韓國컴퓨터情報學會 論文誌
第14卷 第3號, 2009. 3.

2009-14-3-4-1

평가와 선택기법에 기반한 대표패턴 생성 알고리즘

이 형 일*

A Representative Pattern Generation Algorithm Based on Evaluation And Selection

Hyeong-il, Yih*

요 약

메모리 기반 추론 기법은 단순히 학습패턴이나 대표패턴의 형태로 메모리에 저장하며, 테스트 패턴과의 거리 계산을 통하여 분류한다. 이 기법의 가장 큰 문제점은 학습 패턴 전체를 메모리에 저장하거나 학습 패턴들을 대표 패턴으로 대체하는 방법을 사용함으로 다른 기계학습 방법에 비하여 많은 메모리 공간을 필요로 하며, 저장되는 학습 패턴이 증가할수록 분류에 필요한 시간도 많이 소요된다는 단점을 갖는다. 본 논문은 효율적인 메모리 사용과 분류 성능의 향상을 위한 EAS 기법을 제안하였다. 즉, 학습패턴에 대해 분할공간을 생성한 후 생성된 각 분할공간을 MDL기법과 PM기법으로 평가하였다. 그리고 평가 결과 가장 우수한 분할공간만을 취하여 대표패턴으로 삼고 나머지는 다시 분할하여 평가를 반복하는 기법이다. UCI Machine Learning Repository에서 벤치마크 데이터를 발췌한 실험 자료를 사용하여 제안한 기법의 성능과 메모리 사용량에 있어 우수함을 입증하였다.

Abstract

The memory based reasoning just stores in the memory in the form of the training pattern or the representative pattern. And it classifies through the distance calculation with the test pattern. Because it uses the techniques which stores the training pattern whole in the memory or in which it replaces training patterns with the representative pattern. Due to this, the memory in which it is a lot for the other machine learning techniques is required. And as the moreover stored training pattern increases, the time required for a classification is very much required.

In this paper, We propose the EAS(Evaluation And Selection) algorithm in order to minimize memory usage and to improve classification performance. After partitioning the training space, this evaluates each partitioned space as MDL and PM method. The partitioned space in which the evaluation result is most excellent makes into the representative pattern. Remainder partitioned spaces again partitions and repeat the evaluation. We verify the performance of proposed algorithm using benchmark data sets from UCI Machine Learning Repository.

▸ Keyword : Memory-Based Learning(메모리 기반 학습), MDL(Minimum Description Length), Information Gain(정보이득), PM(Probability Measure) 등

• 제1저자 : 이형일
• 투고일 : 2008. 11. 18, 심사일 : 2008. 12. 5, 게재확정일 : 2009. 1. 20.
* 김포대학 인터넷정보과 부교수
• 이 논문은 2009학년도 김포대학의 연구비 지원에 의하여 연구되었음.

Ⅰ. 서 론

데이터마이닝에서 메모리 기반 추론(Memory-Based Reasoning, 이하 MBR) 기법은 다양한 데이터를 효율적으로 분류해 나가기위한 방안으로 널리 사용되고 있다. MBR 기법은 새로운 자료를 분류하기위해 일반화된 규칙이나 모델을 구축하지 않는다. 즉 특징들의 최초 벡터 형태에 의해 표현된 패턴을 단순히 메모리에 저장한다. 그리고 분류 시에 새로운 자료를 메모리에 저장된 패턴들과의 거리를 계산하여 가장 가까운 거리에 있는 저장패턴의 클래스로 분류하는 기법이다. 그러나 방대하고 다양한 문제의 분류에 좋은 결과를 이끌어 낼 수 있지만 자료가 많아지면 학습 패턴 전체를 메모리에 저장하여야 하므로 필요한 메모리의 크기도 커질 뿐만 아니라 추론을 위한 계산도 많아지는 문제점이 있다[1][2].

MBR 기법에서 널리 사용되는 대표적인 분류기는 k-NN(k-Nearest Neighbors, 이하 k-NN) 분류기이다. k-NN분류기는 메모리에 저장된 패턴 중 주어진 입력패턴과 가장 가까운 거리에 있는 k개의 학습패턴을 선택하여 그 중 가장 많은 패턴이 소속된 클래스로 입력패턴을 분류한다. 이 분류기는 성능 면에서 만족할 만한 결과를 보여 다양한 분야에 응용되고 있지만, MBR의 분류기가 가진 문제점을 그대로 내포하고 있는 단점이 있다[3][4].

