ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
논문번호 18-04-32
208

j.inst.Korean.electr.electron.eng.Vol.22,No.4,1093~1098,December 2018
http://dx.doi.org/10.7471/ikeee.2018.22.4.1093

고성능 컴퓨팅을 활용한 뉴럴 네트워크 기반의 휴대용 질병 진단 플랫폼 구현 방법론

Methodology for Implementation of the Portable Disease Diagnosis Platform based on Neural Network Using High Performance Computing

김 상 만*, 박 주 성*★

Sang-man Kim*, Ju-Sung Park*★

Abstract

In this paper, we proposed a methodology for portable disease diagnosis platform using high performance computing. The proposed methodology consists of gathering clinical data, diagnosis and feature selection algorithm, implementation of diagnosis platform. For the algorithm verification, a clinical data which is obtained from 401 people(314 normal subjects and 87 liver cancer patients) using a microarray consists of 1,146 aptamers were used. As the result, we could diagnosis liver cancer with 97.5% accuracy using the 32 selected aptamers. Based on these results, we designed and implemented a portable disease diagnosis platform which has 32 bio-signals as inputs.

요 약

본 논문에서는 고성능 컴퓨팅을 활용한 뉴럴 네트워크 기반의 휴대용 질병 진단 플랫폼 구현 방법론을 제안한다. 제안하는 방법론은 임상 데이터 수집, 진단 알고리즘 및 반응 물질 선정, 진단 플랫폼 구현으로 구성된다. 진단 알고리즘 검증을 위해서 총 401명(정상인 314명, 간암환자 87명)의 혈액과 1,146개의 압타머(aptamer)로 구성된 마이크로 어레이로부터 얻어진 임상 데이터를 사용 하였다. 검증 결과, 최종적으로 32개의 선별된 압타머를 사용하여 97.5%로 간암 여부를 판별 할 수 있었다. 이것을 바탕으로 32개의 생체 신호를 입력으로 가지는 휴대용 질병 진단 플랫폼을 설계 및 구현하였다.

Key words : disease diagnosis, feature selection, neural network, high performance computing, machine learning

Ⅰ. 서론

과거부터 현재에 이르기까지 좀 더 정확하고 신속한 질병 진단을 위한 많은 연구가 수행되어져왔다. 일반적으로 질병 진단은 제한된 수의 마커(marker) 혹은 반응물질(feature)을 통해서 이루어지는데, 이것은 경우에 따라 정확한 진단 결과를 얻지 못 하기도 한다. 최근에는 이러한 한계를 극복 하고자 인공지능 신경망(artificial intelligence)과 마이크로 어레이(micro array)를 활용한 질병진단에 대한 연구가 활발히 이루어지고 있다.[1][2] 본 논문에서는 인공지능 신경망 중 하나인 뉴럴 네

* Dept. of Electronics Engineering, Pusan National University
★ Corresponding author
E-mail : juspark@pusan.ac.kr, Tel : +82-51-510-2444
※ Acknowledgment
This work was supported by a 2-Year Research Grant of Pusan National University.
Manuscript received Dec. 7, 2018; revised Dec. 22, 2018; accepted Dec. 24, 2018

트워크(neural network)를 활용한 질병 진단 플랫폼 구현 방법론을 제안한다. 또한, 제안하는 방법론을 수행 및 검증하는 과정에서 많은 연산량을 필요로 하게 되는데 이것을 효율적으로 처리할 수 있는 고성능 컴퓨팅 활용 방법에 대해서도 제안하고 있다. 고성능 컴퓨팅 활용을 통해 제안 하는 방법론에 대한 정확도 및 실제 활용 가능함을 검증 하였다. 최종적으로 얻어진 진단 알고리즘 및 반응 물질 정보를 활용하여 휴대용 질병 진단 플랫폼을 구현 하였다.

