한국지능정보시스템학회논문지 제8권 제2호 2002년 12월(pp.117~137)

개별 속성의 선택 및 제거효과 순위를 이용한 사례기반 추론의 속성 선정*

이재식
아주대학교 경영대학
(leejsk@ajou.ac.kr)
이혁희
(주) 엑설루션 컨설팅
(jyppa@hanmail.net)

사례기반 추론은 과거의 사례를 기반으로 새로운 사례에 대한 답을 제시하는 기계학습의 한 분야이다. 과거의 사례는 일정한 형식으로 사례 베이스에 저장되는데, 저장의 형식을 결정하는 것이 속성이다. 속성은 사례의 특징을 가장 잘 표현할 수 있는 것들로 구성되며, 속성값간의 유사도 도출을 통해서 유사 사례를 검색하게 된다. 따라서, 사례기반 추론은 사용되는 속성에 따라서 성능이 달라지게 된다. 본 연구에서는 먼저 속성을 하나씩만 사용하여 사례기반 추론을 수행하여 각 속성의 선택효과를 측정하고, 하나씩만 제거하고 사례기반 추론을 수행하여 각 속성의 제거효과를 측정하였다. 이 측정치들을 근거로 속성의 부분집합을 구성하여 사례기반 추론을 구현한 결과, 속성을 전부 사용했을 때보다 성능과 효율성이 우수한 사례기반 추론 시스템을 구축할 수 있었다.

Key words: 속성 선정 방법, 사례기반 추론

1. 서론

기계 학습은 크게 사전학습(Eager learning)과 사후학습(Lazy learning)의 두 가지로 나눌 수 있다. 사전학습 방식은 문제를 풀기 위한 준비 즉, 학습과정이 명백하게 분리되어 있으며 학습을 통한 일반화 과정이 완전히 끝난 뒤에 새로운 사례에 대한 문제를 풀게 된다. 대표적인 기법으로는 인공신경망과 의사결정 나무가 있다. 반면에 사후학습은 학습과정이 명백하게 분리되어 있지 않으며, 새로운 사례를 접하고 나서야 비로소 학습이 진행되어 문제를 해결한다.

사후학습의 대표적인 기법으로는 사례기반 추론(Case-Based Reasoning)이 있으며, 그 장점은 다음과 같다. 사례를 저장함으로써 간단하게 새로운 지식이 추가되며, 그 구조가 이해하기 쉽고 단순하다. 또한 수치형 속성과 범주형(Categorical) 속성 모두를 사용할 수 있고, 제시된 결과에 대한 설명이 가능하며, 복잡한 양상을 갖는 문제 영역에서 비교적 적은 정보만을 이용하여 문제 해결이 가능하다. 반면에 몇 가지 단점이 있는데, 사례를 저장하기 위한 공간이 많이 필요하고, 일반화를 위한 학습 과

* 이 연구는 2000년도 두뇌한국 21 연구비에 의해서 지원되었음.

정과 문제 해결이 동시에 일어나기 때문에 많은 시간이 소요된다는 것이다. 특히 가장 치명적인 단점은 사례를 구성하고 있는 속성이 적절하지 못한 경우에는 사례기반추론 시스템의 성능이 크게 저하된다는 것이다. 이러한 단점을 해결하기 위한 방안은 적절하지 못한 속성을 미리 파악하여 이를 제거하는 것인데, 그 노력의 일환으로 속성 선정과 속성에 대한 가중치 부여에 대한 연구가 수행되어 왔다[Aha, 1998]. 속성 선정은 미리 설정한 값 이하의 가중치를 갖는 속성을 사용하지 않음으로써 부적절한 속성을 제거하는 방법으로서 광의의 속성 가중치 부여에 포함된다.

본 연구의 목적은 사례기반추론의 효과성과 효율성을 동시에 높일 수 있는 속성 선정 방법의 수립과 검증에 있다. 속성을 전부 사용했을 때의 사례기반 추론 시스템의 적중률과 속성의 부분집합을 사용했을 때의 사례기반 추론 시스템의 적중률을 비교함으로써 본 연구에서 개발한 속성 선정 방법의 효과성을 검증하게 된다. 모든 속성을 사용하지 않고 속성 부분집합을 사용하게 되면 그만큼 사례기반 추론의 수행 속도는 빨라진다. 이로 인해서 효율성을 갖게 된다.

본 논문은 다음과 같이 구성되어 있다. 제 2절에서는 본 연구에 사용된 사례기반 추론의 구조에 대해서 기술한다. 제 3절에서는 속성 선정 방법에 대한 선행 연구와 평가방법에 대하여 기술한다. 제 4절에서는 본 연구에서 개발한 속성 선정 방법에 대하여 설명하고, 제 5절에서는 UCI(University of California at Irvine)의 기계학습 데이터 저장소[Blake *et al.*, 1998]에서 획득한 연구용 데이터 3개와 실제 기업의 데이터 1개를 사용하여 본 연구에서 개발한 속성 선정 방법의 유용성을 검증한다. 마지막으로 제 6절에서는 결론 및 향후 연구 과제에 대해서 논의한다.

2. 본 연구에 사용된 사례기반 추론 시스템

사례기반 추론은 인간이 과거 경험에 비추어 사물을 인식한다는 점에 착안하여 나타난 인공지능의 한 분야이며, 과거에 문제를 풀었던 경험을 새로운 문제에 적용하여 해결책을 제시하는 시스템으로 정의할 수 있다[Riesbeck and Schank, 1989; Kolodner, 1993]. 여러 산업 부분의 특정 상황에 맞게 사례기반 추론이 사용되고 있으며[Allen, 1994; Watson 1997], 실제로 파산예측[Jo and Lee, 1997], 고객 서비스[Lee and Xon, 1996], 주가예측[Kim, 1997] 등에 적용된 연구가 있다.

사례기반 추론 시스템은 그 활용 영역에 따라서 다양한 구조를 가지며, 성능을 높이기 위한 노력 여하에 따라서 매우 많은 변형이 나타날 수 있다. 하지만, 모든 경우의 수를 다 고려할 수 없으므로 본 연구에서는 가장 기본적인 사례기반 추론 시스템을 사용한다. 또한 본 연구의 목적은 효과적이고 효율적인 속성 선정 전략의 수립과 그 검증에 있으므로 사례기반 추론 시스템 구축의 과정에서 속성 가중치 설정, 색인부여, 적응규칙과 같은 부분에 대해서는 다양한 변형을 고려하지 않기로 한다. <표

2-1>은 본 연구에서 사용할 기본적인 사례기반 추론 시스템의 특성을 보여주고 있다.

<표 2-1>에서 제시한 유사도 산정 방식을 좀더 자세히 설명하면 (식 2.1), (식 2.2)와 같다.

유사도 산정에 있어서, 연속형 속성의 경우에는 단순히 거리를 측정하는 방식을 사용하였으며, 범주형 속성의 경우에는 완전 일치되는 값을 가지면 1점, 그렇지 않으면 0.5점을 부여하였다. 일치되지 않는 속성에 0.5점을 부여한 이유는, 만일 미세한 차이가 나도 0점을 부여하게 되면 미묘한 차이로 인해서 유사도 점수 산정에 큰 격차를 나타낼 수 있기 때문이다. 유사도 산정 후, 사례를 검색해 오는 단계에서는 k개의 최근접이웃 방법을 사용하였으며, k값으로 5를 사용하였다. 본 연구에서 사용한 데이터는 목표속성이 2개 혹은 여러 개의 값을 가지는 분류 문제이다. 적응(Adaptation) 단계에서는 검색된 사례의 목표속성의 값들 중에서 3개 이상이 새로운 사례의 목표속성의 값과 일치하면 적중한 것으로 정의하는 단순한 투표(Voting) 방법을 사용하였다.

