유전자 발현량 데이터의 클러스터링을 이용한 다중 클래스 분류 모델

김현진*, 안재균*, 박치현*, 윤영미**박상현*
*연세대학교 컴퓨터과학과
**가천의과학대학교 정보공학부
e-mail : chriskim@cs.yonsei.ac.kr

Multi-Class Classification Model Using Gene Expression Data Clustering

Hyun Jin Kim*, Jaegyoon Ahn*, Chihyun Park*, Youngmi Yoon**, Sanghyun Park*
*Dept. of Computer Science, Yonsei University
**Dept. of Information Technology, Gachon University

요 약

본 논문에서는 여러 개의 클래스가 존재할 때, 각 클래스 내에서 샘플들을 클러스터링하고 서로 다른 클래스들과 분산도를 비교하여 클러스터가 가장 겹치지 않는 유전자 쌍들을 찾는다. 각 유전자 쌍에서 테스트 샘플과 가장 가까운 클러스터를 찾음으로써 클래스를 분류하고, 최종적으로 과반수 의결(Majority vote)하여 가장 많이 분류된 클래스를 최종 클래스로 확정한다. 그 결과, 해당 모델이 여러 개의 클래스를 가진 데이터에서 다른 비교 알고리즘의 모델들보다 높은 정확도를 나타내었다.

1. 서론

일반적으로 클래스가 여러 개인 다중 클래스 분류(Multi-class classification)가 클래스가 두 개인 바이너리 클래스 분류(Binary classification)보다 어려운 작업이다. 바이너리 클래스 분류가 두 가지의 클래스만 구분하면 되는 것과 달리 다중 클래스 분류는 테스트 샘플이 주어졌을 때 여러 개의 클래스 중 가장 적합한 하나의 클래스를 선택해야 하기 때문이다. 실제 세상에는 범주가 두 가지로 나누어지는 경우보다는 세 가지 이상으로 나누어지는 경우가 훨씬 많다. 따라서 다중 클래스 분류가 바이너리 클래스 분류보다 중요하다 할 수 있다.

기존 다중 클래스 분류기(Classifier)들의 경우 바이너리 클래스 분류 방법을 다중 클래스에도 적용할 수 있게 확장한 것이 대부분이다. *k*-TSP[1], PST[2] 등이 그 예이다. 이러한 방법들은 대체로 두 가지 스키마(Scheme)를 이용한다. OVR(One Versus Rest)과 OVO(One Versus One)이다. OVR 은 클래스가 *m* 개 있을 때, 각 클래스(Single class)와 나머지 클래스(Rest classes)들로 바이너리 분류기(Binary classifier)를 m 개 만든 후 테스트 샘플이 들어오면 한 개의 클래스로 분류된 것만 가지고 클래스를 결정한다. OVO 의 경우 *m* 개 클래스로 모든 클래스에 대해 $m(m-1)/2$ 개의 쌍(Pair)을 만든 후, 그 쌍으로 바이너리 분류기를 생성한다. 그리고 테스트 샘플을 모든 쌍의 바이너리 분류기에 넣어서 가장 많이 분류된 클래스를 최종 클래스로 확정한다.

반면 본 논문의 모델은 바이너리 클래스 분류 방법을 다중 클래스 분류에 적용할 수 있도록 확장한 것이 아니라, 원칙적으로 바이너리 클래스 분류와 다중 클래스 분류를 모두 수행할 수 있다. 이는 트레이닝 샘플들을 클래스 내 클러스터링(Inner-class clustering)을 하여 그것을 다른 클래스의 클러스터링과 비교하므로써 이루어진다. 클래스 내 클러스터링은 한 클래스 내에서의 다양성을 반영하기 때문에 같은 클래스라도 샘플들이 뭉쳐있지 않고 퍼져있는 경우 유연하게 대응할 수 있다. 이렇게 각 클래스들의 클러스터들이 존재할 때 테스트 샘플은 가장 가까운 클러스터의 클래스로 결정된다.

