Journal of The Institute of Electronics and Information Engineers Vol.53, NO.9, September 2016 http://dx.doi.org/10.5573/ieie.2016.53.9.077
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2016-53-9-10

가중치 세분화 기반의 로지스틱 회귀분석 모델

(Fine-Grain Weighted Logistic Regression Model)

이 창 환*

(Chang-Hwan Lee©)

요 약

로지스틱 회귀분석은 오랫동안 다양한 분야에서 예측을 위한 기술 혹은 변수 간의 관계를 설명하기 위하여 사용되어 왔다. 로지스틱 회귀분석에서 각 속성은 목적 값에 대한 중요도를 가지는데 본 연구에서는 이를 세분화하여 각 속성의 값에 따라서 중요도를 부여하는 새로운 방법을 제시한다. 점진적 하강법을 이용하여 알고리즘의 성능을 최대화하는 각 속성값 가중치의 값을 계산하였다. 제안된 방법은 다양한 데이터를 이용하여 실험하였고 본 연구의 속성값 기반 로지스틱 회귀분석 방법은 기존의 로지스틱 회귀분석보다 우수한 학습 능력을 보임을 알 수 있었다.

Abstract

Logistic regression (LR) has been widely used for predicting the relationships among variables in various fields. We propose a new logistic regression model with a fine-grained weighting method, called value weighted logistic regression, by assigning different weights to each feature value. A gradient approach is utilized to obtain the optimal weights of feature values. We conduct experiments on several data sets and the experimental results show that the proposed method shows meaningful improvement in prediction accuracy.

Keywords : 로지스틱 회귀분석, 속성 가중치, 분류학습

Ⅰ. 서 론

1.

1970년대 이후 로지스틱 회귀분석(Logistic Regression, LR)는 통계학 및 데이터 마이닝 분야를 중심으로 많은 연구가 진행되고 있으며 또한 생명과학 연구, 비즈니스 및 금융, 범죄학, 생태학, 공학, 의료 정책 등 많은 분야에서 널리 사용된다[7]. 로지스틱 회귀 모델은 이진 혹은 다항의 종속 변수(Y)가 어떻게 독립적인 입력 변수의 집합 ($X=\{X_1, X_2, ...\}$)과 관련되는 지를 나타낸다. 이러한 분석은 출력 및 입력 변수들 간의 관계를 탐구한다. 또한 입력변수들의 관측값들의 조합에 기반하여 출력변수의 기대값이 보여지는 선형 회귀와 달리 이진 로지스틱 회귀는 주어진 관측값들에 따른 출력 클래스의 사후 확률 (즉, true'의 확률)을 모델링한다. 이와 같이 로지스틱 회귀는 입력과 출력 사이의 관계가 입력의 선형 조합을 이용하는 로지스틱 방정식의 형태로 추정된다고 가정한다. 이때 입력의 각 속성은 각자의 가중치와 결합된다. 예를 들면, 신용평가에서 로지스틱 회귀는 고객의 신용도를 확률로 모델링하기 위해 사용될 수 있는데, 고객이 제때 청구서를 지불할 확률은 청구 금액, 연간 소득, 직업, 주택 담보 대출 및 채무 의무, 과거 청구서의 지불 비율, 신용 기록의 다른 측면 등의 고객의 금융 정보를 사용할 수 있다.

기존의 로지스틱 회귀모델은 각 속성에 대하여 한 개의 가중치가 정해진다. 따라서 속성이 여러 값을 가지더라도 그 속성의 값들은 동일한 가중치를 가지게 된다. 하지만 이들 속성 각각의 값들은 분류 학습에 있어서 실질적으로 서로 다른 중요도를 가진다. 즉 동일한 속성에 있어서 어떤 속성값은 분류학습에 대하여 높은 중

* 정회원, 동국대학교 정보통신학과 (Dongguk University)
© Corresponding Author (E-mail : chlee@dgu.ac.kr)
※ 본 연구는 한국연구재단 중견연구자지원사업의 지원 (2014R1A2A1A11051011)으로 이루어졌음.
Received ; July 25, 2016 Revised ; August 8, 2016
Accepted ; August 26, 2016

요도를 가지고 어떤 속성값은 낮은 중요도를 가지게 된다. 예를 들어서 임신과 관련한 분류학습에 있어서 성별이 여자인 경우는 많은 중요도를 가지지만 성별이 남자인 경우는 별 의미를 가지지 못한다. 아래 예제 1 은 이와 같이 속성값이 다른 가중치를 가지는 사례를 예시하고 있다.