본 논문의 구성은 다음과 같다. 먼저, II장에서는 제안하는 방법론의 개요, 질병 진단 알고리즘 및 반응 물질 선정 알고리즘에 대해서 소개하고, III장에서는 이 과정에서 요구되는 많은 연산량을 고성능 컴퓨팅을 활용하여 효율적으로 처리하는 방법을 제안할 것이다. IV장에서는 얻어진 알고리즘을 바탕으로 휴대용 질병 진단기기 설계 및 구현에 대한 소개를 하고, 마지막 V장에서 결론을 맺는다.

II. 진단 알고리즘 및 반응 물질 선정

제안하는 방법론은 3단계로 구성된다. 첫 번째 단계에서는 여러 반응 물질을 활용하여 특정 질병과 반응된 임상 데이터를 수집한다. 이 과정에서는 다수의 반응 물질을 통해 많은 반응값을 얻을수록 뒤에 이어질 반응 물질 선정에 유리하다. 두 번째 단계에서는 이렇게 수집한 데이터와 기계학습 및 고성능 컴퓨팅을 활용하여 보다 정확한 진단에 도움이 되는 반응 물질들을 선별하고, 학습을 통해 진단 알고리즘을 완성한다. 이 과정에서는 1단계에서 얻은 임상데이터에 대한 선 처리 및 정렬 과정이 필요하다. 마지막 단계에서는 얻어진 진단 알고리즘을 이식하고 선정된 반응 물질을 활용하여 진단 플랫폼을 설계 및 구현한다.

1. 진단 알고리즘

질병 진단을 위한 알고리즘 및 반응 물질 선정을 위해 본 논문에서는 기계 학습의 한 종류인 뉴럴 네트워크를 사용 하였다. 사용한 뉴럴 네트워크는 입력층, 은닉층, 출력층으로 구성된다. 입력층은 다수의 입력값을 받을 수 있으며, 은닉층은 10개의 노드를 가진다. 출력층은 2개의 출력값을 가지는데 그 구조는 그림 1과 같다. 뉴럴 네트워크에서는 오류 역전파 방식을 이용하여 각 노드들 사이에 연결된 가중치를 조절하는데 이를 학습이라고 한다.[3] 학습의 과정은 주어진 입력 값에 대하여 원하는 출력 값을 얻기 위해 각 가중치들을 조절 하여 그 오차를 줄이는 방향을 진행 된다. 주어진 특정 데이터를 통해 학습이 완료된 신경망은 임의의 입력값에 대하여 스스로 판단이 가능하게 되며, 이를 활용하면 특정 질병에 대한 임상데이터를 학습시킨 후 질병 진단에 활용할 수 있다.

