Journal of the Korean Institute of Industrial Engineers
Vol. 32, No. 2, pp. 133-140, June 2006.

웨이블릿에 기반한 시그널 형태를 지닌 대형 자료의 feature 추출 방법

장우성 · 장우진†

서울대학교 산업공학과

A Wavelet based Feature Selection Method to Improve Classification of Large Signal-type Data

Woosung Jang · Woojin Chang

Department of Industrial Engineering, Seoul National University

Large signal type data sets are difficult to classify, especially if the data sets are non-stationary. In this paper, large signal type and non-stationary data sets are wavelet transformed so that distinct features of the data are extracted in wavelet domain rather than time domain. For the classification of the data, a few wavelet coefficients representing class properties are employed for statistical classification methods : Linear Discriminant Analysis, Quadratic Discriminant Analysis, Neural Network etc. The application of our wavelet-based feature selection method to a mass spectrometry data set for ovarian cancer diagnosis resulted in 100% classification accuracy.

***Keywords*:** classification, wavelet transform, feature extraction

1. 서 론

시그널 형태로 표현되는 자료의 패턴을 구분하는 기술은, 생산과정에서의 공정 이상 유무, 금융 자료의 분석, 바이오 시그널을 이용한 의료진단 등, 일상생활 속에서 널리 유용되고 있다(Amato *et al*., 2005). 그런데, 이러한 시그널 형태의 자료가 비정상적인 과정(non-stationary, 이하 비정상적인 과정은non-stationary를 의미함)을 따르거나 갑작스런 증가와 감소를 가진 경우에 자료 패턴을 찾는 과정에서 많은 어려움이 발생할 수 있다(Jung, 2004). 예를 들면, 공정과정에서 정상 제품과 불량 제품이 생산될 때 시간에 따른 온도 변화가 중요한 분류 기준이 되는 경우나 생물의학 분야에서 치매 발병 가능성 유무를 진단할 때 뇌파가 활용되는 경우에 이러한 형태의 시그널이 생성되어 분석에 어려움이 따른다.

비정상적인 과정을 따르는 시그널 형태의 자료는 분류에 불필요한 부가적인 정보가 많이 있거나, 자료 획득 과정에서 큰 비용이 발생하여 feature의 수가 자료의 개체수를 크게 초과함으로써 차원 축소(dimension reduction)과정 없이 분류할 수 없는 경우에 발생할 수 있다. 이러한 경우 흔히 사용되는 방법은 feature를 선택하는 주성분 분석(principal component analysis, 이하 PCA)과 Linear Discriminant analysis(LDA) 등의 분류기법이다. 하지만, 극단적으로 feature의 수가 크고 자료의 개체수가 적은 경우에는PCA를 사용할 수 없는 경우가 발생한다.

이러한 어려움을 해결하기 위해서 최근 웨이블릿 변환이 많이 사용되고 있다(Lada *et al*., 2002; Vannucci *et al*., 2005; Jeong *et al*., 2003). 웨이블릿 변환은 시그널 자료를 각 주파수에 해당하는 여러 계층(multi-resolution level)들로 분리하고 불필요 하거나 중복적인 정보를 가진 feature를 효과적으로 제거할 수 있어 비정상적 시그널 형태의 자료를 분석하기에 적합하다.

Lada *et al*.(2002)에서는 생산과정에서 발생한 비정상적이거나 갑작스런 증가(sudden jump)가 있는 시그널을 웨이블릿 변환을 이용하여 결함 발견(fault detection)에 필요한 효과적인 자

이 연구는 서울대학교 신임교수 연구정착금으로 지원되는 연구비에 의하여 수행되었음
† 연락저자 : 장우진 교수, 151-742 서울시 관악구 신림동 산 56-1 서울대학교 산업공학과, Fax : 02-889-8560, E-mail : changw@snu.ac.kr
2005년 12월 접수; 2006년 2월 수정본 접수; 2006년 2월 게재 확정.

료 축약(data reduction)을 하였다. 하지만, 분류보다는 자료 축약에 중점을 두고 있어서 자료의 크기가 매우 크고 다양한 정보가 다수 포함된 시그널에서는 효율적이지 못한 단점이 있었다.

Tibshirani *et al*.(2004)은 Protein mass spectrometry를 이용하여 생성된 시그널 형태의 자료로부터 초기 암 진단을 하고자 하였다. 각 집단마다 공통적으로 나타나는 peak을 찾아서 분류에 사용하였는데, 시그널의 peak을 찾는 방법의 기준이 복잡하고 그 방법에 따라 결과가 달라지는 단점이 있다.

Vannucci *et al*.(2005)은 웨이블릿 변환을 사용하여 각각의 개체마다 threshold를 설정하고 차원 축소(dimension reduction)과정을 거친 후에 프로빗(probit) 모형과 베이지안 방법론을 사용하여 예측 성능이 좋은 웨이블릿 계수를 선택 한 다음 그것을 예측 변수로 사용하였다. 이 논문에서 제시한 방법은 집단에 따라 일정한 패턴을 가지는 자료의 차원 축소를 각 개체 마다 한 이후에 다시 복잡한 계산 과정을 반복적으로 시행하여야 하는 단점을 가지고 있기 때문에 계산복잡도(computational complexity) 측면에서 보다 간단한 계수 선택과정 또는 예측변수 선택 과정이 필요하다.

Jeong *et al*.(2003)은 전후 시차와 상관관계(correlation)를 가진 생산 과정에서 측정된 비정상 시그널 형태 자료의 결함 발견(fault detection)에 웨이블릿을 이용하였다. 이 방법은 웨이블릿 변환을 통하여 시간 영역(time domain)의 자료를 웨이블릿 영역(wavelet domain)으로 변환 한 후 여러 주파수 대역의 에너지 크기를 예측 변수로 사용한다. 이산형 웨이블릿 변환(discrete wavelet transform, 이하 DWT)은 시간축의 미세한 이동에도 영향을 받는(shift variant) 단점이 있는데, 이를 개선하여 실시간(real-time) 생산과정 모니터링(process monitoring)을 가능하게 하였고 계산복잡도 측면에서도 우수하였다. 하지만, 반복측정이 가능한 자료에 적용 시 시간에 따른 변화를 잘 반영하지 못하여 분류의 성능이 저하되는 단점을 가지고 있다.