3. 속성 선정 방법

3.1 속성 선정 과정

전통적인 속성 선정 과정은 생성단계, 평가단계, 타당성 검증단계의 세 단계로 구성된다. 생성단계는 탐색 과정이라고 할 수 있는데, 많은 속성들 중에서 어떤 속성들을 가지고 속성 부분집합을 구성하여 성능을 측정할 것인가를 결정하는 단계이다. 평가단계는 생성단계에서 선택한 속성들이 과연 적합한가의 여부를 평가하는 것으로서 거리측정, 상관관계측정, 정

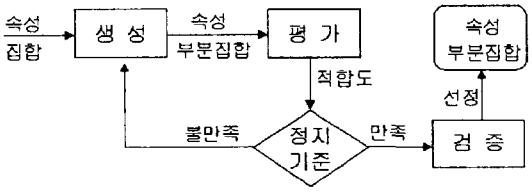
<표 2-1> 기본적인 사례기반 추론 시스템의 특성

단 계	사 용 방 식
사례저장	1개의 데이터베이스 테이블 (Flat File)
속성 가중치 설정	모두 1
유사도 산정	연속형 속성 : 거리(Distance) 범주형 속성 : 완전일치법(Exact Matching)
검색(Retrieval)	k개의 최근접이웃(k-Nearest Neighbors)
적응(Adaptation)	투표(Voting)

$$\text{연속형 속성: 유사도} = 1 - \frac{|\text{새로운 사례의 속성값} - \text{과거 사례의 속성값}|}{\text{해당 속성의 최대값}} \quad \text{(식 2.1)}$$

$$\text{범주형 속성: 유사도} = \begin{cases} 1, & \text{(새로운 사례의 속성값 = 과거 사례의 속성값)인 경우} \\ 0.5, & \text{나머지 경우} \end{cases} \quad \text{(식 2.2)}$$

보측정 등과 같은 평가 방법들이 사용된다. 이러한 평가는 미리 설정한 회수만큼 반복되거나 평가 함수에 비추어 타당한 결과가 나타나면 멈추게 된다. 다음 단계는 검증으로서 최종적으로 선정된 속성의 부분집합이 데이터의 특성을 제대로 반영하고 있으며, 타당한 것인가의 여부를 검증하는 단계이다. 다른 알고리즘을 사용하여 테스트하거나, 인공적인 데이터를 사용하여 결과값을 도출한 후 비교를 통해서 평가를 하게 된다. <그림 3-1>은 이와 같은 속성 선정 과정을 도식화한 것이다[Dash and Liu, 1997].



<그림 3-73> 속성 선정 과정

속성 부분집합을 생성하는 과정은 기본적으로 완전탐색(Complete Search), 순차탐색(Sequential Search), 무작위 탐색(Random Search) 기법 등으로 나눌 수 있다[Doak, 1992]. 완전탐색 기법은 생성할 수 있는 속성 부분집합의 모든 경우를 고려하는 방법으로서 시간과 비용이 막대하게 소요된다. 순차탐색의 가장 일반적인 방법은 전방향 순차 선택(FSS: Forward Sequential Selection)과 역방향 순차 선택(BSS: Backward Sequential Selection) 방식이다. 전방향 순차 선택 방식은 속성을 하나씩 추가해 가면서 속성 추가에 따른 효과를 측정해 나가는 것을 말한다. 이런 과정의 반복을 통해서 좋은 성과를 내는 속성 부분집합을 형성해 나가다가, 더 이상의 성과 향상이 없을 때에 탐색은 멈추게 된다. 역방향 순차 선택 방식은 반대로 속성을 하나씩 제거해 가면서 가장 높은 성과를 보이는 속성 부분집합을 찾아내는 것이다. 무작위 탐색 기법은 무작위로 생성된 수치들 또는 유전적 알고리즘(Genetic Algorithms)[Goldberg, 1989] 등을 이용하여 속성 부분집합을 생성해 내는 방식으로서 너무 많은 속성 부분집합과 왜곡된 부분집합이 생성될 가능성이 높다는 단점이 있다. 이 밖에도 여러 방식을 조합하는 전략에 대한 연구가 있었다[Caruana and Freitag, 1994].

본 연구에서는 순차탐색 기법의 변형된 방식을 사용하였는데, 속성을 하나씩만 사용하거나 하나씩만 제거하고 각 속성의 효과를 측정하였다. 이 내용은 제 4절에서 자세히 설명하기로 한다.

3.2 기존의 속성 선정 방법

제 1절에서 언급한 바와 같이 속성 선정 방법은 속성 가중치 부여 방법의 특수한 경우로 간주된다. 속성 선정 방법이건 속성 가중치 설정 방법이건 간에 이들은 크게 '전역적(Global)'인 방법과 '지역적(Local)'인 방법으로 분류된다. 전역적 방법이란 모든 사례가 동일한 속성 부분집합과 각 속성에 부여된 불변의 가중치에 의해서 유사 사례를 검색하게 되는 것이며, 지역적 방법이란 각 사례별로 서로 다른 속성 부분집합을 사용하거나 또는 동일한 속성 부분집합을 사용하더라도 각 사례별로

상이한 가중치를 사용하여 유사 사례를 검색하는 형식을 말한다.

본 절에서는 대표적인 속성 선정 방법인 Relief[Kira and Rendell, 1992], 의사결정 나무[Quinlan, 1993]를 이용한 방법, 유전적 알고리즘[Goldberg, 1989]을 이용한 방법, RC[Domingos, 1997], 그리고 CBDFW[이재식과 전용준, 2001] 방법을 소개한다.

Relief 알고리즘[Kira and Rendell, 1992]은 전역적 속성 가중치 부여 방법이다. 훈련 집합의 사례로부터 사용자가 결정한 개수만큼 표본 사례들을 선정한다. 이 표본 사례 각각(C_i 라고 하자)에 대해서 다음과 같은 과정을 수행한다. C_i에 대해서 유클리디언 거리 척도를 기반으로 표본 사례들로부터 Near Hit과 Near Miss 사례들을 검색한다. Near Hit은 C_i와 동일한 목표 속성값을 가진 사례들 중에서 가장 가까운 유클리디언 거리를 갖는 사례이고, Near Miss는 C_i와 상이한 목표 속성값을 가진 사례들 중에서 가장 가까운 유클리디언 거리를 갖는 사례이다. 어떤 속성이 C_i와 Near Hit에서 다른 값을 가진다면 그것은 중요하지 않은 속성이고, C_i와 Near Miss에서 다른 값을 가진다면 그것은 중요한 속성이라고 판정할 수 있다. 이 판정에 따라 0으로 초기화되었던 속성들의 가중치를 변화시킨다. 즉, 중요한 속성의 가중치는 양으로 증가시키고 중요하지 않은 속성의 가중치는 음으로 감소시키는 것이다. 이 과정을 표본 사례에 대해서 전부 수행한 후에는 가중치가 일정한 기준치 이상이 되는 속성들을 선정하여 사용하게 된다.

의사결정 나무 기법을 이용하여 전역적 속성 선정을 할 수 있는데, C4.5[Quinlan, 1993]와 같은 기법을 훈련 사례 집합에 대해서 수행하고, 가지치기의 결과로 남은 속성들을 선정된 속성 부분집합으로 간주하게 된다. 의사결정 나무를 생성하는 과정에서 계산되는 엔트로피(Entropy)값을 속성의 가중치를 설정하는데 사용하기도 한다[Cardie and Howe, 1997].

