스트림 데이터에서 회귀분석에 기반한 빈발항목 예측

Frequent Items Mining based on Regression Model in Data Streams

이욱현
한북대학교 컴퓨터정보학과

Uk-Hyun Lee(uhlee@hanbuk.ac.kr)

요약

최근 스트림데이터 환경의 데이터 모델은 데이터의 양이 아주 크고 연속적이며 무한하다. 이에 반해 제한된 용량의 디스크나 메모리 등을 이용해서 질의 처리나 데이터 분석을 처리한다. 이러한 환경에서 트랜잭션 데이터베이스에 대한 전통적인 빈발패턴탐사는 불가능하다고 할 수 있다. 왜냐하면, 연속적으로 들어오는 스트림 데이터에 대해 어떤 항목집합이 빈발항목인지 아닌지에 대한 정보를 계속적으로 유지 관리하기가 어렵기 때문이다. 본 논문에서는 연속적으로 들어오는 스트림 데이터에 회귀모델을 적용하여 빈발항목들을 예측할 수 있는 방법을 제안한다. 스트림 데이터로부터 회귀모델을 생성함으로써 불확실한 항목들에 대한 예측 모델로 사용할 수 있다. 다양한 실험을 통하여 제안하는 방법이 스트림 데이터 환경의 데이터에 효율적으로 사용될 수 있음을 보인다.

■ 중심어 : | 스트림데이터 | 빈발항목 | 회귀모델 | 예측 |

Abstract

Recently, the data model in stream data environment has massive, continuous, and infinity properties. However the stream data processing like query process or data analysis is conducted using a limited capacity of disk or memory. In these environment, the traditional frequent pattern discovery on transaction database can be performed because it is difficult to manage the information continuously whether a continuous stream data is the frequent item or not. In this paper, we propose the method which we are able to predict the frequent items using the regression model on continuous stream data environment. We can use as a prediction model on indefinite items by constructing the regression model on stream data. We will show that the proposed method is able to be efficiently used on stream data environment through a variety of experiments.

■ keyword : | Stream Data | Frequent Item | Regression Models | Prediction |

I. 서 론

빈발 패턴 마이닝(frequent patterns mining)은 데이터마이닝 분야에서 가장 폭넓게 연구되어지고 있는 분야로써 데이터베이스에서 빈번하게 발생하는 패턴들을 찾기 위한 다양한 알고리즘들이 제안되었다[1][12][19]. 기존의 데이터마이닝 방법들은 기본적으로 지식 발견의 대상이 되는 데이터 집합이 마이닝 작업 시작 이전

접수번호 : #081024-001
접수일자 : 2008년 10월 24일
심사완료일 : 2008년 12월 11일
교신저자 : 이욱현, e-mail : uhlee@hanbuk.ac.kr

에 명확히 정의되었다. 또한, 기존의 데이터마이닝 방법들은 대용량의 데이터 집합에 대한 마이닝 결과를 얻는데 있어서 상당한 처리 시간을 요구한다. 일반적으로 응용 환경에 내재된 지식들은 시간의 흐름에 따라 변화하며 최근에 얻어진 의미 있는 마이닝 결과들도 시간이 지남에 따라 무의미한 정보로 변화될 수 있다. 따라서, 최근에 발생된 트랜잭션에 나타나는 정보들을 포함하는 최신의 마이닝 결과를 얻기 위해서는 전체 데이터베이스에 대한 마이닝 작업이 처음부터 다시 수행되어야 한다.

최근 네트워크 트래픽 분석(network traffic analysis), 전력소비측정(power consumption measurement), 센서 네트워크 데이터 분석(sensor network data analysis)과 같은 스트림 데이터분야에서의 다양한 연구 분야들이 등장하고 있다[6]. 센서 네트워크에서 수집이 되는 데이터는 연속적이며 무한한 특징을 가진 스트림 데이터이다. 최근 네트워크와 센서 기술의 발달로 시간과 공간의 제약 없이 실제 환경에서 데이터를 실시간으로 수집하고 분석하여 의사결정에 반영할 수 있는 스트림데이터시스템(Data Stream Management System, DSMS) 환경이 대두 되었다[2]. 스트림 데이터가 수집되는 센서 네트워크는 객체감시(object guarding), 환경감시(environment monitoring), 객체추적(object tracking) 등의 응용 분야가 있으며 현재 많은 연구가 수행되고 있다[5][11][17]. 센서로부터 실시간으로 들어오는 스트림 데이터는 낮은 변화량을 갖으며, 초(seconds)단위, 분(minute)단위로 지속적으로 입력, 발생되는 트랜잭션들로 구성되는 무한 집합으로 정의된다. 이러한 스트림 데이터 시스템은 제한된 메모리와 배터리, 소형 프로세서 및 낮은 통신 대역폭의 특성으로 인하여 센서를 통하여 수집되는 모든 데이터를 완벽하게 전송 및 분석하는 것은 불가능하며, 연속적이므로 스트림 데이터의 모든 트랜잭션들을 별도로 저장하는 것은 불가능하다. 그러므로 스트림 데이터의 각 트랜잭션 정보는 시간 구간(time interval), 슬라이딩 윈도우(sliding window, 데이터 임계값(threshold) 등의 범위에서 오직 한 번 읽고 마이닝 결과를 생성해야 한다. 이러한 이유로 스트림 데이터 마이닝 방법들은 마이닝 결과에 근사(approximate)값을 포함하게 된다.

본 논문에서는 이러한 스트림 데이터의 특성에 따라 데이터의 변화빈도에 따른 일차원 속성의 스트림 데이터로부터 회귀 모델을 적용하여 빈발 항목들을 마이닝하기 위한 기법을 제안한다. 스트림 데이터는 시간에 따라 연속적이고 복잡하여 한시적인 접근만 가능하고 제한된 메모리를 사용하여 동적으로 변화하기 때문에 지속적인 데이터 처리 모델이 요구된다[2][7]. 또한 스트림 데이터는 순서화된 특성을 갖고 있기 때문에 시계열 데이터(time series)로 간주할 수 있다. 이러한 시계열 데이터 예측은 과거 데이터를 통해 미래를 예측하여 유용한 정보를 얻는 것이다. 본 논문에서는 회귀 모델을 수립하기 위해 일차원 속성의 스트림 데이터를 전처리하고 전처리된 스트림 데이터를 이용하여 회귀 모델을 수립한다. 회귀 모델이 생성되면 그 모델을 기반으로 빈발 항목의 가능성 여부에 대한 예측 과정을 수행한다. 본 논문의 구성은 Ⅱ장 관련연구에서 스트림 데이터 시스템과 스트림 데이터 예측과 관련된 기존의 시계열 데이터 예측 기법에 대해 설명한다. 그리고 이 논문에서 제안하는 기법의 기초가 되는 회귀분석에 대해 설명한다. Ⅲ장 회귀분석을 이용한 빈발 항목 추출에서 스트림 데이터에서의 빈발 항목을 추출하기 위한 마이닝 기법에 대해 설명한다. Ⅳ장 실험 및 평가에서 예제 데이터를 이용하여 제안한 빈발 항목 마이닝 기법의 정확도를 실험하고 평가한다. 마지막으로 Ⅴ장에서 연구 결과를 요약한다.

