암의 예후 예측을 위한 그래프 기반의 준지도 학습 방법

(Graph-based Semi-Supervised Learning Method for Predicting Prognosis of Cancer)

박 치 현 † 박 상 현 †† 김 현 진 † 여 윤 구 † 안 재 균 †
(Chihyun Park) (Sanghyun Park) (Hyunjin Kim) (Yunku Yeu) (Jaegyoon Ahn)

요 약 본 논문에서는 준지도 학습 방법에 기반하여 더욱 정확하게 암의 예후를 예측할 수 있는 방법을 제안한다. 제안하는 방법은 유전자 발현을 측정한 마이크로어레이 데이터를 그래프 기반의 준지도 학습에 적용하기 위해서 샘플 기반의 그래프 모델 구축하는 단계와 구축된 그래프 모델에서 비용함수를 이용하여 최종 클래스 레이블을 예측할 수 있는 부분으로 구성되어 있다. 클래스 레이블이 없는 샘플들이 많은 암 예후 관련 데이터에 대해서 결과적으로 비교 방법보다 우수한 분류 정확도를 보임을 확인하였다.

키워드 : 준지도 학습, 마이크로어레이, 암의 예후 예측

Abstract In this paper, we propose a graph-based semi-supervised learning method for accurate prediction of cancer prognosis. Our method consist of two parts, one is about transforming mRNA microarray data into graph data structure for learning and the other is about predicting the class labels of unlabeled samples using cost function. As a result, we achieved that our method has outstanding accuracy compared to other methods in the prognosis related cancer data which have many unlabeled samples.

Key words : Semi-Supervised Learning, Microarray, Prediction of Cancer Prognosis

1. 서 론

암 연구에 있어서 암의 진단과 함께 예후를 예측하는 일은 임상적인 측면에서 매우 중요한 문제 중 하나이다. 특히 암의 진단과 예후에 핵심적인 역할을 하는 유전자들을 찾고 분류할 수 있는 모델을 만드는 것에 대한 많은 연구는 생물정보학에서 기계학습과 데이터 마이닝 기법을 바탕으로 수행되어 오고 있다. 이를 위해 유전자 발현 정도를 측정할 수 있는 마이크로어레이(microarray) 데이터가 가장 많이 사용되고 있다[1]. 마이크로어레이 데이터는 여러 샘플들에 대해서 수 만개의 유전자를 대상으로 mRNA의 발현량을 수치화하여 행렬의 원소로 가지고 있다. 암과 같은 유전적 질병의 경우 생물학적인 발병 메커니즘을 분석하면 유전자 발현 단계에 문제가 있는 경우가 많은데[2], 유전자 발현의 정도를 측정한 마이크로어레이 데이터를 분석하여 암과 관련된 유전자들을 알아낼 수 있고, 이런 분석을 위해서는 다양한 데이터 마이닝 방법이 필요하다.

암의 예후를 “고위험군(high-risk)”과 “저위험군(low-risk)” 혹은 “재발(recurrence)”과 “재발없음(no recurrence)” 등으로 분류하고 예측하기 위해 데이터 마이닝의 분류(classification) 알고리즘 등이 많이 사용되고 있지만 가장 큰 문제는 마이크로어레이 데이터의 샘플 수가 적다는 것이다. 그 이유는 암환자에 대하여 마이크로어레이 실험을 하는 과정이 비용과 시간이 많이 소요되

· 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 도약연구지원사업 지원을 받아 수행된 것임(2012-010775)

† 학생회원 : 연세대학교 컴퓨터과학과
tianell@cs.yonsei.ac.kr
chriskim@cs.yonsei.ac.kr
ajk@cs.yonsei.ac.kr
yyk@cs.yonsei.ac.kr
†† 종신회원 : 연세대학교 컴퓨터과학과 교수
sanghyun@cs.yonsei.ac.kr
(Corresponding author임)

논문접수 : 2012년 10월 16일
심사완료 : 2012년 12월 10일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제19권 제2호(2013.2)

기 때문인데 예를 들어 재발과 재발 없음의 경우는 임상에서 암환자에 대한 수술 혹은 약물치료에 의해서 암을 제거한 후 상당한 시간이 흐른 후 재발여부를 판단해야 하기 때문에 분류를 위한 클래스 레이블(class label)을 부여하기 어려운 측면이 있다. 하지만 분류자(classifier)의 정확도와 신뢰도를 높이기 위해서는 샘플의 수가 많이 필요하고 이 때문에 기존의 여러 가지 데이터 마이닝 방법들로는 정확한 분류와 예측을 하기 어려운 문제점이 존재한다[3].

