스트림 데이터의 윈도우 기반 분류

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

A Window-Based Classification of Stream Data

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

요 약

센서와 모바일 기술의 발달로 인해 다양한 센서에서 수집된 스트림 데이터를 처리하는 연구들이 많이 수행되고 있다. 다차원 속성의 스트림 데이터는 센서에서 주기적으로 수집되어 버퍼링 후 처리되기 때문에 기존의 투플 기반의 데이터 분류 기법에 적합하지 않다. 따라서 이 논문에서는 윈도우 기반의 스트림 데이터 분류를 위해 각 속성의 평균과 표준편차 값을 이용하여 투플 기반으로 변환하는 기법을 제안한다. 제안된 기법의 타당성은 투플 기반 데이터 분류 기법(의사결정트리, 단순 베이지안 분류기, 베이지안 신뢰 네트워크)에 의한 정확도 측정에 기반 한다. 로봇에서 수집된 센서 데이터를 이용한 실험 결과, 높은 정확도로 제안된 기법이 타당함을 증명하였으며 베이지안 신뢰 네트워크 기법이 다른 기법에 비해 우수함을 발견하였다.

1. 서론

최근 센서와 모바일 기술의 발달로 다양한 센서에서 수집된 스트림 데이터를 처리하는 연구들이 수행되고 있다. 스트림 데이터는 다양한 센서로부터 주기적으로 수집되고 통신비용 절감을 위해 센서에서 버퍼링 후 처리되기 때문에 다차원 속성이 윈도우 단위로 분할된다. 각 윈도우 단위의 데이터는 여러 개의 투플이 하나의 클래스로 분류되어야 하기 때문에 기존의 투플 기반의 데이터 분류 기법을 이용하기에는 적합하지 않다. 따라서 윈도우 기반의 데이터를 투플 기반으로 변환한 후 기존의 분류 기법을 적용해야 한다.

이 논문에서는 윈도우 단위 스트림 데이터의 분류를 위해 평균과 표준편차를 대표값으로 사용하여 투플 기반으로 변환하는 기법을 제안한다. 제안된 기법의 타당성을 평가하기 위해 로봇 데이터를 여러 통계 값으로 요약하여 일반적인 분류 기법에 적용한 후 정확도를 비교한다. 또한 실험 결과를 통해 제안된 기법이 적용된 스트림 데이터를 분류하는데 가장 우수한 분류 기법을 비교 평가한다.

2. 관련 연구

윈도우 기반 스트림 데이터와 같은 연속형 데이터를 하나의 값으로 표현해 주기 위해서 위치의 대표값, 산포의 대표값과 같은 일반적인 요약 통계 값[1]을 이용할 수 있다. 위치에 따른 대표값으로는 표본평균, 제1사분위수, 제2사분위수, 제3사분위수, 절사평균, 중앙값 등이 있다. 이런 대표값들은 데이터 값의 위치에 기반 한다. 산포에 따른 대표값으로는 사분위수 범위, 표본 표준편차, 최빈수, 선형 회귀선의 기울기 등이 있다. 이런 대표값들은 데이터 값의 분포에 기반 하기 때문에 데이터가 가지고 있는 산포의 특성을 설명해줄 수 있다.

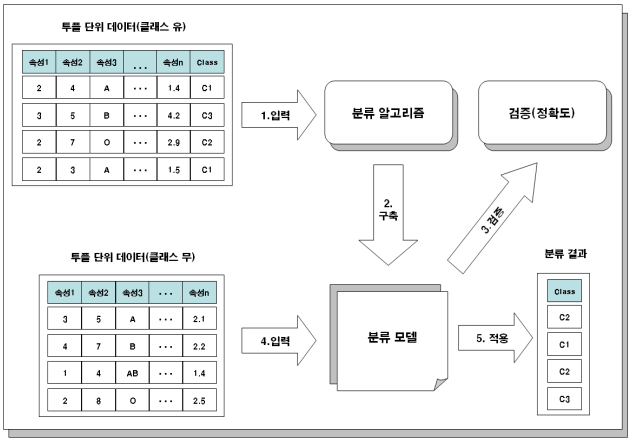
가상 트랜잭션(Virtual Transaction) 기법[2]은 스트림 데이터를 타임 윈도우 기법과 이벤트 윈도우 기법을 사용하여 이벤트 집합을 만드는 방법이다. 스트림 데이터에는 시간 속성과 이벤트 속성이 존재한다. 이러한 스트림 데이터를 시간에 따라 나열한