유전적 알고리즘을 이용한 전역적 속성 선정 방법은 무작위화된(Randomized) 탐색의 대표적인 방법이다[Kim and Shin, 1998; Shin and Han, 1998, 이재식과 차봉근, 1999]. 유전적 알고리즘은 체계화된 탐색기법으로서 단지 난수표만을 사용하는 무작위 탐색(Random Search)과는 다르다. 이 기법에서는 속성의 개수만큼의 유전인자(Gene)로 구성된 염색체(Chromosome)를 미리 결정한 개체군(Population)의 크기로 생성하여 재생산(reproduction), 교배(crossover), 그리고 돌연변이(mutation)를 거치면서 성능이 우수한 속성 부분 집합을 탐색하는 것이다. 유전적 알고리즘을 적용하기 위해서는 개체군의 크기, 재생산 방법, 교배 확률, 돌연변이 확률 등 여러 가지 값들을 정교하게 조정해야 한다.

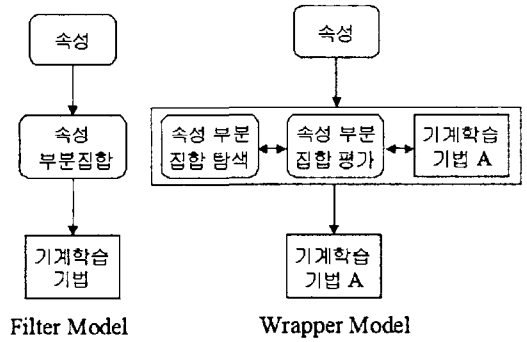
RC 알고리즘[Domingos, 1997]은 지역적 속성 선정 방법이다. 어떤 사례 C_i의 속성 A_{ij}의 값이 최근접이웃 사례의 그 값과 다르고, 또한 그 속성의 제거가 전체적인 분류 오차를 증가시키지 않는다면, A_{ij}를 C_i에서 제거할 수 있다. 모든 사례에 대해서 이와 같은 작업을 수행함으로써 사례별로 상이한 속성 부분 집합을 갖게 된다. 사례별로 상이하게 속성들을 제거하다보면 중복된 사례가 생성될 수 있

으나 이를 제거하지는 않는다.

CBDFW[이재식과 전용준, 2001]는 지역적 속성 가중치 부여 방법이다. 이 방법에서는 각 사례가 몇 개의 후보 가중치들과 함께 저장되어 있다. 이 가중치들은 사례마다 다르게 생성된 것으로서 과거 문제에서 우수한 성능을 보인 것들이다. 새로운 사례가 입력되면, 유사한 과거 사례를 검색할 때에 함께 저장된 가중치들도 검색되어 사용하게 된다. 미리 설정한 임계치 이상의 가중치를 가진 속성들만 사용함으로써 속성 선정의 효과를 얻을 수 있다.

3.3 평가 방법

선정된 속성 부분집합의 평가 방법은 크게 두 가지로 나눌 수 있는데, 그 기준은 선정된 속성 부분집합을 사용하여 얻은 분류결과에 대해서 피드백(Feedback)을 받는가의 여부이다. 피드백을 받지 않는 기법을 Filter Model이라 부르고, 기계학습 기법 자신으로부터 피드백을 받는 기법을 Wrapper Model이라고 부른다[John *et al.*, 1994]. 이 두 Model을 도식화하면 <그림 3-2>와 같다.



<그림 3-2> 속성 부분집합의 두 가지 평가방법

Wrapper Model은 비록 컴퓨터 자원을 훨씬 많이 사용하지만 기계학습 기법 자신이 평가한 속성들을 사용하기 때문에 새로운 사례에 대한 분류의 정확도가 매우 높다[Dash and Liu, 1997]. 이 때문에 John *et al.*은 속성의 부분집합을 선택하는데 있어서 Wrapper Model의 사용을 Filter Model의 사용보다 권장하였고, 일부 연구결과들이 분류의 정확도만을 목표로 하는 경우에는 이러한 주장이 타당함을 뒷받침하고 있다[Wettscherek *et al.*, 1997]. 본 연구에서 제시되는 속성 선정 방법은 사례기반 추론 시스템의 속성을 선정하기 위해서 사례기반 추론 시스템의 피드백을 받으므로 Wrapper Model이다.

4. 개별 속성의 효과 측정을 통한 전역적 속성 선정

4.1 개별 속성의 효과 측정

본 연구에서 수행하는 속성의 효과 측정 방식은 기존의 순차탐색의 방식과 흡사하게 보이지만, 기존의 연구에서는 속성 부분집합의 효과를 측정하는데 비해서 본 연구의 방법은 개별 속성의 효과를 측정한다는 면에서 차이가 있다. 속성 부분집합의 효과를 측정하게 되면, 개별 속성이 갖는 독립적인 기여도를 측정할 수가 없으며, 초기에 선정된 속성에 대한 의존성이 존재하게 된다. 또한, 수많은 조합을 형성할 수 있기 때문에 시간 소모적인 요소가 내재해 있었다.

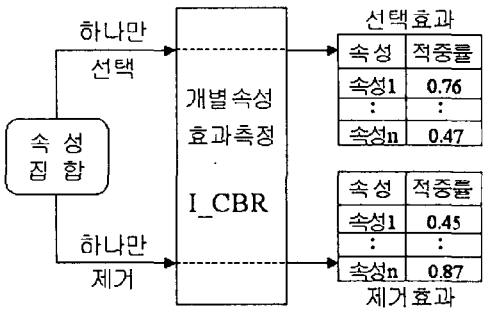
본 연구에서는 단 하나의 속성의 영향력을

평가하기 위해서 다음과 같은 방법을 사용하였다.

(1) 개별 속성의 선택효과(Selection Effect : SE) 측정 : 단 하나의 속성만을 사용하여 사례기반 추론을 수행하여 그 적중률을 측정한다.

(2) 개별 속성의 제거효과(Elimination Effect : EE) 측정 : 전체 속성에서 단 하나의 속성을 제거하였을 때에 사례기반 추론을 수행하여 그 적중률을 측정한다.

이 두 가지 효과측정을 통하여 각 속성의 개별적인 영향력을 평가하고, 그것을 근거로 속성 부분집합을 구성하는 전략을 세우게 된다. <그림 4-1>은 이 과정을 도식화한 것으로서 그림 중앙에 있는 개별 속성의 효과를 측정하기 위한 사례기반 추론 시스템을 I_CBR(CBR for measuring the effects of Individual features)로 명명한다.



<그림 4-1> 개별 속성의 효과 측정

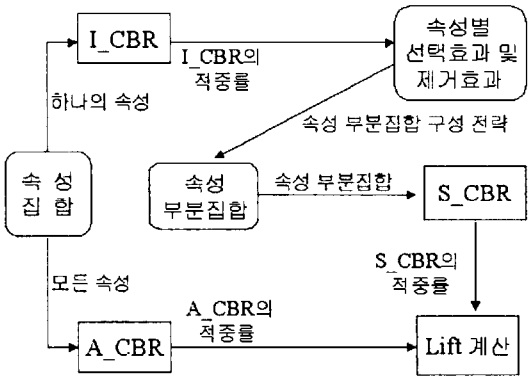
4.2 속성 부분집합의 평가

각 속성의 효과를 측정한 다음에는 그 속성들의 조합을 구성하고 평가하여 속성 부분집합을 선정하게 된다. 본 연구에서는 개별 속성의 효과를 기준으로 9가지 종류의 속성 부분집합을 구성하였는데, 먼저 우수한 효과를 내는 순서로 속성을 나열하고, 일정한 비율로 속성의 개수를 늘려 가며 속성 부분집합을 구성하였다. 한편, 효과가 우수하지 않은 속성들도 다른 속성과의 조합을 통해서 상승 작용을 나타낼 소지가 있으므로, 좋지 않은 효과를 보이는 속성순으로도 일정한 비율로 속성의 개수를 늘려 가며 속성 부분집합을 구성하였다. 또한, 다양한 전략을 구상하기 위해서 상위 30%의 속성과 하위 30%의 속성을 조합하여 속성 부분집합을 구성하는 방식도 사용하였다. 본 연구에서 개발한 방법은 전역적 속성 선정 방법으로서 S&E(feature selection by analyzing Selection effects & Elimination effects) 기법이라고 명명하였다.

