Journal of educational equipment: Applied research, Volume 2, Issue 287 (April 2023)
ISSN 1859 - 0810

# Tìm hiểu một số kỹ thuật khai phá dữ liệu phục vụ dạy học

***Trần Việt****

**ThS. Trường Đại học Hải Phòng*

*Received: 9/2/2023; Accepted: 13/2/2023; Published: 15/2/2023*

***Abstract:*** *In addition to the term knowledge discovery, people also use a number of other terms with similar meanings such as data/pattern analysis, data mining, etc. But in short, in essence, knowledge discovery Knowledge involves analyzing data and using special techniques to find feature patterns in a huge data set. There are many definitions of knowledge discovery that have been given by different authors, according to Fayyad's definition: "KDD is the non-trivial process of identifying valid, novel, and potentially valid latent patterns. useful and understandable in the data".*

*The knowledge mining process includes stages: data preparation, pattern search, data mining, pattern evaluation, and use of the discovered knowledge. Each technique has its own advantages and disadvantages and is applied in different data mining purposes. Decision trees are used in classification and prediction problems. K-means algorithm, EM algorithm is used in data clustering problems.*

***Keywords:*** *Data mining; KDD; Decision trees algorithm; K-means algorithm; EM algorithm;*

## 1. Đặt vấn đề

Những năm gần đây, lượng thông tin được lưu trữ trên các thiết bị (như đĩa cứng, CDROM, bằng từ...) không ngừng tăng lên với một tốc độ bùng nổ. Người ta ước đoán rằng lượng thông tin toàn cầu tăng gấp đôi sau khoảng 2 năm và cùng với nó số lượng cũng như kích thước của các cơ sở dữ liệu cũng tăng lên nhanh chóng.

Trong nhiều lĩnh vực, nhà quản lý đang ngập trong dữ liệu nhưng lại cảm thấy đói tri thức và thông tin hữu ích. Lượng dữ liệu khổng lồ này đang thực sự là một nguồn tài nguyên rất giá trị bởi thông tin là yếu tố then chốt trong các hoạt động đặc biệt là thương mại vì nó giúp người điều hành và quản lý có một cái nhìn sâu sắc, chính xác, khách quan trước khi đưa ra các quyết định. Khai phá dữ liệu là khai thác những thông tin tiềm ẩn mang tính dự đoán từ những cơ sở dữ liệu lớn là hướng tiếp cận có nhiều ý nghĩa và mang tính lịch sử. Các kỹ thuật phát hiện tri thức và khai phá dữ liệu được thực hiện qua nhiều giai đoạn và sử dụng nhiều kỹ thuật: phân lớp (classification),phân cụm (clustering), phân tích sự tương tự (similarity analysis), tổng hợp (summarization), luật kết hợp (association rules),...

Trong khuôn khổ của bài báo này, tác giả trình bày về ba thuật toán trong khai phá dữ liệu là *C*ây quyết định, K-mean, EM.

## 2. Nội dung nghiên cứu

### *2.1. Cây quyết định*

Cây quyết định (CQĐ) là cấu trúc biểu diễn dưới dạng cây. Cấu trúc của một CQĐ bao gồm các nút và các nhánh.

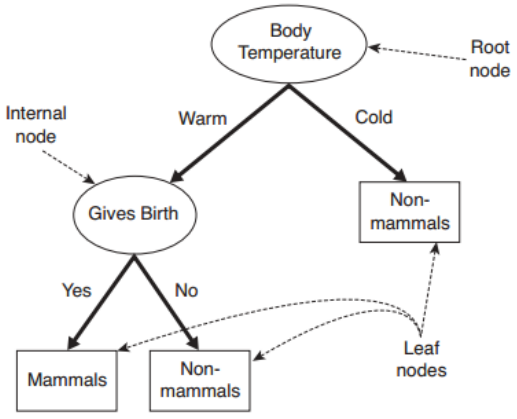
*Nhánh* (branch): biểu diễn giá trị có thể có của thuộc tính.

