클러스터링 알고리즘을 사용한 시계열 데이터 예측

Time Series Prediction using Clustering Algorithm

저자 (Authors)
김진현, 이창형, 심규석
Jinhyun Kim, Changhyung Lee, Kyuseok Shim

출처 (Source)
정보과학회논문지 : 컴퓨팅의 실제 및 레터 20(3), 2014.3, 191-195(5 pages)
Journal of KIISE : Computing Practices and Letters 20(3), 2014.3, 191-195(5 pages)

발행처 (Publisher)
한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE02373597

APA Style
김진현, 이창형, 심규석 (2014). 클러스터링 알고리즘을 사용한 시계열 데이터 예측. 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 20(3), 191-195

클러스터링 알고리즘을 사용한 시계열 데이터 예측

(Time Series Prediction using Clustering Algorithm)

김 진 현 † 이 창 형 ††
(Jinhyun Kim) (Changhyung Lee)
심 규 석 †††
(Kyuseok Shim)

요 약 하드웨어가 급속히 발전하고 SNS와 같이 사용자가 데이터를 생성하는 서비스가 늘어나며 다양한 분야에서 대규모의 시계열 데이터가 생성되고 있고 이들의 분석에 대한 요구가 커지고 있다. 본 논문에서는 다양한 어플리케이션에서 사용되는 시계열 데이터 예측을 위해 mRBF 함수를 사용하여 K-means 클러스터링 알고리즘을 변형한 시계열 데이터 클러스터링(clustering) 기술을 적용한 K-mRBF 모델을 제안한다. 실험에서는 실제 웹 서버 데이터 센터에서 수집된 데이터와 합성 데이터를 이용하여 제안한 시계열 데이터 예측 방식의 정확성을 평가하고 기존의 최신 연구 기법에 비해 나은 성능을 보임을 확인한다.

키워드: 시계열 데이터, 시계열 데이터 예측, 클러스터링

Abstract There is a wide range of applications such as social network services, sensor networks and data centers which generate time series data. Thus, analysis of such time series data has attracted a lot of attention in the recent years. In this paper, we propose a model called K-mRBF which utilizes a modified K-means clustering algorithm with the multivariate radial basis functions (mRBF) to predict future values based on previously observed values. We conduct extensive experiments using synthetic as well as real-life data sets to compare our K-mRBF model to the state-of-the-art model. Experimental results confirm the accuracy of our model compared to state-of-the-art models.

Keywords: time series, time series prediction, time series forecasting, clustering

1. 서 론

시계열 데이터(Time Series data)는 균등한 시간 간격에서 연속적으로 측정된 값들의 시퀀스이다[1]. 소셜 멀티미디어 제공 서비스인 유투브(YouTube)에서 동영상의 시간에 따른 시청수를 시계열 데이터로 표현할 수 있고 대규모 웹서버 클러스터에서 시간마다 측정되는 노드들의 상태 정보를 시계열 데이터로 볼 수 있다.

SNS와 같이 사용자가 데이터를 생성하는 서비스가 활발해지면서 다양한 분야에서 대규모의 시계열 데이터가 생성되고 있고 이에 대한 분석이 꾸준히 요구되고 있다. 특히 주어진 과거 시계열 데이터에 대해 특정 시간의 값을 예측하는 시계열 데이터 예측은 이 분야에서 중요한 문제로 많은 어플리케이션에서 유용하게 사용된다.

유투브의 특정 동영상에 대한 시청수를 시간에 따라 기록한 후 시계열 데이터 예측을 통해 앞으로의 시청수를 예측하고 많은 사람이 볼 것이라 예상되는 동영상에 대한 서버 최적화 작업을 미리 수행할 수 있다. 그리고 대규모 웹 서버 클러스터 환경에서 각 노드의 CPU 사용량 및 네트워크 트래픽 등을 측정해온 데이터를 통해 모델을 생성하고 앞으로의 값을 예측한 후 그에 맞추어 서버의 휴면 상태를 결정하거나 필요한 로드를 분산시킨다면 전체 산업의 에너지 소비에서 큰 비중을 차지하고 있는 데이터 센터의 서버를 유지 관리하는데 사용되는 에너지 소비를 줄일 수 있다[2]. 또는 태양열 발전에서 발전 기기의 시간당 발전량 데이터에 대해 시계열 데이터 예측을 수행한 후 예측값과 많이 다른 발전량을 보이는 발전기기를 고장이라고 판단하는데도 사용할 수 있다.

