논문 2005-42CI-4-5

소속 함수와 유전자 정보의 신경망을 이용한 유전자 타입의 분류

(Classification of Gene Data Using Membership Function and Neural Network)

염 해 영*, 김 재 협*, 문 영 식**

(Jae Hyup Kim, Hae Young Yeom, and Young Shik Moon)

요 약

본 논문에서는 소속 함수와 신경망을 이용한 유전자 발현 정보의 분류 기법을 제안한다. 유전자 발현은 유전자가 mRNA와 생체의 기능을 일으키게 하는 단백질을 만들어내는 과정이다. 유전자 발현에 대한 정보는 유전자의 기능을 밝히고 유전자간의 상관 관계를 알아내는데 중요한 역할을 한다. 이러한 유전자 발현 연구를 위한 정보를 대량으로 신속하게 얻을 수 있는 도구가 DNA 칩이다. DNA 칩으로 얻은 수백~수천개의 데이터는 그 데이터만으로는 의미를 갖지 못한다. 따라서 유전자 발현 정도에 따라 수치적으로 획득된 데이터에서 의미적인 특성을 찾아내기 위해서는 클러스터링 방법이 필요하다. 본 논문에서는 수많은 유전자 데이터 중에서 주요 정보를 포함한 것으로 판단되는 유전자 데이터를 피셔 기준에 의하여 선택한다. 이때 선택된 데이터들이 클러스터링에 효과적인 데이터라고 보장할 수 없으므로, 클러스터링 성능을 저해하는 유전자 데이터의 영향력을 감소시키기 위해서 소속 함수를 이용하여 특징값을 계산하고, 계산된 특징값으로 얻은 특징 벡터들을 적용하여 역전파 신경망 학습을 수행한다. 본 논문에서 제안한 유전자 발현 정보의 분류 결과로 얻은 클러스터링의 성능은 기존의 연구 결과와 비교했을 때 다양한 유전자 데이터에 대하여 향상된 인식율을 보이는 것을 확인할 수 있었다.

Abstract

This paper proposes a classification method for gene expression data, using membership function and neural network. The gene expression is a process to produce mRNA and protains which generate a living body, and the gene expression data is important to find out the functions and correlations of genes. Such gene expression data can be obtained from DNA 칩 massively and quickly. However, thousands of gene expression data may not be useful until it is well organized. Therefore a classification method is necessary to find the characteristics of gene data acquired from the gene expression. In the proposed method, a set of gene data is extracted according to the fisher's criterion, because we assume that selected gene data is the well-classified data sample. However, the selected gene data does not guarantee well-classified data sample and we calculate feature values using membership function to reduce the influence of outliers in gene data. Feature vectors estimated from the selected feature values are used to train back propagation neural network. The experimental results show that the clustering performance of the proposed method has been improved compared to other existing methods in various gene expression data.

Keywords: gene expression data, membership function, neural network

I. 서 론

인체는 다양한 세포로 구성되어 있으며, 각각의 세포는 단일 세포에서 성장 분화된 것이지만 동일한 게놈(genome)에서 만들어지는 메시지(RNA)의 차이에 의해

* 학생회원, ** 정회원, 한양대학교 컴퓨터공학과
(Dept. of Computer Science and Eng., Hanyang Univ.)

※ 본 연구는 대학 IT연구센터 육성지원사업의 연구결과로 수행되었음.
접수일자: 2005년1월20일, 수정완료일: 2005년6월27일

세포의 형태와 기능이 달라진다. 이는 동일한 세포가 인체내의 미세 환경(microenvironment)에 의하여 유전1)자의 발현(gene expression)이 조절되어 장기를 형성하는 세포, 혹은 골격을 만드는 세포등으로 분화(differentiation) 하기 때문이다. 이와 같은 세포 특성 때문에 유전자 발현을 조절하는 물질이 외부 환경에서 인체 내로 유입하게되면 유전자의 발현을 변화시켜 세포 특성을 변화시키고, 세포 본래의 역할에서 탈피함으로 정상적인 활동을 수행할 수 없게 되거나, 정상 세포의 활동을 억제하게 되어 질병을 유발시키게 되는데 이러한 대표적인 사례를 "암"이라 할 수 있다. 즉, 암세포는 환경에 있는 유독한 환경 오염 물질에 의하여 인체 게놈이 파괴되고 정상적인 메시지를 생산할 수 없게 됨으로 정상 세포에서 암세포로 변화되는 것이다. 이러한 유전자 발현에 대한 정보는 유전자의 기능을 밝히고 또한 유전자간의 상관 관계를 알아내는데 매우 중요한 역할을 한다. 또한 이 정보는 병의 원인이 되는 유전자를 찾아내어 병의 정확한 진단 및 조기 진단을 가능케 하며 유전자 치료 및 신약 개발의 중요한 목표를 제공하게 된다.

게놈 프로젝트로 대표되는 유전자 암호 해독 기술(DNA sequencing)의 발달과 유전자 염기 서열 해석은 생체 내 많은 유전자의 역할을 효율적으로 해석하기 위한 새로운 실험 기법이 필요하게 되었다. 이에 정보를 대량으로 신속하게 얻을 수 있는 도구인 DNA 칩이 개발되었고, 유전자 발현 연구를 위한 핵심적인 도구로 사용되고 있다[1]. DNA 칩은 기존의 분자 생물학적 지식과 기계 및 전자공학의 기술을 접목하여 만들어진 것으로 기계 자동화와 전자 제어 기술을 이용하여 만들어졌다. 이러한 DNA 칩은 slide glass상에 수백~수천개의 아주 적은 양의 유전물질인 DNA 유전자를 PCR(polymerase chain reaction) 증폭하여 정렬·고정화해 놓은 것이다. 이렇게 구성한 DNA 칩에 우리가 해석하고자하는 세포에서 추출한 mRNA를 역전사(reverse transcription)시킬 때, 각각 다른 색깔의 형광물질을 띤 염기를 집어넣어 빨간색(Cy5)이나 녹색(Cy3)을 띤 형광표식 cDNA로 합성한다. 이 두 cDNA를 똑같은 양으로 섞어서 DNA 칩에 심어놓은 DNA 유전자와 결합(hybridization, 혼성화반응)하여 각 유전자의 발현 변화를 측정한다. 이렇게 DNA 칩으로 얻어진 수많은 데이터들은 생물학자들이 연구하기에 의미있는 정보로 조직화하고 분석하는 과정이 반드시 필요하다. 따라서 DNA 칩 실험 데이터에 대한 효율적인 클러스터링 알고리즘 개발은 유전자의 기능 분석(functional genomics)과 유전자의 상호 관련성 분석(genomic networks)등의 분야 연구에 크게 기여하고 있다[1].

