지능정보연구 제16권 제4호 2010년 12월(pp.67~84)

퍼지 관계를 활용한 사례기반추론 예측 정확성 향상에 관한 연구

이인호
웅지세무대학 회계정보과
(Fuzzyworld@wat.ac.kr)

신경식
이화여자대학교 경영대학 경영학과
(ksshin@ewha.ac.kr)

미래에 대한 정확한 예측은 경영자, 또는 기업이 수행하는 경영의사결정에 매우 중요한 역할을 한다. 예측만 정확하다면 경영의사결정의 질은 매우 높아질 수 있을 것이다. 하지만 점점 가속화되고 있는 경영 환경의 변화로 말미암아 미래 예측을 정확하게 하는 일은 점점 더 어려워지고 있다. 이에 기업에서는 정확한 예측을 위하여 전문가의 휴리스틱뿐만 아니라 과학적 예측모형을 함께 활용하여 예측의 성과를 높이는 노력을 해 오고 있다.

본 연구는 사례기반추론모형을 예측을 위한 기본 모형으로 설정하고, 데이터 간의 유사도 측정에 퍼지 관계의 개념을 적용함으로써 개선된 예측성과를 얻고자 하였다. 특히, 독립변수 중 기호 데이터 형식의 속성을 가지는 변수들 간의 유사도를 측정하기 위해 이진논리의 개념(일치여부의 판단)과 퍼지 관계 및 합성의 개념을 이용하여 도출된 유사도 매트릭스를 사용하였다.

연구 결과, 기호 데이터 형식의 속성을 가지는 변수들 간의 유사도 측정에서 퍼지 관계 및 합성의 개념을 적용하는 방법이 이진논리의 개념을 적용하는 방법과 비교하여 더 우수한 예측정확성을 나타내었다. 그러나 유사도 측정을 위해 다양한 퍼지합성방법(Max-min 합성, Max-product 합성, Max-average 합성)을 적용하여 예측하는 경우에는 예측정확성 측면에서 퍼지 합성방법 간의 통계적인 차이는 유의하지 않았다. 본 연구는 사례기반추론 모형의 구축에서 가장 중요한 유사도 측정에 있어서 퍼지 관계 및 퍼지 합성의 개념을 적용함으로써 유사도 측정 및 적용 방법론을 제시하였다는데 의의가 있다.

논문접수일 : 2010년 10월 01일 게재확정일 : 2010년 11월 05일 교신저자 : 신경식

1. 서 론

미래에 대한 정확한 예측은 경영자, 또는 기업이 수행하는 경영의사결정에 매우 중요한 역할을 한다. 예측만 정확하다면 경영의사결정의 질은 매우 높아질 수 있을 것이다. 하지만 점점 가속화되고 있는 경영 환경의 변화로 말미암아 미래 예측을 정확하게 하는 일은 점점 더 어려워지고 있다. 이에 기업에서는 정확한 예측을 위하여 전문가의 휴리스틱뿐만 아니라 과학적 예측모형을 함께 활용하여 예측의 성과를 높이는 노력을 해 오고 있다.

일반적으로 예측은 과거의 경험 및 자료를 바탕으로 이루어지며, 이에 따라 정성적 접근(Qualitative Approach)과 정량적 접근(Quantitative Approach)으로 구분할 수 있다. 전통적으로 정량적 접근에 의한 방법론 중 가장 널리 사용되어 왔던 예측모형 구축 방법론은 통계기법에 의한 것이었다. 그러나 1990년대 후반부터 숫자 데이터 뿐만 아니라 기호 데이터의 처리가 가능하고 비선형성을 표현하는데 용이한 인공지능기법이 많이 활용되기 시작하였다. 예측모형을 구축하는데 많이 사용된 대표적인 인공지능기법으로는 인간이 가진 뇌 신경

세포의 구조와 유사하게 학습하고 그 과정을 통해 찾아낸 지식을 사용하여 문제를 해결하는 인공신경망(Artificial Neural Networks), 의사결정나무 분석(Decision Trees), 사례기반 추론(Case-based Reasoning) 기법 등이 있다.

사례기반추론(Case-based Reasoning) 기법은 최근 그 활용정도가 증대되고 있는 인공지능기법으로 현재의 문제를 해결함에 있어 과거의 유사한 사례로부터 그 해결책을 유추한다. 고객지원, 품질보증, 항공기 유지보수, 공정계획, 의사결정 지원 등과 같은 다양한 분야에서 사례기반추론이 활용되고 있으며, 진단, 일정계획 및 설계와 같은 과정들을 자동화하는데 기여하여 기업의 효율성을 증가시키고 원가를 감소시켜 주기도 한다(Watson, 1997; Gardingen and Watson, 1999; Suh et al., 1998; Li, 1999).

사례기반추론은 새로운 문제를 해결하기 위해 유사한 과거 사례를 추출하여 문제해결과정에 사용한다. 따라서 사례기반추론에서 가장 중요한 것 중의 하나는 새로운 사례와 기존의 사례 간의 유사도(Similarity)를 측정하는 것이다. 그런데 일반적으로 각 사례를 구성하는 독립 변수(Independent Variable)들 간의 유사도를 측정하기 위해서는 각 독립변수들이 숫자로 표현될 수 있는 숫자 데이터(Numeric Data)이어야 한다는 한계를 가지고 있다. 물론, 독립변수들이 명목척도(Nominal Scale)로 구성되어 있는 기호 데이터(Symbolic Data)인 경우에도 이진논리(Binary Logic)에 근거한 속성(Feature) 간의 일치여부를 파악하여 그 유사도를 측정하는 방법이 사용되지만, 이와 같은 방식은 실제로 유사도를 측정하는 것이 아니라 일치여부만을 나타내기 때문에 사례 간의 유사도를 측정하는데 충분한 정보를 제공해 주지 못한다는 문제가 있다(Amen and Vomacka, 2001).

본 연구에서는 사례기반추론모형 구축에 있어서 데이터 간의 유사도 측정에 퍼지 관계의 개념을 적용함으로써 개선된 예측성과를 얻고자 하였다. 특히, 독립변수 중 기호 데이터 형식의 속성을 가지는 변수들 간의 유사도를 측정하기 위해 이진논리의 개념과 퍼지 관계 및 합성의 개념을 이용하여 도출된 유사도 매트릭스를 사용하였다.

본 연구의 구성은 다음과 같다. 제 2장에서는 본 연구에서 제시된 모형 구축 방법론과 관련된 주요 개념들과 이론적 배경을 설명하고, 제 3장에서는 본 연구의 프레임워크를 제시한다. 제 4장에서는 제 3장에서 제시한 프레임워크를 기반으로 수행한 실증분석 결과를 제시하고, 제 5장에서는 실증분석 결과를 토대로 연구결과를 도출한다. 마지막으로 제 6장에서는 본 연구의 의의 및 한계점 등을 제시한다.

