J. Soc. Korea Ind. Syst. Eng Vol. 38, No. 1 : 74-82, March 2015
http://dx.doi.org/10.11627/jkise.2014.38.1.74
ISSN : 2005-0461(print)
ISSN : 2287-7975(online)

First-Order Logic Generation and Weight Learning Method in Markov Logic Network Using Association Analysis

Gil-Seung Ahn · Sun Hur†

Department of Industrial and Management Engineering, Hanyang University

연관분석을 이용한 마코프 논리네트워크의 1차 논리 공식 생성과 가중치 학습방법

안길승 · 허 선†

한양대학교 산업경영공학과

Two key challenges in statistical relational learning are uncertainty and complexity. Standard frameworks for handling uncertainty are probability and first-order logic respectively. A Markov logic network (MLN) is a first-order knowledge base with weights attached to each formula and is suitable for classification of dataset which have variables correlated with each other. But we need domain knowledge to construct first-order logics and a computational complexity problem arises when calculating weights of first-order logics. To overcome these problems we suggest a method to generate first-order logics and learn weights using association analysis in this study.

Keywords : Statistical Relational Learning, Markov Logic Network, Association Rule, Knowledge-Based Model, First-Order Logic

1. 서 론

통계적 상관 학습은 데이터에서 같은 형식을 갖는 확률변수들 사이에 존재할 수 있는 상관관계 등 복잡한 확률변수 간의 관계를 네트워크 형태로 표현하고 이를 이용하여 매개변수 및 관계형 모델에 대한 정확한 예측모형을 수립한다[4]. 마코프 논리네트워크(Markov logic network, MLN)는 이러한 통계적 상관 학습의 대표적인 모델 중 하나이다.

통계적 상관 학습을 적용할 때 직면하는 두 과제는 불확실성(uncertainty)과 복잡성(complexity)을 해결하는 것이다. 이들을 다루기 위해 사용하는 표준적인 프레임워크는 각각 확률과 1차 논리(first-order logic)이다. Richardson와 Domingos[12]는 두 프레임워크를 통합하여 확률적 그래프 모델(probabilistic graphical model)이라 할 수 있는 마코프 논리네트워크를 제시하였다. 마코프 논리네트워크는 변수 간의 관계를 나타내기 위하여 데이터를 노드(node)로, 1차 논리의 지식 베이스(first-order knowledge base)에 포함된 공식(formula)을 에지(edge)로 하는 네트워크 모델이다. 여기서 공식이란 논리기호로 표시된 서술문이며, 공식들에는 제약의 강도를 나타내는 가중치가 부여된다.

마코프 논리네트워크는 복잡한 데이터, 즉 데이터들이 서로 다양한 관계를 가지고 있을 때 적용할 수 있으며 설명력이 우수한 예측 모형이다. 또한, 확률과 1차 논리를 결합함으로써 불확실성과 복잡성이라는 단점을 동시에 보완할 수 있다. 마코프 논리네트워크는 관계 예측(link pre-

Received 26 December 2014; Finally Revised 4 February 2015; Accepted 4 February 2015
† Corresponding Author : hursun@hanyang.ac.kr

diction), 관계기반 군집화(link-based clustering), 그리고 사회연결망 분석(social network analysis) 등에 적용될 수 있다. 관계 예측은 두 객체 간에 관계가 있는가를 예측하는 작업이고, 관계기반 군집화는 연결성이 있는 데이터들을 중심으로 군집을 생성하는 작업이다. 마지막으로 사회연결망 분석은 사람을 비롯한 사회적 행위자(Social actor)와 그들 간의 관계를 구축하는 작업이다.

마코프 논리네트워크는 1차 논리 지식 베이스에 포함할 공식을 생성할 때 분석하고자 하는 해당 분야의 깊은 전문지식, 즉 도메인(domain) 지식에 크게 의존한다는 단점이 있다. 또한, 해당 1차 논리 공식의 가중치를 학습하는 데 필요한 계산량이 많고 학습된 가중치가 전역 최적해가 아닌 지역 최적해(Local optimal)에 빠질 수 있다는 단점도 있다[7].

마코프 논리네트워크의 가중치를 학습할 때 발생하는 문제를 해결하기 위한 연구로 Kok과 Domingos[7]는 관계형 데이터베이스를 하이퍼그래프(hypergraph)형태로 변형을 하여 마코프 논리네트워크를 학습하는 알고리즘(Learning via Hypergrpah Lifting, LHL)을 제시하였는데, 이때 레코드를 노드로 하고 레코드 간의 관계를 하이퍼에지(hyperedge)로 하였다. 여기서 하이퍼에지는 하나의 에지가 두 개 이상의 노드를 연결하는 에지이고, 노드들과 하이퍼에지로 구성된 네트워크를 하이퍼그래프라 한다. Lowd와 Domingos[9]는 마코프 논리네트워크를 학습하는 여러 방안을 비교, 평가하였다. Huynh과 Mooney[5]는 SVM(Support Vector Machine)의 기본 아이디어인 여백(margin)을 최대화하는 방법으로 마코프 논리네트워크의 가중치를 학습하는 방법을 제시하였다.

하지만 기존 연구는 1차 논리 공식에 부여할 가중치 학습에 관련된 계산량 줄이기와 지역해 탈피에 초점을 두었으므로 도메인 지식에 기반을 두지 않고 마코프 논리네트워크에 필요한 1차 논리를 생성하는 연구는 전무하다. 데이터 분석을 할 때 도메인 지식이 필요하지 않다면 다양한 분야에 마코프 논리네트워크를 적용할 수 있으며, 표준화된 모형을 생성할 수 있게 해 준다.본 연구에서는 이에 착안하여 연관분석(association analysis)을 통해 얻은 연관규칙을 이용하여 마코프 논리네트워크의 1차 논리 공식을 생성하고 해당 규칙의 가중치를 학습하는 방법을 제시한다.

