사례기반 추론에서 사례별 속성 가중치 부여 방법

A Case-Specific Feature Weighting Method in Case-Based Reasoning

이재식*, 전용준**
*아주대학교 경영학부 교수, (Tel) 0331-219-2719, leejsk@madang.ajou.ac.kr
** 아주대학교 대학원 경영학과 박사과정, (Tel) 0331-219-2910, xyxon@hanmail.net
경기도 수원시 팔달구 원천동 산 5번지 (우: 442-749)

요 약

사례기반 추론을 포함한 Lazy Learning 방법들은 인공신경망이나 의사결정 나무와 같은 Eager Learning 방법들과 비교하여 여러 가지 상대적인 장점을 가지고 있다. 그러나 Lazy Learning 방법은 역시 상대적인 단점들도 가지고 있다. 첫째로 사례를 저장하기 위하여 많은 공간이 필요하며, 둘째로 문제해결 시점에서 시간이 많이 소요된다. 그러나 보다 심각한 문제점은 사례가 관련성이 낮은 속성들을 많이 가지고 있는 경우에 Lazy Learning 방법은 사례를 비교할 때에 혼란을 겪을 수 있다는 점이며, 이로 인하여 분류 정확도가 크게 저하될 수 있다. 이러한 문제점을 해결하기 위하여 Lazy Learning 방법을 위한 속성 가중치 부여 방법들이 많이 연구되어 왔다. 그러나 기존에 발표된 대부분의 방법들이 속성 가중치의 유효 범위를 전역적으로 하는 것들이었다. 이에 본 연구에서는 새로운 지역적 속성 가중치 부여 방법을 제안한다. 본 연구에서 제안하는 속성 가중치 부여 방법(**CBDFW**: 사례기반 동적 속성 가중치 부여)은 사례별로 속성 가중치를 다르게 부여하는 방법으로서 사례기반 추론의 원리를 속성 가중치 부여 과정에 적용하는 것이다. CBDFW의 장점으로는 (1) 수행 방법이 간단하며, (2) 논리적인 처리 비용이 기존 방법들에 비해 낮으며, (3) 신축적이라는 점을 들 수 있다. 본 연구에서는 신용 평가 문제에 CBDFW의 적용을 시도하였고, 다른 기법과의 비교에서 비교적 우수한 결과를 얻었다.

Key words: 사례기반 추론, 속성 가중치 부여, 신용 평가 응용

1. 서론

분류 문제는 주어진 과거 사례들로부터 학습을 한 후 새로운 사례가 입력되면 그에 닷는 클래스를 결정하는 문제이다. 이러한 문제를 풀기 위한 기계 학습 방법은 Eager Learning 방법과 Lazy Learning 방법의 두 종류로 구분된다. Eager Learning 방법은 사례 그 자체의 형태와는 다른, 클래스들에 대한 일반화시킨 결과를 명시적으로 표현한다. 반면, Lazy Learning 방법은 일반화를 분류를 실행하는 시점까지 지연시키며 새로운 사례와 가장 유사한 과거 사례를 찾아 문제를 해결한다. 사례기반 추론 방법(CBR)[Kolodner, 1993]을 비롯한 Lazy Learning 방법들은 의사결정 나무[Quinlan, 1986]나 인공신경망(ANN)[Nelson and Illingworth, 1991]등과 같은 Eager Learning 방법에 비하여 여러 가지 장점을 가지고 있다. 즉, 적은 정보만을 사용하는 경우에도 사례 공간에서 복잡한 의사결정 관계를 형성하는 것이 가능하고, 범주형 및 수치형 출력 모든 형태의 문제에 쉽게 적용이 가능하며, 사례를 단순히 저장해 두는 것으로 학습이 완료되기 때문에 학습과정이 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 Lazy Learning 방법은 그러한 속성으로 인해 혼란을 겪게 되며 결과적으로 정확도가 심히 저하된다. 이에 대한 자연스러운 해결책은 관련성이 낮은 속성을 찾아내서 이를 사용하지 않도록 하는 방법이다. 이러한 목적에서 여러 가지 방법들이 제안되었다[Aha, 1998].