선정된 속성 부분집합은 그 속성들만을 사용하는 사례기반추론 시스템에 의해 그 성능이 측정된다. 이 사례기반추론 시스템을 S_CBR(CBR using the Subset of features)이라고 명명한다. 한편, 모든 속성을 사용하는 사례기반추론 시스템은 A_CBR(CBR using All features)이라고 명명한다. 속성 부분집합 구성 전략의 유용성은 (식 4.1)로 계산하는 상승효과(Lift)에 의하여 검증한다. 모든 분석과 검증은 타당성의 확보를 위해서 사례베이스와 테스트 사례를 무작위로 10번 구성하여 실험하는 10-Fold 테스트로 수행하였다. 즉, (식 4.1)의 분모, 분자는 모두 10-Fold 테스트의 평균값들이다.

$$\text{상승효과(Lift)} = \frac{\text{선정된 속성 부분집합을 사용한 S\_CBR의 성과}}{\text{모든 속성을 사용한 A\_CBR의 성과}} \quad \text{(식 4.1)}$$

속성 부분집합의 구성 및 평가 과정을 도식화하면 <그림 4-2>와 같다.



<그림 4-2> 속성 부분집합의 구성 및 평가 과정

5. S&E 기법의 성능

5.1 사용 데이터

본 연구에서 개발된 S&E 기법의 유용성을 검증하기 위해서 먼저 UCI(University of California at Irvine)의 기계학습 연구소[Blake *et al.*, 1998]에서 제공하는 연구용 데이터인 '신용 평가(Credit Screening)', '당뇨병 진단(Pima Indians Diabetes)', '자동차 연비(Auto MPG)'를 사용하였다. 그리고, 실제 기업의 데이터로는 신용카드 고객의 이탈 예측을 위한 데이터를 사용하였다. 각 데이터는 사례베이스용으로 40%, 테스트용으로 30%가 사용된다. 나머지 30%를 사용하지 않는 이유는, 실험에 사용할 데이터의 규모가 그리 크지 않으므로, 데이터 전부를 사례베이스용과 테스트용으로 사용하게 되면 10-Fold 테스트간에 비슷한 구성을 가진 사례베이스와 테스트 사례가 만들어질 수 있기 때문이다. 즉, 본 연구에서는 전체 데이터에서 70%를 추출하여 이것으로 사례베이스와 테스트 사례를 4:3으로 구성하여 실험을 수행하고, 다시 전체 데이터에서 70%를 추출하여 이것으로 또 사례베이스와 테스트 사례를 4:3으로 구성하여 실험을 수행하는 방식을 총 10번 반복 수행하였다. 본 연구는 속성 선정 전략의 유용성을 검증하기 위한 것이므로 모델 검증에 주로 쓰이는 평가용(Validation) 데이터는 설정하지 않았다.

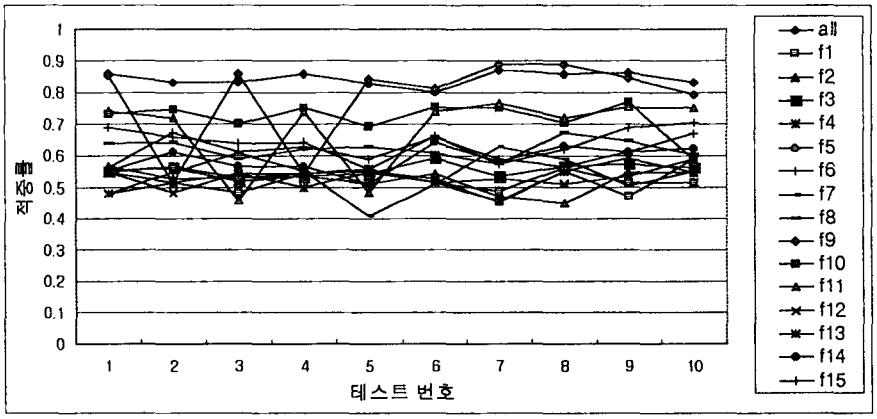
5.2 신용 평가

신용 평가는 고전적인 분류 문제로서, 은행이나 신용카드 회사는 대부나 카드발급에 있어서 그 한도를 결정하거나 여부를 판정하는 업무에 신용 평가에 대한 규칙을 적용하고 있다. 본 연구에 사용된 신용 평가용 데이터는 690개의 사례를 가지고 있다. 여기서 결측치가 있는 사례를 제거하고 651개의 사례를 사용하였다. 각 사례는 15개의 설명 속성(6개의 연속형 속성과 9개의 범주형 속성)과 1개의 목표 속성(이진형 속성)으로 구성되어 있다. 목표 속성은 신용 평가 결과를 나타내는데, 적격자가 55%, 부적격자가 45%로 구성되어 있다. 신용 평가 데이터의 경우에는 개인 정보를 보호하기 위해서 모든 속성이 부호화 되어 있다.

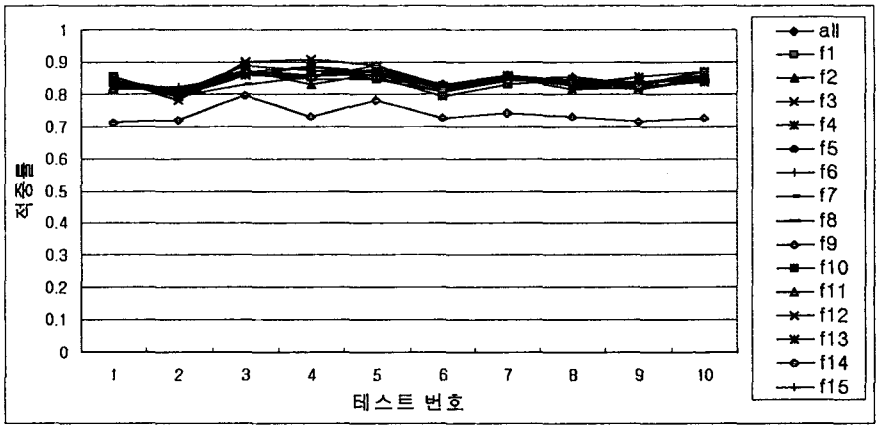
그러므로, 각 속성이 어떤 특성을 갖는지 그 의미를 파악할 수가 없었다.

<그림 5-1>과 <그림 5-2>는 개별 속성의 효과를 10-Fold 테스트로 측정한 결과를 그래프로 나타낸 것이다. 선택효과 측정의 경우에는 테스트별로 차이가 크게 나타났으나, 제거효과 측정의 경우에는 그 격차가 크게 벌어지지는 않았다.

<표 5-1>은 개별 속성의 효과에 따라 순위를 부여한 것인데, 이것을 근거로 속성 부분집합을 선정하게 된다. 선택효과 측정은 속성의 선택시 측정한 적중률을 기준으로 하기 때문에, 높은 적중률을 내는 속성부터 순위를 부여하였다. 반면에 제거효과 측정은 특정 속성을 제거했을 때의 적중률을 기준으로 하기 때문에, 적중률에 대한 기여도가 큰 속성일수록 효과가 낮게 나타나게 된다. 따라서 제거시의 적중률이 낮은 속성부터 높은 순위를 부여하였다.



