논문 2004-41CI-2-9

엔트로피를 기반으로 한 특징 집합 선택 알고리즘

(Feature Subset Selection Algorithm based on Entropy)

홍 석 미*, 안 종 일**, 정 태 충*

(Seok-Mi Hong, Jong-Il Ahn, and Tae-choong Chung)

요 약

특징 집합 선택은 학습 알고리즘의 전처리 과정으로 사용되기도 한다. 수집된 자료가 문제와 관련이 없다거나 중복된 정보를 갖고 있는 경우, 이를 학습 모델생성 이전에 제거함으로써 학습의 성능을 향상시킬 수 있다. 또한 탐색 공간을 감소 시킬 수 있으며 저장 공간도 줄일 수 있다. 본 논문에서는 특징 집합의 추출과 추출된 특징 집합의 성능 평가를 위하여 엔트로피를 기반으로 한 휴리스틱 함수를 사용하는 새로운 특징 선택 알고리즘을 제안하였다. 탐색 방법으로는 ACS 알고리즘을 이용하였다. 그 결과 학습에 사용될 특징의 차원을 감소시킴으로써 학습 모델의 크기와 불필요한 계산 시간을 감소시킬 수 있었다.

Abstract

The feature subset selection is used as a preprocessing step of a learning algorithm. If collected data are irrelevant or redundant information, we can improve the performance of learning by removing these data before creating of the learning model. The feature subset selection can also reduce the search space and the storage requirement. This paper proposed a new feature subset selection algorithm that is using the heuristic function based on entropy to evaluate the performance of the abstracted feature subset and feature selection. The ACS algorithm was used as a search method. We could decrease a size of learning model and unnecessary calculating time by reducing the dimension of the feature that was used for learning.

Keyword : 기계학습(machine learning), 특징 집합 선택(feature subset selection), 분류(classification), 엔트로피(Entropy), ACS(Ant Colony System)

I. 서 론

기계 학습 분야의 많은 알고리즘들은 예제 자료 집합(example data set)으로부터 지식 생성을 위해 사용된다. 그리고 생성된 지식들은 결과가 알려지지 않은 새로운 자료에 대한 클래스를 분류하기 위해 제공될 수 있다.

예제 자료들은 특징(feature)이라 불리는 많은 입력 패턴들로 구성되어 있으며, 예제들이 포함되어 있는 클래스(class)나 범주(category)를 나타내는 특징값(feature value)에 의해 표현된다. 기계 학습은 두 단계로 이루어진다. 학습(learning) 단계에서는 특징들과 클래스간의 관계나 규칙성을 찾기 위한 시도를 하고, 분류(classification) 단계에서는 학습 단계에서 추론된 학습 모델을 이용하여 결과를 알지 못하는 새로운 예제에 대한 클래스를 찾는다.

효과적인 분류를 위해서는 학습하고자 하는 개념과 관련된 많은 특징들이 필요하며, 실제로 문제에 따라 수많은 자료들을 수집할 수 있다. 그러나 수집된 많은 정보들 중에는 학습하고자 하는 개념(concept)과 관련이 없거나(irrelevant) 중복된(redundant) 정보를 가진 경우

* 정회원, 경희대학교 컴퓨터공학과
(Dept. of Computer Engineering, Kyunghee Univ.)

** 정회원, 용인송담대학교 컴퓨터소프트웨어과
(Dept. of Computer Software, YongIn Songdam college)

접수일자: 2003년11월11일, 수정완료일: 2004년2월11일

도 있다. 또한 자료 자체에 노이즈가 있기도 하다. 이와 같이 학습 모델 생성을 위해 수집된 정보가 신뢰할 수 없다면, 학습 과정에서도 정확한 지식의 습득이 어렵다[1].

