사례기반추론 모델의 최근접 이웃 설정을 위한 Similarity Threshold의 사용

이재식, 이진천

[a] *아주대학교 경영대학 e-비즈니스학부*
경기도 수원시 영통구 원천동 산 5번지
Tel: +82-31-219-2719, Fax: +82-31-219-1616, E-mail: leejsk@ajou.ac.kr

[b] *아주대학교 경영학과*
경기도 수원시 영통구 원천동 산 5번지
Tel: +82-31-219-2308, Fax: +82-31-219-1616, E-mail: giny777@empal.com

Abstract

사례기반추론(Case-Based Reasoning)은 다양한 예측 문제에 있어서 성공적으로 활용되고 있는 데이터마이닝 기법 중 하나이다. 사례기반추론 시스템의 예측 성능은 예측에 사용되는 최근접 이웃(Nearest Neighbor)을 어떻게 설정하느냐에 따라 영향을 받게 된다. 따라서 최근접 이웃을 결정짓는 k 값의 설정은 성공적인 사례기반추론 시스템을 구축하기 위한 중요 요인 중 하나가 된다. 최근접 이웃의 설정에 있어서 대부분의 선행 연구들은 고정된 k 값을 사용하는 방식을 채택해왔다. 그러나 고정된 k 값을 사용하는 사례기반추론 시스템은 k 값을 크게 설정할 경우 최근접 이웃 안에 주어진 문제와 유사성이 낮은 사례들이 포함됨으로써 예측 오류를 일으킬 수 있으며, k 값이 작게 설정된 경우에는 유사 사례 중 일부만을 예측에 사용하기 때문에 예측 결과의 왜곡을 초래할 수 있다. 본 연구에서는 이러한 문제를 해결하기 위해 최근접 이웃을 결정함에 있어서 Similarity Threshold를 이용하는 s-NN 방법을 제안하였다. 본 연구의 실험을 위해 UCI(University of California, Irvine) Machine Learning Repository에서 제공하는 두 개의 신용 데이터 셋을 사용하였으며, 실험 결과 s-NN을 적용한 CBR 모델이 고정된 k 값을 적용한 전통적인 CBR 모델보다 더 우수한 성능을 보여주었다.

Keywords:

Case-Based Reasoning, Nearest Neighbors, Classification, Data Mining

1. 서론

데이터마이닝에서 주로 다루어지는 문제 형태는 분류(Classification)와 예측(Prediction)이다. 분류란 새로운 사례(case)가 주어졌을 때 이 사례를 사전에 이미 정의된 범주값(class)들 중 하나에 할당시키려는 것이고, 예측이란 과거 사례의 분석을 통해 주어진 사례가 미래에 어떤 행동 또는 값을 가지게 될 것인가를 추정하는 문제이다[Berry and Linoff, 2005]. 사례기반추론(CBR: Case-Based Reasoning)은 이러한 분류 및 예측 문제 모두에 효과적으로 적용 가능한 기계학습(Machine Learning) 기법이다. CBR은 Exemplar-Based Reasoning, Instance-Based Reasoning, Memory-Based Reasoning, Analogy Based Reasoning등 다양한 용어로 사용되지만, 그 기본 개념은 유사하다[Chanchien & Lin, 2005]. CBR은 두 개의 기본 사상에 기반하는데 하나는 유사한 문제는 유사한 해법을 가진다는 것이고, 다른 하나는 한번 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. 새로운 문제 해결을 위해 과거 사례의 해결책을 재사용한다는 이러한 특성은 CBR이 다른 기계학습 기법들과 구별되는 접근 방식이라고 할 수 있다. CBR의 문제 해결 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 해결에 적용되고 있으며, 고장 진단[Varma and Roddy, 1999; Wang and Wang, 2005; Kuo, R.J., Kuo *et. al.*, 2005], 헬프데스크[Goker and Roth-Berghofer, 1999; Law *et. al.*, 2005], 신용 평가[이재식과 전용준, 2001], 전략 수립[Chanchien and Lin, 2005] 등은 성공적으로 CBR이 적용되었던 대표적인 응용 영역이다.

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

1) 사례베이스의 구성 방법.

2) 유사도 측정에 사용되는 거리 함수.

