Journal of the Korean Institute of Industrial Engineers
Vol. 41, No. 4, pp. 338-343, August 2015.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2015.41.4.338

<Original Research Paper>

회귀나무 분석을 이용한 C-CRF의 특징함수 구성 방법

안길승 · 허 선†

한양대학교 산업경영공학과

Method to Construct Feature Functions of C-CRF Using Regression Tree Analysis

Gil Seung Ahn · Sun Hur

Department of Industrial and Management Engineering, Hanyang University

We suggest a method to configure feature functions of continuous conditional random field (C-CRF). Regression tree and similarity analysis are introduced to construct the first and second feature functions of C-CRF, respectively. Rules from the regression tree are transformed to logic functions. If a logic in the set of rules is true for a data then it returns the corresponding value of leaf node and zero, otherwise. We build an Euclidean similarity matrix to define neighborhood, which constitute the second feature function. Using two feature functions, we make a C-CRF model and an illustrate example is provided.

Keywords: Regression tree, Continuous Conditional Random Field(C-CRF), Feature function, Similarity

1. 서 론

Conditional Random Fields(CRF)는 Hidden Markov Model(HMM)에서 독립성 가정을 완화하여 순차적 데이터를 구분하고 분류하기 위한 확률 모형으로, 조건부 확률 $p(\boldsymbol{y}|\boldsymbol{x})$를 네트워크 형태로 모형화하여 종속변수 간의 복잡한 의존 구조를 표현하는 프레임워크를 제공한다. 즉, CRF는 확률 변수와 확률 변수 간의 상호 의존성을 나타내는 방향이 없는 확률 그래프 모형으로서 독립변수와 종속변수 간의 관계, 종속변수 간의 관계를 토대로 조건부 확률 분포를 추정한다.

CRF는 순차 데이터의 분류를 위해 고안되었으므로 여기서 사용되는 종속변수 $\boldsymbol{y}$는 순차 데이터이다. 이 종속변수를 연속 데이터로 발전시킨 것이 Continuous CRF(C-CRF)이다(Qin *et al*., 2009). C-CRF의 주요 장점은 임의적이고 상관관계가 존재하는 독립변수를 다룰 수 있는 유연성으로, C-CRF에서는 독립변수와 종속변수에 대한 가정이 따로 필요치 않다. 예를 들어, C-CRF 모형을 수립할 때 독립변수의 정규성 가정과 오차항의 iid 가정이 필요하지 않으며, 종속변수 간의 의존성이 있는 경우에 사용하기 적합한 모형화 방법이다(Ahn and Hur, 2015). 이러한 장점 덕분에 다양한 속성이 있으며 속성 간의 관계가 매우 복잡한 데이터의 모형화에 적당하다.

하지만 C-CRF는 그 구성요소인 특징 함수를 구성하는 데에 매우 전문적인 도메인 지식이나 전문가의 직관이 필요하다는 문제, 즉 모형에 전문가의 주관이 들어갈 수 있는 위험을 가지고 있다. C-CRF를 활용한 기존 연구 가운데 Radosavljevic *et al*. (2010)에서는 에어로졸의 광학 깊이(Aerosol Optical Depth, AOD)를 예측하기 위해 C-CRF를 적용하였는데, 에어로졸의 광학 깊이와 관련된 지식을 사용하여 특징함수를 구성하였다. Qin *et al*.(2009)은 문서 간의 순위를 계산하기 위해 C-CRF를 적용하였는데, 역시 도메인 지식을 사용하여 특징함수를 구성하였다.

데이터 분석에서 도메인 지식의 사용은 불가피하지만, 주관적인 판단에 의존하는 우려가 있고 누락되거나 중복되는 특징들이 있을 수 있어 객관적인 특징함수 구성방법이 필요하다. 이를 피하기 위해 데이터마이닝 기법을 적용하는 객관적인 모형을 수립한 연구가 많이 이뤄지고 있다. Lee *et al*.(2014)은 의류산업에서의 수요예측은 디자이너와 전문가의 주관에 의존

† 연락저자 : 허 선 교수, 426-791 경기도 안산시 상록구 한양대학로55 한양대학교 산업경영공학과, Tel : 031-400-5265, Fax : 031-400-5265, E-mail : hursun@hanyang.ac.kr
2014년 12월 16일 접수; 2015년 4월 17일 수정본 접수; 2015년 4월 22일 게재 확정.

