건설적 선택학습 신경망을 이용한 앙상블 머신의 구축

(Building an Ensemble Machine by Constructive Selective Learning Neural Networks)

김 석 준 † 장 병 탁 ††

(Suk-Joon Kim) (Byoung-Tak Zhang)

요 약 본 논문에서는 효과적인 앙상블 머신의 구축을 위한 새로운 방안을 제시한다. 효과적인 앙상블의 구축을 위해서는 앙상블 멤버들간의 상관관계가 아주 낮아야 하며 또한 각 앙상블 멤버들은 전체 문제를 어느 정도는 정확하게 학습하면서도 서로들간에 불일치 하는 부분이 존재해야 한다는 것이 여러 논문들에 발표되었다. 본 논문에서는 주어진 문제의 다양한 면을 학습한 다수의 앙상블 후보 네트웍을 생성하기 위하여 건설적 학습 알고리즘과 능동 학습 알고리즘을 결합한 형태의 신경망 학습 알고리즘을 이용한다. 이 신경망의 학습은 최소 은닉 노드에서 최대 은닉노드까지 점진적으로 은닉노드를 늘려나감과 동시에 후보 데이타 집합에서 학습에 사용할 훈련 데이타를 점진적으로 선택해 나가면서 이루어진다. 은닉 노드의 증가시점에서 앙상블의 후보 네트웍이 생성된다. 이러한 한 차례의 학습 진행을 한 chain이라 정의한다. 다수의 chain을 통하여 다양한 형태의 네트웍 크기와 다양한 형태의 데이타 분포를 학습한 후보 네트웍들이 생성된다. 이렇게 생성된 후보 네트웍들은 확률적 비례 선택법에 의해 선택된 후 generalized ensemble method (GEM)에 의해 결합되어 최종적인 앙상블 성능을 보여준다. 제안된 알고리즘은 한 개의 인공 데이타와 한 개의 실세계 데이타에 적용되었다. 실험을 통하여 제안된 알고리즘에 의해 구성된 앙상블의 최대 일반화 성능은 다른 알고리즘에 의한 그것보다 우수함을 알 수 있다.

Abstract This paper presents a new effective method for building an ensemble machine. Several researchers have shown that a good ensemble machine has a low covariance between its members and has highly-correct member networks that disagree as much as possible. In this paper, we use a hybrid neural network algorithm based on the constructive and active learning paradigm to generate diverse ensemble candidate networks. The learning of this network proceeds by increasing the network capacity as well as selecting new training examples incrementally. A new ensemble candidate network is generated just before the network grows its capacity. This single run is defined as a chain. Multiple chains of the network learning can generate diverse candidate networks which are fitted to the various aspects of the given problem in view of the learning capacity and sample distributions. Proportional selection is used to choose the ensemble candidate networks and the selected networks are combined using generalized ensemble method (GEM) proposed by Perrone. Experiments have been performed on an artificial dataset and a real-world problem. The empirical results show that this algorithm can generate highly diverse member networks and can outperform other existing algorithms in generalization accuracy.

1. 서 론

학습자 (learner)의 일반화 성능은 학습 알고리즘에 있어 가장 중요한 성질 중의 하나이다. 유한개의 데이타가 주어졌을 때 모든 consistent learner들은 임의의 문제에 대해 베이지안 결정 경계선 (Bayes decision boundary)을 추정해 냄으로써 유사한 일반화 성능을 보여 줄 수 있다. 그러나 실제 문제에 있어서는 주어진 학습 공간을 표현하기 위한 무한개의 전체 데이타들 중 일부만을 획득하거나 관찰할 수 있을 뿐이다. 이러한 상황에서는 일반화 성능이 유일하지 않으며, 다른 학습 알

† 비 회 원 : 서울대학교 컴퓨터공학부
sjkim@scai.snu.ac.kr
†† 종신회원 : 서울대학교 컴퓨터공학부 교수
btzhang@comp.snu.ac.kr
논문접수 : 2000년 2월 21일
심사완료 : 2000년 10월 19일

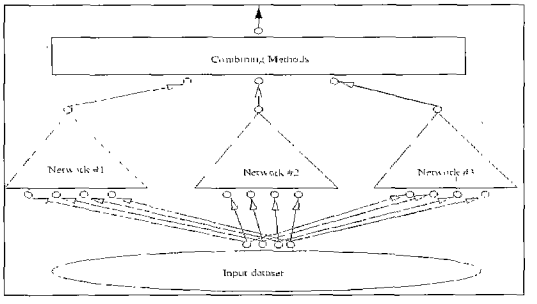

그림 1 3개의 앙상블 멤버를 가진 신경망 앙상블 모델

고리즘은 결정 경계선을 서로 다르게 인식함으로서 다양한 일반화 성능을 보이게 된다. 이렇게 다양한 일반화 성능을 보이는 다수의 학습자들의 출력을 결합함으로서 하나의 학습자에 의존하는 시스템보다 더욱 정확하고 안정적인 일반화 성능을 거둘 수 있음이 많은 연구자들에 의해 밝혀졌다. 특히 독자적으로 훈련된 다수의 신경망을 결합한 시스템은 단일한 신경망에 비해 훨씬 더 뛰어난 일반화 성능과 안정성을 보인다는 사실이 여러 논문들에서 발표되고 있다. 이러한 시스템을 앙상블 머신 (ensemble machine)이라 부른다[1, 2]. 신경망 앙상블의 일반적인 형태는 그림 1과 같다. 그림 1에 표현된 앙상블은 3개의 앙상블 멤버를 가지는 앙상블 머신이다. 앙상블내의 각 네트웍들은 훈련 데이타들을 통하여 훈련되어진다. 이후, 알려지지 않은 새로운 데이타에 대해서 이들 네트웍들에 의해 추정된 예측 값은 앙상블의 최종 예측 값을 만들어내기 위해 결합된다. 이러한 앙상블 머신으로부터 우리가 기대하는 바는 독립적으로 다르게 훈련된 각각의 네트웍들은 에러공간 상에서 서로 다른 지역 국소 (local minima)에 빠지게 되고 이렇게 지역 국소에 빠진 상태에서의 판단을 잘 결합하면 전체적인 성능은 단일 시스템에 비해 훨씬 더 향상되리라는 것이다.