시계열 데이터 예측은 주로 학습 데이터를 이용해 모델을 학습하여 입력 데이터의 특정 시간의 값을 예측하는 방식을 사용한다. 기존에 많이 사용되어 온 선형 회귀 분석 방식은 앞선 시간의 데이터의 추세만을 고려하

· 본 연구는 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단-중견연구자지원사업의 지원을 받아 수행된 연구임(No. NRF-2009-0078828). 또한 서울대학교-한국전력공사-(주)케이디파워 스마트 에코 마이크로그리드 연구센터(SNU-KEPCO-KDPOWER Smart Eco Microgrid Research Center)지원으로 수행하였음
· 이 논문은 2013 한국컴퓨터종합학술대회에서 '클러스터링 알고리즘을 사용한 시계열 데이터 예측 모델 학습'의 제목으로 발표된 논문을 확장한 것임

† 비 회 원 : 서울대학교 전기정보공학부
jhkim@kdd.snu.ac.kr
†† 학생회원 : 서울대학교 전기정보공학부
chlee@kdd.snu.ac.kr
††† 정 회 원 : 서울대학교 전기정보공학부 교수
shim@kdd.snu.ac.kr
(Corresponding author임)

논문접수 : 2013년 10월 7일
심사완료 : 2013년 12월 18일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제20권 제3호(2014.3)

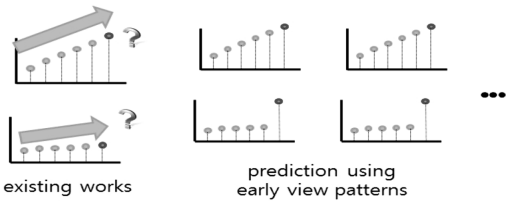

그림 1 시계열 데이터의 패턴을 고려한 예측 방식
Fig. 1 Prediction with the patterns of time series

였지만 최근에는 시계열 데이터들의 앞선 시간의 패턴을 고려하여 더 정확한 예측을 수행하는 모델 학습 방식이 사용되고 있다[3]. 그림 1의 왼쪽처럼 기존의 선형 회귀 분석 방식은 단순히 데이터의 추세만을 고려하여 예측하기 때문에 데이터가 각각 다른 추세를 보이는 그룹들로 구성되어 있다면 정확한 예측을 수행할 수 없다. 그림 1의 오른쪽에 나와 있는 두 가지 추세를 가지는 여러 개의 시계열 데이터 같은 경우 기존의 선형 회귀 분석은 정확한 예측을 수행할 수 없지만 미리 데이터에 숨겨진 두 개의 빈번한 시계열 패턴 정보들을 찾고 이들을 예측에 사용하면 시계열 데이터의 패턴에 따라서 정확한 예측을 할 수 있다. 따라서 본 논문에서는 이와 같은 접근 방향에서 시계열 데이터의 패턴들을 찾기 위한 클러스터링 기술을 적용하여 시계열 데이터 예측을 수행하는 알고리즘을 제안하고 그 성능을 검증한다.

앞으로의 논문 구성은 다음과 같다. 2장에서는 관련 연구를 살펴보고 3장에서는 본 논문에서 풀려는 시계열 데이터 예측 모델 학습 문제를 정의한다. 4장에서는 본 논문을 이해하는데 필요한 mRBF에 대해 자세히 살펴보고 5장에서는 클러스터링 기술을 적용하여 시계열 데이터를 예측하는 K-mRBF 알고리즘을 제안한다. 6장에서는 실험 결과를 통해 제안한 알고리즘의 성능을 보여준다.