후에 정해진 윈도우 길이만큼 이벤트들을 집합으로 묶는다. 스트림 데이터로부터 가상 트랜잭션 집합을 생성해 내는데 있어서 두 가지의 접근 방법을 시도한다. 한 가지는 고정타임 윈도우를 이용하는 방법이고 다른 한 가지는 고정이벤트 윈도우를 이용하는 방법이다. 위의 두 가지 기법에 '겹침깊이(Overlapping Depth)'를 사용하여 가상 트랜잭션을 생성한다. 이 방법으로 만들어진 가상 트랜잭션 데이터 집합은 연관규칙 알고리즘의 입력데이터로 사용하여 이벤트 패턴을 분석하는데 사용한다.

위에서 설명한 여러 요약 통계 값을 사용하여 윈도우 기반 스트림 데이터를 요약할 수 있다. 그러나 변화량이 큰 윈도우를 하나의 요약 통계 값으로만 요약하기는 부족하다. 또한 가상 트랜잭션 기법은 실시간으로 들어오는 데이터를 기존의 마이닝 알고리즘에 적용시키기 위해 윈도우로 묶어서 투플 단위로 변환하기 때문에 윈도우 기반으로 입력되는 스트림 데이터를 처리하는 것은 불가능하다. 따라서 이 논문에서는 위치에 따른 대표값으로 표본평균과 산포에 따른 대표값으로 표본 표준편차를 함께 사용하는 방법을 제안한다.

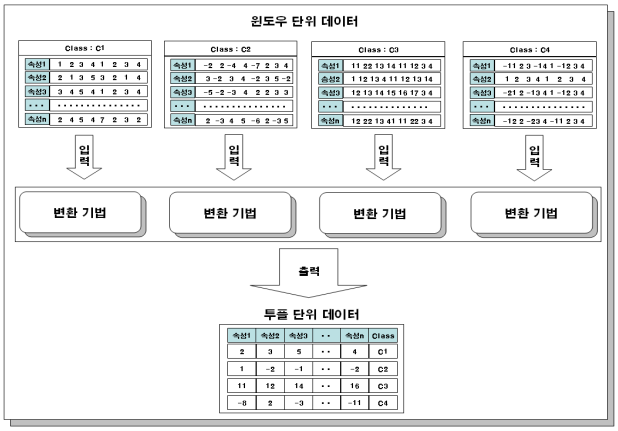
3. 윈도우 기반 데이터의 변환

기존의 대용량의 데이터베이스 기반 기법들은 투플 기반 분류 기법들로서 (그림 1)에서와 같이 2단계 절차로 구성된다. 첫 번째 단계에서는 사전에 정의된 데이터 클래스를 통해 모델을 구축한다. 두 번째 단계는 구축된 모델의 예측 정확도를 추정하여 모델이 타당한지 검증하고 만약 모델이 타당하다면 새로운 데이터의 분류에 사용한다.



(그림 1) 투플 기반 데이터의 분류

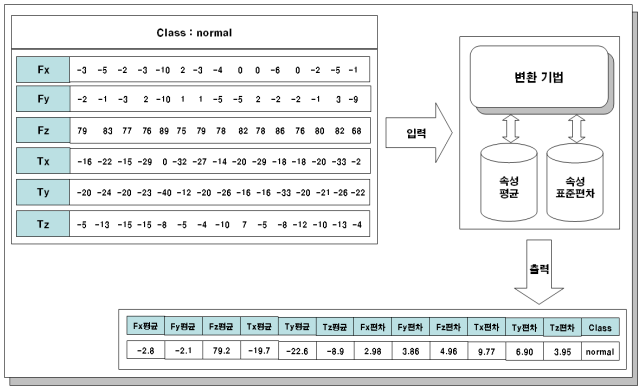
그러나 연속적인 형태를 갖는 스트림 데이터는 주기적인 버퍼링 후에 처리가 되기 때문에 여러 개의 투플들이 모여서 하나의 윈도우를 이룬다. 따라서 하나의 윈도우에 하나의 클래스 레이블이 정의된 스트림 데이터를 기존의 분류 기법에 적용시키기 위해서는 투플 단위로의 변환이 필요하다. (그림 2)는 윈도우 단위의 데이터를 이런 기법에 적용하기 위한 변환 절차를 나타낸다.



