한국지능정보시스템학회논문지 제7권 제1호 2001년 6월(pp. 47~61)

사례기반 추론을 위한 동적 속성 가중치 부여 방법*

이재식
아주대학교 경영대학 교수
(leejsk@madang.ajou.ac.kr)
전용준
(주) 엑셀루션컨설팅
(xyxonxyxon@empal.com)

사례기반 추론과 같은 사후학습 기법은 인공신경망이나 의사결정나무와 같은 사전학습 기법에 비해서 여러 장점을 가지고 있다. 하지만, 사후학습 기법은 사례 표현에 관련성이 적은 속성이 포함된 경우에는 성능이 저하되는 단점을 가지고 있다. 이러한 단점을 극복하기 위해서, 속성 가중치 부여 방법들이 연구되었다. 기존의 속성 가중치 부여 방법들은 대부분 전역적으로 속성 가중치를 부여하는 것이었다. 본 연구에서는 새로운 지역적 속성 가중치 부여 방법인 CBDFW를 제안한다. CBDFW 기법은 무작위로 생성된 속성 가중치들의 분류 성공 여부를 저장하고 있다가, 새로운 사례가 주어졌을 때에 성공적인 분류 결과를 보인 가중치들을 검색하여 동적으로 새로운 가중치들을 생성해낸다. 신용평가 데이터로 CBDFW의 성능을 실험한 결과, 기존의 연구들에서 제시된 분류 적중률보다 우수한 성능을 보였다.

1. 서 론

분류 문제는 주어진 과거 사례들로부터 학습을 한 후 새로운 질의 사례가 입력되면 그에 맞는 클래스를 결정하는 문제이다. 이러한 문제를 풀기 위한 기계 학습 기법은 사전학습(Eager Learning) 방법과 사후학습(Lazy Learning) 방법의 두 종류로 구분된다. 사전학습 방법은 훈련용으로 수집된 사례들로부터 일반화된 규칙 또는 수식으로 표현된 클래스 분류 방법을 미리 도출해 놓는다. 그러므로 새로운 질의 사례가 입력되면 곧바로 이 분류 방법을 적용하여 클래스를 분류한다. 반면, 사후학습 방법은 일반화된 클래스 분류 방법을 미리 도출해 놓지 않고 있다가, 새로운 질의 사례가 입력되면 가장 유사한 과거 사례를 찾아서 그것의 클래스를 사용하여 질의 사례의 클래스를 분류한다. 사례기반 추론(CBR: Case-based Reasoning)[Kolodner, 1993] 기법을 비롯한 사후학습 방법들은 의사결정 나무[Quinlan, 1986]나 인공신경망[Nelson and Illingworth, 1991] 등과 같은 사전학습 방법에 비하여 여러 가지 장점을 가지고 있다. 즉, 적은 정보만을 사용하는 경우에도 사례 공간에서 복잡한 의사결정 관계를 형성하는 것이 가능하고, 입력과 출력이 수치형이든 범주형이든 모든 형태의 문제에 쉽게 적용이 가능하며, 사례를 단순히 저

* 이 논문은 2000년도 두뇌한국 21사업 핵심분야 사업비에 의하여 지원되었음.

장해 두는 것으로 학습이 완료되기 때문에 학습 과정이 간단하다.

하지만, 사례 저장을 위한 공간이 많이 필요하고 적절한 색인 방법을 사용한다해도 분류에 시간이 많이 소요되는 등의 단점도 있다. 그러나 가장 심각한 문제는 관련성이 낮은 속성이 존재하는 경우이다. 관련성이 낮은 속성이 사례에 많이 존재하는 경우에는 사후학습 방법은 그러한 속성들로 인해서 혼란을 겪게 되며 결과적으로 성능이 심하게 저하된다. 이에 대한 자연스러운 해결책은 관련성이 낮은 속성을 찾아내서 이를 사용하지 않도록 하는 방법이다. 이러한 목적에서 여러 가지 방법들이 제안되었는데[Aha, 1998], 속성 선정이나 속성 가중치 부여 방법이 그것들이다. 속성 선정은 연관성이 없는 속성을 사례로부터 삭제시키는 것이고, 속성 가중치 부여는 속성의 연관성 정도에 따라 다른 가중치를 부여하는 것이다. 가중치가 일정 수준 이하인 속성을 사용하지 않는다면 속성 가중치 부여는 곧 속성 선정과 같은 효과를 가져오게 된다. 즉, 속성 선정을 일반화시키면 속성 가중치 부여가 된다고 할 수 있다. 속성 가중치 부여의 자동화는 많은 이점을 제공한다. Caruana and Freitag[1994]는 속성 선정의 장점에 대해 기술하였는데, 우리는 이에 근거하여 속성 가중치 부여의 자동화시 얻을 수 있는 장점을 아래와 같이 파악할 수 있다.

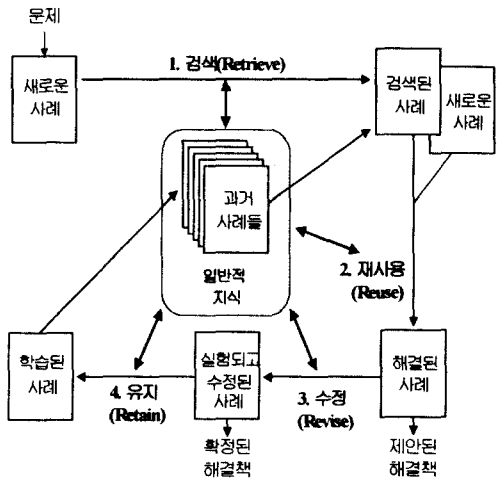
(1) 시스템 설계자는 잠재적으로 유용할 것으로 판단되는 속성을 가능한 한 많이 파악하고, 이 중에서 중요한 속성을 골라내는 작업은 학습 방법이 자동적으로 수행하도록 한다.
(2) 훈련자료가 변화함에 따라서 속성 가중치를 동적으로 변경시킬 수 있다.

