한국지능정보시스템학회논문지 제11권 제2호 2005년 11월(pp.175~195)

효과적인 고객관계관리를 위한 사례기반추론 동시 최적화 모형

안현철
KAIST 테크노경영대학원 경영공학전공
(hcahn@kaist.ac.kr)

김경재
동국대학교 경영대학 정보관리학과
(kjkim@dongguk.edu)

한인구
KAIST 테크노경영대학원 경영공학전공
(ighan@kgsm.kaist.ac.kr)

사례기반추론(case-based reasoning)은 사례간 유사도를 평가하여 유사한 이웃사례를 찾아내고, 이웃사례의 결과를 이용하여 새로운 사례에 대한 예측결과를 생성하는 전통적인 인공지능기법 중 하나다. 이러한 사례기반추론이 최근 적용이 쉽고 간단하다는 장점과 모형의 갱신이 실시간으로 이루어진다는 점 등으로 인해, 온라인 환경에서의 고객관계관리를 위한 도구로 학계와 실무에서 주목을 받고 있다. 하지만, 전통적인 사례기반추론의 경우, 타 인공지능기법에 비해 정확도가 상대적으로 크게 떨어진다는 점이 종종 문제점으로 제기되어 왔다.

이에, 본 연구에서는 사례기반추론의 성과를 획기적으로 개선하기 위한 방법으로 유전자 알고리즘을 활용한 사례기반추론의 동시 최적화 모형을 제안하고자 한다. 본 연구가 제안하는 모형에서는 기존 연구에서 사례기반추론의 성과에 중대한 영향을 미치는 요소들로 제시된 바 있는 사례 특징변수의 상대적 가중치 선정(feature weighting)과 참조사례 선정(instance selection)을 유전자 알고리즘을 이용해 최적화함으로서, 사례간 유사도를 보다 정밀하게 도출하는 동시에 추론의 결과를 왜곡할 수 있는 오류사례의 영향을 최소화하고자 하였다. 제안모형의 유용성을 검증하기 위해, 본 연구에서는 국내 한 전문 인터넷 쇼핑몰의 구매예측모형 구축사례에 제안모형을 적용하여 그 성과를 살펴보았다. 그 결과, 제안모형이 지금까지 기존 연구에서 제안된 다른 사례기반추론 개선모형들은 물론, 로지스틱 회귀분석(LOGIT), 다중판별분석(MDA), 인공신경망(ANN), SVM 등 다른 인공지능 기법들에 비해서도 상대적으로 우수한 성과를 도출할 수 있음을 확인할 수 있었다.

논문접수일 : 2005년 11월 게재확정일 : 2005년 11월 교신저자 : 김경재

1. 서론

최근 인터넷, 데이터웨어하우스 등과 같은 정보기술의 발전은 고객의 교섭력을 강화시키는 동시에, 고객만족을 달성하기 위한 기업 간의 경쟁을 보다 심화시키고 있다. 이에 따라, 축적된 다양한 정보를 활용해 우수고객을 발굴하고, 앞선 정보통신기술을 이용해 이들을 유지, 관리하는 것을 의미하는 고객관계관리(customer relationship management)는 현대 기업들에게 있어서 점차 선택이 아닌 필수로 자리매김해 가고 있다. 고객관계관리를 실무적으로 구현하는 방법에는 다양한 것들이 있으나, 그 중 가장 기본적인 방법은 특정 고객이 어떤 상품 혹은 상품군을 구매할 것인지를 정확히 예측하는 것이다. 기업이 정확한 구매예측모형을 보유하고 있을 경우, 그 기업은 이를 1대1 마케팅이나 다이렉트 메일링(direct mailing), 혹은 전화나 이메일을 통한 프로모션의 수단으로 활용

하여 매출 증대를 가져올 수 있다. 실제로 데이터 마이닝 전문업체인 SAS Institute에 따르면 미국의 자동차 회사인 포드(Ford)나 보험회사 올스테이트(Allstate), 인터넷 꽃배달 회사인 1-800-flowers.com과 같은 선도기업들은 구매예측모형을 실제 마케팅에 적용해, 큰 성공을 거둔 것으로 보고되고 있다(SAS Institute, Success Stories, http://www.sas.com/success/).

이러한 구매예측모형을 구현하기 위한 방법으로는 지금까지 다양한 인공지능 혹은 데이터마이닝 기법들이 제안되어 왔다. 그 중에서, 최근 적용이 쉽고 간단하다는 장점과 모형의 갱신이 실시간으로 이루어져 온라인 환경에 적합하다는 특징으로 인해, 사례기반추론(case-based reasoning)이 학계와 실무 모두로부터 주목을 받고 있다. 사례기반추론은 과거의 경험을 재활용하여, 새로운 문제(혹은 사례)에 대한 예측결과를 생성하는 문제해결기법으로서, 복잡하거나 구조화되지 않은 의사결정문제에 효과적으로 적용될 수 있다는 특징으로 인해 지금까지 생산계획, 재무관리, 마케팅 등 다양한 경영분야에 적용되어 온 인공지능기법이다(Shin & Han, 1999; Kim & Han, 2001; Chiu, 2002; Yin et al., 2002; Chiu et al., 2003 참고).

하지만, 사례기반추론 방법론에는 설계자의 경험과 직관에 의존한 휴리스틱(heuristic)에 의해 설정되어야 하는 설계상의 요소들이 다수 내포되어 있기 때문에, 일반적으로 사례기반추론을 이용해 우수한 예측결과를 도출해 내는 것은 상당히 어렵다. 특히, 사례기반추론의 예측성과는 '유사사례검색(case retrieval) 단계에서 얼마나 주어진 문제와 유사한 사례들을 효과적으로 도출해서, 결합하는가?'에 크게 의존하는데, 전통적인 사례기반추론 방법론에서는 이러한 의문에 대한 구체적이고 명확한 해답이 제시되고 있지 않다. 때문에, 유사사례검색 단계에서의 적절한 특징변수 선정(feature selection), 참조사례 선정(instance selection), 그리고 사례간 유사도 계산시 적용되는 각 특징변수의 상대적 가중치 선정(feature weighting) 등과 같은 주제들이 지금까지 대표적인 사례기반추론 연구의 주요 연구주제로 자리매김해 왔다(Wang & Ishii, 1997; Shin & Han, 1999; Kim & Han, 2001; Chiu, 2002; Kim, 2004 참고).

최근에는 상기 사례기반추론의 설계 요소들을 동시에 최적화 하는 연구들이 연구자들의 관심을 끌고 있다. 이러한 연구의 대표적인 예로 특징변수의 선정과 참조사례의 선정을 동시에 최적화하고자 한 Kuncheva and Jain(1999)의 연구와 Rozsypal and Kubat(2003)의 연구를 들 수 있다. 그런데, 특징변수의 선정(feature selection)은 단순히 0 혹은 1 중 하나의 값으로 특징변수의 가중치를 부여하는 것과 동일하다는 점을 고려해 볼 때, 이는 0에서 1사이의 실수로 특징변수의 가중치를 부여하는 특징변수의 가중치 설정(feature weighting)의 특수한 한 경우라고 할 수 있다. 따라서, 특정변수의 가중치 설정과 참조사례의 선정을 동시에 최적화할 경우, 특징변수의 선정과 참조사례의 선정을 동시에 최적화하려고 한 기존 연구모형에 비해 더 우수한 성과를 가져올 수 있다. 그럼에도 불구하고, 아직까지 기존 문헌에서 이러한 시도를 연구한 예는 거의 찾아보기 어렵다.

이에 본 연구에서는 유전자 알고리즘(genetic algorithm)을 이용하여, 특징변수의 가중치 설정과 참조사례의 선정을 동시에 최적화 하는 새로운 사례기반추론 모형을 제안하고자 한다. 아울러, 본 연구에서는 제안모형을 실제 인터넷 쇼핑몰의 구매예측모형 구축사례에 다른 비교모형들과 함께 적용함으로서, 모형의 유용성을 실증적으로 검증해 보고자 하였다.

본 논문은 다음과 같이 구성된다. 우선 2장에서는 기존 문헌에 대한 간단한 고찰이 있을 것이며, 그 다음 3장에서는 본 연구의 제안모형에 대한 설명이 제공될 것이다. 4장에서는 제안모형을 검증하기 위한 실험설계에 관한 설명이 이루어질 것이며, 5장에서 실증 분석을 통한 실험결과가 종합적으로 제시될 것이다. 끝으로, 마지막 장에서는 본 연구의 의의와 한계점이 토의될 것이다.

