SVM을 이용한 수문 시계열 자료의 예측
Hydrologic Time Series Forecasting using SVM

황석환*, 김중훈**, 정성원***
Seok Hwan Hwang, Joong Hoon Kim, Sung Won Jung

요 지

정확한 수문자료를 예측하기 위한 많은 연구들이 현재까지 진행되어 왔다. SVM(Support Vector Machine)은 그 구조가 신경망과 유사하나 신경망과는 다르게 철저히 통계적, 수학적 이론에 기반을 두고 있고 비선형 예측 모형이며 지역해 문제가 발생하지 않는 다는 점 등으로 인해 상당히 견고한 모형으로 평가받고 있다. 본 연구에서는 두 경우의 수문시계열 자료를 이용하여 전통적인 통계학적 모형과 신경망 모형 그리고 수문학 분야에서는 아직까지 적용된 사례가 매우적은 SVM 모형의 예측 결과 비교를 통해 모형의 장단점을 평가하였다. 비교 결과 SVM 모형은 수문시계열 자료 예측에 있어서 기존의 방법들에 비해 안정적이고 정확한 예측 결과를 보여 주었다.

핵심용어 : SVM, 예측, 시계열

1. 서론

수문학 분야에서 다루어지는 자료는 시간과 밀접한 관련이 있는 시계열 자료가 대부분이다. 이수는 물론 치수적인 면에서 신속하고 정확한 수문자료의 예측은 그 중요도가 매우 높다. 대표적인 것이 신경망(Neural network)을 이용한 수위, 유출 등의 실시간 예측에 관한 연구가 국내외 적으로 많이 이루어져 왔다. 이는 신경망이 기존의 전통적인 통계적, 수학적 모형으로 규명하기 어려웠던 자료의 비선형적인 관계를 여러 개의 중간층과 학습의 과정을 통하여 비교적 적절히 나타낼 수 있었던 장점을 가지고 있었기 때문이다. 더불어 입력 자료의 수에 제한이 없고 처리 과정에 대한 복잡한 고려가 필요하지 않았던 구조나 이론상의 단순함이 예측모형으로 적절하다고 판단되었기 때문이다.

그러나 신경망을 실무에 적용하는 측면에서 보면 실제로 이런 장점들이 현실적인 면에서 단점으로 부각되어 그 적용성에 한계를 보여주었다. 대표적으로 1) 적절한 예측을 위해선 학습을 위한 방대한 양의 자료가 필요하고, 2) 적절한 신경망의 구조나 가중치를 결정하기 위해서 많은 시간과 비용이 소요되며, 3) 이론적 배경이 부족하고 명확한 중간처리과정을 알 수 없다는 점과 4) 지역해 문제 등의 이유들로 인해 신경망 자체의 장점에도 불구하고 수자원 실무에 적절히 적용되지 못하고 있는 것이 현실이다.

이러한 이유로 현재 신경망 자체의 장점과 위에서 열거한 단점을 보완해 줄 수 있는 이론인 SVM(Support Vector Machine)이 국내외에서 활발히 연구가 진행되고 있다. SVM은 신경망과는 다르게 철저히 통계적, 수학적 이론을 근거로 개발되었으나 그 단위 구조가 신경망과 유사하여 상당히 견고한 모형으로 평가받고 있다. SVM은 비선형적인 문제를 풀기위해 저차원의 입력공간(Input space)을 고차원의 특징공간(Feature space)으로 가정을 하게 된다. 다시 말해 단순 평면상에서의 비선형 자료도 적절한 개수의 임의의 차원을 추가하면 충분히 선형화시킬 수 있기 때문이다. SVM은 Vapnik(1995)에 의한 처음 고안되었으며 내부적으로 커널(Kernel) 이란 개념을 적용하여 입력값을 적절한 출력값으로 계산한다. SVM의 장점은 앞서 간략히 언급했듯이 이론적 기반이 두터워 결과에 대한 신뢰도가 높고 더 중요한 점은 가중치 매개변수를 결정하기 위한 함수가 2차식으로 신경망의 최대단점중의 하나인 지역해(Local mimia)에 빠질 우려가 없다는 점이다.

