사례기반예측시스템의 정확한 예측을 위한 최적 결합 사례개수결정방법에 관한 연구*

이훈영
경희대학교 경영학부 조교수
(hylee@nms.kyunghee.ac.kr)
박기남
경희대학교 경영학부 박사과정
(knpark@nms.kyunghee.ac.kr)

사례기반예측(Case-Based Forecasting)은 유사한 사례를 이용하여 미래를 예측하는 유용한 방법이다. 사례기반예측시스템의 예측력은 (1) 사례간의 정확한 유사도 측정, (2) 결합할 유사사례 개수, (3) 결합할 유사사례에 대해 가중치를 주는 결합방식에 달려있다. 이러한 요인들 중 예측력에 가장 큰 영향을 미치는 것은 결합할 유사사례 개수이다[1]. 본 논문은 먼저 사례기반예측시스템의 예측력에 영향을 주는 세가지 요인에 관하여 논하고, 그 중에 가장 중요한 결합할 사례개수 결정에 대하여 보다 깊이 있게 다루었다. 즉, 기존에 발표된 결합할 사례개수 결정에서 사용될 수 있는 여러 가지 방법 (1) 임의의 개수로 결합하는 방법(Fixed Number Combining Methods), (2) 최적의 범위를 탐색하는 방법(Optimal Spanning Methods), (3) 유사도 분포에 따른 최적화 수리모형(Mathematical Programming Model using Similarity Distribution)을 비교하고, 그들의 유효성을 시뮬레이션 자료를 이용하여 비교 분석하였다. 시뮬레이션 결과 사례의 유사도 분포에 따른 최적화 수리모형이 대부분의 경우에서 가장 우수한 것으로 나타났다.

I. 서 론

현대의 경영자들은 성공적인 기업경영을 위하여 많은 문제를 정확하게 예측하도록 요구되고 있다. 그 동안 통계학과 경영과학 등 여러 분야에서 이러한 경영자들의 문제를 도와주기 위해 많은 예측 기법들이 연구되고 개발되어 적용되어 왔다. 그러나 개발된 대부분의 예측방법은 문제영역이 잘 구조화되어있고, 필요한 자료를 꾸준히 이용할 수 있는 경우에는 훌륭한 성과를 낼 수 있지만 자료의 수가 적은 경우, 변수들 간의 관계가 복잡한 경우, 그리고 포함된 오차량이 큰 경우에는 큰 효과를 보지 못하고 있다. 그러나 불행히도 경영상의 많은 예측하여야 할 문제들이 이러한 범주에 포함된다. 그래서 경영자들은 이용 가능한 몇 개의 과거 사례를 기반으로 자신들의 직관적인 판단력을 사용하여 유추를 통한 예측을 시도하곤 했다. 이처럼 현재문제를 해결하기 위하여 과거의 유사한 경험 사례를 분석하고 직관적인 판단 및 유추를 통해 그 결과를 예측해가는 과정을 사례기반에 의한 예측(Case-Based Forecasting)이라고 한다(Lee 1994).

사례기반예측(Case-Based Forecasting)은 예전부터 많은 경영상의 문제에 중요한 수단으로 제

논문 접수일 : 98. 2 게재확정일 : 98. 7
* 본 연구는 1996학년도 경희대학교 교내 연구비로 지원 받아 수행되었음.
1) 연구자의 견해에 따라 각기 중요성에 관한 평가가 다를 수 있으나 결합할 사례개수의 결정이 가장 중요하다는 의견에 대해서는 "*Generalized Additive Models*" Hastie.T.& Tibshirani.R.(1990), pp18을 참고하고, 이 연구를 수행한 논자들의 견해도 같은 의견이며, 기존의 연구에서 해결하지 못한 부분이라는 점을 강조하는 의미에서 가장 중요하다는 표현을 사용함.

공되어왔고, 유사사례에 근거한 추론의 잠재력은 많은 연구자들에 의해 자주 인식되고 토론되어 왔다(Burke 1991; Choffray and Lilien 1986; Easingwood 1989; Mahajan and Wind 1988; Thomas 1987; Wind, Mahajan and Cardozo 1981). 또한 Burke(1991)는 광고 캠페인에 대한 소비자 반응을 예측하는데 유추추론시스템(Case-Based Reasoning System)을 이용한 바 있다.