클러스터 해석은 방대한 데이터를 어떤 기준에 근거하여 그룹화하는 것으로, 주어진 데이터를 의미있는 집단들로 분류하며, 데이터 분석, 시각화, 압축 및 전처리와 관련된 많은 분야에서 널리 응용된다. 이러한 DNA 칩에서 얻은 데이터를 분석하는데 계층적인 클러스터 해석 알고리즘으로 PCA plot, scatter plot등이 사용되고, 비계층적인 클러스터 해석 알고리즘으로 K-means와 자기 조직망등의 알고리즘이 이용되고 있다[2]. 하지만 DNA 칩의 실험시 발생할 수 있는 여러 가지 에러를 포함한 데이터에 대한 분석은 모호하고 불명확한 결과를 초래하여 데이터에 대한 패턴 인식을 적용한 분류를 어렵게 만든다. 그러므로 수백~수천개의 유전자 데이터 전체를 클러스터링하는 것보다 발현 정도가 매우 유사한 일부분의 데이터를 추출하여 클러스터 알고리즘을 적용하는 것이 데이터를 분석하는 데에 좀 더 효과적이다.

그러므로 본 논문에서 제안하는 알고리즘은 수백~수천개의 유전자 데이터를 모두 적용하여 클러스터링하는 것이 시간적으로나 공간적으로 불가능하므로 두 클러스터 서로간의 각각의 분산값의 비율이 작고 전체 유전자 데이터의 분산값이 큰 유전자 데이터들을 추출한다[3,4].

하지만 선택된 유전자 데이터의 분포가 클러스터링 성능에 악영향을 미칠 수 있는 불명확한 데이터일 가능성이 있으므로 이러한 유전자 데이터의 영향을 최소화하고, 나머지 다른 유전자 데이터들의 분류 영향력을 향상시킬 방법이 필요하다.

이에 소속 함수를 이용하여 해당 클러스터의 특징값을 계산하고, 그 차이를 비교한다. 클래스 특징 벡터로 정의한 차이값에 따라 클러스터링 결과를 향상시킬 수 있는 가중치 계산 과정이 필요한데, 이는 신경망을 이용하여 클러스터링이 최적화 될 수 있도록 학습시킨다. 이로써 정확한 클러스터링이 가능하도록 하여, 유전자 발현 데이터의 의미적인 특성들을 파악 할 수 있는 알고리즘을 제안하고자 한다.

본 논문의 구성은, 제 Ⅱ장에서 DNA 칩을 비롯한 관련 연구를 소개하고, 제 Ⅲ장에서 제안하고자 하는 알고리즘을 설명하고, 제 Ⅳ장에서는 실험결과 및 평가를 하며, 마지막 제 Ⅴ장에서는 결론 및 향후과제를 제시하는 것으로 구성하였다.

Ⅱ. 관련 연구

1. DNA 칩

DNA 칩은 기존의 분자 생물학적 지식을 바탕으로 현대에 엄청난 발전을 한 기계 및 전자 공학의 기술을 접목해서 만들어졌다. 이는 기계 자동화와 전자 제어 기술들을 이용하여 적게는 수백 개부터 많게는 수십만 개의 DNA를 아주 작은 공간에 집어넣을 수 있게 만든 것으로 유전자 발현 양상, 유전자 결함 그리고 단백질 분포들을 분석해 낼 수 있는 생물학적 마이크로칩(Biological Micro칩)이라 할 수 있다[1].

이러한 DNA 칩은 유전자 검색용으로써 엄청나게 많은 종류의 DNA를 고밀도로 붙여 놓은 것으로, 이를 대체 할 수 있는 기존의 대표적인 유전공학 방법으로는 Southern과 Northern blot, 돌연변이 검색 그리고 DNA sequencing등이 있다. 이들과 DNA 칩의 가장 큰 차이점은 DNA 칩이 동시에 최소한 수백 개 이상의 유전자를 빠른 시간 안에 검색할 수 있다. 또 하나의 다른 점은 Southern이나 Northern blot의 경우 유전물질을 붙이는 매체로 nitrocellulose막을 사용하는데 반하여 DNA 칩에서는 유리와 같은 고형체를 사용한다. 이러한 차이에 의해 DNA 칩은 아주 적은 양의 유전 물질을 고밀도로 붙일 수 있게 되었고, 동시에 많은 수를 검색할 수 있다. 이는 과학 기술 연구 및 임상, 진단, 검사 등의 분야에 혁신적 변화를 일으킬 것으로 주목받고 있다.

본 논문에서는 단 한번의 실험으로 빠르고 정확하게 수천 개 이상의 유전자 발현 변이를 검색할 수 있는 DNA 칩 데이터를 이용한다. 그림 1은 DNA 칩을 이용해 데이터를 획득하는 실험 과정을 보여준다.

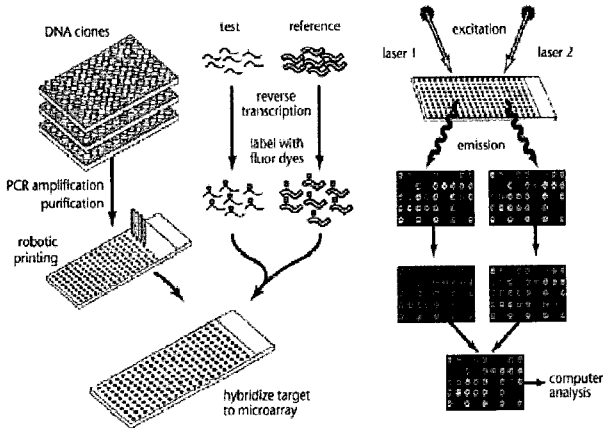

그림 1. cDNA microarray 실험과정
Fig. 1. Steps for cDNA microarray generating.