이와 같은 문제점들을 해결하기 위한 연구가 지금까지 활발히 진행되어 오고 있으며, 대표적인 연구로 학습패턴을 그대로 저장하는 것이 아니라, 인접한 학습패턴들을 포함하는 초월평면(Hyperrectangle)의 형태로 저장하여 이용하는 NGE(Nested Generalized Exemplar) 이론과 학습패턴에 대한 대표패턴을 추출하여 학습하는 FPA기법 등을 들을 수 있다[5][6].

FPA기법은 패턴공간을 고정분할하여 대표패턴을 생성하는 기법이다. 이때 분할된 패턴공간에 같은 클래스의 패턴들이 존재하면 평균하여 대표패턴을 생성하고, 만약 서로 다른 클래스의 패턴들이 존재하면 원래의 패턴 모두를 대표패턴으로 사용하였다. 이렇게 생성된 대표패턴은 개수가 많을 뿐만 아니라 불필요한 것을 포함하게 된다.

본 논문에서는 불필요한 대표패턴의 생성을 방지하여 메모리 사용 효율을 높이고 분류 시간을 단축시키기 위해서 분할공간 중에 평가척도가 높은 분할공간만을 선택하고, 그 외는 다시 분할하는 평가와 선택에 기반한 대표패턴 생성 알고리즘을 제안하고 구현하였다.

제안한 기법의 성능과 대표패턴의 개수의 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, 초월 평면을 사용하는 NGE 이론을 구현한 EACH 시스템과의 실험에서도 탁월한 분류 성능을 보여준다.

Ⅱ. 관련 연구

1. k-NN 기법

k-NN 분류기는 MBR기법의 대표적인 알고리즘이다. 먼저 전체 패턴을 단순히 메모리에 저장한다. 그리고 시험할 패턴과 메모리에 저장된 패턴들과의 거리를 수식 (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.

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

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 Validation기법을 사용하여 결정하며, k=1인 경우를 NN 분류기라 한다[6][7]. 이미 다양한 분야에 응용되고 있다.

표 1. EACH 시스템 알고리즘
Table 1. EACH system Algorithm

① 무작위로 몇 개의 학습패턴을 시드(seed)로 선택하여 예제(Exemplar)로 저장한다.
② 학습패턴을 선택하고, 가장 가까운 예제를 검색한다.
③ 학습패턴의 클래스와 가장 가까운 예제의 클래스가 동일하면, 학습패턴을 이용하여 그 예제를 확장하고 예제의 가중치를 수정한 다음, 단계 ⑥을 수행한다.
④ 클래스가 다를 경우, 가중치를 수정하고 두 번째로 가까운 예제를 선택한다.
⑤ 학습패턴의 클래스와 두 번째로 가까운 예제의 클래스가 동일하면, 예제를 확장하고 가중치를 수정하며, 다를 경우, 학습패턴을 별도의 새로운 예제로 저장한다.
⑥ 학습패턴 집합이 공집합이 될 때까지 단계 ②~⑤를 반복한다.

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][6][7].

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.

3. 고정분할평균기법

고정분할평균(Fixed Partition Averaging : 이하 FPA) 기법은 주어진 패턴공간을 동일한 크기의 분할공간들로 분할한 후 패턴 평균기법을 적용하는 방법이다[8].

패턴공간의 각 특징축을 수식 (2)에서와 같이 같은 크기의 N개로 분할한 후, 분할공간별로 패턴 평균법을 적용한다. 이 때, 여러 클래스의 패턴이 존재하는 분할공간의 경우에는 패턴 평균법을 적용하지 않고 원래의 패턴들을 그대로 저장하며, 단일 클래스의 경우는 해당 분할공간내의 모든 패턴을 평균하여 하나의 대표패턴으로 대체하는 방법을 사용한다.

$$N = \lceil \log_n(0.3 \times |T|) \rceil \quad \cdots\cdots (2)$$

이 때 선정된 N은 특징축의 분할 개수로 사용되며, n은 하나의 패턴을 구성하는 특징 개수이다. 또한 |T|는 전체 학습패턴의 개수이다.

