한국지능정보시스템학회논문지 제13권 제2호 2007년 6월(pp.1~14)

유사도 임계치에 근거한 최근접 이웃 집합의 구성*

이재식
아주대학교 경영대학 e-비즈니스학부
(leejsk@ajou.ac.kr)

이진천
유비쿼터스 컨버전스 연구소
(giny777@empal.com)

사례기반추론은 다양한 예측 문제에 있어서 성공적으로 활용되고 있는 데이터 마이닝 기법 중 하나이다. 사례기반추론 시스템의 예측 성능은 예측에 사용되는 최근접 이웃 집합을 어떻게 구성하느냐에 따라 영향을 받게 된다. 최근접 이웃 집합의 구성에 있어서 대부분의 선행 연구들은 고정된 값인 k개의 사례를 포함시키는 k-NN 방법을 채택해왔다. 그러나 k-NN 방법을 채택하는 사례기반추론 시스템은 k 값을 너무 크게 혹은 작게 설정하게 되면 예측 성능이 저하된다. 본 연구에서는 이러한 문제를 해결하기 위해 최근접 이웃 집합을 구성함에 있어서 유사도의 임계치 자체를 이용하는 s-NN 방법을 제안하였다. UCI의 Machine Learning Repository에서 제공하는 데이터를 사용하여 실험한 결과, s-NN 방법을 적용한 사례기반추론 모델이 k-NN 방법을 적용한 사례기반추론 모델보다 더 우수한 성능을 보여주었다

논문접수일 : 2005년 11월 게재확정일 : 2007년 04월 교신저자 : 이재식

1. 서론

데이터 마이닝은 다량의 데이터로부터 의미있는 패턴과 규칙을 찾아내기 위해서 자동적 또는 반자동적으로 데이터를 탐색하고 분석하는 과정이다[Berry and Linoff, 2004]. 데이터 마이닝에서 주로 다루어지는 문제 형태는 분류(Classification)와 예측(Prediction)이다. 분류란 새로운 사례(case)가 주어졌을 때 이 사례를 사전에 미리 정의된 범주값(class)들 중 하나에 할당하는 것이고, 예측이란 과거 사례의 분석을 통해 주어진 사례가 미래에 어떤 행동 또는 값을 가지게 될 것인가를 추정하는 문제이다. 사례기반추론(CBR : Case-based Reasoning)은 이러한 분류 및 예측 문제 모두에 효과적으로 적용 가능한 기계학습 기법이다. CBR은 Exemplar-based Reasoning, Instance-based Reasoning, Memory-based Reasoning 또는 Analogy-based Reasoning 등 다양한 용어로 사용되지만, 그 기본 개념은 유사하다[Chanchien and Lin, 2005]. CBR은 두 개의 기본 사상에 기반하는데 첫 번째는 유사한 문제는 유사한 해법을 가진다는 것이고, 두 번째는 발생한 문제는 자주 발생할 수 있다는 것이다. 따라서 과거에 현재의 문제와 유사한 문제가 존재하였고 그것이 어떻게 해결됐는지를 안다면, 과거의 경험을 바탕으로 현재 문제의 해결책을 추론할 수 있다는 것이다. CBR의 문제 해결

* 본 연구는 21세기 프론티어 연구개발 사업의 일환으로 추진되고 있는 정보통신부의 유비쿼터스컴퓨팅 및 네트워크 원천기반기술 개발사업의 지원에 의한 것임.

방식은 인간의 문제 해결 방식과 유사하기 때문에 그 결과를 이해하기 쉽고, 새로운 사례를 단순히 저장하는 것만으로도 추가적인 작업 없이 학습이 진행된다는 장점을 가진다.

CBR은 다양한 현실 문제 해결 즉, 고객 분류[Chiu, 2002], 파산 예측[Elhadi, 2000; Park and Han, 2002], 신용 평가[이재식과 전용준, 2001], 판매 예측[Chang and Lai, 2005], 의료 진단[Althoff et al., 1998; Marling and Whitehouse, 2001], 설비 고장 진단[Liao et al., 2000; Tsai et al., 2005; Wang and Wang, 2005], 헬프 데스크[Law et al., 1997; Goker and Roth-Berghofer, 1999], 전자 상거래[Vollrath et al., 1998] 그리고 전략 수립[Chanchien and Lin, 2005] 등에 성공적으로 적용되어 왔다.

CBR 시스템의 예측 성능은 다음과 같은 6개의 요소에 의해 영향을 받게 된다.

1) 사례베이스의 구성 방법.
2) 예측에 사용되는 속성 및 속성에 대한 가중치.
3) 유사도 측정에 사용되는 함수.
4) 최근접 이웃의 수.
5) 예측 결과의 생성 방법.

이와 같은 문제들에 대해 효과적인 CBR 모델을 구축하기 위한 다양한 방법들이 연구되어 왔다. 그러나 과거의 선행 연구들은 주로 1), 2), 3), 5) 영역에 집중되어 왔으며, 4)에 해당하는 최근접 이웃 집합의 구성에 관한 연구는 거의 이루어지지 않았다. 최근접 이웃 집합의 구성에 있어서 대부분의 선행 연구 모델들은 고정된 k 값을 사용하는 방식을 채택하고 있다. 이러한 고정된 k 값에 의한 최근접 이웃 집합의 구성은 k 값을 크게 설정할 경우 최근접 이웃 집합 안에 유사성이 낮은 사례들을 포함시킬 가능성을 증가시키기 때문에 예측 성능을 저하시키는 원인이 된다. 또한 k 값을 작게 설정할 경우에는 유사한 사례들 중 일부만을 가지고 예측을 수행하게 됨으로써 예측 성능을 저하시킬 수 있다. 고정된 k 값의 설정에 따른 이와 같은 문제들을 해결하기 위해 본 연구에서는, 최근접 이웃 집합을 구성함에 있어서 유사도 임계치(Similarity Threshold) 자체를 사용하는 새로운 방법인 s-NN (Similarity based Nearest Neighbors)을 제안한다. s-NN 방법에서는 각 사례마다 다른 개수의 최근접 이웃이 선정될 수 있다. 즉, k 값이 유사도 값에 따라 변하게 되는 것이다.

