한국경영과학회지
제39권 제4호
2014년 11월
http://dx.doi.org/10.7737/JKORMS.2014.39.4.075

Set Covering 기반의 대용량 오믹스데이터 특징변수 추출기법*

마정우[1] · 안기동[2] · 김광수[3] · 류홍서[1†]
[1]고려대학교 산업경영공학과, [2]고려대학교 정보경영공학과, [3]서울대학교 생물정보연구소

Set Covering-based Feature Selection of Large-scale Omics Data

Zhengyu Ma[1] · Kedong Yan[2] · Kwangsoo Kim[3] · Hong Seo Ryoo[1]
[1]School of Industrial Management Engineering, Korea University
[2]School of Information Management Engineering, Korea University
[3]Bioinformatics Institute, Seoul National University

■ Abstract ■

In this paper, we dealt with feature selection problem of large-scale and high-dimensional biological data such as omics data. For this problem, most of the previous approaches used simple score function to reduce the number of original variables and selected features from the small number of remained variables. In the case of methods that do not rely on filtering techniques, they do not consider the interactions between the variables, or generate approximate solutions to the simplified problem. Unlike them, by combining set covering and clustering techniques, we developed a new method that could deal with total number of variables and consider the combinatorial effects of variables for selecting good features. To demonstrate the efficacy and effectiveness of the method, we downloaded gene expression datasets from TCGA (The Cancer Genome Atlas) and compared our method with other algorithms including WEKA embeded feature selection algorithms. In the experimental results, we showed that our method could select high quality features for constructing more accurate classifiers than other feature selection algorithms.

Keywords : Bioinformatics, Feature Selection, Set Covering Problem, Omics Data

논문접수일 : 2014년 09월 06일 논문게재확정일 : 2014년 11월 04일
논문수정일(1차 : 2014년 10월 28일)
* 이 논문은 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. NRF-2013R1A1A2011784).
이 논문은 2014년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2013R1A1A2006592).
† 교신저자 hsryoo@korea.ac.kr

1. Introduction

최근의 생명과학 분야에서는 유전체학(Genomics), 전사체학(Transcriptomics), 단백체학(Proteomics), 대사체학(Metabolomics), 후생유전학(Epigenomics) 등의 다양한 오믹스(omics) 학문분야가 빠르게 발전하고 있다. 이는 마이크로어레이(microarray) 및 차세대시퀀싱(Next Generation Sequencing) 기술의 등장과 발전으로 인해 전 세계적으로 대용량 오믹스 데이터가 폭발적으로 증가하는 것과 그 흐름이 병행한다. 이러한 상황에서 대용량의 오믹스 데이터를 효과적, 체계적으로 분석하여 그로부터 의미 있는 생물학적 지식을 발굴하는 것은 오믹스 학문의 성공을 위한 필수적인 요건이라 할 수 있다.

생물정보학은 생명과학 분야의 문제를 수리과학, 통계학, 전산학 등을 종합적으로 이용하여 다루는 학문이며 대용량의 오믹스 데이터를 다루고 분석하는 것, 특별히 수천 개 이상의 변수를 포함하는 오믹스 데이터로부터 소수의 중요한 변수를 선택하는 특징변수 선택 문제는 대표적인 생물정보학 문제이다(참고, 단일염기다형성(Single Nucleotide Polymorphism, SNP) 데이터 분석[6, 17]; 유전자 발현 데이터 분석[11, 12, 15]; DNA 메틸화 데이터 분석[18, 27]; 단백질 데이터 분석[24, 26] 등). 이때 생물정보학 분야의 특징변수 선택문제는 수 만개 이상의 변수를 포함하는 고차원 데이터를 분석해야 한다는 점에서 기존의 기계학습 및 데이터마이닝 분야에서 다루어져 왔던 특징변수 선택문제와는 특징이 다르다. 예를 들어 수십만 개의 변수를 포함하는 전장유전체 연관분석(Genome-wide Association Study, GWAS) 데이터로부터 변수들의 상호작용을 고려하여 동시에 두 개 이상의 변수를 선택하는 것은 불가능한 것처럼 보인다. 이러한 문제점으로 인해 그동안의 기법들은 t-test[16], wilcoxon rank-sum test[22]와 같이 각 변수들을 한 개씩 독립적으로 선택하는 것이 대부분이었다[21]. 그 외에는 동시에 여러 개의 변수를 고려하는 방법으로 SVM을 이용한 연구와(예 : [13, 25]) 최근에 발표된 선형계획법 기반의 EllipsoidFN[19]과 같은 소수의 연구들이 존재한다.

동시에 두 개 이상의 변수들의 상호작용을 고려하여 특징변수를 선택하고자 할 때 이와 관련된 문제는 조합최적화의 Set Covering(SC) 모형으로 표현될 수 있다[9]. 이때 조합최적화의 SC 문제란 전체집합 S와 S의 부분집합들 T_1, T_2, …, T_n이 주어졌을 때 $S = \cup T_k, k \in K$를 만족시킬 수 있는 최소 개수의 부분집합을 선택하는 것을 목표로 하는 문제이다. 즉 부분집합의 개수 $|K|$를 최소화 하면서 부분집합의 합이($\cup T_k, k \in K$) 전체집합 S가 될 수 있도록 하는 것이다. SC 문제는 승무원 일정계획 결정[20], 응급센터 위치 결정[23] 등의 다양한 현실문제에 적용될 수 있는데 본 논문에서는 특징변수 선택 문제를 SC 문제로 변환하여 다룬다.

