J. of the Korean Sensors Society
Vol. 15, No. 6 (2006) pp. 433 – 441

주기 패턴을 이용한 센서 네트워크 데이터의 이상치 예측

김 형 일†

Outlier prediction in sensor network data using periodic pattern

Hyung-Il Kim†

Abstract

Because of the low power and low rate of a sensor network, outlier is frequently occurred in the time series data of sensor network. In this paper, we suggest periodic pattern analysis that is applied to the time series data of sensor network and predict outlier that exist in the time series data of sensor network. A periodic pattern is minimum period of time in which trend of values in data is appeared continuous and repeated. In this paper, a quantization and smoothing is applied to the time series data in order to analyze the periodic pattern and the fluctuation of each adjacent value in the smoothed data is measured to be modified to a simple data. Then, the periodic pattern is abstracted from the modified simple data, and the time series data is restructured according to the periods to produce periodic pattern data. In the experiment, the machine learning is applied to the periodic pattern data to predict outlier to see the results. The characteristics of analysis of the periodic pattern in this paper is not analyzing the periods according to the size of value of data but to analyze time periods according to the fluctuation of the value of data. Therefore analysis of periodic pattern is robust to outlier. Also it is possible to express values of time attribute as values in time period by restructuring the time series data into periodic pattern. Thus, it is possible to use time attribute even in the general machine learning algorithm in which the time series data is not possible to be learned.

Key Words : Sensor network, outlier, logistic regression, C4.5, pattern analysis

1. 서 론

최근에는 언제 어디서나 어떤 기기로도 네트워크에 접속할 수 있는 유비쿼터스 네트워킹(ubiquitous networking) 환경이 부각되고 있다. 센서 네트워크(sensor network)는 이런 유비쿼터스 네트워킹, 홈 네트워크(home network) 등의 기반이 되는 기술이다[1,2].

일반적으로 센서 네트워크는 무선 센서 네트워크(wireless sensor network)를 말한다. 무선 센서 네트워크는 무선 인터페이스를 활용한 통신 기능과 간단한 처리 기능을 가진 다수의 센서 노드들로 구성된 네트워크이다[3]. 그리고 센서 노드들은 전력 소모량이 작고 비용이 저렴한 특징을 갖는다. 이런 센서 노드들은 목표하는 공간에 무작위로 배치되어 자율적으로 네트워크를 형성한다.

센서 네트워크는 산업, 홈 네트워킹 등 응용에 따라 다양한 분야에 적용될 수 있다는 장점을 갖는다[4-7]. 예를 들어 센서 네트워크는 밀림이나 바다와 같이 사람이 직접 접근하기 어려운 지역의 데이터를 실시간으로 수집할 수 있다. 또한 센서 네트워크는 센서 노드들을 활용하여 목표 감시(surveillance) 및 추적(tracking), 신호처리(signal processing) 등의 기능을 수행할 수 있다.

그러나 센서 네트워크는 패킷 손실에 따른 결측치 발생, 신호 간섭에 따른 이상치 발생, 효율적인 전력 제어의 필요성 등 몇 가지 기술적인 문제점들을 지니고 있다[8-11]. 이 문제점들은 제한적인 전력 자원, 저전력(low power) 소비 등 센서 노드가 갖는 고유한 특성으로부터 발생한 것이다. 실제 인텔 버클리 연구소(Intel Berkeley lab)에서 제공하는 센서 네트워크의 데이터를 분석한 결과 약 20 % 정도의 데이터가 결측치(missing value)거나 이상치(outlier)라 보고하였다[11]. 비록 센서 네트워크의 활용도가 높을지라도 데이터 수

동국대학교 컴퓨터공학과(Dept. of Computer Engineering, Dongguk University)

†Corresponding author: hikim@dongguk.edu
(Received : June 19, 2006, Accepted : September 25, 2006)

집 시에 결측치나 이상치가 빈번하게 발생한다면 수집한 데이터를 이용할 수 없다.

이상치 추출에는 통계적 기법이나 기계학습 기법이 주로 이용된다[8,12-13]. 특히 기계학습 알고리즘은 데이터의 특성에 맞는 모델을 생성하기 때문에 뛰어난 성능을 발휘한다. 그러나 시계열 데이터와 같이 속성 값이 연속적인 경우에는 학습을 수행할 수 없다는 단점이 있다.

본 논문에서 제안한 이상치 예측 방법은 크게 두 단계로 구성된다. 첫 번째 단계에서는 시계열(time series) 특성을 지닌 센서 네트워크 데이터에 대해 시간에 따른 센서 신호의 주기 패턴을 분석한다. 두 번째 단계에서는 결과로 얻어진 주기 패턴을 활용해 시계열 데이터를 주기 패턴 데이터로 변형하여 이상치 예측에 활용한다. 본 논문에서 제안한 이상치 예측 기법은 먼저 시계열 데이터에 양자화(quantization)와 평활화(smoothing)를 적용한 후, 속성 값의 추세 변동에 따라 데이터를 변형한다. 그리고 변형된 데이터에서 주기 패턴을 분석하고, 주기 패턴 내에서 시간 값을 재구성함으로써 주기 패턴 데이터를 생성한다. 이상치 예측은 주기 패턴 데이터를 활용하여 통계적 기법인 로지스틱 회귀(logistic regression)와 기계학습 기법인 C4.5에 적용하였다. 몇 가지 실험을 통해 주기 패턴 데이터를 활용할 경우 이상치 예측에 뛰어난 성능을 보임을 확인할 수 있었다.