이러한 방법들이 속성 선택이나 속성 가중치 부여 방법이다. 속성 선택 방법은 연관성이 없는 속성을 사례로부터 삭제시키는 반면 속성 가중치 부여 방법은 속성의 연관성 정도에 따라 다른 가중치를 부여한다. 가중치가 일정 수준 이하인 속성을 사용하지 않는다면 속성 가중치 부여는 속성 선택과 같은 효과를 가져오게 된다. 즉, 속성 선택을 일반화시키면 속성 가중치 부여가 된다고 할 수 있다.

속성 가중치 부여의 자동화는 많은 이점을 제공한다. Caruana and Freitag[1994]는 속성 선택의 장점에 대해 기술하고 있다. 이에 근거하여 우리는 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, 중요한 속성을 골라내는 작업은 학습시스템 자동적으로 수행하도록 한다.
(2) 훈련자료가 변화함에 따라서 속성 가중치를 동적으로 변경시킬 수 있다.

전방향 순차적 탐색(FSS: Forward Sequential Search)과 역방향 순차적 탐색(BSS: Backward Sequential Search)은 전형적인 속성 가중치 부여 방법이다. 이 방법들을 기본으로 한 변형들이 많이 알려져 있다[Aha and Bankert, 1994]. 이 방법들을 사용하여 우리는 학습시스템의 정확도를 향상시킬 수 있다. 그러나, 이러한 방법들에서 속성의 가중치는 전역적으로(Globally) 사용된다. 즉, 어떤 속성은 어떤 상황(Context)하에서만 관련성이 있다는 사실을[Domingos, 1997] 무시하고 있는 것이다. 전역적으로 관련성이 있는가를 기준으로 속성의 유용성이 판단되기 때문에 사례가 분포되어 있는 공간의 많은 부분에 대하여 성립되지만 어떤 부분에 대해서는 혼란이 발생될 수 있다.

비록 Domingos가 지역적(Local) 속성 선택 알고리즘인 RC[Domingos, 1997]를 제안 했으나 이전의 연구중 지역적 속성 가중치 부여에 대한 것은 매우 희소하다. RC는 어떤 속성들은 문제 공간의 일부 부분에서만 관련성이 있음을 인정하고 있다. BSS와 구동 방식은 비슷하지만 전역적 방법과는 다르게 지역적, 사례별 속성의 관련성에 대한 의사결정을 한다는 데 차이가 있다. 지역적 속성 가중치를 부여하므로 RC는 전역적 가중치 부여의 한계를 극복하고 개선된 분류 정확도를 보였다. RC에 대해서는 제 3절에서 보다 상세히 언급하도록 한다.

본 연구에서는 Lazy Learning 방법을 위한 속성 가중치 부여 방법인 CBDFW를 제시한다. CBDFW는 무작위로 생성된 속성 가중치 벡터들과 이를 사용한 시험 결과로부터 실행시에 새로운 속성 가중치 벡터들을 동적으로 생성시킨다. 제 2절에서는 이 연구에 사용된 사례기반 추론 시스템을 설명한다. 제 3절에서는 속성 가중치 부여 방법을 분류하기 위한 기존 프레임워크을 살펴보고 이를 지역적 속성 가중치 부여 방법을 포함할 수 있도록 확장 시킨다. 이어서 대표적인 속성 가중치 부여 방법들인 FSS, BSS, Relief[Kira and Rendell, 1992] 그리고 RC 등을 살펴본다. 제 4절에서는 사례기반 접근방법에 의한 동적 속성 가중치 부여 방법을 사용하는 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

본 연구에서는 다음과 같은 특성들을 가진 전형적인 분류 문제들에 초점으로 맞추고자 한다.

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

본 연구에서 사용되는 CBR에서는 아래와 같은 k-nearest neighbor(k-NN) 유사도 산출방식을 사용한다.

수치형 속성의 경우, 유사도 = 1-(거리 / 최대 거리), 단, 최소 유사도 = 0.
범주형 속성의 경우, 유사도 = 정확히 일치하는 경우 1, 기타의 경우에는 0

거리는 |입력 속성 값 - 사례 속성 값|을 통해 간단하게 계산된다. 결측된 속성 값의 다른 어떤 값으로 부터의 유사도는 0 대신 0.5로 정의된다. 0은 가장 먼 거리를 의미 하기 때문이다. 이는 확인되지 않은 속성의 값에 대하여 벌점을 부과하지 않으려는 것이다.