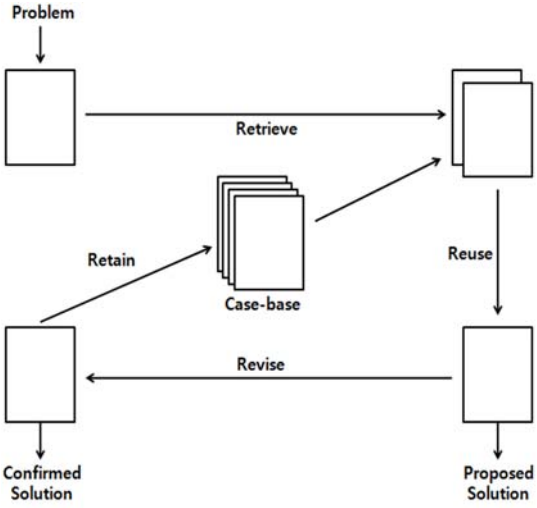
2. 이론적 배경

2.1 사례기반추론

Shank(1982)는 동태적인 인간의 기억에 남아있는 경험을 토대로 하여 학습하는 이론을 개발하였다(Watson, 1997; Barletta, 1991). 사례기반추론은 인간이 과거 경험에 비추어 사물을 인식한다는 점에 착안하여 나타난 인공지능의 한 분야이며, 과거의 문제해결에 대한 경험을 새로운 문제에 적용하여 그 해결책을 제시하는 시스템으로 정의할 수 있다(Riesbeck and Shank, 1989; Kolodner and Mark, 1992).

사례기반추론시스템에서 사례(Case)는 문제(Problem)와 해답(Solution)에 대한 서술을 의미한다(Wang et al., 2003). 사례기반추론의 과정을 살펴보면 <그림 1>과 같다. 새로운 문제가 발생하

면 우선 사례기반(Case-base)으로부터 가장 유사한 사례를 추출하고(Retrieve), 추출된 사례는 새로운 문제의 해답으로 제안된다(Proposed Solution). 물론, 새로운 문제의 해답은 사전의 경험과 지식, 문제의 상황 등에 따라 수정(Revise)될 수 있다. 이렇게 해서 확정된 해답(Confirmed Solution)이 최종적으로 새로운 문제의 해답이 되며, 새로운 문제를 위해 확정된 해답은 사례기반에 추가(Retain)되어 향후에 발생할 또 다른 새로운 문제의 해결에 활용된다.



<그림 1> 사례기반추론과정(Watson, 1997)

2.2 유사도 측정

앞에서 언급한 사례기반추론의 과정에서 사례의 추출(Case Retrieving)은 그 성과에 직접적인 영향을 미치는 과정이다. 사례의 추출이란 새로운 문제와 사례기반에 저장되어 있는 과거의 사례들을 비교하면서 가장 유사한 사례를 찾아가는 과정을 의미한다(Brown and Gupta, 1994). 따라서 이 과정에서 제일 중요한 요소가 바로 사례 간의 유사도를 측정하는 것이다. 사례기반추론에서 유사도를 측정하는 방법은 일반적으로 독립변수의 데이터 속성에 영향을 받게 되는데, 독립변수의 데이터 속성은 크게 숫자 데이터(Numeric Data), 기호 데이터(Symbolic Data), 집합 데이터(Sets)로 구분할 수 있다(Amen and Vomacka, 2001).

독립변수의 데이터 속성이 숫자 데이터인 경우에 유사도를 측정하는 가장 대표적인 방법은 거리(Distance)의 개념을 이용하는 것이고, 일반적으로 유사도를 측정하기 위해 유클리디안 거리(Euclidean-based Distance)의 개념을 이용한다. 여기서 유클리디안 거리란 다음과 같이 정의할 수 있으며, W_i는 i번째 속성의 가중치를 의미하고, f_i^I은 새로운 사례의 속성값이고, f_i^R은 사례기반의 속성값이다(Hsu et al., 2004).

$$DIST = \sqrt{\sum_{i=1}^{n} W_i(f_i^I - f_i^R)^2}$$

이러한 유클리디안 거리의 개념을 일반화시킨 것을 민코우스키 거리(Minkowski Distance)라 하는데 이는 다음과 같이 정의되며, 여기서 $m = 2$일 때 유클리드 거리를 나타내며, $m = 1$일 때는 맨하탄 거리(Manhattan or Rectilinear Distance)를 나타낸다(Burkhard, 2001).

$$DIST = (\sum_{i=1}^{n} |X_i - Y_i|^m)^{1/m}$$

그런데, 일반적으로 유사도는 0과 1사이의 범위를 가지지만, 거리는 음이 아닌 모든 값의 범위를 가진다(Wang et al., 2003). 따라서 거리의 개념을 이용하여 유사도를 측정하기 위해서는 거리의 값을 0과 1사이로 측정이 가능하도록 하는 것이 필

요하며, 이 때 다음과 같은 함수식을 사용할 수 있다. 여기서 X_i와 Y_i는 유사도를 측정하고자 하는 두 개 사례의 i번째 속성값을 의미하고, MAX_i와 MIN_i는 각각 i번째 속성값이 가지는 최대값과 최소값을 의미한다(McLachlan, 2004).

$$DIS(X_i, Y_i) = \frac{|X_i - Y_i|}{|MAX_i - MIN_i|}$$

그리고 거리의 값을 0과 1사이로 측정하여 거리와 유사도 사이의 관계를 살펴보면, 개념적으로 두 사례 간의 거리가 짧아지면 두 사례 간의 유사도는 증가하는 상충관계를 가진다고 할 수 있다. 따라서 유사도와 거리의 관계를 다음과 같이 정의할 수 있다(McLachlan, 2004).

$$SIM(X, Y) = 1 - DIST(X, Y)$$

독립변수의 데이터 속성이 숫자 데이터인 경우에 사용할 수 있는 또 하나의 유사도 측정함수는 다음과 같다(Wang et al., 2003). 여기서 X_i와 Y_i는 각각 두 개 사례의 i번째 속성값을 의미한다.

$$SIM(X_i, Y_i) = \frac{1}{1 + |X_i - Y_i|}$$

독립변수의 데이터 속성이 기호 데이터인 경우에 유사도를 측정하는 대표적인 방법은 두 변수 간의 속성값이 일치하는지의 여부를 유사도 값으로 사용하는 방법이다. 즉 두 변수의 속성값이 일치하면 유사도를 1의 값으로 부여하고, 일치하지 않으면 유사도를 0의 값으로 부여하는 방법이다. 이를 수식으로 표현하면 다음과 같다(Wang et al., 2003; Amen and Vomacka, 2001). 여기서 X_i와 Y_i는 각각 두 개 사례의 i번째 속성값을 의미한다.

$$SIM(X_i, Y_i) = \begin{cases} 1, (X_i = Y_i) \\ 0, (X_i \neq Y_i) \end{cases}$$