<그림 5-1> 선택효과 측정 결과 (신용 평가)

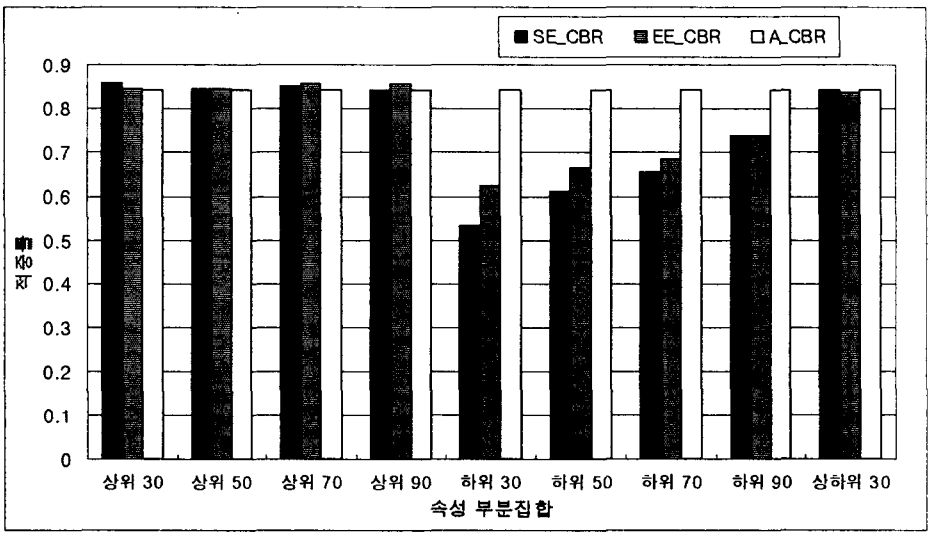


<그림 5-2> 제거효과 측정 결과 (신용 평가)

<표 5-1> 속성 순위 부여 결과 (신용 평가)

선택효과			제거효과		
속 성	적중률	순 위	속 성	적중률	순 위
f9	0.7832	1	f9	0.7367	1
f10	0.7184	2	f7	0.8347	2
f11	0.6872	3	f13	0.8378	3
f15	0.6429	4	f10	0.8383	4
f8	0.6214	5	f8	0.8408	5
f6	0.6153	6	f2		
f14	0.5898	7	f3	0.8418	7
f3	0.5536	8	f5	0.8429	8
f7	0.5372	9	f4		
f4	0.5337	10	f1	0.8434	10
f5			f14		
f12			f11		
f13	0.5245	13	f15	0.8449	13
f1	0.5230	14	f6	0.8459	14
f2	0.5224	15	f12	0.8556	15

속성 부분집합은 <표 5-1>에 부여된 순위에 따라 구성하였는데, 동일 순위를 가진 속성들은 모두 다 포함시키거나, 모두 다 포함시키지 않는 방식으로 부분집합을 구성하였다. <그림 5-3>은 속성 부분집합의 성능 측정 결과를 그래프로 나타낸 것이다. 그림에서 SE_CBR (Selection Effect Case-Based Reasoning)은 선택효과의 순위에 따라 속성 부분집합을 구성했을 때에 사례기반 추론의 적중률이고, EE_CBR(Elimination Effect Case-Based Reasoning)은 제거효과의 순위에 따라 속성 부분집합을 구성했을 때에 사례기반 추론의 적중률이다. 모든 속성을 사용하여 수행한 A_CBR의 적중률은 84.2%이었다. A_CBR보다 적중률이 높게 나타난 속성 부분집합은 SE_CBR의 경우에는 상위 30%, 50%, 70%이었으며, EE_CBR의 경우에는 상위 30%, 50%, 70%, 90%이었다. 가장 적중률이 높은 것은 SE_CBR에서 상위 30%의 속성 부분집합을 사용했을 때인데 85.7%의 적중률을 보여서 A_CBR보다 1.5% 포인트의 적중률 향상을 가져왔다. 각 속



<그림 5-3> 선정된 속성 부분집합의 성능 (신용 평가)

성 부분집합의 Lift 값은 <표 5-2>에 나타나 있는데, Lift 값이 진하게 표기된 부분들이 A_CBR 보다 높은 적중률을 보인 속성 부분집합들이다.

5.3 당뇨병 진단

분류 문제의 가장 대표적인 분야중의 하나는 진단 문제이다. 연구에 사용된 데이터는 Pima 인디언 보호구역에 살고 있는 21세 이상된 여성 중 당뇨병을 보유하고 있는 여성 268명과 당뇨병을 보유하고 있지 않은 여성 500명의 사례로 구성되어 있다. 각 사례는 8개의 설명 속성(연속형 속성)과 1개의 목표 속성(이진형 속성)으로 구성되어 있다. 목표 속성은 당뇨병 보유 여부를 나타낸다. 결측치는 없으며, 속성의 구성은 <표 5-3>과 같다.

<표 5-2> 선정된 속성 부분집합의 Lift 값 (신용 평가)

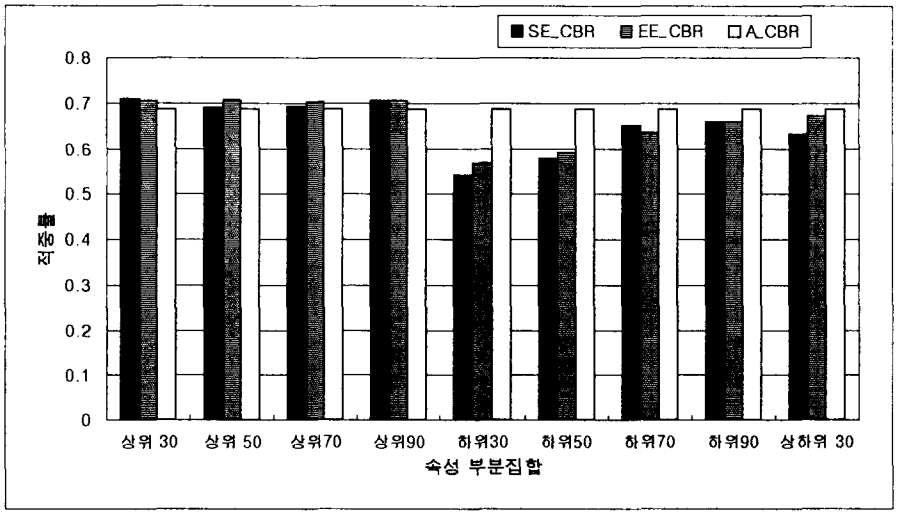
SE_CBR				EE_CBR			
속성 부분집합	적중률		Lift	속성 부분집합	적중률		Lift
	평균	분산			평균	분산	
상위 30	0.857	0.001	1.018	상위 30	0.845	0.001	1.004
상위 50	0.844	0.001	1.002	상위 50	0.844	0.001	1.002
상위 70	0.849	0.001	1.008	상위 70	0.855	0.001	1.015
상위 90	0.841	0.000	0.998	상위 90	0.856	0.001	1.016
하위 30	0.534	0.001	0.634	하위 30	0.621	0.001	0.737
하위 50	0.611	0.001	0.726	하위 50	0.664	0.001	0.788
하위 70	0.652	0.000	0.773	하위 70	0.683	0.002	0.811
하위 90	0.737	0.001	0.875	하위 90	0.737	0.001	0.875
상하위 30	0.840	0.000	0.998	상하위 30	0.836	0.001	0.993

<표 5-3> 당뇨병 진단 데이터의 속성 구성

속성	설 명
f1	임신 회수
f2	구강내 2시간 동안의 글루코오즈 잔류량 테스트
f3	혈압(mm Hg)
f4	삼두박근의 두께(mm)
f5	2시간 동안의 세룸 인슐린의 양(mu U/ml)
f6	비만도(몸무게 kg/(키 m2))
f7	가계 당뇨병 병력 함수
f8	나이
목표속성	양성, 음성 반응(1, 0)

I_CBR로 선택효과와 제거효과를 측정한 결과, 당뇨병 진단 데이터에서는 각 속성의 적중률에 대한 기여도간에 큰 차이는 없었다. 두 가지 효과측정 방법 모두 가장 높은 효과를 보인 속성은 f2 즉, '구강내 2시간 동안의 글루코오즈 잔류량 테스트'였으며, 가장 낮은 효과를 보인 속성은 f7로서 '가계 당뇨병 병력 함수'였다.

