데이터마이닝을 위한 뉴로퍼지시스템에 관한 고찰

손인석[1], 황창하[2], 조길호[3], 김태윤[4]

요약

본 논문에서는 데이터마이닝을 위해 최근에 개발된 뉴로퍼지시스템(nuero-fuzzy system) NEFCLASS 모형을 소개하고 실제 예제에 적용하여 그 성능을 평가한다.

주제어 : NEFCLASS, 뉴로퍼지, 데이터마이닝

1. 서론

신경망(neural network)과 퍼지시스템(fuzzy system)을 결합한 방법들이 공학분야에서 많이 활용되고 있다. 그러나 대부분의 방법들은 매우 다른 네트워크 구조(architecture)와 활성함수(activation function) 그리고 학습알고리즘(learning algorithms)을 사용하기 때문에 비교가 쉽지는 않다. Nauck은 다층퍼지신경망(multilayer fuzzy neural network)의 일반적 모형을 위해 퍼지퍼셉트론(fuzzy perceptron)을 소개했다. 이 퍼지퍼셉트론은 신경망과 퍼지시스템을 결합한 여러 방법들의 비교를 용이하게 하는 뉴로퍼지 구조의 공통 기반으로 사용될 수 있다[3, 4, 5, 6, 7].

본 논문에서는 자료분석의 새로운 방법으로 Nauck[3, 4, 5, 6, 7] 등에 의해 개발된 뉴로퍼지시스템 NEFCLASS 모형을 소개한다. 이 시스템의 목적은 서로 다른 그룹으로 분리될 수 있는 자료들의 집합으로부터 퍼지규칙을 유도하는 것이다. 이때 애매성(fuzziness)은 주로 입력패턴을 정확한 그룹으로 분류하는 것을 어렵게 만드는 입력변수들의 부정확하고 불완전한 측정에 기인한다. 자료를 설명하는 퍼지규칙은 다음과 같은 형태이다.

만약 x_1이 μ_1, x_2가 μ_2, … , x_n이 μ_n이면 패턴 $\boldsymbol{x}$는 그룹 i에 속한다.

여기서 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)'$이고 $\mu_1, \cdots, \mu_n$은 퍼지집합이다.

뉴로퍼지시스템 NEFCLASS는 이런 퍼지규칙을 유도하고 소속함수(membership function)의 형태를 학습한다.

본 논문의 구성은 다음과 같다. 2절에서는 NEFCLASS에 대해 소개한다. 그리고 3절에서는 붓꽃자료(Iris data), 미국 유방암자료 및 우리나라 주가자료에 NEFCLASS를 적용한 결과를 설명하고, 4절에서는 결론을 제시한다.