본 논문에서는 이진논리(boolean logic)를 이용한 SC 기반의 특징변수 선택기법을 제안한다. 이와 관련하여 본 논문의 방법론과 기존 SC 기반 방법론과의 차이점을 설명하기 위해서 SC 기반 방법론에 관한 기본적인 단계를 간략하게 설명할 필요가 있다. [그림 1]에 제시한 것처럼, 두 개의 집단 U와 V로 구성된 이원분류 데이터가 있다고 가정할 때 SC 기반의 특징변수 선택을 위한 첫 번째 단계는 각 집단의 데이터를 이진화(binarization)하는 것이다. 이때 일반적으로 새롭게 생성된 이진변수의 개수 n_b는 원래변수의 개수 n_o보다 훨씬 크다($n_b \gg n_o$). 두 번째 단계는 이진화된 U, V의 모든 데이터 쌍 $U_i \in U$와 $V_j \in V$에 관하여 데이터간의 짝비교(pairwise comparison)를 실시하여 SC 문제의 행렬 A를 생성하는 것이다. 이때 U_i와 V_j에 관련된 A_l의 값은 $A_{lk} = 1(U_{ik} \neq V_{jk})$, $A_{lk} = 0(U_{ik} = V_{jk})$로 정의된다. 따라서 행렬 A는 $m(=|U||V|)$개의 행과 $n_b(>100{,}000)$개의 열을 가지게 된다.

위에서 언급한 것처럼 생물정보학 분야의 SC 문제는 수십만 개 이상의 변수를 가지기 때문에 풀기가 어렵고 데이터 개수가 많은 경우에는 문제 자체를 메모리에서 처리할 수 없는 경우가 발생한다. 이러한 문제를 해결하기 위해서 기존 연구들은 대부분 필터링

✓ 단계 1. 데이터 이진화

(1) 변수 A의 값을 오름차순으로 정렬: $(U_{3A}=1.0,\ V_{2B}=2.3,\ U_{2A}=2.6,\ U_{1A}=V_{1A}=3.5)$;

(2) 인접한 값 U_{iA}, V_{jA}의 집단(표현형)이 다를 경우 cut point를 아래와 같이 설정함.

cut point $=(U_{iA}+V_{jA})/2$ (예: $C_{A1}=(U_{3A}+V_{2B})/2=(1.0+2.3)/2=1.65$)

(3) 남은 변수(B, C)에 대해서 (1), (2) 반복하여 다음과 같이 cut points를 설정함

$C_{A1}=1.65,\ C_{A2}=2.45,\ C_{A3}=3.05;\ C_{B1}=1.85,\ C_{B5}=2.95;\ C_{C6}=1.9,\ C_{C7}=3.3,\ C_{C8}=4.5$

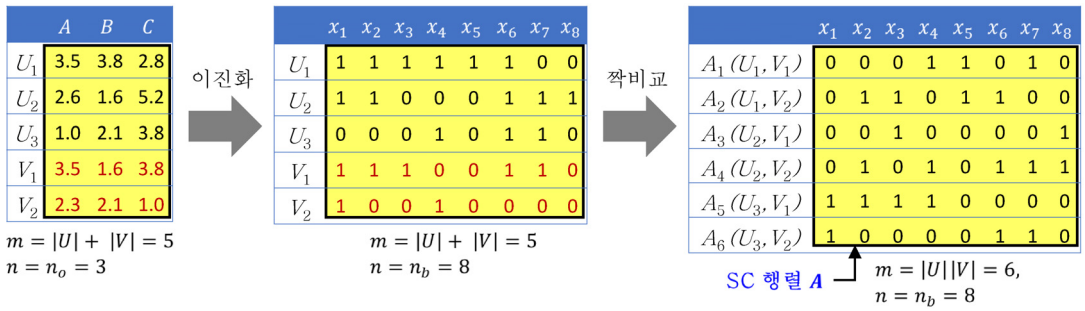
(4) Cut points를 이용하여 아래의 기준을 따라 데이터를 이진화함

$x_j=1(=0)$ if $*\geq C_{*j}(*\leq C_{*j})$ for $*\in\{A,\ B,\ C\}$

✓ 단계 2. 짝비교(Pairwise Comparison)를 통한 Set Covering(SC) 행렬 A 생성

(1) U_i, V_j의 짝비교를 통해 생성되는 데이터를 A_l이라고 할 때 A_l은 아래와 같이 결정됨

$A_{lk}=1(=0)$ if $U_{ik}\neq V_{jk}(U_{ik}=V_{jk})$



[그림 1] 데이터 이진화와 짝비교를 통한 Set Covering 행렬 생성의 예

기법을 사용한다. 예를 들어 entropy와 같은 단순한 계산량을 사용하여 원래 변수의 개수를 100개 내외 정도로 줄인 후에 위에서 언급한 이진화 및 짝비교를 실시함으로써 SC 문제의 크기를 줄인다(참고 [1, 2, 3, 4]). 그 외 방법으로는 짝비교를 다른 방법으로 대체하는 것이다. 집단 U와 V를 완전하게 구분하는 특징변수를 선택하기 위해서는 짝비교를 수행하여 SC의 행렬 A를 만들어야 하지만 짝비교를 수행할 경우 데이터의 개수가 증가함에 따라 행렬 A의 크기가 기하급수적으로 증가한다는 문제점이 발생한다. 따라서 이를 피하기 위해서 overlapping score 등을 정의한 후 행렬 A와 유사한 성격을 가지는 새로운 행렬 A'를 만들고 이에 관한 SC 해법을 통해 U와 V를 불완전하게 구분하는 특징변수를 선택하는 방법들이 제시되었다(참고 [5, 7, 8]). 이 경우 행렬 A'는 $m'(=|U|+|V|)$개의 행으로 구성되기 때문에 문제의 크기가 훨씬 작아지게 된다.