하기 때문에 예측 결과에 대한 신뢰도가 떨어지며 객관적이지 못하다는 한계를 극복하기 위해 k-평균 군집화 기법과 의사결정나무를 이용하여 객관적으로 수요를 예측하는 방법을 제안하였다. Lee *et al.*(2012)은 국내 폐차 발생량에 대한 현황분석이 주관적으로 이뤄진다는 문제를 해결하기 위해, 마코프 체인을 이용하여 국내 폐차 발생량을 예측하였다.

또한, CRF에서의 특징함수를 객관적으로 구성하기 위한 여러 연구가 수행됐다. Stewart *et al.*(2008)은 CRF에서 데이터 내에 포함되는 수많은 관계를 나타내기 위해서 수천 개의 특징함수를 생성해야 한다는 문제를 해결하기 위해, 비 이진형(non-binary) 요소로 구성된 행렬 기반의 매개 변수화된 특징함수 집합 CFOE(conditional field of experts)를 제안하였다. 또한, McCallum(2002)는 문장 처리(text processing)에서 수많은 특징을 사람이 직접 추출한다는 문제(예를 들어 단어가 대문자이다)를 해결하기 위해, 접속사를 이용하여 몇 개의 특징함수로부터 수많은 특징 함수를 효율적으로 추출할 수 있는 방법론을 제안하였다.

Stewart *et al.*(2008)과 McCallum(2002)의 연구에서 제안한 특징 함수 추출 방법은 여전히 도메인 지식에 기반하여 기준이 되는 몇 개의 특징 함수를 생성해야 하므로 여전히 주관적일 수 있으며, 몇 개의 특징 함수의 성능에 따라 추출된 특징함수들 역시 큰 영향을 받을 수 있다.

Zhang and Nebel(2011)은 시뮬레이션 방법론에 기반하여 특징을 추출하는 방법을 세 단계(데이터 발생 단계, 차원 축소 단계, 특징 추출 단계)로 제안하였다. 구체적으로, 시퀀스 생성기(sequence generator)를 이용하여 CRF를 생성하는데, 이때 CRF의 특징 함수와 그 모수는 임의로 생성한다. 그 후 생성된 특징함수 중 예측 정확도를 유지시키는 선에서 성능이 우수한 특징 함수만을 선택하고, 그렇지 않은 함수를 버린다. 이 연구에서 제안한 방법은 임의적인 특징 함수와 모수를 기반으로 특징을 추출하기 때문에, 모형의 안정성을 보장하기 어렵다는 단점이 있다.

본 연구에서는 회귀나무 분석을 이용하여 C-CRF의 특징함수를 객관적이고 효과적으로 추출하는 방법을 제시한다. 회귀나무 분석을 통해 도출되는 결과는 독립변수와 종속변수 간의 관계를 규칙 기반으로 나타낼 수 있어, 객관적인 특징 함수를 생성하는데 적절한 방법이다. 구체적으로, 회귀나무 분석을 통해 생성된 규칙을 바탕으로 논리 함수를 생성하여 C-CRF의 첫 번째 특징함수에 반영한다. 또한, 데이터 간의 유사도를 계산하여, 이를 바탕으로 두 번째 특징함수를 구성한다. C-CRF의 특징함수를 구성할 때 회귀나무에서 생성한 규칙과 데이터 간 유사도를 사용하면 누락되거나 중복됨이 없이 데이터로부터 얻어진 객관적인 지식을 유도할 수 있다.

본 논문은 다음과 같이 구성되어 있다. 제 2장에서는 회귀나무 분석과 C-CRF에 관해 설명한다. 제 3장에서는 본 논문에서 제시하고자 하는 회귀나무 분석을 이용한 C-CRF의 특징함수 구성 방법에 관해 설명한다. 구체적으로, 제 3.1절에서는 회귀나무 분석을 이용한 C-CRF의 첫 번째 특징함수 구성 방법을 제시하고, 제 3.2절에서는 유사도를 이용한 C-CRF의 두 번째 특징함수 구성 방법을 제시한다. 제 4장에서는 제 3장에서 제시하는 방법을 실제 데이터에 적용하고, 그 결과를 명시한다. 마지막으로 제 5장에서는 결론을 정리한다.

2. C-CRF와 회귀나무 분석

2.1 C-CRF

Conditional Random Field(CRF)는 확률 변수와 확률 변수 간의 상호 의존성을 나타내는 방향이 없는 확률 그래프 모형으로, 조건부 확률 분포를 추정하는 데 사용한다. 이 때 조건부 확률분포는 다음과 같이 나타낸다.

