퍼지 가중치 평균 분류기를 위한 통계적 정보 기반의 가중치 설정 방안

신상호*, 조재현**, 우영운***
*동의대학교 디지털미디어공학과
**부산가톨릭대학교 컴퓨터공학과
***동의대학교 멀티미디어공학과
e-mail:false@deu.ac.kr, jhcho@cup.ac.kr, ywwoo@deu.ac.kr

Weight Adjustment Methods Based on Statistical Information for Fuzzy Weighted Mean Classifiers

Sang-Ho Shin*, Jae-Hyun Cho**, Young Woon Woo***
*Dept. of Digital Media Engineering, Dong-Eui University
**Dept. of Computer Engineering, Catholic University of Pusan
***Dept. of Multimedia Engineering, Dong-Eui University

요 약

패턴 인식에서 분류기 모형으로 많이 사용되는 퍼지 가중치 평균 분류기는 가중치를 적절히 설정함으로써 뛰어난 분류 성능을 얻을 수 있다는 장점이 있다. 그러나 일반적으로 가중치는 인식 문제 분야의 특성이나 해당 전문가의 지식이나 주관적 경험을 기반으로 설정되므로 설정된 가중치의 일관성과 객관성을 보장하기가 어려운 문제점을 갖고 있다. 따라서 이 논문에서는 퍼지 가중치 평균 분류기의 가중치를 설정하기 위한 객관적 기준을 제시하기 위하여 특징값들 간의 통계적 정보를 이용한 가중치 설정 기법들을 제안하였다. 제안한 기법들을 이용하여 UCI machine learning repository 사이트에서 제공되는 표준 데이터들 중의 하나인 Iris 데이터 세트를 이용하여 실험하고 그 결과를 비교, 분석하였다.

키워드 : 패턴 인식; 퍼지 가중치 평균 분류기; 가중치 설정; 통계적 정보

I. 서론

정보화 시대에 쉽게 얻을 수 있는 수많은 데이터들을 적절히 표현하는 것은 물론, 다양하고 방대한 자료를 분류하고 처리하기 위한 기술이 다양하게 발전되어 오고 있다. 그 중에서도 인식 기술은 정보를 검색하고 분류, 처리하기 위한 가장 뛰어난 기술이며, 인식기술의 대표적인 것이 패턴인식이다. 패턴인식은 심리학, 컴퓨터과학, 인공지능, 신경생물학, 언어학, 철학을 이용하여 지능적 인식 문제를 다루는 포괄적인 학제적 과학 분야인 인지과학과, 인간의 학습능력과 추론능력을 인공적으로 모델링하여 외부 대상을 지각하는 능력, 나아가 자연언어와 같은 구문적인 패턴까지 이해하는 능력을 컴퓨터 프로그램으로 구현하는 인공지능의 한 분야이다. 패턴인식의 대표적인 응용분야로는 문자인식, 음성인식, 생체인식, 행동패턴 분석, 의료영상 분석, 진단시스템, 도면인식, 예측시스템 등이 있으며, 최근에는 보안과 군사 분야뿐만 아니라 비정보기술 부문을 포함한 다양한 분야에서 활용되고 있다[1].

패턴 인식 기술에서 인식의 정확성을 향상시키기 위해서는 대상의 특성을 파악하여 그 특성에 따른 변별력이 높은 특징값을 추출하는 것과, 추출된 특징값들을 하나의 값으로 집계, 비교하여 인식 결과를 산출하기 위한 분류기 모형을 선택하고 세부 파라미터들을 적절히 구성하는 것이 필요하다. 현재 널리 활용되고 있는 분류기 모형들로는 신경망, 베이스 이론(Bayes theory), SVM(Support Vector Machine), 퍼지(fuzzy), Kernel, Spectral 등의 기법에 기반한 것들을 들 수 있다[2][3][4].

