불완전한 데이터를 처리할 수 있는 분류기

이종찬* · 이원돈**

A Classifier Capable of Handling Incomplete Data Set

Jong Chan Lee* · Won Don Lee**

이 논문은 2009년도 충남대학교 학술연구비에 의해 지원되었음

요 약

본 논문은 변수 값들이나 부류 값을 손실한, 불완전한 데이터를 포함하는 데이터 집합을 가지고 학습하는 문제에 적용될 수 있는 분류 알고리즘을 소개한다. 이 알고리즘은 가중치 값과 확률 기법들을 이용하는 데이터 확장 방법을 사용한다. 이는 휘셔(Fisher)의 식을 기반으로 최적의 투사 면이 되도록 고려된 분류기를 확장함으로써 수행한다. 이를 위해, 데이터 확장에 적용되는 과정으로 부터 몇몇 식들이 유도된다. 제안한 알고리즘의 성능평가를 위해, 데이터에서 하나의 변수를 선택하고 이 선택된 변수에 소실 값과 소실되지 않은 값들의 비율을 변형함에 의해 다른 측정값들의 결과들이 반복적으로 비교된다. 또한 데이터 집합의 객관적인 평가를 위해 기계학습에서 지식 습득 도구로 널리 쓰이는 C4.5의 결과와 비교한다.

ABSTRACT

This paper introduces a classification algorithm which can be applied to a learning problem with incomplete data sets, missing variable values or a class value. This algorithm uses a data expansion method which utilizes weighted values and probability techniques. It operates by extending a classifier which are considered to be in the optimal projection plane based on Fisher's formula. To do this, some equations are derived from the procedure to be applied to the data expansion. To evaluate the performance of the proposed algorithm, results of different measurements are iteratively compared by choosing one variable in the data set and then modifying the rate of missing and non-missing values in this selected variable. And objective evaluation of data sets can be achieved by comparing, the result of a data set with non-missing variable with that of C4.5 which is a known knowledge acquisition tool in machine learning.

키워드

FLDF, 소실치, 확장된 데이터 표현, 최적의 투사면, 엔트로피 함수

Key word

FLDF, Missing value, Extended data expression, Optimized projection plane, Entropy function

* 청운대학교 인터넷학과

** 충남대학교 전기정보통신공학부 컴퓨터전공(교신저자)

접수일자 : 2009. 07. 18

심사완료일자 : 2009. 08. 03

Ⅰ. 서 론

정보통신의 기술이 비약적으로 발전함에 따라, 일부에서는 이미 각 사용자가 몇 백개의 근거리 무선 네트워크로 연결된 컴퓨터들로부터 계속적으로 상호 작용하는 차세대 컴퓨팅 환경에서 일하기 시작했다. 또한 정해진 규칙에 따라 운용되는 컴퓨터 네트워크로 인해 각각의 기기로부터 산출되고 있는 데이터들을 하나로 모으는 결합된 형태의 데이터 집적이 가능해졌다. 이 데이터들은 별개의 분산된 센서를 가지는 입력들로부터 모아진다. 본 논문은 이들 데이터가 여러 가지 운영체제, 네트워크, 기기 등의 서로 다른 환경으로부터 모아진다는 면으로부터 출발한다. 데이터가 다른 목적으로 사용되기 위해 별개의 기기로부터 모아지는 한, 그 포맷이 다를 수밖에 없을 것이다. 예를 들어 오류 등의 원인으로 데이터에 어떤 값이 빠져있다든지 어떤 문제를 처리하기 위한 변수(variable)값이 하나 이상 없는 것 같은 경우이다. 이러한 문제는 사회과학, 컴퓨터 비젼, 생물학 시스템, 원격 탐사 등의 폭넓은 분야의 문제에서 발생한다[15]. 이렇게 불완전한(incomplete) 데이터를 가지고 처리해야 하는 문제를 본 논문에서는 소실(missing) 데이터 문제라고 정의한다. 이 문제는 여러 개의 센서나 정보를 사용하는 미래에 일반화 되어가는 추세이다. 특히 유비쿼터스 환경에서는 이러한 경우가 자주 발생할 것이다.

불완전한 데이터를 가지고 문제를 처리할 경우 좋은 결과를 산출할 수 없기 때문에 과거 수 십년 동안 이를 해결하기 위해 통계학자들을 비롯한 많은 연구자들에 의해 연구가 이루어져 왔다[2]. 이들 연구들 중에 전체 데이터에서 완전한 데이터들만으로 가중치(weight)를 정하고, 이 가중치에 따라 샘플링(sampling)을 함으로써 불완전한 데이터에 대해 추론 해나가는 방법[2] 등이 있으나, 대부분의 연구들에서 데이터에 포함된 소실치를 처리하는 방법들은 두 가지로 크게 분류되어 진다.

