한국지능시스템학회 논문지 2008, Vol. 18, No. 5, pp. 638-641

신경회로망을 이용한 분류모형 개발

Development of Classification Model Using Neural Network

박광박* · 박영만** · 황승국***

Kwang-Bak Park*, Young-Man Park** and Seung-Gook Hwang***

* 경남대학교 응용수리학부 교수

** 경남대학교 경영학부 교수

*** 경남대학교 정보통신공학과 교수

Division of Applied Mathematics and Physics, Kyungnam University, Korea

요 약

본 논문에서는 데이터를 사전처리 한 후 Fuzzy TAM을 이용하여 분류하는 방법을 개발하였다. 사전 처리 방식은 category형 특성인 경우는 그 특성을 이용하여 문제를 분해시키고, 계량형 특성의 경우는 클래스별 영역을 설정하고 겹치지 않는 특성 영역이 있다면 그 영역의 자료를 고정시켜 분류에서 제외시킨다. 이러한 사전 처리를 한 후 Fuzzy TAM을 이용하여 분류를 수행한다.

Abstract

In this paper, a model to classify the method using the fuzzy TAM with preprocessing of data was developed. The preprocessing method can be divide the problem using the characteristics in the case of category type factor. In case of continuous type factor, if there was exist factor's range which is not overlapping by class, the data belong to the range was fixed and eliminated in classification. After these preprocessing of data, classified operation of Fuzzy TAM is performed.

Key Words : Classification Model, Fuzzy TAM

1. 서 론

어떤 자료들을 몇 개의 특성을 이용하여 분류하는 작업은 많은 문제에서 발생할 수 있다. 예로 여러 개의 신체적 측정치를 가지고 특정 질병의 진단, 재무 상태 및 여러 주변 상황을 가지고 기업의 평가, 개인의 신용 평가, 인간의 유형에 분류 등에서 발생한다.

이러한 분류 또는 진단에 있어서의 분류 모형은 간단하면서도 효율적인 측면과 얼마나 잘 분류할 수 있는가가 중요하다. 그러나 대부분의 경우 정확성을 따지면 분류 규칙이 많아지면서 모형이 복잡해지고, 단순화된 모형은 분류 정확성이 낮아진다.

분류 모형을 형성하는 데는 우선 클래스를 구분할 수 있는 특성들은 어떤 것들이 있는가를 찾는 작업과 그 특성 중에서 과연 어떤 것이 그 클래스를 가장 잘 구분할 수 있는가를 찾아야 한다. 전자의 문제는 각 케이스에 종속된 것으로 일반적인 부분이 없는 문제이다. 후자의 문제 즉 클래스의 분류와 관련 있다고 생각되는 많은 수의 특성 중에서 가장 영향력이 큰 특성 또는 특성 군을 찾는 작업이다.

자료들을 각각 동질적인 그룹으로 나눌 수 있도록 특성들을 효율적으로 선택하는 방법으로 χ^2 통계량을 이용한다던지 엔트로피 지수, 상관계수를 이용하는 방법 등 여러 방법이 존재한다. 그러나 근본적으로는 그 특성을 이용한다면 어느 정도 정확하게 분류할 수 있는가의 정도로서 그 특성을 평가하는 방법일 것이다.

본 연구에서 제시하고자 하는 분류 방법은 자료들을 사전 처리를 통해 어느 정도 정리한 후에 분류 모형에서 그런대로 좋은 성과를 보이고 있는 Fuzzy TAM[1-4]을 이용하여 분류한다는 것이다. 자료의 전처리는 명목적 특성이 있다면 그 특성들을 이용하여 자료를 분해하여 각각 처리하고, 계량적 특성들에 대해서는 개별 클래스별로 영역을 설정하고 겹치지 않은 영역이 있다면 그 영역에 속하는 자료를 고정시켜 분류에서 제외함으로써 전체 자료 수를 줄인다.

2. 분류모형

분류 모형에 필요한 특성의 수가 적으면 적을수록 분류 모형은 간단해진다. 물론 적은 수의 특성을 가지고 만족할 만한 수준의 분류가 가능하다면 괜히 복잡한 모형을 만들 이유는 없을 것이다. 그러므로 적당수의 특성을 이용해 효과적인 분류모형을 만들기 위해서는 분류에 적합한 특성들을 골라내야 한다.