이 논문에서는 퍼지 기반의 분류기 모형들 중의 하나인 퍼지 가중치 평균 분류기의 성능을, 문제 분야의 특성에 관계없이 개선시킬 수 있는 방법의 일환으로 가중치를 설정하기 위한 기법들을 제안하였다. 이를 위하여 특징값들 간의 관계를 파악할 수 있는 다양한 통계적 정보들을 활용하였으며, 제안된 기법들을 UCI(University of California, Irvine) machine learning repository 사이트[5]에서 제공되는 표준 데이터들 중의 하나인 Iris 데이터 세트를 이용하여 실험하고 그 결과를 비교, 분석하였다. 논문의 전체 구성은 다음과 같다. 2장에서 제안 기법을 설명하고 3장에서는 제안 기법들을 이용한 실험 및 분석 결과를 제시한 후 4장에서 결론을 맺는다.

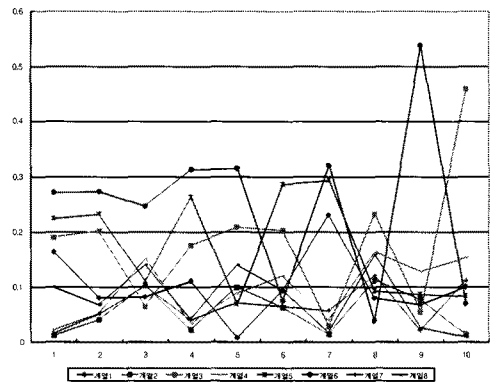
II. 관련 연구 및 제안 배경

패턴 분류 문제들에서 추출된 특징값들을 이용하여 하나의 클래스를 결정짓기 위한 기법들은 보편적으로 유사도 척도(similarity measure)를 이용하여 학습된 기준 데이터와 입력되는 테스트 데이터의 유사도를 비교하여 가장 유사한 클래스에 소속하는 데이터라고 결정짓는 방법을 활용하고 있다. 이 때 기준 학습 데이터의 특징값 벡터와 입력된 테스트 데이터의 특징값 벡터의 유사도를 비교하기 위하여 Minkowsky distance, Euclidean distance, City-block distance, Mahalanobis distance 등의 거리 척도들이 일반적으로 이용되고 있다[2].

그러나 이들 중 Mahalanobis distance를 제외하고는 특징값 벡터의 모든 요소들을 동일한 수식에 의해 계산을 수행한 후 모든 거리값의 합이 가장 작은 클래스를 결과 클래스로 산출하는 방식을 사용하고 있다. Mahalanobis distance는 특징값 벡터의 각 요소가 각 클래스의 해당 차원(dimension)의 클래스 내 분산(within-class variance)을 이용하여 클래스 분포의 특징을 거리에 반영하는 장점을 갖지만, 이 또한 모든 거리값들을 하나의 값으로 결정하기 위하여 개별적으로 산출된 거리값들의 단순 합을 최종 결과값으로 이용하고 있다.

이 논문은 분류를 위해 특징값 벡터의 각 요소들의 거리값들이 산출되었다는 가정 하에 각 요소들의 거리값들을 단순히 합으로 계산하여 비교하지 않고, 각 요소들의 거리값에 기준 클래스들의 평균과 분산 등의 통계적 정보를 가중치로 활용한 가중치 평균 기법으로 최종 결과값을 산출하여 사용한다면 분류 성능을 더욱 높일 수 있을 것이라는 아이디어에서 시작하였다. 예를 들어 8개의 클래스와 각 클래스별로 10개의 요소로 이루어진 특징값 벡터로 이루어진 문제 분야의 학습된 기준 특징값들을 그래프로 나타낸 그림 1을 살펴보도록 하자.

그림 1에 나타나 있는 것처럼 10차원 벡터의 특징값을 갖는 8개 클래스의 특징값 분포를 보면 6번째 특징값들의 클래스간 분포가 8번째 특징값들의 클래스간 분포보다 더욱 넓은 간격으로 벌어져 있음을 알 수 있다. 이는 6번째의 특징 요소가 8번째 보다는 클래스간의 구별을 위한 변별력이 더욱 높은 것으로 생각할 수 있다. 따라서 저자의 초기 연구에서는 이러한 개념을 구현하기 위하여 클래스간의 분포가 더 넓은 정도를 가중치로 활용한 가중치 평균을 산출하면 더욱 분류 성능을 높일 수 있을 것이라고 예상하고, 식 (1)과 같이 표현되는 각 특징값들의 8차원 세로 벡터에 대한 분산을 이용하여 기존 연구들과 비교 실험하여 분류 성능이 더욱 높아짐을 확인하였다[6].



