메모리 기반 추론 기법에 기반한 점진적 다분할평균 알고리즘

An Incremental Multi Partition Averaging Algorithm Based on Memory Based Reasoning

Hyeong-il, Yih

이형일*

Abstract

One of the popular methods used for pattern classification is the MBR (Memory-Based Reasoning) algorithm. Since it simply computes distances between a test pattern and training patterns or hyperplanes stored in memory, and then assigns the class of the nearest training pattern, it is notorious for memory usage and can't learn additional information from new data. In order to overcome this problem, we propose an incremental learning algorithm (iMPA). iMPA divides the entire pattern space into fixed number partitions, and generates representatives from each partition. Also, due to the fact that it can not learn additional information from new data, we present iMPA which can learn additional information from new data and not require access to the original data, used to train. Proposed methods have been successfully shown to exhibit comparable performance to k-NN with a lot less number of patterns and better result than EACH system which implements the NGE theory using benchmark data sets from UCI Machine Learning Repository.

요 약

패턴 분류에 많이 사용되는 기법 중의 하나인 메모리 기반 추론 알고리즘은 단순히 메모리에 저장하고 분류 시에 저장된 패턴과 테스트 패턴간의 거리를 계산하여 가장 가까운 학습패턴의 클래스로 분류하는 기법이기 때문에 패턴의 개수가 늘어나면 메모리가 증가하고 또한 추가로 패턴이 발생할 경우 처음부터 다시 수행해야하는 문제점을 가지고 있다. 이러한 문제점을 해결하기 위하여 이미 학습한 대표패턴을 기억하고 새로 들어오는 패턴에 대해서만 학습하는 점진적 학습 방법을 제안한다. 즉 추가로 학습패턴이 발생할 경우 매번 전체 학습 패턴을 다시 학습하는 것이 아니라, 새로 추가된 데이터만을 학습하여 대표패턴을 추출하여 메모리사용을 줄이는 iMPA(incremental Multi Partition Averaging)기법을 제안하였다. 본 논문에서 제안한 기법은 대표적인 메모리기반 추론 기법인 k-NN 기법과 비교하여 현저하게 줄어든 대표패턴으로 유사한 분류 성능을 보여주며, 점진적 특성을 지닌 NGE 이론을 구현한 EACH 시스템과 점진적인 실험에서도 탁월한 분류 성능을 보여준다.

Key words: *Memory-Based Learning(메모리 기반 학습), Distance-Based Learning(거리기반학습), incremental learning(점진적 학습), Information Gain(정보이득)*

I. 서 론

점진적 학습이란 이미 학습한 대표패턴을 기억하고 새로 들어오는 패턴에 대해서만 학습하는 방법이다. 이 방법은 다음과 같은 특징을 가진다. 첫째, 추가로 발생

* : 김포대학 인터넷정보과 부교수
接受日:2007年 10月 24日, 修正完了日: 2008年 2月 19日

된 새로운 학습자료를 학습할 수 있다. 둘째, 추가로 발생된 새로운 학습자료를 학습할 때, 기존에 학습한 학습자료(original data)를 이용하지 않는다. 셋째, 학습하여 생성된 정보(대표패턴)는 점진적으로 갱신되며 삭제되지 않고 유지된다. 이와 같은 특징은 실시간으로 생성되는 자료에 대하여 학습하고 분류할 때 이용될 수 있다.

메모리 기반 학습은 단순히 모든 학습패턴을 메모리에 저장하고 분류 시에 메모리에 저장된 학습패턴들과의 거리를 계산하여 가장 가까운 거리에 있는 학습패턴의 클래스로 테스트 패턴을 분류하는 기법으로 거리기반 학습(Distance Based Learning) 이라고도 한다[1][2]. 메모리 기반 학습 중에서 가장 널리 알려진 기법은 k-NN(k-Nearest Neighbors) 분류기를 들 수 있으며, 이 분류기는 메모리에 저장된 패턴 중 주어진 입력패턴과 가장 가까운 거리에 있는 k개의 학습패턴을 선택하여 그 중 가장 많은 패턴이 소속된 클래스로 입력패턴을 분류한다[2][3][4]. 이러한 k-NN 분류기는 그 성능면에서 만족할 만한 결과를 보이고 있으며, 이미 다양한 분야에 응용되고 있다. 하지만 이 기법은 패턴의 개수가 늘어나면 메모리가 증가하고 또한 추가로 패턴이 발생할 경우 처음부터 다시 수행해야하는 문제점을 가지고 있다[4]. 메모리 사용 등의 성능을 향상과 점진적 특성을 지닌 다양한 연구들이 발표되었으며, 그 대표적인 예로 NGE(Nested Generalized Exemplar) 이론을 들 수 있다[5][6][7].

본 논문에서는 학습패턴을 단순히 메모리에 저장하지 않고 다분할 평균을 이용하여 대표패턴을 생성하여 메모리에 저장한 후 분류시 저장된 대표패턴을 이용하여 테스트 패턴을 분류하며 실시간으로 발생하는 자료를 처리할 수 있는 점진적 다분할평균 기법을 제안하고 구현하였다. 점진적 다분할 평균기법은 추가로 학습패턴이 발생할 경우 매번 전체 학습 패턴을 이용하여 처음부터 학습을 다시 수행하는 것이 아니라, 새로 추가된 데이터만을 학습하여 사용할 수 있는 점진적 학습 기능을 가진 알고리즘이다.