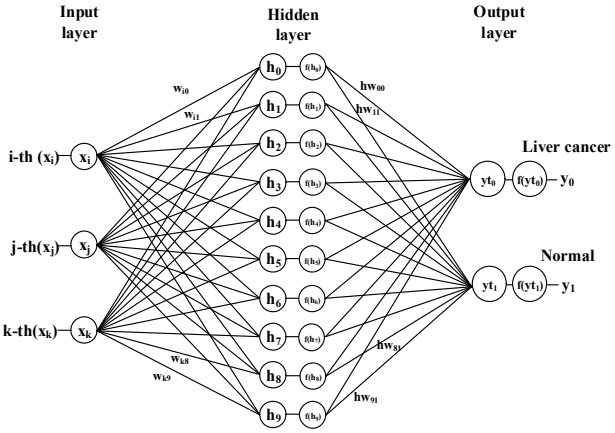

Fig. 1. Structure of used neural network.
그림 1. 사용한 뉴럴 네트워크 구조

2. 반응 물질 선정 알고리즘

본 논문에서 진단에 사용한 뉴럴 네트워크의 경우 그 구조의 특성에 기인하여 사용하는 입력의 조합에 따라 출력값 결과가 달라진다. 이것은 무조건 많은 수의 반응 물질로부터 값을 얻어 진단에 활용하는 것이 더 나은 진단 결과를 보장하지 못 하는 결과를 초래하게 된다. 이에 보다 정확한 질병 진단을 위해 사용하는 반응 물질들 중 가장 정확한 진단 결과를 얻을 수 있는 조합을 찾아내는 것이 필요하다. 이것을 위해 반응 물질을 선정하는 과정이 필요하다. 반응 물질 선정 알고리즘의 기본 개념은 뉴럴 네트워크의 학습 기능을 통해 반응 물질을 하나씩 추가하면서 해당 물질을 추가 할 경우 진단 정확도가 향상되거나 유지되는 경우에만 해당 물질을 반응 물질 조합에 포함시키고 그렇지 않은 경우에는 제외하는 방식으로 그림 2와 같다. 반응물질 선정 알고리즘 수행을 위해 여러 반응물질의 값을 추가하는 순서가 필요한데, 이것은 one-way ANOVA(Analysis of Variance)를 활용하였다.[4] 얻어진 임상 데이터에 대하여 선 처리로 one-way

ANOVA를 통해 반응 물질의 p-value를 구한 뒤, p-value 값을 기준으로 유의미한 데이터만 선별한다. 이후, 선별된 데이터를 p-value 값을 기준으로 오름차순 정렬시킨다.