예제 1: 속성 성별 (Gender) 는 두개의 값 (남성 혹은 여성: male or female)이 있다. 또한 학습하고자 하는 목적 클래스의 값은 “합격 (y)” 또는 “불합격 (n)”을 갖는다.

$P(y) = 0.9, P(n) = 0.1,$
$P(y|female) = 0.1,\ P(n|female) = 0.9,$
$P(y|male) = 0.99,\ P(n|male) = 0.01$로 가정하자.

이 예에서 여성의 성별이 목적 클래스의 변화에 많은 영향을 주는 것을 알 수 있다. 다시 해서 속성의 값이 남성인 경우는 목적 클래스 값의 변화가 미미한 것에 반에 속성값이 여성인 경우는 목적 클래스의 확률분포에 많은 영향을 준다.

위의 예제에서 보듯이 성별이라는 속성은 그 값이 여성인 경우에 결과를 예측하는데 훨씬 중요하다. 하지만 전통적인 로지스틱 회귀에서는 값에 관계없이 하나의 가중치 값이 성별 항목에 할당된다. 따라서 이 경우에 해당 모델은 속성값 간의 차이를 반영하지 못함으로써 좋은 성능을 보이지 못한다.

본 논문은 로지스틱 회귀의 세분화된 가중치 방법인 속성값 기반 로지스틱 회귀 (value weighted logistic regression: VWLR) 모델을 제안한다. 이 논문에서는 각 속성의 값에 각각 다른 가중치를 부여하는 로지스틱 회귀 방법의 새로운 패러다임을 제공한다. 점진적 하강 방법을 이용하여 본 연구에서 제안하는 최적의 속성값 가중치를 계산한다. 다수의 실험 데이터를 이용하여 본 연구 모델의 성능을 비교하였다. 본 문에서는 이분적 분류를 위한 로지스틱 회귀 모델에 중점을 두어 설명하였으며 이는 추후에 다항 분류의 로지스틱 회귀 모델에도 쉽게 확장될 수 있다.

2. 관련 연구

로지스틱 회귀분석은 통계학습에서 오랫동안 사용되어온 분류학습 방법이다. 하지만 로지스틱 회귀분석 방법은 선형 학습 방법으로써 실제 문제가 비선형인 경우에 좋은 성능을 보이기 힘든 단점이 있다. 이와 같은 로지스틱 회귀분석의 단점을 해결하기 위하여 다수의 방법이 제안되고 있다.

첫 번째 방법은 지역화(localized) 로지스틱 회귀분석의 방법이다[1~2]. 이는 하나의 글로벌 로지스틱 회귀법은 예측의 정확도를 보장 할 수 없지만 간단한 국지화 방법은 클러스터링 알고리즘을 이용하여 데이터를 적절하게 여러개의 작은 부분집합으로 분할한다. 각 부분집합은 자신의 특징 공간에서 선형으로 분리 될 가능성이 있다. 이 방식의 주요 아이디어는 각각의 부분집합은 지역적 로지스틱 회귀 (local logistic regression, LLR)에 의해 모델링되는 점이다. 하나의 글로벌 분류법보다 더 나은 국지적 예측 정확도를 달성 할 수 있도록 하는 각 부분집합에 대한 많은 작은 지역적 분류들이 존재한다. 글로벌 회귀 마찬가지로 각 지역 로지스틱 회귀 모델은 그 자신의 지역내에서 로지스틱 함수를 찾는다. 다시 말하자면 지역적으로 가중치된 로지스틱 함수가 존재한다.

이러한 지역 로지스틱 회귀 모델은 주로 K 최근접 방법 (K-nearest-neighbor) 을 사용하는데 이는 오직 이웃한 데이터 포인트만을 선택하고 나머지를 무시한다. 지역적 가중 로지스틱 회귀법은 각 인스턴스에 가중치를 부여하는 것을 제외하고 글로벌 로지스틱 회귀와 매우 유사하다. 특정 도메인에 거리에 대한 다양한 지표들이 제안될 것이다.

두 번째의 방법은 커널함수를 사용하는 방법이다[8~9]. 커널 로지스틱 회귀분석은 확장된 공간에서의 커널함수를 사용하여 비선형 문제를 해결한다. 커널 방법은 비선형 데이터를 원래 입력 공간으로부터 더 높은 차원의 특징 공간으로 변환할 수 있게 도운다. 즉, 데이터 포인트들은 커널 함수에 의해 확장된 공간에서 선형으로 분리될 수 있다. 커널화된 로지스틱 회귀분석은 비선형 확률 분류를 가능하게 하며 여러 벤치마크에서 유망한 결과를 보이고 있다.