분류기의 성능 및 점진적 학습 능력 검증은 UCI Machine Learning Repository에서 벤치마크 데이터를 발췌한 실험 자료를 사용하였다. 제안한 기법은 대표적인 메모리기반 추론 기법인 k-NN 기법과 비교하여 현저하게 줄어든 대표패턴으로 유사한 분류 성능을 보여주며, 점진적 특성을 지닌 NGE 이론을 구현한 EACH 시스템과 점진적인 실험에서도 탁월한 분류 성능을 보여준다.

II. 관련 연구

1. k-NN 기법

k-NN 분류기는 메모리 기반 학습 기법으로 분류되는 대표적인 알고리즘이다. 이 분류기는 학습단계에서는 단순히 학습 패턴을 메모리에 모두 저장하고, 차후 입력패턴의 분류 단계에서 모든 필요한 계산이 수행되어 이를 Lazy learning Algorithm이라고도 부른다[8]. k-NN 분류기는 먼저 전체 패턴을 단순히 메모리에 저장한다. 그리고 시험할 패턴과 메모리에 저장된 패턴들과의 거리를 식 (1)을 이용하여 계산한 다음 계산한 거리를 기준으로 테스트 패턴과 근접한 k개의 저장된 패턴을 선정한다. 이 선정된 k개 중에서 가장 많은 개수의 학습패턴을 포함하는 클래스로 시험 패턴을 분류하는 알고리즘이다.

$$D_{EQ} = \sqrt{\sum_{i=1}^{n}(E_i - Q_i)^2} \quad (1)$$

이때, E는 메모리에 저장된 학습패턴을 나타내며, Q는 주어진 입력패턴이다. 또한 n은 패턴을 구성하는 특징의 개수이며, E_i, Q_i는 각각 학습패턴과 입력패턴의 i번째 특징 값을 나타낸다. 이 때 k값은 분류기의 성능을 최적화하기 위하여 일반적으로 Cross Validation 기법을 사용하여 결정하며, k=1인 경우를 NN 분류기라 한다[2][3]. 또한 위의 과정 중 4번째 단계에서, 입력패턴과의 거리를 이용하여 가중치를 부여하는 방법을 Weight Vote k-NN이라고 하며, 클래스별로 가중치의 합을 구한 후 합이 가장 큰 클래스로 테스트 패턴을 분류한다[2]. 따라서 성능 면에서 만족할 만한 결과를 보이고 있으며, 이미 다양한 분야에 응용되고 있다.

하지만 실시간 처리를 요하는 분야에 있어서는, 추가로 패턴이 발생할 경우 매번 전체 패턴을 이용하여 처음부터 학습을 다시 수행하는 것이 아니라, 새로 추가된 데이터만을 학습하여 사용할 수 있는 점진적 학습 기능이 필수적이라 할 수 있다. 하지만 k-NN과 같은 학습기법의 경우, 새로 추가되는 학습 패턴이 발생할 경우 분류기 성능의 최적화를 위한 k값이 매번 다시 계산되

어야 하며, 분류에 있어서도 기존의 모든 학습패턴이 거리 계산에 사용되므로 점진적 학습이 불가능하다.

2. EACH 시스템

EACH 시스템은 1990년에 Steven Salzberg가 발표한 NGE (Nested Generalized Exemplar) 이론을 구현한 분류기이다. 이 시스템은 주어진 학습패턴을 메모리공간에 초월평면 (hyperrectangle)의 형태로 저장한다.

즉 모든 학습패턴을 그대로 저장하는 것이 아니라, 학습패턴들을 특정 기준에 의하여 군집화한 후, 각 군집을 하나의 인스턴스로 표현함으로써 k-NN과 같은 분류기에 비하여 상대적으로 높은 메모리 효율을 보장한다[5][7]. 또한 기본적으로 시간적으로 간격을 가지고 학습하는 점진적 학습이 가능하다는 특성을 가진다. 즉 먼저 무작위로 몇 개의 학습패턴을 시드(seed)로 선택하여 예제(Exemplar)로 선정하여 저장한 후 학습패턴에서 가장 가까운 예제를 검색하여 학습패턴의 클래스와 가장 가까운 예제의 클래스가 동일하면, 학습패턴을 이용하여 그 예제를 확장하고 예제의 가중치를 수정한 한 다음, 학습패턴이 공집합이 될 때까지 단계 반복하며, 반면 클래스가 다를 경우는 가중치를 수정하고 두 번째로 가까운 예제를 선택하여 클래스를 비교하여 동일하면 예제를 확장하고 가중치를 수정하며, 다를 경우 학습패턴을 별도의 새로운 예제로 저장하는 기법이다. 이 절차는 학습패턴이 공집합이 될 때까지 단계 반복한다.

EACH 시스템의 학습이 종료되면, 학습패턴들은 예제의 집합으로 표현된다. 예제는 점 또는 초월평면의 형태를 취하게 되며 테스트 패턴은 가장 가까운 예제의 클래스로 분류한다. 예제가 점(point)일 경우에는 점과의 거리를 계산하며, 초월평면일 경우에는 가까운 면과의 거리를 계산한다.