〈그림 1〉 8개 클래스와 10개의 특징값으로 구성된 문제 예시

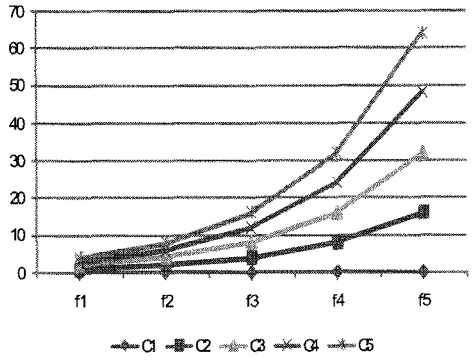
$$vr_j = \frac{1}{c}\sum_{i=1}^{c}(x_{ij} - m_j)^2 \quad \cdots\cdots (1)$$

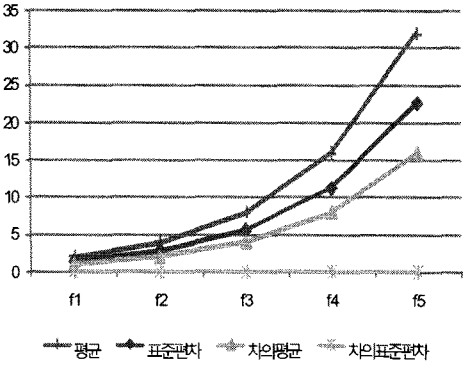
여기서 c는 클래스 수, x_{ij}는 i번째 클래스의 j번째 특징값, m_j는 j번째 특징값들의 평균을 의미한다.

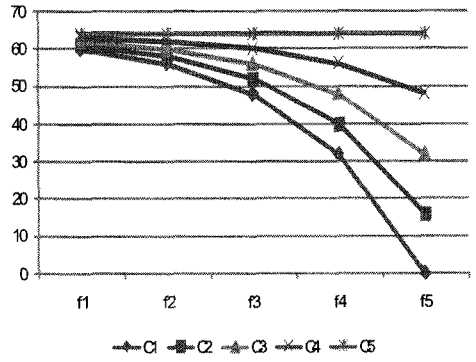
그러나 [6]의 연구에서는 분산에 의한 가중치 개념을 적용하지 않는 결과들보다 좋은 성능을 보이긴 하였으나, 실제 식 (1)에 의해 분산을 계산해 보면 9번째 특징값들에 대한 분산이 4번째나 6번째 보다 훨씬 높게 나옴을 알 수 있다. 즉 모든 특징값들이 고루 넓게 퍼져 있음을 반영하기 위하여 분산을 사용하였지만, 7개의 클래스가 모여 있고 하나의 클래스 값만이 아주 멀리 떨어진 9번째와 같은 경우에는 7개의 클래스에 대한 변별력이 결코 높은 것은 아니기 때문에 9번째 특징값들의 가중치를 낮추고 4번째나 6번째 특징값들의 가중치를 상대적으로 높일 수 있는 새로운 척도가 있으면 분류 성능을 높일 수 있을 것으로 생각하고 가중치 설정을 위한 몇 가지 척도를 제안하고 비교 실험하였다.