본 논문의 목적은 기존의 전통적인 통계적 방법인 MLR(Multiple Linear Regression)과 학습을

* 정회원·한국건설기술연구원 수자원연구부 연구원·E-mail : sukany@kict.re.kr
** 정회원·고려대학교 사회환경시스템공학과 교수·E-mail : jaykim@korea.ac.kr
*** 정회원·한국건설기술연구원 수자원연구부 수석연구원·E-mail : swjung@kict.re.kr

통한 방법인 신경망 이론(NNBP : Neural Network using Back Propagation, NNGA : Neural Network using Genetic Algorithm)을 사용하여 예측된 수문시계열 자료의 비교를 통하여 SVM의 수문학적 적용성을 평가하는 것이다. 본 논문에서는 수문 시계열 자료의 예측을 위한 두 가지 다른 문제에 적용하여 그 성능을 시험하여 보았다. 첫 번째는 일 급수량 자료 예측이고, 두 번째는 충주댐 일 유입량 자료 예측이다. 일 급수량 자료 예측을 위해서는 일 급수량 자료 외에도 일 급수량 자료에 지대한 영향을 미치는 기온, 요일, 일조시간 등이 입력인자로 사용되었고, 충주댐 일 유입량 자료의 예측을 위해서는 일 유입량과 강우자료가 사용되었다. 상대적인 성능을 평가하기 위해 전형적인 통계적 방법인 다중선형회기방법(MLR)과 학습모형인 역전파 신경망모형(NNBP) 그리고 유전자 알고리즘을 사용하여 가중치를 최적화시킨 유전자 신경망모형(NNGA)이 사용되었다.

2. 이론적 배경

여러 미지의 사상들과 연관된 매개변수들로 인해 미래 사상을 예측하는 것은 무척 어렵다. 더욱이 예측의 정확도는 시계열 자료의 잡음들로 인해 떨어지게 되고 시간축이 증가할수록 그 복잡성 또한 증가한다. 이러한 복잡한 비선형적인 현상을 전통적 통계적 방법들로 풀기에는 한계가 있었기 때문에 신경망은 수십 년 동안 수문자료 예측을 위해 광범위하게 사용되어온 가장 대표적인 자료 주도형 모형이었다. 그러나 신경망은 뛰어난 성능에도 불구하고 앞서 언급한 치명적인 단점들로 인해 이러한 단점을 보완할 새로운 모형들이 모색 되었다.

이러한 시도들의 일환으로 미국토목학회(ASCE Task Committee on Application of Neural Network in Hydrology, 2000)는 하천 유량을 예측하는데 통계적 학습 방법에 근간한 새로운 자료 주도형 이론인 SVM(Support Vector Machine)을 소개한 바 있다. Dibike 등(2001)은 SVM을 원격 관측된 영상분류와 회기문제(강우-유출 모형)를 해결하는데 적용한 바 있다, 세 유역에 대하여 SVM과 신경망 그리고 개념적 강우-유출 모형을 비교하였고 SVM이 가장 좋은 결과를 보여 주었다. 또 Liong and Sivapragasam(2002)는 홍수위를 예측에 SVM을 적용하여 신경망에 비해 SVM의 성능이 더 우수함을 보인 바 있다. 이 논문에서는 하류 한 지점의 홍수위를 예측하기 위해 상류 여러 지점에서 홍수위가 입력자료로 사용되었다. Asefa and Kemblowski(2002)는 Monte-Carlo 방법에 기초한 지하수 거동과 초기 지하수 오염물질 검출 모니터링 시스템에 활용하기 위해 이송모형을 재현하는데 SVM을 적용한 바 있다.

