Proceedings of KFIS Autumn Conference 2007 Vol. 17, No. 2.

클래스 구분력이 없는 특징 소거법
Removing non-informative features weakening of class separability

이재성, 김대원

서울시 동작구 중앙대학교 컴퓨터공학과

E-mail : curseor@hotmail.com

요 약

본 논문에서는 불균형 및 Under-sampling된 바이오 데이터에 대하여 클래스 구분력이 없는 특징의 소거를 통해 이후 이어질 FLDA 등 다양한 방법론을 적용할 수 있는 방법을 제안하고자 한다. 제안하는 알고리즘은 평균과 분산을 통해 클래스의 형태를 결정하는 기존 방법론의 문제점을 회피할 수 있는 방법을 제공하며, 클래스 구분력에 중점을 두어 특징을 선별하였을 경우 선별된 특징들의 상관 계수가 높은 문제를 극복할 수 있도록 한다. 이에 따라 알고리즘이 선택한 특징 집합은 서로의 특징에 대해 상관계수가 낮으며, 클래스의 구분력이 높은 특징을 갖게 된다.

Key Words : Hierarchical clustering algorithm, Genetic algorithm, Feature selection, Correlation coefficient

1. 서 론

전통적인 특징 선별 방법들은 일반적으로 데이터에 포함된 클래스의 형태를 평균과 분산을 사용하여 결정한다[2,5]. 이러한 방법론들은 데이터를 이루는 특징들이 서로에 대해 독립임을 가정하고 있으며, 클래스의 형태를 가우시안 형태로 정의를 하고 있고, 각 클래스의 부피가 동일함을 가정하고 있다. 그러나 일반적으로 Microarray를 통하여 얻어내는 대부분의 바이오 데이터들은 샘플의 개수가 적고, 특징으로 표현되는 유전자의 개수가 많다[1]. 또한 유전자 사이에 연관관계가 다수 존재하고, 노이즈가 포함되어 있어 평균과 분산을 사용하는 기존의 방법론들을 적용하기에 많은 문제점이 따르고 있다. 이는 환자 개개인의 유전자 정보를 나타내는 샘플의 개수가 적어 일부 데이터가 과반응(Over-expression)할 경우 발생하는 소수의 데이터가 클래스 전체의 평균과 분산을 좌우하게 되며, 클래스의 분포를 편중되게 계산함에 따라 데이터를 이루는 클래스의 중첩도를 평가할 때 악영향을 끼치게 된다.

본 논문에서는 1)Under-sample된 데이터에서 평균과 분산을 사용하여 소수의 유전자가 클래스의 형태를 좌우하는 문제를 극복하는 방법을 제안하고, 2)특징으로 표현된 유전자의 독립 문제를 다루며, 3)클래스 구분력을 높이기 위하여 선택한 특징들이 서로의 특징에 대해 상관 계수가 높은 문제를 해결하고자 한다.

전체 알고리즘은 다음과 같은 순서로 이루어져 있다.

(1) 주어진 데이터에서 노이즈가 제거된 축약 데이터를 얻는다. 이는 각 특징을 이루고 있는 값들의 비유사성(Dissimilarity)를 계산하여 비유사성이 높은 값들을 제거하여, 이후 클래스 형태 결정에 필요한 평균과 분산 계산에 사용할 수 있도록 한다.(Section 2)
(2) 특징들이 독립이라는 가정을 사용하지 않으므로, 특징 선별 방법이 특징들의 조합 방법으로 변하게 되 는데 유전 알고리즘(Genetic algorithm)을 통해 임의의 특징 부집합(Feature subset)을 제거한다.(Section 3)
(3) 특징 부집합이 제거된 데이터에 대하여 적합함수(Fitness function)를 통해 평가를 한다.(Section 4)

2. HCA를 사용한 노이즈 제거

[표1]은 Golub의 논문[1]에서 사용된 데이터에서 임의의 특징을 선택한 것이며, 소수의 값들이 클래스의 형태를 결정함에 있어서 악영향을 끼치고 있는데 이는 비유사성 계산과 HCA(Hierarchical Clustering Algorithm, 이하 HCA)를 통해 제거하는 것이 가능하다[4].