특징 집합 선택(feature subset selection)의 과정은 종종 학습 알고리즘이 수행되기 전의 전처리 과정으로 사용되기도 한다. 왜냐하면 학습할 개념과 관련이 없거나 중복된 정보를 학습 모델 생성 이전에 제거함으로써 학습 알고리즘의 성능을 향상시킬 수 있기 때문이다. 이러한 과정을 통해서 많은 자료들 중 실제 분류 성능에 영향을 줄 수 있는 특징들을 검증해 낼 수 있다. 또한 학습 모델 생성에 사용될 입력 자료의 수를 줄임으로써 학습 알고리즘이 좀 더 빠르고 효과적으로 동작할 수 있으며 생성된 학습 모델의 크기도 줄일 수 있다.

본 논문에서 제안하는 방법은 어떤 사건이 발생했을 때, 그 사건의 보도 가치를 평가할 수 있는 엔트로피(entropy)를 기반으로 분류에 영향을 줄 수 있는 최적의 특징 집합을 선택하는 것이다. 특징 공간의 탐색 과정에서 계산되는 엔트로피를 통해서 선택된 특징들 간의 중복성이나 학습 개념과의 무관계성을 파악할 수 있다. 그럼으로써 기존의 예제 자료 집합내의 모든 특징들을 사용하여 학습 모델을 생성했을 때보다 분류 성능이 좋으면서도 작은 크기의 모델 생성이 가능하도록 하였다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 연구 배경에 대하여 기술하고, Ⅲ장에서는 본 논문에서 제안하고 있는 새로운 알고리즘에 대하여 설명한다. Ⅳ장에서는 새로운 모델의 성능 평가를 위해 실험과 그 결과를 보이며, Ⅴ장에서 결론 및 향후 연구 과제에 대하여 기술하도록 하겠다.

Ⅱ. 연구 배경

특징 집합 선택의 문제는 기계학습, 통계, 데이터 마이닝 그리고 패턴 인식 분야에서 활발히 연구되고 있으며, 초기의 특징 집합으로부터 중복성이 있거나 문제와 관련성이 적은 특징들을 제거하는 것이 목표이다. 이와 같이 예제 자료 집합으로부터 불필요한 특징들을 추출해냄으로써 학습 모델 생성 시 발생하는 계산 시간이나 많은 자료의 수집 및 관리에 드는 비용 등을 줄일 수 있다. 또한 만들어진 학습 모델로부터 생성되는 규칙들을 보다 쉽게 이해할 수 있다.

특징 선택 알고리즘은 두 가지 범주로 분류된다. wrapper 접근법[2,3]은 특징 집합의 정확성을 평가하기 위해 실제 학습 알고리즘을 사용하는 것으로 학습 알고리즘이 반복적으로 호출됨으로써 속도는 느리지만 그 유용성은 입증되었다. 그러나 학습에 사용되는 특징들이 많은 대량의 자료 집합(large data set)에서는 잘 사용되지 않는다. 반면 filter 접근법[2,3]은 어떤 학습 알고리즘과도 독립적으로 동작하는 방법으로 자료들의 일반적인 특성을 기반으로 한 휴리스틱 함수를 이용하여 선택된 특징 집합을 평가한다. 이 방법은 wrapper 접근법보다 빠르기 때문에 많은 예제 자료 집합이 사용되는 문제에서는 더 효율적이라 볼 수 있다. [그림 1]은 filter 접근법의 구조로써 본 논문에서도 여러 가지 면에서 더 효율적인 filter 접근법 형태의 알고리즘을 제안하였다.

특징 집합 추출은 학습하고자 하는 개념과 관련된 많은 특징들 중 분류에 영향을 줄 수 있는 최적의 특징 집합을 선택하는 것으로 최적화 문제의 한 분야로도 볼 수 있다. 학습에 사용 가능한 n개의 특징이 있는 경우, 이러한 특징들에 의해서 만들어 질 수 있는 특징의 조합은 2^n개가 된다. 물론 조합 가능한 모든 집합들의 탐색을 통해 최적의 특징 집합을 찾는 완전 탐색(exhaustive search) 기법을 이용할 수도 있겠지만 여러 가지 면에서 매우 비효율적이다. 또한 탐색할 특징 공간이 복잡할 경우 실제 수행이 불가능하다. 이와 같은 이유로 인해 유전자 알고리즘, 최적 우선 탐색등과 같은 다양한 휴리스틱 탐색 기법[4,5]들이 특징 집합 탐색을 위해 사용된다.