(1) 소실치들을 무시(ignore)하고 처리하는 방법. [1][4][5][6][7][8].
(2) 여러 방법으로 소실치들을 채워나가는 방법. [3][9][10][11][12][13][14].

그러나 이들 대부분의 연구들은 불완전한 데이터에서 소실치를 처리하는 연구에 주력하였고, 이렇게 처리된 데이터를 가지고 기존의 학습 알고리즘을 이용해 학습을 하는 방법을 택하였다. 따라서 학습이 끝나고 성능을 측정하는 테스트 데이터에 소실치가 포함된 불완전한 데이터인 경우 이를 처리하는 데에 어려움을 가지고 있었다.

본 논문에서는 (2)번 방법의 일종 인, 데이터 확장 기법을 기반으로 이러한 문제를 해결하는 방안을 제안하고자 한다. 데이터 확장은 소실이 발생한 변수 값을 채워 넣는 부분과 각 이벤트 레코드들의 중요도를 나타내는 가중치를 처리하는 부분으로 나누어진다. 다시 말해 데이터 중에 임의 레코드에서 변수 값에 소실이 발생할 경우, 먼저 이 변수의 카디너리티(cardinality)를 이용해 이 변수가 가질 수 있는 균등한 확률 값으로 소실치를 채워 넣는다. 또한 이 과정에서 각 레코드들은 각각의 가중치 값을 가지게 되는데 이는 각 레코드의 중요도를 달리할 수 있다고 가정하고 이를 학습과정에서 반영할 수 있도록 해 준다. 기존의 다른 연구들과 다른 점이 불완전한 데이터를 처리하는 과정에서 각 레코드들마다 중요도를 나타내는 가중치를 가진다는 점이다. 이 가중치는 전문가가 전체 데이터 집합에서 어느 특정한 일부 데이터가 특별히 중요한 의미를 가지고 있어 이들을 전체 시스템에 주요한 의미를 부여하고자 할 때 이를 반영하여 학습이 가능하도록 해 줄 수 있다. 이 점은 AdaBoost 알고리즘[20][21]과 같이 데이터의 가중치를 각 이벤트 레코드에 이용하는 알고리즘에서 유용하게 쓰일 수 있다. 또한 제안한 방법의 장점은 퍼지 이론을 이용할 경우 확률 값이나 멤버쉽의 정도 등과 같이 미리 데이터에 대한 추가 정보가 필요하지 않다는 점이다. 또한 알고리즘이 간단하고 처리하기 쉽다는 점을 들 수 있다.

본 논문에서 이러한 확장된 데이터 표현기법으로 불완전한 패턴 데이터들을 확률적인 기법을 이용해 처리한 후 이를 FLDF(Fisher's Linear Discriminant Function)를 이용해 분류할 수 있도록 FLDF의 확장된 알고리즘을 제안한다. 여기서 확장은 기존의 FLDF 함수를 이용한 분류 알고리즘에 각 레코드들마다 가지고 있는 가중치들을 처리하는 부분을 반영한 것을 말한다. FLDF는 Fisher의 선형분리 함수를 이용하여 패턴을 분류하기 위한 최적의 투사면(projaction plane)을 결정하고, 학습 패턴들

을 이 투사면에 각각 투사한 후, 엔트로피(entropy) 함수를 이용하여 임계치(threshold)를 결정하는 방법을 이용해 학습패턴에 따라 반복적으로 분류해 나가는 방법으로 여러 분류문제에 적용하여 좋은 결과를 산출해 왔다. [18][19].

실험 과정을 통해 정해진 순서에 따라 훈련 데이터에서 한 변수를 선택하고, 앞서 설명된 방법을 적용해 이 변수의 소실치를 채워 넣는다. 실험은 데이터의 각 변수들을 변화시키고, 하나의 변수에서 소실치를 가지는 비율을 여러 가지로 변경해 가며 제안한 알고리즘의 성능을 비교한다. FLDF의 상대적인 평가를 위해 C4.5와 FLDF의 변수에 소실 데이터가 포함되지 않은 훈련 데이터에 대해서도 그 결과를 비교한다.

Ⅱ. 배 경

2.1 불완전한 데이터에서 소실치들의 표현 기법

2.1.1 소실치들을 무시하고 처리하는 방법.

일부 데이터에 특정한 변수 값이 빠져 있을 때 이러한 불완전 레코드를 가지는 데이터를 무시하고 완전한 데이터만을 가지고 분석하는 방법이다. 이 방법은 전체 데이터들 중에 소실 부분이 적을 때는 수행하기 쉽고 만족한 결과를 산출할 수 있다. 그러나 일반적으로 효율적이지 못하며, 부분 데이터만을 가지고 추론해야 하는 경우에는 특별히 비효율적이다.