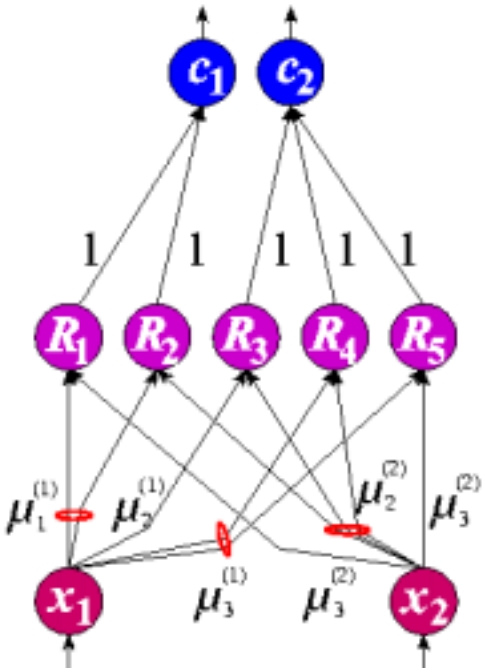
---

1) 경북대학교 통계학과 대학원 석사과정
2) 대구가톨릭대학교 정보통계학과 교수
3) 경북대학교 통계학과 교수
4) 계명대학교 통계학과 교수

2. NEFCLASS 모형

2.1 3층 퍼지퍼셉트론의 NEFCLASS 모형

NEFCLASS는 뉴로(NEuro), 퍼지(Fuzzy), 분류(CLASSification)를 줄여서 표현한 단어이다. NEFCLASS 모형은 주어진 입력패턴의 그룹을 정확하게 결정하는 퍼지규칙을 유도하고 소속함수(membership function)의 형태를 학습한다. 입력패턴은 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)' \in R^n$ 이고 그룹 C는 R^n의 부분집합이다. 패턴의 입력변수들의 값은 퍼지집합으로 표현되며 분류는 언어규칙(linguistic rule)의 집합으로 묘사된다. 각 입력변수 x_i에 q_i개의 퍼지집합 $\mu_1^{(i)}, \cdots, \mu_{q_i}^{(i)}$이 있고, 규칙베이스는 k개의 퍼지규칙 $R_1, \cdots, R_k$을 가진다.



[그림 2.1] NEFCLASS 모형: 입력변수 2개, 규칙 5개, 그룹 2개

그림2.1의 경우와 같이 NEFCLASS 모형은 3층 전방향 망구조를 가진다. 규칙베이스는 분류를 나타내는 함수 $\varphi: R^n \to (0,1)^m$의 근사이다. 여기서 $\varphi(\boldsymbol{x}) = (c_1, \cdots, c_m)$이며 $i, j \in \{1, 2, \cdots, m\}$에 대해 $i \neq j$이면 $c_i = 1$, $c_j = 0$이 성립한다. 즉 입력패턴 $\boldsymbol{x}$는 그룹 C_i에 속한다. 관계식 $\varphi(\boldsymbol{x}) = \psi(\varphi'(\boldsymbol{x}))$에 의해 $\varphi(\boldsymbol{x})$를 얻는다. 여기서 ψ는 NEFCLASS 모형에 의해 유도된 분류결과의 해석을 나타낸다. 그리고 벡터 $\boldsymbol{c}$ 중 가장 큰 성분을 1로 사상(mapping)하고 다른 성분들을 0으로 사상한다.

이런 함수근사를 수행하고 결과적으로 NEFCLASS를 정의하는 퍼지집합과 언어규칙은 학습을 통해 자료로부터 유도된다. 그림1은 5개의 언어규칙을 사용하여 2차원 입력벡터를 2개의 그룹으로 분류하는 NEFCLASS 모형을 설명한다. 3층 퍼지퍼셉트론의 일반적

NEFCLASS 모형을 설명하면 다음과 같다.

[3층 퍼지퍼셉트론의 NEFCLASS 모형]

(1) $U_1 = \{x_1, x_2, \cdots, x_n\}$, $U_2 = \{R_1, R_2, \cdots, R_k\}$ 그리고 $U_3 = \{c_1, c_2, \cdots, c_m\}$.

(2) 노드 $x_i \in U_1$와 노드 $R_r \in U_2$ 사이의 각 연결은 언어적 표현언어적 표현(linguistic term) $A_{j_r}^{(i)}$, $j_r \in \{1, \cdots, q_i\}$로 레이블(label) 된다.

(3) 모든 $R_r \in U_2$, $c \in U_3$에 대해 $W(R, c) \in \{0, 1\}$이 성립한다.

(4) 똑같은 입력노드 x_i로부터 시작되어 동일한 레이블을 갖는 연결(connection)은 항상 똑같은 가중치를 유지한다. 이런 연결을 링크연결(linked connection)이라 부르고 그 가중치를 공유가중치(shared weight)라 부른다.

(5) 노드 $x \in U_1$와 노드 $R \in U_2$ 사이의 연결의 레이블을 $L_{x,R}$로 표기하자. 그러면 모든 $R, R' \in U_2$에 대해 다음 사실이 성립한다.

$$(\forall (x \in U_1) \ L_{x,R} = L_{x,R'}) \Rightarrow R = R'$$

(6) 모든 규칙노드 $R \in U_2$과 모든 노드 $c, c' \in U_3$에 대해 다음 사실이 성립한다.

$$(W(R, c) = 1) \wedge (W(R, c') = 1) \Rightarrow c = c'$$

(7) 모든 출력노드 $c \in U_3$에 대해 $o_c = a_c = \mathrm{net}_c$가 성립한다.

(8) 모든 출력노드 $c \in U_3$에 대해 net_c는 다음과 같이 계산된다.