표1. Leukemia data

Type	$Gene_1$	···	$Gene_i$	···	$Gene_n$	Class Label
:	:	:	:	:	:	:
Normal	-161		-192		16	1
Normal	-48		187		-73	1
Normal	-176	···	**13868**	···	-60	1
Patient	-214		-126		-37	-1
Patient	-139		2267		-14	-1
:	:	:	:	:	:	:

[표1]에서 알 수 있듯이, 특정 값이 클래스의 형태를 결정하고 있다. 이 값을 제거하기 전과 제거한 후의 클래스 평균과 분산은 [표2]와 같다.

표2. 노이즈 제거 전과 후 비교

		평균	분산
제거 전	Normal	**4612.0**	**64166167.0**
	Abnormal	1070.5	2863224.5
제거 후	Normal	**-2.5**	**71820.5**
	Abnormal	1070.5	2863224.5

샘플의 개수가 많은 기존의 방법론에서는 주어진 데이터에 포함된 노이즈의 개수가 적어 평균과 분산을 구

하는 과정에 영향을 크게 주지 못하지만, 샘플의 개수가 적은 바이오 데이터의 경우 소수의 노이즈 데이터가 평균과 분산을 결정하는데 큰 영향을 끼침으로써 클래스 형태 결정에 악영향을 주게 된다. 또한 노이즈 데이터의 제거를 위해 샘플 자체를 제거하는 것은 Under-sample된 상황에서 샘플의 개수를 더욱 줄이므로 적합한 접근 방법이라 할 수 없다. 따라서 본 논문에서는 다음과 같은 접근 방법을 제안한다.

주어진 Dataset을 이루고 있는 임의의 n번째 클래스 w_n이 있을 때, w_n을 k개의 특징 f_k의 집합이라 정의하고, f_k를 이루고 있는 값들을 v_j이라 정의하면 임의의 k번째 특징 f_k는 v_j의 합집합이라 할 수 있으며, 이는 수식 (1)과 같다.

$$f_k = \bigcup_{1}^{j} v_j \tag{1}$$

이 때, f_k를 이루는 임의의 m번째 값 v_m이 노이즈와 노이즈가 아닌 것의 집합으로 이루어져 있다고 가정하면 v_m의 비유사성 측정은 수식 (2)와 같다.

$$Dissimilarity(v_l) = \sum_{1}^{j} (v_m - v_j)^2 \tag{2}$$

이렇게 하여 얻어진 v_m의 비유사성 집합을 DS_k라고 하고, HCA를 통해 노이즈인 v_m을 정의하는 것이 가능하다. 이와 같은 방법으로 노이즈를 제거할 경우, 노이즈가 없는 특징 집합에 대해 노이즈를 과도하게 결정하는 경우가 있으므로, 데이터에 따라 임의의 임계값을 결정하여 병합하는 과정이 필요하다. 이렇게 하여 클래스를 이루는 특징 집합에서 비유사성이 높은 값이 제거된 클래스를 DW_n이라고 정의한다.

표3. 음영으로 표시된 부분이 DW_n이며, 이후 평균과 분산 계산에 있어서 DW_n이 사용됨

Type	f_1	…	f_i	…	f_n	Class Label
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Normal	-161		-192		16	1
Normal	-48		187		-73	1
Normal	-176	…	**13868**	…	-60	1
Patient	-214		-126		-37	-1
Patient	-139		2267		-14	-1
⋮	⋮	⋮	⋮	⋮	⋮	⋮

3. GA를 통한 특징 소거

바이오 데이터를 이루고 있는 유전자 데이터들은 유전자 사이에 연관 관계가 존재하며, 이에 따라 데이터를 이루고 있는 특징들 사이에 상관관계가 일반적으로 높다. 특징들 간의 상관관계는 전통적인 접근 방법에서 특징들은 서로에 대해 독립이라는 가정을 그대로 적용할 수 없게 하지만, 특징들의 독립 가정을 포기하게 될 경우 특징 선별을 할 때, 특징들의 조합 최적화 문제가 된다. 따라서 특징의 개수에 따라 무한에 가까운 조합을 실험해야 하는데, 이는 실세계 문제를 해결하는데 있어서 적합하지 않다[3]. 그러나 GA를 통해 임의의 특징을 소거하고, 임의의 특징이 소거된 데이터를 측정함에 따라 목적에 얼마나 적합한 부분 집합인지 점수를 계산함으로써 점진적으로 더 나은 부분 집합을 얻는 것이 가능하다.