(1) 레코드나 변수를 무시하는 방법[4]
데이터 집합에서 소실치를 포함하고 있는 레코드나 변수를 모두 지우는 방법이다.

(2) 소실 변수 값들을 무시하는 방법[5][6][7][8]
소실치를 널(null) 값으로 간주[6][8]하거나, 변수의 가능한 값 중 하나의 값으로 간주[5][7]하는 방법이다.

2.1.2 여러 방법으로 소실치들을 채워나가는 방법.

(1) 전문가가 직접 채우는 방법.[2]
일반적으로 이 방법은 시간이 많이 소요되고 소실치가 많거나 데이터 집합이 큰 경우 처리하기 쉽지가 않다.

(2) 일반적인 변수 값으로 채우는 방법.[13]
완전한 데이터만의 평균값으로 소실된 변수의 값을 채우거나, 알려진 변수의 값을 이용해 추론해서 소실치를 채워 넣는(imputation) 회귀분석 방법 등이 있다.

(3) 변수의 모든 가능한 값을 할당하는 방법. [10]
특정 변수에 소실이 발생할 경우, 이 변수의 카디너리티에 있는 모든 값이 이 소실치에 균등한 확률을 가지고 채워질 수 있다고 가정하는 방법이다. 일반적으로 이 방법은 일부 레코드에서 실제 데이터와 일치하지 않을 수 있고, 별도로 많은 데이터를 산출해 데이터 사이즈가 많이 증가한다.

(4) Rough Set을 사용하는 방법.[12][14]
(3)의 방법과 같이 초기에 모든 가능한 변수 값을 할당한 후, 불완전한 값의 하부(lower) 근사값과 상부(upper) 근사값을 정해 추정해 나가는 방법이다. 이 방법은 데이터 집합에 변수가 많거나, 변수의 카디너리티가 크다면 시간이 많이 소요된다는 단점이 있다.

(5) EM 알고리즘을 사용하는 방법[2][9]
수치 데이터에서, 모델을 기반으로 하는 과정으로 MLE(Maximum Likelihood Estimation)와 같은 기법을 이용하여 모델링하는 방법이다. 대표적으로 소실치들을 반복적으로 채워나가는 EM 알고리즘이 있다. 이 알고리즘은 많이 이용되면서 유용한 것으로 알려져 있다. 그러나 파라미터의 초기치에 대단히 민감한 알고리즘이어서 다른 클러스터링 기법을 사용하여 초기치를 추정하여 문제점을 완화하는 등의 기법을 사용해야 한다. 또한 수렴 속도가 대단히 느린 알고리즘이다. 시스템 자체의 성능을 향상시키는 알고리즘이 아니어서 제약된 최적화 문제를 풀기 위해 직접적인 방법이 없을 때 사용한다는 점이 지적되고 있다.

2.2 FLDF을 사용하는 점증적 학습 모델[19]

본 논문에서 사용하고 있는 패턴 데이터는 k개의 부류(class)를 가지는 n 차원의 벡터라고 정의한다. 이들 패턴을 각 부류에 따라 분류하기 위한 함수로서 (1)식과 같은 선형분리식(hyperplane)을 정의한다.

$$Hyperplane(H) = \{X | X \in R^n \wedge P^T X = T\} \quad (1)$$

X : 입력 벡터, R^n: 실공간(Real Space),
P : 투사면(Projection Plane),
T : 임계값(Threshold Value).

이 선형분리함수에 따라 입력 패턴 X는 (2)식과 같이 H^L과 H^R로 분리된다.

$$H^R = \{X \mid X \in R^n \wedge P^T X \geq T\}$$
$$H^L = \{X \mid X \in R^n \wedge P^T X < T\} \quad (2)$$

이 모델의 네트워크 구조는 하나의 입력노드, 여러 개의 은닉노드들, 부류수 만큼의 출력노드로 구성되며, 각 노드는 투사면 벡터와 임계값을 가지고 있다. 학습이 이루어지는 동안, 입력 패턴들에 따라 투사면 벡터와 임계값이 결정되고 (1)식에 의해 결정된 초평면은 임의 부류가 이루는 표면의 일부가 되면서 네트워크는 확장된다.

