

# 고차원을 갖는 생체 스펙트럼 데이터의 특징추출 및 분류기법

## Feature Extraction and Classification of High Dimensional Biomedical Spectral Data

조재훈 · 박진일 · 이대종 · 전명근<sup>+</sup>

Jae Hoon Cho, Jin Il Park, Dae Jong Lee and Myung Geun Chun<sup>+</sup>

충북대학교 전기전자컴퓨터공학부

### 요 약

본 논문에서는 비선형 변환에 의해 입력신호를 고차원의 확장공간으로 변환한 후, 주성분분석기법(PCA)에 의해 신호의 특징을 추출하는 기법을 제안한다. 특징추출을 위해 사용되는 기존의 주성분분석기법은 입력데이터가 비선형 특성을 갖는 경우 최적의 변환행렬을 구할 수 없다는 문제점을 가지고 있다. 이러한 문제점을 해결하기 위해, 확장공간상에서 구간별로 입력데이터를 분할한 후 주성분분석기법에 의해 구간별 특징을 추출하는 서브패턴기반 주성분분석기법(SpPCA)을 적용하였다. 다음 단계인 분류단계에서는 MLP 비선형분류기를 이용하여 구간마다 추출된 특징벡터를 이용하여 기준패턴과의 유사도를 산출한다. 최종 분류단계에서는 MLP에 의해서 산출된 유사도에 기반을 둔 융합법칙에 의하여 생체 스펙트럼 패턴을 분류한다. 제안된 방법의 유용성을 보이기 위한 실험결과에서 기존의 방법들에 비해서 향상된 인식결과를 보임을 확인하였다.

**키워드** : PCA, SpPCA, MLP, MRS pattern classification

### Abstract

In this paper, we propose the biomedical spectral pattern classification techniques by the fusion scheme based on the SpPCA and MLP in extended feature space. A conventional PCA technique for the dimension reduction has the problem that it can't find an optimal transformation matrix if the property of input data is nonlinear. To overcome this drawback, we extract features by the SpPCA technique in extended space which use the local patterns rather than whole patterns. In the classification step, individual classifier based on MLP calculates the similarity of each class for local features. Finally, biomedical spectral patterns is classified by the fusion scheme to effectively combine the individual information. As the simulation results to verify the effectiveness, the proposed method showed more improved classification results than conventional methods.

**Key Words** : PCA, SpPCA, MLP, MRS pattern classification

## 1. 서 론

자기공명 스펙트럼 분석기에 의해서 취득된 MRS (magnetic resonance spectra) 생체신호는 높은 입력차원을 가지고 있을 뿐만 아니라, 데이터들간 유사성이 매우 높기 때문에 패턴을 분류하는데 있어서 많은 어려움을 가지고 있다[1]. 또한, 입력차원이 높은 신호를 데이터의 축소 없이 패턴을 분류할 경우 데이터 저장용량의 증가 및 많은 계산 처리에 의한 속도 저하를 초래한다. 이러한 문제점으로 인하여 입력차원이 높을 경우 중요한 특성만을 추출하는 다양한 특징선택 또는 특징추출기법이 제안되고 있다[2],[3].

특징추출기법으로는 주성분분석기법(principal component analysis; PCA)과 선형판별분석기법(linear discrim-

inant analysis; LDA) 등이 있다[4]-[7]. 선형판별분석기법은 분류를 위해 최적의 투영행렬을 찾는 방법 중의 하나로 투영된 분산행렬을 최대화하는 투영행렬을 찾기보다는 클래스간 분산(between-class scatter matrix) 행렬과 클래스내 분산(within-class scatter matrix) 행렬의 비가 최대가 되는 최적의 투영행렬을 찾는 것이다. 또한, 클래스간 데이터가 유사한 분산행렬을 갖지만 평균값이 상이한 경우에 대해서 최적의 해를 구할 수 있다. 그러나, 본 논문에서 고려되는 MRS 패턴과 같은 고차원의 데이터들이 위의 조건을 만족하는 경우는 극히 한정되어 있다.

PCA는 다변수 선형 데이터 분석에서 잘 알려진 대표적인 기법이며 주된 개념은 데이터의 최대 분산 방향을 나타내는 상호직교 기저벡터의 집합을 찾는 것을 목적으로 한다. 구해진 기저벡터행렬을 이용하여 입력공간을 저차원의 특징공간으로 선형적으로 투영시킨다. 이 방법은 차원축소에는 효과적으로 사용되고 있으나 투영된 데이터간의 겹치는 현상으로 인하여 분류문제에서는 효과적이지 못하다. 또한, 기존의 주성분분석기법은 모든 입력차원의 선형성에 기반을 두고 있기 때문에 입력데이터가 구간적으로 선형 특성

접수일자 : 2009년 1월 29일

완료일자 : 2009년 5월 14일

+ 교신저자

본 연구는 보건복지가족부 보건의료기술진흥사업의 지원에 의하여 이루어진 것임 (과제고유번호 : A040032).

을 보이지만 모든 차원에 대해서 비선형 특성을 갖는 경우 최적의 변환행렬을 구할 수 없는 문제점을 지니고 있다[8].