2. 문헌 연구

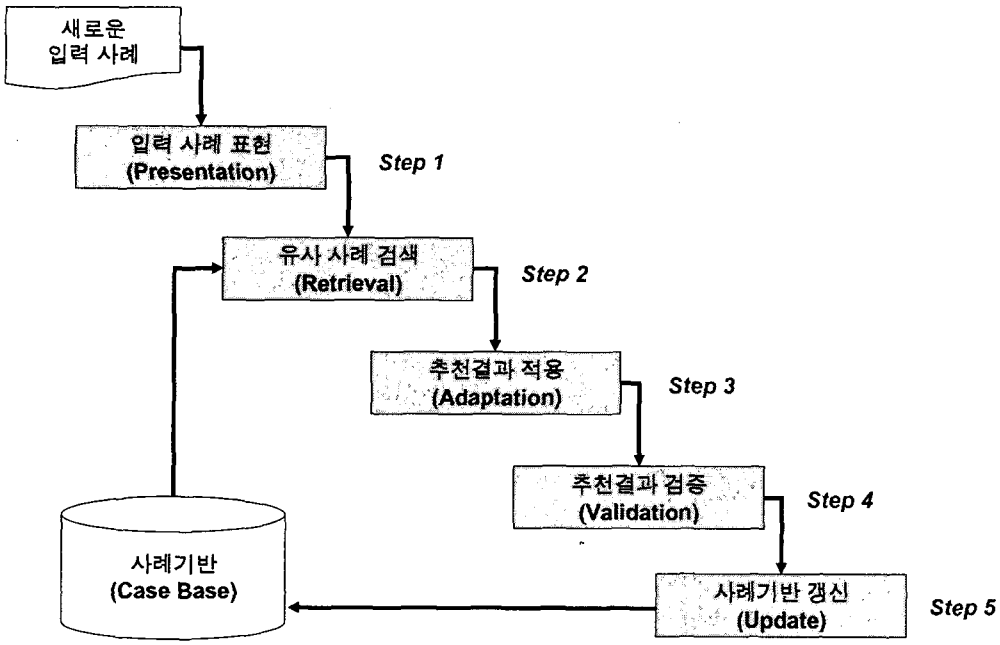
본 장에서는 우선 사례기반추론과 유전자알고리즘의 기본적인 개념에 대해 살펴보고, 이어 사례기반추론의 성과를 개선시키기 위해 시도된 다양한 기존 문헌들을 고찰하고자 한다. 아울러, 비록 그 수가 많지는 않지만, 사례기반추론의 여러 설계요소들을 동시에 최적화 하고자 한 기존연구들에 대해서도 그 내용을 살펴보고자 한다.

2.1 사례기반추론 및 유전자 알고리즘

사례기반추론의 기본 아이디어는 과거에 적용되었던 사례와 그 결과들을 참조하여, 새로운 문제나 사례에 대한 예측결과를 제시한다는 것이다. 따라서, 대부분의 인공지능기법들은 문제를 구성하는 특징변수들과 최종 결과 사이의 일반적인 관계(generalized relationship)를 도출하여, 이를 이용해 문제를 해결하는 방식을 주로 취하고 있으나, 사례기반추론은 과거 사례들 중 일부만 선택하여 그 사례들의 특정 지식을 사용해 문제를 해결하는 방식을 취하게 된다. 이러한 특징으로 인해, 사례기반추론은 축적된 데이터가 많지 않은 분야에도 적용이 가능하며, 복잡하거나 덜 구조화된 분야에도 쉽게 적용이 가능하다. 아울러, 사례기반추론은 새로운 사례가 축적되면 별도의 학습과정 없이 즉시 모형이 갱신된다는 특징도 갖고 있어, 즉각적인 반응을 요구하는 온라인 환경에서의 문제해결 및 예측방법론에 적합하다는 장점도 갖고 있다(Shin & Han, 1999). 이러한 장점들로 인해, 사례기반추론은 지금까지 인터넷 쇼핑몰의 상품 추천, 항공기의 교통 통제, 환자 진단, 심지어는 반도체 설계에 이르기까지 다양한 문제해결 분야에 적용되어 왔다(Turban & Aronson, 2001).

일반적인 사례기반추론의 절차는 [그림 1]에 제시되어 있는 것과 같이, 크게 5단계로 구성된다(Bradley, 1994). 그 중 2단계, 즉, '유사사례검색(case retrieval)' 단계에서 사례기반추론 시스템이 축적된 사례기반을 검색하여 주어진 사례 혹은 문제와 가장 유사한 사례를 찾아 추출하는 작업이 이루어지는데, 이 단계에서 추출된 유사사례의 내용이 최종 문제해결을 위한 예측결과에 중대한 영향을 미치기 때문에, 이 단계가 사례기반추론의 성과를 결정하는 가장 중요한 단계라고 할 수 있다. 이 단계에서 사례기반추론의 성과를 높이기 위해서는 사례간 유사도를 보다 정확하게 측정하는 것이 중요한데, 이를 위해서는 유사도 측정에 활용되는 사례의 특징변수에 대한 선택과 참조사례에 대한 선택, 그리고 특징변수들의 상대적 중요도에 대한 선정을 얼마나 효과적으로 잘 수행하는가가 매우 중요하다. 그 결과 지금까지 이러한 요소들을 최적화하기 위한 많은 연구가 이루어졌다 그리고, 이 중 상당수의 연구는 이러한 매개변수(parameter)들을 최적화하기 위한 방법론으로 유전자 알고리즘(genetic algorithm)을 활용하고 있다.

유전자 알고리즘은 자연 진화 이론에 기반을 둔 최적화 방법론으로서, 생물의 진화과정을 모사하



[그림 1] 사례기반추론의 일반적인 수행절차

여 적응적으로 탐색공간을 탐색하여 최적 또는 유사최적해를 찾아내는 전역 최적화 알고리즘이다. 유전자 알고리즘의 작동단계를 보다 구체적으로 설명하면 다음과 같다. 우선 유전자 알고리즘은 임의의 값을 가진 초기화된 개체집단을 생성한다. 즉, 전체 탐색공간 내에서 임의의 n개의 개체들을 선택하여 개체집단(population)을 형성하는 것이다. 그런 다음, 이렇게 생성된 개체집단이 문제에 얼마나 적합한지를 평가하게 되는데, 이른바 적합도 함수(fitness function)라는 함수를 이용해 각 개체의 적합도를 평가하게 된다. 이렇게 각 개체의 적합도가 평가되고 나면, 유전자 알고리즘은 평가된 개체집단을 확률적으로 선택(selection)하거나, 교배(crossover), 혹은 돌연변이(mutation) 등의 과정을 통해 이전 세대와 다른 새로운 개체들로 구성된 새로운 세대를 생성하게 된다. 이렇게 생성된 새로운 세대의 개체집단은 다시 적합도 함수에 의해 평가되며, 종결조건에 도달할 때까지 즉, 목표로 한 적합도 수준에 도달하거나 사전에 정해진 최대 진화수에 도달할 때까지 반복적으로 다음 세대를 생성하게 된다. 그렇게 하여, 유전자 알고리즘은 전체 생성된 세대 중에서, 가장 최적의 적합정도를 나타낸 개체를 최종적으로 선택하여, 그 결과를 전역 혹은 유사전역 최적해로 도출하게 된다 (Han & Kamber, 2001; Chiu, 2002).

2.2 특징변수 선정 및 상대적 가중치 설정의 최적화

사례기반추론에서 특징변수 선정은 유사도 측정과 관련성이 높은 특징변수들만 선택하고, 관련성이 떨어지거나 불필요한 변수들을 제거하는 일

련의 과정을 의미한다. 반면, 특징변수의 상대적 가중치 설정은 각 변수의 상대적 가중치를 할당하는 과정으로서, 전자의 경우 0 혹은 1의 가중치를 부여하는 것과 동일한 반면, 후자는 0에서 1사이의 실수를 가중치로 부여한다는 측면에서 더 상위의 개념이라고 할 수 있다. 이상의 두 가지 모두 사례기반추론의 성과에 영향을 미치는 중요한 요인들로서, 지금까지 많은 사례기반추론을 연구하는 많은 연구자들이 관련 연구를 진행해 왔다.

특징변수의 선정과 관련해서는 Siedlecki and Sklanski (1989)가 유전자 알고리즘에 기반한 변수선정 방법론을 제안했으며, Cardie (1993)의 연구는 의사결정나무 접근방법을 제안하였다. 아울러 Skalak (1994)과 Domingos (1997) 역시 각각 힐 클라이밍 알고리즘(hill climbing algorithm)과 군집화 기법을 특징변수 선정을 위한 방법론으로 제안하였다.

