BIỂU DIỄN DỮ LIỆU KHAI PHÁ DỮ LIỆU CHUỖI THỜI GIAN: PHƯƠNG PHÁP TIẾP CẬN MIỀN THỜI GIAN

Seunghye J. Wilson, Phòng Thống kê, Đại học George Mason, Mỹ

Tóm tắt:

Trong hầu hết khai phá dữ liệu chuỗi thời gian, cần yêu cầu nhiều hình thức khác nhau cho việc biểu diễn dữ liệu hoặc xử lý dữ liệu vì những đặc tính độc đáo của chuỗi thời gian, ví dụ như nhiều chiều (số lượng điểm dữ liệu), sự xuất hiện của nhiễu ngẫu nhiên và mối quan hệ phi tuyến tính của các phần tử dữ liệu. Do đó, bất kỳ phương pháp biểu diễn dữ liệu nào cũng đều nhằm mục đích giảm đáng kể dữ liệu đến một kích thước có thể quản lý, đồng thời vẫn giữ được các đặc tính quan trọng của dữ liệu ban đầu và sức mạnh với nhiễu ngẫu nhiên. Hơn nữa, việc lựa chọn phương pháp biểu diễn dữ liệu phù hợp có thể dẫn đến khai phá dữ liệu có ý nghĩa. Nhiều phương pháp biểu diễn cấp cao của dữ liệu theo chuỗi thời gian được dựa trên phương pháp tiếp cận miền thời gian. Các phương pháp này xử lý trực tiếp dữ liệu ban đầu trong miền thời gian và hiểu được bản chất của dữ liệu theo thời gian. Phương pháp này dựa trên một số ý tưởng chính của phương pháp xấp xỉ từng đoạn, biểu diễn dữ liệu bằng cách xác định các điểm quan trọng, và biểu diễn ký hiệu hóa đã được sử dụng rộng rãi trong các lĩnh vực khác nhau.

Từ khoá: Khai phá dữ liệu chuỗi thời gian, xử lý dữ liệu, giảm dữ liệu, biểu diễn dữ liệu cấp cao, phương pháp tiếp cận miền thời gian.

1. Giới thiệu

Chuỗi thời gian là một dạng dữ liệu quan trọng trong các lĩnh vực khác nhau của ngành công nghiệp và nghiên cứu. Trong những thập kỷ gần đây, việc khai phá dữ liệu theo chuỗi thời gian đã được quan tâm và phát triển bùng nổ. Tuy nhiên, thật khó để áp dụng kỹ thuật khai phá để lấy dữ liệu trực tiếp vì những đặc tính độc đáo của chuỗi thời gian như: Khối lượng dữ liệu lớn, sự có mặt của nhiễu ngẫu nhiên, và các mối quan hệ phi tuyến tính của các phần tử dữ liệu. Kết quả là, việc biểu diễn dữ liệu chỉ ở dạng đơn giản hóa, hoặc xử lý dữ liệu là một bước thiết yếu trong việc khai phá dữ liệu theo chuỗi thời gian. Mục đích chính của việc biểu diễn dữ liệu là giảm dữ liệu đến một kích thước có thể quản lý hoặc xấp xỉ dữ liệu bằng cách loại bỏ nhiễu ngẫu nhiên. Tuy nhiên, dữ liệu bị giảm đi phải bảo toàn các tính năng quan trọng của toàn bộ dữ liệu ban đầu.

Phương pháp tiếp cận miền thời gian để biểu diễn dữ liệu đặc biệt hữu ích để hiểu được bản chất của dữ liệu theo thời gian. Chúng tóm tắt dữ liệu ban đầu bằng cách ước lượng các khoảng giá trị, xác định các điểm tới hạn, hoặc chuyển đổi dữ liệu số thành các biến rời rạc. Phương pháp xấp xỉ từng đoạn là một trong những phương pháp tiếp cận miền thời gian phổ biến nhất. Các phương pháp này biểu diễn dữ liệu ban đầu dựa trên các khoảng thời gian không chồng chéo. Kết quả trình bày dữ liệu theo phương pháp xấp xỉ từng đoạn có thể là một dãy các