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.

서론에 이은 본 논문의 구성은 다음과 같다. 제2장에서는 사례기반추론시스템의 기본개념과, 사례의 유사도 측정방법, 결합할 유사사례의 개수결정에 관한 방법, 유사사례에 가중치 부가방법등과 같은 사례기반예측의 근본적인 문제들에 관하여 논하고, 이러한 문제를 해결하기 위한 여러 가지 방법들을 제안한다. 제3장에서는 시뮬레이션 자료를 이용하여 2장에서 제안한 다양한 방법들의 유효성을 비교분석 한다. 끝으로 제4장에서 본 논문의 결과를 요약하고 본 연구의 한계점을 지적하며 향후 연구의 방향을 제시한다.

II. 사례기반추론 및 예측시스템의 개념

사례기반예측(Case-Based Forecasting)이란 기존의 통계학적인 방법과 같이 문제영역 내 변수들 간의 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; Han, Park and Kim 1996). 사례기반 예측시스템의 예측력에 영향을 주는 요인으로는 사례간의 정확한 유사도 측정, 최적의 결합할 유사사례의 개수 결정, 결합할 유사사례에 적정한 가중치 부가 등을 들 수 있다. 이들 중 예측력에 가장 큰 영향을 미치는 것은 최적의 결합할 유사사례의 개수를 결정하는 것이다. 따라서 본 논문은 결합 할 유사사례 개수의 결정방법을 개발하는데 중점을 두었다.

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.

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

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[2]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 추정하여 사용하였다.

$$\omega_j = \left\{ \rho_{ij} / \sum_{k=1}^{m} \rho_{jl} \right\} \div \left\{ \sum_{k=1}^{m} \left(\rho_{tk} / \sum_{l=1}^{m} \rho_{kl} \right) \right\}$$

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. 그러나 일반적으로 한 개의 유사사례로 예측하기 보다는 몇 개의 유사사례를 결합하여 예측하는 것이 보다 정확한 예측을 가능케 한다. 왜냐하면 유사한 몇몇 과거사례의 값을 결합하여 예측함으로써 새로운 사례의 타겟값을 보다 작은 분산의 값으로 예측할 수 있기 때문이다. 이러한 사례를 결합하여 예측하는 과정은 여러 전문가들의 예측치나 다양한 방법들의 결과치를 종합하여 예측하는 것과 유사하다. 또한 여러 개의 예측치를 적절히 종합한 예측값이 개별적인 예측치보다 정확하다(Clemen 1989; Winkler 1989; Lee 1996). 따라서 어느 정도 정확한 방법으로 유사도가 측정되었다는 가정하에, 시스템의 예측력에 가장 큰 영향을 미치는 것은 최적의 결합할 유사사례의 개수에 대한 결정이다. 가장 정확한 예측치를 보장하는 사례개수에 관하여 이렇다 할 정해진 답은 없다. 그러나 일반적으로 결합할 사례의 수(n)가 증가할 수록 예측치의 분산(σ^2)은 감소한다(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 있

2) 타겟변수는 사례기반 예측시스템에서 예측하고자 하는 변수를 뜻한다. 통계학에서 일반적으로 말하는 종속변수가 여기에 해당한다.

다. 이와 반대로 결합사례의 수가 감소하면 분산은 증가되고 편의는 감소하는 경향이 있다. 따라서 결합사례의 수는 분산과 편의사이의 균형을 효과적으로 맞추는 방식이 되어야 한다. 예측에서 사용할 수 있는 결합할 유사사례개수의 결정방법은 여러 가지가 있을 수 있으나, 본 논문에서는 결합할 유사사례의 개수를 결정하는 대표적인 세가지 방법을 제시하는데 그 방법들은 다음과 같다.

2.2.1 임의의 개수로 결합하는 방법 (Fixed Number Combining Method)

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(예를 들어, 예측을 위하여 5개나 7개의 특정 개수의 사례를 결합함). 이 방법은 정확한 예측을 위하여 얼마나 많은 사례를 결합해야 하느냐에 대한 아무런 이론적인 기반이 없으나 쉽게 이용할 수 있다는 장점이 있다.

2.2.2 최적의 범위 결정법 (Optimal Spanning Method)

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.