선택된 특징들이 얼만큼 예제 자료 집합을 잘 분류하는지 평가하는 것이 가장 중요하다. 본 논문에서는 발생 확률이 p(a)인 사상 a가 실제로 발생하였을 때 사상 a의 발생 사실을 인지하여 얻을 수 있는 엔트로피를 기반으로 한 휴리스틱 함수를 생성하여 특징 공간을 탐색하고 선택된 특징 집합에 대한 평가 값으로 활용하였다.

특징 공간을 탐색하는 방법으로는 개미들이 최적 경로(tour)를 생성해가는 과정에서 각 노드간의 에지

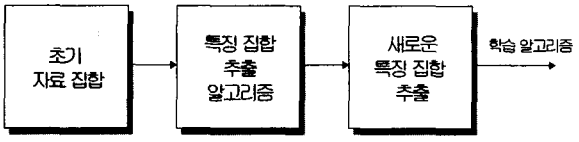


그림 1. filter 접근법의 구조.
Fig. 1. The structure of filter approach.

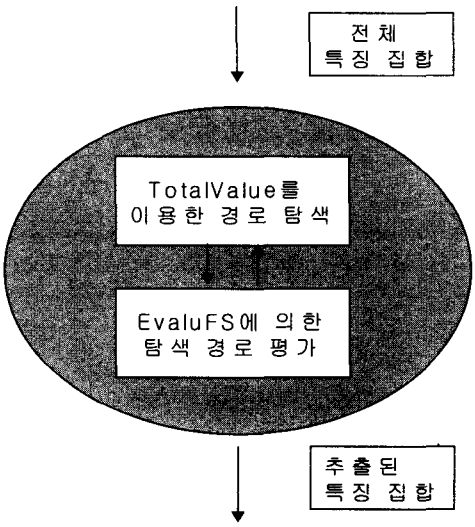


그림 2. 제안된 알고리즘의 구조.
Fig. 2. The structure of proposed algorithm.

(edge)에 대한 페로몬(pheromone) 정보를 이용한 반복적인 경로 생성 과정을 통해 최적 해를 발견하는 ACS(Ant Colony System)[6,7] 알고리즘을 사용하였다.

Ⅲ. 제안된 방법

본 논문에서는 ACS 알고리즘을 이용하여 다양한 특징 공간을 탐색하고, 특징 공간의 탐색과 추출된 특징 집합의 평가를 위하여 엔트로피 기반의 휴리스틱 함수를 이용하는 filter 접근법 형태의 알고리즘을 제안하였다. 알고리즘(acsFS)의 기본 구조는 [그림 2]와 같다.

1. 특징 집합 의 추출 및 평가

여러 가지 학습 방법 중 결정 트리(decision tree)[8,9]를 이용하여 학습 모델을 생성하는 경우, 정보 이론(information theory)에 따른 엔트로피개념을 사용한다. 일반적으로 엔트로피는 트리 생성시 하나의 노드에 속한 예제 집합의 무질서도를 나타내는 정량적인 수치로, 일단 트리가 학습된 상태에서는 각각의 터미널 노드가 하나의 클래스로만 결정되므로 무질서도는 0이 된다. 그러므로 예제가 모두 분류되었다는 것은 엔트로피가 0인 상태를 의미한다.

본 논문에서는 이러한 원리를 기반으로 특징 공간을 탐색한다. 탐색 과정에서 불필요한 특징들의 선택을 배제하기 위하여 하나의 노드에서 이동 가능한 모든 노드들에 대한 엔트로피를 계산한다. 예를 들면 하나의 특징이 선택되어진 후 다른 특징이 선택되어졌을 때 계산된 엔트로피 값이 큰 경우, 즉 트리의 무질서도가 감소하지 않는다면 선택된 특징들 간에 많은 중복요인이 존재하거나 분류에 크게 영향을 주지 못하는 특징이라고 볼 수 있다. 이러한 특징들 간의 엔트로피가 탐색 과정에서 각 노드(특징)들 간의 에지 정보가 되며, 탐색이 완료된 후 특징 간 에지의 페로몬 합이 가장 큰 특징 집합이 최적의 해로 결정된다.

