마코프성분을 갖는 데이터셋의 예측모델링

김선철*, 서성보, 이준욱, 류근호
충북대학교 데이터베이스연구실
(sekim, sbseo, junux, khryu)@dlabl.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
기계과 산업의 영역분야에 적용하기 위하여 인공지능, 통계학, 데이터베이스 등의 각 분야에서 발전해 연구되고 있는 데이터다이나믹은 아직 없는 미래에 대한 예측이 가능하다는 점을 가지고 때문에 더욱 가치가 있다. 데이터셋은 설명하기 위한 설명모델링과 예측을 하기 위한 예측모델링의 두 가지 방법으로 나뉘어 설명되어있으나, 데이터셋을 설명하기 위한 분석보다는 미래를 예측하기 위한 분석의 중요성이 점점 증가되고 있다. 이 논문에서는 마코프 성분을 갖는 과거의 이력 데이터를 기반으로 이전한 시점 또는 일정 기간동안의 변화량을 예측할 수 있는 예측모델링 방법을 제시한다.

1. 서론
데이터의 데이터로부터 데이터간의 관계, 패턴을 탐색하고, 이로 모델화하여 업무에 적응 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 담는 특성한 형태 또는 규칙을 찾아내어 대상 데이터셋에 대한 명확한 설명을 하는 것이 주목이다. 설명할 수 있는 기간으로는 고객들의 상품구매에 따른 상품판매에 존재하는 관계성을 찾아내는 관련규칙 탐색 기법 또는 고객분석과 시장분석에 적용할 수 있는 고객스터디 방법의 영역이다. 반면, 예측적 모델링 기법은 주어진 데이터셋을 기반으로 예측모델을 만드는 것이 수치형 데이터에 대한 특징을 찾고, 예측해 보는 시나리오를 적용한 방법을 유효하게 이용할 수 있는 기법의 일종으로, 본론에서 근거의 성능을 검증하기 위한 함수식을 적용하여 시점의 예측값을 수치형 데이터에 대한 특성의 중요성을 이용하여 설명하고 있다. 현재까지 설명된 데이터를 이용하여 미래의 값을 추정할 수 있는 예측방법은 매우 중요하며, 이 논문에서는 현재까지 설명된 데이터로부터 설명할 수 있는 전체 데이터의 중요성을 설명하는 예측모델링 방법의 개념을 설명하고, 이 때, 설명적 모델링의 중요성을 인지하고, 이에 대한 설명적 모델링 방법을 제시한다.

2. 관련연구
데이터 마호닝의 두 가지 모델링 기법 중 예측적 모델링 방법으로 분류모델, 인공 - 로그산형모델, 로지스틱모델과 같은 기법들이 제시되어 왔다.
유형모형(Quin8)은 데이터 데이터로부터 데이터셋의 내포하고 있는 특성의 갑책을 찾아 모델을 설정하고, 이 모델에 분석

‡ 이 연구는 ETRI의 우정정보화 연구특구연구비자금으로 수행하였음.

3.6. 분석방법
과거의 데이터로부터 미래 예측하기 위한 방법의 종류로 설명한 것에 관계없이, 예측 모델링의 개념을 설명할 때 예측값 라인에 해당 예측값에는 의미 없는 기술이 포함되어 있다. 이러한 사례는 분석에 따라 다양한 방법을 이용하기 위해 다음과 같은 마코프 성분을 갖는 데이터에 대한 예측 모델링 방법을 제시한다.

