휴리스틱 함수를 이용한 feature selection에 관한 연구

홍 석 미[†] · 정 경 숙[†] · 정 태 충[††]

요 약

실생활에서 해결하고자 하는 문제에 대해 수많은 feature들이 수집되어지나 그 feature들을 모두 문제 해결에 활용하는 것은 어렵다. 모든 feature들에 대한 정확한 자료의 수집이 어려우며 관련된 feature들을 모두 학습에 이용할 경우 복잡한 학습 모델이 생성되어지며 좋은 수행 결과도 얻을 수 없다. 또한 수집된 자료들 간에는 상호 관계나 계층적 관계가 존재하는데, 경험적 지식이나 통계적 방법을 이용하여 feature들간의 관계를 분석함으로써 feature의 수를 줄일 수 있다. 휴리스틱 기법은 반복적인 시행 착오와 경험을 통한 학습으로써 미래가 불확실하고 완전한 정보를 갖고 있지 못할 때, 인간의 사고 기능을 통하여 기억이나 경험을 살려, 스스로 해결방안을 모색하면서 점차로 해에 접근해 가는 방법이다. 전문가들은 경험에 의한 의견 수렴 과정을 거쳐 해당 문제 영역에 접근 가능하며, 이러한 특성을 학습에 사용될 feature의 수를 줄이는데 활용할 수 있다. 전문가들은 원시 자료들을 이용하여 새로운 feature들을 생성할 수 있다. 새로이 산출된 feature들과 원시 데이터 내의 feature들을 혼합하여 학습 모델 생성에 이용한다. 본 논문에서는 휴리스틱 함수를 이용하여 학습에 사용될 feature의 수를 줄이고, 추출된 feature들을 신경망의 입력값으로 사용하는 기계 학습 모델을 제시한다. 모델의 성능 평가를 위해 프로야구 경기의 승패 예측 문제를 이용하였다. 실험 결과는 신경 회로망과 휴리스틱 모델을 단독으로 사용했을 때 보다 두 기법을 혼합한 모델이 신경 회로망의 복잡성을 감소시킬 뿐 아니라 분류(classification)의 정확성이 향상되었다.

Research about feature selection that use heuristic function

SeokMi Hong[†] · KyungSook Jung[†] · TaeChoong Chung[††]

ABSTRACT

A large number of features are collected for problem solving in real life, but to utilize all the features collected would be difficult. It is not so easy to collect of correct data about all features. In case it takes advantage of all collected data to learn, complicated learning model is created and good performance result can't get. Also exist inter-relationships or hierarchical relations among the features. We can reduce feature's number analyzing relation among the features using heuristic knowledge or statistical method. Heuristic technique refers to learning through repetitive trial and errors and experience. Experts can approach to relevant problem domain through opinion collection process by experience. These properties can be utilized to reduce the number of feature used in learning. Experts generate a new feature (highly abstract) using raw data. This paper describes machine learning model that reduce the number of features used in learning using heuristic function and use abstracted feature by neural network's input value. We have applied this model to the win/lose prediction in pro-baseball games. The result shows the model mixing two techniques not only reduces the complexity of the neural network model but also significantly improves the classification accuracy than when neural network and heuristic model are used separately.

키워드 : 기계 학습(Machine Learning), 휴리스틱 함수(Heuristic Function), 요소 선택(Feature Selection)

1. 서 론

인공지능은 여러 형태의 지능을 구현하는 시스템을 연구하고 개발하는 컴퓨터 과학의 한 분야로 지식표현, 탐색, 추론, 학습, 인지 그리고 행동들에 대한 연구가 이루어지고 있다. 그 중 학습은 기본적인 인공지능 연구의 한 분야이다. 학습은 사실과 규칙을 반복적인 과정에 의해 지식을 습득하는 일련의 과정으로 이러한 학습 능력을 컴퓨터에게 심어주려는 방법에 관해 많은 연구가 있어 왔다[1, 3, 4].

