퍼지 및 지능시스템학회 논문지 2006, Vol. 16, No. 5, pp. 525-530

고정점 알고리즘의 독립성분분석과 적응분할의 상호정보 추정에 의한 입력변수선택

Input Variable Selection by Using Fixed-Point ICA and Adaptive Partition Mutual Information Estimation

조용현

Yong-Hyun Cho

* 대구가톨릭대학교 컴퓨터정보통신공학부

요 약

본 논문에서는 고정점 알고리즘의 독립성분분석과 적응분할의 상호정보 추정을 조합한 입력변수선택 기법을 제안하였다. 여기서 고정점 알고리즘의 독립성분분석은 할선법에 기반을 둔 방법으로 입력변수 간의 독립성을 빠르게 찾기 위함이고, 적응분할의 상호정보 추정은 입력변수의 확률밀도함수 계산에서 동일한 량의 샘플분할을 가능하게 하여 변수상호간의 종속성을 좀 더 정확하게 구하기 위함이다. 제안된 기법을 인위적으로 제시된 각 500개의 샘플을 가지는 7개의 신호와 특정지역을 대상으로 측정된 각 55개의 샘플을 가진 24개의 환경오염신호를 대상으로 실험한 결과, 빠르고 정확한 변수의 선택이 이루어짐을 확인하였다. 또한 할선법의 고정점 알고리즘 독립성분분석을 수행하지 않을 때와 정규분할의 상호정보 추정 때보다 각각 우수한 선택성능이 있음을 확인하였다.

키워드 : 입력변수선택, 고정점 알고리즘, 독립성분분석, 상호정보, 할선법, 적응분할

Abstract

This paper presents an efficient input variable selection method using both fixed-point independent component analysis(FP-ICA) and adaptive partition mutual information(AP-MI) estimation. FP-ICA which is based on secant method, is applied to quickly find the independence 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(PDF). 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 FP-ICA and regular partition MI estimation.

Key Words : Input Variable Selection, Fixed-Point Algorithm. Independent Component Analysis Mutual Information, Secant Method, Adaptive Partition.

1. 서 론

생체인식, 산업, 환경 시스템 등과 같은 실세계의 모델링에서 선택되는 입력변수에 따라 시스템의 성능은 크게 달라진다. 특히 많은 입력변수들 중에서 모델이 얼마나 많은 또는 어느 입력들을 필요로 하는지 알 수 없는 문제가 있다[1-4]. 이러한 문제는 입력차원이 증가할수록 더욱 더 심각하며, 입력변수선택은 어느 입력변수들이 어떤 모델을 위해 요구되는지를 결정하는데 목적이 있다. 결국 입력변수선택은 어떤 의미에서 최적의 모델을 유도할 입력집합을 선택하는 것이다. 일반적으로 입력변수의 효과적인 선택은 시스템의 차원감소나 특징추출 등 다양한 용도로 이용되며, 특히 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[1-5]. 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 방

접수일자 : 2006년 9월 13일
완료일자 : 2006년 10월 9일

법은 기초모델을 가지지 않는 통계적 종속성 시험에 바탕을 둔 기법으로 입력변수들의 부집합과 원하는 출력사이의 통계적 시험을 수행함으로써 이루어진다. 이때 시험은 이들 결과에 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[1,5].

model-free 방법은 통계적 종속성에 기반을 둠으로써 model-based 방법보다 좀 더 일반화된 방법이다. 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. 이러한 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 할선법(secant method)[6]의 고정점(fixed-point : FP) 알고리즘 독립성분분석(independent component analysis : ICA)[7-10]과 적응분할(adaptive partition : AP) 히스토그램 PDF 근사화에 기초한 상호정보 추정법을 조합한 입력변수선택 방법을 제안한다. 여기서 할선법의 FP-ICA는 입력변수들 사이의 종속성을 빠르고 정확하게 제거하여 과추정을 방지하기 위함이고, 적응분할 히스토그램 PDF 근사화에 기초한 상호정보 추정법은 입력변수의 PDF를 정확하게 계산하여 변수상호간의 종속성을 효과적으로 추정하기 위함이다. 제안된 기법을 인위적으로 제시된 각 500개의 샘플을 가지는 7개의 신호와 특정지역을 대상으로 측정된 각 55개의 샘플을 가진 24개의 환경오염신호를 대상으로 실험하여 타당성을 확인하고, 기존의 FP-ICA의 전처리 없는 입력선택 및 정규적 분할(regular partition : RP)에 기초한 방법의 결과와 비교 · 분석하였다.