2. 클러스터링

클러스터링은 고차원의 데이터를 시각화하거나 압축하는 방법들 중 하나로써, 전체 집단에 대해 각각의 특징을 가지는 몇 개의 부집단으로 분할하는 기법으로 데이터 분석과 패턴 인식을 통해 내재된 클러스터 구조를 추출하려는 기술이다. 클러스터링 알고리즘은 각 데이터의 유사성에 근거하여 데이터들을 그룹화하며 같은 그룹 내의 데이터는 높은 유사도를 갖도록 다른 그룹내의 데이터와는 낮은 유사도를 갖도록 군집화하는 것이다. 이러한 클러스터링 방법은 크게 계층적 클러스터링, 분할 클러스터링, 중복 클러스터링으로 나눌 수 있다.

계층적 클러스터링은 통계학에서 사용되어온 고전적이며 일반적인 클러스터링 알고리즘으로 클러스터들이 더 작은 클러스터로 이루어진 하부 구조를 가지도록 하는 방법이다.

분할 클러스터링은 각 클러스터간에 중복이 없으며, 각 개체를 가장 가까운 클러스터에 할당하는 과정을 반복하며 가장 적합한 클러스터를 구성한다. 이러한 방법은 전체 입력 패턴을 하나의 클러스터로 생각하고 클러스터링 작업을 시작해서 점점 더 많은 클러스터로 분할하는 방법이다.

중복 클러스터링은 클러스터간의 계층적 구조를 갖지 않으며, 클러스터간의 중복을 허락하여 각 개체를 가장 가까운 클러스터로 근접시키는 과정을 반복해서 가장 적합한 클러스터를 구성한다. 즉, 분할 클러스터링이 하나의 개체가 가장 가까운 하나의 클러스터로 할당되는데 반해 중복 클러스터링에서는 하나의 개체가 여러 클러스터에 소속될 수 있으며 근접 정도에 따라 소속 정도를 차등을 두는 방법에 의해 할당된다. 이에 본 논문에서는 신경망을 이용하여 가중치를 적용하므로 소속 정도를 부여하고 학습 결과를 얻고자 한다.

본 논문에서 사용하는 신경망 기술은 불완전하고 잡음이 많은 입력의 해석뿐만 아니라 패턴인식(pattern recognition), 학습, 분류, 일반화, 추상화 등을 위한 활용성에 있다. 신경망은 논리적이고 분석적인 기법을 활용해서도 시뮬레이션하기 어려운 인간의 문제 해결을 지원한다. 이러한 신경망은 규칙이 알려지지 않은 상황에서의 패턴과 특성을 발견하기 위해 대량의 데이터를 분석할 수 있다. 또한 주어진 예제 패턴의 반복 학습을 통해 스스로 지식을 획득하는 특성을 가지므로 이러한 특징을 이용하여 본 논문에서 신경망의 특성을 이용하여 수많은 유전자 데이터를 분류한다.

본 논문에서는, 기존에 제안된 다양한 특징 추출 방

식과 분류 알고리즘을 비교 분석 하였으며, 각각의 알고리즘의 다양한 유전자 데이터에 대한 분류 성능을 비교 분석하였다. 따라서, 다양한 특징 추출 방식과 분류 알고리즘간의 관계를 파악하고, 성격이 다른 다양한 유전자 데이터를 대상으로 어느정도의 일반적인 분류 성능을 나타내는 지를 파악할 수 있다. 실험에 사용된 특징 추출 방식과 분류 알고리즘은 아래와 같다.

가. 특징 추출

(1) 상관 관계법

상관 관계 분석을 이용하면 두 변수간의 선형적 관련성 정도와 관련 방향을 알 수 있다. 상관 관계 분석에 의해 얻어지는 상관계수는 -1과 +1 사이의 값을 가진다. 즉, 학습 데이터의 각 유전자의 발현 수치와 각 샘플이 속하는 클래스를 반영한 클래스 패턴과의 항관 관계를 계산하여 특징 추출의 척도록 사용한다. 사용된 상관 관계법은 피어슨 상관관계(Pearson's Correlation)와 스피어맨 상관관계(Spearman's Correlation)이다[2,5].

(2) 유사도 측정법

두 개의 입력 벡터 X와 Y 사이의 유사성은 이들 변수간의 거리로 볼 수 있으며, 군집 분석에서는 이들 개념들에 의해 클래스들을 집단화할 수 있다. 여기서는 거리 지수로 유클리디언(Euclidean) 거리방법과 코사인(Cosine) 계수방법을 사용하였다.

(3) 정보 이론

전체 데이터로부터 의미있는 정보를 뽑아내는 척도로 정보이론에서 사용하는 정보 이득(IG, Information Gain), 상호 정보(MI, Mutual Information), 신호 대 잡음 비(Signal to nose ration) 등의 방법이 있다[6]. 이를 이용하여 특정 유전자의 샘플이 특정 클래스에 속하는지의 여부와 해당 유전자의 발현 여부를 이용하여 계수를 계산함으로써, 특징 추출의 척도로 삼는다.

(4) 피셔 선택법

클래스내의 분산과 클래스간의 분산의 비율을 고려하여 클래스 내에서 응집도가 높고 클래스 간의 분리도가 높은 공간을 선택하는 방법으로, 유전자마다 샘플간의 분리도와 샘플내 유전자 간의 응집도를 고려하여 잘 분리된 유전자를 선택하는 척도로 삼는다.

나. 분류 알고리즘

(1) 역전파 신경망

연전파 신경망(BP, Backpropagation) 신경망은 매우 다양한 분야에 활용되고 있는 순방향 다층 신경망이다. 이것은 입력 벡터가 나열되어있는 예제를 통해 반복적으로 출력을 각각의 입력에 따라 수정해가며 학습한다. 각각의 학습 입력을 거치는 것이 하나의 주기로 하여 각 주기동안 신경망은 실제의 결과를 가지고 목적화하는 결과와 비교해서 오류를 산출한 뒤, 그 오류를 최소화하기 위해 가중치 값을 조절한다. 지도 학습이라 불리는 이 과정을 통해, 신경망은 올바른 결과와 입력 패턴을 연관시키는 것을 학습한다.