2. 관련 연구

시계열 데이터 예측을 위한 많은 연구가 진행되어 왔다. 시계열 데이터 예측 연구는 주어진 학습 데이터를 가장 잘 설명하는 모델을 학습하고 이를 이용하여 시계열 데이터에 대해 특정 시간의 값을 예측하는 방식을 사용한다. [4]에서는 LMS(Least Mean Square) 적응 필터를 사용하여 평균 에러를 최소로 만드는 필터를 실시간으로 학습하여 예측 시간의 값을 계산하는 연구를 수행하였다. [5]에서는 데이터의 예측 시간 값에 대해 앞선 시간의 값과의 선형성을 가정하여 회귀 분석 모델을 통한 예측을 수행하였다. [6,7]에서는 분류기 학습 알고리즘을 사용하여 모델을 학습하는 방식을 사용하였다.

또한 최근에는 시계열 데이터가 각자 서로 다른 패턴을 보인다는 사실을 고려하여 모든 학습 데이터에 대해 단순히 하나의 모델을 학습하는 방식이 아닌 시계열 데이터의 패턴을 고려한 연구가 진행되고 있다. [8]은 kNN를 이용해 유사한 시계열 데이터를 찾은 후 이를 예측에 적용한다. [9]은 주어진 시계열 데이터를 조각(segment)으로 쪼개어 그 평균을 구한 후 예측에 사용한다. [3]에서는 RBF(Radial Basis Functions)를 사용하여 각 패턴과의 유사도(similarity)를 구하고 이를 통한 시계열 데이터의 예측값을 구한다.

3. 시계열 데이터 예측 모델 학습 문제

시계열 모델 학습 문제는 다음과 같이 정의된다.

문제정의: 시계열 데이터 예측 모델 학습
입력: 학습을 위한 n 개의 길이 t 의 시계열 데이터 집합
$D = \{s_1, s_2, \cdots, s_n\}$ (각 $s_i = (s_i[1], s_i[2], \cdots, s_i[t])$
목표: 모델 Θ 가 시계열 데이터 s_i 의 처음부터 $t-1$ 시간까지의 데이터 $\acute{s_i} = (s_i[1], s_i[2], \cdots, s_i[t-1]$ 을 이용해 t 시간 값을 예측한 것을 $\hat{s_i}[t]$ $(= \Theta(\acute{s_i}))$라 하자. 이 때 모든 학습 데이터에 대해 모델 Θ에 의한 예측값($\hat{s_i}[t]$)과 실제값($s_i[t]$)과의 오차들의 합이 최소가 되는 모델 Θ를 찾는다.

4. 클러스터링 알고리즘을 사용한 시계열 데이터 예측 모델

그림 2는 본 논문에서 제안하는 클러스터링 알고리즘을 사용한 시계열 데이터 예측 모델의 큰 그림을 나타낸 것이다. 학습 시계열 데이터가 주어지면 데이터들의 패턴을 찾기 위해 클러스터링 알고리즘을 수행하여 중심(center) 정보를 찾는다. 회귀 분석 모델을 학습할 때 이 중심 정보를 추가로 고려하여 더 정확한 모델을 생성할 수 있게 된다. 4.1절에서는 시계열 데이터에 대한 클러스터링 알고리즘에 대해 자세히 설명하고 4.2절에서는 이를 이용해 회귀 분석 모델을 학습하고 새로운 질의 데이터에 대해 예측값을 계산하는 방식을 설명한다.

