퍼지 및 지능시스템학회 논문지 2007, Vol. 17, No. 2, pp. 220-225

주요성분분석과 상호정보 추정에 의한 입력변수선택

Input Variables Selection by Principal Component Analysis and Mutual Information Estimation

조용현, 홍성준
Yong-Hyun Cho, Seong-Jun Hong

대구가톨릭대학교 컴퓨터정보통신공학부
E-mail: {yhcho,sjishong}@cu.ac.kr

요 약

본 논문에서는 주요성분분석과 상호정보 추정을 조합한 입력변수선택 기법을 제안하였다. 여기서 주요성분분석은 2차원 통계성에 기반을 둔 기법으로 입력변수 간의 종속성을 빠르게 제거하여 과추정을 방지하기 위함이고, 상호정보의 추정은 적응적 분할을 이용하여 입력변수의 확률밀도함수를 계산함으로써 변수상호간의 종속성을 좀 더 정확하게 측정하기 위함이다. 제안된 기법을 각 500개 샘플의 7개 신호를 가지는 인위적인 문제와 각 55개 샘플의 24개의 신호를 가지는 환경오염신호를 대상으로 각각 실험한 결과, 빠르고 정확한 변수의 선택이 이루어짐을 확인하였다. 또한 주요성분분석을 수행하지 않을 때와 정규분할의 상호정보 추정 때보다 제안된 방법은 각각 우수한 선택성능이 있음을 확인하였다.

키워드 : 주요성분분석, 상호정보추정, 입력변수선택, 적응분할, 정규분할

Abstract

This paper presents an efficient input variable selection method using both principal component analysis(PCA) and adaptive partition mutual information(AP-MI) estimation. PCA which is based on 2nd order statistics, is applied to prevent a overestimation by quickly removing the dependence between input variables. AP-MI estimation is also applied to estimate an accurate dependence information by equally partitioning the samples of input variable for calculating the probability density function. The proposed method has been applied to 2 problems for selecting the input variables, which are the 7 artificial signals of 500 samples and the 24 environmental pollution signals of 55 samples, respectively. The experimental results show that the proposed methods has a fast and accurate selection performance. The proposed method has also respectively better performance than AP-MI estimation without the PCA and regular partition MI estimation.

Key Words : Principal Component Analysis, Mutual Information Estimation, Input Variable Selection Adaptive Partition, Regular Partition

1. 서 론

실세계의 모델링에서 가장 적합한 입력만을 선택하는 것은 시스템 성능에 많은 영향을 미친다[1]. 일반적으로 입력변수의 효과적인 선택은 시스템 차원의 감소나 특징추출 등 다양한 용도로 이용된다[1-4]. 그러나 많은 입력변수들 중에서 모델에 얼마나 많은 또는 어느 입력들이 필요한지 알 수 없으며, 이는 입력차원이 증가할수록 더욱 더 심각하다. 신경망 등에서 불필요한 입력들은 학습을 복잡하게 하고, 과학습 등에 따른 학습성능의 저하도 가져올 수 있다. 입력변수의 잘못된 선택에 여러 가지 문제들이 발생될 수 있다. 먼저, 입력차원의 증가에 따른 계산시간과 메모리의 증가, 다음으로 요구되지 않는 입력들에 의한 학습의 어려움, 추가적인 요구되지 않는 입력에 의한 비수렴과 모델의 정확성 저하, 그리고 복잡한 모델에 따른 해석의 어려움 등의 제약이 있다[2-4].

지금까지 알려진 입력변수선택 기법들은 크게 model-based와 model-free 방법들로 나누어진다[1-4]. 먼저 model-based 방법에 의한 입력선택은 모델을 선정한 후 이용할 입력들을 선택하고, 파리미터들을 최적화한 후 어떤 비용함수를 측정함으로써 이루어진다. 선형모델을 이용한 방법으로 분산의 해석(analysis of variance : ANOVA)에 의해 구현되는 전역 F-test 방법이 잘 알려져 있다. 또한 비선형 모델을 이용한 방법으로는 신경망이나 자동상관성검출(automatic relevance detection : ARD)로 구현되는 방법이 있다[1]. 이러한 model-based 방법들은 입력들이 바뀌면 선택과정은 다시 반복하여야 하는 제약이 있다. model-free 방법은 기초모델을 가지지 않는 통계적 종속성 시험에 바탕을 둔 기법으로 입력변수들의 부집합과 원하는 출력사이의 통계적 시험을 수행함으로써 이루어진다. 이때 시험은 이들 결과

접수일자 : 2006년 10월 21일
완료일자 : 2007년 2월 1일

에 기초하여 어느 입력변수를 선택할 것인가에 이용된다. correlation에 기반을 둔 방법, 고차원의 cross-cumulant에 기반을 둔 방법, 상호정보(mutual information : MI)에 기반을 둔 방법이 통계적 종속성을 시험하는 방법으로 알려져 있다[1,4].

