퍼지 및 지능시스템학회 논문지 2003, Vol. 13, No. 3, pp. 328-333

초월평면 최적화를 이용한 최근접 초월평면 학습법의 성능 향상 방법

An Optimizing Hyperrectangle method for Nearest Hyperrectangle Learning

이형일

Hyeong-il, Lee

Kimpo College

요 약

메모리기반 추론에서 기억공간의 효율적인 사용과 분류성능의 향상을 위하여 제안된 NGE이론에 기반한 최근접 초월평면법은 학습자료를 초월평면상에 투영시켜 생성된 초월평면을 이용한다. 이때 학습자료에 포함될 수 있는 오류자료가 그대로 초월평면에 포함되어 분류의 정확성을 저해하는 요인으로 작용하는 단점을 가지고 있다.

본 논문에서는 기존의 최근접 초월평면의 단점을 보완한 초월평면 최적화(OH : Optimizing Hyperrectangle) 방법을 제안한다. 제안된 방법은 특징가중치 벡터를 초월평면마다 할당하여 학습하고, 학습 후 생성된 모든 초월평면에 대해 특징별 최빈구간을 추출하여 최적초월평면을 구성하여 분류 시 사용한다.

제안된 방법은 EACH시스템과 마찬가지로 k-NN분류기에서 필요로 하는 메모리 공간의 40%정도를 사용하며, 분류에 있어서는 EACH시스템 보다 우수한 인식 성능을 보이고 있다.

Abstract

NGE (Nested Generalized Exemplars) proposed by Salzberg improved the storage requirement and classification rate of the Memory Based Reasoning. It constructs hyperrectangles during training and performs classification tasks. It worked not bad in many area, however, the major drawback of NGE is constructing hyperrectangles because its hyperrectangle is extended so as to cover the error data and the way of maintaining the feature weight vector.

We proposed the OH (Optimizing Hyperrectangle) algorithm which use the feature weight vectors and the ED(Exemplar Densimeter) to optimize resulting Hyperrectangles.

The proposed algorithm, as well as the EACH, required only approximately 40% of memory space that is needed in k-NN classifier, and showed a superior classification performance to the EACH. Also, by reducing the number of stored patterns, it showed excellent results in terms of classification when we compare it to the k-NN and the EACH.

Key Words : 기계학습, 에이전트시스템, 정보검색.

1. 서 론

메모리 기반 학습법의 학습은 주어진 학습패턴을 모두 메모리에 저장하는 것이며, 입력패턴의 분류는 저장된 패턴과 입력패턴사이의 거리를 이용하므로 거리기반 학습(Distance Based Learning) 이라고도 한다[1,2].

메모리 기반 학습 알고리즘에 기반한 분류기로는 k-NN분류기를 들 수 있으며 k-NN (k-Nearest Neighbor)분류기는 메모리 상에 저장된 학습패턴 중 주어진 입력패턴과 가장 가까운 거리에 있는 k개의 학습패턴을 선택하여 그 중 가장 많은 패턴이 소속된 클래스로 입력패턴을 분류하는 방법을 사용한다[2, 3, 4]. 이러한 k-NN분류기는 그 성능 면에서 만족할 만한 결과를 보이고 있으며 이미 다양한 분야에 응용되고 있다. 하지만 이 기법의 가장 큰 문제점은 학습 패턴 전체를 메모리에 저장하여야 하므로 다른 기계학습 방법에 비하여 많은 메모리 공간을 필요로 하며, 저장된 학습 패턴이 증가할수록 분류에 필요한 시간도 많이 소요되는 단점을 갖는다[5, 6].

따라서 메모리 기반 학습 알고리즘이 가지고 있는 문제점을 해결하기 위한 방법에 대한 연구가 지금까지 활발히 진행되어 오고 있으며 대표적인 연구로 IBL(Instance Based Learning)[6] 과 NGE(Nested Generalized Exemplar)[7, 8] 이론을 들 수 있다.

접수일자 : 2002년 12월 20일
완료일자 : 2003년 4월 29일
이 논문은 2003학년도 김포대학에 연구비 지원에 의하여 연구되었음.

2. 관련연구

2.1 k-NN(k-Nearest Neighbors) 기법

k-NN분류기는 메모리기반 학습기법을 사용한 최초의 분류기로 이 방법은 Lazy Learning Algorithm이라고도 하는데, 그 이유는 학습 시에는 단순히 학습 패턴을 메모리에 저장하며, 차후 입력패턴의 분류 시 모든 계산이 수행되기 때

문이다[10].