이를 위해서 본 연구에서는 클래스 레이블이 없는 데이터를 이용하는 새로운 암 예후 예측 방법을 제안한다. 암의 예후와 관련된 마이크로어레이 데이터의 상당수는 유전자 발현 실험은 수행되었지만 임상적으로 클래스 레이블을 결정하지 못한 샘플들이 많이 존재하는데 이런 데이터들도 분류와 예측을 위한 정보들을 포함할 수 있기 때문에 이를 이용할 수 있는 준지도 학습(semi-supervised learning) 방법을 적용하면 기존 방법보다 신뢰성 높은 결과를 도출할 수 있다[4]. 이를 위해 제안하는 방법에서는 마이크로어레이 데이터를 그래프 모델로 변형한 후 이에 기반하여 준지도 학습을 통해 암의 예후를 예측할 수 있는 새로운 방법론은 제시한다.

본 논문은 총 5절로 구성되어 있다. 2절에서는 국내외 주요 관련연구에 대해서 소개하고, 3절에서는 제안하는 준지도 학습 방법에 대해서 기술한다. 4절에서는 제안한 방법에 대한 실험 결과를 기술하여 본 방법의 타당성을 검증하며, 5절에서는 결론을 맺고 향후 연구에 대해서 제시한다.

2. 관련 연구

2.1 데이터 마이닝 기법을 적용하는 암 분류 연구

대표적으로 [5]에서는 데이터 마이닝의 분류화 알고리즘 중 K-TSP(k-Top Scoring Pair)를 이용하여 결정규칙을 생성해내는 방법으로 암을 분류하였다. 또한 실험을 통해서 직장암, 전립선암, 폐암 등 다양한 암에 대해서 K-TSP 방법이 우수한 분류 정확도를 갖는다는 것을 확인하였다. 이 밖에 K-NN, SVM 등 다양한 기계학습 방법이 암 분류에 이용되어 왔으며 그 과정에서 수행하는 속성 선택 등의 중요성을 밝히는 등 많은 연구가 이루어져왔다[6]. 이와 같은 연구는 최근까지도 지속되고 있는데 [7]에서는 계층군집화(hierarchical clustering)을 이용하여 정상인과 폐암 환자를 분류한 연구를 수행하는 등 다양한 기계학습 방법과 데이터 마이닝 기법이 꾸준히 적용되고 있다. 하지만 이와 같은 기계학습 방법들은 모두 지도학습(supervised learning)에 기반하였다는 특징이 있다.

2.2 준지도 학습을 이용하는 분류 방법

준지도 학습은 모든 데이터의 클래스 레이블을 알아야 하는 지도학습과 클래스 레이블을 하나도 모른 채 학습을 하는 자율 학습(unsupervised learning)의 중간 방법이다. SVM과 같이 지도 학습에 기반한 여러 알고리즘들이 준지도 학습을 위하여 변형이 된 연구가 수행되어왔다. [8]에서는 레이블이 있는 데이터와 없는 데이터를 둘 다 이용하여 마진(margin)을 최대화할 수 있는 S^3VM이라는 방법을 제안하였고, [9]에서는 분류하고자 하는 데이터로부터 클래스 레이블이 있는 데이터와 없는 데이터 모두를 이용하여 가중치 그래프를 구축하고 규일화(regularization)를 적용하여 클래스 레이블이 없는 데이터에 새롭게 클래스 레이블을 부여하는 방법을 제시하였다. 그래프 규일화란 초기 설정된 불완전한 클래스 레이블들을 가지고 있는 가중치 그래프 구조에서 그래프 토폴로지(topology)를 따라서 전체 노드에 대해서 클래스 레이블을 부여하는 방법이다. 이 방법은 평활도가정(smoothness assumption)으로부터 도출되며 비용함수(cost function)를 정의하고 그 값이 최소가 되는 경우를 탐색하여 최종 클래스 레이블을 결정한다.

