보상기법을 이용해 불완전한 데이터를 처리할 수 있는 분류기

이종찬*, 한기선**

요 약

수치 데이터의 분류는 기계 학습에서 중요한 연구 주제이다. 그러나 특정 속성이 손실된 불완전한 데이터는 실세계 응용문제에 일반적이다. 이 문제를 풀기위해, FCM 군집화를 기반으로 하는 데이터 보상 기법이 불완전한 데이터를 추정하기 위해 사용한다. 이 접근 방법은 집락들의 중심 벡터를 계산하고 소속 확률을 결정한 다음, 최적의 해를 찾을 때까지 이 과정을 반복한다. 그리고 이 보상된 데이터를 분류하기 위한 알고리즘이 제안되고 이에 대한 뛰어난 성능을 보인다.

언제나 분류 문제는 두 과정으로 나누어질 수 있는데, 학습 과정과 테스트 과정이다. 분류 문제에서 불완전한 데이터를 취급하는 많은 방법들이 제안되어 왔다. 그러나 대부분은 학습과정 중에 불완전한 데이터의 처리에만 초점을 맞추고 있다. 따라서 분류 과정 중에 나타나는 불완전한 값을 위해서는 대부분의 현 접근법의 알고리즘들이 처리하지 못한다. 학습과 분류 과정 모두에서 불완전한 데이터를 해결하기 위한 방법은 중요하며 실생활 문제에 적용하기 위해서는 필요하다.

Classifier Capable of Handling Incomplete Data using Reparation Technique

Jong Chan Lee*, Ki-Sun Han**

ABSTRACT

Classification of the numerical data is a very important research topic in machine learning. But the incomplete data in which certain features are missing, is very common in real world applications. For solving this problem, a data reparation approach base on the Fuzzy c-Means(FCM) clustering is used to estimate the incomplete data. This approach calculates the centroid vectors of the clusters and then determined the membership probability, and repeat this process until the optimum solution is found. Then a new method is proposed to classify the repaired data and it has an outstanding performance.

Usually, classification problem can be separated into two phases: learning phase and classification phase. Many methods dealing with incomplete data in classification problem have been proposed, but most of them only focus on the processing of handling incomplete data in learning phase. For the incomplete value appearing in the classification phase, almost all of the current approaches can not work. So handling incomplete data in both learning and classification phase is important and necessary to be applied for solving the real world problems.

Key Words : Incomplete data, Reparation technique, UChoo, Data Clustering, Classification Problem

---

* 청운대학교 인터넷학과 (✉jclee@chungwoon.ac.kr)
** 강동대학교 방송영상미디어과
· 제1저자(First Author) : 이종찬 · 교신저자(Correspondent Author) : 이종찬
· 접수일(2013년 3월 28일), 수정일(1차 : 2013년 4월 10일), 게재확정일(2013년 4월 18일)

I. 서 론

특정 속성(attribute) 벡터에서 하나이상의 속성 값이 손실된(missing) 불완전한(incomplete) 데이터는 군집화(clustering), 컴퓨터 비전, 생물학적 시스템, 분류(classification) 시스템, 원격 탐사(remote sensing) 등의 폭 넓은 분야에서 발생한다. 여러 분야의 불완전한 데이터는 분류 모델을 학습하는 질을 일반적으로 저하시켜, 해결해야 하는 문제로 대두되었다. 이로 인해 불완전 데이터를 처리하는 분류 문제는 기계 학습(machine learning) 분야에서 중요한 연구 주제이며, 수치 데이터를 처리하기 위한 많은 알고리즘들이 고안되었다[1][2].