이러한 다중 클래스 분류 모델은 생물학적 데이터로 분류 작업을 수행할 때도 유용하게 쓰일 수 있다. 예를 들어 암의 종류, 암의 병기(Tumor stage), 암의 전이도(Tumor metastasis), 그 외 클래스를 여러 개 가질 수 있는 부분에 적용할 수 있다. 본 논문에서는 암환자의 암이 아닌 전립선 세포, 암환자의 전립선 암세포, 암환자의 전이가 일어난 전립선 암세포의 세 가지 클래스로 나누어 분류를 수행하였다.

2. 다중 클래스 분류 방법 및 모델

DNA 마이크로어레이[3] 데이터를 이용하면 생명체

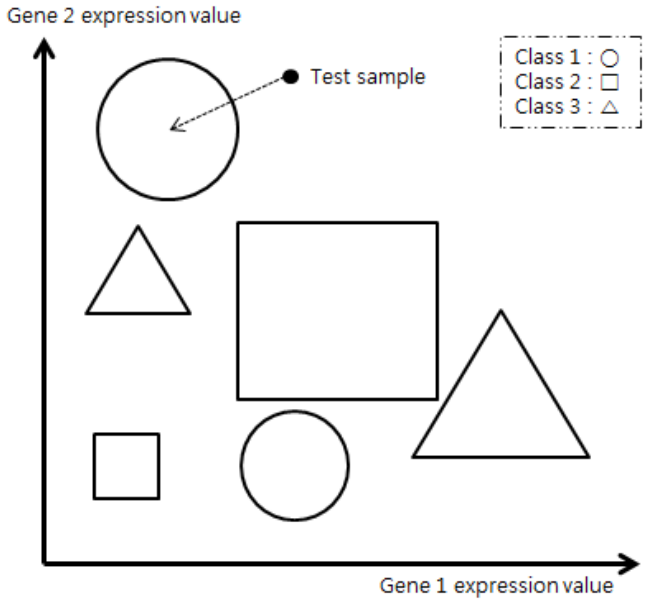
"이 논문은 2011 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(2011-0005154)"

조직 샘플의 유전자 발현 양상을 알 수 있다. 마이크로어레이 데이터는 수 천 개부터 수 만 개까지의 속성들을 가지고 있는데, 이는 기본적으로 유전자의 개수가 굉장히 많기 때문이다. 유전자의 개수가 많으면 분류 작업을 수행할 때도 더 많은 시간이 소요된다.

따라서 본 논문에서는 그러한 시간 복잡도(Time complexity)를 줄이기 위해 Symmetrical Uncertainty [4] 특성 선택(Feature selection) 방법을 사용하였다. Symmetrical Uncertainty 는 어떠한 특성이 클래스를 분류하는데 필요한 정보를 가지고 있다면 그 특성은 클래스와 강한 상관 관계, 즉 상호간 의존성이 높다는 원리를 가지고 유용한 특성을 골라낸다. 또한, Symmetrical Uncertainty 는 여러 개의 클래스를 가진 데이터에도 적용이 가능한 방법이다.

Symmetrical Uncertainty 를 사용한다면, 수 천, 수 만 개의 유전자를 수 백 개로 줄일 수 있다. 이렇게 줄인 유전자의 발현량 데이터에서 두 가지 유전자를 골라 각 클래스 별로 클래스 내 클러스터링을 수행한다. 이 때 클러스터링은 *k*-means 방법[5]을 사용하였다. 이는 주어진 데이터를 *k* 개의 클러스터로 묶는 작업으로써, 구현하기 쉽고 대체적으로 빠른 수행 시간을 보장한다.

각 클래스를 *k*-means 방법으로 클러스터링 했다면, *m* 개 클래스가 있다고 가정 할 때, 한 유전자 쌍에 대해서 총 *m* x *k* 개의 클러스터들이 존재한다. 테스트 샘플이 들어오면 해당 유전자 쌍에 대한 발현량 값들과 클러스터들의 중심 값과의 유클리디안 거리(Euclidean distance)를 비교하여 가장 가까운 클러스터의 클래스로 해당 샘플의 클래스를 예측(Prediction)한다. 예를 들어 클래스가 세 종류이고 두 개 클러스터로 클러스터링을 수행했을 때, 한 유전자 쌍에 대해 총 클러스터는 6 개이고 테스트 샘플은 아래 그림 1 과 같이 가장 가까운 클러스터의 클래스로 결정된다.