연관분석은 방대한 데이터 속에 주목할만한 숨겨진 규칙을 찾는 데 유용한 잘 알려진 방법으로, 연관분석을 통해 생성한 규칙들을 연관규칙이라 한다. 연관규칙은 항목 사이의 관계를 조건부와 결과부로 나누어 {A(조건부) ⇒ B(결과부)}형태로 표현하고, 'A조건에 만족하는 관측치는 B그룹으로 분류된다'와 같이 해석된다[11]. 연관분석은 도메인 지식에 의존하지 않는 방법론이므로 마코프 논리네트워크에 포함될 1차 논리를 생성할 때와 해당 규칙의 가중치를 학습할 때 연관규칙을 적용한다면 도메인 지식이 필요치 않고 또한 가중치 학습에 필요한 계산량이 줄어든다. 본 연구에서 제안하는 방법을 예시하고 성능을 평가하기 위해서 실제 데이터집합(data set)을 이용하여 실험을 수행한다. 수행 결과는 대표적인 분류기법인 k-nn(k-nearest neighbor), 의사결정나무(decision tree), 판별분석(discriminant analysis)을 통한 분류결과와 비교한다.

본 연구는 다음과 같이 구성된다. 제 2장에서는 1차 논리와 마코프 네트워크에 대해 설명하고 이를 결합한 마코프 논리네트워크에 관해 설명한다. 또한 마코프 논리네트워크의 가중치를 학습하기 위한 방법에 대해 설명한다. 제 3장에서는 연관규칙에 대해 설명하고, 연관규칙을 평가하기 위한 관련 연구들을 소개한다. 제 4장에서는 연관규칙을 이용하여 마코프 논리네트워크에서의 1차 논리 규칙을 생성하고 가중치를 학습하는 방법을 제안한다. 제 5장에서는 본 연구에서 제안하는 방법의 성능을 평가하기 위해 실제 데이터에 적용하는 과정에 대해 설명한다. 마지막으로 제 6장에서 본 연구를 정리한다.

2. 마코프 논리네트워크

여기서는 연구내용을 이해하기 위해서 마코프 논리네트워크의 개념과 관련 용어들을 간단히 소개한다. 마코프 논리네트워크의 개념과 적용방법, 적용 예 등에 대한 보다 자세한 설명은 Richardson과 Domingos[12]의 연구를 참고하기 바란다.

2.1 1차 논리

명제논리는 여러 명제의 옳고 그름에 대한 결정을 다루기 위한 논리인데, 명제들을 연결하기 위하여 연산자들을 사용한다. 명제논리에서는 명제가 최소 단위이므로 명제의 내부구조에 대한 분석은 이루어질 수 없다는 한계가 있다. 예를 들어, 'A는 동물이다'와 'B는 동물이다'라는 두 명제는 완전히 별개의 사실이며, 이것으로부터 A와 B는 유사하다는 사실을 발견할 수 없다. 따라서 '모든 동물은 호흡한다'라는 명제를 추가해도, 이 세 명제로부터 A와 B는 호흡한다는 사실을 유도할 수 없다. 즉, 명제논리는 지식표현을 일반화할 수 없다는 문제가 있다[14].

1차 논리는 명제논리를 확장한 개념으로, 최소 단위로 술어(predicate)를 사용함으로써 앞서 지적한 명제논리의 문제를 해결할 수 있다. 술어는 도메인 내에서 객체의 특성(attributes) 혹은 객체 간의 관계를 나타낸다. 'A는 B이다', 'A는 B가 아니다', 'A는 B를 한다' 등의 명제에서

'B'를 술어라고 한다. 예를 들어, 친구 관계, 흡연 여부 등이 술어이다. 앞의 두 명제를 1차 논리식으로 표현하면 다음과 같다 : 'A는 동물이다 : animal(A),' 'B는 동물이다 : animal(B)'. 여기서 A와 B는 모두 animal이라는 공통된 술어의 수식을 받고 있다. 이 두 논리식에 '모든 동물은 호흡을 한다'라고 하는 1차 논리식 ∀x{animal(x) → breathe(x)}이 추가되면 breathe(A)와 breathe(B)가 모두 참임을 유도할 수 있다.

1차 논리 지식베이스는 1차 논리 형태로 나타내어진 공식들의 집합이다[3]. 여기서 공식은 ∀(모든)와 ∃(존재한다) 등의 술어 기호를 사용하여 표현한 논리구조이다. 1차 논리 공식들은 상수(constant), 변수(variable), 함수(function), 그리고 술어를 이용하여 표현한다. 상수는 관심 있는 도메인에서의 객체(object)를 나타내며, '철수', '영희' 등이 상수이다. 변수는 관심 있는 도메인의 객체의 범위를 나타내며, 소문자 x, y, z, … 등으로 나타낸다. 함수는 객체의 집합으로부터 다른 객체의 집합으로의 사상(mapping)이며, '~의 부모', '~의 친구' 등이 함수이다.