손실 데이터의 보다 직관적인 예를 들면, 만일 X1=(1,2,3,4)라 할 때 (?,2,3,4)는 25%의 불완전한 데이터이고, (1,?,?,4)는 50%의 불완전한 데이터라고 할 수 있다. 이러한 불완전 데이터의 처리에 대한 연구는 크게 2가지로 나누어진다. 첫째, 손실 값을 포함하는 즉, 불완전한 데이터를 무시하고 처리하는 방법으로, 이 연구에 대표적인 것이 Quinlan[3], Fridman[4]에 의해 이루어졌다. 둘째, 적절한 대체 방법을 가지고 손실된 값을 보충해 나가는 방법으로 Hathaway[5], Dempster[6], Han[7], Hong[8]의 연구가 대표적이다. 이 들 중 본 논문에서 제안하는 방법은 두 번째 방식을 택하고 있다. 다시 말해 불완전한 데이터가 학습이나 테스트 과정에 입력되었을 때 첫째 손실된 데이터 값을 보상해 놓고, 둘째 이를 학습할 수 있도록 UChoo 알고리즘을 이용하여 학습한 후 이를 테스트하는 과정으로 이루어진다.

첫째, 손실 데이터의 보상 과정을 위해 FCM(Fuzzy c-Means) 군집화 알고리즘을 사용하는데, 이는 집락의 경계가 애매하거나 데이터 요소가 어느 집락에 속한다고 명확하게 구분하기 어려운 경우에 소속 함수(membership function)를 0과 1사이의 연속적인 값으로 표현하는 방법이다. 이 방법은 Dunn에 의해 1973년에 개발되었고 Bezdek에 의해 1981년 개선된 버전이 발표되었다. 이를 바탕으로 Hathaway와 Bezdek은 불완전한 데이터의 군집화 문제를 해결하기 위한 OCS(Optimal Completion Strategy) FCM 알고리즘[5]을 개발하였다. 본 논문은 불완전 데이터의 손실된 속성 값을 예측해 보상하기 위해 이 OCS 알고리즘으로부터 유도된 함수를 사용한다.

둘째, 보상된 데이터의 학습을 위해 UChoo 알고리즘[10-15]이 이용되는데, 이는 입력 벡터에 포함된 불완전한 데이터의 학습을 위해 확장된 데이터[10][11] 표현 방법을 사용하는데, 이에 맞도록 C4.5 알고리즘을 변형한 알고리즘[10]이다. [10][13]에서는 단순히 손실된 속성 값을 카디너리티 정보를 이용해 채워가는 방식을 사용하였고, 학습 중에 데이터 벡터들의 가중치를 고려하며 학습하는 방안과 규칙을 산출하는 방법을 제시하였다. 이 후에 이 알고리즘을 이용해 속성의 수가 서로 다르게 수집된 데이터(A 데이터는 4개의 속성, B 데이터는 3개의 속성)를 규칙을 이용해 함께 학습할 수 있음을 보였다[12]. 이 논문에서는 A 데이터를 이용해 학습을 하여 규칙을 산출 한 후, 이 규칙과 새로운 데이터 B가 서로 같이 학습될 수 있음(규칙+새로운 데이터)을 보였다. 다음으로 손실 값을 보상하는 방법을 좀 더 통계적인 방법으로 개선한 알고리즘이 제안[14]되었다. 여기서는 Hathaway[5]가 제시한 4개의 알고리즘(WDS, PDS, OCS, NPS)들 중에 OCS 알고리즘을 이용해 손실 값을 채운 후 UChoo를 이용해 이를 학습하였다. 이 논문의 주안점은 OCS의 비용 문제를 MFA 알고리즘[15][16]을 이용하여 최소 값을 구하려 하였다는 것이다. 본 논문은 이에 대해 OCS의 최소 에너지(목적함수)를 가지는 방법으로 FCM 알고리즘을 이용하여 반복적으로 비용(cost)를 개선해 나가는 방식을 사용하여 값을 구해 손실 값에 채워나간다.

분류 문제는 학습 과정과 테스트 과정으로 나뉘어 처리된다. 분류 문제에서 불완전한 데이터를 처리하기 위해 많은 방법들이 제안되었는데, 대부분은 학습과정 중에 불완전한 데이터의 처리에 초점을 맞추고 있다. 따라서 테스트 과정 중에 나타나는 불완전한 데이터는 대

부분의 알고리즘들이 처리하지 못한다. 이를 해결하기 위한 방법은 중요하며 실생활 문제에 적용하기 위해서는 필요한 과정이다. 이를 위해 본 논문은 불완전한 데이터의 속성 값을 예측하기 위해 유도된 함수를 사용하고, 이 데이터들을 분류하기 위해 적합한 분류 알고리즘을 제안한다.

