Information Systems Review
Vol. 16, No. 1 April 2014 http://dx.doi.org/10.14329/isr.2014.16.1.037

대용량 데이터를 위한 사례기반 추론기법의 실시간 처리속도 개선방안에 대한 연구: 심장병 예측을 중심으로

A Case-Based Reasoning Method Improving Real-Time Computational Performances: Application to Diagnose for Heart Disease

박 윤 주 (Yoon-Joo Park) 서울과학기술대학교 글로벌경영학과

요 약

사례기반 추론기법(case-based reasoning)은 수많은 데이터 속에서 현재 문제와 유사한 과거데이터를 실시간으로 탐색하고 복원해내야 하기 때문에, 과거에 축적된 데이터의 양이 방대하거나 또는 데이터의 축적 속도가 빠를 경우 계산비용(computational cost)이 급격히 높아지는 확장성(scalability) 문제를 갖는다. 이러한 문제를 해결하기 위하여, 기존의 일부 연구들은 클러스터링(clustering) 기법을 적용하여, 전체 데이타를 사전에 몇 개의 그룹으로 분류한 후, 특정 클러스터 내에서만 과거 사례를 탐색하도록 하는 클러스터링과 사례기반 추론의 하이브리드 기법을 제안하였다. 그러나 이러한 기법은 클러스터 수를 얼마로 설정했는지에 따른 성능편차가 심하고, 또한 기본적인 사례기반 추론기법에 비해 일반적으로 낮은 예측성능을 도출하는 문제점이 있다. 본 연구는 이러한 기존의 클러스터-사례기반 추론기법의 문제점을 실증적으로 분석하고, 이를 극복할 수 있는 새로운 하이브리드(hybrid) 사례기반 추론기법을 제안한다. 제안된 기법은 실제 심장병환자를 예측하는 문제에 적용하였으며, 그 결과 제안된 기법이 기존의 사례기반 추론기법에 비해 현격하게 낮은 계산비용을 사용하면서도, 유사한 수준의 예측성능을 도출할 수 있음을 확인하였다.

키워드 : 사례기반 추론, 클러스터링, 계산성능, K-nearest Neighbors, 대용량 데이터

Ⅰ. 서 론

사례기반 추론기법(CBR)은 과거의 유사사례를 복원하여 현재의 목표사례(target case)를 예측하는 데이터마이닝 기법으로, 도메인에 대한 지식이 충분하지 않거나, 실제 사례가 정형화된 모델보다 중요시 되는 의료분야, 금융예측분야 등에 널리 활용되고 있다(Schmidt and Gierl, 2001; Khan and Hoffmann, 2003; Montani *et al*., 2003; Park *et*

† 이 연구는 서울과학기술대학교 교내 학술연구비 지원으로 수행되었습니다.

al., 2011; Park *et al*., 2007; Park *et al*., 2006; Koton, 1988; Turner, 1988).

그러나, CBR 기법은 예측모델을 수립하지 않기 때문에, 매번 목표사례와 유사한 과거사례들을 실시간으로 탐색해야 한다. 이는 과거에 축적된 데이터 양이 방대하거나, 데이터의 축적속도가 급격히 증가하는 빅데이터 시대에, CBR의 활용성을 저해시키는 주요한 제약요인으로 작용한다.

이러한 문제를 해결하기 위하여, 기존의 일부 연구들은(Qiang and Jing, 2001; Kim and Han 2001; Khan *et al*., 2008), CBR 수행에 앞서, 과거 데이터를 사전에 몇 개의 유사그룹들로 클러스터링(clustering)할 것을 제안하고 있다. 이러한 방법을 사용할 경우, 새로운 목표사례를 예측 시, 해당 목표사례가 소속된 클러스터 내에서만 유사사례를 탐색하게 되어, 예측에 소요되는 실시간 계산비용이 현격하게 절감되는 장점이 있다. 본 연구에서는 이러한 기존의 클러스터링과 사례기반 추론의 하이브리드(hybrid) 기법을 Clustering-CBR(C-CBR)이라고 명명하였다.

그러나, 기존의 C-CBR 기법은 일반적인 CBR에 비해서 부정확한 예측성능을 도출할 가능성이 높다. 이는, 일반 CBR 기법이 전체 과거 데이터 풀(pool)을 검색하는 것과는 다르게, C-CBR 기법이 목표사례가 속한 단일 클러스터 내에서만 이웃사례(neighboring cases)를 복원하기 때문에, 상대적으로 목표사례와의 유사도가 낮은 사례들이 복원되기 때문이다. 이러한 현상은 특히 목표사례가 클러스터의 주변부에 위치했을 때 더욱 심화된다.

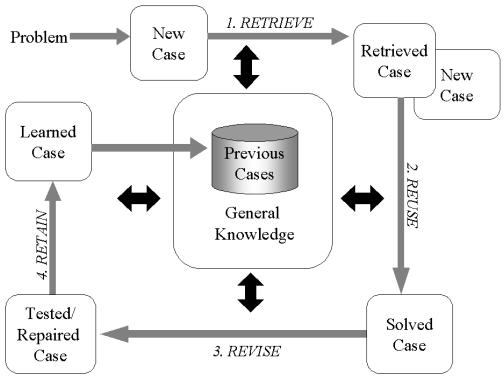
본 연구는 기존의 CBR 기법이 갖는 계산비용에 대한 확장성문제를 해결하면서도, 이와 유사한 수준의 예측성능을 도출할 수 있는 새로운 하이브리드(hybrid) 사례기반 추론기법을 제안한다. 본 연구는 제안된 기법을 Dynamically Merged Clustering-CBR(DMC-CBR)이라고 명명하였다. DMC-CBR 기법은 계산비용을 최소화하는 최적의 클러스터 수를 파악하여, 데이터가 축적되어도 낮은 실시간 계산비용으로 예측이 수행되도록 하였다. 또한, 클러스터 주변부에 위치한 목표사례들의 예측성능 저하문제를 해결하고자, 목표사례 t가 클러스터의 중심부에 위치할 경우는 소속된 클러스터 내에서만 이웃을 탐색하지만, t가 클러스터 주변부에 위치할 경우 소속된 클러스터 이외의 인접한 클러스터의 주변부를 함께 탐색하도록 하여 보다 유사한 이웃사례(neighboring cases)가 복원되도록 하였다.

제안된 기법은 공개 데이터를 제공하는 웹사이트인 UCI repository(Blake and Merz, 1998)의 심장병환자의 데이터에 적용하여 그 효과성과 효율성을 확인하였다. 그 결과, 제안된 DMC-CBR 기법은 기존의 CBR 기법에 비해 현저히 낮은 계산비용을 사용하면서도, 통계적으로 동일한 수준의 예측성능을 도출할 수 있음을 확인하였다. 또한, 기존의 C-CBR 기법이 갖는 클러스터 주변부의 예측성능 저하문제를 극복하고, 클러스터의 주변부에 위치한 목표사례들에 대해서도 정확한 예측을 수행할 수 있음을 확인하였다.