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})}\exp(\sum_i(\sum_j \lambda_j g_j(y_i, y_{i+1}, \boldsymbol{x}, i) + \sum_k \mu_k f_k(y_i, \boldsymbol{x}, i)), \quad (1)$$

여기서, $g(\cdot)$는 종속변수 상호간의 관계를 나타내는 특징함수이고, $f(\cdot)$는 독립변수와 종속변수 간의 관계를 나타내는 특징함수이다. 또한, λ와 μ는 각각 특징함수 $g(\cdot)$와 $f(\cdot)$의 가중치이다. $Z(\boldsymbol{x})$는 정규화 상수로 다음과 같이 계산한다.

$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \exp(\sum_i(\sum_j \lambda_j g_j(y_i, y_{i+1}, \boldsymbol{x}, i) + \sum_k \mu_k f_k(y_i, \boldsymbol{x}, i))) \quad (2)$$

CRF는 순차 데이터의 분류를 위해 고안되었으므로 여기서 사용되는 종속변수는 순차 데이터인데, 이를 확장하여 연속 데이터의 예측을 위해 고안된 것이 C-CRF이다. 독립변수 $\boldsymbol{x} = (\boldsymbol{x_1}, \cdots, \boldsymbol{x_T})^T$와 종속변수 $\boldsymbol{y} = (y_1, \cdots, y_T)^T$에 대해 정의되는 K_1개의 특징 함수의 집합 $\{f_k(y_i, \boldsymbol{x})\}_{k=1}^{K_1}$과 독립변수 $\boldsymbol{x}$와 두 개의 관측치에 대해 정의되는 K_2개의 특징 함수의 집합 $\{g_k(y_i, y_{i+1}, \boldsymbol{x})\}_{k=1}^{K_2}$을 고려하자. 여기서 T는 훈련집합의 데이터의 수이다. C-CRF는 다음과 같은 밀도함수를 가지는 조건부 확률분포이다(Kosta *et al.*, 2013).

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})}\exp(\sum_{i=1}^{T}\sum_{k=1}^{K_1}\alpha_k f_k(y_i, \boldsymbol{x}) + \sum_{i=1}^{T}\sum_{j=1}^{T}\sum_{k=1}^{K_2}\beta_k g_k(y_i, y_{i+1}, \boldsymbol{x})). \quad (3)$$

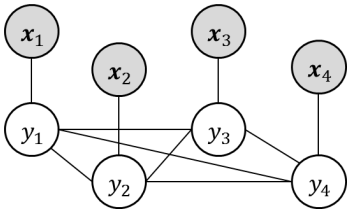

Figure 1. Continuous Conditional Random Field Model

<Figure 1>은 C-CRF 모델 내에서 독립변수와 종속변수 간의 관계를 그래프 형태로 표시한 것이다. 이 그래프에서 독립변수 $\boldsymbol{x_i}$와 종속변수 y_i를 연결하는 선은 특징함수 f를, 종속변수 간의 선은 특징함수 g를 의미한다. 즉, $f_k(y_i, \boldsymbol{x})$는 독립변수인 $\boldsymbol{x}$와 종속변수 y_i와의 의존성을 나타내는 함수이고, $g_k(y_i, y_j, \boldsymbol{x})$는 i번째 데이터와 j번째 데이터의 종속변수 간의 상호관계를 나타내는 함수이다. 이처럼 여러 특징함수를 도입하는 이유는 상호 관계의 특성이 다양할 수 있기 때문이다. $\alpha_k(k=1, 2, 3, \cdots, K_1)$와 $\beta_k(k=1, 2, 3, \cdots, K_2)$는 각 특징함수에 부여되는 모수이며, $Z(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$는 다음과 같이 계산되는 정규화 상수다.

$$Z(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int \exp\left\{\sum_{i=1}^{T}\sum_{k=1}^{K_1}\alpha_k f_k(y_i, \boldsymbol{x}) + \sum_{i=1}^{T}\sum_{j=1}^{T}\sum_{k=1}^{K_2}\beta_k g_k(y_i, y_j, \boldsymbol{x})\right\}dy. \quad (4)$$

Qin *et al.*(2009)에서는 문서들의 순위 점수를 매기기 위해 C-CRF를 사용하였는데, 이 때 문서와 관련된 요인들로는 페이지의 중요도, 특정 단어의 출현 빈도 등을 미리 정의하여 각 요인과 문서들의 순위 간의 관계를 반영하는 특징함수를 구성하였다. 이러한 일련의 작업들은 문서들의 순위점수와 관련된 지식이 없다면 사실상 불가능하다. Radosavljevic *et al.*(2010)에서는 에어로졸의 광학적 깊이(Aeroael Optical Depth)를 예측하기 위해 C-CRF를 적용하였는데, 이 때 i번째 데이터가 특정 영역에 속하면 1, 그렇지 않으면 0을 반환하는 등의 표시함수들을 결합하여 사용하였다. 이러한 표시함수를 구성하고 사용하기 위해서는 해당 분야의 전문 지식이 있어야 한다.

