마코프성분을 갖는 데이터셋의 예측모델링†

김선철°, 서성보, 이준욱, 류근호
충북대학교 데이터베이스연구실
{sckim,sbseo,junux,khryu}@dblab.chungbuk.ac.kr

Predictive Modeling for the Data having Marcov property

Sun Cheul Kim°, Sung Bo Seo, Jun Wook Lee, and Keun Ho Ryu
Dept. of Computer Science, Chungbuk National University

요 약

기업과 산업등 여러분야에 적용하기 위하여 인공지능, 통계학, 데이터베이스등의 각 분야에서 활발히 연구되고 있는 데이터마이닝은 알 수 없는 미래에 대한 예측이 가능하다는 장점을 갖기 때문에 더욱 가치가 있다. 데이터셋을 설명하기 위한 설명모델링과 예측을 하기 위한 예측모델링의 두 가지 범주로 나뉘어 발전되어왔으나, 데이터셋을 설명하기 위한 분석보다는 미래를 예측하기 위한 분석의 중요성이 점점 증가되고 있다. 이 논문에서는 마코프 성분을 갖는 과거의 이력 데이터를 기반으로 일정한 시점 또는 일정 기간동안의 변화량을 예측할 수 있는 예측모델링 방법을 제시한다.

1. 서론

대규모의 데이터로부터 데이터간의 관계, 패턴을 탐색하고, 이를 모형화 하여 업무에 적용 할 수 있는 의미 있는 정보로 변환시키기 위하여 데이터셋에 대한 모형을 설정하는 일은 매우 중요하다. 설명적 모델링과 예측적 모델링과 같이 두 종류의 모델링 방법을 고려할 수 있으며, 설명적 모델링은 주어진 데이터가 지니는 특정한 형태 또는 규칙을 찾아내어 대상 데이터셋에 대한 명확한 설명을 하는 것이 주목적이다. 적용할 수 있는 기법으로는 고객들의 상품구매에 따른 상품간에 존재하는 관련성을 찾아내는 연관규칙 탐사 기법 또는 고객세분화와 시장세분화에 적용할 수 있는 클러스터링 기법 등이 있다. 또한, 예측적 모델링 기법은 주어진 데이터셋을 기반으로 예측모델을 만들고 이 모델을 이용하여 새로운 데이터(셋)에 대한 특정한 경향이나 규칙적 패턴을 표현하는 함수식등을 유도하여 일정 시점의 예측 값 또는 일정 기간동안의 경향 등을 예측하는 것이 주된 목적으로, 분류모형, 로그선형모형, 로지스틱모형 등의 기법들이 적용되고 있다.

현재까지 저장된 데이터를 이용하여 미래의 값을 추정할 수 있는 예측방법들 또한 매우 흥미로운 분야이다. 이 논문에서는 현재까지 저장된 데이터으로부터 일정한 시간 간격 상에 연속적으로 존재하는 조건확률 기반의 추이확률메트릭스를 이용하여 일정한 시점 또는 일정 기간동안의 변화량을 미리 알 수 있는 예측모델링 기법을 제시한다.

2. 관련연구

데이터 마이닝의 두가지 모델링 기법중 예측적 모델링 방법으로 분류모형, 일반·로그선형모형, 로지스틱모형과 같은 기법들이 제시되어 왔다.

분류모형[Quin89]은 예제 데이터셋으로부터 데이터셋이 내포하고 있는 특정한 규칙들을 찾아 모형을 설정하고, 이 모형에 분석하고자 하는 새로운 데이터셋을 적용하여 원하는 정보를 얻어내는 방법이다. 그러나 분석가가 원하는 시점에 발생 할 값을 예측할 수 없는 문제점이 있다.

일반·로그선형모형[Heik96]은 두 변수(또는 요인)가 서로 독립이라는 가정하에

$$y = x\beta + \varepsilon$$

와 같은 선형 모형을 추정·적합시키기 위하여 데이터셋의 특성에 따라 독립변수를 로그 또는 다차원 변환과정을 거친 후 특정한 회귀식을 유도하는 방법이다.