FPA기법은 패턴공간을 일정한 크기로 분할하기 때문에 패턴의 분포를 고려할 수 없어 대표패턴의 개수는 증가하고 성능은 저하되는 문제점을 가지고 있다.

4. 점진적 다분할평균기법

점진적 다분할평균기법(incremental Multi Partition Averaging : 이하 iMPA) 기법은 전체 학습패턴공간을 패턴의 분포를 고려하여 가변 크기의 여러 개의 영역으로 반복해서 분할하면서 대표패턴(Representative Pattern)을 생성하며, 새로운 학습패턴이 추가적으로 발생하면 기존에 학습했던 모든 학습패턴에 대해 다시 학습하지 않고 추가된 학습패턴만 학습하여 대표패턴을 생성하는 기법이다[9]. iMPA기법의 구성은 입력패턴의 정규화와 학습패턴의 특징축 분할점 선정과 분할, 그리고 점진적 다분할평균기법 구현과 분류 등으로 구성된다.

입력패턴의 정규화는 모든 특징의 변화가 패턴의 소속 클래스 결정에 미치는 영향력을 동일하게 하기위해 패턴을 구성하는 모든 특징 값을 0과 1사이의 값으로 정규화 한다.

특징축의 분할점 선정은 iMPA의 성능과 대표패턴의 개수와 밀접한 관계가 있는 요소 중에 하나이다. 따라서 패턴공간의 특징축에 대해 경계값을 선정하여 분할하여, 분할 전후의 정보이득이 가장 큰 경계값을 분할점으로 선택한다. 즉 각 특징에 존재하는 특징값의 분포를 구한 후 특징값의 오름차순으로 분류하고, 특징값과 특징값 사이의 값을 식 (3)와 같이 경계값으로 정한다.

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

b_i는 특징의 i번째 경계값이고, f_i, f_{i+1}는 각각 i번째와 i+1번째 특징값이다. Upperbound는 특징 상한값으로 정규화된 경우는 1이 된다. 구한 경계값들 중에서 결정트리 알고리즘의 결정 노드(Decision Node)에서 특징의 비교 기준을 선정할 때 사용하는 IG(Information Gain) 값을 이용하여 가장 변별력이 좋은 경계값을 분할점으로 선택한다[10][11][12]. IG값은 수식 (4), (5)을 이용하여 계산한다.

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

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

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

I는 분할 이전의 정보량이며, Pi는 분할 이전의 학습 패턴 중, 분할된 각 영역에 포함된 학습패턴의 비율이다. Ii는 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.

점진적 다분할 평균기법의 구현은 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 (6)에 의해 갱신한다.

$$f_{new_i} = \frac{(f_{old_i} \times m) + f_i}{m+1} \quad \cdots\cdots (6)$$

$f_{\neq w_i}$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 분할을 실시한다.

III. 평가와 선택 기법에 의한 대표패턴생성

본 논문에서는 평가와 선택(Evaluating and Selecting) 기법에 의한 대표패턴 생성 알고리즘(이하 EAS기법)은 그림 1과 같이 분할단계와 평가와 선택단계, 그리고 대표패턴생성 단계의 세 가지 단계로 구성된다. 분할단계는 학습패턴에 의해 생성된 패턴공간을 iMPA기법에서 사용한 공간분할기법을 이용한다[9]. 분할 후 모든 분할공간에 대해 평가와 선택단계는 최소코드길이와 확률측도 등의 값을 이용하여 평가하여 최적의 분할공간을 선택하는 단계이다. 그리고 대표패턴생성 단계는 선택된 분할공간에 대해서는 대표패턴을 생성하고, 그 외의 분할공간에 대해서는 다시 패턴공간을 생성한 후 분할하여 대표패턴을 생성하는 과정을 반복하는 기법이다. 이때 대표패턴을 생성하는 학습패턴이 하나인 경우는 대표 패턴의 생성에서 제외하였다. 이 기법은 분할공간의 평가를 통해 불필요한 대표패턴이 생성되는 것을 방지하여 메모리 사용 효율을 높이는 장점이 있다.