đoạn thẳng liên tục hay rời rạc, hoặc các giá trị biểu diễn của tất cả các khoảng với chiều dài giảm đáng kể. Phương pháp tiếp cận phổ biến khác để biểu diễn dữ liệu là xác định các điểm quan trọng để bảo vệ các điểm tới hạn góp phần tiết lộ các tính năng quan trọng, chẳng hạn như hình dạng tổng thể hoặc xu hướng thay đổi các điểm dữ liệu ban đầu. Gần đây, khi sự quan tâm đến việc khai phá dữ liệu có khối lượng lớn, gọi là "dữ liệu lớn" tiếp tục tăng lên, các phương pháp biểu diễn dữ liệu bằng cách biến đổi chuỗi thời gian số sang các biến hoặc ký hiệu rời rạc sẽ trở nên phổ biến hơn. Phương pháp biểu diễn ký hiệu hóa là chuyển đổi ký hiệu cho phép không chỉ giảm dữ liệu mà còn tính toán hiệu quả và sử dụng không gian bộ nhớ để lưu trữ dữ liệu vì yêu cầu ít dung lượng hơn cho dữ liệu chuỗi so với dữ liệu số. Trong bài viết này, chúng ta sẽ xem xét ba phương pháp phổ biến để biểu diễn dữ liệu trong miền thời gian và thảo luận về các thuộc tính của chúng.

2. Phương pháp tiếp cận chung cho xấp xỉ dữ liệu

Các mô hình tổng thể và xấp xỉ từng đoạn. Trong phân tích dữ liệu, các mô hình tổng thể thường được sử dụng để xác định các biểu diễn dữ liệu đơn giản hơn khi mô hình cơ bản quá phức tạp hoặc để ước tính một chức năng không xác định cho dữ liệu được quan sát. Các mô hình tổng thể này rất hữu ích để hiểu các quy trình tạo dữ liệu. Ví dụ, các mô hình hồi quy tuyến tính giữa các biến giải thích (độc lập) và biến kết quả (phụ thuộc) dựa trên một số giả định sao cho phương sai của phần sai số là hằng số độc lập. Hồi quy đa thức là mô hình mở rộng của mô hình hồi quy tuyến tính cho phép các biến giải thích đa thức bậc n - trong mô hình tuyến tính. Mô hình tự hồi quy và trung bình trượt (ARMA), đặc biệt với dữ liệu chuỗi thời gian, mô tả quá trình ngẫu nhiên dưới dạng các đa thức tự hồi quy và chuyển động trung bình. Các mô hình này thường phụ thuộc vào các giả định cụ thể và đủ số lượng các điểm dữ liệu, nhưng trở nên không chính xác khi kích thước dữ liệu tăng lên sẽ không đúng với các điều kiện giả định trong thực tế.

Khi kích thước tăng lên, phương pháp xấp xỉ từng đoạn, chẳng hạn như với đa thức từng đoạn và hàm spline, thường có hiệu quả hơn. Thật vậy, nhiều phương pháp biểu diễn chuỗi thời gian dựa trên phương pháp xấp xỉ từng đoạn do dữ liệu chuỗi thời gian thường được đặc trưng bởi kích thước lớn và sự hiện diện của nhiễu ngẫu nhiên. Theo phương pháp xấp xỉ từng đoạn, tất cả các điểm dữ liệu được chia thành một số phân đoạn không chồng chéo để xây dựng một mô hình cục bộ $\mu_i(t)$ ($b_i - 1 \leq t < b_i$, $b_0 = t_1$) trong từng phân đoạn và dữ liệu ban đầu được biểu diễn bởi một chuỗi các mô hình cục bộ $\{\mu_1(t), \ldots, \mu_i(t), \ldots \mu_n(t)\}$. Do đó, với chuỗi thời gian $X=x_1, \ldots, x_N$ mô hình được viết bằng:

$$\hat{X}(t) = \sum_{i=0}^{n} \mu_i(t), n \ll N \quad (1)$$

Xử lý hàng loạt và trực tuyến

Dữ liệu kích thước lớn có thể được ước lượng hoặc biểu diễn bởi xử lý hàng loạt hoặc xử lý trực tuyến dựa trên tính sẵn có của dữ liệu khi phân tích. Xử lý hàng loạt được sử dụng khi tất cả các điểm dữ liệu có sẵn trong quá trình tính toán, và một khi quá trình xử lý dữ liệu bắt đầu, việc thu thập các điểm dữ liệu mới không thể xảy ra. Do đó, cần phải hiểu cấu trúc dữ liệu trước khi phân tích. Mặt khác, xử lý trực tuyến phân tích dữ liệu là khi tiếp nhận các điểm dữ liệu liên tục và thu thập các điểm dữ liệu mới trong cùng quá trình tính toán. Vì vậy, các kết quả xử lý dữ liệu thu được ngay lập tức trong một thời gian ngắn và yêu cầu lưu trữ dữ liệu ít hơn.