2. 독립성분분석과 상호정보 추정

2.1 할선법의 고정점 알고리즘 독립성분분석

독립성분분석은 미지신호의 혼합인 관찰신호만으로 독립인 신호를 추출하는 신호처리 기법이다[7-10]. 이는 신호의 2차 이상의 고차 통계성을 이용한 것으로 서로 종속인 신호로부터 통계적 독립신호를 만드는 선형기법이다. ICA는 은닉신호의 분리(blind source separation : BSS)나 특징추출에 널리 이용되고 있으며, 여기에서는 BSS에 대해서 알아본다[7-10].

ICA는 m개의 입력신호 **s**로부터 선형적으로 혼합된 n개의 신호 **x**가 알려져 있을 때, 혼합된 신호로부터 역으로 m개의 독립인 입력신호를 찾는 기법이다. 하지만 입력신호들을 혼합할 때의 혼합행렬 **A**는 알려져 있지 않으며, 혼합과정에서 잡음 **n**이 추가 될 수도 있다. 이때 혼합신호와 입력신호와의 관계는 다음 식 (1)과 같다.

$$\mathbf{x} = \mathbf{As} + \mathbf{n} = \sum_{i=1}^{m} s(i)a(i) + \mathbf{n} \qquad (1)$$

여기서 n은 보통 입력신호와 구별되지 않기 때문에 생략할 수도 있으며, **A**=[**a**(1), **a**(2), …, **a**(m)]으로 **a**(i)는 ICA의 basis vector이다. 결국 ICA는 알려진 혼합신호로부터 혼합행렬의 역행렬 $\mathbf{A}^{-1}(=\mathbf{W})$을 찾는 기법이다. 혼합행렬 **A**와 역혼합행렬 W에 대하여 상세히 살펴보면 다음 그림 1과 같은 구성도로 나타낼 수 있다. 여기서 $\mathbf{x} = \mathbf{As}$이고, $\mathbf{y} = \mathbf{Wx}$이다. 그림에서 보면 ICA는 혼합행렬과 일치하는 역혼합행렬을 찾는 과정에서 출력신호가 독립성을 가지도록 하는 기법이다. 결국 ICA는 알려진 혼합신호 **x**로부터 출력신호 **y**를 찾는 기법으로 궁극적으로는 역혼합행렬 W을 찾아서 원신호 **s**의 근사값을 알아내는 것이다.

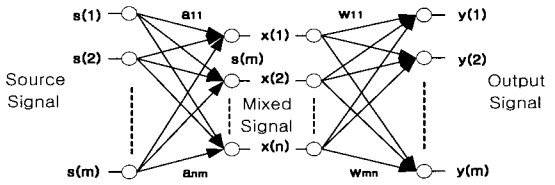


그림 1. 혼합행렬과 역혼합행렬의 상세 설명도
Fig. 1. Detailed diagram of mixing and inverse mixing matrix.

최근 ICA를 위한 다양한 알고리즘들이 연구되었다[7,8]. 그 중에서도 고정점 알고리즘은 신경망이 가지는 병렬성과 분산성, 그리고 더 작은 메모리 요구 등의 제약을 해결하기 위해 제안된 batch mode로 계산되는 ICA 기법이다[8-10]. 특히 FP 알고리즘은 엔트로피 최적화 방법으로부터 유도되고, 지금까지 알려진 기법 중 가장 빠른 학습속도를 가지며 신호 내에 포함된 상호정보를 최소화함으로써 ICA의 해를 구하는 기법이다. 신호벡터 x의 상관행렬 $E\{\mathbf{xx}^T\}=\mathbf{I}$로 whitening되어 있다고 가정할 때, 근사화된 반복기법의 역혼합행렬 **W**를 구하는 뉴우턴(Newton)법[6]은 다음 식 (2)와 같다.