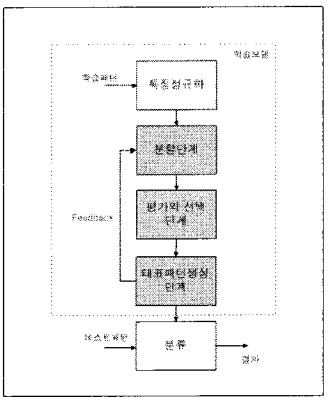


그림 1. EAS기법의 모델
Figure 1. Model of EAS Method

1. MDL(Minimum Description Length)

MDL은 최소코드길이의 값으로 자료를 표현하고자 변수의 값을 코드화하거나 전달을 위해 메시지를 코드화하고자 할 때 필요한 메시지의 길이를 계산할 때 활용되는 값이다[13]. 따라서 분할공간을 구성하는데 필요로 하는 비용을 계산하여 평가의 척도로 활용하고, 가장 적은 비용의 분할공간을 선택하였다. MDL을 구하기 위한 수식 (7)과 같다.

$$\begin{aligned} MDL = & \log(|T|+1) \\ & + fp*(-\log(e/2C)) \\ & + (C-fp)*(-\log(1-e/2C)) \\ & + fn*(-\log(fn/U)) \\ & + (U-fn)*(-\log(1-fn/U)). \end{aligned} \quad \cdots\cdots (7)$$

여기서, T는 전체 학습패턴 개수이며, C는 해당 대표패턴에 포함된 학습패턴 개수이다. 그 외 U와 fp, 그리고 e는 각각 해당 대표패턴에 포함되지 않은 학습패턴 개수, 해당 분할공간의 클래스와 상이한 클래스의 자료개수, 그리고 fp와 해당 분할공간을 제외한 공간에서 분할공간의 클래스와 같은 자료개수의 합이다. 분할공간의 선택조건은 분할된 공간들 중에서 MDL값이 가장 작은 분할공간을 선택한다. 이때 MDL값이 같은 경우에는 많은 학습패턴을 가진 분할공간을, 그리고

공간내 패턴의 Majority Class 클래스의 패턴개수가 큰 순으로 비교하여 분할공간을 선택한다.

표 2는 Breast-Cancer Wisconsin 자료에 대해 첫 번째 생성된 분할공간에 대해 선택 과정을 보여준다. EAS의 선택은 MDL이 647.5인 분할공간번호 22번이다. 이때, 분할공간번호는 iMPA기법으로 분할한 공간에 대해 선택조건으로 분류한 후 지정한 일련번호이다. 표 3은 두 번째 분할공간의 선택과정을 나타낸다. 표 2의 22번 분할공간이 제외되고 다시 생성된 패턴공간에 대해 분할을 실시하여 평가한 것이다. 즉, EAS의 선택은 MDL이 494.9인 분할공간번호 8번이다.

표 2. 최초분할의 분할공간평가
Table 2. partitioned space evaluation after the first partitioning process

분할공간번호	22	8	21	16	7	...
MDL	647.5	699.0	734.3	761.4	755.6	...
학습패턴개수	140	73	47	36	31	...
Majority Class의 패턴개수	131	73	47	33	31	...

표 3. 두 번째 분할의 분할공간평가
Table 3. partitioned space evaluation after the second partitioning process

분할공간번호	8	17	18	7	25	...
MDL	494.9	525.6	567.7	561.4	570.5	...
학습패턴개수	73	54	36	31	15	...
Majority Class의 패턴개수	73	54	33	31	14	...

2. PM(Probability Measure)

확률측도를 나타내는 PM 값은 임의로 생성한 분할공간의 분류 성능이 특정 분할공간의 성능보다 좋을 확률을 의미함으로, PM 값이 작을수록 분할공간의 분류성능이 좋은 것으로 간주된다[14].

$$\Pr(i) = \frac{\binom{P}{i}\binom{T-P}{t-i}}{\binom{T}{t}} \quad \cdots\cdots (8)$$

$$PM(R) = \sum_{i=p}^{\min(t,P)} \Pr(i) \quad \cdots\cdots (9)$$

T는 전체 학습패턴의 개수이며, P는 전체 학습패턴 중 분할공간 R의 클래스에 속하는 학습패턴의 개수이다. t는 분할공간 R에 포함된 학습패턴의 개수이며, p는 분할공간 R에 포함된 학습패턴들 중에서 대표패턴의 클래스와 같은 학습패턴의 개수이다.