II. UChoo 알고리즘

분류기는 데이터의 속성 값들이 주어졌을 때 학습과정을 통해 이에 대한 각각의 부류들을 예측하기 위한 목적으로 사용하는데, 이러한 분류기에서 가장 널리 이용되는 것이 결정 트리기법이다. 결정트리를 사용하는 알고리즘들 중에 C4.5[3]는 최소의 메모리를 가지고 빠르고 좋은 성능을 산출해 내는 것으로 알려져 있다.

C4.5 알고리즘에서는 각 속성(attribute, A)들에 대해 (1) 식과 같은 정보 이득(gain_ratio) 함수를 계산한다. 정보 이득 함수는 해당 속성 값에 따라 분류할 때 부류들이 얼마나 잘 나누어지는지를 측정하는 값이며, 이들 중 가장 큰 정보 이득을 가진 속성을 선택하여 이 속성 값에 따라 하부 노드로 분기한다. 이와 같은 방법을 모든 부류가 분류될 때까지 반복하면 데이터 집합에 따르는 의사 결정 트리를 얻을 수 있다.

$$Gain\_ratio(A) = Gain(A) / Split\_info(A) \quad (1)$$

(1) 식을 정의하기 위한 변수들은 다음과 같다.

- S : 각 노드에서의 데이터 집합
- $|S|$: S의 데이터 개수
- S_{A_j} : 집합 S 중에 속성 A에서 속성 값 j를 가지는 데이터 S의 부분 집합
- $|S_{A_j}|$: S_{A_j}의 데이터 개수
- k : 부류의 개수
- 부류 : $C_1, C_2, \ldots, C_k$
- $freq(C_i, S)$: S에서 부류가 C_i인 레코드의 개수

속성 A의 카디너리티(cardinality)가 n이라 할 때 속성 A의 속성 값들의 집합은 $\{O_{A_1}, O_{A_2}, \ldots, O_{A_n}\}$이다. 만약 A가 연속 값을 가진다면 문턱 값(threshold) Z에 따라 둘로 나누는 기법을 사용할 수 있다. (1) 식에서 Gain(A) 함수는 속성 A를 선택했을 때 데이터의 부류가 분류되는 정도를 나타내는 엔트로피 함수를 사용하여 (2) 식과 같이 나타낼 수 있다.

$$Gain(A) = info(S) - info_A(S) \quad (2)$$

$$info(S) = -\sum_{i=1}^{k} (freq(C_i,S)/|S|) \cdot \log_2 (freq(C_i,S)/|S|)$$

$$info_A(S) = \sum_{j=1}^{n} (|S_{A_j}|/|S|) \cdot info(S_{A_j})$$

$$info(S_{A_j}) = -\sum_{i=1}^{k} (freq(C_i,S_{A_j})/|S_{A_j}|) \cdot \log_2 (freq(C_i,S_{A_j})/|S_{A_j}|)$$

(1) 식에서 Split_info(A) 함수는 속성 A에서 속성 값들의 개수로 인해 측정에 영향이 가지 않도록, 즉 결정 트리가 균등하게 배분되도록 일반화 해주는 역할을 담당한다. 이에 대한 수식이 (3) 식에 나타나 있다.

$$Split\_info(A) = -\sum_{j=1}^{n} (|S_{A_j}|/|S|) \cdot \log_2 (|S_{A_j}|/|S|) \quad (3)$$

UChoo는 C4.5에 확장된 데이터를 처리할 수 있도록 개발된 알고리즘[10]이다. 확장된 데이터 표현이라 함은 데이터 레코드 마다 중요도를 나타내는 가중치를 가지며 속성 값들은 0과 1 사이의 확률 값으로 나타낼 수 있다. C4.5에서 확장된 데이터 표현 방식을 가지는 데이터를 가지고 학습할 수 있도록 변형이 필요하다. 여기서 C4.5와 달라진 정의는 다음과 같다.