III. 점진적 다분할평균 기법

본 논문에서 제안하는 점진적 다분할평균 (iMPA, incremental Multi Partition Averaging) 기법의 학습은 그림 1과 같은 모델을 갖는다. 이 기법은 전체 학습패턴 공간을 패턴의 분포를 고려하여 가변 크기의 여러 개의 영역으로 반복해서 분할하면서 대표패턴(Representative Pattern)을 생성하는 기법으로, 새로운 학습패턴이 추가적으로 발생되어 학습해야 할 때 기존에 학습했던 모든 학습패턴에 대해 다시 학습하지 않고 추가된 학습패턴만 학습하여 생성된 대표패턴을 기존 대표패턴에 추가된다. 이때 대표패턴은 패턴 평균(Pattern Averaging)법을 이용하여 계산한다.

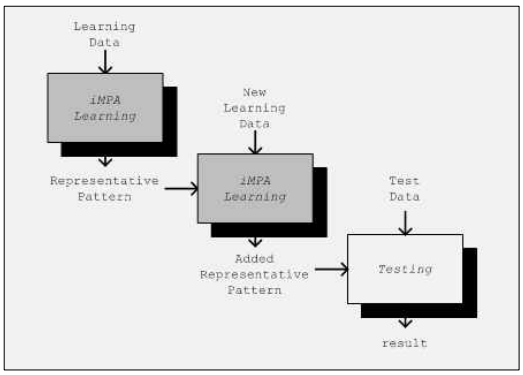

Fig 1. iMPA Learning Model
그림 1. 점진적 다분할(iMPA) 학습모델

다분할 평균기법의 구성은 입력패턴의 정규화 단계와 학습패턴의 특징축 분할점 선정단계, 그리고 특징축 분할점의 선택 및 분할 단계, 점진적 다분할평균기법 구현 단계, 분류단계 등으로 구성된다.

1. 특징의 정규화

인스턴스 기반 추론에서 출력 클래스의 결정은 입력패턴과 메모리에 저장된 학습패턴 사이의 거리를 이용하게 된다. 이 기법에서는 패턴을 구성하는 특징들이 갖는 값의 범위가 판이하게 다를 경우 문제가 발생하게 된다. 예를 들어 (0.9, 400, 0.0004), (0.8, 410, 0.02)와 같은 특징으로 구성된 패턴에서, 두 번째 특징은 다른 두 개의 특징에 비하여 상대적으로 큰 값으로 구성되어있다. 따라서 두 번째 특징이 조금만 차이가 나더라도 나머지 특징간의 차이에 관련 없이 출력 클래스가 결정된다. 이러한 문제점의 해결을 위하여 다음의 식 (2)를 이용하여 특징 값을 정규화 한다. 이 기법은 패턴을 구성하는 모든 특징 값을 0과 1사이의 값으로 정규화 함으로써, 모든 특징의 변화가 패턴의 소속 클래스 결정에 미치는 영향력을 동일하게 한다[10].

$$f_{i_n} = \frac{f_i - f_{i_{\min}}}{f_{i_{\max}} - f_{i_{\min}}} \tag{2}$$

이 때 f_i는 i번째 특징 값, $f_{i_{\max}}$, $f_{i_{\min}}$는 각각 f_i가 가질 수 있는 최대값과 최소값을 나타낸다.

2. 특징축의 분할점 선정

특징축의 분할점 선정은 본 iMPA의 성능과 대표패턴의 개수와 밀접한 관계가 관계가 있는 요소 중에 하나

이다. 패턴공간에 존재하는 패턴의 특징을 파악하여 필요한 구간을 분할하면 무조건 분할하는 방법보다 생성되는 대표패턴의 개수 및 성능이 향상된다.

먼저 각 특징에 존재하는 특징값의 분포를 구한 후 특징값의 오름차순으로 정렬하고, 특징값과 특징값 사이의 값을 식 (3)와 같이 경계값으로 정한다.

$$b_i = \begin{cases} f_{i+1} + \dfrac{f_i}{2}, & f_i < Upperbound \\ Upperbound, Otherwise \end{cases} \quad (3)$$

b_i는 특징의 i번째 경계값이고, f_i, f_{i+1}는 각각 i번째와 i+1번째 특징값이다. $Upperbound$는 특징 상한값으로 정규화된 경우는 1이 된다.

구한 경계값들 중에서 결정트리 알고리즘의 결정노드(Decision Node)에서 특징의 비교 기준을 선정할 때 사용하는 IG(Information Gain) 값을 이용하여 가장 변별력이 좋은 경계값을 분할점으로 선택한다[8]. IG값은 수식 (4), (5)을 이용하여 계산한다.

$$I = -\sum_{i=1}^{C} p_i \log_2 p_i \quad (4)$$

p_i는 학습패턴 집합에서 클래스 i에 소속되는 패턴의 비율이며, C는 클래스의 개수를 의미한다.

$$IG(f) = I - \sum_{i=1}^{N} P_i I_i \quad (5)$$

I는 분할 이전의 정보량이며, P_i는 분할 이전의 학습패턴 중, 분할된 각 영역에 포함된 학습패턴의 비율이다. I_i는 특정 경계값 f를 기준으로 분할했을 때 분할된 각 공간의 정보량을 의미하며, 수식 (4)을 이용하여 계산한다. 이때 I값이 크다는 사실은 올바르게 분류하기 위하여 많은 양의 정보가 필요하다는 것을 의미하며, IG값은 분할 이전의 정보량과 경계값을 기준으로 분할했을 경우 정보량의 차이를 의미한다. 즉, IG값은 분할 이후의 정보량이 작아질 경우에 큰 값을 가지게 되며, 결국 IG값이 큰 경계값을 분할점으로 선택할 때 효율적인 분할이 가능하다.