분할공간의 선택은 분할된 공간들 중에서 PM값이 가장 작은 분할공간을 선택한다. 이때 PM값이 같은 경우에는 많은 학습패턴을 가진 분할공간을, 그리고 공간내 패턴의 Majority Class 클래스의 패턴개수가 큰 순으로 비교하여 분할공간을 선택한다.

표 4는 Breast-Cancer Wisconsin 자료에 대해 첫 번째 생성된 분할공간에 대해 선택 과정을 보여준다. 즉, EAS의 선택은 PM이 0이고 분할공간내의 학습패턴개수가 142인 분할공간번호 22번이다. 이때, 분할공간번호는 iMPA기법으로 분할한 공간에 대해 선택조건으로 분류한 후 지정한 일련번호이다. 표 5은 EAS-PM의 두 번째 분할공간의 선택과정을 나타낸다. 표 4의 22번 분할공간이 제외되고 다시 생성된 패턴공간에 대해 분할을 실시하여 평가한 것이다. 즉, EAS의 선택은 PM이 0이고 분할공간내의 학습패턴개수가 70인 분할공간번호 9번이다.

표 7. 최초 분할의 분할공간평가
Table 7. partitioned space evaluation after the first partitioning process

분할공간번호	22	9	7	21	31	...
PM	0	0	0	0.0001	0.0002	...
학습패턴개수	142	70	46	35	21	...
Majority Class의 패턴개수	132	70	46	35	19	...

표 8. 두 번째 분할의 분할공간평가
Table 8. partitioned space evaluation after the second partitioning process

분할공간번호	9	7	18	20	4	...
PM	0	0	0.0001	0.0005	0.0005	...
학습패턴개수	70	46	45	28	14	...
Majority Class의 패턴개수	70	46	45	26	14	...

3. 대표패턴 생성

대표패턴의 각 특징 값은 선택된 분할공간에 포함된 학습패

턴의 각 특징 값들을 평균하는 패턴평균법을 이용한다. 대표패턴의 클래스는 선택된 분할공간에 포함된 Majority Class Selection으로 결정한다. 그리고 선택된 분할공간에 학습패턴이 하나인 경우는 대표패턴을 생성하지 않는다. 또한 사용된 학습패턴을 학습패턴 집합으로부터 제거하여 이후에 분할할 때 영향이 없도록 한다. 이와 같은 대표패턴을 생성하는 과정을 더 이상 학습패턴이 존재하지 않을 때 까지 반복한다.

이상과 같이 제안한 EAS기법에 의한 대표패턴생성 알고리즘을 그림 2의 상세 구성도과 표 6의 EAS 알고리즘이다.

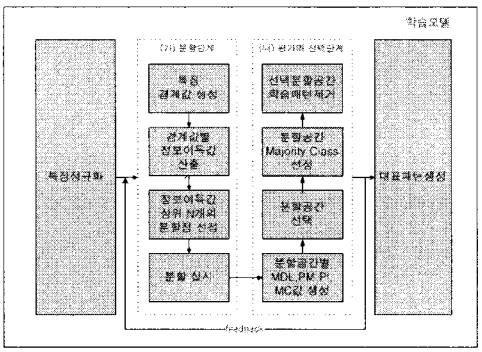


그림 2. EAS기법에 의한 대표패턴생성의 상세 구성도
Figure 2. Detail Modules for the generating algorithm of Representative Pattern by EAS Method

표 4. EAS기법에 의한 대표패턴생성 알고리즘
Table 4. Generation Algorithm of Representative Pattern by EAS Method

(1) 패턴공간을 III.2의 다분할을 실시한다.
(2) 모든 분할영역에 대해 MDL값과 PM값를 계산하여 각각 표 2과 표 4와 같이 분할 공간을 평가와 선택을 실시한다.
(3) 선택된 분할공간에서 하나의 대표패턴을 생성한다.
① 포함된 학습패턴에 대해 Majority Class로 클래스를 선택하고 해당 특징들에 대해 패턴평균법으로 대표패턴을 생성한다. 단, 학습패턴이 1개인경우는 대표패턴을 생성하지 않는다.
② 사용된 학습패턴을 학습패턴 집합으로부터 제거한다.
(4) (1)의 학습패턴 집합에 더 이상의 학습패턴이 존재하지 않을 때까지 단계 (2)~(4)를 실시한다.