생물정보학 분야에서는 마이크로어레이 데이터로부터 암의 예후를 예측을 위해서 준지도 학습 방법 중 하나인 LDS(Low Density Separation)방법을 적용하여 분류자 모델을 구축한 최근에 연구가 수행되었다[4]. 하지만 아직까지 준지도 학습은 대부분 생물정보데이터의 분석이 아닌 일반적인 데이터 마이닝 방법에서도 클래스 레이블을 모르는 데이터가 그렇지 않은 데이터보다 월등히 많을 경우에 대해서 적용되고 있으며 최근에는 그래프 모델에 기반한 연구가 활발히 수행되고 있다.

2.3 암의 예후 및 재발과 관련된 연구

[10]에서는 회귀분석 모델을 이용하여 위암의 재발을 예측하는 연구가 수행되었고, [11]에서는 위와 마찬가지로 마이크로어레이 데이터가 아닌 핵자기 공명 영상법(Magnetic Resonance Imaging, MRI) 이미지 데이터로부터 전립선암의 재발을 예측할 수 있는 모델을 제안하였다. 이렇듯 기존에서는 마이크로어레이 데이터를 이용하여 암의 예후를 예측하는 연구가 많이 없었고, 최근 발표된 [4]에서야 비로소 직장(결장)암과 유방암에 대하여 마이크로어레이 데이터로부터 LDS에 기반한 준지도 학습을 적용하여 암의 재발을 예측하는 방법을 제안하였다.

3. 본 론

제안하는 방법은 그림 1과 같이 전체적으로 6가지 단계로 구성이 되어 있다. 단계 1에서부터 4까지는 마이크로어레이 데이터로부터 그래프 구조로 데이터를 변환하

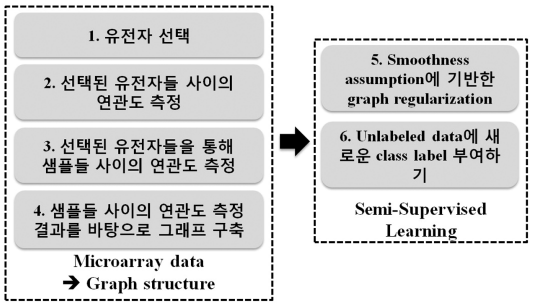

그림 1 전체 방법의 도식도

는 과정을 나타내며, 단계 5와 6은 그래프 기반으로 변환된 데이터에서 클래스 레이블이 없는 데이터도 함께 사용하여 준지도 학습을 통해 새롭게 클래스 레이블을 부여하는 과정을 나타낸다.

3.1 유전자 사이의 연관성 측정

인간에 대한 마이크로어레이 데이터는 약 2만개가 넘는 유전자가 모두가 각각 속성으로 간주되는 매우 고차원의 데이터이다. 따라서 분류 모델을 만들기 위해 모든 유전자를 속성으로 사용한다면 계산 복잡도가 증가하며 중요하지 않은 유전자의 영향으로 전체적인 분류 성능이 감소될 가능성이 존재하기 때문에 중요한 유전자를 선택하여 사용한다. 본 논문에서는 여러 속성 선택(feature selection) 방법 중 Relief-F[12]를 적용한다. Relief-F는 암에 대한 마이크로어레이 데이터에서 유용한 유전자(informative gene)을 뽑아내는데 가장 적합한 속성 선택 방법 중 하나로 많이 사용되고 있기 때문에[13] 본 논문에서도 이를 통해 전체 유전자 N개 중 유용한 유전자를 n개 선택한다.

다음으로 Relif-F를 통해서 선택된 유전자들에 대해서는 모든 가능한 쌍에 대해서 연관성 정도를 계산한다. 두 유전자 사이의 연관성은 아래 식 (1)과 같은 방법으로 측정된다.

$$C(g_i, g_j) = \left| PCC(g_i^{C1}, g_j^{C1}) - PCC(g_i^{C2}, g_j^{C2}) \right| \quad (1)$$

식 (1)에서 PCC는 피어슨상관계수(*Pearson's Correlation Coefficient*, 이하 PCC)를 나타내며, g_i와 g_j는 연관성을 계산하고자 하는 두 유전자를 의미한다. 또한 g_i^{C1}는 유전자 g_i에서 클래스 $C1$에 해당하는 부분을 의미한다. 즉 식 (1)은 두 유전자 사이의 연관성을 측정하기 위해서 각 클래스에 해당하는 요소만을 대상으로 PCC를 측정한 후 두 클래스 사이의 차이의 정도를 계산한다. 그 차이의 정도가 유전자 g_i 와 g_j 사이의 연관도를 의미한다. 식 (1)은 기존 연구 [2]에서 제시된 것으로 두 클래스가 각각 "암" 혹은 "정상"과 같은 서로 상반된 범주에 속해있을 경우 두 유전자에 대해서 서로 다른 클래스에 속하는 샘플들에 대해서는 PCC 값의 차이가 나는 특성을 이용한다. 본 논문에서도 "recurrence"와 "no recurrence" 같은 상반된 서로 다른 2개의 클래스 레이블이 있기 때문에 식 (1)을 적용한다. 제안하는 방법에서는 식 (1)의 값에 대해서 특정 임계치(*thr_g*)를 넘는 유전자 쌍들을 두 클래스를 구분할 수 있는 유용한 유전자로 판단하여 선택하도록 한다.