CBR의 적응 단계에서는, 일련의 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.

본 연구는 분류 문제에서의 속성 가중치 결정에 초점을 두므로 사례색인 부여 방법이나 정교한 적응 방법, 또는 수정 단계와 같은 부분은 고려하지 않았다.

3. 속성 가중치 부여 방법

3.1 속성 가중치 부여 방법의 구분

속성 가중치 부여는 가장 높은 분류 정확도를 보이는 최적의 속성 가중치 벡터를 찾으려 하는 것이다. 즉, 무한의 후보 가중치 벡터들을 탐색하여 어떤 평가 절차에 따라 최적의 가중치 벡터를 찾으려고 하는 것이다. 그러나 최적의 가중치를 찾으려 하는 경우에는 완전 탐색(Exhaustive Search)을 해야 하는 문제점이 생긴다. 그러므로 적은 수의 속성 집합 규모에 대해서도 비용이 커서 실질적으로는 사용하기 어려울 수 있다. 이 때문에 휴리스틱에 기반하거나 무작위 탐색을 하는 다른 방법들은 연산의 복잡성을 감소시키기 위해 성능을 희생하는 방식을 취한다. 이러한 방법들에서는 가중치 벡터들에 대한 무제한의 탐색을 하지 않도록 정지시키기 위한 규칙이 필요하다.

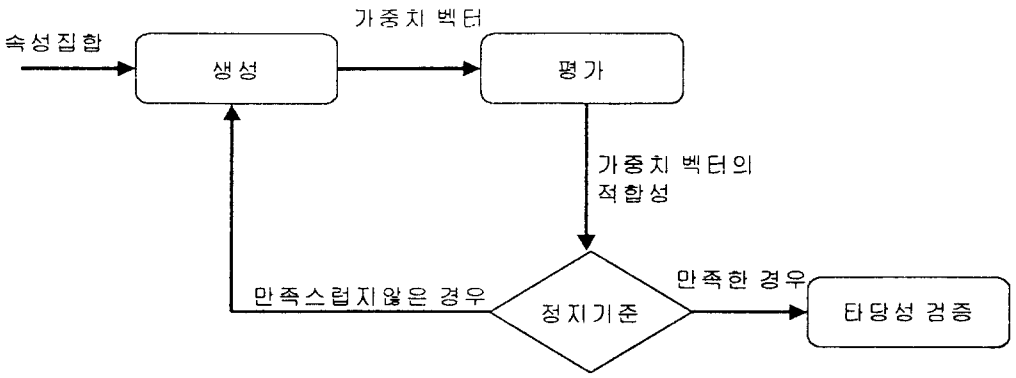
전형적인 속성 가중치 부여 방법에는 다음과 같은 기본적인 구성 요소가 존재한다.

(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.

비록 여러가지 속성 가중치 부여 방법을 구분하기 위한 프레임워크들이 제시되었으나[Dash and Liu, 1997; Aha, 1998], 우리는 Dash and Liu[1997]의 프레임워크를 기반으로 지역적 속성 가중치 부여 방법들을 포함시킬 수 있는 새로운 프레임워크를 제안한다.

Dash and Liu는 속성 선택 방법에 대한 이차원 분류 프레임워크을 제안하였다. 그들의 프레임워크는 생성 절차와 평가 절차를 가장 중요한 차원으로 고려하여, 32가지의 대표적인 속성 선택 방법이 이에 따라 구분되었다. 그러나, 가중치의 범위가 차원으로 고려되지 않았다. 이에 우리는 전역적 속성 가중치 부여 방법과 지역적 속성 가중치 부여 방법을 구분하기 위하여 가중치의 영향 범위를 추가적인 차원으



<그림 3.1> 일반적인 속성 가중치 부여 과정

로 고려한다. <표 3.1>은 수정된 프레임워크와 그에 따른 일부 대표적인 방법들의 구분을 보여 준다.