model-free 방법은 통계적 종속성에 기반을 둠으로써 model-based 방법보다 좀 더 일반화된 방법이다[1,2]. 그러나 통계적 종속성은 입력과 원하는 출력사이의 상호정보를 추정함으로써 구해지며, 이러한 추정과정에는 joint probability density function(PDF)와 marginal PDF의 계산이 요구된다. PDF의 계산방법으로 correlation에 기반을 둔 방법은 변수 사이의 2차원 선형종속성만을 측정하는 방법으로 선형모델에만 적용 가능한 제약이 있다. 고차원의 cross-cumulant에 기반을 둔 방법은 고차원의 통계성을 이용하여 종속성을 측정하는 방법으로 여기에도 입력변수들의 모든 조합들을 조사해야 하는 제약이 있다. 이런 제약을 해결하기 위하여 변수들 간의 정보에 기반을 두고 모든 고차원의 통계성을 이용하여 종속성을 측정하는 상호정보에 기반을 둔 방법이 제안되었다[1]. 특히 상호정보에 기반을 둔 방법은 고차원의 cross-cumulant에 기반을 둔 방법에서 반드시 요구되는 정규화 과정을 제거할 수 있는 장점도 있다. 하지만 서로 종속성이 있는 입력들을 이용할 경우 어떤 선택 방법을 이용하든지 입력 수의 과추정이 발생되어 이를 해결하기 위한 연구가 요구된다.

본 연구에서는 주요성분분석(principal component analysis : PCA)[5-8]과 상호정보에 기반을 둔 방법을 조합한 입력변수선택 방법을 제안한다. 여기서 주요성분분석은 2차원 통계성에 기반을 둔 기법으로 입력변수 간의 독립성을 찾아 종속성에 따른 과추정을 방지하기 위함이고, 상호정보의 추정은 적응적 분할을 이용하여 입력변수의 확률밀도함수를 계산함으로써 변수상호간의 종속성을 좀 더 정확하게 측정하기 위함이다. 제안된 기법을 각 500개의 샘플을 가지는 7개의 신호를 가지는 인위적인 문제와 각 55개의 샘플을 가진 24개의 신호를 가지는 환경오염신호를 대상으로 각각 실험하여 타당성을 확인하고, 기존의 PCA의 전처리 없는 입력선택 및 정규적 분할(regular partition : RP)에 기초한 방법의 결과와 비교 · 분석하였다.

2. 주요성분분석과 상호정보 추정

2.1 주요성분분석에 의한 종속성 제거

주요성분분석은 입력데이터의 특징을 추출하는 기법으로 데이터 내에 포함된 정보를 추출하고 압축하여 통계적 규칙들을 찾아내는 것이다[5-7]. 이는 대용량의 입력데이터를 통계적 독립인 특징들의 집합으로 변환시키는 것이며, n차원 입력공간의 데이터를 k차원 출력공간의 데이터로 투영시키는 것이다. 여기서 k《n이면 입력데이터 벡터가 가지는 대부분의 내부정보를 유지면서도 차원의 감소가 가능하게 된다.

자기상관행렬 Rxx = <xxT>를 가진 평균이 영인 입력벡터 x = [x1, x2, … , xn]T에 대해서 생각해 보자. 여기서 T는 전치를 나타내며, < · >는 기대치를 나타낸다. 또한 $\hat{w}1,\hat{w}2, \cdots ,\hat{w}m$이 Rxx의 고유벡터와 직교되는 연결가중치 벡터라 할 때, $\hat{w}1$ = [$\hat{w}11, \hat{w}12, \cdots , \hat{w}1n$]T는 가장 큰 고유치 λ1과 일치하며, $\hat{w}2$ = [$\hat{w}21, \hat{w}22, \cdots , \hat{w}2n$]T는 두 번째로 큰 고유치 λ2, 그리고 $\hat{w}n$ = [$\hat{w}n1, \hat{w}n2, \cdots , \hat{w}nn$]T는 가장 작은 고유치 λn과 각각 일치한다. 이상의 관계를 행렬방정식으로 나타내면 식 (1)과 같다.

$$Rxx\hat{w}j = \lambda j\hat{w}j, \ (j=1,2,..., n) \tag{1}$$

여기서 λ1 ≥ λ2 ≥ … ≥ λn ≥ 0 이다. 주어진 입력벡터 x를 위한 첫 번째 m개의 주요 특징을 나타내는 고유벡터 y는 다음의 선형변환식 (2)로 나타낼 수 있다.

$$y = \hat{W}x \tag{2}$$

여기서 $\hat{W}$ = [$\hat{w}1,\hat{w}2, \cdots ,\hat{w}m$]T ∈ ℝmxn 이며, 이 식에서 연결가중치행렬 $\hat{W}$의 행은 가장 큰 고유치와 일치하는 상관행렬 Rxx의 고유벡터임을 의미한다.