4. 패턴의 분류

제안한 EAS기법은 테스트 패턴을 분류하기 위하여 대표패턴들과 수식 (10)로 거리 계산을 하여 가장 가까운 대표패턴의 클래스로 출력을 결정하였다. 이때 거리의 계산에는 분류성능 향상을 위하여 학습패턴의 최종적으로 생성된 분할공간에 대응하는 수식 (5)의 IG값 구해 입력패턴과 메모리에 저장된 학습패턴간의 거리계산에 있어 특징의 가중치로 사용한다.

$$D_{EQ} = \sqrt{\sum_{i=0}^{n} IG(i)(E_{f_i} - Q_{f_i})^2} \quad \cdots\cdots (10)$$

IV. 실험 및 분석

본 논문에서 제안한 EAS 기법의 성능을 Stratified 10-fold Cross-validation 기법을 사용하여 k-NN, EACH, FPA 등과 제안한 EAS의 MDL과 PM 알고리즘에 대해 비교 검증하였다. 점진적의로 추가되는 자료를 분류하는 iMPA기법은 실험특성이 맞지 않아 제외하였다.

1. 실험 데이터

본 논문에서는 기계 학습의 벤치마크 자료로 많이 사용되는 UCI Machine learning Database Repository에서 6개의 데이터 셋을 발췌하여 사용하였다[15]. 이들 데이터는 모든 특징이 실수 값을 갖는다. 다음의 표 7는 실험 자료의 분포를 보여주고 있다.

표 5. 클래스별 학습패턴의 분포
Table 5. Training Patterns in Classes

데이터 셋	패턴 개수	특징 개수	클래스별 패턴 개수		
			1	2	3
Breast-Cancer	699	9	458	241	-
Balance-Scale	625	4	49	288	288
Ionosphere	351	34	225	126	-
Iris	150	4	50	50	50
New-Thyroid	215	5	150	35	30
Survival	306	3	225	81	-

Breast-Cancer 데이터 셋은 Wisconsin 대학병원의 William H. Wolberg 박사가 정리한 유방암 진단 자료이며[16], Balance-Scale 데이터 셋은 심리학 실험에 사용하기 위해서 균형정도를 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, Haberman's Survival 데이터 셋(이하 Survival)은 시카고 빌링스 대학병원에서 유방암 수술 후 생존 연구 분석 결과이다.

2. 분류성능

분류 성능 실험에서 k-NN 기법은 Leave-one-out Cross-validation 기법으로 계산한 최적의 k값을 사용하였으며, 특히 EACH시스템은 시드(Seed) 개수 5, 가중치 변화량 0.2를 초기값으로 설정하여 사용하였다. 다음 표 8은 각 데이터 셋에서 사용된 k-NN 기법의 k값과 k값을 계산하기 위하여 사용된 시간을 나타낸다.

표 6. 분류성능 최적화를 위한 k값 및 계산 시간 (Hour)
Table 6. k Value and Hour for kNN Method

데이터셋	breast cancer	Balance-Scale	ionosph-ere	iris	new thyroid	Survival
k값	21	1	1	51	1	1
시간	261	2.26	40.56	0.33	1.61	1.01

표 9은 제안한 기법이 k-NN과 EACH, 그리고 FPA 기법과 비교한 분류성능이다. EACH 시스템이 glass와 Ionosphere에서 저조한 성능을 보이는 것은 무작위(Random)로 설정한 초기 시드(seed)의 영향으로 분석되다.

표 9에서 EAS-MDL기법은 breast-cancer를 제외하고 k-NN과 비슷하거나 우수한 성능을 나타내며, EAS-PM은 유사한 성능을 나타내고 있다. EACH나 FPA기법보다 모든 데이터 셋에서 우수한 결과를 보여주고 있다. 또한 표 10은 표 9의 분류 성능에 대한 표준편차를 보여준 것으로 제안한 EAS-MDL과 EAS-PM이 기존의 kNN과 EACH, 그리고 FPA보다 안정적인 결과를 나타내었다.