Relief[Kira and Rendell, 1992]와 의사결정 나무 생성 방법(DTI)[Cardie, 1993]은 휴리스틱 생성 절차를 사용한다. GA[Kim and Shin, 1998]는 속성 가중치 벡터를 무작위로 생성시킨다.

RC와 본 연구에서 제시하는 CBDFW는 지역 속성 가중치 부여 방법이다. RC는 휴리스틱 생성 절차를 사용하며 CBDFW는 무작위 생성 절차를 사용한다.

<표 3.1> 삼차원적 프레임워크에 의한 속성 가중치 부여 방법 구분

가중치 범위	평가 절차	생성 절차		
		휴리스틱	완전	무작위
전역적	거리	Relief	+	-
	정보	DTI	+	-
	종속성	+	-	-
	일관성	-	+	+
	분류의 오류	+	+	GA
지역적	분류의 오류	RC	-	CBDFW

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

3.2 전방향 순차적 가중치 부여 방법과 역방향 순차적 가중치 부여 방법

이 방법들은 가중치의 초기치가 무엇이냐에 따라 구분된다. 생성 절차에서 (1) 모든 가중치 값들을 0에서, (2) 모든 가중치 값들을 1에서, 또는 (3) 무작위로 생성된 가중치 값에서 출발할 수 있다. Aha and Bankert[1994]와 같이 (1)의 방식을 취하는 방법들을 전방향 방법(FSS)이라고 부르고, (2)의 방식을 취하는 방법들을 역방향 방법(BSS)이라고 부른다.

(1)의 방법에서는 속성 가중치들이 더 이상 개선될 수 없을 때까지 계속적으로 증가되고, (2)의 방법에서는 반대로 감소된다. (3)의 방법에서는 속성 가중치들이 증가 또는 감소 양쪽 방향으로 무작위적으로 변화된다.

3.3 Relief 알고리즘

Relief[Kira and Rendell, 1992]는 통계적 방법을 사용하여 관련성 있는 속성의 가중치를 결정한다. 먼저 훈련 집합의 사례로부터 표본 사례들을 선택한다. 사용자는 표본의 수를 결정 해야한다. Relief는 무작위로 사례의 표본들을 추출한 후, 각 사례에 대한 Near Hit와 Near Miss 사례들을 유클리디언(Euclidean) 거리 척도를 기반으로 선택한다. 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 0값으로 초기화되었던 속성들의 가중치로부터 출발하여 Near Miss와의 속성값에 차이가 존재하는 속성의 가중치를 양으로 증가시키고 Near Hit와의 차이가 존재하는 속성의

가중치를 음으로 감소시키는 과정을 반복하여 속성 가중치를 변화시킨다.

표본에 있는 모든 사례들을 다 사용한 후 이 방법은 모든 속성들중 가중치가 일정한 기준치 이상이 되는 것들을 선택한다. Relief는 잡음이 많거나 속성간의 상관관계가 큰 경우에 잘 작동되며, 속성의 수와 표본의 수에 대하여 수행시간이 선형적으로 증가한다. 한계는 중복된 속성을 제거하지 못한다는 것과 사용자가 적절한 표본의 숫자를 결정하기 어렵다는 점이다.

3.4 의사결정 나무 생성에 의한 속성 가중치 부여

Cardie[1993]는 의사결정 나무를 이용하여 속성 가중치를 부여하여 사례기반 추론의 성능을 개선할 수 있음을 보였다. C4.5[Quinlan, 1993]와 같은 의사결정 나무 생성 방법을 훈련 사례 집합에 대하여 수행하여 가지치기 후의 의사결정 나무에 남은 속성들을 선택된 속성으로 간주한다. 이에 대해서는 여러 가지 변형이 가능한데, 예를 들면 의사결정 나무를 생성한 후 원래의 속성들에 엔트로피(Entrophy) 값을 이용하여 가중치를 부여하는 것이다[Cardie and Howe, 1997].