2.2 마코프 네트워크

앞서 설명한 1차 논리 지식 베이스 내의 공식 간의 결합을 네트워크 형태로 나타내기 위해 마코프 네트워크와 결합한 모델이 마코프 논리네트워크다. 마코프 네트워크(Markov network)는 변수들의 결합확률분포를 네트워크 형태로 나타내는 모델로, 방향성이 없는 그래프(undirected graph) G와 포텐셜 함수(potential function) ϕ_k로 구성된다. 그래프의 노드 i는 확률벡터 $\mathbf{X} = (X_1, X_2, \cdots, X_n)$에서 확률변수 X_i의 값을 나타내고 노드 i와 노드 j를 연결하는 에지는 확률변수 X_i와 X_j에 어떤 관계가 존재함을 의미한다. 에지의 가중치는 노드 간의 관계의 강도를 나타낸다. 색인 k는 네트워크 내에서 클리크(clique) k를 나타낸다. 클리크는 모든 노드가 완전히 연결된 부분 그래프를 나타내는데, 한 클리크 내의 노드들에 해당하는 확률변수들은 서로 의존적이며 클리크 간에는 서로 독립이다. 이러한 클리크의 특성을 이용하여 확률벡터 $\mathbf{X}$에 대한 결합확률분포는 다음과 같이 클리크의 포텐셜 함수의 곱으로 표현한다[8].

$$\mathrm{P}(\mathbf{X}=\mathbf{x}) = \frac{1}{z}\Pi_k \phi_k(\boldsymbol{x}_{\{k\}}), \tag{1}$$

여기서 $x_{\{k\}}$는 클리크 k의 상태(configuration)를 나타내는 값이며, 포텐셜 함수인 $\phi_k(\cdot)$는 사용자가 정의하는 특징함수이다. Z는 정규화 함수(normalization function)이며, 다음과 같이 계산한다.

$$\mathrm{Z} = \sum_{x \in X} \prod_k \phi_k(\boldsymbol{x}_{\{k\}}). \tag{2}$$

정규화 함수는 포텐셜 함수값들의 곱인 $\Pi_k \phi_k(\boldsymbol{x}_{\{k\}})$이 확률값의 범위(0~1)를 벗어나는 것을 방지하기 위해 사용한다. 마코프 네트워크는 대개의 경우 선형로그모형(log-linear model)로 변환하는데, 이때 포텐셜 함수는 개별 클리크의 상태의 가중합을 나타내는 함수를 사용한다.

$$\mathrm{P}(\mathbf{X}=\mathbf{x}) = \frac{1}{z}\exp(\textstyle\sum_j w_j f_j(\boldsymbol{x})). \tag{3}$$

여기서, w_j는 클리크 j에 부여된 가중치이고, f_j는 클리크 j의 상태를 나타내는 실함수이다. 식 (1)에서 보듯이, 가능한 모든 클리크에 각각의 특징을 부여할 수 있는데, 마코프 논리네트워크에서는 클리크의 상태를 논리 함수(logic function) 등을 사용하여 표현한다[2].

2.3 모델 정의

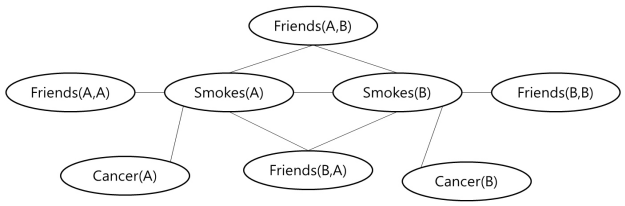
마코프 논리네트워크는 마코프 논리와 마코프 네트워크를 결합한 형태이다. 마코프 논리 L은 (F_i, w_i)의 집합으로 F_i는 1차 논리 i의 공식이며, w_i는 F_i에 실수값으로 주어지는 가중치이다. 마코프 논리네트워크는 L에 있는 모든 1차 논리 공식의 진릿값을 나타내는 이진형 노드(binary node)를 가진다. 즉, 노드의 값은 해당 공식이 참일 때 1, 아니면 0을 가진다. 이 때, 노드의 값이 1인 노드만 네트워크에 나타낸다. w_i는 F_i의 가중치로, 공식의 중요도를 반영한다. 요약하면, 각 술어의 인자에 상수를 적용한 것이 각 노드를 구성하며, 하나의 공식을 함께 구성하는 노드들끼리 마코프 논리네트워크에서 연결된다(즉, 클리크를 형성한다). 결과적으로 마코프 논리네트워크의 결합분포는 식 (3)을 바탕으로 다음과 같이 표현된다.

$$P(\mathbf{X}=\boldsymbol{x}) = \frac{1}{z}\exp(\textstyle\sum_{j=1}^{F} w_j n_j(\boldsymbol{x})). \tag{4}$$

본 연구에서 클리크 j는 1차 논리 공식 F_j를 의미하므로 n_j = (1차 논리식 F_j가 참인 훈련집합 내의 데이터의 수)로 정의된다. 예를 들어, 가중치가 각각 w_1, w_2인 두 개의 공식(F = 2)으로 구성된 마코프 논리네트워크가 있을 때, 데이터셋 $\boldsymbol{x}$가 공식 1은 r회 만족하고 공식 2는 s회 만족한다면 $n_1 = r$, $n_2 = s$가 된다.

L에 포함된 두 상수와 관련된 한 가지의 공식이라도 참이라면, 마코프 논리네트워크의 두 노드 사이에 에지를 생성한다. 예를 들어, '∀xSmokes(x) ⇒ Cancer(x)(흡연은 암을 유발한다)'와 '∀x∀yFriends(x, y) ⇒ (Smokes(x) ⇔

Smokes(y))(친구들 간에는 비슷한 흡연 습관이 있다)'라는 공식이 있고 해당 공식을 '영희(A)'와 '철수(B)'라는 상수에 적용한 마코프 논리네트워크가 있다고 하자. A가 흡연을 하므로, 'Smokes(A) ⇒ Cancer(A)'라는 공식이 참이 되고, Smokes(A)와 Cancer(A)을 연결하는 에지를 생성한다. 이러한 방식으로 마코프 논리네트워크를 도식화하면 <Figure 1>과 같다. 이러한 도식을 통해 '영희가 흡연한다면 영희의 친구인 철수 역시 흡연을 할 것이고, 결국 철수는 암에 걸릴 것이다' 등의 사실을 파악할 수 있다.