속성 가중치 부여 방법에 대한 많은 연구들[Kira and Rendell, 1992; Cardie, 1993; Aha and Bankert, 1994; Kim and Shin, 1998]에서는 생성된 속성의 가중치가 전역적으로(Globally) 사용되고 있다. 전역적으로 관련성이 있는가를 기준으로 속성의 유용성이 판단되기 때문에 사례 공간의 많은 부분에 대하여 어떤 속성의 관련성이 크건 작건 간에 성립되기는 하지만, 어떤 특정한 부분에 대해서는 관련성이 전혀 없어서 혼란이 발생될 수 있다. 즉, 어떤 속성은 어떤 상황(Context)하에서만 관련성이 있다는 사실을 [Domingos, 1997] 무시하고 있는 것이다. 비록 Domingos가 지역적(Local) 속성 선정 방법인 RC[Domingos, 1997]를 제안하긴 했으나, 이전의 연구들 중 지역적 속성 가중치 부여에 대한 것은 매우 희소하다.

본 연구에서는 CBR 기법의 속성들에 가중치를 부여하는 새로운 방법인 CBDFW(Case-Based Dynamic Feature Weighting)를 개발하였다. CBDFW는 무작위로 생성된 속성 가중치 벡터들을 과거에 사용한 실험결과를 사용하여, 새로운 질의 사례에 대한 CBR 수행 시에 새로운 속성 가중치 벡터들을 동적으로 생성시킨다. 제 2절에서는 본 연구에 사용된 CBR 시스템을 설명한다. 제 3절에서는 속성 가중치 부여 방법을 분류하는 기존 프레임워크를 지역적 속성 가중치 부여 방법을 포함할 수 있도록 확장시킨다. 이어서 대표적인 속성 가중치 부여 방법들을 살펴본다. 제 4절에서는 사례별 속성 가중치 부여 방법인 CBDFW를 소개한다. 제 5절에서는 CBDFW를 신용평가 문제에 적용한 결과를 제시하고, 마지막으로 제 6절에서 본 연구의 결론과 한계로부터 향후 연구 방향을 모색한다.

2. 사례기반 추론 시스템

인공지능 분야의 하나인 CBR은 기억 장치인 사례 베이스에서 현재의 문제와 유사한 과거의 문제를 찾고, 과거의 문제와 현재의 문제간의 차이점을 분석하여 과거 문제의 해법을 현재의 문제에 알맞게 수정하여 문제를 풀어 가는 기법이다[Riesbeck and Schank, 1989]. CBR이 갖는 기본적인 아이디어는 인간이 과거의 문제를 해결하기 위해 사용한 해법을 수정하여 새로운 문제의 해결에 사용한다는 것이며, 이와 같은 문제 해결 과정의 재사용을 통하여 자동적인 학습이 가능해진다. 특히 CBR 시스템에서는 지식을 사례의 형태로 저장하기 때문에 기존의 인공지능 기법의 문제점으로 지적 되어온 지식 획득 병목 현상의 문제를 완화할 수 있다는 장점이 있다. Aamodt and Plaza[1996]는 <그림 2-1>과 같이 CBR 과정을 4R이라고 부르는 4 단계로 나누어 보았다.



<그림 2-1> Aamodt and Plaza의 사례기반 추론 과정

1) 검색(Retrieval)

검색은 새롭게 입력된 질의 사례와 가장 유사한 과거의 사례를 찾는 것이다. 검색은 속성 확인, 탐색, 초기 일치, 선택의 순서로 이루어진다. 속성 확인에서는 질의 사례의 어떤 속성 집합을 유사 사례 탐색에 사용할 것인지를 결정하여 탐색의 기준으로 사용한다. 초기 일치는 주어진 유사성의 기준을 넘는 사례들을 제시하는 것이다. 그리고 선택에서는 가장 유사한 사례를 선택한다.

2) 재사용(Reuse)

검색된 사례가 제공하는 해를 재사용할 때에는 두 가지 측면을 고려하여야 한다. 첫째는 질의 사례와 과거 사례간의 차이이고, 둘째는 검색된 사례의 어떤 부분이 질의 사례로 전이될 것인가이다. 재사용의 방법은 검색된 사례의 해를 그대로 복사해 사용하는 방법과 검색된 사례에서 해를 유도하는 방법을 재사용하여 해를 도출하는 방법으로 나눌 수 있다.

3) 수정(Revise)

재사용 단계에서 제공된 해가 질의 사례의 해로서 적합하지 않았을 때에는 발생한 실패로부터 학습할 기회가 발생한다. 이러한 상태를 수정이라고 부르는데 이것에는 두 가지 경우가 있다. 첫째는 재사용에 의해 제안된 해를 평가하여, 만약 성공적이면 성공으로부터 학습하는 경우이고 둘째는 영역 특성 지식을 사용하여 해를 수정하는 것이다.

4) 유지(Retain)

이 과정은 질의 사례에 대해 제안된 해를 지식으로 유지하기 위해 유용한 것들을 합치는 것이다. 제안된 해의 성공 또는 실패로부터 얻어진 해에 대한 평가와 해에 가해진 수정에 의해 연쇄 작용이 학습으로 일어난다. 유지 과

정에는 유지할 정보의 선택, 색인의 부여 여부, 사례 베이스에의 저장 방법 결정 등이 포함된다.

본 연구의 초점은 위의 4 단계 중에서 특히 제1단계인 검색에 맞추어져 있다. 사례의 각 속성에 어느 정도의 가중치가 부여되어 있냐에 따라서 검색되는 유사 사례가 판이하게 달라질 수 있다. 그러므로, 적절한 가중치의 부여는 CBR 시스템의 성능에 거의 절대적인 영향을 미친다고 할 수 있다. 물론, CBR 시스템을 구축할 때에는 속성 가중치 부여 뿐만 아니라, 유사도 산출 방식, 도출된 해의 적응 방법 등 고려할 요인들이 많다. 하지만, 본 연구에서 모든 경우를 고려한 사례기반 추론 시스템을 연구 대상으로 할 수는 없었다. 본 연구에서 사용되는 CBR 시스템에서는 아래와 같은 단순한 k-최근접이웃(k-NN: k-Nearest Neighbors) 유사도 산출방식을 사용한다.

