경영정보학연구
제16권 제1호
2006년 3월

유전자 알고리즘을 이용한 사례기반추론 시스템의 최적화: 주식시장에의 응용

김 경 재*, 안 현 철**, 한 인 구**

Optimization of Case-based Reasoning Systems using Genetic Algorithms: Application to Korean Stock Market

Kyoung-jae Kim, Hyunchul Ahn, Ingoo Han

Case-based reasoning (CBR) is a reasoning technique that reuses past cases to find a solution to the new problem. It often shows significant promise for improving effectiveness of complex and unstructured decision making. It has been applied to various problem-solving areas including manufacturing, finance and marketing for the reason. However, the design of appropriate case indexing and retrieval mechanisms to improve the performance of CBR is still a challenging issue. Most of the previous studies on CBR have focused on the similarity function or optimization of case features and their weights. According to some of the prior research, however, finding the optimal k parameter for the k-nearest neighbor (k-NN) is also crucial for improving the performance of the CBR system. In spite of the fact, there have been few attempts to optimize the number of neighbors, especially using artificial intelligence (AI) techniques. In this study, we introduce a genetic algorithm (GA) to optimize the number of neighbors to combine. This study applies the novel approach to Korean stock market. Experimental results show that the GA-optimized k-NN approach outperforms other AI techniques for stock market prediction.

Keywords : Case-based Reasoning, k-nearest Neighbor, Genetic Algorithm, Stock Market Prediction

* 교신저자, 동국대학교 경영정보학과
** 한국과학기술원 테크노경영대학원

Ⅰ. 서 론

주가지수 예측과 같은 시계열 예측은 예측분야의 여러 문제 중 가장 복잡하고 난해한 것으로 알려져 있다. 특히, Fama[1970]와 Malkiel[1981]의 연구에서 효율적인 시장은 예측이 불가능하다는 주장이 제기된 이후 오랫동안 정확한 시계열의 예측은 불가능한 것으로 여겨졌다. 그러나 Lo and Mackinlay[1988], Fuller and Kling[1990], Brock *et al.*[1992] 등의 연구에서는 시계열의 예측도 어느 정도까지는 가능하다는 주장이 제기되었다. 이에 따라 시계열 예측을 위해 많은 연구들이 이루어져 왔으나 정확한 예측은 여전히 어려운 문제로 인식되고 있다. 이러한 시계열 예측의 어려움의 원인 중 하나는 시계열의 움직임을 제대로 예측하지 못하는 기존 선형 모형들의 약점이다. 최근에는 시계열의 비선형성을 극복하고 보다 정확한 예측 모형을 구축하기 위해 인공신경망, 자기회귀 이분산모형(autoregressive conditional heteroskedasticity; ARCH), 그리고 사례기반추론(case-based reasoning; CBR)을 활용하는 연구가 이루어지고 있다. 이 중 인공신경망과 ARCH는 비교적 많은 연구가 이루어져 왔으나 k-nearest neighbor forecasting model(k-NN)과 같은 사례기반추론 부류의 방법은 상대적으로 시계열 예측에 많이 이용되지 않았다([Poon and Taylor, 1992; Donaldson and Kamstra, 1997; Silvapulle and Choi, 1999; Kim and Han, 2000; Kim and Han, 2001; Kuo *et al.*, 2001; Kim and Lee, 2004; Chun and Park, 2005] 참고).

사례기반추론은 의사결정시 수행하는 논리적 의사결정 과정을 모형화한 문제 해결 기법이다. 사례기반추론은 복잡하거나, 비구조화된 의사결정문제에 특히 잘 응용될 수 있는 특징이 있어서 생산, 재무, 마케팅 등 다양한 분야의 의사결정문제에 적용되어 왔다.

그러나 많은 장점에도 불구하고, 효과적인 사례기반추론 시스템을 설계, 구축하기 위해서는 연구자가 해결해야 할 여러 가지 문제들이 존재한다. 특히, 사례기반추론은 적절한 유사도 측정방법이나 사례 색인(indexing) 방법, 유사사례간의 결합방법 등에 대해 널리 인정된 방법론이나 원리를 제공해 주지 못하고 있다. 이러한 점 때문에 사례기반추론의 각종 설계 요소들은 실험자나 사용자가 자신의 경험이나 직관에 의해 결정해야 하는 어려움이 있었다. 이런 어려움은 주가지수 예측과 같이 정교한 모형을 요구하는 문제에서 부정확한 예측성과를 제시하기도 하였다[Kim, 2004]. 따라서 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([Wang and Ishii, 1997; Shin and Han, 1999; Kim and Han, 2001; Chiu *et al.*, 2003] 참고).

그런데, 이훈영과 박기남[1999], Guiu *et al.*[1999], 그리고 Jarmulak *et al.*[2000]은 사례기반추론 기법에서 결합할 유사사례의 개수를 최적화 하는 것 역시 사례기반추론 기법의 성과를 향상시키는데 있어서, 매우 중요한 요소인 것으로 제시하고 있다. 그러나, 결합할 유사 사례의 개수를 최적화 하려는 기존 연구는 이훈영과 박기남[1999]의 연구 이외에는 찾기 어렵다. 특히, 이 문제를 지능적인 기법을 통해 해결하고자 하는 시도는 더욱 찾아보기 어렵다.

이에 본 연구에서는 사례기반추론 시스템의 유사 결합사례 개수의 최적화를 위한 방법으로 전역탐색기법의 하나인 유전자 알고리즘(genetic algorithm; GA)을 도입한 모형을 제시하고자 한다. 본 연구에서 제안하는 연구모형의 유용성을 확인하기 위해, 본 연구는 연구모형을 한국 주가지수 데이터에 적용하여 그 예측력을 살펴보고자 한다.

본 논문은 다음과 같이 구성된다. 우선 Ⅱ장에서는 기존 문헌을 살펴보고, Ⅲ장에서는 최적 결합 유사사례 개수를 최적화 하기 위한 대안으로

유전자 알고리즘을 제시하는 연구모형을 소개한다. Ⅳ장에서는 앞서 제시한 모형의 유용성을 확인하기 위한 실험 데이터 및 설계 내용을 설명하고, Ⅴ장에서는 실험 결과를 종합적으로 정리해 제시하도록 한다. 끝으로 마지막 Ⅵ장에서는 결론과 함께 본 연구의 한계점이 제시된다.