이러한 문제점을 해결하기 위해 본 논문에서는 주성분분석기법에 의해 특징을 추출하기 전에 입력신호를 비선형변환에 의해 확장공간으로 매핑한 후 특징을 추출하는 기법을 제안하고자 한다. 또한, 기존의 전구간의 데이터를 대상으로 특징을 추출하는 방법 대신에 입력데이터를 구간별로 분할한 후 특징을 추출하는 서브 패턴 기반 주성분분석기법(SpPCA)을 적용하였다[9]-[11]. SpPCA 기법을 이용하여 차원을 축소한 후 최종 분류를 위해 MLP (multi-layer perceptron) 비선형분류기를 사용한다. 또한 SpPCA을 이용하여 구한 특징벡터를 융합하는 문제에 있어서 기존 연구는 전구역 특징벡터(global feature)에 기반을 둔 방법을 채택한 반면에 본 논문에서는 가중치합법칙(sum rule), 우수 후보선출법칙(best candidate vote rule), 다수선출법칙(majority vote rule) 등 다각적인 융합법칙을 적용하여 패턴 특징에 맞는 결정법칙을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 제안된 확장공간상에서 SpPCA와 MLP에 기반을 둔 융합법칙에 의한 분류기법에 대하여 기술하고, 3장에서는 생체 스펙트럼 데이터에 제안된 방법을 이용함으로써 적용된 방법의 우수성을 보인다. 마지막으로 4장에서는 결론을 제시한다.

## 2. 확장공간상에서 부 패턴 특징 추출 및 융합기법

그림 1은 본 논문에서 제안한 확장공간상에서 생체 데이터의 특징 추출 및 분류방법을 나타냈다. 그림 1에서 보는 바와 같이 생체 스펙트럼 데이터는 우선적으로 실험에 의해 결정된 확장공간으로 변환된다. 다음 단계로 확장된 공간을  $k$  개의 부 공간으로 분할한 후 각각의 분할된 영역에서 서브 패턴 기반의 주성분분석기법을 적용하여 지역적 특징을 추출한다. 다음 단계로 부 공간내에서 추출된 특징들을 이용하여 분류기에 의해 각각의 유사도를 산출한다. 최종단계에서는 효과적인 융합법칙에 의해 생체 스펙트럼 데이터를 분류하게 된다. 각각의 과정에 대하여 서술하면 다음과 같다.

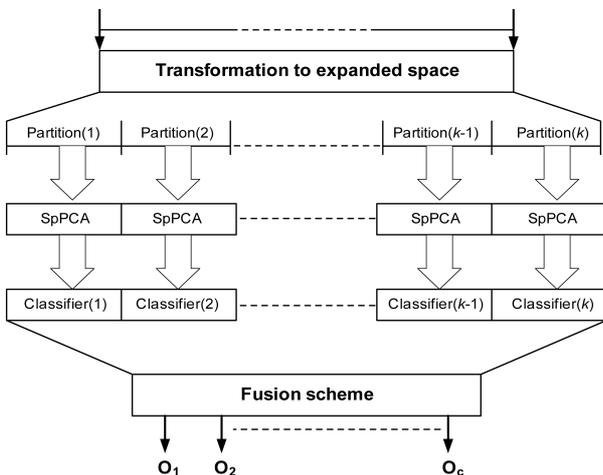


그림 1. 제안된 생체 스펙트럼 데이터의 분류기법  
Fig. 1. Proposed classification method for biomedical spectra data

### 2.1 입력공간의 확장 및 SpPCA에 의한 지역적 특징 추출

PCA 기법은 다변수 선형 데이터 해석에서 잘 알려진 기법이다. PCA 기법은 데이터의 분산이 최대화 되는 변환행렬을 구하고 이 변환행렬에 의해 고차원의 입력공간을 저차원의 공간으로 축소하는 기법으로 선형 특징추출과 데이터 압축 등에 널리 사용되고 있다.

상세한 설명을 위해 다음과 같은  $N$  개의 샘플을 갖는 학습데이터를 고려하자.

$$X = (x_1, x_2, \dots, x_N) \quad (1)$$

여기서,  $x_k = [x_{k1}, x_{k2}, \dots, x_{kd}]^T$ 는  $d$ 차원을 갖는 데이터이다. 식 (1)에서 보인 입력패턴에 대해서 특징추출을 하기 전에 식(2)와 같이 비선형변환을 한다. 이와 같이 입력패턴을 비선형변환에 의해 고차원으로 매핑함으로써 데이터간에 존재할 수 있는 비선형특성을 보상하고자 한다.

$$z_i = T(x_i) \quad (2)$$

비선형 변환된 새로운 공간  $z_i$ 에 대한 공분산행렬은 다음과 같다.

$$R = \frac{1}{N} \sum_{i=1}^N (z_i - \bar{m})(z_i - \bar{m}) \quad (3)$$

$$\bar{m} = \frac{1}{N} \sum_{i=1}^N z_i \quad (4)$$

다음으로 공분산  $R$ 의 고유값과 고유벡터를 계산한다.  $W = (w_1, w_2, \dots, w_r)$ 를  $r$ 개의 가장 큰 고유값에 대응하는  $r$ 개의 고유벡터라 하자. 그러면 원래의 샘플  $Z$ 의 집합에 대해서 그에 대응하는 특징벡터  $Y = (y_1, y_2, \dots, y_N)$ 는 다음과 같이  $Z$ 를 PCA-변환 공간으로 투영함으로써 얻어진다.

$$y_i = W^T(z_i - \bar{m}) \quad (4)$$

여기서,  $W$ 가  $p \times r$  행렬이고  $z_i$ 가  $i$ 번째  $p \times 1$  벡터이면  $y_i$ 는  $i$ 번째  $r \times 1$  벡터이다.