수치형 속성:

$$\text{유사도} = 1 - \frac{|\text{질의 사례의 속성값} - \text{과거 사례의 속성값}|}{\text{해당 속성의 최대값}}$$

범주형 속성:

유사도=1, (질의 사례의 속성값 = 과거 사례의 속성값)인 경우
=0, 나머지 경우

값이 결측된 속성의 유사도는 0이 아닌 0.5로 정의한다. 0은 가장 먼 거리를 의미하기 때문이다. 이는 확인되지 않은 속성의 값에 대하여 벌점을 부과하지 않으려는 의도이다.

CBR은 여러 유형의 문제들에 적용될 수 있지만, 본 연구에서는 다음과 같은 특성들을 가진 전형적인 분류 문제들에 초점으로 맞추고자 한다.

(1) 문제들은 단속적인 출력 클래스를 가진다. 따라서 분류가 맞았는가를 확인함으로써 CBR 시스템의 성과를 판단할 수 있다.
(2) 문제들은 수치형과 범주형 속성 모두를 포함하며 비교적 많은 개수의 속성을 가진다.

그러므로, CBR의 적응 단계에서는 간단한 투표(Voting) 휴리스틱을 사용한다. CBR 시스템이 분류 문제를 해결하는 과정에서 복수의 사례를 조회하여 해를 구하는 경우에는 조회된 사례의 해 간에 갈등이 발생될 수 있다. 이때 빈도수가 가장 많은 클래스를 최종해로 결정하는 방식이 투표 휴리스틱 방법이다. 이러한 투표 휴리스틱 방법에 대한 여러 가지 변형도 존재한다 [Kolodner, 1993]. 본 연구에서는 동일한 가중치를 사용하는 투표 방식을 사용하였다.

3. 속성 가중치 부여 방법

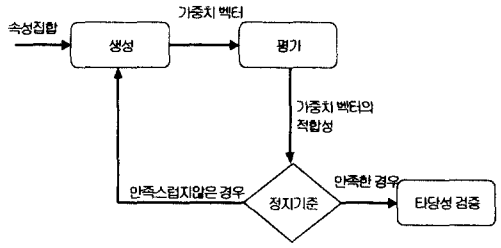
3.1 속성 가중치 부여 방법의 구분

속성 가중치 부여는 가장 높은 분류 적중률을 보이는 최적의 속성 가중치 벡터를 찾으려고 하는 것이다. 즉, 무한개의 후보 가중치 벡터들을 탐색하여 미리 마련해 놓은 평가 절차에 따라 최적의 가중치 벡터를 찾으려고 하는 것이다. 그러나 반드시 최적의 가중치를 찾고자 한다면 완전 탐색(Exhaustive Search)을 해야 한다는 문제점이 생긴다. 비록 속성의 개수가 적더라도 완전 탐색을 하게 되면 탐색 비용이 커지므로 실제적으로는 사용하기 어려울 수 있다. 이 때문에 휴리스틱에 기반하거나 무작위화된 탐색(Randomized Search)을 하는 다른 방법들은 탐

색 비용을 감소시키기 위해 성능을 희생하는 방식을 취한다. 이러한 방법들에서는 가중치 벡터들에 대한 무제한의 탐색을 하지 않도록 정지시키기 위한 규칙이 필요하다. 전형적인 속성 가중치 부여 방법에는 다음과 같은 기본적인 구성요소가 존재한다.

(1) 생성 절차 : 후보 가중치 벡터를 생성시킨다.
(2) 평가 절차 : 가중치 벡터를 평가한다.
(3) 정지 기준 : 정지 시점을 결정한다.
(4) 타당성 검증 절차 : 가중치 벡터의 타당성을 검증한다.

<그림 3-1>은 일반적인 속성 가중치 부여 과정을 보여준다.



<그림 3-1> 일반적인 속성 가중치 부여 과정

Dash and Liu[1997]는 속성 선정 방법에 대한 이차원 분류 프레임워크를 제안하였다. 그들의 프레임워크는 생성 절차와 평가 절차를 가장 중요한 차원으로 고려하여, 32가지의 대표적인 속성 선정 방법이 이에 따라 구분되었다. 가중치의 공간이 얼마나 일반화되는가를 '가중치의 유효 범위'라고 하는데[Aha, 1998], 그들이 제시한 프레임워크에는 가중치의 유효 범위가 차원으로 고려되지 않았다. 이에 우리는 전역적 속성 가중치 부여 방법과 지역적 속성 가중치 부여 방법을 구분하기 위하여 가중치의 유효 범위를 추가적인 차원으로 고려하였다. <표 3-1>은 수정된 프레임워크와 그에 따른 일부 대표적인 방법들의 구분을 보여 주고 있다.

