ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

역퍼지화 기반의 인스턴스 선택을 이용한 파킨슨병 분류

Classification of Parkinson's Disease Using Defuzzification-Based Instance Selection

이 상 홍[1*]
Sang-Hong Lee

요 약

본 논문에서는 분류 성능을 향상하기 위해서 Takagi-Sugeno(T-S) 퍼지 모델 기반의 가중 퍼지소속함수 기반 신경망(Neural Network with Weighted Fuzzy Membership Functions; NEWFM)을 이용한 새로운 인스턴스 선택을 제안하였다. 제안하는 인스턴스 선택은 T-S 퍼지 모델에서의 가중 평균 역퍼지화와 통계학에서 사용하는 정규분포의 신뢰구간과 같은 구간 선택을 이용하여 인스턴스를 선택하였다. 제안하는 인스턴스 선택의 분류 성능을 평가하기 위해서 인스턴스 사용 전/후에 따라서 분류 성능을 비교하였다. 인스턴스 사용 전/후에 따른 분류 성능은 각각 77.33%, 78.19%로 나타났다. 또한 인스턴스 사용 전/후에 따른 분류 성능 간에 차이점을 보여주기 위해서 통계학에서 사용하는 맥니마 검정을 사용하였다. 맥니마 검정의 결과로 유의 확률이 0.05보다 적게 나오므로 인스턴스 선택의 분류 성능이 인스턴스 선택을 하지 않는 경우의 분류 성능보다 우수함을 확인 할 수가 있었다.

☞ 주제어 : 인스턴스 선택, 파킨슨병, 걸음걸이, 퍼지신경망, 웨이블릿 변환, 정규분포

ABSTRACT

This study proposed new instance selection using neural network with weighted fuzzy membership functions(NEWFM) based on Takagi-Sugeno(T-S) fuzzy model to improve the classification performance. The proposed instance selection adopted weighted average defuzzification of the T-S fuzzy model and an interval selection, same as the confidence interval in a normal distribution used in statistics. In order to evaluate the classification performance of the proposed instance selection, the results were compared with depending on whether to use instance selection from the case study. The classification performances of depending on whether to use instance selection show 77.33% and 78.19%, respectively. Also, to show the difference between the classification performance of depending on whether to use instance selection, a statistics methodology, McNemar test, was used. The test results showed that the instance selection was superior to no instance selection as the significance level was lower than 0.05.

☞ keyword : Instance Selection, Parkinson's Disease, Gait, Fuzzy Neural Networks, Wavelet Transforms, Normal Distribution

1. 서 론

최근에는 모든 연구 분야에서 데이터 마이닝이나 기계 학습(Machine Learning)을 수행하는 데 필요한 데이터의 양이 급속히 증가하였다[1][17-18]. 일반적으로 데이터의 양이나 입력이 많으면 어떠한 사실을 분류하거나 판단하는데 좀 더 효율적이라고 생각하지만 오히려 너무 많은 데이터나 입력은 메모리와 시간적인 관점에서 비효율을 초래할 수 있다[2-5]. 또한, 데이터 간의 관련성이 적은 데이터는 잘못된 결과를 야기할 수도 있다[17-20]. 따라서 기존의 퍼지신경망에서는 학습하는데 사용되어지는 데이터의 양을 줄이거나 입력으로 사용되는 입력의 개수를 줄이는 작업이 필요시 되고 있다[6][19-26]. 일반적으로 입력의 개수를 줄이는 방법으로는 특징 선택[21-23]이 있고 그 입력이 가지는 값들을 줄이는 방법으로는 인스턴스 선택[24-26]이 있다. 특징 선택은 중복 또는 관련이 없는 특징을 제거하여 분류 성능을 향상시키고 최소한의 기능을 사용하여 운영비용을 줄임으로써 분류 성능을 향상시킨다[7][21-23]. 인스턴스 선택은 특징이 가지는 실제 값들로부터 의미 있고 좋은 값들을 선택하여 좋은 학습을 유도함으로써 분류 성능을 향상시킨다[8][24-26].

본 논문에서는 기존의 가중 퍼지소속함수 기반 신경망(Neural Network with Weighted Fuzzy Membership Functions; NEWFM)[9][10][11]에 Takagi-Sugeno (T-S) 퍼

[1] Department of Computer Science & Engineering, Anyang University, Anyang-si, 430-714, Korea
* Corresponding author (shleedosa@gmail.com)
[Received 14 March 2014, Reviewed 19 March 2014, Accepted 16 April 2014]

지 모델[12]을 접목한 새로운 인스턴스 선택을 제안하여 기존 NEWFM의 분류 성능을 향상시키고자 하였다. 제안하는 인스턴스 선택은 T-S 퍼지 모델에서 사용하는 가중 평균 역퍼지화와 통계학에서 사용하는 정규분포의 신뢰구간을 이용하였다. 또한 제안하는 인스턴스 선택 과정의 첫 번째 단계에서는 모든 인스턴스를 이용하여 학습을 수행하게 된다. 학습이 끝난 후에는 입력으로부터 학습된 가중 퍼지소속함수의 경계합(Bounded Sum of Weighted Fuzzy Membership functions, BSWFM)을 구하게 된다. 두 번째 단계에서는 첫 번째 단계에서 구한 BSWFM을 T-S 퍼지 모델에서의 가중 평균 역퍼지화와 통계학에서 사용하는 정규분포의 신뢰구간을 이용하여 인스턴스를 선택하게 된다. 이렇게 선택된 인스턴스를 이용하여 새롭운 학습을 수행하게 된다.