본 연구에서 특징함수 $f(\cdot)$는 회귀나무 분석으로 생성된 규칙을 반영하여 구성하고, 특징함수 $g(\cdot)$는 기존 연구에서 사용해 온 데이터 간의 유사도를 반영함으로써 구성한다(Qin *et al.*, 2009, Radosavljevic *et al.*, 2010). 이는 회귀나무 분석을 통해 나오는 결과는 독립변수와 종속변수 간의 관계를 나타내고, 유사도 분석을 통해 나오는 결과는 종속변수 간의 관계를 나타내기 때문이다.

각 특징함수의 가중치인 α_k와 β_k는 최대가능도(maximum likelihood)를 이용하여 추정한다. 즉, 크기가 T인 훈련 집합이 $\{\boldsymbol{x}_i, y_i\}_{i=1}^{T}$와 같이 주어졌다면 가능도 $L(\boldsymbol{\alpha}, \boldsymbol{\beta})$를 최대화하는 $\boldsymbol{\alpha}$와 $\boldsymbol{\beta}$를 추정한다. 이 때, $L(\boldsymbol{\alpha}, \boldsymbol{\beta})$는 식 (5)와 같이 계산한다.

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{N}\log P(y_i|\boldsymbol{x}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}). \quad (5)$$

2.2 회귀나무 분석

회귀나무의 목적은 연속형 또는 명목형의 독립변수를 이용하여 연속형인 종속변수를 예측하는 것이며, 반복적인 분할 알고리즘을 이용해 구축한다. 예측을 위한 회귀나무 모델도 분류 나무 모델과 상당히 유사한 방법으로 작동한다. 이 경우에는 종속변수가 연속형 변수라는 점이 다르고 작동 원리나 절차는 분류나무와 같다. 많은 분할이 시도되고 결과로 나온 나무 모델들의 모든 가지에 대하여 불순도를 측정한다. 다음 단계에서는 불순도의 합이 최소가 되는 분할을 선택한다.

회귀나무에 사용되는 대표적인 알고리즘은 CHAID(Chi-Square Automatic Interaction Detection)와 CART(Classification And Regression Tree)가 있다. CHAID 알고리즘은 카이제곱-검정(이산형 종속변수) 또는 F-검정(연속형 종속변수)을 이용하여 다지 분리(multiway split)를 수행하는 알고리즘이다. CHAID는 변수가 범주형이고 독립변수와 종속변수 간의 관계를 찾아야 할 때 가장 유용하다(Pyle and Dorian, 1998). CART 알고리즘은 나무 모델 중 가장 잘 알려진 방법론 가운데 하나인데, 전체 데이터 셋을 갖고 시작하여 반복해서 두 개의 자식 노드(child node)를 생성하기 위해 모든 독립변수를 사용하여 데이터 셋의 부분집합을 쪼갬으로써 나무 모델을 생성한다.

나무모델은 다음과 같은 두 가지 측면에서 사용자들에게 상대적으로 적은 수고만을 요구한다고 볼 수 있다. 첫째, 변수를 변환시킬 필요성이 없다. 즉, 변수들에 관한 어떠한 단조 변환에 대해서도 나무모델은 같은 결과를 산출한다. 둘째, 변수 중 일부를 선택하는 변수 부분선택이 자동으로 이루어진다(Shmueli *et al.*, 2010).

