CBR에서 최적의 결합사례개수결정을 위한 다양한 알고리즘들의 비교연구

이 훈영
경희대학교 경영학부 조교수
박 기남
경희대학교 경영학부 박사과정

1. 서 론

사례기반예측(Case-Based Forecasting)은 예전부터 많은 경영상의 문제에 중요한 수단으로 제공되어왔고, 유사사례에 근거한 추론의 잠재력은 많은 연구자들에 의해 자주 인식되고 토론되어 왔다(Burke 1991). 그러나 사례기반예측시스템(Case-Based Forecasting System)의 예측력은 (1)사례의 유사도 측정방법 (2)결합할 유사사례 개수의 결정방법 (3)결합 시 유사사례에 가중치를 부여하는 방법들이 얼마나 효과적이냐에 달려있다. 본 논문에서는 이 중에서 특히 결합할 유사사례의 개수를 결정하는 방법을 가장 중요한 문제로 인식하고 결합할 최적의 사례개수를 탐색하기 위한 여러 가지 수리 방법들을 제안하고 개발하였다. 또한 시뮬레이션 자료를 이용하여 제시된 방법들의 유효성을 서로 비교검증 하였다.

2. 사례기반추론 및 예측시스템의 개념

사례기반예측(Case-Based Forecasting)이란 기존의 통계학적인 방법과 같이 문제영역 내 변수들간의 관계에서 유도된 일반식을 통하여 예측치를 구하는 것이 아니라, 현 문제와 유사한 과거의 구체적인 에피소드로부터 문제해결을 위한 지식을 추론하고 예측해 나가는 새로운 방식을 말한다. 즉, 사례기반예측은 유사한 과거 사례를 이용하여 미래를 예측할 때 이용할 수 있는 유용한 예측 방법 중의 하나이다(Kim and Kang 1996).

2.1 사례의 유사도 측정방법

'어떤사례로 예측하는 것이 가장 효과적인가'에 관한 가장 쉽고도 훌륭한 답은 가장 유사한 사례로 예측하는 것이 가장 정확한 예측이 될 것이라는 것이다. 이때 두 사례간의 유사도 측정방법이 문제가 되는데 유사도를 측정하는 가장 명확한 방법 중 하나는 이들 간의 거리를 이용하는 것이다. 가중된 유클리드 거리는 거리함수의 가장 일반적인 형태 중 하나로서 다음과 같이 표현된다.

$$d_{bt}=\left\{\sqrt{\sum_{j=1}^{m}\omega_j(\chi_{bj}-\chi_{tj})^2}\right\}$$

위 식에서 m은 거리 측정에 쓰이는 독립변수들의 개수이고 ω는 유사도를 측정함에 있어서 각 독립변수의 상대적인 중요성을 나타낸다. χ_{bj}와 χ_{tj}는 기반사례(Base Case)와 타겟사례(Target Case)의 j번째 변수의 값을 나타내고, d_{bt}는 기반사례b와 타겟사례t 간의 유클리드 거리를 나타낸다. 또한 다차원 공간에서 각 구성 차원(독립변수)들의 중요성에 따라 서로 다른 가중치를 줌으로써 보다 정확한 거리측정을 할 수 있다. 가중치를 주는 방법에는 전문가의 판단에 의해 임의로 가중치를 주는 방법, 자료에서 주어진 사례들의 독립변수들을 타겟변수[1]에 대하여 회귀 또한 분석을 한 각 계수(Coefficient)의 크기 비율에 따라 가중치로 주는 방법, 타겟변수와의 상관관계의 크기를 고려하여 이에 비례한 가중치를 할당하는 방법 등 다양한 방법이 있다. 본 논문에서는 독립변수와 타겟변수와의 상관관계와 독립변수들 상호간의 상관관계를 모두 고려하는 다음 수식과 같은 방법으로 가중치를 추정하여사용하였다.

$$\omega_j=\left\{\rho_{tj}/\sum_{k=0}^{m}\rho_{jl}\right\}\div\left\{\sum_{k=0}^{m}\left(\rho_{tk}/\sum_{l=0}^{m}\rho_{kl}\right)\right\}$$

1. 타켓변수는 사례기반 예측시스템에서 예측하고자 하는 변수를 뜻한다. 통계학에서 일반적으로 말하는 종속변수가 여기에 해당한다.