본 논문에서는 제안하는 인스턴스 선택의 분류 성능을 평가하기 위해서 족압(foot pressure)의 분석을 기반으로 하여 건강한 사람과 파킨슨병(Parkinson's disease) 환자를 분류하였다. 파킨슨병은 뇌의 흑질(substantia nigra)에 분포하는 도파민이라는 신경세포의 결핍에 의해 발생하는 대표적인 퇴행성 질환이다[13]. 본 논문에서는 파킨슨병을 분류하기 위해서 파킨슨병의 증상 중에서 운동완서(bradykinesia)에 해당하는 발을 끄는 걸음걸이 특징을 이용하였다[14][15]. 건강한 사람의 족압과 파킨슨병 환자의 족압으로부터 NEWFM에서 사용할 입력을 추출하기 위해서 일정 시점에서 족압의 최대값과 최소값의 차이를 계산하였고 이렇게 계산된 값은 웨이블릿 변환(Wavelet Transform, WT)을 이용하여 웨이블릿 계수를 추출하였다. 추출된 웨이블릿 계수들을 주파수 분포와 주파수 변동량을 이용하여 입력으로 사용할 특징을 구하였다.

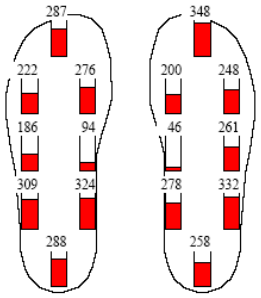
본 논문에서 NEWFM은 족압 데이터를 이용하여 건강한 사람과 파킨슨병 환자를 분류하였을 때 인스턴스 선택 전/후의 분류 성능(accuracy)은 각각 77.33%, 78.19%로 나타났다. 또한 인스턴스 선택 전/후의 분류 성능 간에 차이점을 보여주기 위해서 통계학에서 사용하는 맥니마 검정을 사용하였다. 맥니마 검정의 결과로 유의 확률이 0.031로써 0.05보다 적게 나오므로 인스턴스 선택 후의 분류 성능이 인스턴스 선택 전의 분류 성능보다 우수함을 확인 할 수가 있었다.

2. 파킨슨병 분류 모델의 개요

본 논문에서는 역퍼지화 기반의 인스턴스 선택을 이용하여 파킨슨병을 분류하였다. 센서로부터 수집된 신호는 전처리 과정인 웨이블릿 변환(Wavelet Transform, WT)에서 사용되었다. 전처리 과정 후에 웨이블릿 변환된 신호는 주파수 분포와 주파수 변동량을 이용하여 NEWFM에서 사용할 입력으로 추출되었다.

2.1 실험 데이터(Experimental Data)

본 논문에서는 PhysioBank(http://www.physionet.org/physiobank/database/gaitpdb/)에서 제공하는 족압 데이터를 사용하여 건강한 사람과 파킨슨병 환자를 분류하였다. 이 실험 데이터는 93명의 파킨슨병 환자와 73명의 건강한 사람의 발바닥에 그림 1과 같이 왼쪽/오른쪽에 각각 8개씩의 센서로부터 수집되었다[16].



(그림 1) 발바닥에 위치한 8개의 센서를 통하여 수집되는 압력
(Figure 1) Vertical ground reaction force records for each of eight sensors located under each foot

(표 1) 측정되는 데이터 집합에 대한 설명
(Table 1) Description of measured data

입력 순번	입력 설명
첫 번째 입력	시각
두 번째 입력부터 아홉 번째 입력 $(l_{n1}, l_{n2}, ..., l_{n8})$	그림 1에서 왼쪽발바닥의 센서로부터 수집되는 8가지 데이터
열 번째 입력부터 열일곱 번째 입력 $(r_{n1}, r_{n2}, ..., r_{n8})$	그림 1에서 오른쪽발바닥의 센서로부터 수집되는 8가지 데이터
열여덟 번째 입력	왼쪽발바닥으로부터 수집되는 전체 압력 데이터
열아홉 번째 입력	오른쪽발바닥으로부터 수집되는 전체 압력 데이터

단, n은 입력되는 순서

0.01초 간격으로 수집되는 데이터 집합은 표 1과 같이 총 19개의 입력으로 구성되어있다. 본 논문에서는 표 1의

데이터를 이용하여 표 2에서 설명하고 있는 수식으로 값을 추출하였다. 그 이유는 걸음걸이에 있어서 질질 끄는 특징을 보이는 파킨슨병 환자의 양쪽 족압의 차가 건강한 사람의 양쪽 족압의 차보다는 상대적으로 적다는 특징이 있기 때문이다[16]. 족압 데이터가 저장되어 있는 각각의 파일로부터 순차적으로 2048개씩 추출하였다. 이렇게 구성된 2048개씩의 데이터 집합을 표 2에서 제시한 수식을 이용하여 값을 추출하였다.