본 연구에서는 기존의 방법론이 사용하던 필터링 기법을 사용하지 않으며 두 집단을 불완전하게 구분하는 특징변수를 선택하지도 않는다. 자세한 내용은 제 2장에서 설명할 것이지만 본 논문의 방법론은 첫째, n_b개의 전체 이진변수로부터 특징변수를 선택할 수 있다는 장점을 지닌다. 둘째, 전체 데이터의 짝비교를 수행하는 대신 소수 데이터의 짝비교를 반복적으로 수행함으로써 행렬 A의 크기를 줄이는 동시에 두 집단 U와 V를 완전하게 구분할 수 있는 특징변수를 선택할 수 있도록 하였다. 데이터의 개수가 지속적, 폭발적으로 증가하는 상황이라는 점과 이전 방법론들이 100개 내외의 데이터를 다루는데 그쳤다는 점을 고려할 때 본 연구의 방법론은 생물정보학 문제를 실제적으로 다룰 수 있도록 하였다는 점에서 중요성이 크다고 할 수 있다.

이 논문의 구성은 다음과 같다. 제 2장에서는 본 논문의 특징변수 선택 방법론에 관하여 구체적인 절차 및 최적화 모형을 제시한다. 제 3장에서는 유전자 발현에 관한 실제 오믹스 데이터를 이용하여 다른 방법론들과 특징 변수 선택에 관한 실험 결과를 비교한다. 제 4장에서는 결론 및 향후 연구과제를 제시한다.

2. Methods

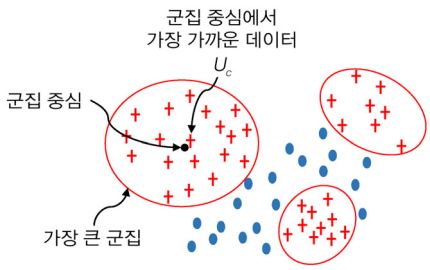
먼저 집합 U와 V에 관한 이진화된 데이터가 있을 때, 편의상 집합 U를 환자들의 데이터, 집합 V를 정상인의 데이터라고 가정하자. 이때 생물정보학에서는 질병을 이해하고 질병의 원인을 파악하는 것이 주목적이므로 대부분의 경우 집합 U가 집합 V에 비해 훨씬 많은 데이터를 포함한다. 더욱이 암과 같은 복잡한 질병의 경우 발병요인이 매우 다양하므로 환자 집합의 데이터 이질성(heterogeneity)이 정상인의 집합보다 훨씬 높다는 점도 집합 U의 크기를 증가시키는 요인이 된다. 이러한 상황에서 본 논문은 생물정보학 분야의 특성을 반영하여 환자 집합 혹은 상대적으로 더 위험한 환자의 집합을 U로 정의하고 집합 V로부터 집합 U를 구별하는데 필요한 특징변수를 추출하도록 하였다.

단계 1 : k-평균 군집화 알고리즘을 이용한 집합 U의 대표 데이터 선택

U, V의 모든 데이터 쌍 $U_i \in U$와 $V_j \in V$에 관하여 데이터간의 짝비교를 실시하면 SC 문제에서 행렬 A의 크기가 너무 커진다는 문제점이 있음을 위에서 언급하였다. 본 논문에서는 이러한 문제점을 해결하고 또한 생물정보학에서 다루는 데이터, 즉 이질성이 높은 집합 U의 특성을 반영하기 위해서 집합 U의 군집분석을 시행하였다. 구체적으로 k-평균 군집화 알고리즘을 적용하여 집합 U의 군집을 구하고 그로부터 가장 큰 군집의 중심에 있는 한 개의 데이터, 즉 집합 U를 대표할 수 있는 데이터 U_c를 구하였다([그림 2] 참조). 이 과정에서 k = 3을 이용하였고, 각 군집의 시작점은 임의로 선택하였다.