이러한 k-NN분류기의 개략적인 알고리즘은 다음과 같다.

① 주어진 학습패턴을 모두 메모리에 저장한다.
② 입력패턴 Q의 분류를 위하여 메모리에 저장된 모든 학습패턴과의 거리(D)를 식(1)을 이용하여 계산한다.

$$D_{EQ}=\sqrt{\sum_{i=1}^{n}(E_{f_i}-Q_{f_i})^2} \quad (1)$$

이때 E는 메모리에 저장된 학습패턴을 나타내며, Q는 주어진 입력 패턴이다. 또한 n은 패턴을 구성하는 특징의 개수이며, $Q_{f_i}E_{f_i}$는 각각 학습패턴과 입력패턴의 i번째 특징 값을 나타낸다.

③ 입력패턴 Q와 가장 가까운 k개의 학습패턴을 선정한다.
④ 선택된 k개의 학습패턴 중 가장 많은 개수의 패턴이 소속되는 클래스로 입력패턴 Q를 분류한다.

위에서 보이는 것처럼 k-NN에서의 학습은 학습패턴을 저장하는 것 이외에 아무런 조치를 취하지 않는다. 이때 k값은 분류기의 성능을 최적화하기 위하여 일반적으로 Cross Validation기법을 이용하여 결정하며, 특히, k=1인 경우를 NN분류기라 한다[2, 3, 4, 9]. 또한 위의 과정에서 4번째 단계에서, 입력패턴과의 거리를 이용하여 가중치를 부여하는 방법을 WeightVote k-NN이라고 한다[3, 4, 9].

2.2 NGE 이론

최근접 초월평면 학습법(Nearest Hyperrectangle Learning Method)은 Steven Salzberg가 1991에 발표한 NGE (Nested Generalized Exemplar)이론에 기반한 거리기반 학습법의 한 종류이며, 실제 사람의 학습법을 모델링 한 EBL(Exemplar Based Learning) 방법으로 분류된다. 이 방법은 표본의 일반화와 초월평면의 가중치, 그리고 가변 거리함수 등의 특징을 찾을 수 있다[3, 9, 10, 11]. 이를 이용하여 EACH시스템이라는 분류기를 구현하였으며, 이 시스템에서는 주어진 학습패턴을 메모리공간에 초월평면(Hyperrectangle)의 형태로 저장한다[4, 8, 12, 13].

다음은 EACH시스템의 학습을 나타내는 순서도이다.

2.2.1 초기화(Seeding)

학습 패턴 중 임의의 순서로 몇 개를 선택하여 메모리에 저장한다. 이 단계에서는 아무런 예측이나 학습이 이루어지지 않는다. 이때 메모리에 저장되는 예제는 특징 벡터로 표현되며 각각의 특징값은 이산값, 연속값 어떤 것이든 사용할 수 있다. 연속값의 경우 같은 값임을 판단하기 위해 오류 한계값 (Error Tolerance)를 사용한다[8].

2.2.2 유사도 측정 (Match)

E와 H를 각각 학습패턴과 메모리상의 초월평면이라 할 때, 두 객체 사이의 거리는 다음에 따라 구한다.

가. H가 점(Point)인 경우.

$$D_{EH}=W_H\sqrt{\sum_{i=0}^{m}W_i\left(\frac{E_{f_i}-H_{f_i}}{\max_i-\min_i}\right)^2} \quad (2)$$

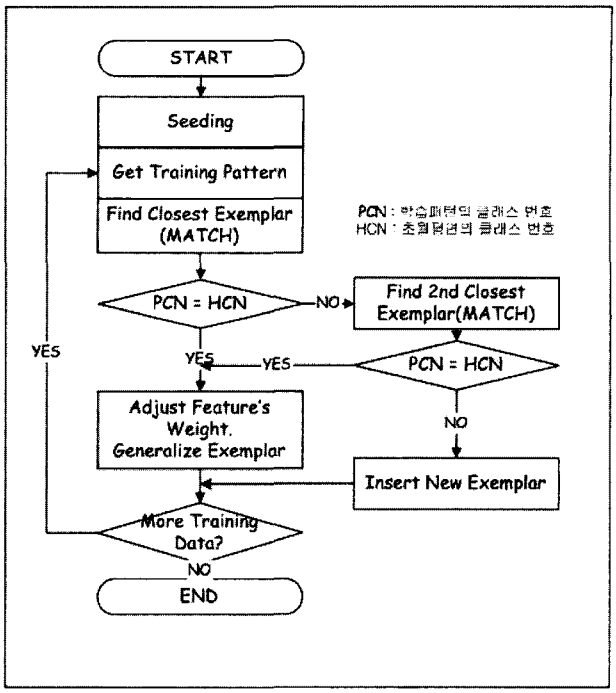

그림 1. EACH 학습 순서도
Fig. 1. Learning Flowchart of the EACH System

나. 초월 평면 H가 초월평면일 경우

$$D_{EH}=W_H\sqrt{\sum_{i=0}^{m}W_i\left(\frac{dif_i}{\max_i-\min_i}\right)^2} \quad (3)$$