3.2 샘플 사이의 연관성 측정

앞선 3.1절에서 구한 유전자 쌍을 이용하여 모든 가능한 샘플 쌍에 대한 연결 가중치를 계산한다. 마이크로어레이 데이터에서 각 샘플은 하나의 클래스 레이블을 가지고 있을 수도 있고 없을 수도 있기 때문에 그래프 구조로 변환을 할 때는 각 샘플이 하나의 노드가 되어야 한다. 이를 위해서 샘플들 사이의 연관성을 계산하는데, 이는 샘플의 속성인 유전자를 통해서 계산할 수 있다. 식 (2)는 샘플들 사이의 연관성을 계산하는 식을 나타낸다.

$$W(s_i, s_j) = \left| PCC(sg_i, sg_j) \right| \quad (2)$$

식 (2)에서 sg는 선택된 유전자들을 의미한다. 식 (2)를 통해서 계산된 W값에 대해서 특정 임계치(*thr_s*)를 넘는 샘플 쌍에 대해서는 그래프 상에 엣지(edge)로 추가를 한다. 식 (2)의 경우 클래스 레이블이 없는 샘플에 대해서도 계산을 하기 때문에 클래스 레이블에 상관없이 모든 샘플이 노드가 될 수 있으며 전체적으로 다음 단계인 그래프 균일화를 할 수 있는 형태로 데이터 변환이 완료 된다. 그림 2는 3.1절과 3.2절에서 기술하고 있는 방법에 대한 간단한 예제이다.

3.3 준지도 학습에 기반한 그래프 균일화

그래프 균일화를 위한 비용 함수는 아래 식 (3)과 같다.

$$\min_{Y \in \{0,1\}^n} \left\{ \sum_{i=1}^{l} (\hat{y}_i - y_j)^2 + \frac{1}{2} \sum_{i,j=1}^{n} w_{i,j} (\hat{y}_i - \hat{y}_j)^2 \right\} \quad (3)$$

여기서 y는 초기상태의 클래스 레이블 벡터를 의미하며, $\hat{y}$는 새롭게 얻어진 클래스 레이블의 벡터를 나타낸다. $w_{i,j}$는 i번째 노드에서 j번째 노드의 가중치를 나타낸다. 총 샘플의 수는 n개 이며 이 가운데 클래스 레이블이 있는 원소의 수는 l개이다. 여기서 첫 번째 다항식은 사실상 값이 0으로 고정 될 수 있기 때문에 두 번째 항목에 대해서만 최소화를 하면 되는데 두 번째 항목은 그래프 라플라시안(graph Laplacian)에 의해서 식 (4)와 같이 변환된다.

$$\min_{Y \in \{0,1\}^n} \hat{Y}^T L \hat{Y} \quad (4)$$

여기서 L은 그래프 라플라시안을 의미한다. 식 (4)를 최소화함으로써 결과적으로 우리가 원하는 것은 $\hat{y}$에서 클래스 레이블이 없는 요소들을 나타내는 $\hat{y_u}$인데 이를 풀기 위해서 식 (4)는 다시 식 (5)로 변환되어 계산될 수 있다.

