Proceedings of Joint Conference of The Korean Data Analysis Society and Korean Data And Information Science Society, October 17-18, 2003, pp. 107-113

전이행렬을 이용한 수익데이터 분석

임승범[1] · 강창완 · 김규곤[2]

요 약

최근 활발히 행하여지고 있는 금융 CRM(customer relationship management)의 주요 목적은 고객의 이해도 증진을 통하여 은행의 수익성을 높이는데 있으며 또한 그 과정에서 높은 수익과 낮은 수익을 주는 고객을 여러 가지 유형으로 나누어 관리에 효율성을 도모한다. 일반적으로 고객 세분화의 중요변수로 고객수익성을 고려하고 이러한 고객 세분화 결과에 의해 마케팅 시사점을 도출하게 된다. 본 연구에서는 고객 세분화 그룹에 따른 수익성 변동과정을 모형화하여 보다 효율적인 고객관계 관리를 가능하게 하는데 있다. 수익성 변동의 모형화 과정은 수익금액에 따라 고객을 몇 개의 범주로 분류하여 여러 기간에 걸쳐 나타내는 고객별 범주의 변화 추이를 전이행렬(transition matrix)로 나타내고 마코프 모형을 이용한 전이 확률의 추정을 통하여 다음 시점에서의 각 범주별 고객의 수를 예측 가능함을 보인다.

주요용어 : 데이터마이닝, 전이행렬, 마코프모형.

1. 서론

최근 들어 금융, 보험, 카드회사와 그 이외의 다양한 분야에서 고객관계관리가 이루어지고 있다. 이는 국내 대부분의 기업들이 성숙기를 맞이하고 있기 때문에 CRM을 통하여 기존 고객에 대한 이해도 증진 및 충성도를 제고시켜 수익성을 극대화하기 위한 노력을 기울이고 있음을 보여주고 있다(한동철, 2001). 금융권 기업의 CRM에서도 수익성 관점에서 목표마케팅, 고객세분화, 신용평가 및 이탈정도 등에 관심을 두고 있으며 이러한 분석을 통하여 수익성을 높일 수 있는 방법을 창출하며, 기존의 분석에 빈번히 사용되는 기법이외에도 다양한 이론을 적용하여 수익성 제고의 방법 및 보다 다각적인 해석을 시도하고 있다. 더욱이 금융 CRM에서 가장 중요한 관심 변수는 수익성이므로 고객별 수익금액의 파악과 그 크기에 따른 고객 분류(segmentation)가 CRM의 주요 목적이라 할 수 있다. 이러한 고객 분류는 분기별 또는 연간 총 수익금액을 기준으로 최상의 수익을 주는 그룹으로부터 손실을 주는 그룹까지 몇 개의 범주로 분류하게 된다. 보통의 경우 이러한 분류 방법을 통하여 연간 수익금액을 기준으로 고객을 범주화하게 되는데 이러한 경우 최상의 수익을 주는 그룹을 살펴보면 12개월간 꾸준히 높은 수익을 주면서 그러한 그룹에 분류된 고객도 있지만,

[1]614-714 부산시 부산진구 가야동 산24, 동의대학교 정보통계학과 대학원 석사과정.
[2]614-714 부산시 부산진구 가야동 산24, 동의대학교 정보통계학과 교수.

12개월 중 대부분을 수익을 주지 못하다가 마지막에 큰 수익을 주어서 분류된 고객도 있을 것이다. 즉 각 그룹별 고객이 속한 그룹에 유지하려는 정도가 강한 고객과 다른 그룹으로의 이탈정도가 강한 고객이 섞여서 구성되어진다. 이러한 고객별 그룹에 대한 유지정도와 이탈정도를 추정 할 수 있다면 각 그룹별 유지정도에 따른 기대되어지는 수익과 이탈정도에 따른 변동되어질 수익을 계산할 수 있으며 더 나아가 다음 시점에서 고객별로 유지 및 이탈을 예측하는 등의 다각적인 해석이 가능하다.

이에 본 논문에서는 금융 CRM에서 주요 관심변수가 되는 수익금액을 기존의 연속형 자료 측면의 분석기법 이외에 범주형 자료분석과 탐색적 자료분석 측면에서 다각적인 해석을 시도하고자 한다. 따라서 수익금액을 기준으로 고객을 몇 개의 범주로 분류하여 여러 기간에 걸쳐 나타내는 고객별 범주의 변화 추이를 전이행렬(transition matrix)로 나타내고 마코프 모형을 이용한 전이 확률의 추정을 통하여 다음 시점에서의 각 범주별 고객의 수의 예측 가능함을 제시하고자 한다.

