http://dx.doi.org/10.3745/KTSDE.2013.2.8.535

Optimizing Similarity Threshold and Coverage of CBR

Ahn, Hyunchul†

ABSTRACT

Since case-based reasoning(CBR) has many advantages, it has been used for supporting decision making in various areas including medical checkup, production planning, customer classification, and so on. However, there are several factors to be set by heuristics when designing effective CBR systems. Among these factors, this study addresses the issue of selecting appropriate neighbors in case retrieval step. As the criterion for selecting appropriate neighbors, conventional studies have used the preset number of neighbors to combine(i.e. k of k-nearest neighbor), or the relative portion of the maximum similarity. However, this study proposes to use the absolute similarity threshold varying from 0 to 1, as the criterion for selecting appropriate neighbors to combine. In this case, too small similarity threshold value may make the model rarely produce the solution. To avoid this, we propose to adopt the coverage, which implies the ratio of the cases in which solutions are produced over the total number of the training cases, and to set it as the constraint when optimizing the similarity threshold. To validate the usefulness of the proposed model, we applied it to a real-world target marketing case of an online shopping mall in Korea. As a result, we found that the proposed model might significantly improve the performance of CBR.

Keywords : Case-based Reasoning, Genetic Algorithm, Similarity Threshold, Coverage

사례기반추론의 유사 임계치 및 커버리지 최적화

안 현 철†

요 약

사례기반추론(CBR)은 많은 장점으로 인해 지금까지 의료진단, 생산계획, 고객분류 등 다양한 분야의 의사결정 지원에 적용되어 왔다. 그러나, 효과적인 CBR 시스템을 설계, 구축하기 위해서는 연구자가 직관적으로 설정해야 할 많은 설계요소들이 존재한다. 본 연구에서는 이러한 CBR의 여러 설계요소들 중 사례 검색 단계에서 결합할 이웃 사례들을 보다 효과적으로 선정할 수 있는 새로운 모형을 제시한다. 기존 연구에서는 결합할 이웃 사례를 선정하는 방법으로 사전에 정해진 이웃사례의 수(k-NN의 k)를 적용하든가, 혹은 최대 유사도의 상대적 비율을 임계치로 사용하는 방식을 적용해 왔다. 하지만, 본 연구에서는 결합할 유사사례를 선택하는 새로운 기준으로 0에서 1사이의 값을 갖는 절대적 유사 임계치를 사용할 것을 제안한다. 이 경우, 임계치 값이 과도하게 작아지게 되면, 예측결과의 생성이 잘 이루어지지 않을 수 있는 문제가 발생할 수 있다. 이에, 전체 학습사례들 중에서 예측결과가 생성된 사례의 비중을 커버리지(coverage)로 정의하고, 이를 유사 임계치 최적화 시 제약조건으로 설정함으로서, 사용자가 원하는 수준의 커버리지는 유지한 상태에서 가장 효과적인 유사 사례를 찾아 추론할 수 있도록 모형을 설계하였다. 제안 모형의 유용성을 검증하기 위해, 본 연구에서는 이 모형을 실존하는 국내 한 온라인 쇼핑몰의 표적 마케팅 사례에 적용하였다. 그 결과, 제안 모형이 CBR의 예측 성과를 유의미하게 개선시킬 수 있음을 확인할 수 있었다.

키워드 : 사례기반추론, 유전자 알고리즘, 유사 임계치, 커버리지

1. 서 론

분석 고객관계관리(Analytic CRM)에 있어 중요한 이슈 중 하나는 기업에서 팔고자 하는 상품을 구매할 가능성이 높은 잠재 구매자를 발굴하는 고객 분류 모형(customer classification model)을 구축하는 것이다. 고객 분류 모형은 다양한 마케팅 기회 창출에 활용될 수 있는데, 예를 들어 일대일 마케팅이나 DM(direct mailing) 발송을 통한 표적 마케팅, 전화나 이메일 등을 이용한 판매 촉진(sales promotion) 등에 있어 대상 고객을 선별하는데 유용하게 활용될 수 있다. 때문에 Ford와 같은 자동차 제조업체나 Allstate와 같은 보험사, 그리고 1-800-flowers.com과 같은 온라인 기업들을 포함한 전세계 유수의 기업들이 고객의 프로필 및 구매 행태를 분석해 상품 구매 가능성을 예측하는 고객 분류 모형을 구축하는데 많은 노력을 기울이고 있다[1].

※ 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 연구되었음(NRF-2011-332-B00104).
† 정 회 원 : 국민대학교 경영정보학부 조교수
논문접수 : 2013년 5월 14일
수 정 일 : 1차 2013년 5월 31일
심사완료 : 2013년 5월 31일
* Corresponding Author : Ahn, Hyunchul(hcahn@kookmin.ac.kr)

이러한 고객 분류 모형을 구축하는데에는 전통적으로 로지스틱 회귀분석(LR, Logistic Regression), 인공신경망(ANN, Artificial Neural Networks), 사례기반추론(CBR, Case-based Reasoning) 등 다양한 방법들이 적용되어 왔다[2-8]. 그 중에서도 특히 CBR은 적용이 쉽고, 유지보수가 상대적으로 편리하며, 실시간으로 연속해서 학습이 이루어진다는 장점으로 인해, 고객 분류 모형 구축[2,3]을 비롯한 여러 경영문제 해결에 널리 활용되어 왔다[9-11].

그러나 많은 장점에도 불구하고, 효과적인 CBR 시스템을 설계, 구축하기 위해서는 해결해야 할 문제들이 상당수 존재한다. 특히, CBR은 적절한 유사도 측정방법이나 사례 인덱싱 방법, 유사사례간의 결합방법 등에 대해 명확한 방법론이나 원리를 제공해 주지 못하고 있다. 때문에 이러한 CBR의 각종 설계 요소들은 실험자나 사용자가 자신의 경험이나 직관에 의해 결정해야 하는 어려움이 있었다. 특히 이러한 어려움은 때때로 연구자를 만족시키지 못하는 예측성과를 산출하는 문제를 야기시켜, 다른 방법론에 비해 CBR의 활용을 저해하는 요소로 작용하기도 하였다. 이런 이유로 지난 오랜 기간 동안 최적 사례 유사도 측정방법이나 사례의 특징을 대표하는 최적 변수군의 선정방법, 혹은 유사사례 결합시 적용할 가중치의 최적화 방법 등이 많은 연구자들에 의해 연구되어왔다[9-11].