여기에서는 식(3)에 따른다.

$$dif_i=\begin{cases}E_{f_i}-H_{upper} & when \quad E_{f_i}>H_{upper}\\ H_{lower}-E_{f_i} & when \quad E_{f_i}<H_{lower}\\ 0 & otherwise\end{cases} \quad (4)$$

식(2), (3)에서 W_H는 초월평면 H의 가중치, W_i는 i번째 특징의 가중치, m은 예제에 속한 특징의 개수, E_{f_i}는 예제에서 i번째 특징의 값, H_{f_i}는 초월평면에서 i번째 특징의 값, $\max_i$, $\min_i$는 i번째 특징이 가질 수 있는 최대/최소값을 각각 나타낸다.

초월평면과의 거리 D_{EH}는 학습패턴과 가장 초월평면의 모서리, 표면, 꼭지점 중 하나에 연결된 수직선의 길이가 된다. 만약 학습패턴이 초월평면의 내부에 속하게 될 경우 0이 되며, 중첩된 초월평면에 포함될 경우 가장 안쪽의 초월평면에 속하는 것으로 본다. 또 식(3)에서 사용된 가중치 W_H는 초월평면 H가 올바른 추론(Prediction)에 사용된 비율의 역수이다. 특징가중치와 초월평면 가중치는 모두 초기값으로 1.0을 사용한다.

2.2.3 학습(Learning)

EACH의 학습은 추론(Prediction)결과에 따라 다른 방법으로 특징가중치 및 초월평면 가중치를 수정하는 것이다.

가) 올바른 예측

E와 H의 클래스가 같은 경우로 초월평면은 새로운 예제

를 포함 할 수 있도록 커진다. EACH시스템에서의 초월평면 확대는 축에 나란하게(Axis-Parallel) 이루어진다. 새로 형성된 초월평면이 기존의 초월평면과 겹칠 경우(Overlap) 겹친 부분에 존재하는 표본은 크기가 작은 초월평면에 속하는 것으로 본다.

나) 잘못된 예측

E와 H의 클래스가 같지 않은 경우는 두 번째로 가까운 초월평면을 이용하여 다시 한번 예측을 시도한다. 두 번째 예측에도 실패할 경우 이 값은 메모리 상에 새로운 표본으로 저장된다.

다) 가중치 수정

EACH시스템에서 특징 가중치의 수정은 수식 (5), (6)에 의해 이루어진다.

$$w_i = w_i(1+\Delta) \tag{5}$$

$$w_i = w_i(i-\Delta) \tag{6}$$

이때는 i번째 특징을 위한 가중치를 나타내며, 이때는 시스템 전체의 효율에 크게 영향을 미치지는 않지만 실험적으로 0.2일 경우 가장 좋은 성능을 나타내고 있다. 특징 가중치의 수정은 올바른 추론과 잘못된 추론 각각 다르게 동작한다.

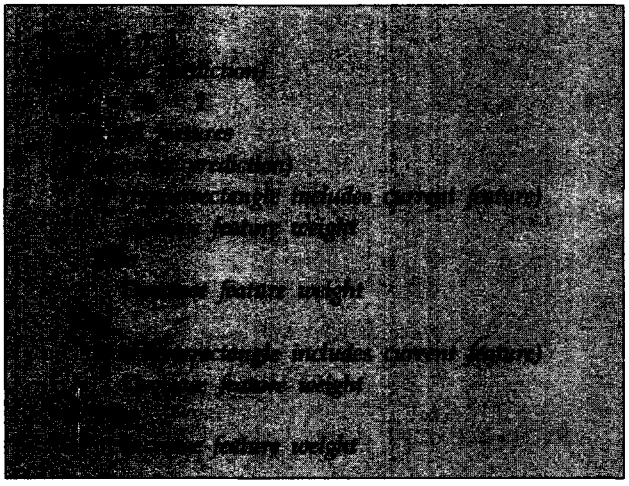

그림 2. EACH 가중치수정 Pseudo Code
Fig. 2. Weighting Pseudo Code

<그림 2>에서 H_t, H_c는 각각 초월평면의 총 사용수와, 올바른 추론에 사용된 회수이다.