Wu *et al*.(2003)은 비정상적인 과정(non-stationary)을 따르며 갑작스런 증가나 감소(sudden jump or drop)가 있고, 예측 변수(feature)의 수가 자료의 개체 수(sample size)보다 월등히 많은 자료를 이용하여 여러 분류 기법(선형 판별 분석, 이차 판별 분석, 의사 결정 나무 모형 등)의 성능을 비교 하였다. 이 논문의 저자들은 결론에서 다양한 분류 기법들을 사용하여 분류기(classifier)의 성능을 높이기보다 자료의 잡음 제거(noise removal)나 적절한 예측 변수 선택 과정이 필요함을 강조하였다.

지금까지의 관련 연구는 크기가 매우 크고 다양한 정보가 다수 포함된 자료 분류 문제의 경우 자료 축약에 중점을 두고 있어 통제해야 할 변수 설정 기준이 모호하여 그 기준에 따라 분류 결과가 달라지고, 계산복잡도 측면에서 복잡한 예측변수 선택 과정을 사용하는 등 여러 문제점을 갖고 있었다. 이와는 반대로 알고리즘이 간단하여 계산복잡도 측면에서 우수하지만, 그로 인해 분류의 성능이 저하되는 단점을 가진 경우도 있었다.

본 논문에서 제시한 방법은 웨이블릿 변환을 통하여 각 집단의 고유한 특징(unique feature)을 나타내는 소수의 웨이블릿 계수들만을 사용하여 분류에 필요한 정보만을 추출하기 때문에 기존의 방법보다 개선된 자료 축약(data reduction)효과를 갖고 있다. 그리고 기존의 웨이블릿 분석은 일반적으로 개개의 자료마다 threshold를 정하는데 반해 본 논문에서 제시하는 방법은 집단에 따라서 일정한 패턴을 가지는 자료를 대상으로 하여 그 집단에 속하는 모든 개체를 고려한 threshold를 정하기 때문에 집단의 공통적인 속성을 찾아내는데 유리한 면이 있다. 그리고 간단하면서 효율적인 feature selection 방법을 개발하였기 때문에 분류를 했을 때의 성능이 비슷하거나 우수하면서도 계산복잡도(computational complexity) 측면에서 타 방법들 보다 우월한 장점을 가지고 있다.

이 후에 전개될 본문의 내용은 다음과 같다. 2장에서는 자료의 웨이블릿 변환에 대해 설명하고, 3장에서는 웨이블릿 변환된 시그널 자료의 패턴 분류를 위한 feature 추출 방법을 제시하고, 4장에서는 추출된 feature들을 이용한 여러 통계적 패턴 분류 기법을 소개하며, 5장에서 이 논문에서 제시된 방법론을 난소암 진단 자료에 적용하여 다른 방법론과의 비교를 통해 이 논문에서 제시된 방법론의 실효성을 판단하고, 그리고 마지막 6장 결론에서 본 논문에서 제시한 방법의 의의와 한계에 대해 언급한다.

2. 웨이블릿 분석

이번 장에서는 이 논문에서 효과적인 feature extraction을 위해 사용되는 웨이블릿 변환에 대해 설명하고자 한다. 웨이블릿은 이름 그대로 가로축의 위아래를 넘나드는 웨이브형의 곡선이다. ϕ으로 표시되는 남성형 웨이블릿(father wavelet)과 ψ로 표시되는 여성형 웨이블릿(mother wavelet)은 적분했을때 각각 1과 0의 값을 갖는다.

$$\int \phi(t)dt = 1, \quad \int \psi(t)dt = 0 \tag{1}$$

두 웨이블릿 함수의 차이를 개략적으로 살펴보면 남성형 웨이블릿은 자료의 smooth한 부분과 저주파수(low-frequency) 부분을 설명하기에 적합하고, 여성형 웨이블릿은 자세한(detail) 부분과 고주파수(high-frequency) 부분을 설명하기에 적합하다. 웨이블릿의 기저(basis)함수는 그 크기(scale)와 위치(location)에 따라 scale index j와 translation index k를 갖는데, 다음과 같은 형식으로 표현될 수 있다.

$$\phi_{j,k}(t) = 2^{-j/2}\phi(2^{-j}t - k) = 2^{-j/2}\phi\left(\frac{t - 2^j k}{2^j}\right) \tag{2}$$

$$\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k) = 2^{-j/2}\psi\left(\frac{t - 2^j k}{2^j}\right) \tag{3}$$

기저 함수(basis function) $\phi_{j,k}(t)$와 $\psi_{j,k}(t)$는 scale index j가 커짐에 따라 남성형 웨이블릿 ϕ와 여성형 웨이블릿 ψ의 폭이 넓어지고 높이가 낮아진 형태를 갖는다. 또한, translation index k가 커질수록 우측으로 이동되는 효과가 있다. 이와 같이 다른 scale index j와 translation index k를 가진 기저함수들을 사용하여 저주파 대역과 고주파 대역은 물론 시간 영역의 모든 부분을 표현할 수 있다. 이와 같은 성질을 웨이블릿의 'time-frequency localization'성질이라고 한다. 앞에서 설명된 기저함수들을 사용하여 임의의 시그널 f(t)를 다음과 같이 근사(approximation)할 수 있다.

$$f(t) \approx \sum_k s_{J,k}\phi_{J,k}(t) + \sum_k d_{J,k}\psi_{J,k}(t) + \sum_k d_{J-1,k}\psi_{J-1,k}(t) + \cdots + \sum_k d_{1,k}\psi_{1,k}(t) \quad (4)$$

여기서 J는 coarest resolution level이다.

