Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 4, pp. 717~724

A Comparison on Independent Component Analysis and Principal Component Analysis -for Classification Analysis-

Daehak Kim[1] · Kilak Lee[2]

Abstract

We often extract a new feature from the original features for the purpose of reducing the dimensions of feature space and better classification. In this paper, we show feature extraction method based on independent component analysis can be used for classification. Entropy and mutual information are used for the selection of ordered features. Performance of classification based on independent component analysis is compared with principal component analysis for three real data sets.

Keywords : 독립성분분석, 분류분석, 엔트로피, 주성분분석, 특징추출

1.서 론

패턴인식(pattern recognition)은 주어진 다양한 형태의 패턴(음성, 문자, 도형, 화상)에 대하여 그것이 무엇인지 혹은 어떤 그룹(class)에 속하는지를 컴퓨터를 이용하여 자동적으로 인식하는 일련의 과정을 말한다. 일반적 패턴인식 시스템은 확인되어지거나 찾아야 할 이미지(image), 혹은 음성신호(sound signals)등을 입력(input)으로 사용하고 특징추출(feature extraction, or making measurement), 전처리(preprocessing), 세그멘테이션(segmentation) 등과 같은 많은 단계를 거쳐 입력을 적절히 분류(classification)해 낸다. 이러한 패턴인식의 여러 과정 중에서 특징추출도 아주 중요한 과정중의 하나이다. 무엇을 입력의 특징으로 선택하느냐에 따라서 그 결과도 상당히 영향을 받을 수 있기 때문이다. 특징추출의 일반적 방법은 각 그룹 내에서의 패턴변동성을 최소화 하면서 각 그룹간의 패턴변동성을 최대화 하는 정보를 추출하는 것이다(Gose등(1996)). 이러한 특징추출에 있어서 주로 사용되는 통계적 방법은 주성분분석과 독립성분분석이다.

---

1) 제 1 저자 : 교수, 경북 경산시 하양읍 금락리 330 대구가톨릭대학교 정보통계학과
E-mail : dhkim@cu.ac.kr
2) 대구가톨릭대학교 정보통계학과 석사졸업

주성분분석(Principal Component Analysis)은 차원의 축소 및 자료 요약에 있어서 매우 중요한 절차에 해당한다. 관련 변수들의 차원이 클수록 데이터를 한눈에 파악하기가 어렵기 때문에 통계적 분석을 수행하기에 앞서서 우선 차원을 축소하거나 자료를 요약할 필요가 있다. 주성분분석은 고차원을 축소하여 자료를 간단한 형태로 요약하는데 이용된다.

한편 독립성분분석(Independent Component Analysis)은 선형적으로 혼합된 입력들로부터 통계적으로 서로 독립적인 입력들로 분리해 내는 분석기법이다.

본 논문에서는 주성분분석 방법과 독립성분분석 방법으로 추출된 특징들을 이용하여 분류분석한 결과를 비교, 연구 하였다. 독립성분분석의 수행은 먼저 주어진 자료에 대하여 특징을 추출하고 엔트로피(entropy)를 이용한 상호정보이론을 바탕으로 추출된 특징들의 중요성을 구하였다. 즉 추출된 특징들을 상호정보이론에 입각하여 중요도에 의해 순서를 정하였고 순서화된 특징들을 분류분석에 사용하였다. 이들 결과들은 주성분을 이용하여 분류분석한 결과와 비교되었다.

2. 독립성분분석과 상호정보

2.1 독립성분분석

d개의 독립변수들을 $s_1, s_2, \cdots, s_d$라 하고 이들의 선형결합으로 만들어진 d개의 변수들을 $x_1, x_2, \cdots, x_d$라 하자. 이들의 관계는 다음 식 (2.1)과 같이 표현할 수 있다

$$x_i = \sum_{j=1}^{d} a_{ij} s_j \tag{2.1}$$

여기서 a_{ij}는 선형결합에 이용된 계수이다. 이제 편의상 x_i와 s_j들을 각각 벡터기호 $\boldsymbol{x}' = (x_1, x_2, \cdots, x_d)$와 $\boldsymbol{s}' = (s_1, s_2, \cdots, s_d)$로 나타내자. 선형결합 계수 a_{ij}로 구성된 정방행렬을 A 라 두면 식(2.1)은

$$\boldsymbol{x} = \mathrm{A}\,\boldsymbol{s} \tag{2.2}$$

와 같이 나타낼 수 있다.

