Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất
Trang điện tử: http://tapchi.humg.edu.vn

Nghiên cứu ứng dụng phương pháp SVM trong dự báo mực nước ngầm tại một số giếng quan trắc vùng Hà Nội

Đặng Hữu Nghị [1,*], Đặng Đình Phúc [2], Bùi Thị Vân Anh [1]

[1] *Khoa Công nghệ Thông tin, Trường Đại học Mỏ - Địa chất, Việt Nam*
[2] *Hội Địa chất thủy văn Việt Nam, Việt Nam*

THÔNG TIN BÀI BÁO

Quá trình:
Nhận bài 20/6/2017
Chấp nhận 20/8/2017
Đăng online 30/10/2017

Từ khóa:
Máy hỗ trợ véc tơ
Học máy
Dự báo mực nước ngầm

TÓM TẮT

Nước ngầm được sử dụng rộng rãi trong nền kinh tế quốc dân, vì thế việc dự báo sự thay đổi lượng trữ nước ngầm mà đặc trưng của nó là mực nước là một việc hết sức cần thiết. Trong bài báo này chúng tôi sử dụng phương pháp SVM (Support Vector Machine) để dự báo mực nước ngầm cho các giếng khoan trong vùng Hà Nội. Việc thử nghiệm được tiến hành với 2 phương án dự báo. Phương án thứ nhất chúng tôi dự báo mực nước tương lai dựa vào mực nước tại thời điểm hiện tại và quá khứ. Phương án thứ hai, theo các nghiên cứu thuỷ văn, dữ liệu về mưa cũng ảnh hưởng rất lớn đến trữ lượng nước ngầm trong tương lai. Việc dự báo cần cả thông số về lượng mưa tại thời điểm hiện tại và lượng mưa trong quá khứ.

1. Mở đầu

Công việc dự báo mực nước dưới đất đóng vai trò hết sức quan trọng nó giúp các nhà quản lý đề ra các biện pháp quy hoạch và sử dụng hợp lý nguồn tài nguyên nước quý giá.

Đã có một số phương pháp dự báo đamg được áp dụng hiện nay (Yannan Zhao et al., 2016). Đó là các phương pháp:

- Dựa trên mô hình vật lý
- Dựa trên mô hình toán học
- Dựa trên mô hình thống kê

Trong bài báo này chúng tôi sử dụng mô hình thống kê và cụ thể là sử dụng các phương pháp học máy (machine learning) để dự báo mực nước ngầm cho một số giếng khoan trong vùng Hà Nội.

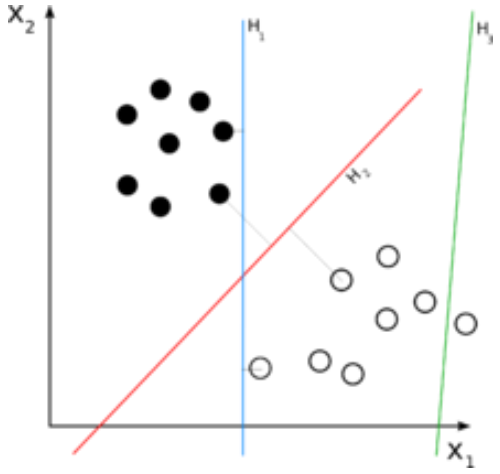
2. Support Vector Machine (SVM)

Support Vector Machine (SVM) là phương pháp mạnh và chính xác nhất trong số các thuật toán nổi bật ở lĩnh vực khai thác dữ liệu. SVM bao gồm hai nội dung chính là: support vector classifier (SVC), bộ phân lớp dựa theo vector hỗ trợ, và support vector regressor (SVR), bộ hồi quy dựa theo vector hỗ trợ. Được phát triển đầu tiên bởi Vapnik vào những năm 1990, SVM có nền tảng lý thuyết được xây dựng trên nền móng lý thuyết xác suất thống kê. Nó yêu cầu số lượng mẫu huấn luyện không nhiều và thường không nhạy cảm với số chiều của dữ liệu. Trong những thập niên qua, SVM đã phát triển nhanh chóng cả về lý thuyết lẫn thực nghiệm.