3. Support Vector machine

3.1 Support Vector machine 소개

Support Vector Machine(SVM)의 개발은 주로 Vapnik과 그의 동료들(Vapnik, 1995, 1998)에 의해 이루어져 왔고, 수많은 매력적인 특징들과 경험적으로 볼 때 신뢰성 높은 성능 때문에 많은 인기를 얻어왔다. SVM 이론은 SRM(Structural Risk Minimization)에 근간을 두고 있으며, 이는 다른 모형화 기법에서 사용되어온 전통적인 ERM(Empirical Risk Minimization)보다 다양한 면에서 우수함을 보여주고 있다(Osuna et al. 1997; Gunn, 1998). SVM 모형의 장점은 통계적 이론의 목표인 일반화 능력이 매우 우수하다는 것이다.

3.2 Support Vector Regression

신경망에 의한 회귀방법과 비교하여 볼 때, SVM은 회귀함수를 추정하기 위해 세 가지 두드러진 특징을 가지고 있다. 첫 번째, SVM은 회귀식을 추정하기 위해 고차원 공간에서 정의된 선형 함수의 조합을 사용한다는 점이다. 두 번째, SVM은 Vapnik의 ϵ-insensitive 손실 함수를 사용하여 오차(risk)를 최소화하여 회귀추정을 수행한다는 것이다. 세 번째, SVM은 SRM(Structural Risk Minimization)으로부터 유도된 일반화 항과 경험적인 오차로 구성된 위험함수(risk function)를 사용한다. 본 논문에서는 LS-SVM(K. Pelckmans et al., 2003)을 적용하였고 기본 이론은 다음과 같다.

주어진 자료 집합 $G=\{(x_i,d_i)\}_i^n$ (x_i는 입력 벡터, d_i는 목표값 그리고 n은 전체 자료사상의 수)일 때 SVM은 다음의 퍼셉트론(Perceptron) 함수를 사용하여 근사화될 수 있다.

$$y = f(x) = w\phi(x) + b, \qquad (1)$$

여기서, $\phi(x)$는 입력공간 x로부터 비선형적으로 사상된 고차원의 특징공간(feature space)이다.

그리고, 계수 w와 b는 최소화 방법에 의해 추정될 수 있다.

$$R_{SVM}(C) = C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(d_i, y_i) + \frac{1}{2}\|w\|^2, \qquad (2)$$

$$L_\varepsilon(d,y) = \begin{cases} |d-y| - \varepsilon & |d-y| \ge \varepsilon \\ 0 & otherwise \end{cases}. \qquad (3)$$

(2)식에서 주어진 일반화된 위험함수에서, 첫 번째 항 $C(1/n)\sum_{i=1}^{n} L_\varepsilon(d_i, y_i)$은 경험적인 오차(error or risk)이고 이 오차는 (3)식에서 주어진 ϵ-insensitive 손실함수에 의하여 측정된다. 두 번째 항 $1/2\|w\|^2$은 일반화 항이다. C는 일반화 상수를 의미하며 경험적인 오차와 일반화 항 사이에서 절충(trade-off)점을 결정하고 C값이 증가하면 일반화항에 대한 경험오차의 상대적인 비중이 증가하게 된다. ε은 튜브의 크기를 말하며 학습 자료가 위치한 지점들에서의 근사 정확도를 결정한다.

w와 b의 추정치를 계산하기 위해서는, (2)식은 (4)식과 같은 최적화 문제(primal function) 형태로 변환되어야 한다.

$$\text{Minimize} \quad R_{SVM}(w, \xi^{(*)}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$\text{Subject to} \quad \begin{aligned} & d_i - w\phi(x_i) - b_i \le \varepsilon + \xi_i, \\ & w\phi(x_i) + b_i - d_i \le \varepsilon + \xi_i^*, \quad \xi_i^* \ge 0. \end{aligned} \qquad (4)$$

결론적으로, 라그랑지 승산자(Lagrange multipliers)와 최적 제약조건(optimality constraint)을 이용하여, (1)식에 주어진 결정함수(decision function)는 다음과 같은 명시적 형태를 가지게 된다.