이러한 여러 설계 요소들 중에 본 연구에서 집중적으로 개선시키고자 하는 요소는 바로 CBR에서 '예측결과를 생성할 때 참조하는 적절한 유사사례의 선택'과 관련된 요소이다. CBR 기법에서 결합할 유사사례의 개수를 최적화 하는 것이 CBR 기법의 성과를 향상시키는데 있어서, 매우 중요한 요소가 될 수 있다는 연구는 기존 연구에서 이미 여러 차례 제시된 바 있다[6,12-14]. 하지만, 기존 연구들에서는 모두 특정 사례수(k-NN의 k)나 유사도의 상대적 비율에 의거해 유사 사례를 선정함으로서, 경우에 따라 절대적 관점에서 유사하지 않은 기존 사례를 참고해서 예측결과를 생성할 수도 있는 구조적 한계점을 갖고 있었다[15].

이에 본 연구에서는 0에서 1사이의 값을 갖는 절대적 유사 임계치(similarity threshold)에 기반한 새로운 CBR 모형을 제안하고자 한다. 그런데, 절대적 유사 임계치는 값이 너무 작게 설정되면, 예측결과의 생성이 과도하게 이루어지지 않을 수 있다. 이에, 본 연구에서는 커버리지(coverage) 변수도 모형에 함께 반영함으로서, 사용자가 원하는 수준의 커버리지는 유지한 상태에서 가장 최적의 절대적 유사 임계치를 탐색하도록 하였다. 상기 두 변수의 동시 최적화 방법으로는, 전통적으로 많이 적용되어 온 유전자 알고리즘(GA, Genetic Algorithms)을 적용하였다. 본 연구는 이 모형의 우수성을 검증하기 위해, 국내 한 온라인 쇼핑몰의 특정 상품에 대한 표적 마케팅 대상 고객 발굴을 위한 데이터에 제안 모형을 적용해 보고, 과연 CBR 성과의 개선을 도모하는지 실증분석하였다.

본 논문의 뒷부분은 다음과 같이 구성된다. 우선 2장에서는 기존 문헌들에서 어떤 관련 연구들이 이루어졌는지 간략히 살펴보고, 3장에서는 본 연구의 제안 모형인 유전자 알고리즘 기반의 유사 임계치 및 커버리지 동시 최적화 CBR 모형을 소개한다. 4장에서는 앞서 제시한 모형의 유용성을 검증하기 위한 실험 데이터 및 실험 설계 내용을 설명하고, 5장에서는 실험 결과를 종합적으로 정리해 제시하도록 한다. 끝으로 마지막 장에서는 결언과 함께 향후 연구계획이 함께 제시된다.

2. 이론적 배경

본 연구에서 제안하는 모형은 기본적으로 CBR과 GA가 결합된 형태로 구성되어 있다. 이에 기존 문헌을 검토하게 될 본 절에서는 우선 CBR의 기본적인 개념과 원리에 대해 먼저 살펴보고, 이어 CBR과 GA를 결합하고자 시도한 다른 기존 연구들을 살펴보도록 한다. 그리고 끝으로, CBR 시스템에서 결합할 유사사례의 개수나 구성을 최적화 하고자 시도했던 기존 연구들을 살펴보고, 그 한계점을 살펴보고자 한다.

2.1 CBR

CBR은 과거 사례나 경험을 이용해, 주어진 문제에 대한 해답을 찾아내는 문제해결 방법론이다. 일반적으로 다른 주요 인공지능기법들은 문제와 해법 사이의 일반적인 관계를 도출하여 이를 기반으로 추론을 하는 원리로 이루어져 있어, 비교적 정형화된 문제 해결에만 적합하고, 지식도 지속적으로 갱신되기 어려운 구조적인 한계를 가지고 있다. 하지만, CBR은 과거에 축적된 정보만 있으면, 어떤 문제든 해결이 가능하므로 복잡하거나 비구조화된 문제를 해결하는데 유리하며, 지식기반을 지속적으로 업데이트 할 수 있다는 측면에서 상대적으로 우수하다고 할 수 있다[11].

CBR은 다음 Fig. 1에 제시되어 있는 것과 같이, 이른바 4R이라 불리는 크게 4단계의 절차에 의해 이루어진다[16]. 이 중에서, CBR 시스템의 효과를 결정짓는 가장 중요한 단계는 바로 1단계인 RETRIEVE 이다. 이 단계에서 시스템이 주어진 문제의 해결에 도움이 될 것으로 추정되는 사례들을 선택

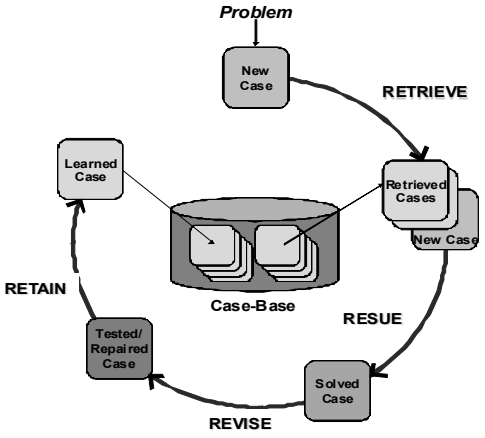

Fig. 1. Process of CBR
(Adopted from Aamodt and Plaza[16])

하게 되는데, '어떤 원리로 유사 사례들을 선별해서, 이들을 어떻게 조합해, 추천 결과를 만들어낼 것인가?' 하는 것에 따라 CBR 시스템의 성능이 크게 변화하기 때문이다. 때문에 사례간 유사도를 어떻게 측정할 것인가, 추천 결과를 도출할 때 유사 사례는 얼마나 결합할 것인가 하는 등의 문제는 전통적으로 주요 CBR의 연구주제로 자리매김하여 왔다[2].

2.2 GA를 이용한 CBR의 최적화

앞서 언급한 CBR의 최적화와 관련한 연구들에서 최적화 수단으로 가장 많이 활용되고 있는 알고리즘 중 하나가 바로 유전자 알고리즘(GA)이다. GA는 유전자, 선택, 교배, 돌연변이 등 생물학의 진화이론에 그 근간을 두고 있다. GA는 방대하고 복잡한 공간을 탐색하면서, 최적 혹은 최적에 가장 가까운 결과를 찾아주는 확률적 검색방법을 이용하는데, 이러한 특징 때문에 다양한 제약식을 포함한 상황에서 목적 함수(objective function)를 최적화 하는 '파라미터(parameter)' 추정에 널리 적용되고 있다[11].