NEFCLASS 모형의 중요한 특징은 어떤 일부의 연결이 가중치를 공유한다는 것이다. 각 언어값(linguistic value) 예컨대 "x_1은 양수로서 크다"는 하나의 퍼지집합을 나타낸다. 그림 2.1에서는 $\mu_1^{(1)}$이다. 즉, 언어값은 모든 규칙노드(예컨데, 그림2.1의 R_1과 R_2)에 대해 한 개의 해석을 가진다. 언제나 가중치를 함께 공유하는 연결은 같은 입력노드에서 생겨난다.

NEFCLASS 모형은 패턴에 대한 부분적인 지식으로부터 만들어질 수 있으며 학습에 의해서 엡데이트될 수 있다. 사용자가 입력변수들의 범위를 분할하는 초기 퍼지집합을 정의해야 하며 은닉층에서 생성되는 규칙노드의 최대수 k를 명시해야 한다.

입력노드 i와 규칙노드 j 사이의 연결을 언어적 표현(linguistic term) $A_j^{(i)}$로 레이블하고 퍼지집합 $\mu_j^{(i)}$로 나타낸다. 언어적 표현 $A_j^{(i)}$는 "작다", "중간이다", "크다" 등이 될 수 있다. 그리고 똑같은 규칙노드 R이 되게끔 하는 연결들의 퍼지집합을 R의 전제(antecedent)라고 부른다.

NEFCLASS 모형은 처음에 은닉노드가 전혀 없이 시작된다. 은닉노드들은 학습과정에서 첫번째 반복을 수행하는 동안 생성된다. 그리고 주어진 입력패턴 $\boldsymbol{p}$에 대해 퍼지집합의 조합을 찾음으로써 규칙을 생성한다. 이때 각 퍼지집합은 대응되는 입력변수에 대해 가장 큰 소속값을 제공한다. 만약 이 조합이 기존의 한 규칙의 전제와 같지 않고 아직 규칙노드의 최대 개수에 도달하지 않았다면 새로운 규칙노드가 생성된다. 이것은 퍼지규칙을 찾는 매우 간단한 방법이다. 만약 입력패턴이 훈련표본으로부터 랜덤하게 선택되고 그룹의 표본수가 거의 같다면 이 방법이 성공적일 수 있다.

규칙베이스가 만들어질 때 학습알고리즘은 전제의 소속함수를 적응적으로 업데이트한다. 일반적으로 소속함수로 3개의 모수를 갖는 다음의 삼각함수가 사용된다.

$$\mu: R \to [0,1], \quad \mu(x) = \begin{cases} \frac{x-a}{b-a} & \text{if } x \in [a,b) \\ \frac{c-x}{c-b} & \text{if } x \in [b,c] \\ 0 & otherwise \end{cases}$$

어떤 한 규칙의 이행정도를 결정하기 위한 t-norm으로, 즉 규칙노드의 활성함수로 min을 사용한다.

2.2 NEFCLASS 학습알고리즘

n개의 입력노드 $x_1, \cdots, x_n$, $k \le k_{\max}$개의 규칙노드 $R_1, \cdots, R_n$ 그리고 m개의 출력노드 $c_1, \cdots, c_m$을 갖는 NEFCLASS 시스템을 생각한다. 훈련표본은 $T = \{(\boldsymbol{p}_1, \boldsymbol{t}_1), \cdots, (\boldsymbol{p}_s, \boldsymbol{t}_s)\}$이다. 여기서 입력패턴 $\boldsymbol{p} \in R^n$이고 목표패턴 $\boldsymbol{t} \in \{0,1\}^m$이다. NEFCLASS 시스템이 k개의 규칙노드를 만들기 위해 사용하는 학습알고리즘은 다음의 **규칙 학습알고리즘**이다.

[규칙 학습알고리즘]

(1) 훈련표본 T로부터 패턴 $(\boldsymbol{p}, \boldsymbol{t})$를 선택한다.

(2) 각 입력노드 $x_i \in U_1$에 대해 $\mu_{j_i}^{(i)}(p_i) = \max_{j \in \{1, \cdots, q_i\}} \{\mu_j^{(i)}(x_i)\}$를 만족하는 소속함수 $\mu_{j_i}^{(i)}$를 찾는다.