독립변수의 데이터 속성이 기호 데이터인 경우에 사용할 수 있는 또 하나의 유사도 측정방법은 유사도 매트릭스(Similarity Matrix)를 사용하는 방법이다. 이 방법은 비교하는 두 변수의 속성값에 대한 유사도를 측정하기 위해 앞에서 설명한 이진논리(Binary Logic)에 의하여 일치여부를 판단하여 일치하는 경우에는 1, 일치하지 않는 경우에는 0의 값을 부여하는 방식이 아니라 각 속성별로 미리 정의된 유사도 매트릭스를 이용하여 두 속성 간의 유사도를 측정하는 방법이다. 이를 수식으로 표현하면 다음과 같다(Wang et al., 2003). 여기서 X_i와 Y_i는 각각 두 개 사례의 i번째 속성값을 의미하고, $value(X_{im}, Y_{\in})$은 X_i가 m의 값을 가지고 Y_i가 n의 값을 가질 때 사전에 정의된 유사도 매트릭스 상의 값을 유사도로 이용한다는 것을 의미한다.

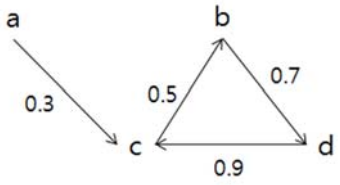
$$SIM(X_i, Y_i) = value(X_{im}, Y_{\in})$$

2.3. 이항관계와 퍼지 관계

컴퓨터가 인공지능을 가지고 인간이 원하는 바를 제대로 수행하기 위해서는 인간이 사용하는 숫자는 물론이고 모호한 표현과 지식도 처리할 수 있어야 한다. 인간의 모호한 표현을 처리할 수 있는 이론적 바탕을 제공하는 것이 바로 퍼지이론(Fuzzy Theory)이다. 퍼지이론은 현상의 불확실한 상태를 그대로 표현해 주는 방법으로서 1965년 미국 버클리 대학의 자데(Zadeh, L. A.) 교수에 의해서 처음 소개되었다. 퍼지이론은 모호하게 표현

된 자료를 유용한 자료로 만들기 위해 퍼지집합(Fuzzy Set), 퍼지논리(Fuzzy Logic), 퍼지숫자(Fuzzy Number) 등의 개념을 포함하고 있으며 수학적인 계산방법도 잘 개발되어 있다. 퍼지이론은 현실세계에서 수학적으로 모형을 구축하는 것이 어려운 문제들에 대해 적용하기 적합한 이론이다(Zadeh, 1978). 퍼지이론이 소개된 이후에 퍼지이론은 과학적인 문제나 공학적인 문제에 폭넓게 적용되어 오고 있다(Ross, 2004; Pires and Hoshiau, 2002; Kosko, 1991). 그러나 퍼지이론의 잠재성에도 불구하고 경영의사결정에 대한 퍼지이론의 적용은 미약하다(Dorsey and Coovert, 2003; Alliger et al., 1993).

현존하는 수학에서 다루는 기본적인 관계인 이항관계(Binary Relation)는 두 집합 X, Y에 각각 속하는 임의 원소 x와 y가 서로 관계가 있거나 또는 서로 관계가 없거나 둘 중의 하나만을 의미하는 이치적 관계이다. 그러나 현실세계에서의 모든 관계를 이치적 관계로 다룰 수는 없다. 따라서 퍼지 관계(Fuzzy Relation)는 이러한 두 원소간의 이치적 관계성을 퍼지의 개념으로 일반화시킨 관계성을 의미한다. 즉 퍼지 관계는 현실세계에서 모든 것들의 관계성의 정도(the Grade of Relationship)를 계량적으로 표현하고자 하는 것이다. 이러한 퍼지 관계는 <그림 2>와 같이 표현할 수 있다.



<그림 2> 퍼지 관계

<그림 2>에서 a, b, c, $d \in X$(전체집합)이며 각 연결선에 표시된 숫자는 퍼지 관계의 정도를 의미한다. 즉 <그림 2>에서 나타난 퍼지 관계의 내용은 a와 c 사이에는 퍼지 관계가 0.3정도 있고, b와 d 사이에는 퍼지 관계가 0.7정도 있고, c와 b 사이에는 퍼지 관계가 0.5정도 있고, d와 c 사이에는 퍼지 관계가 0.9정도 있다는 것을 의미한다.

퍼지 관계는 컴퓨터에서 문제해결에 필요한 지식(Knowledge)을 표현하는데 많이 이용된다. 즉 지식베이스시스템에서 지식은 "*IF A THEN B*" 형태의 규칙으로 표시되는데 이 경우 조건부 A와 결론부 B와의 관계를 나타낼 때 퍼지 관계로써 나타낸다. 퍼지 관계를 표시하는 방법은 여러 가지가 있으나 가장 많이 사용하는 방법은 행렬(Matrix)로 표시하는 방법이다.

2.4. 퍼지 관계의 합성

퍼지 관계의 합성은 이항관계의 합성을 확장한 개념이라고 할 수 있다. 두 개의 퍼지 관계 R과 S가 각각 $R \subseteq X \times Y$이고 $S \subseteq Y \times Z$일 때, 집합 X와 Z 사이의 관계는 두 개의 퍼지 관계 R과 S의 합성으로 얻을 수 있다. Zadeh(1975a, 1975b, 1975c)는 퍼지 관계의 합성에 대해서 다음과 같은 Max-*(Star) 합성을 제안하였다.

$$R(x, y) * S(y, z) = \{((x, z), \mu_{r*s}(x, z)) \mid x \in X,\ y \in Y,\ z \in Z\}$$

$$\mu_{R*S} = \max\{\mu_R(x, y)\ *\ \mu_S(y, z)\}$$

Zadeh의 Max-*(Star) 합성은 다양한 형태로 표현될 수 있는데, *(Star) 연산이 Min 연산인 경우인 Max-Min 합성, 대수곱(Algebraic Product) 연산으로 정의한 합성을 Max-Product 합성, 산술평균(Arithmetic Mean) 연산으로 정의한 합성을

Max-Average 합성이 있다. 각각의 연산은 다음과 같이 표현할 수 있다.