학습 모델 생성을 위한 첫 번째 단계는 해결할 문제와 관련된 자료를 수집하는 것이다. 학습 모델을 만드는데 사용된 요소(feature)에 따라 그 결과가 크게 좌우 될 수 있다. 두 번째 단계는 수집된 학습 자료를 이용하여 문제에 대한 일반적인 학습 모형을 생성하는 것이다. 학습 모형을

† 준 회 원 : 경희대학교 대학원 전자계산공학과
†† 정 회 원 : 경희대학교 컴퓨터공학과 교수
논문접수 : 2002년 7월 15일, 심사완료 : 2003년 5월 30일

생성하기 위해서는 여러 가지 학습 알고리즘(결정트리, 신경회로망 등)들이 사용되어진다. 그리고 마지막으로 학습된 모델은 새로운 문제에 대한 해를 얻기 위해 사용되어진다.

보다 효과적인 학습 모델을 만들기 위해서는 해결하려는 문제와 관련된 많은 자료들이 필요하며, 실제로 많은 자료들의 수집이 가능하다. 그러나 학습을 위해 모아진 자료들이 모두 문제와 관련된 자료가 아닐 수 있으며, 또한 관련이 있다 해도 수집된 자료들 간에는 중복 요소가 존재할 수 있다. 또한 수집된 자료들을 모두 학습에 활용할 경우 학습 모델이 복잡해지고 수행 시간이 오래 걸리는 등의 여러 가지 문제가 발생한다. 그러므로 방대한 양의 자료들 중에서 어떤 요소를 학습에 활용하기 위해 선택할 것인가가 기계학습에 있어서 중요한 문제이다[5]. 아무리 성능 좋은 알고리즘을 학습 모델 생성에 이용한다고 해도 이러한 요소 선택과 관련된 문제가 해결되어지지 않으면 좋은 성능의 학습 모형을 생성할 수 없기 때문이다.

이에 본 논문에서는 실세계의 큰 데이터 공간 내에서의 효과적인 학습을 위해 최적의 요소 집합(feature set)을 생성하고, 학습 알고리즘에 사용하기 위해 차원을 감소시켜 이를 학습 모델 생성에 이용함으로써 학습 알고리즘의 분류 성능을 향상시키기 위한 모델을 제시하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 연구 배경을 설명하고, 3장에서는 본 논문에서 제시한 휴리스틱 함수를 이용한 feature 선택 방법에 대하여 나타낸다. 그리고 4장에서는 실험 결과를 보이고, 마지막으로 5장에서는 결론 및 향후 연구 방향에 대하여 기술한다.

2. 연구 배경

본 논문에서 학습 모델 생성을 위해 선택한 역전파(back-propagation) 알고리즘은 다층 형태로 입력 데이터에 따른 결과값과 각 테스트 데이터의 실제값간의 차를 줄이기 위해서 노드들간의 연결 강도를 조절하면서 학습하는 교사 학습(supervised learning)에 속한다[6, 7]. 역전파 알고리즘을 이용한 학습 네트워크 생성시 입력 노드의 수가 많을 수록 적당한 분류 성능을 얻기 위해 더 큰 네트워크 사이즈를 요구한다. 만약 역전파 알고리즘이 복잡해지면, 수행 시간과 복잡성이 증가한다. 그러므로 가능한 네트워크의 부하를 줄이면서 학습 능력을 높이기 위해 입력 요소의 수를 줄이는 것과 사용될 요소를 어떻게 선택하느냐가 주요 문제이다.

feature 선택 문제에 있어서 두 가지 방법을 생각할 수 있다. 한 가지는 수학적 기법을 이용하여 학습요소를 분석하는 것이고, 또 다른 방법은 인간의 경험적 지식을 사용하는 것이다. 학습요소들의 분석을 위해 회귀분석 등과 같은 통계적 기법이나 유전자 알고리즘, 클러스터링 기법 그리고 사례기반 학스(case-based learning)과 같은 방법을 이용한다. 이러한 분석을 통해 학습요소의 예측 기여도 또는 요소 상호간의 연관 관계 등을 알아낼 수 있다. 만약 두 개의 학습요소들이 서로 강하게 연결되어져 있다면 둘 중 하나의 요소는 입력 값에서 제외할 수 있다. 그러나 이 방법은 feature의 수를 줄이기 위해서는 좋은 전략이나 분석을 위해 많은 자료들이 사용되며 복잡한 처리 과정이 필요하다. feature 선택을 위해 유전자 알고리즘을 이용하는 경우 추출된 feature set의 성능 평가를 위해 매번 각 집합에 대한 학습 모델을 생성하고 모델의 성능을 평가해야 하므로 많은 시간과 노력이 소비되어진다. 그러므로 학습에 사용될 feature가 많고 feature들간의 관계가 복잡하게 얽힌 분야에서 휴리스틱 기법은 학습요소의 수를 줄이기 위한 대안으로 사용될 있다. 경험적 지식은 전문가들이 많은 예제를 경험한 후 문제에 대한 일반화, 특별화 그리고 분석적 추론을 통해 생성되어질 수 있다. 전문가들은 문제와 관련된 여러 개의 feature들을 조합함으로써 좀 더 추상화된 새로운 feature들을 만들 수 있다. 물론, 조합 함수는 feature들간의 내부 관계나 계층적 관계를 반영할 수 있다. 그러므로 학습에 사용될 feature의 수를 줄이면서도 학습 모델 생성에 필요한 정보를 최대한 많이 포함하는 새로운 feature의 생성이 가능하다.