(3) 만약 규칙노드의 수가 $k_{\max}$개 보다 작고, $W(x_1, R) = \mu_{j_1}^{(1)}, \cdots, W(x_n, R) = \mu_{j_n}^{(n)}$을 만족하는 규칙노드 R이 없다면 그런 노드를 생성하고 그 노드를 $t_l = 1$인 출력노드 c_l에 연결한다.

(4) 만약 아직 수행되지 않은 학습패턴이 있고 k가 $k_{\max}$보다 작다면 단계 (1)로 돌아간다. 그렇지 않으면 중단한다.

(5) 다음 세 개의 절차 중 하나를 사용해서 규칙베이스를 결정한다.

- 단순 규칙학습("simple" rule learning): $k_{\max} = k$개의 규칙이 생성되면 멈춘다.
- 최상 규칙학습("best" rule learning): 훈련표본 T에 대해 학습을 수행하고 전파된 패턴들의 각 그룹에 대해 각 규칙노드의 활성값들을 축적한다. 만약 규칙노드 R이 규칙 결론에 의해 명시된 그룹 C_R에 대해서 보다 그룹 C_j에 대해서 더 큰 축적된 활성값을 나타낸다면 규칙 R의 결론을 C_j로 변경한다. 훈련표본 T에 대해 학습을 수행하고 각 규칙노드에 대해 다음의 값을 계산한다

$$V_R = \sum_{\boldsymbol{p} \in T} a_R^{(p)} \cdot e_{\boldsymbol{p}},$$

$$e_{\boldsymbol{p}} = \begin{cases} 1, & \text{만약 패턴 } \boldsymbol{p} \text{가 정확하게 분류되면} \\ 0, & \text{그렇지 않으면} \end{cases}.$$

V_R에 대해 가장 큰 값을 갖는 k개의 규칙노드를 보존하고 NEFCLASS 시스템에서 다른 규칙노드들을 제거한다.

■ 그룹당 최상 규칙학습(**"best per class" rule learning**): 최상 규칙학습에서 처럼 수행된다. 그러나 각 그룹 C_j에 대해 결론이 그룹 C_j를 나타내는 $[\frac{k}{m}]$개의 최상 규칙을 보존한다. 여기서 $[x]$는 x의 정수부분을 의미한다.

퍼지집합을 학습하기 위해 NEFCLASS 시스템의 감독 학습알고리즘은 주어진 종결조건이 만족될 때까지 다음의 **퍼지집합 학습알고리즘**의 단계를 반복함으로써 훈련표본 T에 대해 순환적으로 수행된다.

[퍼지집합 학습알고리즘]

(1) 훈련표본 T로부터 패턴 $(\boldsymbol{p}, \boldsymbol{t})$를 선택하여 그 패턴을 NEFCLASS 시스템을 통해 전파시키고 출력벡터 $\boldsymbol{c}$를 결정한다.

(2) 각 출력노드 c_i에 대해 델타값 $\delta_{c_i} = t_i - a_{c_i}$을 결정한다.

(3) $a_R > 0$인 각 규칙노드 R에 대해

(a) 델타값 $\delta_R = a_R(1 - a_R) \sum_{c \in U_3} W(R, c)\delta_c$ 을 결정한다.

(b) 관계식 $W(x', R)(a_{x'}) = \min_{x \in U_1}\{W(x, R)(a_x)\}$을 만족하는 x'을 찾는다.

(c) 퍼지집합 $W(x', R)$에 대해 학습률 $\sigma > 0$의 다음 델타규칙을 사용하여 모수 a, b, c에 대한 델타값을 결정한다.

$$\delta_b = \sigma \cdot \delta_R \cdot (c - a) \cdot \mathrm{sgn}(a_{x'} - b),$$
$$\delta_a = -\sigma \cdot \delta_R \cdot (c - a) + \delta_b,$$
$$\delta_c = \sigma \cdot \delta_R \cdot (c - a) + \delta_b,$$

그리고 만약 퍼지집합 $W(x', R)$이 주어진 제약조건 Φ를 위배하지 않으면 그 퍼지집합 $W(x', R)$에 변화량을 적용한다.