3.5 유전적 알고리즘을 이용한 속성 가중치 부여

유전적 알고리즘(GA: Genetic Algorithm)을 속성 가중치 부여에 활용하는 여러 연구가 있었다[Shin and Han, 1998; Kim and Shin, 1998]. Kim and Shin의 GA-kNN[1998]은 여러 가지 데이터 집합에 대한 단순한 CBR 모델과의 비교 실험에서 평균적인 분류 정확도를 63%에서 81%로 18% 포인트 가량 증가시키는 결과를 얻었다. 그러나, GA를 사용하여 속성 가중치를 부여하는 방법을 적용하려면 최대 반복 횟수, 초기 모집단 규모, 교배 확율, 돌연변이 확율등의 여러 파라미터들의 값을 적절하게 부여해주는 것이 필요하다.

3.6 평가 절차

평가 절차들은 분류의 결과로부터 피드백을 받는가의 여부에 의해 구분 될 수 있다. 피드백을 사용하지 않는 방법들을 Filter Model이라 부르고, 기계 학습 방법 자신을 평가절차로 사용하는 방법을 Wrapper Model이라고 부른다[John *et al.*, 1994]. Wrapper Model은 비록 컴퓨터 자원을 훨씬 많이 사용하지만 분류기 자신이 선택한 속성들을 사용하기 때문에 새로운 사례에 대한 분류시에 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]. 그러나, Wrapper Model은 컴퓨터 자원의 사용이 더 많으므로 컴퓨터 비용이 문제가 되는 경우에는 Filter Model이나 또는 효율성을 개선한 Wrapper Model의 변형을 사용하는 것을 고려하여야 할 것이다.

3.7 지역 속성 가중치 부여 방법

가중치의 공간이 얼마나 일반화되는가는 가중치의 유효 범위이다[Aha, 1998]. 속성 가중치 설정을 하는 대부분의 알고리즘들은 이 범위가 전역적이며 그 가중치는 사례 공간 전체에 대하여 유효하다. 반면, 속성 가중치를 지역적으로 설정하는 알고리즘들은 속성 가중치를 서로 다른 사례공간의 부분에 대하여 서로 다르게 부여하므로써 보다 폭넓은 문제 유형을 다룰 수 있게 된다.

지역 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 방법등이 존재한다[Aha and Goldstone, 1992; Domingos, 1997].

Domingos[1997]의 RC 알고리즘은 사례별 속성 가중치 부여 방법이다. RC는 여러 면에서 BSS와 유사하지만 사례별로 속성의 관련성에 대한 의사결정을 내린다. 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 이를 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 또는1의 가중치 즉 속성의 선택만 가능한 방법이어서 연속적인 수치의 가중치를 허용할 수 있도록 확장하기 어렵다.

4. 사례별 속성 가중치 부여

4.1 사례별 속성 가중치를 부여한 사례기반 추론

여기서는 먼저 CBDFW 방법을 설명하고 이 방법을 사용하는 이유는 제 5절에서 설명한다. 또다른 CBR인 **CBDFW**를 사용하여 CBR 시스템의 속성 가중치를 부여하는 방법은 다음에 제시하는 바와 같이 간단하다.

> **"무작위로 생성된 속성 가중치들의 훈련시 성능을 희상하여 CBR 시스템이 입력 사례에 따른 속성 가중치 벡터를 동적으로 설계한다."**

CBDFW는 전처리된 결과를 저장해 두었다가 바람직한 속성 가중치 벡터를 실행중에 동적으로 생성한다. CBDFW의 수행 방식은 RC의 방식과 유사하다. 두 방법 모두 입력되는 사례의 상황에 따라 적합한 가중치를 생성하며 Wrapper Model이다. 그러나 CBDFW는 속성 가중치 부여 메커니즘으로 CBR을 사용하므로써 RC에 비하여 Lazy한 방식이다.

CBDFW의 절차에는 두 가지 구성요소가 존재한다. <그림 4.1>과 <그림 4.2>는 CBDFW 알고리즘에 사용된 표기법 및 두 구성요소에 대하여 설명한다. CBDFW 절차는 다음과 같이 수행된다. 새로운 질의 사례가 들어오면 첫 번째 구성요소(**Procedure PreProcessing( )**)가 수행된다. 먼저, 무작위로 모든 사례에 속성의 가중치 값을 생성한다. 그리고, 제 2절에서 설명했던 CBR( ) 절차에 생성된 가중치 벡터를 적용하여 사례베이스 내의 모든 사례를 하나씩 시험한 후 그 분류 결과 즉 성공한 경우에는 1을 실패한 경우에는 0을 사례별로 저장한다. 결과적으로

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 사용되는 표기법

우리는 *m* 개의 속성 값과 *m* 개의 가중치 값 그리고 하나의 시험 결과로 구성된 *n*개의 사례를 사례 베이스 내에 유지하게 된다.