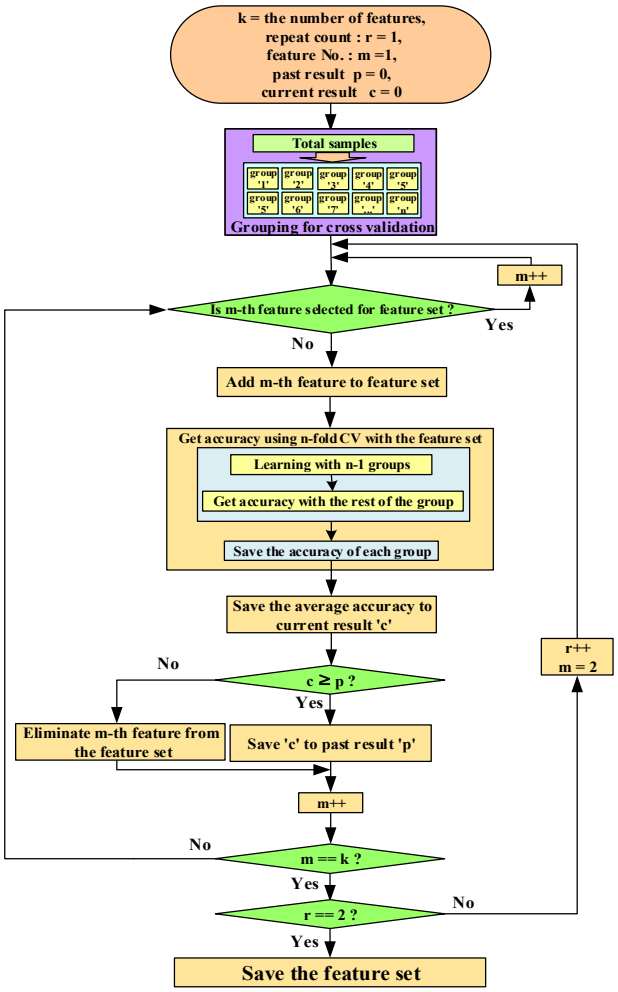

Fig. 2. Flow chart of feature selection algorithm.
그림 2. 반응물질 선정 알고리즘 순서도

여기서 사용되는 p-value는 일반적으로 생체신호 처리에 많이 활용되는 값으로 그 값이 낮을 수록 해당 질병에 대한 분별력이 높다고 할 수 있다.[5] 이렇게 정렬된 순서 정보는 그림 2를 수행하는데 있어 반응물질을 추가하는 순서로 활용된다. 반응 물질의 추가 혹은 제거의 기준이 되는 정확도를 얻기 위해 N-폴드(fold) 교차 검증이 사용된다. N-폴드 교차 검증의 경우 검증을 위해 집단을 나누는(grouping) 방식에 따라 그림 2의 알고리즘에 의해 선택되어지는 반응 물질의 종류가 달라진다. 이를 극복하고 일반화된 결과를 얻기 위해 각각 다른 경우의 수로 집단을 나누고 알고리즘을 여러 번 반복 수행한다. 이를 통해 각각의 경우에 대해 선정된 반응 물질 정보를 누적하여 저장하고, 많이 선택되어진 순으로 반응물질 정보를 재 정렬한다. 이렇게 재 정렬된 반응물질 순서를 바탕으로 최종적으로 반응 물질을 하나씩 추가해가며 정확도를 측정하고 결과값을 얻는다. 사용되는 반응물질 개수 대비 정확도의 향상 정도를 고려하여 적정 수를 찾은 후, 그때 사용한 반응물질 정보는 센서 제작에 활용하고, 학습된 뉴럴 네트워크는 휴대용 진단 플랫폼으로 이식하여 진단 알고리즘으로 사용한다.

III 고성능 컴퓨팅 활용법

2장에서 제안하는 반응 물질 선정 알고리즘은 N-폴드 교차 검증 방법을 사용한다. 이 과정에서 보다 객관적이고 신뢰할 수 있는 결과를 얻으려면 데이터 집단을 각기 다른 경우의 수로 나누어 보다 많은 교차 검증을 수행해야 하는데, 이것은 매우 많은 연산량을 필요로 하게 된다.[6] 이러한 연산을 일반전인 PC(Personal Computer)로 처리하게 된다면 천문학적인 시간을 요하게 되므로 실제로 활용하기에는 많은 어려움이 있다. 이러한 문제를 해결하기 위해 우리는 MPI(Message Passing Interface)라는 병렬 프로그래밍을 활용 하였다. 이것의 기본적인 개념은 전체 연산을 분산하여 연결된 CPU 혹은 thread로 보내어 처리 후, 다시 처리 결과를 모으는 것으로 그림 3과 같다.

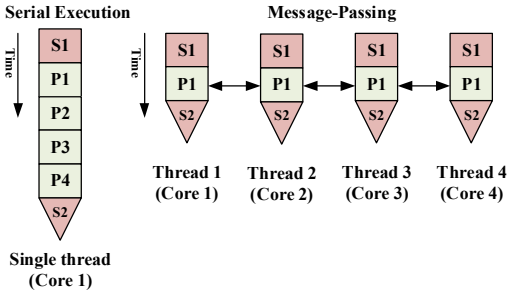

Fig. 3. Concept of MPI.
그림 3. MPI의 기본 개념

본 논문 2장에서 소개한 반응 물질 선정 알고리즘 단계 중 반응 물질 최종 선정 단계에서 MPI를 효율적으로 활용 할 수 있다. 앞서 언급하였듯이 질병 정확도를 얻기 위해서는 수많은 교차 검증과정을 수행하게 된다. 바로 이 교차 검증 과정을 분

산하여 처리하는 것이다. 교차 검증을 위한 의사 코드는 그림 4와 같다.

```
//Get diagnosis accruacy using M features

FOR(i=0;i<cross-validations;i++)
{
 Seed = Get seed value( );
 Generate random number(seed);

 Make N groups using random number( );

 FOR(j=0;j<N;J++)
 {
  Learning using N-1 Groups( );
  Diagnose using j-th Groups( );
  Sum += Save j-th accuracy;
 }
}

Diagnosis accuracy = Sum / (N * cross-validations);
```

(For N-fold cross validation)

