Proceedings of KFIS Autumn Conference 2006 Vol. 16, No. 2.

주요성분분석과 상호정보 추정에 의한 입력변수선택

Input Variable Selection by Principal Component Analysis and Mutual Information Estimation

조용현, 홍성준*

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

요 약

본 논문에서는 주요성분분석과 상호정보 추정을 조합한 입력변수선택 기법을 제안하였다. 여기서 주요성분분석은 2차원 통계성을 이용하여 입력변수 간의 독립성을 찾기 위함이고, 상호정보의 추정은 적응적 분할을 이용하여 입력변수의 확률밀도함수를 계산함으로써 변수상호간의 종속성을 좀 더 정확하게 측정하기 위함이다. 제안된 기법을 인위적으로 제시된 각 500개의 샘플을 가지는 6개의 독립신호와 1개의 종속신호를 대상으로 실험한 결과, 빠르고 정확한 변수의 선택이 이루어짐을 확인하였다.

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

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[1-3]. 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 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[1,4].

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 2

차원 선형종속성만을 측정하는 방법으로 선형 모델에만 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)[5,6]과 상호정보에 기반을 둔 방법을 조합한 입력변수선택 방법을 제안한다. 여기서 주요성분분석은 2차원 통계성을 이용하여 입력변수 간의 독립성을 찾기 위함이고, 상호정보의 추정은 적응적 분할을 이용하여 입력변수의 확률밀도함수를 계산함으로써 변수상호간의 종속성을 좀 더 정확하게 측정하기 위함이다. 제안된 기법을 인위적으로 생성된 각 500개의 샘플을 가지는 6개의 독립 신호로부터 얻어지는 1개의 종속변수를 대상으로 실험하여 결과를 비교 분석하였다.

2. 주요성분분석과 상호정보 추정

주요성분분석은 입력데이터의 특징을 추출하는 기법으로 데이터 내에 포함된 정보를 추출하고 압축하여 통계적 규칙들을 찾아내는 것이다[5,6]. 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 $\mathbf{R}_{xx} = \langle \mathbf{x}\mathbf{x}^T \rangle$를 가진 평균이 영인 입력벡터 $\mathbf{x} = [\, x_1, x_2, \cdots, x_n \,]^T$에 대해서 생각해 보자. 여기서 T는 전치를 나타내며, $\langle \cdot \rangle$는 기대치를 나타낸다. 또한 $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \cdots, \hat{\mathbf{w}}_m$이 $\mathbf{R}_{xx}$의 고유벡터와 직교되는 연결가중치 벡터라 할 때, $\hat{\mathbf{w}}_1 = [\, \hat{w}_{11}, \hat{w}_{12}, \cdots, \hat{w}_{1n} \,]^T$는 가장 큰 고유치 λ_1과 일치하며, $\hat{\mathbf{w}}_2 = [\, \hat{w}_{21}, \hat{w}_{22}, \cdots, \hat{w}_{2n} \,]^T$는 두 번째로 큰 고유치 λ_2, 그리고 $\hat{\mathbf{w}}_n = [\, \hat{w}_{n1}, \hat{w}_{n2}, \cdots, \hat{w}_{nn} \,]^T$는 가장 작은 고유치 λ_n과 각각 일치한다. 이상의 관계를 행렬방정식으로 나타내면 식 (1)과 같다.

$$\mathbf{R}_{xx}\hat{\mathbf{w}}_j = \lambda_j \hat{\mathbf{w}}_j, \quad (j=1,2,\ldots, n) \tag{1}$$

여기서 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ 이다. 주어진 입력벡터 $\mathbf{x}$를 위한 첫 번째 m개의 주요 특징을 나타내는 고유벡터 $\mathbf{y}$는 다음의 선형변환식 (2)로 나타낼 수 있다.