Vì lý do này, xử lý trực tuyến thường được dùng trong việc khai phá luồng dữ liệu lớn.

3. Biểu diễn dữ liệu chuỗi thời gian

Xấp xỉ từng đoạn

Một cách tiếp cận đơn giản và phổ biến để biểu diễn dữ liệu là xấp xỉ từng đoạn. Nhìn chung, các thuật toán xấp xỉ chia toàn bộ tập dữ liệu vào một số khoảng không chồng chéo theo thời gian và đặt các mô hình cục bộ vào các khoảng đó. Theo công thức, $X = \{x_t|t = 1, 2, ..., N\}$, trong đó t là chỉ số thời gian, toàn bộ tập dữ liệu được chia thành các tập con ($k << N$) như là:

$$\begin{aligned} X_1 &= \{x_t|t = 1, \cdots, b_1\} \\ X_2 &= \{x_t|t = b_1 + 1, \cdots, b_2\} \\ &\vdots \\ X_k &= \{x_t|t = b_{k-1} + 1, \cdots, N\} \end{aligned}, \quad (2)$$

Trong đó: $b_1, ..., b_{k-1}$ ($b_i < b_{i+1}$, với mọi i) là các điểm ngắt, và $X_1 \cup ... \cup X_k = X$. Trong xấp xỉ từng đoạn, phân chia dữ liệu theo thời gian và xác định mô hình cục bộ là các mục tiêu chính. Chiều dài của các phân đoạn hoặc số các phân đoạn (k trong phương trình (2)) có thể được xác định bởi một số cố định và được xác định trước theo thời gian. Hoặc, chiều dài của mỗi phân đoạn có thể được xác định dựa trên cơ sở sự đồng nhất của một số thuộc tính đối với dữ liệu tổng hợp, ví dụ như các biến thiên nhỏ hoặc các xu hướng tương tự. Trong trường hợp chiều dài của các phân đoạn thường được xác định bằng cách xác định các điểm ngắt, mà một số thuộc tính của mô hình cục bộ thay đổi đáng kể, thì phương pháp này có thể tập trung vào việc xác định các điểm quan trọng nếu như các điểm tại đó có xu hướng thay đổi, trong khi xấp xỉ từng đoạn với chiều dài không đổi cho tất cả các phân đoạn có thể hữu ích hơn để hiểu xu hướng tổng thể của dữ liệu theo thời gian.

Sự lựa chọn công thức của mô hình cục bộ cho các phân đoạn có thể được xác định bởi một số giá trị mang tính đại diện hoặc bởi một mô hình tham số. Một mô hình cục bộ đơn giản là giá trị trung bình. Sử dụng giá trị trung bình, dữ liệu ban đầu được biểu diễn dưới dạng các hàm hằng số hoặc các hàm bậc thang. Đường tuyến tính và các mô hình đa thức cũng có thể được sử dụng cùng với xu hướng của từng đoạn dữ liệu tổng hợp. Thay vì sử dụng số trung bình, tổng các biến thiên[1] hoặc sự biến động có thể được sử dụng làm giá trị mang tính đại diện của các điểm dữ liệu trong mỗi phân đoạn, do vậy phải xem xét đến mục đích việc phân tích và khai phá.

Ví dụ: Xấp xỉ từng đoạn

Xấp xỉ gộp từng đoạn

Phương pháp xấp xỉ gộp từng đoạn (PAA [2,3]), hoặc xấp xỉ từng đoạn không đổi, sử dụng đơn giản và thực hiện tốt về lập chỉ mục. Lập chỉ mục là một nhiệm vụ khai phá chuỗi thời gian, tìm ra chuỗi thời gian tương tự nhất trong cơ sở dữ liệu với chuỗi thời gian truy vấn và các phép đo tương tự. Thứ nhất, dữ liệu gốc được chuẩn hóa, và sau đó dữ liệu chuẩn hóa được chia thành các khoảng bằng nhau và không chồng chéo khoảng thời gian. Cuối cùng, dữ liệu bị giảm được biểu diễn bởi giá trị trung bình của các điểm dữ liệu trong tất cả các phân đoạn. Cụ thể, một chuỗi thời gian chuẩn hóa $C = \{c_1, c_2, ..., C_N\}$ được biểu diễn như là $\overline{C} = \{\overline{c}_1, \overline{c}_2, ... , \overline{c}_m\}$ ($1 \le m \le N$, trong đó $\overline{c}_i$ là giá trị trung bình của phân đoạn thứ i,

$$\overline{c}_i = \frac{m}{N} \sum_{j=\frac{N}{m}(i-1)+1}^{\frac{N}{m}i} c_j \quad (3)$$