Ⅱ. 문헌 연구

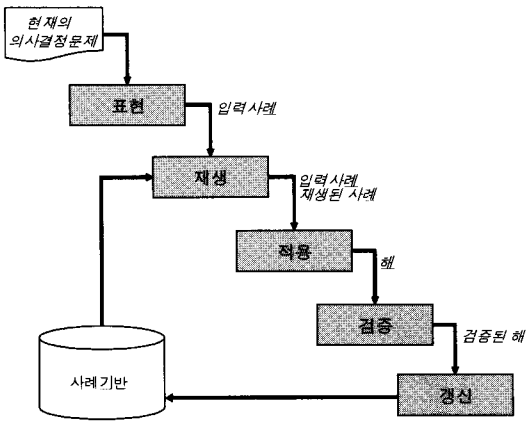
본 연구에서는 사례기반추론과 유전자 알고리즘의 결합 모형을 제시하려고 한다. 이에 본 절에서는 우선 사례기반추론의 기본적인 개념과 원리에 대해 먼저 살펴보고, 이어 사례기반추론과 유전자 알고리즘을 결합하고자 시도한 다른 기존 연구들을 살펴보도록 한다. 끝으로, 사례기반추론 시스템에서 결합할 최적 유사사례 개수를 최적화 하고자 한 기존 연구를 살펴보고, 그 한계점을 살펴본다.

2.1 사례기반추론

사례기반추론은 과거에 축적된 정보만 있으면 어떤 문제든 해결이 가능하므로 복잡하거나 비구조화된 문제를 해결하는데 유리하며, 지식기반을 지속적으로 갱신할 수 있다는 측면에서 상대적으로 우수하다고 할 수 있다[Shin and Han, 1999].

사례기반추론은 아래 <그림 1>에 제시되어 있는 것과 같이 크게 5단계의 절차에 의해 이루어진다[Bradley, 1994].

<그림 1>의 5단계 중에서, 사례기반추론 시스템의 효과를 결정짓는 가장 중요한 단계는 바로 2단계인 재생과 3단계인 적용 단계이다. 이 단계들에서 시스템이 주어진 문제의 해결에 도움이 될 것으로 추정되는 사례들을 선택하게 되는데, '어떤 원리로 유사 사례들을 선별해서, 이들을 어떻게 조합해, 추천 결과를 만들어낼 것인가?' 하는 것에 따라 사례기반추론 시스템의 성능이 크게 변화하기 때문이다. 때문에 사례간 유사도를 어떻게 측정할 것인가, 추천 결과를 도출할 때 유사 사례는 몇 개를 결합할 것인가 하는 등의 문제는 전통적으로 주요한 사례기반추론의 연구주제로 인식되어 왔다[Chiu, 2002].



<그림 1> 일반적인 사례기반추론 프로세스

입력 사례와 기존 축적된 사례간의 유사도를 측정하는 방법에는 다양한 방법이 존재하는데, 그 중에서 가장 널리 사용되고 있는 방법은 바로 각 변수간 차이에 대해 가중합을 도출하여 이를 지표로 활용하는 방법(대표적인 예로 해밍 거리나 유클리드 거리)이다. 그리고 이러한 유사도 기준에 의해 입력 사례와 가장 가까운 기존 사례를 찾아내는 방법으로는 'Nearest-neighbor(NN)' 방법이 가장 널리 활용되고 있다[Jamulak *et al.*, 2000; Chiu, 2002].

Nearest neighbor 방법 중에서도 가장 유사한 과거 사례를 하나만 찾아서, 그것을 기준으로 해를 찾는 방법이 가장 널리 사용되고 있는데, 이러한 방법을 one-nearest neighbor(1-NN)라고 한다. 그러나, 최근에는 유사 사례를 하나가 아닌 다수개 선정하여 투표(voting)나 내삽법(interpolation) 등의 방법론을 이용해 종합한 결과를 최종 해로 제시하는 방법이 더 많이 적용되고 있다. 이러한 방법을 소위 'k-nearest neighbor(k-NN)'

방법이라고 하는데, 여기서의 k는 해를 구하는데 있어 참조할 유사사례의 개수를 의미한다. 이러한 k-NN 방법의 경우, 사례기반추론을 통한 해가 과거의 여러 사례를 보다 폭넓게 반영하여, 보다 일반화된, 그리고 잡음(noise)에 덜 민감한 결과를 도출해 줄 수 있다는 측면에서 큰 장점이 있다. 즉, 일반적으로 k-NN은 1-NN에 비해, 주어진 문제에 대한 보다 정확한 예측 결과를 제시해 줄 수 있다는 것이다. 하지만, k값이 너무 커지게 될 경우에는 반대로, k-NN이 사례기반추론의 성과에 해를 끼칠 수도 있다. k가 필요이상으로 너무 커질 경우, 선택된 유사사례들 중에 잡음이 너무 많이 포함될 가능성이 높기 때문이다. 그렇기 때문에, k-NN에서 최적의 k를 찾아내는 것은 전반적인 사례기반추론 시스템의 성과를 개선하는데 있어서, 매우 중요한 요소라고 할 수 있다.

2.2 GA를 이용한 사례기반추론의 최적화

1-NN이나 k-NN 방법을 이용해 사례기반추론 시스템을 구현할 때, 일반적으로 다음과 같은 2가지 문제가 통상 발생한다. 첫 번째는 '유사도 측정에 사용할 입력 변수를 어떻게 선정할 것인가(feature selection)'하는 것이고, 다른 하나는 '선정된 입력 변수 각각에 대한 가중치를 얼마로 할 것인가(feature weighting)'하는 것이다. 지금까지 이 2가지 문제를 해결하기 위한 다양한 접근이 이루어져 왔지만, 그 중에서도 최근 가장 많이 제시되고 있는 대안이 바로 유전자 알고리즘(genetic algorithm; GA)이다. GA의 작동원리는 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 함수를 최적화 하는 '파라미터' 추정에 널리 적용되고 있다[Shin and Han, 1999].