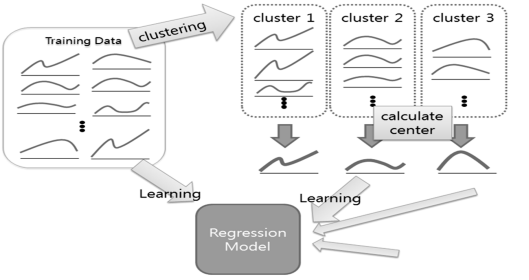

그림 2 클러스터링 알고리즘을 사용한 시계열 데이터 예측 모델
Fig. 2 A time series prediction model using clustering

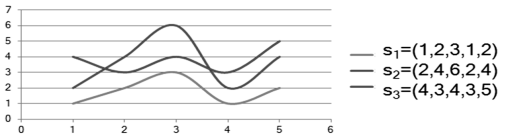

그림 3 시계열 데이터 거리
Fig. 3 A distance for time series data

4.1 K-SC 알고리즘

시계열 데이터의 클러스터링을 위해 [10]에서 제안한 K-SC 알고리즘을 사용한다. 단순히 유클리디언 거리 함수(distance function)를 사용하면 유사한 패턴의 시계열 데이터도 스케일링에 의해 거리값이 크게 나올 수 있다. 그림 3의 시계열 데이터를 보면 s_1 은 s_2 와 매우 유사한 패턴을 보이고 s_3 와는 전혀 다른 패턴을 보이지만 유클리디언 거리 함수에 의하면 두 시계열 데이터와 같은 거리 값을 가지게 된다. 따라서 유클리디언 거리 함수는 시계열 데이터에서의 패턴을 찾기 힘들고 K-SC 알고리즘에서는 시계열 데이터의 스케일링에 영향을 받지 않는 새로운 거리 함수를 제안한다. 두 시계열 데이터 s_1, s_2 에 대한 거리 함수는 다음과 같다.

$$d(s_1, s_2) = \min_{x,y}(\| s_1 - s_2^{(y)} \| / s_1) \quad (1)$$

$s_1^{(y)}$ 는 s_2 를 시간축에 대해 y 만큼 평행이동한 시계열 데이터이다. 즉 이 거리 함수는 s_2 를 x 에 의해 스케일링하고 y 에 대해 평행이동해서 두 시계열 데이터의 차이를 가장 작게 만들고 그 때의 상대적(relative) l_2-norm 값을 사용하기 때문에 시계열 데이터의 스케일링에 영향을 받지 않는다. 고정된 y 에 대해 x 의 최적화 문제를 풀면 $x = s_1 \cdot s_2^{(y)} / \| s_2^{(y)} \|^2$ 가 된다.

K-SC 알고리즘은 식 (1) 거리 함수를 사용하여 K-means 알고리즘과 유사한 방식으로 클러스터를 찾는다. 우선 임의로 K 개의 클러스터 중심(center)을 선택한다. 그 후 식 (1) 거리 함수를 통해 각 포인트를 가장 가까운 클러스터 중심에 할당하고 같은 클러스터 중심에 할당된 포인트들을 하나의 클러스터로 묶는다. 각 클러스터에 할당된 포인트들을 통해 새로운 클러스터 중심을 계산하고 다시 할당하는 작업을 수렴할 때까지 반복해 클러스터를 찾는다. 이 때 클러스터의 새로운 중심은 그 클러스터에 속한 모든 시계열 데이터들과의 거리의 합이 최소가 되는 점으로 결정된다. 즉 k 번째 클러스터 C_k 의 중심 u_k 는 $\mathrm{argmin}_u \sum_{s_i \in C_k} d(u, s_i)^2$ 이고 이는 $\sum_{s_i \in C_k} \left(I - s_i s_i^T / |s_i|^2\right)^T \left(I - s_i s_i^T / |s_i|^2\right)$ 의 최소 고유값에 대응하는 고유 벡터가 된다[10].