두 번째 구성요소인 절차 **RuntimeProcessing(Q)**는 새로운 질의 사례 *Q*가 입력되는 시점에 실행된다. 이 절차는 *Q*에 대한 *K*개의 Nearest Neighbor 사례들을 가중치에 대한 정보 없이 조회한 후에 조회된 사례들이 가지고 있는 각 속성의 가중치 값들을 종합하여 최종적인 CBR( ) 함수가 사용할 속성 가중치 벡터를 산출한다. 조합 방법은 *Ri*를 따른다. *Ri*가 1이면 성공적이었던 가중치 부여 사례만을 조회한다. 본 연구의 현재까지의 구현에서는 *Ri*가 1인 경우만을 사용하고 있다.

본적인 CBR 시스템은 문제 해결에 *O(nm)* 시간이 소요된다. CBDFW-CBR은 훈련에 *O(n2m)* 시간과 문제 해결에 *O(nm)* 시간을 필요로 한다. 가중치를 구하기 위한 사례의 조회에 *O(nm)* 시간이 소요되며 새로운 문제를 풀기 위하여 *O(nm)* 시간이 필요하므로 전체 문제 해결을 위한 시간은 *O(nm)+O(nm)=O(nm)*이 된다. 그러나 이와 같은 방식의 수행시간 추정은 실제 자료 집합에 대한 실제 수행 시간을 그대로 반영하지는 않는다.