(표 2) 족압 데이터를 이용한 특징 추출 방법
(Table 2) Preprocessing formulas using foot pressure

방법	$f_l(1)-f_r(1), \ldots, f_l(2048)-f_r(2048)$

단, $f_l(n)=\max(l_{n1}, l_{n2}, ..., l_{n8})-\min(l_{n1}, l_{n2}, ..., l_{n8})$

$f_r(n)=\max(r_{n1}, r_{n2}, ..., r_{n8})-\min(r_{n1}, r_{n2}, ..., r_{n8})$,

n은 입력되는 순서

2.2 웨이블릿 변환(Wavelet Transforms)과 통계적 기법을 이용한 특징 추출

본 논문에서는 표 2의 수식에 의해 생성된 값을 [11][16]에서 사용한 스케일 레벨 5인 이분 비연속 Haar 웨이블릿 변환을 수행하였다. 이렇게 수행된 레벨 2부터 레벨 5까지의 웨이블릿 계수인 detail coefficient와 approximation coefficient를 [11][16]에서 사용한 다음과 같은 통계적 기법 (1), (2), (3), (4), (5)를 이용하여 본 논문에서 입력으로 사용할 40개의 값을 동일한 방법으로 추출하였다.

(1) 각 레벨 안에 있는 모든 계수들에 대한 절대값의 평균값
(2) 각 레벨 안에 있는 모든 계수들을 제곱하여 구한 평균값
(3) 각 레벨 안에 있는 모든 계수들의 중앙값
(4) 각 레벨 안에 있는 모든 계수들의 표준편차
(5) 인접한 레벨간의 레벨 안에 있는 모든 계수들에 대한 평균값의 절대값 비율

위에서 언급한 통계적 기법 (1), (2), (3)는 신호에 대한 주파수 분포를 의미한다. 또한 통계적 기법 (4), (5)는 신호에 대한 주파수 변동량을 의미한다.

본 논문에서는 통계적 기법에 의해서 추출된 40개의 입력을 구성원으로 하는 실험군을 구성하였다. 건강한 사람의 420개의 실험군과 파킨슨병 환자의 974개의 실험군을 이용하여 표 3과 같이 구성하였다. 저장되어있는 파일의 정렬 순으로 전반부와 후반부로 나누어서 5대5의 비율로 훈련 집합과 테스트 집합으로 나누었다[11][16].

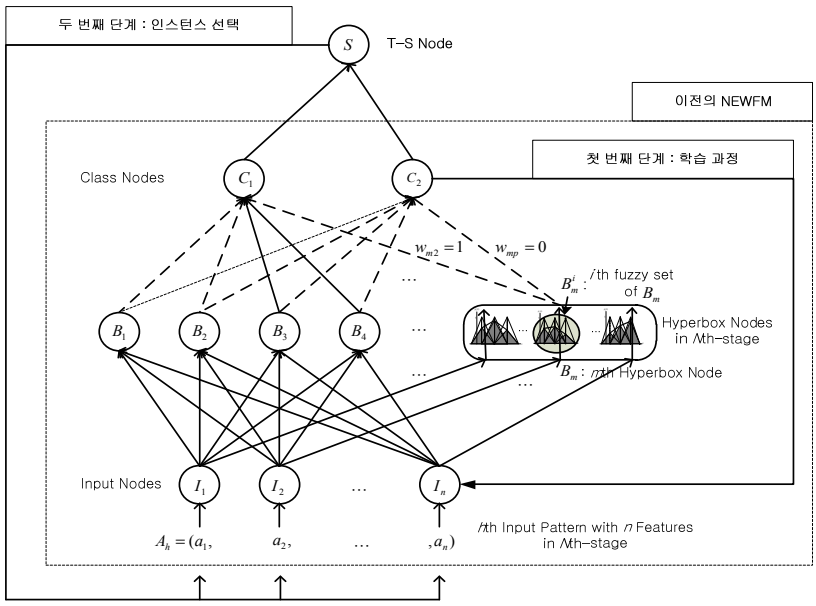
(표 3) 파킨슨병 분류에 사용한 실험군 (훈련 집합과 테스트 집합이 5대5인 비율)
(Table 3) Numbers of training and testing sets

클래스	훈련 집합	테스트 집합	전체 개수
파킨슨병 환자	487개	487개	974개
건강한 사람	210개	210개	420개
전체 개수	697개	697개	1394개