*Tác giả liên hệ
E-mail: nghidanghuu@gmail.com

2.1. Support Vector Classifier – SVC.

Trong mô hình học có giám sát, thuật toán được cho trước một số điểm dữ liệu cùng với nhãn của chúng thuộc một trong hai lớp cho trước. Mục tiêu của thuật toán là xác định xem một điểm dữ liệu mới sẽ được thuộc về lớp nào. Mỗi điểm dữ liệu được biểu diễn dưới dạng một vector p-chiều, và ta muốn biết liệu có thể chia tách hai lớp dữ liệu bằng một siêu phẳng p – 1 chiều. Đây gọi là phân loại tuyến tính. Có nhiều siêu phẳng có thể phân loại được dữ liệu. Một lựa chọn hợp lý trong chúng là siêu phẳng có lề lớn nhất giữa hai lớp. (Nello Cristianini, 2000)



Hình 1. Phân tách dữ liệu tuyến tính.

Như Hình 1, H3 không chia tách hai lớp dữ liệu. H1 phân tách hai lớp với lề nhỏ và H2 phân tách với lề cực đại.

Ta có một tập huấn luyện D gồm n điểm có dạng

$$D = \{(x_i, y_i) \mid x_i \in R^p, y_i \in \{-1,1\}\}_{i=1}^{n} \quad (1)$$

với yi mang giá trị 1 hoặc -1, xác định lớp của điểm x_i. Mỗi x_i là một vectơ thực p-chiều.

Ta cần tìm siêu phẳng có lề lớn nhất chia tách các điểm có $y_i = 1$ và các điểm có $y_i = -1$. Mỗi siêu phẳng đều có thể được viết dưới dạng một tập hợp các điểm x thỏa mãn

$$w.x - b = 0 \quad (2)$$

Đây là một bài toán quy hoạch toàn phương. Cụ thể hơn:

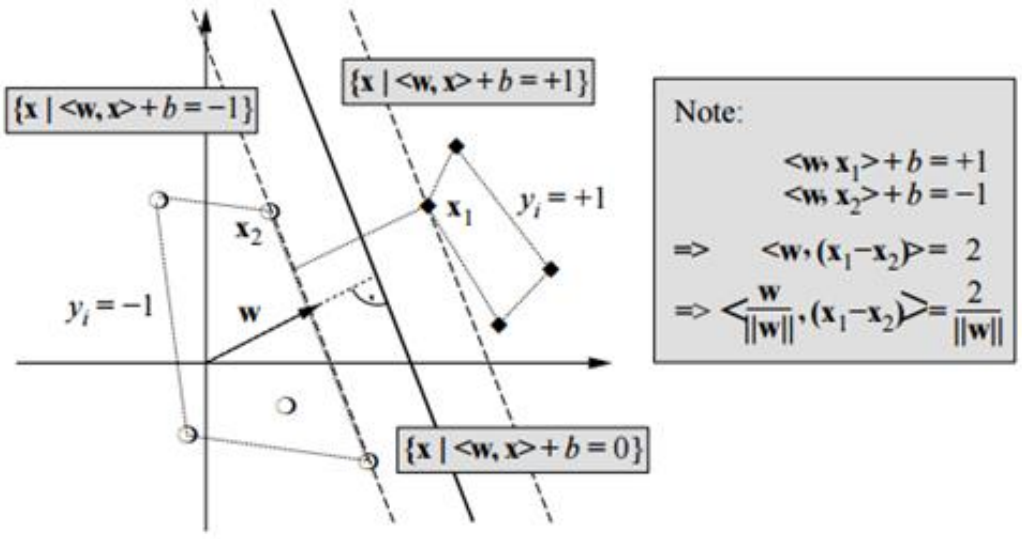
$$\min_{\{w,b\}} \frac{1}{2}\|w\|^2 \quad (3)$$

