2004년도 대한전기학회 하계학술대회 논문집 2004.7.14-16

GA와 중회귀분석을 이용한 부정맥 진단의 최적 웨이블릿 계수의 선택

정갑성*, 김태선**, 이종호*
*인하대학교 정보통신공학과, **가톨릭대학교 정보통신전자공학부

Optimal wavelet coefficient selection for diagnosis of arrhythmia using genetic algorithm and multiple regressions

Kab Sung Chong*, Tae Seon Kim**, Chong Ho Lee*
*Inha University, **Catholic University of Korea

Abstract - 본 논문은 유전알고리즘을 이용하여 부정맥 진단의 최적화된 입력을 구성하는 방법을 제시한다. 심전도 신호의 특징을 추출하기 위해 웨이블릿변환이 널리 사용되고 있지만, 추출된 특징들의 선택과 최적화의 문제에 대해서는 명쾌한 해결책을 제시하지 못하고 있다. 심전도 신호는 연속 웨이블릿 변환을 이용해 5레벨로 분해되었으며, 각 서브밴드에서 추출된 계수들은 부정맥 진단을 위한 특징으로 쓰이게 된다. 웨이블릿변환을 통해 추출된 특징들(feature)은 유전자 알고리즘과 중회귀분석을 통하여 부정맥 진단을 위한 최적화된 특징조합이 결정되었다. 본 연구를 통해 특정레벨의 어떤 계수가 부정맥 진단에 크게 영향을 미치는지 판단할 수 있었으며 입력의 차원감소는 연산시간의 축소를 가져왔고 분류정확도를 향상시켜 분류기의 성능을 증대시켰다.

1. 서 론

심전도의 심장의 전기적인 활동의 기록으로 심전도 신호의 패턴분류는 환자의 심질환을 진단하는데 중요한 수단으로 사용되고 있다. 현재까지 심전도 신호의 패턴을 분류하기위해 여러 가지 방법이 사용되어 왔으나 본 논문에서는 인공지능적인 방법인 신경망을 이용한 패턴분류에 한하여 최적의 입력을 구성하기 위한 방법을 비교 분석한다. 심전도 신호를 분류하기 위해서는 샘플링 과정을 거치게 된다. 본 논문에서는 최근 심전도 특징 추출에서 우수한 효과를 보이고 있는 웨이블릿 변환을 사용하였다[1][2]. 그러나 웨이블릿 변환하여 특징 정보를 추출한다 하더라도 그 정보의 양이 많으므로 추출된 정보에 관한 취사선택 문제가 생기게 된다. 따라서 웨이블릿 변환하여 나온 심전도 신호에 관한 정보 중 양질의 정보만을 취하기 위하여 유전자 알고리즘과 회귀분석을 사용하여 신경망 패턴분류기를 위한 최소의 입력을 구성하는 방법을 제안하고 그 성능을 비교해 보았다.

2. 실험 방법

실험에 사용된 데이터는 MIT_BIH 공개 Database의 심전도 신호를 사용하였으며 신경망은 3가지 패턴(APC, PVC, Normal)을 분류할 수 있도록 구성되었다. 부정맥 진단을 위해서는 역전파신경회로망을 사용하였고, GA와 중회귀분석을 통해 선택된 특징조합들은 신경회로망의 입력으로 사용되어 그 성능을 검증하였다. 전체 시스템의 모식도는 그림 1 과 같다.

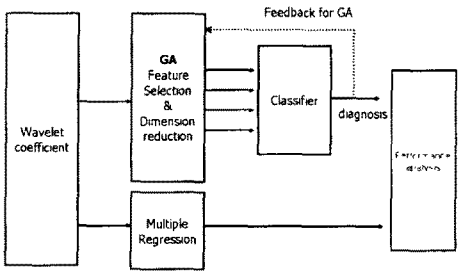

그림 1. 실험을 위한 모식도

2.1 특징 추출을 위한 웨이블릿 변환

전처리 과정으로서 웨이블릿 변환을 하여 특징을 추출하였다. 웨이블릿 변환은 푸리에 변환의 대안으로 제시된 알고리즘으로서 푸리에 변환에 비해 시간과 주파수 정보를 동시에 확인할 수 있다는 장점을 가진다. 본 논문에서는 각 심전도 신호에 대해 연속 웨이블릿 변환을 사용하였으며 기저함수는 'db3'(Daubechies)함수를 사용하였고 5레벨까지 multilevel wavelet decomposition을 수행하였다. 그림 2는 Matlab을 이용하여 심전도 신호 중 정상파형의 한주기를 db3함수로 5레벨로 분해한 그림이다.