특성 선택은 바람직하지 못한 특성들을 추론이 시작되기

접수일자 : 2008년 8월 20일
완료일자 : 2008년 10월 10일
본 연구는 2007학년도 경남대학교 학술진흥연구비 지원에 의하여 이루어졌음.

전에 추출해가는 추출접근법(filter approach)과 실제 추론 방법을 적용해 가면서 영향력이 큰 특성들을 선택해가는 보자기 접근법(wrapper approach)으로 나눌 수 있다.

Hall[5]의 CFS(Correlation-based Feature Selection)은 클래스와 특성들을 명목형 형식으로 변환한 후 특성과 클래스, 특성간의 상관을 이용하여 만든 특성평가함수로 특성을 평가 후 가장 큰 값을 갖는 특성을 순차적으로 추가했다. 김효중,박종신[6]은 명목형 간에는 χ^2통계량에 대한 유의확률, 명목-연속형 간에는 크루스칼-왈리스 χ^2통계량의 유의 확률, 연속-연속형 간에는 스피어만 상관 통계량의 유의확률을 이용한 특성평가함수를 이용하여 특성을 추가하고 유전 알고리즘을 통해 분류 모형을 완성했다.

상관을 사용하든 유의확률을 사용하든 그 목적은 분류에 영향력이 큰 특성을 찾는 것이 목적이다. 그것은 즉 그 특성을 이용할 경우 다른 특성을 이용할 경우보다 분류 정확성이 높을 가능성이 큰 것을 의미한다.

2.1 명목형 특성 처리

명목형 자료는 수치적인 의미가 없기 때문에 일반적인 수리모형에서 처리하기에는 조금 문제가 있다. 본 연구에서는 이와 같은 명목형 특성을 이용하여 자료를 나누는데 사용한다.

명목형 특성과 클래스 간의 crosstab 분석을 한다면 그 관련성의 정도는 χ^2 통계량으로 나타난다. χ^2통계량을 기준으로 몇 개의 특성을 선택하여 각 클래스 수준에 많이 나타나는 부분으로 묶어 문제 자료를 나눈다.

이렇게 나눠진 문제는 특정 클래스 값이 많이 나타나는 문제로 바뀌어 한 클래스의 특성이 많은 자료들의 집합으로 나타나므로 분류 작업이 용이하게 처리될 수 있다.

2.2 계량형 특성 처리

분류에 높은 연관성을 갖는 특성들이 갖는 값들은 어느 정도 각 클래스의 수준 당 특정 범위 내에 움직이는 경우가 많다. 그 특성이 갖는 영역에서 클래스의 각 경우들의 자료가 중복되어 나타난다면 그 특성을 이용하여 각 그룹을 분류하기는 어렵다.

대부분의 자료는 평균을 중심으로 3σ 내에 분포하고 있다. 특이한 경우를 제외한다면 한 특성에 대해서 특정 클래스에 대한 영역을 평균을 중심으로 3σ 내로 보면, 클래스의 각 수준 간 중복되는 영역과 중복되지 않은 영역(고유 영역)으로 나눌 수 있다.

중복되는 영역에서야 분류를 위해 다른 처리가 필요하지만, 중복되지 않은 고유 영역에 대해서는 미리 그와 같은 자료를 고정시켜 분류 작업에서 제외시키는 것이 분류 모형을 간단화 시킬 수 있을 것이다. 분류의 정확성을 조금 높이기 위해 특이 케이스를 포함시켜 분류 모형을 복잡하게 만들 이유는 그렇게 많지 않다.

그림 1에 예로 한 특성에 대해 클래스의 각 수준에 따라 자료를 3개 그룹으로 분류할 수 있을 때 자료들의 분포 영역을 보여주고 있다. 고유 영역1에 속한다면 그룹 1로 분류를 하고, 고유 영역2에 속하는 자료는 그룹3으로 고정시킨다면 나머지 중복 영역의 부분에 속하는 자료들만 고려하여 분류모형을 형성하면 된다.