2. 마코프 모형 (One-step Markov Model)

고정된 패널을 대상으로 여러 번 측정된 동일한 개수의 범주에 대하여 시간의 흐름에 따른 각 패널의 범주 사이의 전이 모습을 전 시점과 현 시점이라는 시차를 가지는 2차원 분할표로 정리할 수 있다<표 1>.

<표 1> 전이행렬

시점	t=k+1					Total
	i \ j	1	2	.....	J	
t=k	1	$y_{11}(k+1)$	$y_{12}(k+1)$	.....	$y_{1J}(k+1)$	$y_1(k)$
	2	$y_{21}(k+1)$	$y_{22}(k+1)$	.....	$y_{2J}(k+1)$	$y_2(k)$
	⋮	⋮	⋮	⋮	⋮	⋮
	I	$y_{I1}(k+1)$	$y_{I2}(k+1)$	.....	$y_{IJ}(k+1)$	$y_I(k)$
						N

$y_{ij}(t)$: $t-1$시점에서 i상태였다가 t시점에서는 j인 개체의 수.
$y_i(t-1)$: $t-1$시점에서 i상태인 모든 개체의 수.
단 $i=1,2,\cdots,I,\ j=1,2,\cdots,J,\ t=1,2,\cdots,T$.

이러한 분할표를 전이행렬(transition matrix)이라 하며, 전이행렬을 통하여 동일한 범주에 머무르는 유지정도와 다른 범주로의 전이정도를 살펴볼 수 있다. 또한 여러 개의 전이행렬이 있는 경우 마코프 모형(one-step markov model)을 통하여 전이확률을 구할 수 있고, 구하여진 전이확률을 통하여 다음 시점의 전이행렬도 예측이 가능하다. 이때 주어진 전이행렬에서의 전이확률을 다음과 같이 표현하면 주어진 전이행렬에서의 전이확률은 최대우도방법에 의해 다음과 같이 구할 수 있다.

$$P_{ij}=\frac{y_{ij}}{y_{i+}}=\frac{\sum_{t=1}^{T} y_{ij}(t)}{\sum_{t=1}^{T} y_i(t-1)} \quad (1)$$

여기서 구한 전이확률의 최대우도추정량이 정상성을 만족한다면 아주 좋은 예측모형이라 알려져 있다(Yvonne M. M. Bishop et. (1975)).

3. 사례연구

2002년 8월부터 12월까지 이루어진 국내 특정 B은행의 CRM에서 수신거래에 의한 수익을 주는 고객의 모집단 22만 명의 고객자료 중에서 외국인과 15세 이하의 비경제활동 고객, 수익금액의 누락값 또는 오류가 있는 자료를 제외하고 최종적으로 선정된 203201명을 모집단으로 정의하고 여기서 1000명을 표본으로 단순임의 추출하였다. 추출된 표본은 모집단의 분포와 유사하게 나타났으며 분석에 사용되어질 수익 자료는 2002년 1월부터 2002년 7월까지의 자료이다. 여기서 2002년 1월의 수익자료의 백분위 수를 기준으로 추출된 1000명의 고객을 분류시킨 결과를 <표 2>과 같이 나타내었다.

<표 2> 표본의 1월 수익자료 백분위 수에 의한 고객분류 및 기술통계

그룹	백분위	N	Min	Max	Mean	S.D.
A	상위 1 ~25	250	1562	268773	15930.07	26785.67
B	상위 25~50	250	-250	1476	234.02	457.37
C	상위 50~75	250	-636	-253	-390.73	98.42
D	상위 75~100	250	-774488	-637	-6116.32	49285.45

<표 2>에서 평균값을 기준으로 그룹별 설명을 하면, A그룹은 B은행에 많은 수익을 주고 있는 고객들이 속하고, B그룹은 적은 수익을, C그룹은 적은 손해를, D그룹은 많은 손해를 입히는 고객들로 구성된다.

이렇게 분류를 하게 되면 초기 시점인 1월에서는 모든 그룹의 고객 수가 250명으로 나타나지만, 1월의 분류기준으로 2월부터 7월까지의 수익자료를 A~D의 범주로 분류하여 현 시점과 전 시점간의 범주화된 수익자료의 교차표를 작성하게 되면 각 그룹별 고객의 수는 달라지게 된다. 이러한 교차표를 전이행렬이라 하며 <표 3>~<표 7>에 정리하였다.