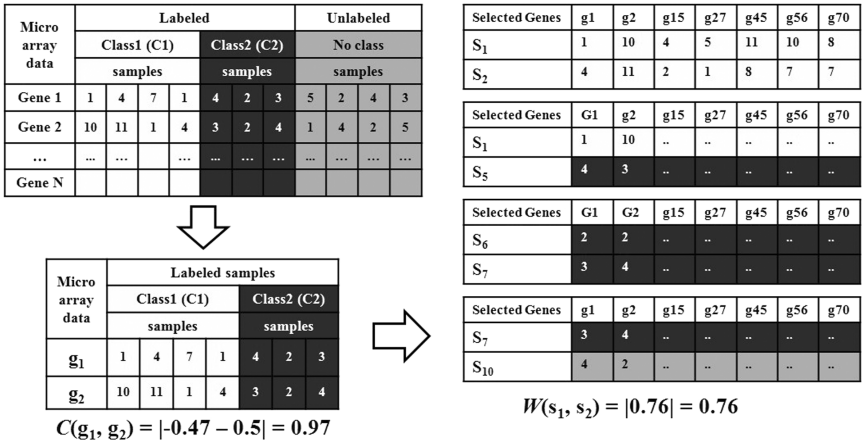

그림 2 유전자와 샘플 사이의 상관도 계산 예

$$\hat{Y}_u = -L_{uu}^{-1} L_{ul} \hat{Y}_l \quad (5)$$

그래프 균일화에서는 결과적으로 식 (5)를 계산하여 그 결과 값에 따라서 클래스 레이블을 모르는 샘플에 대한 클래스 레이블 예측을 하게 된다.

4. 실험 결과 및 분석

본 실험에서는 실제 유방암 환자에 대한 재발 여부를 임상적으로 확인한 마이크로어레이 데이터를 이용해서 제안하는 분류 및 예측 모델의 성능을 측정하고 3장에 기술된 방법의 우수성을 보인다.

4.1 실험 환경 및 실험 데이터

본 실험에서 사용한 데이터는 189명의 유방암 환자에 대해서 유전자 발현 실험을 한 공개 마이크로어레이 데이터이다. 본 데이터는 GEO(Gene Expression Omnibus)에서 아이디 GSE2990을 통해서 다운받을 수 있다. 본 데이터는 189명의 샘플로 구성되어 있으며 그 중 125명은 클래스 레이블이 존재하며 “recurrence”와 “no recurrence”의 형태로 존재하며 64명은 임상적으로 분류되어 있지 않은 샘플이다. 제안하는 방법은 STL을 이용한 C++언어로 구현되었다.

4.2 LOOCV를 이용한 최적의 파라미터 선택

제안하는 방법에서 사용하는 2가지 파라미터인 *thr_g*와 *thr_s*에 대하여 최적의 조합을 찾기 위해서 125명의 클래스 레이블이 존재하는 샘플들만을 대상으로 두가지 파라미터를 변경하면서 LOOCV(Leave One Out Cross Validation)을 수행하였다. 각 파라미터에 대해서 각각 10개의 후보(*thr_g*의 경우 0.35부터~0.53까지 0.02단위로 변경, *thr_s*의 경우 0.50부터~0.95까지 0.05 단위로 변경)를 설정하여 총 100회 LOOCV 실험을 반복하여 최적의 파라미터 집합을 확보하였다. 또한 이와 같은 실험을 Relief-F를 통해 얻어진 유전자 선택의 개수를 변경하면서 수행하였다.

4.3 비교실험 결과 분석

최적의 파라미터 선택 및 비교실험을 위해서 본 실험에서는 정확도(accuracy)를 성능평가 단위로 사용하였다. 정확도는 전체 테스트 실험에 사용된 샘플에 대해서 올바르게 원본 클래스 레이블을 예측한 비율을 나타낸다. 비교 실험을 위해서 대표적인 지도 학습 기반의 분류 방법인 SVM[14]과 Naïve Bayesian[15]을 사용하였고 준지도 학습 기반의 방법 중 SVM기반의 TSVM (Transductive SVM)[8]을 사용하였다. 또한 분류 알고리즘의 경우 트레이닝 과정에서 사용되는 클래스 레이블들의 세부 비율에 따라서 성능 차이가 날 수 있기 때문에 클래스 레이블을 동일 비율로 맞춘 실험도 추가로 수행하였다. 실험 결과는 표 1에 제시하였다. 표 1에서 L은 클래스 레이블이 있는 샘플 수를 나타내고 U는 클래스 레이블이 없는 샘플 수를 의미한다. 표 2는 클래스 레이블을 동일 비율로 설정한 후 동일한 실험을 수행한 결과를 나타낸다. 표 1에 제시되어 있듯이 제안하는 방법은 대표적인 분류 방법인 SVM과 Naïve Bayesian보다 모든 실험에서 정확도가 우수하였으며 준지도 학습 기반의 TSVM보다 더욱 우수한 성능을 보였다. 또한 표 2에 제시한 것 같이 클래스 레이블을 동일한 비율로 설정한 후 수행한 실험에서도 다른 방법보다 우수한 성능을 보임을 확인할 수 있었다.