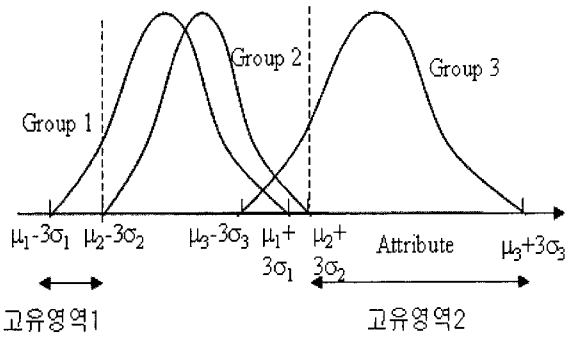

그림 1. 각 클래스에 대한 자료분포의 예
Fig. 1 Example of data distribution for each class

2.3 분류 모형

본 연구에서의 분류 모형은 명목형 특성이 있다면 문제의 자료를 명목형 특성을 이용하여 나누어서 분류를 수행하고, 계량형 자료는 클래스의 수준에 따라 어느 정도 분리된 특성이 있다면 그들 자료는 고정 시킨 후 분류작업을 수행한다.

분류 모형의 처리 절차는 다음과 같다.

1. 명목형 특성이 있으면 문제를 분리한다.
   1.1. 특성과 클래스간의 crosstab 분석 실시.
   1.2 χ^2통계량이 큰 순서로 몇 개의 특성을 혼합하여 가장 크게 반응하는 클래스의 수준과 연관시켜 문제를 분리한다.
2. 계량형 특성을 이용해 자료 개수를 줄인다.
   2.1. 클래스의 각 수준에 대해 각 특성들의 평균과 표준편차를 구한다.
   2.2 각 특성별로 고유 영역이 있는지를 검토하여 있다면 그 영역에 속하는 자료는 고정화 시켜 삭제한다.
3. 사전 처리된 문제에 대해 TAM을 이용해 분류 작업을 수행한다.

3. 사례 적용

UCI 저장소에 있는 몇 가지 실제 자료(Iris 자료, 심장병 자료)를 가지고 이 분류 모형의 효용성을 평가하고자 한다.

3.1 Iris 자료[7]

이 자료는 Sepal length, Sepal width, Petal length, Petal width 등의 연속형의 4개의 특성을 가지고 3개의 종류(Iris Setosa, Iris Versicolour, Iris Virginica)를 가진 총 150개 자료로 구성되어 있다. 표 1에 4개 특성의 최대, 최소값과 평균과 표준편차를 보이고 있다.

그리고 4개 특성들의 자료 분포를 클래스별로 그림 2에 보이고 있다. 이 그림에서 보다시피 특성 Sepal length와 width의 경우에는 자료들이 혼재되어 있지만, Petal length와 width의 경우에는 클래스 별로 자료들이 구분되어 있는 것을 볼 수 있다.

표 1. Iris 종류에 대한 특성들의 최소값-최대값 (평균+표준편차)
Table 1. Maximum and minimum(average+standard deviation of 4 factors for class)

	Sepal l	Sepal w	Petal l	Petal w
Setosa	4.9-7.9 (6.59+0.64)	2.2-3.8 (2.97+0.32)	4.5-6.9 (5.55+0.55)	1.4-2.5 (2.03+0.27)
Vers.	4.9-7.0 (5.94+0.52)	2.0-3.4 (2.77+0.31)	3.0-5.1 (4.26+0.47)	1.0-1.8 (1.33+0.20)
Virgin.	4.3-5.8 (5.01+0.35)	2.3-4.4 (3.43+0.38)	1.0-1.9 (1.46+0.17)	0.1-0.6 (0.25+0.11)

