제25회 한국정보처리학회 춘계학술발표대회 논문집 제13권 제1호 (2006. 5)

시계열 데이터 예측을 위한 점진적인 회귀 분석 모델

김성현, 이용미, 김룡, 서성보, 류근호
충북대학교 전자계산학과
e-mail : hyun@dblab.chungbuk.ac.kr

An Incremental Regression Model for Time Series Data Prediction

Sung Hyun Kim, Yongmi Lee, Long Jin, Sungbo Seo, Keun Ho Ryu
Dept. of Computer Science, Chungbuk University

요 약

기존의 데이터 마이닝 예측 기법 중 회귀분석은 학습 단계에서 생성된 모델을 변경 없이 새로운 데이터에 적용하였다. 그러나 시계열 데이터에 모델 변경 없이 동일하게 적용하면 시간이 지남에 따라 정확도가 낮아지는 단점이 있다. 따라서 이 논문에서는 시간에 따라 변화하는 시계열 데이터의 특성을 고려하여 점진적으로 회귀 모델을 갱신하는 기법을 제안한다. 이 기법은 입력되는 모든 데이터를 회귀 모델에 적용하여 점진적으로 모델을 갱신한다. 제안된 기법의 타당성은 RME(Relative Mean Error)와 RMSE(Root Mean Square Error)를 이용하여 측정하였다. 정확도 측정 실험 결과 제안 기법인 IMQR(Incremental Multiple Quadratic Regression) 기법이 MLR(Multiple Linear Regression), MQR(Multiple Quadratic Regression), SVR(Support Vector Regression) 기법에 비해 RME 가 평균 2%, RMSE 가 평균 0.02 정도 우수한 결과를 얻었다.

1. 서론

데이터 마이닝의 목적은 예측(prediction)과 설명(description)으로 분류할 수 있다. 예측은 목표 속성의 형태가 이산형이면 분류(classification)를 사용하고 연속형이면 회귀분석(regression)을 사용한다[1]. 회귀분석은 주어진 자료의 속성간의 함수관계를 파악하고, 이 함수를 이용하여 입력 속성 값에 대응되는 출력 속성 값을 예측하는 분석기법이다[2]. 예를 들면 광고비와 같은 속성들이 매출액에 미치는 영향을 파악하여 모델을 만든 후에 광고비에 따라 매출액을 예측할 수 있다. 기존의 회귀분석은 학습 데이터로 예측 모델을 만든 후 모델이 타당하면 모델의 갱신 없이 새로운 데이터를 모델에 적용하였다[1].

최근 센서 네트워크 기술의 발달로 실생활에서 대용량의 데이터를 자동적으로 수집이 가능해졌다. 이런 데이터는 발생한 시점을 반영하기 때문에 시계열 데이터(time series)로 간주할 수 있다. 시계열 데이터는 시간이 지남에 따라 데이터의 분포나 특성이 변화될 수 있기 때문에 기존의 회귀분석에 적용하게 되면 시간이 지남에 따라 회귀 모델의 정확도가 낮아질 것이다.

따라서 이 논문에서는 시계열 데이터의 예측을 위해 점진적인 회귀분석 기법을 제안한다. 이 기법은 주기는 고려하지 않고 데이터가 입력될 때마다 모델에 반영하여 모델을 점진적으로 갱신하는 방법이다. 특히 실세계에서는 비선형적인 형태의 데이터가 많이 수집이 되므로 선형 회귀분석 보다 다중 이항 회귀분석을 적용하였다. 또한 갱신 비용을 줄이기 위해 이전 데이터를 유지하지 않고 행렬에 모델식의 최소 정보만을 유지함으로써 공간 비용을 절약한다. 제안하는 기법의 타당성을 검토하기 위해 실험 데이터를 통해 다중 선형 회귀분석(MLR), 다중 이항 회귀분석(MQR), SVR 기법과 제안 기법의 RME 와 RMSE 를 비교 실험하고 결과를 제시한다.