전통적으로 GA는 사례기반추론 뿐만 아니라, 인공신경망, 귀납적 학습(inductive learning), 회귀모형 등 다양한 인공지능 알고리즘에서 입력변수 선정 혹은 가중치 선정에 널리 적용되어 왔다[Kim, 2004]. Siedlecki and Sklansky[1989]는 사례기반추론의 입력변수 선정을 위해 GA를 사용하였으며, Kim and Han[2001]은 입력변수의 범주화에 GA를 사용하였다. 다음으로, GA를 사례기반추론의 입력변수 가중치 선정에 적용한 연구로는 회사채평가에 적용한 Shin and Han [1999], 공장의 만기일 할당문제에 적용한 Chiu *et al.*[2003], 그리고 보험회사의 고객관계관리에 적용한 Chiu[2002] 등을 들 수 있다. Kim[2004]는 사례기반추론의 입력변수와 그 가중치 선정, 입력자료의 전처리 최적화에 GA를 활용하였다.

2.3 k-NN 결합 유사사례 개수의 최적화

사례기반추론 시스템에서 입력변수의 선정과 이의 가중치를 최적화하기 위한 연구는 많이 이루어져 왔으나, k-NN에서 결합할 유사사례의 수를 최적화하기 위한 연구는 거의 이루어지지 않고 있다.

이훈영과 박기남[1999]은 전통적인 k-NN방법으로서 결합할 유사사례의 개수를 임의로 배정하는 방법, 최적의 범위를 탐색하는 방법, 그리고 유사도 분포에 따른 최적화 수리모형 등 세 가지 방법의 성과를 측정하였다. 이 결과 유사도 분포에 따른 최적화 수리모형이 가장 우수한 성과를 도출하는 것으로 나타났다. 아래 식 (1)은 이 최적화 수리모형의 목적함수를 식 (2)는 제약식을 나타내고 있다.

$$Max.\ SF = \frac{\sum_{b=1}^{n} S_{tb} Z_b}{\left(\sum_{b=1}^{n}\sum_{q=1}^{n} S_{tq} Z_b Z_q\right)^p} \quad (1)$$

$$s.t.\ (S_{tb} - S_{tq}) \times (Z_b - Z_q) \geq 0 \quad \forall\ b \text{ and } q \quad (2)$$

$$Z_b = 0 \text{ or } 1$$

$$0 \leq p \leq 0.5$$

n : 전체 사례기반에 축적되어 있는 과거 사례의 수

S_{tb} : 대상 사례 (입력 사례) t와 기반 사례 (과거 사례) b사이의 유사도

S_{bq} : 기반 사례 (과거 사례) b와 다른 기반 사례 q사이의 유사도

Z_b : 기반 사례 (과거 사례) b가 선택이 되었으면 1, 아니면 0이 되는 이진 변수

이러한 수리 모형은 k-NN의 k를 통계학과 선형계획법과 같은 공학적인 접근을 통해 최적화하는 방법을 제시하는 선구적인 연구라는 측면에서 그 의의가 크다고 할 수 있다. 하지만, 다음과 같은 한계점을 갖고 있다.

첫째로, 식 (1)과 식 (2)에서 볼 수 있듯이, 현재 모형은 최적의 k(결합사례수)를 찾아내는 방법론이라기 보다는, 최적의 결합사례를 찾아내는 방법론이라 할 수 있다. 그런데, 특정 입력 사례 t에 대한 최적 결합사례를 찾아내기 위해서는 위 식 (1)에 제시된 목적함수를 최대화하는 Z_b의 조합을 정수계획법(integer programming) 방법을 통해 찾아내야 하는데, n이 큰 값을 가질 경우 정수계획법 최적해(Z_b^*)를 찾아내는데 요구되는 연산량이 매우 커지게 된다. 또한, 식 (1)에 나타난 목적함수 역시 분자의 경우 n번, 그리고 분모의 경우 n^2번 만큼의 사례간 유사도를 계산해야 특정 경우의 목적함수 값을 찾아낼 수 있다. 뿐만 아니라, 이렇듯 정수계획법 및 목적함수 연산에 많은 연산량이 요구되는데, 전체 검증사례에 대한 최적 결합사례를 모두 찾아내려면, 위 언급했던 작업들을 검증사례의 개수만큼 반복 수행해야 한다. 따라서, 적게는 수천, 많게는 수만 건 이상의 레코드수를 갖고 있는 실제 데이터를 이 모형에 적용할 경우, 너무 과도한 연산량으로 인해 원하는 결과를 얻기가 어려울 가능성이 높다.

둘째로, 이 모형에서 최적 결합 사례수는 입력 사례 t에 따라 변하도록 설계되어 있기 때문에, 입력사례가 새로 들어올 때 마다 최적 결합 사례가 정수계획 모형을 통해 매번 새롭게 계산되어야 한다. 즉, 이 모형은 '기억 기반 접근법(memory-based approach)'을 사용하고 있는 것이다. 따라서, 한 번 모형을 구축하면 단순히 적용을 통해 최적해를 도출해 낼 수 있는 '모형 기반 접근법(model-based approach)'에 대해 상대적으로 연산량이 매우 증가하는 약점을 갖고 있다고 할 수 있다.

셋째로, 이 모형에는 파라미터 p라고 하는 여전히 최적화 해야 할 새로운 변수가 존재한다. 연구자들은 논문에서 이 파라미터 p가 결합할 최적 사례의 수를 결정하는데 있어 조정변수의 역할을 한다고 설명하고 있는데, 실질적으로 이 파라미터에 대한 명확한 정의나 값은 제공되고 있지 않다. 또한 이 연구에서 파라미터 p를 최적화 하기 위한 원리나 방법도 제공되고 있지 않다. 때문에, 최적의 파라미터 p를 찾기 위해서는 몇 가지 후보 파라미터 p에 대해 모두 실험해 보고, 가장 좋은 성과를 보이는 결과를 선택해야 하는데, 이 경우 연산량이 다시 기하급수적으로 늘어날 위험성이 있다.

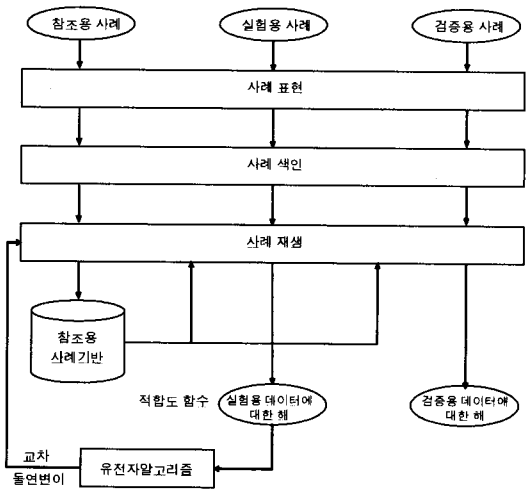
따라서 이상의 3가지 중대한 한계점으로 인해 실제생활의 대용량 데이터에 이훈영과 박기남[1999]의 제안모형을 응용하는 것은 현실적으로 불가능하다.

