이산화 과정을 배재한 실수 값 인자 데이터의 고차 패턴 분석을 위한 진화연산 기반 하이퍼네트워크 모델

하정우° 장병탁
서울대학교 전기·컴퓨터공학부 바이오지능연구실
jwha@bi.snu.ac.kr, btzhaing@bi.snu.ac.kr

Evolutionary Hypernetwork Models for Higher Order Pattern Recognition on Real-valued Feature Data without Discretization

Jung-Woo Ha° Byoung-Tak Zhang
Biointelligence Labs, School of Computer Science and Engineering, Seoul National Univ.

하이퍼그래프 모델[1]은 2개 이하의 인자 간의 연관관계만 표현할 수 있는 기존 그래프 이론의 한계를 극복하기 위하여 제안된 이론이며, 하이퍼네트워크[2]는 하이퍼그래프 모델의 한 종류로서 최근 들어 텍스트, 이미지 분석 및 바이오인포매틱스 등 여러 분야에서 응용되고 있다[3]. 그러나 하이퍼네트워크를 포함한 하이퍼그래프 모델은 데이터를 구성하는 인자들의 값이 범주 형 성격인 경우에만 학습 및 모델링이 가능하다는 한계점이 있다. 그러므로 실수 인자 값으로 구성된 데이터 분석을 위해서는 별도의 전처리 과정을 통한 이산화가 선행되어야 한다. 그런데 데이터를 이산화 하는 과정에서는 필연적으로 정보손실이 발생할 수밖에 없기 때문에 이는 분류 예측 모델의 성능 저하를 유발하는 원인이 될 수 있다. 이러한 하이퍼그래프 모델의 한계점을 극복하기 위해 새로운 하이퍼네트워크 학습모델을 제안하고자 한다.

하이퍼그래프 모델에서는 기존의 그래프 모델과는 달리 그래프를 구성하는 에지(edge)가 3개 이상의 버텍스(vertex)를 동시에 연결하는 것이 가능하며, 이를 기존 그래프의 에지와 구분하기 위하여 하이퍼에지(hyper-edge)라고 부른다. 그러므로 하이퍼그래프를 이용함으로써 다수의 인자들 간의 연관관계를 하나의 하이퍼에지로 표현하는 것이 가능하다. 하이퍼그래프 모델은 버텍스와 하이퍼에지의 의미에 따라 여러 가지 형태로 적용이 가능하며 그 중 버텍스가 데이터를 구성하는 인자들과 그 값의 쌍을 의미하고 하이퍼에지가 버텍스들 간의 연관관계를 의미하는 형태의 하이퍼그래프모델도 존재하는데 이러한 하이퍼그래프 모델을 하이퍼네트워크(hypernetwork)라 부른다. 하이퍼네트워크 모델의 학습은 오른쪽 그림 1과 같은 과정으로 진행된다. 그러나 하이퍼네트워크를 포함한 기존의 하이퍼그래프 모델은 범주형(category; 카테고리) 값을 인자로 갖는 데이터에 대해서만 분석할 수 있다는 한계를 갖고 있었다. 때문에, 실수 값을 인자로 갖는 데이터를 학습시키기 위해서는 먼저 전처리를 통해서 실수 값을 카테고리 값으로 변환하기 위하여 이산화 수행해야 한다. 실수 값을 이산화(discretization)하는 방법은 이미 많은 연구가 진행되어 Fayyad & Irani나 Kononenko가 제안한 MDL(Minimum Description Length)기반의 이산화 방법 등 다양한 알고리즘이 존재한다. 그러나 실수와 이산화 된 값은 표현력의 차이로 인해 필연적으로 정보손실이 발생할 수밖에 없는데, 이러한 데이터의

