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

(Evolutionary Hypernetwork Model for Higher Order Pattern Recognition on Real-valued Feature Data without Discretization)

하 정 우 † 장 병 탁 ††

(Jung-Woo Ha) (Byoung-Tak Zhang)

요 약 하이퍼네트워크는 하이퍼그래프의 일반화된 모델로 학습과정에 있어 진화적 개념을 도입한 확률 그래프 기반의 기계학습 알고리즘으로서 최근 들어 여러 다양한 분야에 응용되고 있다. 그러나 하이퍼네트워크 모델은 데이터와 모델을 구성하는 하이퍼에지 간의 동등비교를 기반으로 하는 학습과정의 특성상 데이터를 구성하는 인자들이 범주형인 경우에만 학습 및 모델링이 가능하고 실수 값으로 표현된 데이터를 학습하기 위해서는 이산화 등의 전처리가 선행되어야 한다는 한계점이 있다. 하지만 데이터 전처리에 있어 이산화 하는 과정은 필연적으로 정보손실이 발생할 수밖에 없기 때문에 이는 분류 예측 모델의 성능 저하를 유발하는 원인이 될 수 있다. 이러한 기존 하이퍼네트워크 모델의 한계점을 극복하기 위해 본 연구에서는 별도의 데이터 전처리 과정을 거치지 않고 실수 인자로 구성된 데이터의 패턴 학습이 가능한 개선된 하이퍼네트워크 모델을 제안한다. 여러 실험 결과를 통해 제안한 하이퍼네트워크 모델은 기존 하이퍼네트워크 모델에 비해 실수형 데이터에 대한 학습 및 분류 결과 성능이 향상되었을 뿐 아니라, 다른 여러 기계학습 방법들에 비해서도 경쟁력 있는 성능이 나타남을 확인하였다.

키워드 : 하이퍼네트워크, 하이퍼그래프, 확률 그래프모델, 실수 데이터, 기계 학습, 분류모델

Abstract A hypernetwork is a generalized hypergraph and a probabilistic graphical model based on evolutionary learning. Hypernetwork models have been applied to various domains including pattern recognition and bioinformatics. Nevertheless, conventional hypernetwork models have the limitation that they can manage data with categorical or discrete attributes only since the learning method of hypernetworks is based on equality comparison of hyperedges with learned data. Therefore, real-valued data need to be discretized by preprocessing before learning with hypernetworks. However, discretization causes inevitable information loss and possible decrease of accuracy in pattern classification. To overcome this weakness, we propose a novel feature-wise L1-distance based method for real-valued attributes in learning hypernetwork models in this study. We show that the proposed

· 본 연구는 지식경제부 및 한국산업기술평가관리원의 IT산업원천기술개발사업(2009-F-051-01, 차세대 맞춤형 서비스를 위한 기계학습 기반 멀티모달 복합 정보 추출 및 추천기술 개발), 교육인적자원부 학술진흥재단(KRF-2008-314-D00377), 교육인적자원부 BK21-IT 사업에 의해 지원되었음

· 이 논문은 2008 한국컴퓨터종합학술대회에서 '이산화 과정을 배제한 실수 값 인자 데이터의 고차 패턴 분석을 위한 진화연산 기반 하이퍼네트워크 모델'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 서울대학교 전기컴퓨터공학부
jwha@bi.snu.ac.kr

†† 종신회원 : 서울대학교 전기컴퓨터공학부 교수
btzhang@bi.snu.ac.kr

논문접수 : 2008년 8월 25일
심사완료 : 2009년 11월 30일

정보과학회논문지: 소프트웨어 및 응용 제37권 제2호(2010.2)

model improves the classification accuracy compared with conventional hypernetworks and it shows competitive performance over other machine learning methods.

Key words : Hypernetwork, Hypergraph, Probabilistic graphical model, Real-valued data, Machine learning, Classifier

1. 서 론

최근 다양한 분야의 문제 해결에 사용되는 기계학습 방법론 중에는 베이지안망이나 CRF(Conditional Random Field)와 같이 그래프를 기반으로 한 방법들이 널리 사용되고 있다. 그런데 실세계에서 발생하는 대부분의 현상은 관련된 인자들 각각이 독립적으로 영향을 끼치는 경우보다 수많은 인자들 간의 복합적 연관관계에 의하여 결정되는 경우가 훨씬 많다. 그러나 기존의 그래프는 두 개 이하의 인자들 간의 연관관계를 표현하는 데 적합하지만 세 개 이상의 연관관계를 표현하는 데는 한계가 존재하며, 비록 그래프 형태로 표현되더라도 정보의 손실이 발생하게 된다. 이러한 측면에서 하이퍼그래프(hypergraph) 모델은 실세계 문제의 표현에 있어 더 적합한 확장된 그래프 모델로 인식될 수 있다[1]. 하이퍼그래프 모델에서는 기존의 그래프 모델과는 달리 그래프를 구성하는 에지(edge)가 3개 이상의 버텍스(vertex)를 동시에 연결하는 것이 가능하며, 이를 기존 그래프에서의 에지와 구분하기 위하여 하이퍼에지(hyperedge)라고 부른다.