Các phân đoạn m chiều dài bằng nhau, được gọi là các khung, được chuyển đổi thành các giá trị trung bình của dữ liệu bên trong, và vector của các giá trị trung bình này biểu diễn độ giảm của C. Do đó, dữ liệu được trình bày giống với dữ liệu ban đầu khi m = N, và giá trị trung bình của dữ liệu ban đầu đạt được khi m = 1. Số phân đoạn m có thể là tham số do người dùng xác định. Do đó nó linh hoạt để điều chỉnh mức độ phân loại của dữ liệu bị giảm. Trong công thức (3), chúng ta giả sử m là một hệ số của N. Trong trường hợp m không phải là một hệ số của N, chiều dài của một chuỗi thời gian nhất định sẽ lớn hơn hoặc nhỏ hơn N, xem Keogh cùng các cộng sự [2], Chakrabarti và Mehrotra[4].

Phương pháp xấp xỉ hằng số từng đoạn thích nghi

Phương pháp xấp xỉ hằng số từng đoạn thích nghi (APCA[5]) giống như phương pháp PAA là xấp xỉ dữ liệu ban đầu thành những đoạn thẳng nằm ngang. Tuy nhiên, phương pháp này khác với PAA là các đoạn ở PAA có kích thước bằng nhau, còn ở APCA thì kích thước của các đoạn là khác nhau tùy theo dữ liệu. Kết quả là, APCA có thể phân đoạn dữ liệu gốc tốt hơn cùng với các lỗi lặp lại nhỏ hơn PAA. Để giảm lỗi lặp lại, APCA có xu hướng có nhiều điểm ngắt trong một phân đoạn dữ liệu biến động cao. Mặt khác, có ít điểm ngắt hơn trong một phân đoạn dữ liệu biến động thấp. Trước hết, các điểm ngắt được xác định bởi phép biến đổi Harr wavelet, đó là giải pháp tối ưu cho việc nén dữ liệu hiệu quả. Sau đó, các giải pháp được chuyển đổi trở lại với biểu diễn miền thời gian. Do đó, dữ liệu đã giảm $\overline{C}$ của chuỗi thời gian gốc C = {c_1, c_2, ..., c_N} chứa giá trị trung bình của dữ liệu trong các phân đoạn và chiều dài của các phân đoạn ghi lại các điểm ngắt của tất cả các phân đoạn như sau:

$$\overline{C} = \{(cv_1, cr_1), \ldots, (cv_n, cr_n)\}, \quad cr_0 = 0, \quad (n \ll N) \quad (4)$$

Trong đó: cv_i là giá trị trung bình của dữ liệu trong phân đoạn i; và cr_i là điểm đầu nút bên phải của phân đoạn i với chiều dài của phân đoạn i là $cr_i - cr_{i-1}$, i = 1, ..., n.

Tính năng tổng các biến thể phân đoạn

Trong khai phá dữ liệu chuỗi thời gian, nhiều biện pháp tương tự được đề xuất dựa trên cơ sở đo độ khoảng cách Euclide. Thông thường, tiêu chuẩn hoá dữ liệu được yêu cầu trước khi áp dụng phương pháp tương tự giữa dữ liệu chuỗi thời gian từ khoảng cách Euclide là nhạy cảm với nhiễu và quy mô dọc của dữ liệu. Lee cùng các cộng sự [1] đề nghị tổng hợp các biến thể (SSV). Phương pháp này được phát triển dựa trên ý tưởng tổng của biến thể là bất biến theo chuyển dịch chiều dọc của dữ liệu. Trước hết, so sánh tập dữ liệu chuỗi thời gian được chia thành các phân đoạn n với chiều dài bằng nhau và sau đó tổng các biến thể cho tất cả các phân đoạn được tính toán. Cụ thể, thuật toán tạo ra n phân đoạn (n << N) của các điểm từ chuỗi thời gian gốc C = {c_1, ..., c_N}, chồng chéo bằng cách chia sẻ một điểm tại ranh giới giữa hai phân đoạn liền kề.

$$(c_{1,1}, \cdots, c_{1,s}), (c_{2,1}, \cdots, c_{2,s}), \cdots, (c_{n,1}, \cdots, c_{n,s}), \quad (5)$$

Trong đó: $c_{i,s} = c_{i+1,1}$ (i = 1, ... n - 1). Lưu ý rằng các điểm ngắt được chia sẻ bởi hai phân đoạn liền kề. Nghĩa là, điểm kết thúc của phân đoạn i cũng trở thành điểm xuất phát của (i + 1) (i = 1, ..., n-1). Tổng các biến thể của phân đoạn thứ i là:

$$\sum_{j=1}^{s-1} |c_{i,j} - c_{i,j+1}| \qquad (6)$$

Do đó, dữ liệu bị giảm được biểu diễn dưới dạng một chuỗi các biến thể cho các phân đoạn có chiều dài n.

