데이터 스트림 시스템에서 인과관계 탐사를 위한 마이닝 방법

한대영, 김대인, 황부현
전남대학교 전자컴퓨터공학부
e-mail:abyo0111@naver.com

A Mining Method for Exploration of Causality on Data Stream System

Dae-Young Han, Dae-In Kim, Bu-Hyun Hwang
Dept of Electronics and Computer Engineering,
Chonnam National University

요 약

일반적으로 이벤트는 발생 시점이라는 시간 속성을 갖는다. 그리고 고객 단위로 이벤트를 축적한 데이터베이스가 있다면 데이터 마이닝을 통하여 유용한 정보를 탐사할 수 있다. 특히 이벤트 발생의 원인과 결과에 대한 관계 규칙을 찾아낼 수 있다면 과거의 정보를 바탕으로 미래를 예측할 수 있는 예측 판단 정보로 사용할 수 있다. 본 연구에서는 데이터 스트림 시스템에서 시간 관계 규칙을 탐사하고 시간 관계 규칙을 구성하는 이벤트 간의 영향력을 측정하기 위한 SM-EC(data Stream Mining for Exploration of Causality)기법을 제안한다. 실험을 통하여 SM-EC가 제공하는 영향력 정보는 다양한 비상 상황에 대처하는 중요한 척도가 될 수 있음을 확인하였다.

1. 서론

연관 규칙 탐사란 데이터베이스에 잠재되어 있는 지식을 발견하기 위한 마이닝 기법의 하나로 최근 데이터 스트림 시스템(DSMS : Data Stream Management System)에서 연관 규칙 탐사에 대한 연구가 활발하게 진행되고 있다[1][2].

데이터 스트림 시스템의 응용의 하나인 ICU와 같은 의료 분야의 경우, 각각의 센서는 환자의 체온, 혈압, 맥박, 심장 박동 수 등 여러 종류의 데이터 스트림을 수집하며 각각의 데이터 스트림은 "체온과 맥박은 비례 관계이다.", "출혈이 발생하는 경우 혈압은 내려가고 맥박은 빨라진다."와 같이 이벤트 또는 환경의 변화에 연관성을 갖고 있다. 그러므로 이러한 데이터 스트림 간의 연관성을 분석함으로써 환자의 추후 발생 가능한 상태를 예측할 수 있다[3].

데이터 스트림 시스템은 센서를 통하여 수집되는 단일 이벤트 관리 및 분석도 중요하지만 하나의 객체에 대한 여러 정보를 수집하는 다차원 센서에서 수집하는 이벤트 간의 연관 관계를 분석함으로써 향후 발생 가능한 이벤트를 예측하는 것도 매우 중요하며 이에 대한 연구가 진행되고 있다[4]. 또한 이러한 응용에서 환자의 정상적인 체온 및 맥박에 대한 정보보다도 환자의 비정상적인 상태를 나타내는 체온 및 맥박의 상승과 같은 이벤트가 그 발생 빈도가 낮아도 중요성은 매우 높다[3]. 그러나 기존의 연구에서 적용하는 지지도 기반의 연관 규칙 탐사 기법들은 발생 빈도가 낮은 이벤트들은 연관 규칙 탐사 과정에서 제외하는 문제를 가지고 있다. 그러므로 본 연구에서는 데이터 스트림 시스템에서 객체의 비정상적인 상태를 나타내는 이상 이벤트(abnormal event) 간의 연관 규칙을 탐사하고 예측하기 위한 SM-EC(data Stream Mining for Exploration of Causality)기법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 데이터 스트림을 사용한 마이닝 기법에 관한 관련 연구를 기술하고 3장에서는 이상 이벤트에 대한 연관 규칙 탐사 알고리즘을 제안한다. 4장에서는 제안한 알고리즘의 성능을 분석하고, 끝으로 5장에서는 본 논문의 결론 및 향후 연구 방향을 기술한다.