- 부류의 소속 값 : $C_1(m), C_2(m), C_3(m), \cdots, C_{k-1}(m), C_k(m)$

Ci(m)는 m번째 사건이 Ci부류에 속한 정도를 나타

낸다. 여기서 i는 부류 값이고, $\sum_{i=1}^{k} C_i(m)=1$이다.

- 속성의 소속값 : $O_{A_1}(m), O_{A_2}(m), \ldots, O_{A_{n-1}}(m), O_{A_n}(m)$

$O_{A_j}(m)$는 m번째 사건의 속성 A에서의 속성 값 j가 가지는 값을 말한다. 여기서 $\sum_{j=1}^{n} O_{A_j}(m)=1$ 이다.

- Weight(m,S) : 집합 S에서 m번째 사건의 가중치 값
- freq(C_i,S) : 집합 S안에서 부류 C_i에 속해있는 사건들의 개수인데 이 경우에는 (4)식과 같이 나타내 진다.

$$freq(C_i,S) = \sum_{m=1}^{|S|} Weight(m,S) \cdot C_i(m) \qquad (4)$$

마찬가지로 freq(Ci,SAj)는 SAj안에서 부류 Ci에 속해있는 사건들의 개수인데 이 경우에는 (5)식과 같이 나타내 진다.

$$freq(C_i,S_{A_j}) \sum_{m=1}^{|S|} Weight(m,S) \cdot C_i(m) \cdot O_{A_j}(m) \qquad (5)$$

$|S_{A_j}|$: S_{A_j} 사건들의 개수. 이 경우에도 위와 마찬가지로 집합 S_{A_j}에 속해 있는 사건들 각각의 속성의 소속 값 $O_{A_j}(m)$에 Weight(m,S_{A_j})을 곱한 후 그것을 모두 더하여 아래와 같이 계산한다.

$$|S_{A_j}| = \sum_{j=1}^{|S_{A_j}|} Weight(m,S_{A_j}) \cdot S_{A_j}(m)$$

따라서 새로 정의된 값들에 의하여 기존의 엔트로피 식을 사용하여 그 노드에서 가장 큰 Gain Ratio를 가진 속성을 결정 할 수 있다.

III. 손실된 속성의 처리

불완전한 데이터에서 하나의 사건에 속성 값이 손실된 경우, 이 손실된 속성 값을 보상(reparation)해 채워가는 방식을 사용한다. 이 보상 기법으로 본 논문에서는 Hathaway와 Bezdek가 제안한 FCM 군집화의 OCS 알고리즘[5]을 사용한다. 이에 대한 간략화 된 설명을 위한 정의되는 기호들은 다음과 같다.

- $X_i=(X_{i1},\ldots,X_{ik},\ldots,X_{ip})$, X_i는 i 번째 p 차원의 데이터 벡터이고, 0 < i < n.
- $X=(X_1,\ldots,X_i,\ldots,X_n)$, X는 완전한 데이터 집합.
- $X_L=(X_i \in X)$, X_i는 불완전한 데이터.
- $Y_i=(Y_{i1},\ldots,Y_{im},\ldots,Y_{ig})$, Y_{im}은 i 번째 데이터가 m 번째 집단에 속할 확률이다. 이 확률은 멤버십(membership)으로 정의되며, 집단의 수는 g이다.
- $C_m=(C_{m1},\ldots,C_{mp})$, m=1,...,g, C_m은 m번째 집단의 중심.