어떤 한 사건 혹은 사상 a가 발생했을 때 그 사상이 시사하는 통보의 보도 가치, 즉 정보량을 I(a)라 한다. 일반적으로 이상한 일이 발생했을 때의 보도가치는 크다. 반면, 흔히 발생하는 일은 보도 가치가 별로 없다. 이러한 사실은 보도의 가치=정보량 I(a)가 사상 a의 발생 확률 p(a)의 함수라는 것을 암시한다. 그러므로 임의의 사상 A=(a_1, a_2,...,a_n)가 발생하여 얻을 수 있는 정보량의 기대치는 평균 정보량으로 식(1)[10]과 같이 나타낼 수 있다.

$$H(A) = -\sum_{i=1}^{n} p(a_i) log_x p(a_i) \quad (1)$$

또한 사상 A가 일어났다는 조건 하에서 사상 B가 일어날 확률은 식(2)와 같다.

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$
$$P(A \cap B) = P(A) \cdot P(B \mid A) \quad (2)$$

그러므로 식(2)에 의하여 하나의 특징을 선택하고 난 후, 다음 특징을 선택할 때 발생하는 두 특징간의 중복된 값은 식(3)으로 정의할 수 있다.

$$R(AB) = H(A) \cdot H(B \mid A) \quad (3)$$

H(B|A)의 값이 H(A)보다 크다면 특징 A로 분류한 후, 속성 B로 분류하는 것이 별 의미가 없다. 그것은 두 속성이 유사하거나 아니면 문제와 관계가 없는 속성일 수 있음을 의미하기 때문이다. 만약 H(B|A)가 0이라면 속성 A로 분류한 후, 속성 B로 분류했을 때 모든 예제들이 분류된 상태를 나타내는 것으로 좋은 특징 집합이라 할 수 있다. 그러므로 선택될 두 특징간의 총 정보량은 식(4)로 표현하고 노드간 에지에 대한 페로몬 정보로 활용한다.

$$TotalValue(A) = H(A) - H(A) \cdot H(B \mid A) \quad (4)$$

탐색 알고리즘에 의해 산출된 여러 개의 특징 집합의 평가를 위해서는 식(5)를 정의한다.

$$EvaluFS = \sum_{i=1}^{n} TotalValue(i) - \frac{n_{nc}}{n_t} \cdot \sum_{i=1}^{n} TotalValue(i) \quad (5)$$

탐색 알고리즘에 의해 선택된 특징 집합들의 TotalValue(A) 합에 전체 학습 예제의 수 nt 중 하나의 클래스로 분류되지 않은 가지의 예제 수 nnc의 비율을 곱함으로써 특징 집합에 대한 평가 시 발생할 수 있는 에러를 방지하고자 하였다. 결과적으로 EvaluFS가 가장 큰 집합이 최적의 특징 집합으로 선택된다. 즉 모든 예제들을 분류하고 또한 동시에 가장 특징의 수가 적은 특징 집합이 최적의 특징 집합으로 선택되어진다.

[그림 3]은 제안된 특징 선택 알고리즘의 전체 수행과정을 나타내는 RunFeatureSubsetSelect 프로시저이다.

```
Procedure RunFeatureSubsetSelect;
Begin
   BestAnt := 최적의 특징 집합을 만든 개미
   AntPath := 개미에 의해 방문된 경로를 기
              록하는 변수
While(종료 조건을 만족할 때까지)
   Initialize AntPath
   Initialize 각 개미의 시작 노드 설정
   SearchFeatureSpace();
   /*  특징 집합을 생성하는 프로시저
       LocalUpdatingRule() 포함 */
   BestAnt := SelectBestFeatureSubset();
   /* 생성된 특징 집합 중
      최적의 특징 집합 선택  */
   GlobalUpdatingRule(BestAnt);
   /* 최적의 특징 집합으로 선택된 경로상의
      모든 에지들에 대하여 페로몬 갱신 */
```

그림 3. 제안된 방법의 전체 프로시저.
Fig. 3. A total procedure of proposed method.