결합할 사례의 범위에 대한 결정은 타겟값으로부터 일정거리 이상인 범위 밖의 사례에는 유사도를 0으로 할당하고, 일정범위 이내의 사례만을 결합하여 예측치를 도출하는 방법으로 이루어진다. 구체적으로 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 측정샘플자료(Estimation Sample Data) 안에 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 수 있다.

$$Minimize\ MSE(\lambda)_{in\ cross-validation} = \frac{1}{n}\sum_{k=1}^{n}\left\{TV_k - \sum_{k=1,k\neq}^{n}\left(\frac{S_{kk}}{\sum_{i=1,i\neq}^{n} S_{ik}}\right)\cdot TV_k\right\}^2$$

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)를 사용하였다.

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

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

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

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

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 수리모형 (Mathematical Programming Model Using Similarity Distribution)에 의한 방법

사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)은 사례간의 유사도 정보를 바탕으로 결합할 최적의 사례개수를 결정하는 수리 모형(Mathematical Programming Model)이다(Lee 1996). 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 유사도 값에 달려있게 된다. 이 모형을 수식으로 나타내면 다음과 같다.

$$Max\ \ SF=\frac{\sum_{b=1}^{n}S_{tb}Z_b}{\left(\sum_{b=1}^{n}\sum_{q=1}^{n}S_{bq}Z_bZ_q\right)^p}$$

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

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

$$0\le p\le 0.5$$

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 기

반사례b가 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.

또한 위의 모형에서 p값을 조정함으로써 결합할 사례의 개수를 어느 정도 조절할 수 있는데, 이때 p값이 증가하면 결합할 사례의 수는 줄어들게 되고 p값이 줄어들면 결합할 사례의 수는 늘어나게 된다. 위 모형에서 목적함수의 분자는 선정된 기반사례와 타겟사례와의 유사도의 합을 나타내고 분모는 선정된 기반사례들 간의 유사도의 합을 나타낸다. 여기서 p값이 작으면 분모가 작아져서 덜 유사한 사례들이 포함될 가능성이 높아지므로 보다 많은 사례가 선정될 수 있고, 반대로 p값이 크면 분모가 커져서 타겟사례와의 유사도가 큰 소수의 기반사례들만이 선정될 수 있다. 따라서 최적의 결합사례 개수를 선택하기 위해서는 적절한 p값을 탐색하는 것이 중요하다. 본 논문에서는 p값을 결정하기 위해서 최적의 범위 결정법에서 λ값을 구하듯이 측정샘플자료를 이용하여 교차검증(Cross Validation)을 통해 예측오차제곱의 평균을 최소화하는 p값을 추정하였다. p값의 범위를 0.1에서 0.99까지로 하여 추정한 결과 p값은 0.43에서 0.47사이가 최적인 것으로 나타났다.

2.3 유사사례의 예측 값에 가중치를 부가하는 방법

유사사례는 각기 다른 하나의 예측치를 가지고 있다. 일반적으로 한 개의 유사사례로 예측하기 보다는 몇 개의 유사사례를 결합하여 예측하는 것이 보다 정확한 예측을 가능하게 한다. 따라서 위에서 언급한 방법을 이용하여 결합할 최적의 유사사례의 개수를 결정하였다. 남은 문제는 최종 선정된 유사사례들의 값을 이용하여 보다 정확한 예측치를 얻는 것이다. 여러 유사사례를 결합할 때 각 예측치에 대해 적절한 가중치를 부가하여 결합하는 방법을 사용할 경우 보다 정확한 예측치를 얻을 수 있다. 가중치를 주는 방법으로는 단순한 등가중법(Equal-Weighting Method), 대수적, 기하학적인 측정에 기반을 둔 가중법, 선형 또는 비선형 가중법 등 다양한 기법들이 가능하다. 그러나 유사사례를 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. 따라서 본 논문에서는 타겟사례와의 유사정도에 따른 비율의 크기에 따라 가중치를 주는 방법을 사용하였다(Lee 1996).