Xác định các điểm tới hạn

Mặc dù xấp xỉ từng đoạn thể hiện dữ liệu bằng cách gắn các mô hình cục bộ hoặc thu thập số liệu thống kê của các phân đoạn, việc biểu diễn dữ liệu bằng cách xác định các điểm tới hạn tập trung vào việc chọn một tập hợp các điểm từ toàn bộ tập dữ liệu. Các điểm dữ liệu đã chọn này góp phần quan trọng vào việc bảo toàn các tính năng của dữ liệu ban đầu. Mặc dù 'tầm quan trọng' của các điểm có thể được xác định tùy thuộc vào tính năng mà người dùng muốn tìm từ dữ liệu, nhiều cách tiếp cận để giảm dữ liệu trong miền thời gian cố gắng tìm ra các điểm góp phần tạo nên hình dạng của dữ liệu ban đầu, ví dụ khi một cú nhảy hoặc rơi đột ngột xảy ra. Nếu tất cả các điểm dữ liệu là có sẵn trước khi xử lý, chúng ta có thể phân tích cấu trúc dữ liệu tổng thể và chọn các điểm quan trọng liên tục cho toàn bộ tập dữ liệu theo các tiêu chí quan trọng (xử lý hàng loạt). Nếu không, chúng ta có thể áp dụng các tiêu chí này cho một nhóm các điểm dữ liệu tuần tự, vì dữ liệu mới được cập nhật để xác định các điểm quan trọng (xử lý trực tuyến). Hai ví dụ sau đây là phương pháp biểu diễn dữ liệu bằng cách xác định các điểm tới hạn bằng xử lý hàng loạt và trực tuyến.

Ví dụ: Xác định các điểm tới hạn

Các điểm tới hạn[6]

Một số điểm dữ liệu trong chuỗi thời gian có thể ảnh hưởng nhiều hơn đến hình dạng của dữ liệu, trong khi một số khác có thể bị bỏ qua ví dụ như nhiễu. Các mẫu được sử dụng trong phân tích kỹ thuật cho các thị trường tài chính thường được xác định dựa trên những điểm có ảnh hưởng như tối thiểu hoặc tối đa cục bộ. Chung và cộng sự[6] đã đề xuất các điểm tới hạn (PIP) là những điểm ảnh hưởng nhất đến hình dạng dữ liệu để giảm dữ liệu. Các PIP này được lựa chọn theo thứ tự dựa trên khoảng cách vuông góc hoặc thẳng đứng từ đường thẳng giữa hai điểm quan trọng trước đó. Đặc biệt, với chuỗi thời gian x_1, x_2, ..., x_n, điểm đầu tiên x_1 và điểm cuối cùng x_n ; P_1 là PIP thứ nhất và P_2 là PIP thứ hai. Sau đó, PIP thứ ba, là P_3 được xác định dựa trên khoảng cách vuông góc hoặc thẳng đứng từ đường thẳng giữa P_1 và P_2. Đó là, các điểm ở khoảng cách tối đa từ $\overline{P_1P_2}$ là P_3. Các điểm trong khoảng cách tối đa từ $\overline{P_1P_3}$ và $\overline{P_2P_3}$ được xác định là PIP thứ tư, P_4. Tương tự, để tìm PIP thứ k, P_k, thuật toán tìm kiếm điểm trong khoảng cách tối đa từ k-2 đường thẳng giữa các PIP lân cận cho đến khi nó xác định một số PIP được xác định trước đó. Cách tiếp cận này rõ ràng là xử lý hàng loạt vì tất cả các điểm dữ liệu được yêu cầu tại thời điểm phân tích để xác định các PIP thứ nhất và thứ hai, x_1 và x_n.