본 논문은 다음과 같이 구성 되었다. 제 2절에서는 CBR에 대한 소개와 CBR 모델의 성능 개선을 위해 시도되었던 선행 연구들에 대하여 고찰한다. 제 3절에서는 본 연구에서 제안하는 s-NN 방법을 소개하고, 제 4절에서는 s-NN 방법을 적용한 CBR 모델의 구현과 그 성능에 대해서 기술한다. 마지막으로 제 5절에서는 본 연구의 결론과 향후 연구 방향을 제시한다.

2. 사례기반추론

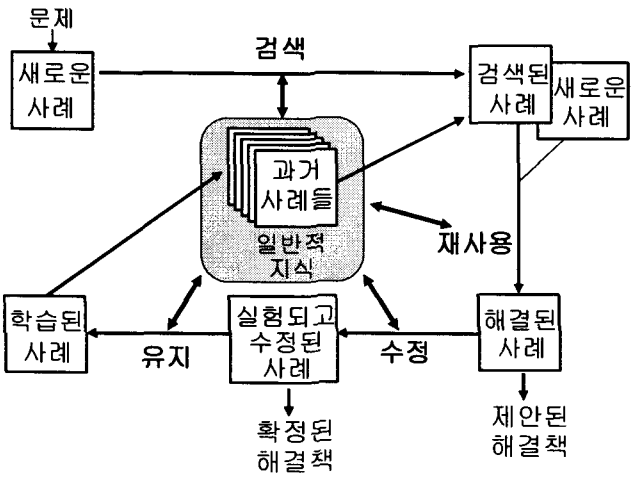
Aamodt and Plaza[1994]는 CBR의 문제 해결 과정을 [그림 1]과 같이 크게 검색, 재사용, 수정, 유지의 4단계로 구분하였다.

2.1 검색(Retrieve)

현재 문제와 가장 유사한 과거 사례들을 사례베이스로부터 찾아내는 것이다.

2.2 재사용(Reuse)

검색을 통해 찾아진 유사 사례들의 해법을 현재 문제 해결을 위해 사용하는 것이다.



[그림 1] Aamodt and Plaza의 사례기반추론 과정

2.3 수정(Revise)

현재 문제의 해결을 위해 검색된 유사 사례들의 해법을 현재 문제에 적합한 형태로 조정 하는 것이다.

2.4 유지(Retain)

새롭게 해결된 문제와 해법을 향후 새로운 문제 해결을 위한 목적으로 사례베이스에 저장하는 것이다.

사례베이스로부터 유사 사례를 찾기 위한 검색 방법으로는 귀납적 검색(Inductive Retrieval)과 최근접 이웃 검색(Nearest Neighbor Retrieval)이 있다. 귀납적 검색은 사례를 가장 잘 구분시켜주는 속성들을 찾아서 이 속성들을 사용하여 유사 사례를 검색 하는 방법이다. 귀납적 검색은 사례의 검색 및 구성을 위해 의사결정나무 형태의 구조를 사용한다. 최근접 이웃 검색은 현재의 새로운 사례와 유사한 사례를 검색하기 위해서, 새로운 사례와 사례 베이스에 있는 과거의 모든 사례와의 유사도를 측정함으로써 유사 사례를 찾는 방법이다. 최근접 이웃 검색의 경우 일반적으로 가장 많이 사용되는 방법은 새로운 사례와 가장 유사한 k개의 과거 사례를 검색해 주는 k-NN(k Nearest Neighbors) 방법이다. 일반적으로 사용되는 유사도 측정의 식은 식 (1)과 같다.

$$Similarity(N, C) = \frac{\sum_{i=1}^{n} f(N_i, C_i) \times W_i}{\sum_{i=1}^{n} W_i} \quad (1)$$

N : 새로운 사례.
C : 사례베이스에 저장된 과거 사례.
n : 사례가 가지는 속성의 개수.
N_i : 새로운 사례의 i번째 속성값.
C_i : 과거 사례의 i번째 속성값.
$f(N_i, C_i)$: N_i와 C_i 사이의 거리 측정 함수.
W_i : i번째 속성에 대한 가중치.

사례간의 유사도는 일반적으로 '0'에서 '1'사이

의 정규화된 실수 값으로 표현되는데, '0'에 가까울수록 두 사례의 유사성이 낮다는 것을 의미하고, '1'에 가까울수록 유사성이 높다는 것을 의미한다. 본 연구에서도 식 (1)을 사용하여 사례간의 유사도를 측정하였다.

CBR 모델의 예측 성능 개선을 위한 다양한 연구들이 시도되었다. 특히, 최적 사례베이스의 구성, 속성의 선정(Feature Selection) 및 속성 가중치 부여(Feature Weighting), 다중모델(Hybrid Model)의 설계 등은 CBR 모델의 예측 성능 개선을 위해 수행된 대표적인 연구 분야들이다.

최적 사례베이스의 구성은 문제 해결에 유용한 사례들만을 선별하여 사례베이스를 구성함으로써 예측 성능을 개선시키고, 사례의 저장공간을 최소화시키는데 있다[Smythe, 1998; Weiss and Indurkhya, 1998]. Brighton and Mellish[2002]는 효과적인 사례베이스를 구성하기 위한 방법으로 사례베이스로부터 중복된 사례와 유해한(harmful) 사례를 제거하는 Instance selection 방법을 제안하였다. Lee and Cho[2005]는 사례베이스로부터 유해한 사례뿐만 아니라 무관한(irrelevant) 사례들도 제거함으로써 사례베이스의 크기를 66%까지 축소하였다.