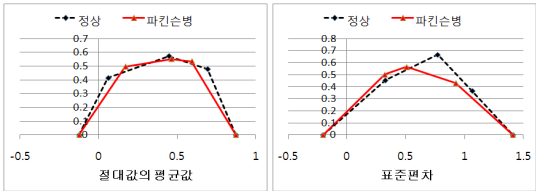
3. 가중 퍼지소속함수 기반 신경망

가중 퍼지소속함수 기반 신경망(Neural Network with Weighted Fuzzy Membership Function, NEWFM)은 입력으로부터 학습된 가중 퍼지소속함수의 경계합(Bounded Sum of Weighted Fuzzy Membership functions, BSWFM)을 이용하여 클래스 분류를 하는 지도학습(supervised) 퍼지 신경망이다[9][10][11][16]. 그림 2의 첫 번째 단계인 NEWFM의 학습 과정에서는 앞에서 설명한 통계적 기법에 의해서 추출된 40개의 특징을 입력으로 사용하여 n개의 입력을 갖는 h번째 입력 Ah = (a1, a2, … , an)으로 사용되어지는 과정을 보여주고 있다. 또한 첫 번째 단계인 학습 과정이 완료된 후에는 그림 3과 같은 BSWFM을 생성하게 된다. 그림 3에서는 통계적 기법 (1)과 (4)에 의해 추출된 입력에 대한 BSWFM의 예를 보여주고 있다.

본 장에서 분류 성능을 향상하기 위하여 제시하는 인스턴스 선택 알고리즘은 다음과 같은 단계들로 구성되어 있다. 첫 번째 단계에서는 기존의 NEWFM에서 학습한 결과로 생성되는 BSWFM을 T-S 퍼지 모델에 적용하여 가중 평균 역퍼지화 값을 구하였다. 두 번째 단계에서는 가중 평균 역퍼지화 값들에 대한 정규분포를 구하고 세 번째 단계에서는 정규분포의 신뢰구간과 같은 구간 선택을 이용하여 가중 평균 역퍼지화 값으로부터 NEWFM의 입력에 사용할 인스턴스의 구간을 선택하였다. 선택된 구간에 있는 인스턴스를 NEWFM의 입력에 사용하여 새롭게 학습을 하였다. 이렇게 하여 새로운 BSWFM을 구하였다. 그림 2는 앞에서 설명한 기존의 NEWFM의 기본 구조에 인스턴스 선택을 추가한 확장된 NEWFM의 구조를 보여주고 있다.



(그림 2) 가중 퍼지소속함수 신경망(NEWFM)의 구조
(Figure 2) Structure of NEWFM



(그림 3) 가중 퍼지소속함수의 경계합의 예 [16]
(Figure 3) Examples of BSWFMs [16]

4. 인스턴스 선택 (Instance Selection)

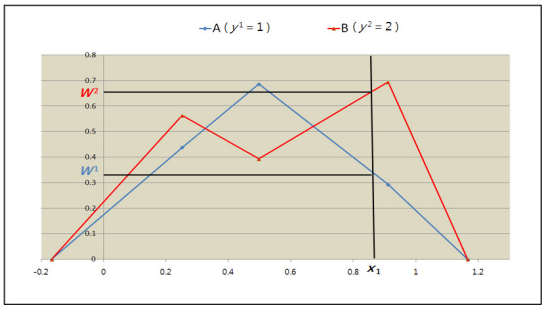
4.1 가중 퍼지소속함수의 경계합에서의 가중 평균 역퍼지화

그림 2에서는 첫 번째 단계에서의 NEWFM 학습 과정에서 제공했던 BSWFM의 가중 평균 역퍼지화 값이 T-S 노드에 저장되는 것을 구조를 보여주고 있다. 그림 4에서는 BSWFM에서 역퍼지화하는 과정을 예를 들어 설명하고 있다. 하나의 입력이 들어오면 BSWFM에서는 두 개의 퍼지소속함수의 값을 구할 수 있는데 이 들 퍼지소속함수의 값을 아래의 식 (1)과 식 (2)를 이용해 가중 평균 역퍼지화 값을 구하게 된다[12].

식 (1)에서 R^i는 i번째 퍼지규칙, x_i는 입력, y는 가중 평균 역퍼지화 값, $A_1^i,\cdots,A_m^i$는 x_i에 대한 i번째 퍼지 집합, $a_0^i,\cdots,a_m^i$는 후건부의 매개변수 집합이다.