(2) 의사결정 트리

의사결정 트리(DT, Decision Tree)는 여러 단계의 복잡한 조건을 갖는 문제에서 각 조건과 그에 따른 해결방안을 트리 형태로 나타낸 것을 말한다[3, 7]. 가장 큰 조건이 트리의 뿌리를 만들고, 세부 조건이 트리의 각 가지를 만들며, 해결 방안은 트리의 잎(leaf) 노드로 나타나게 된다.

(3) 구조적응 자기구성 지도

구조적응 자기구성 지도(SASOM, Structure Adaptive Self-organazing Map)는 일반적인 자기 구성 지도의 구조가 초기에 결정되어 학습이 끝날 때까지 변하지 않는다는 단점을 보완하기 위해 제안되었다[8,9]. 즉, 기존의 SOM 알고리즘을 이용하여 지도를 학습시킨 후, 학습된 지도의 노드들 중 서로 다른 클래스의 데이터가 섞여있는 노드를 반복적으로 분화하여 주어진 데이터에 대하여 최적의 위상을 갖는 지도를 생성한다.

(4) k-최근접 이웃

k-최근접 이웃(kNN, k-Nearest Neighbor) 기법은 기억 기반 추론의 가장 일반적인 방법으로 새로운 데이터에 대한 각각의 모든 기존 데이터와의 유클리디안 거리를 계산한 후, 가장 가까이에 있는 k개의 데이터에 기반하여 새로운 데이터의 부류를 결정한다[10].

Ⅲ. 제안하는 알고리즘

DNA 칩으로 얻은 수백~수천개의 유전자 데이터를 모두 적용하여 클러스터링하는 것은 시간적으로나 공간적으로 불가능하다. 그러므로 클러스터링하기에 적절한 유전자 데이터의 선택 과정이 필요하다.

제안하는 알고리즘으로 클래스 각각의 분산값의 비율이 작고 전체 유전자 데이터의 분산값이 큰 유전자 데이터를 선택하는 방법을 이용한다. 그러나 선택된 유전자 데이터가 클러스터링 성능에 효과적이지 못한 악영향을 주거나 클러스터링을 저해하는 데이터일 수 있다. 이러한 유전자 데이터의 영향을 최소화하고 다른

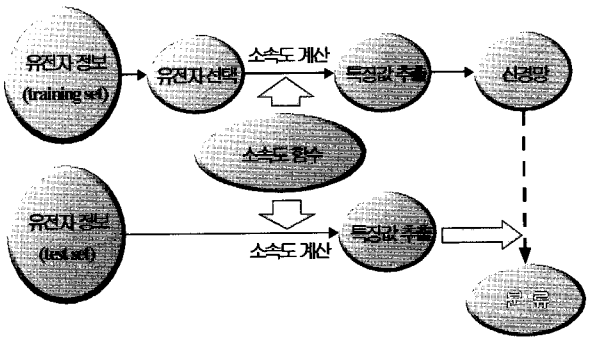

그림 2. 제안하는 알고리즘 개요
Fig. 2. Diagram of proposed algorithm.

유전자 데이터들의 분류 영향력을 향상시킬 수 있는 방법이 필요하다. 이에 각각의 선택된 유전자 데이터를 소속 함수에 대입하여 특징값을 추출한다. 이에 획득된 특징값의 차이를 비교한 샘플 특징 벡터와 클래스 특징 벡터를 이용하여 클러스터링 영향력에 따른 가중치를 고려하여 클러스터링을 수행한다.

이때, 가중치 계산은 역전파 신경망을 이용하여 최적의 클러스터링 성능을 보장할 수 있도록 학습 과정을 거친다. 이러한 알고리즘의 전체 구조를 그림 2에서 보여주고 있다.

1. 유전자 데이터 선택

DNA 칩에서 얻은 수많은 유전자 데이터를 모두 이용하여 클러스터링하는 것은 불가능한 일이다. 또한 얻어진 데이터에서 각 샘플의 특정 클래스와 연관이 있는 유전자의 수는 훨씬 적다. 따라서 유전자 데이터를 이용하여 클래스를 분류하기 위해서는 클래스와의 연관성이 높은 유전자를 추출하는 과정이 필요하다. 이는 엄청나게 많은 데이터 중에서 클러스터링하기에 적당한 데이터를 선택해야하는 이유이다[3]. 유전자와 적은 수의 샘플을 포함하는 데이터로 많은 수의 유전자를 이용하여 적은 수의 샘플을 분류하는 것은 정확한 분류를 위하여 적절하지 않다. 분류에 이용하기 적절한 유전자를 선택하는 과정이 필요한데, 주로 상관 관계와 클러스터링 기법 등을 이용하여 클래스와의 상관 관계가 높은 유전자를 선택한다. 이러한 의미있는 데이터를 선택하는 방법으로 기존의 연구에선 피어슨 상관계수, 스피어맨 상관계수, 유클리드 거리, 코사인계수, 정보이론(정보 이득, 상호 정보, 신호 대 잡음비), 유사도 측정법을 이용하고 있다. 이는 변수간의 관련성과 유사성을 고려한 선택 방법이다.

본 논문에서는 유전자 데이터들의 분포를 고려한 선택 방법을 사용하고 있다. 분산이란 데이터의 흩어진 정도를 말하는데, 흩어진 정도가 크다면 분산이 큰 값을 갖고 클러스터링하기엔 적합하지 않은 결과를 보여준다. 그러므로 식 1과 같이 분산의 비율을 계산하여 클러스터링의 성능을 높일 수 있다.

$$분산 = \frac{전체데이터의분산}{클러스터1의분산 + 클러스터2의분산} \quad (1)$$

그러나 식 1을 이용하여 선택된 유전자 데이터일지라도 클러스터링하기에 부적합하거나 클러스터링을 저해하는 유전자 데이터가 존재할 수 있으므로 이러한 유전자 데이터의 영향을 최소화하고 나머지 유전자 데이터의 클러스터링 성능을 향상시킬 수 있는 방법인 특징값을 추출한다.