이때 여기서 독립성분 $\boldsymbol{s}$가 주어지지 않고 행렬 A에 대해서도 아무런 정보도 없는 상태에서 관측된 변수 $\boldsymbol{x}$만을 가지고 독립성분 $\boldsymbol{s}$와 혼합행렬(mixing matrix) A를 추정하는 것이 독립성분분석의 주된 내용이다. 이러한 독립성분분석은 BSS(Blind Source Separation)라고도 알려져 있다. 최근 많은 연구가 이루어지고 있는 독립성분분석 분야는 선두 주자 격인 Hyvärinen 과 Oja(1997)을 비롯하여 Jutten과 Herault(1991), Hyvärinen등(2001)에 의하여 다양하고도 많은 연구결과가 나타나고 있고 실생활에서의 많은 응용이 이루어지고 있는 실정이다. 식 (2.2)는 다음과 같이 분리행렬(separating matrix) W를 이용하여 다시 표현될 수 있다.

$$\boldsymbol{s} = \mathrm{A}^{-1}\boldsymbol{x} = \mathrm{W}\boldsymbol{x} \tag{2.3}$$

여기서 분리행렬 W는 혼합행렬 A의 역행렬(inverse matrix)이다. 독립성분 $\boldsymbol{s}$는 주어진 관찰값 $\boldsymbol{x}$와 추정된 분리행렬 W를 이용하여 추정할 수 있다. 독립성분의 추정에 필요한 일반적인 가정은 다음과 같다.

1) 독립성분들은 통계적으로 서로 독립 관계에 있다.
2) 독립성분들의 분포는 비정규 분포(non-gaussian)이다
3) $\boldsymbol{s}$와 $\boldsymbol{x}$의 차원은 같다.
4) A (혹은 W)는 정칙(non-singular)행렬이다.
5) 독립성분의 크기는 유일하게 정의되지 않는다.

독립성분들이 서로 독립이라는 것은 독립성분분석의 기본적인 전제이며 이 때 독립성분들 사이의 관계는 통계적 독립의 정의에 의하여 확률밀도함수를 이용하여 다음과 같이 나타낼 수 있다.

$$p(s_1, s_2, \cdots, s_d) = p_1(s_1)p_2(s_2)\cdots p_d(s_d) \tag{2.4}$$

여기서 $p(s_1, s_2, \cdots, s_d)$는 독립성분 $s_1, s_2, \cdots, s_d$의 결합 확률밀도함수이고 $p_i(s_i)$는 s_i의 주변 확률밀도함수 이다.

독립성분의 분포가 비정규분포여야 되는 두 번째 가정은 독립성분의 식별을 위해 필요하다. 두 개 이상의 독립성분들이 정규분포(혹은 가우시안 분포)를 따를 경우 그 독립성분들에 직교행렬을 곱한 결과가 또 다시 서로 독립적인 가우시안 분포가 되기 때문이다. 따라서 이 경우 독립인 관계를 만족하지만 원래의 독립성분과는 다른 추정치가 존재하는 것이다.

$\boldsymbol{s}$와 $\boldsymbol{x}$의 차원이 같다는 것은 A가 정방행렬임을 의미하며, $\boldsymbol{x}$의 차원이 $\boldsymbol{s}$의 차원보다 적거나 A가 비정칙이면 $\boldsymbol{s}$에 관한 정보부족으로 독립성분을 추정 할 수 없게 된다. 이것은 식 (2.3)에서 s_i와 x_j의 관계를 풀어 쓴 다음 식에서 알 수 있다.

$$s_i = \sum_{j=1}^{d} w_{ij}\, x_j \tag{2.5}$$

여기서 w_{ij}는 W의 i행, j열 원소이다. s_i를 추정하기 위해서는 d개의 관찰변수 x_j가 모두 필요하므로 이중 하나라도 부족하게 되면 s_i를 추정할 수 없게 된다.

마지막 가정은 독립성분의 원래 크기는 추정할 수 없음을 의미한다. 예를 들어 식 (2.3) 에서 s_i에 상수 c를 곱하고 A의 i번째 열에 해당하는 원소들을 모두 c로 나누면 $\boldsymbol{x}$가 그대로 나옴을 알 수 있다. 즉 독립성분의 크기 변화는 이와 같이 상쇄되어 원래의 소스 크기를 추정할 수 없게 되기 때문이다.