학습 모델 생성에 활용될 신경망 알고리즘은 처리 노드가 많기 때문에 몇 개의 노드나 연결이 가진 결함이 비교적 시스템 전체에 크게 영향을 주지 않으며(fault tolerance), 새로운 환경에 즉각적으로 프로그램을 갱신하고 유지(adaptability)하는데 용이하다. 잘 정제된 학습요소들을 신경망 학습에 이용할 경우 높은 분류 성능 발휘할 수 있다.

3. Feature 선택을 위한 휴리스틱 모형

3.1 기존 학습 모델에서의 feature 유형

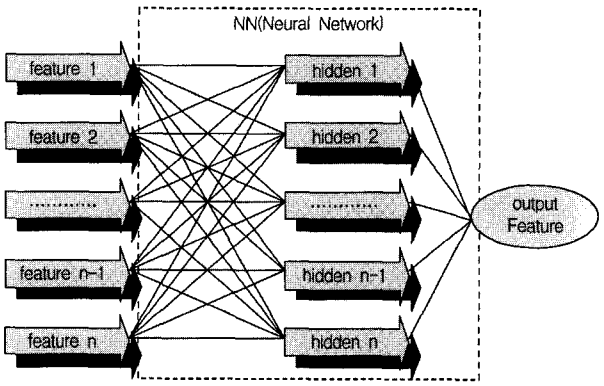
문제와 관련된 자료들의 수집이 쉽거나 관련 학습에 사용될 요소의 수가 적은 경우에는 기존의 학습 알고리즘을 통한 학습 모델 생성이 용이하다. 그러나 많은 자료를 학습에 이용할 경우에는 복잡한 구조의 신경망이나 결정 트리를 요구하며 학습 모델 생성에도 많은 시간이 필요하다. 만약 많은 자료 중 일부만 학습에 활용할 경우에는 해결할 문제에 대한 충분한 자료를 학습 모델에 제공하지 못하므로 좋은 학습 성능을 기대하기 어렵다. 그러므로 충분한 정보를 제공하면서도 사용될 요소의 수를 줄일 수 있다면 적은 비용으로 더 나은 해를 얻을 수 있다.

(그림 1)은 여러 형태의 학습 모형을 보여주고 있다. 신경망과 ID3 알고리즘은 선택된 학습요소들에 대한 학습 자

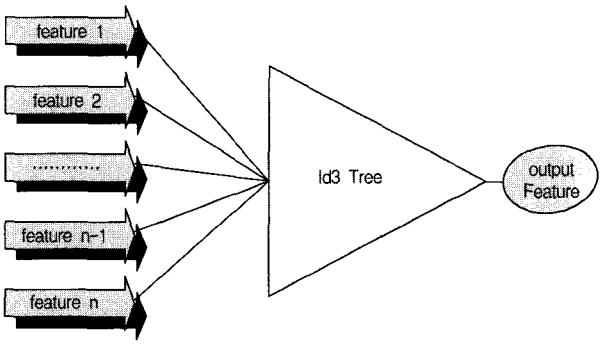
료(training data)를 이용하여 학습 모델을 생성한다. 그러한 알고리즘들은 학습을 위한 입력으로써 많은 요소들을 필요로 한다.

