DOI: 10.3745/KIPSTD.2008.15-D.6.767

순위 비교를 기반으로 하는 다양한 유전자 개수로 이루어진 암 분류 결정 규칙의 생성

윤 영 미[†] · 변 상 재[††] · 박 상 현[†††]

요 약

마이크로어레이 기술은 최근 실험적 분자생물학 분야에서 활발히 사용되고 있는 기술이다. 마이크로어레이 데이터는 한 번의 실험으로 수 만개의 유전자에 대한 발현값을 얻을 수 있으므로, 여러 질병의 발현형질을 연구하는데 매우 유용하게 사용된다. 마이크로어레이 데이터의 문제점은 참여하는 유전자의 수에 비해 참여하는 샘플(생물조직샘플)의 수가 매우 적고, 분류분석 기법을 사용하여 얻어진 분류자의 해석이 어렵다는 점이다.

본 연구에서는 위의 문제점을 해결하고자, 샘플 내 순위를 이용하여 동일한 생물학적 목적으로 수행된 공개 마이크로어레이 데이터를 통합하고, 순위 비교를 기반으로 하는 다양한 유전자 개수로 이루어진 암 분류 결정 규칙들로 이루어진 분류자를 제안한다. 본 분류자는 k개의 규칙으로 이루어진 앙상블 방법을 기반으로 하며, 하나의 규칙은 최대N개의 유전자, 관련유전자간의 순위비교 관계식, 판별클래스로 이루어져 있다. 하나의 규칙에 참여하는 유전자의 수를 다양하게 함으로써 좀더 신뢰성 높은 분류자를 생성할 수 있다. 또한 본 분류자는 생물학적 해석이 용이하며, 분류자를 구성하는 유전자를 명확히 식별할 수 있고, 총 개수가 많지 않으므로 임상환경에서의 사용가능성도 생각해 볼 수 있다.

키워드 : 데이터 마이닝, 분류분석, 지식기반 데이터 마이닝, 마이크로어레이 데이터 분류 분석

Generating Rank-Comparison Decision Rules with Variable Number of Genes for Cancer Classification

Youngmi Yoon[†] · Sangjay Bien[††] · Sanghyun Park[†††]

ABSTRACT

Microarray technology is extensively being used in experimental molecular biology field. Microarray experiments generate quantitative expression measurements for thousands of genes simultaneously, which is useful for the phenotype classification of many diseases. One of the two major problems in microarray data classification is that the number of genes exceeds the number of tissue samples. The other problem is that current methods generate classifiers that are accurate but difficult to interpret.

Our paper addresses these two problems. We performed a direct integration of individual microarrays with same biological objectives by transforming an expression value into a rank value within a sample and generated rank-comparison decision rules with variable number of genes for cancer classification. Our classifier is an ensemble method which has k top scoring decision rules. Each rule contains a number of genes, a relationship among involved genes, and a class label. Current classifiers which are also ensemble methods consist of k top scoring decision rules. However these classifiers fix the number of genes in each rule as a pair or a triple. In this paper we generalized the number of genes involved in each rule. The number of genes in each rule is in the range of 2 to N respectively. Generalizing the number of genes increases the robustness and the reliability of the classifier for the class prediction of an independent sample. Also our classifier is readily interpretable, accurate with small number of genes, and shed a possibility of the use in a clinical setting.

Keywords : Data Mining, Classification, Knowledge-Based Data Mining, Microarray Data Analysis, Microarray Data Classification

1. 서 론

마이크로어레이 기술의 발달로 한 실험에서 대량의 유전자의 발현값을 총체적으로 측정할 수 있게 되었다. 마이크로어레이는 작은 고형체 기판 위에 염기서열을 알고 있는 수 만개의 DNA를 고밀도로 집적한 것이다. 마이크로어레이 데이터는 아래 (그림 1)과 같은 형태를 갖는다. 각각의 행은 하나의 유전자를, 각각의 열은 하나의 샘플을, 하나의 셀 값은 특정 유전자의 특정 샘플에서의 발현값을 의미한다. 하나의 샘플은 유전자 집합으로 이루어지며, 각 샘플은 클래스표지로서 "정상", 또는 "암"을 갖는다.

마이크로어레이 기술을 이용하여, 유전자 발현 프로파일

※ 이 논문은 2006년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. R01-2006-000-11106-0).
† 종신회원 : 가천의과학대학교 IT학과 부교수
†† 준 회 원 : 서울대학교 생물정보학전공 석사과정
††† 종신회원 : 연세대학교 컴퓨터과학과 부교수(교신저자)
논문접수 : 2008년 6월 3일
수 정 일 : 1차 2008년 11월 17일
심사완료 : 2008년 11월 17일