이 밖에 다른 연구로서 Kim *et al.*[2002]은 한국의 종합주가지수와 16개 산업별 주가지수 예측에 k-nearest neighbor 방법을 이용한 예측을 하였다. 이 연구에서는 유사사례의 수를 찾기 위해 교차검정방법(cross validation method)을 사용하여 학습용 자료에서의 평균제곱오차를 최소화하는 방식을 이용하였다. 그러나 이 연구에서 제안한 방법은 유사사례의 개수를 탐색하는 공간이 제한되어 있으므로 최적화된 유사사례의 개수를 제시할 수 없다는 한계를 가지고 있다.

Ⅲ. k-NN의 파라미터 k에 대한 GA 최적화

앞서 언급한 기존 연구들의 한계점을 극복하

기 위해, 본 연구에서는 k-NN의 k를 최적화 하기 위한 알고리즘으로 GA를 제안하고자 하는데, 이러한 새 모형을 앞으로는 GA k-NN(GA-optimized k-Nearest Neighbor algorithm)으로 약칭하고자 한다. GA k-NN의 전체적인 구조를 나타내면, 다음의 <그림 2>와 같다.



<그림 2> GA k-NN의 구조

GA k-NN은 크게 3단계 절차에 의해 구현이 가능한데, 본 연구의 실험대상인 주가지수 예측 사례에 기초하여 각 단계의 수행절차를 설명하면 다음과 같다.

[1단계] 이 단계에서는 우선 k값이 가질 수 있는 전체 탐색 공간 중에서, 초기값 k를 무작위값으로 설정하게 된다. 모집단(최적의 k를 찾아내기 위한 초기 탐색 위치들의 집합)은 탐색과정에 앞서, 무작위값으로 초기화되고, 탐색의 대상이 되는 k는 GA가 인식할 수 있도록 염색체(chromosome) 형태로 코드화 된다. 본 연구에서 진행한 실험에서 염색체의 경우, 우리가 찾아야 할 검색 공간이 1부터 1,056 사이의 정수이므로, 총 11bit의 이진수로 코드화하였다. 그리하여, 11bit로 코드화된 이진수를 십진수로 변경한 값을 x라 할 때, 이 값은 다음의 식에 의해 1부터 1,056사이의 정수로 재해석되게 된다[Michalewicz, 1996].

$$x' = INT\ \frac{(MAX-MIN)}{(2^{11}-1)}x\ + MIN$$
$$= INT\ \frac{(1056-1)}{(2^{11}-1)}x\ +1 \qquad (3)$$

$INT(k)$: k의 소수점 이하 절삭함수
MAX: 검색해야 할 영역의 최대값
MIN: 검색해야 할 영역의 최소값

예를 들어, $10110001101_{(2)}$라는 숫자가 염색체에 코드화된 경우, 이 값은 십진수로 변환하면 $1421_{(10)}$이 되는데, 이 경우 결합할 유사 사례의 갯수인 k는 위 식 (3)에 대입하여 $INT\left(\frac{1055}{2047}\times 1421\right)+1 = INT(733.37)+1 = 734$로 해석될 수 있다.

이렇게 코드화된 염색체는 특정 적합도 함수(fitness function)를 최대화 하는 방향으로 탐색되게 되는데, 이 경우 적합도 함수를 어떻게 정할 것인지가 중요한 문제가 된다. 본 연구의 목표는 k-NN에서 최적의 k를 찾고자 하는 것인데, 최적의 k를 찾고자 하는 이유는 결국 보다 높은 사례기반추론의 예측 결과를 얻기 위함이라고 할 수 있으므로, 본 연구에서 적합도 함수는 실험용 데이터(test data)에 대한 평균예측정확도로 설정한다. 본 연구에서의 적합도 함수를 수식으로 표현하면, 아래의 식 (4)와 같다.

$$\text{Fitness}= \frac{1}{n}\sum_{i=1}^{n} CR_i\ (i=1,2,\cdots,n)$$
$$if\ PO_i = AO_i,\ CR_i = 1$$
$$otherwise,\quad CR_i = 0 \qquad (4)$$

CR_i: i번째 실험용 사례에 대한 예측 결과 확인 (예측결과가 맞으면 1, 틀리면 0)
PO_i: i번째 실험용 사례의 사례기반추론 예측 결과
AO_i: i번째 실험용 사례의 실제 결과

우리가 실험하고자 하는 주가지수 예측 데이

터의 경우에는 결과에 해당하는 종속변수를 주가지수가 오를 경우 1, 그렇지 않은 경우를 0으로 코드화한 이진변수로 설정하였다. 때문에, 위 식에서 사례기반추론의 예측 결과(PO_i)가 개별 레코드에 대해 오르거나(1) 그렇지 않을 것(0)이라고 예측하였는데, 그 결과가 실제 나타난 결과(AO_i)와 일치할 경우, CR_i에 1을 부여하여, 총 개의 데이터에 대한 평균 예측 정확도을 계산해 이를 GA의 적합도 함수로 사용하였다.

이 단계에서, GA는 초기 설정된 염색체에 대해 교배(crossover), 돌연변이(mutation) 등 다양한 탐색과정을 적용하여, 계속 새로운 k값 후보를 생성해 다시 검증용 데이터에 적용해 보게 되며, 중지 조건(stopping condition)이 만족될 때까지 상기 활동을 계속 반복하게 된다.

[2단계] 1단계 과정을 통해 k값의 후보가 도출되면, 여기서는 그 k값을 토대로 k-NN 사례기반추론을 작동하게 된다. 본 연구에서는 최적의 유사사례 결합 개수인 k 파라미터에만 관심을 갖고 있는 상황이므로, 모든 입력변수의 가중치는 대부분의 사례기반추론 시스템에서 적용하는 '1'을 공통적으로 적용하였다. 이 단계에서는 도출된 k를 기준으로 일반적인 사례기반추론과정을 수행하게 되며, 그 연산의 결과를 제시하게 된다.