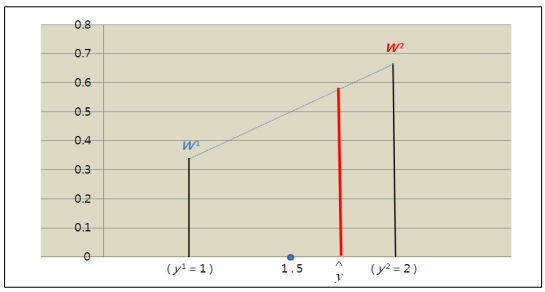
$$R^i : IF x_1 is A_1^i ,\cdots, x_m is A_m^i \ \ THEN y^i = a_0^i + \cdots + a_m^i x_m \quad (1)$$

$$y = \frac{\sum_{i=1}^{c} w^i y^i}{\sum_{i=1}^{c} w^i} \ where\, w^i = Min(A_1^i(x_1),\ldots,A_m^i(x_m)) \quad (2)$$



(그림 4) BSWFM에서 가중 평균 역퍼지화의 예
(Figure 4) The example of weighted average defuzzification in BSWFMs

그림 5는 그림 4에서 가중 평균 역퍼지화 과정을 무게중심법의 형태로 변환하여 설명하고 있다. 그림 4에서는 클래스 분류 1(건강한 사람)과 2(파킨슨병 환자)에 대해서 퍼지소속함수의 값인 w^1과 w^2을 구할 수 있다. 만약 w^1가 커질수록 클래스 분류 1에 속하는 정도(degree)가 커지고 w^2가 커질수록 클래스 분류 2에 속하는 정도가 커진다는 사실을 알 수가 있다. 이러한 사실을 이용하면 기존의 신경망이 가지고 있는 단지 1 또는 2로만 클래스를 분류할 수 있다는 단점을 보완할 수 있다. 예를 들어서 기존의 신경망은 파킨슨병이 '있다' 또는 '없다'로만 분류할 수 있지만 가중 평균 역퍼지화를 이용하면 파킨슨병이 '어느 정도로 있다'로 표현이 가능하다.

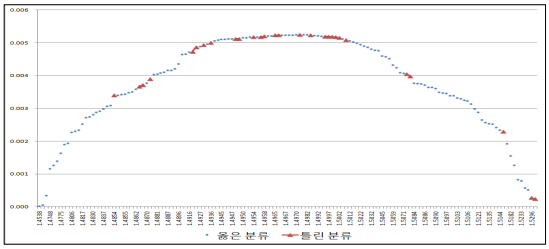


(그림 5) BSWFM에서 가중 평균 역퍼지화 값을 구하는 예
(Figure 5) The example of how to calculate weighted average defuzzification in BSWFMs

4.2 가중 평균 역퍼지화 값의 정규분포

가중 평균 역퍼지화 값은 식 (3)을 이용해 그림 6과 같은 정규분포로 변환되어진다. 그 이유는 정규분포를 구하게 되면 가중 평균 역퍼지화 값에 대한 전체 데이터의 분포도를 구할 수가 있다. 이러한 정규분포에서는 평균값을 중심으로 중심에서 가까울수록 확률이 높은 값(빈도수가 높은 값)이고 멀수록 확률이 낮은 값(빈도수가 낮은 값)을 구할 수가 있다. 또한 이렇게 분포되어있는 가중 평균 역퍼지화 값을 정규분포에서 제공하는 신뢰구간과 같은 구간 선택을 이용하여 가중 평균 역퍼지화 값을 구간별로 선택할 수가 있게 된다. 본 논문에서 제안하는 새로운 인스턴스 선택에서는 확률이 높은 값과 낮은 값을 구분하여 확률이 높은 값을 중심으로 하여 NEWFM에서 새롭게 학습시키고자 하였다. 기존의 NEWFM은 첫 번째 단계의 학습 과정이 완료되면 그림 3과 같은 BSWFMs가 생성이 된다. 따라서 첫 번째 단계의 학습이 완료가 된 후에 테스트(test)를 하는 과정에서 그림 7과 같이 BSWFMs에 의해 입력되는 값에 대해서 1 또는 2의 클래스로 분류되어진다.

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} \quad (-\infty < x < \infty) \qquad (3)$$



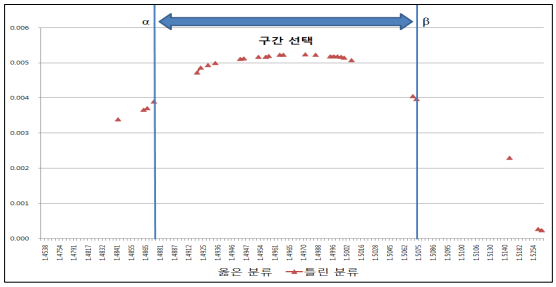
(그림 6) 가중 평균 역퍼지화의 정규분포
(Figure 6) Normal distribution of weighted average defuzzification

