



See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/321264401>

(In Korean) '수준 수 준 수 수: 수 수 수 수 수 수' (Water Level Forecasting based on Deep Learning: A Use Case of Trinity River-Texas-The United States)

Article · June 2017

DOI: 10.5626/JOK.2017.44.6.607

CITATIONS

0

READS

306

2 authors:



[Quang-Khai Tran](#)

Korea Institute of Science and Technology Information (...)

6 PUBLICATIONS **0** CITATIONS

[SEE PROFILE](#)



[Sa-kwang Song](#)

Korea Institute of Science and Technology Information (...)

82 PUBLICATIONS **221** CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Matrix Completion for Predicting Hurricane Damage Levels [View project](#)



Semantic Web [View project](#)

딥러닝 기반 침수 수위 예측: 미국 텍사스 트리니티강 사례연구

(Water Level Forecasting based on Deep Learning: A Use Case of Trinity River-Texas-The United States)

트란 광 카이[†]
(Quang-Khai Tran)

송 사 광^{††}
(Sa-kwang Song)

요약 도시에서 홍수 피해를 방지하기 위한 침수를 예측하기 위해 본 논문에서는 딥러닝(Deep Learning) 기법을 적용한다. 딥러닝 기법 중 시계열 데이터 분석에 적합한 Recurrent Neural Networks (RNNs)을 활용하여 강의 수위 관측 데이터를 학습하고 침수 가능성을 예측하였다. 예측 정확도 검증에 위해 사용한 데이터는 미국의 트리니티강의 데이터로, 학습을 위해 2013년부터 2015년까지 데이터를 사용하였고 평가 데이터로는 2016년 데이터를 사용하였다. 입력은 16개의 레코드로 구성된 15분단위의 시계열 데이터를 사용하였고, 출력으로는 30분과 60분 후의 강의 수위 예측 정보이다. 실험에 사용한 딥러닝 모델들은 표준 RNN, RNN-BPTT(Back Propagation Through Time), LSTM(Long Short-Term Memory)을 사용했는데, 그 중 LSTM의 NE(Nash Efficiency)가 0.98을 넘는 정확도로 기존 연구에 비해 매우 높은 성능 향상을 보였고, 표준 RNN과 RNN-BPTT에 비해서도 좋은 성능을 보였다.

키워드: 침수 예측, 딥러닝, RNN, BPTT, LSTM

Abstract This paper presents an attempt to apply Deep Learning technology to solve the problem of forecasting floods in urban areas. We employ Recurrent Neural Networks (RNNs), which are suitable for analyzing time series data, to learn observed data of river water and to predict the water level. To test the model, we use water observation data of a station in the Trinity river, Texas, the U.S., with data from 2013 to 2015 for training and data in 2016 for testing. Input of the neural networks is a 16-record-length sequence of 15-minute-interval time-series data, and output is the predicted value of the water level at the next 30 minutes and 60 minutes. In the experiment, we compare three Deep Learning models including standard RNN, RNN trained with Back Propagation Through Time (RNN-BPTT), and Long Short-Term Memory (LSTM). The prediction quality of LSTM can obtain Nash Efficiency exceeding 0.98, while the standard RNN and RNN-BPTT also provide very high accuracy.

Keywords: water level forecasting, deep learning, recurrent neural networks, back propagation through time, long short-term memory

· 이 논문은 제43회 동계학술발표회에서 '딥러닝 기반 침수 수위 예측: 미국 텍사스 트리니티강 사례연구'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 과학기술연합대학원대학교 빅데이터과학과 박사과정,
한국과학기술정보연구원 의사결정지원기술연구실
학생연구원
khai.tran@kisti.re.kr

†† 정회원 : 과학기술연합대학원대학교 빅데이터과학과 부교수,
한국과학기술정보연구원 의사결정지원기술연구실 실장
(KISTI)
esmallj@kisti.re.kr
(Corresponding author)

논문접수 : 2017년 1월 18일
(Received 18 January 2017)
논문수정 : 2017년 3월 10일
(Revised 10 March 2017)
심사완료 : 2017년 3월 21일
(Accepted 21 March 2017)

Copyright©2017 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.
정보과학회논문지 제44권 제6호(2017. 6)

1. 서론

최근 텍스트나 이미지, 비디오 등의 데이터 분석에 딥러닝 기법을 적용하여 효과적인 성능향상을 본 연구가 많이 있다[9]. 하지만 이러한 딥러닝 기술은 홍수 예보 등과 같은 강의 수위를 예측하는 문제에는 적용된 사례를 찾기가 어렵다. 일부 도시 침수를 예측하기 위한 수치예측 모델들이 연구되었지만 시계열 데이터를 대상으로 하는 연구는 예측 모델의 연산 복잡성으로 인해, 실시간으로 대응하기는 어려운 문제가 존재한다.