하이퍼네트워크(hypernetwork)는 가중치(weight) 를 갖는 하이퍼그래프로 학습과정 내에 진화적 개념이 도입된 확률 그래프 모델이다[2]. 하이퍼네트워크에서 버텍스는 데이터를 구성하는 인자들과 그 값의 쌍을 의미하고 하이퍼에지는 복수 개 버텍스들의 조합으로 표현된다[2,3]. 그러므로 하이퍼네트워크는 데이터를 구성하는 인자들의 조합의 공간을 표현하게 되며 광대한 조합의 공간을 합리적인 시간 내에 탐색하기 위하여 학습과정에서 랜덤선택기반의 진화연산 방법을 도입한다. 또한 하이퍼에지가 여러 버텍스의 조합으로 구성되기 때문에 하이퍼에지는 인자들의 연관관계를 표현가능하게 되고 그로 인해 SVM(Support Vector Machine)과 같은 커널(kernel) 기반 기계학습방법과는 달리 하이퍼네트워크는 학습된 모델을 통해 인자들 간의 연관관계 분석 및 이해가 용이하다는 장점이 있다[4,5]. 이러한 특징으로 인해 하이퍼네트워크 모델은 이미지나 텍스트 데이터 분석 뿐 아니라 마이크로어레이를 비롯한 생물학 데이터 분석 등 다양한 분야에 적용 가능하며 경쟁력 있는 성능을 보여주는 것으로 알려져 있다[4-7].

그러나 이전까지 제안된 하이퍼네트워크 모델은 범주형(category; 카테고리) 값을 인자(attribute)로 갖는 데이터에 대해서만 학습 및 모델링할 수 있다는 한계를 갖고 있었다. 이로 인해 실수 값을 인자로 갖는 데이터를 학습시키기 위해서는 먼저 실수 값을 범주형 값으로 변환하는 이산화 과정을 전처리로서 수행해야 한다. 그러나 실수와 이산화 된 값은 표현력의 차이로 인해 필연적으로 정보손실이 발생할 수밖에 없기 때문에 이러한 데이터의 정보손실은 학습하는 모델의 성능을 감소시키는 현상을 불러일으킬 수 있다. 이와 같은 기존 하이퍼네트워크의 한계를 극복하기 위하여 본 연구에서는 데이터의 전처리 과정 없이 실수 인자 형태의 데이터 패턴을 직접 학습 가능하도록 하는 개선된 하이퍼네트워크 학습 방법을 제안하고자 한다. 또한 다양한 실수 값으로 표현되는 데이터들에 대한 실험 결과를 통해 본 논문에서 제안된 하이퍼네트워크 모델이 기존의 하이퍼네트워크 모델에 비해 향상된 성능을 보여줄 뿐 아니라 다른 기계학습 방법들에 비해서도 경쟁력 있는 성능을 갖고 있다는 것을 보여준다.

이 논문의 나머지는 다음과 같이 구성된다. 2장에서는 기존의 하이퍼네트워크와 관련된 연구를 제시하고 모델의 학습 메커니즘에 대해서 설명하며, 3장에서는 실수 값을 갖는 데이터를 이용하여 모델을 학습 가능하도록 개선된 하이퍼네트워크 모델을 설명하였다. 또한 4장에서는 제시된 하이퍼네트워크 모델을 기반으로 실수 인자로 구성된 데이터에 대한 학습 결과를 분석 및 설명하였다. 마지막으로 5장에서는 차후 진행될 연구 내용을 제시하고 결론을 도출한다.

2. 하이퍼네트워크 모델

2.1 하이퍼네트워크 모델의 정의

하이퍼네트워크 H는 $H=(V, E, W)$의 형태로 표현할 수 있으며 이 때 V, E, W는 각각 하이퍼네트워크를 구성하는 버텍스 v, 하이퍼에지 e, 그리고 하이퍼에지에 해당하는 가중치 w의 집합을 의미한다. 하이퍼그래프의 정의[1]에 의해 하이퍼에지는 3개 이상의 버텍스들과 연결이 가능하며, 이로 인해 하이퍼에지는 버텍스의 집합으로 표현될 수 있다. 이 때 하이퍼에지가 연결된 버텍스의 수를 하이퍼에지 e의 'cardinality' 혹은 차수(order)라고 부르며 $n(e)$로 표현한다. 이와 같은 내용을 수식으로 표현하면 다음과 같다. m개의 인자 x와 p개의 종류를 가진 클래스 집합 Y로 구성된 n개의 데이터가 주어질 때 데이터집합을 D라 하고, l개의 하이퍼에지로 구성된 하이퍼네트워크를 H라 하면

$X = \{x_1, x_2, \ldots, x_m\}$, $Y = \{y_1, y_2, \ldots, y_p\}$,
$d_i = \{x_{1i}, x_{2i}, \ldots, x_{mi}, y_i\}$, $D = \{d_1, d_2, \ldots, d_n\}$,
$V = \{v_1, v_2, \ldots, v_m\}$, $E = \{e_1, e_2, \ldots, e_l\}$

으로 표현할 수 있고, 임의의 k개의 버텍스들로 이루어진 하이퍼에지 e, e의 차수 $n(e)$, e의 가중치 값 w는 $e = \{v_1, \ldots, v_k, w, y\}$, $n(e) = |e - \{w, y\}|$, $w = f(D)$로 정의된다. 하이퍼네트워크는 확률 그래프 모델로서 수학적으로 표현하면 아래와 같다.

$$P(D|W) = \prod_{n=1}^{N} P(x^{(n)}|W) \quad (1)$$

$$P(x^{(n)}|W) = \frac{1}{Z} exp\left[\sum_{k=1}^{K} \frac{1}{C(k)} \times \sum_{i_1, i_2, \ldots, i_k} w^{(k)} x_{i1}^{(n)} x_{i2}^{(n)} \ldots x_{ik}^{(n)}\right] \quad (2)$$

위 식에서 W는 가중치의 집합이며 하이퍼네트워크는 가중치가 주어질 때 랜덤변수로 표현되는 데이터의 부분정보들의 확률분포의 형태로 표현됨을 의미한다[2].