로지스틱 회귀모형[Heik96]은 성공, 실패 등과 같이 반응범위 수가 두 개(binary)이거나 순서형(ordinal)의 반응범위를 갖는 범주형 자료들을 연속형의 설명변수들을 이용하여 설명하고자 하는 경우에 사용하는 모형으로서 통계학에서 예측을 위해 연구된 기법이다. 그러나 일반,로그선형모형이나 로지스틱 회귀모형은 데이터베이스에 적용하기 보다는 파일로 저장된 분석용 데이터를 이용하여 분석하고 있기 때문에 기업데이터에 적용하기에는 적합하지 않다.

3. 분석방법

과거의 데이터로부터 미래를 예측하기 위한 많은 방법이 존재한다. 그러나 특정한 시간 또는 특정한 시간 구간 동안 발생하는 예측 값(predictive value) 또는 패턴의 발견에 관한 연구는 그리 많지 않다. 우리는 과거의 데이터를 이용하여 일정 시간 이후의 변화된 값에 대한 예측을 할 수 있는데, 어제와 오늘 즉, 과거데이터들간에 존재하는 상태변화를 이용하여 내일의 상태변화를 예측하고자 한다. 주위의 많은 현상들은 한 시점에서 다음 시점으로 연결되는 과정상에 어떠한 종속성을 갖게 된다[최기현91]. 그러므로 동일한 분포 또는 동일한 함수식 등을 이용하는데는 많은 무리가 따른다. 따라서 이러한 문제를 해결하기 위해 다음과 같은 마코프 성질을 갖는 데이터에 대한 예측 모델링 방법을 제시한다.

3.1 모델의 선택

† 이 연구는 ETRI의 우정정보화 연구부 연구비지원으로 수행되었음.

가산적 상태공간 S와 가산적 지수집합 T를 갖는 확률과정{Xn: $n\in T$}에 대하여, 모든 가산적 집합은 자연수의 부분집합과 일대일의 대응관계가 있으므로 T를 자연수의 부분집합으로 나타낼 수 있다. 즉 S와 T가 유한 또는 무한한 것에 관계없이 S={0,1,2,…}, T={0,1,2,…}로 표시할 수 있게 된다. 만일 $X_n=i$이면 이는 n 단계(시점)에서 확률과정이 상태 i에 있는 사상을 의미한다. 이를 마코프연쇄의 정의를 사용하면 임의의 $t_1<\cdots<t_n$과 $j_1,\cdots j_n$에 대하여 다음의 식이 성립된다.

$$P(X_{t_1}=j_1,\cdots,X_{t_n}=j_n)=P(X_{t_1}=j_1)\times P(X_{t_2}=j_2|X_{t_1}=j_1)\times$$
$$P(X_{t_3}=j_3|X_{t_1}=j_1,X_{t_2}=j_2)\times\cdots\times P(X_{t_n}=j_n|X_{t_1}=j_1,\cdots,X_{t_{n-1}}=j_{n-1})$$
$$=P(X_{t_1}=j_1)\times P(X_{t_2}=j_2|X_{t_1}=j_1)\times P(X_{t_3}=j_3|X_{t_2}=j_2)\times\cdots\times X_{t_n}=j_n|X_{t_{n-1}}=j_{n-1})$$

그러므로 모든 결합 밀도 함수는 임의의 m, n에 대한 조건부 확률은 아래(1)의 계산식에 의해 얻어지고 이를 추이확률이라 한다.

$$P_{ij}^{(m,(m+n)}=P(X_{m+n}=j|X_m=i)\ \cdots\cdots(1)$$

주어진 현 상태가 i이고 n단계 이후에 상태가 j가 되는 확률 값을 n단계 추이확률이라 하며 아래 식(2)와 같이 표현한다.

$$P(X_{m+n}=j|X_m=i)=P_{ij}^{(m)}\ \cdots\cdots(2)$$

즉, n=1 이면 추이확률을 P_{ij}는 현 상태가 i일때 다음 상태가 j가 될 확률값이다. 또한, 다음의 정리[1]이 만족된다.