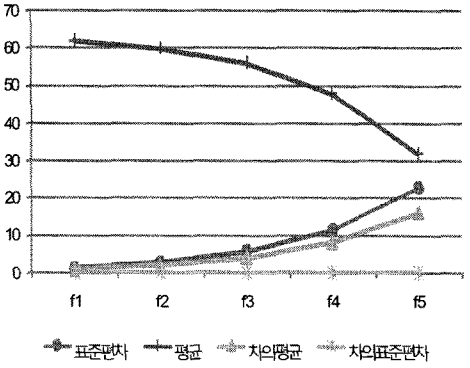
III. 제안한 가중치 설정 기법

특징값들의 분포를 이용하여 가중치를 합리적으로 산출하기 위한 기법을 위하여 그림 2와 그림 3과 같이 특징값들의 분포를 고려한다. 그림 2는 클래스들 간의 특징값 간격이 동일하지만 그 간격이 서로 다른 5가지 경우(f1 ~ f5)를 각각 나타낸 그림이고, 그림 3은 클래스들 간의 특징값 간격이 모두 같은 것부터 하나씩 차이가 나는 5가지 경우(f1 ~ f5)를 각각 나타낸 그림이다.



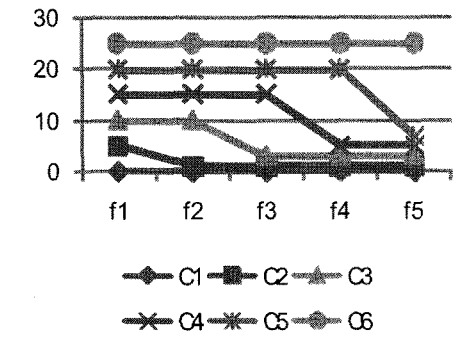


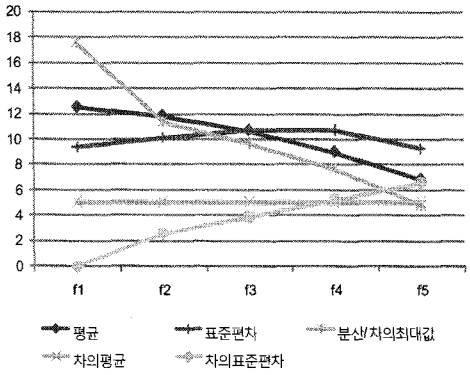


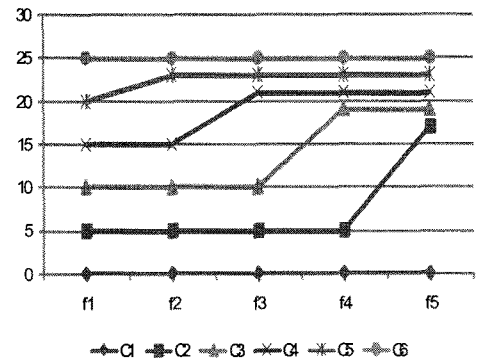


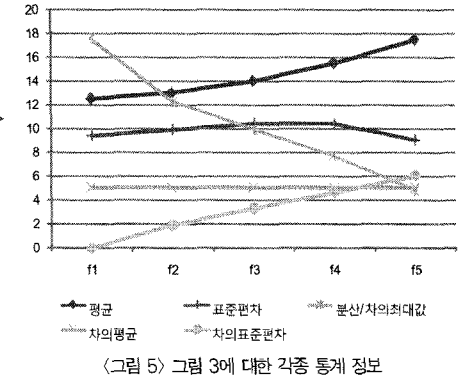
〈그림 2〉 클래스 간에 동등한 간격으로 구성된 특징값 분포 예

〈그림 4〉 그림 2에 대한 각종 통계 정보









〈그림 3〉 클래스 간에 차등 간격으로 구성된 특징값 분포 예

〈그림 5〉 그림 3에 대한 각종 통계 정보

그림 2의 경우에는 각 클래스간의 간격이 동일하므로 전체적으로 간격이 큰 특징값들의 변별력이 우수한 것으로 판단할 수 있으므로 간격이 큰 특징값 세로 벡터가 큰 가중치를 갖도록 하는 척도가 요구된다. 또한 그림 3의 경우에는 각 특징값들의 클래스 간의 간격이 변하는 경우이므로 특징값들의 분포가 고른 것이 변별력이 우수한 것으로 판단할 수 있으므로 간격이 고른 특징값 세로 벡터가 큰 가중치를 갖도록 하는 척도가 필요할 것으로 판단된다.