<표 3-1> 삼차원 프레임워크에 의한 속성 가중치 부여 방법의 구분

가중치 유효 범위	평가 절차	생성 절차		
		휴리스틱	완전	무작위화
전역적	거리	Relief	+	-
	정보	DT	+	-
	종속성	+	-	-
	일관성	-	+	+
	분류의 오류	FSS,BSS,+	+	GA
지역적	분류의 오류	RC	-	CBDFW

+ : 여기에 제시되지 않은 방법들이 존재함
- : 존재하는 방법이 알려져 있지 않음

3.2 전역적 속성 가중치 부여 방법

전방향 순차적 탐색(FSS: Forward Sequential Search)과 역방향 순차적 탐색(BSS: Backward Sequential Search)은 가장 전형적인 속성 가중치 부여 방법이다. 이 방법들을 기본으로 한 변형들이 많이 알려져 있다[Aha and Bankert, 1994]. 이 방법들은 가중치의 초기치가 무엇이냐에 따라 구분된다. 생성 절차에서 모든 가중치 값들을 (1) 0에서, (2) 1에서, 또는 (3) 무작위로 생성된 값에서 출발할 수 있다. (1)의 방식을 취하는 방법들을 FSS라고 부르고, (2)의 방식을 취하는 방법들을 BSS라고 부른다. (1)의 방법에서는 속성 가중치들이 더 이상 분류 적중률을 개선시킬 수 없을 때까지 계속적으로 증가되고, (2)의

방법에서는 반대로 감소된다. (3)의 방법에서는 속성 가중치들이 증가 또는 감소 어느 쪽 방향으로도 변화될 수 있다.

Relief[Kira and Rendell, 1992]에서는 먼저 훈련 사례의 집합으로부터 사용자가 정한 개수만큼의 표본 사례들을 추출한다. Relief는 표본으로 추출된 각 사례에 대하여 'Near Hit'와 'Near Miss' 사례들을 유클리디언(Euclidean) 거리 척도를 기반으로 나머지 훈련 사례들로부터 검색한다. Near Hit은 주어진 사례와 동일한 클래스에 속하는 사례들 중 가장 적은 유클리디언 거리를 가지는 사례이고 Near Miss는 다른 클래스에 속하는 사례들 중 가장 적은 유클리디언 거리를 가지는 사례이다. 이 방법은 0으로 초기화되었던 속성들의 가중치로부터 출발하여 Near Miss와의 속성값에 차이가 존재하는 속성의 가중치는 양으로 증가시키고 Near Hit와의 차이가 존재하는 속성의 가중치는 음으로 감소시키는 과정을 반복하여 속성 가중치를 변화시킨다. 표본에 있는 모든 사례들을 다 사용하고 난 후 이 방법은 모든 속성들 중 가중치가 일정한 기준치 이상이 되는 것들을 선택한다. Relief는 잡음이 많거나 속성간의 상관관계가 큰 경우에 잘 작동되며, 속성의 개수와 표본의 개수에 대하여 수행시간이 선형적으로 증가한다. 이 방법의 한계점은 중복된 속성을 제거하지 못한다는 것과 사용자가 적절한 표본의 개수를 결정하기 어렵다는 점이다.

Cardie[1993]는 의사결정 나무(DT: Decision Tree) 생성 방법을 이용하여 속성 가중치를 부여하여 CBR의 성능을 개선할 수 있음을 보였다. C4.5[Quinlan, 1993]와 같은 의사결정 나무 생성 방법을 훈련 사례 집합에 대하여 수행하고, 가지치기를 한 후 의사결정 나무에 남은 속성들을 선정된 속성으로 간주한다. 이에 대해서는 여러 가지 변형이 가능한데, 예를 들면 의사결정 나무를 생성한 후 원래의 속성들에 엔트로피(Entropy) 값을 이용하여 가중치를 부여하는 것이다[Cardie and Howe, 1997].

유전적 알고리즘(GA: Genetic Algorithm)을 속성 가중치 부여에 활용하는 연구가 있었다[Yang and Honavar, 1996; Kim and Shin, 1998; Shin and Han, 1998]. Kim and Shin의 GA-kNN[1998]은 여러 가지 데이터 집합에 대한 단순한 CBR 모델과의 비교 실험에서 평균 분류 적중률을 63%에서 81%로 18% 포인트 가량 증가시키는 결과를 보였다. 그러나, GA를 사용하여 속성 가중치를 부여하려면 초기 모집단 규모, 세대수, 교배 확률, 돌연변이 확률 등의 여러 파라미터들의 값을 적절하게 부여해주는 것이 필요하다.

3.3 지역적 속성 가중치 부여 방법

전술한 속성 가중치 부여 방법들은 속성 가중치의 유효 범위가 전역적이다. 즉, 가중치는 사례 공간 전체에 대하여 유효하다. 반면, 속성 가중치를 지역적으로 설정하는 방법들은 속성 가중치를 서로 다른 사례공간의 부분에 대하여 서로 다르게 부여함으로써 보다 폭넓은 문제 유형을 다룰 수 있게 된다. 지역적 속성 가중치 부여 방법에는 클래스별 속성 가중치 부여[Aha, 1992; Howe and Cardie, 1997], 속성 값별 속성 가중치 부여[Stanfill and Waltz, 1986], 개별적인 사례별 또는 사례의 부분집합별 속성 가중치 부여 방법[Aha and Goldstone, 1992; Domingos, 1997] 등이 있다.

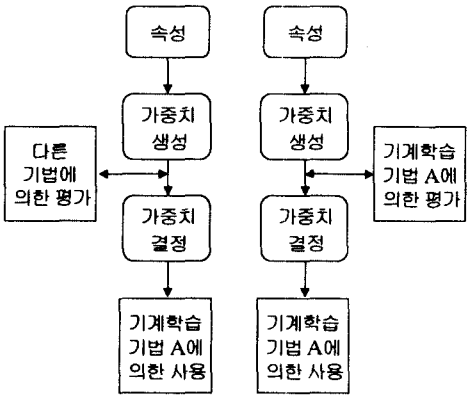
Domingos[1997]의 RC 방법은 사례별 속성 가중치 부여 방법이다. RC는 여러 면에서 BSS

와 구동 방식이 비슷하지만 전역적 방법이 아닌 지역적·사례별 속성의 관련성에 따라 가중치를 부여한다는 데 차이가 있다. 이 방법은 (1) 주어진 사례의 어떤 속성의 값이 가장 가까운 사례의 그 속성의 값과 다르고, (2) 그 속성을 제거해도 전체의 Leave-One-Out-Cross-Validation 오류(LOOCE)가 높아지지 않는 경우에, 사례로부터 그 속성을 제거한다. 속성들을 원래의 사례 집합으로부터 제거하면 중복된 사례가 생성될 수 있으나 이 사례들을 제거하지는 않는다. k=1의 k-NN을 사용하여, RC는 24개의 자료 집합에 대한 실험에서 FSS와 BSS에 비하여 효율성과 속성 관련성의 개선을 보여 주었다. 그러나, RC는 0 또는 1의 가중치만을 부여할 수 있는, 즉 속성의 선정만 가능한 방법이어서 연속적인 수치의 가중치를 허용할 수 있도록 확장하기가 어렵다.