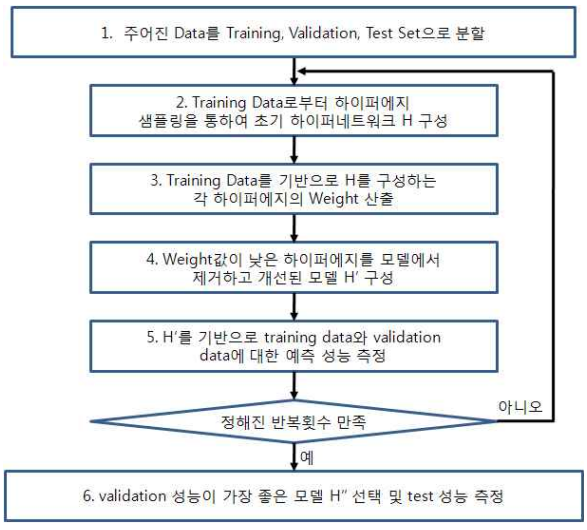

그림 1. 하이퍼네트워크 모델 학습과정의 흐름도

정보손실은 데이터를 학습하는 모델의 성능을 감소시키는 현상을 불러일으킬 수 있다. 하이퍼네트워크 모델의 학습에서 이산화가 필요한 이유는 하이퍼에지와 데이터 샘플간의 패턴 매칭을 실행함에 있어서 실수 인자 값인 경우 기존의 값이 동일한 지 여부를 이용할 수가 없다. 그러므로 본 연구에서는 각각의 인자들에 대해서 평균 L1-거리값을 구하고 이를 기준으로 하여 하이퍼에지와 데이터 샘플의 패턴 매치 여부를 결정한다. 자세한 과정은 오른쪽의 그림 2에 설명되어 있다. 제시한 모델의 성능을 확인하기 위해 본 연구에서는 UCI machine learning repository [4]에서 제공하는 유방암 데이터 (Breast Cancer Wisconsin Diagnostic Data Set)를 사용하였으며 그 결과는 아래의 표 1 및 그림 2와 같다.

```
Dist : 각 인자의 평균거리의 벡터
α : 파라미터
GetDistance(dv, value) = (dv - value)^2
PatternMatchforReal(e, d, Dist)
  d ← D의 임의의 원소 데이터 샘플
  e ← E의 임의의 원소 하이퍼에지
  for i ← 0 to n(e)
    index ← e의 i번째 인자 인덱스;
    value ← index의 인자 값
    dv ← d에서 index번째의 인자 값
    distance ← GetDistance(dv, value)
    threshold ← Dist의 index번째 거리 값
    if distance > α×threshold
      return false
  return true
```

그림 2. 개선된 패턴매칭 알고리즘

결론적으로 새롭게 제시된 거리기반의 하이퍼네트워크 모델 학습방법은 기존의 이산화 과정을 생략하고 실수 인자 데이터를 직접 학습 가능하게 함으로써 하이퍼네트워크 모델이 적용가능한 문제의 폭을 넓혔을 뿐아니라 학습 성능을 개선할 수 있음을 확인하였다.

표 1. 유방암 데이터에 대한 기계학습 방법별 분류 예측 결과. 다른 방법론은 Weka[5]사용

구 분		실수형 HyperNet	이진화 HyperNet	SVM	kNN	Decision Tree	Naive Bayes	Bayesian Net.
조 건		$n(e)$=3 / R=10	$n(e)$=3 / R=10	선형커널	k = 1	-	-	K2 / Parent=3
10 fold CV		**96.66 %**	**91.56 %**	97.71 %	95.95 %	93.32 %	92.97 %	95.08 %
Test	평 균	**95.65 %**	**91.91 %**	96.30 %	94.38 %	92.06 %	93.05 %	94.12 %
	표준편차	**2.21**	**2.53**	0.98	0.84	1.92	0.84	0.93

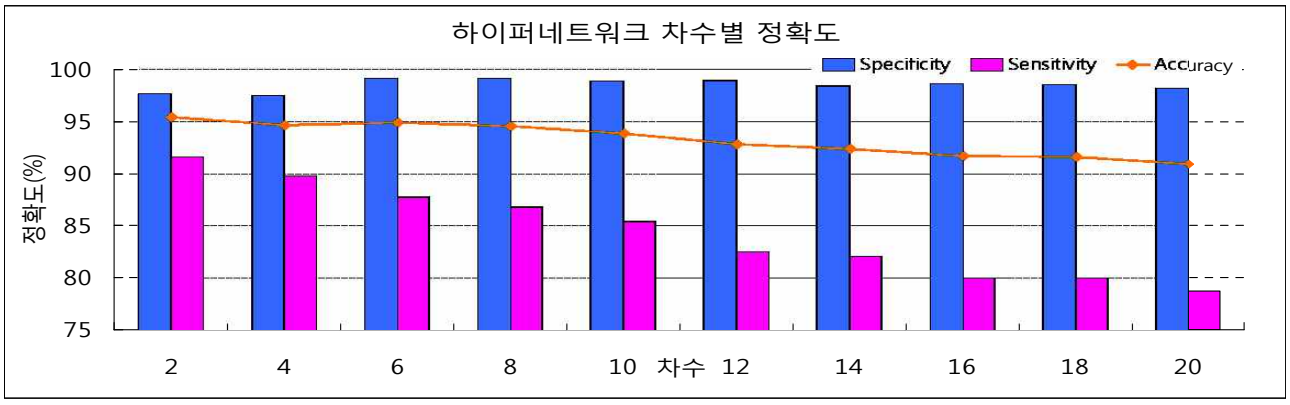

그림 9. 차수별 하이퍼네트워크 모델 학습 결과 변화 (각 차수 별 10 fold Cross validation, 10회 평균)

참고문헌

[1] D. Zhou, J. Huang, and B. Schoelkopf, Learning with hypergraphs: Clustering, classification, and embedding. *Advances in Neural Information Processing Systems (NIPS) 19*. 2007.
[2] B.-T. Zhang, Random hypergraph models of learning and memory in biomolecular networks: shorter-term adaptability vs. longer-term persistency, *The First IEEE Symp. on Foundations of Computational Intelligence (FOCI '07)*, pp. 344-349, 2007.
[3] J.-W. Ha, J.-H. Eom, S.-C. Kim, and B.-T. Zhang, Evolutionary hypernetwork models for aptamer-based cardiovascular disease diagnosis, *Workshop on Medical Applications of Genetic and Evolutionary Computation in GECCO 2007*, pp. 2709-2716, 2007.
[4] University of California, Irvine, UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/.
[5] University of Waikato, Waikato Environmental for Knowledge Analysis (WEKA) ver 3.4.11