với điều kiện

$$y_i(w.x_i - b) \geq 1, \quad i = 1, \dots, n$$

Việc yêu cầu dữ liệu phải phân tách tuyến tính hoàn toàn là nghiêm ngặt và không phù hợp với các bài toán thực tế, đặc biệt là các trường hợp phân lớp phi tuyến phức tạp. Trong khi đó, các mẫu không phân tách tuyến tính hoàn toàn dẫn đến việc không thể giải quyết các bài toán tối ưu để tìm w và b tương ứng. Để giải quyết vấn đề này, có hai cách tiếp cận chính:

- ♣ Lề mềm (Soft-margin)
- ♣ Thủ thuật hàm nhân (hàm Kernel).



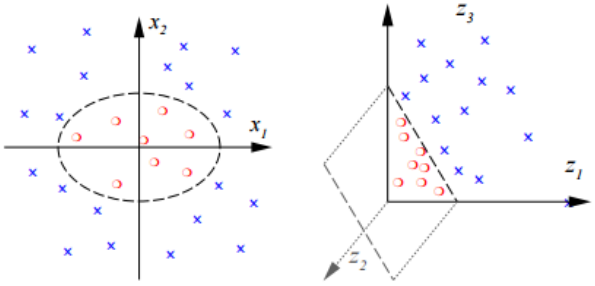
Hình 2. Phân tách dữ liệu bởi siêu phẳng.

Hàm nhân (hàm Kernel) cho phép chuyển đổi phi tuyến dữ liệu đầu vào sang không gian có số chiều cao hơn để có khả năng phân tách tuyến tính.

Gọi $\Phi: X \rightarrow H$ là phép biến đổi phi tuyến từ không gian đầu vào m chiều X vào không gian đặc trưng H mà ở đó các mẫu có thể phân tách tuyến tính.

Ví dụ: ta có phép biến đổi $\Phi: R2 \rightarrow R3$

$(x_1, x_2) \rightarrow (z_1, z_2, z_3) := (x_1^2, \sqrt{2}x_1x_2, x_2^2)$



Hình 3. Ánh xạ từ không gian 2 chiều sang không gian 3 chiều.

Khi đó đường phân cách tối ưu được định nghĩa như sau:

$$w^T.\Phi(x) + b = 0 \quad (4)$$

Tương tự như phần trên thì vector trọng số tối ưu là trong không gian đặc trưng mới sẽ là:

$$w^T = \sum_{i=1}^{n} \alpha_i y_i \Phi(x_i) \quad (5)$$

Theo đó siêu mặt phẳng tối ưu trong không gian đặc trưng mới là:

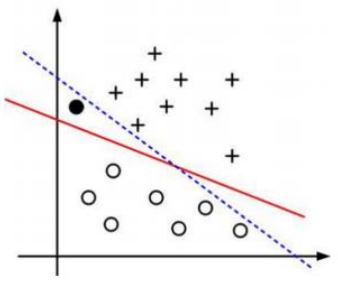
$$\sum_{i=1}^{n} \alpha_i y_i \Phi^T(x_i)\Phi(x) + b = 0 \quad (6)$$

Trong đó $\Phi^T(x_i)\Phi(x)$ là tích vô hướng của hai vector $\Phi(x)$ và $\Phi(x_i)$. Từ đây, chúng ta có thể áp dụng hàm kernel tích vô hướng.