이러한 전이행렬을 통하여 그룹별 전체 고객의 수를 고찰하면 1월에서 6월로 갈수록 A그룹은 변화가 약하며, B와 D 그룹은 증가되는 경향을 보이며, C 그룹은 큰 폭으로 감소하여 시간이 흐를수록 그룹별 고객의 수가 전이하고 있음을 볼 수 있다. 이렇게 1월에서 6월까지 그룹별 고객의 수에 대한 전이정도를 식 (1)을 통하여 확률로서 추정할 수 있으며 <표 8>에 정리하였으며, 추정된 전이확률에서도 같은 결과를 보이고 있음을 확인 할 수 있다. 식 (1)에 의해 추정된 전이확률이 정

상성(stationarity)을 만족하여 예측 모형으로의 사용이 가능하여 <표 7>의 6월의 그룹별 전체 고객수와 <표 8>을 이용하여 6월-7월의 전이행렬을 <표 10>과 같이 예측되었다.

이러한 마코프 모형을 이용하여 추정된 전이확률을 예측모형으로 하여 계산되어진 <표 10>의 모형평가를 위하여 <표 9>의 실제 6월-7월의 전이행렬과의 적합도 검정 결과(d.f.=15, χ^2=17.290, p=0.302)에서도 두 분포가 동일한 것으로 나타나 마코프 모형을 이용하여 추정된 전이확률을 예측모형으로서 사용 가능함을 볼 수 있다.

<표 3> 1월-2월의 전이행렬

1월 \ 2월		A	B	C	D	전체
A	빈도	206	21	2	21	250
	%	82.4	8.4	0.8	8.4	
B	빈도	16	173	24	37	250
	%	6.4	69.2	9.6	14.8	
C	빈도	4	41	160	45	250
	%	1.6	16.4	64	18	
D	빈도	14	20	33	183	250
	%	5.6	8	13.2	73.2	
전체		240	255	219	286	1000

<표 4> 2월-3월의 전이행렬

2월 \ 3월		A	B	C	D	전체
A	빈도	188.0	21.0	4.0	27.0	240
	%	78.3	8.8	1.7	11.3	
B	빈도	26.0	168.0	24.0	37.0	255
	%	10.2	65.9	9.4	14.5	
C	빈도	3.0	52.0	129.0	35.0	219
	%	1.4	23.7	58.9	16.0	
D	빈도	24.0	32.0	55.0	175.0	286
	%	8.4	11.2	19.2	61.2	
전체		241.0	273.0	212.0	274.0	1000

<표 5> 3월-4월의 전이행렬

3월 \ 4월		A	B	C	D	전체
A	빈도	187	25	6	23	241
	%	77.6	10.4	2.5	9.5	
B	빈도	24	178	35	36	273
	%	8.8	65.2	12.8	13.2	
C	빈도	6	26	140	40	212
	%	2.8	12.3	66.0	18.9	
D	빈도	32	29	40	173	274
	%	11.7	10.6	14.6	63.1	
전체		249	258	221	272	1000

<표 6> 4월-5월의 전이행렬

4월 \ 5월		A	B	C	D	전체
A	빈도	214	18	3	14	249
	%	85.9	7.2	1.2	5.6	
B	빈도	22	179	23	34	258
	%	8.5	69.4	8.9	13.2	
C	빈도	6	41	134	40	221
	%	2.7	18.6	60.6	18.1	
D	빈도	23	26	40	183	272
	%	8.5	9.6	14.7	67.3	
전체		265	264	200	271	1000

<표 7> 5월-6월의 전이행렬

5월 \ 6월		A	B	C	D	전체
A	빈도	226	21	1	17	265
	%	85.3	7.9	0.4	6.4	
B	빈도	14	186	25	39	264
	%	5.3	70.5	9.5	14.8	
C	빈도	3	20	146	31	200
	%	1.5	10.0	73.0	15.5	
D	빈도	20	33	40	178	271
	%	7.4	12.2	14.8	65.7	
전체		263	260	212	265	1000

<표 8> 추정된 6월-7월의 전이확률

6월 \ 7월	A	B	C	D	전체
A	0.82	0.09	0.01	0.08	0.25
B	0.08	0.68	0.10	0.14	0.26
C	0.02	0.16	0.64	0.17	0.22
D	0.08	0.10	0.15	0.66	0.27
전체	-	-	-	-	1