3. 초월평면 최적화 방법

본 논문에서는 기존의 EACH시스템에서 사용하던 특징 가중치 방법을 수정하여 학습한 후 특징별 최빈구간 추출로 생성된 최적초월평면을 통하여 분류기의 성능 향상을 제안하였다.

3.1 초월평면별 특징 가중치 할당법

EACH시스템의 경우 전체 분류기에 하나의 특징 가중치만이 사용된다. 예를들어, 입력 패턴이 $\{x_1, x_2, \cdots, x_n\}$ n차원 백터일 경우, 전체시스템에 $\{w, w_2, \cdots, w_n\}$인 n개 값을 가지는 하나의 특징 가중치 벡터가 공통적으로 사용된다. 특징의 가중치는 학습패턴과 초월평면 사이의 거리를 결정하는 중요한 요소로서 어떤 특징이 초월평면(또는 자신이 속한 클래스)에 미치는 영향을 나타내는 값이다. 따라서 전체 시스템에서 하나의 특징 가중치 벡터를 공통적으로 사용하는 것은 모든 초월평면에 어떤 특징이 미치는 영향이 같다는 것을 의미한다. 또 이 방법에서는 n번째 패턴에 의해 수정된 가중치가 n+1번째 패턴의 학습에 영향을 미치게 된다. 이것은 n번째 학습패턴과 n+1번째 학습패턴이 서로 관련이 있다는 가정이 선행되어야 하는 것이다. 그렇지 않을 경우 분류기의 성능을 저하시키는 요인으로 작용할 수 있다. 또한 다른 초월평면의 경우 다른 특징값을 중요한 요소로 사용하게 될 경우 기존의 방법으로는 최적화된 분류효율을 보장하기 어렵게 된다.

본 논문에서는 현재 메모리에 존재하는 모든 초월평면이 자신이 사용하게 될 특징 가중치 벡터를 가지고 있는 방법을 사용하며, 학습에서는 다음의 수식 (7), (8)에 의해 특징 가중치를 수정한다.

$$w_{ij} = w_{ij}(1+\Delta) \tag{7}$$

$$w_{ij} = w_{ij}(1-\Delta) \tag{8}$$

이때 w_{ij}는 j번째 초월평면에서의 i번째 특징 가중치를 나타내며, EACH시스템에서와 같은 0.2를 사용한다.

3.2 특징별 최빈구간 설정을 통한 초월평면 최적화 방법

EACH시스템에서 올바른 예측 시 초월평면은 새로운 예제를 포함할 수 있도록 커진다. 즉, 초월평면 확대는 <그림 3>의 화살표 방향으로 축에 나란하게(Axis-Parallel) 이루어진다. 이것은 <그림 3>에서와 같이 생성된 초월 편면은 실제 자료(Row Data)를 그대로 공간상에 투영을 통해 초월평면을 확장하는 것을 의미한다. 그러나 학습 자료에 노이즈나 오류 자료가 포함된 경우에도 그대로 초월평면의 확장이 이루어져 분류 시 이용하게 된다. 이런 경우 형성된 초월 평면은 오분류를 일으키는 주요 요인으로 작용한다.

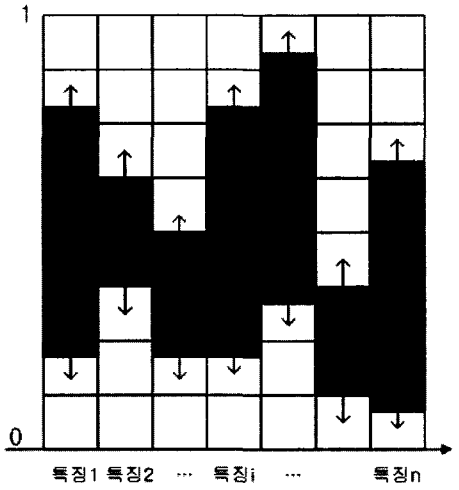

그림 3. EACH시스템의 초월평면 확장기법
Fig. 3. Generalizing Method of the Hyperrectangle in EACH