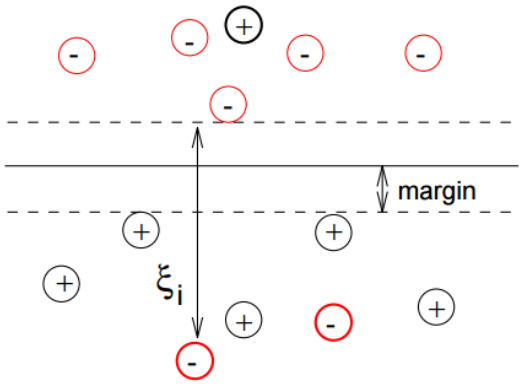
Vì nhiều lý do, do bản chất hoặc do sai sót trong quá trình thu thập dữ liệu, tồn tại một số điểm thuộc lớp này lẫn lộn vào lớp kia, điều này sẽ làm phá vỡ sự phân tách tuyến tính. Nếu ta cố tình phân tách hoàn toàn sẽ làm cho mô hình dự đoán quá khớp. Để chống lại sự quá khớp, người ta mở rộng SVC để nó chấp nhận một vài điểm phân lớp sai. Kỹ thuật này gọi là lề mềm (Soft margin).

Để làm điều này, một biến (gọi là slack variable) ξ_i được thêm vào biểu thức cần tối ưu nhằm cho phép mô hình phân lớp thực hiện phân lớp sai ở mức chấp nhận được:

Với SVM lề mềm L1 việc tìm siêu phẳng lớn



Hình 4. Các điểm nhiễu trong tập dữ liệu.



Hình 5. Phân lớp với biến lỏng.

nhất bằng cách giải bài toán tối ưu sau:

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + \frac{1}{2}C\sum_{i=1}^{N}\xi_i \quad (7)$$

Với các ràng buộc

$$y_i\left(\langle w.x_i\rangle - b\right) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1....N$$

2.2. Support Vector Regressor – SVR

Ý tưởng cơ bản của SVR là ánh xạ không gian đầu vào (mà nếu ta áp dụng trực tiếp hồi qui tuyến tính thì không hiệu quả) sang một không gian đặc trưng nhiều chiều mà ở đó, ta có thể áp dụng được hồi qui tuyến tính. Đặc điểm của SVR là cho ta một giải pháp thưa (sparse solution); nghĩa là để xây dựng được hàm hồi qui, ta không cần phải sử dụng hết tất cả các điểm dữ liệu trong bộ huấn luyện. Những điểm có đóng góp vào việc xây dựng hàm hồi qui được gọi là những Support Vector.

3. Dự báo mực nước ngầm

Một số phương pháp dự báo đang áp dụng hiện nay trong ngành thuỷ văn học là (Lalit Kumar, 2012):

- Dự báo dựa trên mô hình vật lý: Giải bài toán

bằng cách phân tích và tổng hợp các quá trình vật lí cơ sở hình thành nên hiện tượng. Tham số đặc trưng cho các quá trình này được xác định bằng cách đo đạc hoặc bằng con đường lý thuyết.

- Dự báo dựa trên mô hình toán: Mô hình toán học tập trung miêu tả hệ thống dưới dạng toán học. Mô hình toán học là tập hợp các phương trình toán học, các mệnh đề logic thể hiện các quan hệ giữa các biến và các thông số của mô hình để mô phỏng hệ thống tự nhiên, hay nói cách khác mô hình toán học là một hệ thống biến đổi đầu vào (Tầng chứa nước, cách nước, điều kiện biên, hệ số thấm v.v...) thành đầu ra (dòng chảy, mực nước v.v...).

Như vậy để thực hiện các phương pháp này trong thực tế cần phải có các số liệu quan trắc chi tiết các đặc trưng định lượng khác nhau của lưu vực, cấu trúc thuỷ địa chất, hệ số thấm...