2. 관련연구

데이터 스트림의 연관 규칙 탐사 방법은 기존의 데이터베이스 시스템에 비하여 여러 가지 제약사항을 가지고 있다[1][5]. 첫째, 마이닝 기법은 기본적으로 많은 데이터 스캔을 필요로 한다. 그러나 센서로부터 수집되는 데이터 스트림은 연속적이고 크기가 무한하므로 손실 없이 데이터베이스나 메인 메모리에 저장될 수 없으며 데이터가 수집되는 즉시 연관 규칙 탐사에 필요한 요약 정보를 추출하여야 한다. 둘째, 유용한 연관 규칙을 탐사하기 위하여 추출하는 후보 항목 집합(candidate itemsets)의 수가 무

한하다. 데이터 스트림의 크기가 무한하고 연속적이므로 데이터의 조합으로 이루어지는 후보 항목 집합의 수는 무한히 증가한다[6][7]. 그리고 이러한 데이터 스트림의 특성으로 인하여 연관 규칙 탐사 방법은 많은 비용과 노력을 필요로 하며 데이터 스트림의 마이닝 방법으로 일정 시간 단위인 윈도우 동안에 수집된 데이터 스트림의 연관 정보를 분석하는 방법이 제안되었다..

[5]에서는 윈도우 단위로 데이터 스트림간의 최대 빈발 항목 집합(maximal frequent itemsets)을 탐사하는 DSM-MFI(Data Stream Mining for Maximal Frequent Itemsets) 방법을 제안하였다. DSM-MFI 방법은 트리 기반의 SFI(Summary Frequent Itemset)를 구축하여 사용자가 정의한 지지도 X.CL 이상의 발생 빈도를 갖는 데이터만을 추출하여 연관 규칙을 탐사한다. 또한 DSM-MFI 방법은 한 번의 스캔으로 요약 정보를 구축하고 발생 가능한 오류를 고려하여 0과 1 사이의 오류 임계값 ζ를 고려한 X.CL×ζ 이상의 발생 빈도를 갖는 데이터를 포함하여 최대 빈발 항목 집합을 추출한다. 그러나 DSM-MFI 방법은 센서에서 수집되는 단일 데이터 스트림에 대한 연관 규칙만을 탐사하며 다차원 데이터 스트림 간의 연관 규칙 탐사는 고려하지 않는다.

[7]에서는 데이터 스트림의 연관 규칙을 탐사하는 MILE(MIning from muLtiple strEams) 방법을 제안하였다. MILE 방법은 윈도우 단위로 최소 지지도(minimum support) 이상의 발생 빈도를 갖는 토큰에 대한 트리 인덱스를 구축하여 다차원 데이터간의 연관 규칙을 탐사한다. 또한 MILE 방법은 [6]에서 제안한 PrefixSpan 방법의 단점인 반복적인 계산 과정을 줄이기 위하여 이벤트의 선행 관계로 구성된 해시 테이블을 유지함으로써 빠른 연관 규칙 탐사가 가능하다. 그러나 MILE 방법은 최소 지지도 이상의 빈도수를 갖는 데이터의 연관 규칙만 탐사하며 중요도가 높지만 발생 빈도가 낮은 이벤트에 대한 연관 규칙 탐사는 고려하지 않는다.

[1]에서는 윈도우 기반의 이벤트 발생 주기 탐사 대한 방법을 제안하였다. 제안 방법은 윈도우 단위로 이벤트의 발생 지지도를 계산하여 발생 주기를 탐사하고 최대 지지도를 만족하는 이벤트 시퀀스를 추출함으로써 추후에 발생 가능한 이벤트를 예측할 수 있다. 그러나 제안 방법은 이벤트 발생의 시간 간격만을 고려하며 이벤트의 발생 횟수와 센서로부터 동시에 수집되는 다차원 이벤트 간의 연관 관계는 고려하지 않는다.

3. 인터벌 관계, 인터벌 관계 그래프

본 연구에서는 신뢰도에 기반 한 이벤트들 사이의 인과 관계에 대한 영향력 정도를 평가하며 이러한 평가는 더욱 구체적인 정보를 탐사할 수 있으므로 매우 흥미로운 일이라고 말할 수 있다. 그리고 이러한 연구 결과로 탐사된 정보를 의료 분야에 적용하면 보다 효율적인 질병 예측 및 처방을 할 수 있을 것이다.

연관 규칙 탐사는 관계형 데이터베이스에서 함수적 종속성을 추출하는 기법과는 달리 통계적 방법에 의해 연관성이 있는 항목들 사이의 규칙성을 추출하는 과정이다. 연관 규칙 탐사에는 사용자에 의해 주어지는 임계치인 최소 지지도(minimum support: MinSup) 와 최소 신뢰도(minimum confidence: MinConf)가 적용된다. 특히 신뢰도는 이벤트들 사이의 발병 원인 및 관계에 대한 정도를 나타내는 척도로 활용가능하다. 예를 들어 시간 속성을 갖는 이벤트들에 대한 빈발 이벤트 타입을 선별하고 인터벌 이벤트를 요약한 후 빈발 인터벌 이벤트의 시간 인터벌을 사용하여 각각의 환자에 대한 인터벌 이벤트 시간 관계를 계산한다. 그리고 발견된 시간 관계는 before, overlap, during과 같은 의미 있는 인과 관계로 표현되고, 인과 관계가 존재하는 인터벌 이벤트들 사이의 영향을 미치고 영향을 받은 원인 관계에 대한 정도를 평가할 수 있다. SM-EC에서는 인과 관계에 대한 영향력의 정도를 평가할 때 신뢰도에 기반 한다. 다음 표 1은 인터벌 이벤트 시간 관계에 대한 이진 인터벌 관계 연산자들이다.