딥러닝 기술이 주목받기 전에, 수위 예측[1,5] 및 강우로 인한 홍수 관리[6]를 위해 인공 신경망(Artificial Neural Networks - ANNs), Neuron-Fuzzy 및 Support Vector Machines과 같은 머신러닝 기법이 사용되어왔다. 관련 연구[5]에서는 일일 관측 자료에 기반한 24시간 빈도의 강의 유량 예측은 0.822에서 0.887의 Nash Efficiency(NE, 부록)로 예측되었다. [2]에서는 이와 같은 문제를 해결하기 위해 ANN과 Dynamic ANN를 사용한다. 또한, 해당 연구에서는 ANN과 두 개의 Elman 신경망을 결합하여 6단계 수위 예측을 수행하여 10분 단위의 예측에서 NE 지표로 약 0.93으로 가장 좋은 결과를 얻었고, 30분 및 60분 단위의 예측에서는 각각 약 0.72 및 0.67로 가장 좋은 결과를 얻었다. 기존 연구들은 딥러닝과는 다른 방법을 사용하였거나, 딥러닝 방법을 이용할 경우 높은 정확도를 보여주지 못했다(특히 한 시간 앞의 수위를 실시간으로 예측하는 경우). 따라서, 정확하고 신속하게 예측을 수행하는 딥러닝 기법을 이용하는 것이 바람직하다.

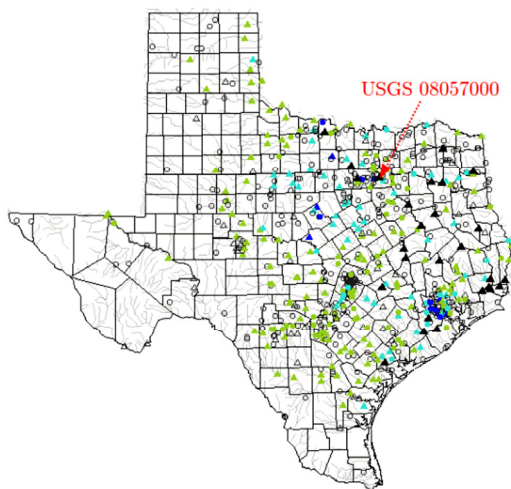


그림 1 연구 장소: 검은 색 삼각형(Texas 지도[13])

Fig. 1 Studied location: the pointed black triangle (Texas map, downloaded from [13])

본 논문에서는 강 수위의 실시간 예측 문제를 해결하기 위해, 시계열 데이터 분석에 강점을 가진 Recurrent Neural Networks(RNNs)[10]를 사용한다. 제안하는 방법의 유용성을 입증하기 위해, 일반 RNN과 Back Propagation Through Time(BPTT)[11]으로 학습하는 RNN(RNN-BPTT) 및 최근 Long Short-Term Memory(LSTM)[3,4,7]의 정확도를 비교한다. 이를 위해, 그림 1의 미국 텍사스 트리니티강(Trinity River - Texas - the U.S)에 있는 수위 관측소(ID: 'USGS 08057000') 사례를 고찰한다[13]. 해당 지역은 침수가 자주 발생하고, 상세한 수치 정보를 제공하기 때문에 제안하는 방법의 성능을 평가하기에 적합하다. 실험 데이터는 15분마다 기록되었으며 각 레코드는 세 종류 - 하천 유량(discharge (feet³/s)), 강수량(precipitation(inches), 수위(water level(feet)) -의 변수들을 가진다. 예측 결과는 예측 모델을 통해 생성된 30분 및 60분 후의 강 수위이다.