Nút (node): mỗi nút mạng một thuộc tính bao gồm 3 loại

Nút gốc (root node) là đỉnh trên cùng của cây.

Nút lá (leaf node) là nút ngoài cùng, mang thuộc tính phân lớp.

Nút trong (internal node) là các nút còn lại, mang thuộc tính phân loại.




*Hình 2.1. Mô hình cây quyết định [4]*

Ưu điểm của CQĐ: CQĐ là tự giải thích vì hai lý do thứ nhất nếu các CQĐ có số lượng nút lá vừa phải, nó có thể được nắm bắt bởi người dùng không chuyên nghiệp. Thứ hai, cây quyết định có thể được chuyển đổi sang tập luật. CQĐ có thể xử lý các dữ liệu số, loại. CQĐ là có khả năng xử lý các tập dữ liệu đó có thể có lỗi, thiếu giá trị.

Nhược điểm của CQĐ: Hầu hết các thuật toán như ID3 và C4.5 yêu cầu các thuộc tính mục tiêu phải là giá trị rời rạc. Cây quyết định nếu sử dụng PP chia để trị sẽ thực hiện tốt khi có một số thuộc tính liên quan chặt chẽ với nhau nhưng sẽ khó khăn nếu một số tương tác phức tạp xuất hiện. Các đặc tính liên quan của CQĐ dẫn đến những khó khăn khác như độ nhạy với tập huấn luyện, các thuộc tính không phù hợp, nhiễu.

CQĐ được xây dựng từ trên xuống. Bắt đầu xây dựng CQĐ tại nút gốc, tất cả các dữ liệu học ở nút gốc. Tiến hành chọn thuộc tính phân hoạch tốt nhất, dữ liệu được chia theo các giá trị của thuộc tính phân hoạch. Quá trình này được lặp lại với tập dữ liệu ở mỗi nút vừa tạo. Điều kiện để dừng phân chia là: Tất cả các mẫu cùng một nút thuộc về cùng một lớp; Không còn thuộc tính nào để thực hiện phân chia tập dữ liệu nữa; Số lượng phần tử của dữ liệu tại nút bằng không.

Theo nguyên tắc xây dựng CQĐ như trên thì với cùng một tập dữ liệu học có thể cho ra các cây có độ rộng, độ sâu, độ phức tạp khác nhau nếu thứ tự chọn thuộc tính triển khai cây khác nhau. Do đó, việc chọn thuộc tính nào để phân hoạch ở mỗi nút mang tính quyết định đến độ phức tạp của CQĐ được tạo. Để đánh giá thuộc tính phân hoạch tốt nhất ta dựa trên độ lợi thông tin (information gain), độ đo information gain ratio, chỉ số gini.

Độ lợi thông tin của thuộc tính A được tính theo công thức:

$$Entropy(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} Info(p_i, n_i)$$

$$Gain(A) = Info(p,n) - Entropy(A)$$

Trong đó:

$$Info(p,n) = Entropy(\frac{p}{p+n}, \frac{n}{p+n}) =$$

$$= -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

P, N là hai lớp.

S là tập dữ liệu có $p$ phần tử lớp $P$ và $n$ phần tử lớp $N$.

Tập {S1, S2…Sv} là một phân hoạch trên tập S, khi sử dụng thuộc tính A. Mỗi Si chứa $p_i$ mẫu lớp P và $n_i$ mẫu lớp N.

Xét ví dụ với tập dữ liệu như bảng 2.1 bên dưới ta có:

| Học lực | Điểm rèn luyện | Khu vực | Điểm tuyển sinh | Thôi học |
|---|---|---|---|---|
| Yếu | Thấp | 1 | TB | Có |
| Trung bình | Thấp | 2 | TB | Có |
| Trung bình | Thấp | 3 | Cao | Có |
| Yếu | Cao | 3 | Cao | Có |
| Khá | Thấp | 3 | Cao | Có |
| Trung bình | Thấp | 2 | Cao | Có |
| Khá | Cao | 2 | Cao | Có |
| Yếu | Cao | 2 | TB | Có |
| Yếu | Thấp | 1 | Cao | Có |
| Khá | Thấp | 1 | TB | Không |
| Khá | Cao | 1 | TB | Không |
| Trung bình | Cao | 3 | Cao | Không |
| Khá | Thấp | 2 | TB | Không |
| Trung bình | Cao | 2 | TB | Không |