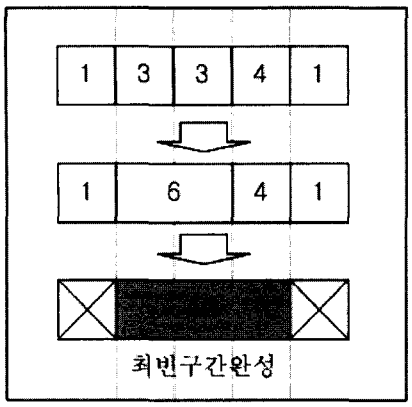

그림 4. ED를 이용한 유효 최빈 구간 계산
Fig. 4. Extracting Valid Ranges with ED

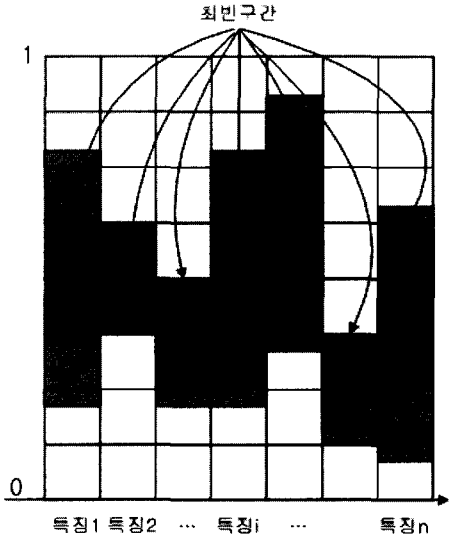

그림 5. ED를 이용한 초월평면 최적화기법
Fig. 5. Optimizing Hyperrectangle using ED

이에 대해 본 논문에서 제안된 ED(Exemplar Densimeter)알고리즘은 학습 후 생성된 모든 초월평면에 대하여 공간상에 분포된 패턴들의 특징별 최빈구간 추출을 실행한다. 이것은 공간에 분포된 패턴의 위상(topology)를 고려하여 각 특징별 유효 구간을 형성하는 과정이다.

ED작업은 연속값을 가지는 특징을 고려한 연관규칙의 추출 시 사용되는 구간 분할 알고리즘과 유사하다[14]. 기본 아이디어는 연속(Continuous)된 구간을 이산(Discrete) 구간으로 분할하는 것으로 분할 구간의 크기는 각 구간에 소속된 패턴의 개수에 따라 가변 적이며, 공간상에 분포된 각 클래스 별로 별도로 이루어진다. 다시 말하면, c개의 클래스로 구성된 패턴공간의 경우 각 특징에 대하여 c번 수행된다는 것이다.

이 알고리즘은 연관규칙 추출에 사용된 최빈구간 알고리즘[14]의 이진 분할 뒤 병합과 달리 최소 분할 후 병합(Divide First and Merge)기법을 사용한다. 즉, 주어진 특징에 대해 일정크기의 작은 구간으로 전체를 분할 한 이후 최소밀도(θ)를 만족하는 연속된 구간에 대해서는 병합 작업을 수행하는 것이다.

$$N_i = |\log_n(0.3 \times T_i)| \times n \quad (9)$$

$$\theta_i = AVE\left(\left|\frac{T_{i_j}}{N_i}\right|\right) \quad (10)$$

식 (9)는 특징 축의 분할 개수로 특징축을 N_{i_j}개로 분할하였을 경우, 구간에 포함되어야 하는 최소 패턴 개수이다. 이때 n은 패턴을 구성하는 특징 개수이고, $|T|$는 전체 학습 패턴 개수이다. 클래스 i에 대한 최소밀도 θ_i는 식(10)로 계산되며, 특징별 구간 분할 수 N과 관련하여 설명하면, 전체 패턴수를 분할한 각 특징의 분할구간 개수의 합으로 나눈 것의 평균값이 된다. 식(9)에서 전체 학습패턴의 30%에 근사한 초월평면을 형성하도록 선택한 이유는 k-NN에 있어 실험적으로 전체패턴의 약 30%만이 실제 분류에 사용되었다는 사실을 기준으로 한 것이다[13].

<그림 4>은 임계값 $\theta_i=2$, 구간수 $N_{i_j}=5$ 일때 ED 알고리즘을 적용하는 과정이다. 이때 셀은 구간을 의미하며 셀 내부의 수치는 각 셀에 소속된 패턴의 개수를 나타낸다.

즉, ED 알고리즘은 N개로 분할된 구간에서 특징축의 패턴밀도가 높은 구간을 찾아내기 위해 필요한 최소밀도(θ)을 만족할 때까지 병합하여 <그림 5>과 같이 패턴밀도가 높은 구간을 유지한다.