이 연구의 주요 공헌 중 하나는 홍수 예측을 목표로 하천 수위 예측을 위해 딥러닝 기술을 도입한 것이다. 다음 섹션에서는 모델을 설계하고 입력과 출력 데이터를 구성하는 방법을 보여주고 실험 결과를 논의한다. 또한, RNN 모델은 문헌 [3,4,7,8]에서와 같이 Vanishing Gradient 문제를 가지고 있는데, LSTM은 이 문제에 대한 좋은 해결책으로 사용되고 있다. 본 연구에서는 침수 수위 예측에서 다양한 RNNs 계열의 딥러닝 방법론이 어떤 효과를 보이는지 검증하고자 한다.

2. 예측 모델

표준 RNN 모델은 그림 2와 같다. 이때, 시간 't'에 있는 활성화 함수들은 다음과 같다:

- 은닉층 (hidden layer):

$$s_t = \tanh(Ux_t + Ws_{t-1} + bias)$$

- 출력층 (output layer):

$$o_t = \text{sigmoid}(Vs_t + bias)$$

'x'는 입력 변수이고 'y'는 출력 변수이며 's'는 은닉

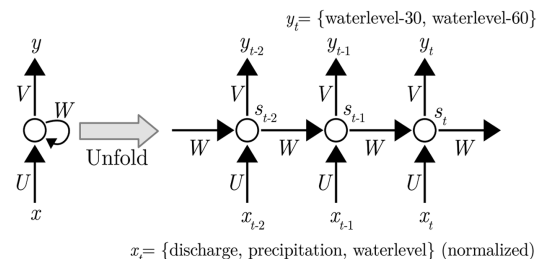


그림 2 표준 Recurrent Neural Network 모델[9]

Fig. 2 Standard RNN model adapted from [9]

층의 상태를 의미한다. 그리고 ‘ U ’, ‘ W ’과 ‘ V ’ 매개변수들은 네트워크의 가중치(‘weights’)들이다. 이 함수들에서는 편차(bias)도 이용하고 이 가중치들과 편차는 임의로 초기화된다. 그런 다음 통계적 경사 하강법(Stochastic Gradient Descent)을 이용한 역전파(Back-propagation) 알고리즘을 사용하여 이러한 가중치를 학습한다. 각 반복 시점에 출력 값과 정답과의 차이인 오류 값은 출력층으로부터 은닉층과 이전 시간 단계로 역방향 전송된다. 이보다 조금 더 진보된 RNN-BPTT는 오류 값을 모든 이전 시간 단계로 역으로 보낸다.

문헌[4]에 LSTM의 많은 변형을 확인할 수 있지만 이 연구에서는 Hochreiter[7]가 소개한 표준 버전을 사용한다. 이 구조에서는 위의 RNN 구조의 은닉층이 메모리 셀 연산으로 대체되며(자세한 내용은 [7] 참조) 출력층은 동일하게 유지된다. 본 연구에서 사용한 LSTM 메모리 셀 연산과 활성화 함수들은 다음과 같다:

$$\begin{aligned} i_t &= \sigma(U_i x_t + W_i s_{t-1} + bias_i) && \text{input gate} \\ f_t &= \sigma(U_f x_t + W_f s_{t-1} + bias_f) && \text{forget gate} \\ o_t &= \sigma(U_o x_t + W_o s_{t-1} + bias_o) && \text{output gate} \\ g_t &= \sigma(U_g x_t + W_g s_{t-1} + bias_g) && \text{block input} \\ c_t &= c_{t-1} * f_t + g_t * i_t && \text{cell state} \\ s_t &= \tanh(c_t) * o_t && \text{block output} \end{aligned}$$

(σ : sigmoid 함수; *: elementwise 곱셈)

3. 실험 및 평가

3.1 실험 설정

수집된 데이터는 다음과 같다. 각 입력 데이터는 15분 간격으로 입력되는 4시간 동안의 16 레코드를 사용한다. 이 시계열 데이터는 그리 길지 않아 RNN-BPTT와 LSTM의 Truncate-step을 1로 설정하였다. 학습 자료는 2013년 5월부터 2015년 12월까지 데이터(91,300레코드: 87.3%)이고 평가 자료는 2016년 1월부터 5월까지 데이터(13,292레코드: 12.7%)이다. 예상되는 값의 범위는 22.16ft에서 35.94ft까지이다. 모든 데이터들은 정규화(normalized)하였고 반복 실험을 통해 조정된 네트워크 설정은 다음과 같다: (1) 은닉층은 6개 노드들로 구성되었고, (2) 학습을 위한 반복 횟수(epochs)는 200이다. (3) 학습율(Learning Rate)은 0.05를 사용하였다. 이 학습율은 학습과정(each epoch)에서 평균 제공된 오차(Root Mean Squared Error(RMSE))로 표현되는 학습 오류가 증가할 때마다 반으로 감소된다.