<Figure 1> Markov Logic Network Applying the Formulas

2.4 추론 및 학습

마코프 논리네트워크에서의 추론은 주어진 근거(evidence), 즉 데이터에 기반해 참인 공식(마코프 네트워크의 클리크에 해당)들의 가중치의 합을 가장 크게 만드는 진릿값 할당을 찾는 것이다[6]. 추론에 많이 사용되는 알고리즘으로는 SAT(Satisfiability) solver 알고리즘과 마코프체인 몬테 카를로(Markov chain Monte Carlo, MCMC) 방법이 있다[1].

마코프 논리네트워크 학습은 구조 및 가중치 학습으로 나눌 수 있는데, 본 논문에서는 가중치 학습 방법에 관해 관심을 두고 있으므로 가중치 학습 방법에 대해서만 언급한다. 가중치 학습은 추론과는 반대로 마코프 논리네트워크에 포함된 노드들의 진릿값이 주어졌을 때 이를 가장 잘 표현할 수 있는 가중치를 계산하는 것이다. 이는 로그 가능도(log likelihood)를 최대화하는 가중치를 구하는 것과 같은데, Hwang et al.[6]에서는 조건부 가능도에 대한 그래디언트(gradient) 방법을 제시하였지만 많은 계산량을 요구하므로 근삿값을 쓰고 있다. 이러한 문제를 해결하기 위해 본 연구에서는 계산량이 많지 않은 가중치 학습 방법을 사용한다.

3. 연관규칙

연관규칙의 중요도를 평가하기 위한 척도로 잘 알려진 지지도(S(A ⇒ B)), 신뢰도(C(A ⇒ B)) 그리고 향상도(L(A ⇒ B))는 다음과 같다.

$$\begin{aligned} \mathrm{S}(\mathrm{A}\Rightarrow\mathrm{B}) &= \Pr(\mathrm{A}\cap\mathrm{B}), \\ \mathrm{C}(\mathrm{A}\Rightarrow\mathrm{B}) &= \frac{\Pr(\mathrm{A}\cap\mathrm{B})}{\Pr(A)}, \\ L(A\Rightarrow B) &= \frac{\Pr(\mathrm{A}\cap\mathrm{B})}{\Pr(A)\times\Pr(B)} \end{aligned} \tag{5}$$

Park[10]은 각 항목집합의 주변 발생확률을 고려하여 객관적이고도 정확한 연관성 정도를 파악하기 위해 표준화된 연관성 평가기준을 다음과 같이 제시하였다.

$$S_{ST}(A\Rightarrow B) = \frac{S(A\Rightarrow B)-\max[P(A)+P(B)-1,\frac{1}{n}]}{\min[P(A),P(B),P(B|A)]-\max[P(A)+P(B)-1,\frac{1}{n}]}, \tag{6}$$

$$C_{ST}(A\Rightarrow B) = \frac{C(A\Rightarrow B)-\mathrm{L.B.C}}{\mathrm{U.B.C}-\mathrm{L.B.C}}, \tag{7}$$

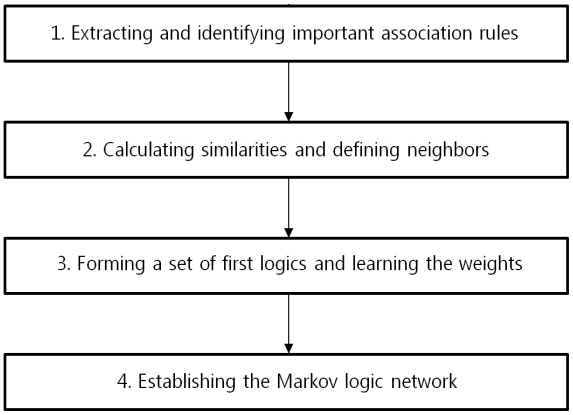
$$L_{ST}(A\Rightarrow B) = \frac{L(A\Rightarrow B)-\mathrm{L.B.L}}{\mathrm{U.B.L}-\mathrm{L.B.L}}, \tag{8}$$

식 (7)에서 U.B.C(Upper Boundary of Confidence)는 $\min\left[1,\frac{\Pr(\mathrm{B})}{\Pr(\mathrm{A})},\frac{\Pr(\mathrm{B}|\mathrm{A})}{\Pr(\mathrm{A})}\right]$, L.B.C(Lower Boundary of Confidence)는 $\max\left[1+\frac{P(B)}{P(A)}-\frac{1}{P(A)},\frac{1}{nP(A)}\right]$이며, 식 (8)에서 U.B.L(Upper Boundary of Lift)은 $\min\left[\frac{1}{\Pr(A)},\frac{1}{\Pr(B)},\frac{\Pr(B|A)}{\Pr(A)\Pr(B)}\right]$, L.B.L(Lower Boundary of Lift)은 $\max\left[\frac{1}{\Pr(b)}+\frac{1}{\Pr(A)}-\frac{1}{\Pr(A)\Pr(B)},\frac{1}{nPr(A)\Pr(B)}\right]$이다. Son과 Kim[13]은 연관규칙의 여러 평가 기준을 동시에 고려하기 위해 Weighted Support-Confidence-Lift(WSCL)이란 척도를 제안하였다. WSCL은 식 (9)와 같이 가중치, 지지도, 신뢰도, 그리고 향상도의 곱으로 나타낸다. 여기서 가중치는 사용자가 정한다.

$$\mathrm{WSCL} = \mathrm{Weight}\times\mathrm{Support}\times\mathrm{Confidence}\times\mathrm{Lift} \tag{9}$$

4. 제안 알고리즘

제 4장에서는 연관규칙을 이용한 마코프 논리네트워크에서의 1차 논리 공식 생성 및 가중치 학습 방법에 관한 구체적인 절차를 <Figure 2>에서 나타난 순서에 따라 제시한다.