Ⅱ. 관련연구

1. 스트림 데이터 관리

스트림 데이터는 연속적이며 크기가 무한하다. 스트림 데이터 시스템은 센서의 제한된 배터리, 무선 통신으로 인한 데이터 손실 등의 특성으로 인하여 수집되는 데이터는 잠재적인 에러를 가지고 있다. 그러므로 스트림 데이터 시스템에서의 데이터 처리는 기존의 데이터베이스 시스템에서의 데이터 처리와는 다르게 수집된

데이터를 축약 및 요약하여 처리한다. 또한 스트림 데이터는 무한한 크기의 연속적인 데이터이므로 스트림 데이터에서 연관 규칙을 탐사하기 위하여 한 번의 데이터 스캔만으로 데이터들의 연관 규칙 정보를 분석하여야 한다[10].

스트림 데이터 시스템에서 수집되는 데이터는 연속적으로 수집되는 무한한 크기의 데이터이다. 스트림 데이터의 처리 및 관리를 위해서는 기존의 데이터베이스 시스템에서의 데이터 분석 및 처리 방법을 스트림 데이터 시스템의 특성을 고려하여 수정하고 적용하여야 한다. 현재 일정 주기 단위로 데이터를 필터링, 근사화 및 축약 방법을 통하여 데이터를 요약하는 방법이 제안되었다[3][8][13]. 스트림 데이터 처리를 위한 일정 주기는 하나의 윈도우로 정의되며, 하나의 윈도우 구간 동안에 센서를 통하여 수집되는 데이터는 기존의 데이터베이스 시스템에서의 트랜잭션에 포함된 항목으로 간주 할 수 있다[3][13].

2. 시계열 데이터 예측

시계열이란 한 사건 또는 여러 사건에 대해 시간의 흐름에 따라 일정한 간격으로 이들을 관측하여 기록한 자료를 말한다[18]. 스트림 데이터 역시 시계열 데이터와 같이 시간의 흐름에 따라 일정한 간격으로 수집되는 데이터이다. 시계열 데이터의 예로 매일 변동하는 종합주가 지수, 특정 소비재의 월별 판매량, 연도별 농작물의 생산량 등이 있다. 이러한 시계열은 어떠한 경제현상이나 자연현상에 관한 시간적 변화를 나타내는 역사적 계열이므로 어느 한 시점에서 관측된 시계열 자료는 그 이전까지의 자료들에 주로 의존하게 된다. 따라서 시계열 분석을 통한 예측에서는 관측된 과거의 자료들을 분석하여 법칙성을 발견하고 이를 모형화하여 추정한다. 이 추정된 모형을 이용하여 미래에 관측될 값들을 예측하게 된다.

기존의 시계열 예측 기법으로는 단순 이동 평균법, 지수 평활법, ARIMA 모형 등이 있다. 단순 이동 평균법은 가장 최근의 m-기간 동안의 자료들 평균을 다음 시점의 예측 값으로 추정한다. 이 방법은 시간의 경과에 따라 평균의 변화가 크지 않을 경우에 적용 가능하다. 지수 평활법은 최근의 자료들에 대해 더 많은 가중치를 부여하는 방법이다. 따라서 빠르게 수집이 되는 데이터에는 적용하기 힘들다. ARIMA 모형은 추세가 있는 경우 이를 제거하여 정상적 데이터로 변환한 후 AR, MA, ARMA 모형 중에서 가장 적합한 모형을 선택할 수 있게 해준다.

시계열 데이터 예측은 [4][9][14]에 의해 연구되었다. [4]는 n개의 입력을 받고 1개의 출력을 내보내는 신경망(neural network) 기법이다. 신경망은 마켓 예측, 기상이나 네트워크 트래픽 예측 등과 같이 시계열 예측에 널리 사용된다. [9]은 웨이블릿(wavelet) 기법과 비모수 회귀분석을 이용하여 짧은 시간에 수집되는 불규칙한 시계열 데이터를 예측한다. [14]은 Kalman Smoother를 이용하여 시계열 예측을 수행하고 [18]은 연관 규칙을 이용하여 시계열 데이터를 예측한다. 이 방법은 시계열 데이터에서 반복되는 패턴을 추출하여 예측을 수행한다.

3. 회귀 분석

회귀분석은 하나의 종속변수가 다른 독립변수들에 의해 어떻게 설명 또는 예측되는지를 알아보기 위해 적절한 함수로 표현하여 자료 분석을 하는 통계적인 기법이다[15]. 표본 데이터를 이용해서 모집단에 존재하는 종속변수 y와 독립변수 x간의 함수관계(functional relationship)를 가장 그럴 듯하게 추론해 보자는 것이 회귀분석의 내용이다.

입력 속성과 출력 속성간의 선형 관계에 관한 분석을 선형 회귀분석이라 하고 입력 속성이 2개 이상일 때를 다중 회귀분석이라고 한다.

회귀분석의 일반적인 목적은 예측에 있다. 회귀분석은 공학, 자연과학, 경제학, 경영학, 생명과학이나 생물학, 사회과학 등 거의 모든 분야에서 이용되고 있다. 회귀분석의 예에서 알 수 있는 것은, 모든 경우에 항상 원인이 되는 변수들이 있고 그 원인에 따라 결과로 나타나는 변수가 있다는 점이다.

두 변수 사이의 함수관계(functional relationship)는 수학식을 통해 표현되는데 보통 다른 요인들로부터 독립적으로 결정되는 X를 독립변수(independent

variable), 그리고 독립변수들의 함수관계에 의해 결정되는 Y를 종속변수(dependent variable)라고 한다.