SpPCA는 전구역 패턴보다는 구간마다 나누어진 분할패턴을 사용한다. SpPCA를 적용하기 위해서 원래의 데이터  $Z = (z_1, z_2, \dots, z_N)$ 는 다음과 같이 동일한 크기의 부패턴들의 집합으로 분할된다.

$$Z_j = (z_{j1}, z_{j2}, \dots, z_{jN}) \quad (5)$$

여기서,  $j$ 는 패턴지수이고,  $j = 1, 2, 3, \dots, L$  ( $L$ 은 분할의 수)이다. 다음으로 식 (6)과 (7)에 의해서  $j$ 번째 분할에 대한 공분산을 계산한다.

$$R_j = \frac{1}{N} \sum_{i=1}^N (z_{ji} - \bar{m}_j)(z_{ji} - \bar{m}_j)^T \quad (6)$$

$$\bar{m}_j = \frac{1}{N} \sum_{i=1}^N z_{ji} \quad (7)$$

일반적인 PCA기법을 이용한 특징추출 과정과 마찬가지로  $j$ 번째 분할에 대한  $r$ 개의 가장 큰 고유값에 대응하는  $r$ 개의 고유벡터  $W_j = (w_{j1}, w_{j2}, \dots, w_{jr})$ 를 결정한다. 마지막으로 식 (8)에서 나타낸 바와 같이 구간마다 나누어진 분할패턴에서 특징벡터를 구할 수 있다.

$$y_{ji} = W_j^T(z_{ji} - \bar{m}_j) \quad (8)$$

따라서,  $j$ 번째 분할  $Y_j = (y_{j1}, y_{j2}, \dots, y_{jN})$ 은 부패된 데이터에서 얻어질 수 있다. 패턴 분류를 위해 전구역 특징 그리고 구간 특징과 같은 두 형태의 특징 조합을 고려할 수 있다. 여기서 전구역 특징은  $Y = (Y_1, Y_2, \dots, Y_L)$ 와 같이 각각의 부패된 특징을 조합함으로써 얻어진다.

2.2 비선형 분류기 및 융합법칙

SpPCA에 의해 각각의 구간별로 특징벡터를 추출한 후 대표적인 선형분류기인 거리기반 측정법칙에 기반을 둔  $k$ -NN과 비선형분류기인 MLP를 구축하였다.  $k$ -NN을 위한 거리척도방법으로는 Euclidean, Hamming, 그리고 Mahananobis 거리척도를 사용하였다. MLP 비선형분류기는 입력층, 은닉층 그리고 출력층의 3층으로 설계된다. 여기서 입력수는 PCA에 의해 계산된 특징의 차원과 같다. 또한 전달함수로 기호함수(logistic function)를 사용한다.

다음 단계인 최종 분류를 위해서 구간별 특징벡터를 효과적으로 융합하여 인식률을 향상시킬 수 있는 융합분류기의 설계가 요구된다. 기존연구에서는 특징벡터를 모두 사용하는 전구간 특징벡터(global feature)에 기반을 둔 방식을 사용한 반면에 본 연구에서는 구간별로 독립적인 유사도를 산출한 후, 최종단계에선 유사도를 융합하는 방식을 채택하였다(그림1). 융합방법으로는 가중치합법칙(sum rule), 우수후보선출법칙(best candidate vote rule), 다수선출법칙(majority vote rule)을 적용하였다.

의사결정법칙의 상세한 설명을 위해,  $\Omega = \{w_1, w_2, \dots, w_c\}$ 는  $c$ 개의 클래스 라벨의 집합이라 하고,  $D_i = [d_1(O_i), d_2(O_i), \dots, d_c(O_i)]$ 는  $c$ -차원 벡터를 갖는 분류기 출력이라 하자. 여기서,  $O_i$ 는  $i$ 번째 입력 특징이다. 예를 들어,  $k$ 개로 입력공간을 분할 한 경우 각각의 구간에 대한 출력값은 다음과 같다.

- 1번째 분할영역:  $o_1[1], o_2[1], \dots, o_c[1]$
- 2번째 분할영역:  $o_1[2], o_2[2], \dots, o_c[2]$
- ⋮
- $k$ 번째 분할영역:  $o_1[k], o_2[k], \dots, o_c[k]$

최종 분류단계에서는 각각의 구간별로 산출된 출력값들을 효과적으로 융합할 수 있는 방법이 요구된다. 논문에서는 가중치합법칙, 우수후보선출법칙과 다수선출법칙의 세 가지 융합법칙을 적용한다.