GA로 문제를 해결하기 위해서는 문제에 대한 모델링이 중요한데, 본 논문에서는 다음과 같이 결정하였다.

임의의 데이터를 이루고 있는 특징 집합을 $\bigcup_{1}^{k} f_k$라고 할 때, 임의의 유전자 G_k는 다음과 같이 정의된다.

$$G_k = \{101010101 \cdots 00010\} \tag{3}$$

이 때, 1은 선택된 특징 f_n이며, 0은 소거된 특징 f_m이다. 또한 G_k의 순서는 주어진 원본 데이터에 있는 특징들의 순서와 동일하다. 또한, G_k에 의해 원본 데이터에서 선택된 특징들의 집합을 S_k, 원본 데이터에서 선택되었으며, 축약된 특징들의 집합을 T_k라고 정의한다.

4. 적합 함수

GA에서는 적합 함수의 결정이 중요한데, 주어진 데이터의 부분집합을 평가하고 점진적으로 개선시켜 나가기 위한 계측을 하기 때문이다.

본 논문에서 사용된 적합 함수는 다음과 같은 형태로 이루어져 있다.

$$Fit(DW_n) = \frac{MCS(W_n)}{CO(DW)} \tag{4}$$

적합함수에서 분자에 해당하는 MCS 함수는 주어진 데이터 중에서 GA에 의해 소거되지 않은 특징 부분 집합 W_n에 대한 평가를 담당한다. 이 함수는 주어진 데이터 Wn을 이루는 특징들이 서로에 대한 비중복도를 평가하는 함수이다.

다음으로 CO 함수는 주어진 축약 데이터 DW를 이루는 특징상에서 각각의 클래스가 보이고 있는 중첩 영역에 대한 평가를 담당하며, 중첩 영역이 클수록 높은 값을 보이게 된다.

4.1 MCS(W_n)

적합 함수의 분자에 해당하는 $MCS(\cdot)$ 함수는 주어진 축약 데이터 DW_n의 특징 비중복도를 나타내며 이는 수식 (5)로 나타낼 수 있다.

$$MCS(W_n) = \frac{\sum_{k=1}^{l} Dr(S_k)}{k} \tag{5}$$

W_n을 이루고 있는 k개의 특징 S_k에 대한 다른 특징

들과의 비중복도를 측정한 평균을 상수로 나타낸다. 이 때, 임의의 k번째 특징 S_k에 대한 특징 비중복도를 측정하는 함수 $Dr(S_k)$는 수식 (6)과 같다.

$$Dr(F_k) = \frac{\sum_{l=1}^{n}(1 - Corrcoef(F_k, F_l))}{n-1} \quad (6)$$

W_n을 이루는 임의의 두 특징 F_k와 F_l에 대해 상관계수를 구하여, 이에 대한 평균을 구한다. 이 작업을 W_n을 이루는 전체 특징에 대해 구하는데, 수식 (6)의 목표가 전체 데이터에 대한 비중복도를 구하는 것이므로, 1에서 상관계수를 뺀 값을 사용한다. 특징 F_k와 자신에 대한 상관계수는 항상 1이므로 데이터 비중복도는 0이 되고 이에 따라 평균을 구하는 수식 (6)에서는 특징 개수에서 1을 뺀 값을 통해 값을 구하게 된다.