이를 위하여 그림 2의 경우를 위한 척도로는 각 특징값들간 차이들의 평균(*MD*)을 고려하였으며, 그림 3의 경우를 위한 척도로는 각 특징값들의 분산(*V*), 각 특징값들 간 차이들의 분산(*VD*), 각 특징값들의 분산을 가장 큰 특징값 차이로 나눈 값(*VPDM*)을 고려하였다. *V*를 산출하는 식은 수식 (1)로 이미 제시되었으며, 수식 (2)는 *MD*, 수식 (3)은 *VD*, 그리고 수식 (4)는 *VPDM*를 구하기 위한 수식이다.

$$MD_j = \frac{1}{c-1}\sum_{i=1}^{c-1}(x_{(i+i)j} - x_{ij}) \quad \cdots\cdots (2)$$

$$VD_j = \frac{1}{c-1}\sum_{i=1}^{c-1}[(x_{(i+i)j} - x_{ij}) - MD_j]^2 \quad \cdots\cdots (3)$$

$$VPDM_j = \frac{\frac{1}{c}\sum_{i=1}^{c}(x_{ij} - m_j)^2}{DiffMax_j} \quad \cdots\cdots (4)$$

여기서, $DiffMax_j = max(x_{(i+1)j} - x_{ij}),\ i = 1..c$

위 수식들에서 c는 클래스 개수이며 j는 특징값 벡터의 차수를 나타낸다. 또한 x_{ij}값들은 i를 기준으로 오름차순으로 정렬되어 수식에 적용된다. 그림 2와 그림 3에 대하여 이상의 수식들과 추가 통계 정보들에 의해 산출된 결과값을 그림 4와 그림 5에 나타내었다. 결과 그림에서 분산 대신 표준편차를 사용하여 나타낸 것은 비교 대상들의 값 차이를 줄여 비교를 용이하게 하기 위함이다.

이상과 같은 척도들을 이용하여 합리적이라고 판단되는 가중치들을 다음과 같은 4가지 수식으로 제안하고 기존에 제안된 분산과 가중치를 고려하지 않는 Euclidean distance 방법을 포함하여 모두 6가지를 비교, 실험하였다. 처음 2가지는 *VPDM* 활용에 중점을 둔 것이며 다음 2가지는 *VD* 활용에 중점을 둔 것이다.

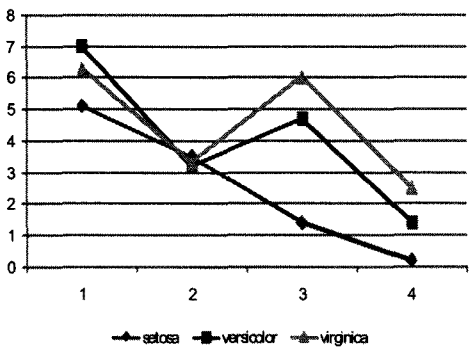
① *VPDM*
② *MD* + *VPDM*
③ 1 - *VD*
④ *MD* + (1 - *VD*)
⑤ *V*
⑥ Euclidean distance

IV. 실험 및 결과 고찰

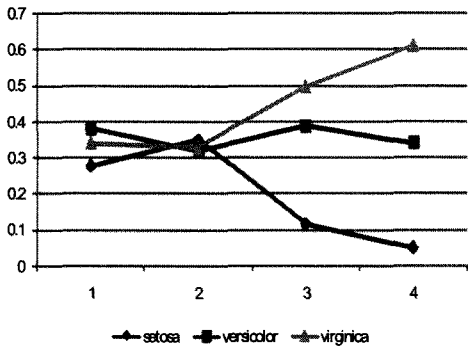
실험을 위하여 UCI(University of California, Irvine) machine learning repository 사이트에서 제공되는 표준 데이터들 중의 하나인 Iris 데이터 세트를 사용하였다. Iris 데이터 세트는 3개의 클래스와 4개의 특징값으로 구성된 벡터가 클래스 별로 50개씩 존재하는 구조이다. 각 클래스의 학습 데이터에 대한 테스트 데이터의 거리는 수식 (5)와 같은 동일한 기울기를 갖는 삼각형 모양의 퍼지 함수 소속도를 이용하여 산출하였다[7]. Iris 데이터 세트를 이용한 실험에서는 학습 데이터가 클래스 별로 복수개의 특징값 벡터가 사용되는데 이 논문에서는 사용되는 특징값 벡터들의 평균값만을 퍼지 함수의 중심값으로 활용하였다.