본 논문에서 제안한 주기 패턴 분석은 데이터의 값 또는 변화량의 크기를 활용하는 것이 아니라 데이터 값의 추세를 분석하여 주기 패턴을 활용하므로 임의로 발생하는 이상치나 결측치에 대하여 강인성을 지닌다. 또한, 시계열 데이터를 주기별로 재구성하여 주기 패턴 데이터를 생성함으로써 일반적인 기계학습 알고리즘에도 적용할 수 있다.

2. 관련 연구

센서 네트워크는 임의 공간에 무작위로 배치된 센서 노드들로 구성된 네트워크이다. 각 센서 노드들은 환경으로부터 온도, 습도, 광량, 방사능뿐만 아니라 생체 기관의 상태, 지질 특성, 지진 진동 등의 다양한 종류의 정보를 감지하는 기능을 지닌 작은 전자 컴포넌트이다[3]. 이런 센서 노드들은 다른 노드들 및 외부 세계와 서로 통신하는 기능을 지닌다. 최근엔 기술의 발전으로 강력한 기능을 제공하는 동시에 작고 에너지 효율적인 센서들이 대량으로 생산되고 있다[14].

이런 특징들을 갖는 센서 네트워크는 광범위한 활용 가능성을 갖는다. 현재 센서 네트워크는 RFID, 유비쿼터스 등과 결합하여 차세대 인프라 구조의 하나로 그 활용성을 주목받고 있다. 또한 디지털 홈 혹은 군사 목적 시스템과 같이 사람 또는 주위의 사물들을 인식하고 위치 정보를 파악하는 등의 특수 목적에도 활용할 수 있다[6,15]. 이 외에도 인텔리전트 빌딩의 자동 환기, 무인 경비 시스템과 오염 물질이 산재된 공장의 작업장 내 환기 및 개폐 장치, 차량 이동 장치 또는 가정용 자동 온도 조절기 등에 이용이 가능하고, 장난감, 게임기, 가전제품 디바이스 및 PC 주변기기에서도 다양하게 이용될 수 있다.

센서 네트워크의 기본 형태는 센서 노드와 센서 노드로부터 무선 통신으로 정보를 수집하는 서버로 구성된다. 그리고 서버가 센서 노드로부터 수집한 데이터를 센서 네트워크 데이터라 한다. 서버는 일정한 시간 간격이나 특정한 이벤트의 발생에 따라 센서 노드로부터 데이터를 수집하므로 센서 네트워크 데이터는 시계열 특성을 갖는다.

센서 네트워크 데이터는 무선 통신을 활용하여 수집할 때 많은 오류 정보를 포함한다[9,16]. 이것은 불완전한 센서 측정, 낮은 전력, 센서 서로간의 혼선 등의 이유로 센서 노드가 잘못된 정보를 보고할 수 있기 때문이다.

결측치나 이상치와 같은 부정확한 데이터는 데이터로부터 정보를 추출하는데 있어서 결과를 왜곡시킬 수 있다. 그러므로 수집한 데이터를 가공하여 유용한 정보를 추출하기 위해서는 데이터의 값들을 검토하는 데이터 정제(data cleaning) 과정이 필요하다[9,13].

Elnahrawy 등은 센서 네트워크에서 발생하는 이상치, 결측치에 관한 데이터 품질 문제를 해결하기 위해 센서 값의 시 · 공간적 상관관계를 나이브 베이즈(Naive Bayes) 분류 알고리즘으로 학습하고 모델링하는 방법을 제안하였다[17].

Janakiram 등은 베이지안 네트워크를 활용하여 이상치를 탐지하는 시스템을 제안하였다[18]. 이 시스템은 센서 노드에서 측정하는 온도, 습도, 기압, 광량, 전압을 속성으로 활용하여 베이지안 네트워크를 학습한다. 이때 각 속성의 센서 값은 연속형 값이므로 센서 값을 구간화하여 클래스로 사용한다. 그리고 이상치 판별은 센서 노드에 대한 베이지안 네트워크의 분류 값과 실제 센서 값의 일치 여부 통해 수행한다.

Tanachaiwiwat 등은 센서 네트워크의 이상치를 보안 측면에서 분석하고 탐색하기 위해 ART(Abnormal Relationships Test) 기법을 제안하였다[19]. 센서 네트워크의 이상치는 센서 노드의 제약에 따라 자연적으로 발생하

는 비정상적인 값일 뿐만 아니라 누군가에 의해 인위적으로 조작된 값일 수도 있기 때문이다. ART 기법은 슬라이딩 윈도우(sliding windows) 구간 내의 센서 값들과 인접한 센서 노드들의 센서 값에 대한 상관관계 분석을 활용하여 이상치를 탐색한다.

센서 네트워크는 활용 가능성과 적용 범위가 다양할지라도 센서 네트워크에서 수집한 데이터에 이상치나 결측치와 같은 노이즈가 존재한다면 수집한 데이터로부터 추출한 정보를 신뢰할 수 없다. 본 논문에서 제안한 주기 패턴 분석 기법은 시계열 데이터의 주기 패턴을 분석하여 주기 패턴 데이터를 생성하고, 주기 패턴 데이터를 이상치 예측에 활용함으로써 우수한 예측 성능을 나타내며, 시계열 데이터도 기계학습에 적용될 수 있도록 가공할 수 있는 장점이 있다.

