chưa khớp). Nếu ví mô hình overfitting là một anh sinh viên học tủ, học vẹt thì underfitting là một anh sinh viên chẳng biết gì, bởi anh ta thậm chí còn chẳng làm tốt trên cả bài tủ (dữ liệu quan sát đã cho) và cũng do chẳng thể làm cả bài cơ bản, nên anh ta cũng chẳng thể vận dụng và làm các bài nâng cao (dữ liệu mới chưa được quan sát).

Overfitting và **Underfitting** xảy ra là do trên thực tế chúng ta không thể biết được dạng của hàm dự đoán là hàm gì mà chỉ có thể biết được một số điểm hữu hạn dữ liệu quan sát (dữ liệu này còn chứa nhiễu do đo đạc, thống kê hoặc sự ngẫu nhiên mặc định trong dữ liệu).

Hiển thị dữ liệu:

- Quan trọng trong quá trình khai mỏ dữ liệu:
    - Trong tiền xử lý dữ liệu: giúp xem sơ lược về dữ liệu, phát hiện một vài tính chất tổng quát, chọn giải thuật để khai mỏ dữ liệu và có một vài ý tưởng cho việc lựa chọn tham số.
    - Trong khai mỏ dữ liệu: có thể thay thế hoặc phối hợp với các phương pháp học tự động (visual data mining), người sử dụng làm trung tâm, sử dụng được khả năng nhận dạng mẫu của con người, có thể sử dụng được ý kiến của chuyên gia khi xây dựng mô hình, giúp người dùng dễ hiểu mô hình xây dựng bởi vì chính họ trực tiếp tham gia xây dựng mô hình.
    - Trong hậu xử lý: giúp giải thích các kết quả sinh ra trong quá trình học tự động.
- Biểu đồ tròn (Pie Chart): so sánh các đối tượng cùng một tiêu chí theo đơn vị phần trăm.
- Biểu đồ cột (Bar Chart): giúp so sánh giá trị các đối tượng theo cùng một tiêu chí. Đồng thời biểu đồ cột còn thể hiện sự tăng giảm qua các năm của các đối tượng.
- Biểu đồ đường (Line Chart): Diễn đạt sự tăng trưởng, thay đổi dịch chuyển theo thời gian của một yếu tố.
- Box Plots: J.Turkey, hiển thị sự phân bố dữ liệu, biểu đồ đơn giản nhưng thể hiện được nhiều thông tin hữu ích; có thể được sử dụng để so sánh các thuộc tính.
- Histograms: thường hiển thị sự phân phối các giá trị của một biến đơn lẻ; chia các giá trị vào các bin và hiển thị một biểu đồ thanh về số lượng các đối tượng trong mỗi bin; Chiều cao của mỗi thanh cho biết số lượng đối tượng.
- Histograms hai chiều: Hiển thị sự phân phối chung của các giá trị của hai thuộc tính.
- Scatter Plots: cho phép hiển thị quan hệ hai chiều (giữa hai thuộc tính) của dữ liệu; Cho phép quan sát (trực quan) các nhóm điểm, các ngoại lai, …; Mỗi cặp giá trị của hai thuộc tính được xét tương ứng với hai tọa độ của điểm được hiển thị trên mặt phẳng.