$$\begin{aligned}
&\mu_{ij}(x_j) = 0.01(x_j - f_{ij}) + 1 \quad \text{if } x_j < f_{ij} \\
&\mu_{ij}(x_j) = -0.01(x_j - f_{ij}) + 1 \text{ if } x_j \geq f_{ij} \quad \cdots\cdots (5) \\
&\mu_{ij}(x_j) = 0 \quad \text{if } \mu_{ij}(x_j) < 0
\end{aligned}$$

그림 6은 Iris 데이터 세트에서 각 클래스의 특징값 벡터 하나씩을 보여주고 있으며, 그림 7은 가중치 계산에 사용되는 통계적 정보를 얻기 위한 정규화된 특징값 벡터를 나타낸다.



〈그림 6〉 Iris 데이터 세트의 특징값 벡터



〈그림 7〉 Iris 데이터 세트의 정규화된 특징값 벡터

이상에서 언급한 6가지 경우의 가중치 설정 방법으로 Iris 데이터를 분류한 실험 결과는 표 1과 같다. 표 1의 결과값들은 각각의 경우에 대하여 10-fold cross validation 방법[8]으로 실험한 평균 결과를 나타낸다. 이 때 가중치들의 설정 방법 차이에 따른 결과들만을 비교하기 위하여 90%의 학습 데이터들로부터 얻어지는 클래스내 분산은 사용하지 않았으며, 학습 데이터들의 평균값만을 클래스 기준 데이터로 사용하여 분류 실험을 수행하였다.

〈표 1〉 Iris 데이터 세트에 대한 각 방법의 분류 실험 결과

가중치 설정 방법	인식률
VPDM	97.3%(146/150)
MD + *VPDM*	96.7%(145/150)
1 - *VD*	94%(141/150)
MD + (1 - *VD*)	94.7%(142/150)
V	95.3%(143/150)
Euclidean distance	90%(135/150)

표 2는 Iris 데이터 세트에 대하여 기존에 많이 사용되는 분류기들을 이용하여 실험한 결과를 나타낸다[9]. 표 1의 결과와 비교해 보면 알 수 있듯이, 이 논문에서 실험한 방법은 대부분의 기존 기법들이 활용하고 있는 클래스 내 분산 정보를 전혀 이용하지 않은 상태에서도 제안한 가중치 설정 기법만으로도 기존 기법들과 유사하거나 더욱 우수한 성능이 나타남을 확인할 수 있었다.

〈표 2〉 Iris 데이터 세트에 대한 기존 분류 기법의 인식률[9]

분류 기법	인식률
1-NN	96%
Naive Bayes	96%
BayesNet	94.7%
C4.5	94.7%

이 논문에서는 클래스 소속도를 산출하기 위하여 모두 동일한 가우거의 퍼지 함수를 사용하였지만, 만약 클래스 내 분산 정보와 각 분산 정보들의 상관 관계 등을 활용하여 학습 데이터와 테스트 데이터들 간의 클래스 소속도를 특징값 벡터 요소에 대해 가변적으로 할당할 수 있는 퍼지 기법과 제안된 가중치 설정 기법이 함께 사용된다면 더욱 우수한 성능이 나올 것으로 기대한다.