Fisher의 선형분리함수를 이용하여 임의의 분포를 가지는 입력 패턴 데이터에 대한 최적의 투사면 벡터를 결정하는데, 그 과정은 다음과 같다. k개의 부류를 가지는 Fisher의 식은 (3)에 나타나 있다. 이 식은 각 패턴들을 투사면에 투사했을 때 투사점들로 이루어지는 부류 간의 상호 거리(B)는 최대로 멀리 떨어지게 하고, 동시에 동일 부류에 속한 투사점(V)들은 가능한 모여 있는 것이 최적의 분류라는 직관을 반영한 것이다.

$$\frac{P^T \cdot B \cdot P}{P^T \cdot V \cdot P} = \frac{P^T \cdot [\sum_i (\overline{X_i} - \overline{X})(\overline{X_i} - \overline{X})^T] \cdot P}{P^T \cdot [\sum_i \sum_j (X_{ij} - \overline{X_i})(X_{ij} - \overline{X_i})^T] \cdot P} \quad (3)$$

여기서 (3)식의 최대값을 얻기 위해 Cauchy-Schwartz 부등식에 따라 $\alpha = V^{1/2}P$라 놓으면 이 식은 $(\alpha^T V^{-1/2} B V^{-1/2} \alpha)/(\alpha^T \alpha)$이 된다. $V^{-1/2}BV^{-1/2}$의 고유값들 중에 가장 큰 고유값(λ_1)에 상응하는 고유 벡터(e_1)를 구한다. Cauchy-Schwartz 부등식에 의해 $\alpha = e_1$일 때 $V^{-1}B$는 최대값을 갖는다. 따라서 투사면 벡터(P)는 $V^{-1/2}e_1$로 정한다.

입력패턴 각각을 투사면(P)에 투사(PTX)하고 투사방향에서 분리면을 결정하는 임계값을 결정하게 되는데 (4)식과 같은 엔트로피 식을 사용한다. 여기서 d는 투사면상의 투사점들을 순서대로 정렬했을 때 각 투사점들 사이의 평균을 말한다.

다시 말해 $d_s = (P^T X_{s-1} + P^T X_s)/2$, s=1,..., (패턴의 총갯수-1).

$$E(C \mid \delta^d) = PL^* E(q_1) + PR^* E(q_2) \quad (4)$$

$PL^* = n_1/(n_1+n_2)$, $PR^* = n_2/(n_1+n_2)$.
$n_1(n_2)$: d를 기준으로 왼쪽(오른쪽) 영역에 속한 패턴의 갯수

$E(q_i) = -\sum q_{ij} \log q_{ij}$, i=d의 왼쪽, 오른쪽, j=1,...,부류의 수

x_j를 각 영역(왼쪽, 오른쪽)에서 각 부류 j의 패턴개수라 정의하면 $q_{ij} = x_{ij} / n_i$가 된다.

투사면에 투사된 각 입력패턴의 점들을 작은 값부터 순서에 따라 d_1 부터 d_{s-1}까지 d를 기준으로 입력패턴을 왼쪽과 오른쪽으로 나누고 (4)식을 이용해 최소의 엔트로피 값을 갖는 위치를 찾아 이 위치의 d값을 임계값으로 정한다.

III. 불완전한 데이터를 처리하기 위한 FLDF 분류 모델

3.1 소실치 처리를 위한 데이터 표현[17]

표 1은 수치 값을 가지는 훈련 데이터의 일반적인 표현 예를 보이고 있다. 이 표는 4개의 변수(V1-V4)와 1개의 부류(Class)로 구성된다. V1과 V3 변수는 2개의 카디너리티, {1, 2}, V2와 V4 변수는 3개의 카디너리티, {1, 2, 3}을 가지고 있다. 그리고 부류 변수는 2개의 카디너리티를 가지고 있다.

표 1. 훈련 데이터의 예
Table 1. The example of training data

V1	V2	V3	V4	Class
2	1	1	2	1
1	2	2	3	2
2	3	2	1	1

표 2는 표 1 데이터가 구성된 이후에 추가로 만들어진 새로운 훈련 데이터의 예이다. 여기서 표 1(기존의 데이터)과 표 2(새로운 데이터)의 데이터를 합해 하나의 데이터 집합을 구성하려고 한다. 그러나 표 2의 (a)는 V2가 소실되었고, (b)는 부류가 소실되었다. 이렇게 불완전한 데이터가 포함될 경우 이를 처리하는 별도의 알고리즘이 선행되어야 한다.

표 2. 추가 훈련 데이터의 예
Table 2. The example of additional Training data

V1	V3	V4	Class
1	1	1	2

(a) V2 변수가 소실

V1	V2	V3	V4	Class
2	2	1	3	?