위의 식에서 m은 변수의 개수이고 ρ_{tj}는 타겟변수 t와 독립변수j 사이의 상관계수(Correlation Coeffient)를 나타내고 ρ_{jl}은 독립변수j와 l간의 상관계수를 나타낸다. 위의 수식에서 독립변수는 타겟변수와 상관관계가 높고 다른 독립변수들과의 상관관계가 낮을수록 가중치가 커진다. 독립변수와 타겟변수가 높은 상관관계가 있다 하더라도 이것이 다른 독립변수와의 상관관계가 높다면 그 효과가 서로 상쇄되어 큰 가중치를 갖지 못하게 된다.

새로운 타겟사례 t와 기반사례 b간의 유사도는 일반적으로 두 사례간의 가중된 유클리드 거리함수의 역함수로 표현된다. 예를들어 지수감소함수를 사용할 경우 유사도 $S_{tb}=exp\{-d_{tb}\}$로 표시 할 수 있다.

2.2 최적의 결합할 유사사례의 개수에 대한 결정방법

유사사례는 각기 다른 하나의 예측치를 가지고 있다. 따라서 정확한 유사도 측정이 사례기반예측에서 중요한 영향을 미친다. 일반적으로 한 개의 유사사례로 예측하기 보다는 몇개의 유사사례를 결합하여 예측하는 것이 보다 정확한 예측을 가능케 한다. 왜냐하면 사례의 수(n)가 증가할 수록 예측치의 분산(σ^2)은 감소하고(σ^2/n) 반면에 결합사례의 수가 증가함에 따라 예측치의 값은 덜 유사한 사례들의 예측치들을 많이 포함하게 되어, 전체 사례의 평균값으로 수렴함에 따라 보다 큰 편의(Bias)를 가질 수 있기 때문이다. 이와는 반대로 결합사례의 수가 감소하면 분산은 증가되고 편의는 감소하는 경향이 있다. 따라서 결합사례의 수는 분산과 편의사이의 균형을 효과적으로 맞추는 방식이 되어야 한다.

2.2.1 임의의 개수로 결합하는 방법

임의의 개수로 결합하는 방법(FNCM)은 유사한 사례들을 임의의 특정한 개수로 결합하는 방법을 말한다(Kim and Kang 1996). 결합사례개수의 결정은 데이터베이스에서 타겟사례와 기반사례 사이의 유사도를 구하여 가장 유사한 사례들의 일정 백분위수를 결합하는 방법과 구체적인 유사사례의 개수를 임의로 지정하여 사용하는 것이 있다.

2.2.2 최적의 범위 결정법

최적의 범위 결정법(OSM)은 일정한 유사도의 범위(Span)를 설정해두고, 정해진 범위 내에 있는 유사사례만을 결합하는 방법이다(Lee 1996). 따라서 이러한 방법의 초점은 최적의 범위를 어떻게 구하느냐에 달려있다. 일정범위(Boundary Distance)를 파라 미터(Parameter) λ로 표시하고, λ의 일정범위 내의 사례들을 적절한 유사도를 가진 유용한 유사사례로 간주한다. 반면, λ경계 밖의 사례들은 유용하지 않다고 판단하여 0의 유사도를 할당한다. λ는 임의적으로 결정될 수도 있지만, 시스템의 Estimation Sample Data안에 있는 사례를 활용하여 교차검증(Cross-Validation)을 통하여 최적의 값을 추정할 수도 있다. 교차검증은 한번에 한 사례씩 제거하면서, 남아있는 사례들을 사용하여 그것의 기대값을 측정하는 방법인데 이 방법을 통하여, 다음의 수식과 같이 예측오차의 평균제곱(Mean Squared prediction Error)을 최소화 하는 파라 미터 λ의 값을 탐색할 수있다.