<Figure 2> Procedure of Suggested Algorithm

4.1 연관규칙 생성 및 주요 연관규칙 파악

1차 논리 공식을 구성하기 위한 첫 단계로, 연관분석을 수행하여 연관규칙 목록을 작성하고 Park[10]이 제안한 표준화 방법을 이용하여 각 연관규칙의 평가기준을 표준화한다. 표준화된 평가기준인 S_{ST}, C_{ST}, L_{ST}를 바탕으로 WSCL을 계산하여 WSCL의 값이 임계치보다 큰 규칙들을 1차 논리 공식으로 선정한다.

4.2 데이터간 유사도 계산 및 이웃 정의

연관규칙의 수가 너무 적으면 모든 데이터를 커버할 수 없으며, 연관규칙의 수가 지나치게 많으면 모형의 복잡도가 증가한다. 연관규칙의 수를 적절하게 유지하면서 동시에 모든 데이터를 만족시키기 위해 데이터 간의 유사도를 계산하고 유사도가 임계치 이상인 경우를 이웃(neighbor)으로 정의하고, 이들 이웃끼리는 결과변수(target variable)의 값이 같다는 공식을 식 (10)과 같이 만들고 이를 1차 논리 지식베이스에 추가한다.

$$\forall x \forall y Ne(x, y) \Rightarrow (Cover\_Type(x) \Leftrightarrow Cover\_Type(y)). \quad (10)$$

이 때, 결과변수의 값은 다수결(majority voting)로 정한다. 동점 상황이 발생한다면 각 결과변수 값을 가지는 이웃들의 유사도를 합하여 결과변수 값을 할당한다.

4.3 1차 논리 공식 생성 및 가중치 학습

제 4.1절에서 생성한 주요 연관규칙과 4.2에서 생성한 공식 (10)을 1차 논리 공식들의 집합으로 구성한다. 이 때, 연관규칙에서 나온 공식의 가중치는 WSCL값으로 설정하고, 공식 (10)의 가중치는 연관규칙에서 나온 공식의 가중치보다 작게 설정한다. 이는 다른 모든 공식을 적용한 후 결과변수의 값을 파악할 수 없는 데이터는 이웃 관계를 이용하여 결과변수의 값을 파악하기 위함이다.

4.4 마코프 논리네트워크 구성

마지막 단계에서는 제 4.3절에서 구성한 1차 논리 공식과 가중치를 바탕으로 마코프 논리네트워크를 구성한다. 즉, 훈련 집합의 데이터에서 각각의 공식을 만족하는 경우의 수를 계산하여 식 (2)에서 제시한 정규화 함수 Z값을 계산하여 식 (3)을 완성한다.

5. 제안 알고리즘의 적용예와 결과 비교

이 장에서는 제안하는 알고리즘을 예시하고 그 성능을 판단하기 위해 실제 데이터에 적용하여 실험하고 기존의 다른 방법과 비교한다.

5.1 데이터

실험에 사용한 데이터 셋은 UCI ML Repository의 'Forest Cover Type'와 'Qualitative Bankruptcy'이다. 'Forest Cover Type'은 땅의 기울기, 고도 등의 총 54개의 설명변수를 이용하여 숲을 구성하고 있는 7가지 지형을 예측하는 것을 목적으로 하고 있으며, 관측치 개수는 총 581,012개이다. 지형이 Type 1과 Type 2인 경우가 대다수를 차지하고 있어, 지형이 Type 1과 Type 2인 데이터만을 이용하여 훈련 집합과 검증 집합을 생성한다. 훈련 집합에는 75,000개의 데이터가 포함되어 있고, 검증 집합에는 50,000개의 데이터가 포함되어 있다. 우선, 모든 변수를 범주형 변수로 변환한다. 변수의 단위가 각도(degree)이면 n(n = 1, 2, 3, 4) 사분면에 속하는 경우 범주형 변수의 n번째 값으로 변환하였다. 그 외의 경우에는 <Table 1>과 같이 변환한다. 그리고 나서 모든 범주형 변수를 이진형 변수로 변환하였다.

<Table 1> Criteria for Data transformation

Conventional value	Transformed value
Minimum value~first quartile	1
First quartile~median	2
Median~third quartile	3
Third quartile~Maximum value	4

'Qualitative Bankruptcy'는 산업 위험, 신용도 등의 총 6개의 설명변수를 이용하여 특정 회사가 파산할 것인지, 혹은 파산하지 않을 것인지를 예측하는 것을 목적으로

하고 있으며, 관측치 개수는 총 250개이다. 훈련 집합에는 총 200개의 데이터가 포함되어 있고, 검증 집합에는 총 50개의 데이터가 포함되어 있다. 모든 설명변수가 범주형(Positive, Average, Negative)이므로 데이터 변환이 필요하지 않다.

5.2 결과

5.2.1 Forest Cover Type

연관분석을 통해 주요 연관규칙을 선정하고 이를 <Table 2>와 같이 1차 논리 형태로 변환하였다. 주요 규칙을 선정하기 위한 임계치는 0.7을 사용하였다.

