제26회 한국정보처리학회 추계학술발표대회 논문집 제13권 제2호 (2006. 11)

Fixed-Point ICA와 상호정보 추정에 의한 입력변수선택

조용현, 홍성준*
대구가톨릭학교 컴퓨터정보통신공학부
e-mail:yhcho,sjishong@cu.ac.kr

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

Yong-Hyun Cho, Seong-Jun Hong*
School of Computer and Information Communications. Eng.,
Catholic Univ. of Daegu

요 약

본 논문에서는 고정점 알고리즘의 독립성분분석과 상호정보 추정을 조합한 입력변수선택 기법을 제안하였다. 여기서 뉴우턴법에 기반을 둔 빠른 분석성능을 가지는 고정점 알고리즘의 독립성분분석은 입력변수 간의 독립성을 빠르게 찾기 위함이고, 입력변수의 확률밀도함수의 계산을 위해 적응적 분할을 이용한 상호정보의 추정은 변수상호간의 종속성을 좀 더 정확하게 정량화하기 위함이다. 제안된 기법을 인위적으로 제시된 각 500개의 샘플을 가지는 6개의 독립신호와 1개의 종속신호를 대상으로 실험한 결과 빠르고 정확한 변수의 선택이 이루어짐을 확인하였다.

1. 서론

생체인식, 산업, 환경 시스템 등과 같은 실세계의 모델링에서 가장 적합한 입력만을 선택하는 것은 시스템의 성능에 많은 영향을 미친다. 일반적으로 입력변수의 효과적인 선택은 시스템의 차원의 감소나 특징추출 등 다양한 용도로 이용된다[1-3]. 그러나 많은 입력변수들 중에서 모델에 얼마나 많은 또는 어느 입력들이 필요한지 알 수 없는 문제가 있다. 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.

입력변수의 부적당한 선택은 여러 가지 제약들을 발생시킨다. 여기에는 입력차원의 증가에 따른 계산시간과 메모리의 증가, 요구되지 않는 입력들에 의한 학습의 어려움과 비수렴 및 성능저하, 복잡한 모델의 어려운 해석 등의 제약이 있다. 지금까지 알려진 입력변수선택 기법들은 크게 model-based와 model-free 방법들로 나눌 수 있다[1-3]. 먼저 model-based 방법에서 입력선택과정은 모델을 선정한 후 이용할 입력들을 선택하고, 파리미터들을 최적화한 후 어떤 비용함수를 측정하여 이루어진다. 가장 잘 알려진 선형모델을 이용한 방법으로 분산의 해석(analysis of variance : ANOVA)에 의해 구현되는 전역 F-test 방법이 있다. 또한 비선형 모델을 이용한 방법으로 신경망이나 자동상관성검출(automatic relevance detection : ARD)로 구현된다[1]. 이러한 model-based 방법들은 입력들이 바뀌면 선택과정은 다시 반복하여야 하는 제약이 있다. 한편 model-free 방법은 기초모델을 가지지 않는 통계적 종속성 시험에 바탕을 둔 기법이다. correlation에 기반을 둔 방법, 고차원의 cross-cumulant에 기반을 둔 방법, 상호정보에 기반을 둔 방법 등이 통계적 종속성을 시험하는 방법으로 알려져 있다[1-3].

model-free 방법은 특별한 모델에 의존하지 않으며 모든 결과가 통계적 종속성에 기반을 둠으로써 좀 더 일반화된 방법이다. 종속성 시험방법 중에서 correlation에 기반을 둔 방법은 2변수 사이의 선형 종속성만을 측정하는 2차원 통계성을 이용함으로써 선형모델에만 적용 가능한 제약이 있다. 고차원의 cross-cumulant에 기반을 둔 방법은 고차원의 통계

성을 이용하여 종속성을 측정하는 방법으로 여기에도 입력변수들의 모든 조합들을 조사해야 하는 제약이 있다. 따라서 이러한 제약을 해결하기 위하여 변수들 사이의 정보에 기반을 두고 모든 고차원의 통계성을 이용하여 종속성을 측정하는 상호정보에 기반을 둔 방법이 제안되었다. 특히 상호정보에 기반을 둔 방법은 cross-cumulant에 기반을 둔 방법에서 반드시 요구되는 정규화 과정을 제거할 수 있는 장점도 가진다.

