Proceedings of KIIS Spring Conference 2008 Vol. 18, No. 1.

마이크로어레이 데이터 분석을 위한 선형 특징 선별 기법
A fast feature selection technique for microarray data

이재성, 김대원

서울시 동작구 중앙대학교 컴퓨터공학과

E-mail : curseor@hotmail.com

요 약

마이크로어레이 데이터는 대량의 유전자들을 짧은 시간에 테스트 하여 얻은 대량의 데이터로 구성되어 있다. 그러나 이렇게 얻은 대량의 데이터에서 특징으로 표현되는 유전자의 수가 매우 많고, 각각의 유전자는 서로에 대해 독립적이지 않기 때문에 전통적인 데이터 마이닝 기법을 적용하여 바이오마커를 찾아내는 작업이 용이하지 않다. 마이크로어레이 데이터에서 나타나는 이러한 특성과 여기에서 파생되는 문제점들을 극복하기 위해 다양한 특징 선별 방법론들이 등장하였으나 다소의 문제점을 가지고 있어 실제 세계의 문제에 적용하기 어렵다. 본 논문에서는 코사인 내적 행렬과 행렬식을 이용하여 직교하지 않는 특징들을 제거하는 방법에 대해 소개하고, 그 결과를 분석하였다.

Key Words : Orthogonality, Determinant, Cosine inner matrix, Feature Selection

1. 서 론

일반적인 데이터 마이닝 기법에서 마이크로어레이 데이터를 다룰 때, 특징 선별 방법을 사용하고 있다. 특징 선별 방법은 매우 다양하지만, 이러한 방법들이 목표로 하고 있는 것은 대개 마이크로어레이 데이터에 포함되어 있는 특징들이 서로 중복된 것들과 상관계수가 높게 나오는 특징에 대해 제거를 하는 것이 일반적이다[1].

이러한 특징 선별 방법론들을 마이크로어레이 데이터에 적용하는 이유는 먼저 마이크로어레이 데이터의 특징(축)을 이루고 있는 유전자들이 서로에 대해 독립적이지 못하고, 상호 연관성을 가지고 있기 때문에 마이크로어레이 데이터에서 표현되는 측정값의 분포가 서로 중복되는 경우가 많고, 또한 모든 유전자들에 대한 정의가 이루어지지 않아 유전자들에 대한 측정값 자체가 동일한 경우가 있어 이후에 적용되는 다양한 기법들이 가정하고 있는 부분을 만족하지 못하기 때문이다.

유전자 간의 중복도를 판단하는 방법 역시 다양한 방법론들이 제안되었으나 대부분 상관계수(Correlation Coefficient)를 통하여 특징 간의 중복도를 측정하고, 중복도가 높은 특징을 제거하는 방법을 사용하고 있다.

2. 상관계수의 문제점

두 특징 간의 상관계수 c를 측정하는 수식은 일반적으로 다음과 같다. 이 때, f_i와 f_j는 마이크로어레이 데이터의 특징(축)을 표현하고 있으며, $cov(f_i,f_j)$는 두 특징의 공분산을, $var(f_i)$는 특징 f_i의 분산을 나타낸다.

$$c(f_i,f_j)=\frac{cov(f_i,f_j)}{\sqrt{var(f_i)var(f_j)}} \quad (1)$$

수식 (1)을 활용하여 상관계수를 통해 중복도를 측정하는 방법은 다양한 장점이 있지만, 마이크로어레이 데이터에 적용하기 위한 특징 선별 방법론으로써는 단점 역시 명확하다. 먼저, 상관계수의 특성에 관한 문제인데 마이크로어레이 데이터에 포함된 유전자들의 경우 그 특성이 완전히 파악되지 않은 상태에서 분석을 시작하게 되므로, 데이터에 포함된 유전자들이 개별적인 유전자인지 명확하지 않으며, 심지어 유전자인지 단백질인지 판별되지 않은 것들이 특징을 이루고 있는 경우가 많다. 이러한 상황에서 상관계수는 서로 관련이 없는 유전자의 경우라도 상관계수를 측정해내기 때문에 수학적으로는 중복도가 있을지라도 실제 유전자 단계에서는 관련이 없는 경우가 있다[2].

이외에 수학적인 기저의 문제가 발생을 하는데, 상관계수는 수식의 특징으로 인해 주어진 샘플이 2개일 경우 측정값이 반드시 1(또는 -1)로 표현된다.

데이터 행렬		
f_1	f_2	f_3
0	1	0
1	0	2

상관계수 행렬		
1	-1	1
-1	1	-1
1	-1	1

그림 1. 상관계수 행렬의 문제점. f_1과 f_2는 서로 직교하지만, 상관계수는 반드시 1로 표현됨.

마지막으로 선별의 문제가 있는데, 중복도가 높은 특징 f_k에 대해서 제거를 하게 될 경우, 향후 만들어질 유전자 부집합에는 f_k가 포함될 방법이 없게 된다. 상관계수는 반드시 두 개의 특징에 대해서 측정을 하게 되기 때문에, 두 개의 특징이 서로 중복도가 높다고 하여 제거를 하게 되었을 경우에 이 두 개의 특징을 제외한 다른 특징들과의 중복도가 고려되지 않거나, 고려를 한다고 하더라도 전체 조합을 모두 보게 되어 실질적으로 실세계의 문제에 적용하기 어렵게 된다.