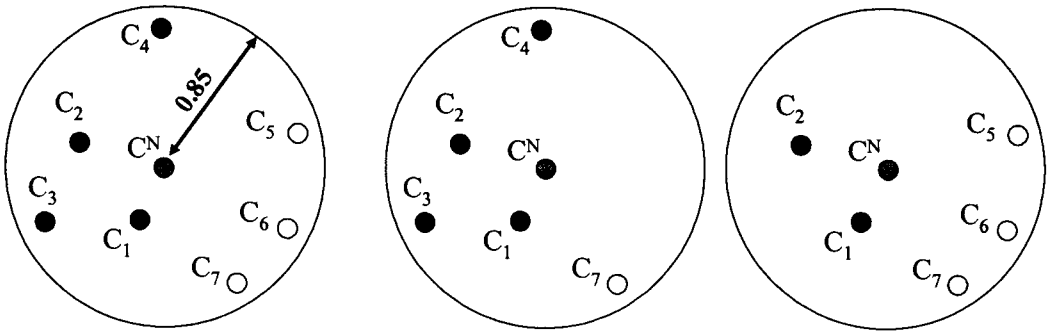
속성 선정 및 속성 가중치 부여는 문제 해결에 있어서 관련성이 낮은 속성들을 제거시키고, 예측에 사용되는 속성에 대해서는 중요도에 따라 적절한 가중치를 부여함으로써 모델의 예측 성능을 개선하기 위한 것이다[Aha and Bankert, 1994; Dash and Liu, 1997; Aha, 1998; 이재식과 전용준, 2001]. 속성 가중치 부여는 문제 해결에 있어서의 각 속성의 중요도에 따라 다른 가중치를 부여하는 것으로서, 특정 속성의 가중치를 '0'에 가깝게 설정할 경우 그 속성을 선정하지 않는 것과 같은 효과를 가져오게 된다. 따라서 속성 가중치 부여는 속성 선정의 일반화된 형태로 볼 수 있다.

CBR 모델의 예측 성능 개선을 위한 또 다른 시도로는 다중모델의 구축이 있다[Shen and Fu, 2004; Chang and Lai, 2005; Chun and Park, 2005; Wang and Wang, 2005; 이재식과 이진천, 2006]. 다중모델의 사용 목적은 하나의 문제를 해결하기 위해 둘 이상의 모델들을 결합하여 사용함으로써 하나의 모델을 사용할 때보다 더 좋은 예측 성능을 얻고자 하는데 있다. Chun and Park[2005]은 주가 예측 문제에 있어서 다중모델의 한 형태인 CBR 앙상블(Ensemble) 구축 방법을 제안하였으며, 이재식과 이진천[2006]은 자동차 보험 이탈 고객을 예측하기 위한 CBR, 의사결정나무 그리고 인공신경망의 다중모델을 제시하였다.

3. 최근접 이웃 집합의 구성을 위한 유사도 임계치의 사용

최근접 이웃 집합의 구성을 위해 본 연구에서는 고정된 k 값을 사용하는 방식이 아닌, 유사도 임계치(Similarity Threshold) 자체를 사용하는 s-NN 방법을 제안한다. s-NN은 사전에 유사도 임계치 값을 미리 설정하여 놓고, 사례베이스의 사례들 중에서 그 유사도 값이 이 임계치보다 큰 사례들을 모두 모아서 최근접 이웃 집합을 구성하는 방법이다. s-NN의 기본적인 아이디어는 새로운 사례와 일정 수준 이상 유사하다고 판정된 모든 사례를 사용하여 문제를 해결함으로써 예측 성능을 개선시키고자 하는 것이다.

[그림 2]는 최근접 이웃 집합을 구성할 때에 고정된 k 값을 사용함으로써 발생할 수 있는 예측 오류의 상황을 보여주고 있다.



새로운 사례와 유사 사례들　　5-NN 집합 A　　5-NN 집합 B

[그림 2] k = 5일 때의 최근접 아웃 집합의 두 가지 경우

[그림 2]에서, 각 작은 원들은 사례를 나타내고, 원의 색깔은 목표 속성의 값을 표시한다. 가운데에 있는 회색 원, C^N은 새로운 사례로서 이 사례는 '검은' 사례로 분류되어야 한다. [그림 2]에서 보듯이, 새로운 사례 C^N과 유사한 사례의 순서는 $C_1 \rightarrow C_2 \rightarrow C_3 = C_4 = C_5 = C_6 = C_7$과 같다. 즉, C_1이 가장 유사하고, C_2가 두 번째로 유사하고, 그 다음 C_3부터 C_7까지의 사례는 모두 동일한 유사도 값을 가지고 있다.

k 값이 5인 k-NN 방법으로 최근접 이웃 집합을 구성한다고 가정해 보자. 이 때에는 최근접 이웃 집합에 어떤 사례가 포함되느냐에 따라서 새로운 사례의 색깔 예측이 달라지게 된다. [그림 2]의 가운데 그림 즉 '5-NN 집합 A'에서 보듯이, C_1, C_2, C_3, C_4, C_7이 최근접 이웃 집합에 포함되면 C^N은 '검은' 색깔로 예측될 것이다. 하지만, [그림 2]의 오른쪽 그림 즉 '5-NN 집합 B'에서 보듯이 C_1, C_2, C_5, C_6, C_7이 최근접 이웃 집합에 포함되면 C^N은 '하얀' 색깔로 예측될 것이다. 이러한 비일관성은 고정된 k 값을 사용하여 최근접 이웃 집합을 구성했기 때문이다. 즉, 동일한 유사도 값을 가지는 사례들이 여러 개 있을 때에 어느 사례가 최근접 이웃 집합에 포함되느냐는 일반적으로 사례들이 사례베이스에 저장된 순서에 따라 결정될 수 밖에 없기 때문에 [그림 2]와 같은 두 가지 상이한 예측결과를 주는 상황이 발생할 수 있는 것이다.