회귀분석 혹은 회귀모형은 데이터의 설명, 모수추정, 예측과 추정, 그리고 품질 관리와 같은 여러 가지 목적으로 사용된다. 회귀분석의 응용사례중 상당수는 반응변수에 대한 예측과 관련이 있다. 예를 들어, 어떤 지역에 풍차에 의한 풍력발전소를 건설하려 한다고 하자. 바람의 세기에 따른 발전량을 예측할 수 있다면 필요한 발전량을 얻기 위해 몇 개의 풍차를 세워야하는지 계산할 수 있을 것이다.

본 논문에서 사용하는 회귀분석은 선형 회귀분석으로써, 표본 데이터를 이용해서 모집단에 존재하는 반응변수 또는 종속변수 y와 설명변수 또는 독립변수 x간의 함수관계를 추론해 보자는 것이다. 이때 선형 회귀모형을 추정하기 위해 최소제곱법을 이용한다. 그럼으로써, 본 논문은 스트림 환경에서는 처음으로 빈발 항목 예측을 위해 선형 회귀 분석을 적용하여 연구 실험해 보고자 한다.

Ⅲ. 회귀 분석을 이용한 빈발 항목 추출

이 장에서는 일차원 스트림 데이터에 대해서 회귀분석 모형을 추정하고 추정된 회귀모형을 이용하여 빈발 항목을 예측하기 위한 기법을 설명한다. 이를 위해 일차원 속성의 스트림 데이터를 전처리하여 회귀분석을 위한 표본값의 형태로 변형하고, 표본값이 생성되면 회귀모형 추정 방법인 최소제곱법을 이용하여 회귀모형을 구하게 된다.

1. 스트림 데이터의 전처리

연관 규칙은 1993년에 처음 소개된 것으로 A→B의 형태를 갖는 패턴이다. 이때 A와 B는 항목의 집합이다. 연관 규칙 탐사는 트랜잭션 데이터베이스에 존재하는 의미 있는 패턴들을 발견하는 것이다. 일반적인 연관 규칙 탐사는 미리 정해져 있는 크기의 트랜잭션 데이터베이스와 많은 자원을 가진 환경에서의 데이터마이닝 기법이다. 그러나 스트림 데이터 환경에서의 데이터에 대한 연관 규칙 탐사는 제한적인 자원과 연속적이고 무한한 데이터에서의 데이터마이닝 기법이라고 정의할 수 있다. 본 논문에서는 일차원 속성의 스트림데이터에서 빈발 항목을 발견하는 문제를 다룬다. 이때, 선형 회귀 모형을 사용하여 빈발 항목을 추정하게 된다. 이 논문에서 설명하는 스트림 데이터는 하나의 속성을 갖는 일차원 속성이기 때문에 회귀분석에 적용하기 위한 적절한 변환이 필요하다.

회귀분석의 목적은 표본으로 관측된 값을 통해서 미래의 추정값을 찾는 것이다. 즉 총 n개의 표본점을 얻었을 때, 이들을 학습하여 x_{n+1}값이 입력으로 들어왔을 때 추정값 y_{n+1}을 구하는 것이다. 따라서 입력으로 들어오는 스트림 데이터를 (x_i, y_i)과 같은 x와 y의 데이터 쌍으로 변환할 필요가 있다.

스트림 데이터는 시계열 데이터의 특성을 가진다. 본 논문에서 사용하는 스트림 데이터는 일차원 속성을 가지기 때문에 [표 1]과 같은 형태를 취하게 된다.

[표 1]에서 보는 것과 같이 기준 시간 T에 따라 연속적으로 스트림 데이터가 입력으로 들어오게 된다. 연속적으로 들어오는 스트림 데이터에서 회귀분석에 적용하기 위해 데이터 전처리 과정을 거친다. 이 전처리 과

표 1. 일차원 속성의 스트림 데이터예

	T-16	T-15	T-14	T-13	T-12	T-11	T-10	T-9	T-8	T-7	T-6	T-5	T-4	T-3	T-2	T-1
A		■		■			■		■				■			■
B	■		■									■				
C					■			■						■		
D											■					
E						■				■					■	

정을 통해 스트림 데이터는 회귀분석에 이용하기 위한 (x_i, y_i)순서쌍의 형태를 취하게 된다. 스트림 데이터 전처리 과정은 다음과 같다. 첫째, 스트림 데이터를 이루는 각각의 속성값에 대하여 각 속성값과 그 속성값이 입력되는 시점으로 스트림 데이터를 재구성한다. 일차원 속성의 스트림 데이터값은 [표 1]의 예와 같이 문자형(character)의 값을 갖거나 아니면 숫자형(numeric)의 데이터 값을 갖게 된다. 이러한 값은 어느 시점에 하나의 속성값을 가지므로 시계열 데이터라고 할 수 있다. 위의 스트림 데이터의 예에서 어떤 속성값이 어떤 시점들에 등장하는지를 알 수 있다.

[표 1]에서 보는 것과 같이, 각 항목과 각 항목이 등장한 시점들로 스트림 데이터를 재구성해 보면, 이들은 트랜잭션 데이터베이스의 형태를 취한다. 여기서의 트랜잭션은 각 항목과 각 항목이 등장한 시점들로 정의되어 진다. [표 1]의 예에서 항목 A는 각각 T-15, T-13, T-10, T-8, T-4, 그리고 T 시점에 입력으로 들어온다. 항목 B는 T-16, T-14, 그리고 T-5 시점에 들어온다. 각 항목에 대한 시점의 차를 구해보자. 항목A의 예를 들면, 속성값 A가 스트림 데이터에 입력으로 들어오는 시점은 각각 T-15, T-13, T-10, T-8, T-4, 그리고 T이다. 이 시점을 해석해보면 스트림 데이터에서 속성값 A라는 데이터는 현재 시점 T보다 15시점 이전에 한 번 입력으로 들어오고 그 다음에 2 {(T-13)-(T-15)} 라는 시간 차이가 지난 다음에 다시 속성값 A가 입력으로 들어온 것이다. 다음엔 3 {(T-10)-(T-13)}, 2 {(T-8)-(T-10)}, 4 {(T-4)-(T-8)}, 그리고 4 {(T-(T-4)}의 시간 차이를 두고 속성값 A가 입력된다. [표 1]에서 등장하는 스트림 데이터의 예에서 T-15 시점 이전에 속성값 A가 언제 입력되었는지는 알 수 없지만, T-15 시점부터 현재 시점 T까지 속성값 A가 등장한 시간의 차를 살펴보면 <2, 3, 2, 4, 4>의 차이를 두고 있음을 알 수 있다. 이러한 시간의 차이를 순서쌍 (2, 3), (3, 2), (2, 4), (4, 4)으로 변환하였다. 이 순서쌍의 의미는 속성값 A가 이전에 입력된 시간의 차와 이후에 입력된 시간의 차의 쌍이다. 이 순서쌍으로부터 절편과 기울기를 구할 수 있고, 그 절편과 기울기를 이용하여 직선을 그을 수 있다. 따라서 우리는 이 순서쌍을 이용하여 회귀모형을 추정할 수 있다.