그리고 예측 실험을 위한 프로그램은 Python-2.7과 Theano 딥러닝 라이브러리[12]로 구현하였고 실험 속도 향상을 위한 가속기로는 GeForce GTX 980M-GPU를 사용하였다.

3.2 결과 및 토론

실험시 학습에 걸린 시간은 표준 RNN이 가장 빨랐고 예상과 같이 LSTM이 가장 느린 결과를 얻을 수 있었다(표 1). 하지만 학습시간과 달리 실제 예측은 지연 없이 바로바로 결과를 얻을 수 있다. 그림 3은 학습과정에서 반복(iteration)이 5 epoch 이후에 진행됨에 따른 오류값을 RMSE으로 표시한 그래프인데, LSTM이 다른 두 알고리즘에 비해 늦게 수렴해 가지만, 결과적으로 가장 좋은 검증(Validation) 정확도를 보여주고 있다 (0.0에서 1.0까지의 정규화 된 값을 예측할 때 RMSE = 0.0045로 최저 오류를 보임(표 1)). RNN-BPTT는 두번째 성능을 보이지만 표준 RNN은 상대적으로 낮은 성능을 보인다. 이 결과는 학습 과정에서 표준 RNN이 RNN-BPTT 및 LSTM과 같은 다른 개선된 방법보다 더 일찍 수렴하지만 전역 최적해(global optimum)보다는 지역 최적해(local optimum)에 빠지는 현상을 보인다. 또한, 이러한 연구 맥락에서 LSTM이 더 나은 최적을 찾기에 더 효율적이라고 보인다. 이러한 현상은 다른 문헌의 딥러닝 연구 결과와 유사하다.

예측 능력의 관점에서 위 세 모델들은 모두 30분(Predicted-30) 후의 예측과 60분(Predicted-60) 후 예측이 0.98 이상의 NE를 보여 잘 예측한다고 할 수 있다. 또한, 이 결과는 [2]와 [5]의 결과보다 더 정확하게

표 1 학습 시간별 검증 오류 및 예측 오류

Table 1 Training time, validation errors and prediction errors

		RNN	RNN-BPTT	LSTM
Training	Training time (hours)	~14	~22	~67
	Validation (RMSE-normalized)	0.0076	0.0064	0.0045
Testing	One prediction (milliseconds)	~1.6	~1.6	~2.8
	RMSE-30 minutes (normalized)	0.0102	0.0091	0.0077
	RMSE-60 minutes (normalized)	0.0105	0.0092	0.0079
	RMSE-30 minutes (feet)	0.4289	0.3839	0.3246
	RMSE-60 minutes (feet)	0.4420	0.3845	0.3317
	NE-30 minutes	0.9831	0.9865	0.9903
	NE-60 minutes	0.9820	0.9864	0.9898

예측하였다. 다만, 30분 예측이 60분 예측보다 더 정확한 것을 보여주는 것은 일반적인 시계열 예측의 현상과 유사하다고 할 수 있다. 세 알고리즘 중 LSTM은 30분 또는 60분의 시간에 상관없이 가장 잘 학습할 뿐만 아니라 예측 능력도 가장 좋은 결과를 보인다. 특히, 그림 4,5,6에서 볼 수 있듯이, 30 feet 이상(점선:범람)의 침수 상황에서도 아주 잘 예측하는 것을 확인할 수 있다. 30분 예측의 RMSE 범위는 0.4289ft에서 0.3246ft까지이고 60분 예측의 RMSE 범위가 0.4420ft에서 0.3317ft까지 이므로, 세 모델들의 예측 가능성은 매우 유효하다(표 1). 또한 그림 7,8,9 산포도를 통해서도 각 모델별 예측결과와 실측값 간에 매우 큰 연관도를 볼 수 있다.