가중치합법칙에 기반을 둔 의사결정은 모든 분할영역에 대한 각 클래스의 유사도의 합으로 계산된 구간별 출력값을 계산한 후 식 (9)에 의해 이들 중에서 가장 우수한 유사도를 갖는 클래스를 선택함으로써 수행된다.

$$O_i = \sum_{j=1}^k o_i[j] \tag{9}$$

우수후보 선출방법은 식 (10)에서 보는 바와 같이 모든 분할영역에 대한 각 클래스의 유사도들 사이의 가장 우수한 유사도를 갖는 클래스를 선택함으로써 수행된다.

$$\begin{aligned} O_i &= \min(o_i[1], o_i[2], \dots, o_i[k]), i = 1, 2, \dots, c \text{ (for } k\text{-NN)} \\ O_i &= \max(o_i[1], o_i[2], \dots, o_i[k]), i = 1, 2, \dots, c \text{ (for MLP)} \end{aligned} \tag{10}$$

다수선출법칙은 식 (11)에서 보는 바와 같이 각각의 분할 영역에 대해서 가장 우수한 유사도를 갖는 클래스의 개수의

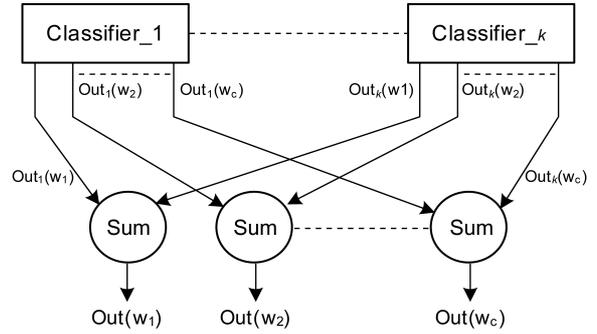
합으로 계산된 지수들 중에서 최대 지수 값을 갖는 클래스를 선택함으로써 수행된다.

$$e_i(j) = \begin{cases} 1 & \text{if } e_i(j) = \min(e_1(j), e_2(j), \dots, e_c(j)) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

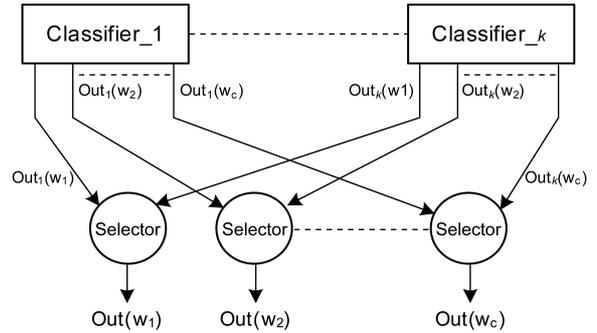
$$O_i = \sum_{j=1}^k e_i[j]$$

마지막으로, 분류는 식 (12)에 의해서 클래스들 사이의 가장 우수한 유사도를 갖는 클래스를 선택함으로써 얻어진다.

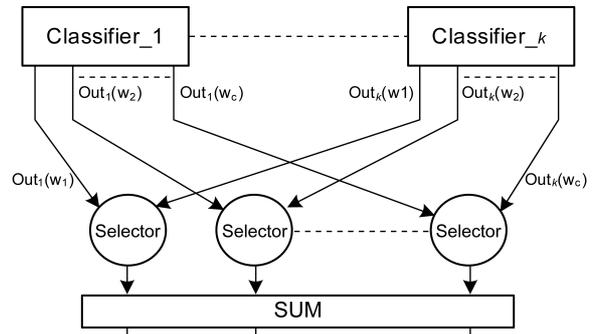
$$i_o = \begin{cases} \arg \min_i (d_i) & \text{for } k\text{-NN}, \forall i, i = 1, 2, \dots, c \\ \arg \max_i (d_i) & \text{for } k\text{-MLP}, \forall i, i = 1, 2, \dots, c \end{cases} \tag{12}$$



(a) 가중치합법칙 (sum rule)



(b) 우수후보선출법칙 (best candidate vote rule)



(c) 다수선출법칙 (majority vote rule)

그림 2. 개별 분류기에 기반을 둔 융합법칙  
Fig. 2. Fusion rules based on individual classifier

### 3. 실험 결과

제안된 방법의 타당성을 보이기 위해 본 논문에서 생체 스펙트럼 데이터를 사용하였다. 사용된 MRS 데이터는 수막종 M(menigiomas), 별아교세포종 A(astorcytomas) 및 간질환자의 뇌조직으로부터 추출한 비종양 E(non-tumourous) 샘플로부터 취득하였다[12]. 3개의 클래스를 갖는 MRS 데이터의 수는 95(M), 74(A), 37(E) 이며 하나의 데이터 당 550의 차원을 갖는다. 그림 3에 각각의 클래스별 패턴 과형을 나타냈다.

데이터 중에서 훈련에 사용된 데이터는 각각 47(M), 37(A), 19(E)로 103개이며 나머지 103개 데이터를 이용하여 제안된 방법의 성능을 분석하였다. 본 실험은 보다 정확한 결과의 해석을 위해서 학습데이터와 검증데이터를 임의로 10번씩 선택하여 총 10회의 실험을 하였다.

특징을 추출하기 전에 입력공간을 비선형변환에 의해 고차원으로 매핑한다. 본 실험에서는 비선형 변환을 위해 다음 식을 사용하였다.

$$z_i = [\sqrt{x_{i1}}, \sqrt{x_{i2}}, \dots, \sqrt{x_{id}}, \sqrt[3]{x_{i1}}, \sqrt[3]{x_{i2}}, \dots, \sqrt[3]{x_{id}}, \sqrt[4]{x_{i1}}, \sqrt[4]{x_{i2}}, \dots, \sqrt[4]{x_{id}}]$$

또한 분류기를 적용하기 전에 주성분분석기법을 이용하여 특징벡터를 추출하고 이 과정에서 미리 축소된 특징벡터의 개수를 설정하여야 한다. 본 연구에서 PCA에 의해서 선택된 특징벡터의 수는 식 (13)과 같이 크기순으로 구해진 고유값의 합계와 선택하고자하는  $p$ 개의 개수만큼 선택된 고유값의 비(thr)가 0.99, 0.995, 0.998을 만족하는 개수  $p$ 를 특징벡터의 수로 설정하였다.