본 논문의 나머지 부분은 다음과 같이 구성된다. 우선, 제 Ⅱ장에서는 관련연구를 소개한다. 제 Ⅲ장에서는 기존의 C-CBR 기법의 장단점을 실증적으로 분석하였으며, 이러한 분석을 바탕으로 새로운 하이브리드(hybrid) 형태의 사례기반 추론기법을 제안하였다. 다음으로 제 Ⅳ장에서는 본 연구의 분석환경 및 결과를 제시하고, 마지막으로 제 Ⅴ장에서는 결론 및 향후 연구방안을 기술하였다.

Ⅱ. 관련 연구

사례기반 추론기법(Case-Based Reasoning, CBR)은 현재의 문제를 해결하기 위해 유사한 과거사례를 복원하여 활용하는 방법으로, 유사한 문제는 유사한 방법으로 해결할 수 있다는 가정을 기반으로 한다(Aamodt and Plaza, 1994). 이러한, 사례기반 추론기법은 일반적으로 <그림 1>과 같은 네 단계로 수행된다.



〈그림 1〉 사례기반 추론 수행과정

첫째, 현재사례와 유사한 과거사례들을 복원한다. 둘째, 유사한 과거사례의 정보를 활용하여 현재 문제에 대한 해결방안을 모색한다. 셋째, 제안된 해결방안을 현재 상황에 맞도록 수정한다. 넷째, 현재의 해결책이 향후 문제해결에 활용될 수 있도록 저장한다.

이러한 사례기반 추론기법은 예측모델을 수립하지 않고, 과거 사례에 대한 복원에 의존하기 때문에 메모리기반추론(memory-based reasoning)으로 분류될 수 있으며, 추론자가 도메인에 대한 지식이 충분하지 않거나, 또는 실제 사례가 정형화된 모델보다 중요시 되는 문제를 해결하는데 효과적으로 활용되고 있다.

그러나 사례기반 추론기법은 현재의 문제를 해결할 때마다, 그와 유사한 과거사례를 복원해야 하기 때문에, 과거에 축적된 데이터 양이 방대하거나, 데이터의 축적속도가 빠를 경우, 실시간 계산비용이 이에 비례하여 증가되는 한계가 있다. 이는 방대한 데이터를 빠른 시간에 처리해야 하는 오늘날, 사례기반 추론기법이 실세계 문제를 효율적으로 해결하는데 커다란 제약조건으로 작용한다.

기존의 일부 연구들은, 이러한 문제를 해결하기 위하여, 클러스터링 기법과 사례기반 추론기법을 결합하는 하이브리드(hybrid) 방안을 제안하였다. 즉, 전체 사례-베이스(case-base)를 사전에 몇 개의 소 그룹으로 클러스터링한 후, 지정된 클러스터 내에서만 과거사례를 복원하도록 하여 과거사례 탐색에 소요되는 실시간 계산비용을 줄이는 것이다. Qiang and Jing(2001)이 제안한 Case Advisor 시스템은 큰 규모의 사례베이스(case-base)를 클러스터링 알고리즘을 활용하여 작은 규모로 줄이도록 제안하고 있으며(Li *et al*., 2006), Kim and Han (2001)의 연구와 Khan *et al*.(2008)의 연구도 전체 데이터를 클러스터링 기법으로 분할 한 후 사례기반 추론을 적용하도록 제안하였다(Khan and Hoffmann, 2003; Khan *et al*., 2008). 그 외에도 Hong and Liou(2008)는 사례복원단계에서 활용되는 속성들을 선택하는데 클러스터링 기법을 활용하도록 제안하고 있다(Hong and Liou, 2008). 이러한 하이브리드(hybrid) 기법들은 의료분야 진단 뿐만 아니라, 금융분야 예측, 식물의 독성분류 등 다양한 분야의 실시간 계산비용 절감을 위해 활용되고 있다.

그러나 이러한 기존 연구들은 클러스터 수 및 클러스터 주변부에 위치한 목표사례의 성능저하 현상을 연구하고 있지 않다는 점에서 본 연구와는 차별성이 있다.

Ⅲ. 동적으로 병합되는 클러스터링-사례기반 추론기법(Dynamically Merged Clustering-CBR)

본 장에서는 기존의 Clustering-CBR(C-CBR) 장단점을 실증적으로 분석하고, 동적인 클러스터 병합기법을 사용하여 이를 극복하는 새로운 하이브리드(hybrid)형 사례기반 추론기법인 Dynamically Merged Clustering-CBR(DMC-CBR)을 설명한다. 우선, 제 3.1절에서는 기존 C-CBR 기법의 장단점을 예측 정확도(accuracy)및 실시간 계산비용(computation)의 측면에서 실증적으로 분석하였으며, 제 3.2절에서는 본 연구에서 제안하는 Dynamically Merged Clustering-CBR(DMC-CBR)의 알고리즘을 제시하였다.

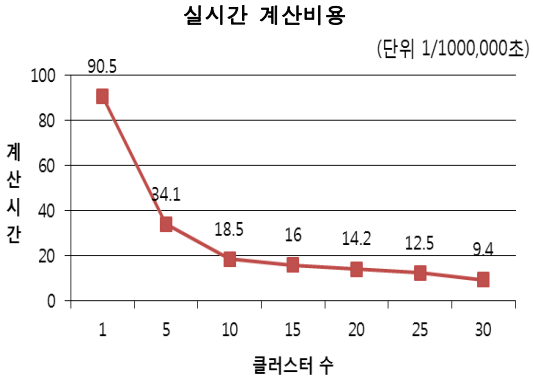
3.1 기존 클러스터링-사례기반 추론기법 (C-CBR)의 한계점

C-CBR 기법은 크게 오프라인 클러스터링 작업과, 온라인 실시간 예측작업으로 구성된다. 오프라인 작업에서는, 전체 데이타 n개가 사전에 지정된 k개의 소그룹(sub-group)으로 클러스터링 된다. 이후, 새로운 목표사례 t가 들어오면, 실시간 온라인 작업으로 예측이 시작된다. 우선, t는 자신이 k개의 그룹들 중 어디에 속하는지 파악한다. 이를 위해, C-CBR은 t와 사전에 분류된 k개 클러스터의 중앙점(centroid)들 간의 거리를 계산하여, 이 거리가 가장 가까운 클러스터를 t의 소속 클러스터인 $Cluster_{Best}$로 설정한다. 일단 t의 소속 클러스터가 결정되면, t는 해당 클러스터인 $Cluster_{Best}$ 내에서만 이웃사례(neighboring cases)를 탐색하게 된다. 이를 위해, t와 $Cluster_{Best}$ 내의 모든 과거사례들 간의 거리를 계산하며, 이중 가장 가까운 과거사례들이 이웃으로 복원된다. 마지막으로, 복원된 이웃사례를 활용하여, 현재의 목표사례인 t에 대한 예측을 수행한다.