이러한 비일관성 문제는 본 연구에서 제시하는 s-NN 방법을 적용함으로써 해소할 수 있다. 유사도 임계치를 0.85로 설정했다고 가정해 보자. 그러면, [그림 2]의 왼쪽 그림에서 보듯이, 유사도 값이 0.85보다 크거나 같은 모든 사례들, 즉 C_1부터 C_7까지의 모든 사례가 최근접 이웃 집합에 포함된다. 이 경우에 새로운 사례 C^N은 '검은' 사례로 올바르게 예측될 것이다. 그러므로 일관성이 유지된다. s-NN 방법을 적용하는 과정은 다음과 같다.

제 1단계 : 유사도 임계치와 k의 디폴트(Default) 값을 설정한다.

1.1 최근접 이웃 집합을 구성할 때에 사용할 유사도 임계치를 설정해 놓는다.

1.2 제 2단계에서 측정된 과거 사례의 유사도 값들이 모두 유사도 임계치보다 작은 경우에 대비하기 위하여, 제 5단계에서 적용할 전통적인 k-NN 방법에 사용할 k 값을 설

정해 놓는다.

제 2단계 : 유사도 값을 계산한다.
새로운 사례와 사례베이스 내의 과거 사례들간의 유사도 값을 계산한다.

제 3단계 : 평가.

3.1 과거 사례들을 유사도 값의 내림차순으로 정돈한다.

3.2 유사도 임계치보다 큰 유사도를 가진 사례가 있는지 검토한다.

3.3 만일 있다면 제 4단계로 진행하고, 없다면 제 5단계로 간다.

제 4단계 : 유사도 임계치를 사용하여 최근접 이웃 집합을 구성하고, 예측을 수행한다.

4.1 유사도 임계치보다 크거나 같은 유사도를 가지는 사례들을 선정하여 최근접 이웃 집합을 구성한다.

4.2 선정된 사례들을 목표 속성의 값에 따라 그룹을 만든다.

4.3 각 그룹내 사례들의 유사도 합계를 계산하다.

4.4 유사도 합계의 값이 가장 큰 그룹의 목표 속성 값을 새로운 사례의 목표 속성 값으로 결정한다.

4.5 만일 유사도 합계의 값에 동점이 발생하면, 유사도 값이 큰 사례를 포함하고 있는 그룹의 목표 속성 값을 새로운 사례의 목표 속성 값으로 결정한다.

4.6 만일 유사도 값에 동점이 발생하면, 먼저 검색된 사례를 포함하고 있는 그룹의 목표 속성 값을 새로운 사례의 목표 속성 값으로 결정한다.

제 5단계 : 디폴트 k 값을 사용하여 최근접 이웃 집합을 구성하고, 예측을 수행한다.

5.1 유사도 값이 큰 k개의 사례들을 선정하여 최근접 이웃 집합을 구성한다.

5.2 다수결 원칙에 근거하여 새로운 사례의 목표 속성 값을 결정한다.

본 연구에서 제안하는 s-NN 방법과 전통적인 k-NN 방법의 차이를 예를 들어 설명해 보자. 지금 5개의 새로운 사례들, C^N_1, C^N_2, C^N_3, C^N_4, C^N_5가 있고 모두 '검은' 사례로 분류되어야 한다고 가정하자. 각 사례에 대해서 유사한 사례들이 검색되

<표 1> 새로운 사례와 유사 사례들

새로운 사례	C^N_1			C^N_2			C^N_3			C^N_4			C^N_5		
유사도 내림차순으로 정돈된 과거의 유사사례들	C	S.S	T	C	S.S	T	C	S.S	T	C	S.S	T	C	S.S	T
	30	0.95	B	3	0.90	W	64	0.94	B	11	0.84	W	34	0.89	B
	4	0.92	W	9	0.88	B	57	0.93	W	73	0.83	W	44	0.87	B
	91	0.90	W	12	0.87	B	46	0.91	W	62	0.83	B	51	0.86	W
	93	0.90	B	6	0.85	W	40	0.89	B	83	0.82	B	4	0.81	W
	84	0.85	W	22	0.82	W	22	0.88	W	19	0.79	W	9	0.78	W
	60	0.85	B	27	0.81	B	34	0.86	B	88	0.72	W	53	0.78	W
	86	0.83	B	4	0.77	B	2	0.85	B	5	0.70	W	27	0.77	W
	41	0.80	B	35	0.73	W	12	0.85	B	59	0.68	B	42	0.76	B
	13	0.79	W	57	0.71	B	48	0.82	B	18	0.60	B	14	0.74	B
	25	0.77	W	17	0.67	B	9	0.74	B	8	0.51	B	13	0.70	W
	43	0.68	B	63	0.58	B	62	0.74	B	57	0.50	W	18	0.70	W
	24	0.65	B	31	0.58	B	71	0.72	W	21	0.49	W	1	0.68	B
	69	0.65	B	44	0.58	W	29	0.71	B	48	0.47	B	28	0.66	B
	6	0.64	W	25	0.56	W	7	0.69	B	77	0.44	W	64	0.65	W

<표 2> k-NN 방법의 결과(k = 5)

새로운 사례	C_1^N			C_2^N			C_3^N			C_4^N			C_5^N		
	C	S.S	T	C	S.S	T	C	S.S	T	C	S.S	T	C	S.S	T
유사도 내림차순으로 정돈된 과거의 유사사례들	30	0.95	B	3	0.90	W	64	0.94	B	11	0.84	W	34	0.89	B
	4	0.92	W	9	0.88	B	57	0.93	W	73	0.83	W	44	0.87	B
	91	0.90	W	12	0.87	B	46	0.91	W	62	0.83	B	51	0.86	W
	93	0.90	B	6	0.85	W	40	0.89	B	83	0.82	B	4	0.81	W
	84	0.85	W	22	0.82	W	22	0.88	W	19	0.79	W	9	0.78	W
	60	0.85	B	27	0.81	B	34	0.86	B	88	0.72	W	53	0.78	W
	86	0.83	B	4	0.77	B	2	0.85	B	5	0.70	W	27	0.77	W
	41	0.80	B	35	0.73	W	12	0.85	B	59	0.68	B	42	0.76	B
	13	0.79	W	57	0.71	B	48	0.82	B	18	0.60	B	14	0.74	B
	25	0.77	W	17	0.67	B	9	0.74	B	8	0.51	B	13	0.70	W
	43	0.68	B	63	0.58	B	62	0.74	B	57	0.50	W	18	0.70	W
	24	0.65	B	31	0.58	B	71	0.72	W	21	0.49	W	1	0.68	B
	69	0.65	B	44	0.58	W	29	0.71	B	48	0.47	B	28	0.66	B
	6	0.64	W	25	0.56	W	7	0.69	B	77	0.44	W	64	0.65	W

k = 5

어, <표 1>에 제시된 것과 같이 유사도 내림차순으로 정돈되었다. <표 1>에서 'C'는 과거 사례의 번호, 'S.S'는 유사도, 'T'는 목표 속성의 값을 나타낸다.