표 7. 분류성능
Table 7. Performance

데이터셋 / 알고리즘	breast cancer	Balance-Scale	ionosph-ere	iris	new thyroid	Survival
k-NN	96.8	88.7	86.7	90.9	96.9	79.4
EACH	94.4	50.8	70.3	94.6	93.9	64.0
FPA	90.7	87.8	86.3	95.6	95.9	77.6
EAS-MDL	96.2	88.1	85.3	96.1	93.0	76.5
EAS-PM	93.4	85.8	83.5	95.3	95.3	74.5

표 8. 표 9에 대한 표준편차
Table 8. Standard Deviations of Table 9

데이터셋 / 알고리즘	breast cancer	Balance-Scale	ionosph-ere	iris	new thyroid	Survival
k-NN	0.4	1.2	1.3	1.4	0.8	0.7
EACH	3.2	2.2	2.3	1.1	1.2	5.1
FPA	0.4	0.6	1.3	1.4	1.0	1.8
EAS-MDL	0.6	2.0	1.7	2.0	1.3	1.3
EAS-PM	0.4	1.7	1.5	2.5	1.3	1.4

3. 메모리 사용량 비교

표 11은 표 9에 대한 메모리의 사용량을 측정한 것이다. 이때 EACH 시스템의 경우는 메모리에 저장된 분할공간의 수 × 2를 저장된 학습패턴의 수로 사용하였는데, 이는 EACH시스템에서 메모리에 저장되는 분할공간이 평면의 범위를 나타내는 상, 하한의 두 개의 패턴으로 표시되기 때문이다. 그리고 FPA와 제안한 EAS-MDL 과 EAS-PM 등은 학습 후에 생성된 대표패턴의 개수로 측정하였다. 표 12에서 EAD-MDL 기법은 breast-cancer를 제외하고 모든 데이터셋에서 메모리 사용율이 매우 우수한 것으로 나타났다. 표 12은 표 11의 메모리 사용량 측정에 대한 표준편차를 보여준다.

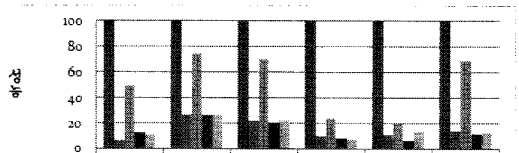
표 9. 메모리 사용량
Table 9. Memory Usage

데이터셋 / 알고리즘	breast cancer	Balance-Scale	ionosph-ere	iris	new thyroid	Survival
k-NN	629.1	562.5	315.9	185	198	275.4
EACH	40.6	149.4	70.8	18.7	20.4	38.9
FPA	308.8	417.3	222.4	43.7	38.9	189.6
EAS-MDL	79.2	146.4	63.3	15.2	13.1	31.7
EAS-PM	69.3	147.5	69.2	13.0	25.7	35.1

표 10. 표 11에 대한 표준편차
Table 10. Standard Deviations of Table 11

데이터셋 / 알고리즘	breast cancer	Balance-Scale	ionosph-ere	iris	new thyroid	Survival
k-NN	0.0	0.0	0.0	0.0	0.0	0.0
EACH	2.2	19.6	1.2	0.3	1.0	3.6
FPA	1.1	1.0	2.2	0.4	0.4	1.7
EAS-MDL	0.8	1.3	1.1	0.8	1.6	0.8
EAS-PM	1.4	1.1	1.0	0.3	1.3	0.6

그림3은 k-NN은 학습패턴 전체를 메모리에 저장함으로 메모리 사용율을 100%로 할 때, 각 기법의 메모리 사용율을 비교한 것이다. EACH 시스템과 FPA 기법이 각각 평균 15%와 51%를 사용하고 있으나, EAS-MDL은 14.2%를 그리고 EAS-PM은 15.3%를 사용하여 우수한 결과를 나타내었다.



	breast-cancer	balance-scale	ionosphere	iris	new-thyroid	survival
■k-NN	100	100	100	100	100	100
■EACH	6.5	26.6	22.4	10.1	10.3	14.1
※FPA	49.1	74.2	70.4	23.6	19.6	68.8
■EAS-MDL	12.6	26.0	20.1	8.2	6.6	11.5
※EAS-PM	11.0	26.2	21.9	7.0	13.0	12.7

그림 3. 메모리 사용율(%)
Figure 3. Memory Ratio(%).