위와 같이 웨이블릿 변환을 하게 되면 각 주파수 대역(frequency bandwidth)에 해당하는 웨이블릿 계수들이 multire-solution level에 따라 나열된다. $s_{J,k}$는 smooth coefficient라고 하며, $d_{j,k}$는 detail coefficient라고 불린다. 각 웨이블릿 계수와 기저함수의 곱의 선형결합으로 시그널을 나타낼 수 있다. 그리고 각 기저에 곱해지는 해당 계수(wavelet coefficient)값은 다음과 같은 식으로 구해진다.

$$s_{J,k} \approx \int \phi_{J,k}(t) f(t) dt \quad (5)$$

$$d_{j,k} \approx \int \psi_{j,k}(t) f(t) dt \quad (6)$$

모든 자료들은 식(4)의 형태로 표현되고, 사용된 기저의 계수는 식(5)와 식(6)에서 보이는 것과 같이 명확하므로 자료 구조를 파악하는 근본적인 방법이 될 수 있다. DWT를 통하여 웨이블릿 계수($w = [w_1, w_2, ..., w_n]$)를 구하는 과정은 이산형 자료($f = [f_1, f_2, \cdots, f_n]$)에 직교행렬(W)을 곱하는 형태로 나타낼 수 있다.

$$w = fW = [w_1, w_2, ..., w_n] = [d_1, ...d_{J-1}, d_J, s_J] \quad (7)$$

$$\begin{aligned} d_1 &= (d_{1,1}, d_{1,2}, ..., d_{1,n/2}) \\ \vdots &= \quad \vdots \quad \vdots \\ d_{J-1} &= (d_{J-1,1}, d_{J-1,2}, ..., d_{J-1,n/2^{J-1}}) \\ d_J &= (d_{J,1}, d_{J,2}, ..., d_{J,n/2^J}) \\ s_J &= (s_{J,1}, s_{J,2}, ..., s_{J,n/2^J}) \end{aligned} \quad (8)$$

DWT는 실제로 복잡한 행렬 계산을 하지 않고, 'pyramid' 알고리즘을 통해 빠른 계산을 통해 웨이블릿 계수를 구하게 된다. 이산형 자료의 총 에너지는 아래 식(9)와 같이 정의 할 수 있는데, DWT는 직교변환이기 때문에 변환 후에도 같은량의 에너지를 갖게 된다.

$$E = \sum_{k=1}^{n} f_k^2 = \sum_{k=1}^{n} w_k^2 \quad (9)$$

웨이블릿 변환 후에는 상대적으로 절대값이 큰 소수의 smooth coefficient나 detail coefficient에 전체 에너지의 대부분이 집중되는 경향이 있다. detail coefficient들 중에는 0이거나 0에 가까운 계수들이 많기 때문에 역변환시에0에 가까운 계수들을 사용하지 않아도 원래 자료로 복원할 때 큰 영향이 없게 된다. 웨이블릿 변환의 이런 성질을 이용하면 원래 자료를 훨씬 적은 수의 웨이블릿 계수로 거의 비슷하게 표현할 수 있기 때문에 정보 축약에 매우 효과적이다.

3. 웨이블릿 기반의 feature selection

이번 장에서는 웨이블릿을 이용하여 비정상적인 과정을 따르며 차원(dimension)이 큰 시그널 형태의 자료의 패턴 분류 방법을 제시하고자 한다.

분류 문제에서 총 예측 변수(feature)의 수가 자료의 크기(sample size)보다 큰 자료를 분류하기 위해서 차원축소과정이 필요하다. 일반적으로 주로 사용되는 모수적인 방법에서는 feature의 수가 극단적으로 많은 경우 복잡한 계산을 해야 하기 때문에 계산 처리과정에서 많은 제약을 갖게 된다. 웨이블릿 변환은 자료 축약에 효과적이고, 'pyramid algorithm'을 통한 빠른 계산이 가능하기 때문에, 이런 자료를 효율적으로 분석할 수 있다.

3.1 웨이블릿 계수 선택 방법

n개의 data point들로 이루어지고 패턴이 비슷한 시그널 형태의 자료가 M개 있다고 했을 때, M개의 관측된 자료들을 size가 $(M \times n)$ 인 행렬 S로 나타낼 수 있다.

$$S = [s_1, s_2, \cdots\cdots, s_M]' \quad (10)$$

여기서 $s_i = (s_{i1}, s_{i2}, s_{i3}, \cdots, s_{in})$는 i번째 관측된 자료의 n개의 data point들로 이루어진 행벡터이고, $s^j = (s_{1j}, s_{2j}, s_{3j}, \cdots, s_{Mj})'$는 행렬 S의 j번째 열벡터이다. 이 M개의 자료의 웨이블릿 변환은 앞장에서 식(7)과 같은 형식의 $D = SW$ 나타낼 수 있다. 이 때 W는 웨이블릿 변환을 수행하는 $n \times n$ 행렬이고, D는 웨이블릿 변환 후 모든 자료들의 웨이블릿 계수들로 구성된 행렬이다. 행렬 D는 행 기준으로 정리했을 때 $D = (d_1, d_2, \cdots, d_M)'$로 나타낼 수 있고, 열 기준으로 나열했을 때 $D = (d^1 \mid d^2 \mid d^3 \mid \cdots \mid d^n)$로 표현할 수 있다.

<Figure 1>에서 나타난 있는 것처럼 행렬 D의 구조를 살펴보면, d^i, i=1,2,...,n는 웨이블릿 변환 후에 i번째 위치에 해당하는

웨이블릿 계수들로 M개의 관측된 자료로부터 얻어진 열벡터(column vector)임을 알 수 있다. 여기서 d_j, j=1,2,...,n은 j번째 자료의 웨이블릿 계수들로 이루어진 행벡터이다.

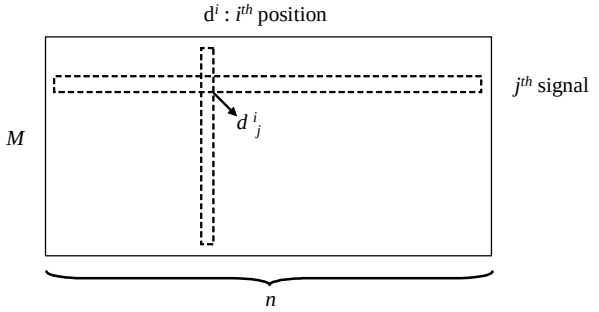

Figure 1. Structure of matrix D.