<표 1> 이진 인터벌 관계 연산자

Relation	Expression
before(x,y)	x.ve < y.vs
equals(x,y)	x.vs=y.vs ∧ x.ve=y.ve
meets(x,y)	x.ve=y.vs
overlaps(x,y)	x.vs<y.vs ∧ x.ve<y.ve
during(x,y)	x.vs<y.vs ∧ y.ve<x.ve

3.1 SM-EC 알고리즘

Input: 환자 진찰 및 센싱 기록 스트림 데이터베이스
Output: 빈발 시간 관계 규칙에서 이벤트간 영향력

Step 1. Preprocessing
Step 2. 빈발 이벤트 타입 계산
Step 3. 빈발 이벤트 타입별 시퀀스 계산
Step 4. 이벤트 시퀀스 분할 및 인터벌 이벤트로 요약
Step 5. 인터벌 이벤트간 시간 관계 규칙 계산
Step 6. 빈발 시간 관계 규칙 계산

제안하는 SM-EC방법에서 센서는 객체에 대한 여러 가지 이벤트를 수집하는 다차원 센서이며, 데이터 스트림은 윈도우 단위로 수집되어 처리된다. 데이터 스트림 시스템에서 수집되는 데이터 스트림은 미리 정의된 값의 범위에 따라 기호화하여 저장하며, 각각의 기호는 센서가 측정하는 객체에 대한 이벤트(event)를 의미한다. 그리고 각각의 이벤트는 센서를 통하여 단위 시간 구간인 윈도우(window) 동안에 수집된 데이터 스트림이며 단일 윈도우

동안에 센서가 수집한 데이터 스트림은 하나의 트랜잭션에 포함된 데이터 항목으로 간주 할 수 있다[6,7,10,12]. 따라서 주기적으로 진료를 받는 환자에 대한 진찰 결과는 시간 속성을 갖는 하나의 이벤트로 간주되어 MDB(Main Data Base)에 축적된다. 또한 진료에 대한 다양한 문서들도 아카이빙 작업을 통하여 MDB에 축적된다. 그러나 MDB의 데이터들은 원시 데이터 형태이므로 오류가 존재할 가능성이 있으며, 또한 마이닝을 하기 위해서는 수집된 데이터는 정제되고 표준화되어야 한다. 따라서 MDB의 데이터는 PP(Preprocessing) 모듈을 통하여 진료 기록의 전산화 작업 중의 오류, 이상치 및 잡음(noisy) 데이터들을 수정 및 보정한다. 그리고 환자에 대한 진찰 기록은 PP 모듈을 통하여 환자 식별 번호, 증상, 증상 발생 시점으로 표준화되며 PP 모듈의 수행 결과물은 TDB(Transaction Data Base)에 저장된다.

3.2 영향력 측정 척도

▪ Eff(A→B): 이벤트 발생에 영향을 미친 정도

$$Eff(A \rightarrow B) = Supp(A \rightarrow B) / Supp(A) \quad (1)$$

식 1은 이벤트 발생에 영향을 미친 정도 Eff(A→B)는 이벤트 A 발생 후 이벤트 B가 연속적으로 실행되는 것을 의미한다.