(4) 만약 한번의 반복이 끝나고 종결조건이 충족되면 중단한다. 그렇지 않으면 단계 (1)로 가서 계속 수행된다.

전술한 바와 같이 NEFCLASS 시스템에 대해 규칙베이스를 생성하기 위한 방법으로 세 가지가 있다. 단순 규칙학습은 학습패턴이 훈련표본으로부터 랜덤으로 선택되고 그룹의 표본수가 거의 같을 때만 성공적으로 사용될 수 있다. 일반적으로 사용자들은 최상 규칙학습 또는 그룹당 최상 규칙학습을 사용한다. 각 그룹의 패턴들이 같은 수의 군집으로 분포되어 있다고 가정할 때는 그룹당 최상 규칙학습을 사용한다. 그리고 다른 그룹들 보다 더 많은 규칙으로 표현되어야 하는 그룹들이 있을 때는 최상 규칙학습이 적합하다. 두 방법에 대해 규칙학습은 훈련자료를 통해 세 번 순환하면 완성된다.

퍼지집합에 대한 학습절차는 간단한 경험적 방법이다. 이 학습절차는 결과적으로 소속함수의 위치를 이동시키고 그 소속함수의 대(support)를 크게 또는 작게 만든다. 학습절차에 대해 제약조건 Φ를 정의하는 것은 쉽다. 예를 들면, 퍼지집합들은 0.5에서 교차해야 한다. 학습절차의 오차값은 수치계산상 0이 될 수 없다. 종결조건으로 주로 오차의 변화량이 사용된다. 단계 (3, a)에서 합기호는 사실상 필요없다. 왜냐하면 각 규칙노드는 단 한 개의 출력

노드에만 연결되기 때문이다. 그러나 이것은 모형을 더욱 더 유연하게 만든다. 왜냐하면 적응적 규칙 가중치를 사용할 수 있게 하기 때문이다. 우리는 이것을 멀리한다. 왜냐하면 NEFCLASS 시스템의 어의(semantics)를 유지하기를 원하기 때문이다. 규칙 가중치들이 좋은 분류결과를 얻기 위해 반드시 필요한 것은 아니다.

3. 적용사례

예제 1. 자료는 1994년 1월부터 2001년 9월까지의 우리나라 주식자료를 사용했으며 훈련자료로 경제위기가 발생된 연도인 1997년 주식자료를 사용했다. 훈련자료를 3구간으로 나누었는데 1월3일부터 9월 18일까지를 정상구간(구간1), 9월 19일부터 10월 21일까지를 위기직전의 불안정구간(구간2), 10월 22일부터 12월 27일까지를 위기구간(구간3)으로 하였다.

구간1은 정상구간으로서 더 이상의 설명이 필요 없다고 생각되며 구간2는 위기상황을 앞두고 변동성의 급작스런 신호가 발생하여 시장이 그것을 인지하여 반응하고 있는 구간이다. 탐색적 자료분석 결과 세 구간이 나름대로 구조적인 차이점이 명백한 것으로 판단되었기 때문에 이들 세 구간을 경제상황 판단의 근거가 되는 패턴(혹은 군집)으로 간주하였다. 그리고 97년 이외의 데이터를 검정자료로 사용하였다.

97년 데이터를 훈련 데이터로 사용하여 NEFCLASS을 훈련시키기 위해 5개의 노드를 갖는 입력층, 3개의 노드를 갖는 출력층으로 구성하였다. 여기서 입력변수는 5개가 사용되었는데 x_1(주가지수), x_2(등락률), x_3(등락률 10일 이동평균), x_4(등락률 10일 이동분산), x_5(10일 이동분산의 변동비) 등이 입력층의 각 노드에 할당되었다. 훈련 후의 분류결과는 292개의 패턴 중 29개가 오분류되었고(표 3.1 참조), 7개의 규칙(표 3.2 참조)이 생성되었다.

가지치기(pruning)를 실행한 후 5개의 입력변수 중 x_1(주가지수), x_5(10일 이동분산의 변동비)가 선택되었다. 선택된 입력변수 두 개를 사용하여 분류한 결과, 292 패턴 중 30개를 오분류되었고(표 3.3 참조), 4개의 규칙이 생성되었다(표 3.4 참조).