여기서 사용되는 상관 계수 함수는 수식 (7)이다.

$$Corrcoef(F_k, F_l) = \frac{|cov(F_k, F_l)|}{\sqrt[2]{var(F_k)var(F_l)}} \quad (7)$$

4.2 CO(DW)

주어진 데이터 DW에서 각 클래스 간의 중첩도를 나타내는 함수 $CO(\cdot)$는 수식 (8)과 같다.

$$CO(DW) = \frac{1}{-Cox(DW)+2} \quad (8)$$

$$Cox(DW) = \frac{Vol(DW) - Vol(DW_{all})}{Vol(DW_{all})} \quad (9)$$

이 때, $CO(\cdot)$ 함수는 중첩 비율에 대한 사상 함수이며, 이는 임의의 클래스들이 완전히 중첩되면 1, 서로 멀리 떨어지면 떨어질수록 0에 수렴하게 된다. 이와 같이 함수를 설정한 이유는 다음과 같다.

(1) 주어진 각각의 특징에서 클래스의 중첩 범위는 거리로 나타낼 수 있다.

(2) 이 때, 중첩된 거리는 각각의 특징에서 나타나게 되며, 모든 중첩 거리의 곱은 주어진 데이터에서 각 클래스들이 충첩된 공간의 부피에 대한 추정치가 된다.

$$Vol(DW) = \sqrt[t]{\prod_{k=1}^{t} Con(F_k)} \quad (10)$$

(3) 클래스 간의 부피를 추정하기 위해서는 각 특징에서 표현된 클래스 간의 중첩 비율의 곱으로 표현해야 하는데 이는 계산된 부피값이 0이 나올 경우 성립하지 않으므로 이러한 상황을 회피하기 위하여 사상함수를 사용한다. 따라서 특징 집합 F의 클래스 중첩도는 수식 (11)을 통해 얻을 수 있다.

$$Con(F) = \frac{1}{-Cop(F)+2} \quad (11)$$

(4) 그러나 실제 주어진 데이터는 Scale이 일정하지 않아 부피에 대한 계산을 할 경우 Scale이 큰 특징들이 저평가 되는 문제가 있으며, 이를 극복하기 위해 주어진 특징에서 전체 분산에 대한 비율을 사용할 수 있다.

$$Cop(F) = \frac{(\sum_{k=0}^{n} Sig(F_k)) - Sig(F)}{Sig(F)} \quad (12)$$

(5) 주어진 축약 데이터 DW을 이루는 임의의 두 클래스를 DW_k, DW_m이라 하고, DW에 포함된 임의의 특징을 F라 할 때, 특징 집합 F 중에서 k 클래스에 포함되는 특징 집합을 F_k, m 클래스에 포함되는 특징 집합을 F_m이라고 정의하자. 이 때, $Sig(F)$함수는 주어진 특징 F의 표준편차이다.

표4. F_k와 F_m 데이터

f_1	f_i	f_n	Class Label	F
:	:	:	:	:
-161	-192	16	1	F_0
-48	187	-73	1	F_0
-176	**13868**	-60	1	
-214	-126	-37	-1	F_1
-139	2267	-14	-1	F_1
:	:	:	:	:

$Cop(F)$ 값이 작은 값일수록 주어진 특징에서 표현된 클래스가 중첩되어 있지 않음을 나타낸다. 또한, 얻어진 중첩 거리에 대한 비율은 음수가 나올 수 있는데, 이는 주어진 특징 F 상에서 클래스들이 서로 중첩되지 않음을 나타낸다.