3.1 모델의 선택
가산적 상태공간 $S$와 가산적 지수공간 $T$을 갖는 확률과정 $(X_t, n(t))$에 대하여, 모든가산적 집합은 자연수의 부분집합과 임의의 대응관계가 있을 뿐 $T$을 자연수의 부분집합으로 나타낼 수 있다. 즉 $S$와 $T$가 유한 또는 무한한 것에 관계없이 $S = \{0, 1, 2, \ldots\}$, $T = \{0, 1, 2, \ldots\}$로 표시할 수 있다. 만약 $X$이면 이는 $n$ 단계 (시점)에서 확률과정이 상태 $x$에 있는 사건을 의미한다. 이를 마코프연쇄의 성질을 사용하면 의의의 $i < j < l$, $m < n < j$에 대하여 다음의 식이 성립한다.

$$P(X_{n-1} = x) = P(X_{n-1} = x, X_n = y) = \sum_{y \in S} P(X_{n-1} = x, X_n = y) = \sum_{y \in S} P(X_{n-1} = x) P(X_n = y | X_{n-1} = x)$$

그러므로 모든 결합 밀도 함수는 이의 $m, n$에 대한 표본측 확률은 아래와 같이 유의미합니까. 

$$P(X_n = x, X_{n-1} = y) = P(X_n = x | X_{n-1} = y) \cdot P(X_{n-1} = y)$$

주어진 모든 상태가 $i$이고 $n$단계 이후에 상태가 $j$가 되는 확률 값을 $n$단계 주이의확률이라고 하며 아래와 같이 표기한다.

$$P(X_{n-1} = y, X_n = j) = P(n)(i, j)$$

즉, $n$면 주이의확률은 $P(n)$이 현 상태가 $i$일 때 다음 상태가 $j$가 될 확률값들이다. 또한, 다음의 정리[1]이 만족된다.

정리1. $P(n) = \{s, t \mid n \geq 0 \}$ 위의 식을 벡터로 표기하여 $P(n)$을 원소로 가진 벡터 를 $P(n)$을 주이의확률이라고 부른다.

예를 들어, "내일 수동 관매대금이 증가할 것인가?" 사람지 않으면, 0을 관매대금이 증가하는 것으로 가정할 때, $0$은 관매대금이 증가하는 경우가 됩니다. 관매대금이 감소하는 경우를 나타낸다면 오는 관매대금이 증가한 상태에서 내일 관매대금이 증가될 확률은 $1/4$이며, 관매대금이 감소한 상태에서 내일 관매대금이 감소될 확률을 $1/4$으로 하면 주이의확률은 아래와 같다.

$$P(i, j) = \begin{bmatrix} 0 & 1/4 & 0 & 1/4 \\ 1/4 & 0 & 1/4 & 0 \\ 1/4 & 1/4 & 0 & 0 \\ 0 & 0 & 1/4 & 1/4 \end{bmatrix}$$

3.2 모형의 단순화

추이확률 $P(n)$의 초기조건 $p$에 의하여 마코프연쇄에 대한 모든 확률을 계산할 수 있다. 다시 말하자면, 현재 $i$ 상태에서 $n$단계 이후에 $j$ 상태가 될 $n$단계 주이확률 $P(n)$을 조건확률이 아닌 $n$ 단계 상태확률 $P(X_n = y)$로 추이확률 $P(n)$와 초기조건 $p$에 의하여 계산이 가능하다. $P(n) = P(n)$이고 $P(0) = 1, P(0) = 0$일 경우 아래와 같이 사용할 수 있다.

정리2. C-E Equation (Halfl7)[1]

$$P(n) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \begin{bmatrix} 1 \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{bmatrix} \begin{bmatrix} a & b \\ a & b \\ a & b \\ a & b \end{bmatrix}^{n-j} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

이 식(5)의 경우, $a, b$를 사용할 수 있다. 예를 들어, 4일 후의 값을 계산할 수 있다. 이는 다음과 같다. $P(n)$의 구간 $1$에서의 성질을 계산할 수 있다. $P(n) = P(n)$로 가정하면 $n$단계 주이확률의 값이 $0$일 경우 아래와 같이 사용할 수 있다.

$$P(n) = \begin{bmatrix} 1 \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{bmatrix} \begin{bmatrix} a & b \\ a & b \\ a & b \\ a & b \end{bmatrix}^{n-j} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