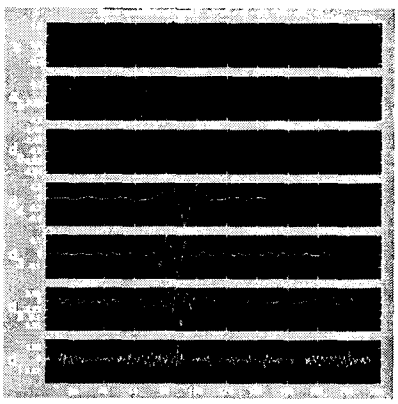

그림 2. db3함수를 이용한 웨이블릿 분해결과

그림 3은 5수준의 상세 레벨중 압축적인 정보를 표현하고 있는 cD5,cD4,cD3레벨에서 추출한 계수의 개수를 나타내고 있다. 심전도 신호 S는 하이패스필터와 로우패스필터를 거치며 모함수에 의거 Approximation level과 Detail level로 분해가 이루어 지는데 cDx로

표현된 부분이 Detail level로 본 실험에서 사용된 계수는 총 61개이다. 이것으로서 각 심전도 신호에 대한 정보는 61개의 변수로 표현되었다고 할 수 있다.

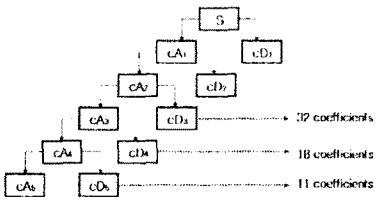

그림 3. 웨이블릿 트리와 실험에 사용된 계수

2.2 특징 인자 선택법

웨이블릿 변환을 사용하여 추출된 61개의 웨이블릿 계수중 의미있는 정보만을 선택하기 위한 방법으로 두가지 방법을 사용하였다. 첫 번째 방법은 유전자 알고리즘을 사용한 것이며 두 번째 방법은 통계적인 방법인 중회귀 분석이다.

2.2.1 유전자 알고리즘과 best subset selection

유전자 알고리즘은 자연계의 법칙을 모사하여 만든 알고리즘으로 홀랜드에 의해 처음 제시되었다. 자연선택과 약육강식에 기초한 유전자 알고리즘은 최적해를 찾기위한 방법으로 사용되고 있다. 본 실험에서는 simpleGA를 사용하였다. 아래그림에서 염색체 유전자의 위치는 정확히 웨이블릿계수에 사상된다.

2.2.1.1 염색체 설계

염색체는 총 61bit의 이진 값을 갖는 스트링으로 구성되었으며 초기 개체수는 30으로 하였다. 세대차이는 계산에 포함되지 않았으며 룰렛휠 방식으로 선택하였다. 그리고 교배와 돌연변이를 통해 세대를 거치며 우수한 형질만이 선택될 수 있도록 하였다. 초기집단을 구성하는 염색체는 난수를 이용하여 생성하였으며 적합도를 만족시키는 조건에서 best chromosome으로 선택된 한 개의 스트링중 1값이 존재하는 부분의 계수를 선택하는 것으로 하였다.
그림4에서 염색체 유전자의 위치는 정확히 웨이블릿계수에 사상된다. 유전자 알고리즘의 연산은 이진문자열로 하게 되지만 이것은 웨이블릿 계수가 있는 곳을 가리키는 주소로 사용되게 된다.

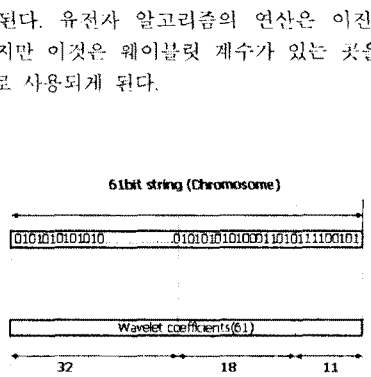

그림 4. 염색체 구성

2.2.1.2 적합도 함수

적합도 함수는 현세대를 평가하여 다음세대로 진화할 해야 할지의 여부를 결정짓는 중요한 부분으로 주어진 문제와 상황에 따라 다르게 적용되어야 한다. 본 논문에서는 그림1의 시스템의 모식도에 나타난 바와 같이 신경망의 재현오차값을 사용하였다. 상기된 방식에 의해 선택된 특징스트링이 우수한 유전자를 가지고 있다고 한다면 신경망의 분류율은 높을것이며 그렇지 않다면 반대의 상황이 될 것이다. 유전자알고리즘은 목표로 하는 신경망의 오차 값에 수렴할 때까지 반복하게 된다.