(그림 1) 테스트 샘플의 클래스 예측

지금까지는 한 유전자 쌍에서의 분류 방법을 설명하였는데, 많은 유전자 쌍들 중 어떤 유전자 쌍을 이용할 것인가도 중요하다. 특성 선택 방법을 통해 *n* 개의 유전자를 골라내었다면, 그 유전자들로 만들 수 있는 유전자 쌍은 *n*(*n*-1)/2 개이다. 이 중 테스트 샘플이 들어왔을 때 좀 더 수월하게 클래스를 예측하려면 해당 유전자 쌍에서 클러스터들 간의 거리가 서로 멀수록 유리하다. 따라서 각 유전자 쌍에서 클러스터간의 모여있는 정도를 파악하기 위해 분산도(Degree of dispersion)를 사용하였다.

분산도는 서로 다른 클래스에 속한 클러스터간 평균값의 유클리디안 거리의 합으로 구할 수 있다.

m 개 클래스 데이터를 클래스 당 *k* 개 클러스터로 클러스터링 했다고 가정할 때, 분산도 *d* 는 아래와 같이 표현할 수 있다.

$$d = \sum_{a=1}^{m}\sum_{b=1}^{k}\sum_{c=a+1}^{m}\sum_{d=1}^{k}\left(C_{ab} \# C_{cd} \right)$$

C_{ij} = *i* 번째 클래스 *j* 번째 클러스터의 평균값 좌표
$I \# J$ = 좌표 *I* 와 좌표 *J* 의 유클리디안 거리

분산도가 가장 높은 유전자 쌍이 클러스터들이 가장 퍼져있다고 볼 수 있고, 따라서 클래스를 예측할 때 더 정확하게 클래스를 확정할 수 있다.

또한 본 논문에서는 정확도를 높이기 위해, 분산도가 가장 높은 하나의 유전자 쌍에서 클래스를 예측하는 것이 아니라 높은 분산도를 가진 여러 개의 유전자 쌍 후보집합을 가지고 과반수 의결(Majority vote)하여 가장 많이 분류된 클래스를 찾는다. 이 때, 가장 많은 득표를 한 클래스가 두 가지 이상 존재한다면 샘플의 개수가 가장 많은 클래스가 최종 클래스가 된다.

3. 다중 클래스 분류 실험 결과

본 논문의 실험에서는 152 명의 전립선 세포 발현량을 분석하여 만든 마이크로어레이 데이터인 Yu[6]를 사용하였다. Yu 데이터 중에서 암 환자의 암이 아닌 전립선 세포(Prostate tissues adjacent to cancer, AT), 전립선 암환자의 전립선 세포(Prostate cancer, PC), 암환자의 전이가 일어난 전립선 암세포(Metastatic prostate cancer, MT)의 3 가지 클래스로 분류를 수행하였다. AT 는 57 개의 샘플, PC 는 62 개의 샘플, MT 는 25 개의 샘플들을 가지고 있다.

트레이닝 데이터를 가지고 분류기를 만드는데 사용한 샘플을 클래스 예측에도 사용한다면, 이미 분류기에 해당 샘플이 영향을 주고 있으므로 결과의 신뢰성이 떨어진다. 따라서 실험에는 LOOCV(Leave One Out Cross Validation) 기법을 적용하였다. LOOCV 는 교차검정(Cross Validation)의 일종으로 하나의 샘플을 테스트 샘플로 사용하고 나머지 샘플들을 트레이닝 데이터로 이용하여 분류기를 구성하는 방법이다.

실험 결과는 정확도(Accuracy)로 평가하였으며 다중

클래스 분류의 경우 바이너리 클래스 분류와 다르게 민감성(Sensitivity)과 특이성(Specificity)은 정의가 모호하여 계산에서 제외하였다. 실험의 정확도는 정확히 분류된 샘플의 개수를 전체 샘플의 개수로 나누어 구하였다.