앙상블 머신의 최종적인 일반화 성능을 결정짓는 중요한 요소들은 크게 다음과 같이 3가지로 요약될 수 있다. 첫째, 각 네트웍의 구조 (노드의 수, 연결의 개수 등)를 어떻게 디자인 할 것인가? 둘째, 각 네트웍들의 출력 값으로부터 최적의 결합 성능을 얻기 위해서는 각 네트웍의 출력 값들을 어떻게 결합해야 하는가? 셋째, 최적의 결과를 얻기 위해 제한된 데이타를 어떻게 이용해야 하는가?

이러한 요소들이 앙상블의 일반화 성능에 미치는 영향에 대한 연구 결과들이 최근 수년동안 활발히 발표되고 있다. 본 논문에서는 이러한 이론적/실험적 연구 결과들을 바탕으로 앙상블을 효과적으로 구축하기 위한 새로운 방안을 제안한다. 제안된 방법에서 각 앙상블 멤버들은 자신의 고유한 네트웍 구조에 따라 주어진 문제 공간에서 다양한 확률 공간을 능동적으로 선택하고 이렇게 선택된 공간에 적합화 되도록 훈련되어진다. 이를 위해 우리는 건설적 선택적 신경망 (constructive selective neural network: CSNN) 학습 알고리즘을 사용하는데 이 알고리즘은 [3]에서 제안된 GENIE framework의 self-developmental learning mode를 계승한 것이다. 이 학습 알고리즘을 통하여 주어진 문제 공간의 다양한 면에 적합화 된 앙상블 멤버들이 생성되어진다. 이 학습 알고리즘은 능동 학습 (active learning)과 건설적 학습 (constructive learning)이 혼합된 형태의 학습 알고리즘이다. 초기의 네트웍은 최소 크기의 구조 즉, 최소한의 은닉노드를 가지고 임의의 초기 데이타로부터 학습을 시작한다. 학습이 진행됨에 따라 그 학습자는 자신의 환경으로부터 새로운 데이타를 능동적으로 선택할 수도 있고, 혹은 학습의 정도에 따라 은닉노드의 수를 늘림으로서 자신의 학습 능력을 높일 수도 있다. 네트웍이 자신의 은닉노드를 늘리는 시점에서 앙상블을 위한 새로운 멤버가 생성된다. 네트웍이 정해진 최대 크기의 은닉노드를 가졌을 때 학습은 종료된다. 이러한 일련의 과정을 본 논문에서는 체인 (chain)이라고 명명한다. 여러 체인의 학습이 수행되는 과정에서 생성된 다양한 네트웍들은 후보자 집합(candidate pool) 에 저장된다. 앙상블 멤버들의 선택과 결합은 각각 확률적 선택 기법과 Perron에 의해 제안된 가중치 결합 방식[2]에 의해 수행된다. 제안된 알고리즘은 한 개의 인공 데이타와 1개의 실세계 데이타에 적용되었다. 실험결과로부터 우리는 제안된 알고리즘이 다른 연구자들에 의해 제안된 방식에 비해 더 좋은 일반화 성능을 낸다는 것을 확인할 수 있다.

본 논문의 구성은 다음과 같다. 2 장에서는 앙상블 머신의 이론적/실험적 연구 결과들을 바탕으로 관련 연구에 대해 설명한다. 3 장에서는 효과적으로 다양한 형태의 앙상블 멤버를 생성시키기 위해 제안된 학습 알고리즘을 기술하고 생성된 후보 네트웍들의 선택과 결합 방식에 대해 기술한다. 4 장에서는 1 개의 인공 데이타와 1 개의 실세계 데이타에 대한 실험 결과를 기술한다. 5 장에서는 본 연구 내용을 요약하고 결론을 기술한다.

2. 관련 연구

2.1 앙상블 머신과 bias-variance dilemma

앙상블 머신의 중요한 이론적 배경중의 하나는 bias-

variance dilemma이다 [1]. $F_i(x, D)$를 i번째 네트웍의 실제 출력 값이라 하면, 앙상블 출력은 다음과 같이 표현된다.

$$F(x, D) = \frac{1}{M} \sum_{i=1}^{M} F_i(x, D)$$

위의 식에서 M 은 네트웍의 수를 의미하고 $x \in R^n$은 입력 패턴을, D는 훈련 데이타 집합을 의미한다. 실제 원하는 출력 값을 d 라 하면 결합된 앙상블 시스템의 평균 제곱 에러의 기대값 (expected mean squared error)은 아래와 같이 각 멤버 네트웍에 대한 항으로 표현 가능하다.

$$E_D[(E[d|x] - F(x, D))^2] = (E_D[F(x, D)] - E[d|x])^2 + E_D\left[\frac{1}{M^2}\sum_{i=1}^{M}(F_i(x, D) - E_D[F_i(x, D)])^2\right] + E_D\left[\frac{1}{M^2}\sum_{i=1}^{M}\sum_{j \neq i}(F_i(x, D) - E_D[F_i(x, D)])(F_j(x, D) - E_D[F_j(x, D)])\right]$$

이 식에서 첫 번째 항은 결합된 시스템의 전체 바이어스 (bias)를 의미하고, 두 번째, 세 번째 항은 각각 멤버들의 출력에 대한 분산 (variance)과 공분산 (covariance)에 해당된다. bias와 variance문제는 모든 학습 알고리즘에서 고유한 문제이므로 앙상블 시스템에서 전체 결합 성능에 가장 영향을 미치는 것은 공분산이라고 할 수 있다. 앙상블 머신이 가장 좋은 일반화 성능을 내기 위해서는 앙상블의 각 멤버 네트웍들간에 공분산이 아주 낮아야 함을 이 식을 통해서 알 수 있다.