3) 예측에 사용되는 속성 및 속성에 대한 가중치 부여 방법.

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 설정에 관한 연구는 거의 이루어지지 않았다. 최근접 이웃 설정에 있어서 대부분의 선행 연구 모델들은 고정된 k 값을 사용하는 방식을 채택하였다. 이러한 고정된 k 값에 의한 최근접 이웃의 설정은 k 값을 크게 할 경우 최근접 이웃 안에 유사성이 낮은 사례들을 포함시킬 가능성을 증가시키기 때문에 예측 성능을 저하시키는 원인이 된다. 또한 k 값을 작게 설정할 경우에는 유사한 사례 중 일부만을 가지고 예측을 수행함으로써 예측 성능을 저하시킬 수 있다. 고정된 k 값의 설정에 따른 이와 같은 문제들을 해결하기 위해 본 연구에서는, 최근접 이웃을 설정함에 있어서 Similarity Threshold 값을 사용하는 새로운 방법인 s-NN(Similarity based Nearest Neighbor)을 제안하였다.

본 논문은 다음과 같이 구성 되었다. 제 2절에서는 CBR 모델의 성능 개선을 위해 시도되었던 선행 연구들에 대한 고찰과 CBR 모델에 대해 설명하였다. 제 3절에서는 본 연구에서 제안한 s-NN 방법에 대해 소개하였으며, 제 4절에서는 s-NN을 적용한 CBR 모델의 구현과 실험에 대한 내용을 다루었다. 마지막으로 제 5절에서는 본 연구의 결론과 향후 연구 방향에 대해 기술하였다.

2. 사례기반추론 모델

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 영역들이다. 최적 사례베이스의 구성은 문제 해결에 유용한 사례만을 선별하여 사례베이스를 구성함으로써 예측 성능을 개선시키고, 사례의 저장공간을 최소화시키는데 있다[Weiss and Indurkhya, 1998; Smythe, 1998]. Brighton & Mellish[2002]는 효과적인 사례베이스를 구성하기 위한 방법으로 사례베이스로부터 중복된 사례와 유해한(hamful) 사례를 제거하는 Instance selection 방법을 제안하였다. Instance Selection은 모델의 학습에 사용되는 사례들로부터 불필요한 사례를 제거하여 모델의 예측 성능을 개선시키고자 하는 데이터 축소(Data Reduction)의 한 방법이다. 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, 1998; Aha and Bankert, Dash and Liu, 1997; 이재식과 전용준, 2001]. 속성 가중치 부여는 문제 해결에 있어서의 각 속성의 중요도에 따라 다른 가중치를 부여하는 것으로, 특정 속성의 가중치를 '0'에 가깝게 설정할 경우 속성 선정과 같은 효과를 가져오게 된다. 따라서 속성 가중치 부여는 속성 선정의 일반화된 형태로 볼 수 있다[이재식과 전용준, 2001].

CBR 모델의 예측 성능 개선을 위한 또 다른 시도로는 다중모델의 구축이다[Shen and Fu, 2004; Wang and Wang, 2005; Chang and Lai, 2005; Chun and Park, 2005]. 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, Kuo *et. al.*[2005]은 장비의 고장 진단 문제에 CBR 과 클러스터링 알고리즘을 결합한 다중모델 구축 방법을 소개하였다.

Aamodt와 Plaza[1996]는 CBR의 문제 해결 과정을 그림 1과 같이 크게 검색, 재사용, 수정, 유지 4 단계로 구분하였다.

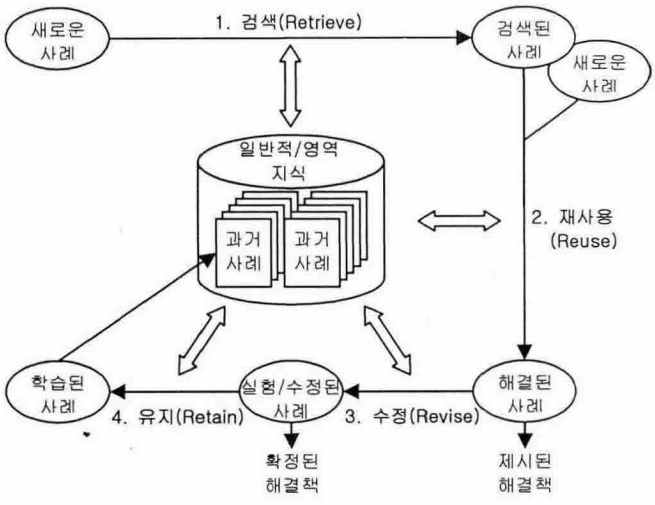

그림 1 - Aamodt and Plaza의 사례기반추론 과정

1. 검색(Retrieve)