<표 2>는 k 값이 5인 k-NN 방법을 적용했을 때의 결과이다. 다수결 원칙에 의하기로 했기 때문에 모든 새로운 사례가 '하얀' 사례로 예측되었다. 즉, 모두 틀린 예측을 한 것이다.

<표 3>은 유사도 임계치를 0.85로 설정하고 s-NN 방법을 적용한 결과이다. <표 3>에서 보듯이, 새로운 사례의 목표 속성을 예측하기 위해서 사용되는 과거 사례의 개수, 즉 k-NN 방법에서의 k 값

<표 3> s-NN 방법의 결과(유사도 임계치 = 0.85)

새로운 사례	C_1^N			C_2^N			C_3^N			C_4^N			C_5^N		
	C	S.S	T	C	S.S	T	C	S.S	T	C	S.S	T	C	S.S	T
유사도 내림차순으로 정돈된 과거의 유사사례들	30	0.95	B	3	0.90	W	64	0.94	B	11	0.84	W	34	0.89	B
	4	0.92	W	9	0.88	B	57	0.93	W	73	0.83	W	44	0.87	B
	91	0.90	W	12	0.87	B	46	0.91	W	62	0.83	B	51	0.86	W
	93	0.90	B	6	0.85	W	40	0.89	B	83	0.82	B	4	0.81	W
	84	0.85	W	22	0.82	W	22	0.88	W	19	0.79	W	9	0.78	W
	60	0.85	B	27	0.81	B	34	0.86	B	88	0.72	W	53	0.78	W
	86	0.83	B	4	0.77	B	2	0.85	B	5	0.70	W	27	0.77	W
	41	0.80	B	35	0.73	W	12	0.85	B	59	0.68	B	42	0.76	B
	13	0.79	W	57	0.71	B	48	0.82	B	18	0.60	B	14	0.74	B
	25	0.77	W	17	0.67	B	9	0.74	B	8	0.51	B	13	0.70	W
	43	0.68	B	63	0.58	B	62	0.74	B	57	0.50	W	18	0.70	W
	24	0.65	B	31	0.58	B	71	0.72	W	21	0.49	W	1	0.68	B
	69	0.65	B	44	0.58	W	29	0.71	B	48	0.47	B	28	0.66	B
	6	0.64	W	25	0.56	W	7	0.69	B	77	0.44	W	64	0.65	W

k = 5

이 사례마다 다르다. 새로운 사례, C^N_1, C^N_2, C^N_3, C^N_4, C^N_5에 대해서 각각 6개, 4개, 8개, 5개, 3개의 유사 사례가 사용되었다. 사례 C^N_4인 경우에는 임계치 0.85와 같거나 큰 유사도를 가지는 과거 사례가 없으므로 디폴트 k를 사용하는 k-NN 방법이 적용되었다. 각 새로운 사례들의 목표 속성은 다음과 같이 예측되었다. C^N_1은 '검은' 사례로(B = 2.70, W = 2.67), C^N_2는 '하얀' 사례로(B = 1.75, W = 1.75; 하지만 W 그룹이 가장 큰 유사도인 0.90을 포함하고 있다), C^N_3은 '검은' 사례로(B = 4.77, W = 2.72), C^N_4는 '하얀' 사례로(B = 1.65, W = 2.46) 그리고 C^N_5는 '검은' 사례로(B = 1.76, W = 0.86) 예측되었다.

<표 4>는 이 간단한 예에 적용된 k-NN 방법과 s-NN 방법의 결과를 비교하여 보여주고 있다. s-NN 방법을 적용함으로써 5개의 새로운 사례 중에서 3개 사례의 목표속성 값을 정확하게 예측하였다.

<표 4> s-NN과 k-NN의 결과 비교

	C^N_1	C^N_2	C^N_3	C^N_4	C^N_5
맞는 목표속성 값	B	B	B	B	B
k-NN 예측 (k = 5)	W	W	W	W	W
s-NN 예측 (유사도 임계치= 0.85)	B	W	B	W	B

4. 실험 및 평가

이 절에서는 본 연구에서 제안하는 s-NN 방법이 유용한지를 검증하기 위한 실험 및 그 결과의 평가에 대한 내용을 다룬다. s-NN 방법을 적용하여 구축된 CBR 모델을 s-NN-CBR로 명명하고, s-NN-CBR과 성능 비교를 하기 위하여 k-NN 방법을 적용하여 구축된 CBR 모델을 k-NN-CBR로 명명한다.

4.1 사용 데이터

본 연구의 실험을 위해 사용한 데이터는 UCI Machine Learning Repository에서 제공하는 Pima Indians Diabetes Data(768개의 레코드), Australian Credit Data(690개의 레코드) 그리고 German Credit Data(1000개의 레코드)이다[Blake and Merz, 1998]. 각 데이터는 <표 5>에 제시된 바와 같이 5 : 3 : 2의 비율로 훈련용(Training), 검증용(Validation), 평가용(Test) 데이터 집합으로 나누었다. 훈련용 데이터 집합은 사례베이스로 사용되고, 검증용 데이터 집합은 k의 값이나 속성의 가중치 등 파라미터를 정하는데 사용되고, 평가용 데이터 집합은 최종 선정된 모델의 성능 평가를 위한 목적으로 사용된다.