대부분의 기존 연구들은 위에서 제시된 이론에 근거한 앙상블 멤버의 효과적인 훈련 방식보다 훈련된 앙상블의 효과적인 결합 방식에 많은 초점을 두어왔다[1, 4, 5]. 이러한 연구들은 주로 상이한 학습 파라미터의 조절을 통하여[6] 혹은 다양한 크기의 네트웍의 생성[4], 다양한 초기 웨이트 값의 설정[7], 훈련 데이타의 분할[8]에 의해 임의의 다양한 네트웍을 생성하였다. 이러한 방법들에서는 앙상블을 전체적으로 최적화 할 수 있는 형태의 멤버 네트웍을 직접적으로 생성할 수 있는 방법들에 대해서는 언급하지 않았다.

본 논문에서는 앙상블 멤버들의 구성에 초점을 둔다. 다양하면서도 신뢰성 있는 앙상블 멤버들을 생성시키기 위하여 본 논문에서는 학습자의 구조와 훈련 데이타를 동시에 변화시키면서 각 멤버들이 스스로 원래의 데이타 분포를 추정해 나가도록 한다. 훈련데이타를 점진적으로 선택함으로서 각 앙상블 멤버들은 자신의 학습 능력 내에서 점차적으로 원래의 데이타 분포를 추정해 나아간다. 따라서 학습이 진행됨에 따라 제안된 알고리즘은 전체 문제의 확률 분포를 세부적으로 추정해 나간다. 이러한 알고리즘을 통하여 구성된 앙상블 멤버들은 동일한 문제를 서로 다른 관점에서 파악함으로서 상호간의 상관관계를 최소화함과 동시에 각 멤버들의 의견에 어느 정도 신뢰성이 존재하게 된다.

2.2 능동학습 (active learning)

일반적인 신경망 알고리즘에서는 학습데이타가 외부 환경이나 외부 전능자에 의해 모두 주어진다고 가정한다. 따라서, 신경망의 학습은 신경망의 자유 변수들의 조정에 초점을 맞추게 된다. 반면에 능동 학습에서는 학습자가 자신의 학습 데이타를 스스로 선택하거나 혹은 자신이 학습 데이타에 어떤 영향력을 행사할 수가 있다고 가정한다 [9, 10]. 일반적으로 학습의 문제는 훈련 데이타 $\{ (x_i, y_i)^m \mid x_i \in X, y_i \in Y \}$에 기반하여 $f : X \rightarrow Y$ 의 대응 관계를 찾아내는 것으로 생각될 수 있다. 능동 학습자는 제한된 데이타 집합으로부터 새로운 입력 x^*를 반복적으로 선택하고 결과 값 y^*를 관찰한 후 새로운 데이타 집합 (x^*, y^*)를 자신의 훈련 데이타에 포함시키는 행위를 반복적으로 하도록 허용된다.

능동 학습에 있어 가장 중요한 문제는 새로운 데이타 (x^*, y^*)를 선택하는 방법이다. 직관에 근거하여 (x^*, y^*)를 선택하는 다양한 경험적인 방법들이 존재한다. 현재 훈련 데이타가 존재하지 않는 곳에서 새로운 데이타를 선택하는 방법, 성능이 나쁘거나 신뢰성이 떨어지는 곳에서 데이타를 선택하는 방법, 현재의 모델을 변경시킬 수 있을 만한 데이타를 선택하는 방법, 총괄적인 분산 (integrated variance)이 최소화될 수 있는 데이타를 선택하는 방법 등이 이 범주에 속한다.

Zhang은 현재까지 훈련된 네트웍 상에서 최대의 에러를 내는 데이타를 최대 임계 데이타 (the most critical data) 라 정의하였다 [11]. 즉, 데이타의 임계성은 평균 에러에 비례하고 다음과 같이 현재까지 훈련된 신경망 (W, A)에 의해 계산된다.

$$e_i(s) = \frac{1}{\dim(y_i)} \sum_{\dim(y_i)} (y_i - f(x_i;\ W, A))^2$$

위 식에서 (x_i, y_i)는 i 번째 학습 데이타에 해당되고 f 는 가중치 W와 네트웍 구조 A를 가진 신경망의 출력을, 그리고 s 는 선택 횟수를 의미한다. 최대 에러를 가지는 데이타가 후보자 집합 C_s 에서 선택되어진다.

$$m^* = \text{argmax}_{(x_i, y_i) \in C} (e_i(s))$$

이러한 선택 방식은 학습 과정을 가속화시키는 성질이 있음이 실험적으로 보여진바 있다[11].

2.3 건설적 학습 (constructive learning)

학습이 진행되어 가는 동안 점차적으로 네트웍의 구

조를 변경하고자하는 시도가 여러 연구가들에 의해 이루어졌다. 이러한 접근들 중의 하나는 필요 이상의 많은 은닉노드를 초기에 할당해 두고 적당한 해가 나올 때까지 학습을 시키는 것이다. 이후에 몇몇 은닉 노드나 연결들이 더 이상 사용되지 않거나 중요한 역할을 하지 않을 경우 이들을 삭제하는 방식이다. 이러한 접근법을 가지치기 알고리즘 (pruning algorithm) [12] 이라 한다. 반면에 아주 작은 크기의 네트웍에서 학습을 시작하여 적당한 해를 얻을 때까지 은닉노드를 하나씩 늘여나가는 접근법도 있다. 이러한 접근법을 건설적 알고리즘 (constructive algorithm) [13] 이라 한다. 이러한 시도들의 목표는 최소 크기의 네트웍을 찾고자 하는 것이 아니라 주어진 문제에 적합한 크기의 네트웍을 생성해 내고자하는 데에 있다. 일반적으로 주어진 문제에 대한 최적의 네트웍 크기를 찾는 문제는 NP-hard 문제임이 증명되었다. 만약 어떤 알고리즘이 은닉노드와 연결 가중치를 점진적으로 네트웍에 추가해 나간다면 다항 함수 시간(polynomial time)에 풀려질 수 있는 문제는 역시 다항 함수 시간에 풀 수 있다는 것이 또한 증명되었다[14]. 만약 신경망의 학습 과정에서 효과적으로 은닉노드와 연결 가중치를 추가해 나갈 수 있다면 신경망 알고리즘은 universal learner가 될 수 있음을 이 증명은 보여준다.