회귀분석은 표본 데이터를 이용해서 모집단에 존재하는 종속변수 y와 독립변수 x의 함수관계를 알고자 할 때 많이 사용된다. (2, 3)이라는 순서쌍은 속성값 A가 2라는 시간이 지난 다음에 스트림 데이터로 입력되고 다시 3 이라는 시간 경과 후에 입력된다는 것을 알 수 있다. 또한, (3, 2)에서 3 이라는 시간이 경과한 후에 속성값 A가 입력되기 위해서는 다시 2 라는 시간이 경과 한다는 것을 알 수 있다. 마찬가지로, (4, 4)는 속성값 A가 4라는 시간이 경과한 후에 다시 등장하는 시간 간격이 4라는 것을 나타낸다. 즉, 이전에 등장한 시간 간격을 독립변수로 설정하고 이후에 등장하는 시간 간격을 종속변수로 설정시, 앞으로 입력되어질 시간간격의 값을 예측할 수 있는 회귀모형을 추정할 수 있게 된다.

2. 스트림 데이터의 회귀 모형 추정

회귀분석은 주어진 자료를 통해 속성간의 함수 관계를 파악하여 입력 속성값에 대응되는 출력 속성값을 예측하는 분석 방법이다. 입력 속성과 출력 속성간의 선형 관계에 관한 분석을 선형 회귀분석이라 하고 입력 속성이 2개 이상일 때를 다중 회귀분석이라고 한다. 본 논문에서는 입력 속성값이 1차원의 형태를 가지므로 선형 회귀분석을 이용하여 회귀모형을 추정한다.

회귀분석을 위해 종속변수(y)와 독립변수(x)를 각각 다음 [그림 1]과 같이 정의한다.

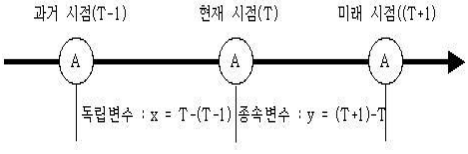

그림 1. 속성값 A의 예에 대한 독립변수 종속변수 관계

독립변수 x는 스트림 데이터로 들어오는 속성값이 현재 입력된 시점과 바로 전에 입력되었던 시점과의 시간 간격을 의미한다. 종속변수 y는 미래에 입력될 시점과 현재 입력된 시점과의 시간 간격을 의미한다. 과거의 이력 데이터로부터 한 시점 이전의 입력 시간 차이와 한 시점 후의 입력 시간 차이로부터 회귀모형을 추

정하고, 생성된 회귀모형을 이용하여 한 시점 이전에 입력된 시간 간격으로부터 미래에 입력될 시간 간격을 예측함으로써 어떤 항목의 향후의 지지도를 예측할 수 있다.

본 논문에서는 선형 회귀분석을 이용하여 회귀직선을 추정하기 위해 최소제곱법을 이용한다. 최소제곱법을 이용한 회귀분석을 설명하기 위해 [표 1]의 속성값 A를 예로 사용한다.

[표 2]는 효율적인 설명을 위해 하나의예제 데이터로써 속성값 A에 대한 입력 시간 간격을 좀 더 추가하였다.

표 2. 속성 값 A에 대한 독립변수 종속변수 예

Sequence	1	2	3	4	5	6	7	8	9
독립변수(x)	2	3	2	4	4	3	5	7	6
종속변수(y)	3	2	4	4	3	5	7	6	5

[표 2]의 데이터를 산점도(scatter plot or scatter diagram)로 표현한 것이 [그림 2]에 나타나있다.

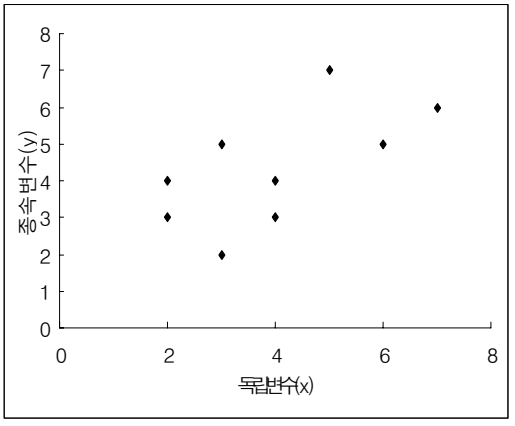

그림 2. 속성 값 A의 데이터에 대한 산점도

산점도는 변수간의 관계를 규명하고 이 관계를 시각적으로 표현하고자 할 때 사용한다. 우리의 목표는 이 산점도에 대한 회귀직선을 만드는 것이다. 직관적으로 가장 좋은 직선은 각 점에서 직선까지의 거리가 가장 가깝게 되도록 직선이 만들어져야 한다. 그러나 모든 점들에 대해서 동시에 거리가 최소가 되는 직선을 만들 수 없으므로 이 거리들의 제곱합이 최소가 되도록 하는 것이 최소제곱법이다[16].

선형 회귀모형은 x가 주어지면 $b_0 + b_1x$라는 함수에 의해 계산되는 y값을 구하는 것이다. 그러므로 일반적인 모회귀모형은 식 (3.1)과 같다.

$$y = b_0 + b_1x + \varepsilon \tag{3.1}$$

이때, 최소제곱법은 거리들의 제곱합을 최소로 하는 기법이다. 여기서 거리는 에러(ε)를 뜻한다. 그러므로 실제 관측값 y_i에서 (3.1) 식의 $b_0 + b_1x_i$ 를 빼준 값(즉 ε_i)의 제곱합을 최소화하는 b_0, b_1을 찾는 것이 된다. ε_i의 제곱합을 S라 하면 다음 식 (3.2)와 같다.

$$S(b_0, b_1) = \sum_{i=1}^{n} \epsilon^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1x_i)^2 \tag{3.2}$$

식 (3.2)에서 $S(b_0, b_1)$이 최소가 되도록 b_0, b_1을 정하는 방법을 최소제곱법이라 하는데, 이렇게 구한 추정량 $\widehat{b_0}, \widehat{b_1}$을 최소제곱추정량이라 한다.