CBR에 GA가 적용된 경우를 살펴보면, 우선 입력변수 선정에 GA를 적용한 Siedlecki와 Sklansky [17]의 연구와 GA를 입력변수의 범주화에 사용한 Kim과 Han[10]의 연구를 들 수 있다. GA를 CBR의 입력변수 가중치 선정에 적용한 기존 연구는 매우 많은데, Shin과 Han[11]의 연구나, Chiu[2], Chiu 등[9]의 연구가 대표적인 기존 연구라고 할 수 있다. 그 외 k-NN의 k를 최적화한 Ahn 등[6]의 연구와 여러 요소들을 GA로 동시에 최적화하고자 한 Kuncheva와 Jain[18], Ahn 등[5,7]의 연구 역시 GA로 CBR을 최적화를 시도한 기존 연구들이다.

2.3 결합 유사사례에 대한 최적화 연구

지금까지 살펴본 바와 같이, CBR 시스템에서 입력 변수의 선정 혹은 가중치 선정의 최적화와 관련해서는 지금까지 많은 연구가 이루어져 왔고, 그 결과 이미 다양한 방법들이 소개된 바 있다. 그러나, 결합 유사사례를 최적화 하기 위한 연구는 아직까지 그 수가 많지 않은 상태이다. 이 부분과 관련해 몇몇 기존 연구들을 살펴보면 다음과 같다.

우선, 이훈영과 박기남[13]의 연구는 k-NN의 k를 최적화하려고 시도하였는데, 유사도 분포에 따른 최적화 수리모형 기법을 제시하였다. 이 연구는 k-NN의 k 최적화를 시도한 첫번째 연구라는 측면에서는 의의가 있으나, 목표사례가 변화할 때마다 최적화 모형이 변화해 새로운 k값을 계속 계산해야 한다는 구조적 한계를 안고 있다.

Kim 등[19]은 유사사례의 수를 찾기 위해 교차검정방법(cross validation method)를 사용하여 학습용 자료에서의 평균제곱오차를 최소화하는 방식을 이용하였는데, 이 연구에서 제안한 방식은 최적유사사례의 수를 탐색하는 공간이 한정적이므로 전역 최적화된 유사사례 수를 제시할 수 없다는 한계를 가지고 있다. 이에 Ahn 등[6]은 상기 두 연구의 한계점을 극복하기 위해, GA를 활용한 유사사례 수 최적화 모형을 제안한 바 있다.

Sun과 Hui[20]는 유사사례 선정에 고정된 사례의 수(k-NN의 k)가 아닌, 상대적 비율값을 갖는 유사 임계치(similarity threshold)을 사용하는 방법을 제안하였다. 아래 Equation (1), (2)가 이러한 방식을 설명하고 있는데, 대상 사례 u_0의 예측결과 생성시 참고되는 유사 사례집단 U^*가 비율값으로 주어지는 상대적 유사 임계치 p에 의해 결정됨을 알 수 있다.

$$U^* = \{u_{i^*} | sim_{oi^*} \geq T\} \text{ where } i^* = 1,2,\cdots,k \quad (1)$$

$$T = p\max(sim_{oi}) \text{ for } \forall i \quad (2)$$

박윤주[14]는 앞서 살펴본 Sun과 Hui[20]와 유사하지만, 다소 차이가 있는 상대적 유사 임계치 사용을 제안하였다. 앞의 연구가 최대 유사도 거리 대비 상대적 비율을 유사 임계치로 사용했다면, 이 연구에서는 전체 학습 사례수 대비 상대적 비율을 유사 임계치로 제안하였다. 예를 들어, 이 연구모형에서는 유사 임계치를 5%로 설정할 경우, 전체 학습 사례 중 대상 사례와 가장 유사한 상위 5%의 결과를 참조해 예측결과를 생성하게 된다.

3. GA를 활용한 유사 임계치와 커버리지의 동시 최적화 모형

앞서 언급한 기존 연구들에서 제시하고 있는 CBR 모형들은 모두 특정 사례수(k-NN의 k)나 유사도의 상대적 비율에 의거해 유사 사례를 선정하고 있다. 때문에, 절대적인 관점에서 주어진 문제를 해결하기 위해 참조하기에 적합한 유사한 사례가 아님에도 불구하고, 다른 사례들과 비교할 때 그나마 상대적으로 더 유사하다는 이유로 다소 억지스러운 예측결과를 생성할 수도 있는 내재적인 문제점을 안고 있다.

스팸 메일(spam mail)의 사례에서 볼 수 있듯이, 표적 마케팅의 경우에는 상품을 원치 않는 사람에게 마케팅을 시도할 경우, 오히려 고객의 불만만 가져올 수 있는 위험을 갖고 있다. 때문에 이 경우에는 정말 확신할 수 있는 엄선된 표적에 대해서만 마케팅을 시도하는 것이 상당히 중요하다. 그런데, 기존의 CBR 시스템처럼 예측결과의 확신 정도에 관계없이 무조건 상대적으로 더 유사한 사례들을 참조해 결과를 생성하는 경우, 그 결과의 효과성은 떨어질 수 밖에 없을 것이다.

이에 본 연구에서는 0에서 1사이의 값을 갖는 절대적 유사 임계치(similarity threshold)에 기반한 새로운 CBR 모형을 제안하고자 한다. 즉, 절대적 유사 임계치를 기준으로 적용할 때, 유사한 사례가 하나도 나오지 않으면, 그 경우에는 예측결과를 생성하지 않고 '모름(don't know)'으로 결과를 회신할 수 있는 CBR 시스템을 제안하고자 하는 것이다.

이 같은 새로운 CBR 모형을 이진분류 문제에 적용할 경우, 기존 모형처럼 0 또는 1의 두가지 경우로만 결과를 예

측하지 않고, 3가지 경우(0, 1, 모름)로 결과를 제공하기 때문에, 확실한 예측 대상을 발굴하는 것이 중요한 의미를 지니는 고객 발굴 분야나 의료 진단 분야에 적용 시, 큰 가치를 지니게 될 수 있다.

그런데, 이처럼 절대적 유사 임계치를 사용할 경우, 그 값이 너무 작아지면 예측결과의 생성이 과도하게 이루어지지 않을 가능성이 있다. 이러한 한계를 보완하기 위해, 본 연구에서는 전체 학습사례들 중에서 예측결과가 생성된 사례의 비중을 의미하는 커버리지(coverage) 개념을 기반으로, 결과의 정밀도와 서로 상충(trade-off) 관계에 있는 커버리지(coverage) 변수를 모형에 함께 반영하였다. 즉, CBR에서 유사 사례 탐색 시, 사용자가 원하는 수준의 커버리지는 유지한 상태에서 가장 효과적인 유사 사례를 찾아, 추론을 수행할 수 있도록 하였다. 그리고, 이러한 서로 다른 특성을 가진 두 변수인 절대적 유사 임계치와 커버리지를 동시에 최적화하기 위해서, 전통적으로 최적화에 많이 적용되어 온 GA를 적용하였다.