3. 주기 패턴 분석과 이상치 예측

시계열 데이터는 시간을 표현하는 속성 값의 범위에 한계가 존재하지 않으므로 기계학습에 적용할 수 없다. 시계열 데이터에 기계학습 알고리즘을 적용하기 위해서는 시간을 표현하는 속성 값을 기계 학습 알고리즘이 활용할 수 있는 형태로 변형해야 한다. 본 논문에서 제안한 기법은 시계열 데이터를 기계학습에 적용할 수 있도록 한다. 본 장에서는 본 논문에서 제안한 주기 패턴 분석 기법을 주기 패턴 분석과 이상치 예측으로 나누어 소개한다.

3.1. 주기 패턴 분석

시계열 데이터에 대한 주기 패턴 분석의 목적은 시간의 흐름에 따라 데이터 값의 추세가 반복하는 시간 주기를 찾는 것이다. 본 논문에서는 센서를 통해 온도, 습도, 광량, 전압을 측정하고 수집한 시계열 데이터를 활용한다. 그리고 수집한 초기 상태의 시계열 데이터를 원시 데이터라 할 때, 원시 데이터의 값이 시간에 따라 변화하는 형태는 그림 1과 같다.

그림 1에 나타난 바와 같이 각 센서에서 측정한 데이터의 값들은 시간 흐름에 따라 유사한 주기를 가지고 반복한다. 시간 주기를 찾기 위해 주기 패턴 분석은 양자화와 평활화를 적용한 뒤 추세 변동 데이터를 생성한다. 그리고 추세 변동 데이터를 분석하여 주기 패턴을 찾는다.

주기 패턴 분석을 위해서는 먼저 일정한 시간 간격으로 데이터를 다시 표현하는 양자화 작업을 선행한다. 양자화를 적용하는 이유는 데이터가 정량적인 시간 구

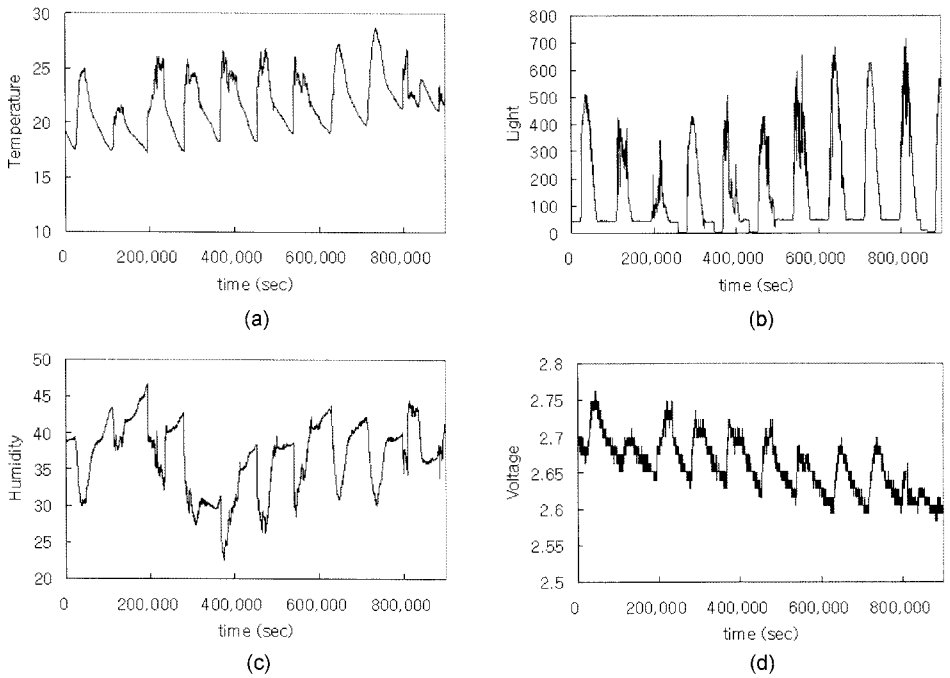


그림 1. 원시 센서 데이터의 주기: (a) 온도, (b) 광량, (c) 습도, (d) 전압
Fig. 1. Period of original sensor data: (a) temperature, (b) light, (c) humidity, (d) voltage.

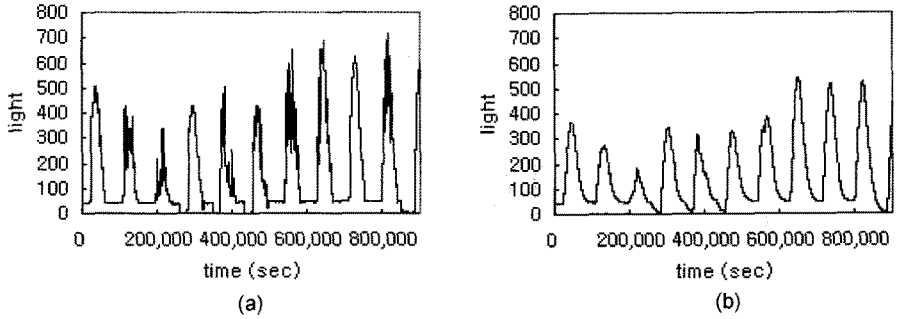


그림 2. 양자화 및 평활화 수행 전과 후의 센서 데이터: (a) 수행 전 (b) 수행 후
Fig. 2. Sensor data before and after applying quantization and smoothing: (a) before (b) after.