C-CBR을 수행하기 위해서는 사전에 적절한 클러스터 수 k값을 사전에 설정해야 한다. 이때, k값이 너무 적게 설정되면, 한 클러스터에 많은 과거사례가 포함되어, t가 이웃사례를 복원하는데 소요되는 실시간 계산비용이 증가한다. 반면, k값이 너무 크게 설정될 경우, 한 클러스터에 소속된 과거사례가 너무 적기 때문에, t가 충분히 유사한 과거사례를 복원하지 못할 수 있다. 본 절에서는 클러스터 수 k를 1, 5, 10, 15, 25, 30으로 조정하면서, 기존 C-CBR 기법의 실시간 계산시간 및 예측정확도(accuracy)의 변화를 살펴보았다. 단, k값이 1일 때에는 클러스터링 과정이 수행되지 않으므로, C-CBR은 일반적인 CBR과 동일하게 수행된다.

본 실험에는 제 Ⅳ장에서 소개한 심장병환자 총 1000명의 데이터가 사용되었으며, 이중 90%를 학습 데이타(training dataset), 10%를 테스트 데이타(test dataset)로 하여 10-fold cross validation을 수행하였다. <그림 2>부터 <그림 4>는 이러한 실험결과를 나타내고 있다.

우선, <그림 2>는 클러스터 수의 변화에 따른 실시간 계산시간의 변화를 1/1000,000초(milli-second) 단위로 나타내고 있다. 이 실험결과, 클러스터 수가 증가함에 따라서, 초기에는 실시간 계산시간이 급격히 감소하였다. 예를 들어, 클러스터를 전혀 수행하지 않은 일반 CBR의 실시간 예측시간은 90milli-seconds였던 것에 비해, 클러스터 5개로 분할했을 때에는 34.1milli-seconds의 실시간 예측시간이 소요되어 무려 62%의 계산시간 단축효과를 도출하였다. 마찬가지로 클러스터가 10개로 증가하면, 5개였을 때보다 약 45%의 실시간 계산시간이 단축된다. 그러나, 클러스터 수가 점차 증가함에 따라서 이러한 실시간 계산비용 절감효과는 <그림 2>의 그래프 우측부분과 같이 점차 둔화됨을 알 수 있었다.

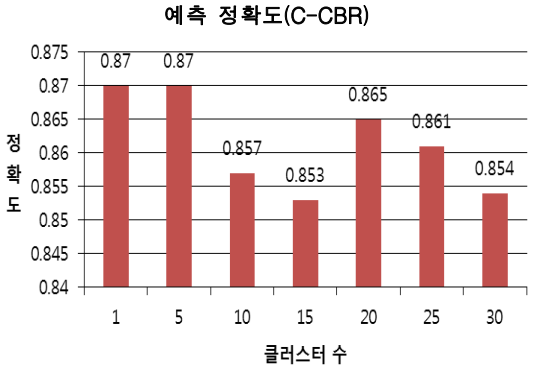


〈그림 2〉 C-CBR 기법의 예측 정확도(Accuracy)

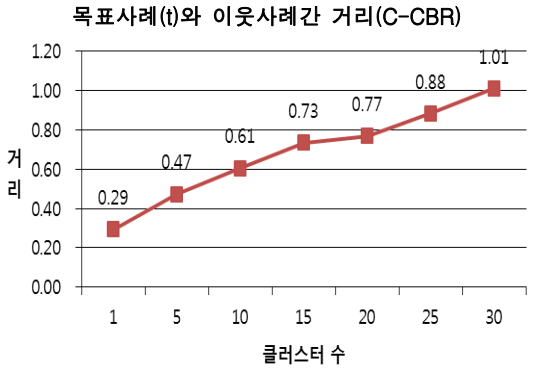
다음으로는, 클러스터 수의 변화에 따른 C-CBR 기법의 예측성능의 변화를 살펴보았다. <그림 3>는 k값을 1, 5, 10, 15, 25, 30으로 조정했을 때의 예측정확도(accuracy) 변화를 나타내고 있다.

<그림 3>의 결과에서 볼 수 있듯이, C-CBR의 예측성능은 클러스터 수에 따라서 달라진다. 본 실험에서는 C-CBR 기법의 예측성능이 k가 1과

5일 때 0.87로 가장 높게 도출되어, 클러스터링을 전혀 수행하지 않거나, 또는 적은 수의 클러스터들로 분할되었을 때, 기존 C-CBR 기법이 많은 수의 클러스터로 분할했을 때에 비해 상대적으로 더욱 정확한 예측을 수행할 수 있음을 알 수 있었다. 그러나, 클러수터 수의 증가에 따라서 예측 정확도가 선형적으로 감소하는 패턴을 보이지는 않는다. 이는 전체 데이터가 적은 수의 클러스터로 분할되면, 한 클러스터의 크기는 반대로 증가하여 목표사례 t가 유사도가 높은 이웃사례를 복원하기 때문에 나타나는 현상으로 생각된다. 그러나, 목표사례와 이웃사례 간 유사도가 반드시 예측성능의 향상을 보장하지는 않기 때문에, 클러스터 수와 예측정확도가 선형적으로 감소하지는 않는 것으로 이해할 수 있다.



〈그림 3〉 C-CBR 기법의 예측 정확도(Accuracy)



〈그림 4〉 C-CBR 기법의 목표사례와 이웃간 거리

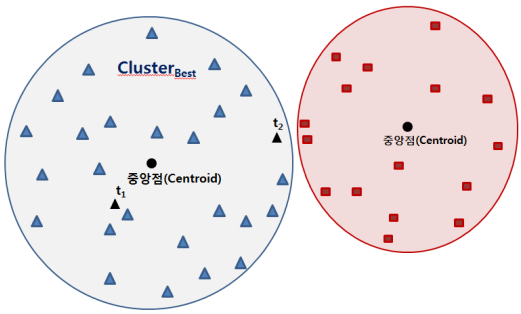
이를 보다 실증적으로 분석하기 위하여, 클러스터 수의 변화에 따른 목표사례 t와 이웃들 간 거리의 변화를 살펴보았다. 즉, 클러스터 수를 1, 5, 10, 15, 20, 25, 30으로 변화시키면서, 기존 C-CBR 기법의 목표사례와 복원된 이웃사례들 간의 평균 거리를 유클리디안 거리계산법을 사용하여 산출하여 <그림 4>와 같이 나타내었다.

