시 계열 데이터에서의 연관성 발견을 위한 기법

이준호° 차재혁
한양대학교 정보통신대학원
haebangin@ihanyang.ac.kr°, chajh@hanyang.ac.kr

The Method of Rule Discovery for Time Series Data

JoonHo Lee° , Jaehyuk Cha
The Graduate School of Information & Communication, Hanyang University

요 약

본 논문은 시 계열 데이터에서의 연관성 발견에 있어서 복잡성과 연산량을 효과적으로 줄이며 연관성을 찾아내는 기법에 대해 기술한다. 기존의 시 계열 데이터에서의 sequence 분할 방법은 복잡한 clustering 기법을 사용하여 많은 시간과 resource를 필요로 하는 제한이 있다. 이에 본 논문에서는 효과적인 sequence 분할을 위한 증감 table을 이용한 방법을 제안하였다.

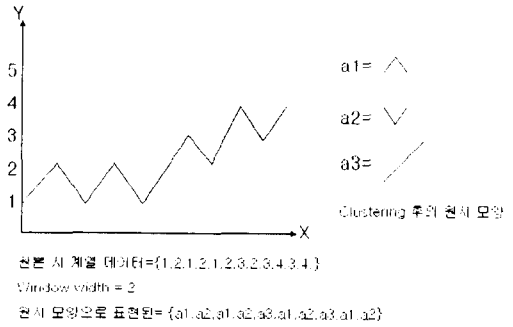
1. 서 론

현대 사회의 여러 분야에서 나타나는 시 계열 데이터에 관한 연구는 꾸준히 진행되어왔고 본 논문 또한 시 계열 데이터 내에서 발생하는 패턴을 찾고 패턴들 간의 연관성을 찾는 데 그 목적이 있다. 본 논문에 사용된 데이터는 IBM사의 1년간의 주식 가격 변동 데이터로 대표적인 시 계열 데이터로서 실험의 목적에 잘 부합한다 하겠다. 시 계열 데이터에서의 패턴을 찾는 일반적인 방법의 순서로 첫 번째 단계는 시 계열 데이터의 전체 sequence를 부분으로 나누어 subsequence[1,2,5,6,7]를 형성하고 두 번째 단계는 subsequence들을 서로 유사한 패턴으로 분류 하여 그룹을 형성하는 것이다. 이때 일반적으로 subsequence들은 Rw차원의 점으로 mapping하고 mapping된 점들을 기존의 clustering[2,7] 기법을 이용해 분류하는 방법이 사용된다. 세 번째 단계는 분류된 그룹간의 연관성을 Association Rule[3,4,7]을 적용해 발견하는 단계이다. 이러한 방법은 많은 연산량을 필요로 하기 때문에 사간과 resource의 제약이 따른다. 이에 본 논문에서는 기존의 방법과 달리하여 패턴이 변하는 domain을 줄여 증감 table을 만들고 그것을 이용하여 기존의 방법으로 얻어지는 결과를 적은 연산량으로 찾아내는 방법을 제시하였다.

2. 기존의 시 계열 데이터 처리기법

2.1 Discretization & Clustering

시 계열 데이터는 sliding window[1] 개념을 써서 subsequence를 형성함으로써 discrete한 형태로 표현될 수 있다. 이렇게 구성된 subsequence들은 적절한 패턴 유사도(similarity)[1,6,7]의 측정을 이용해 clustering 되어진다. 다음의 그림 1에서처럼 연속된 데이터를 일정한 간격(window)으로 잘라 subsequence를 형성하여 primitive shape[1]로 전환되어진 모습을 볼 수 있다.