<Max-Min 합성>

$$R(x, y) \cdot S(y, z) = \{((x, z), \mu_{r \cdot s}(x, z)) \mid x \in X,\ y \in Y,\ z \in Z\}$$

$$\mu_{R \cdot S} = \max\{\min\{\mu_R(x, y),\ \mu_S(y, z)\}\}$$

<Max-Product 합성>

$$R(x, y) * S(y, z) = \{((x, z),\ \mu_{r*s}(x, z)) \mid x \in X,\ y \in Y,\ z \in Z\}$$

$$\mu_{R*S} = \max\{\mu_R(x, y) \times \mu_S(y, z)\}$$

<Max-Average 합성>

$$R(x, y) * S(y, z) = \{((x, z),\ \mu_{r*s}(x, z)) \mid x \in X,\ y \in Y,\ z \in Z\}$$

$$\mu_{R*S} = \max\{[\mu_R(x, y) + \mu_S(y, z)]/2\}$$

이러한 퍼지 합성의 개념은 대부분 최적화의 문제를 해결하기 위해 이용되었는데, Wang(1995), Fang and Li(1999), Ghodousian and Khorram (2008), Wu et al.(2010)은 선형 또는 비선형의 문제를 Max-Min 합성을 이용하여 해결하려고 하였다. 또한, 비선형의 문제를 해결하기 위해 Khorram and Hassanzadeh(2008)은 Max-Average 합성과 유전자 알고리즘(Genetic Algorithm)을 이용하였으며, Khorram and Zarei(2009)는 다목적 계획을 해결하기 위해 Max-Average 합성을 이용하였다. Loetamonphong and Fang(2001), Li and Hu(2010)은 최적해 도출을 위해 Max-Product 합성을 이용하였는데, 특히 Markovskii(2005)와 Lin(2009)은 순회외판원문제(Traveling Salesman Problem)와 같은 NP-hard 문제를 해결하기 위해 Max-Product 합성을 이용하였다.

3. 연구설계

3.1 개요

본 연구에서는 사례기반 추론 모형을 예측을 위한 기본 모형으로 설정하고, 데이터 간의 유사도 측정에 퍼지 관계의 개념을 적용함으로써 개선된 예측성과를 얻고자 하였다. 이를 실증적으로 분석하기 위해 '데이터의 수집 → 모형의 구축 및 실증분석 → 모형의 타당도 검증 → 결론의 도출' 순으로 연구를 진행하였다.

'데이터의 수집' 단계에서는 종속변수가 범주형의 형태를 가지는 데이터를 수집하였다. 수집된 데이터는 횡단면 데이터(Cross-sectional Data)이며, 데이터에 포함되어 있는 독립변수 중에는 기호 데이터를 포함하는 정성 변수(Qualitative Variables)가 다수 포함되어 있다. 횡단면 데이터를 이용하여 연구를 진행한 이유는 본 연구에서 사용되는 기본적인 예측방법인 판별분석과 사례기반 추론이 모두 독립변수와 종속변수의 인과관계를 규명함으로써 예측을 하는 방법들이기 때문이다. '모형의 구축 및 실증분석' 단계에서는 기준모형, 정성 변수의 처리 등을 고려하여 세부 모형을 구축하였으며, '모형의 타당도 검증' 단계에서는 구축된 모형들의 실증분석 결과들을 기준모형과 비교하여 통계적으로 그 유의성을 검증하였다. 실증분석결과의 유의성 검증은 예측값들의 적중률(Hit Ratio)을 도출한 후 맥네마 검정(McNemar Test)을 이용하여 수행하였다(Shin and Kim, 2004). 마지막으로, '결론의 도출' 단계에서는 연구결과를 종합하여 결론을 도출하였다.

3.2 모형 구축과 유사도 매트릭스

사례기반 추론 모형 구축에 있어서 성과에 가장

<그림 3> 유사도 매트릭스의 도출과정 예시[1)]

중요한 영향을 미치는 요소는 유사도의 측정 방식이다. 현재의 문제를 해결하기 위하여 사례베이스(Case-base)에서 가장 가까운, 또는 유사한 사례를 추출하기 위해서는 유사도를 측정하는 함수의 적정성이 보장되어야 한다. 측정하고자 하는 대상 변수가 숫자형일 경우에는 위에서 서술한 유클리디언 거리(Euclidean distance) 등을 활용하는 등 상대적으로 측정이 용이한 측면이 있다. 그러나 기호 형태를 가지는 정성적인 변수일 경우에는 두 변수간의 유사도를 측정하는 일이 매우 어려운 일이 된다. 전술한 바와 같이 많은 연구에서 이진논리(Binary Logic)에 근거하여 속성(Feature) 간의 일치여부를 파악하여 그 유사도를 측정하는 방법이 사용되었으나, 이와 같은 방식은 실제로 유사한 정도를 측정하는 것이 아니라 일치여부만을 나타내기 때문에 사례 간의 유사도를 측정하는데 충분한 정보를 제공해 주지 못한다는 문제가 있었다(Amen and Vomacka, 2001).

본 연구에서는 기호 형태로 표현된 두 정성 변수 간의 속성값 사이의 일치여부를 통해 그 여부를 유사도 값으로 부여하던 방법에서 벗어나 각 속성이 가지는 값들 간의 유사도를 개별적으로 측정하여 유사도 매트릭스를 정의하고, 이를 통해 가장 유사한 사례를 사례베이스로부터 추출함으로써 예측의 성과를 높이고자 하였다. 또한, 유사도 매트릭스에 보다 적절하고 타당한 유사도의 값을 부여하기 위한 방법론으로 퍼지 관계라는 개념을 적용하였다.

유사도 매트릭스는 사전에 정의된 사례기반을 기준으로 기호 데이터를 포함하는 정성 변수들에 대해서 해당 정성 변수와 종속변수 사이의 소속정도(Membership)를 계산하여 1차로 매트릭스를 도출한 후 퍼지 합성(Fuzzy Composition)을 통해 최종적인 유사도 매트릭스를 도출하였다. 유사도 매트릭스를 도출하기 위해 사용된 퍼지 합성은 Max-Min 합성, Max-Product 합성, Max-Average 합성이다. <그림 3>은 유사도 매트릭스를 도출하는 과정을 예시적으로 보여주는 그림이다.