그 결과, <그림 4>에서 보는 바와 같이, C-CBR 기법은 클러스터수가 증가함에 따라서, 점차 유사도가 낮은 과거사례들을 이웃사례로 복원함을 알 수 있었다. 예를 들어, 5개의 클러스터를 사용했을 때는 목표사례와 이웃들 간의 평균 거리가 0.47이지만, 30개의 클러스터로 세분화 되었을 때는 이웃과의 거리가 무려 1.01로, 유사도가 현저히 낮은 사례들이 이웃으로 복원되고 있다. 이는 C-CBR 기법의 예측 정확도를 저하시키는 주요한 원인으로 작용할 수 있다.

본 연구는, 위와 같이 클러스터 수가 점차 증가함에 따라서, 기존 C-CBR 기법이 목표사례와 유사한 이웃사례를 복원하지 못하는 문제가, 클러스터의 '중심부' 보다는 '주변부'에서 나타날 것으로 생각하였다 이는 <그림 5>에서 나타낸 바와 같이, 클러스터의 중심부에 위치한 목표사례 t_1의 경우 자신의 소속 클러스터인 $Cluster_{Best}$에 유사한 이웃사례들이 포함되어 있지만, 클러스터의 경계선 부분에 위치한 목표사례 t_2는 유사사례가 자신이 소속된 클러스터인 $Cluster_{Best}$뿐만 아니라, 인접한 다른 클러스터에 포함될 수 있기 때문이다. 그럼에도, 기존 C-CBR 기법은 이웃의 탐색범주를 t_2 소속 클러스터인 $Cluster_{Best}$로 한정하고 있기 때문에, 클러스터 주변부에서는 유사도가 낮은 사례들이 이웃으로 선택될 수 있다.

이러한 가설을 실증적으로 확인하기 위하여, C-CBR 기법의 성능을 클러스터의 중심부 90% 영역과 주변부 10%로 구분하여 예측성능을 산출하였다. 이때, C-CBR 수행 시 클러스터 수는 30개로 설정하였다. 본 실험 결과, <표 1>에 제시한

바와 같이, 클러스터의 '중심부'에 속하는 목표 사례들의 평균적인 예측 정확도(accuracy)는 0.858인데 반해, '주변부'는 이보다 현저히 낮은 0.812로 도출되어, 클러스터 주변부에서 예측성능 저하 현상이 뚜렷이 나타나고 있음을 확인하였다. 또한, '중심부'에 위치한 목표사례들은 평균적으로 이웃사례들과의 거리가 0.965이지만, '주변부'는 1.436이 도출되어 목표사례와의 유사도가 현저히 낮은 이웃사례들로 예측에 활용하고 있음을 확인하였다.



〈그림 5〉 클러스터내 중심부 및 주변부 영역구분

〈표 1〉 C-CBR 기법의 중심부와 주변부의 예측 성능 및 목표사례와 이웃간 거리

	중심부	주변부
정확도	0.858	0.812
거리	0.965	1.436

본 연구는, 이러한 사전분석을 통하여, 기존 C-CBR 기법이 일반적인 사례기반 추론기법 보다 실시간 계산비용을 절감한다는 측면에서는 장점이 있지만, 목표사례가 클러스터의 주변부에 위치할 경우에는 정확한 예측을 수행하지 못할 뿐만 아니라, 적절한 클러스터 수가 설정되지 않는다면 예측 성능이 크게 저하될 수 있음을 확인하였다. 다음의 제 3.2절에서는 이러한 C-CBR 기법의 한계점을 극복할 수 있는 새로운 하이브리드(hybrid)형 사례기반 추론기법을 제안한다.

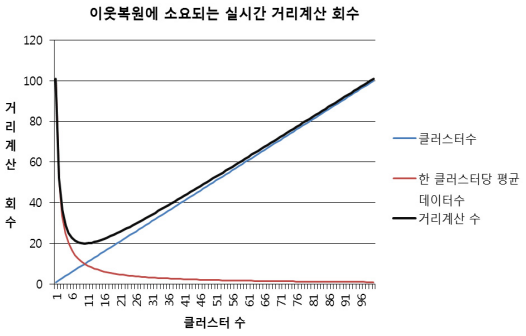
3.2 동적으로 병합되는 클러스터링-사례기반 추론기법(Dynamically Merged-Clustering CBR)

본 절에서는 기존 C-CBR 기법이 갖는 실시간 계산비용 절감효과를 유지하면서도, 클러스터 주변부에서 나타나는 예측성능저하현상을 해결할 수 있는 새로운 하이브리드(hybrid) 기법인 Dynamically Merged Clustering-CBR(DMC-CBR)을 제안한다. 제안된 기법은 전체 데이터를 클러스터링(clustering)한 후에 사례기반 추론(CBR)을 수행한다는 점에서 기존의 C-CBR 기법과 동일하지만, 실시간 계산비용을 최소화하는 최적의 클러스터 수를 사용한다는 점 및, 목표사례의 클러스터 내 위치에 따라서 이웃사례(neighbours)를 복원하는 풀(pool)이 동적으로 변화한다는 점에 차별성이 있다.

제 3.1절의 사전분석에서 클러스터 수와 예측성능 간에 특정한 패턴이 발견되지 않았기 때문에, 제안된 DMC-CBR 기법은 실시간 계산비용을 최소화 시킬 수 있는 클러스터 수 k를 파악하여, 이 값으로 전체 과거데이터를 분할하였다. 이를 위해, 본 연구는 목표사례 t가 이웃사례를 복원하기 위해 수행하는 실시간 연산 회수를 활용하여 실시간 계산비용을 파악하였다. 실시 간 연산은 크게 두 부분으로 구성된다. 첫째, 목표사례 t가 자신의 소속 클러스터인 Cluster$_{Best}$를 파악하기 위해서 실시간 연산을 수행하는 부분이다. 이때, t와 각 클러스터 k개의 중앙점(centroid)들 간 거리가 산출되어야 하므로 k번의 거리계산 연산이 실시간으로 발생한다. 둘째, t의 소속 클러스터인 Cluster$_{Best}$ 내에서 t의 이웃사례들을 탐색하기 위해서, t와 Cluster$_{Best}$의 모든 다른 사례들과의 거리를 계산하는데 n/k번의 연산이 수행된다. 이는 전체 과거 데이터 n개를 k개의 클러스터로 분할하면, 한 클러스터가 평균적으로 n/k개의 사례를 포함하기 때문이다. 이때, 실시간 계산회수에는 정렬에 소요되는 연산 및 클러스터 수와 무관하게 발생되는 연산은 고려하지 않았다. 따라서,