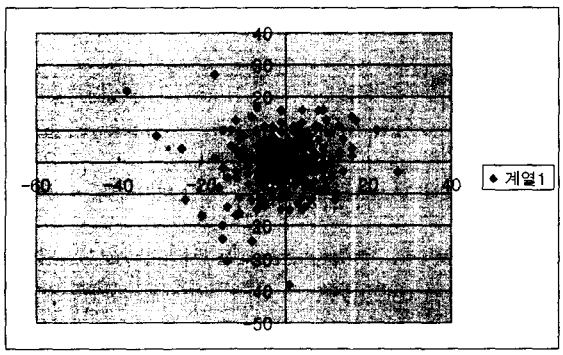


[그림 1] 원시적 모양으로 표현한 시 계열 데이터

위의 그림 1이 유도되는 과정을 좀 더 자세히 살펴보면 전체 sequence를 s, subsequence를 s_i, window width를 w라 가정하자. s = ($x_1, x_2, \ldots\ldots, x_n$)으로 s_i = ($x_i, x_{i+1}, \ldots\ldots x_{i+w-1}$) 으로 나타낸다. 이때 임의의 두 subsequence 사이의 거리는 $d(s_i, s_j)$로 나타낸다. subsequence들 사이의 거리를 측정하기 위해 우선 거리에 대한 정의가 이루어져야 한다. 각각의 subsequence를 Rw 차원의 점으로 mapping 한다. 모든 subsequence들은 Rw 차원의 한 점들과 1:1 로 대응하고 이렇게 mapping된 점들을 이용해 Rw 차원에서 다양한 clustering 기법을 이용하여 분류되어진다. Rw 차원의 점들로 mapping 되어진 결과를 분류 하기위해 본 논문에서 사용한 clustering 기법은 각각의 점들 사이의 거리를 계산해 그 값들을 2차원 Matrix(L_2)로 표현하고 그렇게 계산된 Matrix(L_2)를 이용해 clustering 하게 된다. 이 때 greedy method[7]를 이용해 cluster의 지름을 구

하고 k-means algorithm[7]을 이용해 cluster의 중심을 구하는 방법을 사용하였다. 실제 실험을 하면서 중요한 요소(factor)로 작용하는 것들은 window width, cluster의 지름, 그리고 cluster의 개수인 k값 등이다.

2.2 기존의 처리기법에 의한 실험



[그림 2] R2 차원의 공간에 mapping된 점(point)들

0	148	164	40	81	490	722	293	49	146
148	0	160	164	13	162	490	461	29	50
164	160	0	52	101	178	202	85	101	34
40	164	52	0	85	362	458	121	61	82
81	13	101	85	0	193	461	338	4	29
490	162	178	362	193	0	148	461	245	104
722	490	202	458	461	148	0	293	505	260
293	461	85	121	338	461	293	0	314	225
49	29	101	61	4	245	505	314	0	41
146	50	34	82	29	104	260	225	41	0
117	157	5	25	90	229	269	80	82	41

[그림 3] window width가 2일 때의 Matrix(L_2)

위의 그림 2와 3은 각각 R2 차원의 공간에 mapping된 점들과 그때 mapping된 점들 사이의 거리를 계산한 Matrix(L_2)를 나타낸다. 위의 단계를 거쳐 greedy method와 k-means algorithm을 이용해 k 개의 cluster로 분류하였다.

2.3 Descritized Sequence로부터의 패턴 규칙 찾기

2.2 단계에서 얻어진 cluster들을 각각 A,B,C,.....,K라 하면 A와 B사이에 어떤 연관 관계가 존재하기 위한 조건은 A또는 B가 발생한 빈도수가 minimum support 이상이 되어야 하고 A가 발생한 이후 일정시간 T 이내에 B가 발생할 확률이 minimum confidence 이상이 되어야 한다고 정의한다. 즉 A의 상대적 빈도수 f(A)를 f(A)=F(A)/n (단 n은 sequence의 개수)[1,8]라 하고 F(A,B,T)= $|\{i \mid a_i = A \wedge B \in \{a_{i+w+1}, \cdots, a_{i+w+T-1}\}\}|$ 라 할 때 $C(A \rightarrow B)_T = \frac{F(A,B,T)}{F(A)}$ 는 A가 발생한 다음 최소 w의 시간이 흐르고 T 시간 내에 B가 발생할 확률을 의미한다.[1,8] 여기서 유의할 사항은 window width w 만큼의 시간이 지난 다음부터 고려해야만 한다는 것인데 그 이유는 subsequence가 서로 겹치는 경우를 제외하여야만 하기 때문이다 만약 그러한 사항을 고려하지 않고 실제로 발생할 수 없는 트랜잭션을 발생했다고 가정하면 실험의 결과가 다르게 나타날 수 있다. 따라서 F(A,B,T)= $|\{i \mid a_i = A \wedge B \in \{a_{i+1}, \cdots, a_{i+T-1}\}\}|$ 로 오해하는 일이 발생하지 않도록 유의하여야 한다.