3. 제안 방안

3.1 회귀나무 분석에 의한 특징함수 구성

회귀나무 분석을 수행하여 생성된 규칙을 C-CRF의 첫 번째 특징함수 $\sum_{k}^{K_1}\alpha_k f_k(y_i, \boldsymbol{x})$에 반영한다. 이 특징함수는 특정 규칙을 만족하면 그 결과값을 반환하고 그렇지 않으면 0을 반환하여 그 합을 결과값으로 하는 특징함수이다. 예를 들어, '남성이면 키가 175cm이다.'라는 첫 번째 규칙과 '여성이면 키가 160cm이다'라는 두 번째 규칙을 생성했다고 하자. 또한, 훈련 집합에 어떤 남성의 데이터가 포함되어 있다고 하자. 그렇다면, 그 남성은 첫 번째 규칙을 만족하므로 175cm를 반환하고, 두 번째 규칙을 만족하지 않으므로 0cm를 반환하여 그 합인 175cm가 첫 번째 특징함수의 결과값이 된다. 이와 같은 방법으로 구성한 첫 번째 특징함수는 식 (6)과 같다. 해당 식은 i번째 데이터의 실제 값(y_i)과 회귀나무 분석의 예측 값 ($\sum_{r}^{R}\hat{y}_{ir}$) 간의 오차를 반영한다.

$$f(y_i, \boldsymbol{x}) = (-\alpha)(y_i - \sum_{r}^{R}\hat{y}_{ir})^2, \quad (6)$$

여기서 R은 규칙의 수를 나타내며 $\hat{y}_{ir}$는 i번째 데이터의 r번

째 규칙의 반환 값이다. 한 가지 특징만을 고려하므로 $K_1 = 1$ 이고 α는 상수가 된다. 실제 값과 예측 값의 차이가 작을수록 높은 확률 값을 가지게끔 음의 부호를 사용한다. 본 연구에서는 독립변수와 종속변수 간의 관계를 나타내는 특징을 한 가지만을 고려하였지만, 만일 두 가지 이상의 특징을 고려한다면 첫 번째 특징함수를 $\sum_{k=1}^{K_1} \alpha_k f_k(y_i, \boldsymbol{x})$의 형태로 사용한다.

3.2 유사도 계산에 의한 특징함수 구성

Qin *et al.*(2009)은 사용자의 질의(query)를 포함하는 문서의 순위를 매기기 위해, 문서의 내용과 관련된 특징함수와 더불어 문서 간의 관계를 반영하는 특징함수를 도입하였다. 이는 유사한 문서들은 순위도 유사할 것이라는 기본가정을 바탕으로 하고 있다. 또한, Radosavljevic *et al.*(2010) 역시 에어로졸의 광학적 깊이를 예측하기 위해, 데이터 간의 유사도를 반영하는 특징함수를 도입하였다. 본 연구에서도 데이터 간의 유사도를 반영하는 특징함수를 기존 연구와 동일한 형태인 식 (7)과 같이 도입한다.

$$g(y_i, y_j, \boldsymbol{x}) = (-\beta) E_{i,j} \left(\sum_r^R \hat{y}_{ir} - \sum_r^R \hat{y}_{jr} \right)^2. \tag{7}$$

이 때, $E_{i,j}$는 i번째 데이터와 j번째 데이터 간의 유사도이다. 식 (7)의 의미는 두 종속변수 y_i와 y_j의 예측 값의 차이가 작을수록 서로 유사하다고 간주하는 것이며, β는 0보다 큰 상수이다. 역시 유사도가 높을수록 높은 확률 값을 가지도록 음의 부호를 사용한다. 식 (6)과 식 (7)을 종합하여, 식 (3)을 다음과 같이 변형한다.

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x},\alpha,\beta)} \exp\left\{ \sum_{i=1}^T (-\alpha) \left(y_i - \sum_r^R \hat{y}_{ir} \right)^2 + \sum_{i,j}^T (-\beta) E_{i,j} \left(\sum_r^R \hat{y}_{ir} - \sum_r^R \hat{y}_{jr} \right)^2 \right\}. \tag{8}$$

정규화 상수 $Z(\boldsymbol{x}, \alpha, \beta)$는 식 (9)와 같이 계산한다.

$$\begin{aligned} Z(\boldsymbol{x}, \alpha, \beta) &= \int_{-\infty}^{\infty} \exp\left(\sum_{i=1}^T (-\alpha) \left(y_i - \sum_r^R \hat{y}_{ir} \right)^2 + \sum_{i,j}^T (-\beta) E_{i,j} \left(\sum_r^R \hat{y}_{ir} - \sum_r^R \hat{y}_{jr} \right)^2 \right) dy \\ &= \exp\left(\sum_{i,j}^T (-\beta) E_{i,j} \left(\sum_r^R \hat{y}_{ir} - \sum_r^R \hat{y}_{ir} \right)^2 \right) \\ &\quad \int_{-\infty}^{\infty} \exp\left(\sum_{i=1}^T (-\alpha) \left(y_i - \sum_r^R \hat{y}_{ir} \right)^2 \right) dy. \end{aligned} \tag{9}$$