유사도에 비례하는 가중치를 주어 종합적인 예측치를 만드는 방법은 다음과 같다.

$$E\left(TV_t \middle| \{S_{ti}\}_{i=1..n}\right) = \sum_{b=1}^{n} p\left(TV_k = TV_b \middle| \{S_{ti}\}_{i=1..n}\right) \cdot TV_b$$

$$= \sum_{b=1}^{n} \left(\frac{S_{tb}}{\sum_{i=1}^{n} S_{ti}} \right) \cdot TV_b$$

위의 식에서 n 은 전체 예측을 구성하기 위해서 선택된 사례들의 수이고 St_b는 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.

III. 시뮬레이션에 의한 각 방법의 유효성 검증

본장에서는 2장에서 제시한 최적의 결합할 유사사례의 개수에 대한 결정방법들의 유효성을 시뮬레이션을 통해 검증하였다. 즉, 임의의 개수로 결합하는 방법(FNCM), 최적의 범위 결정법(OSM), 그리고 사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)을 사용했을 때 시스템의 예측 정확성을 계산하여 각 방법의 유효성을 비교 검증하였다.

3.1 자료 생성 및 예측오차 추정

시뮬레이션 자료는 SAS프로그램을 이용하여 만들어졌다. 독립변수들의 값은 0과 1사이의 일양분포로부터 무작위로 생성되었고, 독립변수의 갯수는 5개로 했으며, 종속변수는 독립변수들의 선형 혹은 비선형의 3종류 함수식의 값에 일정한 퍼센트의 오차 값을 포함시켜 만들었다. 포함된 오차량은 그 양에 따라 3가지로 하였다. 따라서 시뮬레이션 자료는 (3가지 독립변수와 종속변수 간의 관계) × (3가지 다른 양으로 포함된 오차)로 9가지 종류의 자료가 생성되었다(〈표 3-1〉 참조). 또한 각기 다른 종류의 조합별(3×3)로 각각 30개씩의 샘플데이터 셋(Sample Data Set)이 만들어져서(3×3×30) 모두 270개의 샘플데이터 셋이 생성되었다. 하나의 샘플데이터 셋은 각기 100개씩의 데이터포인트(Data Point)로 구성된 추정용 샘플데이터 셋(Estimation Sample Data Set)과 검증용 샘플데이터 셋(Holdout Sample Data Set)로 나누어져, 총 200개의 자료포인트로 구성되었다.

본 논문에서는 시뮬레이션된 각 상황에서, 먼저 추정용 샘플데이터 셋(Estimation Sample Data Set)을 이용하여 각 방법의 예측모형을 도출하고, 검증용 샘플데이터 셋(Holdout Sample Data Set)으로 그 타당성을 검증하였다. 따라서 먼저 추정용 샘플데이터 셋을 이용하여 임의의 개수로 결합하는 방법(FNCM), 최적의 범위 결정법(OSM), 사례의 유사도 분포에 따른 최적화 수리모형

〈표 3-1〉 생성된 시뮬레이션 자료에서 고려된 요인

관계	수식	포함된 오차량
선형 (Linear)	$Y_i = \sum_j^m W_j * X_j + \alpha * \varepsilon$	10%, 30%, 50%
다중형 (Multiplicative)	$Y_i = \sum_j^m W_j * X_j * X_{j+1} + \alpha * \varepsilon$	10%, 30%, 50%
자승형 (Square)	$Y_i = \sum_j^m W_j * X_j^2 + \alpha * \varepsilon$	10%, 30%, 50%

(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 〈표 3-2〉와 같다. 여기서 임의의 개수로 결합하는 방법(FNCM)의 MSE는 가장 유사한 사례 하나를 결합했을 때부터 50개를 결합했을 때까지 예측정확성을 비교해서 사후적으로 MSE가 가장 작은 경우를 정리한 것이다. 따라서 이 방법의 결과를 다른 방법과 비교할 때 유의해야 할 사항은 임의의 개수로 결합하는 방법(FNCM)의 결과가 타 방법의 결과보다 우수한 것으로 나타났다고 해서 모든 임의의 개수에서 우수하다고 볼 수는 없다는 것이다.

〈표 3-2〉에서 보듯이 포함된 오차의 양이 증가할수록, 독립변수와 종속변수와의 관계가 복잡할수록 MSE가 증가하고 있다. 이것은 복잡성이 클수록 예측오차가 커진다는 것을 의미한다. 본 논문에서 제시한 여러 가지 방법 중 유사도 분포에 따른 최적화 수리모형(MPMSD)을 이용한 방법이 관계의 복잡성(Complexity)이나 포함된 오차량의 크기(Error Term)에 관계없이 MSE가 가장 작은 것으로 나타났다.