Trong bài báo này chúng tôi sử dụng mô hình thống kê để dự báo. Mô hình này dựa trên một cơ sở dữ liệu thực tế đủ lớn và áp dụng các kỹ thuật tính toán hiện đại đang được dùng nhiều trong lĩnh vực công nghệ thông tin những năm gần đây như mạng nơ-ron nhân tạo, cây quyết định, máy hỗ trợ vector v.v... (Ch. Suryanarayana et al., 2014) và chúng thường được nói tới như các phương pháp học máy. Hướng tiếp cận này đã mở ra một phương pháp mới để mô hình hoá các quá trình tự nhiên phức tạp, phi tuyến mà việc đặc tả các quá trình vật lý của nó gặp nhiều khó khăn. (Suja S Nair et al, 2016).

Cụ thể trong bài báo này chúng tôi sử dụng SVM để dự báo mực nước cho các giếng khoan trong vùng Hà Nội. Vùng nghiên cứu nằm ở phía Đông Nam thành phố Hà Nội, và nằm ở bờ Hữu sông Hồng, bao gồm các huyện Kim Bảng, Lý Nhân và Duy Tiên. Địa hình trong vùng khá bằng phẳng, cao độ mặt đất biến thiên từ 2 tới 5m, có nhiều ao hồ và kênh rạch, khí hậu trong vùng mang đặc trưng của khí hậu nhiệt đới gió mùa, tổng lượng mưa năm từ 1800 tới 2000mm.

Trên khu vực nghiên cứu tồn tại 3 tầng chứa nước: tầng chứa nước Holocen trên (qh2) được cấu tạo bởi cát mịn xen bột có mức độ được chứa nước yếu. Tầng chứa nước Holocen dưới (qh1) được cấu tạo bởi cát mịn tới trung, có mức độ chứa nước trung bình. Tầng chứa nước Pleistocen (qp) được cấu tạo bởi cát mịn trung thô lẫn sạn sỏi có mức độ chứa nước tốt.

Chúng tôi thử nghiệm việc sử dụng SVM để dự báo mực nước tại các giếng Q_83a_0, Q_85b_0 và Q_87_0.

Chúng tôi sử dụng số liệu từ năm 2000 đến năm 2013 tại giếng Q_83a_0, Q_85b_0 và Q_87_0. Số liệu này được đo mỗi tháng một lần, tập dữ liệu này có tất cả 98 mẫu:

- Phần dữ liệu học (training set): Chúng tôi sử dụng 70 mẫu đầu tiên để đưa vào huấn luyện (các mẫu từ cuối năm 2000 đến 6/2009).

- Phần dữ liệu kiểm tra (test set): Chúng tôi sử dụng 28 sau để đưa vào kiểm tra (các mẫu từ 7/2009 đến cuối năm 2013).

Chúng tôi đề xuất sử dụng 2 phương án để dự báo như sau:

+ Phương án 1:

Trong phương án này việc dự báo mực nước trong giếng trong tương lai trước 1 tháng Q(t+1) dựa vào các mực nước tại thời điểm hiện tại và quá khứ. Chúng tôi sử dụng ba giá trị mực nước làm đầu vào gồm:

- Mực nước hiện tại: Q(t)
- Mực nước trước đó 1 tháng: Q(t-1)
- Mực nước trước đó 2 tháng: Q(t-2)

$$Q(t+1) = f(Q(t), Q(t-1), Q(t-2)) \quad (8)$$

+ Phương án 2

Theo các nghiên cứu thuỷ văn, dữ liệu về mưa cũng ảnh hưởng rất lớn đến mực nước trong tương lai.

Bảng 1. Tọa độ và vị trí một số giếng khoan trắc.