검색은 현재 문제와 가장 유사한 과거 사례들을 사례 베이스로부터 찾아내는 것이다.

2. 재사용(Reuse)

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

3. 수정(Revise)

수정은 현재 문제의 해결을 위해 검색된 유사 사례들의 해법을 현재 문제에 적합한 형태로 조정하는 것이다.

4. 유지(Retain)

유지는 새롭게 해결된 문제와 해법을 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 검색하는 방법이다. 귀납적 검색은 사례의 검색 및 구성을 위해 의사결정나무 형태의 구조를 사용한다. 최근접 검색은 현재 문제의 유사 사례 검색을 위해 현재 문제를 표현하는 사례와 사례 베이스에 있는 모든 사례와의 유사도를 측정함으로써 유사 사례를 찾는 방법이다. 최근접 이웃 검색의 경우 일반적으로 가장 많이 사용되는 방법은 주어진 사례와 가장 유사한 k개의 사례를 검색해 주는 k-NN(k Nearest Neighbor) 방법이다. 여기에서 사례간의 유사도는 거리 함수(distance function)에 의해 측정되며, 유사도 측정을 위해 일반적으로 사용되는 함수의 형태는 식 1 과 같다.

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

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

사례간의 유사도 정도를 정의함에 있어서 일반적으로 '0'에서 '1'사이의 정규화된 실수 값으로 표현하는데, '0'에 가까울수록 두 사례의 유사성이 낮다는 것을 의미하고, '1'에 가까울수록 유사성이 높다는 것을 의미한다. 본 연구에서도 식 1을 사용하여 사례간의 유사도를 측정하였다.

3. 최근접 이웃 설정을 위한 Similarity Threshold의 사용

최근접 이웃의 설정을 위해 본 연구에서는 고정된 k 값을 사용하는 방식이 아닌, Similarity Threshold를 사용하는 s-NN 방식을 제안하였다. 즉, s-NN은 최근접 이웃 설정을 위해 사전에 정의한 Similarity Threshold 값을 기준으로 사례베이스에 있는 사례들 중 해당 문제와의 유사도가 설정된 Threshold 값보다 큰 모든 사례들을 최근접 이웃으로 설정하게 된다. s-NN의 기본적인 아이디어는 문제 해결을 위해 해당 문제와 유사한 모든 사례를 사용함으로써 예측 성능을 개선시키고자 하는데 있다.