특징변수의 가중치 설정과 관련한 연구로는 우선 기계학습(machine learning) 문헌에서 자주 소개되고 있는 거리지표(distance metrics) 개념에 기반한 가중치 설정방법이 Wettschereck et al. (1997) 에 의해 제안되었으며, Kelly and Davis (1991)는 유전자 알고리즘에 기반한 가중치 설정 방법을 제안하였다. 특히, Kelly and Davis (1991)의 모형은 이후 다른 사례기반추론 연구에 활발하게 적용되었는데, 대표적인 예로는 기업채권의 등급 예측에 적용한 연구(Shin & Han, 1999), 기계 오작동 예측에 적용한 연구(Liao et al., 2000), 고객 구매 행동 예측에 적용한 연구(Chiu, 2002) 등을 들 수 있다. 아울러, 비록 사례기반추론은 아니지만, Kim and Shin (2000)은 유전자 알고리즘을 인공신경망의 특징변수 가중치 설정에 적용한 모형을 제안한 바 있다.

Cardie and Howe (1997) 그리고 Jarmulak et al. (2001) 은 특징변수의 선정과 가중치 부여를 혼합한 방법을 제안하였다. 전자의 연구에서는 의사결정나무 기법을 이용해 우선 특징변수를 선정하고, 이어 정보 이득(information gain) 값을 활용하여 선정된 변수의 상대적 가중치를 부여하는 방법을 제안하였다. 반면, Jarmulak et al. (2001)의 연구에서는 C4.5와 같은 의사결정나무 기법을 이용해 특징변수를 선정한 다음, 선정된 변수의 상대적 가중치를 유전자 알고리즘을 이용해 부여하는 방법을 제안하였다.

2.3 참조사례 선정의 최적화

사례기반추론에서 참조사례 선정은 전체 사례기반 중에서 유사도 측정과 관련해 대표성이 높은 사례들만 선택하고, 그 외 대표성이 떨어지거나 추론을 왜곡시킬 수 있는 사례는 사례기반에서 제외시키는 과정을 의미한다. 적절한 참조사례의 선정은 사례기반추론의 예측성과를 향상시킬 수 있을 뿐만 아니라, 동시에 사례기반추론의 탐색시간도 감소시킨다는 측면에서 매우 중요한 문제라 할 수 있으며, 그 결과 오랜 시간 동안 사례기반추론 분야의 주요 연구주제 중 하나로 인식되어 왔다.

지금까지 많은 연구자들이 적절한 참조사례를 선택하기 위한 다양한 방법들을 제안하였다. 우선 이 분야의 초기 선도 연구로 Hart (1968)가 농축된 유사 사례 결합 알고리즘(condensed nearest neighbor algorithm)을, 그리고 Wilson (1972)이 윌슨의 방법(Wilson's method)을 제안하였다. 이들 방법들은 간단한 정보이득 개념에 기반하여 구성되었기 때문에, 적용이 쉽고 간단하다는 장점이 있다. 참조사례 선정과 관련한 최근 연구들은 예측성과를 높이기 위해서, 보다 고급화된 수학적 기법이나 인공지능기법을 방법론으로 활용하고 있다.

예를 들어, Sanchez et al. (1997) 은 근접 그래프 접근법(proximity graph approach)을 제안하였으며, Lipowezky (1998)는 선형계획법(linear programming)에 기반한 참조사례 선정기법을 제안하였다. 아울러, Yan (1993)과 Huang et al. (2002)은 인공신경망을 활용한 참조사례 선정기법을 제안하였으며, Babu and Murty (2001)는 유전자 알고리즘에 기반한 기법을 제안하였다.

2.4 사례기반추론의 동시 최적화 연구

앞서 전술했듯이, 지금까지 사례기반추론의 각 요소들 - 즉, 적절한 특징변수의 선정, 상대적 가중치 부여, 적절한 참조사례의 선정 등 - 에 대한 최적화 연구가 활발하게 진행되어 왔다. 하지만 대부분의 기존 연구들은 각 요소들을 개별적으로 최적화 하려는 시도를 하였을 뿐, 이러한 요소들을 동시에 최적화 하려는 시도들은 그리 많지 않았다. 이러한 동시 최적화를 시도한 최초의 연구로는 Kuncheva and Jain (1999)의 연구를 들 수 있다. 그들은 유전자 알고리즘을 이용해 사례기반추론의 특징변수 선정과 참조사례 선정을 동시에 최적화 하는 모형을 제안하였으며, 그들의 제안모형을 기존에 제안된 특징변수 선정 기법 및 참조사례 선정 기법을 순차적으로 결합한 모형과 서로 비교하여, 더 우수한 성과를 보임을 실증적으로 제시하였다. 한편 이후 Rozsypal and Kubat (2003) 역시 유전자 알고리즘을 이용해 특징변수 선정과 참조사례 선정을 동시에 최적화 하는 모형을 제안하였다. 하지만 그들은 Kuncheva and Jain (1999)의 연구가 효율성 및 효과성 측면에서 설계상에 약점이 있음을 지적하고, 이를 개선하기 위해 새로운 유전자 코딩 방법과 적합도 함수를 제시하였다. 아울러, 실험을 통해 그들의 모형이 Kuncheva and Jain (1999)의 결과보다 더 우수한 성과를 보임을 함께 제시하였다.

그런데, 앞서 전술했듯이, 특징변수의 선정은 특징변수의 가중치를 0 혹은 1중 하나로 부여하는 경우로 해석할 수 있으므로, 특징변수의 가중치를 최적화 하는 것은 단순히 특징변수의 선정을 최적화 하는 경우보다 더 나은 결과를 가져올 수 있다. 같은 맥락에서, 특징변수의 가중치와 참조사례를 동시에 최적화 하는 모형은 Kuncheva and Jain (1999)이나 Rozsypal and Kubat (2003)의 연구모형처럼 특징변수의 선정과 참조사례의 선정을 동시에 최적화 하는 모형에 비해 더 우수한 성과를 가져올 수 있다. 이러한 관점에서, Yu et al. (2003)은 사례기반추론과 매우 유사한 원리를 갖고 있는 협업 필터링(collaborative filtering) 분야에 특징변수의 가중치와 참조사례를 함께 최적화 하는 모형을 제안하고, 그들의 모형이 다른 비교모형에 비해 더 우수한 결과를 제공할 수 있음을 실증적으로 분석하였다. 하지만, 그들의 연구는 최적화 기법으로 인공지능 기법이 아닌 정보-이론 접근법(information-theoretic approach)을 적용한 것으로서, 엄밀하게 분류하면 동시최적 기법이라기 보다는 두 요소의 최적화 기법을 순차적으로 결합한 방법이라고 하는 것이 더 타당하다. 아울러, 사례간 유사도를 상관계수를 이용해 계산하는 협업 필터링이 거리를 이용해 계산하는 사례기반추론과는 근본적으로 차이가 있음을 고려해 볼 때, 사례기반추론에 있어서 특징변수의 가중치와 참조사례의 선정을 동시에 최적화 하는 기존 연구는 아직까지 거의 이루어지지 않고 있다고 할 수 있다.