4.2 K-mRBF 알고리즘

앞서 설명했듯이 시계열 데이터들은 서로 다른 패턴들을 보이기 때문에 이런 성질을 고려하지 않은 시계열 데이터 예측은 높은 정확도를 가질 수 없다. [3]에서는 이런 성질을 고려하기 위해 mRBF (Multivariate RBF) 모델을 제안했다. mRBF 모델에서는 예측 모델을 학습할 때 학습 데이터에서 균등하게 K 개의 샘플을 뽑아 패턴 집합 $P = \{p_1, p_2, \cdots, p_K\}$를 만들고 시계열 데이터와 P 의 각 원소들과의 거리 정보를 사용하여 예측을 수행함으로써 시계열 데이터의 패턴 정보를 고려한다.

mRBF 모델에서는 시계열 데이터 s_i 에 대해 t 시간의 예측값 $\hat{s_i}[t]$ 을 다음 계산식을 사용해 계산한다.

$$\hat{s_i}[t] = \Theta_1 \cdot \acute{s}_i + \Theta_2 \cdot \left(rf_{p_1}(\acute{s_i}), \cdots, rf_{p_K}(\acute{s_i}) \right) \quad (2)$$

mRBF 모델에서 학습해야 할 모델 파라미터는 Θ_1, Θ_2로서 Θ_1은 s_i 의 앞선 시간의 데이터 $\acute{s}_i$ 와의 선형성을 고려하는 파라미터($t-1$ 차원 벡터)이고 Θ_2는 각 패턴의 영향을 고려하는 파라미터(K차원 벡터)이다. $rf_{p_j}(\acute{s}_i)$는 $\acute{s}_i$ 와 j 번째 패턴 p_j 와의 거리를 고려하기 위한 값으로서 아래 식 (3)로 정의된다.

$$rf_{p_j}(\acute{s_i}) = \exp\left[-\|s_i - p_j\|^2 / 2\sigma^2 \right] \quad (3)$$

σ 는 이 패턴의 분포 정도를 나타내는 파라미터이다. 결국 두 번째 항의 $\left(rf_{p_1}(\acute{s_i}), rf_{p_2}(\acute{s_i}), \cdots, rf_{p_K}(\acute{s_i}) \right)$ 는 $\acute{s}_i$ 과 P 의 각 패턴들과의 유사도(similarity)를 반영하는 값으로서 $\acute{s}_i$ 이 j 번째 패턴 p_j 와 유사하면 큰 값을 갖고 그렇지 않으면 작은 값을 갖는다.

모델에 의한 예측값 $\hat{s_i}[t]$ 와 실제 $s_i[t]$ 값과의 RSE (Relative Squared Error)[5]를 최소화하기 위해 다음의 목적 함수(objective function)를 최소화하는 모델 파라미터 Θ_1, Θ_2 를 학습한다. 이는 잘 알려진 Ridge 회귀 [11] 문제이다.

$$\mathrm{argmin}_{\Theta_1, \Theta_2} \frac{1}{|D|} \left(\sum_{s_i \in D} \left(\frac{\acute{s}_i[t]}{s_i[t]} - 1 \right)^2 \right) + \alpha \left(\|\Theta_1\|^2 + \|\Theta_2\|^2 \right) \quad (4)$$

이 mRBF는 시계열 데이터의 패턴 정보를 얻기 위해 입력 학습 데이터로부터 균등(uniform)하게 개의 샘플을 뽑는 방식을 사용하였다. 하지만 학습 데이터의 더 정확한 패턴 정보를 이용하면 예측 모델의 성능을 개선시킬 수 있을 것이다. 따라서 본 논문에서는 앞서 설명한 K-SC 클러스터링 방식을 이용해 주어진 학습 데이터가 갖는 패턴들을 찾고 이 정보를 시계열 데이터 예측에 이용하는 K-mRBF 모델을 제안한다.

K-mRBF 모델에서는 앞서 설명한 K-SC 알고리즘을 통해 패턴 정보 P 를 생성한다. $rf_{p_j}(\acute{s}_i)$에서 두 시계열 데이터의 거리를 계산하는 부분도 식 (1)를 사용한다. 다음 알고리즘 1은 K-mRBF 모델 학습 알고리즘의

수도 코드이다.