$$MinimizeMSE(\lambda)_{in\ cross-validation}=\frac{1}{n}\sum_{b=1}^{n}\left\{TV_b-\sum_{k=1_{k\neq b}}^{n}\left(\frac{S_{bk}}{\sum_{i=1_{k\neq b}}^{n}S_{bk}}\right)\cdot TV_k\right\}^2$$

본 논문에서는 위 수식의 교차검증방법을 통하여 기반사례베이스를 구성하는 측정샘플자료의 예측오차의 평균제곱을 최소화 시키는 최적의 λ값을 찾고, 이를 이용하여 유사도의 범위를 결정하는 방법을 선택하였다. 본 연구에서는 유사도 및 λ값을 측정하는 함수로 LT(Linear Transformation), TK(Tricube Kernel), EK(Epanechnikov Kernel), MVK(Minimum Variance Kernel) 사용하였다.

$$S_{ab}=\begin{cases}\frac{\lambda-d_{ab}}{\lambda}, for & d_{ab}\le\lambda \\ 0, otherwise\end{cases} --Linear\ Transformation-(LT)$$

$$S_{ab}=\begin{cases}\left(1-\left(\frac{d_{ab}}{\lambda}\right)^3\right)^3, for & d_{ab}\le\lambda \\ 0, otherwise\end{cases} --Nonlinear\ Transformation\ (TRICUBE\ \ KERNEL)-(TK)$$

$$S_{ab}=\begin{cases}\frac{3}{4}\left(1-\left(\frac{d_{ab}}{\lambda}\right)^2\right), for & d_{ab}\le\lambda \\ 0, otherwise\end{cases} --Nonlinear\ Transformation\ (EPANECHNIKOV\ \ KERNEL)-(EK)$$

$$S_{ab}=\begin{cases}\frac{3}{8}\left(3-5\left(\frac{d_{ab}}{\lambda}\right)^2\right), for & d_{ab}\le\lambda \\ 0, otherwise\end{cases} --Nonlinear\ Transformation\ (MINIMUM\ \ VARIANCE\ \ KERNEL)-(MVK)$$

최적의 범위 결정법(OSM)은 결합할 사례의 개수를 결정하는 이론적인 근거가 있다는 점에서 임의로 유사사례의 개수를 선택하는 것보다 진보된 방법으로 볼 수 있다. 또한 이 방법은 이해가 쉽고 적용이 간편한 장점이 있으나 교차검증을 통하여 파라 미터 λ를 구할 경우 시간이 걸리는 단점이 있다.

2.2.3 유사도 분포에 따른 최적화 수리 모형에 의한 방법

사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)은 사례간의 유사도 정보를 바탕으로 결합할 최적의 사례개수를 결정하는 수리 모형이다. 수리모형은 타겟사례와 높은 유사도를 가지면서 함께 선정될 다른 기반사례들과의 유사도가 낮을수록 효과적인 기반사례라고 판단하여, 선정될 기반사례와 타켓사례간의 유사도합을 선정될 사례들간의 유사도합으로 나눈 값을 최대화하는 사례를 선정한다. 또한 유사도에 따라 사례를 결합할 때 보다 유사한 사례가 덜 유사한 사례보다 항상 우선권을 가져야 한다는 제약조건이 필요하다. 이러한 제약조건으로 인하여 결합사례 개수의 결정은 결합사례의 유이 모형을 수식으로 나타내면 다음과 같다.