2. 특징 공간 탐색방법

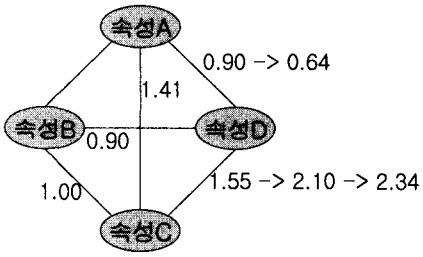
다양한 특징 공간의 탐색은 추출된 특징 집합에 대한 평가 방법 못지않게 중요한 과정이다. 본 논문에서 탐색 방법으로 선택한 ACS는 기존의 Ant System(AS)을 기반으로 한 휴리스틱 기법이다. 실제 개미들은 실질적인 단서 없이 단지 경로상의 페로몬 정보를 이용하여 그들의 둥지(nest)로부터 먹이(food source)가 있는 곳까지의 최단 경로를 찾는 능력을 가진다는 점에서 착안된 방법이다[6,7]. 개미들은 목적지를 향하여 지나가는 경로상에 개미들의 분비물인 페로몬을 분비한다. 그리고 그 이후에 지나가는 개미들은 이전에 축적된 페로몬 양을 정보로 경로를 선택한다. ACS 알고리즘은 상태전이규칙(State Transition Rule), 지역갱신규칙(Local Updat -ing Rule) 그리고 전역갱신규칙(Global Updating Rule)을 탐색 과정에서 각각의 개미들이 생성한 경로에 적용한다. 그럼으로써 지역 최적화에 빠지지 않고, 탐색 공간을 보다 넓고 다양하게 탐색하며 더 효율적으로 최적의 해를 찾도록 구성되었다.

ACS는 주로 순회 외판원 문제(TSP, Traveling Salesman Problem)[7], 배정, 라우팅, 그래프 칼라링과 같은 문제에 활용됨으로 특징 공간의 탐색을 위해서는 파라미터의 변형이 필요하다. 즉, 학습에 사용될 특징은 노드로 표현하고, 탐색될 노드들 간의 에지 정보(페로몬)는 특징들을 선택하는 과정에서 계산되는 TotalValue(A)를 이용한다. 개미들이 모든 경로에 대한 탐색을 완료한 후에는 EvaluFS 함수를 이용하여 탐색된 경로를 평가함으로써 최적의 특징 집합을 얻을 수 있다.

[그림 4]는 ACS를 이용한 특징 공간 탐색의 예이다. 사용된 예제 자료 집합은 전체 4개의 특징을 가지고 있고 특징 C와 D가 있으면 모든 예제 자료들이 분류되는 문제라 가정하자.

예제 자료 집합에 속한 특징들 간의 관계를 분석하기 위하여 모든 특징들은 완전 연결 에지를 가지고 있다. 탐색에 참여한 개미의 수는 4마리이다. 모든 개미들은 초기에 무작위로 시작 지점을 할당 받아 탐색을 시작한다. 각 개미들이 지나간 경로는 그림 아래 기술하였다.

개미들이 각 노드간의 엔트로피 정보를 이용하여 다음 노드를 선택하는 과정에서 지나가는 에지에 대한 페로몬 정보 TotalValue(A)를 수정한다. 즉 그림에서 보이는 바와 같이 최적의 특징 집합에 속하는 특징 C와 D를 연결하고 있는 에지의 페로몬 정보는 탐색 횟수가



Ant 0 : B -> D -> A -> C
Ant 1 : C -> D
Ant 2 : D -> C
Ant 3 : A -> D -> C -> B

그림 4. ACS를 이용한 속성 공간 탐색의 예.
Fig. 4. The example of feature space search using ACS.