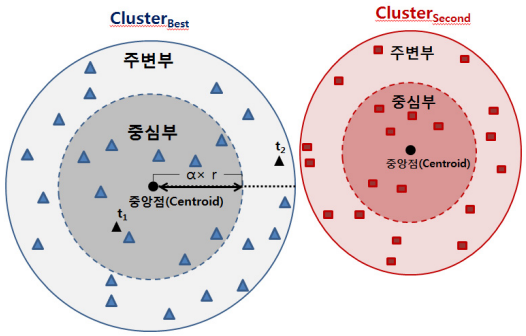
목표사례 t가 이웃사례들을 복원하기 위해 수행하는 실시간 연산회수는 총 $k+(n/k)$가 되며, 이를 그림으로 표현하면 <그림 6>과 같은 형태를 나타낸다. 즉, C-CBR 기법의 실시간 계산비용은 클러스터 수 k가 증가함에 따라 감소하여, k가 $\sqrt{n}$일 때 최소화되고, 그 이후에는 오히려 증가하는 것이다. <그림 6>은 데이터 수가 100개라고 가정했을 때, 클러스터 수의 변화에 따라서, 이웃사례(neighbours) 복원에 요구되는 실시간 거리계산 회수가 변화하는 모습을 나타내고 있다. 데이터 수가 100개이면, 실시간 계산회수는 클러스터 수가 10개(= $\sqrt{100}$)일 때, 총 20회(= k+(n/k))로 최소가 된다. 즉, 목표사례가 자신의 소속 클러스터를 탐색하기 위해, 각 10개의 클러스터 중심점과 수행하는 거리계산회수 10회, 그리고, 소속 클러스터 내에서 이웃사례 복원을 위해 클러스터 내의 다른 사례들과의 거리계산에 소요되는 평균 10회를 합하여, 총 20회의 실시간 거리계산이 필요하다.

따라서 본 연구는 제안된 DMC-CBR 기법 수행 시, 전체 데이터 n개를 $\sqrt{n}$개의 클러스터들로 사전에 분할하도록 제안한다.



〈그림 6〉 클러스터수의 변화에 따른 실시간 계산회수

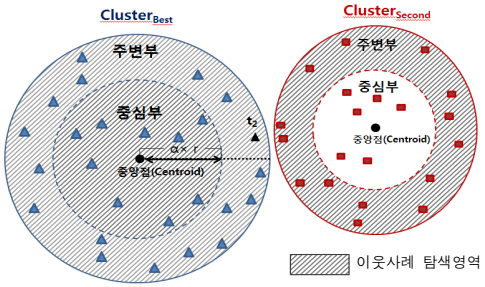
오프라인 작업 시, DMC-CBR은 추가적으로 클러스터 내의 과거사례들을 다시 '중심부'(Center) 영역과 '주변부'(Boundary) 영역으로 구분한다. '중심부'는 클러스터의 중앙점(Centroid)에 인접한 영역을 의미하며, '주변부'는 클러스터의 경계부분 영역을 의미하는 것으로, 다음과 같이 구분된다. 클러스터의 반지름을 r이라고 하면, 중앙점으로부터의 거리가 $\alpha \times r (0 < \alpha < 1)$보다 적은 사례들은 '중심부로 분류되고, 이보다 큰 값을 가지면 '주변부'로 분류된다. 이때, '중심부'와 '주변부'의 경계는 α값을 통하여 조정할 수 있다. 즉, α값이 클수록 더 많은 사례들이 '중심부' 영역에 포함된다. 예를 들어, α가 1일 때에는 모든 사례들이 중심부에 속하게 되며, 반면 0일 때는 모든 사례가 주변부로 분류된다. <그림 7>는 클러스터 내의 사례들이 '중심부'와 '주변부'로 분류되는 모습을 나타내고 있다.



〈그림 7〉 클러스터 중심부와 주변부의 사례 구분

다음으로, 목표사례 t가 들어오면, DMC-CBR 기법은 실시간으로 t가 속하는 클러스터인 $Cluster_{Best}$와 두 번째로 인접한 클러스터인 $Cluster_{Second}$를 파악한다. 이때, 목표사례가 <그림 7>의 t_1과 같이 소속 클러스터의 '중심부'에 위치할 경우, DMC-CBR 기법은 기존 C-CBR 기법과 동일하게 $Cluster_{Best}$ 내에서만 이웃사례를 탐색한다. 그러나, 만약 목표사례가 t_2와 같이 '주변부'에 위치할 경우, DMC-CBR 기법은 <그림 8>의 빗금친 영역에서 나타낸 바와 같이, $Cluster_{Best}$ 뿐만 아니라, $Cluster_{Second}$의 '주변부'에서도 이웃사례를 탐색하게 된다. 이는, 기존의 C-CBR 기법이 주변부에 위치한 목표

사례들에 대해서 충분히 유사한 사례를 복원하지 못하는 문제를 해결할 수 있도록, 인접한 클러스터인 $Cluster_{Second}$로 검색 풀(pool)을 확장하기 위함이다.



〈그림 8〉 목표사례가 클러스터 주변부에 위치할 때의 이웃사례 탐색범위

본 절에서 제안한 DMC-CBR 기법의 전체 알고리즘은 <그림 9>에 제시하였다.

단계 1: 전체 과거 데이타 클러스터링[오프라인]
- 데이터마이닝의 클러스터링 기법을 활용하여, 전체 과거 데이타 n개를 $\sqrt{n}$개의 세부 그룹으로 분할한다.
- 분할된 클러스터들의 각 중앙점으로부터의 반지름 r을 파악하여, 중앙점으로부터의 거리가 α×r 미만인 과거사례는 '중심부', 그 이상인 경우 '주변부' 영역으로 구분한다($0 < \alpha < 1$).

단계 2: 목표사례의 소속 클러스터와 인접클러스터 확인
- 새로운 목표사례 t가 들어오면, t와 단계 1에서 분류해둔 클러스터들의 중심(centroid)간 거리를 계산하여, 가장 가까운 클러스터를 소속 클러스터 $Cluster_{Best}$로 설정하고, 두 번째로 가까운 클러스터를 $Cluster_{Second}$로 설정한다.

단계 3: 소속 클러스터 내에서 목표사례의 위치파악
- 목표사례 t가 소속 클러스터인 $Cluster_{Best}$의 중심부에 위치하는지 주변부에 위치하는지를 다음의 알고리즘을 통해서 파악한다.
  1) 만약, 거리[C_{Best}, t] < $a \times r_{Best}$이면, t는 '중심부' 영역($0 < a < 1$)
  2) 아니면, '주변부' 영역
  (C_{Best}: $Cluster_{Best}$의 중앙점, r_{Best}: $Cluster_{Best}$의 반지름, a: 영역조절 ratio)

단계 4: 목표사례에 대한 이웃사례 복원
  1) 목표사례가 $Cluster_{Best}$의 '중심부'에 위치하는 경우,
  $Cluster_{Best}$내에서만 목표사례 t의 이웃사례를 탐색한다.
  2) 목표사례가 $Cluster_{Best}$의 '주변부'에 위치하는 경우,
  $Cluster_{Best}$와 $Cluster_{Second}$의 주변부에서 목표사례 t의 이웃사례를 탐색한다.