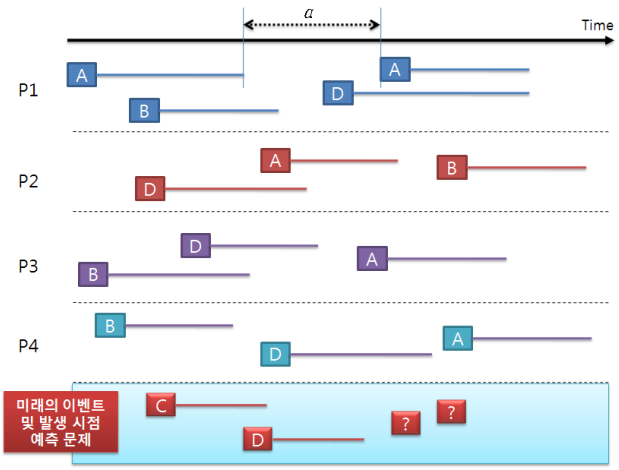
▪ BeEff(A→B): 이벤트 발생에 영향을 받은 정도

$$BeEff(A \rightarrow B) = Supp(A \rightarrow B) / Supp(B) \quad (2)$$

식 2는 이벤트 발생에 영향을 받은 정도 BeEff(A→B)는 이벤트 B 발생에 대한 이벤트 A의 원인 정도를 보여준다.

3.3 시퀀스 분할 방법 및시간 속성을 갖는 이벤트들 사이의 인과 관계

그림 1의 예에서 4명의 환자(P1~P4)의 질병에 대한 증상으로 이벤트 A, B, D가 연속적으로 발생한다. 그리고 각각의 환자는 다양한 시간 인터벌 간격으로 이벤트가 발생하며 환자 P1의 경우 같은 이벤트 타입에 속하는 이벤트 A가 α 만큼의 인터벌을 두고 발생하였다. 그러나 환자 P1의 같은 이벤트 타입 A에 속하는 두 개의 이벤트를 지속적인 하나의 이벤트로 간주할 지 아니면 두 개의 독립적인(연속성이 없는) 이벤트로 간주하여야 합리적인지에 대한 정의 기준이 필요하다. 따라서 본 논문에서는 시간 인터벌 지지도를 정의하여 같은 이벤트 타입에 속하는 두 개의 이벤트 발생 간격이 시간 인터벌 지지도보다 큰 경우에는 두 개의 이벤트를 독립 이벤트로 간주하며, 그렇지 않은 경우에는 지속성을 갖는 연속 이벤트로 정의하여 이벤트들 사이의 인과 관계 정보를 탐사한다.



(그림 1) 시간 속성을 갖는 이벤트들 사이의 인과 관계

예를 들어 다음과 같이 환자에 대한 두 개의 이벤트 시퀀스가 있다고 가정하자.

이벤트 시퀀스 1= <(A,1)(A,3)(A,11)(A,12)>,
이벤트 시퀀스 2= <(B,10)(B,11)>

그리고 이벤트 시퀀스에 포함된 각각의 이벤트 타입에 대한 인터벌 이벤트는 다음과 같이 요약된다.

인터벌 이벤트 1=(A,[1,12]),
인터벌 이벤트 2=(B,[10,11])

위 예와 같은 두 개의 인터벌 이벤트 집합이 빈발하다고 하면 인터벌 이벤트 1 동안에 인터벌 이벤트 2가 발생한다는 시간 관계를 탐사할 수 있다. 즉 이벤트 A가 이벤트 B를 발생시키는 원인이며 이벤트 B는 이벤트 A로 인하여 발생되는 결과라고 판단할 수 있다. 그러나 위 예에서 시점의 단위가 달이라고 한다면(1이라는 시점이 1월을 의미한다면) 이러한 판단은 합리적이지 못하다. 이벤트 A에 대한 이벤트 시퀀스에서 (A,3)과 (A,11)은 각각 3월과 11월에 이벤트 A가 발생하였음을 의미하며 2개의 이벤트 집합 사이에는 많은 시간 간격이 존재하기 때문이다.

같은 이벤트 타입 A에 속하는 이벤트 집합의 발생 시간에 대한 시간 간격이 크다는 것은 이벤트 타입에 존재하는 두 개의 이벤트 시퀀스가 지속적이지 못하고 독립적(independent)으로 해석되어야 함을 의미한다. 그러므로 위의 예에서 이벤트 시퀀스 1과 이벤트 시퀀스 2에 포함된 이벤트 집합 (B,10)과 (A,11)로 인하여 오히려 이벤트 B가 이벤트 A의 원인이라고 판단하는 것이 더욱 합리적이다. 즉 이벤트 A는 1월에서 3월까지 지속되었고, 11월과

12월에 다시 발생하였다. 그리고 4월부터 11월 동안에는 이벤트 A가 발생하지 않았다. 그러므로 이벤트 A는 1월부터 12월까지 지속적으로 발생 하였다기 보다는 1월에서 3월까지, 그리고 11월에서 12월까지 서로 독립적이라고 발생되었다고 판단하여야 한다. 그러므로 이벤트 시퀀스 1은 두 개의 <(A,1)(A,3)> 시퀀스와 <(A,11)(A,12)> 시퀀스로 나누어 인과 관계를 분석 하는 것이 바람직하다.