다시 말하면, 입력데이터 공간에서 k차원의 주요특징을 나타내는 부공간은 Rxx의 k개 주요 고유벡터에 의해 구성된 부공간으로 정의된다. 결국 PCA는 < ‖ $\hat{w}$jT x ‖22 >가 최대인 고유벡터 $\hat{w}1,\hat{w}2, \cdots ,\hat{w}k$의 방향을 찾는 것이다. 일반적으로 얻어지는 고유값은 크기에 따라 정렬하고 고유벡터도 해당 고유값의 위치대로 정렬한다. PCA에서 순서대로 정렬된 고유값의 뒤쪽은 0에 가까운 값을 가지게 되어 이를 삭제할 수 있다. 이는 고유벡터의 작은 값들을 고려하지 않음으로써 입력 데이터의 차원을 줄이기 위함이다.

일반적으로 PCA를 좀 더 효과적으로 수행하기 위해 신호의 영 평균(zero-mean)을 수행한다[5,6]. 이는 신호의 1차적 통계성을 고려한 정규화로 영 평균은 신호벡터 x에서 평균값 x*를 뺀 차로 x = x - x*이다. 따라서 입력되는 변수를 대상으로 PCA를 수행하면 2차원 통계성이 고려된 독립변수를 추출할 수 있다. 이렇게 하면 신호 내에 포함된 종속성을 제거할 수 있어 입력변수 선택에서의 과추정을 방지할 수 있다.

2.2 적응분할 히스토그램 방식에 의한 상호정보 추정

신호들 사이의 종속성을 시험하기 위해 correlation, 고차원의 cross-cumulant, 그리고 상호정보 등에 기반을 둔 여러 가지 방법들이 제안되었다[1-7]. 그 중에서도 상호정보는 변수들 사이의 종속성을 정량화하기 위한 매우 기본적인 통계적 접근방법이다. 결국 상호정보는 입력변수들을 선택하는 가장 자연스러운 척도이며, 그 척도는 입력변수 선택을 위해 미리 이용된다. 하지만 신뢰성 있는 상호정보의 추정은 용이치 않으며, 무슨 방법을 이용하든 충분한 량의 데이터에 의해서만 유효한 결과를 얻을 수 있다.

일반적으로 Shannon의 정의에 따른 입력(독립)신호 x와 출력(종속)신호 y사이의 상호정보 I(x,y)는 joint PDF f(x,y)와 marginal PDF f(x) 및 f(y)의 곱 사이 Kullback-Leibler 거리로 다음 식 (3)과 같이 정의된다[1].

$$I(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y) \cdot \log\left(\frac{f(x,y)}{f(x)f(y)}\right)dxdy \tag{3}$$

여기서 x와 y가 서로 독립이면 상호정보 I(x,y)는 영이 된다. 또 다른 상호정보는 엔트로피(entropy)를 이용하여 다음 식 (4)와 같이 정의 될 수 있다.

$$I(x,y) = H(x) + H(y) - H(x,y) \tag{4}$$

여기서도 H(x)와 H(y)는 각각 신호 x와 y의 엔트로피이고, H(x,y)은 x와 y의 결합엔트로피이다.

식 (3)과 식 (4)에서 각각 상호정보의 계산을 위해서는 복잡한 joint PDF와 marginal PDF의 추정이 요구된다. 이러한 추정법으로 Gram-Charlier 확장에 기초한 방법, 정규분

할 히스토그램 PDF 근사화에 기초한 방법, 적응분할 히스토그램 PDF 근사화에 기초한 방법, 커널변환에 기초한 방법이 있다[1]. Gram-Charlier 확장에 기초한 방법은 PDF의 Gram-Charlier polynomial expansion에 기반을 둔 것으로 계산이 간단하고 빠르며 통계적인 의미가 분명한 장점이 있다. 그러나 PDF의 부적정한 근사화와 Gaussian과 sub-Gaussian 신호에 따라 성능이 달라지는 제약이 있다. 정규분할 히스토그램 PDF 근사화에 기초한 방법은 각 변수들을 샘플을 포함하는 작은 bin들로 일정하게 나누어 PDF를 계산한다. 이 방법은 Gram-Charlier 확장에 기초한 방법보다는 신호들의 성질에 의존하지 않기 때문에 좀 더 일반화된 방법이다. 그러나 이 방법 역시 샘플의 분할과 질에 민감한 제약이 있다. 분할이 너무 조밀하면 샘플을 포함하지 않는 어떤 bin들이 있어 PDF의 평활화에 따른 손실된 분포가 고려되지 않으며, 너무 듬성하면 bin들내의 샘플들이 중요한 PDF를 상세히 잘 반영하지 못하는 제약이 있다[2]. 이러한 분할에 따라 상호정보의 추정 성능이 달라지는 정규분할 히스토그램에 기초한 방법의 제약을 해결하기 위해 각 변수들을 동일한 샘플을 가지는 bin들로 나누어 각 bin의 영향을 평균화하는 적응분할 방법이 제안되었다[1]. 이는 현재 변수의 분포가 균일한지를 시험하기 위해서 공간을 chi-square χ^2에 기초하여 분할하는 반복기법이다. 이 방법의 수행과정을 요약하면 다음과 같다.