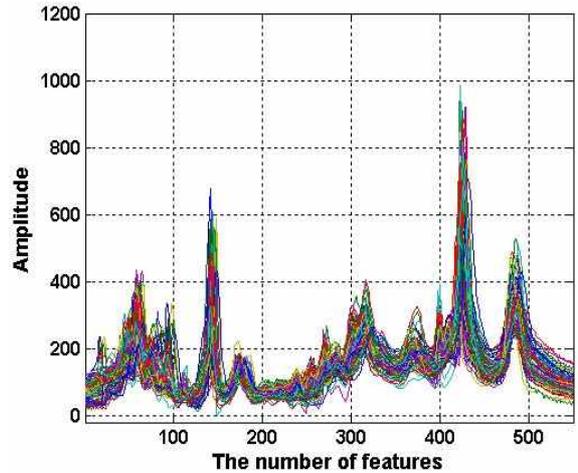
$$thr \geq \frac{\lambda_1 + \lambda_2 + \dots + \lambda_p}{\sum_{i=1}^p \lambda_i} \quad (13)$$

분류기로는 기존의 k-NN 선형분류기 외에 비교를 위해서 MLP 비선형분류기를 사용하였다. 우선, k-NN 분류기는 선택된 k의 값에 따라서 영향을 받기 때문에 k 변화에 따른 선형분류기의 특성을 살펴볼 필요가 있다. 이를 위해서 k의 값을 1, 3, 5 등으로 변경하여 성능을 평가하였다. 비선형 분류기인 MLP의 구조에서 입력차원은 PCA 기법에 의해서 축소된 특징벡터의 차원과 동일하다. 또한, 은닉층의 개수는 20, 학습률은 0.005, 모멘트율은 0.008로 설정하였다. 이러한 파라미터들은 다양한 실험을 통해서 우수한 성능을 보이는 값들로 설정하였다.

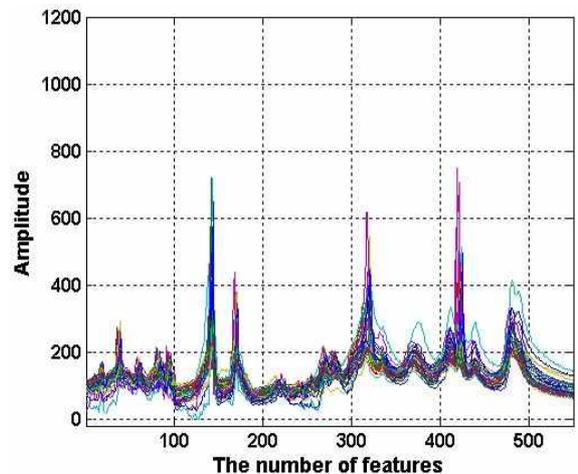
기존의 PCA를 이용하여 특징을 추출한 후 k-NN과 MLP 분류기를 적용한 경우의 인식률을 표 1에 나타냈다. 실험결과 k-NN 선형분류기의 경우 82.6%의 인식률을 나타냈다. 여기서 Hamming 거리척도 방법이 다른 거리척도 방법에 비해서 보다 좋은 결과를 보였다. 반면에 MLP 비선형분류기를 적용한 결과 인식률은 75.5%를 나타냈으며 k-NN 선형분류기에 비해서 약 7%만큼 인식률이 낮아졌다. 이 결과로부터 기존의 PCA 기법에 의해 추출된 특징벡터의 분류는 비선형분류기보다 선형분류기가 보다 우수한 인식률을 보임을 알 수 있다.

SpPCA에 의해서 특징을 추출한 후에 적용된 분류별 인식률을 표 2에 나타냈다. 표 2에서 보는 바와 같이 k-NN 분류기의 경우 분할수가 5이고 k=1 일 때 인식률은 84.3%를 나타냈다. 또한 융합법칙으로는 전구역 특징벡터와 가중치합을 이용한 경우가 동일하게 우수한 결과를 나타냈다. 기준벡터와 입력패턴과의 유사도는 성능이 우수한

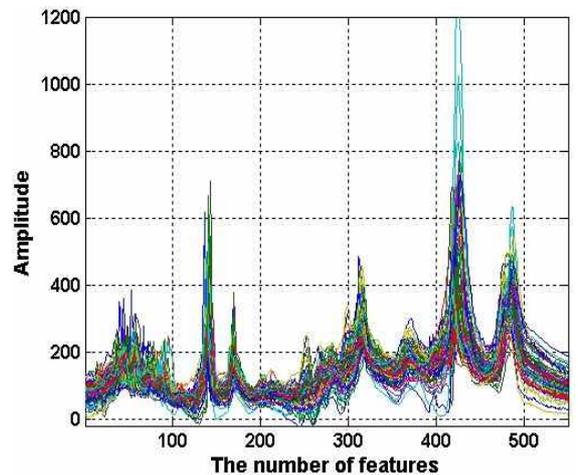
Hamming 거리척도 방법으로 산출하였다. MLP 비선형분류기를 적용한 결과 분할수는 7이고 다수결정법칙에 의해서 융합한 경우 인식률은 88.2%로 가장 좋은 인식률을 보였다.