DWT의 정규 직교 성질(orthonormal property)에 의해서 행렬 D는 식 (9)에서처럼 행렬 S와 같은 양의 에너지를 가진다. 그러나 행렬 D는 행렬 S와 다르게 coarse level의 smooth(혹은 scale) coefficient들이 절대값이 커서 전체 에너지에 기여하는 바가 큰 반면 fine level들의 detail coefficient들은 절대값의 크기가 작아 전체 에너지에 기여하는 바가 작다.

정보량이 많고 적음을 웨이블릿 계수가 갖는 에너지량으로 판단한다면 절대값이 큰 웨이블릿 계수가 절대값이 작은 웨이블릿 계수보다 많은 정보량을 가지고 있는 것으로 생각할 수 있다. 이런 사실은 정보량이 고르게 흩어져 있는 시간영역에서의 data point들이 웨이블릿 변환 후 정보량의 크기가 서로 다른 웨이블릿 계수들로 변화 되었음을 의미한다.

평활(smoothing)이나 잡음을 제거한 원래 자료 복원을 위해 웨이블릿 계수를 선별하는 방법으로 wavelet shrinkage가 있는데, 자료를 웨이블릿 변환 한 후 threshold 이하의 값을 제거하고 의미있는 웨이블릿 계수만 찾아내어 원래 자료를 웨이블릿 역변환을 통해 나타내는 과정을 따른다. 그런데 대개의 wavelet shrinkage 방법에서는 개개의 자료마다 정보량이 큰 계수들을 찾기 위한 threshold를 정하는데 반해 본 논문에서는 그 집단에 속하는 모든 개체를 고려한 threshold를 정한다. 이 방법을 통해 집단의 특징을 결정하는 웨이블 계수들의 위치를 찾아낼 수 있다. 구체적으로 이러한 웨이블릿 계수들을 선택하는 과정을 살펴보면 다음과 같다.

웨이블릿 계수들의 크기에 따른 순서를 나타내는 index set {(1),(2),...,(n)}은 아래 식 (11)의 조건을 만족하는 웨이블릿 계수의 원래 위치를 나타내는 {1,2,...,n}의 permutation set이다.

$$| d^{(1)} | \geq | d^{(2)} | \geq | d^{(3)} | \geq \cdots \geq | d^{(n-1)} | \geq | d^{(n)} | \quad (11)$$

단, 여기서 $| d^i | = \sqrt{\sum_{j=1}^{M} (d_j{}^i)^2}$ 이다.

아래 식 (12)를 만족하는 가장 작은 k를 찾은 후에, 전체 n개의 웨이블릿 계수 위치 index들로 이루어진 집합의 부분집합인 $I = \{(1),(2),\cdots,(k)\}$ 를 정의한다.

$$\sum_{j=1}^{k} | d^{(j)} |^2 \geq \alpha \times \sum_{j=1}^{n} | d^j |^2, \quad 0 < \alpha < 1 \quad (12)$$

I에 해당하는 위치의 웨이블릿 계수들은 동일한 패턴을 가지는 시그널 형태 자료의 집단의 전체 에너지 중 $\alpha \times 100\%$ 를 갖게 된다. 이러한 방법을 통하여 n개보다 매우 작은 k개의 웨이블릿 계수들로 집단 내의 공통적인 특성을 잘 표현할 수 있다. I에 해당하는 웨이블릿 계수들은 아래 행렬과 같이 k개의 위치에 해당하는 값들로 표현되며, 한 집단에 속한 M개의 시그널 형태의 자료를 나타내는 데 쓰인다.

$$I = \begin{pmatrix} d_1{}^{(1)} & d_1{}^{(2)} & d_1{}^{(3)} & & d_1{}^{(k-1)} & d_1{}^{(k)} \\ d_2{}^{(1)} & d_2{}^{(2)} & d_2{}^{(3)} & & d_2{}^{(k-1)} & d_2{}^{(k)} \\ d_3{}^{(1)} & d_3{}^{(2)} & d_3{}^{(3)} & \cdots & d_3{}^{(k-1)} & d_3{}^{(k)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ d_M{}^{(1)} & d_M{}^{(2)} & d_M{}^{(3)} & & d_M{}^{(k-1)} & d_M{}^{(k)} \end{pmatrix}$$

이러한 I에 속하는 웨이블릿 계수들을 그 집단의 특징을 잘 나타내기 때문에, 서로 다른 여러 집단의 자료에서 각 집단의 해당 I들을 찾아내어 미지의 자료를 분류하는 문제에 활용할 수 있다. 이에 관한 구체적 방법은 다음 절에 기술 되어있다.

3.2 Index set을 이용한 Feature 선택

자료의 크기가 큰 경우에 앞 절에서 소개한 것처럼 적은 수의 웨이블릿 계수들로 각 집단의 특징을 표현할 수 있다. 선택된 웨이블릿 계수들은 집단의 특징을 잘 보존하기 때문에 자료의 크기가 큰 자료를 분류할 때 시간영역에서보다 분류가 잘 되는 경우가 많다. 앞 절에서 제시한 특징 추출 방법을 사용하여 여러 집단이 있을 때 분류를 위해 각 집단의 고유한 특징만을 추출하는 방법에 대해 설명하겠다. L개의 집단이 존재한다고 가정했을 때 모든 자료들을 이산형 웨이블릿 변환(DWT)한 후에 앞 절에서 한 방법과 같이 집단에 따라 웨이블릿 계수 index set을 만든다.

$$I_l,\ l = 1, 2, ..., L \quad (13)$$

예를 들어 I_1은 집합 1의 특징을 나타내주는 웨이블릿 계수의 위치들의 집합을 나타낸다. 개별 집단마다 서로 다른 특성을 보일 것이므로 $I_1 \cap (\bigcup_{l \neq 1} I_l)^c \neq \varnothing$ 을 만족한다고 가정할 수 있다. 이런 가정 하에서 집단1의 특성만을 나타내주는 웨이블릿 계수의 위치들의 집단 $\tilde{I}_1$을 식 (14)와 같이 정의할 수 있다.

$$\tilde{I}_1 = I_1 \cap (\bigcup_{l \neq 1} I_l)^c \quad (14)$$

집단 1에서 한 것과 마찬가지로 다른 집단에도 적용하여 고

유한 특징만을 나타내는 $\tilde{I}_l, l = 1, 2, \cdots, L$을 정의 할 수 있다. 이렇게 추출된 각 집단의 고유한 특징들의 위치를 나타내는 $\bigcup_{l=1}^{L} \tilde{I}_l$ 에 해당되는 웨이블릿 계수들(<Figure 2>의 색칠된 부분)을 예측변수로 사용하여 분류를 할 수 있다.