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

여기서 $\hat{\mathbf{W}} = [\, \hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \cdots, \hat{\mathbf{w}}_m \,]^T \in \mathbb{R}^{mxn}$ 이며, 이 식에서 연결가중치행렬 $\hat{\mathbf{W}}$의 행은 가장 큰 고유치와 일치하는 상관행렬 Rxx의 고유벡터임을 의미한다.

다시 말하면, 입력데이터 공간에서 k차원의 주요특징을 나타내는 부공간은 $\mathbf{R}_{xx}$의 k개 주요 고유벡터에 의해 구성된 부공간으로 정의된다. 결국 PCA는 $\langle \| \hat{\mathbf{w}}_j^T \mathbf{x} \|_2^2 \rangle$가 최대인 고유벡터 $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \cdots, \hat{\mathbf{w}}_k$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]. 이는 신호의 1차적 통계성을 고려한 정규화로 영 평균은 신호벡터 $\mathbf{x}$에서 평균값 $\mathbf{x}^*$를 뺀 차로 $\mathbf{x} = \mathbf{x} - \mathbf{x}^*$이다. 따라서 입력되는 변수를 대상으로 PCA를 수행하면 2차원 통계성이 고려된 독립변수를 추출할 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 이용되어 진다. 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 Rn이 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 PCA를 적용함으로써 독립된 변수를 얻을 수 있으며, 확률밀도함수의 계산을 위한 적응적 분할 방법으로 변수 상호간의 정보를 좀 더 정확하게 얻을 수 있어 효과적으로 입력변수의 선택이 가능하다.

3. 실험 및 결과 고찰

PCA와 적응적 분할 히스토그램 PDF 근사화에 기초한 상호정보 추출방법에 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 1개의 cosine 및 impulse noise 신호와 각각 2개의 sine 및 saw-tooth 신호들이다. 이들 신호함수들은 다음 식 (3)과 같다.

$$
\begin{aligned}
x_1 &= \cos(v/4)\\
x_2 &= ((\mathrm{rem}(v,10)-13)/4)\\
x_3 &= \sin(v/9)\\
x_4 &= ((\mathrm{rem}(v,27)-13)/9)\\
x_5 &= ((\mathrm{rand}(1,nt)<.1)*2-1).*\log(\mathrm{rand}(1,nt))\\
x_6 &= \sin(v/3)
\end{aligned}
\qquad (3)
$$

상기 식 (3)에서 x_2와 x_4는 각각 saw-tooth 신호이고 x_5는 impulse noise 신호이다. 또한 nt는 1에서 500까지의 500개 샘플이다. 그림 1은 x_1부터 x_6까지의 신호를 위에서부터 아래로 순차적으로 각각 도시한 것이다.

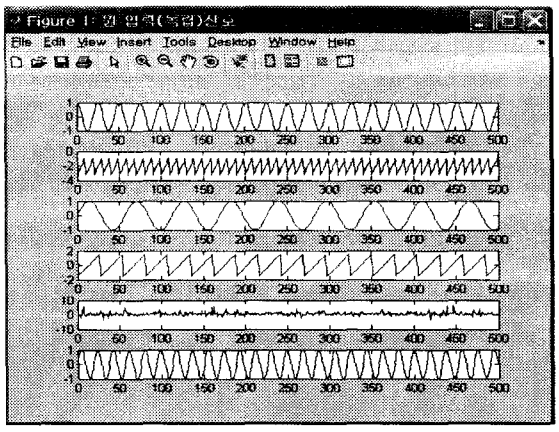

그림 1. 실험에 이용된 6개의 입력신호

그림 2는 입력신호를 대상으로 PCA에 의한 신호를 나타낸 것이다. 고려된 전처리된 신호로 이를 대상으로 상호정보를 추정한다.