단계 1 : 주어진 x와 y의 2차원 범위 Rn이 주어지면 2x2 grid로 나눈다. Rn내의 전체관찰 수는 cRn이고, 각 부분할에서 관찰 수는 $cR_{n+1}^{ij}(1\le i,j\le 2)$이다.(c : 부분할 수)

단계 2 : 4개 부분할의 관찰 쌍에 chi-square χ^2 시험을 행한다.($\chi^2=\frac{4\sum_{i=1}^{2}\sum_{j=1}^{2}(cR_{n+1}^{ij}-cR_n/4)^2}{cR_n}$)

단계 3 : 만약 chi-square χ^2 시험값이 사전 설정값보다 크면, 단계 1과 2는 부분할을 반복한다.

단계 4 : 만약 chi-square χ^2 시험값이 사전 설정값보다 적거나 Rn이 너무 작으면, 분할을 멈추고 정규 분할 히스토그램 PDF 근사화에 기초한 방법과 동일한 과정을 수행한다.

이상의 적응분할 방법은 정규분할에 의한 방법보다 좀 더 정확한 상호정보를 얻을 수 있다. 본 실험에서는 사전 설정값을 7.8로 하였다.

결국 PCA와 적응분할 히스토그램 PDF 근사화 방법을 조합하면 빠르고 정확하게 입력변수를 선택할 수 있을 것이다. 여기서 PCA는 전처리 과정으로 좀 더 빠르게 상호독립인 입력변수를 얻기 위함이고, 적응분할 히스토그램 PDF 근사화 방법은 변수간의 상호정보를 좀 더 정확하게 얻기 위함이다.

3. 실험 및 결과 고찰

PCA와 적응적 분할 히스토그램 PDF 근사화에 기초한 상호정보 추출방법에 의한 제안된 입력변수선택 방법의 성능을 평가하기 위해 각 500개의 샘플을 가지는 7개의 신호를 가지는 인위적인 문제와 각 55개의 샘플을 가진 24 개의 신호를 가지는 환경오염신호를 대상으로 각각 실험하였다. 실험은 펜티엄Ⅳ-3.0G 컴퓨터에서 Matlab 6.5로 구현하였다.

3.1 인위적 신호

제안된 기법의 타당성과 성능을 평가하기 위해 인위적으로 제시된 각각 500개 샘플을 가진 6개의 독립신호와 이에 따른 1개의 종속 신호를 가진 7개 신호를 대상으로 실험하였다. 여기서 6개의 독립신호는 1개의 cosine 및 impulse noise 신호와 각각 2개의 sine 및 saw-tooth 신호들이다. 이들 신호함수들은 다음 식 (5)와 같다.

```
x1 = cos(v/4)
x2 =((rem(v,10)-13)/4)
x3 =sin(v/9)
x4 =((rem(v,27)-13)/9)
x5 = ((rand(1,nt)<.1)*2-1).*log(rand(1,nt))
x6 = sin(v/3)                                   (5)
```

위식 (5)에서 x2와 x4는 각각 saw-tooth 신호이고 x5는 impulse noise 신호이다. 또한 nt는 1에서 500까지의 500개 샘플이다. 그림 1은 x1부터 x6까지의 신호를 위에서부터 아래로 순차적으로 각각 도시한 것이다.

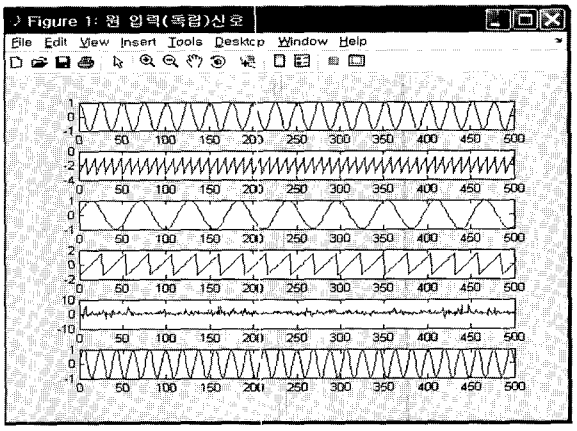


그림 1. 실험에 이용된 6개의 독립신호
Fig. 1. 6 independent signals for experiment

그림 2는 입력신호를 대상으로 PCA에 의해 전처리된 신호를 나타낸 것이다. 이는 2차원의 종속성을 제거한 독립신호로 이를 대상으로 상호정보를 추정한다.