알고리즘 1 모델 학습(D, K, α, σ)
입력 입력 데이터 $D = \{s_1, \cdots, s_n\}$, 클러스터 개수 K, α, σ
출력 학습된 모델 파라미터 Θ_1, Θ_2
1. $\{p_1, p_2, \cdots, p_K\} = K$-SC ($D$, K)
2. **for** $i = 1$ to n do
3. $v_i = \left(s_i[1], \cdots, s_i[t-1], rf_{p_1}(\acute{s}_i), \cdots, rf_{p_K}(\acute{s}_i)\right)$
4. $[\Theta_1 \Theta_2] = (V^T V + \alpha I)^{-1} V^T y$
5. **return** Θ_1, Θ_2

알고리즘 1은 학습 데이터 D와 필요한 파라미터 K, α, σ를 입력으로 받아 목적 함수 식 (4)을 최적화하는 모델 파라미터 Θ_1, Θ_2를 구한다. 우선 입력 데이터에 대해 K-SC 군집화 알고리즘을 수행해 패턴 집합 P를 찾는다(1번째 줄). 시계열 데이터 s_i에 대해 구해진 패턴들과의 RBF $rf_{p_j}(\acute{s}_i)$를 계산하여 학습에 사용될 v_i를 구한다(3번째 줄). 4번째 줄은 식 (4)의 최적화 문제의 해로 V는 i 열이 v_i인 $(n \times (t-1+K))$ 행렬이고 y는 i 행의 값이 $s_i[t]$인 $(n \times 1)$ 행렬이다.

6. 실 험

제안한 모델의 정확성을 보이기 위한 실험을 수행하였다. 실험은 Intel Pentium Dual core 3.3G 프로세서와 4GB 메모리를 지닌 PC에서 진행되었다. 실험을 위해 다음 기존 연구와 제안한 알고리즘을 구현하였다.

- ML [5]: 앞선 시간 값과의 선형성을 가정하여 회귀 분석 모델을 통해 예측을 수행한다.
- mRBF [3]: 학습 데이터에서 균등하게 샘플을 뽑아 이를 패턴 정보로 사용하여 예측값을 계산한다.
- eu-mRBF: 클러스터링을 이용해 패턴 집합 P를 찾을 때 K-Means 알고리즘을 사용한 모델이다. 제안한 거리 측도의 우수성을 위해 성능을 구현하였다.
- K-mRBF: 4.2절에서 제안한 모델로 식 (1)의 거리 측도를 사용하여 클러스터링을 수행하고 그 결과를 이용하여 시계열 데이터 예측을 수행한다.
- sK-mRBF: 식 (1)의 거리 측도는 두 시계열 데이터가 시간 축에 의해서만 평행 이동되었으면 그 거리를 0으로 본다. 하지만 어플리케이션에 따라서 시간 축에 의해 평행 이동한 두 시계열 데이터를 다르게 인식해야 할 필요가 있을 수 있다. 따라서 식 (1)에서 시간 축에 의한 평행이동을 고려하지 않은 거리 측도를 사용한 sK-mRBF모델을 구현하였다.

실험은 k-fold 교차타당법 (cross validation)을 사용하여 RSE를 구하였고 (k=10) 이 실험을 20번 반복 수행하여 그 평균값을 사용하였다. 즉 매 실험마다 데이터를 중복이 없게 k개의 그룹으로 나누고 그 중 1개의 그룹을 테스트 데이터로 사용하고 k-1개의 그룹을 모두 합쳐서 학습 데이터로 사용하여 모델의 정확도를 측정하는 작업을 k번 반복하였다. 최적 파라미터 α, σ, K 값을 찾기 위해 $\alpha \in \{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$, $\sigma \in \{2, 1, 1/2, 1/22, 1/23, 1/24, 1/25, 1/26\}$, $K \in \{3, 5, 7, 15, 30\}$의 각 조합에 대해 각각 실험해보고 가장 좋은 성능을 보이는 조합의 값을 사용하였다.