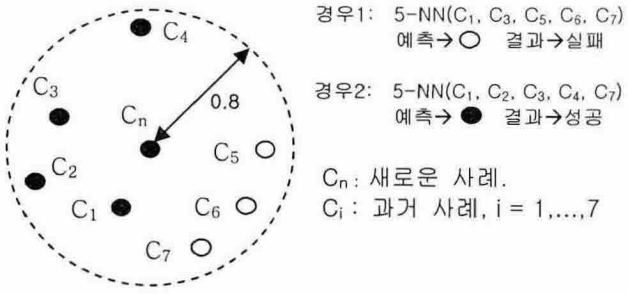

그림 2 - k = 5일 때의 예측 결과 생성 예

그림 2는 최근접 이웃 설정을 위해 고정된 k를 사용함으로써 발생되는 예측 오류에 대한 예를 보여준다. 각각의 원은 하나의 사례를 나타내며, 원의 색깔은 해당 사례가 가지는 목표 속성 값을 의미한다. 그림 2에서 볼 수 있듯이 새로운 사례와 가장 유사한 사례는 $C_1 \rightarrow C_3 \rightarrow C_2 = C_4 = \ldots = C_7$ 순이다. 여기에서 C_2와 C_4~C_7 사례들은 새로운 사례에 대해 동일한 유사도 거리를 가지는 사례들이다. 따라서 k가 5로 설정된 경우, 이 사례들 중 어떤 사례가 최근접 이웃에 포함되느냐에 따라 예측 결과가 달라지게 된다. 즉, 최근접 이웃으로 C_1, C_2, C_3, C_5, C_6, C_7이 선정된 경우와 C_1, C_2, C_3, C_4, C_7가 선정된 경우의 예측 결과가 서로 달라지게 된다. 이러한 예측 결과의 비일관성 문제는 고정 k값을 사용함으로써 발생하는 근본적인 문제라고 할 수 있다. s-NN은 유사도 값을 기반으로 최근접 이웃을 결정하기 때문에 이와 같은 문제를 피하게 된다. 즉, 그림 2에서와 같이 Similarity Threshold를 0.85로 설정한 경우 유사도 점수가 0.85 이상인 모든 사례들이 예측에 사용됨으로써 예측 결과에 대한 일관성이 유지된다.

s-NN에 의한 최근접 이웃 선정 과정의 다음과 같다.

단계1: Similarity Threshold 설정.

최근접 이웃 설정을 위한 Similarity Threshold를

정의한다.

단계2: 유사도 측정.

유사도 함수를 사용하여 신규 사례와 사례베이스 내의 모든 사례들과의 유사도를 측정한다.

단계3: 평가.

단계 2에서 측정된 유사도 점수가 단계 1에서 설정된 Similarity Threshold 값을 충족하는지를 확인한다.

단계4: 최근접 이웃의 형성.

Similarity Threshold를 충족하는 사례들을 추출하여 최근접 이웃을 형성한다. 만약 Similarity Threshold를 충족하는 사례가 없을 경우, 사전에 정의된 default k에 의해 k 최근접 이웃을 형성한다.

4. 실험 및 평가

이 절에서는 CBR 모델의 구축에 있어서 제안된 s-NN 방법이 유용한지를 검증하기 위한 실험 및 결과 평가에 대한 내용을 다루었다. 본 연구에서는 제안 방법에 의해 구축된 CBR 모델을 s-NN-CBR로 명명하였으며, s-NN-CBR 성능의 타당한 비교 평가를 위해 고정된 k 값을 사용한 CBR 모델(k-NN-CBR), C5.0 의사결정나무 모델(C5.0 DT), 인공신경망 모델(ANN)을 benchmarking 모델로 구축하였다.

4.1 사용 데이터

본 연구의 실험을 위해 사용한 데이터는 UCI Machine Learning Repository에서 제공하는 German Credit Data와 Australian Credit Data이다. 두 데이터 모두 개인의 신용 평가 문제를 다룬 것으로, 목표 속성 값이 두 개로 이루어져 있다. German Credit Data는 총 21개의 속성을 가지고 있으며, 1,000개의 사례를 포함하고 있다. Australian Credit Data의 경우 14개의 속성을 가지고 있으며, 690개의 사례를 포함하고 있다.

4.2 실험 설계

4.2.1 실험데이터 구성

모델의 학습과 평가를 위해 표 1과 같이 실험 데이터를 구성하였다. 훈련 데이터 셋(Training Data Set)은 모델의 학습을 위해 사용하였으며, 검증 데이터 셋(Validation Data Set)은 모델의 과잉학습(Overfitting) 방지와 최적 모델의 선정을 위해 사용하였다. 그리고 평가 데이터 셋(Test DataSet)은 최종 선정된 모델의 성능 평가를 위한 목적으로 사용하였다. CBR 모델의 구축에 있어서 훈련 데이터 셋은 사례 베이스(Case Base)로 사용하였다. 본 연구에서는 이들 3개의 데이터 셋 비율을 50:30:20으로 구성 하였다.