OCS는 집단의 중심 C_{mk}와 멤버십 Y_{im}을 가지고, (6) 식의 $E(\cdot)$목적함수를 반복적으로 최적화하는 과정이다.

$$E(g,r,Y)=\sum_{i=1}^{n}\sum_{m=1}^{g}(Y_{im})^r\sum_{k=1}^{p}(X_{ik}-C_{mk})^2 \qquad (6)$$

$$C_{mk}^{t+1}=\sum_{i=1}^{n}((Y_{im}^{t+1})^r \times X_{ik}) / \sum_{i=1}^{n}(Y_{im}^{t+1})^r \ ,$$

m=1,...,g, k=1,...,p

$$Y_{im}=\frac{D_{im}^{1/(1-r)}}{\sum_{i=1}^{g} D_{im}^{1/(1-r)}}$$

$$D_{im}=\| X_m - C_i^t \|, 1 \leq i \leq g, 1 \leq m \leq n$$

$\| \cdot \|$는 노름이며, 제약식으로 $Y_{im} \geq 0$, $\sum_{i=1}^{n} Y_{im}=1$이다.

여기서 r=1일 때 $E(\cdot)$목적함수를 최소화하면 얻어지는 결과가 일반적인 배타적 집락분석(crisp partition)인 k-mean 방법을 말한다. 또한 r>1이면 애매성을 표시하는 퍼지 집락화를 의미한다. 이때 보통 사용하는 r=2이며 이는 분석적인 결과가 아니고 1 이상의 값들에 대한 모의실험 결과 r=2 근처에서 분석 결과가 수렴되는 것으로 나

타나기 때문에 일반적으로 받아들여지고 있다.

각 반복 과정에서 손실 값은 (7)식을 사용해 계산된다.

$$X_{mk}^{t+1} = \sum_{i=1}^{q}((Y_{im}^{t+1})^r \times C_{ik}) / \sum_{i=1}^{q}((Y_{im}^{t+1})^r \qquad (7)$$

OCS 알고리즘은 다음과 같다.

과정 1) Y_i를 초기화 한다.

과정 2) $E(\cdot)$목적함수를 계산한다.

$$E(g,r,Y)^{t+1} = \sum_{i=1}^{n}\sum_{m=1}^{q}(Y_{im}^{t+1})^r \sum_{k=1}^{p}(X_{ik} - C_{mk}^{t+1})^2$$

$$C_{mk}^{t+1} = \sum_{i=1}^{n}((Y_{im}^{t+1})^r \times X_{ik}) / \sum_{i=1}^{n}(Y_{im}^{t+1})^r \ ,$$

m=1,...,g, k=1,...,p

$$Y_{im} = \frac{D_{im}^{1/(1-r)}}{\sum_{i=1}^{g} D_{im}^{1/(1-r)}}$$

과정 3) 만일 $\| Y^{t+1} - Y^t \| < \epsilon$ 이면 알고리즘을 멈추고, 그렇지 않으면 과정 4)으로 간다.

과정 4) $\forall X_{mk} \in X_L$ 에 대해

$X_{mk}^{t+1} = \sum_{i=1}^{q}((Y_{im}^{t+1})^r \times C_{ik}) / \sum_{i=1}^{q}((Y_{im}^{t+1})^r$를 계산하고, 과정 2)로 간다.

OCS 알고리즘의 수행의 예를 들면 <표 1>의 (a)는 보상이 이루어진 후 (b)와 같이 된다. 이러한 과정은 훈련 데이터와 테스트 데이터에 똑 같이 적용될 수 있다. 따라서 테스트 데이터가 불완전한 데이터라 하더라도 처리할 수 있다. <표 1>의 (b)가 C4.5에서 사용되는 일반적인 훈련 데이터의 예라고 한다면, <표 1>의 (b)를 UChoo의 표현 방식[10]으로 변환이 필요하다. 여기서 각 속성 값들은 0과 1사이의 확률 값으로 채운다는 것이 확장된 데이터 표현 기법의 기본이다. 각 사건은 가중치 값을 가지게 된다. 이것은 그 사건이 얼마만큼의 중요도를 가지는가를 나타낸다. 일반적인 레코드의 중요도가 1이라고 보았을 때, 가중치 20인 사건은 다른 가중치 1인 레코드의 20개에 해당하는 중요도를 가진 사건이라는 뜻이다. 따라서 전체 사건의 개수와 레코드들의 개수는 서로 다른 값일 수 있다.