다양한 건설적 학습 알고리즘의 크게 다음과 같은 대표적인 알고리즘으로 구분된다. 그림 2 에서 보는 바와 같이 dynamic node creation algorithm (DNC), projection pursuit regression (PPR), cascade correlation (Casco), resource allocation network (RAN), group method of data handling (GMDH) 등으로 크게 구분된다. Dynamic-Node-Creation (DNC) 알고리즘은 Ash에 의해 제안되어졌다. 이 알고리즘에서 시그모이드 형태의 은닉 노드들은 한번에 하나씩 증가하고 이들은 모두 동일한 은닉 층에 추가된다. 전체 네트웍은 각 은닉노드 추가 시 다시 새롭게 훈련된다. Projection - Pursuit - Regression (PPR)은 Friedman에 의해 제안된 통계학적인 방법에 기반 한 알고리즘이다. DNC와 마찬가지로 이 알고리즘에서도 은닉노드는 한번에 하나씩 증가하고 추가된 은닉노드들은 모두 동일한 은닉 층에 존재한다. 그러나 이 알고리즘에서는 단순한 시그모이드 형태의 은닉 노드를 사용하는 것이 아니라 좀 더 복잡한 형태의 은닉 노드를 사용하고 또 새로운 노드 추가 시 전체 네트웍을 다시 훈련시키는 것이 아니라 추가된 노드만을 새로이 훈련시킨다. Cascade-Correlation (Casco) 알고리즘은 Fahlman과 Lebiere에 의

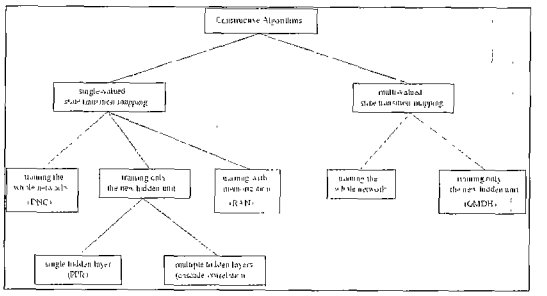

그림 2 건설적 학습 알고리즘의 종류 및 구분
괄호 안에 기입된 것은 해당 분류의 대표적인 알고리즘

해 제안되었다. 이 알고리즘에서는 새로이 추가된 노드가 하나의 새로운 은닉 층을 구성한다. 따라서 전체적인 네트웍은 여러 개의 은닉 층을 가진 신경망 형태가 된다. 이러한 구조는 아주 복잡한 특성을 파악해 내는데 아주 적합한 반면 주어진 데이타에 overfitting되는 경향이 심한 단점이 있다. Resource Allocation Network (RAN) 알고리즘 또한 한번에 한 노드씩 동일한 층에 추가한다. 그러나 훈련 데이타들을 기억함으로서 학습에 필요한 시간을 단축시킨다는 데에 기존의 알고리즘과 차이점이 있다. RAN 알고리즘의 핵심은 학습하기 쉬운 데이타만을 전체 네트웍이 훈련하고 학습이 어렵거나 새로운 데이타는 기억에 의해 이들 데이타를 수용한다는데 있다. Group Method of Data Handling (GMDH)는 Ivakhnenko에 의해 제안되었으며 수많은 알고리즘들이 이를 바탕으로 개발되었다. 이 알고리즘에서는 각 은닉 노드가 고정된 수의 입력 연결을 유지하지만 이들 입력 연결의 소스가 되는 노드들은 고정되지 않는다. 이 입력 노드는 네트웍에 존재하는 다양한 조합의 입력 또는 은닉 노드들이 될 수 있다. 따라서 새로운 노드가 추가될 때 다양한 후보 네트웍들이 존재한다. 이들 후보 네트웍들에 대한 탐색과 선택은 다양한 방법으로 이루어진다.

3. 건설적 선택 학습 신경망에 의한 앙상블의 학습

3.1 건설적 선택 학습 신경망에 의한 앙상블 멤버 생성

건설적 선택적 신경망의 학습은 능동 학습과 DNC 기반의 건설적 학습의 혼합형 학습 방법이다. 이 학습 알고리즘에 의해 앙상블 멤버를 생성하는 알고리즘이 그림 3개 기술되어 있다. 초기에 네트웍은 최소의 은닉 노드를 가지고 임의로 선택된 최소한의 훈련 데이타에 의해 학습을 시작한다. 학습이 진행됨에 따라 네트웍은

훈련 데이타 집합 C 로부터 새로운 데이타를 2.2에서 정의된 critical data로부터 선택하거나 혹은 훈련 데이타 집합 D 에 존재하는 훈련데이타를 학습하기 위해 은닉노드를 추가 할 수 있다. 네트웍 A의 학습은 다음과 같은 일반적인 gradient descent 방식에 의해 훈련 데이타 집합 D의 에러를 최소화시키기 위한 가중치 조절의 과정이다.

$$W_{l+1} \leftarrow W_l + \delta W_l$$

$$\delta W_l \leftarrow -\varepsilon \nabla E_D + \eta \Delta W_{l-1}$$

위의 식에서 E_D 는 현재의 훈련 데이타 집합 D 에 대한 에러의 총합을 나타내고 에러 그래디언트 ∇E_D 는 오류 역전파 알고리즘에 의한 오차 최소화의 방향을 의미한다. ε , η 는 각각 step size와 momentum을 나타낸다. 만약 현재 훈련 데이타 집합 D 에 대한 총 에러가 정의된 오차 수준 $\varepsilon(A)$ 이하로 줄어든다면

현재 훈련 데이타에 대한 훈련 과정은 종료되고 새로운 데이타를 선택하게 된다. 네트웍의 학습 용량은 네트웍 내의 자유 변수의 수와 밀접한 관계가 있으므로[15] 오류 허용 값을 다음과 같이 정의할 수 있다.