2. 특징값 추출

가. 소속 함수 설계

소속 함수는 선택된 데이터가 서로 다른 클래스에 속하는 정도를 수치적으로 나타냄으로, 모호한 데이터들이 나타내는 서로 유사한 데이터 분포를 데이터들의 소속 정도를 이용하여 클러스터링하기에 적합한 유전자 데이터들의 영향력을 높여주는 방법이다. 이에 그림 3은 소속 함수 계산 과정을 도표로 보여주고 있다.

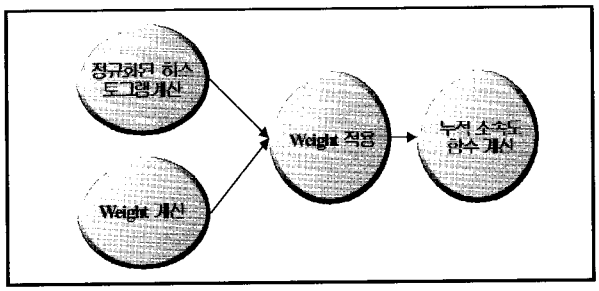

그림 3. 소속 함수 계산 과정
Fig. 3. Steps for membership function.

(1) 정규화된 히스토그램 계산

선택된 유전자가 존재하는 구간을 N(정규화된 히스토그램 레벨)으로 나누어 데이터의 분포를 반영하는 히스토그램을 식 2와 같이 계산한다. S는 선택된 유전자에 해당하는 데이터이고, i는 두 개의 클래스, j는 선택된 유전자 데이터이고, m은 선택한 유전자 데이터의 개수이다. k는 정규화하는 레벨 N의 크기를 의미한다.

$$H_j^i(k) = histogram\ (S_j^i) \quad \begin{matrix} 1 \le i \le 2 \\ 1 \le j \le m \\ 1 \le k \le N \end{matrix} \quad (2)$$

(2) 평균위치를 기준으로 거리를 고려한 가중치 계산

N레벨로 정규화된 히스토그램을 기반으로 클래스 각각에 평균위치 C를 기준으로 식 3을 적용하여 거리 관계를 계산한 가중치를 정의한다. 이는 평균위치인 C로부터 멀수록 데이터의 분산값이 큼을 의미한다.

$$W_j^i(k)=\begin{cases}1/\alpha\left\{\left(\dfrac{C_j^i-k}{C_j^i-1}\right)^2+1\right\} & k\le C_j^i\\ 1/\alpha\left\{\left(\dfrac{C_j^i-k}{C_j^i-N}\right)^2+1\right\} & k\le C_j^i\end{cases} \tag{3}$$

여기서, C_j^i는 i 클래스의 j번째 유전자 데이터의 평균을 의미한다.

(3) 가중치 적용

정규화된 히스토그램의 데이터를 평균위치를 고려하여 계산한 가중치로 나누면 평균위치 C로부터 멀리 떨어질수록 히스토그램은 낮은 가중치를 갖는다. 이는 선택된 유전자 데이터가 평균위치로부터 멀리 떨어진 클러스터링하기에 모호한 데이터들의 소속 정도를 낮춰주고, 평균 위치에 가까운 데이터들은 높은 가중치를 부여함으로 선택된 유전자 데이터의 소속 정도를 가늠할 수 있도록 한다.

$$H_j^{i'}(k)=\frac{H_j^i(k)}{W_j^i(k)}\quad \begin{matrix}1\le i\le 2\\ 1\le j\le m\\ 1\le k\le N\end{matrix} \tag{4}$$

(4) 누적소속 함수 계산

누적소속 함수는 식 5로 주어지며, 가중치를 적용하여 얻어진 데이터들의 주변 데이터들과의 관계를 고려한 계산 결과이다.

$$F_j^i(k)=\begin{cases}H_j^{i'}(k)+H_j^{i'}(k-1) & 2\le k\le C_j^i\\ H_j^{i'}(k)+H_j^{i'}(k+1) & C_j^i\le k\le N\end{cases} \tag{5}$$

이러한 소속 함수를 이용하여 선택된 데이터를 적용한 특징값 벡터를 계산한다.

나. 샘플의 특징 벡터

선택된 유전자 데이터를 소속 함수 F^1과 F^2 각각에 대입한 차이를 계산한 샘플의 특징 벡터 SV를 식 6과 같이 정의한다. i,j,k는 앞서 정의한 변수들을 의미하고 선택된 유전자를 가진 조직의 개수를 $p+q$로 나타낸다. 이는 클래스 각각에 선택된 유전자 개수 m만큼의 샘플의 특징 벡터를 얻고, 특징 벡터는 서로 다른 클래스에 p개, q개의 조직을 갖음을 의미한다.

$$SV_j^i(t)=F_j^1(S_j^i(t))-F_j^2(S_j^i(t))$$
$$1\le i\le 2\ \ 1\le k\le N\ \ 1\le j\le m\ \ 1\le t\le p+q \tag{6}$$

다. 클래스의 특징 벡터

각각의 클래스에 존재하는 유전자 데이터 개수 만큼의 샘플의 특징 벡터(SV)는 p개, q개의 조직으로 구성된다. 선택된 유전자에 해당하는 샘플 특징 벡터들의 값들의 평균은 다음 식으로 구성한다. 이를 각각의 클래스에 선택한 유전자 개수를 가진 클래스의 특징 벡터 (CV)라 한다.

$$CV_j^i=\begin{cases}\dfrac{\sum_{t=1}^{p}SV_j^i(t)}{p} & i=1\\ \dfrac{\sum_{t=p+1}^{p+q}SV_j^i(t)}{q} & j=1\end{cases} \tag{7}$$

이는 클래스 특징 벡터를 포함하는 유전자가 선택된 유전자들에 의해 두 클래스 특징 벡터에 각각의 유전자의 차이를 계산하면 두 분포가 클러스터링하기에 적합한 분포를 가진 유전자인지 아닌지를 판별할 수 있다.