$$\mathbf{W}^+ = \mathbf{W} - [E\{\mathbf{x}g(\mathbf{W}^T\mathbf{x})\} - \beta\mathbf{W}]/[E\{g'(\mathbf{W}^T\mathbf{x})\} - \beta]$$
$$\mathbf{W}^* = \mathbf{W}^+/\|\mathbf{W}^+\| \qquad (2)$$

여기서 $\mathbf{W}^*$는 **W**의 새롭게 경신된 값이고, $\beta = E\{\mathbf{W}^T\mathbf{x}g(\mathbf{W}^T\mathbf{x})\}$이다. 결국 식 (2)는 뉴우턴법에 기초를 둔 ICA를 위한 FP 알고리즘이다. 또한 식 (2)의 좌측식 양쪽에 $\beta - E\{g'(\mathbf{W}^T\mathbf{x})\}$를 곱해 구해지는 더욱 간단해진 뉴우턴법의 FP 알고리즘은 다음의 식 (3)과 같다.

$$\mathbf{W}^+ = E\{\mathbf{x}g(\mathbf{W}^T\mathbf{x})\} - E\{g'(\mathbf{W}^T\mathbf{x})\}\mathbf{W}, \quad \mathbf{W}^* = \mathbf{W}^+/\|\mathbf{W}^+\| \qquad (3)$$

위의 경신식에서 $g(\cdot)$는 비선형 함수이며 일반적으로 $(\cdot)^3$과 $\tanh(\cdot)$의 함수가 이용된다[7-9].

이상에서 유도된 FP 알고리즘에서는 역혼합행렬 **W**의 경신을 위해 비선형 방정식을 푸는데 이용되는 뉴우턴법을 이용하고 있다. 이는 방정식의 해를 직접적으로나 단순한 방법으로 구할 수 없을 때 쉽게 해결할 수 있는 다른 문제로 근사화하는 방법이다. 하지만 뉴우턴법은 주어진 정확도를 달

성하는데 걸리는 반복수가 적어 빠른 학습속도를 가지나 계산과정에 미분계산의 복잡함에 의해 계산시간이 상대적으로 길어지는 제약이 있다. 따라서 뉴우턴법에서 발생되는 계산의 복잡성을 줄일 수만 있다면 좀 더 빠른 효과적인 FP 알고리즘의 ICA 기법을 수행할 수 있을 것이다.

본 연구에서는 할선법[6]을 이용한 새로운 역혼합행렬 $\mathbf{W}$의 경신을 위한 FP 알고리즘을 제안한다. 할선법은 현재의 방정식 값과 바로 이전의 방정식 값만을 이용하여 해를 구할 수 있어 별도의 1차 미분이 요구되지 않는다. 이는 뉴우턴법에 비해 할선법이 하나의 반복에 소요되는 시간이 적게 걸리는 이유이다. 따라서 β를 $\mathbf{W}_0$ 대신에 $\mathbf{W}$의 현재값으로 나타낸 근사화된 반복방법의 역혼합행렬 $\mathbf{W}$를 구하기 위한 할선법은 다음 식 (4)와 같다.

$$\begin{aligned} f(\mathbf{W}^{\#}) &= [E\{\mathbf{x}g(\mathbf{W}^{\#T}\mathbf{x})\} - \beta\mathbf{W}^{\#}] \\ f(\mathbf{W}) &= [E\{\mathbf{x}g(\mathbf{W}^{T}\mathbf{x})\} - \beta\mathbf{W}] \\ \mathbf{W}^{+} &= \mathbf{W} - f(\mathbf{W})[\{\mathbf{W}-\mathbf{W}^{\#}\}/\{f(\mathbf{W})-f(\mathbf{W}^{\#})\}] \\ \mathbf{W}^{*} &= \mathbf{W}^{+}/\|\mathbf{W}^{+}\| \end{aligned} \quad (4)$$

여기서 $\mathbf{W}^{*}$는 $\mathbf{W}$의 새롭게 경신된 값이고, $\mathbf{W}^{\#}$은 바로 전에 계산된 $\mathbf{W}$의 값이다.