2.2 상호정보

주어진 자료에 대하여 독립성분분석 방법을 이용하여 특징을 추출하였다고 가정하여 보자. 주성분분석의 결과는 어떤 주성분이 얼마만큼 더 중요한 역할을 하는지 쉽게 알 수 있지만 독립성분분석의 결과는 추출된 특징들 중에서 어떤 것이 더 중요한 특징인지 판단할 수가 없다. 이런 경우에 Torkkola(2003)는 추출된 특징들에 대하여 엔트로피를 이용한 상호정보(mutual Information)이론을 바탕으로 추출된 특징들의 중요도에 의해 순서를 부과한 바 있다. 이는 s_i성분들의 상호정보가 최소가 되도록 행렬 A 를 결정한다. 상호정보는 확률변수의 구성에서 다른 확률변수에 대하여 가지고 있는 정보를 의미한다. 상호정보를 I라 두면 확률변수 x_i들 사이에 대한 상호정보는 다음과 같이 정의된다.

$$I(x_1, x_2, \cdots, x_n) = \sum_{i=1}^{n} H(x_i) - H(\boldsymbol{x}) \tag{2.6}$$

이때 $H(\cdot)$는 엔트로피 함수로서 $H(x_i) = -\int p_i \log p_i(t)\, dt$이다.

3. 특징추출과 자료분석

독립성분분석과 주성분분석 방법의 특징추출의 성능을 비교하기 위하여 분류분석에서 많이 이용되고 있는 붓꽃자료, 유방암자료 그리고 피마인디안 당뇨병 자료 등의 세 가지 실제자료를 고려하였다. 특징추출에서 주성분분석을 통해서 추출된 특징(주성분)들은 자동적으로 중요도가 높은 순서대로 뽑혀져 정렬된다. 그러나 독립성분분석을 통해서 추출된 특징(독립성분)들은 그러한 정보가 없기 때문에 상호정보를 통한 순서화를 고려하였다. 비교를 위하여 얻어진 주성분들은 순서대로 $pc_1, pc_2, \cdots$로 표시하였고 독립성분분석으로부터 얻어진 독립성분들은 $ic_1, ic_2, \cdots$로 또 상호정보이론에 의해 중요도가 높은 순서대로 재배열된 독립성분들은 $ic_1^*, ic_2^*, \cdots$ 로 표현하였다.

Mardia 등(1979)의 붓꽃자료(Iris data)는 일반적으로 Setosa, Vesicolour와 Virginica 등의 세 그룹으로 구분된다. 하지만 이들 자료는 너무 명확하게 잘 구분되기 때문에 Setosa를 그룹에서 제외하고 나머지 두 그룹만을 가지고 특징추출을 고려하였다. 붓꽃 자료에서 꽃받침 조각의 넓이와 길이, 꽃잎의 넓이와 길이 (ic_1, ic_2, ic_3, ic_4)등의 네 개 변수에 대하여 각각의 그룹으로부터 훈련자료 50개와 시험자료 50개를 구성하였다. 위스콘신 유방암 자료의 경우에는 9개$(ic_1, ic_2, \cdots, ic_9)$의 변수에 대하여 훈련자료 300개와 시험자료 383개를 구성하였고 피마 인디안 당뇨병 자료는 8개의 변수$(ic_1, ic_2, \cdots, ic_8)$에 대하여 훈련자료 300개와 시험자료 368개를 구성하여 실험하였다.