Fig. 4. Pseudo code for determination step.
그림 4. 반응 물질 선정 알고리즘의 의사코드

```
//Get diagnosis accruacy using M features with MPI

MPI_init( );
MPI_Comm_size(...,nProcs);
MPI_Comm_rank(...,nRank);

nQuotient = cross-validations / nProcs;
nRemainder = cross-validations % nProcs;
nMySatart = nRank * nQuotient + 1;
nMyEnd = nMyStart + nQuotient – 1;

FOR(i=nMyStart;i<MnMyEnd;i++)
{
 Seed = Get seed value( );
 Generate random number(seed+nProcs);
 .
 .
 .
}

IF(nRank = Root)
{
 MPI_Rev(nProcs_Result,...);
 Final_accuracy += nProcs_Result;
}
ELSE
{
 MPI_Send(nProcs_Result,...);
}
```

(MPI setting; Divide iteration cases for each processer; For N-fold cross validation in Fig.4; Root processor: MPI result gathering; Other processor: MPI result sending)

Fig. 5. Code for determination step using MPI.
그림 5. MPI를 활용한 반응 물질 선정 알고리즘의 코드

보다 신뢰성 있는 결과를 얻기 위해 가장 중요한 것은 서로 다른 많은 경우에 대하여 교차 검증을 진행 하는 것이다. 이것을 위해 랜덤 함수를 활용하였다. 앞서 소개한 알고리즘을 병렬 처리하기 위해 그림 4에 소개한 의사코드에 MPI를 위한 몇 가지 API(Application Programming Interface)를 추가하는 것으로 구현이 가능하며, 이것은 그림 5에 나타내었다. MPI를 활용하는 경우, 랜덤 함수의 임의성을 최대한 활용하기 위해 각 thread의 ID 값을 랜덤 함수의 초기 값으로 사용 하였다.

Ⅳ. 휴대용 질병 진단 플랫폼 설계 및 구현

1. 진단 알고리즘 실험 및 검증

제안하는 알고리즘 검증을 위해 1,146개의 압타머를 반응 물질로 가지는 마이크로 어레이를 사용하였으며, 사용한 마이크로 어레이의 기본 구조 및 실제 혈액과 반응시킨 후 이미지 스캐너를 통해 얻은 이미지는 그림 6과 같다.

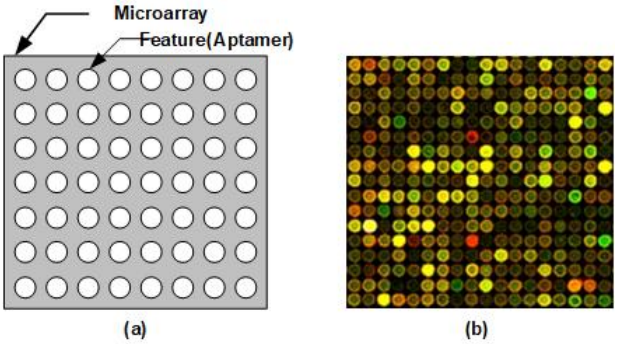

Fig. 6. (a) Structure of microarray, (b) Microarray image obtained from chip-scanner after reaction.
그림 6. (a) 마이크로 어레이의 기본 구조, (b) 반응 후 칩 스캐너를 통해 얻은 마이크로 어레이 이미지

마이크로 어레이에 사용된 압타머는 각 종류에 따라 특정 단백질과 결합하여 반응하게 되는 물질로서 환자의 혈액에 반응시킨 후, 혈액 속에 포함된 특정 물질을 감지하는데 사용된다.[7] 총 314명의 정상인과 87명의 간암 환자로부터 얻은 마이크로 어레이 임상데이터를 이용하여 2장에서 소개한 반응 물질 선정 알고리즘을 수행하였으며, 그 결과는 그림 7에 나타내었다.

