

giải thuật, việc làm duy nhất là thay thế các tích vô hướng của hai véc tơ trong các công thức bởi một trong các hàm nhân cơ bản được dùng phổ biến như:

– Đa thức bậc d : $K(u, v) = (u.v + c)^d$ (15)

– Radial Basis Function (RBF):

$K(u, v) = \exp(-\gamma||u - v||^2)$ (16)

3.7 Mô hình hồi quy phân cấp

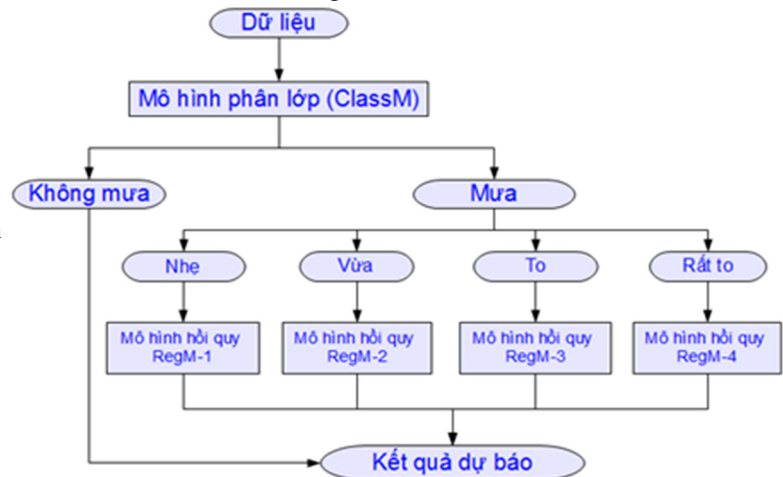
Chúng ta có thể sử dụng trực tiếp các mô hình hồi quy vừa được trình bày để dự báo lượng mưa trong ngày. Mỗi mô hình đều có ưu điểm và khuyết điểm khác nhau. Chẳng hạn mô hình hồi quy tuyến tính thì rất đơn giản, thời gian xây dựng mô hình và dự báo nhanh, điều tất yếu là độ chính xác cũng không cao. Riêng mô hình kNN cũng đơn giản, chỉ sử dụng duy nhất tham số là $k = 1, 2, \dots$ là số láng giềng, tuy nhiên thời gian dự báo lâu hơn do phải tìm kiếm láng giềng của phần tử cần dự báo. Mô hình cây quyết định chỉ cần duy nhất tham số $minobj = 1, 2, \dots$ là số phần tử tối thiểu tại mỗi nút lá, thời gian xây dựng mô hình và dự báo nhanh, đạt được độ chính xác tương đối cao so với kNN và hồi quy tuyến tính. Bagging và rừng ngẫu nhiên thì cần thêm tham số là số lượng cây $T = 50, 100, \dots$ riêng, rừng ngẫu nhiên còn sử dụng thêm tham số là số biến ngẫu nhiên sử dụng cho phân hoạch (trong khoảng $\left[\sqrt{n}, \frac{n}{2} \right]$ với n là số biến của dữ

liệu). Cho dù phức tạp, nhưng Bagging, rừng ngẫu nhiên vẫn đơn giản hơn khi so với máy học SVM. Xây dựng mô hình SVM cần thiết ba tham số là hằng số $c > 0$ (để chỉnh độ rộng lề và lỗi), độ lệch chuẩn là σ và tham số của hàm nhân. Thời gian xây dựng mô hình và dự báo rất cao (ít nhất là bậc 2 so với số lượng phần tử). Mặc dù phức tạp, nhưng Bagging, rừng ngẫu nhiên và SVM là mô hình phi tuyến, nên xử lý tốt cho các vấn đề phi tuyến, đặc biệt là dự báo lượng mưa đang xét ở đây.

Hình 8 minh họa mô hình hồi quy phân cấp. Dữ liệu được phân lớp (ClassM) vào một trong năm lớp như: không mưa (lượng mưa = 0), mưa nhẹ (lượng mưa: 0-2,5 mm), mưa vừa (lượng mưa: 2,5-7,6 mm), mưa to (lượng mưa: 7,6-50mm), rất to (lượng mưa trên 50 mm). Tương ứng với từng lớp, một mô hình hồi quy được xây dựng cho phép dự báo tốt các phần tử thuộc lớp đó (RegM-i).

Xét về độ phức tạp, xử lý vấn đề phân lớp đơn giản hơn rất nhiều so với bài toán hồi quy. Hơn nữa, quá trình xây dựng mô hình hồi quy càng phức tạp hơn khi cần dự báo lượng mưa từ tập dữ liệu, có mối quan hệ phi tuyến giữa biến phụ thuộc (lượng mưa) với nhiều biến độc lập (bức xạ mặt trời, hướng gió, tốc độ gió, nhiệt độ). Từ phân tích trên, chúng tôi đề xuất mô hình hồi quy phân cấp, kết hợp giữa mô hình phân lớp và nhiều mô hình hồi quy cục bộ để nâng cao hiệu quả xử lý của dự báo lượng mưa.

Hình 8: Mô hình phân cấp (phân lớp + hồi quy)



4 KẾT QUẢ THỰC NGHIỆM

Để tiến hành đánh giá hiệu quả của các mô hình dự báo lượng mưa, chúng tôi tiến hành cài đặt tất cả các chương trình dự báo bằng ngôn ngữ R (Ihaka and Gentleman, 1996) có sử dụng các gói thư viện FNN, rpart, ipred, randomForest, e1071.

Chương trình bao gồm các mô hình: Hồi quy tuyến tính (LM), k láng giềng (kNN), Cây quyết định (DT), Bagging (BagDT), Rừng ngẫu nhiên (RF), Máy học véc tơ hỗ trợ cho hồi quy SVR, Mô hình phân cấp: RF phân lớp và RF hồi quy (RFC-RFR), Mô hình phân cấp: SVC phân lớp và SVR hồi quy (SVC-SVR) để dự báo lượng mưa.