Lớp ***P***: Thôi học = "Có"

Lớp ***N***: Thôi học = "Không"

Thông tin cần thiết để phân lớp một mẫu được cho là:

$$Info(p,n) = Info(9,5) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$$= 0.940$$

Áp dụng công thức 2.1 và 2.2 ở trên ta có

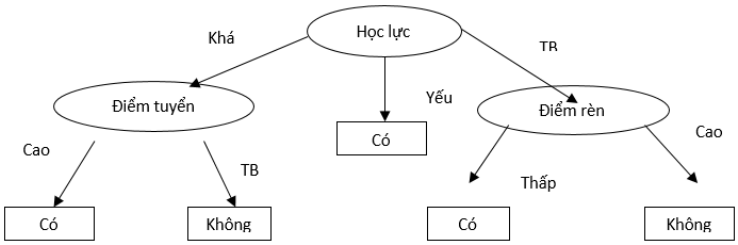
Gain ("Học lực")=0.246

Gain ("Điểm tuyển sinh")=0.151

Gain ("Điểm rèn luyện")=0.048

Gain ("Khu vực")=0.029

Kết quả trên cho thấy thuộc tính "Học lực" có độ lợi thông tin lớn nhất nên được chọn làm thuộc tính để phân tách. Tiếp tục tương tự ta sẽ được cây quyết định cuối cùng có dạng:




*Hình 2.2 Kết quả cây quyết định với tập dữ liệu học trong bảng 2.1*

### *2.2. Thuật toán K-means*

Thuật toán k-Means được xếp vào lớp thuật toán phân cụm phẳng (phân cụm phân vùng), ý tưởng chính của thuật toán là lặp lại nhiều lần quá trình bố trí lại vị trí của đối tượng dữ liệu để phân hoạch một tập dữ liệu D thành các cụm và cực tiểu địa phương giá trị bình phương trung bình khoảng cách giữa các các đối tượng tới tâm cụm tương ứng. Việc quyết định phân một đối tượng dữ liệu vào một cụm dựa

vào độ tương đồng của đối tượng đó với trọng tậm của các cụm.

Giải thuật K-Means

Đầu vào:

Tập các đối tượng dữ liệu S = {di}.

Số nguyên k > 0 các cụm cho trước.

Đầu ra:

Các cụm Si phân hoạch tách rời sao cho hàm tiêu chuẩn E đạt giá trị tối thiểu.

*Thuật toán:*

1. Khởi tạo: Chọn ngẫu nhiên k dữ liệu trong S làm trọng tâm đại diện cho các cụm Si= {ci: ci ∈ S}, ∀ i = 1, ...,k

2. While (điều kiện dừng chưa thỏa mãn) do

2.1 Si = ∅ //Các cụm mới là rỗng

2.2 ∀ d ∈ S

2.2.1 Tính sim (d, ci), ∀ i = 1, ..., k

2.2.2 St = St ∪{d } nếu sim (d, ct ) = max {sim (d, ci), ∀ i = 1, ..., k }

2.3. For i = 1..k do Tính lại trọng tâm các cụm Si

Ưu điểm của thuật toán K-Mean:

Đơn giản, dễ sử dụng, dễ cài đặt và được dùng phổ biến nhất. Thường cho tối ưu cục bộ. Tối ưu toàn cục rất khó tìm.

Hiệu quả về thời gian: tuyến tính O (tkn) với t số lần lặp, k số cụm, n là số phần tử. Nếu cả 2 giá trị k và t đều nhỏ, thì giải thuật k-Means được xem như là có độ phức tạp ở mức tuyến tính.