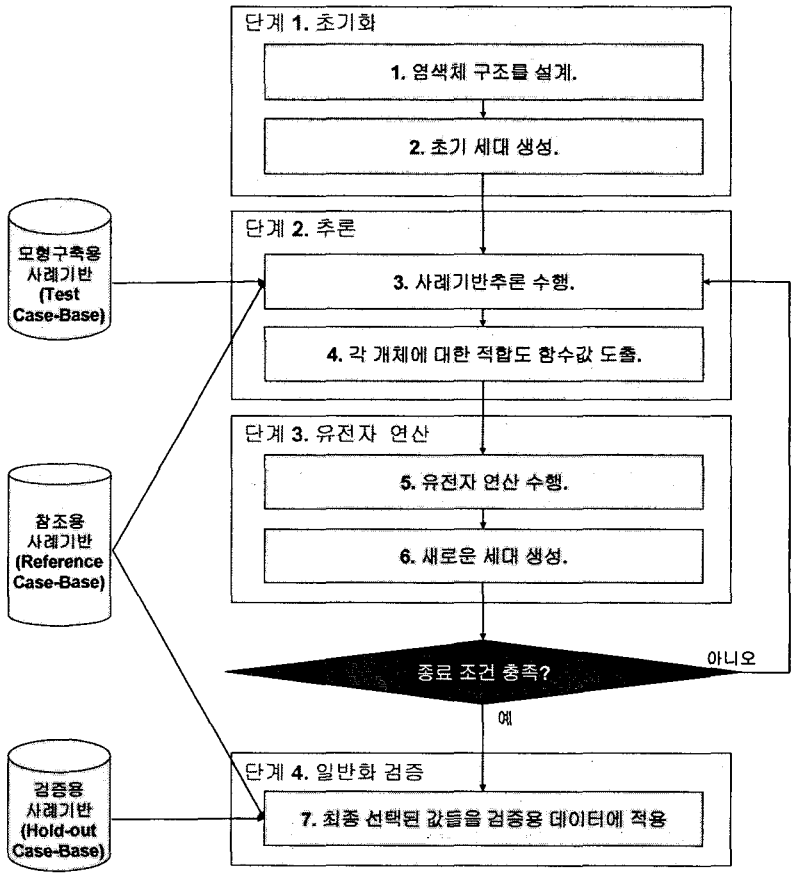
3. 연구 모형

본 연구에서는 전통적인 사례기반추론 시스템의 성과를 개선하기 위한 방법론으로, 유전자 알고리즘을 활용한 특징변수의 가중치 및 참조사례 선정의 동시 최적화 모형을 제안한다. 편의상 본 연구에서는 우리의 제안모형을 FWISCBR(Feature Weighting and Instance Selection simultaneously for CBR)으로 호칭하였다. FWISCBR의 전체적인 진행체계는 [그림 2]에 나타나 있는 구조도와 같다.

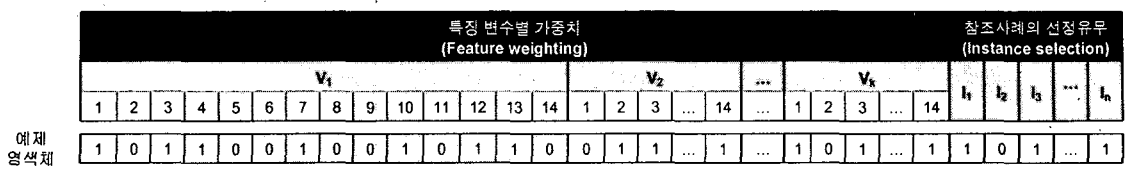
FWISCBR의 작동 원리에 대한 상세한 설명은 아래와 같다.

3.1 초기화

FWISCBR 시스템은 최적의 매개변수들(즉, 각 특징변수별 상대적 가중치 및 각 참조사례별 선택 유/무 여부)을 찾기 위해 전체 탐색 공간(search space)을 효율적으로 탐색해야 하는데, 이러한 탐색을 유전자 알고리즘을 통해 수행하기 위해서는, 먼저 2진 문자열 방식의 염색체(chromosome) 형



[그림 2] FWISCBR의 진행체계



[그림 3] FWISCBR의 염색체 구조

태로 최적화 하고자 하는 매개변수들을 코딩해야 한다. 이에 본 연구에서는 FWISCBR에 적용하기 위한 염색체의 구조를 설계하였는데, 그 결과가 [그림 3]에 제시되어 있다.

[그림 3]에 나타나 있는 바와 같이, 사례를 표현하는데 사용되는 전체 특징변수의 개수를 k개라고 하고, 전체 주어진 참조사례의 개수를 n개라고 할 때, FWISCBR의 염색체는 총 $14 \times k + n$ 비트의 길이를 갖게 된다. 그 과정과 원리를 자세히 살펴보면, 다음과 같다.

우선, 각 개별 특징변수에 대한 상대적 중요도의 경우, 0부터 1사이의 값을 갖는 소수로 부여하되 1/10,000의 정밀도를 갖게끔 설정하고자 하였다. 이 때, 이 값을 2진 비트값으로 표현하기 위해서는, $8192 = 2^{13} < 10000 \leq 2^{14} = 16384$ 이므로 각 개별변수당 14비트씩 요구됨을 알 수 있다 (Michalewicz, 1996). 이러한 14비트의 이진수는 다음의 식 (1)에 의해 0에서 1사이의 값을 갖는 십진 소수로 변환될 수 있다.

$$x' = \frac{x}{2^{14}} = \frac{x}{16384} \qquad (1)$$

(x: 각 변수의 가중치로 할당된 이진코드를 십진 정수로 변환한 값)

예를 들어, [그림 3]에 나와 있는 특징변수 1(V_1)에 대한 가중치는 $(10110010010110)_2$로 제시되어 있다. 이 값을 변환하고자 한다면, 우선 이 이진코드의 십진 정수값이 $(11414)_{10}$이 되므로, 이를 위 식 (1)에 대입하면 $\frac{11414}{16384} = 0.696655273 \approx 0.6967$ 이 됨을 최종적으로 확인할 수 있다.

한편, 참조사례 선정과 관련한 코드값은 사례가 선정될 경우 1을, 배제하고자 할 경우 0을 부여하는 형태로 설정하였다. 이렇게 참조사례의 선정에는 한 사례당 1비트밖에 요구되지 않기 때문에, 총 n개의 참조사례가 존재할 경우, 모두 n비트 길이의 염색체가 요구된다. 따라서, 이 두가지 요소를 모두 최적화 하기 위한 염색체는 $14 \times k + n$비트의 길이를 필요로 하게 되는 것이다.

이러한 염색체에 대한 설계가 끝나면, 개체집단(최적 매개변수들을 찾기 위한 일련의 염색체 집합)을 초기화하는 작업이 이루어지게 된다. 개체집단의 초기화는 일반적으로 염색체의 각 코드값을 무작위값으로 설정하는 것이 보통이다. 이렇게 초기화가 마무리되면, 유전자 알고리즘은 적합도 함수값을 극대화 하는 방향으로 개체집단에 대한 진화를 진행하게 된다. 본 연구에서는 사례기반추론 시스템의 최적 특징변수별 가중치와 참조사례를 설정하는 것이 최종 목적이므로, 모형구축용 데이터셋(test data set)에 대해서 가장 높은 예측 정확도를 보이는 변수별 가중치와 참조사례 선택결과를 찾고자 하였다. 즉, 본 연구에서는 유전자 알고리즘을 위한 적합도 함수로서 모형구축용 데이

터셋에 대한 예측 정확도를 활용하였다(Shin & Han, 1999; Kim, 2004). 모형구축용 데이터셋을 T라고 할 때, 적합도 함수 fr는 다음의 식 (2)와 같이 표현할 수 있다.

$$f_T = \frac{1}{n}\sum_{i=1}^{n} CA_i \qquad (2)$$

I_i에 대해 $PO_i = AO_i$를 만족할 경우, $CA_i = 1$, 그렇지 않을 경우, $CA_i = 0$

(CA_i: i번째 사례인 I_i에 대한 예측결과와 실제결과의 일치여부, PO_i: i번째 사례의 예측결과, AO_i: i번째 사례의 실제결과, T: $\{I_1, I_2, I_3, \Lambda, I_n\}$으로 구성된 모형구축용 데이터셋.

3.2 추론

2단계에서는 1단계에서 만들어진 후보 매개변수들(특징변수별 상대적 가중치 및 참조사례의 선정 결과)를 사례기반추론 시스템에 적용하여, 실제적인 추론을 진행하는 과정이 이루어진다. 본 연구에서는 사례간 유사도를 측정하기 위한 지표로 각 사례에 대한 가중평균 유클리드 거리(weighted average of Euclidean distance)를 사용하였다. 그리고, 유사사례탐색 방법으로는 입력사례와 가장 유사한 참조사례 중 총 3개를 찾아, 이들을 결합해 최종 예측결과를 생성하는 3-NN(3-nearest neighbor) 방법을 적용하였다. 이렇듯 k-NN의 k 값을 3으로 설정한 이유는, 1-NN부터 9-NN까지 홀수의 경우에 대하여 모든 경우를 실험해 본 결과, 3-NN이 가장 우수한 성과를 보임을 확인할 수 있었기 때문이다. 이상의 설정을 이용한 사례기반추론의 추론과정이 모두 끝나고 나면, 전체 모형구축용 사례집합에 대한 적합도 함수값(fr)이 새로 갱신된다.

3.3 유전자 연산

3단계에서는 유전자 알고리즘의 진화과정이 적합도 함수값을 최대화하는 방향으로 진행하게 되는데, 이러한 진화과정은 선택(selection), 교배(crossover), 돌연변이(mutation)등과 같은 유전자 연산의 적용을 통해 이루어진다. 이를 통해 일반적으로 보다 우수한 형질을 가지고 있는 새로운 세대(generation)가 생성되는데, FWISCBR은 이상의 2단계와 3단계의 작업을 종료조건이 만족될 때까지 계속해서 반복하게 된다.