3. 제안하는 알고리즘

3.1 코사인 내적 행렬(Cosine inner matrix)

특징 f_i와 f_k에 대한 코사인 내적을 구하는 수식은 다음과 같다. 이 때, f_i^T는 f_i의 전치행렬을 의미한다.

$$s(f_i, f_k) = \frac{f_i^T f_k}{\sqrt{|f_i||f_k|}} \quad (2)$$

이렇게 얻어진 코사인 내적 행렬은 다음과 같은 형태를 보이게 된다.

데이터 행렬			코사인 내적 행렬		
f_1	f_2	f_3	1	0	1
0	1	0	0	1	0
1	0	2	1	0	1

그림 2. 코사인 내적 행렬의 형태. 샘플의 개수에 관계없이 중복도를 측정해주며, 여기서의 중복도는 직교성을 보장함.

3.2 행렬식(Determinant)

상관계수의 문제점에서 언급하였듯이 상관계수는 반드시 한 쌍의 특징에 대해서 중복도를 측정하기 때문에 이후에 특징의 선별에 따른 부집합의 결정 형태가 달라진다. 이러한 문제를 회피하기 위해서 행렬로 표현된 중복도를 하나의 상수로 표현할 필요가 있는데, 행렬을 상수로 표현하는 대표적인 방법이 바로 행렬식이다.

코사인 내적 행렬을 통해서 데이터의 중복도를 표현할 때, 가장 이상적인 경우는 모든 특징들이 서로에 대해 모두 직교할 경우이며, 이때의 코사인 내적 행렬은 단위행렬(Identity matrix)이 되므로, 행렬식의 값은 최대 1이 된다. 또한 부집합을 구성하기 위하여 각각의 특징들이 부집합에서 중복도에 어떠한 영향을 끼치는지 측정하여 각 특징들의 중복도의 최대 임계값(Upper bound) $t(f_1)$을 구한다.

코사인 내적 행렬			f_1이 중복도에 미치는 영향 측정		
1	0.5	0.3	1	0.5	0.3
0.5	1	0.7	0.5	1	0
0.3	0.7	1	0.3	0	1
f_1이 데이터의 중복도에 영향을 주는 값(Entry)			행렬식 최대 임계값 $t(f_1) = 0.66$		

그림 3. 전체 데이터에서 특징 f_1이 중복도에 끼치는 영향에 대한 최대 임계값을 구함.

3.3 부집합 구성을 위한 중복도 측정

부집합을 구성할 때는 각 특징의 중복도를 표현하는 모든 값이 사용될 필요는 없다. 부집합에 포함될 특징의 개수에 따라 어떠한 값을 알아내는가는 부집합이 결정되기 전에는 알 수 없으므로 모든 조합을 분석해야 하지만 최대 임계값을 활용하여 이러한 문제를 일부 해결할 수 있다. 이에 대한 방법은 그림 3과 같으며, 각 특징과 연관되어 있는 값(Entry)들을 정렬하여 평가할 수 있다. 그림 4는 특징이 3개로 구성되어 있는 데이터에 대해서 특징이 2개로 구성된 부집합을 작성한다고 하였을 때, 부집합에 f_1이 포함될 경우 중복도가 가장 낮을 수 있는 상황은 0.3이 코사인 내적 행렬에 포함되었을 경우임을 보이고 있다.

코사인 내적 행렬			f_1이 부집합에서 중복도에 미치는 영향 측정		
1	0.5	0.3	1	0.3	0.5
0.5	1	0.7	0.3	1	0
0.3	0.7	1	0.5	0	1
부집합을 구성할 경우 f_1에 의해 결정되는 값			행렬식 최대 임계값 $t(f_1') = 0.91$		

그림 4. 각 특징의 부분 최적값을 찾는 방법

4. 실험 결과 및 생물학적 분석

제안한 알고리즘의 분석을 용이하게 하기 위해 백혈병(Leukemia) 데이터에 대해 Chris D.가 제안한 방법론에서 선별된 유전자 62개 중에서 실제로 백혈병과 관련이 있음이 밝혀진 유전자 23개와 간접적인 관련이 있음이 밝혀진 유전자 14개, 직·간접적인 관련이 밝혀지진 않았으나 Chris D.가 제안한 알고리즘에 의해 선별된 유전자 10개, 거의 무관하거나 알려지지 않은 유전자 13개를 포함하여 총 60개의 유전자로 전체 데이터를 작성하고 제안한 알고리즘을 적용하였다[1,3]. 이에 대한 결과는 표 1과 같다. 실험 결과를 분석해보면 제안한 알고리즘이 간접적인 연관성이 있는 유전자들에 대한 선별율이 낮은 것을 볼 수 있는데, 이는 각 유전자들에 대한 개별적인 측정을 했기 때문이라고 추정할 수 있다. 또한 부집합에 포함될 특징의 개수에 따라 적중률이 높아지는 것을 알 수 있는데, 특징의 개수에 따라 부집합의 최적화가 달라지므로, 특징 선별에서 유효하게 작용되었음을 알 수 있다.

표 1. 실험 결과. 전체 측정은 부집합에 포함된 특징의 개수에 관계없이 전체 값(Entry)에 대해 평가를 하였고, 부분 측정은 부집합에 포함된 특징의 개수만큼의 값(Entry)에 대해 평가한 결과임

측정법	직접	간접	추정	무관
전체	7/23	3/14	4/10	6/13
부분	11/23	3/14	2/10	4/13

5. 참고 문헌

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

[2] J. R. de Haan et al, "Interpretation of ANOVA models for microarray data using PCA," Bioinformatics, Vol. 23, No. 2, pp. 184-190, Jan 2007

[3] C. Ding et al, "Minimum Redundancy feature selection from microarray gene expression data," Journal of Bioinformatics and Computational Biology, Vol. 3, No. 2, pp. 185-205, Apr 2005