(그림 7) 첫 번째 단계에서의 분류 과정
(Figure 7) Classification process in the first step

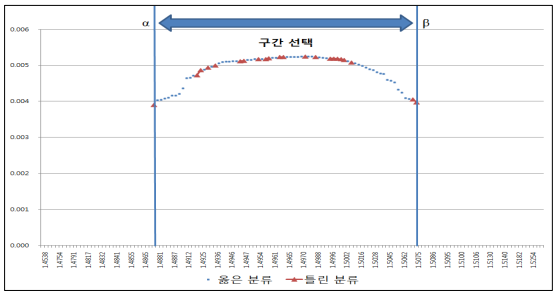
4.3 정규분포의 신뢰구간 기반의 인스턴스 선택

그림 8은 그림 6에서 클래스 분류를 틀리게 한 경우에 해당하는 가중 평균 역퍼지화 값만 표시하였다. 그 이유는 정규분포에서 제공하는 신뢰구간과 같은 구간 선택을 이용하여 전체 인스턴스에서 일부 구간을 선택할 때 그림 8과 같이 틀리게 분류한 가중 평균 역퍼지화 값을 기준으로 선택하였기 때문이다. 그림 8에서는 구간의 경계값을 α, β로 정하여 α와 β사이에 있는 가중 평균 역퍼지화 값과 매칭이 되는 인스턴스를 두 번째 단계에서 NEWFM의 입력으로 사용하여 새롭게 학습을 하게 된다. 여기서 α, β로 정할 때 본 논문에서는 틀린 분류를 중심으로 전체 틀린 분류에서 약 10%씩 감소하면서 구간 α, β를 선택하였다. 이렇게 구간이 정해지면 그림 9와 같이 구간 안에 존재하는 모든 가중 평균 역퍼지화 값과 매칭이 되는 인스턴스를 식 (4)와 같이 구하였다. 식 (4)에서 $x_i (1 \le i \le m)$은 식 (1)에서의 입력이고 y는 식 (2)의 가중 평균 역퍼지화 값이다. α와 β는 그림 6에서 구간의 경계 값이다.

$$Select\, x_1, x_2, \cdots, x_i \; where \; y \ge \alpha \; and \; y \le \beta \qquad (4)$$

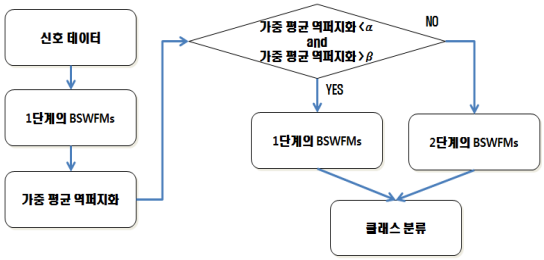


(그림 8) 정규분포에서 인스턴스 구간을 선택하는 예
(Figure 8) The example of how to select instance interval in normal distribution



(그림 9) 구간 안에 있는 가중 평균 역퍼지화 값
(Figure 9) Weighted average defuzzification values in interval

두 번째 단계가 완료된 후에 그림 9의 α와 β사이에 있는 가중 평균 역퍼지화 값과 매칭이 되는 인스턴스를 NEWFM의 입력으로 사용하여 다시 학습하게 된다. 그 결과로 1 또는 2의 클래스를 분류하기 위한 두 번째 단계에서의 BSWFMs가 새롭게 생성이 된다. 따라서 두 번째 단계의 학습이 완료가 된 후에 테스트(test)를 하는 과정에서 그림 10과 같은 과정을 통해 입력되는 값에 대해서 1 또는 2의 클래스로 분류하게 된다.



(그림 10) 두 번째 단계에서의 분류 과정
(Figure 10) Classification process in the second step

5. 실험 결과 (Experimental Results)

본 장에서는 4장에서 설명한 인스턴스 선택의 분류 성능의 우수성을 보이기 위해서 사례 연구로써 파킨슨병 환자의 족압 데이터를 이용하였다. 또한 통계학적 검증을 이용하여 인스턴스 선택의 분류 성능의 우수함을 평가하였다.

표 4에서는 성능 평가를 위해 그림 2의 NEWFM 구조에서 첫 번째 단계의 학습 과정에서 사용된 인스턴스의 개수와 두 번째 단계의 인스턴스 선택 과정에서 사용된 인스턴스의 개수에 대해서 설명하고 있다. 훈련용 데이터인 경우에는 전체의 인스턴스에서 약 29%의 인스턴스가 선택 되었고 테스트용 데이터인 경우에는 전체의 인스턴스에서 약 32%의 인스턴스가 선택 되었다.

(표 4) 성능 평가를 위해 사용한 인스턴스의 개수
(Table 4) Number of instances for performance evaluation

	훈련 개수		테스트 개수	
	파킨슨병 환자	건강한 사람	파킨슨병 환자	건강한 사람
인스턴스 선택 전	487	210	487	210
인스턴스 선택 후	64	140	55	166

(표 5) 분류성능 비교
(Table 5) Comparisons of performance results

	분류 성능
인스턴스 선택 전	77.33 %
인스턴스 선택 후	78.19 %

(표 6) 짝진 범주형 자료의 분할표
(Table 6) McNemar's test contingency table

		인스턴스 선택 전		합계
		옳은 분류	틀린 분류	
인스턴스 선택 후	옳은 분류	166	6	172
	틀린 분류	0	49	49
합계		166	55	221

(표 7) 맥니마 검정에 의한 유의 확률
(Table 5) p-value of McNemar's test

	값	유의 확률
유효 케이스 개수	221	0.031

표 5에서는 인스턴스 선택 전/후 간의 성능 차이를 보여주고 있다. 인스턴스 선택을 사용한 경우에 분류 성능에 있어서 향상됨을 알 수가 있었다.