선택된 유전자 데이터가 클러스터링하기에 모호한 경우의 데이터 영향력을 감소시키기 위한 방법으로 이용되었던 소속 함수를 적용한 특징값 추출 방법이 클러스터링하기에 효율적이라고 말하기에는 아직 부족하다. 이는 한 번에 수많은 유전자 데이터를 얻을 수 있고, 그 분포가 일정한 패턴을 가지고 있지 않으므로 불규칙한 데이터를 이용하여 클러스터링하는 것이 쉽지 않기 때문이다.

이로써 계산된 소속 함수의 결과가 선택된 유전자 데이터를 클러스터링하기에 적절하지 않을 수 있으므로 데이터들의 클러스터링을 향상시킬 수 있는 가중치 계산이 필요하다. 이러한 가중치 계산은 클러스터링 방법 중에 신경망 학습을 적용한다.

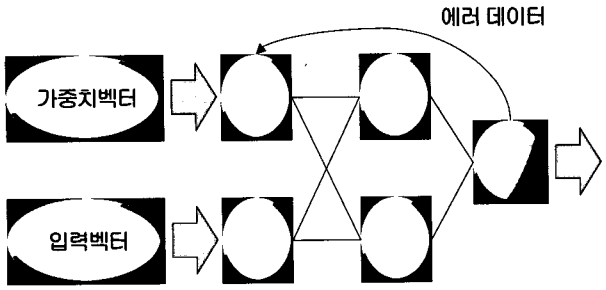

그림 4. 제안한 알고리즘의 신경망 구조
Fig. 4. Structure of neural networks.

3. 클러스터링

소속 함수를 이용하여 특징값을 추출하여 얻은 결과가 선택된 유전자 데이터의 클러스터링 성능을 향상 시킬 수 있는 데이터 집합일 수도 있지만 클러스터링을 저해하는 구분이 모호한 선택된 유전자 데이터 집합일 수도 있다. 이에 클러스터링하기에 적절하지 않은 데이터들의 클러스터링 결과를 향상시킬 수 있는 가중치 계산을 이용하기 위해 역전파 신경망을 사용한다. 본 논문에서 사용한 역전파 신경망은 그림 5와 같이 구성하였다. 앞서 소개한 소속 함수를 이용하여 계산한 클래스 특징 벡터의 차이값을 가중치 벡터의 입력값으로 대입하고, 유전자 데이터의 소속 함수의 벡터값을 대입하여 신경망 학습을 수행한다. 이는 선택된 유전자 데이터에 따른 소속 정도에 따라 가중치를 부여하므로 클러스터링이 가능하도록 한다.

Ⅳ. 실험 결과

본 논문에서는 7129개의 유전자로 구성된 72개의 백혈병 샘플, 2000개의 유전자로 이루어진 62개의 대장암 샘플, 그리고 4026개의 유전자로 구성된 69개의 림프종 샘플을 실험에 사용하였다. 실험에서는 피어슨 상관관계, 스피어맨 상관관계, Euclidean거리, Cosine 계수, MI(상호정보), IG(정보 이득), SN(신호대 잡음비), 피셔 계수를 이용한 특징 선택 방법과 역전파 신경망, 의사결정 트리, 구조적응 자기구성 지도, k-최근접 이웃 분류 알고리즘을 사용하여, 각각의 특징 선택 방법과 다양한 분류 알고리즘에 따른 성능을 비교하고 다양한 종류의 유전자 데이터에 따른 성능을 제안하는 알고리즘의 성능과 비교하였다. 이는 특징 선택 방법과 분류 방법간의 상호 관계를 이해하고, 선택 가능한 기존의 알고리즘들과 제안하는 알고리즘이 유전자 데이터의 종류에 따라 어느 정도의 적응성을 가질 수 있는지를 비교 분석 하는데 적합하다.

각각의 유전자 데이터에 대한 실험은, 주어진 특징 선택 방법에 의해 순차적으로 100개 까지의 유전자를 선택한 후, 선택된 유전자를 대상으로 분류 알고리즘을 적용하였으며, 이에 따른 분류 성능중 가장 우수한 분류 성능을 비교 분석하였다.

또한, 각각의 유전자 데이터들의 샘플에 대해서 1개의 샘플부터 전체 샘플의 50% 에 대당하는 샘플까지 순차적으로 제외한 후 나머지 샘플을 학습에 이용하고, 제외된 샘플을 통해 분류 성능을 실험하였으며, 여기서 구해진 분류 성능의 평균치를 해당 알고리즘의 분류 성능으로 사용하였다.

1. 백혈병 데이터의 분류

표 1에서는 제안하는 알고리즘에서 소속 함수의 퍼지 레벨과 선택된 유전자 데이터의 개수에 대한 백혈병 데이터의 분류 성능을 보여주고 있다.

표 2는 제안하는 알고리즘과 비교 알고리즘들을 백혈병 유전자 데이터[3,7,11]를 대상으로 분류 성능을 비교한 결과이다.

실험 결과에 의하여, 제안된 알고리즘은 백혈병 유전자 데이터의 분류에 있어서, 소속 함수의 퍼지 레벨이 5이고, 선택된 유전자 개수가 10개일때의 분류 성능이 가장 우수함을 알 수 있다. 동일한 유전자 데이터를 이용하여 비교 알고리즘들과 성능을 비교해 보면, 분류 성능의 차이가 큰 경우 30% 가량 차이가 나는 경우를 볼 수 있으나, 피어슨 계수와 역전파 신경망을 이용했을 경우와 같이 분류 성능의 차이가 거의 없는 결과를 볼 수 있었다. 그림 5에서는 제안하는 방법과 다른 분

표 1. 소속 함수의 퍼지 레벨과 선택된 유전자 개수에 대한 성능 비교
Table 1. Classification rates on fuzzy levels and number of genes.