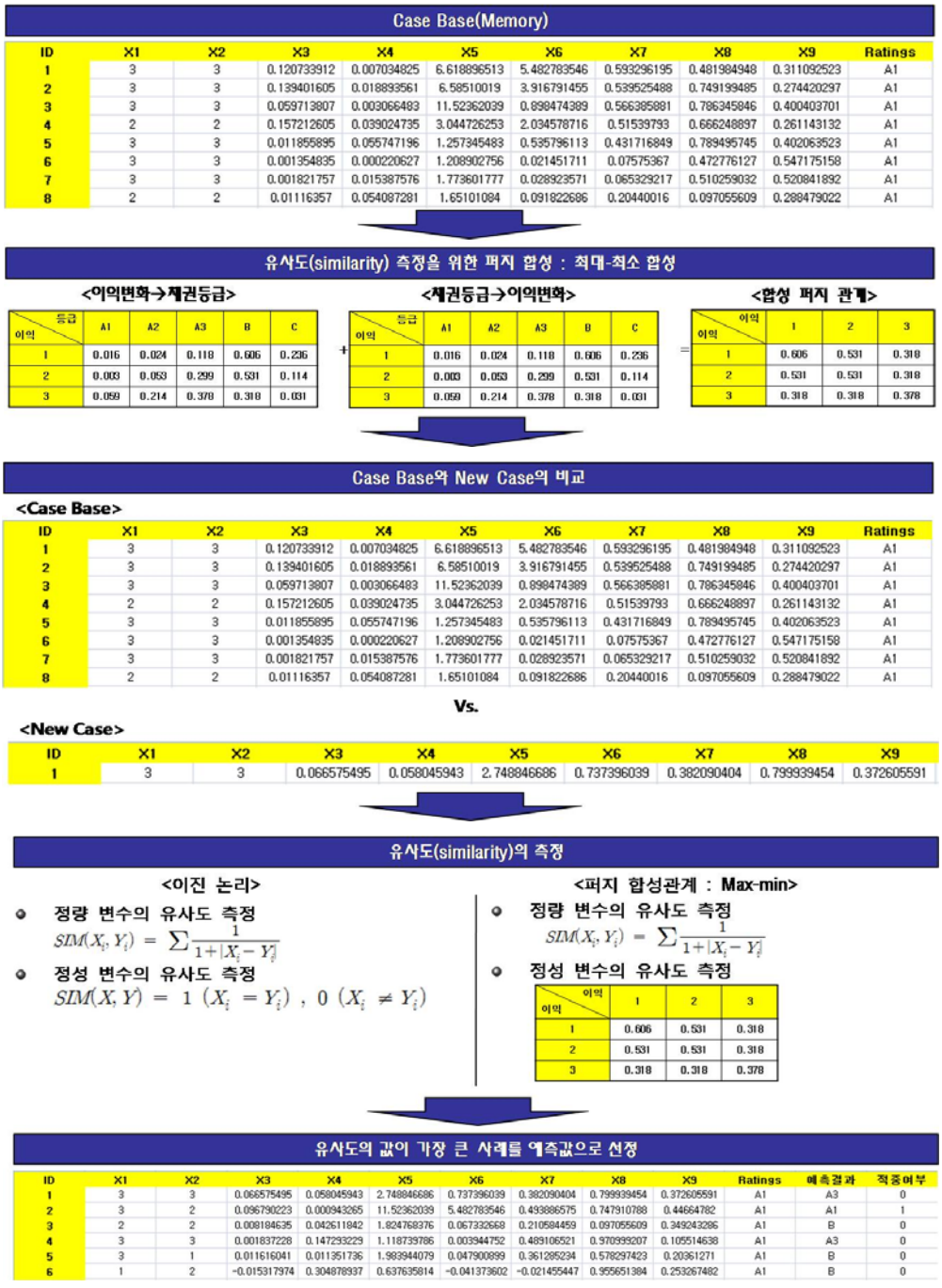
이렇게 구축된 유사도 매트릭스는 기호형 정성변수의 변수간 거리를 제공한다. 숫자형 변수와 기호형 정성변수가 혼재되어 있다는 가정 하에 본 연구에서 제시하고 있는 유사도의 측정 및 사례추출을 통한 예측 과정은 <그림 4>에 제시하고 있다.

4. 실증분석

4.1 데이터의 구성

종속변수의 예측에 사용된 데이터는 A신용평가사가 1997년부터 2000년에 걸쳐 수행한 한국기업

1) 본 예시는 Max-Min 합성을 통해 유사도 매트릭스를 도출한 예시임.


Case Base(Memory)
유사도(similarity) 측정을 위한 퍼지 합성 : 최대-최소 합성
<이익변화→채권등급>
<채권등급→이익변화>
<합성 퍼지 관계>
Case Base와 New Case의 비교
<Case Base>
Vs.
<New Case>
유사도(similarity)의 측정
<이진 논리>
정량 변수의 유사도 측정
정성 변수의 유사도 측정
<퍼지 합성관계 : Max-min>
정량 변수의 유사도 측정
정성 변수의 유사도 측정
유사도의 값이 가장 큰 사례를 예측값으로 선정

<그림 4> 예측과정

<표 1> 독립 변수의 구성

독립 변수	변수 정의	변수형태	변수값	속성값
X1	경상이익의 변화	정성	1 2 3	(–, –, –) (–, –, +) ~ (+, +, –) (+, +, +)
X2	영업현금흐름의 변화	정성	1 2 3	(–, –, –) (–, –, +) ~ (+, +, –) (+, +, +)
X3	자산수익률	정량		
X4	순이자 보상배율	정량		
X5	이자 보상배율	정량		
X6	자본금 순이익율	정량		
X7	자기자본비율	정량		
X8	고정자산비율	정량		
X9	유동부채비율	정량		

의 채권등급평가결과로, 종속변수는 A1, A2, A3, B, C와 같은 5개의 등급으로 정의되어 있는 범주형 데이터이다. 데이터의 수는 1,816개이며, 재무비율을 비롯한 채권평가 관련 정보를 포함하고 있는데, 본 연구를 위한 독립변수로 총 9개를 선정하였다. 이 중 기호형 정성 변수는 2개가 해당된다. 본 연구에 사용된 독립변수들과 그 구성은 <표 1>과 같다.

본 데이터에 포함되어 있는 2개의 정성 변수는 경상이익의 변화(Transition of Ordinary Profit)와 영업현금흐름의 변화(Transition of Operating Activities Cash Flows)이다. 이는 각각 3개년도 동안에 경상이익이나 영업현금흐름의 변화를 정성화한 변수인데, (–, –, –)는 3개년 동안 계속 경상이익이나 영업현금흐름이 감소했다는 것을 의미하고, (–, +, –)는 1차 년도와 3차 년도에는 경상이익이나 영업현금흐름이 감소했지만 2차 년도에는 경상이익이나 영업현금흐름이 증가했음을 의미한다. 정성 변수 외에 자산수익률(Net Income to Total Asset), 순이자 보상배율(Net Interest Coverage Ratio), 이자 보상배율(Times Interest Earned, Interest Coverage Ratio), 자본금 순이익율(Net Income to Capital Stock), 자기자본비율(Equity to Total Asset), 고정자산비율(Fixed Assets to Total Asset), 유동부채비율(Current Liabilities to Total Asset) 등의 정량 변수가 예측에 사용되었다.

실증분석에 사용된 데이터의 총 수량은 1,816개였으며, 이 중 80%에 해당하는 1,454개의 데이터는 사례기반 구축용으로 사용하였고, 20%에 해당하는 362개의 데이터는 모형의 타당도를 검증하기 위한 검증용 데이터로 사용하였다. 해당 데이터를 종속변수 기준으로 사례기반 구축용 데이터와 검증용 데이터를 구분하면 <표 2>와 같으며, 종속변수의 범주(채권 등급)별 비율은 사례기반 구축용

데이터와 검증용 데이터에서 유사한 비율을 보이도록 구분하였다. 즉 사례기반 구축용 데이터와 검증용 데이터를 범주(채권 등급)별로 80 : 20의 비율로 구분한 것과 동일하게 구분한 것이다.