<Table 2> Result of Association Rule Analysis 1

Association rule	WSCL
∀x Elevation 3(x) ⇒ Cover_Type 1(x)	0.709
∀x Elevation 4(x) ⇒ Cover_Type 1(x)	1.449
∀x Wilderness_Area_1(x) ⇒ Cover_Type 2(x)	0.881
∀x Elevation 2(x) ⇒ Cover_Type 2(x)	2.087
∀xHillshade_Noon 4(x) ⇒ Cover_Type 2(x)	0.789

이 실험에서 사용한 데이터는 비대칭 이진 속성(asymmetric binary attribute)을 가지기 때문에 이웃을 정의하기 위한 두 레코드 x와 y간의 유사도는 다음의 자카드(Jaccard) 계수를 이용한다[15].

$$J(x, y) = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}, \tag{11}$$

여기서 f_{01}은 데이터 x가 0일 때 데이터 y가 1인 값을 가지는 속성의 수를, f_{10}는 데이터 x가 1일 때 데이터 y가 0인 값을 가지는 속성의 수를, f_{11}은 데이터 x가 1일 때 데이터 y도 1인 값을 가지는 속성의 수를 나타낸다. 자카드 계수가 0.9 이상인 경우를 이웃으로 정의한다. 즉, 두 노드 x와 y의 유사도가 0.9 이상이면 $\mathrm{Ne}(x, y)$이 참이 된다. 이웃인 경우는 91%의 확률로 같은 지형을 갖는다는 사실을 확인하였으며, 이를 바탕으로 식 (12)와 같은 공식을 생성하였다. 이 공식의 가중치는 다른 규칙의 가중치의 최소값인 0.79보다 적당히 작은 0.5로 설정하였는데, 이는 다른 모든 공식을 적용한 후 지형을 파악할 수 없는 데이터는 이웃 관계를 이용하여 지형을 파악하기 위함이다.

$$\forall x \forall y \mathrm{Ne}(x, y) \Rightarrow (\mathrm{Cover\_Type}(x) \Leftrightarrow \mathrm{Cover\_Type}(y)) \tag{12}$$

훈련 집합 내에서 공식이 참인 경우의 수를 계산하여 이를 n_i로 하여 식 (2)의 Z값을 계산한 후 완성된 모델은 식 (13)과 같다.

$$P(\boldsymbol{X} = \boldsymbol{x}) = \exp\left(\sum_{i=1}^{F} w_i n_i(\boldsymbol{x}) - 67326.12\right). \tag{13}$$

이 모델로 어떤 레코드의 결과변수 값(지형이 Type 1이면 $z=1$, Type 2이면 $z=2$)을 예측하는 것은 $z=1$일 때 만족하는 공식들의 가중치 합 $\sum_{i=1}^{F} w_i n_i(z=1)$과 $z=2$일 때 만족하는 공식들의 가중치 합 $\sum_{i=1}^{F} w_i n_i(z=2)$을 비교하는 것과 같다. 이와 같은 방식으로 검증 데이터에 포함된 숲의 지형을 예측하였고 그 결과를 나타내는 혼동행렬은 <Table 3>과 같다.

<Table 3> Confusion Matrix of Suggested Method Result 1

Actual/Predict	CoverType 1	CoverType 2
CoverType 1	10952	9092
CoverType 2	4999	24957

k-nn, 의사결정 나무, 판별분석을 수행하여 본 연구에서 제안하는 방법과 비교하였다. 정분류율(Accuracy)과 재현율(Recall)을 기준으로 비교한 결과는 <Table 4>와 같다.

<Table 4> Overall result1

Model	Accuracy	Recall
k-nn	60.95%	48.33%
Decision tree	69.73%	51.21%
Discriminant analysis	80.37%	61.37%
Suggested method	71.82%	54.64%

제안한 방법은 k-nn과 의사결정 나무보다 정분류율과 재현율이 높지만 판별분석보다는 낮았다. 하지만 본 연구에서 제안하는 방안은 판별분석에 비해 좋은 설명력을 가지고 있다. 예를 들어, 네트워크 형태의 도식을 통해 종속 변수가 특정한 값을 가지는 이유에 관해 설명이 가능하고 특정한 값을 가질 확률을 제시할 수 있다. 이에 반해, 판별 분석은 차원이 증가하면 도식으로 표현이 어렵다는 단점이 있다. 또한, 판별 분석은 데이터의 정규성(normality)을 가정하지만, 마코프 논리네트워크에서는 데이터에 대한 가정이 필요치 않다. 이러한 점을 고려하였을 때, 본 연구에서 제안하는 방안이 분류기로서 양호한 성능을 나타냄을 알 수 있다.

완성된 모델을 사용하는 방법을 예시하기 위하여 임의로 선정한 데이터(레코드 #18)를 적용해 본다. 레코드 #18의 속성값은 다음과 같다.

Elevation = 2, Wildness_Area = 1, HillShade_Noon = 3, Aspect = 2, Slope = 2, … , Soil_Type = 30

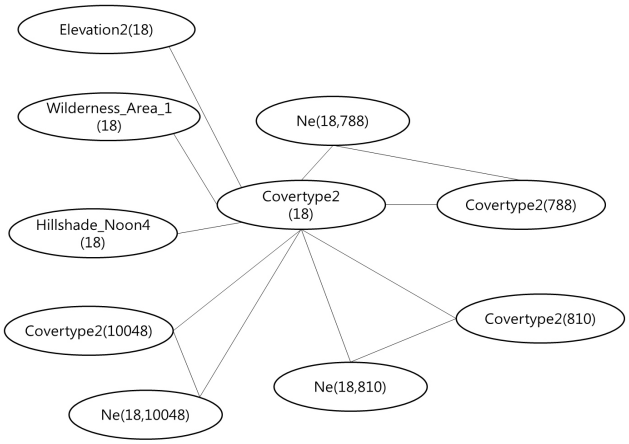
#18을 <Table 2>의 공식들과 식 (12)의 공식에 적용한 결과 만족여부는 <Table 5>와 같다. 이에 따라 $\sum_{i=1}^{F} w_i n_i (z = \text{Type 1})$과 $\sum_{i=1}^{F} w_i n_i (z = \text{Type 2})$를 비교해보자 $z = \text{Type 1}$라고 가정했을 때 만족하는 공식은 없으므로 $\sum_{i=1}^{F} w_i n_i (z = \text{Type 1}) = 0$이고, $z = \text{Type 2}$라고 가정했을 때 3, 4, 6번째 공식을 만족하므로 $\sum_{i=1}^{F} w_i n_i (z = \text{Type 2})$는 3, 4, 6번째 공식의 가중치의 합인 3.467538과 같다. 따라서 #18의 z값은 2라고 보는 것이 합당하므로 #18의 결과변수 값은 2(즉, Type 2)이다.