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

〈표 3-2〉 각 방법의 MSE비교

관계	오차량	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

오차량이 증가할수록 사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)과 타 방법들과의 MSE 차이가 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. 반면에 최적의 범위 결정법(OSM) 중에는 포함된 오차량의 변화에 상관없이 TK(Tricube Kernel)에 의한 결과가 상대적으로 가장 작은 MSE를 보이고 있다.

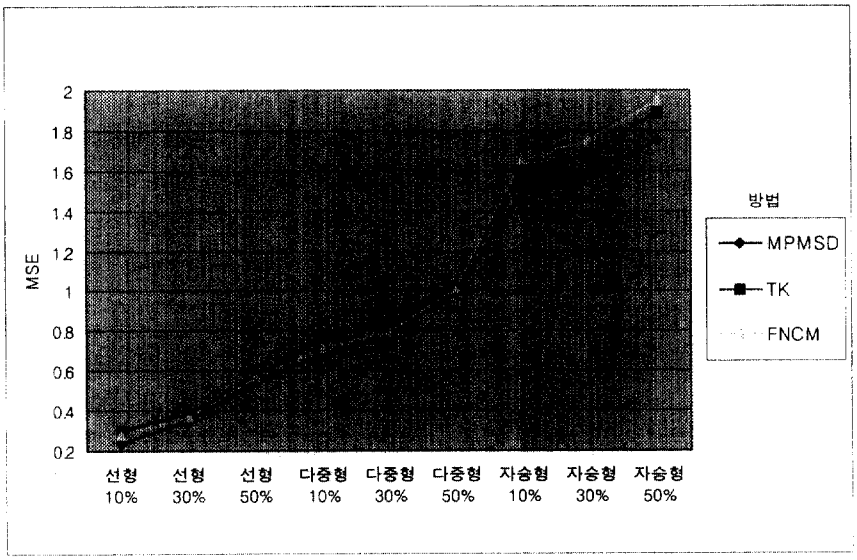
독립변수와 종속변수 간의 관계가 다중형일 경우에서도 선형관계에서와 마찬가지로 오차량의 크기에 관계없이 사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)을 이용한 방법의 결과가 가장 작은MSE를 보여주고 있으며, 임의의 개수로 결합하는 방법(FNCM)이 두 번째, 최적의 범위 결정법(OSM)에 의한 결과가 가장 큰 MSE를 나타내고 있다. 또한 최적의 범위 결정법(OSM) 중에서는 포함된 오차량이 10%인 경우 LT(Linear Transformation)가 가장 우수했으나 TK(Tricube Kernel)와의 MSE차이는 0.001에 불과했고, 30%와 50%의 경우에는 역시 TK가 가장 작은 MSE를 보였으며, 포함된 오차량이 늘어남에 따라 다른 최적의 범위를 탐색하는 방법들(LT, EK, MVK)과의 MSE차이가 점차 커지고 있다. 이것은 다중형의 관계일 경우 포함된 오차량이 증가할수록 TK에 의한 방법이 다른 방법보다 우수하다는 것을 나타낸다.

독립변수와 종속변수 간의 관계가 자승형관계(Square Relationship)인 경우, 방법들간의 MSE를 비교해보면 사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)을 이용한 방법에 의한 결과가 여전히 포함된 오차량에 관계없이 가장 작은 MSE를 보였고 최적의 범위 결정법(OSM)에 의한 결과가 그 다음으로 작은 MSE를, 임의의 개수로 결합하는 방법(FNCM)에 의한 결과가 가장 큰 MSE를 갖는 것으로 나타나 앞의 선형관계나 다중형관계에서의 결과와는 다른 결과가 나왔다. 또한 4가지의 최적의 범위 결정법(OSM) 중에서는 포함된 오차량이 10%와 30%인 경우에는 여전히 TK(Tricube Kernel)가 가장 작은 MSE를 보이고 있고, 50%인 경우에는 EK(Epanechnikov Kernel)가 가장 작은 MSE를 나타냈으나 TK와의 차이는 0.001로서 큰 차이는 없는 것으로 나타났다. 따라서 최적의 범위 결정법(OSM) 4가지 중에서는 관계의 복잡성과 포함된 오차량의 크기를 고려할 때 TK(Tricube Kernel)를 이용한 방법이 가장 우수한 것으로 나타났다.