5. 결론 및 향후 연구

암 환자의 데이터에 대해서 임상적으로 클래스 레이블을 부여하는 것은 비용과 시간이 많이 소요되는 작업이기 때문에 클래스 레이블이 없는 데이터가 많이 존재하며 이런 데이터를 이용하는 준지도 학습 기반의 정확

표 1 비교 실험 결과

사용한 데이터 특징		제안하는 방법	TSVM [8]	SVM [14]	Naïve Bayesian [15]
사용한 데이터의 샘플 수	Relief-F에 의하여 선택된 유전자 수	LOOCV	10-fold CV	10-fold CV	10-fold CV
L: 125 (76, 49) U: 64	1000	0.736 (thr_g=0.53, thr_s=0.80)	0.563	0.496	0.592
L: 125 (76, 49) U: 64	2000	0.672 (thr_g=0.53, thr_s=0.80)	0.521	0.512	0.592
L: 125 (76, 49) U: 64	3000	0.760 (thr_g=0.53, thr_s=0.80)	0.563	0.536	0.584
L: 125 (76, 49) U: 64	4000	0.760 (thr_g=0.53, thr_s=0.80)	0.521	0.560	0.584
L: 125 (76, 49) U: 64	5000	0.760 (thr_g=0.53, thr_s=0.80)	0.542	0.544	0.584

* TSVM: 파라미터 P는 훈련데이터 상의 두 클래스 레이블의 비율
* SVM: PolyKernel -C 250007 -E 1.0, The complexity parameter C(1.0), epsilon(1.0E-12), filterType(Normalize training data)
* Naïve Bayesian: 파라미터 설정 없음

표 2 클래스 레이블 비율을 동일하게 한 후 비교 실험 결과

사용한 데이터 특징		제안하는 방법	TSVM [8]	SVM [14]	Naïve Bayesian [15]
사용한 데이터의 샘플 수	Relief-F에 의하여 선택된 유전자 수	LOOCV	10-fold CV	10-fold CV	10-fold CV
L: 98 (49, 49) U: 64	1000	0.765 (thr_g=0.53, thr_s=0.80)	0.494	0.561	0.622
L: 98 (49, 49) U: 64	2000	0.908 (thr_g=0.45, thr_s=0.80)	0.494	0.459	0.592
L: 98 (49, 49) U: 64	3000	1.000 (thr_g=0.53, thr_s=0.80)	0.493	0.428	0.571
L: 98 (49, 49) U: 64	4000	1.000 (thr_g=0.53, thr_s=0.75)	0.497	0.561	0.541
L: 98 (49, 49) U: 64	5000	1.000 (thr_g=0.53, thr_s=0.75)	0.506	0.479	0.561

* 비교 방법에 대해서 표 1 실험과 동일한 파라미터를 적용함

한 예측모델을 제시한 점은 본 논문의 가장 큰 특징이다. 또한 제안하는 방법에서 결과적으로 얻어지는 최적의 파라미터를 적용하여 재발 여부를 가장 정확하게 분류할 수 있는 유전자 쌍의 집합을 얻을 수 있고 이용해서 유방암의 재발에 특이적인 유전자 네트워크를 구축할 수 있다. 이를 기반으로 기능적인 분석을 수행하고 검증한다면 임상적으로 더욱 의미 있는 결과를 얻을 수 있기 때문에 의학적으로도 유용하게 참고될 수 있다. 하지만 동일한 클래스 비율을 갖는 데이터로 실험을 수행하였을 경우 몇 가지 경우에서 정확도가 1.0이 나온 점은 본 논문에서 사용한 유방암 데이터에 과적합(over-fitting)되었을 가능성이 있으며 향후 연구에서는 다양한 암 데이터로 실험을 확장할 예정이다. 또한 속성 선택 방법을 다양화하고 결과로 나오는 유전자 네트워크에 대한 검증도 수행할 계획이다.

참 고 문 헌

[1] S. Mitra, Y. Hayashi, "Bioinformatics with soft computing," *IEEE Trans. Syst., Man, Cybern. C*, vol.36, no.5, pp.616-635, Sep. 2006.

[2] J. G. Ahn et al., "Integrative Gene Network Construction for Predicting a Set of Complementary Prostate Cancer Genes," *Bioinformatics*, vol.27, no.13, pp.1846-1853, 2011.