2.2.2 중회귀 분석과 인자 추출

다중회귀분석이란 종속변수의 변화를 설명하기 위하여 두 개 이상의 독립변수가 사용되는 선형회귀모형을 말한다. 본 논문에서 다루는 문제는 여러항의 독립변수가 종속변수에 미치는 영향을 밝히고 어떤 변수들의 조합으로 사건의 결과가 설명가능한지를 찾는 것 이므로 중회귀분석을 사용한다. 회귀분석은 회귀식을 구성하여 주는데 회귀식이란 독립변수와 종속변수간의 관계를 통계적으로 분석하여 얻은 일종의 방정식이다. 심전도 신호에서 추출된 특징정보들은 독립변수가 되며 질환에 대한 진단결과는 종속변수로 쓰여진다. 분석을 위해서는 통계패키지인 minitab을 이용하였다.

2.2.2.1 다중 공선성

2개의 인자가 서로 상관관계가 아주 높고, 각각의 인자가 특성 값에 영향을 미치는 정도가 서로 촉매작용을 하여 그것보다 훨씬 더 영향이 있는 것처럼 보일 때 다중공선성(Variance Influence Factor)이 있다고 한다. 문제는 회귀분석에서 다중공선성이 유발하는 문제점이다. 다중공선성이 있을 경우 결정계수의 값이 과대하게 나타날 수 있으며 분산이 커져 회귀모형의 적합성이 떨어지고 다른 중요한 인자가 모형에서 제거될 가능성이 높다. 실험결과 cD3 level의 27th 인자가 다중공선성을 유발하여 이를 제거후 실험하였다.[4]

2.2.2.2 단계별 회귀(Stepwise regression)

단계별 회귀기법은 종속변수에 기여도가 가장 높은 변수를 선택한 후 나머지 변수 중에서 새로이 회귀모형에 추가될 때 기여도가 가장 높은 변수를 선택하여 새로이 추가 또는 삭제되는 변수가 없을 때까지 반복하여 분석하는 방법이다. 단계별 회귀(stepwise) 방식에서는 독립변수 중 영향력이 큰 순서대로 산출되어 실험 후 결과를 쉽게 판단할 수 있는 장점이 있다.[4]

3. 실험결과

3.1 GA를 이용한 best subset selection

실험결과 소수의 입력인자로 선택되는 것을 확인 할 수 있었지만 신경망과 유전자 알고리즘의 결합은 소수의 인자로 입력의 차원을 크게 감소시키지는 못하였다. 그림

6은 진화연산 후 얻은 결과로 30개의 염색체의 신경망 시뮬레이션 에러를 나타내는 종합적인 결과이다.

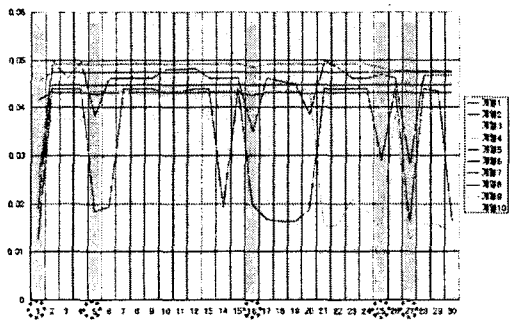

그림 6. 진화 연산후 각 염색체의 분류정확도

즉 표시된 1, 5, 16, 25, 27번 염색체가 공통적으로 좋은 성능을 보임을 알 수 있으며 염색체들의 공통적인 유전자는 x1, x11, x33, x57, x60, x4, x18 등 18개의 인자로 나타났다.

3.2 중회귀분석 이용한 best subset selection

웨이블릿 변환을 통해 얻어진 61개의 변수중 4, 6, 18, 3, 49, 17, 15, 7, 5, 52번째의 계수들이 best subset으로 선택되었으며 얻어진 계수들로 아래와 같은 회귀식을 얻을 수 있었다.

$$y1 = 1.06 + 0.261\,x4 - 0.462\,x6 + 0.772\,x18 + 0.553\,x3 + 0.153\,x49 + 0.735\,x17 + 0.263\,x7 + 0.252\,x5 - 0.698\,x52$$

결과 분석을 위해 최종 결과 그림 7을 보면 ($p<0.05$)의 값이므로 선형 회귀모형이 존재한다는 사실을 알 수 있었고 결정계수 설명력을 나타내고 있는 R-sq(adj) = 97.5%로 위의 10가지 변수로 결과를 설명할 수 있는 비율은 97.5%가 된다는 사실을 확인할 수 있다.