반복될수록 증가하고 있는 반면 그 외의 에지의 페로몬 정보는 감소됨을 볼 수 있다. 모든 탐색이 종료된 후에는 특징 C와 D만을 가진 최적의 특징 집합을 생성하게 된다.

이와 같이 ACS를 이용한 특징 공간 탐색을 통해 서로 유사하거나 관련이 없는 특징들 즉, 엔트로피를 낮추지 못하는 특징들 간의 에지 정보를 줄여감으로써 특징 집합으로 선택될 가능성을 감소시키면서 탐색을 진행한다.

Ⅳ. 실험 및 결과

본 논문에서 제시한 모델의 성능을 평가하기 위해서 기계 학습용 데이터베이스 중 UCI 저장소[11]로부터 10개의 표준 자료 집합을 추출하였다. 자료 집합내의 특징 값들은 모두 명명형 자료(nominal data)이며, 클래스는 2개로 이루어져 있다. 각 자료 집합에 대하여 학습 모델 생성을 위해 사용할 학습 자료로 80%, 생성된 학습 모델의 성능을 평가하기 위한 테스트 자료로 20%를 활용하였다. 생성된 속성 집합의 분류 성능을 평가하기 위한 학습 알고리즘으로는 자체적으로 가지치기(pruning)를 하지 않는 ID3를 사용하였다. 그럼으로써 특징 부분 집합의 효과를 좀 더 확실하게 보일 수 있다.

ID3 알고리즘은 많은 예제들로부터 그 예제들이 암시적으로 포함하고 있는 개념을 추출하여 결정 트리의 형태로 일반화하고 이를 이용하여 새로운 예제들을 분류하는 기능을 가지고 있다.

<표 1>은 성능 평가를 위해 사용된 초기 자료 집합

표 1. 초기 예제 자료 집합.
Table 1. Original Example Data Set.

자료 집합명	학습 자료의 수	테스트 자료의 수	초기 속성의 수	클래스의 수
lympho-graphy	109	31	17	2
balloon	80	20	4	2
chess	400	100	36	2
connect	400	100	42	2
credit	530	136	9	2
gene	258	65	14	2
germen	400	100	17	2
mush-room	400	100	22	2
tictactoe	400	100	9	2
voting	185	47	16	2

에 대하여 간단히 설명한 것이다.

실험은 다음과 같은 순서로 진행하였다. 첫 번째, 초기 예제 자료 집합내의 모든 특징들을 이용하여 학습 모델을 생성하고 생성된 모델에 대한 분류 성능을 평가하였다. 두 번째, 제안된 모델(acsFS)을 적용하여 새로운 특징 집합을 추출하였다. 세 번째, 초기 예제 자료 집합에 대하여 acsFS에 의하여 생성된 특징들로만 이루어진 예제 자료 집합을 구성하여 학습 모델을 생성하고 분류 성능을 평가하여 보았다. 그리고 특징 선택을 위해 사용하는 TotalValue(A) 중 H(B|A) 값이 0인 경우 이미 모든 예제 자료들이 현재까지 탐색된 특징들에 의하여 완전 분류된 경우이므로 아직 방문하지 않은 노드가 있다고 해도 탐색을 종료하도록 하였다. 이러한 종료 조건은 기존의 TSP 문제의 해결을 위해 고안된 ACS의 탐색 방법에는 위배되어진다. 그러나 특징 집합 선택의 문제는 더 적은 수의 특징이 더 높은 분류 성능을 가지도록 하기위해서 수행되는 것이므로 모든 예제들이 분류된 후의 특징 공간의 탐색은 의미없다. 그러므로 기존의 ACS와는 다른 종료 조건을 제시하여 수행하였다.

<표 2>는 초기 예제 자료내의 특징 집합을 그대로 이용한 경우와 제안된 모델인 acsFS 적용 후 생성된 특징 집합을 이용하여 학습 모델을 생성하였을 경우에 사용

표 2. acsFS 모델 적용 후 분류성능 및 속성의수 비교
Table 2. The comparison of classification rate and the number of feature after appling acsFS model.