$$Max \quad SF = \frac{\sum_{b=1}^{n} S_{tb} Z_b}{(\sum_{b=1}^{n} \sum_{q=1}^{n} S_{bq} Z_b Z_q)^r}$$

$$s.t. \quad (S_{tb} - S_{tq}) \times (Z_b - Z_q) \geq 0 \quad \forall \quad b \quad and \quad q$$

$$Z_b = 0 \quad or \quad 1$$

$$0 \leq p \leq 0.5$$

위의 모형에서 n은 선택된 사례의 개수이고, S_{tb}는 타겟사례 t와 기반사례b 사이의 유사도이며, S_{bq}은 기반사례b와 기반사례q 사이의 유사도이다. Z_b는 기반사례b의 선택 여부를 나타내는 변수로서 0혹은 1의 이진 값을 갖는다. 변수 Z_b가 1이면 기반사례b가 선택되고 그렇지 않으면 선택되지 않는다. 여기에서 제약식 $(S_{tb}-S_{bq})\times(Z_b-Z_q)\geq 0$은 사례 선택 시 항상 유사도가 높은 것이 유사도가 덜 높은 것에 우선하도록 하는 제약 조건이다. 위의 모형에서 함수값 SF를 최대화 하는 수준에서 결합할 사례들이 결정되는데, 이때의 사례의 개수가 최적의 결합할 사례개수가 되는 것이다.

2.3 유사사례의 예측값에 가중치를 부가하는 방법

남은 문제는 최종 선정된 유사사례들의 값을 이용하여 보다 정확한 예측치를 얻는 것이다. 유사 사례를 기반으로 한 예측에 있어서는 타겟사례와 유사한 사례일수록 그 예측값이 현재 사례의 타겟값을 예측하는데 보다 정확하고 효과적이라고 생각되기 때문에 기반사례의 유사도 정도에 따라 가중치를 주어 결합하는 것이 어떠한 형태의 가중법보다 합리적이다. 유사도에 비례하는 가중치를 주어 종합적인 예측치를 만드는 방법은 다음과 같다.

$$E(TV_t|\{S_{tt}\}_{t=1,n}) = \sum_{b=1}^{n} P(TV_t = TV_b|\{S_{tt}\}_{t=1,n}) TV_b = \sum_{b=1}^{n} \left(\frac{S_{tb}}{\sum_{i=1}^{n} S_{ti}} \right) \cdot TV_b$$

위의 식에서 n 은 전체 예측을 구성하기 위해서 선택된 사례들의 수이고 S_{tb}는 새로운 사례인 t와 유사사례 b 사이의 유사도이다. 그리고 TV_b는 기반사례 b의 타겟값(Target Value)을 의미한다. 위의 수식에서 유사도 비율(즉, 모든 사례의 유사도의 총합에 대한 새로운 타겟사례와 각 기반사례의 유사도의 비율)은 각 사례를 결합할 때 가중치로 사용된다. 결국 현재 사례의 타겟값에 대한 예측치는 현재 사례와의 유사도 비율로 가중된 기반사례 타겟값들을 선형결합한 값이 된다.

3. 시뮬레이션에 의한 각 방법의 유효성 검증

시뮬레이션 자료는 SAS프로그램을 이용하여 만들어졌다. 독립변수들의 값은 0과 1사이의 일양분포로부터 무작위로 생성되었고, 독립변수의 갯수는 5개로 했으며, 그 구성은 다음과 같다.

(표3-1)생성된 시뮬레이션 자료에서 고려된 요인

관계	수 식	포함된 오차량
선형	$Y_i = \sum_{j=1}^{m} \omega_j \cdot x_j + \alpha \cdot \varepsilon$	10%, 30%, 50%
다중형	$Y_i = \sum_{j=1}^{m} \omega_j \cdot x_j \cdot x_{j+1} + \alpha \cdot \varepsilon$	10%, 30%, 50%
자승형	$Y_i = \sum_{j=1}^{m} \omega_j \cdot x_j^2 + \alpha \cdot \varepsilon$	10%, 30%, 50%

먼저 추정용 샘플데이터 셋을 이용하여 임의의 개수로 결합하는 방법(FNCM), 최적의 범위 결정법(OSM), 사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)을 이용하여 최적의