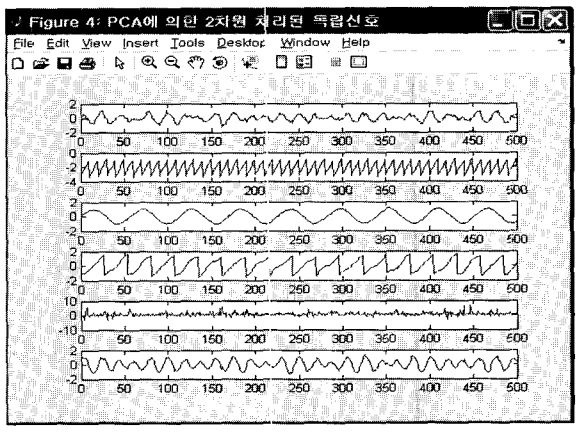


그림 2. 전처리된 6개의 독립신호
Fig. 2. 6 preprocessed independent signals

그림 3은 6개의 입력신호인 독립신호로부터 인위적으로 생성된 종속신호 y = x12 + 2x3 + x5를 도시한 것이다. 여기서 종속신호는 독립신호 중에서 x1, x3, 그리고 x5의 3개 신호에 의해서 생성되도록 하였다.

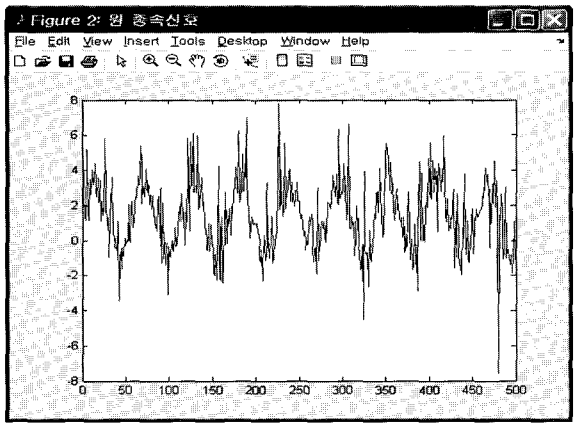


그림 3. y = x12 + 2x3 + x5의 종속신호
Fig. 3. Dependent signal of x12 + 2x3 + x5

한편 그림 4는 그림 2의 6개 독립신호 x와 그림 3의 종속신호 y를 대상으로 PCA와 적응분할의 상호정보(PCA+AP-MI) 추정을 조합한 제안된 방법과 단순히 AP-MI만을 수행한 결과를 각각 도시한 것이다. 여기서 chi-square 시험을 위한 사전 설정값은 7.8로 하였다. 그림 4에서 보면 제안된 PCA+AP-MI 방법에서 종속변수 y와 6개의 독립변수 중 x1, x3, x5와의 상호정보량은 각각 0.058396, 0.736867, 0.062279로 비교적 큰 값을 가지나 x2, x4, x5는 각각 0.000128, 0.003203, 0.000032의 작은 값을 가짐을 알 수 있다. 이는 6개의 입력변수 중에서 x1, x3, x5가 종속변수 y와 관계되는 변수임을 나타내는 것이고 나머지 3개의 입력변수는 상대적으로 종속변수에 영향을 미치지 못함을 알 수 있다. 한편 전처리 과정인 PCA를 수행하지 않는 AP-MI만에 의한 추정에서 독립변수x1, x4, x5는 각각 0.001843, 0.005125, 0.030248의 상호정보량을 가지며, x2, x3, x6은 각각 0.052128, 0.077745, 0.081154의 상호정보량을 가진다. 따라서 전체적으로 종속변수에 영향을 미치는 독립변수를 선택하기가 상대적으로 어려움을 알 수 있다. 특히 독립변수 x2, x3, x6은 상대적으로 높은 상호정보량을 가지나 x3를 제외한 x2와 x6은 실제로 종속변수 y에 영향을 미치지 않으며, 이는 잘못된 입력선택의 결과이다. 결국 제안된 방법은 인위적으로 제시된 문제에서 정확하게 입력변수를 선택하나 단순히 AP-MI만을 이용한 기존 방법은 1개의 변수만을 선택하며, 나머지 2개의 변수는 선택하지 못함을 알 수 있다. 특히 AP-MI만을 이용한 방법에서 3개의 입력변수를 선택할 경우 2개의 잘못된 선택결과를 가짐을 알 수 있다. 이는 입력변수 상호간의 종속성에 의한 과추정으로 PCA의 전처리가 입력변수 상호간의 종속성을 감소시킴을 보여 주는 것이다. 따라서 제안된 조합기법은 입력변수선택을 위한 우수한 성능이 있음을 알 수 있다.