<Figure 2>의 교집합에 해당되는 웨이브릿 계수 index들은 각 집단의 특성을 나타내기보다 모든 집단의 공통적인 특징을 나타내는 것으로 상대적으로 분류에 이용하기 어렵기 때문에 분류문제에서는 제외시키는 것이 바람직하다.

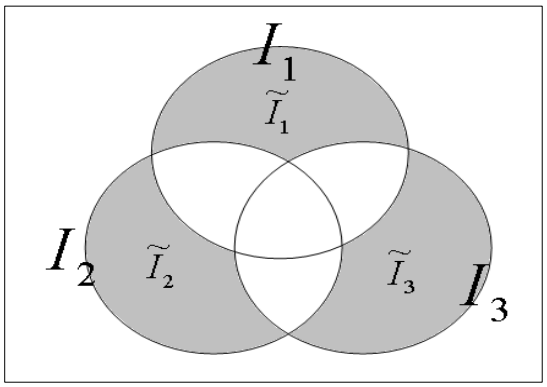

Figure 2. Venn diagram of wavelet coefficients index sets.

분류기법에 사용될 feature matrix를 X로 정의하고, 총 feature의 개수가 p라고 가정하자. 이때 $\bigcup_{l=1}^{L} \tilde{I}_l$ 의 원소의 개수는 p와 같게 된다. L개의 집단이 각각 $M_i, i = 1, 2, \cdots, L$ 개의 자료로 이루어진 집단이라고 보았을 때, feature matrix, X는 size가 $(N \times p), N = \sum_{i=1}^{L} M_i$ 인 행렬이 된다. 즉, $\bigcup_{l=1}^{L} \tilde{I}_l$ 에 해당되는 위치의 웨이블릿 계수들을 N개의 자료로부터 추출하여

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & x_{N3} & \cdots & x_{Np} \end{pmatrix} \tag{15}$$

가 이루어진다.

다음 장에서는 3장의 모형으로부터 추출된 feature들을 사용하여 실제 분류기를 만들 때 사용된 판별 분석에 대해 설명하고자 한다.

4. 판별분석

판별분석은 두 개 이상의 집단을 분류하는 데 있어서 그 오류를 최소화할 수 있는 변수들의 결합을 도출하고, 이를 통해 판별함수를 구축하여 소속된 집단을 모르는 자료를 분류하는데 목적을 두고 있다. 판별분석(discriminant analysis)은 집단간의 차이를 식별하는 데 사용되는 여래 개의 예측 변수와 각 개체가 소속되어 있는 집단을 나타내는 하나의 분류변수를 가진 다변량 자료를 그 대상으로 한다.

첫 번째로 주어진 예측변수들로부터 어떠한 집단에 속했는지 잘 나타낼 수 있는 판별함수를 구축하고 두 번째로 소속집단이 알려지지 않은 새로운 개체를 판별과정에서 구축된 판별함수를 통하여 여러 부분집단으로 분류한다. 가장 간단한 형태인 선형판별분석은 예측변수들의 선형결합을 통하여 자료의 분류를 가장 잘 해주는 선형판별함수를 찾아 분류를 한다.

Fisher의 선형판별함수(Linear discriminant analysis, LDA)는 1936년 처음 소개되었는데, 3장에서 구축된 feature matrix를 사용하여 LDA가 집단을 어떻게 분류하는가를 설명하면 다음과 같다. 앞서 정의된 $N \times p$ feature matrix, $X = (x_1, x_2, ..., x_p)$ 에서 x_i는 i번째 관측치의 예측 변수들의 열벡터, p는 예측변수의 개수이다. 이때 $a = (a_1, a_1, \cdots, a_p)'$가 x와의 선형결합에 사용되는 계수들의 행벡터, B가 집단 간의 공분산 행렬(between-classes covariance matrix), C가 집단 내 공분산 행렬(within-class covariance matrix) 이라면 LDA는 a'Ba/a'Ca를 최대화 하는 선형 결합 xa를 찾아내는 과정을 의미한다. 이는 집단 내에서의 분산은 최소화하면서, 집단간의 분산은 최대화 하는 계수를 추정하는 것이다.

다변량 정규분포 따르고, 각 집단에 있어서 변수들 간의 공분산 행렬이 동일하다는 가정을 하면 식(15)를 통하여 계산이 간단한 판별함수를 구축할 수 있다. 그리고 구축된 판별함수 $p(c \mid x)$가 최대인 집단으로 분류하면 된다.

$$p(c \mid x) = \frac{\pi_c p(x \mid c)}{p(x)} \propto \pi_c p(x \mid c) \tag{16}$$

π_c : 집단 c의 사전확률(prior probability)
$p(x|c)$: x의 집단 c에 대한 조건부확률
(class-conditional density)
$p(c|x)$: 집단 c의 사후확률(posterior probability)

집단 c의 반응변수가 평균 μ_c 공분산 Σ_c 인 다변량 정규분포를 따른다고 가정한다면, 판별식 Q_c는 식 (17)과 같이 된다.

$$\begin{aligned} Q_c &= -2log(p(x \mid c)) - 2log(\pi_c) \\ &= (x - \mu_c)\Sigma_c^{-1}(x - \mu_c)^T + log(|\Sigma_c|) - 2log(\pi_c) \end{aligned} \tag{17}$$

이 경우에는 판별식이 예측 변수 x의 이차형식으로 표현되기 때문에 Quadratic discriminant analysis, QDA라고 부른다. x space에서 decision region의 경계는 quadratic surface로 나타난다. LDA는 QDA의 한 예로서 집단간에 공통된 공분산행렬을 가진 경우이다. LDA와 QDA의 자세한 설명은 Hastie *et al.* (2001)을 참고하기 바란다. 이번 장에서 소개된 LDA와 QDA는 3장의 모형을 통해 추출된 특징을 예측변수로 하였을 때 상당히 좋은 결과를 나타내었다.