		퍼지 레벨				
		5	10	30	50	100
유전자 개수	5	88.7%	85.5%	85.4%	84.3%	85.3%
	10	97.2%	95.4%	92.3%	95.4%	90.8%
	30	93.5%	94.3%	93.5%	92.3%	92.3%
	50	90.8%	92.3%	95.4%	95.4%	95.4%
	100	85.3%	88.7%	90.8%	89.2%	81.8%

표 2. 백혈병 데이터(ALL/AML)의 분류 성능 비교
Table 2. Classification rates of leukemia(ALL/AML) gene data.

	BP	DT	SASOM	kNN
Pearson	97.1%	97.1%	88.2%	25.4%
Spearman	70.6%	82.4%	82.4%	44.0%
Uclidean	97.1%	91.2%	82.4%	44.0%
Cosine	94.1%	83.5%	70.6%	35.7%
IG	88.2%	47.1%	64.7%	58.5%
MI	67.6%	55.9%	64.7%	58.8%
SNR	94.1%	91.2%	94.1%	20.0%
제안하는 알고리즘	97.2%			

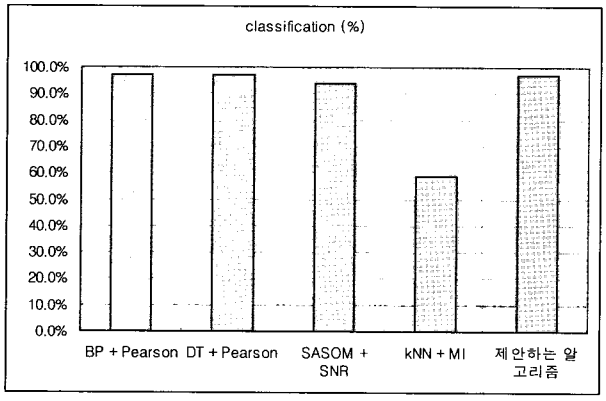

그림 5. 백혈병 유전자 데이터의 성능 비교
Fig. 5. Classification rates of leukemiagene data.

표 3. 대장암 데이터의 분류 성능 비교
Table 3. Classification rates of colon gene data.

	BP	DT	SASOM	kNN
Pearson	65.7%	70.0%	65.7%	45.3%
Spearman	72.0%	62.8%	50.4%	48.0%
Uclidean	58.3%	72.0%	60.4%	44.0%
Cosine	57.5%	50.5%	64.6%	35.7%
IG	76.3%	57.5%	68.3%	58.5%
MI	50.3%	65.7%	66.7%	52.8%
SNR	78.7%	50.8%	72.1%	44.0%
제안하는 알고리즘	80.2%			

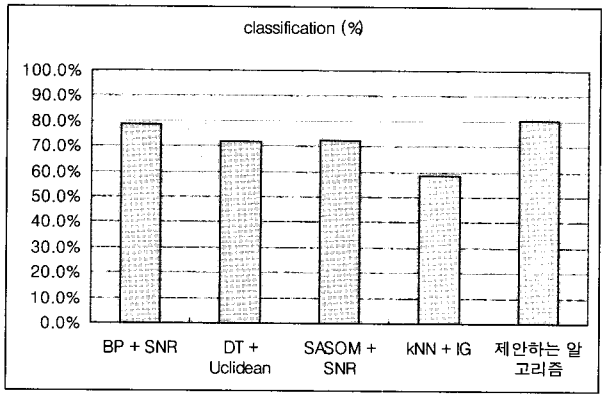

그림 6. 대장암 데이터의 성능 비교
Fig. 6. Classification rates of colon gene data.

류 알고리즘간의 가장 우수한 분류 성능을 보여주고 있다.

2. 대장암 데이터의 분류

표 3은 제안하는 알고리즘과 비교 알고리즘들을 대장암 유전자 데이터[3]를 대상으로 분류 성능을 비교한 결과이다. 제안하는 알고리즘의 경우 퍼지 레벨이 10, 선택된 유전자의 개수가 10개일 때 가장 우수한 분류 성능을 얻을 수 있었다.

실험 결과에 의하여, 제안된 알고리즘은 백혈병 유전자 데이터의 분류에 있어서, 동일한 유전자 데이터를 이용하여 비교 알고리즘들과 성능을 비교해 보면, 분류 성능이 전체적으로 향상되는 결과르 확인할 수 있었다. 실험 1에서 사용된 백혈병 유전자 데이터의 경우 유전자 간의 클래스 구별이 상대적으로 잘 되어져있는 반면, 대장암 유전자 데이터의 경우, 클래스간의 유전자 데이터의 구분이 상대적으로 불명확하고, 이에 따른 분류 오류 유전자의 개수가 많기 때문에, 단순히 유전자 데이터의 수치를 이용하여 분류할 경우 상대적으로 낮은 분류 결과를 나타낸다고 볼 수 있다. 이는 유전자의 선택의 특징값에서도 두드러진 분류 성능의 변화를 볼 수 없는 점에서도 확인할 수 있다.

그림 6에서는 제안하는 방법과 다른 분류 알고리즘간의 가장 우수한 분류 성능을 보여주고 있다.

3. 림프종 유전자 데이터의 분류

표 4는 제안하는 알고리즘과 비교 알고리즘들을 림프종 유전자 데이터를 대상으로 분류 성능을 비교한 결과이다. 제안하는 알고리즘의 경우 퍼지 레벨이 10, 선택된 유전자의 개수가 10개일 때 가장 우수한 분류 성

표 4. 림프종 데이터의 분류 성능 비교
Table 4. Classification rates of lymphoma gene data.