3. 분할점의 선택 및 분할

특징축 분할점의 선택은 각 특징마다 해당 특징의 경계값들의 신뢰할 만한 큰 IG값을 가진 경우 선택을 한다. 그림 2은 iris 학습자료의 첫 번째 특징에 대해 정규화 과정을 실시한 후 경계값에 대한 IG값을 구한 예이다.

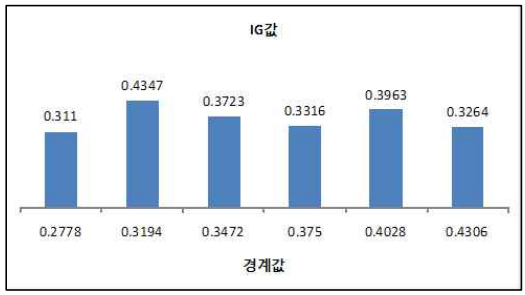

Fig 2. *IG* Values in Border Values
그림 2. 경계값에 따른 IG값

구해진 IG값을 최대로 하는 경계값 들로부터 N개를 선정하여 분할점으로 선정하여 다분할(multi partition)한다. 이 때 선정된 N은 식 (6)에서 특징축의 분할 개수로 n은 하나의 패턴을 구성하는 특징 개수이며, |T|는 전체 학습패턴의 개수이다. 또한 전체 학습패턴의 30%에 근사한 초월평면을 형성하도록 선택하였다[10].

$$N = \lceil \log_n (0.3 \times |T|) \rceil \quad (6)$$

4. 점진적 다분할평균기법

제안한 점진적 다분할 평균기법(iMPA, incremental Multi Partition Averaging)은 다음과 같은 특징을 갖는다. 첫째, 추가로 발생된 새로운 학습자료를 학습할 수 있다. 둘째, 추가로 발생된 새로운 학습자료를 학습할 때, 기존에 학습한 학습자료(original data)를 이용하지 않는다. 셋째, 학습하여 생성된 정보(대표패턴)는 점진적으로 갱신되며 삭제되지 않고 유지된다. 즉 세 가지 특징을 갖는 iMPA 기법은 새로운 학습패턴이 발생할 경우 매번 전체 학습 패턴을 이용하여 처음부터 학습을 다시 수행하는 것이 아니라, 새로 추가된 데이터만을 학습하여 사용할 수 있는 점진적 학습이 가능한 알고리즘이다.

그림 3는 제안한 점진적 다분할 평균기법의 학습절차로 새로운 자료가 발생하였을 때, 기존 학습패턴의 분할영역 내에 존재하는 경우와 그렇지 않은 경우로 나누어 처리된다. 새롭게 추가된 학습패턴이 분할 영역 내인 경우는 이전 학습 수행 시 패턴 평균법을 적용한 대표패턴과 새로운 학습패턴의 특징값을 다시 평균하여 기존의 대표패턴을 식 (7)에 의해 갱신한다.

$$f_{new_i} = \frac{(f_{old_i} \times m) + f_i}{m+1} \quad (7)$$

f_{new_i}는 갱신되는 대표패턴의 i번째 특징값, f_{old_i}는 갱신 이전 대표패턴의 i번째 특징값이며, f_i는 추가로 학습되는 패턴의 i번째 특징값이다. 또한 m은 이전 대표패턴 작성 시 사용된 패턴의 개수를 나타낸다. 그렇지 않은 경우에는 선정된 분할점에 대해 모든 특징에 대해 주어진 학습패턴공간을 분할하고 각 분할된 영역에 포함된 현재의 학습패턴이 속한 클래스를 검사하여 학습패턴의 클래스가 동일한 경우는 대표패턴을 생성하고 종료하며, 서로 다른 클래스에 속하는 패턴들이 혼재되어있는 경우는 다시 다분할을 실시한다. 표 1은 본 논문에서 제안한 iMPA 기법의 알고리즘을 보여준다.