<Table 5> 18th Data Property

ID	Association rule	Contentment
1	∀x Elevation 3(x) ⇒ Cover_Type 1(x)	X
2	∀x Elevation 4(x) ⇒ Cover_Type 1(x)	X
3	∀x Wilderness_Area_1(x) ⇒ Cover_Type 2(x)	O
4	∀x Elevation 2(x) ⇒ Cover_Type 2(x)	O
5	∀xHillshade_Noon 4(x) ⇒ Cover_Type 2(x)	X
6	∀x∀yNe(x, y) ⇒ (Cover_Type(x) ⇔ Cover_Type(y))	O

#18을 중심으로 마코프 논리네트워크의 일부를 <Figure 3>과 같이 도식화할 수 있다. 우선 #18은 공식 3을 만족하므로, 공식 3에 포함된 Wilderness_Area_1(18) 노드와 Cover_Type 2(18) 노드를 생성한다. 같은 방법으로 공식 4, 5에 포함된 노드들을 생성한다. 공식 6을 이용하여 #18과 이웃 관계에 있는 데이터들을 나타낸다. 예를 들어, #788은 #18과 이웃 관계에 있으므로 Ne(18,788)이라는 노드를 생성한다. 한편 #788의 결과변수 값은 2이므로 Covertype 2 (18) 노드와 연결한다. 이때, 노드는 중복되지 않게 생성한다.

이렇게 만들어진 네트워크는 다음과 같이 해석할 수 있다. #18은 Wilderness_Area_1(18), Elevation2(18)와 Hillshad_Noon4(18)을 만족하므로 결과변수가 Covertype 2값을 갖는다고 해석할 수 있다. 물론, 이웃들이 Covertype 2를 만족하므로 18번 데이터 역시 Covertype 2값을 갖는다고 해석할 수도 있다. 특정 노드를 시발점으로 하여 에지를 따라 순차적인 해석 역시 가능하다. 예를 들어, #10048의 결과변수가 Covertype 2값을 가지므로 그 이웃인 #18 역시 Covertype 2값을 가지며, Covertype 2를 가지므로 Elevation 2 (18)를 만족한다고 할 수 있다. 또한, 가중치는 에지의 강도를 나타내므로, 가중치가 가장 높은 4번 공식을 만족할 확률이 다른 공식을 만족할 확률보다 높다.



<Figure 3> Part of Markov Logic Network Model

5.2.2 Qualitative Bankruptcy

연관분석을 통해 주요 연관규칙을 선정하고 이를 <Table 6>과 같이 1차 논리 형태로 변환하였다. 주요 규칙을 선정하기 위한 임계치는 1.3을 사용하였다.

<Table 6> Result of Association Rule Analysis 2

Association rule	WSCL
∀x Competitiveness = N(x) ⇒ Class_B(x)	2.011
∀x Competitiveness = P(x) ⇒ Class_NB(x)	1.764
∀x Credibility = N(x) ⇒ Class_B(x)	1.599
∀x Credibility = P(x) ⇒ Class_NB(x)	1.343

모든 설명 변수가 N(Negative), A(Average), P(Positive) 값을 가져, N을 -1로, A를 0으로, P를 1로 변형하여 데이터 맨하탄 유사도를 다음과 같이 계산하였다.

$$M(p, q) = \frac{1}{1 + \sum_{i}^{n} |p_i - q_i|} \tag{14}$$

맨하탄 유사도가 1인 경우를 이웃으로 설정하였고, 그 결과 각 레코드당 평균 이웃수는 2.97개로 나타났다. 또한, 이웃끼리 같은 클래스를 가질 확률은 1로 나타났다. 즉, 이웃인 경우에는 모두 같은 클래스를 갖는다. 이를 바탕으로 식 (15)와 같은 공식을 생성하였다. 이 공식의 가중치는 다른 규칙의 가중치의 최소값인 1.343보다 적당히 작은 1로 설정하였는데, 이는 다른 모든 공식을 적

용한 후 클래스를 파악할 수 없는 데이터는 이웃 관계를 이용하여 클래스를 파악하기 위함이다.

$$\forall x \forall y Ne(x, y) \Rightarrow (Class(x) \Leftrightarrow Class(y)) \qquad (15)$$

훈련 집합 내에서 공식이 참인 경우의 수를 계산하여 이를 n_i로 하여 식 (2)의 Z값을 계산한 후 완성된 모델은 식 (16)과 같다.

$$P(\boldsymbol{X}=\boldsymbol{x}) = \exp\left(\sum_{i=1}^{F} w_n n_i(\boldsymbol{x}) - 606.0786\right). \qquad (16)$$

이 모델로 어떤 레코드의 결과변수 값을 예측하는 것은 class=NB일 때 만족하는 공식들의 가중치 합 $\sum_{i=1}^{F} w_i n_i (class = NB)$과 class =B일 때 만족하는 공식들의 가중치 합 $\sum_{i=1}^{F} w_i n_i (class = B)$을 비교하는 것과 같다. 이를 이용하여 검증 데이터에 포함된 class를 예측하였고 그 결과를 나타내는 혼동행렬은 <Table 7>과 같다.