식 (9)의 우변의 적분항 $\int_{-\infty}^{\infty} \exp(\sum_{i=1}^T (-\alpha)(y - \sum_r^R \hat{y}_{ir})^2 dy$ 부분에서 $y - \sum_r^R \hat{y}_{ir}$를 t로 치환한 후, $s = t\sqrt{2\alpha}$로 다시 치환하면 표준정규분포의 누적확률분포와 유사하다. 결과적으로 식 (9)는 식 (10)과 같이 변형할 수 있다.

$$\begin{aligned} Z(\boldsymbol{x}, \alpha, \beta) = \exp\left((-\beta) \sum_{i,j}^T E_{i,j} \left(\sum_r^R \hat{y}_{ir} - \sum_r^R \hat{y}_{jr} \right)^2 \right) \\ \sqrt{\frac{\pi}{\alpha}} \left(1 - \Phi\left(-\sqrt{2\alpha} \sum_{i=1}^T \sum_r^R \hat{y}_{ir} \right) \right), \end{aligned} \tag{10}$$

여기서 $\Phi(\cdot)$는 표준정규분포의 누적확률분포함수이다.

4. 적용 예제

본 연구에서 제안하는 방법을 실제 데이터에 적용하여 예시한다. 사용한 데이터 셋은 Concrete Compressive Strength Data set으로 UCI Machine Learning Repository(https://archive.ics.uci.edu/ml/datasets.html)에서 획득하였다. 해당 데이터는 콘크리트의 압축강도를 예측하기 위한 데이터 셋으로, 총 1,030개의 레코드를 포함하고 있으며, 정규화하여 사용한다. 해당 데이터에 포함된 변수 목록은 <Table 1>과 같다.

Table 1. Concrete compressive strength data set

변수명	유형	설명
Cement	연속형	콘크리트에 포함된 시멘트의 양(kg/m³)
Blast Furnace Slag	연속형	콘크리트에 포함된 고로재의 양(kg/m³)
Fly Ash	연속형	콘크리트에 포함된 시멘트 혼합제의 양(kg/m³)
Water	연속형	콘크리트에 포함된 물의 양(kg/m³)
Superplasticizer	연속형	콘크리트에 포함된 감수제의 양(kg/m³)
Coarse Aggregate	연속형	콘크리트에 포함된 굵은 골재의 양(kg/m³)
Fine Aggregate	연속형	콘크리트에 포함된 잔골재의 양(kg/m³)
Age	연속형	콘크리트의 재령(day)
Concrete compressive strength(Target Variable)	연속형	콘크리트의 압축 강도(MPa)

Table 2. Rules obtained by regression tree

ID	규칙	결과값
1	Age ≤ -2.488 and Cement ≤ 0.707	-2.504
2	Age ≤ -2.488 and Cement > 0.707	-2.344
3	Age > -2.488 and Cement > 0.736	-2.148
4	Age > -2.488, and Cement ≤ -1.110	-2.434
5	Age > -2.488, Cement ≤ 0.736, Cement > -1.110 and BlastFurnacneSlage ≤ -2.552	-2.358
6	Age > -2.488, Cement ≤ 0.736, Cement > -1.110, BlastFurnacneSlage > -2.552 and Cement ≤ -0.064	-2.298
7	Age > -2.488, Cement ≤ 0.736, Cement > -1.110, BlastFurnacneSlage > -2.552 and Cement > -0.064	-2.188

회귀 나무의 성장방법으로는 CHAID와 CART를 사용하였고, CHAID 기법을 적용한 회귀 나무의 평균 제곱근 편차(Root Mean Square Error, RMSE) 값은 2.450이었고 CART 기법을 적용한 회귀 나무의 RMSE 값은 2.457로 CHAID 기법을 사용한 모델의 RMSE 값이 더 작아, 본 연구에서는 CHAID 기법을 적용한 회귀 나무를 이용하여 특징함수를 구성한다. 생성된 회귀 나무를 규칙으로 나타내면 <Table 2>와 같다. 본 예제에서는 보다 많은 규칙을 생성하기 위해 회귀나무를 완전히 성장시켰으나, 반드시 완전히 성장시킬 필요는 없으며, 필요에 따라 가지치기(pruning)방법을 사용할 수 있다.