LSTM이 학습과 예측 모두에서 단순 RNN과 RNN-BPTT를 능가하는 이유는 VG(Vanishing Gradient) 문제 해결 능력 여부에 달린 것으로 보인다. 비록 이 연구에서 사용된 서열 길이가 상당히 길지는 않지만(16 레코드) VG 문제는 여전히 전통적인 RNN 모델의 성능향상에 방해가 되는 것으로 보인다. LSTM에는 시퀀스의 장기 종속성을 연결할 수 있으므로 VG 문제는 학습 시에 조금 줄어든다. 향후 연구에서는 더 긴 입력 데이터 시퀀스와 예측 시간의 단계를 증가시켜 이 문제를 더 자세히 연구할 것이다. 이 상황에서 그래디언트가 급증하거나 유실될 수 있다. 이를 통해 LSTM이 VG의 영향을 크게 줄일 수 있는지 확인할 수 있을 것이다.

제다가 모든 알고리즘은 하나의 예측(순방향 전파)(표 1)을 수행하는데 단지 수 밀리 초가 필요하다는 점에 유의해야 한다. LSTM은 학습 과정과 마찬가지로 복잡한 구조로 인해 예측을 수행하는데 시간이 많이 걸린다. 그러나 이러한 상황(~2.8 밀리 초)에서도 제안된 모델은 예측된 값을 거의 즉시 제공할 수 있다. 따라서 이 모델은 실시간 상황에서 수위 예측에 사용될 수 있다.

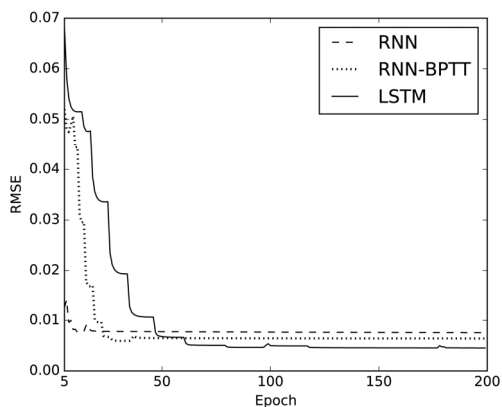


그림 3 세 알고리즘별 수렴 과정(5 epoch 이후)
Fig. 3 Convergence process(after the 5th epoch)

4. 결론 및 향후 연구

본 연구에서는 최근 좋은 성능을 보이는 표준 RNN, RNN-BPTT, LSTM 등의 다양한 딥러닝 알고리즘 사용하여 시계열 침수센서 데이터를 분석하여 높은 예측 정확도를 얻을 수 있었다. 이러한 3가지 종류의 딥러닝 네트워크를 활용하여 미래 수위 패턴을 학습 및 예측한 결과, NE값이 0.98 이상의 높은 값을 얻었고, 30분 수위예측에서 60분 예측보다 높은 정확도를 얻을 수 있었다. 특히, 시계열 데이터에 좋은 성능을 보이는 LSTM 알고리즘이 미래의 수위 예측에서도 가장 좋은 예측 정확도를 보였다. RNN-BPTT와 표준 RNN이 뒤를 이었다. 이러한 모델들의 학습 데이터의 양이나 학습 설정에 따라 학습 시간이 상대적으로 길지만, 실제 적용을 위한 예측에서는 거의 실시간으로 침수 예측이 가능하기 때문에 실제 예측 활용도 가능한 결과이다.

향후 연구로는, 좀 더 빠른 수위 예측을 위해 기상 예측 정보를 활용하여 30분, 60분 예측을 넘어 3시간, 6시간, 더 나아가 24시간, 48시간 등의 긴 시간 예측을 수