4.2.2 실험

본 연구에서는 최적의 s-NN-CBR 모델의 선정을 위해서 먼저, 검증용 데이터 셋을 대상으로 Similarity Threshold 값을 변화시켜 가면서 적중률의 변화를 살펴본 후 가장 좋은 적중률을 보인 모델을 최종 모델로 선정하였다. German Credit Data의 경우 Similarity Threshold 값을 0.9 ~ 0.75까지, Australian Credit Data의 경우 0.95 ~ 0.75까지 0.01씩 변화시키면서 실험을 수행하였으며 그림 3에 그 결과가 나타나 있다. 그림 3에 나타나 있는 Hit Ratio 1은 s-NN-CBR 모델에 있어서 Similarity Threshold 기반으로 예측된 사례 중 예측이 적중한 사례의 비율을 나타내며, Hit Ratio 2는 전체 검증 데이터 셋에 대한 s-NN-CBR 모델의 예측 적중률을 나타낸다. 마지막으로 Coverage는 전체 검증 데이터 셋 중 Similarity Threshold 기반으로 예측이 수행된 사례의 비율을 나타낸다. 그림 3에서 알 수 있듯이 German Credit Data의 경우 Similarity Threshold 값이 0.85 일 때 Hit Ratio1이 80.17%, Hit Ratio2가 77.33%로 가장 좋은 성능을 보여주었으며, 이때의 Coverage 값은 77.3%였다. 이것은 전체 검증 데이터 셋의 77.3%가 Similarity Threshold 기반으로 예측이 수행되었음을 의미한다. Australian Credit Data의 경우에는 Similarity Threshold 값이 0.87에서 가장 좋은 예측 성능을 보여주고 있으며, 이때의 Hit Ratio1, Hit Ratio2, Coverage는 각각 91.54%, 91.3%, 97.1% 였다.

표 1 - 실험 데이터

	Data Set				Input Features		
	All	**Training**	**Validation**	**Test**	**All**	**Numeric**	**Categorical**
German	**1,000**	500	300	200	**20**	7	13
Australian	**690**	345	207	138	**14**	6	8

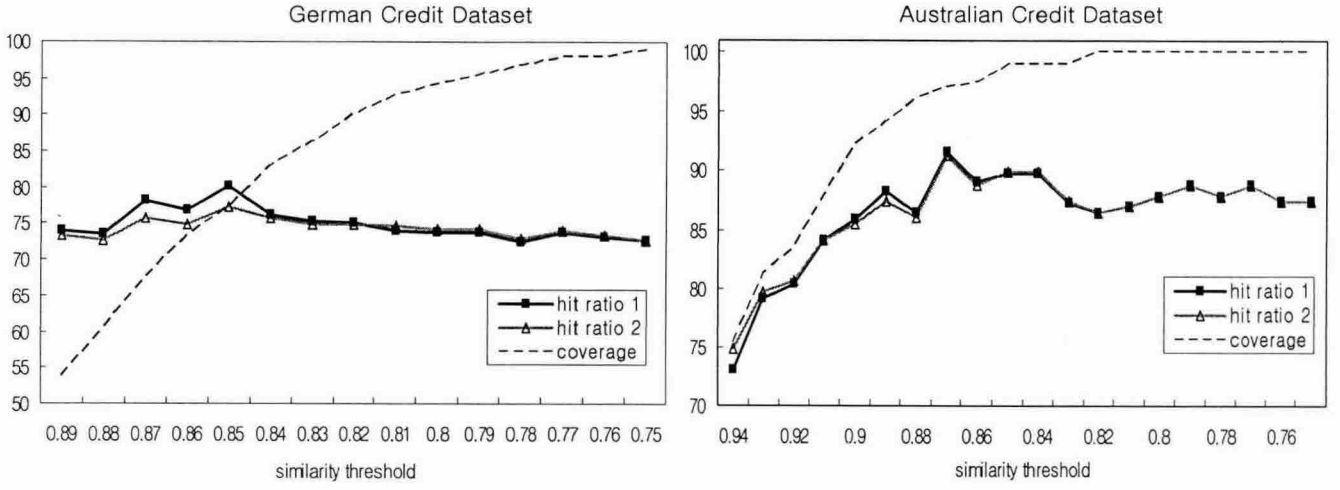

그림 3 - Threshold 값 변화에 따른 *s*-NN-CBR 모델의 실험 결과

표 2 - 실험 결과

	German		Australian	
	Validation Data Set	**Test Data Set**	**Validation Data Set**	**Test Data Set**
C5.0 DT	77.00 %	72.00 %	88.89 %	84.78 %
ANN	***78.00 %***	***76.00 %***	89.73 %	86.23 %
***k*-NN-CBR**	74.00 %	73.50 %	88.89 %	85.51 %
***s*-NN-CBR**	***77.00 %***	***76.00 %***	***91.30 %***	***88.41 %***