본 연구에서 제안하는 새로운 CBR 모형의 전체적인 프레임워크를 설명하면, 다음의 Fig. 2와 같다.

Fig. 2에서 볼 수 있듯이, 본 연구는 GA를 CBR 파라미터 최적화에 적용한 기존 연구를 참고하여 4개의 단계로 제안 모형을 구성하고 있다[21].

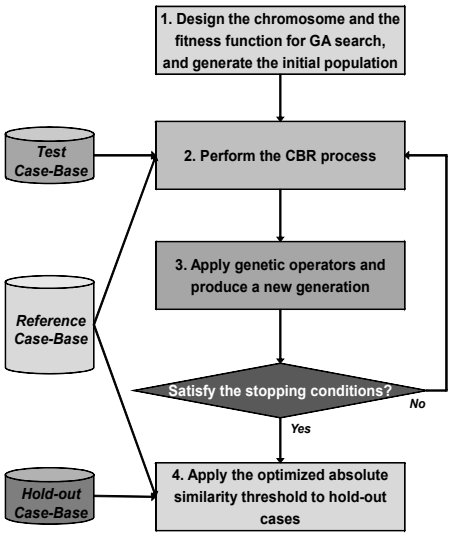

Fig. 2. Framework of the proposed model

1단계. GA 탐색을 위한 염색체 설계 및 초기 개체군 생성

1단계에서는 우선 절대적 유사 임계치 값이 가질 수 있는 전체 탐색 공간 중에서, 가장 최적 혹은 최적에 가까울 것으로 추정되는 유사 임계치 값들을 임시로 설정하게 된다. 개체군(최적의 유사 임계치를 찾아내기 위한 초기 탐색 위치들의 집합)은 본격적인 탐색 프로세스에 앞서, 무작위값으로 초기화되고, 탐색의 대상이 되는 유사 임계치는 GA가 인식할 수 있도록 염색체(chromosome) 형태로 코드화 된다. 이렇게 코드화된 염색체는 특정 적합도 함수(fitness function)를 최대화 하는 방향으로 진화해 나가게 된다. 본 연구에서 염색체는 유사 임계치가 1/10,000의 정확도를 갖는 0에서 1사이의 값을 가질 수 있도록, 14비트의 이진수로 모델링하였다. 적합도 함수는 CBR연구에서 일반적으로 가장 많이 적용되는 실험용 데이터(test data)에 대한 평균 예측 정확도로 설정하였다.

2단계. 사례기반추론의 적용

1단계 과정을 통해 개체군이 도출되면, 여기서는 개체군에 속한 개별 염색체들을 대상으로 CBR을 작동하게 된다. 본 연구에서 절대적 유사 임계치를 적용하기 위해서는, 유사도 값이 어떤 경우에도 0에서 1사이의 절대값을 갖도록 산출하는 것이 기본적으로 요구된다. 이에, 본 연구에서는 모든 입력변수들(a_k)에 최소-최대 정규화(min-max normalization)를 적용한 뒤, 아래 Equation (3)과 같이 유클리드 거리(Euclidean distance) 기반의 유사도를 계산하였다.

$$sim(u_0,u_i) = \frac{\sum_{k=1}^{m}(a_k^0 - a_k^i)^2}{m} \quad (3)$$

$$\text{where } u_i = (a_1^i, a_2^i, \cdots, a_m^i),\ k = 1,2,\cdots,m$$

이 때, 모든 a_k는 $0 \le a_k \le 1$을 만족하므로, 어떤 경우에도 유사도 값은 0에서 1사이의 값을 갖게 된다. 본 연구에서는 최적의 결합 유사사례에만 관심을 갖고 있는 상황이므로, 입력변수의 가중치 등과 같은 다른 CBR의 설계요소들은 대부분의 CBR 시스템에서 적용하는 것과 같이 동일 가중치를 적용하였다[9].

3단계. 유전 조작 수행을 통한 새로운 개체군 생성

3단계에서는 2단계의 결과들, 즉 각 후보 유사 임계치 적용시 도출되는 실험용 데이터의 평균 예측 정확도에 의거해, 우수한 유사 임계치들을 선별하고 이들에게 각종 유전 조작을 수행해 새로운 개체군을 생성하는 작업이 이루어지게 된다. 이 단계에서 기설정된 목표 커버리지를 달성하지 못하는 유사 임계치들 역시 선택에서 제외되게 된다. 여러 유전 조작 기법 중 본 연구에서는 선택(selection), 교배(crossover), 돌연변이(mutation)의 3가지를 적용하며, 사전에 설정한 중지 조건이 만족될 때까지 2단계, 3단계 작업을 계속 반복한다.

4단계. 검증용 데이터에 적용 및 성능 확인

앞의 단계들이 모두 끝나면, GA의 중지 조건이 만족되는 시점에서 최적 혹은 최적에 근접한 유사 임계치가 도출되게 된다. 4단계에서는 이렇게 도출된 유사 임계치를 이용, CBR 모형 구축에 사용하지 않은 검증용 데이터(hold-out data)에 적용함으로서, 그 결과를 살펴보게 된다. 본 단계를 통해, 연구의 제안모형의 실제적인 적용가능성(general applicability)을 확인해 볼 수 있다.

4. 실험 및 결과

4.1 실험 데이터

제안된 연구모형의 유용성을 검증하기 위해, 본 연구에서는 국내 한 다이어트 인터넷 쇼핑몰의 고객 분류 모형 구축 사례에 제안 모형을 적용하였다. 본 연구의 대상이 된 쇼핑몰은 다이어트와 관련한 정보 제공, 커뮤니티 서비스, 쇼핑몰 등 원스톱(one-stop) 서비스를 제공하는 다이어트 전문 포털 사이트이다. 이러한 다이어트 사이트의 경우, 보다 정확하고 맞춤화된 서비스를 받기 위해 고객이 본인에 대한 상세한 정보를 입력해야만 하는데다, 대체로 사이트에 대한 이용 목적이 분명한 고객들이 주로 방문하기 때문에, 많은 고객들이 양적인 측면이나 질적인 측면에서 우수한 본인의 개인정보를 서비스 제공업체에 기꺼이 제공하는 경향이 있다. 따라서, 본 사례의 대상업체는 고객들에 대한 상당히 자세하고도 정확한 정보를 많이 보유하고 있는 상황이며, 이로 인해 이를 기업의 새로운 마케팅 기회로 활용하고자 하는 동기를 가지고 있다. 특히, 이 쇼핑몰에서는 여러 상품을 취급하고 있지만, 그 중에서 수익이 가장 큰 '다이어트 보조식'의 판매에 큰 관심을 갖고 있다. 이에, 회원가입시 입력되는 고객 관련 정보를 활용해, '다이어트 보조식'의 구매여부를 예측하는 고객 분류 모형의 구축에 많은 관심을 갖고 있다. 이에 본 연구에서는 '다이어트 보조식 제품에 대한 표적 마케팅'에 적절한 대상 고객을 산출할 수 있도록, 본 연구의 제안모형을 적용해 보고자 하였다.