표 6은 인스턴스 선택 전/후에 있어서 분류 성능의 차이가 있음을 보여주는 맥니마 검정용 짝진 범주형 자료의 분할표이다. 표 6에서 인스턴스 선택 전에서는 틀리게 분류하였지만 인스턴스 선택 후에서는 옳게 분류한 인스턴스가 6개, 인스턴스 선택 전에서는 옳게 분류하였지만 인스턴스 선택 후에서는 틀리게 분류한 인스턴스가 0개로써 6개의 분류 성능 향상이 되었음을 알 수가 있다. 표 7에서는 맥니마 검정에 의한 유의 확률을 보여주고 있다. 맥니마 검정에 의한 유의 확률이 0.05보다 작기 때문에 5% 유의 수준에서 귀무가설을 기각할 수 있다. 이것은 인스턴스 선택 후의 분류 성능이 인스턴스 선택 전의 분류 성능과 차이가 있음을 보여주고 있고 이러한 사실은 표 5에서 보여주고 있는 인스턴스 선택 후의 분류 성능이 인스턴스 선택 전의 분류 성능보다 우수하다는 것을 통계학적으로 보여주고 있다.

6. 결 론

본 논문에서는 기존의 NEWFM에 새로운 인스턴스 선택을 지원하여 분류 성능을 향상시키는 방안을 제안하였다. 새롭게 제안하는 인스턴스 선택은 T-S 퍼지 모델에서의 가중 평균 역퍼지화와 통계학에서 사용하는 정규분포의 신뢰구간과 같은 구간 선택을 이용하였다. 첫 번째 단계에서는 모든 인스턴스를 이용하여 학습 과정을 수행하였고 두 번째 단계에서는 첫 번째 단계에서 학습된 결과를 기반으로 T-S 퍼지 모델에서의 가중 평균 역퍼지화와 통계학에서 사용하는 정규분포의 신뢰구간과 같은 구간 선택을 이용하여 인스턴스를 선택하였다.

본 논문에서는 인스턴스 선택의 분류 성능의 우수함을 평가하기 위해서 사례 연구로써 파킨슨병 환자의 족압 데이터를 이용하였다. 제안하는 인스턴스 선택의 성능을 평가하기 위해서 첫 번째 단계인 학습 과정 단계에서의 분류 성능과 두 번째 단계인 인스턴스 단계에서의 분류 성능을 비교하였으며 통계학적 검증인 맥니마 검정을 이용하여 인스턴스 선택의 분류 성능이 우수함을 평가하였다.

파킨슨병 환자의 족압 데이터를 적용하였을 때 인스턴스 선택 전/후의 분류 성능에 있어서 각각 77.33%과 78.19%로 나타났다. 따라서 인스턴스 선택에 있어서의 성능이 우수하다는 사실을 확인을 할 수가 있었다. 또한 인스턴스 선택 전/후 간의 성능 차이를 맥니마 검정으로 확인한 결과 유의 확률이 0.031로 나타났다. 따라서 맥니마 검정에 의한 유의 확률이 모두 0.05보다 작기 때문에 5% 유의 수준에서 귀무가설을 기각할 수 있다. 이것은 인스턴스 선택의 분류 성능이 인스턴서 선택을 하지 않은 경우의 분류 성능과 차이가 있음을 통계학적으로 보여주고 있다.

참 고 문 헌(Reference)

[1] Bell, G., Hey, T., and Szalay, A., "Beyond the data deluge," Science 323, pp.1297-1298, 2009.

[2] Yi Hong, Sam Kwong, Yuchou Chang, and Qingsheng Ren, "Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm," Pattern Recognition, Vol.41, pp.2742-2756, 2008.

[3] Minh Hoai Nguyen and Fernando de la Torre, "Optimal feature selection for support vector machines," Pattern Recognition, Vol.43, pp.584-591, 2010.

[4] José Martínez Sotoca and Filiberto Pla, "Supervised feature selection by clustering using conditional mutual information-baseddistances," Pattern Recognition, Vol.43, pp.2068-2081, 2010.

[5] Patricia E.N. Lutu and Andries P. Engelbrecht, "A decision rule-based method for feature selection in predictive data mining," Expert Systems with Applications, Vol.37, pp.602-609, 2010.

