상호정보 추출에 기초한 효과적인 신호분류

An Efficient Signal Classifications by Extracting Mutual Information

조용현, 홍성준[0]
대구가톨릭대학교 컴퓨터정보통신공학부
Yong-Hyun Cho, Seung-Jun Hong[0]
School of Computer and Information Comm. Eng., Catholic Univ. of Daegu
E-mail : {yhcho,sjishong}@cu.ac.kr

요 약

본 논문에서는 신호의 상호정보 추출에 의한 효율적인 군집화 방법을 제안하였다. 여기서 상호정보는 신호상호간의 상관관계를 나타내는 정보로 군집화를 위한 유사성을 나타내는 척도이다. 특히 적응적 분할을 이용하여 신호의 확률밀도를 계산함으로써 신호 상호간의 종속성을 좀 더 정확하고 빠르게 측정하였다. 제안된 방법을 500개 샘플을 가진 6개의 인위적인 신호군집화에 적용한 결과 정확한 분류성능이 있음을 확인하였다.

1. 서론

실세계의 모델링에서 가장 적합한 입력만을 선택하는 것은 시스템 성능에 많은 영향을 미친다. 일반적으로 입력변수의 효과적인 선택은 시스템 차원의 감소나 특징추출 등 다양한 용도로 이용된다[1-5]. 또한 입력변수선택은 분류, 예측, 자료해석 등을 위한 알고리즘으로도 이용되고 있다. 그러나 많은 입력변수들 중에서 모델에 얼마나 많은 또는 어느 입력들이 필요한지 알 수 없으며, 이는 입력차원이 증가할수록 더욱 더 심각하다. 특히 입력차원의 증가에 따른 계산시간과 메모리의 증가, 다음으로 요구되지 않는 입력들에 의한 학습의 어려움, 추가적인 요구되지 않는 입력에 의한 비수렴성과 모델의 정확성 저하, 그리고 복잡성에 따른 해석의 어려움 등의 제약이 있다[2-4].

지금까지 알려진 입력변수선택 기법들은 크게 model-based와 model-free 방법들로 나누어진다[2-4]. 먼저 model-based 방법에 의한 입력선택은 모델을 선정한 후 이용할 입력들을 선택하고, 파라미터들을 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[2-4].

model-free 방법은 통계적 종속성에 기반을 둠으로써 model-based 방법보다 좀 더 일반화된 방법이다[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 선형종속성

만을 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 상호정보에 추출에 기반을 둔 입력신호의 효과적인 분류방법을 제안한다. 여기서 상호정보의 추정은 적응적 분할을 이용하여 입력신호의 확률밀도함수를 계산함으로써 신호상호간의 종속성을 좀 더 정확하게 측정하기 위함이다. 제안된 기법을 각 500개의 샘플을 가지는 6개의 신호를 가지는 인위적인 문제를 대상으로 실험하여 타당성을 확인하였다.

2. 적응분할 방식에 의한 상호정보 추출

신호들 사이의 종속성을 시험하기 위해 correlation, 고차원의 cross-cumulant, 그리고 상호정보 등에 기반을 둔 여러 가지 방법들이 제안되었다[1-4]. 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 (1)과 같이 정의된다[2].

$$I(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\cdot\log\left(\frac{f(x,y)}{f(x)f(y)}\right)dxdy \quad (1)$$

여기서 x와 y가 서로 독립이면 상호정보 I(x,y)는 영이 된다. 또 다른 상호정보는 엔트로피(entropy)를 이용하여 다음 식 (2)와 같이 정의될 수 있다.

$$I(x,y) = H(x) + H(y) - H(x,y) \quad (2)$$

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

식 (1)과 식 (2)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[2,3]. 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[5]. 이는 현재 변수의 분포가 균일한지를 시험하기 위해서 공간을 chi-square χ^2에 기초하여 분할하는 반복 기법이다. 이 방법의 수행과

정을 요약하면 다음과 같다.

단계 1 : 주어진 x와 y의 2차원 범위 Rn이 주어지면 2x2 grid로 나눈다. Rn내의 전체 관찰 수는 cRn이고, 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.$(\chi^2 = \frac{4\sum_{i=1}^{2}\sum_{j=1}^{2}(cR_{n+1}^{ij} - cR_n/4)^2}{cR_n})$

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. 따라서 적응분할 히스토그램 PDF 근사화 방법을 이용한 상호정보 추출은 좀 더 정확한 신호분류를 가능하게 한다.

3. 실험 및 결과분석

적응적 분할 히스토그램 PDF 근사화에 기초한 상호정보 추출방법에 의한 제안된 신호분류 방법의 성능을 평가하기 위해 각각 500개 샘플을 가진 6개의 신호를 대상으로 실험하였다. 실험은 펜티엄Ⅳ-3.0G 컴퓨터에서 Matlab 6.5로 구현하였다.

인위적인 문제에서 6개의 신호는 1개의 cosine 및 impulse noise 신호와 각각 2개의 sine 및 saw-tooth 신호들이다. 이들 신호함수들은 다음 식 (3)과 같다.

```
x1 = sin(v/6)
x2 =((rem(v,30)-15)/4)
x3 =cos(v)
x4 =((rand(1,nt)<0.5)*2-1).*log(rand(1,nt))
x5 = ((rem(v,20)-10)/5)
x6 = sin(v/3)                                    (3)
```

위식 (3)에서 x1과 x6는 sine 신호, x2와 x5는 각각 saw-tooth 신호, x3는 cosine 신호, 그리고 x4는 impulse noise 신호이다. 또한 nt는 1에서 500까지의 500개 샘플이다. 그림 1은 x1부터 x6까지의 신호를 위에서부터 아래로 순차적으로 각각 도시한 것이다.