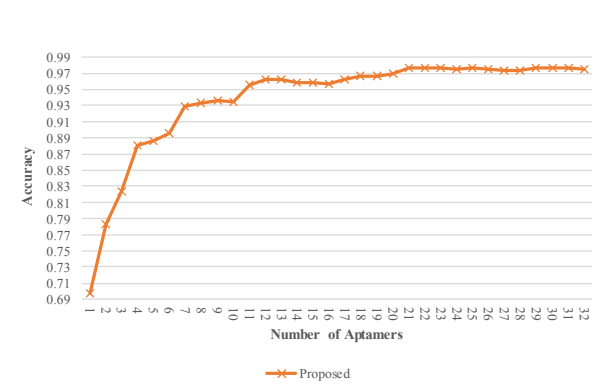

Fig. 7. Accuracy of the number of used aptamers.
그림 7. 사용된 압타머 수에 따른 진단 정확도

최종적으로 32개의 압타머를 사용할 때, 97.5%의 진단 정확도를 얻을 수 있었다. 제안하는 알고리즘의 수행 및 검증을 위해 4장에서 소개한 MPI를 활용하여 연산을 처리 하였다. 검증에 사용 된 장비는 KIST(Korea Institute of Science and Technology)에 있는 슈퍼컴퓨터로, 해당 장비는 Intel Xeon 2.6Ghz CPU-512 nodes(threads)로 구성되어 있으며 MPI 환경을 지원한다.

2. 휴대용 질병 진단 플랫폼 설계 및 구현

앞서 얻어진 실험 결과 및 휴대용 진단기기에 필요한 부가 IC(Integrated Circuit)등을 고려하여 32개의 센서로 부터 값을 받아 처리 할 수 있는 장치를 설계 및 구현하였다. 설계한 장치는 32 비트 RISC 프로세서를 기반으로 SRAM, Flash memory, UART, TFT-LCD 및 ADC 센서 보드로 구성되어 있으며 그림 8과 같다.

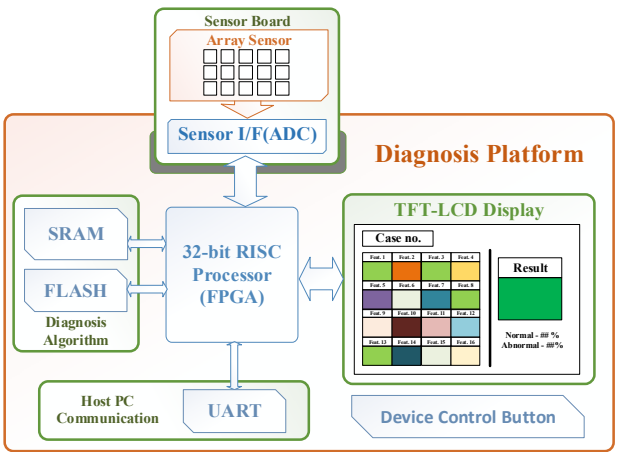

Fig. 8. Block of the designed diagnosis platform.
그림 8. 질병 진단 기기의 구성도

센서 보드는 16개의 2채널-ADC(Analog to Digital Converter) IC로 구성되어 있으며, 총 32개의 생체 신호를 입력으로 처리할 수 있다. 센서 보드에는 다양한 센서들이 연결 가능한데 본 논문에서 설계한 플랫폼에서는 박막 캔틸레버 어레이 센서를 장착 하였으며, 이를 통해 생체 신호값을 얻을 수 있다.[8] 실제 구현된 장치는 그림 9와 같다.

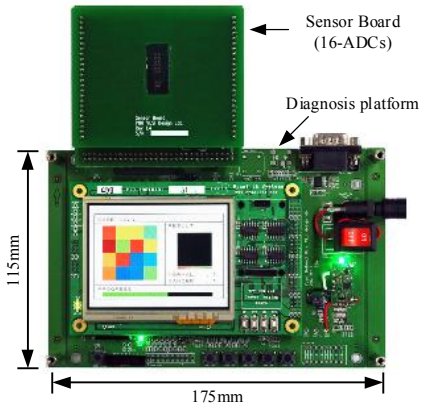

Fig. 9. Picture of the implemented diagnosis platform.
그림 9. 실제 구현된 질병 진단 기기의 모습

32개의 생체 신호를 얻은 후, 질병 진단을 위해 학습된 뉴럴 네트워크 기반의 진단 알고리즘을 수행하는 경우 필요로 하는 연산량은 표 1에 나타내었다. 실제 학습되어진 뉴럴 네트워크를 통해 진단 결과를 얻는데 필요한 연산량은 표 1과 같이 아주 적은 수준으로 저-전력 RISC 프로세서를 활용하여 처리하기에 충분하다.