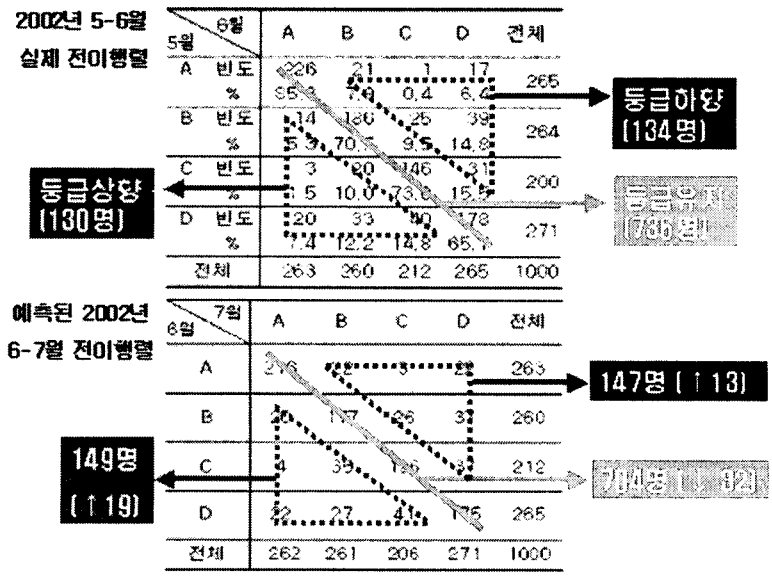
<표 9> 6월-7월의 전이행렬

6월 \ 7월		A	B	C	D	전체
A	빈도	227	18	1	17	263
	%	86.3	6.8	0.4	6.5	
B	빈도	25	170	34	31	260
	%	9.6	65.4	13.1	11.9	
C	빈도	3	25	143	41	212
	%	1.4	11.8	67.5	19.3	
D	빈도	13	24	39	189	265
	%	4.9	9.1	14.7	71.3	
전체		268	237	217	278	1000

<표 10> 예측된 6월-7월의 전이행렬

6월 \ 7월	A	B	C	D	전체
A	216	22	3	22	263
B	20	177	26	37	260
C	4	35	136	37	212
D	22	27	41	175	265
전체	262	261	206	271	1000

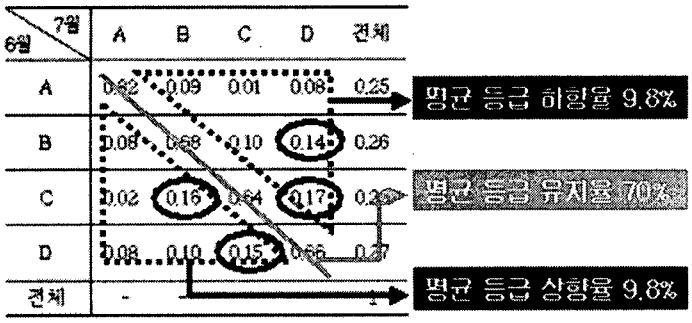
또한 전이행렬은 마코프 모형을 통하여 예측을 가능케 하는 것 이외에도 또 다른 해석을 가능하게 한다. 첫 째로 현시점의 전이행렬과 추정된 다음 시점의 전이행렬을 통하여 고객의 변동에 대한 예측을 해보도록 한다. <그림 1>은 실제 5-6월의 전이행렬과 예측된 6-7월의 전이행렬을 비교한 것인데 5-6월의 전이행렬에서 A-A, B-B, C-C, D-D로의 5월과 6월간 그룹간의 변동이 없이 등급을 유지한 고객의 수가 1000명중 736명으로 나타났으며, A-B, A-C, A-D, B-C, B-D, C-D로의 그룹의 변화를 보이며 등급이 하향한 고객의 수가 134명, 그리고 B-A, C-A, C-B, D-A, D-B, D-C로의 등급이 상향한 고객의 수가 130명으로 나타나 B은행의 고객의 수익금액에 따른 분류그룹의 변동은 유지율이 73.6%, 변동율이 26.4%로 나타나며 변동율 중 상향으로의 변동은 13%, 하향은 13.4%로 유사하게 나타났다. 예측된 6-7월 전이행렬로부터 다음 시점에서의 그룹별 고객의 변동을 살펴보면 유지율은 704명으로 5-6월에 비하여 3.2% 감소 할 것으로 예측되고, 하향 변동율은 147명으로 1.3% 증가할 것으로 보이며, 상향 변동율은 149명으로 1.9% 증가할 것으로 예측되어진다. 종합



<그림 1> 5월-6월 전이행렬과 예측된 6월-7월 전이행렬

적으로 유지율은 현 시점보다 3.2% 감소할 것으로 보이며 변동율은 3.2% 증가할 것으로 예측되어진다.