3.4 평가 절차

평가 절차들은 분류의 결과로부터 피드백을 받는가의 여부에 의해 구분될 수 있다. 얻어진 속성 가중치들을 사용하게 될 기계 학습 기법 자신의 평가를 피드백으로 사용하는 방법을 Wrapper Model이라 부르고, 그 속성 가중치들을 사용하게 될 기계 학습 기법 자신으로부터 전혀 피드백을 받지 않는 방법을 Filter Model이라고 부른다[John *et al.*, 1994]. 이 두 모델의 분류는 원래 속성의 선정 과정에서 생성된 속성 부분집합을 평가하는 방식을 구분하기 위하여 John *et al.*이 제안한 용어이다. Wrapper Model은 비록 컴퓨터 자원을 훨씬 많이 사용하지만 기계 학습 기법 자신이 선정한 속성들을 사용하기 때문에 새로운 사례에 대한 분류 시에 적중률이 매우 높다[Dash and Liu, 1997]. 이 때문에 John *et al.*은 속성의 부분집합을 선정하는데 있어서 Wrapper Model의 사용을 Filter Model의 사용보다 권장하였고 일부 연구결과들이 분류의 적중률만을 목표로 하는 경우에는 이러한 주장이 타당함을 뒷받침하고 있다[Wettscherek *et al.*, 1997].

비록 John *et al.*이 제안한 두 모델이 속성 선정의 평가 절차에 대한 것이기는 하지만, 본 연구에서는 이를 속성 가중치 부여 방법의 평가 절차에 원용하고자 한다. 즉, <그림 3-2>와 같이 두 모델을 표현할 수 있다.



<그림 3-2> 속성 가중치 부여 방법의 평가 절차

4. 사례별 속성 가중치 부여

4.1 사례별 속성 가중치를 부여한 사례기반 추론

여기서는 본 연구에서 개발한 CBDFW 방법에 대해서 설명한다. CBDFW는 일종의 CBR이다. 즉, CBR 시스템의 속성 가중치를 부여하기 위해

서 또다른 CBR인 CBDFW를 사용하는 것이다. CBDFW의 기본 개념은 아래에 제시하는 바와 같이 간단하다.

> "무작위로 생성된 속성 가중치 벡터들의 과거 훈련 성능을 회상하여, 새로운 질의 사례에 대하여 CBR 시스템이 속성 가중치 벡터를 동적으로 생성한다."

CBDFW는 과거에 성능이 실험된 가중치 벡터들을 저장해 두었다가 새로운 질의 사례에 바람직한 속성 가중치 벡터를 CBR 수행 중에 동적으로 생성한다. CBDFW의 수행 방식은 RC의 방식과 유사하다. 두 방법 모두 입력되는 질의 사례의 상황에 따라 적합한 가중치를 생성하며 Wrapper Model이다. 그러나 CBDFW는 속성 가중치 부여 메커니즘으로 CBR을 사용하므로 RC와는 달리 사후학습적인(Lazy) 방식이다.

CBDFW 방법으로 가중치를 생성하여 CBR을 수행하는 시스템을 CBDFW-CBR이라고 명명한다. CBDFW-CBR의 절차에는 두 개의 구성요소가 있다. <그림 4-1>과 <그림 4-2>는 CBDFW-CBR 절차에 사용된 표기법 및 두 구성요소에 대하여 기술하고 있다.

n : 사례의 수
m : 속성의 수
p : 사례당 무작위로 생성되는 속성 가중치 벡터의 수
CB : 사례 베이스
Ci : 하나의 사례, Ci ∈ CB, i = 1, ... , n
CB-i : CB(Ci, 즉 CB-i는 CB에서 Ci 를 제거한 임시 사례 베이스, i = 1, ... , n
Wivj : i번째 사례의 v번째 속성 가중치 벡터의 j번째 속성의 가중치, i = 1, ... , n, v = 1, ... , p, j = 1, ... , m
Wiv* : i번째 사례의 v번째 속성 가중치 벡터, i = 1, ... , n, v = 1, ... , p
W**j : j번째 속성 의 가중치, j = 1, ... , m
I : Identity 벡터
CBR(X, Y, Z) : 주어진 질의 사례 X, 속성 가중치 벡터 Y, 사례 베이스 Z에 대하여 이 함수는 CBR 프로세스를 수행한 후, 분류 실패시에는 0을, 성공시에는 1을 반환한다.
Retrieve(X, Y, Z, K) : 주어진 질의 사례 X, 속성 가중치 벡터 Y, 사례 베이스 Z에 대하여 이 함수는 K개의 Nearest Neighbor 사례들을 조회하여, 그 사례들의 인덱스 집합을 반환한다.
NN(K) : K Nearest Neighbor 알고리즘에 의하여 선택된 사례들의 인덱스 집합
Combine(Y, R) : 이 함수는 R로 주어진 기준을 사용하여 Y에 주어진 가중치들을 합성한다.
Result(Q) : 주어진 질의 사례 Q에 대한 CBDFW-CBR 절차의 최종결과