Giếng	Tọa độ			Vị trí		
	X	Y	z	Xã	Huyện	Tỉnh
Q_83a_0	2372204	18594719	3.86	Châu Sơn	Kim Bảng	Hà Nam
Q_85b_0	2373976	18597556	3.18	Lâm Hạ	Duy Tiên	Hà Nam
Q_87_0	2378278	18505188	3.87	Hùng Lý	Lý Nhân	Hà Nam

Việc dự báo cần cả thông số về lượng mưa tại thời điểm hiện tại và lượng mưa trong quá khứ (do ảnh hưởng đến mực nước ngầm). Trong phương án này, dự báo mực nước tương lai trước 1 tháng. *Q(t+1)* không những chỉ dựa vào các mực nước quá khứ và hiện tại *(Q(t), Q(t -1), Q(t - 2))* như phương án 1 mà còn phụ thuộc vào lượng mưa trong quá khứ và hiện tại tại khu vực đó *(X (t), X (t -1), X (t - 2)).*

$$Q(t+1) = f(Q(t), Q(t-1), Q(t-2), X(t), X(t-1), X(t-2)) \quad (9)$$

Chúng tôi đã tiến hành cho SVM học các mối quan hệ này theo thủ tục sau:

Tỉ lệ dữ liệu của các thuộc tính về miền giá trị [-1, 1]. (Nghi Dang Huu et al., 2011)

Lựa chọn mô hình: Trong bài báo này chúng tôi sử dụng hàm nhân RBF cho SVM (Chih-Wei Hsu, 2008). Để lựa chọn giá trị tối ưu cho các tham số C và γ chúng tôi sử dụng phương pháp tối ưu bầy đàn. (Nghi Dang Huu et al., 2011)

Sử dụng các tham số C và γ tốt nhất để huấn luyện tập huấn luyện (70 mẫu).

Kiểm thử trên tập kiểm thử (28 mẫu).

Và cho kết quả dự báo như Bảng 2.

Biểu đồ kết quả dự báo và thực nghiệm được thể hiện như Hình 7, Hình 8, Hình 9.

4. Kết luận

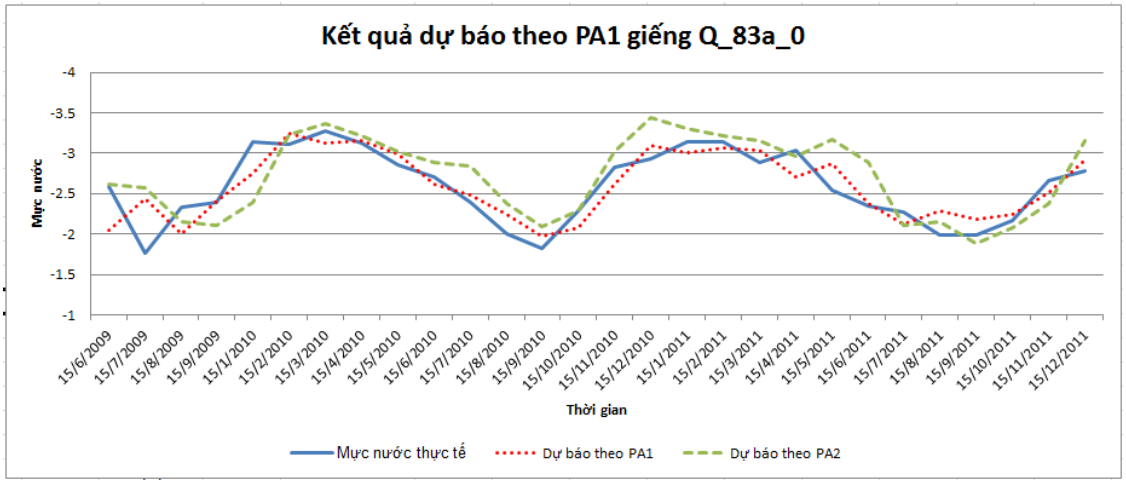
Trong bài báo này chúng tôi đã sử dụng phương pháp SVM để dự báo mực nước ngầm tại một số giếng quan trắc trong vùng Hà Nội. Chúng tôi đưa ra 2 phương án dự báo, phương án thứ nhất là dự báo mực nước trong giếng trong tương lai dựa vào mực nước tại thời điểm hiện tại và quá khứ.