<표 2> 종속변수의 구성

등급	사례기반 구축용 데이터		검증용 데이터	
	빈도	비율	빈도	비율
A1	48	3.30%	10	2.76%
A2	192	13.20%	50	13.81%
A3	470	32.32%	116	32.04%
B	624	42.92%	156	43.09%
C	120	8.25%	30	8.29%
합계	1,454	100.00%	362	100.00%

4.2 분석과정

분석을 위한 벤치마크로 다중판별분석을 실시하였다. 제시된 9개의 독립변수를 활용하여 도출된 판별함수식은 다음과 같다.

$$Y = 0.435X_{11} + 0.538X_{12} + 0.792X_{21} + 0.515X_{22} + 2.237X_3 + 6.155X_4 - 0.127X_5 - 0.204X_6 - 1.134X_7 - 2.329X_8 + 0.674X_9 + 0.653$$

사례기반 추론 모형은 유사도의 측정방법에 따라 다음과 같은 2가지 방식에 의해 구축되었다. 하나는 이진논리를 적용하여 유사도를 측정하는 방법이고, 또 하나는 퍼지 합성을 이용한 유사도 매트릭스에 의거하여 유사도를 측정하는 방법이다.

이진논리를 적용하여 유사도를 측정하는 방법은 두 변수 간의 일치여부를 판단하여 일치하면 1의 값을 유사도 값으로 부여하고, 일치하지 않으면 0의 값을 유사도 값으로 부여하는 것이다. 따라서 이러한 이진논리를 유사도 매트릭스로 표현하면 <그림 5>와 같다.

위와 같이 이진논리를 적용하여 유사도를 측정하는 방법은 각 속성의 일치 여부만을 판단하여 유사도 매트릭스를 완성할 수 있지만, 퍼지합성을 이용하여 유사도를 측정하는 방법을 사용하면 각 속성간의 유사 정도에 따라 상대적으로 정교한 유사도 매트릭스를 도출할 수 있다. 본 연구에서는 매트릭스에 정의되는 각 속성 간의 유사도 값을 도출하기 위해 퍼지합성의 개념을 사용하였으며, 퍼지합성의 방법은 Max-Min 합성, Max-Product 합성, Max-Average 합성의 세 가지 방법을 사용하였다. 각 합성방법에 따라 도출된 유사도 매트릭스는 <그림 6>~<그림 8>과 같다.

사례추출을 위해 사용된 변수별 가중치는 독립변수와 종속변수간의 피어슨(Pearson) 상관계수를 사용하였다.

<경상이익의 변화>

이익 \ 이익	1	2	3
1	1.0	0.0	0.0
2	0.0	1.0	0.0
3	0.0	0.0	1.0

<영업 현금흐름의 변화>

현금 \ 현금	1	2	3
1	1.0	0.0	0.0
2	0.0	1.0	0.0
3	0.0	0.0	1.0

<그림 5> 이진논리를 적용한 유사도 매트릭스

<경상이익의 변화>

이익 \ 이익	1	2	3
1	0.606	0.531	0.318
2	0.531	0.531	0.318
3	0.318	0.318	0.378

<영업 현금흐름의 변화>

현금 \ 현금	1	2	3
1	0.614	0.526	0.264
2	0.526	0.526	0.279
3	0.264	0.279	0.428

<그림 6> Max-Min 퍼지합성을 적용한 유사도 매트릭스

<경상이익의 변화>

이익 \ 이익	1	2	3
1	0.368	0.322	0.193
2	0.322	0.282	0.169
3	0.193	0.169	0.143

<영업 현금흐름의 변화>

현금 \ 현금	1	2	3
1	0.377	0.323	0.162
2	0.323	0.277	0.139
3	0.162	0.139	0.183

<그림 7> Max-Product 퍼지합성을 적용한 유사도 매트릭스

<경상이익의 변화>

이익 \ 이익	1	2	3
1	0.606	0.568	0.462
2	0.568	0.531	0.424
3	0.462	0.424	0.378

<영업 현금흐름의 변화>

현금 \ 현금	1	2	3
1	0.614	0.570	0.439
2	0.570	0.526	0.395
3	0.439	0.395	0.428

<그림 8> Max-Average 퍼지합성을 적용한 유사도 매트릭스

4.3 분석결과

본 연구에서 사용한 모형들을 통해 예측한 값의 적중률은 <표 3>과 같다. <표 3>에서 보는 바와 같이 각 예측모형별 적중률은 판별분석모형이 29.28%로 가장 낮고, Max-Min 합성을 이용한 사례기반 추론 모형에서 43.65%로 가장 크게 나타나고 있다.

<표 3> 각 예측모형별 적중률

예측모형		적중률
판별분석		29.28%
사례기반 추론	이진논리	36.46%
	Max-Min 합성	43.65%
	Max-Product 합성	43.37%
	Max-Average 합성	43.37%

<표 4> 각 모형 간 유의성 검증결과

	판별분석	이진논리	Max-Min	Max-Product	Max-Average
판별분석	-	0.043**	0.000***	0.000***	0.000***
이진논리		-	0.030**	0.037**	0.037**
Max-Min			-	1.000	1.000
Max-Product				-	1.000

<표 4>는 모형별 성과간의 차이를 검증한 결과로 맥네마 검정(McNemar Test)을 수행하였다. <표 4>에서 *는 유의수준 10%를 의미하며, **는 유의수준 5%를 의미하고, ***는 유의수준 1%를 의미한다. <표 4>에서 제시된 바와 같이, 퍼지 합성 방식을 적용하여 도출한 유사도 매트릭스를 적용한 사례기반 모형이 벤치마크인 다중 판별분석 모형과 이진 논리를 적용한 사례기반 추론 모형에 비해 그 성과가 통계적으로 유의할 정도로 우수한 것을 알 수 있다. 그러나 퍼지 합성 방식에 따른 성과 차이는 있다고 볼 수 없었다.

5. 연구결과

본 연구는 기존의 숫자 데이터에 기반한 통계모형과 차별화된 예측모형으로 사례기반 추론 모형을 제시하고, 기호 데이터 간의 유사도 측정에 퍼지 관계의 개념을 적용함으로써 더 높은 예측성과를 달성할 수 있는 방법을 찾기 위한 목적으로 수행되었다. 이를 위해 본 연구에서는 범주형의 종속변수를 가지고 있는 예측용 데이터를 수집하고, 독립변수 중 기호 데이터 형식의 속성들 각각에 대해서 이진논리 및 퍼지합성의 개념을 이용하여 유사도 매트릭스를 도출하였다. 그리고 이렇게 도출된 유사도 매트릭스를 이용하여 사례기반 추론에 의한 예측활동을 수행하였다.

본 연구에서 도출된 첫 번째 결과는 예측정확성의 향상이라는 관점에서 통계모형과 사례기반 추론 모형의 차이에 관한 것이다. 그 결과는 <표 3>과 <표 4>에서와 같이 기존의 통계모형보다 사례기반 추론 모형에 의한 예측정확성이 더 좋다는 것을 보여주고 있다.