3.2 휴리스틱 함수를 이용한 feature 선택

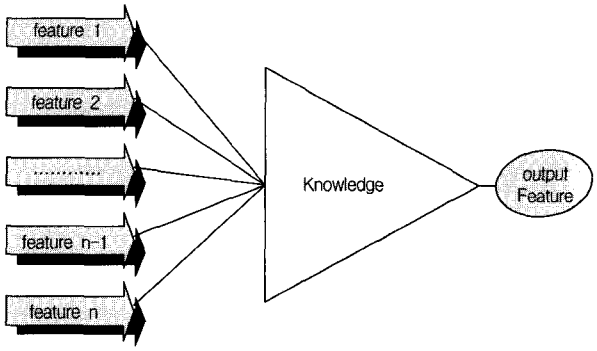
본 논문에서는 학습요소의 수를 줄이기 위한 방법으로 전문가들의 의견을 사용하는 휴리스틱 기법을 선택하였다. 기본적으로 인간은 기억의 한계를 갖고 있으므로, 어떠한 문제에 대한 해를 찾는데 함축된 의미를 갖는 적은 수의 요소들을 사용한다. 그러므로 전문가들이 선택한 요소들을 학습에 이용함으로써 학습요소의 수를 줄일 수 있다.



(a) 신경망 알고리즘



(b) ID3 알고리즘



(c) 전문가 지식(where m <= n)

(그림 1) 여러 가지 기계 학습 모형

그러나 전문가의 의견을 듣는 데는 비용도 많이 들고 쉽게 접하기가 힘들다. 그러므로 전문가들의 지식을 얻기 위해서는 다른 접근법이 필요하다. 만약 스포츠와 같이 해결하고자 하는 문제 영역이 대중적이라면 비록 전문가가 아니더라도 문제와 관련된 지식을 손쉽게 얻을 수 있다. 다수의 사람들은 경기 결과에 대하여 자신들의 생각이나 지식을 가지고 있다. 비록 그들의 지식이 완벽하지 않다고 하더라도 자료가 없는 것보다는 경기 결과를 예측하는데 도움이 된다. 그런 사람들은 요소들간의 관계를 기록하고, 초기 요소로부터 새로운 학습요소를 산출하고 요소에 대한 트리나 그래프를 그릴 수 있다. 특히 학습요소들간의 상대적인 가중치를 알고 있으며 중복된 요소들을 추출할 수도 있다. 몇몇 사람들은 효과적인 지식을 구성할 수도 있다. 그러므로 일반적인 사람들의 지식을 테스트한 후, 선택된 자료들을 일종의 전문가 지식으로써 사용할 수 있다. 본 논문에서는 전문가의 지식을 표현할 수 있는 새로운 휴리스틱 함수를 다음과 같이 정의하였다.

$$F_{old} = \{x_1, x_2, \cdots, x_n\}$$
$$^{\forall}A_i \subset F_{old}, A_1, A_2, \cdots, A_m$$
$$Def)\ f_{new_i} = h_i(A_i)$$
$$F_{new} = \{f_{new_i}\} \cup \{x_j \mid {}^{\forall}x_j \notin A_i\}$$

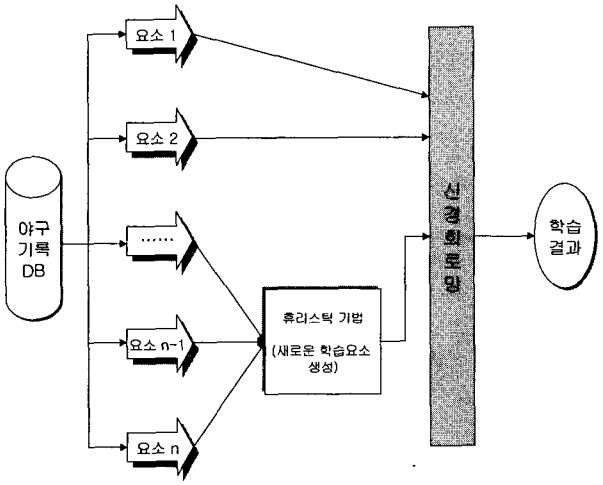
F_{old} : 원시 데이터베이스로부터 추출된 1차 feature set
A_i : 중복 요인을 가진 feature들의 집합
h : heuristic function
f_{new_i} : A_i로부터 생성된 새로운 feature
F_{new} : 새로이 구성된 학습 feature set

F_{old}는 데이터베이스로부터 추출되어지는 학습요소들의 집합으로 예측에 많은 영향을 줄 것으로 예상되는 요소들을 선택한다. A_i는 F_{old}로부터 추출되는 서로 중복 요인을 가지고 있는 요소들의 집합이다. h는 휴리스틱 함수로써, A_i 집합의 요소들을 이용하여 새로운 feature f_{new_i}를 생성한다. 새로이 생성된 feature f_{new_i}와 F_{old}에서 A_i에 속하지 않는 학습요소들로 실제 학습에 활용할 feature set F_{new}가 구성된다.