$$f(x, a_i, a_i^*) = \sum_{i=1}^{n}(a_i - a_i^*)K(x, x_i) + b. \qquad (5)$$

여기서 $K(x_i, x_j)$는 커널함수(kernel function)로 정의되며, 커널값은 특징공간 $\phi(x_i)$과 $\phi(x_j)$에서의 두 벡터 X_i와 X_j의 내적(inner product)과 같으며 다음과 같이 표현된다.

$$K(x_i, x_j) = \phi(x_i) * \phi(x_j) \qquad (6)$$

변환함수 $\phi(x)$는 x에 대한 새로운 특징을 추출하는 것이라 할 수 있는데, 실제로는 구체적인 형태를 알 필요 없이 별도로 정의되는 커널함수 (kernel function)를 활용한다.

4. 적용 결과

4.1 예측 오차 산정에 의한 모형의 검증 방법

서로 다른 모형에 의해 산정된 예측의 정확도를 객관적으로 평가하기 위해서는 적절한 오차 산정법이 필요하다. 본 논문에서는 예측의 정확도를 수치적으로 평가, 비교하고자 (7)식과 같은 시계열자료의 예측오차 계산에 가장 일반적으로 쓰이는 다음의 네 가지 오차 산정 방법을 사용하였다.

$$AMB = \frac{1}{M}\sum_{m=1}^{M}|\hat{z}(x_m) - z(x_m)| \qquad RMSE = \sqrt{\frac{1}{m}\sum_{m=1}^{M}[\hat{z}(x_m) - z(x_m)]^2} \qquad (7)$$

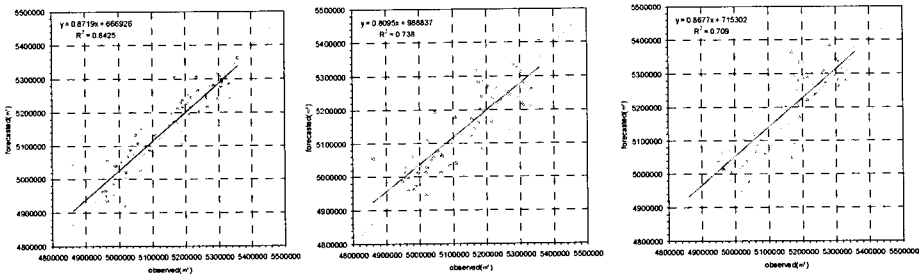
$$RRMSE = \sqrt{\frac{\sum_{m=1}^{M}[\hat{z}(x_m) - z(x_m)]^2}{\sum_{m=1}^{M} z(x_m)^2}} \qquad MAPE = \frac{1}{M}\sum_{m=1}^{M}\frac{|\hat{z}(x_m) - z(x_m)|}{z(x_m)}$$

AMB : Absolute Mean Bias
RMSE : Root Mean Square Error
RRMSE : Relative Root Mean Square Error
MAPE : Mean Absolute Percentage Error

4.2 일급수량 예측

한국의 경우 1년 중 고온인 7, 8월의 경우에 기온과 용수수요가 높은 상관관계를 보인다. 그리고 특히 이시기에 관심을 두는 이유는 용수사용량이 가장 많고 하절기인 관계로 정확한 용수수요량의 예측이 중요시되기 때문이다. 그래서 이 시기의 일별 용수수요량에 대한 정확한 예측은 도시의 급수시스템을 운영하는데 필수조건 이라 하겠다.

따라서 본 연구에서는 1992년~1996년 7월과 8월의 서울시 일급수량 자료를 이용하여 일 예측모형을 구현하여 보았다. 1992년~1995년 자료는 모형의 학습을 위해 사용하였고 1996년 자료는 모형의 검증을 위해 사용하였다. 일급수량 자료 외에 기온, 요일(평일 or 휴일), 일조시간을 입력 자료로 함께 사용하였다. 그림 1은 SVM(Support Vector Machine)과 NNGA(Neural Network using Genetic Algorithm), 그리고 NNBP(Back-Propagation Neural Network) 모형으로 예측된 결과와 실측치를 비교한 그림이다. 그림 1에서 보듯이 SVM의 상관도가 NNGA나 NNBP보다 높고 기울기도 가장 1에 근사함을 알 수 있다.