(a) Class 1 (M)



(b) Class 2 (E)



(c) Class 3 (A)

그림 3. 클래스별 생체 스펙트럼 데이터

Fig. 3. Biomedical spectral data for each class

표 1. PCA에 의한 인식결과.

Table 1. Classification results obtained by PCA

[단위 : %]

Classifier		thr = 0.99		thr = 0.995		thr = 0.998	
k-NN	Distance	k = 1	k = 3	k = 1	k = 3	k = 1	k = 3
	Euclidean	79.7±3.5	77.7±4.2	79.4±3.8	77.3±4.4	79.3±4.0	77.4±4.0
	Hamming	<b>82.6±4.1</b>	81.0±4.3	81.8±4.8	80.8±4.6	82.1±4.7	81.3±4.1
	Mahananobis	75.1±4.4	72.5±4.4	75.3±4.9	71.8±4.4	75.1±4.8	72.2±4.8
MLP	Training		Test	Training	Test	Training	Test
		97.0±1.7	74.1±4.9	96.7±2.0	<b>75.5±3.1</b>	96.3±1.2	75.3±5.8

표 2. SpPCA에 의한 인식결과.

Table 2. Classification results obtained by SpPCA

[단위 : %]

Number of Partition	Fusion Method	k-NN			MLP	
		k = 1	k = 3	k = 5	Training	Test
3 (thr = 0.998)	Global Feature	83.7±2.8	83.0±3.6	79.8±3.9	97.1±1.5	75.1±2.5
	Sum Rule	83.7±2.8	83.0±3.6	79.8±3.9	99.8±0.4	81.4±3.4
	Best Vote	75.3±2.2			99.7±0.7	79.5±3.1
	Majority Vote	79.9±2.3			100±0.0	81.9±2.8
5 (thr = 0.998)	Global Feature	<b>84.3±3.3</b>	81.8±2.6	81.0±3.6	98.3±1.5	77.2±5.1
	Sum Rule	<b>84.3±3.3</b>	81.8±2.6	81.0±3.6	99.5±0.5	81.4±3.6
	Best Vote	63.0±2.3			99.9±0.3	82.6±1.8
	Majority Vote	80.1±2.4			99.9±0.3	84.2±2.6
7 (thr = 0.998)	Global Feature	83.6±3.3	81.6±3.1	80.7±3.2	97.7±1.8	76.7±4.8
	Sum Rule	83.6±3.3	81.6±3.1	80.7±3.2	99.5±0.7	84.6±3.5
	Best Vote	67.1±4.2			100±0.0	86.7±2.7
	Majority Vote	80.2±2.4			100±0.0	<b>88.2±2.0</b>

표 3. 확장된 공간상에서 PCA에 의한 인식결과.

Table 3. Classification results obtained by in the expanded space

[단위 : %]

Classifier		thr = 0.99		thr = 0.995		thr = 0.998	
k-NN	Distance	k = 1	k = 3	k = 1	k = 3	k = 1	k = 3
	Euclidean	81.5±3.6	79.3±3.8	81.4±3.7	79.9±3.7	81.7±3.7	79.7±3.7
	Hamming	84.2±3.9	83.1±3.0	84.1±3.3	82.2±4.3	<b>84.7±4.1</b>	83.0±5.0
	Mahananobis	77.2±3.9	74.8±3.7	76.8±4.2	74.7±3.6	76.8±3.9	74.7±3.2
MLP	Training		Test	Training	Test	Training	Test
		99.9±0.3	86.7±3.2	99.8±0.4	<b>87.1±2.6</b>	100±0.0	86.5±3.0

표 4. 확장된 공간상에서 SpPCA에 의한 인식결과.

Table 4. Classification results obtained by SpPCA in the expanded space

[단위 : %]

Number of Partition	Fusion Method	k-NN			MLP	
		k = 1	k = 3	k = 5	Training	Test
3 (thr = 0.998)	Global Feature	<b>85.0±4.0</b>	82.0±4.4	82.1±3.7	100±0.0	88.3±3.1
	Sum Rule	<b>85.0±4.0</b>	82.0±4.4	82.1±3.7	100±0.0	89.8±3.8
	Best Vote	80.2±4.6			100±0.0	89.9±3.4
	Majority Vote	83.1±3.6			100±0.0	89.8±3.2
6 (thr = 0.998)	Global Feature	83.4±4.3	81.4±3.7	80.2±3.5	99.8±0.6	87.8±3.5
	Sum Rule	83.4±4.3	81.4±3.7	80.2±3.5	100±0.0	89.3±2.6
	Best Vote	70.3±3.8			100±0.0	89.2±2.5
	Majority Vote	76.2±3.0			100±0.0	89.9±2.6
9 (thr = 0.998)	Global Feature	83.1±3.9	82.1±4.2	82.1±3.4	100±0.0	87.0±2.5
	Sum Rule	83.1±3.9	82.1±4.2	82.1±3.4	100±0.0	89.2±2.9
	Best Vote	68.5±3.7			100±0.0	89.9±2.9
	Majority Vote	80.9±2.5			100±0.0	<b>90.2±2.7</b>

확장된 공간내에서 PCA를 이용하여 특징을 추출한 후 k-NN과 MLP 분류기를 적용한 경우의 인식률을 표 3에 나타냈다. 실험결과 k-NN 선형분류기의 경우 84.7%의 인식률을 나타냈다. 여기서 Hamming 거리척도 방법이 다른 거리척도 방법에 비하여 보다 좋은 결과를 보였다. MLP 비선형 분류기를 적용한 결과 인식률은 87.1%를 나타냈으며 k-NN 분류기에 비해서 약 3%의 인식률이 향상되었다. 이는 표 1에서 보는 바와 같이 확장되지 않은 공간에서는 MLP 분류기보다 k-NN 분류기의 성능이 우수한 반면에, 확장된 공간내에서는 k-NN 분류기보다 MLP 분류기의 성능이 우수하다는 점을 주목할 만 한다.