이 정의에서 $K(A)$는 현재 네트웍 A 내에 존재하는 모든 연결의 개수를, 그리고 상수 τ는 오류에 대한 민감도를 나타낸다. 새로운 훈련 데이타로는 후보 집합 C 에서 가장 큰 에러를 보이는 데이타 λ 개가 선택되고 이들 데이타들은 현재의 훈련데이타 집합에 추가된다.

$$D \leftarrow D \cup (x^*, y^*) \quad , \quad C \leftarrow C - (x^*, y^*)$$

만약 최대 반복 횟수 $N_{\max}$ 까지 훈련을 수행해도 만족할 만한 성능을 얻지 못한다면 네트웍은 자신의 은닉 노드 개수를 증가시킴으로서 자신의 학습 능력을 향상시킨다. 이때, 이전 크기의 네트웍에서 가장 높은 일반화 성능을 보이는 네트웍이 새로운 앙상블 멤버로 생성되어 앙상블 후보 집합에 저장된다. 이러한 일련의 학습 과정, 즉 가중치 학습, 데이타 선택 그리고 네트웍의 크기 증가의 과정들은 정의된 최대 사이즈 $A_{\max}$ 까지 수행되거나 혹은 더 이상 선택할 후보 데이타가 존재하지 않을 때까지 수행된다. 이러한 학습의 시작에서 종료 시점까지를 본 논문에서는 chain이라 정의한다. 한 chain의 학습이 이루어 질 때마다 앙상블 후보 집합에는 $A_{\max}$개의 네트웍이 생성, 저장된다. 이 과정을 K 번 반복함으로서 $K*A_{\max}$개의 앙상블 후보 집합이 생성된다.

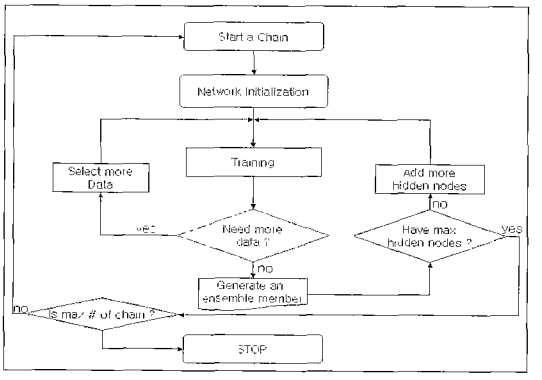

그림 3 능동적 학습 신경망에 의한 앙상블 멤버의 구성

3.2 앙상블 멤버의 선택과 결합

건설적 선택적 신경망의 학습을 여러 chain 수행시킴으로서 다양한 형태의 앙상블 멤버들이 생성된다. 이들을 결합하여 최적의 일반화 성능을 얻기 위하여 몇 개의 적절한 네트웍들이 선택되고 또 이들을 결합하기 위한 최적의 연결 가중치가 계산 되어져야한다.

$$E_l(W;\ A, D) \le \varepsilon(A)$$

$$\varepsilon(A) = \tau^{-1} K(A)$$

3.2.1 앙상블 멤버의 선택

본 논문에서 앙상블 멤버의 선택은 각 네트웍의 평균 제곱 에러에 기반 한 확률적 선택 방법이 사용된다. 에러가 작은 네트웍이 뽑힐 확률을 높게 함과 동시에 에러가 큰 네트웍도 뽑힐 확률을 적절히 부여함으로서 다양한 네트웍이 앙상블의 멤버로 사용될 수 있도록 한다. 각 네트웍의 적합도 G 를 측정하기 위해서 다음과 같이 평균 제곱에러의 지수함수 형태를 정의한다.

$$G(\mathrm{CM}_{i,j}) = \mathrm{Exp}\left\{ -\alpha \frac{\mathrm{MSE}(\mathrm{CM}_{i,j})}{\sum_{i=1}^{m}\sum_{j=1}^{n} \mathrm{MSE}(\mathrm{CM}_{i,j})} \right\}$$

위 식에서 α 는 각 네트웍 사이의 상대적인 적합도 차이를 조절하기 위한 상수이고 $\mathrm{CM}_{i,j}$ 는 i번째 chain에 해당하는 신경망에 의해 생성된 네트웍들 중 j 개의 은닉노드를 가지고 있는 앙상블 멤버를 의미한다. m 과 n 은 각각 신경망 chain의 최대 개수와 한 chain 내에서 각 네트웍이 가질 수 있는 최대 은닉노드의 개수를 의미한다. 적합도 함수인 G 가 높다는 것은 해당 네트웍이 주어진 문제의 전체 분포를 상대적으로 더 잘 추정하고 있음을 의미한다. 확률적 비례 선택 방식은 성능이 낮은 네트웍도 최종적인 앙상블 출력에 참여할 기

회를 효과적으로 부여하기 위한 방식이다.

3.2.2 선택된 멤버의 결합

일반적인 앙상블 멤버의 결합은 앙상블 출력과 실제 값 사이의 오차를 최소화시키기 위한 방향으로 이루어진다. 본 논문에서는 이러한 목적을 달성하기 위하여 Perrone에 의해 제안된 generalized ensemble method (GEM) [1]를 사용한다. GEM은 실험적 에러(empirical MSE)에 기반하여 각 네트워들에 가중치를 두어 결합하는 선형 가중 결합 방식이다. 즉 앙상블 출력 f_{GEM}는 각 네트워의 출력 f_i 와 각 네트워의 가중치 v_i의 곱으로 정의된다.

$$f_{GEM}(x) = \sum_{i=1}^{K} v_i f_i(x)$$

각 네트워의 가중치 v_i는 $\sum_i v_i = 1$의 조건을 만족한다. 가중치 v_i 는 목적 함수인 $y(x)$에 대한 MSE를 최소화시킬 수 있도록 설정되어야 한다. 다음과 같이 멤버 네트워 i 의 에러 $e_i(x)$ 와 i, j 번째 네트워간의 상관관계 행렬 (correlation matrix) $C_{i,j}$ 를 정의한다면

$e_i(x) = y(x) - f_i(x)$, $C_{i,j} = E[e_i(x)\, e_j(x)]$

최소화되어야 하는 앙상블의 출력에 대한 에러는 다음과 같이 정의될 수 있다.

$$\mathrm{MSE}[\overline{f}] = \sum_{i,j} \alpha_i \alpha_j C_{i,j}$$

각 가중치 v_i 는 따라서 다음과 같이 주어진다.

$$v_i = \frac{\sum_{j=1}^{K} C_{i,j}^{-1}}{\sum_{k=1}^{K} \sum_{j=1}^{K} C_{k,j}^{-1}}$$