이 연구는 ETRI 의 “센서 데이터 처리를 위한 스트림 데이터 관리기술에 관한 연구” 사업의 연구비 지원으로 수행되었음

2. 관련 연구

점진적인 모델링은 새로운 데이터가 입력되면 모델에 반영이 되는 형태이다. 이것은 사례 기반 학습(instance-based learning)과 유사하다고 할 수 있다. 사례 기반 학습은 [3-7]에 의해 조사되고 연구되었다.

[3]은 분류를 위한 사례 기반 학습의 개념을 소개하였다. 이것은 저장된 사례를 기반으로 가장 가까운 이웃으로 분류하는 기법이다. 사례 기반 예측은 [4]에 의해 소개되었다. 예측을 위해 로컬 선형 회귀분석 형태의 기법을 사용하였다. 로컬 선형 회귀분석은 [5]에 의해 세부적으로 조사되었다. 이것은 데이터베이스의 크기를 증가하는 것은 제한하지 못한다. [6] 은 메모리 기반의 LWR(Locally Weighted Learning)과 이전 데이터를 기억할 필요가 없는 점진적인 LWR 을 논의 하였다. [7]은 제어 작업과 같이 연속적인 상태를 유지해야 하는 응용을 위해 적은 양의 데이터를 빠르고 안전하게 학습하는 HEDGER 알고리즘을 제안했다. HEDGER 알고리즘 LWR 에 기반한 사례 기반 알고리즘이다. 이것은 일반적인 선형 회귀분석을 변형한 것이다.

시계열 데이터에 대한 점진적인 모델링은 [8]에 의해 연구되었다. 이 기법은 순차적인 베이지안 진화 연산을 이용하여 예측을 수행한다. 이전 단계의 모델을 바탕으로 예측을 수행하고 새로운 데이터가 주어지면 현재의 예측 모델을 평가하여 더 좋은 모델을 생성하도록 하는 것이다. 예측된 값의 실제 값이 주어지면 최근 k 개의 데이터와 새로 주어진 데이터를 사용하여 생성된 모델의 성능을 평가한다.

기존의 연구는 모델의 재계산을 위해 이전의 데이터를 유지해야 할 필요가 있다. 또한 모델 갱신의 최적 주기를 판단하기 어려운 단점이 있다. 이 논문에서는 기존의 데이터를 유지하지 않고 입력되는 모든 데이터에 대해 모델을 갱신하고 최소 정보만 유지하는 기법을 제안한다.

3. 다중 회귀분석

선형 회귀분석(linear regression)은 입력 속성과 출력 속성간의 선형 관계에 관한 분석을 말하고, 다중 회귀분석(multiple regression)은 입력 속성이 2 개 이상일 때를 일컫는다. 따라서 입력 속성이 k 개인 다중 선형 회귀분석(MLR)은 식 (1)과 같다.

$$y = b_0 + b_1 x_1 + \cdots + b_k x_k + \varepsilon \qquad (1)$$

식 (1)과 같은 함수식에 의해 입력되는 속성 x 의 값들을 통해서 미지의 값 y 를 예측할 수 있다. 함수식의 계수인 b 값은 오차를 최소로 하는 최소제곱법 [2]에 의해 계산될 수 있고 식 (2)에 의해 수행된다. 식 (2)의 (a)에 의해 기존 데이터의 속성 x 와 y 를 행렬로 변환하고, 식 (2)의 (b)에 의해 행렬 X 의 전치행렬과 역행렬을 통해서 b 의 집합인 행렬 B 가 계산된다.