3.3 OH 학습기법 패턴분류

OH 방법의 학습 패턴 분류는 메모리 기반 알고리즘에서 사용하는 거리 기반 기법이며, 거리의 계산에는 분류 성능 향상을 위하여 식 (7)과 식 (8)로 주어진 wij 값을 입력패턴과 메모리에 저장된 학습패턴간의 거리계산에 있어 특징의 가중치로 사용한다. 다만 식 (7)과 식 (8)에서 고려 대상이 되는 분할 공간과 대상패턴은 전체이지만, OH에서는 wij가 OH 방법을 적용하여 형성된 최적 초월 평면에 포함되는 영역만을 계산에 사용한다.

OH 방법에서 패턴의 분류는 가장 인접한 초월평면과 그 다음으로 가까운 패턴의 클래스가 같을 경우 k=1인 NN분류기와 같은 방법으로 분류하게 된다. 만일 가까운 두 패턴의 클래스가 다를 경우, 데이터를 구성하는 모든 클래스에서 적어도 하나의 초월 평면이 추출될 때까지 거리 순서로 패턴을 추출하게 되며 이것이 k값이 된다. 따라서 OH 방법에서 분류 대상 초월평면의 수는 현재 입력패턴과 가까운 초월 평면의 개수에 따라 가변적이 된다. 그 후 입력패턴의 분류는 가장 많은 초월평면이 소속된 클래스로 분류한다.

4. 실험 및 고찰

OH 방법을 이용한 분류기의 성능을 k-NN, EACH기법과 비교하여 검증하였다. 실험은 기계학습의 벤치마크 자료로 사용되는 7개의 데이터 셋을 이용하였으며, 실험 방법은 70:30 법(전체 데이터 셋을 기준으로 70%는 학습패턴으로 30%는 평가패턴으로 사용하는 방법)을 사용하였다[13]. 이때 70%의 학습패턴은 전체 패턴의 클래스별 분포를 고려하여 모든 클래스에서 같은 비율로 추출하였다. 실험은 Windows 2000를 적재한 Pentium IV-2.0 컴퓨터를 사용하였으며, 모든 실험결과는 25회 반복측정 한 후 평균값으로 나타내었다.

4.1 실험 데이터 셋

본 논문에서는 기계학습의 벤치마크 자료로 사용되는 자

료 7개의 데이터 셋을 UCI Machine Learning Database Repository에서 가져와 사용하였으며, 이들 7개의 데이터 셋은 Breast-Cancer Wisconsin, Glass, Ionosphere, Iris, New-Thyroid, Sonar, Wine이며, 이들 데이터 셋은 <표1>와 같이 구성되며 모든 특징이 실수값을 가진다[5].

표 1. 실험 데이터셋의 구성
Table 1. Data Set

데이터 셋	전체 패턴개수	특징개수	클래스 개수
Breast-Cancer Wisconsin	699	10	2
Glass	214	10	6
Ionosphere	351	34	2
Iris	150	4	3
New-Thyroid	215	5	3
Sonar	208	60	2
Wine	178	13	3

표 2. 클래스별 학습패턴 분포
Table 2. the Training Patterns by class Data Set

데이터 셋	전체 학습패턴 개수	클래스별 학습패턴 개수					
		C1	C2	C3	C4	C5	C6
Breast-Cancer Wisconsin	488	320	168	×	×	×	×
Glass	148	53	11	0	9	6	20
Ionosphere	245	157	88	×	×	×	×
Iris	105	35	35	35	×	×	×
New-Thyroid	150	105	24	21	×	×	×
Sonar	144	67	77	×	×	×	×
Wine	123	41	49	33	×	×	×

<표 2>는 7개의 데이터 셋을 70:30법을 이용하여 분할하였을 경우, 클래스별 학습패턴의 분포를 보여주고 있다.

4.2 분류성능 실험

<그림 6>에서 k-NN, EACH, OH의 분류성능을 보면, 본 논문에서 제안한 OH 방법은 Glass, Sonar 두 개의 데이터 셋에서는 k-NN에 비하여 우수한 분류성능을 보이고 있는 반면, Ionosphere 데이터 셋에서는 다소 저조한 분류성능을 보이고 있으며, 나머지 4개의 데이터 셋에서는 두 분류기가 비슷한 성능을 나타내고 있음을 볼 수 있다. EACH시스템의 경우 Glass 데이터 셋에서 다른 기법에 비하여 우수한 분류성능을 보이기는 하지만, Breast Cancer, Ionosphere, Sonar 3개의 데이터 셋에서는 아주 저조한 분류성능을 보이고 있다. 이처럼 EACH시스템이 저조한 분류성능을 보이는 것은, [8]에서는 최초 시드(Seed)의 개수가 전체 학습에 큰 영향을 미치지 않는다고 기술되어 있다. 그러나 본 실험과 [12]에서 볼 수 있듯이 시스템에서 사용하고 있는 초기 시드(Seed) 개수의 영향에 의한 것으로 볼 수 있다.