일반적으로 하이퍼네트워크에서 버텍스는 데이터를 구성하는 인자(attribute)와 그 값의 쌍(pair)를 의미하며 하이퍼에지는 이러한 버텍스들 간의 조합을 의미한다. 또한 하이퍼에지의 가중치 값은 각각의 하이퍼에지가 학습 데이터를 예측하는 정확도의 함수로 표현된다. 그러므로 하이퍼네트워크는 데이터를 구성하는 인자들의 조합의 공간을 표현하게 되며 하이퍼네트워크를 학습하는 과정은 조합의 공간을 탐색하는 과정으로 인식될 수 있다. 하이퍼네트워크가 표현하는 조합의 공간은 매우 넓기 때문에 모든 공간을 탐색하는 것은 실질적으로 불가능한 일이다. 그러므로 효과적인 탐색을 위하여 학습과정에 있어서 랜덤선택(random selection)을 기반으로 한 진화 연산을 도입하였다[2]. 이러한 진화연산적 특성으로 인해 연관관계가 높은 인자들의 조합으로 구성된 하이퍼에지가 살아남을 확률이 높아지게 된다. 하이퍼네트워크는 데이터의 부분정보인 하이퍼에지로 표현된다는 모델의 특성상 SVM과 같은 커널(kernel)기반 기계학습 방법이 제공하지 못하는 이점을 갖고 있다. 먼저 학습된 하이퍼네트워크에 대하여 하이퍼에지를 구성하는 인자들의 동시출현빈도(co-occurrence) 분석 등의 방법을 통하여 유의미한 인자 조합 혹은 인자들 간의 연관관계를 직접적으로 추론가능하며[4] 인자 선택(feature selection) 방법으로도 활용 가능하다[5]. 이러한 장점은 특히 분류성능 만큼 인자들의 의미 분석이 중요한 생물정보학을 비롯한 생물학 및 의학 등 여러 분야에 있어 매우 중요한 장점으로 활용될 수 있다. 또한 다양한 모달리티의 특성을 갖는 인자들을 한꺼번에 모델링 가능하므로 다양한 모달리티가 동일한 개체(object)를 표현하는 멀티미디어 데이터 모델링 및 추론에도 활용될 수 있다[7]. 위의 그림 1은 하이퍼네트워크 모델의 개념을 설명하고 있다. 다음 절에서는 이전 제안된 일반적 하이퍼네트워크 모델의 생성 및 학습 방법에 대해서 설명한다.

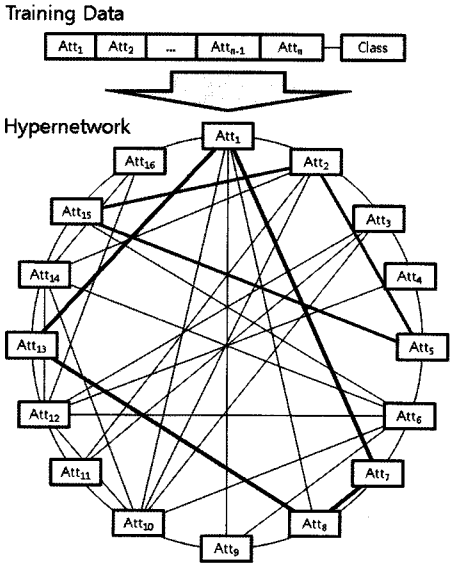


그림 1 하이퍼네트워크 모델

2.2 하이퍼네트워크 모델의 생성 및 학습 과정

하이퍼네트워크 모델에서의 학습이 이루어지는 과정은 위의 그림 2와 같다. 그림 2에서 단계 1과 단계 6은 한번만 수행되며, 단계 2에서 단계 5까지의 과정이 일정한 횟수만큼 반복 수행되는 동안 데이터에 대한 학습이 진행되며 이를 통해 최적의 하이퍼네트워크가 생성된다. 즉, 학습과정에 있어서 일련의 데이터가 주어지면 그 데이터 집합은 먼저 훈련(training), 검증(validation), 그

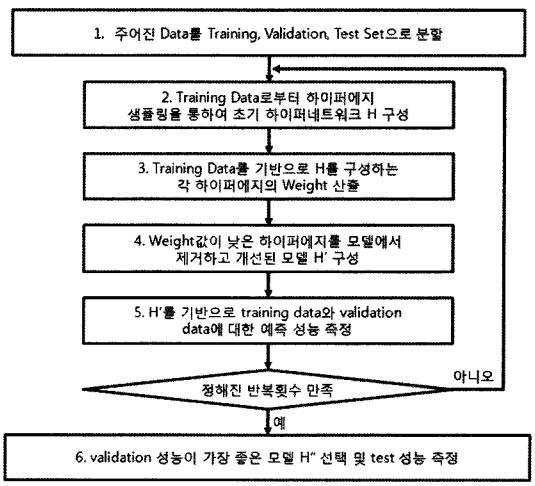


그림 2 하이퍼네트워크 모델 학습과정의 흐름도

리고 테스트(test) 데이터 집합으로 분할된다. 훈련 데이터 집합은 하이퍼네트워크 모델을 생성하기위해 하이퍼에지를 샘플링(sampling)하고 샘플링 된 하이퍼에지의 가중치 값을 산출하는 데 사용되며, 검증 데이터 집합은 반복 수행되어 생성된 여러 개의 하이퍼네트워크 모델의 예측 성능을 산출하여 가장 성능이 좋은 하이퍼네트워크 모델을 선정하는 데 사용된다. 마지막으로 테스트 데이터는 최적 하이퍼네트워크로 선정된 모델의 일반화 성능을 검증하는 데 활용된다.