(a) 속성 x 와 속성 y 의 행렬

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

(b) 전치행렬과 역행렬 계산에 의해 유도

$$B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = (X^T \times X)^{-1} \times X^T \times y \qquad (2)$$

입력 속성 x 의 변화에 따라 출력 속성 y 의 변화가 직선적인 관계를 가질 때 선형 회귀분석을 사용하였다. 그러나 실세계의 많은 데이터들은 선형적인 관계보다는 곡선이나 불규칙한 분포를 갖는다. 입력 속성과 출력 속성의 곡선적인 관계에 관한 분석을 다항 회귀분석(polynomial regression)이라 한다. 특히, 이차항에 대한 분석을 이항 회귀분석(quadratic regression)이라 하고 입력 속성이 2 개인 다중 이항 회귀분석(MQR)은 식 (3)과 같다.

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1^{\,2} + b_4 x_2^{\,2} + b_5 x_1 x_2 + \varepsilon \qquad (3)$$

다중 이항 회귀분석의 b 값은 식 (4)의 (a)와 같이 각 속성을 이차 형태로 변환한 후에 다중 선형 회귀분석의 b 값 계산법인 식 (2)의 절차에 의해서 계산된다.

(a) 각 속성을 이차 형태로 변환

$$x_1^{\,2} = x_1 \times x_1 \qquad x_2^{\,2} = x_2 \times x_2$$
$$x_{12} = x_1 \times x_2 \qquad (4)$$

(b) 변환된 값으로 행렬 X 생성

$$X = \begin{bmatrix} 1 & (x_1)_1 & (x_2)_1 & (x_1^{\,2})_1 & (x_2^{\,2})_1 & (x_{12})_1 \\ 1 & (x_1)_2 & (x_2)_2 & (x_1^{\,2})_2 & (x_2^{\,2})_2 & (x_{12})_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & (x_1)_n & (x_2)_n & (x_1^{\,2})_n & (x_2^{\,2})_n & (x_{12})_n \end{bmatrix}$$

4. 점진적인 다중 이항 회귀분석

기존의 회귀분석에서는 초기 한번의 학습으로 모델을 생성하였다. 시계열 데이터는 시간이 지남에 따라 데이터의 특성이 변경될 수 있기 때문에 기존의 회귀분석을 그대로 적용하게 되면 시간이 지날수록 예측 정확도가 낮아질 것이다.

이 논문에서는 입력되는 시계열 데이터를 반영하여 회귀 모델을 갱신하는 기법인 점진적인 다중 이항 회귀분석(IMQR)을 제안한다. 이 기법은 바로 이전 시점까지의 데이터를 이용하여 학습한 후 모델을 갱신하는 기법이다. 최근 데이터를 반영하여 모델을 만들기 때문에 높은 정확도가 나올 것이라 예상된다. 제안하

는 IMQR 기법은 적은 비용으로 식 (3)을 갱신하여 함수식을 최신으로 유지하는 기법이다. 이 기법에서는 함수식의 갱신을 위해 이전의 모든 데이터를 유지할 필요가 없다. 행렬 X의 전치행렬과 행렬 X의 곱인 $X^{\mathrm{T}}X$와 행렬 X의 전치행렬과 행렬 y의 곱인 $X^{\mathrm{T}}y$만 유지하면 된다. 따라서 항상 같은 크기의 최소 정보만을 유지하므로 공간 복잡도를 줄일 수 있는 장점이 있다. 식 (5)의 (a)는 초기 데이터를 행렬 X, 행렬 X의 전치행렬 X^{T}, 행렬 y로 나타내고, 식 (5)의 (b)는 행렬 X^{T}와 X의 곱, 행렬 X^{T}와 y의 곱의 결과를 보여준다. 식 (6)과 식 (7)은 새로운 데이터 입력으로 행렬 $X^{\mathrm{T}}X$와 $X^{\mathrm{T}}y$가 갱신되는 과정을 보여준다.