당뇨병 진단 데이터에서는 A_CBR이 68.7%의 적중률을 보였는데, SE_CBR에서 상위 30%의 속성 부분집합을 사용했을 때에 71.0%의 적중률을 보여서 2.3% 포인트의 성능 향상을 보였다. 선정된 속성 부분집합의 실험 결과는 <그림 5-4>와 <표 5-4>에 나타나 있다.



<그림 5-4> 선정된 속성 부분집합의 성능 (당뇨병 진단)

<표 5-4> 선정된 속성 부분집합의 Lift 값 (당뇨병 진단)

SE_CBR				EE_CBR			
속성 부분집합	적중률		Lift	속성 부분집합	적중률		Lift
	평균	분산			평균	분산	
상위 30	0.710	0.000	1.034	상위 30	0.704	0.000	1.026
상위 50	0.689	0.000	1.003	상위 50	0.707	0.001	1.030
상위 70	0.691	0.001	1.007	상위 70	0.703	0.001	1.023
상위 90	0.705	0.001	1.027	상위 90	0.705	0.001	1.027
하위 30	0.541	0.011	0.788	하위 30	0.567	0.010	0.826
하위 50	0.577	0.009	0.841	하위 50	0.590	0.010	0.860
하위 70	0.649	0.001	0.946	하위 70	0.636	0.001	0.927
하위 90	0.657	0.001	0.957	하위 90	0.657	0.001	0.957
상하위 30	0.631	0.011	0.920	상하위 30	0.670	0.002	0.977

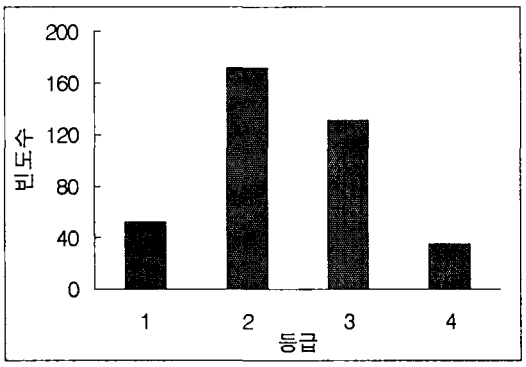
Lift 값이 1이상인 속성 부분집합은 8개인데, 모두 효과가 높은 속성들을 기준으로 부분집합을 구성했을 때이다. 효과가 낮은 속성을 기준으로 속성 부분집합을 구성했을 때에는 포함되는 속성의 개수가 많아질수록 좋은 결과가 나타나고 있다.

5.4 자동차 연비

UCI에서 제공하는 자동차 연비 데이터의 각 사례는 <표 5-5>와 같이 7개의 설명 속성(4개의 연속형 속성과 3개의 범주형 속성)과 1개의 목표 속성(연속형 속성)으로 구성되어 있다. 목표 속성이 연속형 값을 가지므로 분류문제화 하기 위해서는 몇 개의 군으로 묶어야 했다. 본 연구에서는 차량의 연비를 10단위까지 반올림하여 5개의 연비 등급군을 형성하였는데, 이 과정에서 사례의 개수가 충분하지 못한 등급 한 개와 결측치를 갖는 사례들은 제거되었다. 결과적으로 총 398개의 사례 중에서 7개가 제거되어 391개의 사례를 사용하게 되었다. <그림 5-5>는 등급별 사례의 빈도수를 나타낸 것이다. 등급 1은 연비가 가장 낮은 집단이며, 등급 4는 연비가 가장 좋은 집단이다.

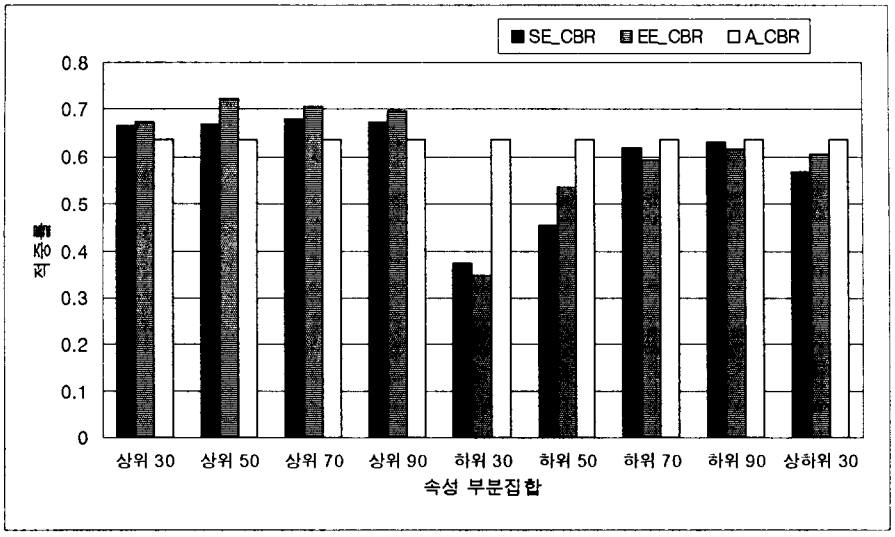
<표 5-5> 자동차 연비 데이터의 속성 구성

속 성	설 명
f1	실린더 수 : 이산형
f2	배기량 : 연속형
f3	마력 : 연속형
f4	중량 : 연속형
f5	가속 : 연속형
f6	연식 : 이산형
f7	제조국가 : 이산형
목표속성	MPG(Miles per Gallon) : 연속형 (이산형으로 변환)



<그림 5-5> 등급별 사례의 빈도 (자동차 연비)

<그림 5-6>과 <표 5-6>은 실험 결과로서, 자동차 연비 데이터에서도 인디언의 당뇨병 진단 데이터와 마찬가지로 A_CBR 보다 나은 성능을 보인 속성 부분집합은 SE_CBR의 경우 상위 30%, 50%, 70%, 90%, 그리고 EE_CBR의 경우에도 역시 상위 30%, 50%, 70%, 90%로 총 8개이었다. 이 데이터에서는 A_CBR이 63.7%의 적중률을 보였는데, EE_CBR에서 상위 50%의 속성 부분집합을 사용했을 때에 72.1%의 적중률을 보여서 8.4% 포인트의 성능 향상을 보였다.