본 논문에서 제안한 OH 방법과 Salzberg의 EACH시스템은 학습패턴으로 주어진 패턴 중 일부만을 저장하여 입력 패턴의 분류에 사용한다. 위의 결과에서 보듯 EACH시스템의 경우 테이터 셋에 따라서 분류기의 성능변화가 심한 것에 반하여, OH 방법은 안정된 분류성능을 보이고 있다.

위 실험에서 k-NN분류기 성능은 Leave-One-Out Cross Validation기법을 사용하여 계산한 k값을 사용한 것이며. EACH시스템과 OH 방법의 분류성능은 초기 시드(Seed) 개수 5, 가중치 증가량 0.2를 사용하여 측정한 결과이다.

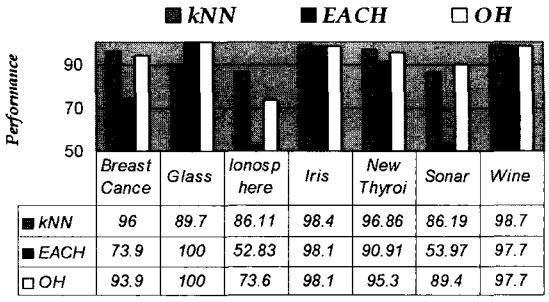

그림 6. kNN, EACH, OH 분류성능
Fig. 6. Classification Result of kNN, EACH, OH

다음 <표 3>는 k-NN분류기에서 각 데이터 셋에서 사용된 k값을 보여주고 있다.

표 3. 분류성능 최적화를 위한 k값
Table 3. The k data for the k-NN Classifier

데이터 셋	Breast Cancer	Glass	Ionosp here	Iris	New Thyroid	Sonar	Wine
k값	21	1	1	51	1	1	19

4.3 메모리 사용량 비교 실험

<그림 7>의 실험결과에서는 k-NN, EACH, OH 세 가지 방법을 이용한 분류기의 메모리 사용량을 보여주고 있으며, 표에 나타난 수치는 메모리에 저장된 학습 패턴의 개수를 의미한다. 이때 EACH시스템과 OH 방법의 경우는 메모리에 저장된 초월평면의 수×2를 저장된 학습패턴의 수로 사용하였는데, 이는 EACH시스템에서 메모리에 저장되는 초월평면이 평면의 범위를 나타내는 상, 하한의 두 개의 패턴으로 표시되기 때문이다.

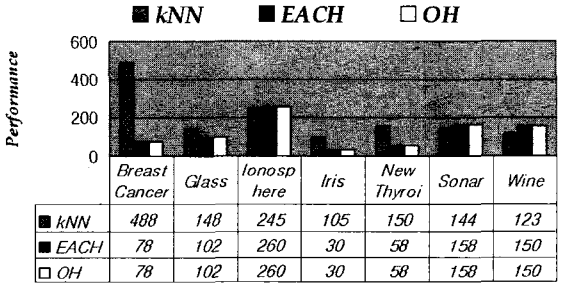

그림 7. 메모리 사용량의 비교
Fig. 7, Comparison of Memory Usage

결과에서 보면 k-NN의 경우 모든 학습패턴을 메모리에 저장하고 분류 시 입력패턴을 모든 학습패턴과 비교한다. 하

지만 EACH시스템과 OH 방법의 경우 구성된 초월평면을 사용함으로써 우수한 메모리 사용효율을 보장하고 있다.

본 논문에서 제안한 OH 방법에서는 Iris데이터의 경우 약 30%정도의 메모리만을 사용하고 있는 것을 볼 수 있으며 나머지 6개의 데이터 셋에서도 k-NN의 약 60~80% 정도만의 학습패턴만을 메모리에 저장하는 것을 볼 수 있으며, EACH 시스템과의 비교에 있어서 같은 효율을 보이는 것은 최빈구간 형성이 분리되어 나타나지 않은 것을 나타낸다. 여기에서 주목할 만한 것은, EACH시스템의 경우, Ionosphere, Sonar, Wine 3개의 데이터 셋에서 전체 학습패턴으로 주어진 패턴 수 보다 많은 패턴을 메모리에 저장하는 결과를 보이는데 이것은 앞에서 언급한 바와 같이 EACH시스템에서는 메모리에 저장되는 패턴을 초월평면의 형태로 표현하며, 이때의 초월평면은 상, 하한을 나타내는 2개의 패턴으로 구성이 되기 때문이다.