(a) 행렬 X^{T}, X, y

$$X^T=\begin{bmatrix} 1 & 1 & \cdots & 1 \\ (x_1)_1 & (x_1)_2 & \cdots & (x_1)_n \\ (x_2)_1 & (x_2)_2 & \cdots & (x_2)_n \\ (x_1^2)_1 & (x_1^2)_2 & \cdots & (x_1^2)_n \\ (x_2^2)_1 & (x_2^2)_2 & \cdots & (x_2^2)_n \\ (x_{12})_1 & (x_{12})_2 & \cdots & (x_{12})_n \end{bmatrix}$$

$$X=\begin{bmatrix} 1 & (x_1)_1 & (x_2)_1 & (x_1^2)_1 & (x_2^2)_1 & (x_{12})_1 \\ 1 & (x_1)_2 & (x_2)_2 & (x_1^2)_2 & (x_2^2)_2 & (x_{12})_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & (x_1)_n & (x_2)_n & (x_1^2)_n & (x_2^2)_n & (x_{12})_n \end{bmatrix}$$

$$y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad (5)$$

(b) $X^{\mathrm{T}}\times X$, $X^{\mathrm{T}}\times y$

$$X^TX=\begin{bmatrix} z_{11} & z_{12} & z_{13} & z_{14} & z_{15} & z_{16} \\ z_{21} & z_{22} & z_{23} & z_{24} & z_{25} & z_{26} \\ z_{31} & z_{32} & z_{33} & z_{34} & z_{35} & z_{36} \\ z_{41} & z_{42} & z_{43} & z_{44} & z_{45} & z_{46} \\ z_{51} & z_{52} & z_{53} & z_{54} & z_{55} & z_{56} \\ z_{61} & z_{62} & z_{63} & z_{64} & z_{65} & z_{66} \end{bmatrix}$$

$$X^Ty=\begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \end{bmatrix}$$

• 데이터 입력으로 $X^{\mathrm{T}}X$ 갱신

$$x_{n+1}=\{x_1,x_2\}\rightarrow x_{n+1}=\{1,x_1,x_2,x_1^2,x_2^2,x_{12}\}$$

$$X^TX=\begin{bmatrix} z_{11}+1\times 1 & z_{12}+1\times x_1 & \cdots & z_{16}+1\times x_{12} \\ z_{21}+x_1\times 1 & z_{22}+x_1\times x_1 & \cdots & z_{26}+x_1\times x_{12} \\ \vdots & \vdots & \vdots & \vdots \\ z_{61}+x_{12}\times 1 & z_{62}+x_{12}\times x_1 & \cdots & z_{66}+x_{12}\times x_{12} \end{bmatrix} \qquad (6)$$

• 데이터 입력으로 $X^{\mathrm{T}}y$ 갱신

$$x_{n+1}=\{x_1,x_2\}\rightarrow x_{n+1}=\{1,x_1,x_2,x_1^2,x_2^2,x_{12}\},\quad y_{n+1}=y$$

$$X^Ty=\begin{bmatrix} c_1+1\times y \\ c_2+x_1\times y \\ c_3+x_2\times y \\ c_4+x_1^2\times y \\ c_5+x_2^2\times y \\ c_6+x_{12}\times y \end{bmatrix} \qquad (7)$$

b 값의 집합인 행렬 B를 구하기 위한 역행렬 계산은 행렬이 대칭이고 양의 정수로 되어 있을 때 편리한 Cholesky LU 분해법[9]을 사용하였다. 행렬 $X^{\mathrm{T}}X$는 행렬 X의 전치행렬과 행렬 X의 곱이기 때문에 결과는 대칭이고 양의 정수로 구성된다. 간단한 역행렬 계산 절차는 식 (8)과 같다.

$$\begin{gathered} X^T\times X=M:\text{ 대칭 양정치 행렬} \\ M=L\times L^T:\text{ L은 하삼각 행렬} \\ (L\times L^T)\times x=b,\quad L^T\times x=y \\ L\times y=b\text{로 }y\text{ 계산} \\ L^T\times x=y\text{로 역행렬인 }x\text{ 계산} \end{gathered} \qquad (8)$$

<표 1>은 새로운 시계열 데이터가 입력되면 데이터 값이 기존의 행렬 $X^{\mathrm{T}}X$, $X^{\mathrm{T}}y$와 함께 계산되어 새로운 b 값의 집합인 행렬 B가 계산되는 절차를 보여주는 알고리즘이다.