일부 다른 속성 가중치 부여 방법들에 비한다면 CBDFW는 신축적이다. 하나 이상의 사례를 해의 도출을 위해 사용할 수 도 있으며 하나의 속성 가중치 벡터 생성을 위해 여러 개의 과거 가중치 부여 경험을 사용할 수도 있고 속성별 가중치를 결정하는 대

```
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

4.2 CBDFW의 장점

CBDFW는 Wrapper Model이다. 따라서 John *et al.*[1994]에서 설명된 Wrapper Model들이 가지는 일반적인 특성들을 가지고 있다. 그러나 CBDFW는 다른 전역적인 Wrapper Model들뿐만 아니라 지역적 방법인 RC에 비해서도 상대적으로 빠르게 수행된다. 기신 속성의 선택여부를 산출할 수도 있다. 현재의 CBDFW 버전은 연속적인 가중치를 사용하고 있으나 손쉽게 속성의 선택여부를 산출하는 방식으로 변경시킬 수도 있다.

CBDFW의 속성 가중치 부여 능력은 현재 계속해서 시험 중이다. 그러나 여러 데이터 집합을 사용한 일부 실증적인 실험 결과에서 이 방법의 가능성이

발견되었다. 그 외에도 이 방법이 사용하는 알고리즘이 간단하고 자연스럽다는 점도 장점이라 할 수 있을 것이다.

5. 실증적인 평가

이 절에서 우리는 제 4절에서 제시한 새로운 속성 가중치 부여 방법 즉 CBDFW의 유용성과 성능에 대한 실증적인 시험의 결과를 살펴본다. CBDFW는 각 응용의 문제 구조가 분류 문제라는 공통점을 가지도록 설계된 서로 다른 몇 가지 문제영역들에 대해 적용된다.

본 연구의 주된 목표가 CBR 시스템의 분류 정확도를 향상시키는데 있기 때문에 우리는 CBDFW를 포함한 몇 가지 다른 가중치 부여 방법에 따른 CBR 시스템의 분류 정확도를 측정한다. 속성에 대한 가중치 정보를 사용하지 않는 CBR 시스템이 비교상의 기저(baseline)가 된다. 우리는 CBDFW 이외의 대안적인 속성 가중치 부여 방법으로 의사결정 나무 생성에 의한 속성 가중치 부여 방법을 사용한다.

5.1 신용 평가 응용

신용 평가는 고객의 신용도가 예를 들어 은행에서의 대출과 같은 어떠한 조치를 취하기에 적합한 수준인가를 판단하는 것이다. 이 문제는 분석적인 모델링의 응용 문제 영역으로 전형적으로 사용되어왔으며 많은 선행 연구들과 산업계 응용 노력이 이루어져 왔다.

은행이나 신용카드사와 같은 금융기관들은 잠재적인 문제를 가지고 있는 고객을 미리 선별해 내어 신용 위험을 감소시킬 수 있다. 신용 위험의 감소는 금융기관의 수익성을 개선시키는 효과를 가져오게 된다. 이 문제 영역에서의 핵심 이슈는 지원자중 건전한 고객을 잃지 않으면서 문제 있는 고객을 어떻게 찾아내는가 하는 것이다. 이 문제가 전형적인 분석 모델링 문제영역이므로 이에 대한 응용 연구는 많이 찾을 수 있다. 예를 들면 Quinlan[1987]은 가지치기를 기반으로 한 의사결정 나무 생성 방법을 적용하여 82.6%에서 87.1%에 이르는 분류 정확도를 얻었다. 그는 가지치기 방법을 적용하여 가지치기를 사용하지 않은 방법에 비하여 3%포인트에서 7%포인트 정도의 분류 정확도 개선 효과를 얻었다고 보고하였다.

우리는 제 3절에서 소개된 기본적인 CBR 시스템에 대하여 본 연구에서 제시하는 속성 가중치 부여 방법인 CBDFW를 사용하는 신용 평가를 위한 CBR 시스템을 구현하였다. 우리가 응용에 사용한 데이터는 UCI Machine Learning Database Repository[Blake *et al.*, 1998]로부터 획득된 것으로서 Quinlan[1987]과 Domingos[1996]가 사용한 데이터와 동일한 데이터 집합이다. 이 데이터 집합에는 690개의 실제 신용 평가 기록들이 포함되어 있다. 일부 속성 항목의 값에는 결측치가 포함되어 있으며, 6개의 연속형 속성과 9개의 범주형 속성을 가지고 있어서 총 15개의 속성이 존재한다.

기본적인 CBR과 CBDFW-CBR 그리고 의사결정 나무에 의한 속성 가중치 부여 방법을 비교한 결과는 <표 5.1>, <그림 5.1>과 같다.

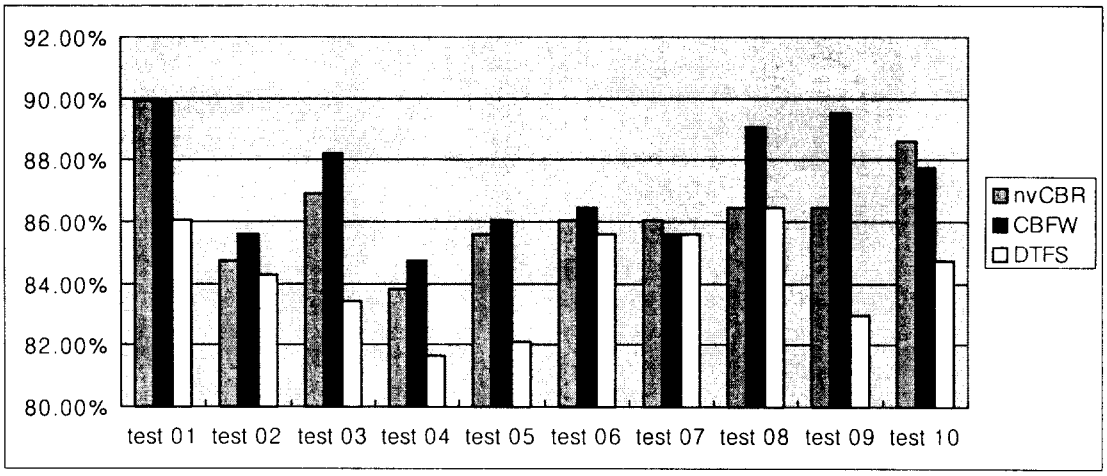