한편, 전문가들은 지식을 구성하는데 그리 많은 요소들을 사용하지 않으므로 그들로부터 얻어지는 지식들은 불안정하다. 그러므로 전문가들로부터 얻은 함축된 의미를 가진 요소들과 그 외 학습에 필요한 원시 자료들을 혼합하여 학습 자료를 구성하여 학습 알고리즘의 입력으로 활용한다.

(그림 2)는 본 논문에서 제시한 구조로써 학습요소의 수

를 줄이기 위한 휴리스틱 함수와 전문가들로부터 얻은 지식들에 대한 불안정성을 보완할 수 있는 신경망 알고리즘으로 구성되어 있다[8, 9]. 정제된 feature set을 학습에 사용함으로써 좀 더 나은 예측율을 얻을 수 있다.



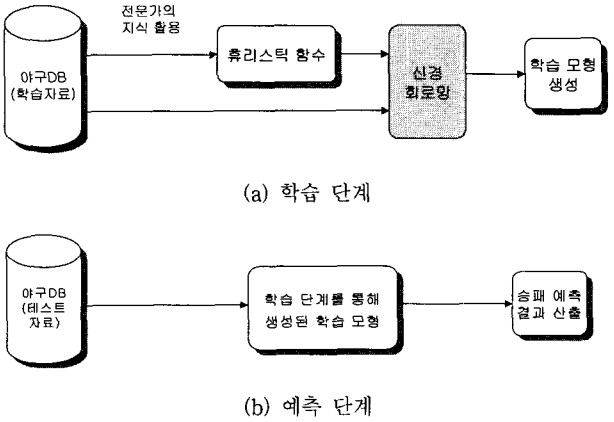
(그림 2) 본 논문에서 제안한 모형

4. 실 험

본 논문에서 제시한 학습 모델의 성능 평가를 위해 프로야구 경기의 승패 예측을 이용하였다. 경기의 승리 팀을 예측하는 것은 쉽지 않다. 왜냐하면 경기에 영향을 미치는 요인들은 아주 많으나, 이러한 것들을 모두 예측에 사용하는 것은 시스템을 복잡하게 할뿐만 아니라 학습요소의 수가 많다고 해도 좋은 예측율을 기대하기 어렵기 때문이다. 게다가 선택된 요소들에 대한 정확한 자료의 수집도 용이하지 않다. 승패 예측에 활용하기 위해 KBO(한국 야구위원회)의 데이터베이스 자료를 사용하였다.

먼저 함축적 의미를 가진 새로운 학습요소 산출을 위해 top-down 접근법으로 상호 관련된 자료들을 조합하였다. 홈팀의 승리 정도는 홈팀과 원정 팀에 대한 예상 점수로 계산되며, 각 팀의 현재 경기의 예상 점수는 경기 기록에서 여러 가지 형태로 보여지는 상대팀에 대한 타율과 투수의 방어율을 이용하여 생성하였다. 이처럼 새로운 요소의 산출을 위해서 데이터베이스 내의 원시 자료들을 사용한다. 문제와 관련된 자료들은 순서화된 그래프를 구성하며, 상위 요소는 원시 자료로부터 산출된 새로운 자료 즉, 홈팀의 승리 정도를 나타낸다. 만들어진 그래프는 휴리스틱 함수로써 (그림 3)에서 삼각형 부분에 위치한 일종의 전문가 지식이 된다.

프로야구 경기 예측은 두 단계로 이루어진다. (그림 3)(a)와 (그림 3)(b)에서 보여준 것처럼 학습 단계와 테스트(예측) 단계. KBO 데이터베이스로부터 1998년도 경기에 대한 자료 1022개를 학습 데이터로 선택했고, 그 자료들은 학습 단계에서 신경망으로 학습되어졌다. 테스트 단계에서는 126개의 테스트 데이터를 사용하였고, 그 결과 본 논문에서 제시한 학습 모형이 84.92%로 경기의 승패를 올바르게 예측했다.