<그림 4-1> CBDFW 절차에 사용된 표기법

```
Procedure PreProcessing( ):
    모든 가중치 Wivj들을 무작위로 생성된 0과 1 사이의 숫자로 초기화한다.
    For i = 1 to n
      For v = 1 to p
        Riv = CBR( Ci , Wiv* , CB-i )

Procedure RuntimeProcessing(Q):
    NN(K) = Retrieve( Q, I, CB, K)
    For j = 1 to m
      For i ∈NN(K)
        For v = 1 to p
          W**j = Combine( Wivj , Riv )
    Result(Q) = CBR( Q, W**j , CB )
```

<그림 4-2> CBDFW의 절차

CBDFW 절차는 다음과 같이 수행된다. 첫 번째 구성요소인 **Procedure PreProcessing( )**은 후보 가중치 벡터들을 만들어 놓는 과정이다. 먼저 무작위로 모든 사례에 속성의 가중치 벡터들을 생성한다. 그리고, 제 2절에서 설명했던 CBR( ) 절차에 생성된 가중치 벡터를 적용하여 사례베이스 내의 모든 사례를 하나씩 시험한 후 그 분류 결과가 성공한 경우에는 1을, 실패한 경우에는 0을 가중치 벡터별로 저장한다. 결과적으로 우리는 n개의 각 사례에 대해서 m 개의 속성값과 $m \times p$ 개의 가중치 값, 그리고 p개의 가중치 벡터 각각에 대한 시험에서 얻어진 p개의 시험결과를 사례 베이스 내에 유지하게 된다.

두 번째 구성요소인 **RuntimeProcessing(Q)**는 새로운 질의 사례 Q가 입력되는 시점에 실행된다. 이 절차는 Q에 대한 K개의 최근접이웃 사례들을 가중치에 대한 정보 없이 조회한 후에 조회된 사례들이 가지고 있는 각 속성의 가중치 값들을 종합하여 최종적으로 CBR( ) 함수가 사용할 속성 가중치 벡터를 산출한다. 조합 방법은 R_{iv}를 따르는데, R_{iv}가 1이면 성공적이었던 가중치 부여 사례만을 조회한다. 본 연구의 현재까지의 구현에서는 R_{iv}가 1인 경우만을 사용하고 있다. 예를 들어, m=4, p=2, K=3인 경우에 <표

4-1>과 같이 가중치 값들과 성공여부가 검색되었다고 하자.

<표 4-1> 검색된 가중치 값들과 성공여부

검색된 사례	속성				성공 여부
	속성 1	속성 2	속성 3	속성 4	
사례 1	0.42	0.19	0.25	0.33	0
	0.12	0.27	0.74	0.24	1
사례 2	0.08	0.18	0.69	0.34	1
	0.57	0.24	0.12	0.79	0
사례 3	0.31	0.48	0.17	0.05	0
	0.27	0.38	0.63	0.21	1

본 연구에서는 성공적이었던 가중치 벡터만을 사용하므로, <표 4-1>에서 2번째, 3번째 그리고 6번째의 가중치 벡터만을 사용하여 최종적으로 CBR( ) 함수가 사용할 속성 가중치 벡터를 산출한다. 예를 들어, 단순 평균으로 최종 속성 가중치 벡터를 산출한다면, 그 결과는 (0.16, 0.28, 0.69, 0.26)이 된다.

4.2 CBDFW의 장점

CBDFW는 Wrapper Model이다. 따라서 John *et al.*[1994]에서 설명된 Wrapper Model들이 가지는 일반적인 특성들을 가지고 있다. 또한 CBDFW는 다른 전역적인 Wrapper Model들뿐만 아니라 지역적 속성 가중치 부여 방법인 RC에 비해서도 상대적으로 빠르게 수행된다. 일반적인 CBR 시스템은 하나의 질의 사례를 해결하는데 $O(nm)$ 시간이 소요된다. CBDFW-CBR은 <그림 4-2>의 첫 번째 구성요소에서 $O(n^2m)$ 시간이 소요되고, 두 번째 구성요소에서는 가중치를 구하기 위한 사례의 조회에 $O(nm)$ 시간, 구한 가중치를 사용하여 분류를 수행하는데 $O(nm)$ 시간이 소요된다. 즉, 후보 가중치 벡터를 준비하는 첫 번째 구성요소의 수행 시간이 사례 베이스 크기의 제곱에 비례한다. 하지만, 후보 가중치 벡터의 준비는 매 질의 사례마다 수행하는 것이 아니고, 한 번 준비해 놓으면 모든 질의 사례에 대해서 일정 기간동안 수행할 필요가 없다. 다시 말해서, 일단 $O(n^2m)$ 시간을 소요하여 후보 가중치 벡터들이 준비되면, 그 다음에는 일반적인 CBR과 마찬가지로 매 질의 사례마다 $O(nm)$ 시간만이 소요되는 것이다.

다른 속성 가중치 부여 방법들에 비한다면 CBDFW는 매우 신축적이다. 하나의 사례에 대해서 여러 개의 후보 가중치 벡터를 저장해 놓을 수 있고, 속성 가중치 벡터 생성을 위해 유사 사례를 검색할 때에 거기에 딸린 여러 개의 후보 가중치 벡터를 가져올 수도 있고, 속성별 가중치를 산출하는 대신 속성의 선정 여부를 결정할 수도 있다. 현재의 CBDFW 버전은 연속적인 가중치를 사용하고 있으나 손쉽게 속성의 선정 여부를 결정하는 0 또는 1의 방식으로 변경시킬 수가 있다.

5. 실증적인 평가

이 절에서는 CBDFW의 유용성과 성능을 신용평가 데이터에 적용하여 평가해 본다. 본 연구의 주된 목표가 CBR 시스템의 분류 적중률을 향상시키는데 있기 때문에 우리는 CBDFW를 포함한 몇 가지 다른 가중치 부여 방법을 사용하여 CBR 시스템의 분류 적중률들을 측정하고 이것들을 서

로 비교한다.