3.4 일반화 검증

상기 3단계 과정을 통해 유전자 알고리즘을 이용한 학습이 모두 마무리되면, 마지막 4단계에서 FWISCBR 시스템은 최종 선정된 각 특징변수별 가중치와 참조사례의 선정 결과를 검증용 데이터셋에 적용해 봄으로서, 유전자 알고리즘을 통해 선정된 매개변수의 값들이 과연 일반화된 결과인지를 마지막으로 검증하게 된다. 자주 발생하는 현상을 아니지만, 유전자 알고리즘은 주어진 데이터에 대해서는 잘 적용되지만, 새로운 데이터에 대해서는 잘 적용되지 않는 이른바 과적합화(overfitting) 현상을 보이기도 한다. 이러한 결과의 과적합화를 확인, 검증하기 위해 연구 모형의 마지막 단계로 이 검증 과정이 포함된다.

4. 실험 설계

4.1 사례 소개

본 연구에서는 제안모형의 우수성을 검증하기 위해, 이를 국내의 한 인터넷 쇼핑몰의 구매예측 모형 구축 사례에 적용하였다. 본 연구의 대상이 된 쇼핑몰은 다이어트와 관련한 정보 제공, 커뮤니티 서비스, 쇼핑몰 등 원스톱(one-stop) 서비스를 제공하는 다이어트 전문 포털 사이트이다. 이러한 다이어트 사이트의 경우, 보다 정확하고 맞춤화된 서비스를 받기 위해 고객이 본인에 대한 상세한 정보를 입력해야만 하는데다, 대체로 사이트에 대한 이용 목적이 분명한 고객들이 주로 방문하기 때문에, 많은 고객들이 양적인 측면이나 질적인 측면에서 우수한 본인의 개인정보를 서비스 제공업체에 기꺼이 제공하는 경향이 있다. 따라서, 본 사례의 대상업체는 고객들에 대한 상당히 자세하고도 정확한 정보를 많이 보유하고 있는 상황이며, 이로 인해 이를 기업의 새로운 마케팅 기회로 활용하고자 하는 동기를 가지고 있다. 이에 본 연구에서는 제안된 모형을 활용해 본 대상업체를 위한 특정 상품의 구매예측모형을 구축함으로서, 전체 사이트 회원 중 예상 구매자를 발굴해 이들에 대한 일대일 마케팅을 수행할 수 있도록 지원하고자 하였다.

본 연구를 위해 업체로부터 수집된 데이터는 구매 및 비구매고객이 1:1의 비율로 혼합된 총 980건의 데이터였다. 종속변수는 대상업체에게 가장 높은 마진을 제공하는 다이어트 식품관련 상품의 구매여부 변수로서, 구매한 고객의 경우 1을, 구매하지 않은 고객의 경우 0을 값으로 부여하였다. 종속변수를 예측하기 위해 활용한 독립변수로는 회원가입시 입력되는 성별, 나이, 체중, 키 등 다이어트

<표 1> 실험에 사용된 특징변수

변수명	설명	측정값
AGE	나이	연속형 변수 (단위:년)
ADD0	거주지가 '서울'인지의 여부	0 : 아니다 / 1: 그렇다
ADD1	거주지가 '대도시'인지의 여부	0 : 아니다 / 1: 그렇다
OCCU0	직업이 '회사원'인지의 여부	0 : 아니다 / 1: 그렇다
OCCU2	직업이 '학생'인지의 여부	0 : 아니다 / 1: 그렇다
OCCU4	직업이 '자영업'인지의 여부	0 : 아니다 / 1: 그렇다
SEX	성별	0 : 남성 / 1: 여성
LOSS4	허벅지 부위의 살을 빼고 싶은지 여부	0 : 아니다 / 1: 그렇다
PUR0	'미용' 목적으로 다이어트를 하는지 여부	0 : 아니다 / 1: 그렇다
HEIGHT	키	연속형 변수 (단위:m)
BMI	비만도(Body Mass Index)는 다음 식에 의해 계산 $BMI\,(kg/m^2) = \frac{weight\,(kg)}{(height\,(m))^2}$	연속형 변수 (단위: kg/m^2)
E01	다이어트를 위한 '기능성 식품' 의 경험 여부	0 : 없다 / 1: 있다
E02	'다이어트 약물' 의 경험 여부	0 : 없다 / 1: 있다
E05	'원푸드 다이어트'의 경험 여부	0 : 없다 / 1: 있다

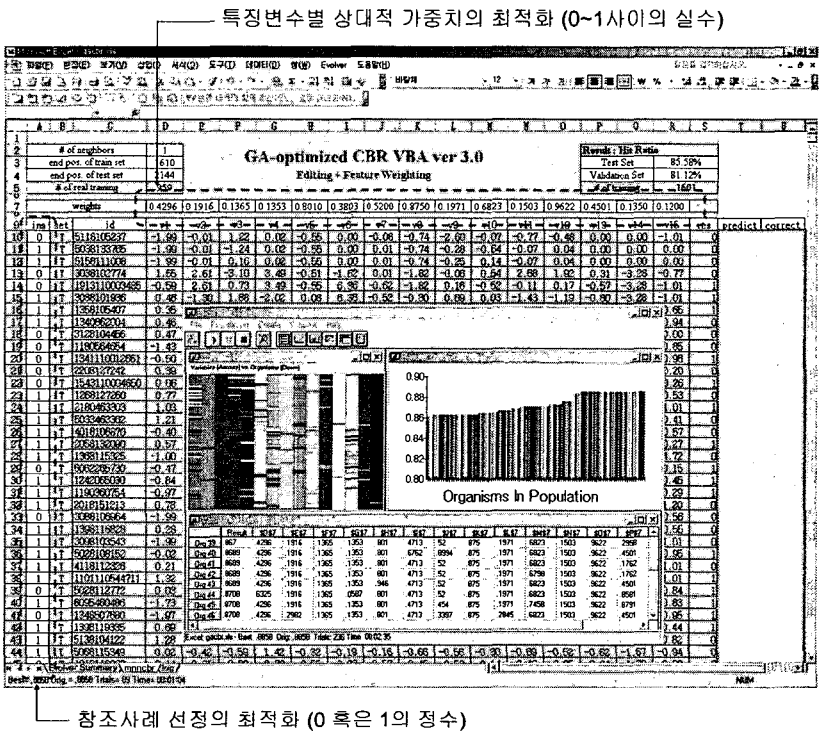
와 관련한 인구통계적인 변수들 중 총 46개가 수집되었다. 이러한 입력변수 중 종속변수의 예측에 관련성이 없는 변수를 사전에 제거하기 위해 이표본 t검정(two-sample t-test)과 카이제곱 검정(chi-square test)을 적용해 총 14개의 변수만 사례기반추론의 특징변수로 최종 선정하였다. <표 1>은 선택된 특징변수에 대한 상세한 정보를 설명하고 있다. 아울러 본 연구에서는 제안모형을 구축, 검증하기 위해 전체 수집된 데이터를 참조용, 모형구축용, 검증용 사례기반 등 총 3개의 그룹으로 구분하였다. 이 3가지 사례기반(데이터셋)은 각각 전체 데이터의 60% (588건), 20% (196건), 20% (196건)의 비중을 차지하도록 적절하게 배분되었다.

4.2 실험 설계 및 시스템 개발

유전자 알고리즘의 각종 통제변수들과 관련해서, 본 연구에서는 탐색해야 할 공간이 상당히 넓은 공간임을 고려해, 개체집단의 크기를 200개체로 설정하고, 교배율과 돌연변이율은 각각 70%와 10%로 설정하였다. 그리고, 종료조건으로는 총 4000회의 연산(즉, 20세대)을 시도하게끔 하였다.