〈그림 3-1〉은 위의 결과들을 종합하여 꺾은선 그래프로 나타낸 것이다. 〈그림 3-1〉에서 보듯이, 독립변수와 종속변수의 관계가 선형에서 다중형 그리고 자승형으로 점차 그 복잡성(Complexity)이 증가함에 따라 MSE가 커지고, 포함된 오차량이 증가할수록 MSE가 커지고 있다. 또한 자료의 복잡성이나 오차의 양에 관계 없이 전반적으로 사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)을 이용할 때, 예측력이 가장 좋은 것으로 나타났다. 이것은 사례기반예측에서는 결합할 사례를 일정하게 유지시키는 것보다는 가능하면 각 사례별로 기반사례들과의 유사도 정도에 따라 결합사례의 개수를 변화시키는 것이 효과적임을 시사하고 있다.

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 통계

〈그림 3-1〉 각 방법별 MSE비교



적으로 유의할 만큼 작은지 쌍체비교(Paired T-test)를 통하여 검정하여, 그 결과를 〈표 3-3〉에 요약하였다.

사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)을 이용한 방법은 임의의 개수로 결합하는 방법(FNCM)보다 다중형 관계이면서 포함된 오차량이 50%인 경우를 제외한 모든 비교에서 통계적으로 유의한 차이를 보이고 있다. 따라서 사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)을 이용한 방법이 임의의 개수로 결합하는 방법

〈표 3-3〉 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%	0.0219**	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

**는 95%신뢰구간, *는 90%의 신뢰구간에서 유의함을 나타냄

(FNCM)보다 통계적으로 유의하게 더 우수하다고 말할 수 있다. 한편 사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)과 최적의 범위 결정법(OSM) 간의 통계적인 차이를 살펴보면 선형관계이면서 포함된 오차가 10%인 샘플과 30%인 샘플 그리고 다중형 관계이면서 포함된 오차량이 10%인 경우의 EK(Epanechnikov Kernel)와 MVK(Maximum Variance Kernel)에서는 통계적으로 유의한 차이가 있는 것으로 나타났으나 그 밖의 모든 경우에서는 통계적으로 유의한 차이가 없는 것으로 나타났다. 이것은 대부분의 경우 사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)의 예측력이 최적의 범위 결정법(OSM)의 것 보다 우수하긴 하지만 통계적으로 유의할 만큼은 아니다는 것을 의미한다.

3.3 각 방법별 결합한 사례의 개수에 대한 분석

다음은 제안된 각 방법에 따라 예측치를 구할 때 사용된 사례의 개수를 분석하여 〈표 3-4〉에 요약하였다. 〈표 3-4〉에서 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 사례결합개수를 나타낸다.

일반적으로 포함된 오차의 크기가 클수록 결합하는 사례의 개수는 늘어남을 알 수 있다. 또한 임의의 개수로 결합하는 방법(FNCM)의 경우, 각 개별 관계에서 포함된 오차량이 증가할수록 결합할 유사사례의 개수가 많아져야 예측정확성이 높아짐을 알 수 있다. 최적의 범위를 탐색하는 방법(LT, TK, EK, MVK)은 관계가 선형에서 비선형으로 복잡해질수록 오히려 결합하는 사례의 개수가 줄어들고 있다. 사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)의 경우, 선형에서 비선형으로 관계의 복잡성이 증가할수록 또 포함된 오차량이 증가할수록 결합하는 사례의 개수가 많아져서 복잡성의 증가에 따른 결합사례의 개수가 가장 민감하게 반응하고 있음을 알 수 있다.

위의 결과를 종합해보면, 사례기반예측시스템은 문제의 복잡성이 클수록 결합하는 유사사례의 개수

〈표 3-4〉 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

가 많아짐을 알 수 있다. 이를 실제 경영자의 의사결정과정에 비추어 볼 때, 경영자는 복잡성이 큰 문제일수록 한 두개 사례만이 아니라 보다 많은 사례를 고려해보고 의사결정 하고자 하는 경향과 일맥상통한다고 보아, 사례기반시스템은 합리적인 예측방법을 사용하고 있다고 생각할 수 있다.