자료 집합명	ID3 적용		acsFS 적용 후 ID3 적용	
	분류성능 (%)	사용된 특징의 수	분류성능 (%)	사용된 특징의 수
lympho-graphy	51.6	17	61.3	7
balloon	85	4	90	2
chess	95	36	97	15
connect	81	42	82	15
credit	82.3	9	84.5	9
gene	95.3	14	93.8	14
germen	70	17	71	12
mush-room	100	22	100	2
tictactoe	73	9	75	8
voting	87.2	16	89.3	2

된 특징의 수와 분류 성능을 비교한 것이다.

<표 2>의 결과를 보면 acsFS를 적용하기 전과 후의 분류 성능에 있어서는 큰 차이를 보이지 않는다. 그 이유는 예제 자료 집합에 속한 모든 특징들을 이용할 경우에도 유사한 형태의 학습 모델의 생성이 가능하므로 분류 성능에 있어서는 많은 영향을 주지는 않는다. 그러나 사용된 특징의 수에 있어서는 acsFS의 적용 전에는 평균 18.6개를 사용하였고, 적용 후에는 평균 8.6개를 사용함으로써 특징의 수에 있어서는 약 60% 정도 특징의 수를 감소시켰다. 그러므로 비슷한 분류 성능을 보여도 학습 모델 생성에 사용될 특징의 수를 현저히 줄임으로써 학습 모델 내에 존재하던 불필요한 가지들을 제거하는 효과를 얻을 수 있었으며 저장 공간이나 학습 속도면에서도 좋은 성능을 보였다.

학습에 사용된 특징의 수를 줄이지 못한 학습 자료의 경우에도 특징 공간 탐색 시 산출된 최적의 탐색 경로가 존재하므로 그 정보를 이용하여 추가의 계산 과정 없이 바로 학습 모델의 생성이 가능하다. 또한 분류 성능의 향상을 위해서는 더 많은 속성이 필요함을 알 수 있다.

V. 결론 및 향후 연구 과제

보다 정확한 학습 모델 생성을 위해 어떤 특징들을 사용할 것인가가 기계 학습 분야에 있어서 가장 중요한 문제이다. 왜냐하면 아무리 좋은 학습 알고리즘이 있다고 해도 학습 모델 생성 시 사용할 자료가 학습할 개념에 적절하지 않다면 좋은 분류 성능을 가진 학습 모델을 생성할 수 없기 때문이다.

실제로 학습을 위해 수집되는 자료들 중에는 문제와 관련이 없는 것들도 있고, 서로 중복된 정보를 가진 특징들도 존재하게 된다. 이러한 자료들을 이용하여 학습 모델을 생성하기 전에 미리 최적의 특징 집합을 산출하여 활용함으로써 보다 효과적인 학습 모델의 생성이 가능하다.

본 논문에서 제안하고 있는 특징 집합 추출 알고리즘은 filter 접근법의 한 형태로써 특징 추출과 추출된 특징 집합에 대한 평가를 위해 어떤 하나의 사건이 발생했을 때의 정보량을 나타내는 엔트로피 기반의 휴리스틱 함수를 이용한다. 학습 트리의 분류 정도를 나타내는 엔트로피 값, 즉 무질서도가 0에 가까워지면 모든 예제들이 완전 분류되었다고 볼 수 있다. 하지만 무질서도가 커지게 되면 트리의 각 가지에 여러 개의 클래스들이 혼합되어 있는 상태로 볼 수 있다.

하나의 특징을 이용하여 예제 집합을 분류하고, 다시 다른 특징을 이용하여 분류하게 될 경우의 무질서도가 감소하지 않는 경우는 예제 집합을 분류한 두 특징이 서로 유사하거나 분류에 영향을 주지 못하는 자료로 간주할 수 있다. 왜냐하면, 아무리 많은 특징들이 학습 모델 생성에 사용한다고 해도 그 특징들이 서로 비슷하거나 문제와 관련이 없다면 분류 성능에는 그다지 큰 변화가 없을 것이기 때문이다. 이러한 원리를 이용하여 초기 자료 집합으로부터 최적의 특징 집합을 추출하는데 유용한 휴리스틱 함수를 생성하였다.