<그림 5-6> 선정된 속성 부분집합의 성능 (자동차 연비)

<표 5-6> 선정된 속성 부분집합의 Lift 값 (자동차 연비)

SE_CBR				EE_CBR			
속성 부분집합	적중률		Lift	속성 부분집합	적중률		Lift
	평균	분산			평균	분산	
상위 30	0.664	0.001	1.043	상위 30	0.673	0.002	1.056
상위 50	0.667	0.001	1.047	상위 50	0.721	0.001	1.133
상위 70	0.679	0.002	1.067	상위 70	0.706	0.001	1.109
상위 90	0.672	0.003	1.055	상위 90	0.697	0.002	1.094
하위 30	0.374	0.002	0.587	하위 30	0.347	0.006	0.545
하위 50	0.455	0.004	0.714	하위 50	0.533	0.003	0.838
하위 70	0.619	0.004	0.972	하위 70	0.592	0.004	0.930
하위 90	0.631	0.005	0.991	하위 90	0.615	0.004	0.965
상하위 30	0.568	0.002	0.891	상하위 30	0.604	0.002	0.949

5.5 고객 이탈 예측

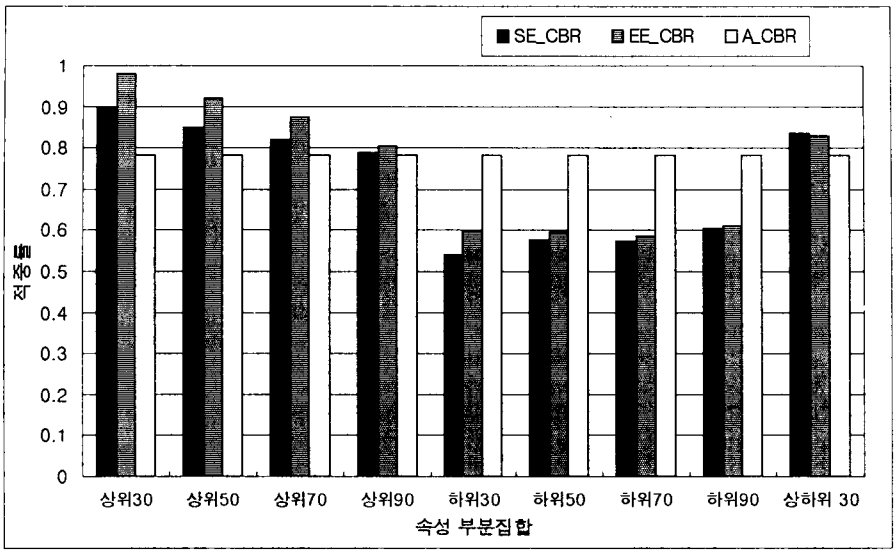
고객 관계 관리(Customer Relationship Management: CRM)는 현대의 경영 환경에서 매우 중요한 위치를 차지하고 있다. 고객을 지속적으로 유지·관리하여 기업의 이윤과 연결시키려는 노력들이 분주하게 진행되고 있는데, 현재 대부분의 기업에서 수행하고 있는 고객 관계 관리는 고객의 구매 성향이나 선호도 등을 통해서 고객의 가치를 계산하고, 이탈 가능

성 여부를 예측하는 것이다. 본 절에서는 실제 기업의 데이터를 사용하여 본 연구에서 제시한 속성선정 전략의 검증을 시도하였는데, UCI에서 제공한 연구용 데이터 보다 뛰어난 성능 향상을 보였다. 각 사례는 <표 5-7>과 같이 30개의 설명 속성(12개의 연속형 속성, 16개의 범주형 속성, 2개의 문자형 속성)과 1개의 목표 속성(이진형 속성)으로 구성되어 있다. 문자형 속성인 '카드번호'와 '성명'은 사용하지 않았다. <표 5-7>에서 물음표로 표시한 것들은 고객들의 개인 정보 보호를 위해서 그 내용을 공개하지 않은 속성들이다.

실험 결과는 <그림 5-7>과 <표 5-8>에서 볼 수 있는 바와 같이 적중률의 매우 높은 향상을 보였다. UCI 연구용 데이터에서와 마찬가지로 효과가 높은 속성을 기준으로 상위 30%, 50%, 70%, 90%에서 성능의 향상을 보였으며, 상위 30%와 하위 30%를 조합하여 속성 부분집합을 구성한 결과도 A_CBR 보다 나은 적중률을 보였다. A_CBR의 적중률이 78.3%인데, EE_CBR의 경우 상위 30%의 속성 부분집합을 사용했을 때에 적중률이 97.9%로서 19.6% 포인트의 성능 향상을 보였다. 이 때에 포함된 속성들은 f3, f4, f6, f8, f20, f21, f23, f25, f26, f28이었다. 분류 정확도의 향상에 가장 높게 기여하는 속성은 f28인 '행사 참여 빈도'로 나타났으며, 기여도가 가장 낮은 속성은 선택효과의 경우 f9인 '가족의 수'로 나타났고, 제거효과의 경우에는 f17인 '취미 코드'로 나타났다.

<표 5-7> 고객 이탈 예측 데이터의 속성 구성

속 성	설 명	데이터 형식	속 성	설 명	데이터 형식
f1	카드 번호	문자형	f17	취미코드	범주형
f2	성명	문자형	f18	아파트 코드	범주형
f3	직장인 여부	범주형	f19	직위	범주형
f4	결혼 기간	연속형	f20	카드 구분	범주형
f5	연령	연속형	f21	한도 금액	연속형
f6	아파트 평수	연속형	f22	연체회수	연속형
f7	가입 기간	연속형	f23	?	연속형
f8	고객 등급	범주형	f24	?	연속형
f9	가족의 수	연속형	f25	?	연속형
f10	청구지 구분	범주형	f26	?	연속형
f11	결혼 여부	범주형	f27	주 구매점	범주형
f12	성별	범주형	f28	행사 참여 빈도	연속형
f13	?	범주형	f29	집 우편지역코드	범주형
f14	?	범주형	f30	직장 우편지역코드	범주형
f15	?	범주형	목표속성	이탈 여부	이진형
f16	직종 코드	범주형			



<그림 5-7> 선정된 속성 부분집합의 성능 (고객 이탈 예측)

<표 5-8> 선정된 속성 부분집합의 Lift 값 (고객 이탈 예측)

SE_CBR				EE_CBR			
속성 부분집합	적중률		Lift	속성 부분집합	적중률		Lift
	평균	분산			평균	분산	
상위 30	0.897	0.000	1.146	상위 30	0.979	0.000	1.251
상위 50	0.851	0.000	1.087	상위 50	0.920	0.000	1.174
상위 70	0.818	0.001	1.044	상위 70	0.874	0.000	1.116
상위 90	0.790	0.000	1.008	상위 90	0.804	0.000	1.027
하위 30	0.541	0.001	0.691	하위 30	0.598	0.000	0.763
하위 50	0.577	0.000	0.737	하위 50	0.595	0.000	0.760
하위 70	0.573	0.001	0.732	하위 70	0.587	0.000	0.749
하위 90	0.604	0.001	0.771	하위 90	0.610	0.000	0.779
상하위 30	0.835	0.000	1.066	상하위 30	0.828	0.000	1.058

5.6 S&E 기법의 성능 평가

본 연구에서 사용한 4개 데이터 모두에서 A_CBR 보다는 SE_CBR과 EE_CBR이 더 나은 적중률을 보이고 있다. 특히, 실제 기업 데이터인 고객 이탈 예측 데이터에서는 적중률이 19.6% 포인트 향상된 괄목할 만한 성과를 보였다. <표 5-9>는 FSS, BSS, RC 기법 [Domingos, 1997], 그리고 본 연구에서 제시한 S&E 기법의 적중률을 비교한 것이다.

Domingos가 제시한 실험결과들 중에서 '신용평가'와 '당뇨병 진단'의 데이터가 본 연구에서 사용한 데이터와 일치했으므로 이 두 가지 데이터의 연구 결과만을 비교할 수 있었는데, 본 연구에서 제시한 S&E 기법이 FSS, BSS, RC 기법들보다 더 나은 성능을 보였다.