단계 2 : SC 문제 생성 및 풀이를 통한 중요 이진변수 선택

단계 1에서 제시된 것처럼 k-평균 군집화 알고리즘을 수행함으로써 집합 V의 데이터와 비교할 수 있는 한 개의 중요한 데이터 U_c를 선택하였다. 단계



[그림 2] 군집화 알고리즘을 이용한 대표 데이터 선택

2에서는 U_c와 모든 $V_j \in V$에 대해 짝비교를 시행하여 SC 풀이를 위한 행렬 A^{U_c}를 생성하였다. 이때 행렬 A^{U_c}는 $|V|$개의 행으로 구성되며 각 행 $A_j^{U_c}$의 값은 $A_{jk}^{U_c}=1(U_{ck} \neq V_{jk})$, $A_{jk}^{U_c}=0(U_{ck}=V_{jk})$로 정의된다. 본 논문에서는 A^{U_c}를 바탕으로 아래에 제시된 SC 문제의 해법을 통해 데이터 $U_c \in U$와 집단 V를 분류할 수 있는 이진변수를 선택하였다.

$$\text{minimize} \quad \sum_{k=0}^{n} c_k x_k$$

$$\text{s.t.} \quad \sum_{k=0}^{n} a_{jk} x_k \geq b_j, \qquad j=1,\cdots,|V|$$

$$^{\forall} x_k \in \{0,1\}$$

위 모델에서 $n=n_b$, $a_{jk}=A_{jk}^{U_c}$이다. 또한 $c_k=1/(d_k s_k)$이고, 이때 d_k는 집합 V에서 U_{ck}의 값과 다른 값을 가지는 데이터의 개수($\sum_{j=1}^{|V|} A_{jk}^{U_c}$), s_k는 집합 U에서 U_{ck}의 값과 같은 값을 가지는 데이터의 개수를 의미한다. 본 논문에서는 c_k의 값을 $1/(d_k s_k)$로 정의함으로써 U, V의 모든 데이터에 대한 짝비교를 수행하지 않았음에도 불구하고 U_c를 V 집합의 데이터와 잘 구별시킬 뿐 아니라, U 집합의 데이터와 공통된 특성을 가질 수 있도록 특징변수를 추출하였다. 나아가 본 논문에서는 생물정보학 문제의 특성(예 : 생물학적 실험의 오류로 인하여 생기는 데이터의 오류 등)을 고려하여 b_j의 값을 1보다 큰 값을 가지도록 함으로써 U_c를 집합 V의 데이터로부터 보다 확실하게 구분할 수 있게 하였다. 이때 $b_j>1$인 경우를 general set covering

(GSC)라 하는데, GSC 문제의 해법으로 Chvatal[10]의 알고리즘을 사용하였다.

단계 3 : SC에서 선택된 이진변수로 설명되는 집합 U의 데이터 제거

SC 해법으로 얻어진 해를 x^{U_c}라 하고 $x_k^{U_c}=1$을 만족하는 이진변수 x_k의 인덱스 집합을 K라 하자. U_c를 제외한 집합 U의 임의의 데이터를 U_i라 할 때, 모든 $k \in K$에 대하여 $U_{ck}=U_{ik}$가 성립할 때 U_i는 x^{U_c}에 의해 집합 V의 모든 데이터와 분리된다. 이처럼 x^{U_c}에 의해 집합 V와 분리되는 집합 U의 데이터를 집합 U로부터 제거한 후 다시 단계 1로 돌아간다. 이때 집합 U에 남아 있는 데이터가 없다면 단계 4로 넘어간다.

단계 4 : 선택된 이진변수와 관련된 원래변수들을 특징변수로 선택

단계 1~단계 3을 반복적으로 수행함으로써 얻어진 SC의 해집합은 이진변수로 구성된다. 따라서 마지막으로 선택된 이진변수들의 정보를 바탕으로 원래변수를 선택하는 단계가 필요하다. 이때 선택된 이진변수들과 많이 관련될수록 원래변수의 중요성은 높아진다고 할 수 있지만, 본 논문에서는 단 한 개의 이진변수와 관련된 원래변수도 특징변수로 선택하였다.

3. Results

〈표 1〉 TCGA에서 다운받은 유전자발현 데이터

데이터 이름	원래변수 개수	암 샘플 데이터 개수	정상 샘플 데이터 개수
COAD	20500	434	41
LUSC	20500	483	50
KIRC	20500	518	72
BRCA	20213	975	92

본 논문에서는 제안된 방법론의 우수성을 입증하기 위해서 실제 생물정보학 데이터를 이용하여 다양한 특징변수 선택방법론과 비교를 시행하였다. 이때 본 논문에서 사용한 데이터는 미국 국립인간게놈연구소와 미국 국립암연구소가 주도하는 TCGA(The Cancer Genome Atlas) 프로젝트에서 공개한 유전자 발현 데이터이다. TCGA 프로젝트를 통해 십여 개가 넘는 암에 관해서 대규모의 암 세포 관련 데이터가 축적되고 있는데 본 연구에서는 <표 1>에 제시된 것처럼 Colon adenocarcinoma(COAD), Lung squamous cell carcinoma(LUSC), Kidney renal clear cell carcinoma(KIRC), Breast invasive carcinoma(BRCA)에 관한 유전자 발현 데이터를 사용하였다. 이때 사용된 유전자 발현 데이터를 간략하게 설명하면 차세대 시퀀싱 기법을 사용하여 유전자의 발현량을 측정한 것으로서 각 유전자의 발현량 값을 변수로 가진다. 따라서 전체 변수의 개수는 발현량을 측정한 유전자의 개수와 같다. 또한 각 변수는 유전자의 발현량 값을 측정한 것으로서 숫자형 변수이며 0 이상의 실수값을 가진다.