따라서 제안된 할선법의 FP-ICA를 이용하면 좀 더 빠르고 정확하게 독립성을 가지는 변수로의 변환이 가능하다. 결국 FP-ICA는 입력변수 x로부터 독립인 변수 s를 추정하는 전처리과정으로 이용하며, s를 대상으로 원하는 입력변수들을 선택하기 위한 통계적인 시험을 수행한다. 이렇게 하면 통계적 종속성 측정에 기반을 둔 빠르면서도 정확한 model-free 입력변수의 선택이 유도될 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[1-5]. 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 (5)와 같이 정의된다[1].

$$I(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\cdot\log\left(\frac{f(x,y)}{f(x)f(y)}\right)dxdy \quad (5)$$

여기서 x와 y가 서로 독립이면 상호정보 I(x,y)는 영이 된다. 또 다른 상호정보는 엔트로피(entropy)를 이용하여 다음 식 (6)과 같이 정의 될 수 있다.

$$I(x,y) = H(x) + H(y) - H(x,y) \quad (6)$$

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

식 (5)와 식 (6)에서 각각 상호정보의 계산을 위해서는 복잡한 joint PDF와 marginal PDF의 추정이 요구된다. 이러한 추정법으로 Gram-Charlier 확장에 기초한 방법, 정규분할 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. 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 R_n이 주어지면 2x2 grid로 나눈다. R_n내의 전체관찰 수는 cR_n이고, 각 부분할에서 관찰 수는 cR_{n+1}^{ij} $(1\leq i,j\leq 2)$이다.(c : 부분할 수)

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

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 R_n이 너무 작으면, 분할을 멈추고 정규분할 히스토그램 PDF 근사화에 기초한 방법과 동일한 과정을 수행한다.

이상의 적응분할 방법은 정규분할에 의한 방법보다 좀 더 정확한 상호정보를 얻을 수 있다.

따라서 할선법의 FP-ICA와 적응분할 히스토그램 PDF 근사화 방법을 조합하면 빠르고 정확하게 입력변수를 선택할 수 있을 것이다. 여기서 할선법의 FP-ICA는 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

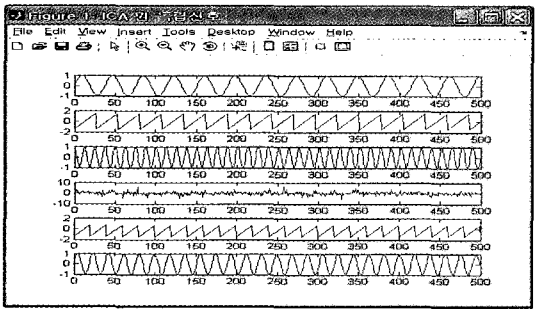
전처리 과정으로 할선법의 FP-ICA와 적응분할 히스토그램 PDF 근사화에 기초한 상호정보 추출법을 조합한 제안된 입력변수선택 방법의 성능을 평가하기 위해 2가지 문제를 대상으로 실험하였다. 대상 신호는 인위적으로 제시된 각 500개의 샘플을 가지는 7개의 신호와 특정지역을 대상으로 측정된 각 55개의 샘플을 가진 24개의 환경오염 신호이다. 실험은 펜티엄Ⅳ-3.0G 컴퓨터에서 Matlab 6.5[11]로 구현하였다.