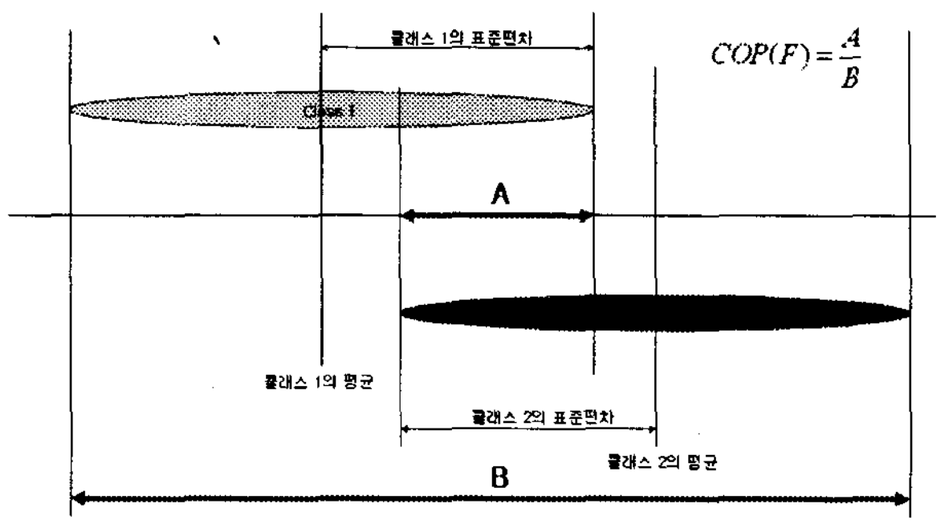
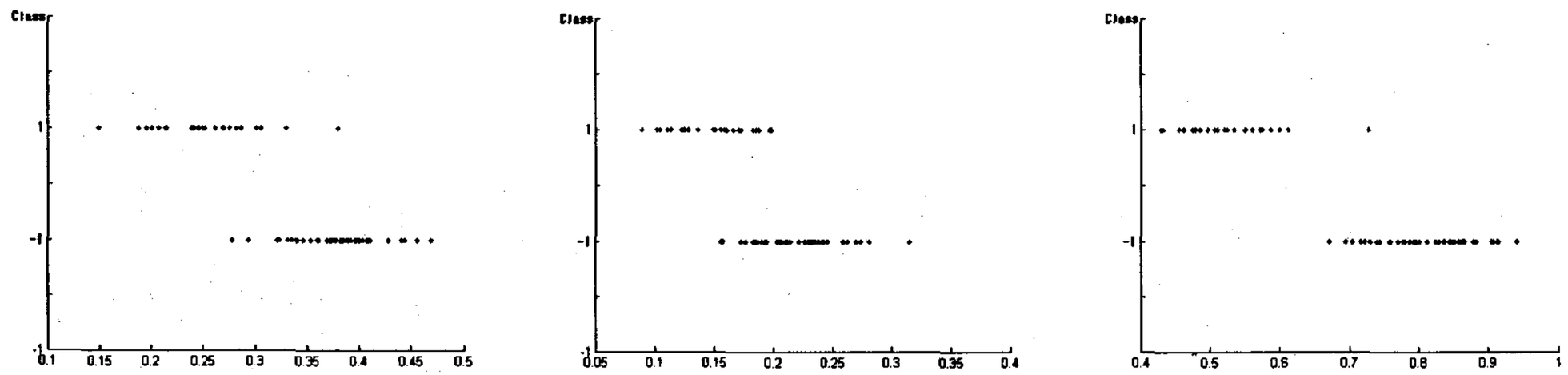


그림 1. $Cop(\cdot)$ 함수의 의미

정리하면, 주어진 데이터를 표현하는 각각의 특징상에서 나타나는 클래스의 중첩도를 비율로 측정하고, 비율들의 곱으로 주어진 특징에서의 클래스 중첩도 부피를 측정하고, 이를 기반으로 주어진 특징으로 이루어진 데이터에서 클래스 간 중첩도를 전체적으로 측정한 상수를 만든다. 이렇게 만들어진 상수를 축약하지 않은 원본 데이터에서 나타난 중첩도에서 얼마나 감소하였는지, 혹은 증가하였는지를 비율로 계산하도록 한다. 이를 통해 GA가 선택한 데이터의 구분력을 증가시키도록 한다.



a) MCS : 0.8625 b) MCS : 0.8655 c) MCS : 0.8587

그림 2. 선택된 유전자들을 FLDA로 사상한 결과와 평균 비중복도(MCS)

5. 실험 결과 및 결론

그림 2는 GA를 통해 얻어진 유전자(선택된 특징들의 집합) 중에서 상위 3개의 유전자에 대해 FLDA(Fisher Linear Discriminant Analysis)를 적용한 결과이다. 유전자간의 비중복도를 나타내는 MCS 값이 높아질수록 클래스 사이의 중첩도가 높아짐을 알 수 있다. 또한, 그림 2-c에서 Class 1을 사상한 결과를 분석하면, 별개로 떨어진 1개의 샘플이 있으나, 이 샘플로 인하여 FLDA가 편중된 축을 찾지 않음을 알 수 있다.