실험을 위해 합성 데이터와 실제 데이터를 준비하였다. 합성 데이터의 생성을 위해 정규 분포를 따르는 K개의 시계열 데이터를 생성해 각 클러스터의 센터 정보로 사용하여 총 490개의 시계열 데이터를 생성하였다. 각 시계열 데이터를 생성할 때는 우선 어느 클러스터 C_j에 속할지를 결정하고 그 클러스터의 센터 p_j에 대해 0.1~5의 스케일링 팩터를 곱하고 각 차원마다 25%까지의 오차를 가지게 만들었다. 실제 데이터는 현재 서비스 중인 데이터 센터에서 수집된 서버들의 네트워크 트래픽 양을 6개월간 기록한 58,400개의 데이터에 대해 한 시계열 데이터의 길이를 12시간(t=12)으로 잘라 데이터 셋을 만들어 사용하였다.

6.1 합성 데이터 실험 결과

그림 4~6은 K = 3일 때 각 모델에서 찾은 클러스터 센터이다. 그림 4는 합성 데이터를 생성할 때 처음 주어진 센터 시계열 데이터이고 그림 5는 eu-mRBF에서 찾은 센터 집합이고 그림 6은 제안한 K-mRBF에서 찾은 센터 집합이다. K-mRBF은 원본 센터에 스케일 팩터만 곱해진 정확한 센터를 찾는데 반해 eu-mRBF은 원래

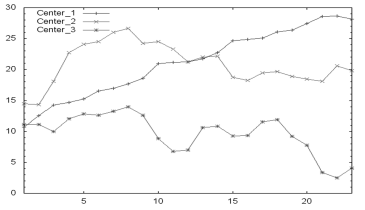

그림 4 합성 데이터 중심들
Fig. 4 Cluster centers of synthetic data

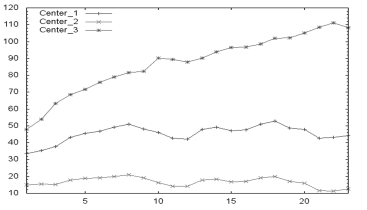

그림 5 eu-mRBF가 찾은 중심들
Fig. 5 Cluster centers found by eu-mRBF

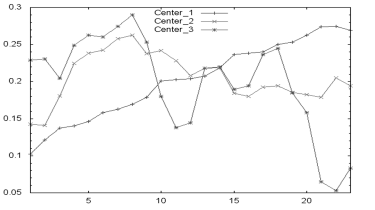

그림 6 K-mRBF가 찾은 중심들
Fig. 6 Cluster centers found by K-mRBF

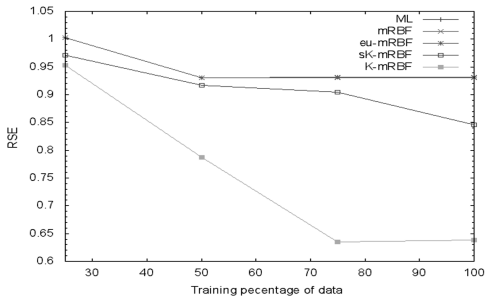

그림 7 합성 데이터 실험 결과

Fig. 7 Experimental results for synthetic data

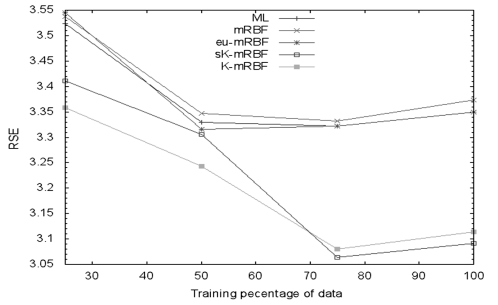

그림 8 실제 데이터 실험 결과

Fig. 8 Experimental results for real-life data

모양과 전혀 다른 센터를 찾는다.

그림 7은 합성 데이터에서 실험 결과 그래프이다. x축은 시계열 데이터의 길이(12시간)에 대해 학습에 사용된 데이터 길이(tr)의 비율이고 y축은 RSE를 나타낸다. 더 많은 데이터를 사용할수록 예측의 정확도가 높아져 더 낮은 RSE를 갖는 경향을 알 수 있다. mRBF와 eu-mRBF가 ML의 모델에 비해 성능 향상이 거의 없는 것은 mRBF와 eu-mRBF에서 찾아진 클러스터 센터 정보에 의한 성능 개선이 거의 없음을 의미한다. 본 논문에서 제안한 K-mRBF가 기존 연구에 비해 평균 20% 가량의 성능 향상을 보이는 것을 알 수 있다.