아울러, 제안모형이 기존 문헌에 제안된 사례기반추론의 개선모형에 비해 얼마나 더 개선된 성과를 보여줄 수 있는지 검증하기 위해서, 동일한 데이터셋에 총 5개의 사례기반추론 기반 비교모형을 함께 적용해 보았다. 우선, COCBR(COnventional CBR)로 명명한 첫번째 비교모형은 아무런 최적화 과정을 수행하지 않고, 단순히 전통적인 사례기반추론을 수행하는 모형이다. 두번째 모형은 FSCBR(Feature Selection for CBR)로 명명하였는데, 이 모형에서는 유전자 알고리즘을 이용해 최적 특징변수를 선정하도록 설계하였다. 하지만 이 모형에서 참조사례의 선정이나 특징변수의 가중치 선정과 같은 작업은 이루어지지 않는다. 세번째 모형은 두번째 모형보다 약간 확장된 모형으로서, 여기서는 유전자 알고리즘을 이용해 특징변수의 최적 가중치를 설정하도록 하였다. 하지만, FWCBR (Feature Weighting for CBR)로 명명한 본 모형 역시 앞서 FSCBR과 마찬가지로 최적 참조사례의 선정과 관련해서는 고려하지 않는다. 기존 연구 중 Kelly and Davis (1991), Shin and Han (1999), Kim and Shin (2000) 및 Liao et al. (2000) 연구가 바로 이 세번째 모형과 동일한 모형을 제안한 바 있다. 네번째 모형은 앞의 두 모형과 반대로, 오직 참조사례의 선정만 유전자 알고리즘을 이용해 최적화하는 모형으로서, 여기서는 특징변수의 선정이나 가중치에 대한 최적화를 고려하지 않도록 설계하였다. 이러한 네번째 모형은 ISCBR (Instance Selection for CBR)이라 명명하였는데, 기존 연구 중 Babu and Murty (2001)가 유사한 모형을 제안한 바 있다. 마지막 다섯번째 비교모형은 유전자 알고리즘으로 최적 특징변수와 최적 참조사례를 동시에 최적화 하는 FISCBR (Feature and Instance Selection simultaneously for CBR)이라 명명된 모형이다. 이 모형은 Kuncheva and Jain (1999) 및 Rozsypal and Kubat (2003) 연구에서 제안된 모형과 동일하며, 다섯 비교모형 중 유일한 동시최적화 모형이라는 특징이 있다.

이상의 사례기반추론과 관련한 비교모형들과 본 연구의 제안모형인 FWISCBR은 Microsoft Excel 2003의 VBA (Visual Basic for Application)로 개발된 사례기반추론 소프트웨어와 유전자 알고리즘을 수행해 주는 상용 소프트웨어인 Palisade Software社의 Evolver Industrial Version 4.08를 결합하여 만든 실험용 프로토타입 시스템을 활용해 실험을 진행하였다. [그림 4]는 개발된 실험용 프로토타입 시스템의 작동 예시화



[그림 4] 실험용 프로토타입 시스템의 작동 화면 예시

면을 보여주고 있다.

추가적으로 본 연구의 제안모형인 FWISCBR이 다른 통계 및 인공지능기법에 비해서도 얼마나 우수한 성과를 보이는지 알아보기 위해, 우리는 추가적으로 로지스틱 회귀모형(logistic regression), 다중판별분석(multiple discriminant analysis), 인공신경망(artificial neural network), SVM(support vector machine) 등 총 4개의 비교모형을 확보된 데이터에 적용해 보았다. 로지스틱 회귀분석의 경우, 전진선택법(forward selection procedure)을 사용하였으며, 이 때, 단계별 변수입력 확률은 0.05로 설정하였다. 다중판별분석의 경우, Wilks' lambda를 활용한 입력변수의 단계별 선택방법을 활용하였는데, 이 때 변수의 입력 혹은 제거의 기준으로는 F값을 사용하였다. 이상의 통계기법들의 경우, SPSS for Windows 13.0을 이용해 실험을 수행하였다.

인공신경망에 대해서는 입력층과 출력층 사이에 은닉층을 1개 포함하는 3계층 역전파 네트워크(three-layer back propagation network)를 적용하였다. 인공신경망의 학습율과 모멘텀율은 각각 0.1씩 설정하였으며, 은닉층과 출력층의 노드들은 시그모이드 전이함수(sigmoid transfer function)를 사용하게끔 설계하였다. 은닉층의 노드수와 관련해서는 7, 14, 21, 28등 4가지 경우를 모두 대입해 보고 실험해 보았으며, 그 중에서 가장 우수한 결과를 보이는 은닉층의 노드수를 설정하고자 하였다. 아울러, 학습중지조건으로는 총 150차례 전

체 학습데이터에 대한 학습을 반복하게끔 설정하였다. 인공신경망과 관련한 실험은 상용 인공신경망 소프트웨어인 Neuroshell R4.0을 활용해 실험을 수행하였다.

SVM의 경우, 주어진 자료들을 고차원 공간으로 맵핑(mapping)하는 커널함수를 어떤 함수로 사용하는가에 따라 성과가 달라질 수 있기 때문에, 일반적으로 여러가지 커널함수를 모두 실험해 보고 가장 우수한 성과를 보이는 커널함수를 선정하는 것이 보통이다. 본 연구에서도 선형, 다항식, 그리고 가우시안(Gaussian) RBF 등 총 3개의 커널함수를 적용하여, 가장 우수한 성과를 보이는 결과를 최종적으로 선정하였는데, 이 3가지 커널함수에 대한 수식이 다음의 식 (3), (4), (5)에 제시되어 있다.

선형 커널함수: $K(\mathbf{x}_i,\mathbf{x}_j)= \mathbf{x}_i^T\mathbf{x}_j$ (3)

d차원의 다항식 커널함수:

$K(\mathbf{x}_i,\mathbf{x}_j)= (1+ \mathbf{x}_i^T\mathbf{x}_j)^d$ (4)

가우시안 RBF 커널함수:

$K(\mathbf{x}_i,\mathbf{x}_j)= \exp(-1/\delta^2(\mathbf{x}_i - \mathbf{x}_j)^2)$ (5)

Tay and Cao (2001)는 SVM의 성과를 결정짓는데 있어서, 상한계수 C나 d, δ^2와 같은 커널함수 내 매개변수들의 값에 대한 설정이 중대한 영향을 미칠 수 있음을 지적하였다. 만약 이러한 매개변수 값들이 적절하게 설정되지 않은 경우, SVM은 과적합화(overfitting)나 혹은 불충분적합(underfitting) 될 수 있기 때문이다. 때문에, 본 연구에서는 상기 매개변수들의 값을 다양하게 바꾸어가면서 실험하여, 가장 우수한 성과를 보이는 매개변수 값들을 최종적으로 선택하고자 하였다. SVM 실험을 위한 실험도구로는 공개 소프트웨어인 LIBSVM version 2.8 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/)을 활용하였다.

5. 실험 결과

우선 COCBR의 실험결과를 도출하기 위해, 우리는 k-NN의 k값으로 1에서 9사이의 홀수값을 모두 대입해 본 다음, 가장 우수한 성과를 보이는 k를 선택하고자 하였다. 그 결과, <표 2>에 제시되어 있는 바와 같이, 3-NN이 가장 우수한 성과를 보임을 확인할 수 있었다. 이에 본 연구에서는 FSCBR, FWCBR, ISCBR, FISCBR 그리고 FWISCBR 등 본 연구에서 적용한 모든 사례기반추론 관련 모형들에 3-NN을 적용하였다.

<표 3>은 FSCBR, FWCBR, ISCBR, FISCBR 및 본 연구의 제안모형인 FWISCBR의 모든 실험결과를 요약, 정리한 결과를 보여주고 있다. <표 3>에 따르면, ISCBR 이나 FISCBR 그리고 FWISCBR이 전체 참조사례 중에서 약 60-90%만 활용할 때, 최적의 예측결과를 보이고 있음을 확인할 수 있다. 이는 본 연구에 사용된 데이터셋이 쓸모없거나 잘못된 참조사례를 상당수 포함하고 있음을 보여주는 증거라고 할 수 있다.

