Title	Feature selection with the Fisher score followed by the Maximal Clique Centrality algorithm can accurately identify the hub genes of hepatocellular carcinoma

Paper Summary

본 연구는 Hepatocellular carcinoma(HCC)에서 hub genes를 정확하게 추출하기 위해 fisher score을 포함한 5개의 다양한 feature selection method를 사용한다. 이 후 Maximal Clique Centrality(MCC) 알고리즘을 각 method에 적용하여 각 성능의 차이를 비교한 논문이다. hub genes 란 여러 개의 다른 유전자의 영향을 받아, 상호 작용을 하게 되는 유전자를 의미한다.Wisconsin breast cancer 과 같이 오래된 데이터셋은 노후 된 세포 분석 기술로 인해 데이터 의 품질이 떨어짐을 지적하였고, 이를 위해 본 연구에서는 더 많은 양의 데이터를 주기적으로 업데이트 해주는 Gene Expression Omnibus(GEO) 데이터베이스를 사용하였다. 해당 데이터로 feature selection 알고리즘을 수행한 뒤, 이미 연구된 5가지의 HCC 유전자 발현 데이터셋과의 비교를 통해 (1)feature selection 성능 평가, (2)개별 유전자에 의한 촉매, 결합 활동 수준을 나타내는 Molecular function(MFs) 과 같은 세포 수준에서의 평가, (3)MCC 알고리즘을 적용하여 상위 top 10의 hub genes 추출. 이 세 가지 케이스를 기준으로 feature selection 방식 중 filter method를 활용하여 데이터셋을 평가한다. 본 연구에서는 Filter method와 MCC 알고리즘을 사용하여 선택된 hub genes의 과발현이 Survival analysis 방법론에 따라 HCC환자의 전체 생존 시간의 감소와 유의미한 상관관계를 갖음을 보이게 된다. Survival analysis란 특정 시간 t라는 변수가 정해지고, 사망에 이르는 시점을 표현하는 T 변수 값을 통해 t보다 오래 생존할 확률을 계산해주는 함수이다. 이 함수를 기반으로 P-value값이 구해지며 해당 값이 유의수준보다 작다면 생존 기간 증가, 크다면 생존 기간 감소로 예측을 하게 된다. 본 연구에서 사용한 다양한 filter method들이 (3)에서 추출한 10개의 hub gene들을 얼마나 잘 예측했는지 시각화했다. 그 결과, Lasso 그리고 ReliefF는 모든 hub gene들이 생존 기간을 증가시키는 것으로 예측했고, 반면, WGCNA와 Random forest는 6개의 유전자들이 생존 기간을 증가시키는 걸로 예측했다. 이 6개의 유전자는 Fisher score를 기반으로 추출한 7개 유전자와 비슷한 결과를 보였다.

Criticism for Research

HCC 데이터에서 hub 유전자들을 가장 잘 표현하는 개별 유전자의 부분 집합 추출이 목적 이기 때문에, 본 연구에서는 초고차원 HCC 데이터에서 차원 축소를 위해 여러 가지 feature selection 방법 중 filter method를 선택했다. Feature selection 방법 중 가장 정확하게 변수를 선택할 수 있는 Wrapper method의 경우 모든 feature들의 조합으로 생기는 경우의 수만큼 모델을 생성한다. 본 연구에서 사용한 데이터셋은 54,613개의 feature를 가지기 때문에 wrapper method 적용 시, 시간면에서 효율이 매우 떨어지기 때문에 적합하지 않다. 반면, filter method를 통해 각 개별 유전자들을 분석하고 MCC 알고리즘으로 유전자들을 조합하여 hub genes를 추출한 본 연구의 방법론을 고려해 본다면, feature extraction을 통해서도 추출이 가능하다고 생각한다.

Fisher score 기반의 feature selection을 적용한 후, MCC 알고리즘을 적용한 결과가 PCA와 같은 수십 가지의 feature extraction 방법들보다 hub genes를 더 잘 추출할 수 있는지에 대한 연구가 더 필요하다고 생각한다.

Idea sketch

본 연구에서 다룬 Feature selection 알고리즘보다 Feature extraction 이 효과적일 수 있다고 가정하였으나, 이러한 가정에는 이유가 필요하다. feature selection은 기존 feature들로부터 개별 유전자 혹은 그 유전자들의 조합으로 hub genes를 잘 추출하는 method를 선택하는 것이다. 본 연구에서는 Fisher score를 최적의 method로 선택하고,MCC 알고리즘을 통해 hub genes를 추출하여 성능을 증명하였다. 반면에 feature extraction의 경우는 기존 feature로 새로운 feature를 구성하기 때문에 어떤 유전자에 의해 새롭게 생성된 feature인지 파악하기 어렵다는 단점이 있다. 하지만 알고리즘의 원리와 구성된 새로운 feature가 생성되기까지 기존의 유전자들의 기여 정도를 파악할 수 있다면 이 방법론이 더 좋은 성능을 낼 수도 있다. 그렇다면 유전자들의 기여 정도를 어떻게 파악할 수 있는지 알아야 한다. 먼저, feature extraction 에는 projection기법과 manifold기법으로 나뉘게 된다. Projection의 경우 PCA처럼 특정 축을 기준으로 투영시키는 기법이다. R로 PCA를 수행해보면 알 수 있듯이 새롭게 생성된 주축들에 따라 기존 feature들의 기여도를 바로 알 수 있다. 다른 예로 특이값 분해(SVD) 역시 M=Udv 공식에 따라 특이 값을 나타내는 대각행렬 d를 통해 중요도를 파악할 수 있다. 분류 알고리즘으로 클래스 사이를 가장 잘 구분하는 축으로 투영시키는 LDA 역시 판별함수식의 계수를 통해 기존 feature의 기여도를 대략적으로 파악할 수 있다. 이처럼 projection은 알고리즘의 원리를 제대로 이해하고 R에서 패키지를 사용할 줄만 안다면 기여도를 파악하는 것은 어렵지 않다. 문제는 manifold기법이다. 대표적인 예로 t-sne 가 있다. 고차원에서의 벡터들의 유사성을 저차원에서도 유지하려는 기법으로 embedding에 해당한다. 또 다른 예로 LLE가 있으며, 서로 인접한 데이터들을 보존하면서 저차원으로 embedding을 하는 알고리즘이다. 이처럼 embedding을 기반으로 하는 manifold 기법은 기존 feature들의 기여도를 파악하기 어렵기 때문에 다른 통계적 방법론을 생각해보아야 한다. 기여도를 파악했더라도 한 가지 문제가 더 존재한다. 다양한 차원 축소 알고리즘들의 결과와 MCC 알고리즘을 융합하여 사용하는 방법론에 대한 문제 역시 풀어나가야 할 과제 중 하나이다.