4. 실 험

제안된 알고리즘은 한 개의 인공 데이타와 한 개의 실세계 데이타에 적용되었다. 인공 데이타는 cloud dataset으로 2개의 입력 벡터와 2개의 클래스를 가지고 있다. 이 데이타는 비 선형적인 특성을 강하게 지니고 있다. cloud dataset은 ESPRIT basic research project (ELENA)에서 사용된 데이타이다. 실세계 데이타는 phoneme recognition dataset 이다. 이 데이타는 비모음 (nasal vowel) 과 구모음 (oral vowel)을 구분해 내는 것이 목적으로 5개의 입력 벡터와 2개의 클래스로 구성되어 있다. 이들 데이타는 ELENA project ftp site[1]에서 받을 수 있다. 제안된 앙상블 구축 방식은 다양한 단일 학습 알고리즘에 의한 성능 및 동일한 네트워 구조에서 초기 값만을 달리하여 앙상블을 구축하는 Maclin의 방법 [7]에 의한 성능과 비교를 하였다. 공정한 성능 비교를 위하여 1 개부터 5 개까지의 은닉 노드를 가지는 다양한 신경망 구조에 대해 Maclin의 방법을 적용하였다.

4.1 cloud dataset

이 데이타를 학습하기 위하여 10개의 chain을 이용하였다. 각 chain에서 가질 수 있는 최대 은닉 노드의 개수는 5개로 한정하였으므로 한 chain 당 5개의 후보 네트워이 생성된다. 따라서 학습이 끝난 후 앙상블 후보 네트워은 총 50개가 생성된다. 생성된 앙상블 후보 네트워들의 통계적 특성은 다음 표 1 과 같다.

표 1 clouds dataset을 학습한 앙상블 후보 네트워들의 통계적 특성

네트워의 총 수	50
상이한 네트워 구조의 종류	5
동일한 구조의 네트워 수	10
평균 일반화 율	0.6742
일반화 율의 표준 편차	0.1454
일반화 율의 skewness	-0.7415
최소 일반화 율	0.2913
최대 일반화 율	0.8660

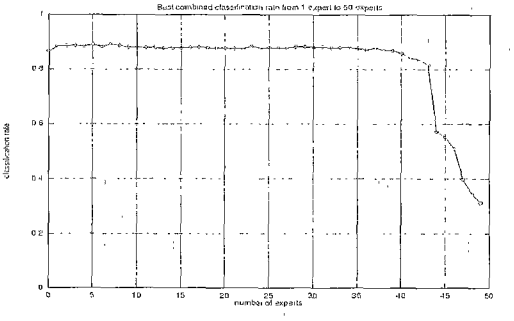

그림 4 앙상블 멤버의 개수에 따른 최대 일반화 성능 비교 (clouds data)

네트워의 선택과 결합은 확률적 비례 선택법에 의해 1개부터 50 개까지 선택하여 이들을 GEM으로 결합하였다. 동일한 개수의 앙상블 멤버에 대해 매번 500회식 선택과 결합을 수행한다. 그림 4 는 앙상블 멤버의 개수 (1 ~ 50) 당 500 회의 선택 결합 중 최대의 성능을 보이는 조합의 일반화 율의 변화 양상을 보여준다. 이 결과로 볼 때 총 50개의 후보 네트워 중 40개 이상의 네트워을 결합하면 앙상블의 일반화 율이 현저하게 떨어

1) *ftp://pegasus.ece.utexas.edu/pub/bench/datasets/ELENA*

지는 것을 알 수 있다. 이 모든 결합들 중 가장 좋은 성능은 7개의 네트웍이 결합하였을 때 나타나며 이때의 일반화 율은 0.8916 이다. 이는 베이지안 일반화 성능인 0.9034에 근사한 값으로 오차는 0.011 정도이다. 반면 최저의 성능은 50 개 후보 네트웍이 모두 결합하였을 때 나타나며 이때의 일반화 율은 0.3131로 최대 성능과 최저 성능간의 차이가 엄청나게 남을 알 수 있다. 이 실험을 통하여 많은 수의 네트웍을 결합하는 것보다는 적절한 개수의 네트웍을 효과적으로 결합하는 것이 앙상블의 일반화 성능 향상에 매우 중요함을 알 수 있다.

표 2 는 앙상블 멤버를 1 개에서 10 개까지 결합하였을 때 나타나는 일반화 율의 통계적 수치이다. 각 멤버 개수마다 동일하게 500번의 조합을 반복하였다. 모든 경우에 있어 500번의 평균값은 최대값에 비해 일반화 성능이 훨씬 떨어지는 현상을 보인다. 표에 의하면 500번 선택 결합 시에 그 중 최대성능과 최소 성능간의 차는 심한 경우 50 % 이상의 성능 차이를 보인다. 따라서 앙상블 멤버의 선택이 결합 성능에 결정적인 역할을 함을 알 수 있다. 표 3 는 Maclin의 방법에 의해 구현된 앙상블의 일반화 성능을 결합된 멤버 개수별로 정리한 것이다. 표 2 의 결과와 비교해 볼 때 이 방법에 의한 평균 일반화 성능은 오히려 더 나아 보인다. 반면 최대 성능은 표 2 에 의한 방법이 조금 더 나은 현상을 보인다. 이는 Maclin의 방법이 제안된 방법에 비해 비교적 다양한 네트웍을 생성해 내지 못하는데 그 이유가 있으리라 짐작된다. 표 3 에서 최대값과 최소값 항목의 ( ) 안의 숫자는 해당 성능을 내는 네트웍의 은닉노드의 수를 나타낸다. 이를 통하여 결합된 일반화 성능은 네트웍의 크기에 비례하는 것이 아님을 알 수 있다. 그림 5은 이 표 2와 표 3에 나타나는 최대 일반화 성능을 결합 네트웍의 개수에 따라 나타낸 것이다. 이 그림을 통하여

표 2 결합된 네트웍의 개수 (1~10)에 따른 일반화 율의 통계적 특성 (clouds dataset)