간으로 이루어져야 규칙적으로 반복하는 데이터의 주기 패턴을 결정하기 용이하기 때문이다.

평활화는 변화 폭이 심한 데이터를 평탄하고 변화가 완만한 값으로 변환시킴으로써 변화에 대한 추세를 보다 명확히 표현하기 위해 수행한다. 평활화 기법으로는 식 (1)에 표현한 지수 평활법(exponential smoothing method)을 사용한다. 지수 평활법은 현재 시점의 데이터 값과 과거의 데이터 값들에 대한 함수 관계로 표현한다. 식 (1)에서 t는 양자화 데이터의 타임스탬프를 표현하며, Q_t는 타임스탬프 t 시점에서 데이터 값을 표현한다. S_t는 지수 평활 값을 표현하며, 초기값 S_1은 Q_1을 사용한다. a는 평활상수(smoothing constant)로 0과 1 사이의 값을 갖는다. 평활 상수 a가 0에 가까울수록 S_t는 t 시점에서 멀리 떨어진 과거의 데이터 값에 영향을 받으며, 1에 가까울수록 가까운 과거의 데이터 값에 영향을 받는다. 그림 2는 광량 센서 데이터에 양자화와 평활화를 수행하기 전과 후의 결과이다. 원시 데이터보다 추세의 표현이 명확하다는 것을 확인할 수 있다.

$$S_t = aQ_t + (1-a)S_{t-1} \quad (0<a<1,\ t>2) \tag{1}$$

다음 단계는 추세 변동 데이터로 변형하는 과정이다. 데이터 변형의 목적은 데이터 값의 크기가 변화하는 패턴을 찾기 위함이 아니라 데이터 값의 추세가 변화하는 시간적 반복 주기를 찾기 위함이다. 데이터 변형은 두 개의 단계를 거쳐 수행된다. 첫 번째 단계는 양자화한 타임스탬프의 구간별로 데이터의 변화량을 측정한다. 데이터의 변화량은 변동(fluctuation)이라 표현하며, 변화의 크기에 상관없이 상승은 '1', 하강은 '-1', 변동 없음은 '0'으로 나타낸다. 두 번째 단계는 그림 3과 같이 인접한 세 개의 변동 값을 결합하여 새로운 추세 변동 값으로 변환한다.

그림 3에서 i는 양자화 타임스탬프 사이의 구간을

```
for i = 1 to n-1
  if(f_{i-1} + f_i + f_{i+1}) > 0 then transform to 1
  else if(f_{i-1} + f_i + f_{i+1}) = 0 then transform to 0
  else if(f_{i-1} + f_i + f_{i+1}) < 0 then transform to -1
  3 added to i
end
```

그림 3. 양자화 구간별 변동을 이용한 데이터 변형
Fig. 3. Data transform using fluctuation of quantization interval.

표현하며, f_i는 양자화 구간별 변동 값을 표현한다. 양자화를 통하여 데이터의 시간 구간이 일정하게 나뉘었으므로 세 변동의 합으로 표현된 추세 변동 값은 일정 구간에 대한 변동을 나타낸다. 그리고 변화량의 크기를 일반화했기 때문에 데이터 값의 크기에 민감하지 않은 변동의 추세를 구할 수 있다. 즉, 데이터를 추세 변동 데이터로 변형함으로써 이상치의 영향을 적게 받는다.

그림 4는 추세 변동 데이터로 변형한 결과이다. 각 그림의 위쪽에서부터 온도, 습도, 광량, 전압의 속성 값에 대한 추세 변동 그래프를 보여준다. 각 그래프의 구성 요소는 변동의 상승(+1), 하강(-1), 변화 없음(0)으로 구성되지만 실질적으로 변화 없음의 경우는 거의 발생하지 않는다. 상승은 각 추세 변동 그래프의 위쪽 라인으로 표현되며, 하강은 아래쪽 라인으로 표현되었다.

주기는 추세 변동 데이터에서 상승 추세와 하강 추세가 한번씩 나타나는 시간 구간을 나타낸다. 주기 패턴은 1개 이상의 주기들로 구성된 주기 집합으로써 추세 변동 데이터에서 연속적이고 반복적으로 나타나는 주기 집합들 중에서 최소 시간 구간을 갖는 경우이다. 한 가지 속성만 존재하는 추세 변동 데이터에서 주기 패턴은 일반적으로 하나의 주기를 나타낸다. 만약 여러

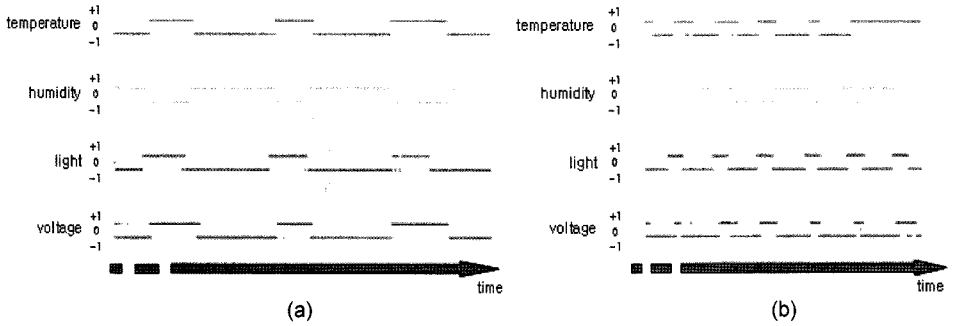


그림 4. 데이터 변형 후 주기 패턴: (a) 안정적인 주기 패턴 (b) 불안정한 주기 패턴
Fig. 4. The periodic patterns of fluctuation after data transform: (a) stable periodic pattern (b) unstable periodic pattern.