Nhược điểm của thuật toán K-Mean:

Giá trị k phải được xác định trước. Cần xác định cách tính điểm trung bình (centroid) của một nhóm. Đối với các thuộc tính định danh (nominal attributes), giá trị trung bình có thể được xác định là giá trị phổ biến nhất.

### *2.3. Thuật toán EM (Expectation Maximization)*

Thuật toán EM được xem như là mở rộng của thuật toán K-means, thuật toán này nhằm tìm ra sự ước lượng về khả năng lớn nhất của các tham số trong mô hình xác suất. Thuật toán EM gán các đối tượng cho các cụm đã cho theo xác suất phân phối thành phần của đối tượng đó. Phân phối xác suất được sử dụng là phân phối xác suất Gaussian nhằm khám phá lặp các giá trị tốt cho các tham số của nó bằng hàm tiêu chuẩn là hàm logarit khả năng của đối tượng dữ liệu, đây là hàm tốt để mô hình xác suất cho các đối tượng dữ liệu.

Thuật toán EM có thể khám phá ra nhiều hình dạng cụm khác nhau, nhưng do thời gian lặp của thuật toán nhiều nhằm xác định các tham số tốt nên chi phí tính toán của thuật toán cao.

Thuật toán phân cụm EM được mô tả như sau:

Cho X={$x_1$, $x_2$…$x_n$} là tập hợp các điểm dữ liệu.

V= { $ì_1, ì_2, ì_3 ... ì_c$ } là tập các giá trị trung bình.

P= {$p_1$, $p_2$, $p_3$…$p_c$} là tập hợp các xác xuất.

Bước 1: Khởi tạo lần lặp thứ i

$$ë(t) = \{ì_1(t), ì_2(t) ... ì_c(t), \sum_1(t), \sum_2(t) ... \sum_1(t), p_1(t), p_2(t) ... p_c(t)\}$$

Bước 2 (bước E): Tính toán các lớp của tất cả các điểm dữ liệu sử dụng cho mỗi lớp:

$$P(w_t | x_k, ë_t) = \frac{P(x_k | w_t, ë_t) P(w_t | ë_t)}{P(x_k | ë_t)}$$

$$= \frac{P(x_k | w_i, ì_i(t), \sum_i(t)) P_i(t)}{\sum_{j=1}^{c} P(x_k | w_j, ì_j(t), \sum_j(t)) P_j(t)} \quad (2.4)$$

Bước 3 (bước M): Tính $ì$ lớn nhất cho lớp dữ liệu bằng công thức:

$$ì_i(t+1) = \frac{\sum_k P(w_t | x_k, ë_t) x_k}{\sum_k P(w_t | x_k, ë_t)} \quad (2.5)$$

$$P_i(t+1) = \frac{\sum_k P(w_t | x_k, ë_t)}{R} \quad (2.6)$$

Trong đó R là số điểm dữ liệu

## 3. Kết luận

Các kỹ thuật: cây quyết định, thuật toán K-means, thuật toán EM; mỗi kỹ thuật đều có ưu nhược điểm riêng và được áp dụng trong những mục đích khai phá dữ liệu khác nhau. CQĐ được dùng trong các bài toán phân lớp, dự báo. CQĐ được dùng rộng rãi vì những ưu điểm của nó như cây quyết định cho kết quả trực quan, dễ hiểu. CQĐ xử lý được dữ liệu kiểu số và rời rạc, dữ liệu thiếu. Thuật toán K-means, thuật toán EM được dùng trong các bài toán phân cụm dữ liệu. Thuật toán K-means được sử dụng rộng rãi vì giải thuật đơn giản, cho ra kết quả dễ hiểu.

**Tài liệu tham khảo**

[1] U.Fayyad, G (1996). *Piatetsky-Shapiro, P.Smyth From Data Mining to Knowledge Discovery in Databases, AI Magazine.*

[2] Bing Liu (2007). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Spinger.*

[3] U. M. Fayyad, G. P (1996). *Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park, CA.*

[4] D. Hand, H (2001). *Mannila, and P. Smyth. Principles of Data Mining. The MIT Press, London, England.*