Nén bằng cách trích xuất các cực trị

Với ý tưởng rằng giá trị cực tiểu và giá trị cực đại cục bộ có thể tốt cho những điểm quan trọng ảnh hưởng đến hình dạng của dữ liệu, Fink và Gandhi đề xuất nén hiệu quả bằng cách điều tra cực trị (minima và maxima). Trong số tất cả các điểm cực trị, thuật toán chọn các điểm tới hạn góp phần tạo ra một mức độ dao động lớn hơn và loại bỏ các điểm dữ liệu còn lại. "Mức độ quan trọng" của cực trị được xác định bởi một tham số ngưỡng R>0, là một mức độ dao động "quan trọng" tối thiểu. Ví dụ, cho một chuỗi thời gian x_i, ..., x_j và R>0, x_k (i <k <j) là một cực tiểu (cực đại) quan trọng nếu (1) x_k = min {x_i, ..., x_j} (x_k = max {x_i, …, x_j}), và (2) khoảng cách (x_k, x_i) ≥ R và khoảng cách (x_k, x_j) ≥ R, trong đó khoảng cách (a, b) là khoảng cách giữa a và b sao cho $|a-b|$, $\frac{|a-b|}{|a|+|b|}$ hoặc $\frac{|a-b|}{\max(|a|,|b|)}$. Như vậy, một giá trị lớn của R hàm ý một tỷ lệ nén cao, nghĩa là, lựa chọn một vài số cực trị.

Thuật toán này có thể được sử dụng không chỉ cho việc xử lý hàng loạt mà còn cho xử lý trực tuyến để lập chỉ mục nhanh.

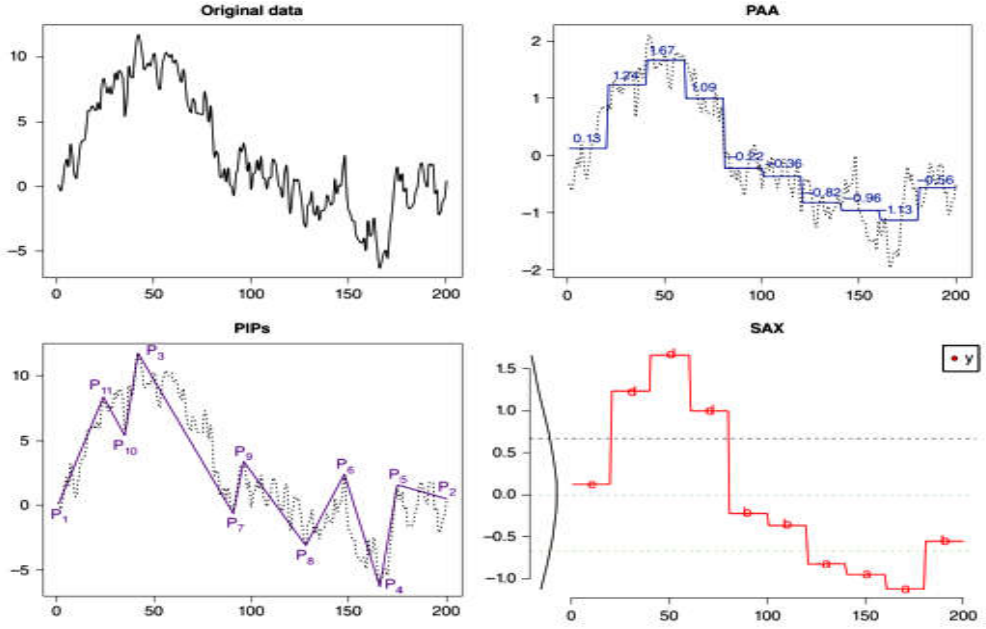
Biểu diễn dữ liệu ký hiệu hóa

Một cách tiếp cận phổ biến khác cho việc biểu diễn chuỗi thời gian là chuyển đổi dữ liệu số thành một số hữu hạn các biến rời rạc, thường là các biến ký hiệu. Chuyển đổi các giá trị số thành các chuỗi giúp tiết kiệm không gian bộ nhớ và cho phép tính toán nhanh. Phương pháp thứ nhất đơn giản là biểu diễn dữ liệu ký hiệu hóa trong một dải giá trị nhất định. Cho một chuỗi thời gian $X = \{x_i | x_i \in R, i = 1,...,N\}$, nó được ánh xạ tới chuỗi ký hiệu $S = \{s_i | s_i \in C, i = 1,...,N\}$, trong đó C là tập hợp các ký hiệu. Một phương pháp phổ biến khác là làm rời rạc dữ liệu từng đoạn và sau đó chuyển đổi những dữ liệu từng đoạn vào chuỗi. Tức là, dữ liệu biểu diễn bao gồm hai bước: Đầu tiên là xấp xỉ từng đoạn và sau đó, chuyển đổi các dữ liệu thu được từ bước đầu tiên thành các ký hiệu. Phương pháp thứ hai cho phép giảm dữ liệu cũng như tiết kiệm không gian bộ nhớ và tính toán hiệu quả hơn trong khi kích thước của dữ liệu ban đầu không thay đổi theo phương pháp cũ. Hai ví dụ tiếp theo mô tả chi tiết về biểu diễn dữ liệu ký hiệu hóa.