Table 1. The amount of required computations using 32 features.
표 1. 32개의 입력값을 사용할 경우 필요한 연산량

	Input-Hidden	Hidden node	Hidden-Output	Output node	Total
MUL /DIV	60	20	20	4	684
ADD	0	640	20	2	662

V. 결론

본 논문에서 우리는 뉴럴 네터워크 기반의 질병 진단기기를 개발하는 방법론을 제안 하였다. 또한 이것을 위해 필요로 하는 많은 연산량은 MPI를 통

해 효율적으로 처리하는 방법을 제안하였다. 알고리즘 검증을 위해서는 1,146개의 압타머로 구성된 마이크로 어레이를 활용하여 실제 401명(정상인 314명, 간암 환자 87명)으로 구성된 표본들의 혈액과 반응시킨 생체 데이터를 활용하였으며, 검증 결과 알고리즘에 의해 선정된 32개의 반응물질을 활용하여 97.5%의 정확도로 간암 여부를 검출 할 수 있었다. 또한, 이렇게 얻어진 정보를 활용하여 32개의 센서 입력을 처리할 수 있는 휴대용 질병 진단 플랫폼을 설계 및 구현 하였다. 본 논문에서 제안하는 방법론은 여러 질병에 대하여 반응물질 선정 및 휴대용 진단 플랫폼 개발에 활용될 수 있을 것으로 기대된다.

References

[1] Cho. S. B. and Won. H. H, "Machine Learning DNA Microarray Analysis for cancer Classification," *conferences in Research and Practice in Information Technology*, pp.523-527, 2003.

[2] Leung. Y. F. and Cavalieri D, "Fundamentals of cDNA microarray data analysis," *Trends in Gnetics*, vol.19, no.11, pp.649 - 659, 2003. DOI:10.1016/j.tig.2003.09.015

[3] W. G. Baxt, "Use of an artificial neural network for data analysis in clinical decision-making: the diagnosis of acute coronary occlusion," *Neural computation*, vol.2, no.4, pp.480-489, 1990. DOI:10.1162/neco.1990.2.4.480

[4] V. Bewick, L. Cheek and J. Ball, "Statistics review 9: one-way analysis of variance," *Critical care*, vol.8, 2004. DOI:10.1186/cc2836

[5] P. Pavlidis, "Using ANVOA for gene selection from microarray studies of the nervous system," *Methods,* vol.31, pp.282-289, 2003. DOI:10.1016/S1046-2023(03)00157-9

[6] Ron. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1995, pp.1137-1143.

[7] A. D. Keefe, S. Pai and A. Ellington, "Aptamers as therapeutics," *Nature,* vol.418, pp.252-258, 2002.

[8] Seung-pyo Jung, Jun-Kyu Choi, Jung-hoon Lee, Ju-sung Park, "Design and Implementation of the Diseases Diagnosis System Using The Cantilever Micro-Arrays," *Journal of IKEEE,* vol.19, no.1, pp.52-57, 2015.

BIOGRAPHY

Sang-man Kim (Member)

2011 : BS degree in Electrical Engineering, Pusan National University.
2013 : MS degree in Electrical Engineering, Pusan National University.
2013~Present : Ph.D candidate, Pusan National University.

Ju-Sung Park (Member)

1976 : BS degree in Electrical Engineering, Pusan National University.
1978 : MS degree in Electrical Engineering, KAIST.
1989 : Ph.D degree in Electrical Engineering, University of Florida.
1978~1991 : Senior/Principal Engineer/Director, ETRI
1991~Present : Professor of Electronics Engineering, Pusan National University.