2.2.1 하이퍼네트워크 모델의 생성

주어진 데이터가 일정비율의 훈련 집합 및 검증 집합, 그리고 테스트 데이터 집합으로 분할되면 훈련 데이터로부터 하이퍼에지를 샘플링함으로써 초기 하이퍼네트워크를 구성한다. 하이퍼에지는 하나의 훈련 데이터로부터 여러 번 샘플링하는 것이 가능한데 이를 우리는 'sampling rate'라고 부른다. 즉 훈련 데이터의 개수가 n개라고 하고 sampling rate값이 R이라고 하면 생성되는 하이퍼에지의 개수 $|E|$는 $n \times R$이 된다.

```
for i → 1 to n
    d ← i번째 훈련 데이터;
    for j → 1 to R
        e ← RandomSampling(d)
        E ← E∪{e};
```

그림 3 하이퍼에지 생성 알고리즘

하이퍼에지의 생성은 위의 그림 3과 같이 진행되며, 하나의 하이퍼에지가 샘플링되는 과정을 의미하는 *RandomSampling*(·)은 아래의 그림 4와 같은 과정을 따른다. 이렇게 생성된 하이퍼에지의 집합을 초기 하이퍼네트워크라 정의한다.

2.2.2 하이퍼네트워크 모델의 학습

훈련 데이터로부터 생성된 하이퍼에지들을 약한 학습자(weak learner)로 가정하면 초기 하이퍼네트워크를 구성하는 하이퍼에지들 중에는 분류 성능을 감소시키는 것들이 존재할 수 있다. 그러므로 하이퍼네트워크의 학습과정을 통해 분류 성능에 악영향을 줄 수 있는 하이퍼에지들을 제거해야 한다. 하이퍼네트워크의 학습과정은 아래 그림 5와 같이 진행된다. 그림 5에서 *PatternMatch*는 하이퍼에지가 훈련 데이터를 얼마나 정확하게 반영하고 있는 지를 측정하는 과정이다. 즉 하나의 훈련 데이터 샘플과 임의의 하이퍼에지를 구성하는 인자 값이 서로 같은 지를 확인하는 과정이다. 하이퍼에지를 구성하는 버텍스가 인자의 인덱스와 인자 값의 쌍의 형태로 되어 있기 때문에 훈련 데이터 샘플의 인자 값과 비교하는 것이 가능하며, 하이퍼에지를 구성하는 모든 버텍스에 대해서 훈련 데이터내의 인자 값과 같을 때 *PatternMatch*는 true값을 리턴 한다. 그리고 리턴 값이 true일 때에 한해서 하이퍼에지와 훈련 데이터의 클래스 값이 서로 같은 경우와 다른 경우를 각각 카운트하여 전체 훈련 데이터에 대하여 누적한 후 이를 바탕으로 하이퍼에지의 가중치 값을 산출한다. *GetWeight* 함수는 이를 구현한 부분이며, 맞게 예측한 경우가 많을수록, 틀리게 예측한 경우가 적을수록 가중치 값이 크게 책정되도록 함수를 정의한다.

```
for j → 1 to |E|
    correct ← 0;
    wrong ← 0;
    e ← 하이퍼에지 집합 E의 j번째 원소
    for k → 1 to n
        d ← k번째 훈련 데이터;
        if PatternMatch(e, d) = true
            if e의 클래스 값 = d의 클래스 값
                correct ← correct + 1;
            else
                wrong ← wrong + 1;
    w = GetWeight(correct, wrong);
/* 가중치 기반으로 하이퍼에지 필터링 */
E' = EliminateBadEgdes(E);
```

그림 5 하이퍼네트워크 모델의 학습과정

그리고 *EliminateBadEdges*는 초기 하이퍼네트워크 모델을 구성하는 하이퍼에지들의 가중치 값을 기준으로 내림차순으로 정렬하여 일정비율 만큼 제거하는 기능을 담당한다. 이를 통해 실제 데이터의 분류에 사용되는 하이퍼에지들의 집합인 E'가 만들어지며 이를 기반으로 새로운 하이퍼네트워크 모델 H'이 생성된다.

2.2.3 하이퍼네트워크 모델의 분류 및 예측 성능 측정

학습의 결과로 만들어진 하이퍼네트워크 모델 H'를 이용하여 주어진 데이터를 분류하는 과정은 학습과정과 거의 유사하다. 하나의 훈련 데이터 샘플이 주어지면 이

```
하나의 훈련 데이터 샘플이 주어질 때
1. 데이터 인자의 집합을 X라 할 때 0에서 |X|-1사이의 임의의 값을 랜덤하게 선택한다.
2. 1에서 선택된 값을 인덱스로 하여 인덱스에 해당하는 데이터의 인자 값을 저장한다.
3. 하이퍼에지의 차수만큼 인덱스가 중복되지 않도록 1과 2를 반복 수행한다.
4. 초기화된 가중치 값과 데이터의 클래스 값을 저장한다.
```