실험 4.2와 4.3에서 보는 것처럼 본 논문에서 제안한 OH 방법이 분류성능 대비 메모리 사용효율의 측면에서 기존의 k-NN분류기 및 EACH시스템에 비하여 우수한 성능을 보이고 있는 것을 볼 수 있다.

5. 결론 및 향후 연구과제

본 논문에서는 메모리 기반학습에서의 효율적인 메모리 사용과 분류성능을 향상시킬 수 있는 OH 방법을 제안하였다. 이 방법은 메모리 상에 존재하는 모든 초월평면이 사용할 특징가중치 벡터를 각각 유지한다. 따라서 전체 시스템에서 단 하나의 특징가중치 벡터를 사용할 경우 패턴의 학습순서가 전체 시스템의 성능에 미치는 문제를 제거하고, 학습패턴 간 상호 밀접한 관련을 가지고 있을 경우에도 효율적인 학습을 보장하고 있다.

또한 학습 후 생성된 모든 초월 평면들에 대하여 각 특징들이 갖는 구간에 대해 최빈구간을 추출하여 원시 데이터 셋에 존재하는 노이즈를 제거하고 패턴의 특성을 파악하여 효율적인 분류를 가능케 하였다.

연구 과제로는 점진적 학습시의 초월 평면 최적화 측면과 초월평면 내의 최빈구간이 분리되어 나타날 경우 메모리 사용효율 등에 연구가 필요한 것으로 사료된다.

참 고 문 헌

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

[2] D. Wettschereck and T. Dietterich, "Locally Adaptive Nearest Neighbor Algorithms", Advances in Neural Information Processing Systems 6, pp. 184-191, Morgan Kaufmann, San Mateo, CA. 1994.

[3] D. Wettschereck, "Weighted k-NN versus Majority k-NN A Recommendation", German National Research Center for Information Technology, 1995.

[4] S. Cost and S. Salzberg, "A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features", Machine Learning, Vol. 10, No. 1, pp. 57-78, 1993.

[5] D. Aha, "A Study of Instance-Based Algorithms for Supervised Learning Tasks: Mathematical, Empirical, and Psychological Evaluations", Ph. D. Thesis, Information and Computer Science Dept., University of California, Irvine, 1990.

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

[7] D. Wettschereck and T. Dietterich, "An Experimental Comparison of the Nearest-Neighbor and Nearest-Hyperrectangle Algorithms", Machine Learning, Vol. 19, No. 1, pp. 1-25, 1995.

[8] S. Salzberg, "A Nearest hyperrectangle learning method", Machine Learning, no. 1, pp. 251-276, 1991.

[9] D. Wettschereck, et al., "A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms", Artificial Intelligence Review Journal, 1996.

[10] T. Kohonen, "Learning vector quantization for pattern recognition (Technical Report TKK-F-A601), Espoo, Finland : Helsinki University of Technology, Department of Technical Physics, 1986.

[11] S. Salzberg, "On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach", Data Mining and Knowledge Discovery, Vol. 1, pp. 1-11, 1997

[12] 이형일, 정태선, 윤충화, "EACH시스템에서의 새로운 가중치 적용법", 한국 정보과학회 "98 춘계학술대회 논문집(B), 제25권 1호, pp288-290, 1998.

[13] 이형일, 정태선, 윤충화, 강경식, "재귀 분할 평균기법을 이용한 새로운 메모리 기반 추론 알고리즘", 한국정보처리학회 논문지 제6권 제7호, pp1849-1857, 1999

[14] 최영희, 장수민, 유재수, 오재철, "수량적 연관규칙탐사를 위한 효율적인 고빈도 항목열 생성기법", 한국정보처리학회 논문지 제6권 제10호, pp2597-2607, 1999

저 자 소 개

이형일(Hyeong-il Lee)

1985년 명지대학교 전자계산학과 졸업 (학사)

1994년 명지대학교 대학원 전자계산과 (석사)

2000년 명지대학교 대학원 컴퓨터공학과(박사)

기타양력

1985년-1990년 (주)쌍용정보통신

1990년-1996년 (주)시에치노컨설팅

1997년 현재 김포대학 컴퓨터계열 조교수

관심분야 : 기계학습, 에이전트시스템, 정보검색

Phone : 031-999-4173

Fax : 031-999-4216

e-mail : hilee@kimpo.ac.kr