$$\widehat{b_0} = \overline{y} - \widehat{b_1}\overline{x}$$

$$\widehat{b_1} = \frac{\Sigma x_i y_i - n\overline{x}\,\overline{y}}{\Sigma x_i^2 - n\overline{x}^2} = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\Sigma(x_i - \overline{x})^2} \tag{3.3}$$

여기서 $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$이다. 또 $\widehat{b_0}$은 절편에 대한 추정량이며, $\widehat{b_1}$은 직선의 기울기에 대한 추정량이다. 따라서 우리가 추정하고자하는 표본 회귀식은 식 (3.4)와 같다.

$$\widehat{y} = \widehat{b_1}x + \widehat{b_0} \tag{3.4}$$

한편 식 (3.3)의 분자, 분모는 모두 평균으로 수정된 제곱합이므로 식 (3.5)와 같이 간단하게 쓸 수 있다.

$$S_{xx} = \Sigma(x_i - \overline{x})^2 \quad (3.5)$$
$$S_{xy} = \Sigma(x_i - \overline{x})(y_i - \overline{y})$$
$$\widehat{b}_1 = \frac{S_{xy}}{S_{xx}}$$

여기서 관측된 값 y_i와 추정된 값 $\widehat{y}_i$사이의 차를 잔차(residual)라 하는데 i번째 잔차는 식 (3.6)과 같이 나타낼 수 있다. 잔차는 회귀분석에서 모형의 타당성이나 가정의 타당성을 검토하는 데에 중요한 역할을 수행한다.

$$e_i = y_i - \hat{y_i} = y_i - (\hat{b_0} + \hat{b_1} x_i), \ i = 1, \cdots, n \quad (3.6)$$

위의 식들을 이용하여 [표 3]에 있는 독립변수(x)와 종속변수(y) 데이터에 대한 회귀직선을 추정하기 위한 회귀직선은 $\widehat{y} = 0.6x + 2$ 이다. 이 회귀직선에서 기울기는 0.6이다. 이 기울기의 의미는 속성 값 A가 이전의 입력 시점에서 0.6만큼 증가된 후에 속성 값 A가 입력된다고 해석할 수 있다. 관측된 값 y_i와 추정된 종속변수의 값 $\widehat{y}_i$, 그리고 관측된 값과 추정값의 차이를 나타내는 잔차 e_i는 아래의 [표 3]에서 확인할 수 있다.

표 3. 속성값 A에 대한 관측값, 예측값, 잔차

순서	독립변수(x)	종속변수(y)	예측 값($\widehat{y}_i$)	종속변수(y)-예측값($\widehat{y}_i$)=잔차(e_i)
1	2	3	3.2	−0.2
2	3	2	3.8	−1.8
3	2	4	3.2	0.8
4	4	4	4.4	−0.4
5	4	3	4.4	−1.4
6	3	5	3.8	1.2
7	5	7	5	2
8	7	6	6.2	−0.2
9	6	5	5.6	−0.6

[표 3]에서 관측값과 추정값의 차이인 에러값들의 전체 합은 식 (3.7)과 같다.

$$\sum_{i=1}^{n}(y_i - \widehat{y}_i) = \sum_{i=1}^{n} e_i = -0.6 \quad (3.7)$$

여기서 잔차들의 전체 합인 -0.6의 의미는 속성값 A가 입력되는 시점을 예측하였을 때, 속성값 A가 입력되는 시점을 더 크게 보거나 더 작게 볼 수는 있지만 결국에 속성값 A의 입력 시점에 대한 예측 에러는 -0.6이라는 것이다. 이때, 속성값 A가 입력되는 하나의 시점은 스트림 데이터에서의 하나의 지지도를 뜻한다. 그러므로 전체 스트림 데이터에서의 속성값 A에 대한 지지도는 -0.6만큼의 에러를 갖는다는 것이다. 즉, 속성값 A에 대한 지지도를 하나 더 적게 예측한다는 의미로 해석할 수 있다. 이것은 전체 지지도에서 아주 작은 값을 나타내므로 이 회귀직선은 정확도를 가지고 있다고 평가할 수 있다.

회귀모형이 추정된 후의 빈발 항목을 찾는 과정은 다음과 같다. 먼저, 현재 시점 t에 속성값 A가 입력이 됐을 때, 이전 시점 (t-1)과 현재 시점의 시간의 차를 d_1이라 하고 d_1을 회귀모형에 입력했을 때 출력되는 종속변수 y의 값을 d_2라 하자. 이때, d_2의 의미는 현재 시점 t와 속성값 A가 앞으로 입력될 시점인 (t+1) 사이의 차이를 뜻한다. 이렇게 출력된 종속변수 y의 값인 d_2가 출력되면 속성값 A의 지지도가 하나 증가하는 것이다. 다시 d_2는 독립변수가 되어 회귀모형에 입력되면 원하는 종속변수 y의 값이 출력된다. 이러한 과정을 되풀이하면 미래의 어느 시점에 대한 속성값 A의 지지도를 알 수가 있다. 만약, 속성값 A의 지지도가 사용자가 정의한 최소지지도를 만족하면 속성값 A는 빈발 항목이 되는 것이다.

Ⅳ. 실험 및 평가

본 논문에서는 연속적으로 들어오는 스트림데이터를 일차원 속성으로 변형 후 데이터에서 빈발항목을 추출하기 위한 방법을 제안하였다. 이 장에서는 제안한 FIMR(Frequent Item Mining Regression) 방법의 타당성을 비교 실험한다.