표 4는 본 논문에서 제안한 확장공간상에서 SpPCA에 의해서 특징을 추출한 후 실험한 결과를 나타냈다. 표 4에서 보는 바와 같이 k-NN 분류기의 경우 분할 수가 3이고 k=1일 때 인식률은 85.0%를 나타냈다. 융합법칙으로는 전 구역 특징벡터와 가중치합을 이용한 경우가 동일하게 우수한 결과를 나타냈으며, 이 때 유사도는 Hamming 거리척도를 이용하였다. 한편, MLP 비선형 분류기를 적용한 결과 분할수는 9이고 다중결정법칙에 의해서 융합한 결과 인식률이 90.2%로 가장 좋은 인식률을 나타냈다.

그림 4는 10개의 랜덤 검증데이터에 대한 적용 기법별 인식률을 나타냈다. 그림 4에서 보는 바와 같이 9번째의 랜덤 검증 데이터를 제외한 모든 검증 데이터에 대하여 제안된 방법이 다른 기법에 비하여 우수한 결과를 보임을 알 수 있다. 평균 인식률 측면에서 비교해 보면, 제안된 방법은 기존의 PCA+kNN 기법에 비하여 약 8% 높게 나타나 제안된 방법의 우수성이 검증된다. 또한, 고차원 입력패턴을 분할 없이 전차원에 대하여 PCA를 이용하여 특징을 추출한 후 MLP 비선형 분류기에 의하여 패턴을 분류한 결과 인식률이 매우 낮게 나타났다. 반면에, 구간별로 분류한 후 각각의 구간에 대해서 특징을 산출하고, MLP 비선형분류기를 적용한 결과 인식률이 매우 향상됨을 알 수 있었다. 이것은 그림 5에서 보는 바와 같이 다중결정법칙에 의해 구간별로 산출된 특징들을 효과적으로 융합하였기 때문이라 분석된다.

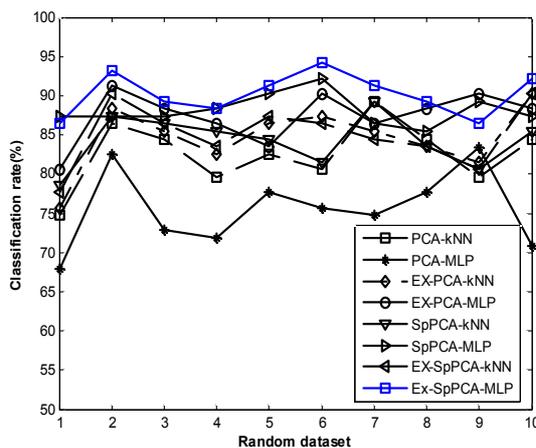


그림 4. 적용기법에 따른 인식 성능

Fig. 4. Classification accuracies obtained via different methods.

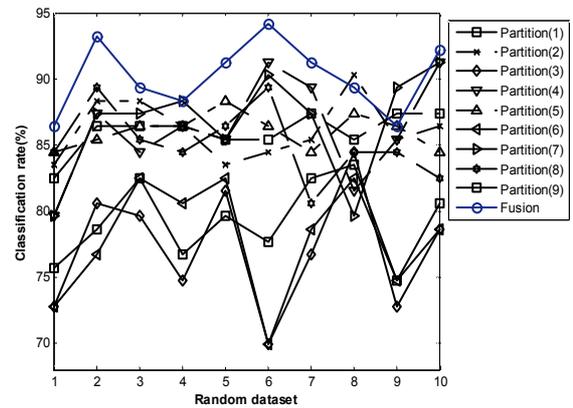


그림 5. 서브 패턴 및 융합 인식률

Fig. 5. Classification accuracies in each subpattern and fusion scheme

#### 4. 결 론

본 논문에서는 확장공간상에서 SpPCA와 MLP에 기반을 둔 융합기법을 이용한 생체 스펙트럼 패턴 분류 기법을 제안하였다. 제안된 방법은 우선 입력공간을 비선형변환에 의해 확장공간으로 변환한 후 확장공간을 몇 개의 부 공간으로 분할하고 SpPCA에 의해 분할 영역별 지역 특징값을 추출한다. 다음 단계에서는 추출된 특징벡터를 MLP 비선형 분류기를 이용하여 구간별 유사도를 산출했다. 최종 분류단계에서는 효과적인 융합법칙에 의해서 패턴을 분류하였다.

실험결과, 구간을 분할하지 않고 PCA에 의해 특징을 추출한 경우 k-NN 선형 분류기는 비선형 분류기인 MLP에 비하여 보다 나은 결과를 나타냈다. 반면에 확장공간상에서 구간을 분할하여 SpPCA에 의해 특징을 추출한 결과 MLP가 k-NN에 비하여 성능이 우수하게 나타났다. 제안된 방법은 기존의 PCA만을 적용한 경우와 비교하면 인식률이 7% 향상되었으며 이로부터 제안된 방법은 고차원을 갖는 생체 스펙트럼 데이터를 효과적으로 분류할 수 있음을 확인할 수 있었다.