CBDFW-CBR과 기본적인 CBR 모두 k=5를 쓰고 있으며 CBDFW-CBR은 p=3의 인수를 사용하였다. DTFS는 의사결정 나무를 이용한 속성가중치 부여 방법으로서 SAS의 E-Miner 시스템을 이용하여 C4.5와 같은 구조를 구현하였고 그 결과를 CBR을 이용하여 실험한 결과이다. 실험을 위해서는 하나의 데이터 집합으로부터 10개의 서로 다른 사례 베이스를 생성하여 실험결과의 안정성을 파악하는 방법(10-Fold Cross Validation)을 적용하였다. Quinlan[1987]과 Domingos[1996] 역시 같은 데이터 집합에 의사결정 나무를 적용하였다. Domingos[1996]는 C4.5를 사용하였고 Quinlan[1987]은 가지치기를 이용한 ID3를 사용하였다. 실험 결과 CBDFW-CBR은 기본적인 CBR이나 다른 방법들에 비해 비교적 높은 분류 정확도를 보여 주었다.

6. 결론 및 연구의 한계

지금까지 신축적이고 지역적인 속성가중치 부여를 위한 일부 연구들이 이루어져 왔다[Aha and Goldstone, 1992; Domingos, 1997]. 그러나 그중 Wrapper Model을 이용하는 지역적 속성 가중치 부여 [Domingos, 1997] 방법에 대한 연구는 많이 이루어지지 않았다. 우리는 사례기반 추론의 속성 가중치 부여를 위한 하나의 새로운 지역적 Wrapper Model으로 사례기반 추론 자체를 적용하는 **CBDFW**라는 방법을 제안하였다. 이 방법은 비교적 간단하지만 Wrapper Model을 기반으로 한 방법중 상대적으로 효율적이었다. 비록 아직까지는 광범위한 영역의 문제에 대한 실증적인 평가가 이루어지지는 못하였지만, 신용평가 문제에 대한 적용에서 기존의 방법들에 비하여 우수한 결과를 보여 주므로 써 그 유용성에 대한 가능성을 보여 주었다. CBDFW의 기반이 되는 사상이

<표 5.1> 학습 방법에 따른 정확도 비교

	평균정확도(%)	표준편차	비 고
기본적인 CBR	86.46	1.77	k=5
CBDFW 사용 CBR	87.29	1.86	k=5, p=3
DTFS	84.28	1.68	SAS E-Miner를 이용한 C4.5
Domingos[1996]	84.5	2.5	C4.5
Quinlan[1987]	82.6~87.1		가지치기를 이용한 ID3



* nvCBR: 기본적인 CBR; CBFW: CBDFW 사용 CBR; DTFS: 의사결정 나무 이용한 방법

<그림 5.1> 학습 방법에 따른 시험 자료 집합별 정확도 비교

"CBR을 해본 경험 자체를 재사용 한다"는 매우 자연스러운 것이기 때문에 우리는 다양한 문제영역에서 이 방법이 효과적일 수 있을 것으로 기대하고 있다.

본 연구의 한계는 다음과 같은 몇 가지 측면으로 나누어질 수 있다. (1) 여러 가지 서로 다른 문제 영역에 적용하지 못하였고, (2) 사례 베이스내에 중복적인 속성이 많이 존재하는 경우에 대하여 시험하지 못하였고, (3) CBDFW 방법을 사용하는 충분히 다양한 CBR 설계 방법을 비교해 보지 못하였다. 그 외에도 향후 연구에서는 전역적인 방법과 지역적 방법을 혼합하는 방식의 가능성도 연구될 수 있을 것으로 판단된다.

참고문헌

Adriaans, P. and D. Zantinge, *Data Mining: Syllogic, Addison-Wesley*, 1996.

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

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

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.

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.

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. pp.97-108.

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," *Proceedings of the Korean Expert Systems Society Conference*, Fall, 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.

Lee, H. Y., "A Case-Based Forecasting System," *J. of the Korean OR/MS Society*, V.19 No.2, 1994. pp.199-216.

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 J. 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," *Proceedings of the KMISS/KESS Joint Conference*, Spring 1998. pp.106-109.