[3단계] 앞의 1, 2단계를 반복하면, GA의 중지조건이 만족되는 시점에서 최적 혹은 최적에 근접한 k가 도출되게 된다. 3단계에서는 이렇게 도출된 k를 기반으로, 1, 2단계에서 사용하지 않은 검증용 데이터(hold-out data)에 적용하여, 그 결과를 살펴보게 된다. 본 단계가 필요한 이유는 GA가 실험용 데이터에 대해 평균예측정확도를 최대화 하는 방향으로 k를 최적화 하려고 하는데, 이 경우 최적화된 k가 일반화될 수 있는 값이 아니라서, 실험용 데이터에 대해서는 예측을 잘 하지만 새로운 데이터에 대해서는 예측을 잘 하지 못하는 과도적합문제(overfitting problem)가 발생할 수도 있기 때문이다. 이러한 이유로, 모형 구축에 사용되지 않은 검증용 데이터에 적용해, 이 같은 과도적합문제가 발생한 상황인지 아닌지를 최종적으로 확인해 볼 필요가 있다.

Ⅳ. 실험 설계

4.1 실험 데이터

제안된 연구모형의 유용성을 확인하기 위해 한국종합주가지수의 과거 자료를 이용하여 예측모형을 구축한다. 모형 구축에 사용된 표본은 한국종합주가지수의 1989년부터 1998년까지의 한국종합주가지수의 일별 종가자료이다. 표본의 추출은 임의추출방식에 의해 수행되었으며 총 표본의 크기는 2,218개이다.

기존의 사례기반추론 모형은 일반적으로 모형 구축을 위한 데이터셋과 구축된 모형의 일반화 정도를 측정하기 위한 데이터셋의 두 가지 데이터셋으로 구성되는 것에 반해 본 연구에서 제안하는 모형은 구축과정이 최적화 과정을 통하여 이루어 지게 되므로 최적화에 따른 과도적합문제를 최소화하기 위해 모형구축을 위한 데이터셋을 참조용 데이터셋과 실험용 데이터셋으로 구분한다. 참조용 데이터셋은 일반적인 모형구축을 위해 파라미터 추정을 위해 사용되고, 실험용 데이터셋은 인공신경망 모형에서의 구축과정과 유사하게 최적화 모형에서의 과도적합문제를 통제하기 위한 목적으로 사용된다. 따라서 일반적인 사례기반추론이 두 개의 데이터셋으로 모형을 구축하는 데에 반해 본 연구에서 제안하는 모형은 참조용과 실험용 데이터셋, 그리고 구축된 모형의 일반화 정도를 측정하기 위한 검증용 데이터셋 등 세 개의 데이터셋으로 모형을 구축하게 된다. 이는 최적화를 통한 학습과정에 의해서 사례기반추론 모형을 구축하는 본 연구의 제안모형의 특징에 따른 것이다. 각 데이터셋의 크기와 상대비중은 <표 1>과 같다.

<표 1> 데이터셋 별 데이터 크기와 비율

데이터	데이터 개수	전체 데이터에 대한 비율
참조용	1,056	48%
실험용	581	26%
검증용	581	26%
전 체	2,218	100%

본 연구에서는 모형의 구축을 위해 일반적으로 주가지수 예측에 많이 사용되는 기술적 지표를 중심으로 입력변수를 선정하였다. 본 연구에서 사용된 기술적 지표는 주가지수 예측과 관련된 선행연구에서 사용된 지표 중 한국 주식시장의 특성을 반영하기 위하여 주식시장 예측 전문가인 투자전문기업의 5인의 투자전문가들의 검토를 거친 후 이들의 추천에 의해 선정하였다. 선정된 독립변수의 목록과 관련된 선행연구는 <표 2>에 정리된 바와 같다.

<표 2> 선택된 변수와 관련 선행연구

변 수 명	산 식	관련 연구
Stochastic %K	$\frac{C_t - LL_{t-n}}{HH_{t-n} - LL_{t-n}} \times 100$	[Achelis, 1995]
Stochastic %D	$\frac{\sum_{i=0}^{n-1} \% \ k_{t-i}}{n}$	[Achelis, 1995]
Stochastic Slow %D	$\frac{\sum_{i=0}^{n-1} \% \ D_{t-i}}{n}$	[Gifford, 1995]
Momentum	$C_t - C_{t-4}$	[Chang *et al.*, 1996]
ROC	$\frac{C_t}{C_{t-n}} \times 100$	[Murphy, 1986]
Williams' %R	$\frac{H_n - C_t}{H_n - L_n} \times 100$	[Achelis, 1995]
A/D Oscillator	$\frac{H_n - C_{t-1}}{H_n - L_t}$	[Chang *et al.*, 1996]
Disparity5	$\frac{C_t}{MA_5} \times 100$	[Choi, 1995]
Disparity10	$\frac{C_t}{MA_{10}} \times 100$	[Choi, 1995]
OSCP	$\frac{MA_5 - MA_{10}}{MA_5}$	[Achelis, 1995]
CCI	$\frac{(M_t - SM_t)}{0.015 \times D_t}$	[Achelis, 1995; Chang *et al.*, 1996]
RSI	$100 - \frac{100}{1 + \frac{\sum_{i=0}^{n-1} UP_{t-i}/n}{\sum_{i=1}^{n-1} DW_{t-1}/n}}$	[Achelis, 1995]

주) C_t: t 거래일의 종가, L_t: t 거래일의 저가, H_t: t 거래일의 고가, MA_t: 직전 t 거래일의 이동평균, LL_t: 직전 t 거래일 동안의 최저가, HH_t: 직전 t 거래일 동안의 최고가, UP_t: t 거래일에서의 상향가격변동치, DW_t: t 거래일에서의 하향가격변동치, M_t: $\frac{(H_t + L_t + C_t)}{3}$, SM_t: $\frac{\sum_{i=1}^{n} M_{t-i+1}}{n}$, D_t: $\frac{\sum_{i=1}^{n} |M_{t-i+1} - SM_t|}{n}$

4.2 실험 설계 및 실험용 시스템 개발

GA 탐색을 위한 제어 파라미터들과 관련해서는 모집단을 50개체(organisms)로 설정하였으며, 교배 및 돌연변이 비율에 대해서는 각각 0.7, 0.1로 설정하였다. 아울러 중지 조건으로는 700회 반복, 즉 14세대만큼 k값에 대한 탐색을 반복하도록 설정하였다.