예측 모형을 구하였다. 그리고 각각의 예측모형을 검증용 샘플데이터 셋에 적용하여 얻어진 예측치를 그 실제 값과 비교하여 예측오차를 측정하고, 이것을 제곱함으로써 예측오차의 제곱(Squared Prediction Error)을 얻었다. 또한 각 샘플데이터 셋 마다 검증용 샘플을 통하여 계산한 예측오차의 제곱을 합하고, 이것을 검증용 샘플의 사례개수로 나누어 각 추정용 샘플데이터 셋의 예측오차제곱의 평균(Mean Squared prediction Error)을 구했다. 그리고 각각의 모형을 통하여 얻어진 예측오차제곱의 평균(이하MSE로 표기)을 상호 비교함으로써 각 방법들의 예측정확성을 분석하였다.

3.2 예측 오차에 따른 각 방법의 유효성 분석

시뮬레이션 결과로 얻은 각 방법의 MSE를 요약하면 표3-2와 같다.

(표3-2)각 방법의 MSE비교

관계	오차	FNCM	OSM				MPMSD
			LT	TK	EK	MVK	
선형	10%	0.276	0.320	0.303	0.325	0.325	0.238
	30%	0.373	0.404	0.385	0.411	0.385	0.345
	50%	0.569	0.582	0.574	0.594	0.592	0.545
다중형	10%	0.709	0.772	0.774	0.787	0.787	0.680
	30%	0.795	0.831	0.830	0.841	0.836	0.771
	50%	1.008	1.032	1.009	1.036	1.038	0.998
자승형	10%	1.656	1.619	1.606	1.629	1.612	1.492
	30%	1.754	1.700	1.682	1.712	1.703	1.574
	50%	1.945	1.891	1.891	1.890	1.899	1.752

표 3-2에서 보듯이 독립변수와 종속변수 간의 관계가 선형일 경우, 방법별로는 사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)에 의한 방법이 가장 예측정확성이 높은 것으로 나타났고, 두 번째가 임의의 개수로 결합하는 방법(FNCM), 세 번째가 최적의 범위 결정법(OSM)의 순서로 나타났다. 또한 포함된 오차량이 증가할수록 사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)과 타 방법들과의 MSE차이가 줄어드는 것을 알 수 있다. 이것은 사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)이 타 방법에 비해서 예측정확성이 높지만 선형관계의 경우, 포함된 오차량이 증가할수록 이 방법의 우수성은 복잡성의 증가로 인해 상쇄된다는 것을 나타낸다. 시뮬레이션 결과, 사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)은 어떠한 상황에서도 가장 예측정확도가 높은 것으로 나타났다. 그러나 이 방법에 의한 오차가 타 방법에 의한 오차보다 통계적으로 유의할 만큼 작은지 쌍체비교(Paired T-test)를 통하여 검정하여, 그 결과를 표3-3에 요약하였다.

(표3-3)사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)과 타 방법들 간의 Paired T검증 결과

관계	오차량	OSM				MVK
		FNCM	LT	TK	EK	
선형	10%	**0.0053**	**0.0002**	**0.0008**	**0.0001**	**0.0001**
	30%	**0.0001**	**0.0079**	**0.0588**	**0.0034**	**0.0588**
	50%	**0.0002**	0.2704	0.3792	0.1547	0.1759
다중형	10%	**0.0088**	0.1306	0.1147	**0.0824**	**0.0787**
	30%	**0.021**	0.2834	0.2939	0.2188	0.2415
	50%	0.3937	0.6237	0.8675	0.5797	0.5633
자승형	10%	**0.0001**	0.1576	0.2125	0.1313	0.2134
	30%	**0.0001**	0.2358	0.3013	0.1966	0.2361
	50%	**0.0001**	0.1898	0.1956	0.1894	0.1821

3.3 각 방법별 결합한 사례의 개수에 대한 분석

다음은 제안된 각 방법에 따라 예측치를 구할 때 사용된 사례의 개수를 분석하여 표3-4에 요약하였다. 표 3-4에서 FNCM은 1개의 유사사례를 결합했을 때부터 50개의 유사사례를 결합했을 때까지 중 가장 작은 MSE를 가졌을 때의 사례의 개수를 나타낸 것이며, OSM(LT,TK,EK,MVK)과 MPMSD의 경우는 검증용 샘플의 사례를 예측할 때 이용한 사례 결합개수를 나타낸다.