(그림 2) 윈도우 기반 데이터 변환 절차

연속 값을 갖는 윈도우내의 각 속성들은 평균, 중앙값, 모드 등을 이용하여 대표값으로 표현될 수 있다. 중앙값은 데이터를 크기 순서로 정렬했을 때 중앙에 해당하는 값이고 모드는 빈도수가 가장 많은 값을 말한다. 스트림 데이터는 부분적으로 값들의 변화량이 커질 수 있고 그 변화가 클래스에 중요한 영향을 줄 수 있기 때문에 중앙값이나 모드를 통한 대표값은 적합하지 않다. 평균을 통해 윈도우를 하나의 투플로 변환을 하는 방법은 같은 평균 값 이라도 윈도우 내부의 편차에 따라 의미가 달라질 수 있기 때문에 평균만으로 대표값을 표현하기에는 부족하다. 따라서 이 논문에서는 평균값과 함께 표준편차도 고려하여 윈도우의 내부 분포를 반영하는 윈도우 단위의 스트림 데이터 변환 기법을 제안한다.

(그림 3)은 [4]의 normal 이라는 클래스 레이블을 갖고 있는 윈도우를 각 속성의 평균과 표준편차를 이용하여 하나의 투플로 변환한 예이다. 하나의 윈도우에는 {Fx, Fy, Fz, Tx, Ty, Tz}의 6개 속성과 하나의 클래스가 있고 각 속성에는 15개의 스트림 값이 존재한다. 변환 결과 왼쪽의 윈도우가 총 12개의 속성과 1개의 클래스로 구성된 투플로 변환이 되었다.



(그림 3) 변환 기법

4. 분류 기법

제안한 기법의 타당성 및 정확도 측정은 아래의 일반적인 분류 기법의 비교를 통해 측정이 가능하다.

의사결정트리는 흐름도와 유사한 트리구조로 중간 노드에는 속성에 대한 검사를 표시하고 가지는 검사의 결과를 나타내며, 잎 노드는 클래스나 클래스의 분포를 나타낸다. 클래스 레이블을 알 수 없는 투플을 분류하기 위해서는 투플의 속성 값들을 의사결정트리로 검사한다. 검사는 루트에서부터 해당 투플에 대한 클래스의 예측을 갖는 잎 노드까지의 경로를 따라 진행한다. 의사결정트리는 사용자가 해석이 용이하고 두 속성 간 상호작용의 해석이 가능한 장점을 가지고 있다.

베이지안 분류기는 주어진 샘플이 특정 클래스에 속할 확률을 예측한다. 베이지안 분류는 베이즈 이론에 기반하며 대규모 데이터베이스에 적용되어도 높은 정확성과 속도를 보여준다.

단순 베이지안 분류기는 주어진 클래스의 한 속성 값이 다른 속성 값과의 상호 독립을 가정한다. 만약 이 가정을 만족한다면 단순 베이지안 분류기는 다른 분류기와 비교할 때 최소 오류율을 갖는다.

베이지안 신뢰 네트워크는 단순 베이지안 분류기와 다르게 속성의 부분집합들이 가지는 종속 관계를 표현할 수 있는 그래픽 모델이다.