Ví dụ: Biểu diễn ký hiệu hóa[8]

Mô tả hình dạng chữ cái

Mô tả hình dạng chữ cái (SDA[8]) được đề xuất cho việc tìm kiếm tương đối trong cơ sở dữ liệu chuỗi thời gian lớn. Phương pháp này biến đổi sự khác biệt giữa hai điểm lân cận, x_i và x_{i+1}, đó là $d_i = x_{i+1} - x_i$, đến một tập hợp các chữ cái hữu hạn. Ví dụ, nó sử dụng a, u, s, d, và e tương ứng với các biến tăng cao, tăng nhẹ, ổn định, giảm nhẹ, và giảm nhiều. Các điểm cắt, lvalue (cận dưới) và hvalue (cận trên), để xác định một giá trị ký hiệu cho mỗi d_i được lấy dựa trên sự phân bố của d_i. Do đó, kiến thức về d_i là cần thiết để tìm điểm cắt tối ưu. SDA không phù hợp với dữ liệu nhiễu vì sự khác biệt d_i bị ảnh hưởng lớn bởi các nhiễu ngẫu nhiên và kết quả là không nắm bắt được hình dạng chung của dữ liệu ban đầu[9].



Hình 1 biểu diễn chuỗi thời gian theo phương pháp PAA, PIPs, và SAX. Kích thước của dữ liệu gốc đã được giảm từ N = 200 xuống n = 10 bằng phương pháp PAA và SAX, và còn n=11 bởi phương pháp PIP.

Xấp xỉ gộp ký hiệu hóa

Xấp xỉ gộp ký hiệu hóa (SAX[10]) biểu diễn dữ liệu chuỗi thời gian qua hai bước. Trước hết, SAX sử dụng dữ liệu bình thường để biểu diễn bởi PAA, và sau đó các hệ số thu được từ PAA được chuyển thành các chuỗi chữ cái. Do đó, cần phải có hai tham số để biểu diễn SAX: Số ký hiệu (kích thước chữ cái) và kích thước của dữ liệu bị giảm (chiều dài của dữ liệu bị giảm). Cho chuỗi thời gian C = {c_1, ..., c_N}, hệ số của dữ liệu giảm $\overline{C} = \{\overline{c}_1,...,\overline{c}_n\}$ (n<< N) bởi PAA được chuyển đổi dựa trên cơ sở các giá trị số lượng của $c_i's$. Cụ thể, với ký hiệu được xác định trước tập hợp {L_1, ..., L_a} (kích thước ký hiệu = a), SAX tìm điểm ngắt {β_1, ..., β_{a-1}} để xác định các giá trị ký hiệu sao cho P (Z <β_1) = P (β_1 ≤ Z ≤ β_2) = ... = P (β_{a-1} ≤ Z), trong đó Z ~ N (0,1). Sau đó, mỗi hệ số $\overline{c}_i$ trong phép tính xấp xỉ PAA được chuyển thành một ký hiệu $\hat{c}_i$ bằng:

$$\hat{c}_i = L_j \text{ khi và chỉ khi } \overline{c}_i \in \left[\beta_{j-1}, \beta_j\right), \quad (7)$$

Trong đó: i = 1, ..., n và j = 1, ..., a. SAX được sử dụng rộng rãi trong việc khai phá dữ liệu theo chuỗi thời gian do lợi thế của nó là tính toán nhanh và giảm kích thước đáng kể.

4. Kết luận

Mục tiêu cuối cùng của việc biểu diễn dữ liệu là giảm kích thước và trích xuất các tính năng quan trọng từ dữ liệu để cho phép thực hiện các công việc khai phá dữ liệu, chẳng hạn như phân loại, phân nhóm, lập chỉ mục, vv… Hai thuộc tính giảm dữ liệu và khai phá tính năng được trình bày trong tất cả các phương pháp biểu diễn dữ liệu. Mặc dù có rất nhiều phương pháp đã được đề xuất, không có phương pháp nào vượt trội hoàn toàn so với tất cả những phương pháp khác. Thay vào đó, các tính năng mà người sử dụng muốn truy cập dữ liệu, nên được xem xét để chọn một phương pháp biểu diễn dữ liệu thích hợp. Hình 1 minh họa biểu diễn chuỗi thời gian bằng ba phương pháp khác nhau.