[6] S-M Zhou and J. Q. Gan, "Constructing L2-SVM-Based Fuzzy Classifiers in High-Dimensional Space With Automatic Model Selection and Fuzzy Rule Ranking," IEEE Trans. on Fuzzy Systems, Vol. 15, No. 3, pp. 398-409, 2007.

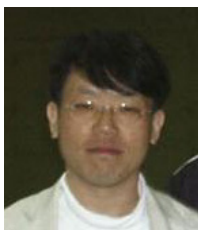
[7] Kudo, M. and Sklansky, J., "Comparison of algorithms that select features for pattern classifiers," Pattern Recognition 33, pp.25-41, 2000.

[8] Kuncheva, L.I., "Editing for the k-nearest neighbors rule by a genetic algorithm," Pattern Recognition Letters 16, pp.809-814, 1995.

[9] Joon S. Lim, "Finding Features for Real-Time Premature Ventricular Contraction Detection Using a Fuzzy Neural Network System," IEEE TRANSACTIONS ON NEURAL NETWORKS, vol.20, Issue 3, pp.522-527, 2009.

[10] Sang-Hong Lee and Joon S. Lim, “Forecasting KOSPI based on a neural network with weighted fuzzy membership functions,” Expert Systems with Applications, vol.38, Issue 4, pp.4259-4263, 2011.
[11] Sang-Hong Lee and Joon S. Lim, ″Parkinson’s disease classification using gait characteristics and wavelet-based feature extraction,″ Expert Systems with Applications, vol.39, Issue 8, pp.7338-7344, 2012.
[12] Takagi T. and Sugeno M., Fuzzy identification of system and its applications to modeling and control, IEEE Trans. Syst., Man, Cybern., SMC-15, (1985), 116-132.
[13] Koller, W.C., et al., ″Falls and Parkinson’s disease,″ Clin Neuropharmacol, vol.12, pp.98-105, 1989.
[14] C.-N. Lee, G.-M. Eom, K.-W. Park, S.-B. Koh, B.-J. Kim, K.-M. Oh,H.-J. Kim, and D.-H. Lee, ″Dynamic Foot Pressure Measurement in Parkinson’s Disease with Foot Scan System,″ J Korean Neurol Assoc, vol.25, No.2, pp.172-179, 2007.
[15] J.-W. Kim and G.-M Eom, ″Comparison of the Total Stance Time And the Phase Ratio in Parkinson’s Disease Patients And Normal Subjects,″ Journal of Biomedical Engineering Research, vol.27, No.6, pp.351-356, 2006.
[16] Sang-Hong Lee, Joon S. Lim, and Dong-Kun Shin, “Features Extraction for Classifying Parkinson’s Disease Based on Gait Analysis,” Journal of Internet Computing and Services, vol.11, No.6, pp.13-20, 2010.
[17] Kabir M, Shahjahan, and Murase K, "A new local search based hybrid genetic algorithm for feature selection," Neurocomputing 74, pp.2914-2928, 2011.
[18] Lee CP and Leu Y, "A novel hybrid feature selection method for microarray data analysis," Applied Soft Computing 11, pp.208-213, 2011.
[19] Tapia E, Bulacio P, and Angelone L, "Sparse and stable gene selection with consensus SVM-RFE," Pattern Recognition Letters 33, pp.64-172, 2012.
[20] Mejdoub M and Amar CB, "Classification improvement of local feature vectors over the KNN algorithm," Multimed Tools Appl 64, pp.197-218, 2011.
[21] Krishnamoorthy P and Kumar S, "Hierarchical audio content classification system using an optimal feature selection algorithm," Multimed Tools Appl 54, pp.415-444, 2011.
[22] Bing Xue, Mengjie Zhang, Will N. Browne, "Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms," Applied Soft Computing 18, pp.261-276, 2014
[23] Monami Banerjee and Nikhil R. Pal, "Feature selection with SVD entropy: Some modification and extension," Information Sciences 264, pp.118-134, 2014.
[24] Chih-Fong Tsai, Zong-Yao Chen, and Shih-Wen Ke, "Evolutionary instance selection for text classification," Journal of Systems and Software 90, pp.104-113, 2014.
[25] Tingting Zhai and Zhenfeng He, "Instance selection for time series classification based on immune binary particle swarm optimization," Knowledge-Based Systems 49, pp.106-115, 2013.
[26] Chih-Fong Tsai, William Eberle, and Chi-Yuan Chu, "Genetic algorithms in feature and instance selection," Knowledge-Based Systems 39, pp.240-247, 2013.

◐ 저 자 소 개 ◑



이 상 홍
1999년 경원대학교 전자계산학과(공학사)
2001년 경원대학교 일반대학원 전자계산학과(공학석사)
2012년 경원대학교 일반대학원 전자계산학과(공학박사)
2013년~현재 : 안양대학교 컴퓨터공학과 조교수
관심분야 : neuro-fuzzy system을 이용한 전문가 시스템.
E-mail : shleedosa@gmail.com, shleedosa@anyang.ac.kr