German Credit Data의 실험 결과를 보면 Similarity Threshold 값이 0.9에서 0.85가 될 때까지는 Hit Ratio1과 Hit Ratio2 모두 증가하는 현상을 보이다가 그 이후부터는 두 값 모두 감소하는 현상을 보이고 있다. 또한 0.85 이상에서는 Hit Ratio1이 Hit Ratio2보다 더 높은 수치를 보이다가 0.85 이하에서는 두 값이 점차 수렴하였다. 이것은 Similarity Threshold 값을 0.85 이하로 설정할 경우 관련성이 낮은 사례들이 최근접 이웃에 포함됨으로써 예측에 편의(bias)를 준 것으로 해석할 수 있다. Australian Credit Data에서도 Similarity Threshold 값 0.87을 기준으로 이와 유사한 현상이 나타남을 알 수 있다. 따라서 본 연구에서는 최종 *s*-NN-CBR 모델의 Similarity Threshold 값을 German Credit Data의 경우 0.85로, Australian Credit Data의 경우 0.87로 설정하였다.

4.3 결과 및 평가

s-NN-CBR 모델의 성능 개선 효과를 평가하기 위해, 본 연구에서는 3개의 benchmarking 모델 즉, *k*-NN-CBR, C5.0 DT, ANN을 구축하였다. 본 연구에 사용된 C5.0 DT와 ANN 모델은 SPSS사의 Clementine 8.2를 사용하여 구축하였으며, *s*-NN-CBR과 *k*-NN-CBR 모델은 Microsoft사의 Visual Basic 6.0을 사용하여 구축하였다. 표 2는 두 신용 데이터에 대해 이들 모델들이 보여준 검증 데이터 적중률과 평가 데이터 적중률이다. 표 2에서 알 수 있듯이 German Credit Data의 경우 본 연구에서 제안된 방법에 의해 구축된 *s*-NN-CBR 모델과 ANN 모델이 가장 우수한 성능을 보여주었으며, Australian Credit Data의 경우 *s*-NN-CBR 모델이 평가 데이터 적중률에 있어서 88.41%로 가장 우수한 성능을 보여주었다. 우리는 위의 실험결과에서 나타난 *s*-NN-CBR 모델과 *k*-NN-CBR 모델의 예측 성능 차이가 통계적으로 유의한 것인지를 분석하기 위해 McNemar 분석을 사용하였다.

표 3 - s-NN-CBR과 k-NN-CBR의 McNemar 분석 결과

	***k*-NN-CBR → *s*-NN-CBR**	
	German	**Australian**
McNemar Value	56.0057*	79.0775*

표 3에서 볼 수 있듯이 유의수준 1% 본 연구에서 제안한 *s*-NN-CBR 모델이 *k*-NN-CBR 모델보다 우수한 성능을 보여주었다.

5. 결론

CBR 모델의 구축에 있어서 예측 결과 생성을 위해 사용되는 최근접 이웃을 어떻게 선정하느냐는 모델의 예측 성능에 직접적인 영향을 주는 중요한 요인이 된다. 본 연구에서는 Similarity Threshold를 기반으로 하는 효과적인 최근접 이웃 설정 방법인 *s*-NN을 제안하였다. 이 방법은 고정된 *k*의 사용에 이해 발생할 수 있는 예측 결과의 비일관성 문제를 제거해줌으로써 예측 결과의 신뢰성을 높여준다는 장점을 가지고 있다. *s*-NN 방법의 유용성을 평가하기 위해 UCI Machine Learning Repository로부터 제공된 두 개의 신용데이터를 대상으로 실험을 수행한 결과, *s*-NN을 적용한 *s*-NN-CBR 모델이 전통적인 방식의 고정된 *k*를 적용한 *k*-NN-CBR 모델보다 더 우수한 성능을 보여주었다. 이 결과가 모든 영역에서 발생되는 일반적인 현상이라고 결론 짓기에는 무리가 따르지만, *s*-NN의 유용성에 대한 가능성을 보여주었다는데 의의가 있다.