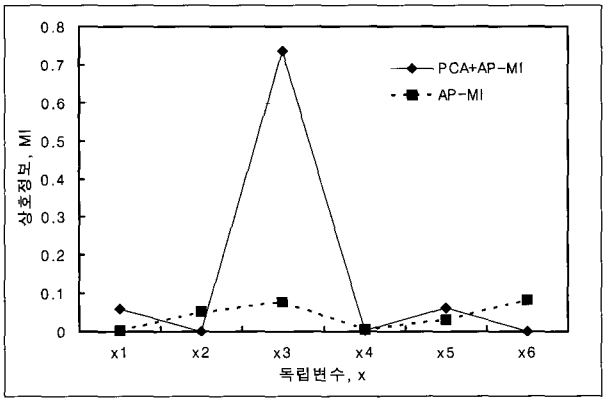


그림 4. 6개 독립신호와 종속신호와의 상호정보량
Fig. 4. Mutual informations between 6 independent signals and dependent signal

3.2 환경 오염신호

입력변수로 포항을 중심으로 여름과 겨울 동안에 실제 측정된 23개의 중금속과 그에 따른 직경 10㎛ 이하의 물질(particular material 10㎛ : PM10) 1개를 가진 24개의 환경 오염신호를 대상으로 실험하였다. 실험에서는 23개의 중금속 중에서 PM10에 가장 영향을 많이 미치는 주요인을 분석하였으며, PCA에 의한 전처리를 수행한 후 AP-MI와 RP-MI를 각각 조합한 방법으로 실험하였다.

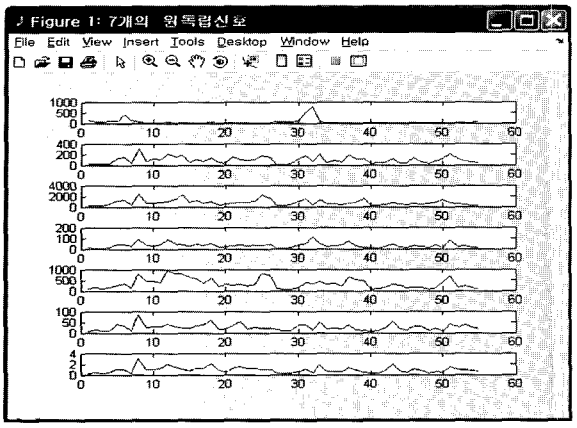
표 1은 실험에 이용된 독립변수 23개의 중금속 중에서 7개와 종속변수 PM10 각각에 대해서 55개의 샘플 중 10개의 샘플만을 나타낸 것이다.

표 1. 55개 중 10개의 샘플을 가진 7개 독립변수와 1개 종속변수
Table 1. 7 independent variables and 1 dependent variable of 10 samples in 55 samples

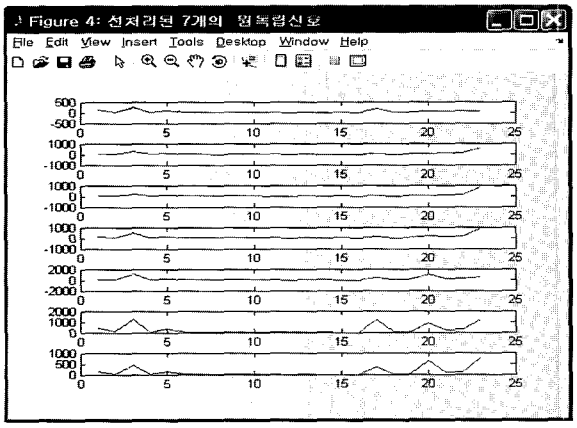
샘플	독립변수						…		종속변수
	Cu	Zn	Fe	Mn	Al	Pb	…	Cd	PM10
1	124.83	17.37	232.75	7.87	96.34	6.95	…	0.23	36.01
3	66.00	15.75	223.95	6.02	89.63	8.80	…	0.23	24.51
7	130.62	34.74	434.47	26.86	151.23	9.03	…	0.23	32.22
11	25.71	102.60	786.72	44.47	377.73	30.34	…	1.16	189.95
13	20.15	160.96	1482.90	51.18	841.84	27.79	…	1.39	44.86
15	33.12	51.65	833.74	28.25	625.77	19.45	…	0.69	80.79
21	27.79	141.97	747.35	29.41	252.21	32.42	…	1.16	66.30
28	64.30	12.49	178.44	4.47	62.15	8.20	…	0.19	14.79
50	11.05	108.66	1370.59	11.22	405.71	7.45	…	0.36	56.93
55	112.43	36.00	172.88	9.02	43.71	15.89	…	0.47	28.61

그림 5는 표 1의 23개의 독립변수들을 대상으로 PCA를 적용하여 전처리한 결과 중에서 7개의 독립변수들과 그 전처리 결과를 나타낸 것이다. 그림 5(a)는 7개의 독립변수 Cu, Zn, Fe, Mn, Al, Pb, Cd를 각각 위에서부터 아래로 차례로 도시한 것이다. 그림 5(b)는 PCA를 적용하여 전처리된 상호

독립인 그림 5(a)의 변수들을 차례로 도시한 것이다.