그림 4 *RandomSampling* 과정

샘플에 대하여 E'를 구성하는 모든 하이퍼에지들과 패턴 매칭(pattern matching)을 수행한다. 이 때 사용되는 방법은 2.2.2절에서 언급한 *PatternMatch*와 동일하며 그 결과가 true인 경우에 대하여 하이퍼에지의 클래스 값을 기준으로 카운트 한 후 다수결의 원칙에 의해 정해진 클래스 값을 데이터 샘플의 클래스로 분류 하며 이를 실제 데이터 샘플의 클래스 값과 비교하여 성능을 측정한다(그림 6). 또한 검증 데이터 집합에 대해서도 동일한 과정을 수행하여 성능을 측정한다. 하이퍼네트워크 H'의 성능 측정이 완료되면 한 번의 사이클이 완료된다. 그리고 모델의 학습과정에서 제거된 수만큼의 하이퍼에지를 2.2.1에서 설명한 과정을 통해 다시 생성한다. 이는 다음 학습 턴에서 제거된 만큼 다시 하이퍼에지를 생성함으로써 하이퍼네트워크의 다양성을 유지하는 역할을 한다. 이러한 과정을 정해진 횟수만큼 반복수행하고 반복 수행이 종료되면 반복하는 동안 검증 분류 성능이 가장 좋았던 하이퍼네트워크 모델을 이용하여 테스트 데이터 집합에 대한 성능을 측정한다.

```
p ← |Y|;
accuracy ← 0;
M = {m1, m2, ..., mp}; /*클래스 값별 카운트*/
for i ← 1 to |D|;
  d ← 데이터 집합 D의 i번째 샘플
  yd ← d의 클래스 값
  for j ← 1 to n
     e ← 하이퍼에지 집합 E의 j번째 원소;
     y ← e의 클래스 값;
     if PatternMatch(e, d) = true
        my ← my + 1;
  ye ← argmax mi / Σ_{i=1}^{p} mi ; /*예측된 결과값*/
  if yd = ye
     accuracy ← accuracy + 1;
```

그림 6 하이퍼네트워크 모델의 분류 및 성능 측정

3. 실수인자 데이터 모델링 하이퍼네트워크 모델

3.1 기존 모델의 한계점

기존 하이퍼네트워크 모델에서는 하이퍼에지와 데이터의 패턴매칭을 하는 데 있어서 각각 인자의 값이 동일한지 비교한다. 인자 값들 간의 동일 여부를 판단하기 위해서는 데이터가 이산화되었거나 범주 성격인 경우에만 가능하고 실수 값을 인자로 갖는 데이터의 경우에는 패턴 매칭이 불가능해진다. 그러므로 실수 인자 값을 갖는 데이터를 적용하기 위해서는 학습과 분류 과정에서 사용되는 동등비교 기반의 패턴 매칭하는 부분이 개선되어야 한다.

3.2 개선된 패턴 매칭 방법

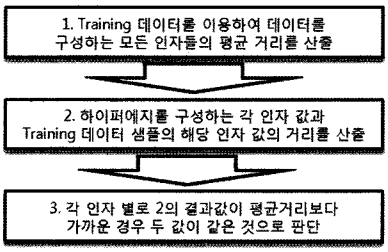


그림 7 개선된 패턴 매칭 방법

실수 값이 동일한 것인지를 판단하기 위해서 이번 연구에서 하이퍼에지와 데이터 샘플 간 각 인자간의 유사성 측정도구로서 거리를 기반으로 한 방법을 적용하였다. 위 그림 7은 패턴 매칭하는 방법에 흐름을 설명하고 있다.