두 번째로 각 그룹별 유지 및 변동의 전이확률을 추정하게 되면 각 그룹간의 유지정도와 전이정도를 비교할 수 있으며 이는 <그림 2>와 같다.



<그림 2> 추정된 6월-7월의 전이확률

<그림 2>의 전이확률에서 대각은 그룹을 유지하는 정도를 나타내는데 평균적인 등급유지율은 약 70%로 나타나며 A 그룹을 제외한 나머지 그룹의 유지율은 평균유지율 보다 낮았으며 특히 C 그룹의 유지율이 낮아서 다른 그룹으로의 변동이 많아져 고객의 수가 작아질 것으로 보이며, A그룹에서의 변동은 높은 유지율 때문에 없을 것으로 예측된다. 상 삼각의 모양을 나타내고 있는 하위 그룹으로의 변동은 평균 9.8%의 하향율을 보이며 등급의 변화가 예상되며 특히 C=>D의 변동이 가장 크며 다음으로 B=>D로의 변동이 높은 것으로 나타났다. 하 삼각의 모양을 나타내고 있는 상위 그룹으로의 변동은 평균 9.8%의 상향율을 보이며 등급의 변화가 예상되며 특히 C=>B그룹으로의 변동이 가장 크며 다음으로 D=>C로의 변동이 높은 것으로 나타났다.

<그림 1>과 <그림 2>을 통하여 종합적인 해석을 하면 2002년 6월을 기준으로 7월의 그룹별 고객의 비율을 예측하면 6월에 속하였던 그룹에서 계속 유지하는 고객은 3.2% 감소할 것으로 나타나 다른 그룹으로의 변동이 있을 것이라 예측된다. 특히 C 그룹은 가장 낮은 유지율을 보이며 많은 변동이 있을 것이라 예측된다. 다른 그룹으로의 변동에 있어서 하위 그룹으로의 변동은 전체 9.8%의 증가가 기대되며 이중 B=>D, C=>D로의 변동이 높게 나타나 7월의 D그룹 고객의 수가 증가 할 것으로 예상된다. 상위 그룹으로의 변동 역시 전체 9.8%의 증가가 기대되며 특히 C=>B, D=>C로의 변동이 가장 크게 나타났다. 변동에 대하여 그룹별로 해석을 하면 B그룹은 D그룹으로의 등급이 하향되는 변동이 크게 나고 있으며 이러한 고객의 성향을 파악하여 고객 등급 유지정책을 통한 수익성 제고의 마케팅이 필요할 것으로 여겨진다. D그룹은 C그룹으로의 등급 상향의 변동이 크게 나타나고 있으며 이러한 고객의 분석을 통하여 등급 상향정책(up-selling)의 적용이 가능할 것이다. C그룹은 등급의 상향 및 하향의 변동이 동시에 나타나는 그룹으로서 상향 및 유지정책 모두가 필요한 그룹이라 할 수 있다.

4. 결 론

본 연구에서는 금융권 기업의 CRM에서 핵심이 되는 수익금액을 기준으로 분류된 고객의 등급을 범주화하여 여러 번 측정된 동일 범주에 대하여 시간의 흐름에 따라 고객이 보이는 범주 사이에서의 변동을 전이행렬로 정리하여 파악할 수 있었고, 이러한 전이행렬의 정보를 토대로 마코프 모형에 적용하여 다음 시점에서의 고객 등급 범주별 변동 및 유지정도도 추정과 예측이 가능함을 볼 수 있었다.

참고문헌

[1] 강현철, 한상태, 최종후, 김차용, 김은석, 김미경 (1999), *SAS Enterprise Miner를 이용한 데이터마이닝*, 서울, 자유아카데미.

[2] 한동철 (2001), *고객관계관리 CRM*, 서울, 우용출판사.

[3] Yvonne M. M. Bishop, and Stephen E. Fienberg, and Paul W. Holland (1975), *Discrete Multivariate Analysis*, pp. 257~267.