(a) 독립신호



(b) 전처리된 신호

그림 5. 실험에 이용된 7개 독립신호와 전처리된 신호
Fig. 5. 7 independent signals for experiment and preprocessed signals

그림 6은 23개의 독립변수와 연계되어 생성되는 종속변수 PM10을 도시한 것이다. 종속변수 PM10은 또 다른 환경오염 요인을 포함하고 있지만, 여기에서는 23개의 중금속만으로 구성된다고 가정하였다.

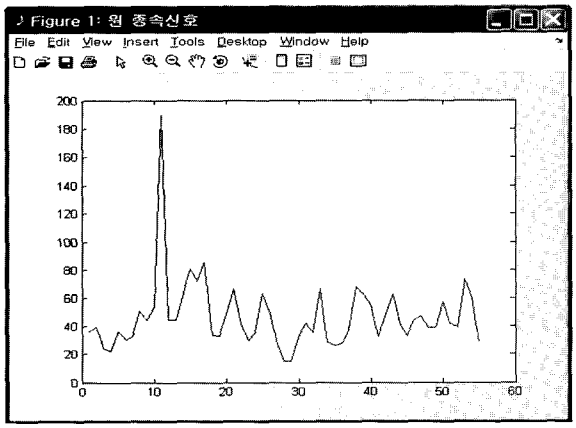


그림 6. 23개의 독립변수에 따른 종속변수 PM10
Fig. 6. Dependent signal PM10 by 23 independent signals

그림 7은 23개 독립변수와 종속변수를 대상으로 PCA로 전처리를 수행한 후, 적응분할의 상호정보(AP-MI)와 정규분할의 상호정보(RP-MI)를 각각 수행하여 추출된 상호정보량을 도시한 것이다. 여기서도 chi-square 시험을 위한 사전 설정값은 7.8로 하였다. 그림 7에서 보면 AP-MI와 RP-MI의 평균 상호정보량은 각각 0.06과 0.03으로 AP-MI가 RP-MI보다 평균적으로 큰 상호정보 값을 가지며, 상호정보량이 0.06이상의 중금속만을 살펴보면, AP-MI의 경우 Fe, Al, Pb, Ba, Sb, Ti, Ca, Mg, Na의 9개이고, RP-MI 경우는 Zn, Fe, Cd, Ti, Si, Mg의 6개임을 알 수 있다. 이는 상대적으로 AP-MI 방법이 RP-MI 방법보다 좀 더 정확한 정보추출 능력이 있기 때문으로 추측된다. 결국 입력변수인 독립변수 23개 중에서 9개의 중금속들이 주로 종속변수 PM10에 영향을 미침을 알 수 있다. 한편 입력의 선택을 위한 문턱값은 필요한 입력의 선택기준을 설정하는 것으로 일반적으로 문턱값보다 더 높은 상호정보 값을 가지는 입력을 선택하게 된다. 하지만 추출을 위한 문턱값의 설정은 지금까지 주로 휴리스틱하게 이루어지며, 적응적 설정과 같은 연구가 추가적으로 이루어져야 할 것이다. 따라서 환경 오염신호 분석에서 보면, PCA+AP-MI가 PCA+RP-MI보다 좀 더 정확하게 입력변수를 선택할 수 있음을 알 수 있다.

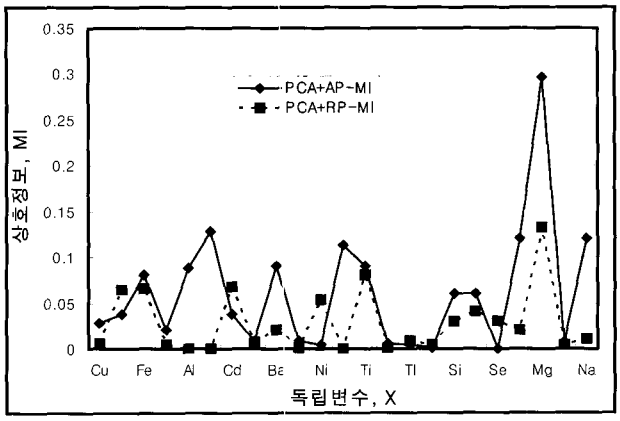


그림 7. 23개 독립변수와 종속변수와의 상호정보량
Fig. 7. Mutual informations between 23 independent signals and dependent signal

이상의 인위적인 생성신호와 환경 오염신호의 실험결과로부터 PCA에 의한 전처리는 입력변수를 좀 더 정확하고 빠르게 추출할 수 있도록 하며, AP-MI 역시 RP-MI보다 좀 더 정확한 추출을 성능이 있음을 확인하였다. 따라서 PCA와 AP-MI를 조합한 상호정보 추출법을 이용하면 빠르고 정확하게 입력변수들을 선택할 수 있을 것이다.