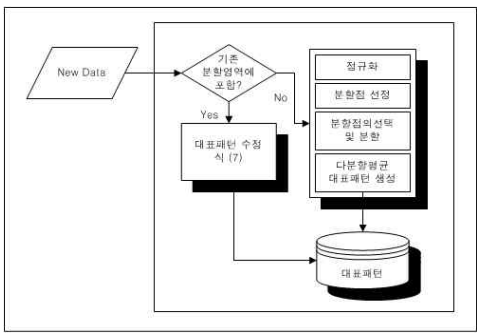

Fig 3.iMPA Learning Process
그림 3. 점진적 학습과정

Table 1. iMPA Algorithm
표 1. iMPA 알고리즘

① 전체 패턴 집합을 식 (2)와 같이 정규화 한다.
② 전체 학습패턴 집합을 포함하는 영역을 식 (6)의 패턴공간을 구성하는 특징축의 분할 개수 N을 결정한다.
③ 새로운 자료가 기존 대표패턴에 속하면 대표패턴을 식 (7)과 같이 패턴의 개수를 수정하여 저장한다. 그렇지 않으면 다음절차를 따른다.
④ 모든 특징축에 대해
(ㄱ) 식 (3)과 같이 경계값을 구한다.
(ㄴ) 구해진 경계값들을 기준으로 분할할 경우의 IG값 (그림 2)을 구한다.
(ㄷ) IG값을 크기 역순으로 N개를 선택하여 각각의 구간값을 선택한다.
⑤ N개의 구간값을 기준으로 패턴공간을 다분할을 실시한다.
⑥ 모든 분할영역에 대해 서로 다른 클래스의 학습패턴이 같은 분할영역에 존재하는지 검사한다.
⑦ 포함된 학습패턴의 클래스가 동일하면 패턴평균법으로 대표패턴을 추출할 때 점진적 학습을 위하여 사용된 패턴의 개수를 저장하고 종료한다.
⑧ 만약 클래스가 다른 학습패턴이 존재하면, 같은 클래스의 학습패턴이 될 때까지 단계 ③~⑥를 실시한다.

대표패턴의 생성은 표 1의 학습알고리즘 단계 ⑦의 패턴평균법은 같은 클래스의 학습패턴들을 평균하여 하나의 대표패턴을 만들어 대체하는 방법으로 각각의 특징값 들에 대해 평균을 한다. 학습이 종료되면 하나의 대표패턴 집합으로 유지관리된다.

5. 패턴의 분류

테스트 패턴을 분류하기 위하여 대표패턴들과 수식 (7)로 거리 계산을 하며, 가장 가까운 대표패턴의 클래스를 출력으로 결정한다. 거리의 계산에는 분류성능 향상을 위하여 학습패턴의 최종적으로 생성된 분할영역에 대응하는 식 (5)의 IG값 구해 입력패턴과 메모리에 저장된 학습패턴간의 거리계산에 있어 특징의 가중치로 사용한다.

$$D_{EQ} = \sqrt{\sum_{i=0}^{n} IG(i)(E_{f_i} - Q_{f_i})^2} \quad (7)$$

IV. 실험 및 분석

본 논문에서 제안한 iMPA 기법의 성능을 Stratified 10-fold Cross-validation 기법을 사용하여 k-NN, EACH, iMPA 등의 알고리즘에 대해 비교 검증하였다.

1. 실험 데이터

본 논문에서는 기계 학습의 벤치마크 자료로 많이 사용되는 UCI Machine learning Database Repository에서 6개의 데이터 셋을 발췌하여 사용하였다[14]. 이들 데이터는 모든 특징이 실수 값을 갖는다. 다음의 표 2는 실험 자료의 분포를 보여주고 있다.

Table 2. Training Patterns in Classes
표 2. 클래스별 학습패턴의 분포

데이터 셋	패턴 개수	특징 개수	클래스 별 패턴 개수					
			1	2	3	4	5	6
Breast-Cancer	699	10	458	241	-	-	-	-
Glass	214	10	70	76	17	13	9	29
Ionosphere	351	34	225	126	-	-	-	-
Iris	150	4	50	50	50	-	-	-
New-Thyroid	215	5	150	35	30	-	-	-
Wine	178	13	59	71	48	-	-	-

Breast-Cancer 데이터 셋은 Wisconsin 대학병원의 William H. Wolberg 박사가 정리한 유방암 진단 자료이며[13], Glass 데이터 셋은 범죄 수사 연구에 사용하기 위해서 유리를 분석한 자료이다. Ionosphere 데이터 셋은 Goose Bay에서 수집된 레이더 데이터이며, Iris 데이터 셋은 패턴인식 분야에서 가장 많이 사용되는 꽃잎과 꽃받침의 길이와 너비 수치를 기반으로 식물의 종류를 판별하는 데이터 셋이다. New-Thyroid 데이터 셋은 갑상선 진단 자료이며, Wine 데이터 셋은 이탈리아의 동일 지역에서 세 가지 다른 품종으로 재배된 와인의 화학적 분석 결과이다.

2. 분류성능

분류 성능 실험에서는 표 3과 같이 학습패턴의 개수를 10%씩 증가시켜 가면서 테스트패턴의 분류성능을 EACH 시스템과 iMPA에 대해 검사하였다. iris 데이터를 예로 들면, 처음에는 전체 학습패턴 135 개중 10%인 13개를 학습한 후 테스트패턴 15개로 분류성능을 측정하고, 그 후에는 학습패턴 개수를 10%씩 증가시켜 가면서 점진적 학습을 수행하였다. 표 3은 전체 학습패턴이 점진적으로 사용되는 상태를 나타낸다. 또한 6개의 데이터 셋에 대한 EACH 시스템 성능은 초기 시드 개수 5, 가중치 증가량 0.2를 사용하여 측정한 결과이다.