Phương án thứ hai là dự báo mực nước trong giếng trong tương lai dựa vào lượng mưa, mực nước tại thời điểm hiện tại và quá khứ.

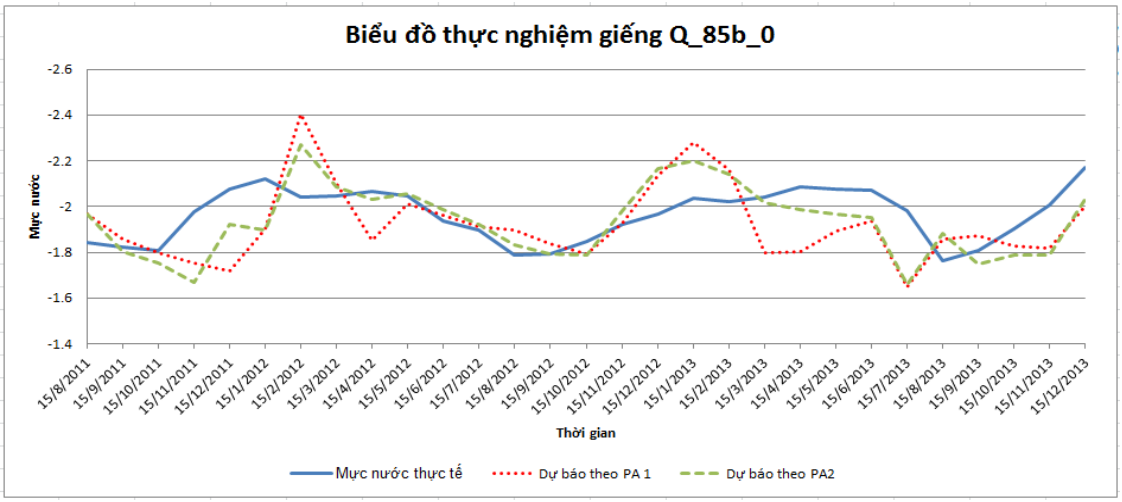
Qua thực nghiệm cho thấy việc áp dụng SVM cho bài toán dự báo mực nước ngầm cho kết quả có độ chính xác gần tương đương với việc dự báo sử dụng mô hình vật lý hoặc toán học mà không cần phải có các số liệu quan trắc chi tiết các đặc trưng định lượng khác nhau của lưu vực, cấu trúc thuỷ địa chất, hệ số thấm... Kết quả dự báo cũng cho thấy dự báo theo phương án hai chính xác hơn phương án 1. Trong thời gian tới chúng tôi sẽ thu thập và đưa thêm vào các tham số như bốc hơi để nhằm nâng cao hơn độ chính xác, chúng tôi cũng sẽ tiến hành áp dụng mô hình trên cho các vùng khác như Tây Nguyên.

Bảng 2. Kết quả dự báo.