가지 속성이 존재할 때는 속성별로 주기 패턴을 찾고, 그 중에서 가장 긴 주기 패턴을 속성들의 기본 주기 패턴으로 설정한다. 주기가 규칙적인 경우에는 주기 패턴이 명료하게 나타나며, 주기 패턴이 안정되었다고 표현한다. 반면 각 주기들이 불규칙적이어서 주기 패턴이 명료하지 않은 경우에는 주기 패턴이 불안정하다고 표현한다. 그림 4(a)의 경우 안정적인 주기 패턴을 찾을 수 있었으며, 추출한 주기 값은 1일에 근사하였다. 그림 4(b)는 센서의 전압이 낮아져 광량을 제외한 대부분의 값이 이상치로 나타나는 영역으로 주기가 불안정하다.

주기 패턴 분석은 먼저 전체 속성들이 안정적인 주기를 보이는 구간을 활용하여 주기 패턴을 분석하고 분석한 주기 패턴을 나머지 데이터에 대하여 확장하여 적용한다. 만약 전체 속성들이 안정적인 주기를 보이는 구간이 없을 때는 안정적인 주기를 보이는 몇 개의 속성만을 이용하여 주기 패턴을 분석한다. 그림 4(b)에서는 광량 속성만을 이용하여 주기를 분석할 수 있다.

주기 패턴 데이터는 시계열 데이터를 주기별로 재구성한 데이터이다. 먼저 데이터를 주기 패턴의 주기별로 분할한다. 그리고 분할한 데이터의 시간을 표현하는 속성 값을 주기 내의 상대적인 시간 값으로 재구성한다. 즉, 주기의 시작 시에는 시간 값을 0으로 초기화한다. 그림 5는 하나의 주기에 표현된 데이터들이다. 그림 5(a)는 광량의 센서 값을 나타낸 것이며, 거의 일정 영역에서만 값이 오르고 내림을 보인다. 그림 5(b)는 온도의 센서 값을 나타낸 것이며, 불규칙적인 신호 모양을 갖지만 전체적으로 일률적인 신호 구간을 갖는다. 그림 5(b)에 나타낸 온도 데이터의 경우 서로 간에 불규칙적인 모양을 하고 있기 때문에 그래프 패턴만으로는 이것들을 구분하기 어렵다. 그러나 주기 패턴을 기준으로 처리하면 규칙적인 의미 구간을 발견할 수 있다. 그리고 주기라는 규칙성 있는 구간으로 시간을 재구성할 경우, 시계열 데이터를 기계학습에 활용할 수도 있다.

3.2. 이상치 예측

본 논문에서는 교사학습(supervised learning) 알고리즘인 로지스틱 회귀와 C4.5 결정트리를 활용하여 이상치 예측을 수행한다.

로지스틱 회귀는 선형 회귀를 기반으로 한 분류 알고리즘이다. 선형 회귀는 모든 속성 값이 수치 값이고

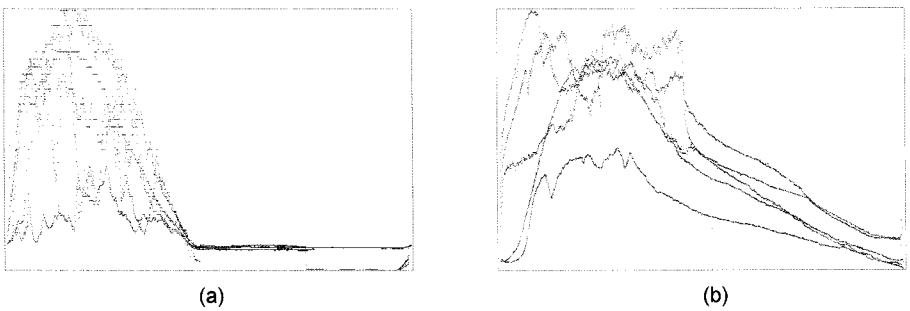


그림 5. 하나의 주기에 표현된 광량과 온도의 센서 데이터: (a) 광량 센서 데이터 (b) 온도 센서 데이터
Fig. 5. Sensor data of light and temperature expressed in a period: (a) sensor data of light. (b) sensor data of temperature.

수치 값 형태의 결과를 예측할 때 사용되는 대표적인 통계 기법 중에 하나이다. 선형 회귀는 속성들의 선형 조합에 따라 결과를 예측한다. 만약 k개의 속성 a가 존재할 때, 예측 결과 x에 대한 선형 회귀는 식 (2)와 같이 표현된다. 식 (2)에서 w는 가중치이다. 선형 회귀는 결과로 연속적인 수치 값만을 도출할 수 있다. 이 단점을 보완한 것이 로지스틱 회귀이다.

$$x = x_0 + w_1 a_1 + w_2 a_2 + \ldots + w_k a_k \quad (2)$$

로지스틱 회귀는 차량 구매 여부, 보험 가입 여부와 같이 두 개의 분류 결과를 예측하는 통계 기법이다. 로지스틱 회귀는 내부적으로 선형 회귀를 기반으로 한다. 분류하고자 하는 두 가지 범주가 c_1, c_2일 때, 식 (3)와 식 (4)과 같이 동일한 속성 집합 a에 대하여 각각의 선형 회귀 모델을 형성한다. 그리고 새로운 데이터가 들어왔을 때 각각의 모델을 통하여 결과를 예측한다. 분류는 예측 결과가 큰 쪽으로 이루어진다.