이러한 확장된 데이터 표현을 사용하는 UChoo 분류기는 데이터의 일부나 부류가 손실된 불완전한 데이터를 처리할 수 있다는 장점을 가진다. 이는 일부분의 손실에 대해 이 사건을 학습에서 제외하는 방법에 비해 정보의 손실을 줄임으로서 보다 정확한 학습을 할 수 있도록 한다.

표 1. 훈련 데이터의 예
Table 1. Training data set in example
(a) 손실된 속성 값을 가지는 원 데이터
(a) Origin data set with missed attribute values.

사 건	속성 1	속성 2	속성 3	부류
1	5.1	?	1.3	2
2	?	3.5	1.4	2
3	4.7	3.2	?	1
4	4.7	3.0	?	1
5	5.1	?	1.3	1
6	?	3.5	1.4	1
7	4.9	?	1.3	2

(b) (a) 데이터의 보상된 결과
(b) Repaired result of (a) data set

사 건	속성 1	속성 2	속성 3	부류
1	5.1	3.5	1.3	2
2	5.1	3.5	1.4	2
3	4.7	3.2	1.3	1
4	4.7	3.0	1.4	1
5	5.1	3.0	1.3	1
6	4.7	3.5	1.4	1
7	4.9	3.5	1.3	2

Ⅳ. 실 험

제안된 방법의 결과를 평가하기 위해 UCI 기계 저장소[17](machine repository)의 데이터 집합들을 사용하였다. 데이터 집합에서 주어진 백분율만큼을 무작위로 골라 지움으로서 불완전한 데이터를 만들었다. 이 과정은 다음과 같이 정하였다.

1. 각 사건에서 최소한 하나의 속성 값은 남아 있다.
2. 각 속성에서 최소한 하나의 속성 값은 남아 있다.

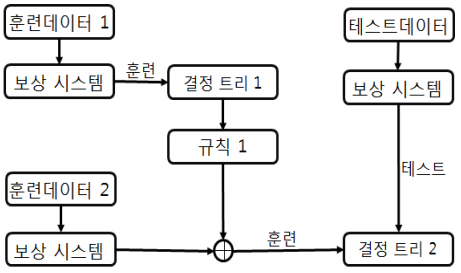

그림 1 불완전 데이터의 분류 과정
Fig.1 The process of classifying incomplete data

실험 방법으로 10-fold cross validation을 사용하였다. 즉, 데이터들을 10개의 블록으로 나눈 다음, 9개의 블록은 훈련을 위해, 나머지 1개의 블록은 테스트를 위해 사용한다. 이 과정을 10번 실행하여 결과들의 평균을 사용한다.

실험은 두 가지 방법으로 나누어지며, 그 과정이 <그림 1>에 나타나 있다. 첫 번째로 테스트 데이터는 완전한 데이터를, 훈련 데이터는 불완전한 데이터를 사용한다. <표 2>에 이에 대한 결과가 나타나 있다. 여기서 C4.5는 불완전 데이터를 무시하고 훈련을 해 산출한 결과이고, "제안된 분류기"는 보상 시스템으로 손상된 데이터를 보상한 후 결정트리 1에 테스트 데이터를 입력해 산출한 결과이다. 불완전한 데이터의 비율에 따라 차이는 있으나 결과로부터 제안된 분류기의 결과가 보다 우수한 것으로 나타났다. 이는 보상의 효과로 해석된다. 그리고 표 2의 "규칙 + 데이터"는 다음과 같이 만들어 졌다. 우선 훈련 데이터, 9개의 블록을 하나로 합한 후 이를 다시 임의로 2개의 블록(훈련데이터 1, 훈련데이터 2)으로 나눈다. 그리고 하나의 블록(훈련데이터 1)은 규칙(규칙 1)을 얻기 위한 훈련을 하기 위해 사용하고, 나머지 하나의 블록(훈련데이터 2)은 이 규칙에 추가되는 데이터로 사용한다.