GA k-NN의 성과를 보다 정밀하게 검증하기 위해, 본 연구에서는 1-NN, 전통적 k-NN 모형도 함께 실험하였다. 1-NN은 기존문헌연구에서도 소개했듯이, 유사사례를 단 1개만 검색해서, 그 결과를 토대로 사례기반추론이 예측결과를 생성하는 원리이다. 전통적인 k-NN의 경우, k값을 임의의 고정된 수치로 갖는 k-NN 방법론인데, 여기서는 일반적으로 많이 사용되는 3, 5, 7, 9 등 홀수값들을 k값에 대입해 그 결과를 살펴보았다. 홀수값을 이용하는 이유는 이진분류문제에서 k개의 과거사례의 결과를 다수결 방법에 의해 분류결과를 결정할 때 홀수개의 과거사례를 이용하는 것이 짝수개의 과거사례를 이용할 때 발생하는 동수의 이진분류결과를 배제할 수 있기 때문이다. 또한 k를 9까지만 실험한 이유는 연구자들이 k-NN을 사용할 때, 10 이내의 k를 이용하여 모형을 구축하는 경우가 많기 때문이다. 본 연구는 연구자들이 많이 사용하는 수준에서의 1-NN과 k-NN을, 제안한 연구모형과 비교하고자 한다.

1-NN과 k-NN의 경우, Microsoft Excel 2002의 Excel VBA(Visual Basic for Applications)를 이용해 사례기반추론 알고리즘을 구현해 실험을 하였으며, GA k-NN 시스템은 Microsoft Excel 2002과 Palisade Software사의 Evolver Version 4.06를 결합하여 개발하였다. 즉, k-NN은 Excel의 VBA로 구현하고, k 파라미터를 GA로 최적화 하는 부분은 Evolver가 수행하도록 하였다.

V. 실험 결과

본 장에서는 GA k-NN과 1-NN, 전통적인 k-NN의 예측 성과를 비교하고자 한다. 실험 결과, GA k-NN은 k가 63일 때 가장 최적인 것으로 나타났으며, 전통적인 k-NN의 경우에는 k가 7일 때 가장 최적인 것으로 나타났다. <표 3>에 전체 실험결과가 종합적으로 제시되어 있다.

<표 3>에서 볼 수 있듯이, 검증용 데이터셋에서 GA k-NN은 1-NN 및 전통적인 k-NN에 비해 약 4.65~8.78% 정도 높은 성과를 보이고 있다.

이상의 성과 차이가 통계적으로 유의한 것인지를 파악하기 위해, McNemar 검정을 수행하였다. 이 검정은 명목 데이터에 적용되는 기법으로, 동일한 주제에 대해 전-후 측정을 통한 비교를 할 때 특히 유용한 기법이다[15]. <표 4>는 McNemar 검정 결과를 정리하고 있다.

<표 4>에서 볼 수 있듯이, GA k-NN은 1-NN이나 k가 3, 5, 9인 경우의 k-NN보다 1% 유의수준 하에서 유의한 성과를 보이고 있으며, k가 7인 경우의 k-NN에 대해서는 10% 유의수준 하에서 유의한 성과를 보이고 있음을 알 수 있다.

<표 3> 전체 모형의 예측력 평가 결과

데이터	1-NN	전통적 k-NN				GA k-NN (k=63)
		k=3	k=5	k=7	k=9	
실험용(Test)	-*	-	-	-	-	56.80%
검증용(Holdout)	49.23%	47.16%	48.88%	51.29%	50.43%	55.94%

주) * 1-NN과 전통적 k-NN은 최적화과정이 없으므로 실험용 데이터셋을 이용하지 않음.

<표 4> 검증용 데이터에 대한 McNemar 검정 결과

	1-NN	Conv. k-NN (k=3)	Conv. k-NN (k=5)	Conv. k-NN (k=7)	Conv. k-NN (k=9)
GA k-NN	5.36803***	10.20408***	7.174888***	3.144186*	4.757426***

주) * 10% 유의수준, ** 1% 유의수준

<표 5> 검증용 데이터에 대한 Two sample test for proportions의 p values 결과

	1-NN	Conv. k-NN (k=3)	Conv. k-NN (k=5)	Conv. k-NN (k=7)	Conv. k-NN (k=9)
GA k-NN	0.01088	0.001342	0.00796	0.05614	0.02996

실험결과의 일반화 정도를 확인하기 위하여 McNemar 검정 외에도 비율에 대한 이표본검정(two-sample test for proportions)을 함께 수행하였다. 이 검정 역시 McNemar 검정과 마찬가지로 두 기법간 성과 비교에 널리 사용되는 기법이다[13]. <표 5>는 GA k-NN과 기타 다른 모형 사이의 성과 비교에 대한 검정의 P value가 정리되어 있다.

<표 5>에서 볼 수 있듯이, GA k-NN은 k가 3, 5인 경우의 k-NN보다 1% 유의수준 하에서 통계적으로 유의한 성과 향상을 보였으며, 1-NN과 k가 9인 경우의 k-NN보다 5% 유의수준 하에서 통계적으로 유의한 성과 향상을 보이고 있다. 아울러, k가 7인 경우의 k-NN에 대해서도 10% 유의수준 하에서 통계적으로 유의한 성과의 차이를 보이고 있다.