5. 실험 및 결과 검증

앞 장에서 소개된 방법론의 평가를 위해 난소암 진단에 사용되는 Mass spectrometry 자료(http://clinicalproteomics.steem.com)를 이용하였다. Mass spectrometry는 암 진단에 상당히 효과적인 자료이지만 불필요한 정보가 많고 잡음이 많이 섞여있어 효과적인 자료 축약이 이루어져야 분류를 할 수 있다. Mass spectrometry 자료 분류 기법은 Alexe *et al.*(2004), Tibshirani *et al.*(2004), Wu *et al.*(2003) 등에서 연구가 되었고, 현재도 활발하게 연구가 진행 중이다. 본 논문에서 사용된 Mass spectrometry 자료는 162명의 난소암 환자와 91명의 건강한 사람의 혈액 serum sample로부터 15154개의 mass/charge ratio에서 질량 분석법(mass spectometry)을 사용하여 intensity를 표현한 자료이다. Mass spectrometry에 대한 보다 자세한 설명은 Alexe *et al.*(2004)과 Vannucci *et al.*(2005)을 참고하기 바란다.

<Figure 3>과 <Figure 4>는 난소암 환자와 건강한 사람으로부터 추출된 자료의 평균값을 그린 것이다. 자료의 특성에 대해 살펴보면 253명이 각각 15154개의 변수를 가지고 있는 형태이므로 sample size보다 feature의 수가 월등히 많기 때문에 분류를 하기 위해서 자료 축약이 필요하다.

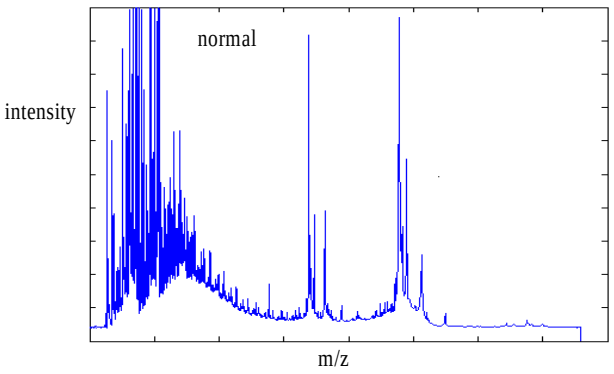

Figure 3. Average MS spectrum of healthy individuals.

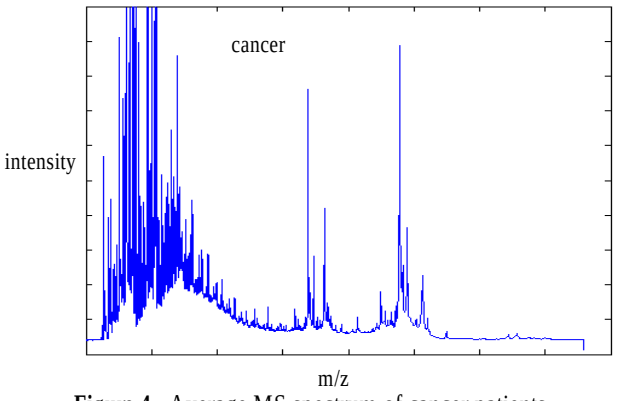

Figure 4. Average MS spectrum of cancer patients.

모든 15154개의 data point를 예측변수로 하여 의사결정나무 모형과 같은 방법을 사용할 수 있는데 원래 자료의 intensity 값들로 실험해 본 결과 training set에서는 성능이 굉장히 좋은 반면에 test set에서는 결과가 크게 좋지 못 하였다. 왜냐하면, training set에서 단 한 개의 변수로 분류가 완벽하게 이루어졌기 때문에 training set에서는 한 개의 변수로 분류를 할 수 있었지만, test set에서는 오분류율이 30% 이상 나오는 경우가 많았다. 이러한 결점을 보완한 의사결정나무모형 기반의 방법은 Alexe *et al.*(2004)의 논문에서 제시되었는데 의사결정나무모형에서처럼 적은 수의 변수를 사용한 것이 아니라 건강한 사람을 잘 분류해주는 변수와 난소암 환자를 잘 분류해주는 여러 변수를 찾아내어 자료 축약을 한 후에 분류를 하였다.

Mass spectrometry 자료의 잡음과 불필요한 정보를 제거하고자 3장에서 소개된 방법을 이용하여 난소암 자료를 'symmlet8' 웨이블릿 함수를 사용하여 웨이블릿 변환한 후 웨이블릿 영역에서 각 집단의 고유한 특징을 추출하였다. 162명의 난소암 환자들로 구성된 집단 1과 91명의 건강한 사람들로 구성된 집단 2의 mass spectrometry 자료를 최대한 균등하게 각각 10개의 set으로 나눈 후, 두 집단에서 하나의 set을 선택해 test set으로 하고 나머지를 training set으로 만드는 방식으로 10-fold Cross Validation(이하 CV)을 시행하였다. 이때 선택된 test set에서 feature set을 만드는 방법은 다음과 같다. 웨이블릿 변환된 mass spectrometry 자료들에서 각 위치에 해당하는 웨이블릿 계수들을 추출하여 그 제곱합을 3장에서 설명한 방법대로 순서통계량(order statistics)으로 만든 후에 시그널의 전체 에너지 중 95% 이상(식 (12) $\alpha = 0.95$)을 차지하는 최소 개수의 웨이블릿 계수들을 각 집단마다 찾아내었다. 그 후에 전체 웨이블릿 계수 index set의 부분집합인 각 집단에 해당되는 index set I_1과 I_2가 식 (13)과 같이 정의 되었고, 식 (14)에서 제시한 방법처럼 전체 웨이블릿 계수 index set에서 그 index set들의 교집합을 제외한 부분을 이용해 식 (15)의 feature matrix(X)를 정의하였다.