하지만 실험에서 사용된 Iris 데이터 세트는 2가지 클래스가 비교적 잘 구별되는 특성을 갖고 있으며, 데이터 샘플수가 적은 편이어서 기법에 따른 성능이 크게 나타나지 않는 단점이 있다. 따라서 제안한 기법의 성능을 더욱 정확하게 평가할 수 있는 anneal, soybean 등의 표준 데이터 세트를 활용할 필요가 있을 것으로 생각한다.

V. 결론

이 논문에서는 퍼지 가중치 평균 분류기의 성능을, 문제 분야의 특성에 관계없이 개선시킬 수 있는 한 가지 방법으로 가중치를 산출하는 기법을 제안하였다. 수학적으로 합리적이라고 판단될 수 있는 가중치를 산출하기 위하여 특징값들 간의 관계를 나타내는 통계적 정보들을 활용하였다. 이를 위하여 여러 종류의 특징값 분포에 대한 모형을 활용하여 특징을 분석하였으며 가중치에 사용될 수 있는 통계적 척도 3가지를 제안하였다. 첫 번째 척도로 "각 특징값들 간 차이들의 평균", 두 번째 척도로 "각 특징값들의 분산을 가장 큰 특징값 차이로 나눈 값", 그리고 마지막 척도로 "각 특징값들 간 차이들의 분산" 등을 제안하였다. 제안된 기법을 이용하여 가중치 설정을 위한 수식 4가지를 제안하였으며, 이 4가지 기법들과 기존에 사용되었던 단순 분산과 Euclidean 거리 기법의 2가지를 실험하고 그 결과를 비교, 분석하였다.

실험에 사용된 표준 데이터 세트로는 UCI(University of California, Irvine) machine learning repository 사이트[5]에서 제공되는 Iris 데이터 세트를 이용하였다. 실험 결과 클래스 소속도를 나타내는 거리 척도로 동일한 퍼지 함수의 소속도 값을 사용함에도 불구하고, 최종 분류를 위한 결과값 산출에 제안된 가중치를 적용한 평균을 활용함으로써 기존에 사용되는 기법들과 유사하거나 그 이상의 성능이 달성됨을 알 수 있었다.

따라서 기존의 기법들에서 활용하고 있는 클래스 내 분산 정보를 이용하여 소속도를 산출한 다음 그 소속도에 제안된 가중치 평균 기법을 적용한다면 더욱 우수한 성능이 산출될 것으로 예상한다.

향후 연구 과제로는 클래스 내 통계 정보를 활용하여 더욱 합리적인 클래스 소속도를 할당하는 기법을 보완하는 것과 더욱 다양한 표준 실험 데이터 세트를 이용하여 제안한 기법의 정당성과 일반성을 검증하는 것이 필요할 것으로 생각한다.

References

[1] 한학용, 패턴인식개론, 한빛미디어, 2005.

[2] Rui Xu and Donald Wunsch II, "Survey of Clustering Algorithms," IEEE Transactions on Neural Networks, Vol.16, No.3, pp.645-678, 2005.

[3] Sanjeev R. Kulkarni, Gabor Lugosi, and Santosh S. Venkatesh, "Learning Pattern Calssification - A Survey," IEEE Transactions on Information Theory Vol.44, No.6, pp.2178-2206, 1998.

[4] Anil K. Jain, Robert P. W. Duin, and Jianchang Mao, "Statistical Pattern Recognition: A Review," IEEE Transactions on Pattern Analysis and Machine Intelligence Vol.22, No.1, pp.4-37, 2000.

[5] http://archive.ics.uci.edu/ml/datasets.html

[6] Young Woon Woo, Imgeun Lee, Jongkeuk Lee, "Optimal Feature Selection and Performance Analysis on 2D Object Recognition System," Proceedings of MTIA2003, 2003.

[7] Timothy J. Ross, Fuzzy Logic with Engineering Applications(2nd ed.), Wiley, 2004.

[8] http://www.cs.cmu.edu/~schneide/tut5/node42 .html

[9] A. Küçükyilmaz, Pattern Classification: A Survey and Comparison, Department of Computer Engineering, Bilkent University, Ankara, Turkey. April 2005.