신용 평가 문제는 고객의 신용도가, 예를 들어, 은행에서의 대출승인 결정을 내리기에 적합한 수준인가 아닌가를 판단하는 것이다. 이를 통해 은행이나 신용카드사와 같은 금융기관들은 잠재적인 문제를 가지고 있는 고객을 미리 선별해 내어 신용 위험을 감소시킬 수 있으므로 수익성을 개선시키는 효과를 얻을 수 있다. 이 문제 영역에서의 핵심적인 관심사는 대출 지원자 중 건전한 고객을 잃지 않으면서 문제 있는 고객을 어떻게 찾아내는가 하는 것이다. 이 문제는 전형적인 분류 문제이므로 이에 대한 응용 연구는 많이 찾을 수 있다. 예를 들면 Quinlan[1987]은 가지치기를 기반으로 한 의사결정 나무 생성 방법을 적용하여 82.6%~87.1%에 이르는 분류 적중률을 얻었다. 그는 가지치기 방법을 적용함으로써 가지치기를 사용하지 않은 방법에 비하여 3%~7% 포인트 정도의 분류 정확도 개선 효과를 얻었다고 보고하였다.

우리는 제 2절에서 소개된 CBR 시스템에 CBDFW를 적용하여 신용 평가를 위한 CBR 시스템인 CBDFW-CBR을 구현하였다. 우리가 평가에 사용한 데이터는 UCI Machine Learning Database Repository[Blake *et al.*, 1998]로부터 획득한 것으로서 Quinlan[1987]과 Domingos [1996]가 사용한 데이터와 동일한 데이터 집합이다. 이 데이터 집합에는 690개의 신용 평가 기록들이 포함되어 있다. 6개의 수치형 속성과 9개의 범주형 속성을 가지고 있어서 총 15개의 속성이 존재하는데, 일부 속성의 값에는 결측치가 포함되어 있다.

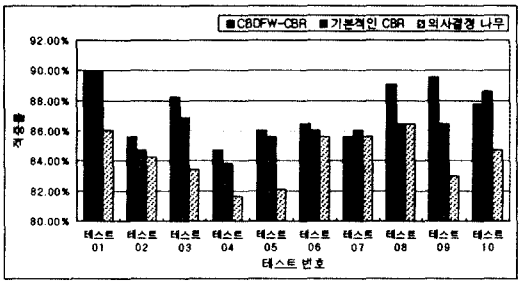
속성에 대한 가중치 정보를 사용하지 않는, 즉 모든 속성에 동일한 가중치가 부여된 CBR 시스템을 '기본적인 CBR'이라고 명명한다. 이것이 분류 적중률 비교의 기저(Baseline)가 된다. CBDFW-CBR, 기본적인 CBR, 의사결정 나무에 의한 속성 가중치 부여 방법, 그리고 기존 연구들의 성능과 비교한 결과는 <표 5-1>과 같다.

<표 5-1> CBDFW-CBR과 다른 기법과의 적중률 비교

	평균 적중률 (%)	표준 편차	비 고
CBDFW-CBR	87.3	1.9	K=5, p=3
기본적인 CBR	86.5	1.8	K=5
의사결정 나무	84.3	1.7	SAS E-Miner를 이용한 C4.5
Domingos[1996]	84.5	2.5	C4.5
Quinlan[1987]	82.6~87.1	-	가지치기를 이용한 ID3

<표 5-1>에서 K는 '최근접이웃 방법으로 검색해 오는 유사 사례의 개수'이고, p는 <그림 4-1>에 기술되었듯이 '사례당 무작위로 생성되는 속성 가중치 벡터의 수'이다. CBDFW-CBR과 기본적인 CBR에서 K값은 모두 5를 사용하였으며, CBDFW-CBR에서 p의 값은 3을 사용하였다. 의사결정 나무를 이용한 속성 가중치 부여의 적중률은 SAS E-Miner 시스템의 C4.5를 통하여 얻은 결과를 CBR에 적용하여 얻은 것이다.

우리의 실험에서는 하나의 데이터 집합으로부터 10개의 서로 다른 사례 베이스를 생성하여 실험결과의 안정성을 파악하는 방법(10-Fold Cross Validation)을 적용하였다. 10번의 테스트 각각의 적중률들은 <그림 5-1>과 같다.



<그림 5-1> 학습 방법에 따른 테스트 자료 집합별 적중률 비교

Domingos[1996]와 Quinlan[1987]도 같은 데이터 집합에 의사결정 나무를 적용하여 실험하였는데, Domingos는 C4.5를 사용하였고 Quinlan은 가지치기를 이용한 ID3를 사용하였다. <표 5-1>에서 볼 수 있듯이, CBDFW-CBR은 기본적인 CBR이나 다른 방법들에 비해 비교적 높은 분류 적중률을 보여 주고 있다. Quinlan의 논문에서는 평균 적중률과 표준편차를 제시하지 않고, 적중률의 범위를 82.6%~87.1%로 제시하고 있다. 이 중에 가장 좋은 적중률과 CBDFW-CBR의 평균 적중률인 87.3%와 비교하면 CBDFW로 인한 적중률 향상이 매우 적은 것 같이 인식된다. 하지만, CBDFW-CBR의 적중률도 범위로 나타내면, 84.7%~90.0%가 된다. 즉 범위로 표기된 적중률로 비교하면 CBDFW-CBR과 Quinlan의 연구 결과의 성능 차이가 적다고 할 수는 없다. <표 5-1>에서 하나 더 언급할 것은, 속성들의 가중치를 동일하게 부여한 기본적인 CBR의 평균 적중률이 비교적 높다는 것이다. 기본적인 CBR의 평균 적중률은 Domingos의 연구 결과보다 무려 2% 포인트나 높았다. CBDFW-CBR보다는 0.8% 포인트가 낮았지만 이것은 그리 큰 차이라고 볼 수는 없다. 이러한 결과로 볼 때에 신용 평가 데이터는 속성들간의 중요도에 별 차이가 없다고 판단할 수 있다.