#### 참 고 문 헌

- [1] R. Baumgartner, R. Somorjai, C. Bowman, T.C. Sorrell, C.E. Mountford, U. Himmelreich, "Unsupervised feature dimension reduction for classification of MR spectra," *Magnetic Resonance Imaging* 22, pp. 251-256, 2004.
- [2] Y.Y.B. Lee, Y. Huang, W. El-Deredy, P.J.G. Lisboa, C. Arus, P. Harris, "Robust methodology for the discrimination of brain tumours from in vivo magnetic resonance spectra: Science, Measurement and Technology," *IEEE Proc.*, Vol. 147, No. 6, pp. 309-314, 2000.
- [3] Yu Wang, S. Van Huffel, E. Heyvaert, L. Vanhamme, N. Mastronardi, P. Van Hecke, "Magnetic resonance spectroscopic quantitation

via complex principal component analysis," *WCCC-ICSP 2000*, Vol. 3, pp. 2074-2077, 2000.

- [4] M. Turk and A. Pentland, "Face recognition using eigenfaces," *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 586-591, 1991.
- [5] Hong Yang, Joseph Irudayaraj and Manish M. Paradkar, "Discriminant analysis of edible oils and fats by FTIR, FT-NIR and FT-Raman spectroscopy," *Food Chemistry*, Vol. 93, No. 1, pp. 25-32, 2005.
- [6] Marina Vannucci, Naijun Sha and Philip J. Brown, "*NIR and mass spectra classification: Bayesian methods for wavelet-based feature selection*," *Chemometrics and Intelligent Laboratory Systems*, In Press, Corrected Proof, Available online 4 March 2005.
- [7] E.K. Tang, P.N. Suganthan, X. Yao and A.K. Qin, "Linear dimensionality reduction using relevance weighted LDA," *Pattern Recognition*, Vol. 38, No. 4, pp. 485-493, 2005.
- [8] N.P. Hughes, S.J. Roberts, L. Tarassenko, "Semi-supervised learning of probabilistic models for ECG segmentation," *Engineering in Medicine and Biology Society, EMBC 2004*, Vol. 1, pp. 434-437, 2004.
- [9] Songcan Chen, Yulian Zhu, "Subpattern-based principal component analysis," *Pattern Recognition* 37, pp. 1081-1083, 2004.
- [10] Rajkiran Gottumukkal, Vijayan K. Asari, "An improved face recognition technique based on modular PCA approach," *Pattern Recognition Letters* 25, pp. 429-436, 2004.
- [11] Yang-Lang Chang, Chin-Chuan Han, Fan-Di Jou, Kuo-Chin Fan, K.S. Chen, Jeng-Horng Chang, "A modular eigen subspace scheme for high-dimensional data classification," *Future Generation Computer System* 20, pp. 1131-1143, 2004.
- [12] Nicolino J. Pizzi, "Fuzzy pre-processing of gold standards as applied to biomedical spectra classification," *Artificial Intelligence in Medicine*, Vol. 16, No. 2, pp. 171-182, 1999.

## 저자 소개



**조재훈(Jae Hoon Cho)**  
 2002년 : 한밭대학교 제어계측공학과(학사)  
 2004년 : 한밭대학교 제어계측공학과  
 (공학석사)  
 2005년~현재 : 충북대학교 제어계측공학과  
 박사과정

관심분야 : 지능시스템, 다중생체인식, 퍼지이론  
 E-mail : mmi8988@lycos.co.kr



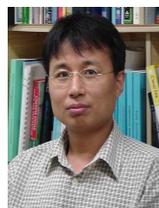
**이대종(Dae Jong Lee)**  
 1995년 : 충북대학교 전기공학과(학사)  
 1997년 : 충북대학교 전기공학과(공학석사)  
 2002년 : 충북대학교 전기공학과(공학박사)  
 2004년~2005년 : University of Alberta,  
 Postdoc  
 2006년~현재 : 충북대학교 BK21 충북정  
 보기술단 초빙조교수

관심분야 : 음성신호처리, 얼굴인식, 다중생체인식  
 E-mail : djmidori@empal.com



**박진일(Jin Il Park)**  
 2001년 : 한밭대학교 제어계측공학과(학사)  
 2003년 : 한밭대학교 제어계측공학과  
 (공학석사)  
 2005년~현재 : 충북대학교 제어계측공  
 과 박사과정

관심분야 : 지능시스템, 다중생체인식, 퍼지이론  
 Phone : 043) 261-2388  
 Fax : 043) 268-2386  
 E-mail : moralskr@yahoo.co.kr



**전명근(Myung Geun Chun)**  
 1987년 : 부산대학교 전자공학과(학사)  
 1989년 : 한국과학기술원 전기 및 전자공  
 학과(공학석사)  
 1993년 : 한국과학기술원 전기 및 전자공  
 학과(공학박사)  
 1993년~1996년 : 삼성전자 자동화연구소  
 선임연구원

2000년~2001년 : University of Alberta 방문교수  
 1996년~현재 : 충북대학교 전기전자 및 컴퓨터공학부 교수

관심분야 : Biometrics, 패턴인식, 데이터 마이닝, 얼굴인식  
 E-mail : mgchun@chungbuk.ac.kr