(그림 3) 프로야구 경기의 승패 예측을 위한 모형

예측을 위해 데이터베이스로부터 경기에 영향을 미치는 기록 요소 12개를 추출하고, 이 중 중복 요인을 가지고 있는 9개의 기록을 이용하여 하나의 새로운 학습요소인 '홈팀이 승리할 확률'을 산출하였다. 이렇게 새로이 산출된 feature와 나머지 원시 데이터 2개(현재 경기 이닝, 상대팀에 대한 승률)를 신경망의 입력값으로 이용하였다. 그리고 나머지 한 개, 홈팀의 기록상 승률은 신경망에서 연결 강도 조정을 위해 활용될 목표값으로 활용되었다. 그러므로 신경망과 휴리스틱 함수를 혼합한 모델에 사용된 학습요소는 4개이지만 실제로 예측에 영향을 주는 요소는 12개이다.

제시한 학습 시스템의 성능은 신경망과 휴리스틱 방법 각각을 예측에 이용했을 경우와 비교하였다. 두 방법에 대해서도 혼합형 모델에서와 동일한 데이터를 학습과 테스트에 이용하였다. 신경망은 12개의 초기 feature들을 그대로 학습에 이용하였고, 휴리스틱 함수는 9개의 feature를 사용하였다. <표 1>은 세 가지 방법으로 실험한 결과를 보여준다.

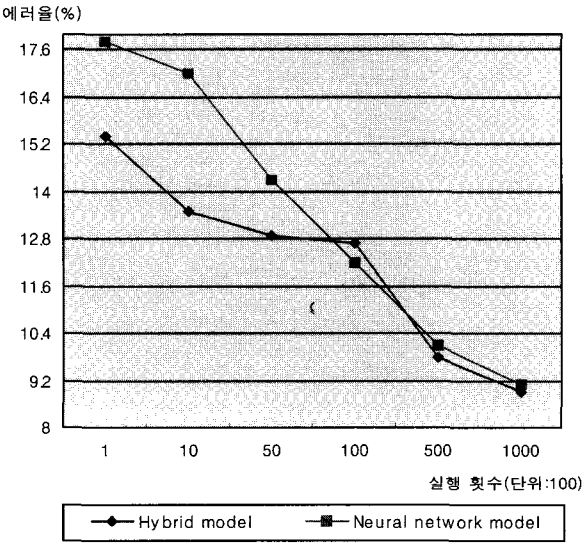
예측 결과를 보면 신경망이 가장 나쁜 예측율을 보이고 있다. 사용된 요소의 수는 많으나 서로 관련된 요소들이 중복되어 학습에 사용되므로 복잡도는 증가하였다. 휴리스틱 기법의 경우는 사용된 요소의 수는 적으나 신경망 보다는 좋은 결과를 보이고 있다. 이는 전문가의 지식이 경우에 따라 기계 학습보다 더 강력할 수 있음을 보이고 있다.

학습 단계에서 신경망과 휴리스틱 함수를 학습요소 추출에 활용한 혼합 모델의 실행 속도를 비교해보자. (그림 4)는

에러율과 안정 상태까지 필요한 시간을 보여준다. 혼합형 모델은 꾸준히 에러율이 감소하고 있음을 볼 수 있다. 이러한 이유는 혼합형 모델이 학습요소를 적게 사용했고 경험적 지식을 구성하는데 시간이 들지 않기 때문이다.

〈표 1〉 세 가지 방법에 대한 사용된 feature와 예측결과 비교

알고리즘	Feature의 수	예측 결과
신 경 망	12	69.84%
휴리스틱 기법	9	77.77%
혼 합 모 형	4	84.92%



(그림 4) 에러율과 학습 속도

5. 결 론

본 논문에서는 학습에 사용할 요소의 수를 줄이고 예측율을 증가시키기 위해서 휴리스틱 기법과 신경망을 이용하는 기계 학습 모델을 소개하였다.