<표 1> 점진적인 다중 이항 회귀분석 알고리즘

Algorithm IMQR
입력: 시계열 데이터(xadd[], yadd)
출력: 함수의 계수(b)
Begin
Step 1: 행렬 $X^{\mathrm{T}}X$에 입력된 데이터 xadd[]를 적용
For i = 1 to r Do // r = $X^{\mathrm{T}}X$의 행 길이
For j = 1 to c DO // c = $X^{\mathrm{T}}X$의 열 길이
matrix[i][j] += xadd[i] * xadd[j];
End of inner For
End of outer For
Step 2: 행렬 $X^{\mathrm{T}}y$에 입력된 시계열 데이터 xadd[], yadd를 적용
For i = 1 to r Do // r = $X^{\mathrm{T}}y$의 행 길이
ymatrix[i] += xadd[i] * yadd;
End of For
Step 3: $X^{\mathrm{T}}X$의 역행렬 계산(Cholesky LU 분해법)
inverse[][] = LU(matrix[][]);
Step 4: 모델식의 계수 b 계산
For i = 1 to r Do // r = inverse의 행 길이
For j = 1 to c DO // c = inverse의 열 길이
beta[i][j] += inverse[i][j] * ymatrix[j];
End of inner For
End of outer For
End

5. 실험 및 평가

제안한 기법인 IMQR 의 비교 실험을 위해 [10]에서 사용된 불규칙한 특성을 갖는 Mackey-Glass Time Series[11] 데이터를 사용하였다. 실험 데이터는 총 6 개의 입력 속성과 1 개의 출력 속성을 갖는 1385 개의 샘플로 구성되어 있다.

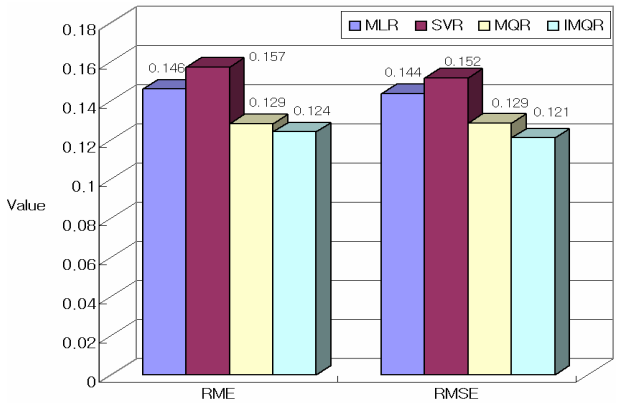
실험 방법은 실험 데이터를 MLR, MQR, SVR과 제안한 기법인 IMQR에 적용하여 에러율을 측정하였다. 초기 모델을 위한 학습 데이터는 500개의 샘플을 사용하였고, 검증 데이터는 885개의 샘플을 사용하였다. SVR 실험을 위해서는 mySVM[12]을 사용하였고 에러율의 측정은 식 (9) 같이 RME와 RMSE를 사용하였다.

$$RME = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y^*_i}{y_i}\right| \qquad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y^*_i)^2} \qquad (9)$$

실험의 결과는 <표 2>와 (그림 2)에서와 같이 제안한 기법인 IMQR 기법이 가장 우수하였다.