Fig 4. Incremental Learning Performance of iMPA, EACH (Breast-Cancer)
그림 4. iMPA, EACH의 점진적 학습성능 (Breast-Cancer)

Table 3. Training Patterns increases of 10%
표 3. 10%씩 추가된 학습패턴개수

데이터 셋	Breast Cancer	Glass	Ionosphere	Iris	New Thyroid	Wine
10%	63	19	32	13	19	16
20%	126	40	64	26	40	32
30%	189	60	96	39	60	48
40%	252	80	128	52	80	64
50%	315	100	160	65	100	80
60%	378	120	192	78	120	96
70%	441	140	224	91	140	112
80%	504	160	256	104	160	128
90%	567	180	288	117	180	144
100%	630	198	324	135	198	162

Fig 5. Incremental Learning Performance of iMPA, EACH (Glass)
그림 5. iMPA, EACH의 점진적 학습성능 (Glass)

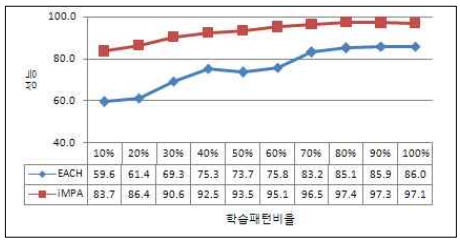

Fig 6. Incremental Learning Performance of iMPA, EACH (ionosphere)
그림 6. iMPA, EACH의 점진적 학습성능 (ionosphere)

Fig 7. Incremental Learning Performance of IMPA, EACH (iris)
그림 7. iMPA, EACH의 점진적 학습성능(iris)

Fig 8. Incremental Learning Performance of IMPA, EACH (new-thyroid)
그림 8. iMPA, EACH의 점진적 학습성능 (new-thyroid)

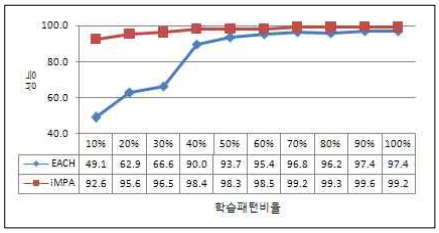

Fig 9. Incremental Learning Performance of IMPA, EACH (wine)
그림 9. iMPA, EACH의 점진적 학습성능(wine)

그림 4에서 그림 9에 걸쳐 각 데이터 셋에 대한 학습패턴의 점진적 증가에 대한 성능변화를 고찰한 것으로, 각 그림에서 x축의 표는 학습패턴비율에 대한 성능값을 나타낸다. 그림 5의 Glass 데이터 셋을 제외하고 iMPA가 EACH 시스템보다 우수한 것으로 나타났다. 또한 EACH시스템은 성능이 50에서 98.4까지 학습 자료의 양에 따라 증가 폭이 크게 측정되었고, iMPA는 66.7에서 99.2까지 그 변화에 비교적 적게 안정된 모습으로 측정되었다. 따라서 성능을 기준으로 신뢰할 수 있는 성능은 학습 자료가 EACH시스템은 70% ~ 80%이상부터, iMPA는 30% ~ 40%부터 신뢰할 수 있는 성능을 측정할 수 있었다.

표 4와 표 5는 그림 4부터 그림 9에 걸쳐 나타낸 점진적 학습의 EACH 시스템과 iMPA의 분류성능 비교에 대한 표준편차를 보여준다.

Table 4. Standard Deviations of EACH Performances
표 4. EACH 분류 성능에 대한 표준편차

데이터 셋	Breast Cancer	Glass	Ionosphere	Iris	New Thyroid	Wine
10%	6.21	7.32	16.13	7.23	6.89	8.48
20%	5.02	7.35	16.88	5.78	7.12	8.32
30%	5.03	7.78	15.27	4.23	6.58	8.19
40%	4.30	7.46	15.92	5.41	6.31	7.98
50%	4.78	8.34	14.56	5.34	5.61	7.23
60%	4.23	7.23	16.71	4.22	5.76	7.65
70%	3.18	7.14	15.43	5.62	4.39	6.91
80%	3.37	6.38	16.17	5.37	4.86	6.32
90%	3.45	6.37	15.38	5.12	4.53	6.67
100%	3.78	6.18	15.46	5.27	4.87	6.30

Table 5. Standard Deviations of iMPA Performances
표 5. iMPA 분류 성능에 대한 표준편차

데이터셋	Breast Cancer	Glass	Ionosphere	Iris	New Thyroid	Wine
10%	3.21	13.60	10.53	4.30	8.61	7.69
20%	4.02	12.17	9.33	4.62	7.59	6.61
30%	3.03	11.49	7.31	4.25	4.50	6.89
40%	2.63	12.75	7.61	3.46	4.42	4.95
50%	2.45	13.58	6.25	3.51	5.03	5.18
60%	2.96	11.56	5.53	2.90	5.40	5.61
70%	2.80	12.14	6.58	3.80	3.59	5.96
80%	2.78	13.31	4.94	3.36	4.04	6.55
90%	2.70	12.23	5.13	3.69	3.50	6.41
100%	1.66	11.88	5.97	4.07	4.36	6.49

그림 10은 학습패턴이 점진적이 아닌 일괄적으로 처리되었을 때, 제안한 iMPA가 kNN 기법과 EACH 시스템과 비교하여 유사한 성능을 보여주고 있어 일괄적인 환경에서도 신뢰할만한 성능을 보인다. 표 6은 그림 10에 대한 표준편차이다. 이때 k-NN 기법은 Leave-one-out Cross-validation 기법으로 계산한 최적의 k값을 사용하였으며[9], 가중치 변화량 0.2를 초기값으로 설정하여 실험하였다. 다음 표 7은 각 데이터 셋에서 사용된 k-NN 기법의 k값과 k값을 계산하기 위하여 사용된 시간을 나타낸다.