하이퍼에지를 구성하는 모든 인자들에 대해서 위와 같은 과정으로 기존의 패턴 매칭 부분을 대신하게 된다. 실수 데이터 기반 패턴 매칭의 구체적인 알고리즘은 아래 그림 8과 같다. 두 인자 값들 간의 *distance*를 산출하기 위해 사용되는 거리계산 방법은 유클리드 거리(Euclidean distance), *Sigmoid* 함수 등 다양한 방법이 적용될 수가 있다. 본 연구에서는 거리를 산출하는 방법으로 두 값의 제곱 값을 측정기준으로 선택하였다. 이러한 동일한 인자 값들 간의 거리를 이용한 방법을 기하학적으로 설명하면 다음과 같다(그림 8). 첫 번째 인자와 세 번째 인자로 구성된 차수가 2인 하이퍼에지와 m 차원 벡터 형태로 주어진 임의의 데이터 샘플을 패턴 매칭을 실행한다고 가정하자. 먼저 하이퍼에지는 인자 1(*d1*)-인자3(*d3*)의 2차원 공간상의 한 점으로 표현될 수 있다. 그리고 공간상의 한 점을 중심으로 각 축별로 각 축에 해당하는 '*인자의 평균거리 (Dist(Att1))*α*2*' 크기의 2차원 영역이 형성된다. 이 때 m-차원 벡터 형태의 데이터 샘플이 주어지면 이 데이터는 2차원 공간

```
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 8 개선된 패턴매칭 알고리즘

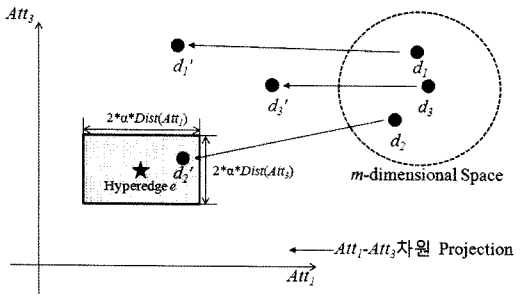


그림 9 하이퍼네트워크 패턴 매칭의 기하학적 의미

상으로 투영되는데 만약 영역 안으로 투영되면 하이퍼에지와 데이터 샘플은 매치가 되는 것이다. 그렇지 않고 영역 밖으로 투영되는 경우에는 매치가 되지 않은 것으로 이해할 수 있다. 그림 9에서 살펴보면 인자1(d_1)과 인자3(d_3)은 매치되지 않고 인자2(d_2)만 매치되는 것을 알 수 있다. 거리측정에서 적용된 파라미터 α는 영역의 크기를 조절하는 기능을 하게 되며 α값에 따라 하이퍼에지와 매치되는 데이터 샘플의 개수가 변동되므로 성능에 중요한 역할을 하게 된다.

특히 파라미터 α는 하이퍼에지의 차수와 밀접한 연관이 있다. 일반적으로 차수가 높아지는 경우 패턴매치가 될 확률은 기하급수적으로 감소하게 된다. 이로 인하여 높은 차수의 하이퍼에지들은 매치되기가 어려워진다. 이때 α값을 증가시키면 하이퍼에지가 매치될 확률을 높이는 것이 가능하게 되어 다양한 차수의 하이퍼에지로 구성된 하이퍼네트워크를 생성하는 것이 가능해진다.

3.3 성능 개선 메커니즘

본 연구에서는 하이퍼네트워크 모델의 성능을 개선하기 위하여 기존 연구에서 성능 개선에 효과가 있는 것으로 확인된 진화연산과 부스팅의 두 가지 개념을 도입하였다[6].

기존의 하이퍼네트워크 모델에서는 한번 샘플링 된 하이퍼에지의 구성은 변하지 않고 반복주기 마다 가중치 값만 지속적으로 업데이트를 해주었다. 그러나 이러한 방법은 처음 샘플링 될 때 결정된 모델 구성을 변화시키는 것이 불가능 하다는 단점을 갖고 있다. 그리하여 더 넓은 문제 공간을 탐색하기 위하여 가중치 값이 작은 하이퍼에지들을 버리고 새로운 하이퍼에지들을 생성하는 방법을 도입하였다. 이것은 하이퍼네트워크 모델 측면에서 보았을 때 하이퍼에지들에 대한 돌연변이를 구현한 것이며, 이를 통하여 하이퍼네트워크 모델의 다양성 유지하는 것이 가능해 진다.

본 연구에서 부스팅 개념을 적용하기 위해 기존 연구와 마찬가지로 훈련 데이터 집합을 두 가지 종류로 구분하였다. 즉 이전 반복 실행에서 하이퍼네트워크 모델에 의해 옳게 분류된 데이터와 잘못 분류된 데이터를 별도로 저장하고, 다음 반복 실행 시 하이퍼에지를 샘플링할 때와 가중치를 산출할 때 잘못 분류한 데이터에 가중치를 주어 분류 성능을 개선하는 것이다.

4. 실험 및 결과

4.1 데이터 및 전처리

제시한 모델의 성능을 확인하기 위해 본 연구에서는 UCI machine learning repository[8]에서 제공하는 유방암 데이터 WDBC (Breast Cancer Wisconsin Diagnostic Data Set)와 Sonar Rock vs. Mine 데이터, 그리고 RNA 압타머(Aptamer) 마이크로어레이(microarray) 기법을 이용한 질병 분류 데이터를 사용하였다. 압타머는 일종의 뉴클레오타이드 서열로서 SELEX(systematic evolution of ligands by exponential enrichment)라는 과정을 통해 만들어지는 물질이다. SELEX 과정에 의해 압타머는 특정 물질에 매우 높은 친화력을 갖게 생성되며 이로 인해 목적물질의 존재여부 확인에 유용하게 활용되고 있다[9]. 데이터의 정보는 아래 표 1과 같다.

그리고 성능 개선 여부를 확인하기 위하여 기존 하이퍼네트워크 모델을 이용하여 이산화된 데이터를 학습한 결과와 새롭게 제안된 모델을 이용하여 실수 데이터를 학습한 결과를 비교하였다. 실험에서 데이터 이산화를 위한 전처리로서 Fayyad와 Irani가 제안한 이진화 방법[10]을 사용하여 이산화를 실행하였으며, 실수 형태의 데이터는 각 인자별 평균 정규화된 값을 사용하였다. 그리고 다른 기계학습 방법들과의 성능 비교를 통하여 제시된 모델의 성능이 경쟁력 있음을 확인하였다. 또한 본 논문에서 제안하고 있는 모델의 학습에 있어 중요 역할을 하는 파라미터 α와 하이퍼네트워크의 차수 간의 관계를 분석함으로써 파라미터가 학습 성능에 어떠한 영향을 주는지 분석하였다.

표 1 실험 데이터의 구성

구분	인자 수	인자 형태	클래스 종류	샘플 수
WDBC	30	실수형	2 (양성/악성)	569 (357/212)
Sonar Rock vs. Mine	60	실수형	2 (바위/광산)	208 (104/104)