<표 5-9> 선행 연구와의 성능 비교

단위: %

사용 데이터	FSS	BSS	RC	S&E
신용 평가	80.9	81.2	83.7	85.7
당뇨병 진단	69.6	69.2	70.5	71.0

* FSS : Forward Sequential Selection (전방향 순차 선택)
* BSS : Backward Sequential Selection (역방향 순차 선택)
* RC : Relevance in Context

본 절에서 제시된 4개의 데이터에 대한 실험결과에서 볼 수 있듯이, S&E 기법에서 선정된 속성 부분집합은 선택효과와 제거효과의 순위에서 상위를 차지하는 속성들로 구성되어 있다. 즉, 본 연구에서 측정한 개별 속성의 효과 수치들이 속성 선정의 기준으로서 유용성을 가진다는 것을 알 수 있다. 그러므로, 본 연구에서 제시한 개별 속성의 효과 측정 방법으로 각 속성들의 효과를 측정한 후에 상위의 속성들을 선정함으로써, 모든 속성을 사용할 때보다 효과적·효율적인 속성 부분집합을 선정할 수 있는 것이다.

6. 결론 및 향후 연구과제

사례기반 추론에 있어서 속성 선정과 속성 가중치 설정에 관한 많은 연구가 수행되어 왔다. 그러나 선행 연구들은 속성의 부분집합을 구성하여 그 효과를 측정하는 것들이었으며, 개별 속성의 효과를 측정한 연구는 없었다. 본 연구에서는 개별 속성별로 선택효과와 제거효과를 측정하고, 이를 이용하여 속성 부분집합을 구성하는 S&E 기법을 개발하였다. 연구용 데이터 및 실제 기업 데이터에 S&E 기법을 적용하여 속성 선정을 한 후에 사례기반 추론을 수행해 본 결과, 모든 속성을 전부 사용했을 때보다도 적중률이 좋은 속성 부분집합을 얻을 수 있었다.

본 연구에서 사용한 사례기반 추론 시스템은 제 2절에서 기술하였듯이 매우 단순한 것이다. 특히, 속성들의 가중치가 모두 1로 고정되어있다. 가중치를 모두 동일하게 하였는데도 선정된 속성 부분집합의 Lift 값이 1보다 크게 나왔다. 그러므로, 만일 가중치를 변화시킨다면, 본 연구에서 보여준 적중률보다 향상된 적중률을 얻을 수 있을 것이다.

본 연구의 한계점 및 향후 연구에서 보완할 수 있는 부분은 다음과 같다. 첫째, 범주형 속성의 경우에 유사도 점수를 단순하게 1과 0.5의 두 가지 값으로 부여한 것을 한계점으로 들 수 있다. 범주형 속성의 유사도 점수를 부분일치(Partial Matching)에 의하여 부여할 수 있도록 영역 지식이나 전문가를 적극적으로 활용하여야 한다. 둘째, 선택효과와 제거효과의 결과를 조합하여 새로운 속성 부분집합 구성에 대한 전략을 모색하여야 한다. 예를 들어, 효과가 우수한 속성 또는 속성 부분집합을 고정적으로 포함시킨 상태에서 다른 속성들의 선택효과와 제거효과를 측정함으로써 새로운

측정치를 도출할 수 있을 것이다.

참고문헌

[1] 이재식, 전용준, "사례기반 추론을 위한 동적 속성 가중치 부여 방법," 한국지능정보시스템학회지, 제 7권 1호, (2001), 47-61.

[2] 이재식, 차봉근, "유전적 알고리즘을 이용한 인공신경망의 구조 설계," 한국경영과학회지, 제 24권 4호, (1999), 49-62.

[3] Aha, D. W., "Feature Weighting for Lazy Learning Algorithms," in Liu, H. and H. Motoda(eds.), *Feature Extraction, Construction and Selection: A Data Mining Perspective*, Kluwer, Norwell MA, 1998.

[4] Allen, P. B., "Case-Based Reasoning: Business Applications," *Communications of the ACM*, Vol. 37, No.3, March, (1994), 40-42.

[5] Blake, C., E. Keogh and C. J. Merz, *UCI Repository of Machine Learning Databases*, [http://www.ics.uci.edu/~mlearn/MLRepository.html], Dept. of Information and Computer Science, Irvine, CA: Univ. of California, 1998.

[6] Cardie, C. and N. Howe, "Improving Minority Class Prediction Using Case-Specific Feature Weights," *Proceedings of the Fourteenth International Conference on Machine Learning*, (1997), 57-65.

[7] Caruana, R. and D. Freitag, "Greedy Attribute Selection," *Proceedings of the Eleventh International Conference on Machine Learning*, (1994).

[8] Dash, M. and H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis*, Vol.3 No.3, (1997), 131-156.

[9] Doak, J., *An Evaluation of Feature Selection Methods and Their Application to Computer Security*, Technical Report CSE-92-18, Dept. of Computer Science, Davis, CA: Univ. of California, 1992.

[10] Domingos, P., "Context-Sensitive Feature Selection for Lazy Learners," *Artificial Intelligence Review*, Vol.11, (1997), 227-253.

[11] Goldberg, D. E., *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Pub. Co., Inc., 1989.

[12] Jo, H., I. Han and H. Lee, "Bankruptcy Prediction using Case-Based Reasoning, Neural Networks, and Discriminant Analysis," *Expert Systems With Applications*, Vol.13, No.2, (1997), 97-108.

[13] John, G. H., R. Kohavi and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," *Proceedings of the Eleventh International Conference on Machine Learning*, (1994), 121-129.

[14] Kim, S. H., *Smart Finance: Forecasting Chaotic Markets and Economies through Knowledge Discovery*, KAIST, Korea, 1997.

[15] Kim, S. H. and S. Shin, "Optimizing Retrieval of Precedents in Case-Based Reasoning through a Genetic Algorithm," 1998 한국전문가시스템학회 추계 학술대회 논문집, (1998), 123-129.

[16] Kira, A. and L. A. Rendell, "A Practical Approach to Feature Selection," *Proceedings of the Ninth International Conference on Machine Learning*, (1992), 249-256.

[17] Kolodner, J., *Case-Based Reasoning*, Morgan Kaufman Publishers, 1993.

[18] Lee, J. S. and Y. X. Xon, "A Customer Service Process Innovation using the Integration of Data Base and Case Base," *Expert Systems with Applications*, Vol.11, No.4, (1996), 543-552.

[19] Quinlan, J. R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA., 1993.

[20] Riesbeck, C. K. and R. L. Schank, *Inside Case-Based Reasoning*, Lawrence Erlbaum

Associates, 1989.

[21] Shin, K. and I. Han, "A Hybrid Approach using Case-based Reasoning and Genetic Algorithm for Corporate Bond Rating," 1998 한국경영정보학회/한국전문가시스템학회 춘계 공동 학술대회 논문집, (1998) 106-109.

[22] Watson, I., *Applying Case-Based Reasoning: Techniques for Enterprise Systems*, Morgan Kaufmann, 1997.

[23] Wettschereck, D., D. W. Aha and T. Mohri, "A Review and Empirical Comparison of Feature Weighting Methods for a Class of Lazy Learning Algorithms," *AI Review*, Vol.11, (1997), 273-314.

Abstract

Feature Selection for Case-Based Reasoning using the Order of Selection and Elimination Effects of Individual Features

Jae Sik Lee* · Hyuk Hee Lee**

A CBR(Case-Based Reasoning) system solves the new problems by adapting the solutions that were used to solve the old problems. Past cases are retained in the case base, each in a specific form that is determined by features. Features are selected for the purpose of representing the case in the best way. Similar cases are retrieved by comparing the feature values and calculating the similarity scores. Therefore, the performance of CBR depends on the selected feature subsets. In this research, we measured the Selection Effect and the Elimination Effect of each feature. The Selection Effect is measured by performing the CBR with only one feature, and the Elimination Effect is measured by performing the CBR without only one feature. Based on these measurements, the feature subsets are selected. The resulting CBR showed better performance in terms of accuracy and efficiency than the CBR with all features.

* School of Business Administration, Ajou University
** Xsolution Consulting