한편 본 연구에서 제안한 연구모형의 효율성을 평가하기 위해 전통적인 1-NN과 k-NN, 그리고 본 연구에서 제안하는 GA k-NN의 수행시간을 비교하였다. 1-NN을 수행하는 데에는 5분 18초가 소요되었으며, k-NN은 1-NN을 4회 반복 수행하는 과정이므로 20분 32초가 소요되었는데, 이는 1-NN 수행시간의 약 4배 정도의 실험시간이 사용된 것이다. 또한, GA k-NN은 3582분 43초가 소요되었는데, 이는 1-NN을 700회 반복 수행하는 과정으로 볼 수 있으므로 대략 680배 이상의 실험시간이 소요된 것이다. 본 실험은 Pentium IV 2.8GHz의 CPU와 1GB의 메모리를 가진 개인용 컴퓨터에서 Microsoft Excel Version 2003의 VBA 매크로로 작성된 본 연구용 프로그램을 작동시켰을 때, 시작시간과 종료시간을 VBA에 기록하게끔 설계하여 측정한 결과이다. 이 결과를 통해 볼 때, 1-NN보다는 k-NN이, k-NN보다는 GA k-NN이 훨씬 많은 시간을 필요로 함을 알 수 있다. 하지만, GA k-NN의 경우 비록 상대적으로 많은 시간이 소요되지만 한 번 실험을 통해 최적의 k가 발견되면, 구해진 k를 사용하여 추론을 하게 되므로 이후부터는 1-NN과 똑같은 실험시간만이 소요되므로, 보다 정확한 사례기반추론의 결과를 요구하는 영역에서는 충분히 적용할 가치가 있다고 할 수 있다. 또한 본 연구에서 개발한 GA k-NN 프로토타입은 GA k-NN의 매 연산과정, 즉, 700회의 1-NN을 연산하는 과정에서 매번 각 사례와의 거리를 계산하는 과정을 반복하였기에 실험시간이 많이 소요되었으나 이는 Microsoft Excel의 기술적 한계로 인해 발생한 문제이며, C++과 같은 전문 프로그래밍 언어를 이용하여 프로토타입을 구현하면 700회의 연산과정에서 단 한 차례만 각 사례 간의 거리를 계산하여 이를 계속 이용하므로 실제 실험에 소요되는 시간은 현격히 감소할 것으로 기대된다.

Ⅵ. 결 론

본 논문에서는 기존의 사례기반추론 시스템의

성과 향상을 위해 GA와 사례기반추론을 결합한 새로운 시스템을 제안하였다. 본 연구는 GA를 k-NN의 k를 최적화 하는 방법으로 도입하여, 전통적인 k-NN보다 예측 성과를 개선할 수 있는 GA k-NN 모형을 제시하였다. 그리고 이 새로운 모형을 KOSPI 주가지수 데이터에 적용하여, GA k-NN이 1-NN이나 전통적인 개념의 k-NN에 비해 통계적으로 유의한 성과의 향상이 나타남을 살펴보았다. 본 연구의 의의는 그 동안 많은 연구가 이루어지지 못한 사례기반추론의 최적 결합 유사사례의 수를 인공지능 기법을 이용해 최적화하고, 이를 통해 사례기반추론의 성과를 개선시키고자 한 최초의 시도라는 점에 있다고 할 수 있다.

그러나 본 연구도 여러 가지 한계점을 갖고 있다. 우선, 사례기반추론의 성과를 개선하는 데에는 k-NN의 k 이외에도 다른 여러 요인들이 존재하고 있음이 이미 기존 문헌을 통해 제시되었음에도 불구하고, 현재 본 연구에서는 실험설계의 간결성을 위해 이러한 다른 요인들에 대해서는 전혀 고려하고 있지 않다. 때문에 본 연구에서 제시한 k-NN의 k와 함께, 입력변수의 선정이나 가중치 등 여러 가지 요인들을 동시에 최적화하는 연구를 수행한다면, 그것은 상당히 큰 의미가 있을 것으로 전망되며, 향후 좋은 연구주제가 될 수 있을 것으로 예상된다. 또한 연구모형에 대한 프로토타입의 개발에 있어서 보다 효율적인 프로그래밍 언어를 이용한다면 실험시간을 대폭 감소시켜 효율적인 프로토타입을 개발할 수 있을 것이다.

또한 사례기반추론의 다른 요소를 GA로 최적화 하려는 연구 역시 지속적으로 진행될 필요가 있다. GA를 사례기반추론에 적용하는 새로운 예로는 GA를 사례기반 추론의 기반 사례 선택(instance selection) 등에 적용하는 것을 들 수 있다([Kuncheva and Jain, 1999; Rozsypal and Kubat, 2003] 참고). 이처럼 GA가 사례기반추론과 결합될 수 있는 부분이 많고, 더불어 아직 연구가 충분히 이루어지지 않은 분야가 많이 남아 있음을 고려할 때, 앞으로 사례기반추론의 다양한 요소들을 각각 혹은 여러 개를 동시에 최적화하는 연구가 다양하게 이루어질 것으로 전망된다. 끝으로, 본 연구모형은 주가지수 데이터에만 적용, 검증된 상태이므로, 앞으로 본 모형을 더 폭넓은 다양한 분야의 데이터에 적용하여, 모형의 범용성을 추가로 검증 받을 필요가 있다.

〈참 고 문 헌〉

[1] 이훈영, 박기남, "사례기반예측시스템의 정확한 예측을 위한 최적 결합 사례개수 결정 방법에 관한 연구," *경영학연구*, 제27권, 1999, pp. 1239-1252.

[2] Achelis, S.B., *Technical analysis from A to Z*, Probus Publishing, Chicago, 1995.

[3] Bradley, P., "Case-based Reasoning: Business Applications," *Communication of the ACM*, Vol. 37, No. 3, 1994, pp. 40-43.

[4] Brock, W.A., Lakonishok, J., and LeBaron, B., "Simple Technical Trading Rules and the Stochastic Properties of Stock Returns," *The Journal of Finance*, Vol. 47, 1992, pp. 1731-1764.

[5] Chang, J., Jung, Y., Yeon, K., Jun, J., Shin, D., and Kim, H., *Technical Indicators and Analysis Methods*, Jinritamgu Publishing, Seoul, 1996.

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

[7] Chiu, C., Chang, P.C., and Chiu, N.H., "A Case-based Expert Support System for

Due-date Assignment in a Water Fabrication Factory," *Journal of Intelligent Manufacturing*, Vol. 14, 2003, pp. 287-296.

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

[9] Choi, J., *Technical indicators*, Jinritamgu Publishing, Seoul, 1995.

[10] Donaldson, R.G. and Kamstra, M., "An Artificial Neural Network-GARCH Model for International Stock Return Volatility," *Journal of Empirical Finance*, Vol. 4, 1997, pp. 17-46.

[11] Fama, E.F., "Efficient Capital Markets: A Review of Theory and Empirical work," *The Journal of Finance*, Vol. 25, 1970, pp. 383-417.

[12] Fuller, R.J. and Kling, J.L., "Is the Stock Market Predictable?," *The Journal of Portfolio Management*, Vol. 16, 1990, pp. 28-36.

[13] Garrell i Guiu, J.M., E. Golobardes i Ribé, E. Bernadó i Mansilla, and X. Llorà i Fàbrega, "Automatic Diagnosis with Genetic Algorithms and Case-based Reasoning," *Artificial Intelligence in Engineering*, Vol. 13, 1999, pp. 367-372.