3. 속성값 기반의 로지스틱 회귀분석

입력 데이터 ($D_k \in R^n$)의 학습 데이터 ($\{D_k, Y_k\}^{N_{k=1}}$)이 있고 데이터 포인트 k 에 해당하는 클래스 값을 Y_k 라고 하자. 출력 값 $Y_k \in \{0,1\}$ 의 이진 분류 문제를 고려하면 성공의 확률 $P(Y_k = 1|D_k)$ 는 k 번째 입력 데이터 D_k 가 클래스 1 에 속하고 클래스 0 일 확률 $P(Y_k = 0|D_k)$ 가 $1 - P(Y_k = 1|D_k)$ 인 경우이다. 전통적 로지스틱 회귀에서 회귀선은 다음의 형태로 표현된다.

$$Y = w_0 + w_1 X_1 + \cdots + w_n X_n$$

LR의 기본적인 가정은 각 입력 특징 값이 출력값에 대하여 상이한 의미를 갖는다는 것이다. 특정 입력 항목이 중요하거나 의미있을 때 그 항목의 중요성은 그 입력 값으로 분석될 수 있다고 믿는다.

특정 속성 X_i 가 m 개의 입력 값 ($\{x_{i1}, x_{i2}, ..., x_{im}\}$)으로 이뤄졌다고 가정하자. 이때 m 은 i 번째 데이터의 입력 값의 수 ($|X_i|$)를 의미한다. 본 연구에서 제안하는 VWLR 모델에서 각 값들은 다른 가중치를 부여 받는다. 따라서 회귀선 Y 는 아래와 같이 표현된다.

$$Y = w_{01} + (w_{11}x_{11} + w_{12}x_{12} + \cdots + w_{1|x_i|}x_{1|x_i|}) + \cdots \\ (w_{n1}x_{n1} + w_{n2}x_{n2} + \cdots + w_{n|x_n|}x_{n|x_n|})$$

여기서 x_{ij} 는 i 번째 데이터의 j 번째 입력 값을 의미하고 w_{ij} 는 x_{ij} 의 가중치를 나타낸다.

여기서 로지스틱 회귀분석의 정의에 따라 k 번째 관측 D_k 에 대응하는 출력 값 Y_k 의 확률은 다음과 같이 주어진다.

$$P(y_k = 1|D_k, W) = \frac{1}{1 + \exp(w_{01} + \Sigma_{x_{ij}\epsilon D_k} w_{ij}x)}$$

$$P(y_k = 0|D_k, W) = \frac{\exp(w_{01} + \Sigma_{x_{kj}\epsilon D_k} w_{ij})}{1 + \exp(w_{01} + \Sigma_{x_{ij}\epsilon D_k} w_{ij}x)}$$

VWLR 에서는 속성의 값마다 다른 가중치를 부여한다. 따라서 그 결과로 속성값 기반의 로지스틱 회귀평면은 그 행태에 있어 급격한 변화를 보인다.

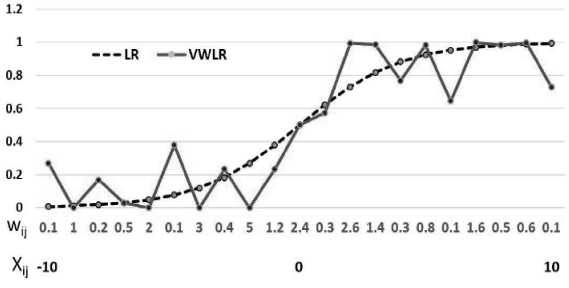

그림 1. VMLR 과 기존 LR의 차이
Fig. 1. VMLR vs traditional LR.

위의 그림 1 은 VWLR과 전통적 로지스틱 회귀의 차이점을 직관적으로 보인다. 그래프의 X-축 의 숫자는 x_{ij} 값들의 가중치를 의미한다. 그림 1 에서 보여지듯이 속성값 가중 방법이 VWLR을 비선형 함수로 만들어주어 VWLR이 가설 공간에서 보다 복잡하고 또한 더 세분화된 분류 경계를 나타내는 비선형 분류기가 된다. 반면에 전통적 로지스틱 회귀법은 선형 예측을 한다[4]. VWLR은 특징 값의 역할을 극대화하여 예측 정밀성에서 최적의 성능을 기대한다. 결과적으로 VWLR은 두개 이상의 클래스 값들의 클러스터가 있는 경우에도 효과적으로 문제를 해결할 수 있다.