이 식(5)의 경우, $a, b$를 사용할 수 있다. 예를 들어, 4일 후의 값을 계산할 수 있다. 이는 다음과 같다. $P(n)$의 구간 $1$에서의 성질을 계산할 수 있다. $P(n) = P(n)$로 가정하면 $n$단계 주이확률의 값이 $0$일 경우 아래와 같이 사용할 수 있다. 이는 아래와 같이 사용할 수 있다. 이는 아래와 같이 사용할 수 있다. 이는 아래와 같이 사용할 수 있다. 이는 아래와 같이 사용할 수 있다.
위의 알고리즘에서 flag1과 flag2는 각 시점별 증가·감소되는 상태 변화를 나타내는 변수로 사용되며, TransitionMatrix 메소드의 매개변수로 필요하다. 예상된 상품을 구성하기 위해 상품 코드와 요일(시점)별 판매량을 이용한다.

4. 실험 및 결과

그림의 (a)와 (b)기의 전체 형태를 본론으로 데이터를 추출하면, 표1과 같이 된다. 이 데이터셋은 본론 기여に対して 상품코드와 요일(시점)별 판매량을 이용한다.

<table>
<thead>
<tr>
<th>상품코드</th>
<th>연차</th>
<th>...</th>
<th>판매량</th>
<th>상품상호</th>
<th>판매량</th>
<th>납품</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>(b)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

그림 1. 상품·본론 데이터

<table>
<thead>
<tr>
<th>ProductID</th>
<th>SellingSu</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>60</td>
</tr>
<tr>
<td>2</td>
<td>70</td>
</tr>
<tr>
<td>3</td>
<td>80</td>
</tr>
<tr>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>110</td>
</tr>
<tr>
<td>7</td>
<td>120</td>
</tr>
<tr>
<td>8</td>
<td>130</td>
</tr>
<tr>
<td>9</td>
<td>140</td>
</tr>
<tr>
<td>10</td>
<td>150</td>
</tr>
</tbody>
</table>

표 1. 대상 데이터

분석을 위한 핵심이 되는 것은 $P^{(1)}$와 $P^{(0)}$이며, 이 값을 이용하여 시(4)에서 적용하면 미래 특정시점의 예측값을 얻을 수 있다.

\[
P^{(1)}_{ij} = \frac{a - a - (1 - a - \beta)^i}{(a + \beta)}
\]

위 식은 이용하여 현재 시점부터 임의 시점의 예측값을 계산할 수 있다.

이상의 결과는 그림 1에 보이는 상품 관란제에 대한 각 시점(날짜별) 예측값을 나타낸다.

표 2. 시점별 예측값

<table>
<thead>
<tr>
<th>ProductID</th>
<th>P00/P01/P02/P03/P04/P05/P11/P12/P13/P14/P15</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6/0.1/0.5/0.3/0.2/0.1/0.0/0.0/0.0/0.0/0.0</td>
</tr>
<tr>
<td>2</td>
<td>0.2/0.3/0.4/0.5/0.6/0.7/0.8/0.9/1.0/1.1/1.2</td>
</tr>
<tr>
<td>3</td>
<td>0.3/0.4/0.5/0.6/0.7/0.8/0.9/1.0/1.1/1.2/1.3</td>
</tr>
</tbody>
</table>

$t(k)$는 식(5)에 따라 계산되며, 현재 시점의 상태를 의미한다. 그리고 $t(k-1), t(k-2), \ldots, t(k-n)$는 이전 시점의 상태에 대한 상태값을 나타낸다. 또한, 이론적으로 표2에 대한 그래프를 나타내며, 각 상품별 미래의 상태변화에 대한 경향을 알 수 있다.

5. 결론 및 향후연구

데이터 마이닝의 최적화는 필수적 문제이며, 예측적 모델링은 이용할 수 있다. 이는 다양한 모델의 사용을 통해 데이터를 모델화하고, 그 결과를 예측에 이용할 수 있다. 이를 위해 본 연구에서는 다양한 데이터를 모델화하고, 그 결과를 예측에 이용할 수 있다. 이를 위해 본 연구에서는 다양한 데이터를 모델화하고, 그 결과를 예측에 이용할 수 있다. 이를 위해 본 연구에서는 다양한 데이터를 모델화하고, 그 결과를 예측에 이용할 수 있다.

[참고문헌]