정리1. $P_{ij}\geq 0, i,j\in S;\ \sum_{j\in S}P_{ij}=1, i\in S$ [Ralp97]

위의 식을 메트릭스로 표현하여 P_{ij}를 원소로 지닌 메트릭스 P = { P_{ij} }를 추이확률메트릭스라 부른다.
예를 들어, "내일 상품 판매량이 증가할 것인가?, 그렇지 않으면 감소할 것인가?"라는 문제가 오늘의 판매결과에 의존한다고 가정할 때, 0은 판매량이 증가하는 경우이고, 1은 판매량이 감소하는 경우를 나타낸다면 오늘 판매량이 증가한 상태에서 내일 판매량이 증가될 확률이 $1-\alpha$ 이고, 오늘 판매량이 감소한 상태에서 내일 판매량이 감소될 확률을 β라고 한다면 추이메트릭스은 아래와 같다.

$$P=\begin{bmatrix} P_{00}, P_{01} \\ P_{10}, P_{11} \end{bmatrix}=\begin{bmatrix} 1-\alpha, \alpha \\ \beta, 1-\beta \end{bmatrix}$$

3.2 모델의 단순화

추이확률 P_{ij}와 초기분포 pi에 의하여 마코프연쇄에 대한 모든 확률을 계산 할 수 있다. 다시 말하면, 현재 i 상태에서 n단계이후에 j 상태가 될 n-단계 추이확률 $P_{ij}^{(n)}$와 조건확률이 아닌 n단계 상태확률 $P(X_n=j)$도 추이확률 P_{ij}와 초기분포 pi에 의하여 계산되어 진다. $P_{ij}^{(1)}=P_{ij}$이고 $P_{ii}^{(0)}=1, P_{ij}^{(0)}=0 (i\neq j)$일 경우 아래 정리2를 사용할 수 있다.

정리2. C-K Equation [Ralp97]

$$P_{ij}^{(n+m)}=\sum_{k=0}^{\infty}P_{ik}^{(n)}P_{kj}^{(m)},\ n,m>0, i,j=0,1,2,\cdots,\infty$$

위 정리는, 현 상태가 i일 때 (n+m)단계에서 j로 가기 위해서는 처음 n단계에서 어떠한 상태 k를 거쳐 나머지 m단계에서 j로 가야 한다는 것이다. 그러므로 처음 n 단계에 대한 확률은 $P_{ik}^{(n)}$이 되고 후의 m 단계가

$$P_{kj}^{(m)}=P(X_{n+m}=j|X_n=k)=P(X_m=j|X_0=k)$$

임을 나타낸다. 다시 말해서 $P^{(n)}$가 $P_{ij}^{(n)}$을 원소로 하는 메트릭스라면, 다음과 같은 메트릭스의 곱을 의미한다.

$$P^{(n+m)}=P^{(n)}\cdot P^{(m)}\cdots\cdots(3)$$

3.2 모형으로의 적합

기존의 데이터의 값을 이용하여 미래의 일정 시점후의 예측 값을 얻기 위해서는, 적용가능한 데이터 모델이 필요하며, 이 논문에서는 상태공간이 둘 즉 0과 1인 경우에 한해서 기술한다.
상태공간이 둘인 마코프연쇄의 추이메트릭스 P는 한가지 형태만 존재한다. 이런 경우 $P^{(n)}$를 다음과 같이 구할 수 있다.

$$P_{00}=1-\alpha,\ P_{01}=\alpha$$
$$P_{10}=\beta\ ,\ P_{11}=1-\beta$$

이고, 정리2를 사용하면 아래와 같이 식(4)를 유도 할 수 있다.,

$$P_{00}^{(1)}=1-\alpha,$$
$$P_{00}^{(n)}=(1-\alpha)P_{00}^{(n-1)}+\beta P_{01}^{(n-1)}, n>1$$
$$=\beta+(1-\alpha-\beta)P_{00}^{(n-1)}, n>1$$
$$=\frac{\beta}{(\alpha+\beta)}-\frac{\alpha(1-\alpha-\beta)^n}{(\alpha+\beta)}\ \cdots\cdots(4)$$
$$P_{01}^{(n)}=\frac{\alpha}{(\alpha+\beta)}-\frac{\alpha(1-\alpha-\beta)^n}{(\alpha+\beta)}\ \cdots\cdots(4')$$

즉, $P^{(1)}$는 현재상태를 의미하고 $P^{(n)}$은 현재 상태에서부터 n-1번째의 상태변화를 의미한다. 따라서 현재의 시점으로부터 다음시점에의 상태변화는 $P^{(2)}$가되고, 임의 구간 i번째 후의 상태변화는 $P^{(1+i)}$가 되며 다음과 같다.