본 연구에서 도출된 두 번째 결과는 예측정확성의 향상이라는 관점에서 정성적인 독립변수를 구성하는 기호 데이터 간의 유사도를 측정하기 위한 방법으로 이진논리를 이용한 사례기반 추론 모형과 퍼지합성을 이용한 사례기반 추론 모형의 차이에 관한 것이다. 그 결과는 <표 3>과 <표 4>에서와 같이 이진논리를 이용한 사례기반 추론 모형보다 퍼지합성을 이용한 사례기반 추론 모형에 의한 예측정확성이 더 우수하다는 것을 보여주고 있다. 따라서 정성적인 독립변수를 구성하는 기호 데이터 간의 유사도를 측정하는 방법으로 이진논리를 적용하는 경우보다 퍼지합성을 이용하는 경우가 예측정확성의 향상이라는 관점에서 우월하다는 결론을 얻을 수 있다.

본 연구에서 도출된 세 번째 결과는 예측정확성의 향상이라는 관점에서 정성적인 독립변수를 구성하는 기호 데이터 간의 유사도를 측정하기 위해 이용되는 퍼지 관계 합성방법 간의 차이에 관한 것이다. 그 결과는 <표 3>과 <표 4>에서와 같이 Max-Min 합성, Max-Product 합성, Max-Average

합성을 이용한 개별 사례기반 추론 모형 사이에는 예측정확성의 통계적인 차이가 없다는 것을 보여주고 있다. 따라서 정성적인 독립변수를 구성하는 기호 데이터 간의 유사도를 측정하기 위해 퍼지합성을 이용하는 경우에 퍼지합성방법 간의 차이는 없다는 결론을 얻을 수 있다.

6. 연구의 의의 및 한계점

본 연구는 사례기반 추론 모형을 예측을 위한 기본 모형으로 설정하고, 데이터 간의 유사도 측정에 퍼지 관계의 개념을 적용함으로써 개선된 예측성과를 얻고자 하였다. 특히, 독립변수 중 기호 데이터 형식의 속성을 가지는 변수들 간의 유사도를 측정하기 위해 이진논리의 개념과 퍼지 관계 및 합성의 개념을 이용하여 도출된 유사도 매트릭스를 사용하였다.

연구 결과, 기호 데이터 형식의 속성을 가지는 변수들 간의 유사도 측정에서 퍼지 관계 및 합성의 개념을 적용하는 방법이 이진논리의 개념(일치여부의 판단)을 적용하는 방법과 비교하여 더 우수한 예측정확성을 나타내었다. 그러나 유사도 측정을 위해 다양한 퍼지합성방법(Max-min 합성, Max-product 합성, Max-average 합성)을 적용하여 예측하는 경우에는 예측정확성 측면에서 퍼지합성방법 간의 통계적인 차이는 유의하지 않았다. 본 연구는 사례기반 추론 모형의 구축에서 가장 중요한 유사도 측정에 있어서 퍼지 관계 및 퍼지 합성의 개념을 적용함으로써 유사도 측정 및 적용방법론을 제시하였다는데 의의가 있다.

그러나 본 연구는 하나의 데이터 집합(Data Set)을 통해서만 제안 모형의 성과를 검증했다는 점, 통제를 위한 것이었지만 적용된 독립변수별 가중치를 피어슨(Pearson) 상관계수로 한정하였다는 점, 정성변수의 수가 2개에 국한되어 있었다는 점 등 많은 한계를 가지고 있다. 이러한 문제점은 추후 연구를 통해 보완될 예정이다.

참고문헌

Alliger, G. M., S. L. Feinzig and E. A. Janak, "Fuzzy Sets and Personnel Selection : Discussion and an Application", *Journal of Occupational and Organization Psychology*, Vol.66, No.2(1993), 163~169.

Amen, R. and P. Vomacka, "Case-based Reasoning As a Tool for Materials Selection", *Materials and Design*, Vol.22(2001), 353~358.

Barletta, R., "An Introduction to Case-based Reasoning", *AI Expert*, Vol.6, No.8(1991), 42~49.

Brown, C. E. and U. G. Gupta, "Applying Case-based Reasoning to the Accounting Domain", *Intelligent Systems in Accounting, Finance and Management*, Vol.3(1994), 205~221.

Burkhard, H. D., "Similarity and Distance in Case-based Reasoning", *Foundamenta Informaticae*, Vol.47(2001), 201~215.

Dorsey, D. W. and M. D. Coovert, "Mathematical Modeling of Decision Making : A Soft and Fuzzy Approach to Capturing Hard Decision", *Human Factors*, Vol.45, No.1(2003), 117~135.

Fang, S. C. and G. Li, "Solving Fuzzy Relations Equation with a Linear Objective Function", *Fuzzy Sets and Systems*, Vol.7 (1999), 89~101.

Gardingen, D. and I. Watson, "A Web Based CBR System for Heating Ventilation and Air Conditioning Systems Sales Support", *Knowledge-based Systems*, Vol.12(1999), 207

~214.

Ghodousian, A. and E. Khorram, "Fuzzy linear optimization in the presence of the fuzzy relation inequality constraints with max - min composition", *Information Sciences*, Vol.178(2008) 501~519.

Hsu, C., C. Chiu and P. L. Hsu, "Predicting Information Systems Outsourcing Success Using a Hierarchical Design of Case-based Reasoning", *Expert Systems with Applications*, Vol.26(2004), 435~441.

Khorram, E. and R. Hassanzadeha, "Solving nonlinear optimization problems subjected to fuzzy relation equation constraints with max - average composition using a modified genetic algorithm", *Computers and Industrial Engineering*, Vol.55(2008), 1~14.

Khorram, E. and H. Zarei, "Multi-objective optimization problems with Fuzzy relation equation constraints regarding max-average composition", *Mathematical and Computer Modelling*, Vol.49(2009), 856~867.

Kolodner, J. L. and W. Mark, "Case-based Reasoning", *IEEE Expert*, Vol.7, No.5(1992), 5~6.

Kosko, B., Neural Networks and Fuzzy Systems, Prentice-Hall, Upper Saddle River, NJ., 1991.

Li, L. L. X., "Knowledge-based Problem Solving : An Approach to Heath Assessment", *Expert Systems with Applications*, Vol.16 (1999), 33~42.

Li, J. and G. Hu, "A new algorithm for minimizing a linear objective function subject to a system of fuzzy relation equations with max-product composition", *Fuzzy Information and Engineering*, Vol.2(2010), 249~267.

Lin. J., "On the relation between fuzzy max-Archimedean t-norm relational equations and the covering problem", *Fuzzy Sets and Systems*, Vol.160(2009), 2328~2344.