$$c_1 = w_0 + w_1 a_1 + w_2 a_2 + \ldots + w_k a_k \quad (3)$$

$$c_2 = w_0' + w_1' a_1 + w_2' a_2 + \ldots + w_k' a_k \quad (4)$$

C4.5 결정트리는 대표적인 기계학습 알고리즘들 중 하나이다. 결정트리 학습 알고리즘은 트리 구조로 형성되며, 데이터를 분할 정복(divide and conquer) 방식으로 분류하여 최종적인 분류 모델을 얻는 알고리즘이다. 결정트리 학습 알고리즘은 여러 가지가 존재하며, C4.5는 그 중에서 가장 대표적인 알고리즘 중에 하나이다. C4.5는 정보이론(Information Theory)의 불확실성(Entropy)와 정보 획득(Information Gain)을 기반으로 한다. 불확실성은 현재 상태에서 서로 다른 분류에 속하는 데이터들이 섞여있는 정도를 나타낸다. 그리고 정보 획득은 현재 상태에서 임의 속성 하나를 기준으로 데이터를 분류하였을 때 얻을 수 있는 불확실성의 감소량이다. 식 (5)와 식 (6)은 불확실성과 정보 획득을 구하는 방법이다. 수식에서 S는 전체 데이터 집합이고 c는 클래스를 나타내며 A는 속성을 나타낸다. p_i는 전체 데이터 집합 S에 대한 i번째 클래스 집합이 나올 확률이며, v는 A의 속성 값들을 의미한다. 그리고 S_v는 속성 A의 속성 값이 v인 데이터들의 집합이다.

$$Entropy(S) \equiv \sum_{i=0}^{c} (-p_i \log p_i) \quad (5)$$

$$Information\ Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (6)$$

4. 실 험

4.1. 실험 데이터

실험에는 Intel Berkeley Lab에서 제공하는 공개 데이터를 사용하였다[10]. 이 데이터는 건물 한 층의 곳곳에 54개의 센서 노드들을 배치하고 약 30초 간격으로 추출한 센서 값을 약 한 달 동안 수집한 데이터이다. 사용한 센서는 Mica2Dot이며 TinyOS와 TinyDB 플랫폼을 기반으로 하여 데이터를 수집하였다. 각각의 센서 노드에서 추출한 요소는 검사 시간, 노드 ID, 온도, 습도, 광량, 전압이다.

시간은 날짜와 시, 분, 초 단위로 구성되어 있다. 노드 ID는 1에서 54의 값 범위를 갖는다. 온도는 섭씨를 단위로 하며 습도는 퍼센트(%)를 단위로 사용한다. 광량은 룩스(lux)를 그 단위로 하고 있다. 각 센서 노드에서 약 30초마다 데이터 수집을 시행하였지만 장애물의 간섭, 신호의 세기 등에 따라 이상치가 빈번하게 포함되었으며, 값이 수집되지 못한 노드도 상당히 존재한다.

본 논문에서는 날짜와 시, 분, 초 단위로 구성된 시간을 초 단위로 모두 계산한 뒤에 실험에 적용하였으며, 나머지 노드, 온도, 습도, 광량, 전압 데이터는 변형 없이 사용하였다. 센서 노드는 총 54개의 노드 중에 10개의 노드를 선택하여 실험에 사용하였다. 특히 서로 물리적 거리를 갖는 4개의 공간에서 2~3개 노드들을 임의적으로 선택함으로써, 동일 공간 내의 노드들은 유사한 센서 값 특성을 보이고, 상이한 공간의 노드들 사이에는 조금씩 다른 센서 값 특성을 갖도록 하였다.

실험에 사용하는 실험용 데이터는 크게 두 가지로 나뉜다. 하나는 원시 시계열 데이터로 가공되지 않은 센서 네트워크 데이터이고, 다른 하나는 주기 패턴 데이터로 주기에 따라 타임스탬프 값을 재구성한 데이터이다. 기계학습에 사용되는 학습 데이터는 원시 시계열 데이터에 양자화 및 평활화를 적용하고 주기에 따라 타임스탬프 값을 재구성한 주기 패턴 데이터를 사용한다. 이상치 검출을 위한 테스트 데이터는 원시 시계열 데이터를 사용한다. 실험용 데이터를 구성하는 인스턴스들은 임의의 측정시간에 한 센서에서 수집한 시간, 온도, 습도, 광량, 전압의 속성 값들로 구성된다.