1. 실험 환경 및 방법

제안한 방법은 AMD 3000+ CPU 1.83GHz와 1GB 메

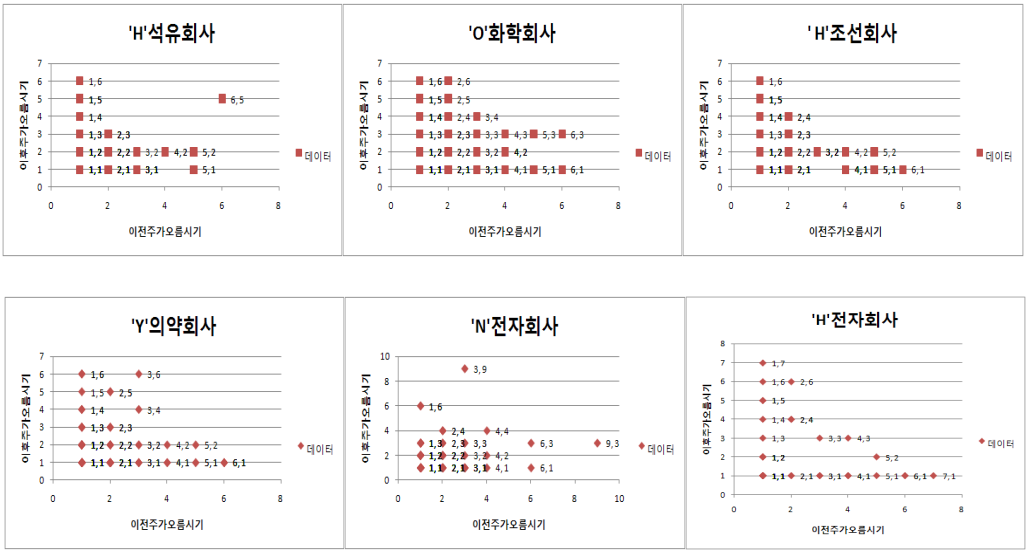

그림 3. 6개사의 주식데이터에 대한 산점도

모리를 가진 시스템 환경에서 실험되었다. 시스템 플랫폼은 Windows XP professional을 사용하였고 프로그래밍 구현 환경은 JDK 1.4.2를 사용하였다. 회귀모형을 추정하기위해 통계 패키지인 SPSS for Windows Version 12.0을 사용하였다. 통계 패키지 SPSS를 통해 회귀모형이 추정되면 JDK를 이용하여 회귀모형의 정확도를 측정하였다.

실험에 쓰인 데이터는 W증권회사에서 실시간으로 입력되어지는 6개 회사의 주식 데이터로써 7개월 동안 하루 전의 주식 가격과 비교하여 얼마만큼 내려가고 얼마만큼 올라갔는지에 대한 결과를 기록한 데이터를 이용하였다. 본 논문에서 제안하는 방법은 주식 가격의 예측이 아닌 실시간데이터에 대한 빈발 항목 예측에 대한 문제로 주식 가격이 오른 시점을 한 번의 지지도로 간주하고 실험하였다. 실험에 쓰인 6개 회사의 독립변수와 종속변수와의 산점도는 [그림 3]에 나타난다.

x축 “이전 주가 오름일”은 독립변수로써 현시점에서 며칠 전에 주가가 올랐던가를 나타내고 y축 “이후 주가 오름일”은 종속변수로써 현시점에서 며칠 후에 주가가 올랐는지를 나타낸다. [그림 3]에서 볼 수 있듯이, 표본으로 쓰인 6개의 주식 가격 오름은 10일 이내의 간격에서 다양한 차이를 보이고 있다. 실험 방법은 6개 회사, {O화학회사, H석유회사, H조선회사, Y의약회사, H전자회사, N전자회사}, 의 과거 7개월 동안의 주식 데이터를 가지고 실험하였다. 7개월 동안의 주식 데이터에서 4개월 동안의 주식 데이터를 가지고 회귀모형을 추정하고 추정된 회귀모형의 정확도를 평가하기 위해 이후 3개월 동안의 데이터를 이용하였다. 회귀분석을 위한 독립변수와 종속변수의 정의는 다음과 같다. 먼저, 독립변수는 어느 회사의 주식이 현 시점에서 며칠 전에 올랐을 때, 그 며칠이 독립변수가 된다. 그리고 종속변수는 현 시점에서 며칠 후에 올랐을 때, 그 며칠이 종속변수가 된다.예를 들어, O화학회사의 주식이 3일 전에 올랐고 오늘 올랐으면 독립변수의 값은 3이 된다. 또, 만약 O화학회사의 주식이 5일 뒤에 올랐다면 그때의 종속변수의 값은 5가 된다. 회귀모형의 정확도 측정은 과거 4개월 동안의 주식 가격 변동 데이터로부터 추정된 회귀모형을이용하여 이후 3개월 동안 주식 가격이 오른 횟수를 계산하고 계산된 횟수와 실제 주식 가격의 오른 경우의 수를 비교하였다. [표 4]는 6개 회사의 주식 데이터로부터 회귀모형을 추정한 것이다.

[표 5]는 관측된 값 y_i와 추정된 값 $\widehat{y}_i$, 그리고 잔차 e_i를 나타낸다. 이러한 실험값들은 추정된 회귀모

형이 좋은 모형인지 아닌지를 판단하기 위해 사용될 수 있다.

표 4. 6개사의 추정된 회귀모형

주식회사	절편 추정량 ($\widehat{b}_0$)	기울기 추정량($\widehat{b}_1$)	회귀직선
O'화학회사	2.213	−0.099	$\widehat{y} = -0.099x + 2.213$
H'석유회사	1.952	−0.056	$\widehat{y} = -0.056x + 1.952$
H'조선회사	2.335	−0.226	$\widehat{y} = -0.226x + 2.335$
Y'의약회사	2.175	−0.139	$\widehat{y} = 0.139x + 2.175$
H'전자회사	1.630	0.138	$\widehat{y} = 0.138x + 1.630$
N'전자회사	2.370	−0.171	$\widehat{y} = -0.171x + 2.370$

2. 실험 결과

모형의 진단 여부에 대하여 잔차는 데이터와 추정된 값 사이의 차이이기 때문에 회귀모형에 의해 설명되지 않는 변동을 나타내는 척도로써, 잔차를 분석함으로써 오차항에 대한 것을 포함하여 모형의 타당성 여부에 대한 정보를 얻을 수 있다. 이때 많이 사용되는 잔차통계량으로는 표준화잔차(standardized residual)와 스튜던트화잔차(studentized residual) 등이 있다[19]. 잔차를 분석하여 잔차통계량인 표준화잔차나 스튜던트화잔차를 얻었을 때, 정당한 모형은 표준화잔차와 스튜던트화잔차가 각각 근사적으로 평균이 0이고 분산은 1이 된다. [표 6]은 6개 표본 주식데이터에 대한 회귀분석을 통해 추정된 각 회귀모형이 정당하게 추정된 것인지를 확인하기 위해 표준화잔차와 스튜던트화잔차를 나타낸 것이다.