3.1 인위적 생성신호

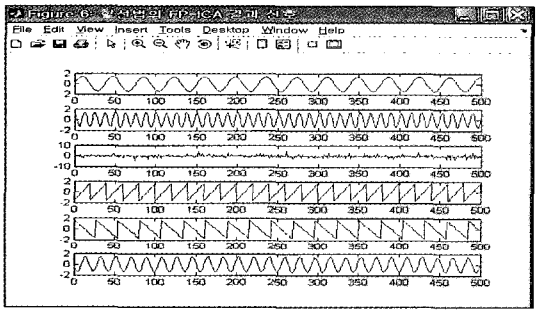
제안된 기법의 타당성과 성능을 평가하기 위해 인위적으로 제시된 각각 500개 샘플을 가진 6개의 독립신호와 이에 따른 1개의 종속 신호를 가진 7개 신호를 대상으로 실험하였다. 여기서 6개의 독립신호는 각각 2개의 sine과 saw-tooth 신호, 1개의 cosine과 impulse noise 신호들이다. 이들 신호 함수들은 다음 식 (7)과 같다.

$$
\begin{aligned}
x_1 &= \sin(r/6) \\
x_2 &= ((\mathrm{rem}(r,27)-13)/9) \\
x_3 &= \cos(r/2) \\
x_4 &= ((\mathrm{rand}(1,nt)<.5)*2-1).*\log(\mathrm{rand}(1,nt)) \\
x_5 &= ((\mathrm{rem}(r,20)-13)/9) \\
x_6 &= \sin(r/3)
\end{aligned} \tag{7}
$$

여기서 x_2와 x_5는 각각 saw-tooth 신호이고 x_4는 impulse noise 신호이다. 또한 r은 1에서 500까지의 변수이고 nt는 상수값 500이다. 그림 2는 식 (7)의 함수들에 의해 생성된 6개의 독립신호와 이를 대상으로 전처리된 신호들을 나타낸 것이다. 그림 2(a)는 x_1부터 x_6까지 6개의 독립신호를 위에서부터 아래로 차례로 도시한 것이고, 그림 2(b)는 그림 2(a)의 입력된 독립신호에 할선법의 FP-ICA를 적용하여 전처리된 상호 독립인 신호들을 도시한 것이다. 여기서 보면 신호의 추출순서와 부호가 각각 바뀐 신호들을 볼 수 있다. x_1과 x_6만 제외한 모든 추출신호는 순서가 바뀌었고, x_2, x_4, x_6는 부호가 바뀌었음도 알 수 있다. 이는 신호의 추출순서와 부호의 변화를 고려하지 않는 ICA 고유의 속성 때문이다.



(a) 독립신호



(b) 전처리된 신호

그림 2. 실험에 이용된 6개 독립신호와 전처리된 신호
Fig. 2. 6 independent signals for experiment and preprocessed signals

그림 3은 6개의 독립신호로부터 인위적으로 생성된 종속신호 y로 $y = x_1^2 + 2x_3 + x_5$를 도시한 것이다. 여기서 종속신호는 독립신호 중에서 x_1, x_3, 그리고 x_5의 3개 신호에 의해서 생성되도록 하였다.

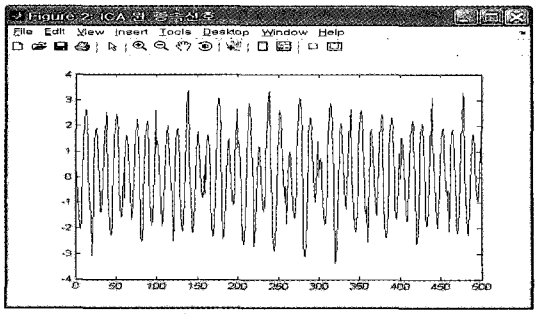


그림 3. $x_1^2 + 2x_3 + x_5$의 종속신호
Fig. 3. Dependent signal of $x_1^2 + 2x_3 + x_5$