[3] Y. Lai et al., "A mixture model approach to the tests of concordance and discordance between two large-scale experiments with two-sample groups," *Bioinformatics*, vol.23, no.10, pp.1243-1250, 2007.

[4] M. Shi et al., "Semi-supervised learning improves gene expression-based prediction of cancer recurrence," *Bioinformatics*, vol.27, no.21, pp.3017-3023, 2011.

[5] A. Tan et al., "Simple decision rules for classifying human cancers from gene expression profiles," *Bioinformatics*, vol.21, no.20, pp.3896-3904, 2005.

[6] M. H. Asyali et al., "Gene Expression Profile Classification: A Review," *Current Bioinformatics*, vol.1, pp.55-73, 2006.

[7] J. Hou et al., "Gene Expression-Based Classification of Non-Small Cell Lung Carcinomas and

Survival Prediction," *PLoS One*, vol.5, no.4, e10312, 2010.
[8] O. Chapelle et al., "Optimization Techniques for Semi-Supervised Support Vector Machines," *Journal of Machine Learning Research*, vol.9, pp.203-233, 2008.
[9] M. Belkin et al., "Regularization and semi-supervised learning on large graphs," *Proc. of the 17th Annual Conference On Learning Theory*, 2004.
[10] D. Marrelli et al., "Prediction of Recurrence After Radical Surgery for Gastric Cancer," *Annals of Surgery*, vol.241, no.2, pp.247-255, 2005.
[11] A. Shukla-Dave et al., "Prediction of Prostate Cancer Recurrence Using Magnetic Resonance Imaging and Molecular Profiles," *Clinical Cancer Research*, vol.15, no.11, pp.3842-3849, 2009.
[12] M. Robnik-Sikonja et al., "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Machine Learning Journal*, vol.53, pp.23-69, 2003.
[13] Y. Wang et al., "HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data," Bioinformatics, vol.21. no.8, pp.1530-1537, 2005.
[14] B. Schölkopf et al., "Advances in kernel methods: support vector learning," MIT Press, 1999.
[15] John, G. H., Langley, P. "Estimating continuous distributions in Bayesian classifiers," *Proc. of the Eleventh conference on Uncertainty in artificial intelligence*, pp.338-345, 1995.

박 치 현

2007년 홍익대학교 컴퓨터공학과 졸업(학사). 2009년 연세대학교 컴퓨터과학과 졸업(석사). 2009년~현재 연세대학교 컴퓨터과학과 박사과정. 관심분야는 시스템 생물학, 바이오인포매틱스, 데이터마이닝, 데이터베이스 시스템

박 상 현

1989년 서울대학교 컴퓨터공학과 졸업(학사). 1991년 서울대학교 대학원 컴퓨터공학과(공학석사). 2001년 UCLA 대학원 컴퓨터과학과(공학박사). 1991년~1996년 대우통신 연구원. 2001년~2002년 IBM T. J. Watson Research Center Post-Doctoral Fellow. 2002년~2003년 포항공과대학교 컴퓨터공학과 조교수. 2003년~2006년 연세대학교 컴퓨터과학과 조교수. 2006년~2011년 연세대학교 컴퓨터과학과 부교수. 2011년~현재 연세대학교 컴퓨터과학과 교수. 관심분야는 데이터베이스, 데이터 마이닝, 바이오인포매틱스, 적응적 저장장치 시스템, 플래쉬메모리 인덱스, SSD

김 현 진

2010년 연세대학교 컴퓨터과학과 졸업(학사). 2010년~현재 연세대학교 컴퓨터과학과 통합과정. 관심분야는 바이오인포매틱스, 데이터 마이닝, 텍스트 마이닝, 그래프 마이닝, 데이터베이스

여 윤 구

2009년 연세대학교 컴퓨터과학과 졸업(학사). 2011년 연세대학교 컴퓨터과학과 졸업(석사). 2011년~현재 연세대학교 컴퓨터과학과 박사과정. 관심분야는 바이오인포매틱스, 데이터 마이닝, 데이터베이스 시스템

안 재 균

2006년 연세대학교 컴퓨터과학과 졸업(학사). 2009년 연세대학교 대학원 컴퓨터과학과 졸업(석사). 2009년~현재 연세대학교 대학원 컴퓨터과학과 박사과정 관심분야는 바이오인포매틱스, 데이터 마이닝, 데이터베이스 시스템