표 5. 관측값(y_i), 추정된 값($\widehat{y}_i$), 에러(e_i)

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	비고
O화학회사	y_i	2.0	1.0	3.0	1.0	5.0	1.0	1.0	1.0	2.0	2.0	3.0	2.0	3.0	2.0	1.0	1.0	1.0	1.0	5.0	3.0	
	$\widehat{y}_i$	1.82	2.02	2.11	1.92	2.11	1.72	2.11	2.11	2.11	2.02	2.02	1.92	2.02	1.92	2.02	2.11	2.11	2.11	2.11	1.72	
	e_i	0.18	−1.02	0.89	−0.92	2.89	−0.72	−1.11	−1.11	−0.11	−0.02	−0.98	0.08	0.98	0.08	−1.02	−1.11	−1.11	−1.11	2.89	1.28	
H석유회사	y_i	2.0	1.0	1.0	1.0	4.0	2.0	2.0	2.0	1.0	5.0	2.0	1.0	1.0	1.0	3.0	1.0	1.0	1.0	1.0	2.0	
	$\widehat{y}_i$	1.78	1.84	1.9	1.9	1.9	1.73	1.84	1.84	1.84	1.9	1.67	1.84	1.9	1.9	1.9	1.78	1.9	1.9	1.9	1.9	
	e_i	0.22	−0.84	−0.9	−0.9	2.1	0.27	0.16	0.16	−0.84	3.1	0.33	−0.84	−0.9	−0.9	1.1	−0.79	−0.9	−0.9	−0.9	0.1	
H조선회사	y_i	2.0	1.0	1.0	5.0	1.0	1.0	1.0	2.0	1.0	1.0	3.0	2.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	5.0	
	$\widehat{y}_i$	1.89	1.89	2.11	2.11	1.2	2.11	2.11	2.11	1.89	2.11	2.11	1.66	1.89	2.11	2.11	2.11	1.89	2.11	2.11	2.11	
	e_i	0.12	−0.89	−1.11	2.89	−0.2	−1.11	−1.11	−0.11	−0.89	−1.11	0.89	0.34	−0.88	−1.11	−1.11	−0.11	−0.88	−1.11	−1.11	2.89	
Y의약회사	y_i	1.0	3.0	1.0	2.0	1.0	1.0	1.0	2.0	1.0	2.0	5.0	2.0	1.0	1.0	1.0	2.0	1.0	6.0	1.0	1.0	
	$\widehat{y}_i$	2.04	2.04	1.76	2.04	1.9	2.04	2.04	2.04	1.9	2.04	1.9	1.48	1.9	2.04	2.04	2.04	1.9	2.04	1.34	2.04	
	e_i	−1.04	0.96	−0.76	−0.04	−0.9	−1.04	−1.04	−0.04	−0.9	−0.04	3.1	0.52	−0.9	−1.04	−1.04	−0.04	−0.9	3.96	−0.34	−1.04	
H전자회사	y_i	2.0	1.0	1.0	1.0	1.0	1.0	1.0	3.0	1.0	2.0	1.0	2.0	1.0	2.0	1.0	1.0	2.0	1.0	1.0	2.0	
	$\widehat{y}_i$	1.91	1.91	1.77	1.77	1.77	1.77	1.77	1.77	2.04	1.77	1.91	1.77	1.91	1.77	1.91	1.77	1.77	1.91	1.77	1.77	
	e_i	0.09	−0.91	−0.77	−0.77	−0.77	−0.77	−0.77	1.23	−1.04	0.23	−0.91	0.23	−0.91	0.23	−0.91	−0.77	0.23	−0.91	−0.77	0.23	
N전자회사	y_i	1.0	1.0	1.0	1.0	1.0	3.0	3.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	7.0	1.0	1.0	2.0	6.0	1.0	
	$\widehat{y}_i$	2.2	2.2	2.2	2.2	2.2	2.2	1.86	1.86	2.2	2.2	2.2	2.03	2.2	2.2	2.2	1.17	2.2	2.2	2.03	1.34	
	e_i	−1.2	−1.2	−1.2	−1.2	−1.2	0.8	1.14	−0.86	−1.2	−1.2	−0.2	−1.03	−1.2	−1.2	4.8	−0.17	−1.2	−0.2	4.97	−0.34	

표 6. 6개 표본에 대한 잔차통계량

통계치 / 표본	표준화 잔차		스튜던트화 잔차	
	평균	표준편차	평균	표준편차
O'화학회사	0.000	0.994	0.000	1.004
H'석유회사	0.000	0.994	0.003	1.014
H'조선회사	0.000	0.993	0.000	1.003
Y'의약회사	0.000	0.993	−0.001	1.003
H'전자회사	0.000	0.994	0.000	1.005
N'전자회사	0.000	0.988	−0.001	1.004

[표 6]에서 볼 수 있듯이, 각 표본의 주식데이터에 대한 표준화잔차와 스튜던트화잔차의 평균과 분산이 각각 0과 1에 근사하다는 것을 알 수 있다. 표에는 표준편차를 표시하였지만 분산의 제곱이 표준편차이므로 결국 모두 1에 근사하다는 것을 알 수 있다. 그러므로 6개 표본의 주식데이터에 대해 추정된 회귀모형은 모두 정당하다고 할 수 있다. [그림 4]는 실제 6개 표본 데이터에서 매달 주가가 오른 회수와 표본 데이터로부터 추정된 회귀모형을 이용하여 추정한 매달 추정 회수를 비교한 것이다.

V. 결론

본 논문에서는 스트림데이터 환경에서 입력되는 데이터를 통해 전송되는 일차원 속성의 스트림 데이터로부터 회귀 모델을 적용하여 빈발 항목들을 마이닝하기 위한 기법을 제안하였다. 회귀분석은 두 개 혹은 그 이상의 변수들 간의 관계를 연구하고, 이 관계를 모형화함으로써 다른 변수들을 이용하여 어느 변수에 대한 예측을 가능하게 해주는 통계 방법 중의 하나이다. 본 논문에서의 회귀모형은 선형 회귀모형으로써 두 개의 변수의 변화에 대한 회귀모형을 추정하였다.

스트림 데이터는 시간에 따라 연속적으로 전송되어지는 특징을 가지고 있기 때문에 시간에 따른 변화량을 두 개의 변수로 설정하였다. 두 개의 변수는 각각 독립변수와 종속변수를 나타내는데, 독립변수는 현시점에서 한 시점 이전의 데이터 입력 시간과의 차이를 나타내고 종속변수는 현시점에서 한 시점 이후의 데이터 입력 시간과의 차이를 나타낸다. 이 두 변수간의 관계로부터 회귀모형을 추정함으로써, 어떤 값이 입력이 됐을 때, 그 값의 이후의 입력 시점을 예측할 수 있다. 그러므로 스트림 데이터 항목에 대한 지지도는 이 회귀모형을 이용함으로써 계산될 수 있다.