한편 그림 4는 그림 2의 6개 독립신호 x와 그림 3의 종속신호 y를 대상으로 할선법의 FP-ICA와 적응분할의 상호정보(AP-MI) 추정을 조합한 제안된 방법과 단순히 AP-MI만을 수행한 결과를 도시한 것이다. 여기서 chi-square 시험을 위한 사전 설정값은 7.8로 하였다. 그림 4에서 보면 제안된 FP-ICA+AP-MI 방법에서 종속변수 y와 6개의 독립변수 중 x_1, x_3, x_5와의 상호정보량은 각각 0.174386, 0.046383, 0.722109로 비교적 큰 값을 가지나 x_2, x_4, x_5는 각각 0.001022, 0.00086, 0.000028의 작은 값을 가짐을 알 수 있다. 이는 6개의 입력변수 중에서 x_1, x_3, x_5가 종속변수 y와 관계되는 변수임을 나타내는 것이고 나머지 3개의 입력변수는 종속변수에 영향을 미치지 못함을 알 수 있다. 한편 전처리 과정인 FP-ICA를 수행하지 않는 AP-MI만에 의한 추정에서 독립변수 x_1와 x_6는 종속변수 y와 각각 0.000128과 0.000032, x_2와 x_4는 각각 0.000425, 그리고 x_3와 x_5는 각각 0.700125와 0.017025의 상호정보량을 가진다. 여기에서는 독립변수 x_3와 x_5만이 종속변수 y에 영향을 미치며, 나머지 4개의 독립변수는 영향을 거의 미치지 못함을 알 수 있다. 결국 제안된 방법은 인위적으로 제시된 문제에서 정확하게 입력변수를 선택하나 단순히 AP-MI만을 이용한 기존 방법은 2개의 변수만을 선택하며, 나머지 1개의 변수는 선택하지 못함을 알 수 있다. 특히 AP-MI만을 이용한 방법에서 3개의 입력변수를 선택할 경우 x_2와 x_4가 동일한 값을 가져 4개가 선택된다. 이는 입력변수 상호간의 종속성에 기인한 과추정으로 FP-ICA의 전처리가 입력변수 상호간의 종속성을 감소시킴을 보여 주는 것이다. 따라서 제안된 조합기법은 입력변수선택을 위한 우수한 성능이 있음을 알 수 있다.

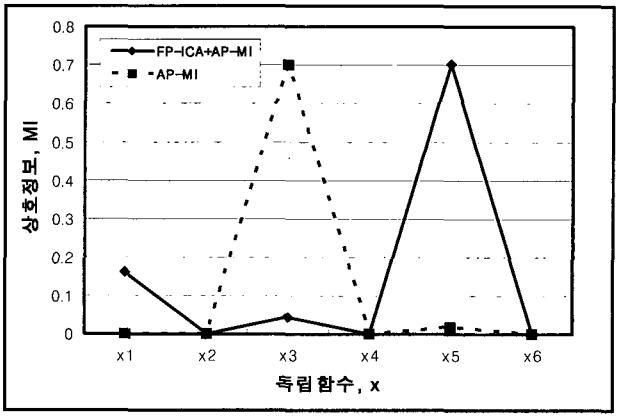


그림 4. 6개 독립신호와 종속신호와의 상호정보량
Fig. 4. Mutual informations between 6 independent signals and dependent signal

3.2 환경 오염신호

입력변수로 특정지역을 대상으로 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, 할선법의 FP-ICA에 의한 전처리를 수행한 후 AP-MI와 RP-MI를 각각 조합한 방법으로 실험하였다.

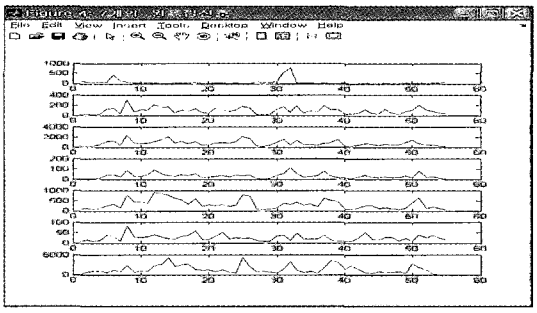
표 1은 실험에 이용된 독립변수 23개의 중금속 중에서 7개와 종속변수 PM10 각각에 대해서 55개의 샘플 중 10개의 샘플만을 나타낸 것이다.

