

K-means 클러스터링을 이용한 데이터 분류

임신자*, 윤성대*

*부경대학교 컴퓨터공학과공학과
lsja76@gmail.com, sdyoung@pknu.ac.kr

Data classification using K-means clustering

Seon-Ja Lim*, Sung-Dae Youn*

*Dept. of Computer Engineering, Pukyong National University

요 약

본 논문에서는 특징 추출 분석, 관심 영역을 추출하기 위한 몇 가지 종래의 이미지 전처리 방법과 K-means 클러스터링 및 이미지 분할방법을 통해서 얻어진 결과를 정상적인 세포와 비정상 세포를 추출하는 기법을 제안한다. 그 결과 97.8% 분류로 우수한 성능을 보여주었다.

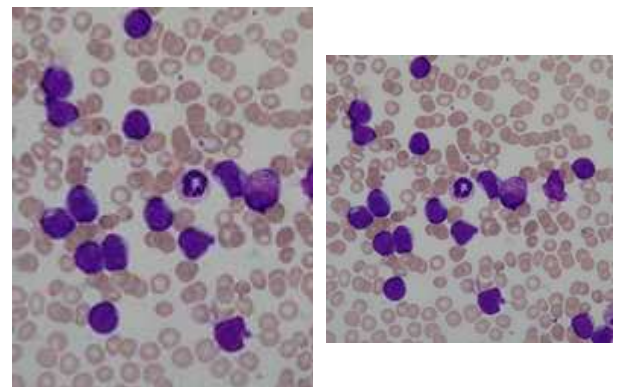
1. 서론

백혈병은 백혈구를 공격하는 치명적인 질병이다. 체내에서 백혈구는 질병 예방과 면역체계를 담당한다. 혈액세포는 골수 내부에 형성되며, 초기 단계를 미성숙 세포라고 한다. 일반적인 혈액세포와는 달리 정상적인 혈소판과 백혈구, 적혈구, 혈액 성분의 활동을 방해 한다.[1] 세포의 분화 정도에 따라 급성 백혈병, 만성백혈병으로 나뉘어진다. 세포의 기원에 따라 골수성과 림프구성으로 나뉜다. 백혈병의 분류는 4가지 급성골수성 백혈병(AML), 급성 림프성 백혈병(ALL), 만성골수성 백혈병(CML) 및 만성림프성 백혈병(CLL)이다.[1]

분류 프로세스에서 우수한 성능을 제공하는 순차적신경망 분류기를 이용하여 정상, 급성림프성 백혈병 및 급성골수성 백혈병을 정확하게 분류 한다.[2-3] 백혈병 분류에 관한 연구가 많은데, 그중 가장 일반적인 접근방식들은 이미지 사전처리, 클러스터링, 마스크 작성, 세분화, 기능선택 또는 추출, 분류 및 확인과 같은 엄격한 단계로 구성된다. 백혈병은 매우 치명적인 질병이며 생존을 향상을 위해 신속한 판단과 의사결정이 요구된다. 본 논문에서는 백혈구와 핵 영역을 추출하고, 주요 성분분석을 통해 그 특징을 추출하며, 분류 프로세스에서 97.8% 분류로 우수한 성능을 보여주었다.

2. 제안하는 방법

본 연구에서는 밀라노대학교의 Fabio Scotti교수의 데이터로 47개의 비정상 셀이미지와 53개의 일반 셀이미지를 포함하는 100개의 정규화된 이미지 샘플 이미지를 선택하였다.



(a) (b)
(그림 1) (a) 1712 × 1368 해상도의 원본 이미지, (b) 512 × 512 픽셀 해상도의 정규화 된 이미지

① 이미지 전처리

이미지 전처리는 RGB에서 CIE L * a * b로 색 공간을 변환한다.