<Table 2>에서 나타난 규칙들을 바탕으로 첫 번째 특징함수를 다음과 같이 구성한다. 규칙의 수가 7개이므로, 식 (8)에서 $R=7$이며 특정 규칙을 충족하면 대응되는 결과값을 반환하고 그렇지 않으면 0을 반환한다. 예를 들어, k번째 데이터가 7번째 규칙을 만족하고 그 외의 규칙을 모두 만족하지 못한다면, 첫 번째 특징함수는 다음과 같이 구성된다.

$$f(y_k, \boldsymbol{x}) = -\alpha(y_k + 2.188)^2. \tag{11}$$

또한, 유클리디안 거리를 이용하여 두 번째 특징함수의 유사도를 구성한다. 즉, 식 (7)에서 제시된 두 데이터 간의 유사도는 두 데이터 사이의 유클리디안 거리의 역수이다.

이제 식 (5)를 이용하여 모수를 추정한다. 식 (10)에서 구한 $Z(\boldsymbol{x}, \alpha, \beta)$를 식 (8)에 대입하면 β가 약분되어 사라지므로 로그가능도의 값은 α값에만 영향을 받는다. 따라서 β를 임의로 0.05로 고정 후 α값을 추정하였다. 모수 추정은 시뮬레이티드 어닐링(simulated annealing)을 사용하였다. 시뮬레이티드 어닐링은 복잡한 제약식을 가지는 조합 최적화 문제에서 최적값에 근접한 해를 구하기 위하여 확률적 탐색을 기반으로 하는 메타 휴리스틱(meta-heuristic) 기법이다. α의 초기값은 β와 같은 0.05로, 초기 온도는 100으로 설정하였을 때 α는 18.54로 추정되었다.

완성된 모델을 사용하여 예측을 수행하기 위한 예시로서, 임의로 선정한 데이터(15번 데이터)의 콘크리트 압축 강도를 종속변수로 하는 확률 분포를 계산한다. 15번 데이터의 특성을 정규화하면 다음 <Table 3>에서 제시하였다.

Table 3. Property of no. 15 data

변수	#15 데이터
Cement	1.35809
Blast Furnace Slag	0.38885
Fly Ash	-0.8538
Water	-1.3029
Superplasticizer	1.74237
Coarse Aggregate	-1.5664
Fine Aggregate	1.42115
Age	-0.7172

회귀나무 분석을 통해 예측된 15번 데이터의 콘크리트 압축 강도는 -2.148이다. 이는 $\sum_r^R \hat{y}_{15,r}$가 -2.148임을 의미한다. 또한, 15번 데이터와 모든 데이터 간의 유사도와 모든 데이터의 콘크리트 압축 강도 예측 값을 바탕으로 식 (8)의 $\sum_{i,j}^T(-\beta)E_{i,j}(\sum_r^R \hat{y}_{ir} - \sum_r^R \hat{y}_{jr})^2$을 계산하여 -3.172라는 값을 얻었다. 추정한 모수 값 α와 β를 바탕으로 $Z(\boldsymbol{x}, \alpha\,\beta)$를 계산하여 확률 분포를 완성하였다. 이 때 계산된 $Z(\boldsymbol{x}, \alpha\,\beta)$은 1/57.94이고 이를 바탕으로 완성한 확률 밀도함수는 식 (12)와 같으며 이를 그래프로 나타내면 <Figure 2>와 같다.

$$P(y \mid x) = 57.94 \times \exp(-18.54 \times (y + 2.148)^2 - 3.172). \tag{12}$$

이를 바탕으로 15번 데이터의 콘크리트 압축 강도에 대한 예측 구간을 확률 값과 함께 제시할 수 있다. 또한, 확률밀도함수를 최대화하는 값(15번 데이터의 경우 -2.148)을 예측 값으로 두어 사용할 수 있다.

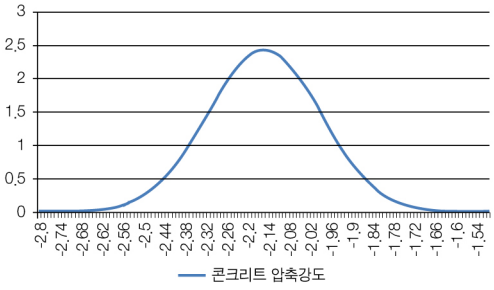

Figure 2. Probability density function of no. 15 data

본 연구에서 제안하는 방법으로 C-CRF를 구성하였을 때의 성능을 평가하기 위해, 다중회귀분석(Multiple Linear Regression, MLR)과 인공신경망(Artificial Neural Network, ANN)을 적용하여 콘크리트 압축 강도를 예측하였고 그때의 RMSE 값을 비교하여 그 결과를 <Table 4>에 제시하였다.