(b) 부류 값이 소실

표 3은 표 1과 표 2의 훈련 데이터 집합을 결합된 표현 결과를 보이고 있다. 본 연구에서 제안하는 데이터 표현은 소실치가 포함되지 않은 변수의 값들은 해당 엔트리에 간단히 0 또는 1을 채워 넣는다. 반면에 소실된 각 엔트리는 0에서 1사이의 확률적 값으로 채워 넣는다. 즉, 표 3에서 사건 1, 2, 3은 소실치가 포함되지 않은 경우의 처리 예이다. 또한 사건(event) 4는 표 2의 (a)의 경우를 처리한 예로, V2가 소실치를 가지고 있으므로 V2의 카디너리티인 3을 엔트리에 균등하게 1/3씩 배치한다. 마지막으로 사건 5는 표 2의 (b)의 경우를 처리한 예로, 부류의 카디너리티가 2이므로 1/2씩 균등하게 배치한다.

이 과정에서 가중치(weight, W) 개념을 추가하여 각 레코드를 재 정의한다. 가중치 값은 각 레코드들의 사건이 얼마나 중요한 지를 나타내는 것으로 전문가가 정한 값일 수도 있고, AdaBoost 알고리즘과 같이 알고리즘에서 산출해낸 값일 수도 있다. 임의의 레코드의 중요도를 나타내는 가중치가 1이라고 한다면 이 가중치는 다른 사건과 비교해서 1만큼의 중요도를 가진다고 할 수 있다. 예를 들어 표 3의 첫 번째 사건의 가중치가 10이라고 가정한다면 이 사건은 10개의 인스턴스와 같은 중요도를 가진다고 볼 수 있다. 가중치가 1인 사건은 그 자신의 인스턴스로 생각될 수 있다. 그러므로 사건의 수는 모든 인스턴스의 수와 같지 않을 것이다. 예를 들어 표 3에서 모든 인스턴스의 수는 14이지만 사건의 수는 5이다. 표 3의 예로부터 하나의 변수가 생략되었을 때 이 변수에 생략된 엔트리에 똑같은 확률을 가지는 값을 할당함으로서 해결함을 볼 수 있다.

3.2 확장된 데이터 표현을 가지는 FLDF 모델

앞 절에서 소실치를 가지는 불완전한 데이터 집합을 처리하기 위해 확장된 데이터 표현 방법을 살펴보았다. 이 방법에서는 새롭게 가중치 개념이 포함되었고 각 데이터 엔트리의 표현 방법도 변화하였다. 따라서 FLDF

표 3. 확장된 데이터 표현.
Table 3. The extended data representation.

Event	Weight	V1		V2			V3		V4			Class	
		1	2	1	2	3	1	2	1	2	3	1	2
1	10	0	1	1	0	0	1	0	0	1	0	1	0
2	1	1	0	0	1	0	0	1	0	0	1	0	1
3	1	0	1	0	0	1	0	1	1	0	0	1	0
4	1	1	0	1/3	1/3	1/3	1	0	1	0	0	0	1
5	1	0	1	0	1	0	1	0	0	0	1	1/2	1/2

모델에서 추가로 이들에 맞게 투사면과 엔트로피 함수를 새롭게 정의해야 한다. 이를 위해 관계되는 항들을 다음과 같이 정의한다.

- cardty[j] : 변수 j(Vj)의 카디너리티 수.
- num[i] : 부류 i를 가지는 사건의 수.
- vanum : 훈련데이터에서 변수의 수(n).
- evnum : 훈련데이터에서 전체 사건의 수.
- clanum : 훈련데이터에서 부류의 수(k).

가중치 값을 고려한 실제 데이터 값을 구하기 위해 다음과 같이 $M(\bullet)$ 함수를 (5)식과 같이 정의한다.

$$M(X, j, k) = \sum_{m}^{cardty[k]} X_{jkm} \cdot W_j \cdot m \tag{5}$$

(5)식을 이용해, 학습패턴에서 각 부류의 평균 벡터는 다음과 같이 유도한다.

$$\overline{X_i} = \sum_{j}^{vanum} \sum_{k}^{num[i]} \sum_{m}^{cardty[j]} X_{ikjm} \cdot W_{ik} \cdot m \Big/ \sum_{k}^{num[i]} W_{ik}$$
$$= \sum_{j}^{vanum} \sum_{k}^{num[i]} M(X_i, k, j) \Big/ \sum_{k}^{num[i]} W_{ik} \tag{6}$$

그리고 훈련 데이터의 전체 평균 벡터를 다음과 같이 변형된 수식을 이용해 구한다.

$$\overline{X} = \sum_{j}^{vanum} \sum_{k}^{evnum} \sum_{m}^{cardty[j]} X_{kjm} \cdot W_k \cdot m \Big/ \sum_{k}^{evnum} W_k$$
$$= \sum_{j}^{vanum} \sum_{k}^{evnum} M(X, k, j) \Big/ \sum_{k}^{evnum} W_k \tag{7}$$