동일한 데이터로 다른 방법들과도 비교 실험을 하였는데, 비교에 사용된 다른 알고리즘들은 *k*-Nearest neighbor[7], Decision tree[8], AdaboostM1[9], Grading[10], StackingC[11]이다.

<표 1> 다중 클래스 분류 비교 실험 결과

	정확도(Accuracy)
제안하는 방법 (Cluster=2, Vote set = 500)	**65.27 %**
k-Nearest Neighbor	61.81 %
Decision Tree	60.42 %
Adaboost M1	61.81 %
Grading	43.06 %
StackingC	43.06 %

표 1 을 보면, 본 논문에서 제안하는 방법이 다른 비교 알고리즘들보다 높은 정확도를 보이고 있음을 알 수 있다.

4. 결론

본 논문에서 제안하는 방법은 기존에 없던 새로운 다중 클래스 분류 방법(Novel approach)이다. 클래스 내 클러스터링을 통하여 클래스 내의 다양성을 반영함으로써 한 클래스 내에서 여러 가지 유형의 발현량을 가지는 경우 효과적으로 클래스를 분류할 수 있다. 또한 그 클러스터들을 다른 클래스의 클러스터들과 비교하여 다중 분류를 수행한다.

이러한 특성을 이용하여 기존 일반 세포와 암세포의 바이너리 분류(Tumor vs normal)가 아닌, 구분하기 힘든 AT, PC, MT 데이터를 다른 알고리즘들보다 높은 정확도로 분류하였다.

또한 분류기에서 분산도가 높은 유전자 쌍을 조사해본 결과, 거의 모든 유전자 쌍에서 ACTA2 라는 유전자가 빈번하게 등장하였다. ACTA2 유전자는 Alpha smooth muscle actin 이라고도 알려져 있으며 세포의 구조(Structure), 운동성(Motility), 보전 문제(Integrity) 등과 관련이 있다. 이 ACTA2 유전자는 전립선 암의 발암(Prostate carcinogenesis)과 관련이 있으며 전립선 암의 조직학적 등급(Histologic grade)와 역 상관 관계(Inversely correlation)에 있다[12].

본 논문의 방법 및 모델은 다중 클래스 분류를 높은 정확도로 수행할 수 있고, 클래스 사이에 분산도가 큰 특성 쌍(Feature pair)를 찾아내기 때문에 의미가 있는 특성을 찾아낼 수 있다.

참고문헌

[1] Aik Choon Tan et al, "Simple decision rules for classifying human cancers from gene expression profiles", Bioinformatics, Vol. 21, pp. 3896-3904, 2005.

[2] Wensheng Zhang et al, "A jackknife-like method for classification and uncertainty assessment of multi-category tumor samples using gene expression information", BMC Genomics, Vol. 11, pp. 273, 2010.

[3] Duggan et al, "Expression profiling using cDNA microarrays", Nature Genetics Supplement, vol. 21, pp.10-14, 1999.

[4] Press et al, Numerical recipes in C, Cambridge University Press, 1988.

[5] J. B. MacQueen, “Some Methods for classification and Analysis of Multivariate Observations”, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability.Berkeley, University of California Press, pp. 281-297, 1967.

[6] Yan Ping Yu et al, "Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy", Jounal of Clinical Oncology, Vol. 22, pp. 2790-2799, 2004.

[7] D. Aha et al, “Instance-based learning algorithms”, Machine Learning, Vol. 6, pp. 37-66, 1991.

[8] Ross Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[9] Yoav Freund et al, “Schapire: Experiments with a new boosting algorithm”, In: Thirteenth International Conference on Machine Learning, San Francisco, pp. 148-156, 1996.

[10] A.K. Seewald et al, “An Evaluation of Grading Classifiers”, In: Advances in Intelligent Data Analysis: 4th International Conference, pp. 115-124, 2001.

[11] A.K. Seewald et al, “How to Make Stacking Better and Faster While Also Taking Care of an Unknown Weakness”, In: Nineteenth International Conference on Machine Learning, pp. 554-561, 2002.

[12] Wong, Y. C et al, “Dedifferentiation of stromal smooth muscle as a factor in prostate carcinogenesis”, Differentiation, Vol. 70, pp. 633-645, 2002.