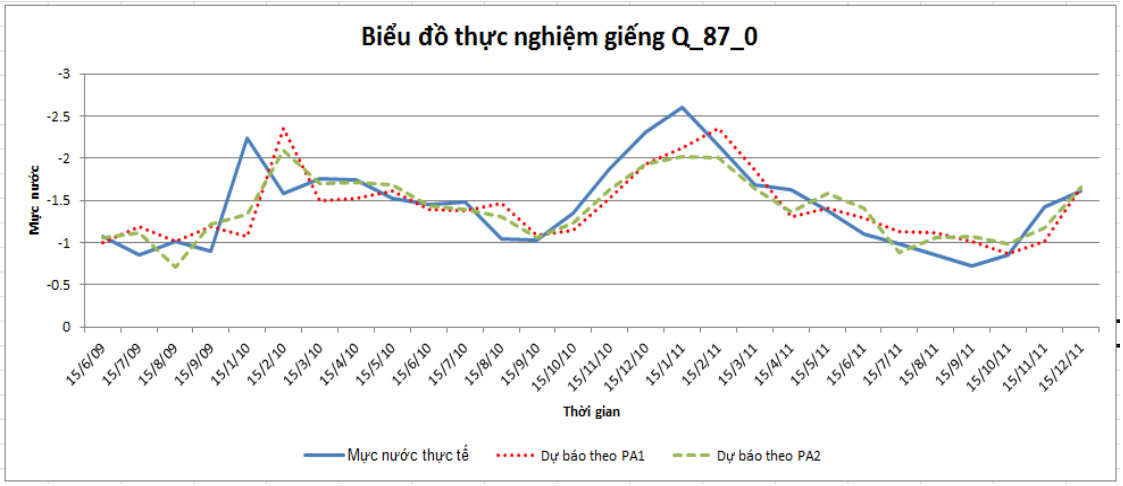
Giếng	Kết quả phương án 1		Kết quả phương án 2	
	Sai số trung bình bình phương	Hệ số tương quan bình phương	Sai số trung bình bình phương	Hệ số tương quan bình phương
Q_85b_0	0.030	0.58	0.021	0.70
Q_83a_0	0.103	0.518	0.05	0.746
Q_87_0	0.085	0.623	0.063	0.72



Hình 7. Kết quả dự báo của giếng Q_83b_0.



Hình 8. Kết quả dự báo của giếng Q_85b_0.



Hình 9. Kết quả dự báo của giếng Q_87_0.

Một trong những hướng phát triển tiếp theo của chúng tôi là nghiên cứu, cải tiến và thử nghiệm các phương pháp học máy tiên tiến khác để có thể nâng cao được kết quả dự báo và thời gian dự báo.

Tài liệu tham khảo

Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, 2008, A practical guide to Support Vector Classification.

Lalit Kumar, 2012. Temporal Models for Groundwater Level Prediction in Regions of Maharashtra, Department of Computer Science and Engineering Indian Institute of Technology Bombay.

Nello Cristianini, John Shawe-Taylor, 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, *Cambridge University Press*.

Nghi Dang Huu, Mai Luong Chi, 2011. "A New Model of Particle Swarm Optimization for Model Selection of Support Vector Machine", *ACIIDS, volume 351 of Studies in Computational Intelligence*, 167-173, Springer.

Nghi Dang Huu, Mai Luong Chi, 2011. "Training Data Selection for Support Vector Machines Model" *International Conference on Information and Electronics Engineering IPCSIT vol.6 IACSIT Press*, Singapore.

Suja S Nair, Dr. Sindhu G, 2016. Groundwater level forecasting using Artificial Neural Network, *International Journal of Scientific and Research Publications*.

Suryanarayana, C., Sudheer, C., Vazeer Mahammood, Panigrahi, B. K., 2014. India, Neurocomputing 145, 324–335.

Yannan Zhao, Yuan Li, Lifen Zhang, Qiuliang Wang, 2016. Groundwater level prediction of landslide based on classification and regression tree, *Geodesy and geodynamics*.

ABSTRACT

Study and applications the svm method in groundwater level forecast at some well in ha noi region

Nghi Huu Dang [1], Phuc Dinh Dang [2], Van Anh Thi Bui [1]

[1] *Faculty of Information Technology, University of Mining and Geology, Vietnam*
[2] *Vietnam Hydrological Association, Vietnam*

Groundwater is widely used in the national economy, so it is important to forecast the change in groundwater reserves that characterize it as water level. In this paper we use SVM (Support Vector Machine) method to forecast groundwater table for wells in Hanoi area. The test was conducted with 2 forecasting options. The first is that we predict future water levels based on current and past water levels. The second option, according to hydrological studies, data on rain, also has a huge impact on future groundwater reserves. Forecasting needs both current rainfall and past rainfall.

Keywords: Support Vector Machine, Machine learnging, Groundwater Level Prediction.