4. 결 론

본 논문에서는 주요성분분석과 적응분할 히스토그램 PDF 근사화에 기초한 상호정보 추정을 조합한 입력변수선택 방법을 제안하였다. 여기서 PCA는 2차원 통계성을 이용하여 입력변수들 사이의 종속성을 빠르고 정확하게 제거하여 과추정을 방지하기 위함이고, 적응분할 히스토그램 PDF 근사화에 기초한 상호정보 추정은 입력변수 상호간의 종속성을 효과적으로 추정하기 위함이다.

재안된 기법을 인위적으로 제시된 각 500개의 샘플을 가지는 7개의 신호와 포항지역을 대상으로 측정된 각 55개의 샘플을 가진 24개의 환경오염신호를 대상으로 실험한 결과, PCA에 의한 전처리는 입력변수를 좀 더 정확하고 빠르게 추출할 수 있도록 하며, 적응분할이 정규분할보다 더 정확한 변수 추출성능이 있음을 확인하였다. 따라서 PCA와 적응분할의 상호정보 추출을 조합한 제안된 방법을 이용하면 빠르고 정확하게 입력변수들을 선택할 수 있다.

향후 제안된 PCA를 이용한 방법을 좀 더 다양한 분야와 큰 규모의 문제에 적용하는 연구와 입력의 선택을 위한 문턱값의 적응적 설정을 위한 연구가 추가적으로 이루어져야 할 것이다. 또한 독립성분분석의 기법과도 비교연구가 이루어져야 할 것이다.

참 고 문 헌

[1] T. Trappenberg, J. Ouyang, and A. Back, "Input Variable Selection : Mutual Information and Linear Mixing Measures", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 1, No. 8, pp. 37-46, Jan. 2006

[2] A. Back and A. Cichocki, "Input Variable Selection Using Independent Component Analysis and Higher Order Statistics", *Proc. of ICA99*, Jan. 1999

[3] A. Back and T. Trappenberg,"Input Variable Selection Using Independent Component Analysis," *International Joint Conference on Neural Networks*, pp. 1-5, Washington, 1999

[4] A. Back and T. Trappenberg, "Selecting Inputs for Modelling Using Normalized Higher Order Statistics and Independent Component Analysis," *IEEE Transactions on Neural Networks*, Vol. 12, No. 3, pp. 612-617, March. 2001

[5] K. I. Diamantaras and S. Y. Kung, *Principal Component Neural Networks : Theory and Applications, Adaptive and learning Systems for Signal Processing, Communications, and Control*, John Wiley & Sons, Inc., 1996

[6] S. Haykin, Neural Networks : *A Comprehensive Foundation*, Prentice-Hall, 2ed, London, 1999

[7] A. Cichock and R. Unbehaun, *Neural Networks for Optimization and Signal Processing*, John Wiley & Sons., New York, 1993

[8] P. Foldiak, "Adaptive Network for Optimal Linear Feature Extraction," *International Joint Conference on Neural Networks*, Washington D.C., Vol. 1, pp. 401-406, June 1989

저 자 소 개

조용현(Yong-Hyun Cho)

1979년 : 경북대학교 전자공학과(공학사)
1981년 : 경북대학교 대학원 전자공학과 (공학석사)
1993년 : 경북대학교 대학원 전자공학과 (공학박사)
1983년~1984년 : 삼성전자(주)
1984년~1987년 : 한국전자통신연구원
1987년~1997년 : 영남이공대학 전자과 교수
1997년~현재 : 대구가톨릭대학교 컴퓨터 정보통신공학부 교수

관심분야 : 신경회로망, 영상신호처리 및 인식, 상황 인식, 전전자교환기 등
Phone : +82-53-850-2747
Fax : +82-53-850-2740
E-mail : yhcho@cu.ac.kr

홍성준(Seong-Jun Hong)

1986년 : 경북대학교 전자공학과(공학사)
1988년 : 경북대학교 대학원 전자공학과 (공학석사)
2004년~현재 : 대구가톨릭대학교 대학원 컴퓨터정보통신공학과 박사과정 수료
1993년~현재 : 대구산업정보대학 컴퓨터 정보계열 부교수

관심분야 : 신경회로망, 영상 및 음성 신호 처리 및 인식, e-Learning 시스템, 전자상거래 등
Phone : +82-53-749-7215
Fax : +82-53-749-7218
E-mail : sjishong@tpic.ac.kr