<표 1>에 제시된 데이터를 사용한 이유는 크게 두 가지 측면에서 살펴볼 수 있다. 첫째로 특징변수 선택문제에 관한 생물정보학 분야의 기존 연구는 대부분 수십 개 내외의 데이터를 사용하였지만, 최근에는 생물정보학 데이터가 급격하게 증가하고 있으며 그러한 상황에서 데이터의 축적 및 공유를 통한 훨씬 많은 수의 데이터를 분석해야 할 필요성이 증대되었기 때문이다. 따라서 본 연구에서는 현실 상황에서 이미 발생하고 있는 대용량 데이터 분석 문제를 실제적으로 다룰 수 있음을 보이기 위하여 수백에서 최대 천개에 이르는 데이터를 이용하였다. 둘째로 생물정보학 데이터는 대부분 실험군 데이터(예: 암 샘플 데이터)는 개수가 많고 대조군 데이터(예: 정상 샘플 데이터)는 개수가 매우 작다(대부분 20개 미만)는 특성과 고차원 데이터라는 특성을 동시에 갖고 있으므로 특징변수로 선택된 것들 중에는 위양성오류(false positive)가 필연적으로 많이 포함될 수밖에 없기 때문이다. 따라서 본 연구에서는 대조군 데이터의 개수가 큰 문제를

선택함으로써 위양성오류를 최소화 할 수 있도록 하였고 이를 바탕으로 방법론들의 우수성을 보다 객관적으로 비교하였다.

아래 부분에는 편의를 위해 본 논문의 특징변수 선택 방법론을 SCFSB(Set Covering-based Feature Selection in Bioinformatics)라고 지칭하였다. SCFSB의 비교대상으로 WEKA[14]에서 특징변수 선택 방법론을 4개(ChisquareAttributeEval, FilteredAttributeEval, ReliefFAttributeEval, SymmetricalUncertAttributeEval)를 선택하였다. 또한 SCFSB 기법이 수리계획법을 기반으로 변수 간의 상호작용을 고려하여 특징변수를 선택한다는 점에서 최근에 발표된 선형계획법 기반의 ellipsoidFN[19]을 추가로 선택하였다.

선택된 특징변수의 우수성을 측정하는 기준으로 데이터 분류 정확성을 이용하는 것은 특징변수 선택 목적에 부합하기 때문에 합리적이다. 따라서 본 연구에서는 WEKA의 데이터 분류 방법론을 이용하여 특징변수만을 포함한 데이터의 분류 정확도를 측정하였다. 이때 WEKA에는 30여 가지가 넘는 분류 방법론이 포함되어 있는데 본 논문에서는 NaiveBayes (NB), SMO, IBK, Random forest(RF)를 이용하였다. 아래의 <표 2>는 <표 1>의 데이터를 이용하여 10-묶음 교차타당성(10-fold cross validation) 분석을 수행한 결과이다. 몇 개의 특징변수를 선택할 것인가에 관해서는 여러 가지 논의가 있는데 본 논문에서는 기존의 연구결과들이 대부분 10개 내외의 특징변수를 제공했다는 점에 기반하여 SCFSB가 최소 10개의 특징변수를 선택할 수 있도록 SC 모델의 b_j 값을 결정하였고, 다른 기법들은 SCFSB에서 선택된 것과 같은 개수의 특징변수를 선택하도록 하였다. 참고로, 표에 제시된 시간은 특징변수를 선택하는데 소요된

〈표 2〉 특징변수 선택 알고리즘의 계산 비용(소요 시간)과 분류 정확도를 비교한 결과

데이터	특징변수 선택방법	시간(초)	특징변수 개수	WEKA 알고리즘의 데이터 분류 정확도(%)				
				NB	SMO	IBK	RF	평균
COAD	LAD	95.6	10.0	100.00	100.00	100.00	100.00	100.00
	EllipsoidFN	2011.3	10.0	97.68	98.32	98.31	99.31	98.41
	Chisquare	3.8	10.0	100.00	100.00	100.00	99.79	99.95
	Filtered	3.6	10.0	100.00	100.00	100.00	99.79	99.95
	ReliefF	66.6	10.0	97.92	100.00	100.00	100.00	99.48
	SymmetricalUncert	3.7	10.0	100.00	100.00	100.00	100.00	100.00
LUSC	LAD	121.1	10.0	99.25	99.63	99.63	99.44	99.49
	EllipsoidFN	4069.7	10.0	98.31	97.56	98.31	97.94	98.03
	Chisquare	4.8	10.0	98.68	99.63	99.44	99.25	99.25
	Filtered	4.6	10.0	99.25	99.44	99.44	99.44	99.39
	ReliefF	82.3	10.0	96.81	99.62	99.43	99.81	98.92
	SymmetricalUncert	5.1	10.0	98.87	99.63	99.44	99.25	99.30
KIRC	LAD	96.9	13.7	97.45	98.98	99.32	99.49	98.81
	EllipsoidFN	9866.9	13.7	95.93	97.44	97.45	98.65	97.37
	Chisquare	4.2	13.7	96.43	99.32	98.98	99.32	98.52
	Filtered	4.2	13.7	96.77	98.98	98.98	99.15	98.47
	ReliefF	91.4	13.7	96.27	99.15	98.98	99.15	98.39
	SymmetricalUncert	4.4	13.7	96.94	99.15	98.81	99.32	98.56
BRCA	LAD	486.4	10.2	99.16	98.03	98.50	99.16	98.71
	EllipsoidFN	4394.3	10.2	96.54	94.76	96.16	98.04	96.38
	Chisquare	8.1	10.2	98.78	98.87	98.69	99.16	98.87
	Filtered	8.3	10.2	98.60	98.87	98.97	98.97	98.85
	ReliefF	296.7	10.2	94.93	97.10	97.00	96.82	96.46
	SymmetricalUncert	8.1	10.2	98.41	98.31	98.97	98.97	98.67

시간을 의미하며 24GB 메모리와 i7-2600 4 core CPU를 탑재한 개인용 컴퓨터를 이용하여 얻은 결과이다. 또한 ellipsoidFN의 결과는 홈페이지에서 제공되는 matlab 코드보다 c++를 이용하여 자체적으로 구현한 것이 더 빠른 속도를 나타냈기 때문에 자체적으로 구현한 코드의 결과를 제시하였다.