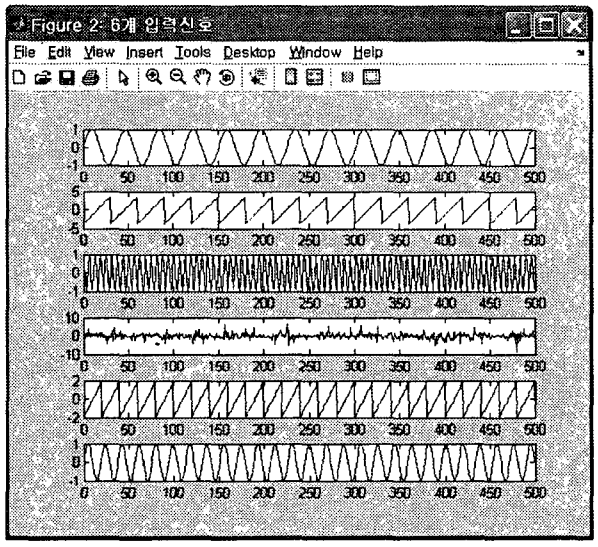

그림 1. 실험에 이용된 6개의 입력신호

그림 2는 입력신호를 대상으로 제안된 기법에 의해 실험한 결과를 나타낸 것이다. 여기서 상호정보의 계산은 6개 신호 각각의 조합에 따른 만

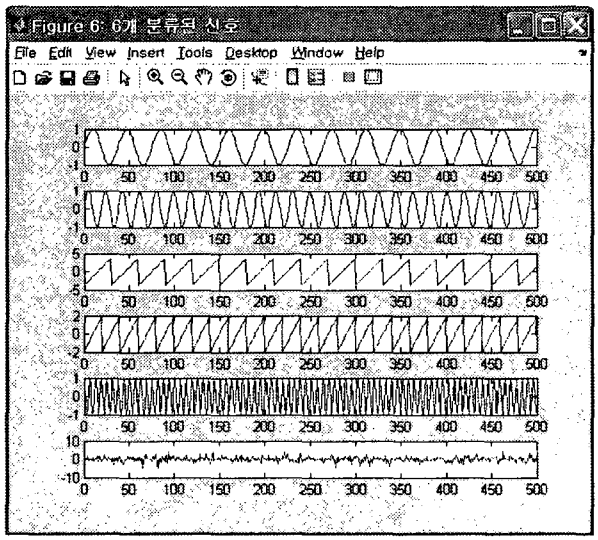

그림 2. 전처리된 6개의 독립신호

큼의 회수로 총 36개의 상호정보 값을 얻을 수

있다. 따라서 그림 2는 계산된 상호정보 값에 따라 분류된 신호를 차례로 도시한 것으로 x3와 x6, x2와 x4, x1, 그리고 x5의 순으로 군집됨을 알 수 있다.

한편 <표 1>은 제안된 방법에 의해서 추출된 각 신호 간의 상호정보 값을 나타낸 것이다. 여기서 신호 자신과의 재귀상호정보 값은 모두 4.734298로 높은 값을 가진다. 또한 신호 x1의 경우는 자기 자신을 제외하고 상대적으로 신호 x6와 가장 높은 상호정보 값을 가져 서로 유사한 신호임을 알 수 있다. 신호 x2의 경우도 자기 자신 외에 x5와, 신호 x3와 x4는 각각 자기 자신을 제외한 다른 신호들과의 상호정보량이 높지 않아 관련이 없음을 알 수 있다. 신호 x5와 x6는 각각 x2와 x1과의 상호정보 값이 상대적으로 큰 값이어서 밀접한 관계가 있음을 알 수 있다. 특히 표 1의 상호정보량 행렬은 대칭행렬이며, 신호 x2와 x3의 경우는 상호정보 값으로 보면 전혀 무관한 신호임을 추측할 수 있다. 따라서 상호정보량 추출에 의한 제안된 방법을 이용한 신호의 군집화는 간단한 계산으로 빠르고 정확하게 신호를 분류할 수 있음을 확인할 수 있다.

<표 1> 6개의 신호 간의 상호정보 값

신호	x1	x2	x3	x4	x5	x6
x1	4.734298	0.000128	0.000288	0.000512	0.000032	0.818232
x2	0.000128	4.734298	0.000000	0.000288	1.682299	0.000288
x3	0.000288	0.000000	4.734298	0.003877	0.000512	0.000032
x4	0.000512	0.000288	0.003877	4.734298	0.000032	0.000128
x5	0.000032	1.682299	0.000512	0.000032	4.734298	0.000288
x6	0.818232	0.000288	0.000032	0.000128	0.000288	4.734298

4. 결론

본 논문에서는 상호정보에 추출에 기반을 둔 입력신호의 효과적인 분류방법을 제안하였다. 여기서 상호정보의 추정은 적응적 분할을 이용하여 입력신호의 확률밀도함수를 계산함으로써 신호상호간의 종속성을 좀 더 정확하게 추정하기 위함이다.

제안된 방법을 각 500개의 샘플을 가지는 6개의 신호를 가지는 인위적인 문제를 대상으로 실험한 결과, 빠르고 정확한 신호분류의 성능이 있음을 확인하였다.

향후 제안된 방법을 좀 더 큰 규모의 분류문제와 다양한 분야에 적용하는 연구가 지속적으로 이루어져야 할 것이다.

5. 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. 2002

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

[3] 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

[4] I. Guyon and A. Elisseeff,"An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, Vol. 3, pp. 1157-1182, March. 2003

[5] G. A. Darbellay and I. Vajda,"Estimation of the Information by an Adaptive Partitioning of the Observation Space", *IEEE Transactions on Information Theory*, Vol.45, No. 4, pp. 1315-1321, May. 1999