	BP	DT	SASOM	kNN
Pearson	68.0%	57.4%	63.0%	55.8%
Spearman	52.0%	50.3%	48.4%	48.0%
Uclidean	45.3%	38.3%	40.1%	42.5%
Cosine	40.1%	42.3%	44.0%	44.0%
IG	62.1%	58.3%	68.3%	55.5%
MI	50.3%	52.7%	55.8%	47.3%
SNR	54.1%	55.8%	60.5%	58.3%
제안하는 알고리즘	75.8%			

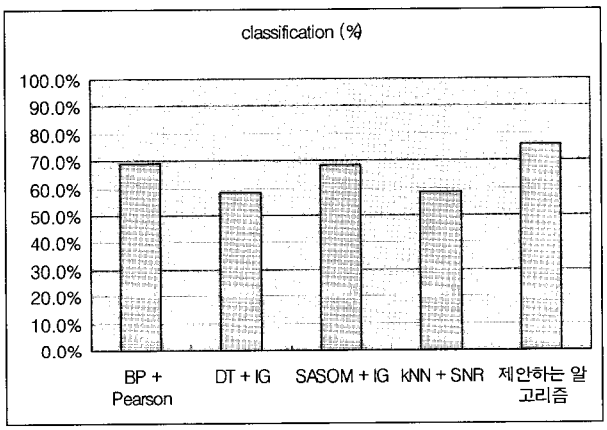

그림 7. 림프종 데이터의 성능 비교
Fig. 7. Classification rates of lymphoma gene data.

능을 얻을 수 있었다.

림프종 유전자 데이터를 이용한 실험 결과에서도 제안하는 알고리즘이 비교 알고리즘들에 비해 우수한 분류 성능을 나타내는 것을 볼 수 있다. 그림 7에서는 제안하는 방법과 다른 분류 알고리즘간의 가장 우수한 분류 성능을 보여주고 있다.

V. 결론 및 향후과제

본 논문에서는 소속 함수(membership function)와 신경망(neural network)을 이용한 유전자 발현 정보를 분류하는 기법을 제안하였다. DNA 칩에 의해 얻어진 수많은 데이터를 분석하기 위해 의미있는 적절한 유전자 데이터를 선택하고 소속 함수을 사용하여 특징값을 구하고, 신경망에 적용하여 클러스터링하였다. 본 논문에서 제안하는 알고리즘의 결과는 기존의 연구와 비교하여 유전자 데이터의 종류와 알고리즘에 따라 다소 차이는 있으나 약 5~10% 향상된 분류 성능을 확인할 수 있었다. 그러나, 제안하는 알고리즘이 잘 분리되어있지 않은 유전자 데이터에서도 상대적으로 우수한 분류 성능을 가지긴 하지만, 제안하는 알고리즘 자체도 유전자 데이터의 분류 정도에 따라 점차적으로 분류 성능이 하락되는 결과도 또한 보이고 있다. 즉, 유전자 발현 데이터들 간의 의미있는 상관 관계를 찾는 일은 클러스터링으로만 적용하기에는 어려운 일이다. 앞으로 유전자 발현 데이터들간의 의미있는 상관 관계와 유전자의 기능과 특성들을 발견하기 위해서는 더욱 많은 유전자 데이터를 통한 실험이 필요하고, 기존의 연구에 대한 비교 검증이 필요하며, 더 효과적인 클러스터링을 위한 학습 방법과 유전자 발현 데이터 분석에 대한 연구가 계속되어야 한다.

참 고 문 헌

[1] M. B. Eisen and P. O. Brown, "DNA arrays for analysis of gene expression," Method Enzymbol, vol. 303, pp. 179-205, 1999.
[2] 유시호, 조성배, "전진선택법에 의해 선택된 부분 상관관계의 유전자들을 이용한 암 분류," 전자공학회논문지, 제41권 SP., 제3호, 83-92쪽, 2004년 5월.
[3] U .Alon, N. Barkai, D. A. Notterman, K. Gish, S.Ybarra, D.Mack, A.J.Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", Proceedings of National Academy of Sciences of the USA, vol.96, pp.6745-6750, June 1999.
[4] Eisen MB, Spellman PT, Brown PO et al. "Cluster analysis and display of genome-wide expression patterns", Proceedings of National Academy of Sciences of the USA, vol.95, pp.14863-14868, 1998.
[5] W. D. Shanon, M. A. Watson, A. Perry, and K. Rich, "Mantel statistics to correlate gene expression levels from microarrays with clinical covariates," Genetic Epidemiology, vol. 23, no. 1, pp 87-96, 2002.
[6] F. Sebastiani, "Machine learning in automated text categorisation: A Survey," Technical Report IEI-B4-31-1999, Istituto DI Elaborazione dell'Informazione, C.N.R, Pisa, 1999.
[7] D. Berrar, W. Dubitzky, M. Granzow, "A Practical Approach to Microdata Analysis,"

Kluwer Academic Publishers, Boston, Dec. 2002.
[8] P. Tamayo, "Interpreting patterns of gene expression with self-organizing map: Methods and application to hematopoietic differentiation," Proc. of National Academy of Sciences, vol. 96, pp. 2907-2912, 1999.
[9] 김현돈, 조성배, "비교사 학습과 교사 학습 알고리즘을 결합한 구조 적응형 자기구성 지도", '99 추계 정보 과학회, 1999.
[10] L. Li, C. R. Weinberg, T. A. Darden and L. G. Pedersen, "Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method," Bioinformatics, vol. 17, no. 12, pp. 1131-1142, 2001.

저 자 소 개

염 해 영(정회원)
2002년 2월 한양대학교 전자컴퓨터공학부 졸업(공학사).
2004년 2월 한양대학교 컴퓨터공학과 졸업(석사).
2003년~현재 삼성전자 연구원.
<주관심분야 : 영상처리, 컴퓨터 비전, 패턴인식 등 >

김 재 협(학생회원)
2001년 2월 한양대학교 전자계산학과 졸업(공학사).
2003년 2월 한양대학교 컴퓨터공학과 졸업(석사).
2003년~현재 한양대학교 컴퓨터공학과 박사과정.
<주관심분야 : 영상처리, 컴퓨터 비전, 패턴인식 등>

문 영 식(정회원)
1980년 2월 서울대학교 공과대학 전자공학과 졸업(학사).
1982년 2월 한국과학기술원 전기 및 전자공학과 졸업(석사).
1990년 University of California at Irvine Dept. of Electrical and Computer Engr. (박사).
1982년~1985년 한국전자통신연구소 연구원.
1989년~1990년 Inno Vision Medical 선임연구원.
1990년~1992년 생산기술연구원 선임연구원.
1992년~현재 한양대학교 전자계산학과 부교수.
<주관심 분야 : 영상처리, 컴퓨터 비전, 패턴인식 등>