(a) SVM (b) NNGA (c) NNBP

그림 1. 실측 vs 예측 일급수량 자료

표 1은 세 모형의 예측오차를 비교한 결과이다. 오차비교 결과에서도 SVM이 다른 모형보다 정확도가 높음을 알 수 있다. 그림 2는 1996년 7월과 8월의 서울시 실측 일급수량 자료와 SVM에 의해 예측된 일급수량자료이다. 현저하게 이상거동을 보이는 실측구간을 제외하면 저수요 부분과 고수요 부분 전체적으로 비교적 정확히 예측하고 있음을 알 수 있다.

표 1. 일급수량 예측 오차 비교

Error	SVM	NNGA	NNBP
AMB	37107.3	50168.5	61840.7
RMSE	50550.6	65587.8	79755.6
RRMSE	0.0098	0.0128	0.0155
MAPE(%)	0.726	0.976	1.209
CC	0.918	0.859	0.842

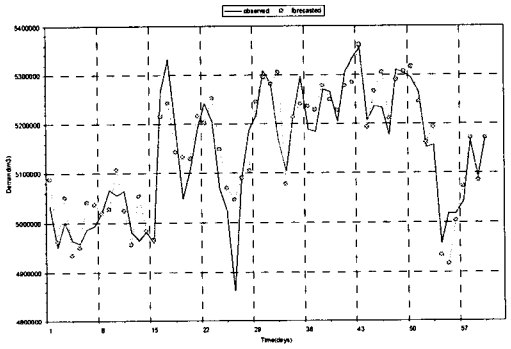


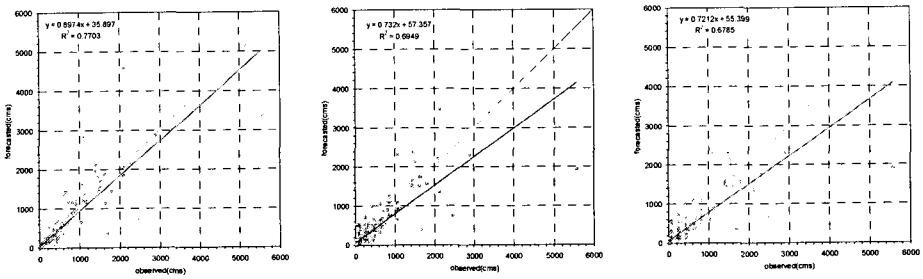
그림 2. SVM에 의한 일급수량 예측결과(1996년 7월~8월)

4.3 충주댐 일유입량 예측

댐으로의 유입량 자료는 댐의 효율적인 운영을 위해 매우 중요한 자료이다. 특히 일유입량 자료는 치수적인 측면뿐만 아니라 이수적인 측면에서도 매우 중요하다. 이러한 이유로 댐의 일유입량 자료를 정확히 예측하는 것은 댐의 안정적이고 효율적인 운영을 위해서 필수적이다. 그러나 실제로 기존의 통계적인 방법으로 댐의 일유입량을 저유입량에서 고유입량까지 전반적으로 정확도 높게 예측하는 것은 매우 어렵고 더욱이 예측의 정확도를 안정적으로 유지하기는 거의 불가능하다.