마지막으로 이렇게 합해진 규칙(규칙 1)과 새로운 데이터(훈련데이터 2)로 학습하여 결정트리(결정트리 2)를 산출한다. 그리고 이 결정트리에 테스트 데이터를 입력하여, 최종 테스트 결과를 산출한다. 결과로부터 비록 규칙을 이용해 데이터를 산출하고, 이를 새로운 데이터와 결합하는 방법을 사용하였으나 결과에 커다란 영향을 미치지 않았음을 보이고 있다.

두 번째 실험에서는 훈련과 실험 데이터로 불완전한 데이터가 사용되었다. 우선 10%의 불완전한 데이터가 포함된 훈련데이터가 사용되고, 여러 불완전한 데이터 비율로 테스트 데이터를 만든 후 실험하였다.

<표 3>은 이에 대한 결과를 보이고 있다. 첫 번째와 두 번째 실험에서 공통적으로 C4.5의 결과에 비해 "제안된 분류기"와 "규칙+데이터"의 결과가 다소 우수하였다. 특히 "규칙+데이터"는 학습에 사용된 본래의 데이터가 분실되었거나 손상된 경우에도 이를 통해 산출된 규칙만 남아 있다면 새로운 데이터와 결합하여 성능에 커다란 영향을 미치지 않고 유용하게 사용될 수 있음을 보인다.

본래의 데이터에 비해 규칙이 비교가 안 될 정도로 간단하다. 따라서 저장 공간이 부족한 경우, 규칙만을 보존하여도 어느 정도 변화되는 환경에 적응하도록 할 수 있다고 볼 수 있다.

표 2. 완전한 테스트 데이터를 가지고 IRIS 데이터를 분류한 에러 비율(%)
Table 2. Error rate(%) of classifying IRIS data with complete testing data

불완전한 훈련데이터 비율	C4.5	제안된 분류기	규칙 + 데이터
0%	4.7	4.7	5.3
5%	5.69	4.7	5.3
10%	6.16	5.3	6.0
20%	7.7	5.3	6.7
30%	12.09	7.3	7.9

표 3. 10%의 불완전한 데이터가 포함된 IRIS 데이터를 분류한 에러 비율(%)
Table 3. Error rate(%) of classifying IRIS data with 10% incomplete training data

불완전한 테스트데이터 비율	C4.5	제안된 분류기	규칙 + 데이터
0%	4.7	4.7	5.3
5%	10.74	6.0	9.3
10%	13.46	8.6	12
20%	19.7	14.6	15.9
30%	29.4	19.3	21.6

V 결 론

본 논문은 불완전한 데이터를 포함한 데이터는 학습 결과를 매우 저조하게 한다는 점에서 이를 해결하는 방안을 제안하였다. 그 과정으로 첫째 FCM을 기반으로 하는 OCS를 이용해 손실된 데이터를 보상하고, 둘째 보상된 데이터를 각각의 데이터 레코드 마다 다른 가중치를 가질 수 있고 학습 후에 산출된 규칙을 데이터 형식으로 변형이 가능하여 규칙과 새로운 데이터를 결합하는데 유용하게 개발된 알고리즘인 UChoo를 이용해 학습하는 방안을 제시하였다. 그 결과를 테스트하는 실험에서 제안된 방법의 결과가 우수함을 보았다.

또한 알고리즘의 전개 과정에서 부수적으로 규칙의 개선과 중요도 선호 문제를 고려하였다. 규칙 개선 문제는 이미 완성된 규칙과 새로운 데이터가 결합될 수 있도록 하며, 규칙에 새로운 데이터가 보강되어 새로운 정보가 산출될 수 있음을 보였다. 또한 중요도 선호 문제는 서로 다른 환경에서 수집된 데이터 간의 중요도를 학습에서 반영할 수 있는 방법을 제안하였다.