$$P_{00}^{(i+1)}=P_{00}^{1}\cdot P_{00}^{i}$$

3.3 적 용

상품 및 거래 데이터베이스에서 시간(시점)단위별로 각각의 상품에 대한 판매량의 등락을 이용하여, 미래의 특정 i시간값 이후의 상태를 예측할 수 있다. 이때 가장 중요한 문제는 P_{00}^{1}을 구하는 것이다.
아래 그림3로부터 P_{00}^{1}은 t_k 시점까지의 조건확률의 기대값이므로

$$P_{00}^{1}=\sum_{i=1}^{k}\left(\frac{P_{00}^{i+1}}{P_{00}^{i}}\right)\ \cdots\cdots(5)$$

이 식(5)가 되며, 식(4)을 사용할 수 있다. 예를 들어 4일 후의 값에 대한 예측을 하고자 할 경우, P_{00}^{1}는 식(4)를 사용하여 계산할 수 있고 값이 ($1-\alpha$)이라고 가정한다면, 4일 후 예측값은

$$P_{00}^{(4)}=\frac{\beta}{(\alpha+\beta)}-\frac{\alpha(1-\alpha-\beta)^4}{(\alpha+\beta)},\ P_{01}^{(4)}=1-P_{00}^{(4)}$$
$$P_{10}^{(4)}=\frac{\alpha}{(\alpha+\beta)}-\frac{\alpha(1-\alpha-\beta)^4}{(\alpha+\beta)},\ P_{11}^{(4)}=1-P_{10}^{(4)}$$

와 같이 계산될 수 있고, 따라서, 원하는 시간단위 이후의 값을 예측 할 수 있다. 이러한 예측기법의 알고리즘을 아래에 기술 하였다.

추이확률메트릭스 생성 알고리즘

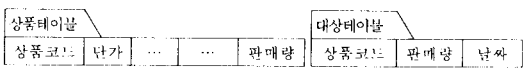
```
//추이확률 메트릭스의 생성
Vector TransitionMatrix(){
    ......
    for(i=0;i<Items.size();i++){
       /*각 변수의 선언 및 Item별 정렬된 데이터셋의 추출*/
       ......
       /* 추이확률의 계산 */
       for(j=2;j<itemSet[].size();j++){
          flag1 = itemSet[j-1]-itemSet[j-2];
          flag2 = itemSet[j-2]-itemSet[j-1];
          if (flag1>0 and flag2>0){
            p00++ ;
          }else if(flag1>0 and flag2<0){
```

```
            p01++;
          }else if(flag1<0 and flag2>0) {
            p10++;
          }else if(flag1<0 and flag2<0){
            p11++;
          }
        }
        ......
    }
}
/* n시점 후  마코프연쇄과정 */
Vector MarcovChain(Vector matrix, int n)
{
    ........
    //n번째 시점의 상태에 따른 각 구성요소의 계산
    a = matrix.elementAt(1);
    b = matrix.elementAt(2);
    P00_n = b/(a+b) - a(1-a-b)^n/(a+b);
    P01_n = 1 - P00_n;
    P10_n = a/(a+b) - a(1-a-b)^n/(a+b);
    P11_n = 1 - P10_n;
    ........
}
```

위의 알고리즘에서 flag1과 flag2는 각 시점별 증가·감소되는 상태 변화를 나타내는 변수로 사용되며, TransitionMatrix 메소드에 의해 생성된 상품(Item)별 추이메트릭스를 MarcovChain메소드에 적용하면 n단계이후에 변화되는 상태 값을 얻을 수 있다.

4. 실험 및 결과

그림1의 (a)테이블에서 (b)와 같은 분석용 테이블로 데이터를 추출하면, 표1과 같이 된다. 이 데이터셋을 분석하기 위하여 상품코드와 요일(시점)별 판매량을 이용한다.



(a) (b)

그림1. 상품·분석 테이블

표1. 대상 테이블

ProductID	SellingSum	Date	ProductID	SellingSum	Date
1	60	99-10-01	2	130	99-10-08
1	70	99-10-02	2	120	99-10-09
1	80	99-10-03	2	110	99-10-10
1	90	99-10-04	2	100	99-10-11
1	100	99-10-05	2	90	99-10-12
1	80	99-10-06	2	80	99-10-13
1	50	99-10-07	3	55	99-10-01
1	130	99-10-08	3	50	99-10-02
.....	.....	.....	.....	.....	.....

분석을 위한 핵심이 되는 값은 P_{00}와 P_{01}이며, 이 값을 이용하여 식(4)에 적용하면 미래 특정시점의 예측값을 얻을 수 있다.