이와 관련하여, [표5]는 각각의 알고리즘에 의해 특징 선별이 이루어지고, 이에 대한 클래스 예측기의 정확도를 나타내고 있다. 클래스 예측기는 K-nn(K-nearest neighbor) 알고리즘을 사용하여, 무작위 선택(Random sampling) 1000회 반복하고, 각각 선택된 데이터의 정확도를 구한 다음, 평균을 취한 것이다.

알고리즘들이 모두 비슷한 결과를 보이고 있는데, 이는 유전자를 표현하고 있는 특징 간의 Scale이 정규화(Normalization) 되어 있지 않아서 일부 특징에 편중된 거리 척도에 의한 정확도의 하향평준화로 분석된다. 그림 2에서 보는 바와 같이 사상 방법을 적용하면, 실험에 포함된 모든 알고리즘이 높은 정확도를 보였다.

이에 반해, 각각의 알고리즘을 통해 얻은 특징들을 상관 계수를 통해 중복도를 측정하면, 제안된 알고리즘의 비중복도가 가장 높았으며, T-test의 중복도가 높아 전통적인 알고리즘의 한계를 알 수 있었다. 또한 mRMR과 제안된 알고리즘의 비중복도 측정에서는 거의 유사한 결과를 보였는데, 알고리즘 사이의 우위를 결정하기 위해서는 유전자 간의 상관관계 정보를 사용하는 알고리즘의 개발이 필요할 것으로 보인다.

실험 결과를 정리하면, 제안한 알고리즘은 기존의 알고리즘들이 보이는 클래스 예측기 적용에서의 정확도를 잃지 않으면서 선택된 유전자간의 중복도가 낮은 특징을 선별하였다. 이 데이터들은 Scale이 일정하지 않으므로, 선별된 유전자들을 별다른 변형없이 클래스 예측기를 통해 적용하기는 무리가 있었으나, FLDA 등 다양한 특징 추출(Feature extraction) 알고리즘을 적용하여 찾아낸 벡터 상에서 클래스 예측기를 적용할 경우, 매우 높은 적중률을 얻을 수 있었으므로 실제 선택된 데이터가 초평면(Hyper-plane) 상에서 잘 구분되고 있음을 알 수 있었다.

표5. 정확도 실험 결과

Dataset	Feature Selection	Accuracy (%)
Leukemia	Proposed	76.5
	T-test	79.1
	mRMR	80.0
Colon	Proposed	78.5
	T-test	74.0
	mRMR	78.5

6. 향후 연구

향후 연구에서는 1) GA에서의 검색 능력 개선과 2) GA에 의존하여 검색하지 않고 수학적인 방법으로 특징 부분 집합을 찾는 방법을 구하며, 3) 소거된 특징들에 대한 평가를 통해 현재 평가 중인 유전자 집합에 대한 검증 작업을 거치도록 하며, 4) 유전자 간의 상관관계를 활용할 수 있는 방법을 추가하여 현재의 알고리즘이 가지고 있는 문제점을 극복해야 할 것이다.

7. 참고 문헌

[1] T. R. Golub, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," Science, Vol 286, pp.531-537, Oct 1999

[2] Danh V, "Tumor classification by partial least squares using microarray gene expression data," Bioinformatics, Vol. 18, No. 1, pp.39-50, Jun 2001

[3] Il-Seok Oh, "Hybrid Genetic Algorithms for Feature Selection," IEEE Trans. Pattern analysis and Machine intelligence, Vol. 26, No. 11, pp.1424-1437, Nov 2004

[4] M. S. Yang, K. L. Wu, "A Similarity-Based Robust Clustering Method," IEEE Trans. Pattern analysis and Machine intelligence, Vol. 26, No. 4, pp.434-448, Apr 2004

[5] Chris. D, H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," Journal of Bioinformatics and Computational Biology, Vol. 3, No. 2, pp. 185-205, Jun 2004