Loetamonphong, J. and S. Fang, "Optimization of Fuzzy Relation Equation with Max-product Composition", *Fuzzy Sets and Systems*, Vol.118(2001), 509~517.

Markowskii, A. V., "On the Relation between Equations with Max-product Composition and the Covering Problem", *Fuzzy Seta and Systems*, Vol.153(2005), 261~273.

McLachlan, G., Discriminant Analysis And Statistical Pattern Recognition, John Wiley and Sons Inc, New York, (2004).

Pires, G. and Y. Hoshiau, "A Wheelchair Steered through Voice Commands and Assisted by a Reactive Fuzzy Logic Controller", *Journal of Intelligent and Robotics Systems*, Vol.34, No.3(2002), 301~314.

Riesbeck, C. K. and R. Schank, Inside Case-Based Reasoning, Lawrence Eribaum Associates, (1989).

Ross, T. J., Fuzzy Logic with Engineering Applications, John Wiley, New York, (2004).

Shin, K. S. and H. J. Kim, "Optimization of Symbolic Similarity by Genetic Algorithm for Case Based Approach to Corporate Bond Rating", *The 9th Asia-Pacific Decision Science Institute Conference*, July, Seoul, Korea, (2004).

Suh, M. S., W. C. Jhee, Y. K. Ko and A. Lee, "A Case-based Expert System Approach for Quality Design", *Expert Systems with Applications*, Vol.15(1998), 181~190.

Wang, H. F., "A Multiple-objectives Mathematical Programming Problem with Fuzzy Relation Constraints", *Journal of Multi-criteria Decision Analysis*, Vol.4(1995), 23~35.

Wang, Z., Z. Liu and X. Ai, "Case Representation and Similarity in High-speed Machin-

ing", *International Journal of Machine Tools and Manufacture*, Vol.43(2003), 1347~1353.

Watson, I., Applying Case-Based Reasoning : Techniques for Enterprise Systems, Morgan Kaufman Publishers, San Francisco, CA., (1997).

Wu, Y., C. Liu, Y. Lur and S. Guu, "One Simple Procedure to Finding the Best Approximate Solution for a Particular Fuzzy Relational Equation with Max-min Composition", Computational Science and Optimization, Third International Joint Conference(2010), 141~144.

Zadeh, L. A., "The Concept of a Linguistic Variable and its Application to Approximate Reasoning-I", *Information Sciences*, Vol.8 (1975a), 199~249.

Zadeh, L. A., "The Concept of a Linguistic Variable and its Application to Approximate Reasoning-II", *Information Sciences*, Vol.8 (1975b), 301~357.

Zadeh, L. A., "The Concept of a Linguistic Variable and its Application to Approximate Reasoning-III", *Information Sciences*, Vol.9 (1975c), 43~80.

Zadeh, L. A., "Fuzzy Sets as a Basis for a Theory of Possibility", *Fuzzy Sets and Systems*, Vol.1(1978), 3~28.

Abstract

A Study on Forecasting Accuracy Improvement of Case Based Reasoning Approach Using Fuzzy Relation

In Ho Lee* · Kyung-shik Shin**

In terms of business, forecasting is a work of what is expected to happen in the future to make managerial decisions and plans. Therefore, the accurate forecasting is very important for major managerial decision making and is the basis for making various strategies of business. But it is very difficult to make an unbiased and consistent estimate because of uncertainty and complexity in the future business environment. That is why we should use scientific forecasting model to support business decision making, and make an effort to minimize the model's forecasting error which is difference between observation and estimator. Nevertheless, minimizing the error is not an easy task.

Case-based reasoning is a problem solving method that utilizes the past similar case to solve the current problem. To build the successful case-based reasoning models, retrieving the case not only the most similar case but also the most relevant case is very important. To retrieve the similar and relevant case from past cases, the measurement of similarities between cases is an important key factor. Especially, if the cases contain symbolic data, it is more difficult to measure the distances.

The purpose of this study is to improve the forecasting accuracy of case-based reasoning approach using fuzzy relation and composition. Especially, two methods are adopted to measure the similarity between cases containing symbolic data. One is to deduct the similarity matrix following binary logic(the judgment of sameness between two symbolic data), the other is to deduct the similarity matrix following fuzzy relation and composition.

This study is conducted in the following order; data gathering and preprocessing, model building and analysis, validation analysis, conclusion. First, in the progress of data gathering and preprocessing we collect data set including categorical dependent variables. Also, the data set gathered is cross-section data and independent variables of the data set include several qualitative variables expressed symbolic data. The research data consists of many financial ratios and the corresponding bond ratings of Korean companies. The ratings we employ in this study cover all bonds rated by one

* Assistant Professor, Dept. of the Accounting Information, Woongji Accounting & Tax College
** Associate Professor, College of Business Admin., Ewha Womans University

of the bond rating agencies in Korea. Our total sample includes 1,816 companies whose commercial papers have been rated in the period 1997~2000. Credit grades are defined as outputs and classified into 5 rating categories(A1, A2, A3, B, C) according to credit levels. Second, in the progress of model building and analysis we deduct the similarity matrix following binary logic and fuzzy composition to measure the similarity between cases containing symbolic data. In this process, the used types of fuzzy composition are max-min, max-product, max-average. And then, the analysis is carried out by case-based reasoning approach with the deducted similarity matrix. Third, in the progress of validation analysis we verify the validation of model through McNemar test based on hit ratio. Finally, we draw a conclusion from the study.

As a result, the similarity measuring method using fuzzy relation and composition shows good forecasting performance compared to the similarity measuring method using binary logic for similarity measurement between two symbolic data. But the results of the analysis are not statistically significant in forecasting performance among the types of fuzzy composition. The contributions of this study are as follows. We propose another methodology that fuzzy relation and fuzzy composition could be applied for the similarity measurement between two symbolic data. That is the most important factor to build case-based reasoning model.

Key Words : Forecasting, Symbolic Data, Fuzzy Relation, Fuzzy Composition, Case-Based Reasoning, Similarity Matrix

저 자 소 개

이인호

이인호 교수는 웅지세무대학 회계정보과 조교수로 재직 중이다. 연세대학교에서 경영학사 및 석사, 생산관리 전공으로 박사학위를 취득하였다. 주요 연구분야는 인공지능 응용, 수요예측, 공급사슬관리 등이다.

신경식

신경식 교수는 이화여자대학교 경영대학 부교수 겸 지식시스템 연구센터장으로 재직 중이다. 연세대학교에서 경영학사, George Washington University에서 MBA, 한국과학기술원에서 경영정보학으로 박사학위를 취득하였다. 주요 연구분야는 지능형 의사결정지원시스템, 인공지능과 데이터마이닝, 지식기반 시스템 등이며 이와 관련한 다수의 연구논문 및 산학연구를 수행하였다. 최근에는 가상화에 따른 인간, 조직 및 사회 변화연구, 사회관계망 분석 등에 관련된 연구를 수행하고 있다.