10-fold CV에서는 샘플들의 조합 형태에 따라서 매번 모형의 성능이 조금씩 달라지기 때문에, 서로 다른 10-fold CV를 20회 반복 실행하여 보다 robust한 결과를 얻을 수 있었다. 10-fold cross validation을 20번 반복하는 과정에서 feature 선택과정이 총 200번 시행되었는데, 이 과정에서 총 feature의 수는 29개에서 31개사이로 나타났고, 평균 29.60개가 선택되었다. feature의 수가 feature 선택과정 마다 다른 이유는 training 과정에서 사용되는 개체가 매번 다르기 때문이다. 15154개의 원래 data point 들 중 약 30개 정도의 feature(wavelet coefficient)들이 추출되었다는 것은 본 논문의 방법이 효과적인 자료 축약 방법임을 의미하는 결과이다.

<Figure 5>는 웨이블릿 변환 후에 선택된 index들에 해당하는 training set의 웨이블릿 계수들의 분포를 집단에 따라 도식화 해본 것이다. training set에서는 선택된 index 하나하나 만으로도 거의 완벽한 분류가 이뤄진 것을 알 수 있다.

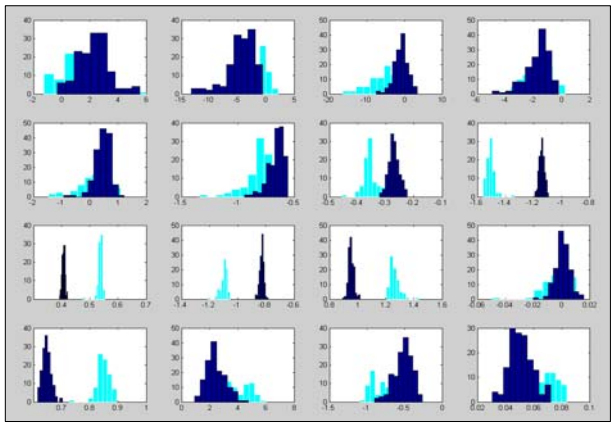

Figure 5. Histograms of wavelet coefficients in selected positions(training set).

<Figure 6>은 training set에서 선택된 웨이블릿 계수들의 위치가 새로운 자료에도 적용될 수 있는지 확인하기 위하여 test set의 웨이블릿 계수들도 각 집단 마다 히스토그램을 그려 도식화 해 보았다. 다음 그림을 보아 training set에서 분류를 잘 하는 index들이 test set에서도 분류를 잘 하는 것으로 나타났다.

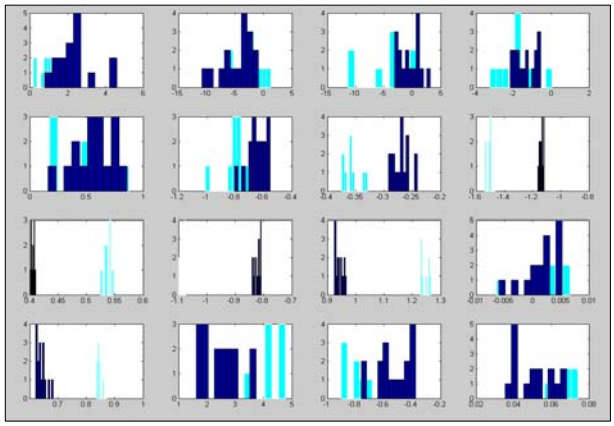

Figure 6. Histograms of wavelet coefficients in selected positions(test set).

Feature selection 과정을 거친 후에 LDA, QDA, Neural Network 등을 사용하여 난소암 환자와 건강한 사람을 분류하였다. 보다 구체적이고 객관적인 모형 평가를 하기 위해 질병의 유무를 진단하는 문제에서 널리 사용되는 sensitivity와 specificity를 사용하였다.

$$Sensitivity = Prob(predicting\ disease \mid true\ state\ is\ disease) \quad (18)$$

$$Specificity = Prob(predicting\ non-disease \mid true\ state\ is\ non-disease) \quad (19)$$

Sensitivity는 질병을 가지고 있는 사람이 질병을 가졌다는 것을 예측하는 확률이고, specificity는 질병을 가지고 있지 않은 사람이 질병을 가지지 않았다는 것을 예측하는 확률이다.

선택된 idnex의 웨이블릿 계수들로 만들어진 feature set을 LDA, QDA, Neural Network를 사용하여 건강한 사람과 난소암 환자로 분류한 후에 sensitivity와 specificity를 기준으로 모형을 평가하였다. <Table 1>은 웨이블릿 영역에서 이루어지는 feature 추출과 시간영역에서의 feature 추출 방법을 평가기준에 의해서 비교한 것이다. 10-folding QDA, 10-folding LDA, 10-folding NN는 각각 본 연구의 방법을 이용 웨이블릿 영역에서추출한 feature들로 20번의 10 fold CV를 시행한 결과이다. 여기서 LDA는 specificity와 sensitivity가 100%로 완벽하게 분류를 하였고, QDA는 5000번의 예측 중에 단 한번의 오류가 있을 만큼 성능이 좋았다. 하지만 Neural Network는 기대만큼 좋은 결과를 갖지는 못하였다. <Table1>의 마지막 열에 나와 있는 95% 신뢰구간은 평균에서 ±1.96×표준편차의 오차를 갖는 구간으로 100%를 초과하는 값은 불가능한 것이지만, 비교를 위해 Alexe *et al.*(2004)에서의 표기를 따랐다.