네트웍 수	평균	표준편차	최소값	최대값
1	0.7935	0.0586	0.5399	0.8650
2	0.7848	0.0834	0.4545	0.8843
3	0.7911	0.0893	0.4536	0.8852
4	0.7815	0.0944	0.4555	0.8871
5	0.7666	0.1072	0.4564	0.8843
6	0.7678	0.1086	0.3893	0.8898
7	0.7656	0.1091	0.3921	0.8825
8	0.7614	0.1093	0.5005	0.8916
9	0.7595	0.1121	0.3563	0.8871
10	0.7614	0.1077	0.4536	0.8806

표 3 Maclin의 방법에 의한 앙상블의 일반화 성능 ( )안의 숫자는 해당 성능을 내는 네트웍의 은닉 노드 수

네트웍 수	평균	표준편차	최소값	최대값
1	0.7941	0.0767	0.6823 (1)	0.8650 (3)
2	0.8107	0.0826	0.6878 (2)	0.8742 (4)
3	0.8202	0.0855	0.6887 (1)	0.8815 (4)
4	0.8176	0.0831	0.8126 (1)	0.8879 (4)
5	0.8152	0.0871	0.6979 (1)	0.8852 (4)
6	0.8151	0.0884	0.6924 (1)	0.8806 (3)
7	0.8002	0.1024	0.6667 (1)	0.8797 (4)
8	0.7842	0.1180	0.6409 (1)	0.8788 (4)
9	0.7484	0.1657	0.5216 (2)	0.8779 (4)
10	0.7295	0.1805	0.4793 (2)	0.8724 (4)

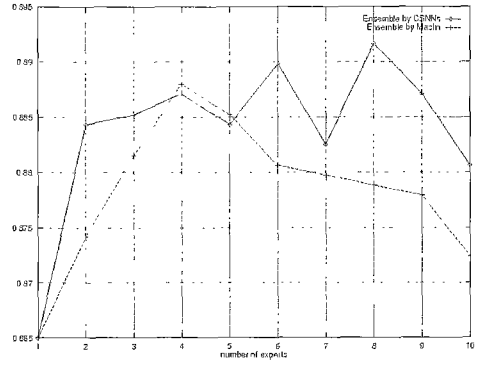

그림 5 CSNN에 의한 Ensemble과 Maclin에 의한 Ensemble의 최대 일반화 율의 비교

각 네트웍의 개수에 대한 최대 일반화 성능 면에서 본 논문에서 제안한 방법 (CSNN)이 기존의 앙상블 알고리즘보다 더 좋은 성능을 보인다는 것을 알 수 있다. clouds 인공 데이타에 대한 실험을 통하여 본 논문에서 제안한 방법이 다른 방법보다 훨씬 더 다양한 형태의 네트웍을 만들어 낼 수 있으며 이들을 잘 결합함으로서 각 네트웍의 단순한 성능 합 이상의 결과를 얻어 낼 수 있음을 알 수 있다.

4.2 phoneme dataset

이 데이타의 학습을 위하여 10 개의 chain을 사용하여 총 50개의 후보 네트웍을 생성하였다. 이렇게 생성된 50 개의 후보 네트웍들의 통계적 특성은 아래 표 4 와 같다.

이 음성데이타에 대한 다양한 알고리즘들의 일반화 성능은 표 5에서 보여진다. 이 성능은 Kegelmeyer에 의해 보고된 수치이다 [16]. 표 4와 5에서 알 수 있듯

표 4 phoneme dataset을 학습한 앙상블 후보 네트웍들의 통계적 특성

네트웍의 총 수	50
상이한 네트웍 구조의 종류	5
동일한 구조의 네트웍 수	10
평균 일반화 율	0.6209
일반화 율의 표준 편차	0.1458
일반화 율의 skewness	-0.9161
최소 일반화 율	0.2948
최대 일반화 율	0.7878

표 5 음성데이타에 대한 다양한 알고리즘의 일반화 율 [Kegelmeyer16]

알고리즘	일반화 율
Neural Networks	0.7921
C4.5 Decision Tree	0.8392
Quadratic Bayes	0.7541
Linear Bayes	0.7300

이 제안된 알고리즘에 의해 생성된 후보 네트웍의 일반화 성능은 표 5에서 제시된 다른 단일 알고리즘의 성능보다 상대적으로 낮다. 특히 50개의 총 후보 네트웍들 중 최대 성능인 0.7878은 동일 알고리즘인 신경망의 보고된 성능 0.7921 에 미치지 못한다. 또한 이들 후보 네트웍은 최대 일반화 성능을 보이는 네트웍과 최저 성능을 보이는 네트웍간의 성능 차이가 거의 50% 정도의 성능차이를 보일 정도로 다양한 네트웍들로 구성되어 있음을 표 4을 통해 알 수 있다.

앙상블의 멤버 선택은 1개에서 50개까지 각 멤버 수에 대해 확률적 비례 선택법을 적용하여 각 멤버수 당 최대 500번의 조합을 선택한다. 현재 구성하고자 하는 앙상블 멤버의 수를 k 라 가정하면 각 멤버 수 k 별로 선택되는 조합의 횟수 $N_{\max}$는 다음 식과 같다.

$$N_{\max} = \min\left\{500 \ , \ \binom{50}{k}\right\}$$

이렇게 선택된 멤버들의 결합은 GEM에 의해 이루어진다. 그림 6에서는 이렇게 $N_{\max}$ 만큼 선택 결합된 앙상블 출력들 중 최대 성능을 보이는 결합들을 멤버 변수의 변화에 따라 보여준다. 이 그림으로부터 역시 인공 데이타 때와 마찬가지로 약 40개 이상의 후보 네트웍 결합 시 결합된 일반화 성능은 급격히 감소함을 알 수 있다. 최대 일반화 율은 0.8417로 17개의 후보 네트웍이 결합되었을 때 나타난다. 반면 최저의 일반화 성능은 0.3012로 50개의 후보 네트웍을 모두 결합하였을 때 나타난다. 하지만 결합된 후보 네트웍의 수가 2개에서 40까지는 최고 일반화 성능에 그다지 큰 변화는 나타나지 않음을 이 표에서 알 수 있다.