5. 실험 및 평가

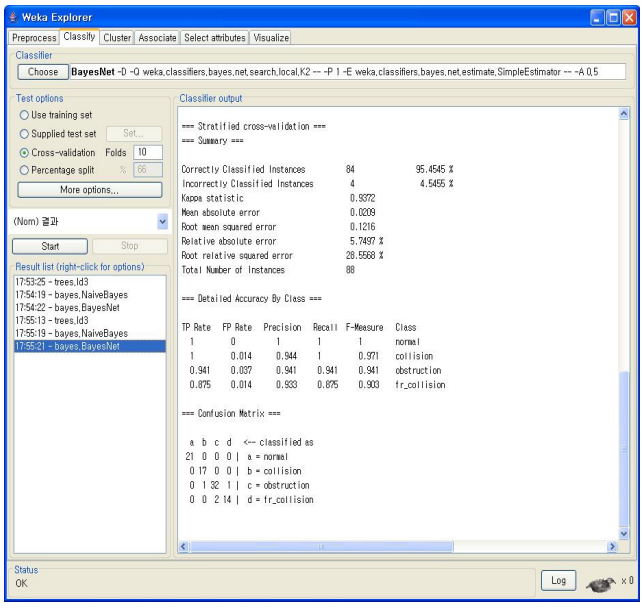
5.1 실험 과정

실험은 윈도우XP 환경에서 데이터 변환을 위해 MS-Excel의 함수와 내장 Visual Basic(매크로)을 사용하였고, 데이터의 이산화 및 정확도 측정을 위해 여러 분류 기법을 지원해 주는 Weka[3] 프로그램(그림 4)을 통해 실시하였다. 실험에 사용된 데이터는 로봇이 작동 중 외부의 간섭에 의해 발생한 에러 데이터로 로봇의 Force/ Torque에 의해 측정된 LP1, LP3~LP5의 4개의 데이터이다[4]. 다속성의 스트림 데이터는 연속형의 값을 갖고 총 6개의 속성으로 이루어져 있으며 각 속성에는 15개의 값이 포함되어 있다. 과정은 <표 1>과 같다.

<표 1> 실험 절차

단계	내용
1단계	• 여러 변환 기법에 의해 로봇 데이터를 투플 단위로 변환 - 평균 / 표준편차 / 중앙값 / 회귀선 기울기 / 절사평균 / 사분위수 / 사분위 범위 - 제안된 기법(평균과 표준편차를 동시 고려)
2단계	• 분류 기법에 적용하기 위해서는 연속형 데이터를 이산형으로 변환 - Fayyad와 Irani가 제안한 엔트로피 기반 이산화 방법인 MDL 방법[5]을 사용하는 Weka 프로그램을 통해 변환
3단계	• 3가지 분류 기법에 의해 정확도 측정 - 의사결정트리 - 단순 베이지안 분류기 - 베이지안 신뢰 네트워크
4단계	• 4개의 실험 데이터에 1~3 단계 적용

실험에서 정확도의 측정은 k-fold cross validation 방법을 사용하였다. 이 방법은 일반적으로 k가 10일 때 상대적으로 적은 편향과 분산을 갖기 때문에 분류기 정확도를 추정하는데 많이 사용된다.



(그림 4) Weka 프로그램 이용한 실험

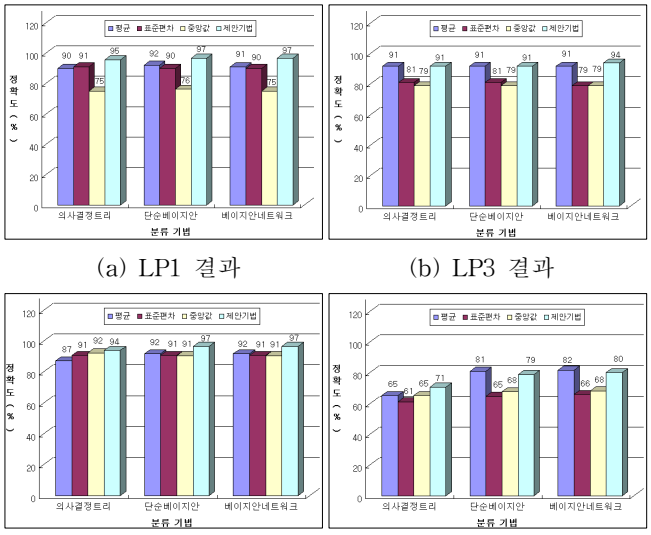
5.2 실험 결과

4개의 로봇 데이터에 대한 정확도 실험 결과는 <표 2>과 같다. 절사평균, 사분위수, 사분위 범위, 회귀선 기울기에 의한 정확도는 다른 방법에 비해 매우 낮아서 표에서는 나타내지 않았다.