<Table 7> Confusion Matrix of Suggested Method Result 2

Actual/Predict	Class B	Class NB
Class B	21	0
Class NB	0	29

k-nn, 의사결정 나무, 판별분석을 수행하여 본 연구에서 제안하는 방법과 비교하였다. 정분류율과 재현율을 기준으로 비교한 결과는 <Table 8>과 같다.

<Table 8> Overall result2

Model	Accuracy	Recall
k-nn	100%	100%
Decision tree	93.5%	100%
Discriminant analysis	98%	100%
Suggested method	100%	100%

k-nn과 본 연구에서 제안하는 방법은 정확도와 재현율 모두 100%였고, 의사결정나무와 판별분석은 정확도가 각 93.5%와 98%였다. 도식화 방법과 해석 방법은 제 5.2.1절을 참고하기 바란다.

6. 결 론

본 연구에서는 마코프 논리네트워크의 1차 논리 공식을 구성하기 위하여 연관규칙을 적용하였다. 연관규칙들이 모든 데이터를 커버할 수 있도록 연관규칙의 수를 적절히 조절하기 위해 데이터 간 자카드 계수에 의한 유사도를 계산하여 이웃을 정의하였다. 연관규칙으로부터 생성된 1차 논리의 가중치를 학습하기 위해 WCSL을 사용하였다. 구성된 1차 논리를 통하여 마코프 논리네트워크를 완성하였다.

본 연구에서는 마코프 논리네트워크를 학습할 때, 발생하는 두 가지의 문제점을 해결하기 위한 새로운 방안을 제시하였다는 것에 의의가 있다. 첫 번째 문제점은 1차 논리의 지식베이스를 구성할 때 도메인 지식에 기반하여 생성해야 한다는 점과 두 번째 문제점은 1차 논리 공식의 가중치를 학습하는 데 많은 계산량이 필요하다는 점이다. 이를 해결하기 위해서 연관규칙과 이웃의 정의를 이용하여 1차 논리를 구성하고 해당 가중치를 계산하였다.

하지만 본 연구에서 제안하는 방법은 클래스 분포가 한 클래스로 치우친 데이터 셋을 사용할 때, 좋은 결과를 내지 못할 수 있다는 한계가 있다. 이는 적은 빈도의 클래스를 포함하고 있는 연관 규칙의 지지도가 낮아, 주요 연관 규칙으로 선정되지 않을 수 있기 때문이다. 또한, 유사도에 적용된 가중치가 객관적으로 설정되지 않았다는 한계점도 가지고 있다.

추후 연구과제로는 각 규칙을 만족하지 않을 때 감점을 매겨, 클래스 분포가 균일하지 않더라도 좋은 결과를 낼 수 있는 모형 개발이 필요하다. 또한, 유사도에 적용된 가중치를 설정하는 객관적인 방법을 개발하는 것이 있다.

Acknowledgements

This work was supported by the research fund of Hanyang University(HY-2014-G).

References

[1] Domingos, P. and Lowd, D., *Markov Logic : An Interface Layer for Artificial Intelligence*, Morgan and Claypool, 2009.

[2] Domingos, P., Kok, S., Lowd, D., and Poon, H., Richardson M., and Singla, P., Probabilistic Inductive Logic Programming, Springer, 2008.

[3] Geneserth, M.R. and Nilsson, N.J., Logical Foundations of Artifical Intelligence, Morgan Kaufmann, San Mateo, 1987.

[4] Getoor, L. and Tasker, B., Introduction to Statistical Relational Learning, MIT Press, Cambridge, MA, USA, 2007.

[5] Huynh, T. and Mooney, R., Max-Margin Weight Learning for Markov Logic Networks, In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2009, pp. 564-579.

[6] Hwang, K.B., Bong, S.Y., Ku, H.S., and Paek, E.O., Semantic Document-Retrieval Based on Markov Logic. *KIISE Transactions on Computing Practices*, 2010, pp. 663-667.

[7] Kok, S. and Domingos, P., Learning Markov Logic Networks Structure via Hypergraph Lifting, Procceding ICML of the 26th Annual International Conference on Machine Learning, 2009, pp. 505-512.

[8] Kolaczyk, E., Statistical Analysis of Network Data, Springer, Boston, USA, 2009.

[9] Lowd, D. and Domingos, P., Efficient Weight Learning for Markov Logic Networks, Knowledge Discovery in Database : PKDD, 2007, pp. 200-211.

[10] Park, H.C., Standardized Evaluation Method Based on the Association Rule Mining. *Journal of the Korean Data and Information Science Society*, 2010, Vol. 21, pp. 891-899.

[11] Park, J.S., You, W.K., and Hong, K.H., Association rule discovery and its application. *Journal of KIISE*, 1998, Vol. 16, pp. 37-44.

[12] Richardson, M. and Domingos, P., Markov Logic Networks. *Machine Learning*, 2005, Vol. 62, pp. 107-136.

[13] Son, J.E. and Kim, S.B., Rule Selection Method in Decision Tree Models. *Journal of the Korean Institute of Industrial Engineers*, 2014, Vol. 40, pp. 375-381.

[14] So, G.H., Symbolic logic, Kyungmoon Publishers, Seoul, Korea, 1990.

[15] Tan, P.N., Steinbach, M., and Kumar, V., Introduction to data mining, Addison Wesley, New York, 2006.

ORCID

Gil-Seung Ahn | http://orcid.org/0000-0003-1024-8897

Sun Hur | http://orcid.org/0000-0003-1832-046X