② 특징벡터 선택 및 추출

생성된 단일핵마스크 이미지는 이제 신경망 분류기에서 입력노드로 일부기능을 추출하는 관심 영역이된다. 이 단계에서는 첫 번째 분류 프로세스의

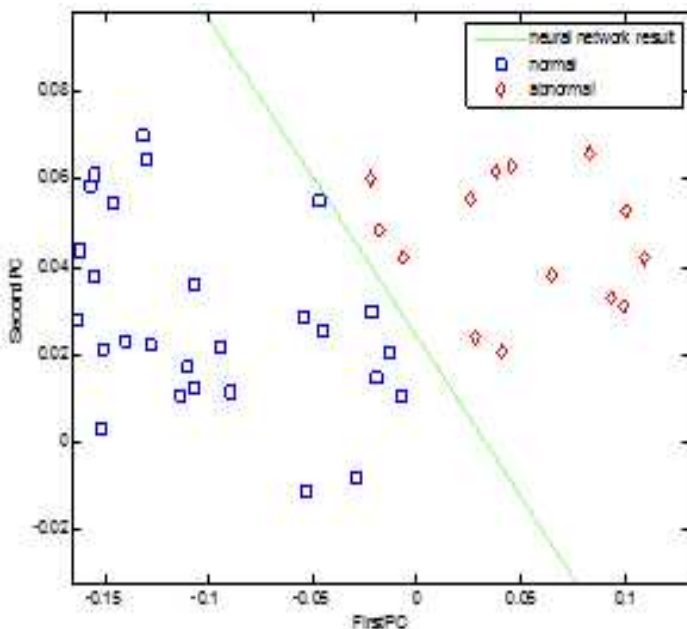
경우, 정상세포 이미지와 비정상 세포이미지만 구별하면 되므로 에너지, 엔트로피, 대비, 상관관계 및 백혈병의 프랙탈 차원과 같은 정상세포와 비정상 세포 간 차이를 보여주는 기능을 향상 시킨다

③ K-평균 군집

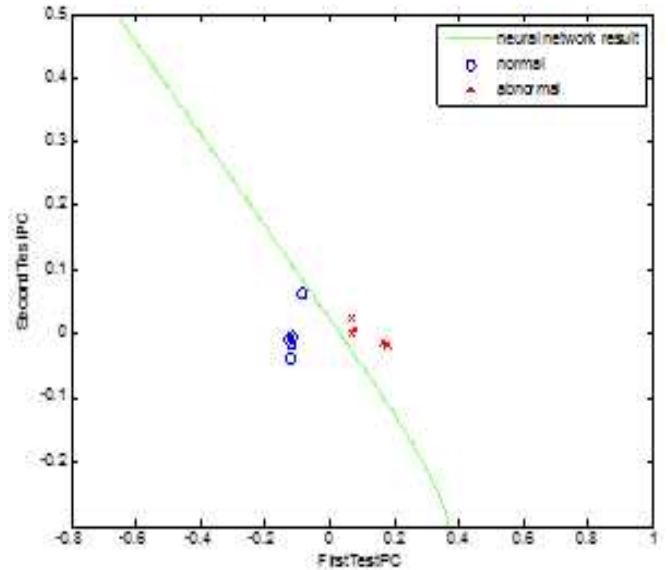
결과 이미지는 색상 정보에 따라 이미지를 3개의 다른 클래스로 분리한 다음 k-평균 군집화에 사용된다. 군집화 알고리즘과 간단한 후처리 알고리즘으로 거짓 긍정 오류를 제거한다.[4].

3. 실험결과 및 분석

100개의 훈련 이미지를 사용한 훈련 성능은 (그림 2)에서 나타난 바와 같이 검증 결과가 97.8 % 로 매우 우수하였다. 여기서 파란색 사각형은 일반 하위 클래스를 나타내며 빨간색 다이아몬드는 비정상 하위 클래스를 나타낸다. 검증절차 후에 훈련 이미지와 다른 10개의 이미지(5개 정상, 5개 비정상)훈련된 테스트 결과는 (그림 3)과 같다. 여기서 파란색 원은 정상이미지를 나타내며 빨간색 십자가는 비정상 이미지를 의미한다. 이 그림에서 분류테스트 프로세스는 소규모 테스트 데이터 세트에 대해 100% 정확도로 정상 및 비정상을 분류하여 성능이 매우 우수함을 알 수 있다



(그림 2) 훈련 이미지 검증 결과



(그림 3) 테스트 이미지 결과

4. 결론

본 논문에서는 정상, 급성림프성 백혈병 및 급성골수성 백혈병을 정확하게 분류하기 위해 100개의 훈련 이미지를 사용한 결과 97.8%와 10개의 테스트 이미지 분류 결과가 100 % 정확도를 보였다. 즉, 소규모 데이터 셋에 대해 정상세포와 비정상세포의 분류에서 성능이 매우 우수한 것으로 나타났다.

참고문헌

- [1] 위키백과, 백혈병, <https://en.wikipedia.org/wiki/백혈병>
- [2] E. Berner, Clinical Decision Support Systems: State of the Art, Agency for Healthcare Research and Quality Publication, June 2009.
- [3] V. F. R.D. Labati, "ALL-IDB: The Acute Lymphoblastic Leukemia Image Database for Image Processing," 18th IEEE International Conference on Image Processing, Vol. 18, pp. 2045-2048, Sept. 2011.
- [4] Y. K. Park, J. K. Park, H. I. On, and D. J. Kang, "Convolutional Neural Network based System for Vehicle Front Side Detection", Journal of Institute of Control, Robotics and Systems, vol. 21, no. 11, pp. 1008-1016, 2015.