이러한 이유로, 본 논문에서는 충주댐 유역의 일유입량을 예측하기 위해 1986년~2003년 충주댐 일유입량 자료와 강우량 자료를 사용하여 SVM 일 예측모형을 구축하여 보았다. 이중 1986년~2002년 자료는

모형의 학습을 위해 사용하였고 2003년 자료는 모형의 검증을 위해 사용하였다. 일유입량의 경우는 4.2절의 일급수량 자료와는 다르게 NNGA모형 대신 MLR(Multiple Linear Regression) 모형을 사용하였다. 이는 본 모형에서 유입량의 예측에 사용되는 입력 자료가 비교적 단순하고, NNGA 모형의 경우 자료의 양이 많으면 예측시간이 현저히 증가하여 효율성이 떨어지는 단점이 있기 때문이다. 다음의 그림 3은 SVM(Support Vector Machine), NNBP(Back-Propagation Neural Network) 그리고 MLR(Multiple Linear Regression) 모형으로 예측된 결과와 실측치를 비교한 그림이다. 그림에서 보듯이 SVM의 상관도가 NNBP나 MLR보다 월등히 높음을 알 수 있다. 표 2는 세 모형의 예측오차를 비교한 결과이다. 오차비교 결과에서도 SVM이 다른 모형보다 높은 정확도를 보여주고 있다. 그림 3은 2003년 충주댐 실측 일유입량과 SVM에 의해 예측된 충주댐 일유입량이다. 저유입량에서 고유입량까지 비교적 정확하게 예측하고 있다는 점에서 모형의 예측 신뢰도와 안정성이 높고 적용성이 뛰어남을 보여준다.



(a) SVM (b) NNBP (c) MLR

그림 3. 실측 vs 예측 충주댐 일유입량 자료

표 2. 일유입량 예측 오차 비교

	SVM	NNBP	MLR
AMB	74.6	106.3	113.4
RMSE	254.0	281.2	289.2
RRMSE	0.444	0.492	0.506
MAPE(%)	35.9	99.7	104.1
CC	0.878	0.834	0.824

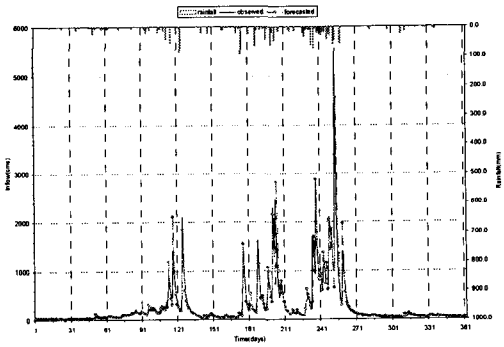


그림 4. SVM에 의한 충주댐 일유입량 예측결과(2003년)

5. 결 론

본 연구에서는 서울시 일급수량과 충주댐 일유입량 수문시계열 자료를 이용하여 기존의 전통적인 통계학적 모형과 신경망 모형, 그리고 수문학 분야에서는 아직까지 적용된 사례가 매우적은 SVM 예측모형을 구축하여 결과 비교를 통해 모형의 예측 정확도를 비교 평가하였다. 전형적인 오차 산정 방법에 의해 두 가지 경우의 수문시계열 예측 오차를 산정하여 비교해 볼 때 SVM 모형이 수문시계열 자료 예측에 있어서 기존의 방법들에 비해 안정적이고 정확한 예측 결과를 보여 주었다.

참 고 문 헌

1. Dibike, Y. B., Velickov, S., Solomatine, D. P. & Abbott, M. B. 2001 Model induction with support vector machines: introduction and applications. J. Comput. Civil Engng. ASCE 15 (3), 208-216.
2. Liong, S. Y. & Sivapragasm, C. 2002 Flood stage forecasting with SVM. J. Am. Wat. Res. Assoc. 38 (1), 173-186.
3. Matterra, D. & Haykin, S. 1999 Support vector machines for dynamic reconstruction of a chaotic system. In Advances in Kernel Methods (ed. Scho¨ lkopf, B., Burges, C. J. C. & Smola, A. J.), pp. 211-241. MIT Press, Cambridge, MA.
4. Vapnik, V. N. 1995 The Nature of Statistical Learning Theory. Springer Verlag, New York.