단계 5: 목표사례 예측
- 유사사례 k개를 사용하여, 목표사례 t의 종속변수 예측한다.

단계 6: 성능평가
예측정확도(accuracy)와 실시간 계산회수를 평가한다.

〈그림 9〉 DMC-CBR 기법의 전체 알고리즘

Ⅳ. 사례분석

다음에서는 본 연구의 분석환경을 설명하고, 논문에서 제안한 Dynamically Merged Clustering-CBR(DMC-CBR) 기법을 일반적인 사례추론기법인 Case-Based Reasoning(CBR)및 기존의 클러스터링-사례기반 추론기법인 Clustered-Case Based Reasoning(C-CBR)과 비교 분석한 결과를 제시한다. 각 기법은 목표사례가 이웃사례를 복원하는 풀(pool)을 구성하는 방식에 차이가 있다. 일반 CBR 기법은 전체 과거 데이터 풀을 검색하며, 기존 C-CBR 기법은 목표사례가 소속된 단일 클러스터만을 검색한다. 제안된 DMC-CBR 기법은 목표사례가 소속클러스터의 중심부에 위치하면, 기존 C-CBR 기법과 같이, 목표사례가 소속된 단일 클러스터만 검색하지만, 목표사례가 클러스터의 주변부에 위치할 경우 탐색영역이 인접한 주변 클러스터로 확장된다.

4.1 분석환경

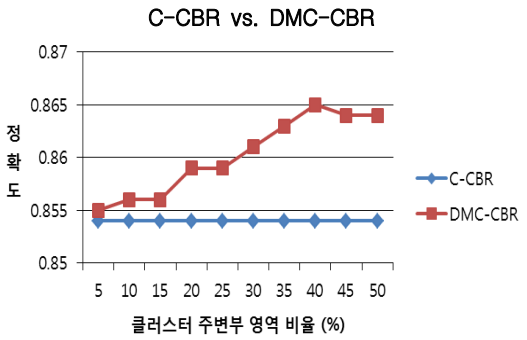
본 연구에는 UCI repository(Blake and Merz, 1998)의 심장병 환자 데이터 1,000개가 사용되었다. 이 데

이터는 Biomedical Engineering 기관 및 Porto 의과대학으로부터 수집된 실제 환자에 대한 23개의 속성들로 구성되어 있으며, 의료진에 의한 진료결과가 '정상', '의심', 또는 '발병'의 형태로 포함되어 있다.

본 연구에서 하이브리드(hybrid) 기법을 사용하는 C-CBR 및 DMC-CBR의 클러스터링은 k-means clustering 알고리즘(Montani *et al.*, 2003)을 사용하여 수행되었으며, 사례기반 추론 시 복원된 이웃사례들은 세 개로 설정하였다. C-CBR의 클러스터 수는 클러스터 수를 5, 10, 15, 20, 25, 30으로 바꿔가면서, training dataset에서 가장 좋은 성능을 도출한 클러스터 수가 test dataset에 적용되도록 하였으며, DMC-CBR 기법의 클러스터 수는 실시간 계산회수가 최소화 되는 30개(= $\sqrt{900}$)로 고정하였다. 성능평가는 예측정확도와 계산비용 두 가지 측면에서 수행하였다. 예측성능은 정확도(accuracy)를 사용하여 평가하였고, 계산비용은 실시간 거리계산회수로 평가하였다. 이러한 실험은 프로그래밍 언어 Java와 무료 데이터마이닝 툴(tool)인 Weka(Montani *et al.*, 2003)를 활용하여 구현하였다.

4.2 분석결과

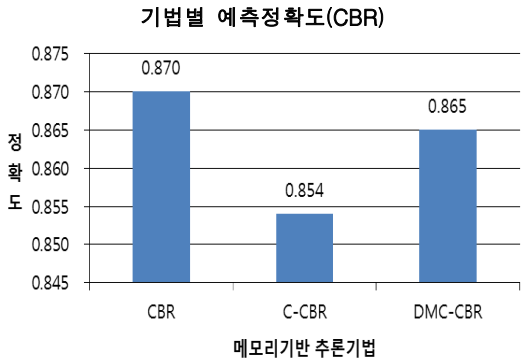
본 절에서는 제 Ⅲ장에서 제안한 DMC-CBR 기법을 심장병환자 예측에 적용한 결과를 CBR 기법 및 기존 C-CBR 결과와 비교 분석하여 제시하였다. 우선, DMC-CBR에서 각 클러스터의 '중심부'와 '주변부' 영역을 적절하게 구분하기 위하여, <그림 10>과 같이 영역조절 값인 α를 조정하면서, 예측성능의 변화를 살펴보았다. <그림 10>은 클러스터의 주변부 영역비율을 5%에서 50%까지 증가시키면서, 제안된 DMC-CBR 기법의 예측정확도(accuracy) 변화를 측정한 결과를 나타내고 있다. <그림 10>에서 보는 바와 같이, DMC-CBR 기법은 '주변부' 영역을 확장시킴에 따라서 예측정확도가 점차 증가하지만, '주변부' 영역이 40%를 넘어서면 더 이상 예측정확도가 향상되지는 않는다. 이는 주변부 영역이 점차 클러스터의 중심부로 확장되면, 목표사례들이 이미 유사한 이웃사례를 소속클러스터인 $Cluster_{Best}$에서 복원할 가능성이 높아지기 때문에, 탐색영역 확장으로 인한 성능 향상효과가 제한적인 것으로 생각된다. 또한, '주변부' 영역을 증가하면, 이웃 사례를 확장된 영역에서 탐색하는데 소요되는 실시간 비용이 증가하기 때문에, 본 연구에서는 명시적인 성능향상이 나타나는 중심부 60%(주변부 40%)로 각 영역이 설정될 수 있도록, 영역조절 ratio인 α = 0.6으로 설정하였다.



