biến đổi khí hậu và sự không chắc chắn trong các dòng chảy của sông. Thêm vào đó, SDSM thường được so sánh với các phương pháp downscaling thống kê.

Trong nghiên cứu dự báo lượng mưa, chúng tôi trước tiên sử dụng mô hình hồi quy tuyến tính. Tiếp đến, nghiên cứu tập trung vào hướng tiếp cận dựa trên các mô hình máy học tự động như: k láng giềng (k Nearest Neighbors) (Fix and Hodges, 1952), cây quyết định (Decision Trees) (Breiman *et al*., 1984), bagging (Breiman, 1996), rừng ngẫu nhiên (Random Forests) (Breiman, 2001) và máy học véc tơ hỗ trợ (Support Vector Machines) (Vapnik, 1995). Chúng tôi cũng đề xuất mô hình học phân cấp kết hợp giữa mô hình phân lớp và mô hình hồi quy dựa trên rừng ngẫu nhiên và máy học véc tơ hỗ trợ.

3 MÔ HÌNH DỰ BÁO

3.1 Mô hình hồi quy tuyến tính (linear regression - LM)

Hồi quy là phương pháp toán học được áp dụng thường xuyên trong thống kê để phân tích mối liên hệ giữa các hiện tượng kinh tế xã hội. Hồi quy tuyến tính được sử dụng rộng rãi trong thực tế do tính chất đơn giản hóa của hồi quy.

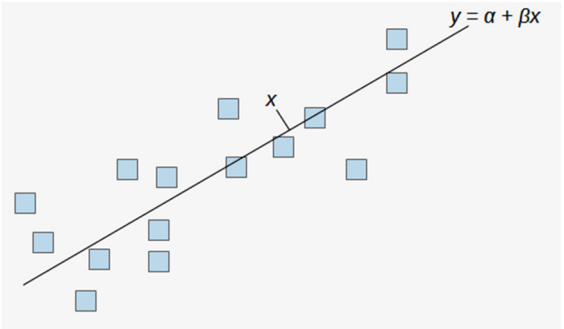
Phân tích hồi quy là phân tích thống kê để xác định mối quan hệ giữa biến phụ thuộc y với một hay nhiều biến độc lập x. Mô hình hồi quy đơn giản nhất là hàm tuyến tính (bậc 1) dùng để mô tả mối quan hệ giữa biến phụ thuộc và biến độc lập là tuyến tính. Mô hình hồi quy tuyến tính có dạng:

$$y = \alpha + \beta x \quad (1)$$

với α là chặn (intercept), β là độ dốc (slope)

Các tham số α, β của mô hình được ước lượng từ dữ liệu quan sát. Xét tập dữ liệu gồm m phần tử $x_1, x_2, \ldots, x_m$ trong không gian n chiều (biến độc lập, thuộc tính), có giá trị tương ứng của biến phụ thuộc (cần dự báo) là $y_1, y_2, \ldots, y_m$. Các tham số α, β của mô hình được ước lượng bằng phương pháp bình phương bé nhất (least squares):

$$Min\left(\sum_{i=1}^{m}[y_i - (\alpha + \beta x_i)]^2\right) \quad (2)$$



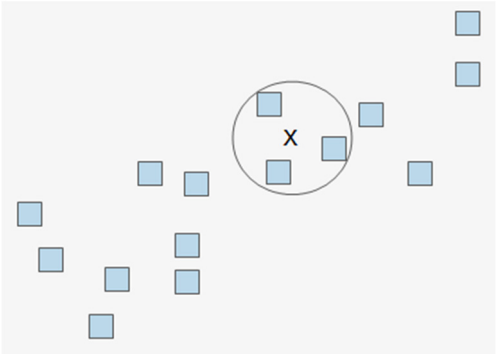
Hình 1: Hồi quy tuyến tính

Giá trị dự báo cho phần tử mới x dựa vào công thức (3):

$$\hat{y} = \alpha + \beta x \quad (3)$$

3.2 k láng giềng (k Nearest Neighbors - kNN)

Giải thuật k láng giềng (kNN) được Fix và Hodges đề xuất từ những năm 1952. Đây là phương pháp rất đơn giản nhưng cũng cho hiệu quả cao trong khai mỏ dữ liệu (Hastie *et al*., 2009; Wu and Kumar, 2009). Giả sử có tập dữ liệu bao gồm m phần tử $x_1, x_2, \ldots, x_m$ trong không gian n chiều, có giá trị tương ứng của biến phụ thuộc là $y_1, y_2, \ldots, y_m$.



Hình 2: Giải thuật k láng giềng

Giải thuật kNN không có quá trình học. Khi dự đoán giá trị biến phụ thuộc của phần tử dữ liệu x mới đến, giải thuật đi tìm k láng giềng ($k=1, 2, \ldots$) của x từ tập dữ liệu học là các phần tử $\{(x_1,y_1), \ldots, (x_k,y_k)\}$, sau đó thực hiện:

– Phân lớp với bình chọn số đông trong các giá trị $\{y_1, \ldots, y_k\}$,

– Hồi quy với giá trị trung bình của các $\{y_1, \ldots, y_k\}$.