실험은 UCI 저장소로부터 얻는 10개의 자료 집합을 이용하였다. 실험 결과 대부분의 집합에서 학습에 사용할 특징의 수를 줄일 수 있었으며 분류 성능에서도 좋은 결과를 보임을 알 수 있었다.

같은 문제에 대하여 초기보다 적은 수의 특징을 사용함으로써, 보다 간단한 학습 모델의 생성이 가능하였다.

비록 특징의 수를 줄이지 못한 경우에도 특징 집합 추출 과정에서 생성한 경로를 이용하여 별도의 계산 과정 없이 학습 모델의 생성이 가능하였다.

하지만 현재 실험에서 사용된 특징들은 명명형 자료들로써 연속형 자료들에 대한 추가적인 연구와 실험이 필요하다.

참 고 문 헌

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

[2] G. H. John, R. Kohavi, and P. Pfleger. Irrelevant features and the subset selection problem, In Proc. of 11th Int'l Conf. On Machine Learning, p 121-129, San Mateo, CA, 1994, Morgan Kaufmann.

[3] K. kira and L. A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In 10th National Conference on Artificial Intelligence, p 129-134. MIT Press, 1992.

[4] D. Opitz. Feature selection for ensemble. In 16thNational Conf. on Artificial Intellignecr (AAAI), pp 379-384, Orlando, FL, 1999.

[5] Y-S Kim, W. N. Street and F. Menczer. Meta-Evolutionary Ensembles. In Proc. 2002 Int'l Joint Conf. on Neural Networks(IJCNN -02), pp 2791-2796, 2002.

[6] R. Beckers, J. L. Deneubourg and S. Goss, Trails and U-turns in the selection of the shortest path bye the ant Lasius Niger, Journal of Theoretical Biology, vol. 159, p 397-415, 1992.

[7] A. Colorni, M. Dorgio and V. Maniezzo, Distributed optimization by ant colonies, Proceedings of ECAL91-Euripean Conference on Artificial Life, Paris, France, F. Vardla and P. Bourgine(Eds.), Elsevier Publishing, p 134-142, 1991.

[8] J. R. Quinlan. Induction of decision trees. Machine Learning, 1:81-106, 1986.

[9] Utgoff P. An improved algorithm for incremental induction of decision trees. In Proceedings of the Eleventh International Conference on Machine Learning, p 318-325, 1994.

[10] 원동호 역, 정보와 부호이론, 도서출판 ohm, 1997.

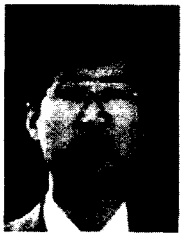
[11] UCI Repository of Machine Learning Data-bases.http://www.ics.uci.edu/~mlearn/MLRepository.html].

저 자 소 개

홍 석 미(정회원)

1994년 상지대학교 전자계산학과 이학사
1997년 경희대학교 컴퓨터공학과 공학 석사
1998년~현재 경희대학교 컴퓨터공학과 박사과정.

<주관심분야 : 기계학습, 데이터마이닝, 에이전트, 정보보호>



정 태 충(정회원)

1980년 서울대학교 전자공학과 학사
1982년 한국과학기술원 전자공학전공 공학석사
1987년 한국과학기술원 전자공학전공 공학박사

1987년~1988년 KIST 시스템 공학센터 선임연구원
1988년~현재 경희대학교 컴퓨터공학과 교수
<주관심분야: 기계학습, 보안, 최적화, 에이전트>

안 종 일(정회원)

1994년 경희대학교 전자계산공학과 공학석사
1998년 경희대학교 전자계산공학과 공학박사
1999년-2000년 혜전대학 전임강사.
2004년 현재 용인송담대학 컴퓨터정보과 조교수

<주관심분야 : 인공지능, 최적화알고리즘, 기계학습 등>