표 1. 55개 중 10개의 샘플을 가진 7개 독립변수와 1개 종속변수

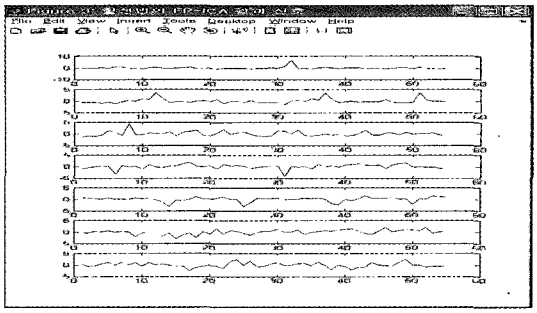
Table 1. 7 independent variables and 1 dependent variable of 10 samples in 55 samples

샘플	독립변수								종속변수
	Cu	Zn	Fe	Mn	Al	Pb	…	Na	PM10
1	124.83	17.37	232.75	7.87	96.34	6.95	…	69.25	36.01
3	66.00	15.75	223.95	6.02	89.63	8.80	…	805.95	24.51
7	130.62	34.74	434.47	26.86	151.23	9.03	…	794.37	32.22
11	25.71	102.60	786.72	44.47	377.73	30.34	…	898.58	189.95
13	20.15	160.96	1482.90	51.18	841.84	27.79	…	2654.30	44.86
15	33.12	51.65	833.74	28.25	625.77	19.45	…	2139.93	80.79
21	27.79	141.97	747.35	29.41	252.21	32.42	…	715.63	66.30
28	64.30	12.49	178.44	4.47	62.15	8.20	…	900.13	14.79
50	11.05	108.66	1370.59	11.22	405.71	7.45	…	2882.53	56.93
55	112.43	36.00	172.88	9.02	43.71	15.89	…	76.00	28.61

그림 5는 표 1의 7개 독립변수들을 대상으로 할선법의 FP-ICA를 적용하여 전처리한 변수들을 나타낸 것이다. 그림 5(a)는 7개의 독립변수 Cu, Zn, Fe, Mn, Al, Pb, Na를 각각 위에서부터 아래로 차례로 도시한 것이다. 그림 5(b)는 그림 5(a)의 입력된 독립변수에 할선법의 FP-ICA를 적용하여



(a) 독립신호



(b) 전처리된 신호

그림 5. 실험에 이용된 7개 독립신호와 전처리된 신호

Fig. 5. 7 independent signals for experiment and preprocessed signals

전처리된 상호 독립인 변수들을 도시한 것이다. 여기에서도 그림 2에서처럼 변수의 추출순서와 부호의 변화와 같은 ICA 고유의 속성을 확인할 수 있다.

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

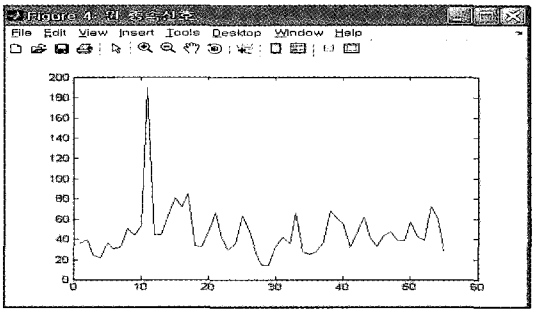


그림 6. 23개의 독립변수에 따른 종속변수 PM10

Fig. 6. Dependent signal PM10 by 23 independent signals

그림 7은 23개 독립변수와 종속변수를 대상으로 할선법의 FP-ICA로 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-MI가 RP-MI보다 평균적으로 큰 상호정보 값을 가지며, 상호정보량이 0.02이상인 중금속만을 살펴보면, AP-MI의 경우 Cu, Cd, Cr, Sb, Mo, Ca, Mg, Na의 8개이고, RP-MI 경우는 Cu, Cd, Cr, Ni, Na의 5개임을 알 수 있다. 이는 상대적으로 AP-MI 방법이 RP-MI 방법보다 좀 더 정확한 정보추출 능력이 있기 때문으로 추측된다. 결국 입력변수인 독립변수 23개 중에서 8개의 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. 지금까지의 설정방법은 주로 휴리스틱하게 이루어지며, 적응적 설정과 같은 연구가 추가적으로 이루어져야 할 것이다. 따라서 환경 오염신호 분석을 통해 FP-ICA+AP-MI가 FP-ICA+RP-MI보다 좀 더 정확하게 입력변수를 선택할 수 있음을 알 수 있다.