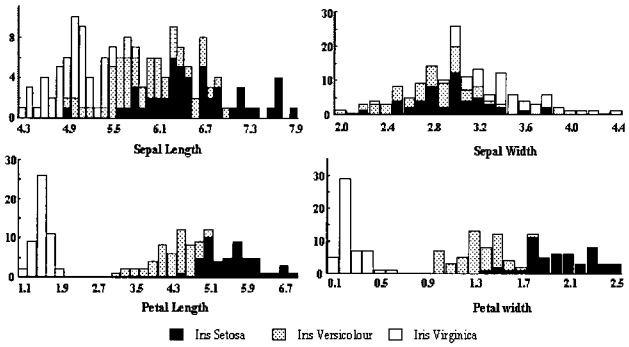

그림 2. 각 클래스별 Iris 자료 분포도
Fig. 2. Iris data distribution for each class

특성 Petal length의 경우 iris Setosa는 [3.9, 7.2], Iris Versicolour는 [2.85, 5.67], Iris Virginica는 [0.91, 1.97]의 영역을 가진다. 그러므로 iris Setosa의 고유영역은 [5.67, 7.2]이고, Iris Versicolour는 [2.85, 3.9], Iris Virginica는 영역 전체가 고유 영역이 된다. 4개 특성 모두에 대해 고유영역을 찾은 후 그것에 속하는 자료를 고정화하여 제거한 결과 150개 자료 중 100개를 제거할 수 있었다.

표 2. Iris자료의 분류 결과
Table 2. Results of classification

방법	Training Data	Checking Data
FTAM	98.67%(74/75)	94.67%(71/75)
영역처리 +FTAM	전처리 50개 고정 FTAM:98%(24/75)	전처리 50개 고정 FTAM:88%(22/75)

150개 자료 중 75개를 훈련 자료로 나머지를 검토자료로 사용하여 사전 처리를 하지 않고 Fuzzy TAM을 이용하여 분류한 결과와 사전 처리를 한 후의 분류 결과를 표 2에 보여 준다. 사전 처리의 결과(총 에러 4/150)가 하지 않은 것(총 에러 5/150)보다 월등히 좋은 결과는 보여주지 못했다.

3.2 심장병 자료[8]

이 자료는 13개의 특성과 심장병 존재 유무를 판정하는 1개의 클래스로 총 270개의 사례로 구성되어 있다. 13개의 특성은 age (F1), sex (F2), chest pain type (F3), resting blood pressure (F4), serum cholestoral in mg/dl (F5), fasting blood sugar>120 mg/dl (F6), resting electrocardiographic results (F7), maximum heart rate achieved (F8), exercise induced angina (F9), oldpeak = ST depression .induced by exercise relative to rest (F10), the slope of the peak exercise ST segment (F11), number of major vessels (0-3) colored by flourosopy (F12), thal (F13)으로 그 중 7개(F2, F3, F6, F7, F9, F11, F13)는 명목형 특성이다.

표 3. 카테고리 특성의 χ^2값
Table 3. χ^2 value of category factors

명목 특성	χ^2통계량
F2	23.93
F3	68.59
F6	0.07
F7	8.98
F9	47.47
F11	40.37
F13	74.57

표 3에는 명목형 특성의 교차분석의 χ^2통계량을 보여 주고 있다. 가장 값이 큰 2개의 특성 F3, F13을 가지고 자료들을 분리한 자료 분포를 표 4에 보여 주고 있다. 자료 1은 심장병이 발견되지 않은 자료가 많은 그룹으로, 자료 2는 심장병이 나타난 자료 비중이 높은 그룹으로 분리되었다.

표 4. F3과 F13에 의해 분리된 자료분포
Table 4. Distribution of dividing data by factors F3 and F13

자료 1	심장병 여부		자료 2	심장병 여부	
F13-F3	No	Yes	F13-F3	No	Yes
3-1	9	3	6-2	1	1
3-2	31	2	6-3	0	2
3-3	51	5	6-4	3	5
3-4	28	23	7-2	3	4
6-1	2	0	7-3	11	10
7-1	4	2	7-4	7	63
합계	125	35	합계	25	85

총 270개의 자료 중 135개를 훈련 자료로 나머지를 체킹 자료로 하여 Fuzzy TAM을 수행한 결과를 표 5에 보여주고 있다. 자료 분해를 한 경우는 각각 81개, 54개를 훈련자료로 사용했다. 이 예제에서는 Fuzzy TAM을 그대로 적용하는 것 보다는 문제를 분리하고 난 후의 모형의 일치율이 약간 높아졌다.