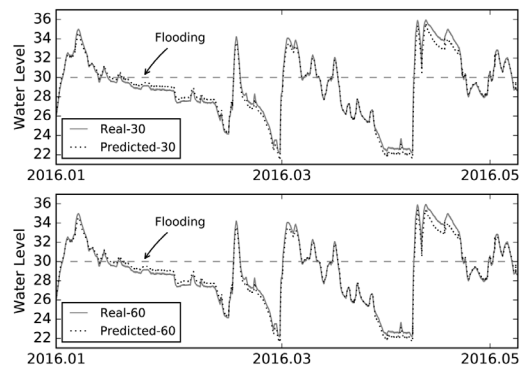


그림 4 표준 RNN의 예측한 패턴
Fig. 4 Predicted patterns of standard RNN

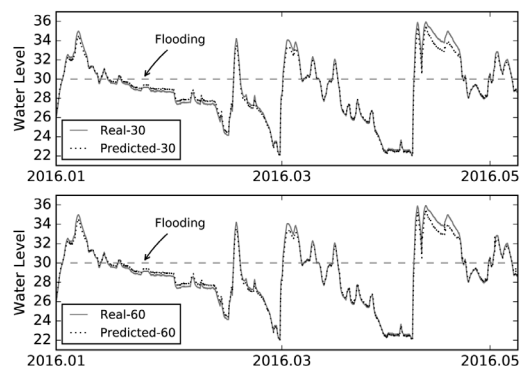


그림 5 RNN-BPTT의 예측한 패턴
Fig. 5 Predicted patterns of RNN-BPTT

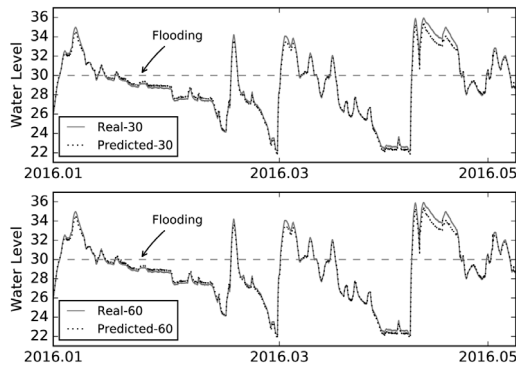


그림 6 LSTM의 예측한 패턴
Fig. 6 Predicted patterns of LSTM

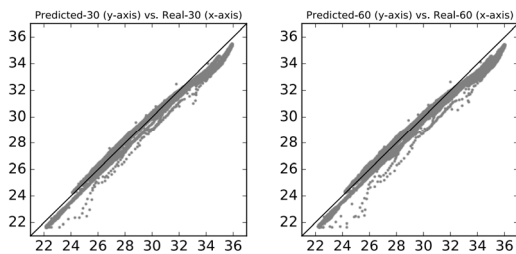


그림 7 표준 RNN의 예측결과의 산포도
Fig. 7 Scatter plot of values predicted by standard RNN

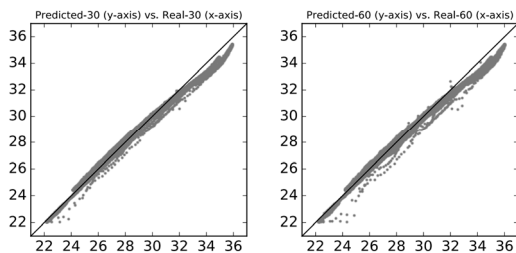


그림 8 RNN-BPTT의 예측결과의 산포도
Fig. 8 Scatter plot of values predicted by RNN-BPTT

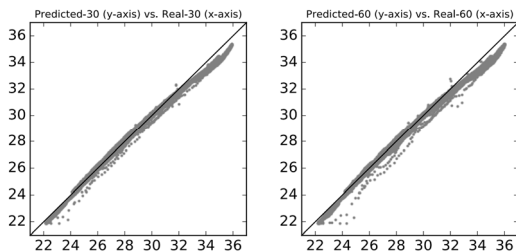


그림 9 LSTM의 예측결과의 산포도
Fig. 9 Scatter plot of values predicted by LSTM

행하고자 한다. 이를 위해서 위성영상, 레이더 영상 등을 활용하여 예측 정확도를 높이하고자 한다.