<표 2> 각 실험 데이터 셋에 대한 정확도

구분		정확도(%)			
		평균	표준편차	중앙값	제안기법
LP1	트리	89.77	90.91	75.00	95.45
	단순	92.05	89.77	76.14	96.59
	네트워크	90.91	89.77	75.00	96.59
LP3	트리	91.49	80.85	78.72	91.49
	단순	91.49	80.85	78.72	91.49
	네트워크	91.49	78.72	78.72	93.62
LP4	트리	87.18	90.73	92.31	94.02
	단순	91.82	90.73	90.60	96.58
	네트워크	91.82	90.73	90.60	96.58
LP5	트리	65.24	60.98	65.24	70.73
	단순	81.10	64.63	67.68	78.83
	네트워크	81.71	65.85	68.29	80.05

(그림 5)는 <표 2>에서 보여주는 정확도를 막대 그래프로 표현한 것이다.



(a) LP1 결과 (b) LP3 결과

(c) LP4 결과 (d) LP5 결과

(그림 5) 각 실험에 따른 정확도 그래프

실험 결과 (그림 5)와 같이 LP1, LP3~LP5에서 제안된 기법의 정확도가 평균, 표준편차, 중앙값 등을 이용한 방법에 비해 정확하고 신뢰할만한 수준인 것으로 나타났다. 또한 일반적인 분류 기법(의사결정트리, 단순 베이지안 분류기, 베이지안 신뢰 네트워크)을 이용한 측정에서도 제안된 변환 기법이 우수한 성능을 보임을 알 수 있다. 그 중 베이지안 신뢰 네트워크 기법은 다른 분류 기법보다 일반적으로 높은 정확도를 보였다. 그 이유는 데이터 변환에서 1개의 속성을 2개(평균, 표준편차)의 속성으로 도출해냈기 때문에 각 속성 간에는 종속성이 존재할 가능성이 있고, 베이지안 신뢰 네트워크는 종속성을 가지는 데이터에 적합한 모델이기 때문이다.

6. 결론 및 향후 연구

윈도우 기반 스트림 데이터를 기존의 분류 기법에 적용하기 위해 평균과 표준편차를 이용하여 투플 기반으로 변환하였다. 로봇 데이터를 통해 정확도를 측정한 결과 제안된 기법이 다른 요약 통계 값에 비해 우수하고 타당함이 증명되었다. 또한 제안된 기법을 적용한 스트림 데이터의 분류 기법으로는 베이지안 신뢰 네트워크가 최적임이 발견 되었다.

이 논문에서는 윈도우 기반 스트림 데이터를 변환하기 위해 평균과 표준편차를 사용하였지만 이 방법은 윈도우 내부의 정확한 변화 파악에는 어려움이 있다. 향후에는 윈도우 내부 변화를 잘 반영하는 방법을 통해 정확도를 향상시키는 연구가 필요하다.

참고문헌

[1] B.W. Lindgren, “Statistical Theory”, CRC Press, January 1993.

[2] 김민수, 김철환, 김응모, “가상 트랜잭션을 이용한 시계열 데이터의 데이터 마이닝”, 한국정보처리학회논문지 제9-D권 제2호, 2002.

[3] Waikato Environment for Knowledge Analysis, Version 3.4.5, Available at http://www.cs.waikato.ac.nz/ml/weka/, University of Waikato, New Zealand.

[4] Robot Execution Failures, The UCI KDD, Available at http://kdd.ics.uci.edu, University of California, Irvine, Department of Information and Computer Science, 1999.

[5] U.M. Fayyad, K.B. Irani, “Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning”, Proceedings 13th International Joint Conference on Artificial Intelligence, pp. 1022-1027, 1993.