Việc biểu diễn nguồn dữ liệu là một thách thức vì quy mô và tốc độ của nó, tuy nhiên lĩnh vực đầy hứa hẹn vì sự quan tâm đến "dữ liệu lớn" tiếp tục tăng lên trong thời gian gần đây. Hơn nữa, lựa chọn một biện pháp phù hợp là điều cần thiết cho việc khai phá dữ liệu và biểu diễn dữ liệu. Do tính chất độc đáo của dữ liệu chuỗi thời gian, kích thước lớn, nhiều giá trị gây nhiễu, và các phép đo tương tự thường được sử dụng, ví dụ như các quy tắc L_p không khả thi để đo hai dữ liệu chuỗi thời gian. Do đó hầu hết các phương pháp biểu diễn chuỗi thời gian thường được đề xuất với các biện pháp tương tự trong bài viết này. Vì vậy, khả năng áp dụng biện pháp tương tự đối với dữ liệu đã bị giảm cũng là một cân nhắc quan trọng trong việc biểu diễn dữ liệu.

Tài liệu tham khảo:

1. Lee S, Kwon D, Lee S, *Giảm kích thước cho chuỗi thời gian lập chỉ mục dựa trên khoảng cách nhỏ nhất,* J Inf Sci Eng, 2003, 19:697–711;

2. Keogh E, Chakrabarti K, Pazzani M, Mehrotra S, Giảm kích thước để tìm kiếm tương tự trong các cơ sở dữ liệu chuỗi thời gian trong *Kiến thức và Hệ thống thông tin,* tập 3, New York: Springer, 2001, 263–286;

(Xem tiếp trang 13)

Tiếp theo trang 41

3. Yi B, Faloutsos C, Lập chỉ mục chuỗi thời gian nhanh cho các chỉ tiêu tùy ý trong *Kỷ yếu của Hội nghị quốc tế lần thứ 26 về Cơ sở dữ liệu rất lớn,* San Francisco, Morgan Kaufmann Publishers Inc, 2000, VLDB'00: 385–394;

4. Chakrabarti K, Mehrotra S, Cây hybrid: một cấu trúc chỉ mục cho không gian đặc trưng trong *Kỷ yếu Hội thảo quốc tế về Kỹ thuật dữ liệu lần thứ 15,* IEEE, 1999, 440-447;

5. Keogh E, Chakrabarti K, Pazzani M, Mehrotra S, *Giảm kích thước thích ứng cục bộ để lập chỉ mục các cơ sở dữ liệu chuỗi thời gian lớn,* ACM SIGMOD Record 2001, 30:151–162;

6. Chung F, Fu T, Luk R, Ng V, Sự kết hợp chuỗi thời gian linh hoạt dựa trên các điểm tới hạn trong *Hội thảo quốc tế về Hội thảo Trí thức nhân tạo về học hỏi từ dữ liệu tạm thời và không gian,* 2001, 1–7;

7. Fink E, Gandhi H, *Sự nén của chuỗi thời gian bằng cách trích xuất các extrema lớn,* J Exp Theor Artif Intell 2011, 23:255–270;

8. André-Jönsson H, Dushan ZB, *Sử dụng tệp chữ ký để truy vấn dữ liệu theo chuỗi thời gian,* New York:Springer, 1977, 211–220;

9. Lin J, Keogh E, Wei L, Lonardi , Trải nghiệm SAX: Một biểu diễn biểu tượng cho chuỗi thời gian trong *Khai phá dữ liệu và Khám phá kiến thức,* tập 15, New York: Springer; 2007, 107–144;

10. Lin J, Keogh E, Wei L, Lonardi S, Chiu B. Một biểu diễn biểu tượng chuỗi thời gian, có liên quan đến thuật toán phát trực tuyến trong *Kỷ yếu hội thảo ACM SIGMOD lần thứ 8 về các vấn đề nghiên cứu trong khai phá dữ liệu và khám phá kiến thức, ACM, 2003.*

Thái Học (lược dịch)

Nguồn: Data representation for time series data mining: time domain approaches, http://onlinelibrary.wiley.com/doi/10.1002/wics.1392/epdf