본 연구에서는 고정점(fixed-point : FP) 알고리즘의 독립성분분석(independent component analysis : ICA)[4-6]과 상호정보에 기반을 둔 방법을 조합한 입력변수선택 방법을 제안한다. 여기서 FP-ICA는 입력변수들 사이의 종속성을 빠르고 정확하게 제거하기 위함이고, 적응적 분할을 이용한 상호정보에 기반을 둔 방법은 입력변수의 확률밀도함수를 계산하여 변수상호간의 종속성을 효과적으로 정량화하기 위함이다. 제안된 기법을 인위적으로 제시된 각 500개의 샘플을 가지는 6개의 독립신호로부터 얻어지는 1개의 종속변수를 대상으로 실험하여 결과를 비교분석하였다.

2. 독립성분분석과 상호정보 추정

ICA는 m개의 입력신호 **s**로부터 선형적으로 혼합된 n개의 신호 **x**가 알려져 있을 때, 혼합된 신호로부터 역으로 m개의 독립인 입력신호를 찾는 기법이다[3-6]. 하지만 입력신호들을 혼합할 때의 혼합행렬 **A**는 알려져 있지 않으며, 혼합과정에서 잡음 **n**이 추가 될 수도 있다. 이때 혼합신호와 입력신호와의 관계는

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

로 정의된다. 여기서 **n**은 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 **W**에 대하여 상세히 살펴보면 다음 그림 1과 같은 구성도로 나타낼 수 있다. 여기서 **x** = **As**이고, **y** = **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 **W**을 찾아서 원 신호 **s**의 근사값을 알아내는 것이다.

최근 ICA를 위한 다양한 알고리즘들이 연구되었다[5]. 그 중에서도 고정점 알고리즘은 신경망이 가지는

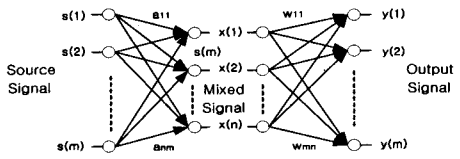

그림 1. 혼합행렬과 역혼합행렬의 상세 설명도

병렬성과 분산성, 그리고 더 작은 메모리 요구 등의 제약을 해결하기 위해 제안된 batch mode로 계산되는 ICA 기법이다[4,5]. 특히 FP 알고리즘은 엔트로피 최적화 방법으로부터 유도되며, 지금까지 알려진 기법 중 가장 빠른 학습속도를 가지며 신호 내에 포함된 상호정보를 최소화함으로써 ICA의 해를 구하는 기법이다. 신호벡터 **x**의 상관행렬 $E\{\mathbf{xx}^T\}=\mathbf{I}$로 whitening되어 있다고 가정할 때, 근사화된 반복기법의 역혼합행렬 **W**를 구하는 뉴우턴법[4]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 함수값을 가진다. 본 연구에서는 $\tanh(\cdot)$ 함수를 이용하였다.

따라서 FP-ICA는 입력변수 **x**로부터 독립인 변수 **s**를 추정하는 전처리과정으로 이용하며, **s**를 대상으로 원하는 입력변수들을 선택하기 위하여 통계적인 시험을 수행한다. 이렇게 하면 통계적 종속성 측정에 기반을 둔 빠르면서도 정확한 model-free 입력변수의 선택이 유도될 수 있다.

한편 신호들 사이의 종속성을 시험하기 위한 여러 가지 방법들이 제안되었다[1]. 그 중에서 상호정보는 신호들 사이의 종속성을 정량화하기 위한 가장 자연스러운 방법으로 입력변수 선택을 위해 사전에 이용되어진다. 그러나 랜덤변수의 표본화 데이터로부터 상호정보를 추정하는 것은 데이터의 분포를 가장 나타내는 확률밀도함수(probability density function : PDF)의