본 연구를 위해 정제한 데이터는 구매 및 비구매고객이 1:1의 비율로 혼합된 총 980건의 데이터였다. 종속변수는 대상업체에게 가장 높은 마진을 제공하는 다이어트 보조식품 관련 상품의 구매여부 변수로서, 구매한 고객의 경우 1을, 구매하지 않은 고객의 경우 0을 값으로 부여하였다. 종속변수를 예측하기 위해 활용한 독립변수로는 회원 가입시 입력되는 성별, 나이, 체중, 키 등 다이어트와 관련한 인구통계적인 변수들 중 총 46개가 수집되었다. 이러한 입력변수 중 종속변수의 예측에 관련성이 없는 변수를 사전에 제거하기 위해 독립표본 t검정(independent samples t-test)과 카이제곱 검정(chi-square test)을 적용해 총 14개의 변수를 CBR의 입력변수로 최종 선정하였다. 다음의 Table 1은 선택된 입력변수에 대한 상세한 정보를 설명하고 있다.

아울러 본 연구에서는 제안모형을 구축, 검증하기 위해 전체 수집된 데이터를 참조용, 테스트용, 검증용 사례기반 등 총 3개의 그룹으로 구분하였다. 이 3가지 사례기반(데이터셋)은 각각 전체 데이터의 60%(588건), 20%(196건), 20%(196건)의 비중을 차지하도록 적절하게 배분되었다.

4.2 실험 설계

GA 탐색을 위한 제어 파라미터들과 관련해서는 개체군의 규모를 50개체(organisms)로 설정하였으며, 교배 및 돌연변이 비율에 대해서는 각각 0.7, 0.1로 설정하였다. 아울러 중지 조건으로는 1000회 반복, 즉 20세대만큼 탐색을 반복하도록 설정하였다.

Table 1. Selected features and their descriptions

Feature name	Description	Range
AGE	Age	Continuous (years)
ADD0	I live in Seoul, Korea	0 : False 1 : True
ADD1	I live in metropolitan areas in Korea	0 : False 1 : True
OCCU0	I am a company employee	0 : False 1 : True
OCCU2	I am a student	0 : False 1 : True
OCCU4	I am a business owner	0 : False 1 : True
GENDER	Gender	0 : Male 1 : Female
LOSS4	I hope to lose weight around my legs and thighs	0 : False 1 : True
PUR0	I am on diet for beauty	0 : False 1 : True
HEIGHT	Height	Continuous (cm)
BMI	Body Mass Index(BMI) is the measure of body fat based on height and weight that applies to both adult men and women. It is calculated as follows: $BMI(kg/m^2) = \frac{weight(kg)}{height(m)^2}$	Continuous (kg/m^2)
E01	I've experienced 'functional diet food'	0 : False 1 : True
E02	I've experienced 'diet drugs'	0 : False 1 : True
E05	I've experienced 'one food diet'	0 : False 1 : True

제안모형이 기존의 다른 통계 및 인공지능 기법들에 비해 얼마나 더 개선된 성과를 보여줄 수 있는지 검증하기 위해, 동일한 데이터셋에 로지스틱 회귀모형(logistic regression), 다중판별분석(multiple discriminant analysis), 인공신경망(artificial neural network), SVM(support vector machine) 등 총 4개의 비교모형을 확보된 데이터에 적용하였다. 전통적인 CBR, 즉 k-NN 역시 적용해 보고, 그 성과를 제안모형과 비교했다[5].

로지스틱 회귀분석의 경우, 전진선택법(forward selection procedure)을 사용하였으며, 이 때, 단계별 변수입력 확률은 0.05로 설정하였다. 다중판별분석의 경우, Wilks' lambda를 활용한 입력변수의 단계별 선택방법을 활용하였는데, 이 때 변수의 입력 혹은 제거의 기준으로는 F값을 사용하였다. 이상의 통계기법들의 경우, IBM SPSS Statistics 20.0을 이용해 실험을 수행하였다.

인공신경망에 대해서는 입력층과 출력층 사이에 은닉층을 1개 포함하는 3계층 역전파 망을 적용하였다. 인공신경망의 학습율과 모멘텀율은 각각 0.1씩 설정하였으며, 은닉층과 출

력층의 노드들은 시그모이드 전이함수(sigmoid transfer function)를 사용하게끔 설계하였다. 은닉층의 노드수와 관련해서는 7, 14, 21, 28등 4가지 경우를 모두 대입해 보고 실험해 보았으며, 그 중에서 가장 우수한 결과를 보이는 은닉층의 노드수를 설정하고자 하였다. 아울러, 학습중지조건으로는 총 150차례 전체 학습데이터에 대한 학습을 반복하게끔 설정하였다. 인공신경망과 관련한 실험은 상용 인공신경망 소프트웨어인 Neuroshell R4.0을 활용해 실험을 수행하였다.

SVM의 경우, 주어진 자료들을 고차원 공간으로 맵핑(mapping)하는 커널함수를 어떤 함수로 사용하는가에 따라 성과가 달라질 수 있기 때문에, 일반적으로 여러가지 커널함수를 모두 실험해 보고 가장 우수한 성과를 보이는 커널함수를 선정한다. 본 연구에서도 선형, 다항식, 그리고 가우시안(Gaussian) RBF 등 총 3개의 커널함수를 적용하여, 가장 우수한 성과를 보이는 결과를 최종적으로 선정하였는데, 이 3가지 커널함수에 대한 수식이 다음의 식 (4)-(6)에 제시되어 있다.

$$\text{선형 커널함수: } K(\mathrm{x}_i,\mathrm{x}_j) = \mathrm{x}_i{}^{\mathrm{T}}\mathrm{x}_j \quad (4)$$

$$d\text{차원의 다항식 커널함수: } K(\mathrm{x}_i,\mathrm{x}_j) = (1+\mathrm{x}_i{}^{T}\mathrm{x}_j)^d \quad (5)$$

$$\text{Gaussian RBF 커널함수: } K(\mathrm{x}_i,\mathrm{x}_j) = e^{-\frac{(\mathrm{x}_i-\mathrm{x}_j)^2}{\sigma^2}} \quad (6)$$