6.2 실제 데이터 실험 결과

그림 8은 실제 측정된 데이터에 대해 각 모델들이 예측한 값의 RSE를 비교한 그래프이다. 입력 데이터를 전부 사용해 학습한 모델이 (100%) 일부를 사용해 학습한 모델 (75%) 보다 더 안 좋은 성능을 보였다. 이는 모델이 학습 데이터에 너무 과적합 (overfitting) 되었기 때문이라 분석할 수 있다. 논문의 공간을 절약하기 위해서 실험 결과 그래프를 넣지 않았지만 학습 데이터에 대한 RSE는 x축이 증가할 때 감소하는 것을 볼 수 있었다. 결국 생성된 모델이 학습 데이터에 너무 과적합되어 일반적인 테스트 데이터에 대해서는 약간 안 좋은 성능을 보이게 된 것으로 보인다. 하지만 이 경우에도 제안한 k-mRBF의 성능이 mRBF에 비해 평균적으로 5.9%, 최대 7.6% 좋게 된다.

7. 결론 및 앞으로 수행할 연구

본 논문에서는 시계열 데이터 예측을 위해 학습 데이터들을 군집화하고 거기서 발견된 중심과의 거리를 이용하는 k-mRBF 모델을 제안하고 기존 연구와의 성능 비교를 통해 그 적합성을 검증하였다.

차후에는 본 논문에 사용한 시계열 데이터의 거리 측도를 개량해 시간 축에 대한 스케일링과 평행 이동에 영향을 받지 않을 뿐만 아니라 데이터 값 축에 대한 변형에 대해서도 영향을 받지 않는 거리 측도에 대한 연구를 수행하여 더 나은 클러스터링 알고리즘을 개발하고 이를 적용해 더 높은 정확도를 보이는 모델링을 학습하는 연구를 수행할 계획이다.

References

[1] http://en.wikipedia.org/wiki/Time_series [online]

[2] McKinsey & Company, White Paper: Revolutionizing Data Center Efficiency, 2009.

[3] P.Henrique, M.Jussara, A.Marcos, "Using early view patterns to predict the popularity of youtube videos," *proc. of the WSDM*, pp.365-374, 2013.

[4] Y. Kim, T. Kim, K. Shim, "Saving Energy in Data Centers by Forecasting Resource Usage," *proc. of the KCC fall conference*, vol.38, no.2(A), pp.214-217, 2011. (in Korean)

[5] S. Gabor, A. H. Bernardo, "Predicting the popularity of online content," *Communications of the ACM*, vol.53, no.8, pp.80-88, 2010.

[6] E. Cadenas, W. Rivera, "Short term wind speed forecasting in La Venta, Oaxaca, Mexico, using artificial neural network," *Journal of Renewable Energy*, vol.34, issue 1, pp.274-278, 2009.

[7] Y. Radhika, M, Shashi, "Atmospheric Temperature Prediction using Support Vector Machines," *International Journal of Computer Theory and Engineering*," vol.1, no.1, 2009.

[8] R. Nayak, P. Braak, "Temporal Pattern Matching for the Prediction of Stock Prices," *proc. of the AIDM*, pp.99-107, 2007.

[9] N. T. Son, D. T. Anh, "Time Series Similarity Search based on Middle Points and Clipping," *proc. of the DMO*, pp.13-19, 2011.

[10] Y. Jaewon, L. Jure, "Patterns of Temporal Variation in Online Media," *proc. of the WSDM*, pp.177-186, 2011.

[11] J. Friedman, T. Hastie, R. Tibshirani, "The elements of statistical learning," *Springer Series in Statistics*, 2001.