[표 3.1] 훈련자료의 분류

	예측된 그룹			
	정상구간	불안정구간	위기구간	합계
정상구간	209 (100%)	0 (0%)	0 (0%)	209
불안정구간	24 (88.9%)	1 (3.7%)	2 (7.4%)	27
위기구간	3 (5.4%)	0 (0%)	53 (94.6%)	56
합계	236	1	55	292

정확한 분류: 263 (90.07%), 오분류: 29 (9.93%)

표3.1은 정상구간은 정확하게 분류되고, 위기구간은 비교적 정확하게 분류되나 불안정구간은 잘 분류되지 못함을 보여준다. 이것은 그룹별 자료의 수가 불균등하기 때문인 것으로 생

각된다. 이와 같이 그룹별 자료의 수가 많이 차이가 날 경우에는 NEFCLASS 모형이 만족할 만한 결과를 보여주지 못하는 단점이 있다.

[표 3.2] 5개의 입력변수에 대한 규칙

R_1: **if** (x_1=l, x_2=m, x_3=m, x_4=s, x_5=s) **then** 정상구간
R_2: **if** (x_1=m, x_2=m, x_3=m, x_4=s, x_5=s) **then** 위기구간
R_3: **if** (x_1=l, x_2=m, x_3=m, x_4=s, x_5=m) **then** 정상구간
R_4: **if** (x_1=l, x_2=m, x_3=m, x_4=s, x_5=l) **then** 불안정구간
R_5: **if** (x_1=m, x_2=m, x_3=m, x_4=s, x_5=m) **then** 정상구간
R_6: **if** (x_1=s, x_2=m, x_3=m, x_4=s, x_5=s) **then** 위기구간
R_7: **if** (x_1=s, x_2=m, x_3=m, x_4=s, x_5=m) **then** 위기구간

여기서 s는 small, m은 medium, l은 large를 의미한다.

[표 3.3] 가지치기 후의 훈련자료의 분류

	예측된 그룹			
	정상구간	불안정구간	위기구간	합계
정상구간	207 (99%)	0 (0%)	2 (1%)	209
불안정구간	23 (85.2%)	1 (3.7%)	3 (11.1%)	27
위기구간	2 (3.6%)	0 (0%)	54 (96.4%)	56
합계	232	1	59	292

정확한 분류: 262 (89.73%), 오분류: 30 (10.27%)

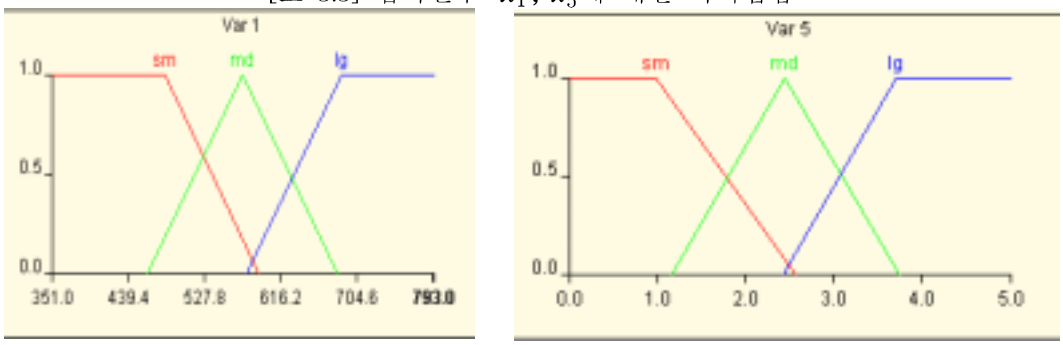
표3.3은 가지치기를 한 후에 선택된 2개의 입력변수 x_1, x_5를 사용하여 훈련자료에 대해 다시 분류를 한 결과를 보여주는데 5개의 입력변수에 대한 훈련자료의 분류결과보다 오분류율이 조금 커졌다.