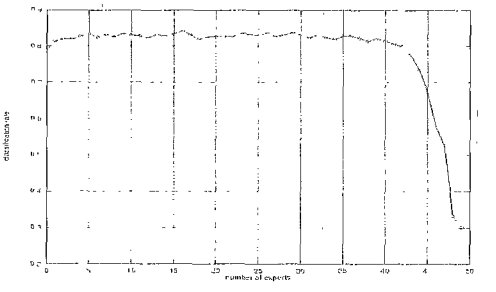

그림 6 결합된 네트웍 수의 변화에 따른 앙상블 최대 일반화 성능 변화 (phoneme dataset)

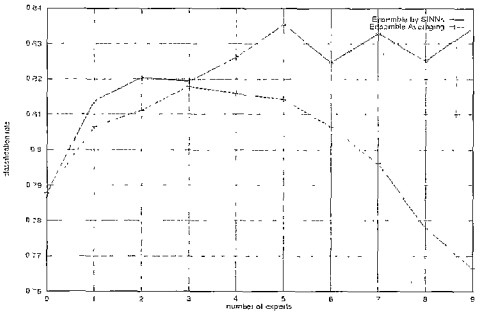

그림 7 CSNN에 의한 앙상블과 Maclin에 의한 앙상블의 최대 일반화 성능 비교 (phoneme dataset)

또한 인공 데이타의 경우와 마찬가지로 Maclin에 의한 방법과 제안된 방법에 의해 얻어진 최대 일반화 성능의 비교 그래프 (앙상블 멤버의 수에 따른 변화)가 그림 7 에서 보인다. 이 그림에서 알 수 있듯이 결합된 최대 일반화 성능 또한 본 논문에서 사용한 방법이 Maclin에 의한 방법보다 모든 경우에 더욱 좋음을 알 수 있다.

5. 결론 및 고찰

본 논문에서는 능동 학습과 건설적 학습을 결합한 신경망 학습 방식을 통하여 효과적인 앙상블 구축을 위한 알고리즘을 제안하였다. 앙상블의 일반화 성능 향상을 위해서는 앙상블 멤버들간의 상관관계가 아주 낮아야 된다는 것이 여러 논문들에서 발표되었다. 따라서 본 알고리즘은 앙상블의 멤버가 될 후보 네트웍들을 서로 다

른 데이타 분포에 대해 서로 다른 학습 능력을 통하여 학습시킴으로서 전체 분포를 다양한 각도에서 파악할 수 있는 다양한 네트웍을 생성시킨다. 이렇게 생성된 다양한 네트웍들을 확률적 비례선택 방식을 통하여 선택하고 이들을 GEM에 의해 결합함으로서 다양한 형태의 앙상블을 구축할 수 있음을 보여주었다. 이렇게 구성된 앙상블은 평균적으로는 낮은 성능을 보이나 최대 성능을 내는 앙상블은 다른 단일 알고리즘과 Maclin에 의해 제안된 앙상블 알고리즘에 비해 훨씬 뛰어난 성능을 보임을 한 개의 인공 데이타와 한 개의 실세계 데이타에 대한 실험을 통하여 보여 주었다.

참 고 문 헌

[1] M.P. Perrone, "Improving regression estimation: Averaging methods for variance reduction with extensions to general convex measure optimization," Ph.D.Thesis, Brown University, Rhode Island, 1993.

[2] M.P. Perrone and L.N. Cooper, "When networks disagree: Ensemble methods for hybrid neural networks," *Artificial Neural Networks for Speech and Vision*, pp. 126-142. 1994.

[3] B.-T. Zhang, "Self-development learning: constructing optimal size neural networks via incremental data selection," Arbeitspapiere der GMD, No 768, German National Research Center for Computer Science (GMD), St. Augustin/Bonn, July 1993.

[4] S. Hashem, "Optimal linear combinations of neural networks," *Neural Networks*, vol. 10, pp.599-614, 1997.

[5] G. Rogova, "Combining the results of several neural network classifiers," *Neural Networks*, vol. 7, no. 5, pp. 777-781, 1994.

[6] E. Alpaydin, "Multiple networks for function learning," *Proceedings of the IEEE International Conference on Neural Networks*, vol. 1, pp. 27-32, 1993.

[7] R. Maclin and J. Shavlik, "Combining the predictions of multiple classifiers: Using competitive learning to initialize neural networks," *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1995.

[8] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.

[9] M. Plutowski and H. White, "Selecting concise training sets from clean data," *IEEE Transactions on Neural Networks*, vol. 4, pp. 305-318, 1993.

[10] B.-T. Zhang, "Convergence and generalization properties of active learning with growing neural nets," *Journal of Korea Information Science Society* (B), vol. 24, no. 12, pp. 1382-1390, 1997.

[11] B.-T. Zhang, "Accelerated learning by active example selection," *International Journal of Neural Systems* 5, no. 1, pp. 67-75, 1994.

[12] R. Reed, "Pruning algorithms - A survey," *IEEE Transactions on Neural Networks*, vol. 4, pp. 740-747, 1993.

[13] T.-Y. Kwok and D.-Y. Yeung, "Constructive algorithms for structure learning in feedforward neural networks for regression problems," *IEEE Transactions on Neural Networks*, vol. 8, no. 3, pp. 630-644, 1997.

[14] E.B. Baum, "A proposal for more powerful learning algorithms," *Neural Computation*, vol. 1, no. 2, pp. 201-207, 1989.

[15] M. Anthony, "Probabilistic analysis of learning in artificial neural networks: the PAC model and its variants," *Neural Computing Survey*, vol.1, pp.1-47, 1997.

[16] W.P. Kegelmeyer Jr. and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 405-410, 1997.

[17] L. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 12, pp. 993-1001, 1990.

[18] X. Yao, and Y. Liu, "A new evolutionary system for evolving artificial neural networks," *IEEE Transactions on Neural Networks*, vol. 8, no.3, pp. 694-713, 1997.

김 석 준

1998년 서울대학교 컴퓨터공학과 학사 졸업. 2000년 서울대학교 컴퓨터공학과 석사 졸업. 현재 한국증권전산 근무. 관심분야는 인공지능, 기계학습, 신경망.

장 병 탁

정보과학회논문지 : 소프트웨어 및 응용 제 27 권 제 3 호 참조