4.2. 이상치 예측 실험

실험에 사용한 센서 노드의 총 개수는 10개이며, 실험은 각 센서 노드별로 수행한다. 로지스틱 회귀는 학습 데이터를 학습하여 선형 회귀 모델을 생성하고, C4.5는 결정트리 모델을 생성한다. 실험 결과에 대한 비교에는 이상치 분류의 정확도와 증가량을 이용하였

표 1. 27번 노드에 대하여 로지스틱 회귀를 이용한 이상치 예측 정확도
Table 1. The accuracy of logistic regression which predict outlier in node 27

Training size (주기 수)	원시 시계열 데이터	주기 패턴 데이터	증가량
1	47.04 %	56.09 %	9.05 %
3	51.44 %	62.24 %	10.80 %
5	51.59 %	62.49 %	10.90 %
All	85.83 %	86.05 %	0.20 %

표 2. 37번 노드에 대하여 C4.5를 이용한 이상치 예측 정확도
Table 2. The accuracy of C4.5 which predict outlier in node 37

Training Size	원시 시계열 데이터	주기 패턴 데이터	증가량
1	67.67 %	78.82 %	11.15 %
3	81.77 %	84.36 %	2.59 %
5	87.54 %	93.28 %	5.74 %
All	97.95 %	97.95 %	0.00 %

다. 증가량은 주기 패턴 데이터의 정확도와 원시 시계열 데이터의 정확도의 차이로써 주기 패턴 데이터의 성능 향상을 나타낸다.

표 1은 27번 노드에 대하여 로지스틱 회귀를 활용하여 이상치를 예측한 정확도이다. 원시 시계열 데이터에서 학습 데이터의 주기가 1일 때 이상치 분류 정확도는 47.04 %이며, 전체 주기일 때 85.83 %이다. 주기 패턴 데이터의 주기가 1일 때 이상치 분류 정확도는 56.09 %이며, 전체 주기일 때 86.03 %이다. 학습 데이터의 주기가 1일 때 원시 시계열 데이터의 정확도에 대한 주기 패턴 데이터의 정확도 증가량은 9.05 %이고, 전체 주기일 때의 증가량은 0.20 %이다. 본 실험으로 주기 패턴 데이터를 활용하는 것이 이상치 예측에 효과적임을 알 수 있다.

표 2는 37번 노드에 대하여 C4.5를 이용한 이상치 예측 정확도이다. 37번 노드의 실험 결과에서도 로지스틱 회귀에서와 마찬가지로 학습 데이터의 양이 증가함에 따라 각 데이터의 정확도가 증가한다. 그리고 본 실험에서도 주기 패턴 데이터가 원시 시계열 데이터보다 높은 정확도를 나타내는 것을 확인할 수 있다.

일반적으로 주기 패턴 데이터를 학습한 알고리즘이 높은 정확도를 보이지만 항상 그런 것만은 아니다. 표 3에서 노드 17번의 3주기 분량의 주기 패턴 데이터 학습의 경우 증가량이 −0.26 %와 −0.41 %로 감소하는 경우도 발생한다. 이것은 분할하여 학습한 주기 안의 데이터 분포가 다른 주기 안의 데이터 분포와 상반되

표 3. 17번 노드에 대하여 노지스틱 회귀와 C4.5를 이용한 이상치 예측 정확도
Table 3. The accuracy of logistic regression and C4.5 which predict outlier in node 17

Training size	원시 시계열 데이터		주기 패턴 데이터		증가량	
	Logistic regression	C4.5	Logistic regression	C4.5	Logistic regression	C4.5
1	39.59 %	74.55 %	53.69 %	77.38 %	14.10 %	2.83 %
3	63.67 %	81.38 %	63.41 %	80.97 %	−0.26 %	−0.41 %
5	61.75 %	81.56 %	62.51 %	85.08 %	0.76 %	3.52 %
All	81.06 %	97.42 %	79.13 %	97.43 %	−1.93 %	0.01 %

표 4. 10개 노드의 이상치 예측 평균 정확도
Table 4. The average accuracy of outlier prediction in which 10 nodes were used

Training size	원시 시계열 데이터		주기 패턴 데이터		증가량	
	Logistic	C4.5	Logistic	C4.5	Logistic	C4.5
1	43.11 %	69.53 %	57.12 %	71.91 %	14.01 %	2.38 %
3	60.60 %	73.27 %	66.01 %	77.89 %	5.41 %	4.62 %
5	67.07 %	84.43 %	66.90 %	85.16 %	-0.17 %	0.73 %
All	85.84 %	97.77 %	85.63 %	97.82 %	-0.21 %	0.05 %

게 나타나기 때문이다. 이 경우에 학습 모델의 정확도가 낮아지는 결과가 발생한다. 만약 주기 패턴에 따라 데이터를 주기별로 분할했을 경우, 분할한 데이터 분포를 분석하여 데이터를 학습하면 전체 데이터를 대표할 수 있는 우수한 모델을 얻을 수 있을 것이다.

표 4는 10개 노드의 이상치 예측 평균 정확도이다. 각 실험에서 로지스틱 회귀보다는 C4.5 결정트리의 실험 결과가 우수하게 나타났다. 결정트리는 속성이 연속하는 수치 값으로 표현될 때 값의 구간을 나누어 처리한다. 그러므로 회귀 분석보다 값의 범위에 민감하게 반응한다. 또한 실험에 사용된 센서 노드들의 속성 값이 규칙적인 형태로 발생하기 때문에 C4.5 결정트리가 높은 성능을 나타낸다.

5. 결 론

센서 네트워크는 그 활용성과 적용 범위가 다양하다. 그러나 센서 노드들 간의 데이터에 이상치나 결측치가 존재한다면 데이터로부터 추출한 결과를 신뢰할 수 없다. 그러므로 센서 노드들을 통해 수집하는 데이터를 활용하여 정보를 생성하기 위해서는 먼저 이상치나 결측치에 대한 데이터 정제를 수행하는 것이 필요하다.