Tay와 Cao[22]는 SVM의 성과를 결정짓는데 있어서, 상한계수 C나 d, σ^2와 같은 커널함수 내 매개변수들의 값에 대한 설정이 중대한 영향을 미칠 수 있음을 지적하였다. 만약 이러한 매개변수 값들이 적절하게 설정되지 않은 경우, SVM은 과적합화(overfitting)나 혹은 불충분적합(underfitting) 될 수 있기 때문이다. 때문에, 본 연구에서는 상기 매개변수들의 값을 다양하게 바꾸어가면서 실험하여, 가장 우수한 성과를 보이는 매개변수 값들을 최종적으로 선택하고자 하였다. SVM 실험을 위한 실험도구로는 공개 소프트웨어인 LIBSVM version 2.8을 활용하였다.

CBR을 위한 실험용 소프트웨어는 Microsoft Excel 2010과 그 안에 내장된 Visual Basic for Application(VBA)를 이용해 개발하였다. 전통적인 CBR(k-NN) 모형의 경우에는 IBM SPSS Statistics 20.0버전에 내장된 가장 가까운 이웃 기능을 이용해 실험하였다. 그리고 GA의 경우, MS-Excel plug-in 형태의 상용 소프트웨어인 Palisade Software사의 Evolver Industrial Version 5.5를 이용해 실험하였다.

4.3 실험 결과

제안모형이 앞서 제안한 비교모형들에 비해 얼마나 우수한 예측결과를 생성하는지에 대한 확인은 검증용 데이터에 대한 구매 예측 정확도를 비교해 보는 방식을 통해 수행하고자 한다. 이 때, 예측 정확도(Hit)는 다음의 식 (7)을 통해 산출된다.

$$Hit = \frac{1}{n}\sum_{i=1}^{n} CA_i \text{ where } CA_i = 1 \text{ if } PO_i = AO_i \quad (7)$$
$$CA_i = 0 \text{ if } PO_i \neq AO_i$$

PO_i : i번째 검증용 데이터에 대해 모형이 예측한 결과값 (1:구매, 0:비구매)

AO_i : i번째 검증용 데이터에 대한 실제 결과값 (1:구매, 0:비구매)

우선 전통적인 CBR 실험을 위해, 우리는 k-NN의 k값으로 1에서 9사이의 홀수값을 모두 대입해 본 다음, 가장 우수한 성과를 보이는 k를 선택하고자 하였다. 그 결과, Table 2에 제시되어 있듯이, 3-NN이 가장 우수한 성과를 보임을 확인할 수 있었다.

Table 2. Experimental results of conventional CBR(k-NN)

k of k-NN	1	3	5	7	9
Prediction accuracy of hold-out case base	52.04%	56.12%	55.61%	54.08%	53.57%

방금 소개한 k-NN을 비롯한 모든 비교모형과 제안모형에 대한 실험결과를 종합적으로 정리한 결과는 다음의 Table 3에 제시되어 있다. 제안모형의 경우, 제약조건으로 최소 충족 커버리지를 얼마로 설정하는가에 따라 결과값이 달라질 수 있다. 이에 본 연구에서는 제안모형을 모형A ~ C로 구분하고, 각각 최소 충족 커버리지를 80%와 50%, 그리고 20%로 나누어 설정한 뒤 실험을 수행하였다.

이 표에서 확인할 수 있듯이, 전통적인 CBR은 예측성과가 56%대로 나타나, 60% 초반대 성과를 나타내는 ANN, SVM은 물론 통계기반의 다른 비교모형들과 비교해도 상대적으로 낮은 예측성과를 보이고 있다. 즉, 연구의 배경과 목적에서 설명했던 바와 같이, 전통적인 CBR을 예측의 용도로 활용할 경우, 낮은 정확도로 인한 손실은 불가피함을 본 연구의 실험결과를 통해서도 다시 한 번 확인할 수 있다.

하지만, 본 연구에서 제안하는 절대적 유사임계치를 적용하게 되면, 기존 모형과 대등하거나 훨씬 향상된 예측성과를 나타냄을 확인할 수 있다. 본 연구의 실험결과를 보면, 본 연구의 제안모형이 최대 70% 초반대까지 예측정확도를 향상시킬 수 있음을 확인할 수 있다. 물론 비교모형과 달리, 본 연구의 제안모형은 100% 모든 예측대상에 대해 예측결과를 제시한 것은 아니라는 한계점이 있지만, 본 연구의 적용대상이 되고 있는 마케팅(CRM) 분야라든가, 의료분야 등 일부 특수한 분야의 경우 오히려 이러한 특징이 더 유익하게 활용될 수 있을 것으로 기대된다.

다음의 Table 4에는 제안 모형 A ~ C에 대한 보다 구체적인 실험결과가 제시되어 있다. 이 표에서 볼 수 있듯이, 절대 유사 임계치를 낮게 가져갈수록 커버리지는 감소하지만, 정확도는 대체로 상승하는 패턴을 보임을 알 수 있다. 그리하여, 목표 커버리지를 20% 이상으로 설정할 경우, 제

Table 3. Overall results of the proposed models and the comparative models

Model	Reference data set	Test data set	Hold-out data set	Remarks
LOGIT	63.30%		62.76%	Forward feature selection
MDA	63.10%		62.76%	Stepwise feature selection (byWilks'lambda)
ANN	68.37%	66.84%	61.73%	No. of the nodes in the hidden layer: 14
SVM	65.82%		63.27%	Kernel function: Gaussian RBF C=1 and σ^2=75
CBR		62.24%	56.12%	k of k-NN = 3
Proposed Model A (80%Cov.)		70.37%	66.00%	Threshold = 0.05668
Proposed Model B (50%Cov.)		73.26%	67.11%	Threshold = 0.03571
Proposed Model C (20%Cov.)		71.74%	71.43%	Threshold = 0.01806

Table 4. Results of the proposed models

Models		Proposed Model A	Proposed Model B	Proposed Model C
Target coverage		80%	50%	20%
Optimized absolute similarity threshold		0.05668	0.03571	0.01806
Coverage	Test data set	91.53%	73.73%	38.98%
	Hold-out data set	84.75%	64.41%	35.59%
Prediction accuracy	Test data set	70.37%	73.56%	71.74%
	Hold-out data set	66.00%	67.11%	71.43%
Average number of neighbors (k of k-NN)	Test data set	8.31356	3.55085	0.78814
	Hold-out data set	7.87288	2.89831	0.55085