먼저 <표 2>에 제시된 결과를 통해 SCFSB 방법론과 다른 방법론들을 계산 비용 측면에서 비교해 볼 수 있다. 이 과정에서 고려해야 할 점은 SCFSB가 수리계획법의 조합최적화 기반의 방법론으로서 단순한 점수함수(score function) 기반의 WEKA 방법론들에 비해 훨씬 난이도가 높은 문제라는 점이다. 비록 SCFSB가 WEKA의 특징변수 선택 방법론에 비해 시간 측면에서 우수하다고 말할 수는 없지만 대용량의 오믹스 데이터를 다루는 상황에서 문제의 난이도가 훨씬 높음에도 불구하고 수행시간이 비교적 짧다는 점, WEKA의 방법론과는 달리 변수 간의 상호관계를 고려할 수 있다는 점은 SCFSB가 계산 비용 측면에서 비효율적인 방법론이 아님을 알려준다. 나아가 조합최적화 기반의 SCFSB가 선형계획법 기반의 EllipsoidFN 보다 약 10~100배 이상 빠르다는 점을 고려할 때 SCFSB가 계산 비용 측면에서 경쟁력이 있는 방법론임을 알 수 있다.

다음으로 <표 2>의 결과를 통해서 SCFSB가 데이터 4개 중 3개(COAD, LUSC, KIRC)에서 가장 좋은 분류 정확도를 나타냈음을 살펴볼 수 있다. 이때 본 논문에서 사용한 유전자 발현 데이터는 암 샘플의 개수가 정상 샘플의 개수에 비해서 훨씬 많

〈표 3〉 혼동 행렬과 민감도, 특이도, 정확도를 이용하여 데이터 분류 우수성을 비교한 결과

데이터	특징변수 선택방법	데이터 개수*				민감도*	특이도*	정확도*
		TP	FN	FP	TN	(Sensitivity)	(Specificity)	(Accuracy)
COAD	SCFSB	434.00	0.00	0.00	41.00	100.00	100.00	100.00
	EllipsoidFN	428.00	6.00	1.25	39.75	98.62	96.95	98.41
	Chisquare	433.75	0.25	0.00	41.00	99.94	100.00	99.95
	Filtered	433.75	0.25	0.00	41.00	99.94	100.00	99.95
	ReliefF	431.50	2.50	0.00	41.00	99.42	100.00	99.48
	SymmetricalUncert	434.00	0.00	0.00	41.00	100.00	100.00	100.00
LUSC	SCFSB	481.00	2.00	0.75	49.25	99.59	98.50	99.49
	EllipsoidFN	477.25	5.75	4.75	45.25	98.81	90.50	98.03
	Chisquare	480.75	2.25	1.75	48.25	99.53	96.50	99.25
	Filtered	481.50	1.50	1.75	48.25	99.69	96.50	99.39
	ReliefF	477.50	5.50	0.25	48.25	98.86	99.50	98.92
	SymmetricalUncert	481.00	2.00	1.75	48.25	99.59	96.50	99.30
KIRC	SCFSB	512.75	5.25	1.75	70.25	98.99	97.57	98.81
	EllipsoidFN	508.50	9.50	6.00	66.00	98.17	91.67	97.37
	Chisquare	511.75	6.25	2.50	69.50	98.79	96.53	98.52
	Filtered	512.25	5.75	3.00	69.00	98.89	95.83	98.47
	ReliefF	510.00	8.00	1.50	70.50	98.46	97.92	98.39
	SymmetricalUncert	512.25	5.75	2.75	69.25	98.89	96.18	98.56
BRCA	SCFSB	968.00	7.00	6.75	85.25	99.28	92.66	98.71
	EllipsoidFN	957.00	18.00	20.75	71.25	98.15	77.45	96.38
	Chisquare	967.25	7.75	4.25	87.75	99.21	95.38	98.87
	Filtered	967.00	8.00	4.25	87.75	99.18	95.38	98.85
	ReliefF	952.75	22.25	15.50	76.50	97.72	83.15	96.46
	SymmetricalUncert	966.25	8.75	5.50	86.75	99.10	94.02	98.67

* : 4가지 WEKA 데이터 분류 방법론(NB, SMO, IBK, RF) 결과의 평균 값.
TP : True Positive, FN : False Negative, FP : False Positive, TN : True Negative.
Positive : 암 샘플 데이터, Negative : 정상 샘플 데이터.