한편 FWCBR, FISCBR 및 FWISCBR에서 도

<표 2> COCBR의 실험결과

k	1	3	5	7	9
검증용 데이터셋의 예측성과	52.04%	56.12%	55.61%	54.08%	53.57%

<표 3> 사례기반추론 최적화 모형의 최종 변수별 가중치 및 참조사례 선정 결과

변수명		FSCBR	FWCBR	ISCBR	FISCBR	FWISCBR
특징변수별 가중치						
	AGE	0	0.5678	1	1	0.3264
	ADD0	1	0.9035	1	0	0.0102
	ADD1	1	0.8532	1	1	0.3997
	OCCU0	0	0.9715	1	0	0.0135
	OCCU2	0	0.9313	1	1	0.8071
	OCCU4	0	0.7782	1	1	0.8974
	SEX	1	0.8097	1	0	0.0124
	LOSS4	1	1.0000	1	1	0.4335
	PUR0	0	0.8836	1	1	0.8093
	HEIGHT	0	0.6249	1	1	0.8475
	BMI	1	0.8093	1	0	0.3038
	E01	1	0.8836	1	1	0.7450
	E02	1	0.5022	1	1	0.2632
	E05	1	1.0000	1	1	0.6381
참조사례 선택결과						
	선택된 사례수	588	588	524	478	347
	비율 (%)	100%	100%	89.12%	81.29%	59.01%

출된 상대적 가중치들을 서로 비교해 볼 때, 우리는 FWISCBR의 가중치 패턴이 FWCBR보다는 FISCBR의 가중치 패턴과 더 유사함을 발견할 수 있다. 특히 ADD0, OCCU0, SEX의 가중치를 비교해 볼 때, FWISCBR과 FISCBR에서는 낮거나 0으로 도출된 반면, 상대적으로 FWCBR에서는 다소 높게 도출되었음을 알 수 있다. 이러한 현상으로 미루어 볼 때, 참조사례에 대한 선정결과는 특징변수의 선정 혹은 가중치 설정에 밀접한 영향을 미치는 것으로 추정되며, 이는 사례기반추론에서 '여러 요소들에 대한 동시 최적화'가 왜 중요한지를 보여주는 하나의 근거라고 할 수 있다.

<표 4>는 FWISCBR 및 다른 비교모형들의 평균 예측 정확도를 보여주고 있다. 이 결과를 보면, 본 연구의 제안모형인 FWISCBR이 64.29%의 정확도로 가장 우수한 성과를 보이고 있으며, 이어 ISCBR과 SVM이 63.27%의 정확도를 보여 두번째로 우수한 예측 성과를 보임을 알 수 있다. 특히, FWISCBR이 전통적인 COCBR에 비해 무려 11.73%나 더 우수한 성과를 보인다는 점은 제안모형이 사례기반추론의 예측 성과를 상당한 수준으로 개선시킬 수 있다는 점을 시사하고 있다. 아울러, 본 연구에서는 ISCBR이 FSCBR 및 FWCBR에 비해 더 우수한 성과를 보이고 있음을 확인할 수 있는데, 이는 본 연구에서 적용한 사례의 경우, 적절한 참조사례에 대한 선택이 적절한 특징변수의 선택이나 가중치 설정에 비해 성과개선에 더 큰 영향을 미친다고 해석할 수 있다.

<표 4> 각 모형별 평균 예측정확도

모형	학습(참조)용 데이터셋	모형구축용 데이터셋	검증용 데이터셋	비 고
LOGIT	63.30%		62.76%	전진선택법
MDA	63.10%		62.76%	단계별 선택법 (Wilks' lambda)
ANN	68.37%	66.84%	61.73%	은닉층 내 최적 노드수: 14
SVM	65.82%		63.27%	커널함수: 가우시안 RBF C=1 and δ^2=75
COCBR		62.24%	56.12%	k of k-NN = 3
FSCBR		62.76%	60.20%	
FWCBR		65.82%	61.73%	
ISCBR		66.33%	63.27%	
FISCBR		70.92%	64.29%	
FWISCBR		72.96%	67.86%	

상기 <표 4>에서 보여진 FWISCBR과 다른 비교모형들간의 예측성과 차이가 통계적으로 유의한지를 검증하기 위해, 본 연구에서는 two-sample test for proportions를 수행하였다. 이 방법을 이용하면 0에서 1사이의 확률로 측정되는 두 모형의 성과 차이가 통계적으로 유의한지에 대해 검증이 가능하다(Harnett & Soni, 1991). 본 연구에서는 이 검정을 단측검정의 형태로 수행하였는데, 이 경우 p_k가 k번째 모형의 예측성과라고 할 때, 귀무가설은 $H_0 : p_i - p_j = 0$ (i=1,…,8, j=2,…,9), 대립가설은 $H_a : p_i - p_j > 0$ (i=1,…,8, j=2,…,9)가 된다. <표 5>는 각 모형별 상대비교 검정결과를 Z값 형태로 제시하고 있다. <표 5>에 제시된 바와 같이, FWISCBR는 COCBR보다 99% 신뢰수준 하에서, 그리고 FSCBR보다는 95% 신뢰

<표 5> Two-sample test for proportions 수행 결과

	LOGIT	ANN	SVM	COCBR	FSCBR	FWCBR	ISCBR	FISCBR	FWISCBR
MDA	0.0000	0.2084	0.1046	1.3372**	0.5189	0.2084	0.1046	0.3148	1.0611*
LOGIT		0.2084	0.1046	1.3372**	0.5189	0.2084	0.1046	0.3148	1.0611*
ANN			0.3130	1.1293*	0.3106	0.0000	0.3130	0.5231	1.2690*
SVM				1.4416**	0.6235	0.3130	0.0000	0.2102	0.9566*
COCBR					0.8191	1.1293*	1.4416**	1.6510**	2.3932***
FSCBR						0.3106	0.6235	0.8335	1.5787**
FWCBR							0.3130	0.5231	1.2690*
ISCBR								0.2102	0.9566*
FISCBR									0.7467

(*: 90% 신뢰수준, **: 95% 신뢰수준, ***: 99% 신뢰수준)

수준 하에서 통계적으로 유의한 성과차이를 보임을 확인할 수 있다. 또한 FWISCBR은 FISCBR을 제외한 모든 비교모형에 대해 90% 신뢰수준 하에서 통계적으로 유의한 성과차이를 보이고 있음을 알 수 있다.

6. 연구의 의의 및 한계점

본 연구에서는 일반적인 사례기반추론 시스템의 성과를 개선하기 위한 대안으로 FWISCBR이라 명명한 유전자 알고리즘 기반의 새로운 사례기반추론 방법론을 제안하였다. 본 연구에서는 유전자 알고리즘을 이용해 사례간 유사도 측정에 사용되는 특징변수의 가중치 및 참조사례의 선정을 동시에 최적화 함으로서, 사례기반추론의 가장 큰 한계점으로 지적되어 온 낮은 예측력의 문제를 극복하고자 하였다. 아울러 제안된 연구모형을 검증하기 위해, 실제 고객관계관리 사례에 모형을 적용해 그 성과를 살펴보았으며, 그 결과 제안모형인 FWISCBR이 COCBR, FSCBR, FWCBR, ISCBR, FISCBR 등 기존 연구를 통해 제안된 다른 모든 사례기반추론 관련 최적화 모형들에 비해 더 우수한 성과를 보이고 있음을 확인할 수 있었다. 뿐만 아니라, 로지스틱 회귀분석, 다중판별분석, 인공신경망, SVM 등 다른 통계기법이나 인공지능기법들에 비해서도 FWISCBR이 더 우수한 성과를 보인다는 사실도 함께 확인할 수 있었다.

앞서 <표 4>에서 볼 수 있듯이, 본 연구에서 수행한 실험에서는 유전자 알고리즘을 이용해 적절한 참조사례를 선택한 ISCBR이나 FISCBR, 그리고 FWISCBR이 타 모형들에 비해 유독 우수한 성과를 보였음을 알 수 있다. 아울러, 전체 학습데이터 중에서 이른바 서포트 벡터(support vector)에 해당되는 데이터만으로 학습을 진행하는 SVM 기법 역시 타 기법들에 비해 우수한 성과를 보이고 있음을 확인할 수 있다. 이러한 결과는 본 연구에 사용한 참조용 데이터셋이 상대적으로 많은 불필요한 사례나 오류가 포함된 사례를 다소 많이 갖고 있는 것으로 해석될 수 있으며, 이 경우 적절한 참조사례를 선택하는 것이 성과개선에 매우 큰 영향을 미칠 수 있음도 함께 확인할 수 있다.

하지만, 본 연구는 다음과 같은 몇 가지 한계점을 갖고 있다. 우선 첫째로, FWISCBR은 최적 매개변수들, 즉 최적 특징변수별 가중치와 참조사례의 선정결과를 도출하는데 많은 연산을 필요로 한다는 점을 들 수 있다. 일반적으로 사례기반추론은 문제가 주어져야 비로소 학습을 시작하게 되는 게으른 학습방법 중 하나로서, 사례가 많아지면 많아질수록 그 연산량이 기하급수적으로 증가한다는 단점이 있다. 그런데 FWISCBR은 이러한 사례기반추론을 유전자 알고리즘의 진화과정에 따라 지속적으로 반복 수행하여야 하기 때문에, 그 요구되는 연산량이 상당히 크다고 할 수 있다. 때문에, 향후 FWISCBR을 보다 효율화 하는 연구가 이루어져야 할 것이다.