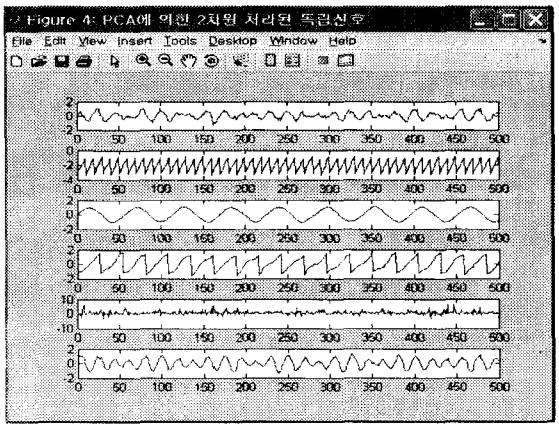

그림 2. PCA에 의한 6개의 독립된 신호

그림 3은 6개의 입력신호인 독립신호로부터 인위적으로 생성된 종속신호 $y = x_1^2 + 2x_3 + x_5$를 도시한 것이다.

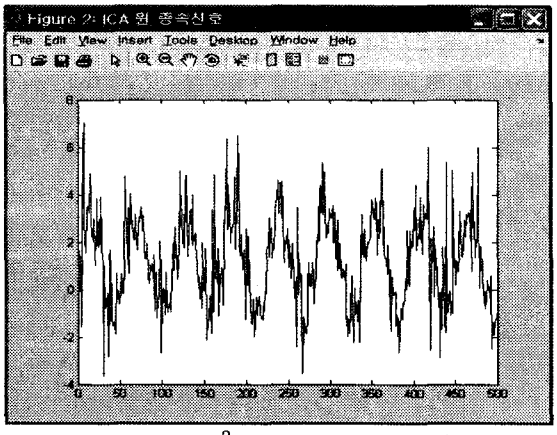

그림 3. $y = x_1^2 + 2x_3 + x_5$의 종속신호

한편 그림 4는 그림 2의 6개 독립인 종속신호 x를 대상으로 적응적 분할에 의한 종속신호 y와의 상호정보량을 각각 도시한 것이다. 여기서 chi-square 시험을 위한 사전 설정값은 7.8로 하였다.

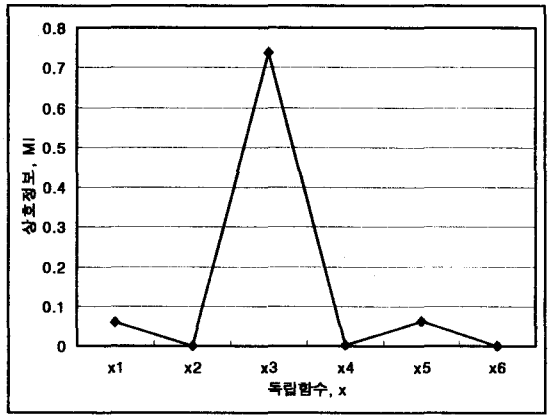

그림 4. 6개 독립신호와 1개 종속신호와의 상호정보량

그림 4에서 x_1, x_3, x_5와 y와의 상호정보량은 각각 0.058396, 0.736867, 0.062279로 큰 값을 가지나 x_2, x_4, x_5는 각각 0.000128, 0.003203, 0.000032의작은 값을 가짐을 알 수 있다. 이는 6개의 입력변수 중에서 x_1, x_3, x_5가 종속변수 y와 관계되는 변수임을 의미하는 것이다. 그리고 나머지 3개의 입력변수는 종속변수에 영향을 미치지 못함을 알 수 있다. 따라서 제안된 조합기법은 입력변수선택을 위한 우수한 성능의 기법임을 알 수 있다.

4. 결 론

논문에서는 주요성분분석과 상호정보 추정을 조합한 입력변수선택 기법을 제안하였다. 여기서 주요성분분석은 2차원 통계성을 이용하여 입력변수 간의 독립성을 찾기 위함이고, 상호정보의 추정은 적응적 분할을 이용하여 입력변수의 확률밀도함수를 계산함으로써 변수상호간의 종속성을 좀 더 정확하게 측정하기 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.

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," IJCNN99, 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