추정이 요구되어 매우 어렵다. 잘 알려진 상호정보 추정으로는 Gram-Charlier 확장에 기초한 방법, 규칙적 히스토그램 PDF 근사화에 기초한 방법, 적응적 분할 히스토그램 PDF 근사화에 기초한 방법, 커널변환에 기초한 방법이 있다. 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. 또한 일정한 분할을 가지는 규칙적인 히스토그램 PDF 근사화에 기초한 방법은 Gram-Charlier 확장에 기초한 방법보다는 신호들의 성질에 의존하지 않기 때문에 좀 더 일반화된 방법이다. 그러나 이 방법은 샘플의 분할과 질에 민감한 제약이 있다. 분할이 너무 조밀하면 샘플을 포함하지 않는 어떤 부분이 있어 PDF의 평활화에 따른 손실된 분포를 고려하지 않으며, 너무 듬성하면 샘플들이 중요한 PDF를 상세히 설명하지 못하는 제약이 있다. 이러한 분할에 따른 상호정보의 추정 성능변화를 가진 히스토그램에 기초한 방법의 제약을 해결하기 위해서 동일한 양의 분할을 얻기 방법이 제안되었다[1]. 이는 적응적으로 동일한 분할을 이용한 상호정보에 기반을 둔 방법이다. 이 방법의 수행과정을 요약하면 다음과 같다. 즉,

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.

단계 2 : 4개 부분할의 관찰 쌍에 chi-square 시험을 행한다.

단계 3: 만약 chi-square 시험값이 사전 설정값보다 크면, 단계 1과 2를 다음 부분할에 대해서 수행한다.

단계 4: 만약 chi-square 시험값이 사전 설정값보다 적거나 R_n이 너무 작으면, 분할을 멈추고 규칙적인 히스토그램 PDF 근사화에 기초한 방법과 동일한 과정을 수행한다.

이상의 적응적 분할 방법은 규칙적 히스토그램 분할 방법보다 좀 더 정확한 상호정보를 얻을 수 있다. 본 실험에서는 사전 설정값을 7.8로 하였다.

따라서 입력된 변수를 대상으로 FP-ICA를 적용함으로써 빠르게 독립된 변수를 얻을 수 있으며, 확률밀도함수의 계산을 위한 적응적 분할 방법으로 변수 상호간의 정보를 좀 더 정확하게 얻을 수 있어 효과적인 입력변수 선택이 가능하다.

3. 실험 및 결과고찰

전처리 과정으로 FP-ICA와 적응적 분할 히스토그램 PDF 근사화에 기초한 상호정보 추출 방법에 의한 제안된 입력변수선택 방법의 성능을 평가하기 위해 입력신호로 각각 500개 샘플을 가진 6개의 독립신호와 이에 따른 1개의 종속 신호를 대상으로 실험하였다. 실험은 펜티엄Ⅳ-3.0G 컴퓨터에서 Matlab 6.5로 구현하였다.

한편 6개의 독립신호는 각각 2개의 sine 및 saw-tooth 신호와 1개의 cosine 및 impulse noise 신호들이다. 이들 신호함수들은 다음 식 (4)와 같다.

$$\begin{aligned} x_1 &= \sin(nt/6) \\ x_2 &= ((\mathrm{rem}(nt,27)-13)/9) \\ x_3 &= \cos(nt/2) \\ x_4 &= ((\mathrm{rand}(1,nt)<.5)*2-1).*\log(\mathrm{rand}(1,nt)) \\ x_5 &= ((\mathrm{rem}(nt,20)-13)/9) \\ x_6 &= \sin(nt/3) \end{aligned} \quad (4)$$

상기 식 (4)에서 x_2와 x_5는 각각 saw-tooth 신호이고 x_4는 impulse noise 신호이다. 또한 nt는 1에서 500까지의 500개 샘플이다. 그림 2는 x1부터 x6까지의 신호를 위에서부터 아래로 순차적으로 각각 도시한 것이다.

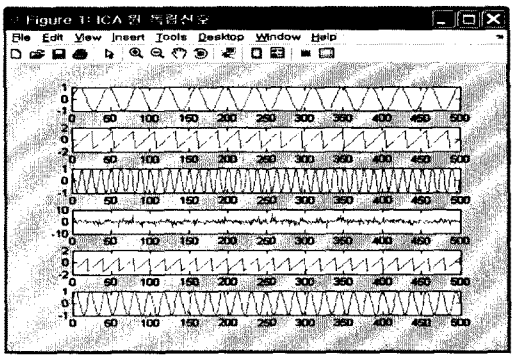

그림 2. 실험에 이용된 6개의 입력신호

그림 3은 6개의 입력신호에 FP-ICA를 적용한 상호 독립인 변수들을 도시한 것이다. 여기서 보면 신호의 추출순서와 부호가 각각 바뀐 신호들을 볼 수 있다. x_6만 제외한 모든 신호는 순서가 바뀌었으며, x_1, x_3, x_5는 부호가 바뀌었음을 알 수 있다. 이는 신호의 추출순서나 부호의 변화와 같은 ICA 고유의 속성을 보여 준 것이다.