(표3-4) 각 방법의 결합사례의 개수

관계	오차량	FNCM	OSM				MPMSD
			LT	TK	EK	MVK	
선형	10%	5	19.70	22.77	18.77	18.73	7.410
	30%	6	21.67	25.70	21.43	25.70	8.100
	50%	8	23.73	28.07	21.73	22.17	11.00
다중형	10%	4	19.13	22.07	17.50	17.00	7.790
	30%	4	18.43	22.67	17.73	16.60	9.150
	50%	7	19.03	23.97	18.63	19.40	12.54
자승형	10%	3	18.40	21.67	17.26	13.67	9.690
	30%	5	18.06	21.56	18.00	13.53	10.10
	50%	5	22.73	25.06	21.36	17.90	10.20

위의 결과를 종합해보면, 사례기반예측시스템은 문제의 복잡성이 클수록 결합하는 유사사례의 개수가 많아짐을 알 수 있다. 이를 실제 경영자의 의사결정과정에 비추어 볼 때, 경영자는 복잡성이 큰 문제일수록 한 두개 사례만이 아니라 보다 많은 사례를 고려해보고 의

사결정 하고자 하는 경향과 일맥상통한다고 보아, 사례기반시스템은 합리적인 예측방법을 사용하고 있다고 생각할 수 있다.

4.결 론

본 연구에서는 결합사례 개수의 결정에 관한 여러 가지 방법을 제시하고 각 방법의 유효성을 시뮬레이션 자료를 이용하여 비교분석하였다. 시뮬레이션 결과 사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)의 예측정확성이 가장 우수한 것으로 나타났다. 본 연구의 한계점 및 향후연구과제는 다음과 같다. 우선 본 연구에서 이용한 자료가 실제자료를 이용하여 검증한 것이 아니라 시뮬레이션을 위해 가공된 자료를 사용하였다는 점이 일차적인 한계점으로 지적될 수 있다. 또한 본 논문에서 사용된 유사도 측정 방법이외에도 다양한 유사도 측정방법이 있는데 이러한 방법들을 모두 비교하고 분석하지 못한 아쉬움이 있다.

Reference

Burke, R. R. (1991), "Reasoning with Empirical Marketing Knowledge, *International Journal of Research in Marketing* ", 8 (1), 75-90.

Clemen, R. T. (1989), "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting*, 5, 559-583.

Choffray, J. M. and G. L. Lilien (1986), "A Decision-Support System for Evaluating Sales Prospects and Launch Strategies for New Products," *Industrial Marketing Management*, 15, 75-85.

Han, I., C. Park and C. Kim (1996), "Bankrupcy Predictions for Korea Medium-Sized Firms using Neural Nerwork and Case Based Reasoning," *Conference Proceedings, The Korean OR/MS Society,* 203-206.

Kim, S. and D. Kang (1996), "Composite Neighbors for Case Based Prediction: Structural Effects of Stock Price Forecasting," *Conference Proceedings, The Korean OR/MS Society*, 207-210.

Lee, H.(1994), "A Case-based Forecasting System," *Jounal of the Korean Operations Research and Management Science Society*, 19(2), 199-215.

Lee, H.(1996), "Combining Judgments for Better Decisions: A Study for Investigating Effective Combining Schemes," *Jounal of the Korean Operations Research and Management Science Society* , 21(3), 159-174.

Sjoberg, L. (1980), "Similarity and Correlation," *Similarity and Choice*, Hans Huber publishers Bern, eds., Lantermann and Feger, 70-87.

Winkler, R. L. (1989), "Combining Forecasts: A Philosophical Basis and Some Current Issues," *International Journal of Forecast.*