종합적으로 (3)식과 같은 휘셔(Fisher)의 수식은 가중치가 포함된 확장된 데이터 집합에 적용하기 하기 위해 (8)식과 같이 변형된다.

$$\frac{P^T \cdot \sum_{i}^{clanum} (\overline{X_i} - \overline{X})(\overline{X_i} - \overline{X})^T \cdot P}{P^T \cdot \sum_{i}^{clanum} \left(\sum_{j}^{vanum} \sum_{k}^{num[i]} M(X_i, j, k) - \overline{X_i} \right)\left(\sum_{j}^{vanum} \sum_{k}^{num[i]} M(X_i, j, k) - \overline{X_i} \right)^T \cdot P} \tag{8}$$

투사 과정에서 $(P^T X)$ 수식은 $(P^T WX)$으로 변경된다. 또한 엔트로피 함수에서는 왼쪽(오른쪽) 지역의 사건 수를 변경해야 한다. leftnum(rightnum)은 (4)식에서 d를 기준으로 왼쪽(오른쪽) 사건의 수를 의미한다면, n_1 (n_2)는 초평면에 의해 분리되는 왼쪽(오른쪽) 지역에서의 인스턴스의 수이다. n_1 (n_2)는 다음과 같이 얻는다.

$$n_1 = \sum_{i}^{leftnum} W_i, \quad n_2 = \sum_{i}^{rightnum} W_i$$

마지막으로 (4)식에서 부류가 j일 확률 q_i는 다음과 같이 구한다.

$$q_{ij} = (WX_{ij}) / n_i, \quad i = leftnum, rightnum, \quad j = 1, \cdots, clanum$$

Ⅳ. 실 험

실험 데이터로 UCI Machine Repository에서 Balance Scale Weight & Distance Database와 Car Evaluation Database를 사용하였다. 또한 뇌파를 측정해 이를 양자화(quantization)한 데이터인 sleep stage scoring 데이터도 사용하였다. 실험과정에서 10-fold cross validation을 각 데이터 집합에 적용하였다. 즉, 각 데이터 집합을 임의대로 10개의 블록으로 나눈 후, 9개의 블록은 훈련데이터로 사용하고 나머지 1개의 블록은 제안한 시스템의 정확도를 측정하기 위한 테스트 데이터로 사용하였다. 이들 데이터 집합(훈련, 테스트)을 이용해 실험을 한 후, 10개의 블록을 하나로 합친 후 같은 방법으로 다시 2개의 데이터 집합(훈련, 테스트)으로 나누었다. 이 과정을 각 데이터에서 하나의 변수에 대해 10번을 수행한 후, 그 실험 결과의 평균을 산출하였다. 그 결과를 표 4에서 확인할 수 있다.

표 4. 실험 결과들.
(a) Balance Scale Weight & Distance 데이터베이스 : 3 변수, 3 부류, 625 사건
(b) Car Evaluation 데이터베이스 : 6 변수, 4 부류, 1728 사건
(c) Sleep Stage Scoring 데이터 : 11 변수, 6 부류, 1236 사건
Table 4. The experiment results.
(a) Balance Scale Weight & Distance Database : 3 Variable, 3 Class, 625 Event
(b) Car Evaluation Database : 6 Variable, 4 Class, 1728 Event
(c) Sleep Stage Scoring Database : 11 Variable, 6 Class, 1236 Event

	45%	30%	15%	5%	FLDF	C4.5
V1	88.5826	88.7799	91.2093	91.1681	93.0796	71.9190
V2	84.4284	86.7234	90.3909	93.1593		
V3	88.8074	90.0377	90.1375	92.5325		
Average	87.2728	88.5137	90.5726	92.2866		

(a)

	45%	30%	15%	5%	FLDF	C4.5
V1	83.6331	85.5836	87.7643	88.9465	92.8085	90.5457
V2	80.4634	83.3710	86.1877	87.5078		
V3	83.2549	87.0320	87.3406	88.0528		
V4	85.0521	86.9948	87.4164	90.2059		
V5	82.4627	86.3853	88.5790	90.3278		
V6	80.3012	86.4611	86.0248	88.8070		
Average	82.5279	85.9713	87.2188	88.9746		

(b)