〈표 4〉 분류 정확도를 순위 값으로 변경한 결과, 순위가 같은 경우 평균값을 사용하였음

특징변수 선택방법	분류 정확도의 순위															
COAD	LUSC				KIRC				BRCA							
N	S	I	R	N	S	I	R	N	S	I	R	N	S	I	R	
SCFSB	2.5	3	3	2	1.5	1	1	2.5	1	4	1	1	1	4	4	1
EllipsoidFN	6	6	6	6	5	6	6	6	6	6	6	6	5	6	6	5
Chisquare	2.5	3	3	4.5	4	2.5	3	4.5	4	1	2.5	3	2	1.5	3	2
Filtered	2.5	3	3	4.5	1.5	5	3	2.5	3	5	2.5	4.5	3	1.5	2	4
ReliefF	5	3	3	2	6	4	5	1	5	2.5	4	4.5	6	5	5	6
SymmetricalUncert	2.5	3	3	2	3	2.5	3	4.5	2	2.5	5	2	4	3	1	3

N : NaiveBayes, S : SMO, I : IBK, R : Random forest.

다는 점에서 클래스 불균형(Class Imbalance) 문제를 유발하므로 본 논문에서는 혼동 행렬과 민감도, 특이도, 정확도에 관한 결과를 종합적으로 제시한다(<표 3> 참조). 참고로 <표 3>에서 제시한 값은 NB, SMO, IBK, RF로부터 얻은 결과의 평균값을 의미한다. 본 논문에서 제시한 SCFSB의 결과와 다른 방법론들의 결과는 분류 정확도의 절대적인 값에서 큰 차이가 난다고 말할 수 없지만 <표 4>~<표 5>의 결과를 통해 SCFSB의 우수성을 추가로 살펴볼 수 있다. <표 4>는 <표 2>의 분류 정확도를 순위 값으로 변경한 것이며 <표 5>는 각 방법론의 순위 값 16개를 이용하여 대응표본 T 검정(Paired T-test)을 수행한 결과이다. <표 5>를 살펴보면 SCFSB는 다른 모든 알고리즘에 비해 유의한 수준(P-value < 0.05)에서 우수한 분류 정확도를 나타낸다는 것을 알 수 있다.

이상에서 언급한 계산 비용과 분류 정확도 측면 이외에 SCFSB는 사용자 편의성 측면에서도 장점을 지닌다. 예를 들어 ellipsoidFN은 두 개의 파라미터를 사용하는데 격자 탐색(grid search) 기법을 바탕으로 $\alpha \in \{0.1, 0.5, 1, 2, 5, 10, 100\}$과 $C \in \{10, 100, 1000, 10000\}$의 다양한 값을 실험해봐야 한다는 문제점이 있다. WEKA의 기법들 역시 파라미터들의 기본 값이 설정되어 있지만 최적의 결과를 얻기 위해서는 데이터에 적합한 값을 찾기 위한 노력이 추가적으로 필요하다. 하지만 SCFSB는 모든 특징변수 선택 기법이 요구하는 파라미터, 즉 몇 개의 특징변수를 선택할 것인가에 관련된 파라미터만 설정해 줄 필요가 있다는 점에서 다른 방법론들과는 구별된다.

〈표 5〉 분류 정확도의 순위 값을 이용한 SCFSB 방법론의 우월성 검증

SCFSB와 비교된 특징변수 선택방법	대응표본 T 검정(Paired T-test) P-value
EllipsoidFN	4.7183E-10
Chisquare	0.0481
Filtered	0.0179
ReliefF	0.0010
SymmetricalUncert	0.0473

4. Conclusion

본 논문은 생물정보학 분야에서 최근 급격하게 증가하고 있는 오믹스 데이터로부터 특징변수를 선택하는 기법을 개발하였다. 특별히 생물정보학 분야 데이터는 대조군 데이터(예: 정상인 데이터)에 비해 실험군 데이터(예: 환자 데이터)의 이질성이 훨씬 높다는 점, 실험군 데이터의 특성을 분석하는 것이 주목적이라는 점, 데이터가 급격하게 증가·축적·공유·공개되고 있다는 점을 고려하여 생물정보학 분야 데이터를 효과적으로 다룰 수 있는 기법을 개발하였다. 구체적으로 본 논문은 실험군 데이터에 대한 군집분석을 실시하여 실험군 데이터의 특성을 잘 보유한 대표 데이터를 선택함으로써 실험군의 대표 데이터와 대조군의 전체 데이터를 효과적·효율적으로 비교할 수 있도록 하였다. 그 결과 데이터가 급격하게 증가하는 상황에서 대용량

문제의 어려움을 해결할 수 있도록 하였다. 또한 본 논문은 조합최적화의 SC 기법을 바탕으로 변수 간의 상호관계를 고려하되 전체 변수를 후보군으로 사용하여 그 중에서 최적의 특징변수를 선택할 수 있도록 하였다. 따라서 필터링 기법에 의존하거나 변수 간의 상호관계를 고려하지 않았던 기존의 방법론보다 우수한 특징변수를 선택할 수 있도록 하였다. 이때 SC 기반의 특징변수 선택기법은 기존의 다른 기법들이 여러 개의 파라미터를 포함하는 것과는 다르게 특징변수 개수와 관련된 단 하나의 파라미터만을 포함한다는 점에서 사용성이 편리하다는 장점 역시 지니고 있다. 향후 군집분석에 관한 연구, SC 해법에 관한 연구, 두 기법을 보다 효과적으로 결합하는 연구가 필요할 것으로 예상되며 급격히 데이터가 증가하고 있는 생물정보학 분야에서 유용한 도구로 사용될 수 있을 것으로 기대된다.