본 연구의 한계점으로는 다음의 몇 가지를 지적할 수 있다. 첫째, 고정된 *k*를 사용하는 전통적인 방식과 마찬가지로 *s*-NN도 유사도 측정에 사용되는 함수를 어떻게 정의하느냐에 따라 민감한 영향을 받는다는 점이다. 따라서 이러한 문제를 완화시킬 수 있는 추가적인 연구가 필요하다. 둘째, 다양한 문제 영역에 대한 추가적인 실험을 통해 *s*-NN 방법의 유용성에 대한 추가적인 검증이 필요하다. 그리고 Similarity Threshold를 만족하는 사례가 사례베이스에 존재하지 않는 신규 사례를 위한 효과적인 최근접 이웃 설정 방법에 대한 연구 또한 필요하다고 판단된다.

Acknowledgement

본 연구는 21세기 프론티어 연구개발 사업의 일환으로 추진되고 있는 정보통신부의 유비쿼터스컴퓨팅및네트워크원천기반기술개발사업의 지원에 의한 것임.

참고문헌

[1] 이재식, 전용준 (2001), "사례기반 추론을 위한 동적 속성 가중치 부여 방법," *한국지능정보시스템학회 논문지*, 제7권 제1호, pp. 47-61.

[2] Aamodt, A. and E. Plaza (1994). "Case-based reasoning: fundamental issues, methodological variations, and system approaches," *Artificial Intelligence Communication*, Vol. 7(1), pp. 39-59.

[3] Aha, D. W. (1998). "Feature Weighting for lazy learning algorithms," *Feature Extraction, Construction and Selection: A Data Mining Perspective*, Nowell MA: Kluwer, 1998.

[4] Aha, D. W. and R. L. Bankert (1994). "Feature selection for case-based classification of cloud type: An Empirical Comparison," *Proceedings of AAAI-94 Workshop on CBR*.

[5] Berry, M. J. A. and G. S. Linoff (1999). *Mastering Data Mining*, Wiley Publishers.

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

[7] Chanchien, S. W. and M, Lin (2005). "Design and implementation of a case-based reasoning system for marketing plans," *Expert Systems with Applications*, Vol. 28, pp. 43-53.

[8] Chang, P.C., and Lai, C.Y. (2005). "A hybrid system combining self-organizing maps with case-based reasoning in wholesaler's new-release book forecasting," *Expert Systems with Applications*, Vol. 29, pp. 183-192.

[9] Chun, S.H., and Park, Y.J. (2005). "Dynamic adaptive ensemble case-based reasoning: application to stock market prediction," *Expert Systems with Applications*, Vol. 28, pp. 435-443.

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

[11] Goker, M.H., and Roth-Berghofer, T. (1999). "The development and utilization of the case-based help-desk support system HOMER," *Engineering Application of Artificial Intelligence*, Vol.12 (6), pp. 665-680.

[12] Kuo, R.J., Kuo, Y.P., and Chen, K.Y. (2005). "Developing a diagnostic system through integration of fuzzy case-based reasoning and fuzzy ant colony system," Expert Systems with Applications, Vol. 28, pp. 783-797.

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

[14] Smyth, B. (1998). "Case-base maintenance," *Proceedings of the 11th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pp. 507-516.

[15] Park, C. S. and I, Han (2002). "A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction," *Expert Systems with Applications*, Vol. 23, pp. 255-264.

[16] Shen, R., and Fu, Y. (2004). "GA based CBR approach in Q&A system," *Expert Systems with*

Applications, Vol.26, pp. 167-170.

[17] Varma, A., and Roddy, N. (1999). "ICARUS: design and development of a case-based reasoning system for locomotive diagnostics," *Engineering Applications of Artificial Intelligence*, Vol.12 (6), pp. 681-429.

[18] Wang, H.C., and Wang, H.S. (2005). "A hybrid expert system for equipment failure analysis," *Experts Systems with Applications*, Vol. 28, pp. 615-622.

[19] Weiss, S. M. and N. Indurkhya (1998). *Predictive Data Mining: A practical guide*, CA: Morgan Kaufmann Publishers.