우리의 목적은 VWLR을 훈련시켜 조건부 데이터의 확률을 최대화하는 최적의 가중치 값을 선택하도록 하는데 있다. 조건부 데이터 확률은 해당하는 D_k 값의 조건하에 훈련 데이터 내의 목격되는 Y 값들의 확률을 의미한다. 즉 다음을 만족하는 가중치 집합 W를 선택해야 한다.

$$\arg max_W \prod_{k=1}^{N} P(y_k|D_k, W)$$

로그함수는 학습의 결과를 바꾸지 않기 때문에 우리는 로그 함수 형태의 값을 최대화한다. 조건부 로그 가능도 함수를 최대화하는 것은 최적의 가중치 값들을 가질 수 있도록 한다. 따라서 최적의 가중치 W 는 다음과 같다.

$$l(W) = \arg max_W \sum_{k=1}^{N} yx \ln P(yx = 1|D_k, W) + \\ (1 - y_k)\ln P(y_k = 0|D_k, W)$$

VWLR에서는 특정한 부분의 값과 연관된 훈련 데이터의 수는 매우 작을 수 있다. 게다가 전통적 로지스틱 회귀법과 비교해서 VWLR의 파라미터 개수는 증가한다. 즉 전통적 로지스틱 회귀의 파라미터 수는 n 이다. 하지만 VWLR 에서 그 파라미터의 수는 $\sum_{i=1}^{n}|X_i|$ 으로 증가하게된다. 여기서 n은 입력 항목의 수, $|X_i|$ 이 i 번째 항목의 값들의 수를 의미한다. 파라미터 수의 증가는 VWLR이 각 값들을 효과적으로 대표할 수 있도록 한다. 반면 큰 가중치를 가질 가능성도 증가한다. 이러한 이유 때문에 특히 데이터가 아주 다차원이고 훈련 데이터가 드문 경우에 VWLR에서 훈련 데이터의 오버 피팅 문제가 종종 발생할 수 있다.

VWLR에서 가능한 오버 피팅을 줄이는 한 가지 방법은 정형화(regularization)이다[3]. 정형화는 특히 작은 수의 훈련 예제만이 있거나 학습되어야 할 많은 수의 파라미터들이 있을 때 가중치의 크기를 줄이고 오버 피

팅을 피할 수 있게 하는 기법으로 알려져 있다. 따라서 W 의 큰 값들을 불리하게 하는 수정된 처벌적 로그 가능성 함수를 만들었다. 본 연구에서는 릿지 페널티 (ridge penalty) $\frac{\lambda}{2}\|W\|^2$ 이며 이러한 이차적 정형화 항목을 l(W) 에 추가한다.

따라서 l(W) 함수는 다음과 같이 수정된다. 여기에서 λ 는 이 페널티 용어의 강도를 결정하는 상수이다.

$$l(W)=\sum_{k=1}^{N} yk\ln P(y_k=1|D_k,W)+ (1-y_k)\ln P(y_k=0|D_k,W)-\frac{\lambda}{2}\parallel W\parallel^2$$

위의 식은 다음과 같이 변환할 수 있다.

$$l(W)=\sum_{k=1}^{N} y_k\ln\frac{P(y_k=1|D_k,W)}{P(y_k=0|D_k,W)}+ \ln P(y_k=0|D_k,W)-\frac{\lambda}{2}\parallel W\parallel^2$$

$$=\sum_{k=1}^{N} y_k\left(w_{01}+\sum_{x_{kj}\epsilon_{Dk}} w_{kj}x_{kj}\right)- \ln\left(l+\exp\left(w_{01}+\sum_{x_{ij}\epsilon_{Dk}} w_{kj}x_{ij}\right)\right)-\frac{\lambda}{2}\parallel W\parallel^2$$

따라서 위의 수식을 만족하는 최적의 w_{ij} 집합을 구하는 것이 학습의 목표이다. 본 연구에서는 점진적 상승 방법을 이용하여 최적의 w_{ij} 집합을 구한다. 점진적 상승에서의 계산하는 벡터의 ij 번째 요소는 다음과 같다.