	45%	30%	15%	5%	FLDF	C4.5
V1	79.6353	83.3312	85.1103	86.0165	87.6280	83.1252
V2	84.0057	84.1607	86.4302	87.3428		
V3	80.5831	84.4522	87.3312	86.8167		
V4	80.0975	84.4503	86.2013	87.1393		
V5	81.0005	82.8289	85.1468	85.8426		
V6	79.3323	83.6702	86.3931	87.4169		
V7	83.0633	83.5973	85.9385	86.6343		
V8	82.9915	84.1544	86.4113	87.2170		
V9	78.7792	82.6970	84.8054	85.6872		
V10	74.8968	81.1854	83.7778	87.0163		
V11	79.3516	82.5671	84.5602	87.1010		
Average	80.3397	83.3723	85.6460	86.8127		

(c)

표 4에서 45%는 훈련 데이터를 임의로 2 부분으로 나누는 비율을 의미한다. 즉, 첫 번째, 선택된 변수에서 소실치를 가지는 퍼센트가 45%이고, 두 번째, 선택된 변수에서 소실되지 않은(완전한) 퍼센트가 55%라는 것이다. 예를 들어 표 4 (c)의 경우에 순서에 따라 V1에서 V11까지 변수를 차례로 선택한 후, 이 변수에 소실치들을 각각 45%, 30%, 15%, 5%씩을 포함시키고 결과를 산출한다. 이 과정을 각 퍼센트와 각 변수에 대해 각각 10번씩을 계산한 후 이 결과 값들의 평균을 보이고 있다. 이들 결과로부터 데이터나 변수에 따라 약간의 차이는 있지만 소실치를 적게 포함하고 있는 실험에서 보다 나은 결과가 나오고 있는 것을 확인할 수 있다. 그러나 다른 면에서는 소실 부분의 비율이 어느 정도 높아지더라도 실험 결과의 성능에 많은 차이를 보이지 않고 있다.

이는 소실치를 채워가는 데이터 확장 과정을 통해 소실치에 의한 정보의 손실을 최소화하는 결과라고 해석된다.

또한 FLDF와 C4.5의 결과는 소실치를 포함하지 않고 산출한 결과로 데이터에 대한 객관적인 정확도를 비교하기 위해 실험하였다. FLDF의 결과가 C4.5의 결과 보다 다소 우수한 것으로 나왔는데 이는 FLDF가 C4.5에서는 고려하고 있지 않는 최적의 투사면을 고려한 결과라고 해석된다. 그림 1은 표 4의 (c)에서 보인 결과를 보다 명확하게 확인할 수 있도록 표시하였다.

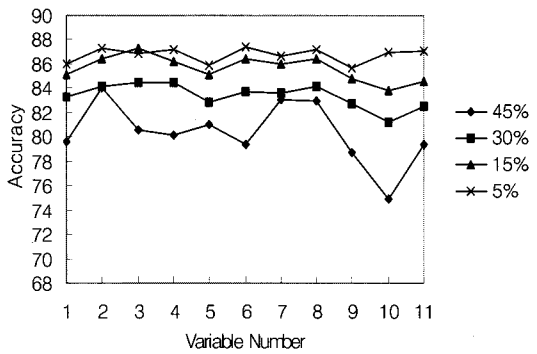

그림 1. 표 4의 (c)에서 소실치 백분율에 따르는 결과들의 비교.
Fig 1. The result comparison according to the missing value percentage, in Table 4 (c).

이 그림에서도 확인할 수 있듯이 변수 값에 따라 차이는 있지만 소실율이 높아질수록 정확도는 떨어진다. 그러나 그 차이는 그다지 크지 않은 것을 확인할 수 있다.

V. 결 론

지금까지 소실치들을 포함하는 데이터에 확장된 표현 방법을 사용한 분류 알고리즘을 소개하였다. 실험 결과들로부터 훈련데이터 안에 포함된 소실치의 퍼센트가 증가할수록 성능은 저하됨을 확인할 수 있었다. 이는 불완전한 데이터가 시스템에 미친 영향으로 피할 수 없는 결과이다. 본 논문에서 제안한 알고리즘은 불완전한 데이터가 포함된 데이터를 가지고 학습을 수행해야 할 때 그 성능저하의 범위를 줄일 수 있는 보상 방법을 사용했으며 이를 학습할 수 있는 새로운 알고리즘을 제시했다는 면에서 의미를 찾을 수 있다.

제안한 알고리즘의 장점으로는 소실치를 보상하는 방법이 간단해서 처리하기 쉽다는 점과 처리하는 과정에서 추가 정보가 필요 없다는 점을 들 수 있다. 또한 이 과정에서 데이터 집합의 각 레코드마다 가중치를 부여할 수 있다는 점이 AdaBoost와 같이 가중치를 이용하는 시스템으로 발전시키는데 유리한 점을 가지고 있다고 볼 수 있다.