[14] Gifford, E., *Investor's Guide to Technical Analysis: Predicting Price Action in the Markets*, Pitman Publishing, London, 1995.

[15] Harnett, D.L. and Soni, A.K., *Statistical Methods for Business and Economics*, Addison-Wesley, Massachusetts, 1991.

[16] Jarmulak, J., Craw, S., and Rowe, R., "Self-optimizing CBR Retrieval," *Proc. of the 12th IEEE International Conference on Tools with Artificial Intelligence*, 2000, pp. 376-383.

[17] Kim, K., "Toward Global Optimization of Case-based Reasoning Systems for Financial Forecasting," *Applied Intelligence*, Vol. 21, 2004, pp. 239-249.

[18] Kim, K. and Han, I., "Genetic algorithms Approach to Feature Discretization in Artificial Neural Networks for the Prediction of Stock Price Index," *Expert Systems with Applications*, Vol. 19, 2000, pp. 125-132.

[19] Kim, K. and Han, I., "Maintaining Case-based Reasoning Systems Using a Genetic Algorithms Approach," *Expert Systems with Applications*, Vol. 21, 2001, pp. 139-145.

[20] Kim, K. and Lee, W.B., "Stock Market Prediction using Artificial Neural Networks with Optimal Feature Transformation," *Neural Computing & Applications*, Vol. 13, 2004, pp. 255-260.

[21] Kim, T.S., Yoon, J.H., and Lee, H.K., "Performance of a Nonparametric Multivariate Nearest Neighbor Model in the Prediction of Stock Index Returns," *Asia Pacific Management Review*, Vol. 7, 2002, pp. 107-118.

[22] Kuncheva, L.I. and Jain, L.C., "Nearest Neighbor Classifier: Simultaneous Editing and Feature Selection," *Pattern Recognition Letters*, Vol. 20, 1999, pp. 1149-1156.

[23] Kuo, R.J., Chen, C.H., and Hwang, Y.C., "An Intelligent Stock Trading Decision Support System through Integration of Genetic Algorithm based Fuzzy Neural Network and Artificial Neural Network," *Fuzzy Sets and Systems*, Vol. 118, 2001, pp. 21-45.

[24] Lo, A. and Malkiel, A., "Stock Market Prices do not Follow Random Walks: Evi-

dence from a Simple Specification Test," *The Review of Financial Studies*, Vol. 1, 1988, pp. 41-66.

[25] Malkiel, B., *A Random Walk Down Wall Street*, Norton, New York, 1981.

[26] Michalewicz, Z.b., *Genetic Algorithms + Data Structures = Evolution Programs*, Springer-Verlag, Berlin, 1996.

[27] Murphy, J.J., *Technical Analysis of the Futures Markets: A Comprehensive guide to Trading Methods and Applications*, Prentice-Hall, New York, 1986.

[28] Poon, S.-H. and Taylor, S.J., "Stock Returns and Volatility: An Empirical Study of the UK Stock Market," *Journal of Banking & Finance*, Vol. 16, 1992, pp. 37-59.

[29] Rozsypal, A. and Kubat, M., "Selecting Representative Examples and Attributes by a Genetic Algorithm," *Intelligent Data Analysis*, Vol. 7, 2003, pp. 291-304.

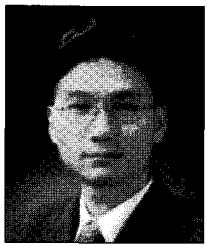
[30] Shin, K.S. and Han, I., "Case-based Reasoning Supported by Genetic Algorithms for Corporate Bond Rating," *Expert Systems with Applications*, Vol. 16, 1999, pp. 85-95.

[31] Siedlecki, W. and Sklanski, J., "A Note on Genetic Algorithms for Large-scale Feature Selection," *Pattern Recognition Letters*, Vol. 10, 1989, pp. 335-347.

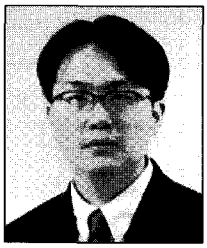
[32] Silvapulle, P. and Choi, J.-S., "Testing for Linear and Nonlinear Granger Causality in the Stock Price-volume Relation: Korean Evidence," *The Quarterly Review of Economics and Finance*, Vol. 39, 1999, pp. 59-76.

[33] Wang, Y. and Ishii, N., "A Method of Similarity Metrics for Structured Representations," *Expert Systems with Applications*, Vol. 12, 1997, pp. 89-100.

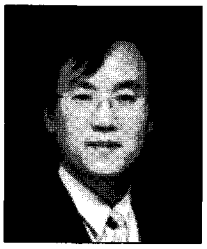
◆ 저자소개 ◆



김경재 (Kim, Kyoung-jae)
현재 동국대학교 경영대학 경영정보학과 교수로 재직 중이다. 중앙대에서 경영학사를, KAIST에서 경영정보시스템을 전공하여 공학석사와 박사를 취득하였다. 주요 관심분야는 데이터마이닝, 고객관계관리, 지식경영, 전사적자원관리 등이며, *Applied Intelligence, Expert Systems, Expert Systems with Applications, Intelligent Data Analysis, Intelligent Systems in Accounting, Finance and Management, Neural Computing & Applications, Neurocomputing* 등에 논문을 발표하였다.



안현철 (Ahn, Hyunchul)
현재 한국과학기술원 테크노경영대학원 박사과정에 재학 중이다. KAIST에서 산업경영학사 및 경영공학석사를 취득하였다. 주요 관심분야는 인공지능 및 데이터마이닝을 이용한 재무예측, 고객관계관리, m-CRM 등이다.



한인구 (Han, Ingoo)
현재 한국과학기술원 테크노경영대학원 교수로 재직 중이다. 서울대학교 국제경제학사, KAIST 경영과학석사를 취득하였고, University of Illinois at Urbana-Champaign에서 회계정보시스템을 전공하여 경영학박사를 취득하였다. 주요 관심분야는 지능형 신용평가시스템, 인공지능을 이용한 재무예측, 지식자산 가치평가, 정보시스템 감사 및 보안 등이다.

이 논문은 2005년 1월 17일 접수하여 1차 수정을 거쳐 2006년 1월 10일 게재확정되었습니다.