〈그림 10〉 주변부 영역에 따른 예측정확도

다음으로는, 제안된 DMC-CBR의 예측정확도를 일반 CBR 및 기존 C-CBR 결과와 비교하여 <그림 11>과 같이 나타내었다. 본 실험결과, 제안된 DMC-CBR 기법의 예측정확도는 0.865로 도출되었으며, 이는 전체 데이터를 탐색한 CBR 기법의 정확도 0.87보다는 다소 낮은 수치이나, 윌콕슨 순위합 검정(Wilcoxon rank-sum test)을 수행한 결과, 이러한 차이가 통계적으로 유의하지는 않은 것으로 나타났다. 윌콕슨 순위합 검정은 비교대상인 두 집단이 이분산이고, 정규분포 가정을 둘 수 없을 경우 사용되는 검정방법으로 <표 2>과 같이 각 기법의 평균을 비교하였다. 반면, DMC-CBR 기법의 예측정확도 기존 하이브리드 기법인 C-CBR의 성능인 0.854보다 90%의 유의수준에서 통

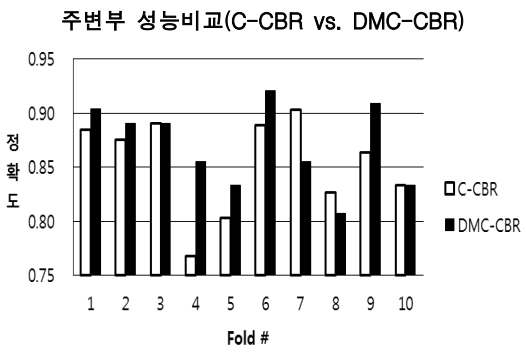
계적으로 유의하게 높은 것으로 나타나서, DMC-CBR 기법이 기존 C-CBR보다 더욱 정확한 질병예측을 수행할 수 있음을 확인하였다. 이러한, 두 하이브리드 기법 C-CBR과 DMC-CBR간 예측성능의 차이는 제 3.2절에서 설명한 바와 같이 클러스터의 주변부에서 나타난 것으로, <그림 12>와 같이 10 fold cross validation 실험결과를 비교하면 DMC-CBR 기법이 총 8fold에서 C-CBR 기법보다 높은 예측정확도를 도출함을 볼 수 있다.



〈그림 11〉 사례기반 추론기법별 예측정확도

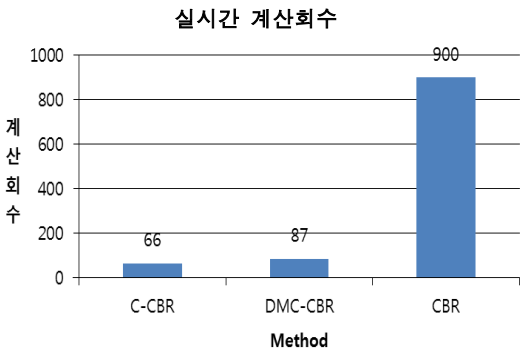
〈표 2〉 윌콕슨 순위합 검정결과

H_0	P-value
CBR ≤ DMC-CBR	0.2059
DBC-CBR ≤ C-CBR	0.0655



〈그림 12〉 C-CBR과 DMC-CBR 기법의 주변부 예측정확도 비교 결과

다음으로는, 위 세 기법의 실시간 계산비용을 비교하여 <그림 13>과 같이 나타내었다. 그 결과, 전체 과거 데이타를 이웃사례의 탐색영역으로 하는 CBR 기법은, 이웃사례를 복원하는데 학습데이타 수와 동일한 총 900번의 실시간 거리계산 연산이 수행되었다. 반면, 사전에 전체데이타를 클러스터링 하여, 지정된 클러스터 내에서만 이웃사례를 탐색하도록 한 C-CBR 기법과 DBC-CBR 기법은 각 66번, 87번의 거리계산 연산이 수행되어, 일반적인 CBR 기법보다 현저히 낮은 실시간 계산비용을 사용하여 이웃사례를 복원할 수 있음을 확인하였다.



〈그림 13〉 사례기반 추론기법별 실시간 계산회수

마지막으로, 위 세 기법의 예측정확도 및 실시간 계산성능 결과를 순위분석을 통해 <표 3>와 같이 비교분석 하였다. 본 분석결과, 제안된 DMC-CBR 기법은 기존의 C-CBR 기법보다 통계적으로 유의하게 높은 예측성능을 도출하고 있으며, 특히 C-CBR 기법의 주변부의 예측성능문제를 효과적으로 해결할 수 있는 방법임을 알 수 있었다. 이러한 예측성능 향상은 일반적인 CBR 기법보다 현저하게 낮은 실시간 계산비용으로 도출된 것으로, 제안된 DMC-CBR 기법이 기존 CBR 기법의 실시간 계산비용문제를 효과적으로 해결하면서도, 유사한 수준의 예측성능을 도출할 수 있음을 알 수 있다.

〈표 3〉 기법 별 성능에 대한 순위분석

순위	1	2	3
정확도	CBR (0.87)	BM-CBR (0.865)	C-CBR (0.854)
실시간 계산회수	C-CBR (66)	BM-CBR (87)	CBR (900)

Ⅴ. 결 론

본 연구는 사례기반 추론기법(CBR)이 과거데이터의 양이 방대하거나 또는 데이터의 축적 속도가 빠를 경우 계산비용(computational cost)이 급격히 높아지는 문제점을 지적하고, 이를 해결하기 위해 제안된 기존의 클러스터링 사례기반 추론기법(C-CBR)에 대한 장단점을 실증적으로 분석하였다. 특히, 기존의 C-CBR 기법이 클러스터의 주변부에 위치한 목표사례들에 대해 정확한 예측서비스를 제공하지 못할 뿐만 아니라, 적절한 클러스터 수가 사용되지 않았을 경우에는 예측성능이 현저히 저하될 수 있음을 심장병 환자에 대한 사례분석을 통하여 확인하였다.

본 연구는 이러한 문제점을 극복하기 위하여, 이웃사례의 탐색영역을 동적으로 조정하는 새로운 하이브리드(hybrid) 사례기반 추론기법인 Dynamically Merged-Clustering CBR(DMC-CBR) 기법을 제안하였다. 제안된 기법은 목표사례의 클러스터 내 위치에 따라서 이웃사례(neighbours)를 복원하는 풀(pool)을 동적으로 변화시킨다. 본 기법을 실제 심장병 환자 1,000명의 데이터에 적용하여 실험한 결과, 제안된 DMC-CBR 기법이 기존의 CBR 기법에 비해 현저히 낮은 계산비용을 사용하면서도, 이와 동일한 수준의 예측성능을 도출함을 확인하였다. 또한 기존의 C-CBR 기법이 클러스터 주변부의 목표사례들의 예측이 정확히 수행되지 못하는 문제를 극복하고, 보다 정확한 예측을 수행할 수 있음을 확인하였다. 이러한 연구는 사례기반 추론기법이 대용량 데이터에서도 효율적인 실시간 예측을 수행할 수 있도록 하여, 실제 사례가 정형화된 모델이나 도메인 지식보다 중요시 되는 실세계의 많은 문제에 효과적으로 활용될 수 있도록 하였다는 점에서 의의가 있다.