전체 데이터들의 평균을 구한다든지 각 부류에 속하는 데이터들 사이의 거리를 구하는 문제 등의 FLDF의 특성에 의해 제안한 알고리즘은 한정된 숫자 데이터의 처리에 국한된다. 앞으로 연속된 값을 가지는 데이터도 처리할 수 있도록 하는 알고리즘의 확장을 연구 중에 있다.

참고문헌

[1] N.H.Nie, C.H.Hull, J.G.Jenkins, K Steinbrenner, Bent D.H, SPSS, 2nd ed. NewYork: McGraw -Hill, 1975.

[2] Roderick J. A. Littile, Donald B. Rubin, Statistical Analysis with Missing Data, 2ED, John Wiley & Sons, 2002

[3] J.M.Robins, A.Rotnitzky, L. P. Zhao," Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data", J. Am. Statist. Assoc. 90, PP 106-121, 1995.

[4] J.H.Friedman,"A recursive partitioning decision rule for non-parametric classification", IEEE Transactions on Computer Science, PP404- 408, 1977.

[5] J. W. Grzymala-Busse,"Rough set strategies to data with missing attribute values", Workshop on Foundations & New Directions in Data Mining, PP19-22, Nov. 2003.

[6] R. J. Hathaway, J. C. Bezdek,"Fuzzy c-means clustering of incomplete data", IEEE Trans. on Systems, Man, Cybernetics-part B: Cybernetics, Vol.31, No. 5, 2001.

[7] M. Kryszkiewicz, "Rough set approach to incomplete information systems", Information Science, Vol.112, PP39-49, 1998.

[8] J. R. Quinlan, "C4.5:Program for Machine Learning," San Mateo, Calif, Morgan Kaufmann, 1993.

[9] A. P. Dempster, N. M. Laird, D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, Vol.B39, PP1-38, 1977.

[10] J. W. Grzymala-Busse,"On the unknown attribute values in learning from examples" ,ISMIS-91, 6th International Symposium on Methodologies for Intelligent Systems, PP368-377, Oct. 1991.

[11] J. Han, M. Kamber, Data Mining : Concept and Techniques, Morgan Kaufmann publishers, 2001.

[12] T.P.Hong, L.H. Tseng, B.C. Chien,"Learning fuzzy rules from incomplete numerical data by rough sets",IEEE international Conference on Fuzzy Syatems, PP1438-1443, 2002.

[13] I. Koninenko, I. Brtko, E. Roskar, "Experiments in automatic learning of medical diagnostic rules", Technical Report, Jozef Stefan Institute, Ljubljana, 1984.

[14] R.Slowinski, J. Stefanowski,"Handling various types of uncertainty in the rough set approach",International Workshop on Rough Sets and Knowlege Discovery, PP366-376, 1993

[15] M. Weiser, "Some Computer Science Issues in Ubiquitous Computing," Com. ACM, Vol. 36, No.7, PP.75-84, July. 1993

[16] Mehmed Kantardzic,"Data Mining:Concepts, Models, Methods, and Algorithms," Wiley- IEEE Press, PP.139-161, 2002.

[17] D. Kim, D. Lee, W. D. Lee, "Classifier using Extended Data Expression," IEEE Mountain Workshop on Adaptive and Learning Systems, July. 2006

[18] J. C. Lee, Y. H. Kim, W. D. Lee, S. H. Lee, "Pattern Classifying Neural Network Based on Fisher's Linear Discreminant", Inter'l Joint Conference on Neural Networks (IJCNN), Vol. 1, PP743~748. July 1992.

[19] J. C. Lee, Y. H. Kim, W. D.Lee, S. H. Lee, "A method to find the structure and weights of layered neural networks",World Congress on Neural Networks, Vol III,July 1993.

[20] Ronny Kohavi, J.R.Quinlan, "Data mining tasks and methods: Classification: Decision-tree discovery," Handbook of data mining and knowledge discovery, Oxford University Press, PP.267-276, 2002.

[21] Thomas G. Dietterich,"An Experimental Com -parison of three methods for constructing emsembles for decision trees: Bagging, Boosting and randomization.", Machine Learning, Vol.40, NO. 2, PP139-157, August, 2000.

저자소개

이종찬(Jong Chan Lee)

1988년 충남대학교(학사)
1990년 충남대학교 대학원(석사)
1996년 충남대학교 대학원(박사)
2006년~현재 : 청운대학교 인터넷학과 교수

※ 관심분야 : 신경회로망, 패턴분류, 정보보호, 데이터압축

이원돈(Won Don Lee)

1979년 서울대학교(학사)
1982년 U. of Illinois 대학원(석사)
1986년 U. of Illinois 대학원(박사)
1987년~현재 : 충남대학교
전기정보통신공학부 교수

※ 관심분야 : 신경회로망, 멀티미디어, 데이터마이닝