Regression Analysis:
y1 versus x4, x6, x18, x3, x49, x17, x7, x5, x52

Predictor	Coef	SE Coef	T	P
Constant	1.0573	0.1727	6.12	0.000
x4	0.2605	0.2755	0.95	0.349
x6	-0.46248	0.04978	-9.29	0.000
x18	0.7718	0.1042	7.41	0.000
x3	0.5535	0.3313	1.67	0.100
x49	0.1529	0.1264	1.21	0.231
x17	0.7354	0.3368	2.18	0.033
x7	0.26264	0.05159	5.09	0.000
x5	0.25180	0.05368	4.69	0.000
x52	-0.6982	0.1897	-3.68	0.000

S = 0.147007 R-Sq = 96.9% R-Sq(adj) = 96.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	9	42.6385	4.7376	219.22	0.000
Residual Error	63	1.3615	0.0216		
Total	72	44.0000			

그림 7. 중 회귀 분석의 최종 결과

이는 곧 61개의 입력으로 구성될 61-x-x 구조의 신경망이 10-x-x로도 가능하다는 것을 말해주고 있으며 이를 확인하기 위해 Matlab으로 시뮬레이션 해 보았다.

recall error는 Mean Squared Error 0.01로 변수 10개로 충분히 설명가능하다는 결과를 얻었다.

3.3 실험결과 비교

유전자 알고리즘과 중회귀분석을 이용하여 61개의 신경망 입력의 차원을 감소시킨 결과 중회귀분석이 유전자 알고리즘을 사용했을 때보다 나은 결과를 보였다. 문제의 특성상 변수의 개수가 많지 않았으므로 이 경우에는 통계적인 방법이 유용했다. 유전자 알고리즘으로는 18개의 인자가 선택되었으며 중회귀 분석은 총 61개중 10개의 인자로 축약시켜주어 유전자알고리즘에 비해 약 55% 향상된 결과를 보였다.

4. 결 론

유전자 알고리즘과 통계적 기법인 중회귀 분석을 사용하여 신경망의 입력이 될 변수를 선택하는 실험을 수행한 결과 본 논문에서 다루어진 문제에서는 중회귀 분석이 유전자 알고리즘을 사용했을 때보다 나은 결과를 얻었다. 신경망과 유전자 알고리즘의 조합은 더 큰 복잡도를 야기하는 것으로 사료된다. 또한 선택된 인자의 개수도 전체 61개의 인자 중 18개로 중회귀 분석의 약 두 배 정도의 인자로 축약되었다. 반면 회귀분석은 실험의 신뢰도를 높이기 위한 10회 반복수행시 동일한 인자를 선택해줌으로써 신뢰도와 정확성이 높았음을 확인 할 수 있었다.

심전도 신호를 웨이블릿 변환하여 특징을 추출하는 경우 중요도를 알 수 없는 많은 계수 중 결과 값에 큰 영향을 미치는 인자를 선택함으로써 입력의 차원감소를 하였다. 이는 곧 패턴분류 작업을 효율적으로 할 수 있는 기반이 될 것이며 연산시간을 줄여 분류기의 성능을 증대 시켰다.

[참 고 문 헌]

[1] Stefan Pittner and Sagar V. Kamarithi, "Feature extraction from wavelet coefficients for pattern recognition tasks, IEEE transaction on pattern analysis and machine intelligence, vol.21, 1999

[2] Castro, B., Kogan, D., Geva, A.B.,"ECG feature extraction using optimal mother wavelet", Electrical and ELectronic Engineers in Israel, 2000. The 21st IEEE Convention of the, Pages:346 - 350, 11-12 April 2000,

[3] de Chazal, P., Celler, B.G., Reilly, R.B.,"Using wavelet coefficients for the classification of the electrocardiogram", Engineering in Medicine and Biology Society, 2000. Proceedings of the 22nd Annual International Conference of the IEEE , Pages:64 - 67 vol.1 , 23 28 July 2000

[4] 최용석,정광모, "실무자를 위한 MINITAB 다변량 분석",이레테크,2001