참 고 문 헌

[1] Alexe, G., S. Alexe, D.E. Axelrod, P.L. Hammer, and D. Weissmann, "Logical analysis of diffuse large B-cell lymphomas," *Artificial Intelligence in Medicine*, Vol.34 (2005), pp.235-267.

[2] Alexe, G., S. Alexe, D.E. Axelrod, T.O. Bonates, I.I. Lozina, M. Reiss, and P.L. Hammer, "Breast cancer prognosis by combinatorial analysis of gene expression data," *Breast Cancer Research*, Vol.8, No.4(2006), p.R41.

[3] Alexe, G., S. Alexe, L.A. Liotta, E. Petricoin, M. Reiss, and P.L. Hammer, "Ovarian cancer detection by logical analysis of proteomic data," *Proteomics*, Vol.4(2004), pp.766-783.

[4] Alexe, G., S. Alexe, P.L. Hammer, and B. Vizvari, "Pattern-based feature selections in genomics and proteomics," *Annals of Operations Research*, Vol.148(2006), pp.189-201.

[5] Apiletti, D., E. Baralis, G. Bruno, and A. Fiori, "MaskedPainter: Feature selection for microarray data analysis," *Intelligent Data Analysis*, (2012), pp.717-737.

[6] Ayers, K.L. and H.J. Cordell, "SNP selection in genome-wide and candidate gene studies via penalized logistic regression," *Genetic epidemiology*, Vol.34, No.8(2010), pp.879-891.

[7] Baralis, E., G. Bruno, and A. Fiori, "Maximum number of genes for microarray feature selection," *30th Annual International IEEE EMBS Conference*, 2008.

[8] Bertolazzi, P., G. Felici, P. Festa, and G. Lancia, "Logic classification and feature selection for biomedical data," *Computers and Mathematics with Applications*, (2008), pp.889-899.

[9] Boros, E., P.L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, and I. Muchnik, "An implementation of logical analysis of data," *Knowledge and Data Engineering, IEEE Transactions on*, Vol.12, No.2(2000), pp.292-306.

[10] Chvatal, V., "A greedy heuristic for the set-covering problem," *Mathematics of operations research*, Vol.4, No.3(1979), pp.233-235.

[11] Díaz-Uriarte, R. and S.A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, Vol.7, No.1(2006), p.3.

[12] Ding, C. and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, Vol.3, No.2(2005), pp.185-205.

[13] Guyon, I., J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines,"

Machine learning, Vol.46, No.1-3(2002), pp.389-422.

[14] Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The weka data mining software : an update," *ACM SIGKDD explorations newsletter*, Vol.11, No.1(2009), pp.10-18.

[15] Li, L., 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(2001), pp.1131-1142.

[16] Liu, H., J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," *Genome Informatics Series*, (2002), pp.51-60.

[17] Long, N., D. Gianola, G.J.M. Rosa, K.A. Weigel, and S. Avendano, "Machine learning classification procedure for selecting SNPs in genomic selection : application to early mortality in broilers," *Journal of animal breeding and genetics*, Vol.124, No.6(2007), pp.377-389.

[18] Model, F., P. Adorjan, A. Olek, and C. Piepenbrock, "Feature selection for DNA methylation based cancer classification," *Bioinformatics*, Vol.17, No.1(2001), pp.S157-S164.

[19] Ren, X., Y. Wang, L. Chen, X. Zhang, and Q. Jin, "ellipsoidFN : a tool for identifying a heterogeneous set of cancer biomarkers based on gene expressions," *Nucleic acids research*, Vol.41, No.4(2013), pp.e53-e53.

[20] Rubin, J., "A technique for the solution of massive set covering problems, with application to airline crew scheduling," *Transportation Science*, Vol.7, No.1(1973), pp.34-48.

[21] Saeys, Y., I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, Vol.23, No.19(2007), pp.2507-2517.

[22] Thomas, J.G., J.M. Olson, S.J. Tapscott, and L.P. Zhao, "An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles," *Genome Research*, Vol.11, No.7(2001), pp.1227-1236.

[23] Toregas, C., R. Swain, C. ReVelle, and L. Bergman, "The location of emergency service facilities," *Operations Research*, Vol.19, No.6(1971), pp.1363-1373.

[24] Wang, Z., I.C. Yuan-chin, Z. Ying, L. Zhu, and Y. Yang, "A parsimonious threshold-independent protein feature selection method through the area under receiver operating characteristic curve," *Bioinformatics*, Vol.23, No.20(2007), pp.2788-2794.

[25] Zhang, H.H., J. Ahn, X. Lin, and C. Park, "Gene selection using support vector machines with non-convex penalty," *Bioinformatics*, Vol.22, No.1(2006), pp.88-95.

[26] Zhang, X., X. Lu, Q. Shi, X. Xu, E.L. Honchiu, L.N. Harris, J.D. Iglehart, A. Miron, J.S. Liu, and W.H. Wong, "Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data," *BMC bioinformatics*, Vol.7, No.1(2006), p.197.

[27] Zhuang, J., M. Widschwendter, and A.E. Teschendorff, "A comparison of feature selection and classification methods in DNA methylation studies using the illumina infinium platform," *BMC bioinformatics*, Vol.13, No.1(2012), p.59.