제안된 방법에서 사용된 데이터 보상 방법은 훈련 데이터는 물론 테스트 데이터에도 적용이 되어, 일반적으로 불완전한 데이터를 처리하는 알고리즘들이 고려하지 못한 테스트 데이터에 불완전한 데이터가 포함된 경우에도 자연스럽게 처리할 수 있음을 보였다.

참고문헌

[1] P. N. Tan, M. SteinBach, V. Kumar, "Introduction to data mining", 2005

[2] M. Kantardzic, "Data Mining : Concepts, Models, Methods, and Algorithms", Wiley-IEEE Press, 2002.

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

[4] J. W. Friedman, "A recursive partitioning decision

rule for non parametric classification", IEEE Transaction on Computer Science, pp404-408, 1977
[5] R. J. Hathaway, J. C. Bezdek, "Fuzzy c-means clustering of incomplete data", IEEE Transaction on systems, Man and Cybernetics-part B: Cybernetics, Vol.31, No. 5, 2001
[6] 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
[7] J. Han, M. Damber, "Data mining : Concept and Techniques", Morgan Kaufmann Publishers, 2001
[8] T.P.Hong, L.H.Tseng, B.C.Chien, "Learning fuzzy rules from incomplete numerical data by rough sets", IEEE international conference on Fuzzy Systems, pp1438-1443, 2002
[9] J.W.Grzymala-Busse, "Incomplete Data and Generation of Indiscernibility Relation, Definability, and Approximations", Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing Lecture Notes in Computer Science, Vol.3641, pp244-253, 2005
[10] D. H. Kim, D. H. Lee, W. D. Lee, "Classifier using Extended Data Expression", IEEE Mountain Workshop on Adaptive and Learning Systems. pp.154-159, July 2006.
[11] J.C.Lee, D.H.Seo, C.H.Song, W.D.Lee, "FLDF based Decision Tree using Extended Data Expression", The 6th International Conference on Machine Learning and Cybernetics, Hong Kong, pp3478-3483, Aug. 2007.
[12] D.H.Seo,C.H.Song,W.D.Lee, "A Classifier Capable of Handling New Attributes", Computational Intelligence and Data Mining, IEEE Symposium on CIDM, pp323-327, 2007.
[13] J.Wu, Y.S.Kim, C.H.Song, W.D.Lee,"A New Classifier to Deal with Incomplete Data", 9th ACIS Internetional Conference on Software Engineering, Artificial Intelligence, Networking, pp105- 110, 2008.
[14] J. Wu, D.H.Seo, C.H.Song, W.D.Lee, "A New Classifier for Numerical Incomplete Data", IEEE international conference on Intelligence and Security Informatics, pp273-274, 2008.
[15] J. Wu, C.H.Song, J.M.Kong, W.D.Lee, "Extended Mean Field Annealing for Clustering Incomplete Data", International Symposium on Jeonju, KOREA, pp8-22, 2007.
[16] Y.H.Kim, J.C.Lee, S.H.Lee, "A Clustering Method with an Ambiguous Class", 5th Conference of the International Federation of Classification Societies(IFCS), pp282-284, 1996.
[17] http://www.ics.uci.edu/~mlearn/MLRepository.html, UCI Machine Learning, 1989

감사의 글

본 논문은 청운대학교 학술연구조성비 지원에 의해 연구되었습니다.

저자소개

이 종 찬(Lee, Jong Chan)

1988년 충남대학교 계산통계학과(학사)
1990년 동대학원(석사)
1996년 충남대학교 대학원 전산학과(박사)

2006년~현재 : 청운대학교 인터넷학과 교수
※ 관심분야 : 신경회로망, 패턴분류, 정보보호, 데이터압축

한 기 선(Han, Ki Sun)

1988년 충남대학교 계산통계학과(학사)
1990년 동대학원(석사)
2004년 충남대학교 대학원 정보통신공학과 박사 수료

1998년~현재 : 강동대학교 방송영상미디어과 교수
※ 관심분야 : 영상처리, 정보통신, 데이터압축