안모형이 검증용 데이터 기준으로 대략 36% 정도의 대상 고객에 대해 약 71.74% 수준의 높은 정확도로 고객의 구매 여부를 예측할 수 있음을 확인할 수 있다. 본 연구의 제안 모형은 절대적 유사 임계치를 기준으로 참조할 유사 사례를 선정하기 때문에, 매번 참조하는 사례의 수(k-NN의 k)가 변화하게 된다. 각 제안모형의 유형별로 그 평균값을 구해 보면, 검증용 데이터셋을 기준으로 모형 A에서는 약 7.9개, B에서는 2.9개, C에서는 0.6개를 참조하고 있음을 알 수 있다. 결국, 목표 커버리지가 낮아질수록 가장 유사한 것으로 판단되는 엄선된 소수의 사례만으로 예측결과를 생성하고 있음을 알 수 있다.

5. 결 론

본 논문에서는 이른바 '절대적 유사 임계치'와 '커버리지' 개념을 도입하여, 사용자가 요구하는 정확도 수준에 따라 예측결과를 생성하는 새로운 개념의 CBR 시스템을 제안하였다. 본 연구의 제안모형은 이진분류 문제에서, 확실한 예측결과를 생성하기 어려운 경우, '모름'으로 회신함으로서 보다 확실한 경우에 대해서만 결과를 생성하는 것이 가능하도록 설계되어 있다. 이러한 본 연구의 제안모형은 확실한 예측 대상을 발굴하는 것이 중요한 의미를 지니는 표적 마케팅 분야나 의료 분야에 특히 유익할 것으로 전망된다.

학술적으로, 본 연구는 기존에 존재하지 않았던 새로운 CBR 방법론을 제시한다는 의의가 있다. 앞서 소개된 실증분석 결과를 통해 확인했듯이, 제안된 알고리즘을 이용해 보다 정밀하게 CBR의 유사 결합사례를 찾아낼 경우, CBR의 예측력 개선에 크게 기여할 수 있다. 일반적으로 CBR은 적용이 쉽고 간편하며, 적은 수의 학습용 자료만으로도 예측결과 생성이 가능할 뿐만 아니라, 예측결과에 대한 설명력(explainability)도 갖고 있어 상당히 매력적인 인공지능 기법이다. 하지만, 상대적으로 낮은 예측 정확도가 현업에서 이 기법을 활용하는데 주요한 걸림돌이 되고 있다. 따라서 CBR의 예측성과를 개선할 수 있는 새로운 방법론을 제안하는 것은 상당히 높은 학술적 가치를 지닌 연구라고 사료된다.

실무적으로도 본 연구의 기대효과는 상당히 높다고 하겠다. 최근 고객정보의 저장 비용이 급격히 저렴해지고, 웹, 스마트폰 등 고객과의 직접적인 커뮤니케이션을 가능하게 하는 매체 역시 빠르게 확대되면서 CRM, 그 중에서도 특히 표적 마케팅에 대한 기업들의 관심이 높아지고 있다. 하지만, 표적 마케팅을 잘못 수행할 경우, 당초 기대했던 매출 증대는커녕 오히려 고객들에게 스팸으로 인식되어, 그들의 충성도를 떨어뜨리는 역효과를 가져올 수도 있다. 이러한 상황에서 본 연구가 제안하는 새로운 CBR 알고리즘은 이른바 '절대적 유사 임계치'와 '커버리지' 개념을 도입하여, 사용자가 요구하는 정확도 수준에 따라 예측결과를 생성하는 유연한 방법론을 제안하고 있다. 때문에 이러한 본 연구의 제안 알고리즘은 확실한 예측 대상을 발굴하는 것이 중요한 의미를 지니는 표적 마케팅 분야에 상당히 유용하게 활용될 수 있을 것으로 전망된다.

예를 들어, 본 연구의 제안모형을 기업의 표적 마케팅에 적용할 경우, 확실히 구매를 하지 않을 것으로 예상되는 대상 고객에 대해서는 아무런 반응을 하지 않고, 살지 말지 잘 모를 대상에 대해서는 최대한 비용 효율적인 수단(예. SMS 혹은 E-mail 전단지)을 사용하는 방안을 도입할 수 있다. 물론 제안모형의 예측결과 확실히 구매할 것 같은 대상에 대해서는 콜센터 상담원이 직접 전화해서 설득하든가, 필요시 방문상담도 추진해 볼 수 있다.

비단 표적 마케팅 분야 뿐 아니라, 확실한 위험 환자를 식별, 관리해야 하는 의료분야나 확실한 주가 상승 또는 하락 패턴 발견 시 거래를 수행하는 것이 바람직한 트레이딩 시스템과 관련된 금융 분야에서도 제안 알고리즘을 적용하는 것이 가능하다. 예를 들어 질병 예측 분야에 적용할 경우, 확실하게 그 질병에 감염된 경우 바로 집중적인 치료 조치단계를 취하고, 거의 확실하게 감염되지 않은 것으로 확인되는 경우는 그 사람이 일상생활을 자유롭게 즐길 수 있도록 보장하며, 감염 여부가 불확실한 경우 일단 대상자

의 일상생활을 보장하지만 보다 상세한 검사를 실시하는 등의 보완책을 구사할 수 있다[23]. 또한, 주가지수에 대한 등락 예측에 제안모형을 적용할 경우, 익일 지수가 확실하게 오를 것으로 예상되는 경우에는 콜옵션 매입 포지션을 취하고, 익일 지수가 확실하게 내릴 것으로 예상되는 경우에는 풋옵션 매도 포지션을 취하며, 익일 지수 향방이 불확실한 경우 특별한 액션을 취하지 않도록 함으로서 수익 극대화를 추구할 수 있다[24]. 이러한 관점에서 볼 때, 본 연구의 실무적 기대효과는 매우 높으며, 향후 타 경영분야의 적용 연구로 자연스럽게 확장, 발전될 수 있을 것으로 기대된다.