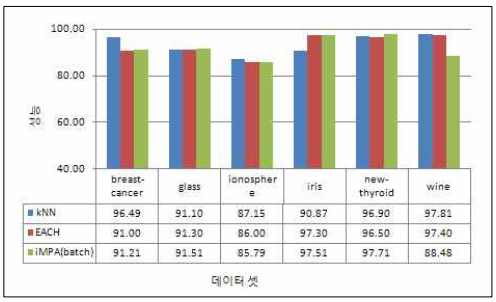

Fig 10. Batch Learning Performances (kNN, EACH, iMPA(batch))
그림 10. 일괄학습의 분류 성능 (kNN, EACH, iMPA (batch))

Table 6. Standard Deviations of Batch Learning Performances
표 6. 일괄학습의 분류 성능에 대한 표준편차

자료명	Breast-cancer	Glass	Ionosphere	Iris	New-thyroid	Wine
k-NN	2.24	5.37	5.08	7.16	3.66	3.57
EACH	3.66	5.19	18.13	5.58	4.84	6.29
iMPA (batch)	2.16	11.81	5.97	4.07	6.30	7.77

Table 7. k Value and Hour for kNN Method
표 7. 분류성능 최적화를 위한 k값 및 계산 시간 (Hour)

데이터셋	Breast Cancer	Glass	Ionosphere	Iris	New Thyroid	Wine
k값	21	1	1	51	1	19
시간	261	2.26	40.56	0.33	1.61	1.29

3. 메모리 사용량 비교

그림 11에서 그림 16까지에 걸쳐 학습패턴이 점진적으로 추가되어 학습될 때 EACH 시스템과 iMPA에 대해메모리의 사용량을 측정하였으며 각 그림에서 x축의 표는 학습패턴비율에 대한 메모리 사용량을 나타낸다. 이때 EACH 시스템의 경우는 메모리에 저장된 분할영역의 수 × 2를 저장된 학습패턴의 수로 사용하였는데, 이는 EACH시스템에서 메모리에 저장되는 분할영역이 평면의 범위를 나타내는 상, 하한의 두 개의 패턴으로 표시되기 때문이다. 또한 iMPA는 생성된 대표패턴의 개수로 측정하였다.

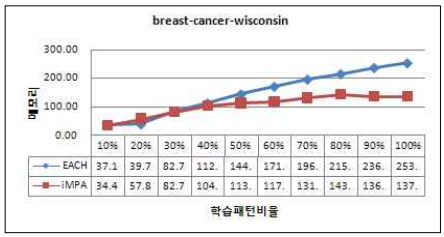

Fig 11. Memory Usages(Breast-cancer)
그림 11. 메모리 사용량(Breast-cancer)

Fig 12. Memory Usages(Glass)
그림 12. 메모리 사용량(Glass)

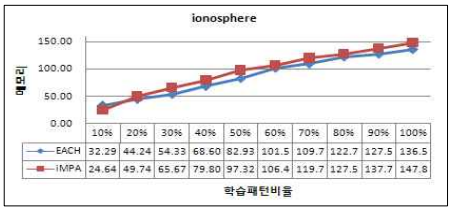

Fig 13. Memory Usages(Ionosphere)
그림 13. 메모리 사용량(Ionosphere)

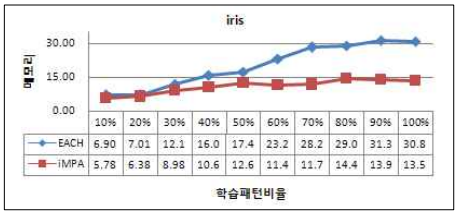

Fig 14. Memory Usages(Iris)
그림 14. 메모리 사용량(Iris)

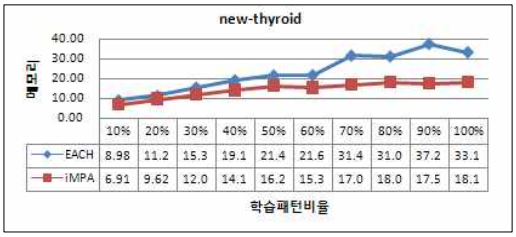

Fig 15. Memory Usages(New Thyroid)
그림 15. 메모리 사용량(New Thyroid)

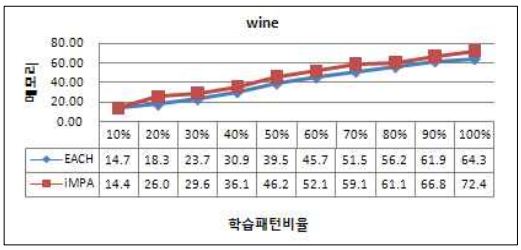

Fig 16. Memory Usages(Wine)
그림 16. 메모리 사용량(Wine)

제안한 iMPA기법이 EACH 시스템 기법보다 메모리 사용량이 유사하거나 우수한 것으로 나타났으며, 메모리 사용량도 학습패턴의 개수가 약 80% 정도부터 증가가 둔화되어 안정된 것을 나타내고 있다.