Table 4. RMSE of MLR, ANN and suggested method

Method	RMSE
MLR	4.020
ANN	2.312
Suggested Method	2.324

이 예제에서는 본 연구에서 제안하는 방법이 다중회귀분석보다 좋은 성능을 보였으며, 인공신경망과는 비슷한 성능을 나타내었다.

5. 결 론

본 연구에서는 회귀나무를 이용하여 데이터 기반의 객관적이고 효과적인 C-CRF의 특징 함수구성방법을 제안하였다. 즉, 회귀나무의 규칙들을 C-CRF의 첫 번째 특징 함수에 반영하고 두 번째 특징함수는 유사도를 반영하여 구성한다. 콘크리트의 압축 강도를 예측하기 위한 데이터에 적용하여 C-CRF를 구성하였고, 다른 예측 방법과 비교 평가하여 본 방법이 적절함을 설명하였다.

본 연구에서 제시하는 방법의 예측 성능은 회귀나무에 큰 영향을 받을 수밖에 없다. 하지만 객관적인 특징함수를 사용하여 종속변수의 점추정과 구간추정을 비롯한 확률적 특성을 파악할 수 있다는 장점을 가진다. 이러한 장점으로 인해 도메인 지식에 의한 특징 함수 구성이 쉽지 않을 때, 데이터로부터 특징 함수를 구성하는데 적용할 수 있다.

참고문헌

Ahn, G.-S. and Hur, S. (2015), Prediction of new customer's degree of loyalty of interest shopping mall using continuous conditional random field, *Journal of the Korean Institute of Industrial Engineers*, **41**(1), 10-16.

Jeon, J.-W. and Lee, Y.-H. (2005), Iterative simulated annealing for graph coloring problem, *In proceedings of the 2005 Korean Institute of Industrial Engineers Fall Conference*, 226-229.

Kim, S.-G. and Park, S.-Y. (1999), A study on the comparison of data mining techniques' performances, In proceedings of the 1999 *Korean Society of Management Information Systems Spring Conference*, 371-383.

Kosta, R., Vladan, R., Slobodan, V., and Zoran, O. (2013), Continuous conditional random fields for efficient regression in large fully connected graphs, *Proc. 27th AAAI Conf. on Artificial intelligence*, 840-846.

McCallum, A. (2002), Efficiently inducing features of conditional random fields, *Proc. 19th Conf. on Uncertainty in Artificial Intelligence*, 403-410.

Lee, E.-A., Choi, H.-R., and Lee, H.-C. (2012), A study on the forecasting of the number of end of life vehicles in Korea using Markov Chain, *Journal of the Korean Institute of Industrial Engineers*, **38**(3), 208-219.

Lee, S.-K., Kang, J.-H., Lee, H.-K., Joo, T.-W., Oh, S.-H., Park, S.-W., and Kim, S.-B. (2014), Prediction of product life cycle using data mining algorithms : a case study of clothing industry, *Journal of the Korean Institute of Industrial Engineers*, **40**(3), 291-298.

Pyle, D. (1998), Putting data mining in its place, *Database Programming and Design*, **11**(3), 32-36.

Qin, T., Liu, T.-Y., Zhang, X.-D., Wang, D.-S., and Li, H. (2009), Global ranking using continuous conditional random fields, *Proc. Conf. on the Advances in Neural Information Processing Systems*, 1281-1288.

Radosavljevic, V., Vuetic, S., and Obradovic, Z. (2010), Condtinuous conditional random fields for regression in remote sensing, *Proc. 19th Int. Conf. on Artificial Intelligence*, 809-814.

Shmueli, G., Patel, N. R., and Bruce, P. C. (2010), *Data Mining for Business Intelligence : Concepts, Techniques, and Applications in Microsoft Office Excel With XLMiner, Second Edition*, WILEY, Hoboken, New Jersey, USA.

Stewart, L., He, X., and Zemel, R. S. (2008), Learning flexible features for conditional random fields, *Pattern Analysis and Machine Intelligence*, **30**(8), 1415-1426.

You, Y.-J., Lim, B.-M., Park, J.-S., and Baek, J.-G. (2012), Non-normal regression tree learning model, *Proc. Conf. on Korean Society of Management Information Systems*, 1503-1510.

Zhang, D. and Nebel, B. (2011), Feature Induction of Linear-chain Conditional Random Fields, *Proc. Int. Conf. on Agents and Artificial Intelligence*, 230-235.