<표 5> 모델 데이터 집합

	레코드 개수				입력속성 개수		
	전체	훈련용	검증용	평가용	전체	수치형	범주형
Pima Indians Diabetes Data	768	383	231	154	8	8	0
Australian Credit Data	690	345	207	138	14	6	8
German Credit Data	1000	500	300	200	20	7	13

4.2 유사도 임계치

최적의 s-NN-CBR 모델을 구축하기 위하여, 우리는 먼저 100개의 속성 가중치 벡터를 생성한 후에, 이를 이용하여 구축한 s-NN-CBR 모델에서 유사도 임계치를 변화시키면서 검증용 데이터 집

합에 대한 성능의 변화를 관찰하였다. 속성 가중치는 0과 1사이의 실수값으로서 무작위로 생성된다. 제 3절의 제 5단계에서 기술하였듯이, s-NN-CBR 모델의 구축 과정에 디폴트 k-NN 단계가 포함되어 있기 때문에 검증용 데이터 집합을 사용한 예측에 대해서 <표 6>과 같은 결과를 얻게 된다. <표 6>의 각 셀은 사례의 개수를 표시한다.

<표 6> 방법별 예측 결과의 구분

	검증용 데이터 집합 (V)			
방법	s-NN (S)		디폴트 k-NN (K)	
예측 결과	맞음(S_C)	틀림(S_W)	맞음(K_C)	틀림(K_W)

예측의 성능은 다음과 같은 세 가지 지표를 사용하여 측정하였다.

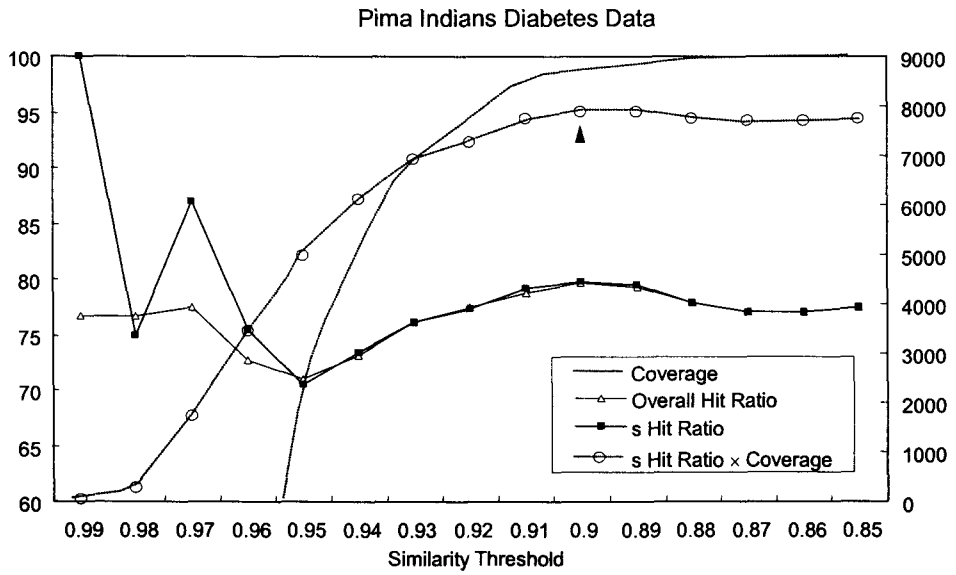
Coverage = (S / V) × 100%

Overall Hit Ratio = ((S_C + K_C)/V) × 100%

S Hit Ratio = (S_C/S) × 100%

Coverage는 전체 검증용 데이터 집합의 사례들 중에서 몇 %가 s-NN 방법을 이용하여 예측되었나를 나타내는 지표이다. 즉, 제 3절의 s-NN 적용 과정에서 제 4단계까지로 예측을 끝낸 경우의 사례의 비율이다. Overall Hit Ratio는 전체 검증용 데이터 집합의 사례들 중에서 s-NN 방법과 디폴트 k-NN 방법을 모두 사용하여 올바르게 예측한 사례의 비율이다. S Hit Ratio는 s-NN 방법으로 예측을 수행하게 된 사례들 중에서 올바르게 예측된 사례의 비율이다.

제 3절의 제 1단계에서 기술한 바와 같이, 유사도 임계치와 k의 디폴트 값을 설정해 놓아야 한다. k의 디폴트 값은 5로 설정하였고, 유사도 임계치는 실험을 거쳐서 설정되었는데, 유사도 임계치를 0.99에서 0.75까지 0.01씩 감소시키면서 검증용 데이터



[그림 3] 유사도 임계치의 설정

집합을 대상으로 실험을 하였다. 각 유사도 임계치마다 세 가지 성능 지표를 측정하였다. 최종의 유사도 임계치는 S Hit Ratio와 Coverage를 동시에 고려하여 설정되어야 한다. 그러므로, S Hit Ratio와 Coverage의 곱이 가장 큰 값을 가질 때의 유사도 임계치를 최종의 유사도 임계치로 설정하였다. Pima Indians Diabetes Data에 대한 실험 결과는 [그림 3]과 같다. [그림 3]의 X 축은 유사도 임계치, 왼쪽 Y 축은 Coverage, Overall Hit Ratio 그리고 S Hit Ratio, 오른쪽 Y 축은 S Hit Ratio × Coverage를 나타낸다. [그림 3]에서 작은 검정색 삼각형으로 표시되어 있는 것과 같이, Pima Indians Diabetes Data에 대해서는 유사도 임계치가 0.9로 설정되었다. 동일한 방법으로, Australian Credit Data의 유사도 임계치는 0.78, German Credit Data의 유사도 임계치는 0.75로 설정되었다.