3. 증감 table을 이용한 시 계열 데이터 처리기법

3.1 기존의 방법의 문제점 제시

위에서 설명했듯이 기존의 방법은 각각의 window width를 결정하고 subsequence를 Rw 차원의 공간으로 mapping 한 다음 복잡한 방법의 clustering 작업을 거쳐 Association Rule을 적용하여 패턴간의 연관성을 찾게 된다. 여기서 window width를 바꿀 때 마다 차원이 바뀌게 되고 그것에 따른 여러 번의 연산량을 필요로 하게 된다. 즉 window width는 패턴의 범위를 결정하는데 중요한 요소(factor)이고 그것을 여러 번 반복해서 바꾸어 주면서 유용한 규칙을 찾아내야 하는데 w값이 바뀔 때마다 mapping 되는 공간이 바뀌고 유사도(similarity)[6]의 측정도 복잡해짐과 동시에 clustering에도 많은 연산량을 필요로 한다. 따라서 본 논문에서는 기존의 방법과는 다른 새로운 접근 방법을 제시하였다.

3.2 증감 table을 이용한 방법의 제안

먼저 독립변수 x가 변함에 따라 종속적으로 변하는 f(x) 값을 증가는 양수(+) 감소는(-) 변하지 않고 일정하면 0의 값으로 놓고 이것을 증감 table로 만든다. 또한 양수 혹은 음수의 값을 정할 때 각각의 변화되는 크기에 따라 그 값의 정도(granularity)를 다르게 하였다. 예를 들어 1~3만큼 커지면 +1 4~6만큼 작아지면 -2 이런 방법으로 각각의 값을 정하여 변화량에 따른 보정치를 꾀하였다. 또한 value width란 개념을 써서 독립변수 x의 증가량 즉 x가 1→2, 2→3 등으로 변하면 그 값이 1이 되고 1→3, 2→4 등으로 변하면 그 값은 2가됨을 의미한다. 여기서 value width는 window width보다 클 수 없고 반드시 결정한 window width의 약수가 되어야 한다. 예를 들어 window width가 8일 때 value width는 1,2,4,8 이렇게 네 가지만 가능하게 된다. 위와 같은 방법을 적용해 window width와 value width가 결정되면 value width에 맞춰서 증감 table을 작성하고 window width를 value width로 나눈 값이 그룹을 나누는 한 단위가 된다. 이 단위가 결정되면 각각의 subsequence들은 각각의 그룹에 속하게 된다. 이때 유사도(similarity)가 사용되는데 기존의 방법에서는 mapping된 점들 사이의 거리를 측정해 유사도를 결정한 반면 증감 table을 이용한 방법에서는 각각의 자릿수에서 값(value)의 차를 구한 후 이 값들의 합으로 유사도가 결정된다. subsequence $S_1 = a_1, a_2, a_3, a_4, a_5$ $S_2 = b_1, b_2, b_3, b_4, b_5$ 이라 하고 S_1과 S_2의 유사도를 E_q 라 하면 $E_q = \sum_{i=1}^{5} |a_i - b_i|$ 로 정의된다. E_q의 값이 작을수록 subsequence간의 유사도는 높다. 위와 같이 증감 table을 이용해 sequence를 간략히 표현하고 표현된 sequence를 유사도를 이용해 패턴 그룹으로 나누면 그 이후의 패턴 그룹들 사이의 관계를 유도하는 단계는 기존의 방법과 일치한다.

3.3 증감 table을 이용했을 때의 실험

[표 1] 증감 table

-1	-2	3	1	-1	2	•
-3	1	3	0	1	•	•
-1	2	3	2	•	•	•
•	•	•	•	•	•	•

표 1은 value width가 각각 1,2,3 일 때의 증감을 나타내는 증감 table이다. 위의 table에서 value width가 1일 때의 값을 참조해 보면 실제 value 들이 증감을 어떻게 나타내는지 알 수 있고 value width가 2일 때의 실제 value 들이 어떻게 변하는지 알기 위해서는 table에 나타난 값들을 하나씩 건너뛰어 읽으면 value의 변화를 알 수 있다. 즉 -3→1→3→0→1→..... 의 방식이 아니고 -3→3→1→..... 의 방식으로 참조해야만 한다. 같은 방법으로 value width가 3일 때는 -1→2→..... 의 방식으로 참조하면 연속된 value의 변화를 알 수 있다.