References

- [1] B. Bazartseren, G. Hildebrandt, K. P. Holz, "Short-term water level prediction using neural networks and neuro-fuzzy approach," *Neurocomputing*, Vol. 55, No. 3, pp. 439-450, 2003.
- [2] F. J. Chang, P. A. Chen, Y. R. Lu, E. Huang, K. Y. Chang, "Real-time Multi-Step-Ahead Water Level Forecasting by Recurrent Neural Networks for Urban Flood Control," *Journal of Hydrology*, Vol. 517, pp. 836-846, 2014.
- [3] A. Graves, J. Schmidhuber, "Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures," *Neural Networks*, Vol. 18, No. 5, pp. 602-610, 2005.
- [4] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, J. Schmidhuber, "LSTM: A Search Space Odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, Vol. PP, Issue 99, 2016.
- [5] Z. He, X. Wen, H. Liu, J. Du, "A Comparative Study of Artificial Neural Network, Adaptive Neuro Fuzzy Inference System and Support Vector Machine for Forecasting River Flow in the Semiarid Mountain Region," *Journal of Hydrology*, Vol. 509, pp. 379-386, 2014.
- [6] N.Q. Hung, M.S. Babel, S. Weesakul, N.K. Tripathi, "An Artificial Neural Network Model for Rainfall Forecasting in Bangkok, Thailand," *Hydrology and Earth System Sciences*, Vol. 13, No. 8, pp. 1413-1425, 2009.
- [7] S. Hochreiter, J. Schmidhuber, "Long Short Term Memory," *Neural Computation*, Vol. 9, No. 8, pp. 1735-1780, 1997.
- [8] S. Hochreiter, "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 6, No. 02, pp. 107-116, 1998.
- [9] Y. LeCun, Y. Bengio, G. Hinton, "Deep Learning," *Nature*, Vol. 521, No. 7553, pp. 436-444, 2015.
- [10] D. E. Rumelhart, G. Hinton, R. J. Williams, "Learning Representations by Back-propagating Errors," *Cognitive Modeling*, Vol. 5, No. 3, p. 213-220, 1988.
- [11] P. J. Werbos, "Backpropagation Through Time: What It Does and How To Do It," *Proc. of the IEEE*, Vol. 78, No. 10, pp. 1550-1560, 1990.
- [12] Theano [Online]. Available: <http://deeplearning.net/software/theano/> (last accessed: Sep. 2016)
- [13] USGS-WaterWatch [Online]. Available: <http://water-watch.usgs.gov/index.php?m=flood&r=tx&w=map> (last accessed: Sep. 2016)

부 록

'Nash Efficiency'(NE) 또는 'Nash-Sutcliffe efficiency coefficient'는 수위 예측 성능을 측정하는 데 사용되며

다음과 같이 계산한다[2,5]:

$$NE = 1 - \frac{\sum_{i=1}^n (Q_i^o - Q_i^p)^2}{\sum_{i=1}^n (Q_i^o - \overline{Q^o})^2}$$

‘n’는 샘플 수이고 Q_i^o 와 Q_i^p 는 관측값과 예측값들이다. $\overline{Q^o}$ 는 관측 평균 값이다. NE 값이 1.0에 가까울 수록되면 예측 정확도가 높아진다.



트란 광 카이

2007년 Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City, Vietnam(학사). 2011년 Faculty of Computer Science and Engineering, University of Technology, Vietnam National University - Ho Chi Minh City, Vietnam(석사). 2011년~2013년 Faculty of Information Technology, Eastern International University(EIU), Binh Duong, Vietnam(강사) 2013년~2014년 Department of Water Management, Unesco-IHE, Delft, The Netherlands(박사과정 학생). 2015년~현재 과학기술연합대학원대학교 빅데이터과학과(박사과정 학생). 2015년~현재 한국과학기술정보연구원 의사결정지원기술연구실(학생연구원). 관심분야는 빅데이터, 인공지능, 딥러닝, 물정보과학, 시맨틱 웹



송 사 평

1997년 충남대학교 통계학과 졸업(학사) 1999년 충남대학교 대학원 컴퓨터과학과 졸업(석사). 2011년 KAIST 대학원 전산학과 졸업(박사). 1999년~2000년 한국전자통신연구원 정보검색연구팀 연구원. 2000년~2003년 (주)서치캐스트 정보검색팀 팀장. 2005년~2010년 한국전자통신연구원 바이오인포매틱스팀 선임연구원. 2010년~현재 한국과학기술정보연구원 의사결정지원기술연구실 실장. 2014년~현재 과학기술연합대학원대학교 빅데이터과학과 부교수. 관심분야는 빅데이터, 인공지능, 텍스트 마이닝, 자연어처리, 정보검색, 시맨틱 웹