압타머 기반 질병 진단	142	실수형	8 (정상, 간암, 간염 간경화, 폐암, 심혈관1, 2, 3)	372 (55/92/42/37/30/34/32/50)

4.2 결과 및 분석

아래의 표 2는 기존 하이퍼네트워크 모델, 개선된 하이퍼네트워크 모델, Support Vector Machine(SVM), k-Nearest Neighbor(kNN), Decision Tree(C4.5), Naive Bayes, Bayesian Network을 이용하여 데이터를 분류한 정확도를 나타낸 것이다. 분류정확도는 데이터를 훈련: 검증: 테스트 = 7: 1: 2로 분할한 후 10회 실행한 후 테스트 데이터의 분류 정확도의 평균값을 계산한 것이다. 그리고 각각의 하이퍼네트워크 모델은 같은 차수를 가진 에지들로만 구성되도록 모델링 하였고 50회 반복 학습 한다. 다른 알고리즘들은 기계학습 공개 소프트웨어인 Weka 3.6.1[11]을 이용했다.

위의 표 2에서 알 수 있듯이 실수 그대로 사용한 데이터를 분석한 경우가 이진화를 수행한 경우보다 분류 정확도가 더 높게 나오는 것을 알 수 있다. 또한 본 논문에서 제안하는 개선된 하이퍼네트워크 모델은 다른 기계학습 방법들과 비교했을 때 경쟁력 있는 성능을 보여주었다.

그림 10은 WDBC데이터에 대하여 하이퍼에지의 차수에 따른 모델의 분류 정확성을 그래프로 표현한 것이다. 분석된 데이터의 경우 2개 인자의 조합으로 분석할 경우 가장 좋은 성능을 보이고 있다. 특이할만한 점은 차수가 증가함에 따라 민감도(sensitivity)가 감소하며, 이로 인해 정확도(accuracy)가 감소하는 것을 알 수 있다. 결국 양성 샘플은 상대적으로 고차 하이퍼네트워크에서 잘 분류되고 악성 샘플은 상대적으로 저차 모델에서 잘 분류됨을 알 수 있다. 이는 양성을 분류할 때 더 많은 인자를 확인하는 과정이 필요하며 악성은 중요한 소수의 인자만 확인하면 판별 가능하다는 것을 의미한다.

아래의 표 3은 하이퍼네트워크 모델은 높은 분류 성능과 함께 직관적으로 결과 해석이 가능함을 보여주는 실험 결과로 암타머 질병 진단 데이터에 대하여 하이퍼에지를 구성하는 인자들의 동시출현빈도 분석을 통해서 간과 관련된 질병과 관련정도가 높은 압타머들의 인덱스목록이다. 표의 결과에서 발현된 정도는 하이퍼에지를 구성하는 인자의 값을 데이터 전체의 인자의 샘플평균값과 비교할 때 에지를 구성하는 인자 값이 더 큰 경우에는 높게 발현된 것으로 판단하고 그 반대의 경우에는 낮게 발현된 것으로 판단한다. 하이퍼에지 분석 결과에 따르면 112번 인덱스를 갖는 압타머에 친화적인 물질은 간 질환과 높은 연관성을 가진 물질일 가능성이 높다는 것을 추측할 수 있으며 압타머 구성 분석(identification)을 통해서 어떠한 물질인지 확인하는 것이 가능하다.

다음 그림 11은 Sonar Rock vs. Mine 데이터 분류 문제에 대한 결과의 분석을 통하여 파라미터 α와 하이퍼에지의 차수간의 관계를 설명하고 있다. 그림 11(a)는

표 3 간과 관련된 질병과 관련 있는 압타머 추출 결과

질병	높게 발현된 압타머	낮게 발현된 압타머
간암	12, 31, 60, 73, 85, 89, **112**	**3**, 34, 37, 103
간염	2, **3**, 11, 17, 83, 103, **112**	12, 73, 96, 101, 123
간경화	**3**, 22, 70, 107, **112**	48, 60, 93, 133, 138

표 2 기계학습 방법별 분류 예측 결과. 하이퍼네트워크의 조건은 WDBC와 Sonar는 차수가 3 sampling rate은 10, 압타머는 차수가 12, sampling rate은 50으로 하였다.

구 분		실수형 HyperNet	기존 HyperNet	SVM	kNN	Decision Tree	Naive Bayes	Bayesian Net.
조 건		.	.	선형커널	k = 1	-	-	K2 / Parent=3
WDBC	평 균	**95.7 %**	**91.9 %**	96.3 %	94.4 %	92.1 %	93.1 %	94.1 %
	표준편차	**2.21**	**2.53**	0.98	0.84	1.92	0.84	0.93
Sonar	평 균	**87.0 %**	**84.5 %**	82.6 %	87.1 %	70.6 %	68.3 %	73.2 %
	표준편차	**3.64**	**3.49**	4.01	2.94	3.61	6.44	6.88
압타머	평 균	**89.3 %**	**81.6 %**	95.1 %	90.5 %	69.9 %	77.4 %	86.4 %
	표준편차	**3.09**	**4.41**	1.28	1.75	5.4	3.94	4.39

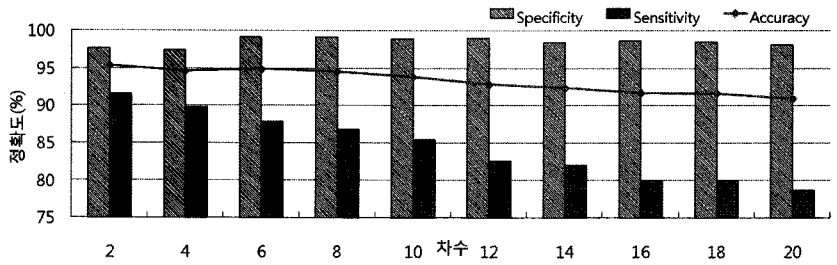
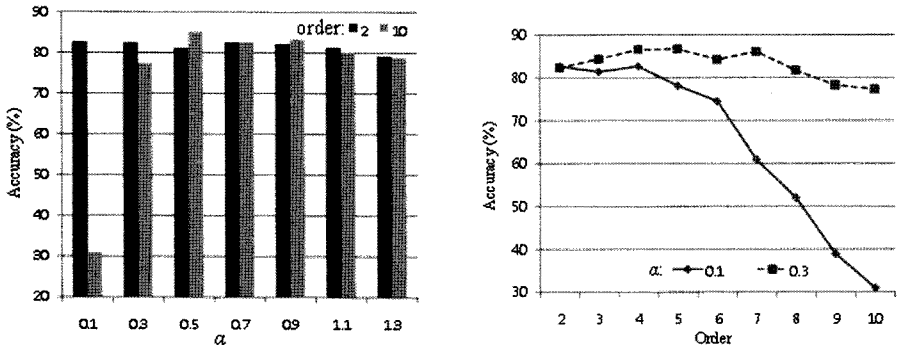


그림 10 WDBC데이터에 대한 차수 증가에 따른 하이퍼네트워크 항목별 분류 성능 변화



(a) (b)

그림 11 파라미터 α와 하이퍼네트워크 차수간의 관계 그래프

α가 증가함에 따라 낮은 차수의 하이퍼에지로 구성된 하이퍼네트워크와 높은 차수 하이퍼네트워크의 분류정확도가 어떻게 달라지는 지 보여주고 있다. 상대적으로 높은 차수의 하이퍼네트워크의 경우 α값이 작으면 분류정확도가 매우 낮은 것을 확인할 수 있다. 이는 하이퍼에지가 데이터와 매치되는 가능성이 현저히 낮아져서 분류가 불가능하기 때문이다. 그림 11(b) 또한 비슷한 맥락으로 이해될 수 있다. α값이 작은 경우(0.1) 차수가 증가함에 따라 급격히 분류 정확도가 감소하는데 비해 상대적으로 큰 α값(0.3)을 갖는 경우 분류정확도가 차수가 증가하더라도 어느 정도 유지됨을 알 수 있다. 그러므로 하이퍼네트워크의 차수에 따라 적합한 파라미터 α 값을 설정하는 것이 분류성능에 큰 영향을 주는 것을 알 수 있다.