본 논문에서 제안하는 회귀모델 기반 빈발항목 마이닝 방법에 대한 타당성과 정확성을 검증하기 위해 실시간으로 입력되어지는 주식데이터를 이용하여 실험하였다. 실험을 통해 회귀모델에 대한 유효성을 검증하고 이 회귀모델을 이용하여 스트림 데이터 항목의 지지도를 예측하는데 큰 에러가 발생하지 않는다는 것을 실험을 통해 확인하였다.

향후에는 일차원 속성뿐만 아니라 다차원 속성 데이터에 대한 빈발 항목 마이닝 방법에 대한 연구가 이어

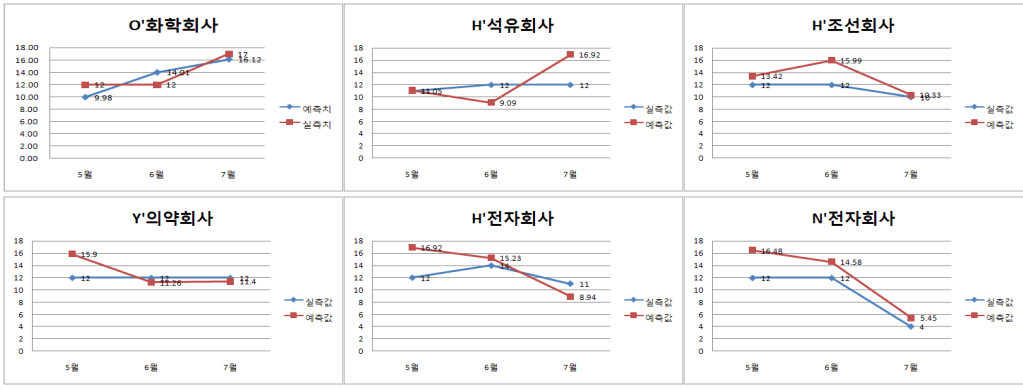

그림 4. 6개사의 주식데이터에 대한 산점도

져야 한다. 또한, 통계적 방법을 이용하여 스트림 데이터에 대한 빈발 항목집합 마이닝 방법으로의 연구가 진행 중에 있다. 본 논문에서는 일차원 스트림 데이터로부터 하나의 회귀모형을 추정하고 이를 이용하였지만, 회귀모형의 정확도를 위해서는 시간의 변화에 따라 변화되는 데이터의 분포를 고려하여 회귀모형도 수정이 되어져야 한다. 따라서 스트림 데이터의 분포를 고려하여 점진적으로 회귀모델을 변화시키는 방법도 연구가 되어져야 한다.

참 고 문 헌

[1] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," In Proc. of Very Large Data Bases, pp.487-499, 1994.

[2] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and Issues in Data Stream Systems," In Proc. of PODS, 2002(3).

[3] G. Chen, X. Wu, and X. Zhu, "Mining Sequential Patterns Across Data Streams," Univ. of Vermont Computer Science Technical Report(CS-05-04), 2005(3).

[4] N. Davey, S. P. Hunt, and R. J. Frank, "Time Series Prediction and Neural Networks," In Journal of Intelligent and Robotic Systems, 2001.

[5] M. J. Franklin and S. R. Jeffery etc., "Design Considerations for High Fan-In System: The HiFi Approach," Conference on Innovative Data Systems Research, pp.290-304, 2005.

[6] C. Giannella, J. Han, J. Pei, X. Yan, and P. S. Yu, "Mining Frequent Patterns in Data Streams at Multiple Time Granularities," In H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yeshar(eds.), Next Generation Data Mining, AAAI/MIT, 2003.

[7] L. Golab, M. Tamer Ozsu, "Issues in Data Stream Management," In SIGMOD Record, Vol.32, No.2, 2003.

[8] H. Han, H. Ryoo, and H. Patrick, "An Infrastructure of Stream Data Mining, Fusion and Management for Monitored Patients," In Proc. of 19th IEEE International Symposium on CBMS 2006, pp.461-468, 2006(6).

[9] X. Hao and D. Xu, "Time Series Prediction based on Non-Parametric," In SIGMOD Record, Vol.32, No.2, 2003.

[10] H. Li, S. Lee, and M. Shan, "Online Mining (Recently) Maximal Frequent Itemsets over Data Streams," In Proc. of RIDE-SDMA'05, pp.11-18, 2005(4).

[11] R. C. Olover and K. Smettem, "Field Testing a Wireless Sensor Network for Reactive Environmental Monitoring," Intelligent Sensors, Sensor Networks and Information Processing, pp.7-12, 2004.

[12] J. Pei, J. Han, and R. Mao, "CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets," In Proc. of 2000 ACM SIGMOD International Workshop Data Mining and Knowledge Discovery, pp.11-20, 2000.

[13] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, "Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach," IEEE Transactions on Knowledge and Data Engineering, Vol.16, No.11, 2004(11).

[14] S. Sarkka, A. Vehtari, and J. Lampinen, "Time Series Prediction by Kalman Smoother with Cross-Validated Noise Density," In Proc. of IJCNN, pp.1653-1658, 2004.

[15] D. F. Specht, "A General Regression Neural Network," IEEE Trans. on Neural Networks, Vol.2, No.6, pp.568-576, 1991(11).

[16] M. J. Zaki and C. J. Hsiao, "CHARM: An Efficient Algorithm for Closed Itemset Mining,"

In Proc. 2002 SIAM International Conference Data Mining, pp457-473, 2002.

[17] B. Xu. and O. Wolfson, "Time-Series Prediction with Application to Traffic and Moving Objects Databases," ACM Workshop on Data Engineering for Wireless and Mobile Access, pp.56-60, 2003.

[18] O. B. Yaik, C. H. Yong, and F. Haron, "Time Series Prediction using Adaptive Association Rules," In Proc. of DFMA05, pp.310-314, 2005.

[19] 김현철, "SPSS for Windows에 의한 실용회귀분석"

저 자 소 개

이 욱 현(Uk-Hyun Lee) 정회원

- 1992년 2월 : 이화여자대학교 전자계산학과(이학사)
- 1997년 2월 : 한국과학기술원 정보및통신공학과(공학석사)
- 2003년 2월 : 전남대학교 전산학과(공학박사)
- 2004년 3월 ~ 현재 : 한북대학교 컴퓨터정보학과 교수

<관심분야> : 데이타베이스, 데이터마이닝, 콘텐츠IT 기술