6. 결론 및 연구의 한계

신축적이고 지역적인 속성 가중치 부여를 위한 연구들이 이루어져 왔으나, 그 중 Wrapper Model을 이용하는 지역적 속성 가중치 부여 방법에 대한 연구는 많이 이루어지지 않았다. 우리는 CBR의 속성 가중치 부여를 위한 새로운 지역적 Wrapper Model로서 CBR 자체를 활용하는 CBDFW라는 방법을 제안하였다. 이 방법은 비교적 간단하지만 Wrapper Model을 기반으로 한 방법 중 상대적으로 효율적이었다. 비록 아직까지는 광범위한 영역의 문제에 대한 실증적인 평가가 이루어지지는 못하였지만, 신용평가 문제에 대한 적용에서 기존의 방법들에 비하여 우수한 결과를 보여 줌으로써 그 유용성에 대한 가능성을 보여 주었다. CBDFW의 기반이 되는 사상이 "CBR을 해본 경험 자체를 재사용 한다"는 매우 자연스러운 것이기 때문에 우리는 다양한 문제영역에서 이 방법이 효과적일 수 있을 것으로 기대하고 있다.

본 연구의 한계점으로 다음의 몇 가지를 지적할 수 있다. 첫째, 여러 가지 서로 다른 문제 영역에 적용하지 못하였고, 둘째, 사례 베이스 내에 중복적인 속성이 많이 존재하는 경우에 대하여 실험하지 못하였고, 셋째, CBDFW 방법을 사용하는 다양한 CBR 설계 방법을 비교해 보지 못하였다. 그 외에도 향후 연구에서는 전역적인 방법과 지역적 방법을 혼합하는 방식의 가능성도 연구될 수 있을 것으로 판단된다.

참고문헌

Aamodt, A. and E. Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches," *Artificial Intelligence Communications*, Vol. 7, No. 1, 1996, pp.9-13.

Aha, D. W., "Generalizing from Case Studies: A Case Study," *Proceedings of the Ninth International Workshop on Machine Learning*, 1992. pp.1-10.

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

Aha, D. W. and R. L. Bankert, "Feature Selection for Case-Based Classification of Cloud Types: An Empirical Comparison," *Proceedings of AAAI-94 Workshop on CBR*, 1994.

Aha, D. W. and R. L. Goldstone, "Concept Learning and Flexible Weighting," *Proceedings of the Ninth National Conference on the Cognitive Science Society*, 1992, pp.534-539.

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

Cardie, C., "Using Decision Trees to Improve Case-Based Reasoning," *Proceedings of the Tenth International Conference on Machine Learning*, Morgan Kaufman, 1993, pp.25-32.

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

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

Dash, M. and H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis*, Vol.3, No.3, 1997, http://www.elsevier.nl/cite/show/.

Domingos, P., "Unifying Instance-Based and Rule-Based Induction," *Machine Learning*, Vol.24, 1996, pp.141-168.

Domingos, P., "Context-Sensitive Feature Selection for Lazy Learners," *Artificial Intelligence Review*, Vol.11, 1997, pp.227-253.

Howe, N. and C. Cardie, "Examining Locally Varying Weights for Nearest Neighbor Algorithms," *Case-Based Reasoning Research and Development: Second International Conference on Case-Based Reasoning*, Leake, D. and E. Plaza (eds.), Lecture Notes in Artificial Intelligence, Springer, 1997, pp.445-466.

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, pp.121-129.

Kim, S. H. and S. Shin, "Optimizing Retrieval of Precedents in Case-Based Reasoning through a Genetic Algorithm," '98 한국전문가시스템학회 추계 학술대회 논문집, 1998, pp.123-129.

Kira A. and L. A. Rendell, "A Practical Approach to Feature Selection," *Proceedings of The Ninth International Workshop on Machine Learning*, 1992, pp.249-256.

Kolodner, J., *Case-Based Reasoning*, Morgan

Kaufman Publishers, 1993.

Nelson, M. M., and W. T. Illingworth, *A Practical Guide to Neural Nets*, Addison-Wesley, 1991.

Quinlan, J. R., "Induction of Decision Trees," *Machine Learning*, Vol.1 No.1, 1986.

Quinlan, J. R., "Simplifying Decision Trees," *International Journal of Man-Machine Studies*, Vol.27, 1987, pp.221-234.

Quinlan, J. R., *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufman, 1993.

Shin, K. and I. Han, "A Hybrid Approach Using Case-based Reasoning and Genetic Algorithm for Corporate Bond Rating," '98 한국경영정보학회/한국전문가시스템학회 춘계 공동학술대회 논문집, 1998, pp.106-109.

Stanfill, C. and D. Waltz, "Toward Memory-Based Reasoning," *Communications of the ACM*, Vol.29, 1986, pp.1213-1228.

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, pp.273-314.

Yang, J. H. and V. Honavar, "Feature Subset Selection Using a Genetic Algorithm," *IEEE Intelligent Systems*, 1996, pp.44-49.

Abstract

A Dynamic Feature Weighting Method for Case-based Reasoning

Jae Sik Lee*
Yong Xune Xon**

Lazy learning methods including CBR have relative advantages in comparison with eager learning methods such as artificial neural networks and decision trees. However, they are very sensitive to irrelevant features. In other words, when there are irrelevant features, lazy learning methods have difficulty in comparing cases. Therefore, their performance can be degraded significantly. To overcome this disadvantage, feature weighting methods for lazy learning methods have been studied. Most of the existing researches, however, were focused on global feature weighting. In this research, we propose a new local feature weighting method, which we shall call CBDFW. CBDFW stores classification performance of randomly generated feature weight vectors. Then, given a new query case, CBDFW retrieves the successful feature weight vectors and designs a feature weight vector for the query case. In the test on credit evaluation domain, CBDFW showed better classification accuracy when compared to the results of previous researches.

Key words: Feature Weighting, Case-based Reasoning, Credit Evaluation.

* School of Business Administration, Ajou University
** x-solution consulting