5. 결론 및 향후과제

이 연구에서는 실수형 데이터를 직접 학습하는 하이퍼네트워크 모델을 제시함으로써 범주형 데이터만 학습 가능한 기존 모델의 한계점을 극복하였다. 제시된 모델은 기존 모델에 비해 성능이 개선되었을 뿐 아니라 다른 기계 학습 방법론들에 비해서도 경쟁력 있는 성능을 보여준다는 것을 확인하였다. 연구의 향후과제로서 모델 학습에 필요한 다양한 파라미터들에 대한 최적화된 조합을 찾는 것과 학습 속도 개선에 대한 연구가 필요하다.

참 고 문 헌

[1] Zhou, D., Huang, J., and Schoelkopf, B., "Learning with hypergraphs: Clustering, classification, and embedding," *Advances in Neural Information Processing Systems (NIPS) 19*. 2007.

[2] Zhang, B.-T., "Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory," *IEEE Computational Intelligence Magazine*, 3(3), pp.49-63, 2008.

[3] Zhang, B.-T. and Kim, J.-K., "DNA hypernetworks for information storage and retrieval," *Lecture Notes in Computer Science, DNA 12* (4287) pp. 298-307, 2006.

[4] Ha, J.-W., Eom, J.-H., Kim, S.-C., and Zhang, B.-T., "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.

[5] Ha, J.-W., Jang J. H., Kang, D.-H., Jung, W. H., Kwon, J. S., and Zhang, B.-T., "Gender Classification with Cortical Thickness Measurement from Magnetic Resonance Imaging by Using a Feature Selection Method Based on Evolutionary Hypernetworks," *2009 IEEE International conference on Fuzzy Systems (FUZZ-IEEE 2009)*, pp. 41-46, 2009.

[6] Kim, J. -K. and Zhang, B. -T., "Evolving hypernetworks for pattern classification," *IEEE Congress on Evolutionary Computation (CEC 2007)*, pp.1856-1862, 2007.

[7] Ha, J.-W., Kim, B.-H., Kim, H.-W., Yoon, W.C., Eom, J.-H., and Zhang, B.-T., "Text-to-image cross-modal retrieval of magazine articles based on higher-order pattern recall by hypernetworks," *The 10th International Symposium on Advanced Intelligent Systems (ISIS 2009)*, pp.274-277, 2009.

[8] University of California, Irvine, UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/.

[9] Jayasena, S. D. Aptamers: an emerging class of molecules that rival antibodies in diagnostics, Clinical Chemistry, 45(9), pp.1628.1650., 1999.

[10] Fayyad, U. M. and Irani, K. B., "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning," *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pp.1022-1027, 1993.

[11] University of Waikato, Waikato Environmental for Knowledge Analysis (WEKA) ver 3.6.1

하 정 우

2004년 서울대학교 전기컴퓨터공학부 학사. 2006년~현재 서울대학교 전기컴퓨터공학부 석박사통합과정. 관심분야는 기계학습, Computational Intelligence, 멀티미디어마이닝, 정보추출 및 추천, 생물정보학, 의료정보학

장 병 탁

1986년 서울대학교 컴퓨터공학 학사. 1988년 서울대학교 컴퓨터공학 석사. 1992년 독일 Bonn대학교 컴퓨터과학 박사. 1992년~1995년 독일국립정보기술연구소(GMD) 연구원. 1995년~1997년 건국대학교 컴퓨터공학과 조교수. 1997년~현재 서울대학교 전기컴퓨터공학부 교수, 인지과학, 뇌과학, 생물정보학 협동과정 겸임. 2001년~현재 서울대학교 바이오지능기술연구센터(CBIT) 센터장. 2003년~2004년 MIT CSAIL & Brain and Cognitive Sciences 방문 교수. 2005년~2006년 겨울 Bernstein Center for Comp. Neuroscience Berlin 방문 교수. 관심분야는 Biointelligence, Cognitive Machine Learning, Molecular Evolutionary Computation