4.3 결과 및 평가

s-NN-CBR의 성능과 비교하기 위하여 전통적인 *k*-NN 방법을 적용하여 *k*-NN-CBR을 구축하였다. *k*-NN-CBR에 사용할 최적의 *k* 값은 실험을 통하여 설정되었는데, Pima Indians Diabetes Data의 경우에는 11, Australian Credit Data의 경우에는 5, 그리고 German Credit Data의 경우에는 23으로 설정되었다. 속성 가중치 벡터는 *s*-NN-CBR 구축에서 사용한 것을 그대로 *k*-NN-CBR 구축에 사용하였다.

평가용 데이터 집합에 대하여 *s*-NN-CBR의 Overall Hit Ratio와 *k*-NN-CBR의 Hit Ratio의 비교는 <표 7>과 같다. *k*-NN-CBR의 Hit Ratio는 평가용 데이터 집합의 사례들 중에서 *k*-NN-CBR이 올바르게 예측한 사례의 비율이다.

<표 7>에서 보듯이, *s*-NN-CBR이 *k*-NN-CBR

<표 7> 평가용 데이터 집합에 대한 두 모델의 성능
(단위 : %)

	Pima Indians Diabetes Data	Australian Credit Data	German Credit Data
k-NN-CBR	74.7	87.7	72.5
s-NN-CBR	77.3	90.6	75.5

보다 우수한 성능을 보인다. Pima Indians Diabetes Data의 경우에는 2.6% 포인트, Australian Credit Data의 경우에는 2.9% 포인트, 그리고 German Credit Data의 경우에는 3.0% 포인트만큼 *s*-NN-CBR이 우수한 성능을 보였다.

<표 7>에 제시된 실험 결과가 통계적으로 유의한지를 검증하기 위하여 <표 8>과 같이 McNemar 테스트를 수행하였다. 이 테스트는 동일 대상에 대한 사전 및 사후의 측정치의 변화를 검증하는데 사용된다[Cooper and Emory, 1995].

<표 8> 성능비교를 위한 McNemar 값

	k-NN-CBR → *s*-NN-CBR		
	Pima Indians Diabetes Data	Australian Credit Data	German Credit Data
McNemar 값	110.00*	88.36*	52.36*

* McNemar 값 > 6.63이므로 1% 수준에서 유의함.

<표 8>에 나타난 바와 같이, 유의 수준 1%에서 *s*-NN-CBR이 *k*-NN-CBR 보다 우수한 성능을 보인다고 할 수 있다. 즉, 본 연구에서 얻은 결과는 통계적으로 유의하고 그러므로 의미가 있다.

만일 디폴트 *k*-NN 과정, 즉 제 3절에서 기술한 제 5단계를 제거하고 *s*-NN 방법만을 사용하였다면 성능이 어떻게 나왔을까? 이 질문에 답하기 위해서 다음과 같은 새로운 성능 지표를 하나 더 측정해 보자. 여기서 T는 평가용 데이터 집합내의

사례의 개수이다.

s-NN-only Hit Ratio = (S_C/T) × 100%

<표 9>는 s-NN 방법이 예측 성능에 얼마나 공헌하고 있는지를 보여준다. 세 개의 실험 데이터 모두에 있어서, s-NN-only Hit Ratio가 k-NN-CBR Hit Ratio보다 우수하다. 이것은 s-NN 방법 혼자서 디폴트 k-NN 과정의 도움이 없이도 충분히 우수한 성능을 보여줄 수 있다는 것은 의미한다.

<표 9> s-NN 방법의 예측 성능에 대한 공헌

(단위 : %)

	Coverage	Overall Hit Ratio	S Hit Ratio	s-NN-only Hit Ratio	k-NN-CBR Hit Ratio
Pima Indians Diabetes Data	98.7	77.3	77.0	76.0	74.7
Australian Credit Data	99.3	90.6	91.2	90.6	87.7
German Credit Data	95.5	75.5	77.0	73.5	72.5

5. 결론

CBR 모델의 구축에 있어서 최근접 이웃 집합을 어떻게 구성하느냐는 모델의 예측 성능에 직접적인 영향을 주는 중요한 요인이다. 본 연구에서는 유사도 임계치를 이용하여 최근접 이웃 집합을 구성하는 방법인 s-NN을 제안하였다. 이 방법은 고정된 개수의 유사 사례로 최근접 이웃 집합을 구성하는 k-NN 방법에서 발생할 수 있는 예측의 비일관성 문제를 제거해줌으로써 예측 결과의 신뢰성을 높여준다는 장점을 가지고 있다. s-NN 방법의 유용성을 평가하기 위하여 UCI Machine Learning Repository에서 제공하는 세 개의 데이터를 대상으로 실험을 수행한 결과, s-NN 방법을 적용한 CBR 모델이 k-NN 방법을 적용한 CBR 모델보다 우수한 성능을 보여주었다. 이 결과를 모든 영역의 데이터 마이닝 문제에 일반화시키는 것은 무리가 따르지만, s-NN 방법이 k-NN 방법의 대안으로 사용될 수 있다는 가능성을 보여주었다는데 의의가 있다.

본 연구의 한계점 및 추가 연구 방향은 다음의 세 가지를 지적할 수 있다. 첫째, k-NN 방법에서와 마찬가지로 s-NN 방법도 유사도 측정에 사용되는 함수를 어떻게 정의하느냐에 따라 민감한 영향을 받는다. 따라서 이러한 문제를 완화시킬 수 있는 추가적인 연구가 필요하다. 둘째, 다양한 문제 영역에 대한 실험을 통해 s-NN 방법의 유용성에 대한 추가적인 검증이 필요하다. 마지막으로, 제 4.3절에서 기술하였듯이 s-NN 방법만으로도 충분히 우수한 성능을 보이고는 있지만, 유사도 임계치 이상의 유사도를 갖는 과거 사례가 사례베이스에 존재하지 않는 경우에 효과적으로 대처하기 위한 추가적인 연구 또한 필요하다고 판단된다.

참고문헌

[1] 이재식, 이진천. "다중모델을 이용한 자동차 보험 고객의 이탈예측." *한국 지능정보 시스템학회 논문지*, 12권 2호(2006), 167~183.