본 논문에서 제안한 주기 패턴 분석 기법은 센서 네트워크의 원시 시계열 데이터에 대해 주기 패턴을 분석하고, 추출한 시간 주기로 원시 시계열 데이터를 재구성하여 주기 패턴 데이터를 생성한 후, 이상치 예측을 위해 주기 패턴 데이터를 이용하여 로지스틱 회귀와 C4.5에 적용한다. 주기 패턴 데이터가 이상치 예측에 높은 정확도를 나타낸다는 것을 몇 가지 실험을 통해 확일할 수 있었다. 주기 패턴 분석은 평활화 과정을 거쳐 반복하는 시간주기를 먼저 찾고, 분석된 시간 주기로 시간 속성값을 재구성함으로써 기계학습 알고리즘의 정확도를 향상시킬 수 있다. 주기 패턴 데이터는 센서 네트워크의 원시 시계열 데이터가 가지는 이상치와 결측치에 강인한 특징을 가진다. 반면 학습에서 단일 센서 노드 자신의 과거 시계열 데이터만을 이용하여 주기 패턴을 분석하기 때문에 센서 노드의 환경이 변경될 경우에 주기 패턴 분석을 적용하기 어렵다.

향후 단일 센서 노드의 원시 시계열 데이터만을 활용했을 때 나타나는 단점을 보완하기 위해서 센서들 간의 연관성을 활용한 연구가 필요하다.

참고 문헌

[1] 정보통신부, u-센서 네트워크 구축 기본계획, 2004.

[2] 정보통신연구진흥원, IT 차세대 성장동력 기획보고서(RFID/USN), 2004.

[3] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks", *IEEE Communications Magazine*, vol. 40, pp. 102-104, 2002.

[4] E. Callaway, V. Bahl, P. Gorday, J. A. Gutierrez, L. Hester, M. Naeve, and B. Heile, "Home networking with IEEE 802.15.4: A developing standard for low-rate wireless personal area networks", *IEEE Communications Magazine, Special Issue on Home Networking*, vol. 40, pp. 70-77, 2002.

[5] 정덕진, 송병철, 이승열, 조위덕, "상황인지 센서네트워크 기술동향", 한국정보과학회 정보통신연구회 정보통신기술지, 제18권, 제1호, pp. 2-30, 2004.

[6] 장병준, "RFID/USN 기술개발 동향 및 발전전망", 한국인터넷정보학회지, 제5권, 제3호, pp. 77-83, 2004.

[7] 조위덕, 이상학, 강정훈, "센서 네트워크 기술 개요", 한국정보과학회 정보통신연구회 정보통신기술지, 제17권, 제1호, pp. 0101-0118, 2003.

[8] G. Bontempi and Y. L. Borgne, "An adaptive modular approach to the mining of sensor network data", *Proceedings of the 1st International Workshop on Data Mining in Sensor Networks*, pp. 3-9, 2005.

[9] F. Zhao and L. Guibas, *Wireless Sensor Networks: An Information Processing Approach*, Morgan Kaufmann, 2005.

[10] Intel Lab Data, http://db.lcs.mit.edu/labdata/labdata.html

[11] I. Davidson and S. S. Ravi, "Distributed pre-processing of data on networks of berkeley motes using non-parametric EM", *Proceedings of the 1st International Workshop on Data Mining in Sensor Networks*, pp. 17-27, 2005.

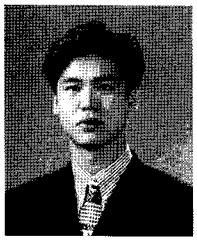
[12] C. Guestrin, P. Bodik, R. Thibaux, M. Paskin, and S. Madden, "Distributed regression: an efficient framework for modeling sensor network data", *Information Processing in Sensor Networks*, pp. 1-10, 2004.

[13] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2005.

[14] 이상학, 조위덕, 정태충, "무선 센서 네트워크의 네트워킹 기술", 한국정보과학회 정보통신연구회 정보통신기술지, 제18권, 제1호, pp. 31-47, 2004.

[15] 김선진, 박석지, 구정은, 김내수, "RFID/USN 산업 동향 및 발전전망", 전자통신동향분석, 제20권, 제3호 통권 93권, 2005년.

[16] G. Gupta and M. Younis, "Load-balanced clustering of wireless sensor networks", *Proceedings of IEEE International Conference on Communications*, vol. 3, pp. 1848-1852, 2003.
[17] E. Elnahrawy and B. Nath, "Context-aware sensors", *Proceedings of 1st European Workshop on Wireless Sensor Networks*, pp. 77-93, 2004.
[18] D. Janakiram, V. A. Reddy, and A. V. U. P. Kumar, "Outlier detection in wireless sensor networks using bayesian belief networks", *Proceedings of the 1st International Conference on Communication System Software and Middleware*, pp. 1-6, 2006.
[19] S. Tanachaiwiwat and A. Helmy, "Correlation analysis for alleviating effects of inserted data in wireless sensor networks", *Proceedings of the The 2nd Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services*, pp. 97-108, 2005.



김 형 일

- 1996년 목원대학교 수학과 졸업(이학사)
- 1996년~1998년 (주)경기은행
- 2001년 동국대학교 대학원 컴퓨터공학과 (공학석사)
- 2004년 동국대학교 대학원 컴퓨터공학과 (공학박사)
- 2005년~2006년 동국대학교 컴퓨터공학과 IT분야 교수요원(정보통신부)
- 주관심분야 : 센서 네트워크, 지능형 로봇, 기계학습, 지능형 에이전트, e-learning