두 번째 한계점으로는, 본 연구에서 최적화하고자 한 2가지 요소 외에 다른 사례기반추론의 매개변수들도 함께 최적화함으로서 더 성과를 개선시킬 수 있을 것이라는 점을 들 수 있다. 본 연구에서는 사례기반추론의 각 특징변수별 가중치와 참조사례의 선택만을 유전자 알고리즘을 이용해 최적화 하고 있지만, 그 외에도 k-NN의 k값(즉, 결합할 최적 유사사례의 개수)이나 각 참조사례별 상대적 가중치 같은 매개변수들도 사례기반추론의 성과를 개선시킬 수 있는 설정요소가 될 수 있다 (이훈영과 박기남, 1996; Ahn et al., 2003). 따라서 FWISCBR에서 제안된 요소들과 더불어, 상기

다른 요소들까지 모두 최적화 하는 전체적인 동시 최적화 모형을 구축한다면, 기존의 사례기반추론의 성과를 더 크게 개선할 수 있을 것으로 사료된다.

마지막으로 비록 본 연구의 검증 결과가 전체 확보된 데이터셋 중에서 검증용 데이터셋을 따로 분리해 별도의 일반화 검증과정을 거치기는 하였지만, 단 한가지 경우의 사례 적용을 통해서만 제안모형의 유용성을 검증했다는 점을 한계점으로 들 수 있다. 특히 본 연구의 검증에 활용된 데이터는 상대적으로 규모가 작고, 오류사례를 다수 포함하고 있는 것으로 추정되는데, 이로 인해 각 모형의 성과차이를 극명하게 비교하는데 다소 어려움이 있었다. 따라서, 본 연구에서 제안한 모형을 추후 다른 분야에 적용함으로서, 모형의 일반화 가능성을 추가로 검증해 볼 필요가 있다.

참고문헌

[1] 이훈영, 박기남, "사례기반예측시스템의 정확한 예측을 위한 최적 결합 사례개수결정방법에 관한 연구", *경영학연구*, 27권 5호(1999), 1239-1252.

[2] Ahn, H., K.-j. Kim and I. Han, "Determining the optimal number of cases to combine in an effective case-based reasoning system using genetic algorithms", *Proceedings of International Conference of Korea Intelligent Information Systems Society 2003 (ICKIISS 2003)*, 178-184, 2003.

[3] Babu, T.R. and M.N. Murty, "Comparison of genetic algorithm based prototype selection schemes", *Pattern Recognition*, Vol. 34, No. 2(2001), 523-525.

[4] Bradley, P., "Case-based reasoning: Business applications", *Communication of the ACM*, Vol. 37, No. 3(1994), 40-43.

[5] Cardie, C., "Using decision trees to improve case-based learning", *Proceedings of the Tenth International Conference on Machine Learning*, San Francisco, CA, 25-32, 1993.

[6] Cardie, C. and N. Howe, "Improving minority class prediction using case-specific feature weights", *Proceedings of the Fourteenth International Conference on Machine Learning*, San Francisco, CA, 57-65, 1997.

[7] Chiu, C., "A case-based customer classification approach for direct marketing", *Expert Systems with Applications*, Vol. 22, No. 2(2002), 163-168.

[8] Chiu, C., P.C. Chang and N.H. Chiu, "A case-based expert support system for due-date assignment in a water fabrication factory", *Journal of Intelligent Manufacturing*, Vol. 14, No. 3-4(2003), 287-296.

[9] Domingos, P., "Context-sensitive feature selection for lazy learners", *Artificial Intelligence Review*, Vol. 11, No. 1-5(1997), 227-253.

[10] Han, J. and M. Kamber, *Datamining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, CA, 2001.

[11] Harnett, D.L. and A.K. Soni, *Statistical methods for business and economics*, Addison-Wesley, Massachusetts, MA, 1991.

[12] Hart, P.E., "The condensed nearest neighbor rule", *IEEE Transactions on Information Theory*, Vol. 14, No. 3(1968), 515-516.

[13] Huang, Y.S., C.C. Chiang, J.W. Shieh and E. Grimson, "Prototype optimization for nearest-neighbor classification", *Pattern Recognition*, Vol. 35, No. 6(2002), 1237-1245.

[14] Jarmulak, J., S. Craw and R. Rowe, "Self-optimizing CBR Retrieval", *Proceedings of the Twelfth IEEE International Conference on Tools with Artificial Intelligence*, Vancouver, Canada, 376-383, 2000.

[15] Kelly, J.D.J. and L. Davis, "Hybridizing the genetic algorithm and the k nearest neighbors classification algorithm", *Proceedings of the Fourth International Conference on Genetic Algorithms*, San Diego, CA, 377-383, 1991.

[16] Kim, K., "Toward global optimization of case-based reasoning systems for financial forecasting", *Applied Intelligence*, Vol. 21, No. 3(2004), 239-249.

[17] Kim, K. and I. Han, "Maintaining case-based reasoning systems using a genetic algorithms approach", *Expert Systems with Applications*, Vol. 21, No. 3(2001), 139-145.

[18] Kim, S.H. and S.W. Shin, "Identifying the impact of decision variables for nonlinear classification tasks", *Expert Systems with Applications*, Vol. 18, No. 3(2000), 201-214.

[19] Kuncheva, L.I. and L.C. Jain, "Nearest neighbor classifier: Simultaneous editing and feature selection", *Pattern Recognition Letters*, Vol. 20, No. 11-13(1999), 1149-1156.

[20] Liao, T.W., Z.M. Zhang and C.R. Mount, "A case-based reasoning system for identifying failure mechanisms", *Engineering Applications of Artificial Intelligence*, Vol. 13, No. 2(2000), 199-213.

[21] Lipowezky, U., "Selection of the optimal prototype subset for 1-NN classification", *Pattern Recognition Letters*, Vol. 19, No. 10(1998), 907-918.

[22] Michalewicz, Zb., *Genetic Algorithms + Data Structures = Evolution Programs (3rd edition)*, Springer-Verlag, Berlin, 1996.

[23] Rozsypal, A. and M. Kubat, "Selecting representative examples and attributes by a genetic algorithm", *Intelligent Data Analysis*, Vol. 7, No. 4(2003), 291-304.

[24] Sanchez, J.S., F. Pla and F.J. Ferri, "Prototype selection for the nearest neighbour rule through proximity graphs", *Pattern Recognition Letters*, Vol. 18, No. 6(1997), 507-513.

[25] Shin, K.S., and I. Han, "Case-based reasoning supported by genetic algorithms for corporate bond rating", *Expert Systems with Applications*, Vol. 16, No. 2(1999), 85-95.

[26] Siedlecki, W. and J. Sklanski, "A note on genetic algorithms for large-scale feature selection", *Pattern Recognition Letters*, Vol. 10, No. 5(1989), 335-347.

[27] Skalak, D.B., "Prototype and feature selection by sampling and random mutation hill climbing algorithms", *Proceedings of the Eleventh International Conference on Machine Learning*, New Jersey, NJ, 293-301, 1994.

[28] Tay, F.E.H. and L.J. Cao, "Application of support vector machines in financial time series forecasting", *Omega*, Vol. 29, No. 4(2001), 309-317.

[29] Turban, E. and J.E. Aronson, *Decision support systems and intelligent systems*

(6th edition), Prentice-Hall, Upper Saddle River, NJ, 2001.

[30] Wang, Y., and N. Ishii, "A method of similarity metrics for structured representations", *Expert Systems with Applications*, Vol. 12, No. 1(1997), 89-100.

[31] Wettschereck, D., D.W. Aha and T. Mohri, "A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms", *Artificial Intelligence Review*, Vol. 11, No. 1-5(1997), 273-314.

[32] Wilson, D.L., "Asymptotic properties of nearest neighbor rules using edited data", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 2, No. 3(1972), 408-421.

[33] Yan, H., "Prototype optimization for nearest neighbor classifier using a two-layer perceptron", *Pattern Recognition*, Vol. 26, No. 2(1993), 317-324.

[34] Yin, W.J., M. Liu and C. Wu, "A genetic learning approach with case-based memory for job-shop scheduling problems", *Proceedings of the First International Conference on Machine Learning and Cybermetics*, Beijing, China, 1683-1687, 2002.

[35] Yu, K., X. Xu, M. Ester and H.-P. Kriegel, "Feature weighting and instance selection for collaborative filtering: an information-theoretic approach", *Knowledge and Information Systems*, Vol. 5, No. 2(2003), 201-224.