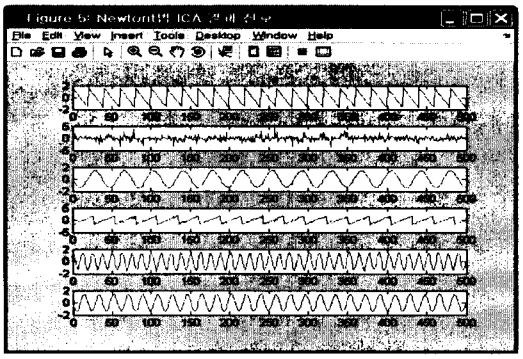

그림 3. FP-ICA에 의한 6개의 독립된 신호

그림 4는 6개의 입력신호로부터 인위적으로 생성된 종속신호로 $y = x_1^2 + 2x_3 + x_5$를 도시한 것이다.

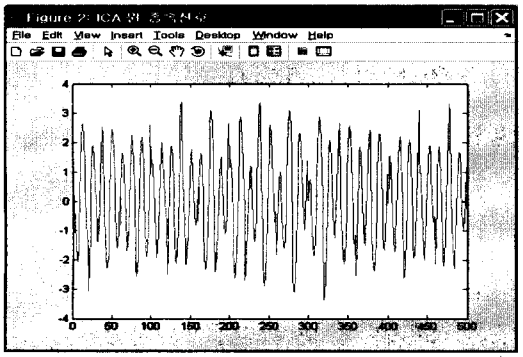

그림 4. $x_1^2 + 2x_3 + x_5$의 종속신호

한편 그림 5는 그림 3의 6개 독립인 종속신호 x를 대상으로 적응적 분할에 의한 종속신호 y와의 상호정보를 각각 구한 결과값을 도시한 것이다. 여기서 chi-square 시험을 위한 사전 설정값은 7.8로 하였다.

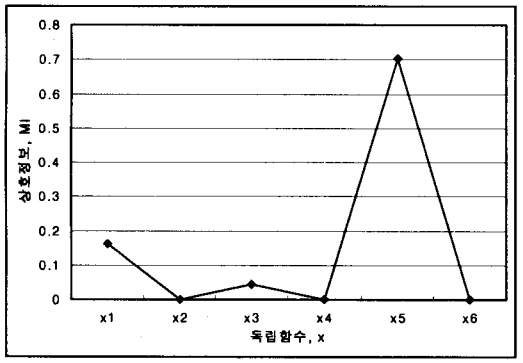

그림 5. 6개 독립신호와 1개 종속신호와의 상호정보량

그림 5에서 x_1, x_3, x_5와 y와의 상호정보량은 각각 0.162124, 0.044668, 0.703209로 높은 값을 가지나 x_2, x_4, x_5는 각각 0.001152, 0.000128, 0.000032의 낮은 값을 가짐을 알 수 있다. 이는 6개의 입력변수 중에서 x_1, x_3, x_5가 종속변수 y와 관계되는 변수임을 나타내는 것이다. 그리고 나머지 3개의 입력변수는 종속변수에 영향을 미치지 못함을 알 수 있다. 따라서 제안된 조합기법은 입력변수선택을 위한 우수한 성능의 기법임을 알 수 있다.

4. 결론

본 논문에서는 고정점 알고리즘의 독립성분분석과 상호정보 추정을 조합한 입력변수선택 기법을 제안하였다. 여기서 고정점 알고리즘의 독립성분분석은 뉴우턴법에 기반을 둔 방법으로 입력변수 간의 독립성을 빠르게 찾기 위함이고, 적응적 분할에 기반을 둔 상호정보 추출은 좀 더 정확한 정보의 추출을 위함이다.

제안된 기법을 인위적으로 제시된 각각 500개의 샘플을 가지는 6개의 독립신호와 1개의 종속신호를 대상으로 실험한 결과 빠르고 정확한 변수의 선택이 이루어짐을 확인하였다.

향후 제안된 방법을 다양한 분야에 좀 더 큰 규모의 문제에 적용하는 연구가 뒤따라야 할 것이다.

참고문헌

[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," *IJCNN99*, pp. 1-5, Washington, 1999

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

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

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