Ⅳ. 결 론

본 논문은 사례기반예측시스템을 통하여 예측을 할 때 예측력에 영향을 미치는 중요한 세가지 요인 즉, 사례간의 정확한 유사도 측정, 결합할 유사사례 개수, 결합할 유사사례에 대한 가중치를 주는 방법에 대한 연구를 수행하였다. 이러한 요인들 중 예측에서 특히 중요한 요인은 결합할 유사사례의 개수를 결정하는 것이다. 따라서 본 논문에서는 결합사례 개수의 결정에 관한 여러 가지 방법을 제시하고 각 방법의 유효성을 시뮬레이션 자료를 이용하여 비교분석 하였다. 시뮬레이션 결과 사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)의 예측정확성이 가장 우수한 것으로 나타났다.

한편, 결합할 유사사례의 개수는 관계의 복잡성이 커지고, 포함된 오차량이 늘어남에 따라 증가하고 있다. 따라서 임의의 개수로 결합하는 방법(FNCM)에서 처럼 결합할 사례의 개수를 정하거나, 최적의 범위 결정법(OSM)에서처럼 유사도 범위를 고정하여 예측치를 구하는 것보다, 사례의 유사도 분포에 따른 최적화 수리모형(MPMSD)을 이용한 방법에서처럼 가능하면 각 사례를 예측할 때마다 사례간 유사도의 분포에 따라 결합할 유사사례의 개수를 다르게 하는 것이 보다 효과적인 방법으로 나타났다.

본 논문의 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. 따라서 향후 이러한 방법들을 실제문제에 적용시켜 검증하는 연구가 필요하다. 또한 본 논문에서 사용된 유사도 측정 방법이외에도 다양한 유사도 측정방법이 있는데 이러한 방법들을 모두 비교하고 분석하지 못한 아쉬움이 있다. 그리고 본 연구에서는 유사도 측정 시 사용할 변수선정문제에 대하여 언급하지 않았는데, 이 문제는 본 연구에서 집중적으로 다룬 결합할 사례의 수에 대한 결정 문제와 더불어 사례기반시스템의 효과적인 적용에 있어서 매우 중요한 연구분야로서 향후 지속적인 연구를 필요로 한다. 마지막으로 유사사례가 존재하는 경우에는 유사도의 적용이 가능하지만 유사한 사례가 없는 경우는 사례기반예측을 활용할 수 없다는 한계점이 있다. 따라서 향후 이러한 한계점들을 극복한 연구가 이루어질 수 있기를 기대한다.

참 고 문 헌

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.

Easingwood, C. J. (1989), "An Analogical Approach to The Long Term Forecasting of Major New Product Sales," *International Journal of Forecasting*, 5, 69-82.

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.

Mahajan, V. and J. Wind (1988), "New Product Forecasting Methods: Direction for Research and Implementation," *International Journal of Forecasting*, 341-358.

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

Thomas, J. R. (1985), "Estimating Market Growth for New Products: An Analogical Diffusion Method Approach," *Journal of Product Innovation Management*, 2, 45-55.

Wind, Y., A. Mahajan, and R. Cardozo (1981), "New Product Forecasting: Methods and Applications," Lexington, Mass.

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

Methods for Determining the Optimal Number of Cases to Combine in An Effective Case-Based Forecasting System

Hoon Young Lee* · Kinam Park**

Abstract

A case-based forecasting system has been used to predict the outcome of current problem using those of past analogous cases. However, the effectiveness of its forecasting depends on the three factors: (1)accurate measure of similarity, (2)appropriate determination of the number of cases to combine, and (3)appropriate weighting of similar cases when combining them to produce a forecast. Among these three, determining the number of cases to combine is the most critical to the accuracy of forecasting. In this paper, we thus focused on the second, though briefly addressing these all three factors. We suggested several methods to determine the number of cases to combine, such as (1)fixed number combining methods, (2)optimal spanning methods, and (3)mathematical programming model using similarity distribution. A simulation study was conducted to test their effectiveness for accurate forecasting. It proved mathematical programming model using similarity distribution the best among the suggested.

* Assistant Professor, School of Business Administration, Kyunghee University
** Ph.D. Candidate, School of Business Administration, Kyunghee University