[표 3.4] 가지치기후의 규칙

R_1: **if** (x_1=l, x_5=s) **then** 정상구간
R_2: **if** (x_1=l, x_5=l) **then** 불안정구간
R_3: **if** (x_1=s, x_5=s) **then** 위기구간
R_4: **if** (x_1=s, x_5=m) **then** 위기구간

여기서 s는 small, m은 medium, l은 large를 의미한다.

[표 3.5] 입력변수 x_1, x_5에 대한 퍼지집합



여기서 sm은 small, md는 medium, lg는 large를 의미한다.

표3.6은 1997년 이외의 검정자료에 대한 분류결과를 보여준다. 검정자료의 분류결과에 대한 정확한 해석은 경제학적인 해석과 더불어 가능하다고 생각되어 본 논문에서는 검정자료의 분류결과에 대한 해석을 고려하지 않겠다.

[표 3.6] 검정자료의 분류

예측 그룹				
정상구간	불안정구간	위기구간	미분류	합계
1322	8	486	1	1817

예제 2. 붓꽃자료(Iris data)는 3개의 그룹 Setosa, Virginica, Versicolour로 이루어져 있으며 각 그룹은 50개의 훈련자료를 가진다. 검정자료는 없기 때문에 훈련자료에 대한 결과만 설명한다. 표3.7과 표3.8은 각각 분류결과와 규칙을 보여준다. 붓꽃자료는 세 그룹의 표본수가 50으로 같기 때문에 NEFCLASS 모형이 만족할 만한 결과를 보여준다고 생각된다.

[표 3.7] 훈련자료의 분류결과

	예측 그룹			
	Setosa	Virginica	Versicolour	합계
Setosa	50	0	0	50
Virginica	0	48	2	50
Versicolour	0	4	46	50
합계	50	52	48	150

정확한 분류: 144 (96.0%), 오분류: 6 (4.0%)

[표 3.8] 4개의 입력변수에 대한 규칙

R_1: **if** (x_1=s, x_2=m, x_3=s, x_4=s) **then** Setosa
R_2: **if** (x_1=s, x_2=s, x_3=s, x_4=s) **then** Setosa
R_3: **if** (x_1=m, x_2=s, x_3=l, x_4=l) **then** Virginica
R_4: **if** (x_1=l, x_2=s, x_3=l, x_4=l) **then** Virginica
R_5: **if** (x_1=l, x_2=m, x_3=l, x_4=l) **then** Virginica
R_6: **if** (x_1=m, x_2=s, x_3=m, x_4=m) **then** Versicolour
R_7: **if** (x_1=m, x_2=s, x_3=m, x_4=s) **then** Versicolour

여기서 s는 small, m은 medium, l은 large를 의미한다.

예제 3. Wisconsin 유방암 자료는 699개의 관측치를 가진다. 그중 16개의 관측치는 결측치를 포함하고 있기 때문에 분석에서 제외되었다. 관측치들은 두 개의 그룹 양성(benign), 악성(malign)으로 나누어진다. 각 관측치는 9개의 입력변수를 가진다.

표3.9과 표3.10은 각각 분류결과와 규칙을 보여준다. 이 자료에 대해서 NEFCLASS 모형은 만족할 만한 결과를 보여준다.

[표 3.9] 훈련자료의 분류결과

	예측 그룹		
	악성	양성	합계
악성	225	14	239
양성	11	433	444
합계	236	447	683

정확한 분류: 658 (96.3%), 오분류: 6 (3.7%)

[표 3.10] 9개의 입력변수에 대한 규칙

R_1: **if** (s, s, s, s, s, s, s, s, s) **then** 양성
R_2: **if** (l, s, s, s, s, s, s, s, s) **then** 양성
R_3: **if** (l, l, l, l, l, l, l, l, l) **then** 악성
R_4: **if** (l, l, l, l, s, l, l, l, s) **then** 악성

여기서 s는 small, m은 medium, l은 large를 의미한다.