표 5. 심장병 자료에의 분류결과
Table 5. Results of classification

방법	Training Data	Checking Data
FTAM	88.1%(119/135)	80.0%(108/135)
자료분해 +FTAM	FTAM1:92.6%(75/81) FTAM2:94.4%(51/54)	FTAM1:83.5%(66/79) FTAM2:82.1%(46/56)

4. 결 론

본 연구에서 카테고리형 특성에 대해서는 클래스 수준에 대응하는 자료가 많은 그룹으로 자료를 분리시켰고, 연속형 특성에 대해서는 각 클래스 수준이 단독으로 차지하고 있는 영역이 고유영역을 찾아 그곳에 속하는 자료는 미리 분류함으로써 자료 숫자를 줄인 다음 신경망 분류법 중의 하나인 Fuzzy TAM을 적용시켰다.

자료를 그대로 Fuzzy TAM을 적용하여 분류한 것보다는 사전에 자료의 특성을 이용하여 정리한 자료를 사용하는 것이 분류 일치성을 높게 만들었다.

실제 분류를 위한 많은 분류 모형이 있고 그 중 하나가 신경망 모형을 이용한 분류모형이다. 어떤 문제에서든 우수한 결과를 보이는 방법은 없지만 미리 자료의 성질을 이용하여 자료를 정리한다면 그런대로 좋은 결과를 보일 수 있다고 판단된다.

참 고 문 헌

[1] I. Hayashi, J.R. Williamson : "Acquisition of Fuzzy Knowledge from Topographic Mixture Networks with Attentional Feedback", *The International Joint Conference on Neural Networks(IJCNN '01)*, pp. 1386-1391, 2001.

[2] J.R. Williamson : "Self-Organization of Topographic Mixture Networks Using Attentional Feedback", *Neural Computation*, Vol. 13, pp. 563-593, 2001.

[3] Isao Hayashi, Hiromasa Maeda, "A Formulation of Fuzzy TAM Network with Gabor Type Receptive Fields", *2003 International Symposium on Advanced Intelligent Systems*, pp. 620-623, 2003.

[4] 林 勳, James R. Williamson, "TAM Networkのブルーニング手法の提案",システム制御情報學會論文誌, Vol. 17, No. 2, pp. 81-88, 2004.

[5] 김효중, 박종선,"순차적으로 선택된 특성과 유전프로그래밍을 이용한 결정나무", *경영과학*, Vol. 23, No 1, pp. 63-74, 2006.

[6] M. Hall, "Correlation-based Feature Selection of Discrete and Numeric Class Machine Learning", *Proceedings of the International Conference on Machine Learning*, pp. 359-366, 2000

[7] archive.ics.uci.edu/ml/machine-learning-data-base/iris/

[8] rchive.ics.uci.edu/ml/machine-learning-data-base/heart/

저 자 소 개

박광박(Kwang-Pak Park)
1969년 : 부산대학교 수학 학사
1977년 : 부산대학교 수학 석사
1986년 : 경상대학교 수학 박사
현재 : 남대학교 응용수리학부 교수

관심분야 : 함수해석학
Phone : +82-55-249-2207
Fax : +82-55-244-6504
E-mail : kppark@kyungnam.ac.kr

박영만(Young-Man Park)
1980 : 서울대학교 산업공학 학사
1982 : 서울대학교 산업공학 석사
1999 : 일본 동아대학교 정보시스템 박사
현재 : 경남대학교 경영학부 교수

관심분야 : 경영과학, 의사결정
Phone : +82-55-249-2704
Fax : +82-55-223-1655
E-Mail : youngman@kyungnam.ac.kr

황승국(Seung_Gook Hwang)
1981년 : 동아대학교 산업공학 학사
1983년 : 동아대학교 산업공학 석사
1991년 : Osaka Prefecture University 경영공학 박사
현재 : 경남대학교 정보통신공학과 교수

관심분야 : 퍼지모델링 및 평가
Phone : +82-55-249-2705
Fax : +82-55-249-2463
E-mail : hwangsg@kyungnam.ac.kr