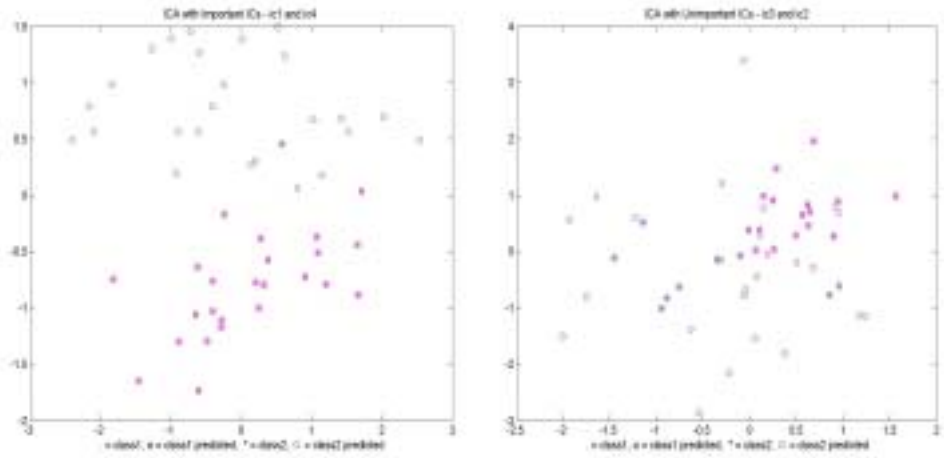
[표 3.1]은 붓꽃자료에 대한 분석 결과를 나타내고 있다. 독립성분분석을 이용하여 추출된 특징 $ic_1, ic_2, \cdots, ic_4$에 대하여 한 개, 두 개, 세 개 그리고 모두 다 사용하여 분류한 결과가 나타나 있다. 한 개씩만을 사용한 경우에는 분류율이 모두다 50%로 나타났다. 두 개의 특징을 사용한 경우 분류율은 약 80% 정도 되었고 세 개의 특징을 사용한 경우 약 85% 정도로 상향되었다. 마지막으로 특징들을 모두 분류에 사용하였을 때는 90%의 분류율을 보였다.

한편 이 결과들은 추출된 특징들에 대한 중요도의 정보가 없는 경우이다. [표 3.2]는 붓꽃 자료에 대하여 독립성분 분석을 이용하여 얻어진 특징들을 엔트로피를 이용한 상호정보이론을 바탕으로 중요도가 높은 순서대로 특징들을 추출하여 순서화된 특징들을 이용하여 분류한 결과이다. 여기서 분류율은 주성분분석을 이용한 결과와 비교하였다. 한 개의 특징이 사용된 경우에는 주성분분석이 86%, 독립성분분석(ic_1^*)은 96%의 분류율을 보였다.

[표 3.1] 붓꽃자료의 분류분석 결과(독립성분분석, 단위 %))

이용된 특징 수							
1		2		3		4	
이용된 특징	분류율	이용된 특징	분류율	이용된 특징	분류율	이용된 특징	분류율
ic_1	50	ic_1, ic_2	62	ic_1, ic_2, ic_3	82	$ic_1, \cdots, ic_4$	90
		ic_1, ic_3	84				
ic_2	50	ic_1, ic_4	82	ic_1, ic_2, ic_4	88		
ic_3	50	ic_2, ic_3	82				
		ic_2, ic_4	86	ic_1, ic_3, ic_4	86		
ic_4	50	ic_3, ic_4	80				

[표 3.2]로부터 한 개의 특징을 이용하여 분류할 경우 주성분분석에서 얻어진 첫 주성분 보다는 독립성분분석에서 상호정보를 이용하여 얻어진 성분이 더 잘 분류시킴을 확인할 수 있다. 두 개의 특징을 사용할 경우에는 두 방법 모두 비슷한 결과를 보이고 있다.



[그림 3.1] 붓꽃자료의 독립성분 산점도

(왼쪽 : 가로=ic_1^*,세로=ic_2^*, 오른쪽 : 가로=ic_3^*, 세로=ic_4^*)

또한 상호정보 이론을 적용하여 중요도 순으로 순서화된 독립성분들을 사용할 경우 일반적인 독립성분분석을 이용하는 경우보다 더 잘 분류하는 것을 확인할 수 있다.

[그림 3.1]의 왼쪽 그림은 붓꽃 자료에 대하여 독립성분분석으로부터 얻어진 성분 중 중요도가 높은 순서인 ic_1^*와 ic_2^*를 이용하여 산점도로 나타내었고 오른쪽 그림은 중요도가 낮은 순서인 ic_3^*와 ic_4^*의 성분의 산점도이다. 왼쪽 그림으로부터 Versicolour와 Virginica에서 단 두개만이 잘못 분류되고 오른쪽 그림으로부터 Versicolour와 Virginica에서 20개가 잘못 분류됨을 알 수 있다.

[표 3.2] 붓꽃자료의 분류결과(상호정보 이용)

이용된 특징 수	분류율(%) / 오분류 개수	
	주성분 분석	독립성분 분석
1	0.86 / 7	0.96 / 2
2	0.98 / 1	0.96 / 2
3	0.98 / 1	0.96 / 2
4	0.98 / 1	0.94 / 3

위스콘신 유방암 자료의 경우도 붓꽃 자료와 같은 과정을 거쳐 분석하였다. 상호정보이론에 입각한 중요도가 높은 순서대로 특징들을 추출하고 이 특징들을 분류분석에 적용한 결과가 [표 3.3]에 나타나 있다. 중요도가 가장 높은 한 개의 특징을 사용했을 때는 주성분 분석이 독립성분분석보다는 조금 더 잘 분류된 것을 볼 수 있고 6개의

특징을 사용한 경우 독립성분분석이 주성분 분석보다는 분류가 잘되는 것을 확인 할 수 있다.