4. 결론

본 논문에서는 데이터마이닝을 위해 최근에 개발된 뉴로퍼지분류시스템(nuero-fuzzy classification system) NEFCLASS 모형을 소개하고 실제자료에 적용하여 모형의 성능을 평가하였다. 사전지식에 의해 초기화 될 수 있는 NEFCLASS 모형은 퍼지의 **if-then** 규칙을 사용하여 학습을 한 후 해석이 가능하다는 강한 장점이 있다. 즉 일반적 신경망과 같은 블랙박스 방법은 아니고 해석가능한 퍼지분류 방법이다. NEFCLASS 모형은 훈련자료에 대해 한번 학습한 후 퍼지규칙을 생성할 수 있다. 퍼지규칙을 생성한 후 NEFCLASS 모형은 감독 학습알고리즘을 사용해 소속함수의 모수들을 적응적으로 추정하여 최종적으로 퍼지규칙을 완성한다. 유도된 규칙은 다른 뉴로퍼지 방법처럼 가중평균을 사용하여 계산되지 않는다. 따라서 의미론적 문제(semantical problem)를 피하고 학습결과를 간단하게 만든다.

세 가지 실제자료인 우리나라 주식자료, 붓꽃자료, Wisconsin 유방암자료에 적용한 결과에 의하면 NEFCLASS 모형은 붓꽃자료와 Wisconsin 유방암자료에 대해서는 만족할 만한 결과를 보여주었다. 그러나 우리나라 주식자료에 대해서는 NEFCLASS 모형이 만족할 만한 결과를 보여주지 못했다. 그룹별 자료의 수가 많이 차이가 날 경우에는 NEFCLASS 모형의 성능이 떨어짐을 보여주는 결과로 생각된다. 따라서 이런 문제점이 해결되어야 하겠다.

데이터마이닝을 위해 적절한 모형을 선택할 때는 모형의 정확성(accuracy)과 해석가능성(interpretability)을 고려하는데 해석력을 강조하는 자료분석을 위해서는 주로 의사결정트리(decision tree)를 사용하고 예측력(prediction)을 강조하는 자료분석을 위해서는 주로 신경망 또는 SVM(support vector machines)을 사용한다. NEFCLASS 모형은 신경망의 역전파 알고리즘과 같이 반복학습알고리즘을 사용하여 예측력을 높이고 궁극적으로는 모형의 해석력도 높이고자 개발된 모형이다. 그러나 최근에 개발된 Wu 등(1999)의 MOC1 모형은 의사결정트리이며 예측력도 상당히 높은 모형이다. 따라서 NEFCLASS 모형이 해석력과 예측력을 겸비한 데이터마이닝을 위한 모형으로 자리잡기 위해서는 이론적으로 더 많은 연구가 이루어져야 한다고 생각된다.

참고문헌

[1] 이상배 (1999). 퍼지뉴로 제어시스템, 교학사.

[2] Cherkassky, V. and Mulier, F. (1998). Learning from data. John Wiley & Sons, Inc.

[3] Nauck, D., Nauck, U. and Kruse, R. (1999) NEFCLASS for JAVA - New Learning Algorithms. In Proc. 18th International Conference of the North American Fuzzy Information Processing Society (NAFIPS'99), New York, pp. 474-476, 1999.

[4] Nauck, D. (2000). Adaptive Rule Weights in Neuro-Fuzzy Systems. Neural Computing and Applications, 9:60-70

[5] Nauck, D. (2000). Knowledge Discovery with NEFCLASS. In Proc. Fourth Int. Conf. Knowledge-Based Intelligent Engineering Systems & Allied Technologies (KES'2000), Brighton, pp. 158-161.

[6] Nauck, D. and Kruse, R. (2000). NEFCLASS-J - A JAVA-based Soft Computing Tool. In: Benham Azvine, Nader Azarmi, Detlef Nauck (eds.): Intelligent Systems and

Soft Computing: Prospects, Tools and Applications. Lecture Notes in Artifical Intelligence 1804, pp. 143-164, Springer-Verlag, Berlin, 2000.

[7] Nauck, D. (2001). Fuzzy Data Analysis with NEFCLASS. In Proc. Ninth IFSA World Congress, Vancouver, to appear.

[8] Vapnik, V. (1998). Statsistical leraning theory. John Wiley & Sons, Inc.

[9] Wu, I. G., Bennett, K., Cristianini, N. and Shawe-Taylor, J. (1999). Large margin decision trees for induction and transduction. In Proceedings of the Sixteenth International Conference on Machine Learning(ICML99).