Table 1. Comparison of feature selection efficiency in wavelet and time domains

	Validation method		Accuracy	Confidence Interval(95%)
Wavelet Domain	10-folding QDA	Sensitivity	100%	100%~100%
		Specificity	99.9994%	99.46%~100.43%
	10-folding LDA	Sensitivity	100%	100%~100%
		Specificity	100%	100%~100%
	10-folding NN	Sensitivity	97.89%	94.15~102.18%
		Specificity	98.17%	91.23~104.55%
Time Domain	10-folding LAD* (complete)	Sensitivity	100%	100%~100%
		Specificity	99.6%	97%~101%
	10-folding LAD* (fortified)	Sensitivity	99.6%	96%~102%
		Specificity	99.6%	98~101%

* indicates results from Alexe *et al.*(2004)

웨이블릿에 기반한 feature 추출 방법에 의해 얻어진 위와 같은 결과를 시간영역(time domain)에서 이루어지는 기존 방법의 결과와 서로 비교하기 위해 Alexe *et al.*(2004)의 결과를 <Table 1>에 제시하였다. 이 결과는 본 논문에서 사용한 것과 같은 난소암 mass spectrometry 자료를 이용해 10번의 10-fold CV를 통해 얻어진 것으로, 이는 두 결과의 비교가 타당함을 의미한다.

Alexe *et al.*(2004)의 logical analysis of data(LAD) complete의 결과는 10-fold CV에서 sensitivity는 100%이고 specificity는 99.6%로 나타났으며 specificity의 95%의 신뢰구간은 97%에서 101%였다. 10-folding LAD(complete)는 LAD통해 찾아낸 20개의 positive pattern과 21개의 negative pattern을 모두 사용하여 분류한 것이고, 10-folding LAD(fortified)는 20개의 positive pattern 중 7개의 pattern과 21개의 negative pattern 중 8개의

patten을 사용하여 분류한 것이다. 이 외에 웨이블릿 변환 기법을 사용해 본 논문과 동일한 난소암 mass spectrometry 자료를 분류한 Vannucci *et al.*(2005)의 방법에서는 가장 좋은 결과가 100%의 sensitivity와 97%의 specificity를 가진 것이었다.

<Table 1>에서 결과를 살펴보면, 본 논문 3장에서 제시한 웨이블릿 영역에서 이루어지는 자료 축약 방법을 사용한 경우, 하나의 hidden layer를 가진 Neural Network를 분류기로 사용하였을 때는 약간 좋지 않은 결과가 나왔으나 QDA나 LDA를 사용한 분류에서는 Alexe *et al.*(2004)과 Vannucci *et al.*(2005)에서 제시한 방법보다 분류 성능이 우수한 것을 알 수 있다.

분류 과정의 계산적 측면에서도 본 논문의 방법은 대부분의 정보량(energy)이 적은 수의 wavelet 계수들로 표현 가능하기 때문에 계산량이 적은데 반해 시간영역에서 이루어지는 feature 추출 방법들은 정보량이 각 data point마다 흩어져 있어 간단한 feature 추출과정 알고리즘이라 하더라도 반복적으로 수행되어야 하므로 계산량이 크다. 예를 들면, Alexe *at al.*(2004)에서는 LAD를 사용하여 난소암 환자 진단에 사용되는 pattern을 찾기 위해 약 20000개의 제약식과 2백만에서 3백만개의 이진 변수(binary variable)들로부터 분류에 적합한 20개의 positive pattern(난소암환자 진단에 적합한 변수들의 조합)과 21개의 negative pattern(건강한 사람 분류에 적합한 변수들의 조합)을 찾아내는 복잡한 과정이 필요하였다.

6. 결 론

일반적으로 시그널 형태의 자료가 자료의 크기가 커서 불필요한 정보가 많고 잡음이 많이 섞여 있거나 비정상적인 과정을 따르고 갑작스런 증가와 감소를 가진 자료의 경우에 자료 처리과정에서 많은 어려움이 발생할 수 있다. 본 논문에서는 웨이블릿 영역에서 이루어지는 시그널 형태를 가진 대형 자료의 분류에 대해 살펴보았는데 비정상적이고 크기가 큰 자료의 패턴 분류가 복잡하지 않으면서도 효율적으로 이루어 질 수 있었으며 분류에 필요한 정보만을 추출하기 때문에 기존의 방법보다 개선된 자료 축약(data reduction)효과를 갖고 있고 있는 것으로 판단되었다. 그러나, 본 논문에서 제시하는 방법은 집단에 따라 일정한 패턴을 가지지 않거나 패턴의 차이가 크지 않은 자료에 적용하기 어렵기 때문에, 보다 개선된 웨이블릿 분류 기법 연구가 필요하다.

참고문헌

Alexe, G., Alexe, S., Liotta, L. A., Emanuel, P., Reiss, M., and Hammer, P. L. (2004), Ovarian cancer detection by logical analysis of proteomic data, *Proteomis*, **4**, 766-783.

Amato, U. and Sapatinas, T. (2005), Wavelet shrinkage approaches to baseline signal estimation from repeated noisy measurements, *Advances and Applications in Statistics*, **5**, 21-50.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The elements of statistical learning*, Springer, USA.

Jeong, M. K., Chen, D., Lu, J. C. (2003), Thresholded scalogram and its applications in process fault detection, *Applied stochastic models in business and industry*, **19**(3), 231-244.

Jung, U. (2004), Wavelet-based data reduction and mining for multiple functional data, Ph.D. dissertation, Georgia Institute of Technology, USA.

Lada, E. K., Lu, J. C., and Willson, J. R. (2002), A wavelet- based procedure for process fault detection, *IEEE Transactions on Semiconductor Manufacturing*, **15(**1), 79-90.

Raimondo, M. (2002), Wavelet shrinkage via peaks over threshold, *Intersat*, *May*, 1-19.

Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., and Le, Q. (2004), Sample classification from protein mass spectrometry by peak probability contrasts, *Bioinformatics*, **20**(17), 3034-3044.

Vannucci, M., Sha, N., Brown, J. P. (2005), NIR and mass spectra classification : Bayesian methods for wavelet-based feature selection, *Chemometrics and intelligent laboratory systems*, **77**(1/2), 139-148.

Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., and Zhao, H. (2003), Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data, *Bioinformatics*, **19**(13), 1636-1643.