예측이나 분류를 위해 정보의 표현이 풍부하고 입력값과 출력값이 주어지면 원하는 결과를 학습할 수 있는 신경망 알고리즘을 학습 모델 생성에 사용하였다. 문제에 대한 해를 얻기 위해 전문가의 지식을 사용하는 휴리스틱 기법은 그 적용 대상이 특정한 문제에 한정되고 문제에 따른 해법의 구축도 용이하지 않다는 문제점을 지니고 있지만, 이러한 문제점을 극복하기 위해 실제 경기 자료를 바탕으로 휴리스틱 함수를 만들면 자료의 상관 관계와 계층적 구조를 고려하여 입력으로 활용할 수 있는 수식을 이끌어 낼 수 있다. 그러므로 학습에 사용될 요소의 수를 줄이기 위한 방법으로 사용할 수 있다. 이와 같이 각각의 방법에 대한 단점을 보완하여 더 강력한 분류 능력을 갖는 모형을 제안하였다.

모델의 성능을 테스트하기 위한 분야로써는 한국 프로야구 경기 승패 예측을 이용했다. 본 논문에서 제시한 학습 시스템의 성능을 보이기 위해서 동일한 학습 자료와 테스트 자료를 사용하여 신경망과 휴리스틱 기법으로도 수행되어졌다. 신경망은 69.84%의 낮은 예측 율을 보였다. 휴리스틱 기법은 77.77%의 예측 율을 보인다. 혼합된 모델은 휴리스틱 기법과 신경망보다 나은 84.92%의 예측 율을 보인다. 요소 선택에 있어서 휴리스틱 함수를 이용한 방법은 적은 수의 요소를 학습에 사용하고 경험적 지식을 구성하는데 시간을 소비하지 않으므로 효율적이다.

휴리스틱 함수를 이용한 요소 선택 기법이 예측 율이 좋음을 일반화하기 위해서 더 많은 예제를 이용한 실험이 앞으로 계속 연구되어져야 한다.

참 고 문 헌

[1] 서재순, "귀납적 추론을 이용한 프로야구 승패 예측 시스템 개발에 관한 연구", 경희대학교 석사학위논문, 1994.

[2] 홍석미, "프로야구 승패 예측을 위한 게임 시뮬레이터 개발에 관한 연구", 경희대학교 석사학위논문, 1997.

[3] Tom M. Mitchell, "Machine Learning," The McGraw-Hill Companies, Inc., 1997.

[4] Patric Henry Winston "Artificial Intelligence," Addison wesley, 1992.

[5] M. A. Hall, "Correlation-based Feature Selection for Machone Learning," Ph. D. diss. Hamilton, NZ : Waikato University, Department of Computer Science.

[6] M. Riedmiller, "Advanced supervised learning in multi-layer perceptrons-from backpropagation to adaptive learning algorithm," International Journal of computer standards and Interfaces, 16(5), pp.265-278, 1994.

[7] W. S. Sarle, Neural networks and statistical models, In Proc. of 19th Annual SAS Users Group International Conference, SAS Institute, pp.1538-1550, 1994.

[8] Abraham Kandel and Gideon Langholz, "Hybrid Architectures for Intelligent Systems," CRC Press, Inc. 1992.

[9] J. Bala, Huang and H. Vafaie, "Hybrid Learning Using Genetic Algorithms and Decision Trees for Pattern Classification," IJCAI conference, Montreal, pp.19-25, August, 1995.

홍 석 미

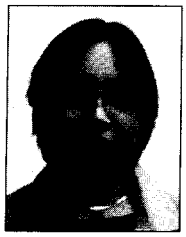
e-mail : smhong@iislab.kyunghee.ac.kr

1994년 상지대학교 전자계산학과(이학사)

1997년 경희대학교 대학원 전자계산공학과 (공학석사)

1998년~현재 경희대학교 대학원 전자계산공학과 박사과정

관심분야 : 기계학습, 데이터마이닝, 에이전트, 정보보호



정 경 숙

e-mail : jungks@iislab.kyunghee.ac.kr

1995년 경희대학교 수학과(이학사)

1997년 경희대학교 대학원 전자계산공학과 (공학석사)

1999년~현재 경희대학교 전자계산공학과 박사 과정

관심분야 : 인공지능, 정보보호, 암호화, 데이터마이닝

정 태 충

e-mail : tcchung@khu.ac.kr

1980년 서울대학교 전자공학과(공학사)

1982년 한국과학기술원 대학원 전자계산 공학과(공학석사)

1987년 한국과학기술원 대학원 전자계산 공학과(공학박사)

1987년~1988년 KIST 시스템 공학센터 선임 연구원

1988년~현재 경희대학교 컴퓨터공학과 정교수

관심분야 : 인공지능, 자연어처리, 로봇에이전트, 최적화, 정보보호 등