4. 실험 결과의 비교

4.1 그룹으로 분류하는데 걸린 시간 비교

4.1.1 시간 table

[표 2] data 크기가 100일 때의 시간 단위(MS)

	기존의 방법 적용	증감 table 이용
window width=2	7.119	0.050
window width=3	9.391	0.053

[표 3] data 크기가 200일 때의 시간 단위(MS)

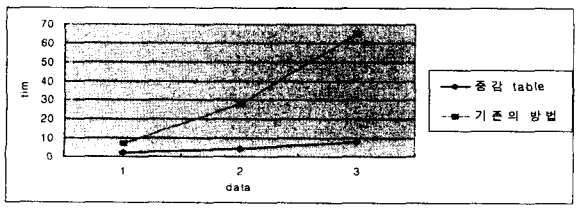
	기존의 방법 적용	증감 table 이용
window width=2	28.473	0.099
window width=3	37.662	0.103

[표 4] data 크기가 300일 때의 시간 단위(MS)

	기존의 방법 적용	증감 table 이용
window width=2	65.266	0.147
window width=3	86.119	0.154

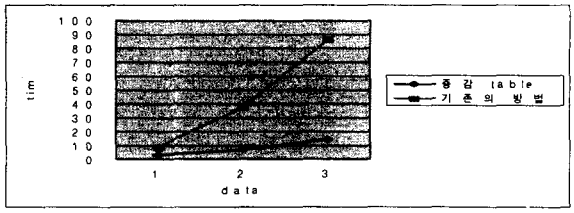
4.1.2 시간 table을 이용한 그래프

단위(data=100개 time=MS)



[그림 4] window width가 2일 때의 그래프

단위(data=100개 time=MS)



[그림 5] window width가 3일 때의 그래프

4.2 Association Rule을 적용했을 때의 결과 비교

window width=2 일 때 증감 table에서 2→2 혹은 그 subsequence와의 유사도가 2이하인 패턴이 발생하고 시간 T가 5이하일 때 -2→-2 혹은 그 subsequence와의 유사도가 2이하인 패턴이 발생할 확률을 기존의 clustering 방법과 비교했을 때 각각 0.50과 0.48로 나타났다. 같은 방법으로 window width=3 일 때 증감 table에서 2→2→2 혹은 그 subsequence와의 유사도가 3이하인 패턴이 발생하고 시간 T가 7이하일 때 -2→-2→-2 혹은 그 subsequence와의 유사도가 3이하인 패턴이 발생할 확률을 기존의 clustering 방법과 비교했을 때 각각 0.30과 0.29로 나타났다.

5. 결 론

본 논문에서 보인바와 같이 시 계열 데이터에서의 패턴들 간의 연관성을 찾는데 기존의 방법과 비교해 증감 table을 이용하는 방법이 더 효율적임을 알 수 있었다. 이러한 현상은 data의 양이 늘어날수록 두드러지게 나타난다. 추후 연구 사항으로 window width를 크게 증가시켰을 때 value width를 달리하여 결과를 비교하고 다양한 패턴들에 대해 실험을 확대함으로써 복잡한 연관성을 알아내는데 더욱 많은 연구가 계속되어야 할 것이다.

6. 참 고 문 헌

[1] Gautam Das, et. al, Rule discovery from time series. Proc. of the 4th Int'l Conference on KDD NY, Aug 27-31. pp 16-22. 1998

[2] R. Agrawal, et. al, Mining Sequential Patterns, In ICDE95, Taipei, Taiwan, 1995

[3] R. Agrawal, et. al, Fast Algorithms for Mining Association Rules in Large Databases, Proc. of the 20th Int'l Conference on VLDB, p.487-499, Sep 12-15, 1994

[4] R. Agrawal, et. al, The Quest Data Mining System. Proc. of the 2nd Int'l Conference on KDD, Portland, Oregon, August, 1996.

[5] Eamonn Keogh, et. al, A probabilistic approach to fast pattern matching in time series databases. In Proc. of the 3rd Int'l Conference of KDD, pages 24-30, 1997

[6] R. Agrawal, et. al, Efficient similarity search in sequence databases. In FODO93, Chicago, 69-84, 1993.

[7] Jiawei Han, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, Aug 2000.