[2] 이재식, 전용준. "사례기반 추론을 위한 동적 속성 가중치 부여 방법", *한국 지능정보 시스템학회 논문지*, 7권 1호(2001), 47~61.

[3] Aamodt, A. and E. Plaza, "Case-based Reasoning : Fundamental Issues, Methodological Variations, and System Approaches", *Artificial Intelligence Communication*, Vol.7, No.1(1994), 39~59.

[4] Aha, D. W., "Feature Weighting for Lazy Learning Algorithms", *Feature Extraction, Construction and Selection : A Data Mining Perspective*, Nowell, MA : Kluwer, 1998.

[5] Aha, D. W. and R. L. Bankert, "Feature Selection for Case-based Classification of Cloud Type : An Empirical Comparison", *Proc. AAAI-94 Workshop on CBR*, 1994.

[6] Althoff, K. D., R. Bergmann, S. Wess, M. Manago, E. Auriol, O. I. Larichev, A. Bolotov, Y. I. Zhuravlev and S. I. Gurov. "Case-based Reasoning for Medical Decision Support Tasks : The INRECA Approach", *Artificial Intelligence in Medicine*, Vol.12(1998), 25~41.

[7] Berry, M. J. A. and G. Linoff, *Data Mining Techniques -For Marketing, Sales, and Customer Support*, 2nd ed. John Wiley & Sons, Inc, 2004.

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

[9] Brighton, H. and C. Mellish, "Advances in Instance Selection for Instance-Based Learning Algorithms", *Data Mining and Knowledge Discovery*, Vol.6(2002), 153~172.

[10] Chanchien, S. W. and M. Lin, "Design and Implementation of a Case-based Reasoning System for Marketing Plans", *Expert Systems with Applications*, Vol.28(2005), 43~53.

[11] Chang, P. C. and C. Y. Lai, "A Hybrid System Combining Self-organizing Maps with Case-based Reasoning in Wholesaler's New-release Book Forecasting", *Expert Systems with Applications*, Vol.29(2005), 183~192.

[12] Chiu, C. . "A Case-based Customer Classification Approach for Direct Marketing", *Expert Systems with Applications*, Vol.22 (2002), 163~168.

[13] Chun, S. H. and Y. J. Park, "Dynamic Adaptive Ensemble Case-based Reasoning : Application to Stock Market Prediction", *Expert Systems with Applications*, Vol. 28(2005), 435~443.

[14] Cooper, D. R. and C. W. Emory, *Business Research Methods*. Irwin, Chicago, 1995.

[15] Dash, M. and H. Liu, "Feature Selection for Classification", *Intelligent Data Analysis*, Vol.3(1997).

[16] Elhadi, M. T., "Bankruptcy Support System : Taking Advantage of Information Retrieval and Case-based Reasoning", *Expert Systems with Applications*, Vol.18 (2000), 215~219.

[17] Goker, M. H. and T. Roth-Berghofer, "The Development and Utilization of the Case-based Help-desk Support System HOMER", *Engineering Application of Artificial In-*

telligence, Vol.12, No.6(1999), 665~680.

[18] Law, Y. F. D., S. B. Foong and S. E. J. Kwan, "An Integrated Case-Based Reasoning Approach for Intelligent Help Desk Fault Management", *Expert Systems with Applications*, Vol.13(1997), 265~274.

[19] Lee, J. S. and Y. C. Cho, "Improvement of the Classification Performance by Removing Harmful or Irrelevant Cases", *Proc. 2005 KMIS International Conference*, (2005), 591~595.

[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(2000), 199~213.

[21] Marling, C. and P. Whitehouse, "Case-based Reasoning in the Care of Alzheimer's Disease Patients", *Lecture Notes in Computer Science*, Vol.2080(2001), 702~715.

[22] Park, C. S. and I, Han, "A Case-based Reasoning with the Feature Weights Derived by Analytic Hierarchy Process for Bankruptcy Prediction", *Expert Systems with Applications*, Vol.23(2002), 255~264.

[23] Shen, R. and Y. Fu, "GA based CBR Approach in Q&A System", *Expert Systems with Applications*, Vol.26(2004), 167~170.

[24] Smyth, B., "Case-base Maintenance", *Proc. the 11th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, (1998), 507~516.

[25] Tsai, C. Y., C. C. Chiu and J. S. Chen, "A Case-based Reasoning System for PCB Defect Prediction", *Expert Systems with Applications*, Vol.28(2005), 813~822.

[26] Vollrath, I., W. Wilke and R. Bergmann "Case-based Reasoning Support for Online Catalog Sales", *IEEE Internet Computing Online*, Vol.2(1998), 47~54.

[27] Wang, H. C. and H. S. Wang, "A Hybrid Expert System for Equipment Failure Analysis", *Experts Systems with Applications*, Vol.28(2005), 615~622.

[28] Weiss, S. M. and N. Indurkhya, *Predictive Data Mining : A Practical Guide*, CA : Morgan Kaufmann Publishers, 1998.

Abstract

Formation of Nearest Neighbors Set Based on Similarity Threshold

Jae Sik Lee* · Jin Chun Lee**

Case-based reasoning (CBR) is one of the most widely applied data mining techniques and has proven its effectiveness in various domains. Since CBR is basically based on *k*-Nearest Neighbors (NN) method, the value of *k* affects the performance of CBR model directly. Once the value of *k* is set, it is fixed for the lifetime of the CBR model. However, if the value is set greater or smaller than the optimal value, the performance of CBR model will be deteriorated. In this research, we propose a new method of composing the NN set using similarity scores as themselves, which we shall call *s*-NN method, rather than using the fixed value of *k*. In the *s*-NN method, the different number of nearest neighbors can be selected for each new case. Performance evaluation using the data from UCI Machine Learning Repository shows that the CBR model adopting the *s*-NN method outperforms the CBR model adopting the traditional *k*-NN method.

Key words : Nearest Neighbors, Case-based Reasoning, Classification, Data Mining

* School of Business Administration, Ajou University
** Ubiquition Convergence Research Institute