[표 3.3] 위스콘신 유방암 자료의 분류결과(상호정보이용)

사용된 특징 수	분류율(%) / 오분류 개수	
	주성분 분석	독립성분 분석
1	0.973 / 8	0.943 / 17
2	0.976 / 7	0.956 / 13
3	0.966 / 10	0.963 / 11
6	0.966 / 10	0.973 / 8
9	0.966 / 10	0.966 / 10

마지막으로 피마 인디안 당뇨병자료에 대한 분류분석의 결과가 [표 3.4]에 나타나 있다. 중요도가 가장 높은 한 개의 특징을 사용했을 때는 주성분분석이 50%, 독립성분분석(ic_1^*)은 61.3%의 분류율을 보임으로서 중요도가 가장 높은 한 개의 특징을 사용했을 때는 독립성분분석이 주성분 분석보다는 더 잘 분류된 것을 볼 수 있고 중요도가 높은 순서대로 추출된 모든 특징들을 다 사용한 경우에는 두 방법 모두 약 70% 정도의 분류율을 보이고 있음을 알 수 있다.

[표 3.4] 피마 인디언의 당뇨병 자료의 분류분석 결과비교(상호정보이용)

이용된 특징 수	분류율(%) / 오분류 개수	
	주성분 분석	독립성분 분석
1	0.5 /150	0.613 / 116
2	0.656 /103	0.63 / 111
3	0.653 /104	0.596 / 121
5	0.703 / 89	0.686 / 94
8	0.7 / 90	0.7 / 90

4. 결론 및 토의

본 논문에서는 독립성분분석과 주성분분석을 이용한 분류분석에 대하여 비교, 연구하였다. 특히 독립성분분석을 이용하는 경우 추출된 독립성분들을 엔트로피의 함수인 상호정보를 이용한 중요도에 의하여 순서화를 적용하여 보았다. 분류분석에 자주 이용되는 대표적인 세 가지의 자료에 대한 분류분석의 결과 독립성분분석 방법으로 추출된 특징들을 바로 이용하기보다는 상호정보를 이용하여 추출된 특징을 순서화하여

이용하는 방법이 더 낳은 분류결과를 보임을 확인할 수 있었다. 주성분분석과 독립성분분석을 비교하여 본 결과 분류율은 한개(ic_1^*)의 특징을 사용했을 때의 결과가 첫 주성분(pc_1)을 사용한 경우보다 잘 분류됨을 알 수 있었고 두개 이상일 때는 거의 근소한 결과를 나타냄을 알 수 있었다. 통계적 분류분석방법이 적절하게 적용될 수 없는 경우, 예로서 거대한 영상 자료(huge image data)의 경우나 통계적 가정이 적절하지 않은 경우, 본 논문의 자료분석의 결과를 토대로 영상처리(image recognition)에 있어서의 특징추출이나 얼굴인식(face recognition)등과 같은 영역에서 분류의 한 가지 대응방안으로서 그 활용이 가능할 것으로 기대된다.

참고문헌

1. Gose, E., Johnsonbaugh, R. and Jost, S. (1996), *Pattern Recognition and Image Analysis*, New Jersey : Prentice Hall.
2. Hyvärinen, A. and Oja, E. (1997), A Fast Fixed-Point Algorithm for Independent Component Analysis, *Neural Computation, vol. 9*, 7 1483 - 1492
3. Hyvärinen, A., Karhunen, J. and Oja, E. (2001), *Independent Component Analysis*, Wiley-Interscience
4. Jutten, C. and Herault, J. (1991), Blind Separation of Sources, Part I : An Adaptive Algorithm based on Neuromimetic Architecture, *Signal processing, vol. 24*, 1 - 10
5. Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979), *Multivariate analysis*, Academic Press, New York.
6. Torkkola, K. (2003), Feature Extraction by Non-Parametric Mutual Information Maximization , *Journal of Machine Learning Research, 3*, 1415-1438.

[2005년 5월 접수, 2005년 7월 채택]