그림2. 시간단위의 표현

그림2에서 t_i $(i = 0,1,\cdots,n)$는 각 시점을 나타내며, t_k는 현재시점을 의미한다. 따라서, t_k의 추이확률메트릭스를 계산해야 하며, 다음과 같이 계산할 수 있다.

$$P_{00}^{(1)} = \frac{\sum_{i=1_0}^{k}(P_{i+1}=0|P_i=0)}{k} = 1-\alpha$$

이며, $P_{10}^{(1)}$도 쉽게 계산될 수 있다.

$$P_{00}^{(i+1)} = \frac{\alpha}{(\alpha+\beta)} - \frac{\alpha(1-\alpha-\beta)^i}{(\alpha+\beta)} = \beta$$

위 식을 이용하여 현재시점으로부터 임의 i시점 후의 예측값을 계산할 수 있다.
아래 표는 그림1 테이블로부터 상품의 판매량에 대한 각 시점(날짜)별 예측값을 나타낸다.

표2. 시점별 예측값

ProductID	P00/P01/P10/P11	t(k)	t(k+1)	t(k+2)	t(k+3)	t(k+4)	t(k+5)
1	0.9/0.1/0.5/0.5	0.9	0.77	0.81	0.82	0.83	0.83
2	0.2/0.8/0.43/0.57	0.2	0.18	0.00	0.04	0.03	0.04
3	0.333/0.667/0.5/0.5	0.3	0.52	0.41	0.43	0.43	0.43

t(k)는 식(5)에 따라 계산되며, 현시점의 상태를 의미한다. 그리고 t(k+i), i=1,…,5 는 현시점으로부터 i번째 이후의 시점에 대한 상태를 나타낸다. 또한, 아래 그림3은 표2에 대한 그래프를 나타내며, 각 상품별 미래의 상태변화에 대한 경향을 알 수 있다.

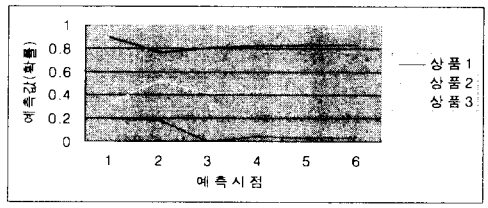


그림3. 예측 그래프

5. 결론 및 향후연구

데이터 마이닝의 지식탐사 모델로는 설명적 모델링과 예측적 모델링이 있다. 설명적모델링은 데이터셋이 지니는 특정한 형태 또는 규칙을 설명하기 위하여 적용하며, 예측적모델링은 데이터 셋을 기반으로 예측모델을 만들어 특정한 경향이나 규칙적 패턴을 표현하는 함수식 등을 유도하여 일정 시점의 상태변화를 예측하기 위한 분석기법이다. 과거의 데이터를 이용하여 미래의 값을 추정할 수 있는 예측방법들 또한 매우 흥미로운 분야이다. 이 논문에서는 마코프연쇄의 성질을 지니는 데이터셋에 대한 예측모델을 제시하였다. 따라서, 이 제안 모델은 과거 이력데이터를 사용하여 미래의 값을 추정하는데 사용할 수 있다.
현실 세계의 데이터셋은 일반적으로 다차원적이다. 따라서 다차원 상태공간에 적용 가능한 예측방법들이 필요하다. 그러나 상태공간이 확장되면 위에서 설명한 추이확률메트릭스의 시점별 순환관계를 유도하기가 매우 어렵다. 따라서 적률생성함수나 확률생성함수를 이용한 다차원상태공간에서 적용 가능한 예측기법에 대한 연구가 필요하다.

[참고문헌]

[Quin89] Quinlan. J. R "Induction of decision trees using minimum description length principle, Information and Computation", 1989.
[Agra92] Agrawal. R. et.al., "An Internal Classifier for Database Mining Application", In Proc of VLDB,1992.
[Heik96] Heikki Mannila: "Data mining: machine learning, statistics, and databases", In Proc of 8th International Conference on Scientific and Statistical Database Management, Stockholm June 18-20, 1996.
[Ralp97] Ralph. L. and Brace Clarke, "Probability and Random process", pp 185-188 published JohnWiley & Sons, 1997.
[최기현91] 최기현, "응용확률론 입문", pp157-159 자유아카데이, 1991.