V. 결 론

점진적 학습은 몇 가지 특징을 갖는다. 첫째, 추가로 발생된 새로운 학습자료를 학습할 수 있다. 둘째, 추가로 발생된 새로운 학습자료를 학습할 때, 기존에 학습한 학습자료(original data)를 이용하지 않는다. 셋째, 학습하여 생성된 정보(대표패턴)는 점진적으로 갱신되며 삭제되지 않고 유지된다. 이 세 가지 특징은 새로운 학습패턴이 발생할 경우 매번 전체 학습 패턴을 이용하여 처음부터 학습을 다시 수행하는 것이 아니라, 새로 추가된 데이터만을 학습하여 사용할 수 있는 알고리즘의 조건이다.

본 논문에서 제안한 iMPA기법은 새로운 자료가 발생하였을 때, 기존 학습패턴의 분할영역 내에 존재하는 경우는 이전 학습 수행 시 패턴 평균법을 적용한 대표패턴과 새로운 학습패턴의 특징값을 다시 평균하여 기존의 대표패턴을 갱신한다. 그렇지 않은 경우에는 모든 특징에 대해 주어진 학습패턴공간을 분할하고 각 분할된 영역에 포함된 현재의 학습패턴이 속한 클래스를 검사하여 학습패턴의 클래스가 동일한 경우는 대표패턴을 생성하고 종료하며, 서로 다른 클래스에 속하는 패턴들이 혼재되어있는 경우는 다시 다분할을 실시한다.

기존의 메모리 기반 추론 기법이 외부 파라미터의 최적화와 전체 학습패턴을 저장 등의 문제로 인해 점진적 특징을 만족시키지 못하며, 신경회로망이나 결정트리기법과 같은 기법도 본 논문에서 실현한 점진적 학습기능은 없는 실정이다. 그러나 본 논문의 iMPA 기법에서는 외부파라미터를 전혀 사용하지 않으며, 전체 학습패턴 중 각 초월평면의 대표패턴만을 추출하고, 그에 대한 통계자료만을 이용한 점진적 학습을 가능케 한다. 그리고, EACH 시스템의 시간 복

턴간의 거리 계산으로 $O(n^2)$이며, iMPA의 시간 복잡도 계산은 분할점 선택이 가장 큰 영향을 주며, 경계값과 패턴과의 단순비교 회수인 $O(n^2)$이다. 이는 EACH 시스템과 유사한 시간 복잡도를 가지지만, 실제 학습에 필요한 시간은 EACH 시스템에 비하여 적다고 볼 수 있다.

참고문헌

[1] T. Dietterich, "A Study of Distance-Based Machine Learning Algorithms," Ph. D. Thesis, computer Science Dept., Oregon State University, 1995.

[2] D. Wettschereck, "Weighted kNN versus Majority kNN : A Recommendation," *German National Research Center for Information Technology*, 1995.

[3] D. Wettschereck, "A Hybrid Nearest-Neighbor and Nearest-Hyperrectangle Algorithm," *Proceedings of the 7th European Conference on Machine Learning*, pp. Pages: 323 - 335, 1995.

[4] D. Aha, "Instance-Based Learning Algorithms," *Machine Learning*, Vol. 6, No. 1, pp. 37-66, 1991.

[6] D. Wettschereck and T. Dietterich, "Locally Adaptive Nearest Neighbor Algorithms," Advances in Neural Information Processing Systems 6, pp. 184 ~ 191, 1994.

[7] S. Salzberg, "A Nearest Hyperrectangle Learning Method," Machine Learning, Vol. 6. No. 3. pp. 251-276, 1991.

[8] J.R. Quinlan, "Induction of Decision Trees," Machine Learning, Vol. 1, pp. 81-106, 1986.

[9] 심범식, 정태선, 윤충화, "최근접 초월평면 학습법에서 시드개수의 영향에 대한 분석", 한국정보처리학회, '98 춘계학술대회, 1998.

[10] 정태선, 이형일, 윤충화, "고정 분할 평균알고리즘을 사용하는 새로운 메모리 기반 추론," 한국정보처리학회논문지, 제6권 제6호, pp. 1563-1570, 1999.

[11] 이형일, "RPA분류기의 성능 향상을 위한 OHC알고리즘," 한국멀티미디어학회논문지, 제6권 제5호, pp. 824-830, 2003

[12] O. L. Mangasarian and W. H. Wolberg. "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[13] http://www.ics.uci.edu/~mlearn

저 자 소 개

이 형 일(정회원)

1985년 2월 명지대학교 전자계산학과 학사
1994년 2월 명지대학교 대학원 전자계산학과 석사
2000년 8월 명지대학교 대학원 컴퓨터공학과 박사
1984년 12월~1989년 11월 (주)쌍용정보통신
1990년 5월~1994년 8월 (주)시에치노컨설팅
2005년 9월~김포대학 인터넷정보과 부교수
관심분야 : 미디어 영상인식, 패턴인식, 에이전트시스템, 정보검색, 기계학습