본 연구의 한계는 다음과 같다. 현재 제안된 모형은 전통적인 CBR에 비해 예측성과를 크게 개선시키고 있음을 알 수 있지만, 절대적 관점에서 볼 때 아주 획기적으로 개선시키지는 못하고 있는 것으로 판단된다. 특히, 본 연구의 제안모형은 전체 모든 사례에 대해 해답을 제공하지 못하고 커버리지에 포함되는 일부에 대해서만 해답을 제공하는데, 이러한 손실 대비 예측성과의 개선효과는 기대에 못 미치게 나타나고 있는 것이 사실이다. 따라서, 현재 제안된 연구모형은 개선이 요구되는 상황인데, 저자는 '참조되는 사례기반(즉, 학습사례)에 대한 정제 과정의 부재'가 현 제안모형의 가장 큰 문제점인 것으로 예상하고 있다. 유사 임계치에 의한 유사사례 선정이 의미를 갖기 위해서는 기본적으로 모든 참조사례(학습사례)가 대표성을 갖는 정제된 사례라는 가정이 선행되어야 하는데, 현실세계의 참조사례들은 소위 이상치(outlier)들을 상당수 포함하고 있을 가능성이 있다. 때문에 앞으로 현 모형에 이와 같은 결점을 보완할 수 있는 요소를 결합하여, 모형을 보완하는 연구가 추후 이루어져야 할 것으로 예상된다.

참 고 문 헌

[1] H. Ahn, K.-j. Kim, and I. Han, "Simultaneous Optimization Model of Case-Based Reasoning for Effective Customer Relationship Management," *Journal of Intelligence and Information Systems,* Vol.11, No.2, pp.175-195, 2005.

[2] C. Chiu, "A case-based customer classification approach for direct marketing," *Expert Systems with Applications*, Vol.22, No.2, pp.163-168, 2002.

[3] G.-H. Lee and D.-H. Lee, "Recommending System of Products on e-shopping malls based on CBR and RBR," *The KIPS Transactions: Part D*, Vol.11D, No.5, pp.1189-1196, 2004.

[4] V. Kumar and W. J. Reinartz, *Customer Relationship Management: A Databased Approach*, New Jersey: John Wiley & Sons, 2006.

[5] H. Ahn, K.-j. Kim, and I. Han, "Hybrid Genetic Algorithms and Case-based Reasoning Systems for Customer Classification," *Expert Systems*, Vol.23, No.3, pp.127-144, 2006.

[6] H. Ahn, K.-j. Kim, and I. Han, "Global optimization of feature weights and the number of neighbors that combine in a CBR system," *Expert Systems*, Vol.23, No.5, pp.290-301, 2006.

[7] H. Ahn, K.-j. Kim, and I. Han, "A Case-based Reasoning System with the Two-Dimensional Reduction Technique for Customer Classification," *Expert Systems with Applications*, Vol.32, No.4, pp.1011-1019, 2007.

[8] H. Ahn, J. J. Ahn, K. J. Oh, and D. H. Kim, "Facilitating Cross-selling in a Mobile Telecom Market to develop Customer Classification Model based on Hybrid Data Mining Techniques," *Expert Systems with Applications*, Vol.38, No.5, pp.5005-5012, 2011.

[9] C. Chiu, P. C. Chang, and N. H. Chiu, "A case-based expert support system for due-date assignment in a water fabrication factory," *Journal of Intelligent Manufacturing*, Vol.14, No.3-4, pp.287-296, 2003.

[10] K.-j. Kim and I. Han, "Maintaining case-based reasoning systems using a genetic algorithms approach," *Expert Systems with Applications*, Vol.21, No.3, pp.139-145, 2001.

[11] K. S. Shin and I. Han, "Case-based reasoning supported by genetic algorithms for corporate bond rating," *Expert Systems with Applications*, Vol.16, No.2, pp.85-95, 1999.

[12] J. M. Garrell i Guiu, E. Golobardes i Rib?, E. Bernad? i Mansilla, and X. Llor? i F?brega, "Automatic diagnosis with genetic algorithms and case-based reasoning," *Artificial Intelligence in Engineering*, Vol.13, No.4, pp.367-372, 1999.

[13] H. Y. Lee and K. Park, "Methods for Determining the Optimal Number of Cases to Combine in An Effective Case-Based Forecasting System," *Korean Management Review*, Vol.27, No.5, pp.1239-1252, 1999.

[14] Y. J. Park, "Case-Based Reasoning Methods based on Statistical Analysis," Doctoral Thesis, Division of Management Engineering, Korea Advanced Institute of Science and Technology, Seoul, Korea, 2006.

[15] H. Ahn, "Simultaneous optimization model of similarity threshold and coverage of the CBR system for target marketing," in *Proceedings of 2007 KMIS Fall Conference*, Seoul, Korea, pp.605-610, 2007.

[16] A. Aamodt and E. Plaza, "Case-based reasoning; Foundational issues, methodological variations, and system approaches," *AI Communications*, Vol.7, No.1, pp.39-59, 1994.

[17] W. Siedlecki and J. Sklanski, "A note on genetic algorithms for large-scale feature selection," *Pattern Recognition Letters*, Vol.10, No.5, pp.335-347, 1989.

[18] L. I. Kuncheva and L. C. Jain, "Nearest neighbor classifier: Simultaneous editing and feature selection," *Pattern Recognition Letters*, Vol.20, No.11-13, pp.1149-1156, 1999.

[19] T. S. Kim, J. H. Yoon, and H. K. Lee, "Performance of a nonparametric multivariate nearest neighbor model in the prediction of stock index returns," *Asia Pacific Management Review*, Vol.7, No.1, pp.107-118, 2002.

[20] J. Sun and X.-F. Hui, "Financial Distress Prediction Based on Similarity Weighted Voting CBR," *Lecture Notes in Artificial Intelligence*, Vol.4093, pp.947-958, 2006.

[21] H. Ahn and K.-j. Kim, "Bankruptcy Prediction Modeling with Hybrid Case-Based Reasoning and Genetic Algorithms Approach," *Applied Soft Computing*, Vol.9, No.2, pp.599-607, 2009.

[22] F. E. H. Tay and L. J. Cao, "Application of support vector machines in financial time series forecasting," *Omega*, Vol.29, No.4, pp.309-317, 2001.

[23] Bichindaritz and C. Marling, "Case-based reasoning in the health sciences: What's next?," *Artificial Intelligence in Medicine*, Vol.36, No.2, pp.127-135, 2006.

[24] S.-W. Kim and H. Ahn, "Development of an Intelligent Trading System Using Support Vector Machines and Genetic Algorithms," *Journal of Intelligence and Information Systems*, Vol.16, No.1, pp.71-92, 2010.

안 현 철

e-mail : hcahn@kookmin.ac.kr
1999년 KAIST 산업경영학과(학사)
2002년 KAIST 경영공학전공(석사)
2006년 KAIST 경영공학전공(Ph.D.)
2009년~현 재 국민대학교 경영정보학부 조교수
관심분야 : 지능형의사결정지원시스템, 재무정보시스템, CRM