<표 2> 정확도 측정 실험 결과 비교

구분	MLR	SVR	MQR	IMQR
RME	0.1462	0.1574	0.1285	0.1243
RMSE	0.1438	0.1519	0.1287	0.1213



(그림 2) 정확도 측정 실험 결과 그래프

실험 결과 MLR 과 SVR 은 각각 14.6%와 15.7%의 상대 에러율을 보였다. 실험 데이터는 선형적인 관계가 아닌 실세계를 반영한 불규칙한 데이터를 사용했기 때문에 선형적인 관계를 고려한 MLR 과 SVR 기법이 다른 기법보다 에러율이 높게 나타남을 알 수 있다. 또한 MQR 기법의 RME 는 12.9%로 한번의 학습으로 생성된 모델에 새로운 데이터를 적용했기 때문에 제안한 IMQR 기법의 12.4% 보다 에러율이 높았다. 만약 입력되는 데이터의 시점이 점점 더 멀어진다면 MQR 기법의 에러율은 점진적으로 증가할 것이라고 예상할 수 있다. <표 2>와 (그림 2)의 결과와 같이 RMSE 결과도 RME 결과와 유사한 정확도를 보였다.

6. 결론 및 향후 연구

시간 속성을 갖는 시계열 데이터는 시간이 지남에 따라 데이터 특성이 변화될 수 있기 때문에 기존의 회귀분석을 적용하는 것은 적합하지 않다. 이 논문은 시간에 따라 데이터의 분포가 변화될 가능성이 있는 시계열 데이터를 위해 점진적인 다중 이항 회귀분석 기법을 제안하였다. 제안한 기법은 점진적으로 변화되는 데이터 분포를 모델에 반영할 수 있기 때문에 에러율의 증가를 최소화 시킨다. 또한 모델을 갱신하기 위해 이전 데이터를 유지하지 않고 일정 크기의 행렬만 유지하여 공간 복잡도를 줄여준다. 타 기법과의 비교 실험 결과 가장 낮은 에러율을 보여 타당함을 증명하였다. 제안한 기법은 비선형적이며 시간에 따라 데이터의 분포가 변화될 수 있는 도메인에 적용 가능하다.

향후 연구로는 모델 갱신에 적용할 데이터 선택 기법, 특정 과거 시점의 이전 데이터 제거 후 모델 갱신 기법, 최적의 모델 갱신 주기 선택 기법 등의 연구가 필요하다.

참고문헌

[1] P. N. Tan, M. Steinbach, and V. Kumar, “INTRODUCTION TO DATA MINING”, Addison Wesley, 2005.

[2] 박성현, “회귀분석”, 민영사, 1997.

[3] D. W. Aha, D. Kibler, and M. K. Albert, “Instance-Based Learning,” In Machine Learning, pp.37-66, 1991.

[4] D. Kibler, D. W. Aha, and M. Albert, “Instance-based prediction of real-valued attributes,” In Computational Intelligence, pp.51-57, 1989.

[5] C. G. Atkeson, A. W. Moore, and S. Schaal, “Locally weighted learning,” In Artificial Intelligence Review, pp.11-73, 1997.

[6] S. Schaal, C. G. Atkeson, and S. Vijayakumar, “Real-Time Robot Learning With Locally Weighted Statistical Learning,” In Proc. of the IEEE International Conference on Robotics and Automation, pp.288–293, 2000.

[7] W. D. Smart, L. P. Kaelbling, “Practical reinforcement learning in continuous spaces,” In Proc. of the 17th International Conference on Machine Learning, pp.903-910, 2000.

[8] 조동연, 장병탁, “순차적 베이지안 진화 연산을 이용한 시계열 예측”, 한국정보과학회 가을 학술발표 논문집, Vol.27, No.2, pp.311-313, 2000.

[9] S. C. Chapra, and R. P. Canale, “Numerical Methodd for Engineers, Third Edition”, McGraw-Hill Korea, 1999.

[10] G. W. Flake, and S. Lawrence, “Efficient SVM Regression Training with SMO,” In Machine Learning, 271-290, 2002.

[11] C. C. Chang, and C. J. Lin, LIBSVM, http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html

[12] S. Ruping, mySVM, Computer Science Dep. AI Unit Univ. of Dortmund, 2000.