본 연구의 다음과 같은 한계점이 있다. 첫째, 실시간 계산비용 산출 시, 거리계산회수만을 사용하여, 정렬에 소요되는 연산 및 복원된 사례를 통해 예측을 수행하는데 소요되는 연산을 고려되지 않았다는 점이다. 이는 정렬 및 복원된 사례를 활용한 예측에 요구되는 실시간 계산비용이 사례복원에 소요되는 비용에 비해 상대적으로 적기 때문이나, 더욱 정확한 성능평가를 위해서는 이에 대한 부분을 함께 고려할 필요가 있다. 둘째, 제안된 DMC-CBR 기법은 이웃사례 탐색 풀(pool)을 최대 두 개의 클러스터로 한정하였다는 점이다. 이는 탐색 영역 확장에 따른 계산비용 증가를 방지하기 위함이나, 향후 상황에 따라서 추가적으로 탐색 영역을 확장하는 연구를 수행할 필요가 있다. 마지막으로, 본 연구에 사용된 데이터가 심장병 환자 1,000명으로 한정되었다는 점도 한계로 생각할 수 있다. 본 연구는 데이터의 증가에 따른 확장성을 연구하였기 때문에, 데이터 수의 절대적인 크기 보다는, 각 기법별 실시간 거리계산회수의 차이에 중점을 두고 연구를 수행하였으나, 향후 보다 다양한 도메인의 대용량 데이터 셋에 본 기법을 적용하여, 보편적인 결과를 도출할 필요가 있겠다.

참 고 문 헌

Aamodt, A. and E. Plaza, “Case-based reasoning: foundational issues, methodological variations, and system approaches”, *AI communications: the European journal on artificial intelligence*, Vol.7, No.1, 1994, pp. 39-59.

Blake, C. L. and C. J. Merz, *UCI Repository of Machine Learning Database*, Department of Information and Computer Science, University of California, Irvine, CA, http://www.ics.uci.edu/~

mlearn/MLRepository.html, 1998.

Hochbaum, S. D. and B. D. Shmoys, *A Best Possible Heuristic for the K-Center Problem*, Math. Operational Research, 1985.

Hong, T.-P. and Y.-L. Liou, *Case-based reasoning with feature clustering*, 7th IEEE International Conference on Cognitive Informatics, 2008, pp. 449-454.

Khan, A. S. and A. Hoffmann, "Building a case-based diet recommendation system without a knowledge engineer", *Artificial intelligence in medicine*, Vol.27, No.2, 2003, pp. 155-179.

Khan, M. J., M. M. Awais, and S. Shamail, "Self-Configuration in Autonomic Systems Using Clustered CBR Approach", *ICAC International Conference on Autonomic Computing*, 2008, pp. 211-212.

Kim, K.-S. and I. Han, "The cluster-indexing method for case-based reasoning using self-organizing maps and learning vector quantization for bond rating cases", *Expert systems with applications*, Vol.21, No.3, 2001, pp. 147-156.

Koton, P., "Reasoning about evidence in causal explanations", *Case-Based Reasoning*, 1988, pp. 260-270.

Li, Y., S. C. Shiu, and S. K. Pal, "Combining feature reduction and case selection in building CBR classifiers", *IEEE Transactions on Knowledge and Data Engineering*, Vol.18, No.3, 2006, pp. 415-429.

Montani, S., P. Magni, R. Bellazzi, C. Larizza, A. V. Roudsari, and E. R. Carson, "Integrating model-based decision support in a multi-modal reasoning system for managing type 1 diabetic patients", *Artificial intelligence in medicine*, Vol.29, No.1/2, 2003, pp. 131-151.

Park, Y. J., S. H. Chun, and B. C. Kim, "Cost-sensitive case-based reasoning using a genetic algorithm: Application to medical diagnosis", *Artificial Intelligence in Medicine*, Vol.51, No. 2, 2011, pp. 133-145.

Park, Y. J., B. C. Kim, and S. H. Chun, "New knowledge extraction technique using probability for case-based reasoning: application to medical diagnosis", *Expert Systems*, Vol.23, No.1, 2006, pp. 2-20.

Park, Y. J., B. C. Kim, and S. H. Chun, "An interactive case-based reasoning method considering proximity from the cut-off point", *Expert Syst. Appl.*, Vol.33, No.4, 2007, pp. 903-915.

Porter, B. W., R. Bareiss, and R. C. Holte, "Concept learning and heuristic classification in weak-theory domains", *Artificial Intelligence*, Vol.45, No.1, 1990, pp. 229-263.

Qiang, Y. and W. Jing, "Enhancing the Effectiveness of Interactive Case-Based Reasoning with Clustering and Decision Forests", *Applied intelligence*, Vol.14, No.1, 2001 pp. 49-64.

Schmidt, R. and L. Gierl, "Case-based reasoning for antibiotics therapy advice: an investigation of retrieval algorithms and prototypes", *Artificial intelligence in medicine*, Vol.23, No.2, 2001, pp. 171-186.

Turner, R., "Organizing and Using Schematic Knowledge for Medical Diagnosis", *Case-Based Reasoning*, 1988, pp. 435-446.

Witten, I. H. and E. Frank, *Data mining: practical machine learning tools and techniques with Java Implementations*, San Francisco: Morgan Kaufmann, 2000.

Information Systems Review
Volume 16 Number 1
April 2014

A Case-Based Reasoning Method Improving Real-Time Computational Performances: Application to Diagnose for Heart Disease

Yoon-Joo Park*

Abstract

Conventional case-based reasoning (CBR) does not perform efficiently for high volume dataset because of case-retrieval time. In order to overcome this problem, some previous researches suggest clustering a case-base into several small groups, and retrieve neighbors within a corresponding group to a target case. However, this approach generally produces less accurate predictive performances than the conventional CBR. This paper suggests a new hybrid case-based reasoning method which dynamically composing a searching pool for each target case. This method is applied to diagnose for the heart disease dataset. The results show that the suggested hybrid method produces statistically the same level of predictive performances with using significantly less computational cost than the CBR method and also outperforms the basic clustering-CBR (C-CBR) method.

Keywords: ***Case-Based Reasoning, Clustering, Computational Cost, K-Nearest Neighbors, High-Volume Dataset***

* Department of Business Administration, Seoul National University of Science and Technology

◐ 저 자 소 개 ◑

박 윤 주 (yjpark@seoultech.ac.kr)

고려대학교 컴퓨터학과에서 학부 및 석사학위를 취득하였으며, 한국과학기술원에서 경영공학 박사학위를 취득하였다. 이 후, New York University의 Stern Business School에서 초빙연구원으로 근무하였으며, 삼성생명 정보기획부서에서 과장으로 근무한 바 있다. 현재는 서울과학기술대학교 글로벌경영학과에서 조교수로 재직 중이다. 기존 연구는 IEEE Transactions on Knowledge and Data Engineering, Artificial Intelligence in Medicine, Expert Systems with Applications 등의 논문지에 게재되었다. 주요 연구 분야는 데이터마이닝을 이용한 질병 예측, 개인화 시스템, 그리고 온라인 매칭시스템 등이다.

논문접수일 : 2014년 01월 06일 **게재확정일 : 2014년 04월 09일**
1차 수정일 : 2014년 04월 06일