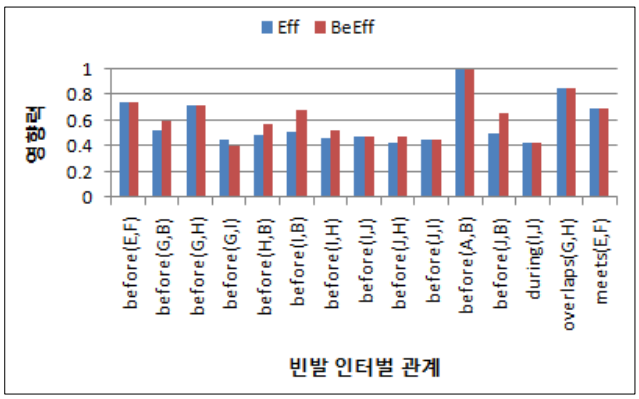
5. 시뮬레이션

시뮬레이션을 위해 가상의 환자 5,000명에 대한 총 40,000건의 트랜잭션 데이터를 생성하였다. 시뮬레이션에 사용한 파라메터는 표 2와 같다.

<표 2> 시뮬레이션 파라메터

파라메터	설명	값
MinSup	최소 지지도	50%(2,500명 이상)
Event	증상	A~J(랜덤생성)

그림 2는 시뮬레이션 결과를 그래프형태로 표현한 것이다.



(그림 2) 시뮬레이션 데이터에 대한 영향력 계산 결과

Eff 그래프에서 관계를 구성하는 두 이벤트를 A, B라 하면 이벤트 A와 이벤트 B에 대한 관계를 포함하는 트랜잭션의 지지도와 이벤트 A를 포함하는 트랜잭션 지지도가 모두 0.5라면 신뢰도에 기반하여 이벤트 A가 이벤트 B에 영향에 주는 정도 Eff(A→B)는 1이 된다. before(A→B)에서 Eff값은 1을 나타내고 있다. 이것은 이벤트 A가 발생하면 이벤트 B가 반드시 발생한다는 것을 의미한다. 그러나 이러한 정보는 이벤트 B의 발생 원인이 100% 이벤트 A라는 것을 의미하지는 않는다. 만약 이벤트 B를 포함하는 트랜잭션 지지도가 0.9라면 이벤트 B의 발생 원인에 대하여 0.4만큼은 이벤트 A가 아닌 다른 이벤트로 인하여 이벤트 B가 발생되었음을 알 수 있다.

한편 BeEff에서는 이벤트 A와 이벤트 B에 대한 관계를 포함하는 트랜잭션의 지지도와 이벤트 B를 포함하는 트랜잭션 지지도가 모두 0.5라면 신뢰도에 기반하여 이벤트 B가 이벤트 A에 대하여 영향을 받은 정도 BeEff(A→B)는 1이 된다. 즉 이벤트 B의 발생 원인이 100% 이벤트 A라는 것을 의미한다. 그러나 이벤트 A가 발생한다고 해서 항상 이벤트 B가 발생한다는 것을 의미하지는 않는다.

5. 결론 및 향후 연구

본 논문에서는 신뢰도에 기반한 이벤트들 사이의 인과관계에 대한 영향력 정도를 평가하며 이러한 평가는 더욱 구체적인 정보를 탐사할 수 있으므로 매우 흥미로운 일이라고 말할 수 있다. 그리고 이러한 연구 결과로 탐사된 정보를 의료 분야에 적용하면 보다 효율적인 질병 예측 및 처방을 할 수 있을 것이다. 향후로는 실제 환자 진찰 기록 데이터를 사용하여 다양한 인과 관계에 대한 척도를 제시할 수 있는 연구를 진행하고자 한다.

참고문헌

[1] H. Li, Z. Lu, and H. Chen, "Mining Approximate Closed Frequent Itemsets over Stream," SNPD '08. Ninth ACIS International Conference, pp.405-410, 2008.

[2] H. Thakkar, B. Mozafari, and C. Zaniolo, "A Data Stream Mining System," ICDMW '08. IEEE International Conference, pp.987-990, 2008.

[3] K. Kuramitsu, "Finding Periodic Outliers over a Monogenetic Event System," In Proc. of UDM05, pp.97-104, April 2005.

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

[5] 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, April 2005.

[6] 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, Nov. 2004.

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