V. 결 론

메모리 기반 추론 기법은 단순히 학습패턴이나 대표패턴의 형태로 메모리에 저장하며, 테스트 패턴과의 거리 계산을 통하여 분류한다. 이 기법의 가장 큰 문제점은 학습 패턴 전체를 메모리에 저장하거나 학습 패턴들을 대표 패턴으로 대체하는 방법을 사용함으로 다른 기계학습 방법에 비하여 많은 메모리 공간을 필요로 하며, 저장되는 학습 패턴이 증가할수록 분류에 필요한 시간도 많이 소요된다는 단점을 갖는다. 본 논문은 효율적인 메모리 사용과 분류 속도의 향상 기법을 제안하였다. 즉, 학습패턴에 대해 분할공간을 생성한 후 생성된 각 분할공간을 분류를 위한 척도로 평가하였다. 그리고 평가결과 가장 우수한 분할공간만을 취하여 대표패턴으로 삼고 나머지는 다시 분할을 실시하여 평가를 반복하는 기법이다. 이로써, 분할된 공간에 포함된 패턴들에 대해 변별력의 특성을 비교, 평가하여 최선의 분할공간을 대표패턴으로 선택하는 방법이다.

실험 결과에서 볼 수 있는 것처럼 EAS-MDL과 EAS-PM은 분류 성능 및 메모리 사용량에 있어 기존의 k-NN 기법과 EACH시스템에 비하여 아주 적은 개수의 대표패턴을 이용하여 비교대상의 기법에 유사한 분류성능을 보여주고 있다. 특히, EAS-MDL은 분류성능에 있어서 k-NN과 유사하고, 메모리 사용량은 평균 14.2%로 EACH시스템 15.3%(k-NN을 100%로 하였을때) 보다 우수한 수치를 나타냈다. 따라서 본 논문에서 제안한 EAS-MDL은 기존 기법들에 비해 대폭 감소된 대표패턴을 이용하여 우수한 분류성능을 안정적으로 보장한다.

참고문헌

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

[2] 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.

[3] 김경재, "자료편집기법과 사례기반추론을 이용한 한국종합주가지수 예측," 한국컴퓨터정보학회논문지, 제12권, 제6호, 287-295쪽, 2007년 11월.

[4] Cindy Marling, Edwina Rissland and Agnar Aamodt, "*Integrations with Case-Based Reasoning*," The Knowledge Engineering Review (2005), Cambridge University Press, pp. 241-245, 2006.

[5] 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.

[6] Tan S. "Neighbor-Weighted k-Nearest Neighbor for Unbalanced Text Corpus," Expert Systems with Applications, 28(4), pp. 667-671, 2005.

[7] Song, Y., Huang, J., Zhou, D., Zha, H., and Giles, L. "IKNN : Informative K-Nearest Neighbor Pattern Classification," In the Proceedings of PKDD, pp. 248-264, 2007.

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

[9] 이형일, "메모리기반 추론기법에 기반한 점진적 다분할 평균 알고리즘," 한국전기전자학회논문지, 제12권, 제1호, 65-74쪽 2008년.

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

[11] 노창현, 조규철, 마용범, 이종식, "의사결정 트리 기법을 이용한 그리드 자원선택 시스템," 한국컴퓨터정보학회논문지, 제13권, 제1호, 1-10쪽, 2008년 1월.
[12] 신성윤, 문형윤, 이양원, "전역적 결정트리를 이용한 샷 경계 검출," 한국컴퓨터정보학회논문지, 제13권, 제1호, 75-80쪽, 2008년 1월
[13] J.R. Quinlan, "MDL and categorical theories (continued)," Proceedings of 12th International Conference on Machine Learning , 1995, pp. 464-470.
[14] Witten, I. H. & Frank, E., *"Data Mining: Practical Machine Learning Tools and Techniques,"* Second Edition, Morgan Kaufmann, 2005.
[15] *http://www.ics.uci.edu/~mlearn*
[16] O. L. Mangasarian and W. H. Wolberg, "Cancer diagnosis via linear programming", SIAM News, Vol. 23, NO. 5, pp 1 & 18, Sept. 1990.

저 자 소 개

이 형 일

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)시예치노컨설팅
1997년 3월~김포대학 인터넷정보과 부교수
관심분야 : 기계학습, 미디어영상인식, 비디오요약