이상의 인위적 생성신호와 환경 오염신호의 실험결과로부

터 할선법의 FP-ICA에 의한 전처리는 입력변수를 좀 더 정확하고 빠르게 추출할 수 있도록 하며, AP-MI 역시 RP-MI 보다 좀 더 정확한 추출을 성능이 있음을 확인하였다. 따라서 할선법의 FP-ICA와 AP-MI를 조합한 상호정보 추출법을 이용하면 빠르고 정확하게 입력변수들을 선택할 수 있을 것이다.

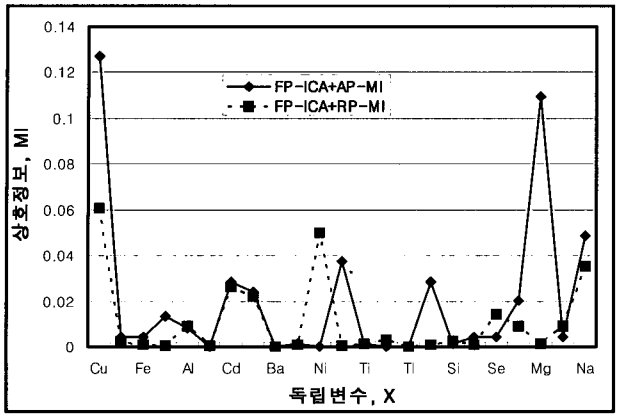


그림 7. 23개 독립변수와 종속변수와의 상호정보량
Fig. 7. Mutual informations between 23 independent signals and dependent signal

4. 결 론

본 논문에서는 할선법의 고정점 알고리즘 독립성분분석과 적응분할 히스토그램 PDF 근사화에 기초한 상호정보 추정을 조합한 입력변수선택 방법을 제안하였다. 여기서 할선법의 FP-ICA는 입력변수들 사이의 종속성을 빠르고 정확하게 제거하여 과추정을 방지하기 위함이고, 적응분할 히스토그램 PDF 근사화에 기초한 상호정보 추정은 입력변수 상호간의 종속성을 효과적으로 추정하기 위함이다.

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

향후 제안된 FP-ICA를 이용한 방법을 좀 더 다양한 분야와 큰 규모의 문제에 적용하는 연구와 입력의 선택을 위한 문턱값의 적응적 설정을 위한 연구가 추가적으로 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.

[2] A. Back and T. Trappenberg, "Input Variable Selection Using Independent Component Analysis," *IJCNN99,* pp. 1-5, Washington, 1999

[3] B. Blinnikiov and A. Weigend, "Selecting Input Variables Using Mutual Information and Nonparametric Density Estimation," *Pro. of ISANN'94,* pp. 42-50, Taiwan, Oct. 1994

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

[5] 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

[6] K. Atkinson, *Elementary Numerical Analysis,* John Wiley & Sons, Inc., New York, 1993

[7] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis,* John Wiley & Sons, Inc., New York, May 2001.

[8] T. W. Lee, *Independent Component Analysis : Theory and Applications,* Kluwer Academic Pub., Boston, Dec. 1998.

[9] A. Hyvarinen and E. Oja, "A Fast Fixed Point Algorithms for Independent Component Analysis," *Neural Computation,* 9(7), pp. 1483-1492, Oct. 1997

[10] J. Karhunen, "Neural Approaches to Independent Component Analysis and Source Separation," *ESANN96,* Burges, Belgium, pp. 249-266, Apr. 1996

[11] J. Wesley Hines, MATLAB *Supplement to Fuzzy and Neural Approaches in Engineering,* John Wiley & Sons, Inc., June 1997.

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 컴퓨터정보통신공학부 교수

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