$$\frac{\partial l(W)}{\partial w_{kj}}=\sum_{k=1}^{N}=y_y x_{kj}-\left(\frac{1}{1+\exp(w_{01}+\Sigma_{aij}\epsilon_{Dk}w_{ij}x_{kj})}\right)x_i$$

$$=\sum_{k=1}^{N} x_{kj}(yk-P(yk=1|D_k,W))-\lambda w_{kj}$$

따라서 가중치 값 들을 계산하는 최종식은 다음과 같이 계산된다. 즉 임의의 초기값으로 시작하여서 다음 수식을 이용하여 계속적으로 가중치 값을 수정해서 최종적인 가중치 값을 계산한다. (η 는 스텝의 크기를 의미한다)

$$w_{ij=w_{ij}}+\eta\sum_{k=1}^{N} x_{kj}(yk-P(y_k=1|D_k,W))-\eta\lambda$$

4. 실험 결과

이 섹션에서 VWLR, 전통적 로지스틱 회귀법 (LR), 더미 코딩 (dummy coding) 로지스틱 회귀법 (DCLR)[5] 의 세 가지 방법이 많이 사용되는 통계 컴퓨팅 프로그램 R 을 이용하여 구현되었다. DCLR은 로지스틱 회귀법의 명목 변수들을 처리하기 위해 더미 코딩 방법을 적용한다. 모든 모델에 대하여 동일한 시스템 환경 아래에서 예측 정확도들을 조사한다. VWLR의 실험 결과들은 LR과 DCLR의 결과들과 비교된다.

총 8 개의 데이터를 UCI[10] 에서 이용하여 성능을 비교하였다. 이 데이터들은 기계학습 알고리즘의 실험에 많이 사용되었던 데이터들이다. 실험은 전체 데이터 세트들에 대하여 10-fold cross-validation 검증을 사용하여 처리되었고 모든 예측 정확도를 기록하였다. 실험에서 연속 변수의 값은 R 패키지의 calm 방법[6] 을 이용하여 그 값을 이산화 시켰다. 이 실험에서 λ 값은 0.001 의 값으로 설정하였다. 너무 큰 λ 값은 짧은 계산 시간과 나쁜 모델 성능을 야기한다. 또한 너무 작은 값은 좋은 성능을 갖지만 아주 긴 계산 시간이 발생한다.

표 1 은 본 연구의 VWLR 방법과 기존의 LR 및 DCLR 과의 성능을 비교한 내용이다. 표 1에서 * 표시는 최고의 수치를 의미하고 + 표시는 두 번째 높은 수치를 의미한다. 표 1에서 보는 것과 같이 VWLR 알고리즘은 전체 8 개의 데이터에서 5 번의 경우에 최고의 정확도를 보이고 있으며 2번의 차점을 기록하고 있다. 즉 8개의 데이터에서 7번에 대하여 1-2 등의 정확도를 보여주고 있다. 이는 본 알고리즘이 기존의 LR과 DCLR 의 성능을 상당한 수준으로 향상 시키고 있음을 알 수 있다.

표 4. VWLR 의 성능비교
Fig. 4. Performance of VWLR.

dataset	LR	VWLR	DCLR
ionosphere	0.841 +	0.751	0.849 *
wdbc	0.827	0.947 *	0.856 +
breast-w	0.821	0.974 *	0.953 +
agaricus	0.818	0.964 *	0.848 +
pendigits	0.753	0.786 +	0.792 *
nursery	0.798	0.891 *	0.878 +
uva	0.751	0.919 *	0.851 +
adult	0.791	0.811 +	0.881 *

5. 결론 및 추후 연구

본 논문은 각각의 특징 값에 적응 가중치를 할당하는 값어치 가중 로지스틱 회귀(value weighted logistic regression -- VWLR) 모델을 제안했다. 최적의 파라미터를 얻기 위해 기울기 상승 방법을 이용했다. 실험 결과는 제안된 방법은 각각의 특징 값에 동일한 가중치를

부여하는 기존의 로지스틱 회귀 모델에 비해 예측 정확도를 크게 개선하는 점을 보인다. 대부분의 경우에 있어 VWLR 이 성공적 결과를 보이고 이러한 결과는 값어치 가중법이 로지스틱 모델 성능을 향상시킬 수 있음을 시사한다.

추후 연구는 본 알고리즘을 다양한 종류의 데이터들에 적용하여 실제적인 성능의 향상을 검증하고 또한 정형화의 기능에 있어서 VWLR 그래프의 좀더 smoothing을 위한 추가적인 정형화 내용을 연구하여 알고리즘의 성능을 더욱 향상시키고자 한다.

REFERENCES

[1] Atkeson, Christopher G., Andrew W. Moore, and Stefan Schaal. “Locally weighted learning for control.” Lazy learning. Springer Netherlands, 1997. 75-113.
[2] Cleveland, William S., and Susan J. Devlin. “Locally weighted regression: an approach to regression analysis by local fitting.” Journal of the American Statistical Association 83.403 (1988): 596-610.
[3] Goeman, Jelle, Rosa Meijer, and Nimisha Chaturvedi. “L1 and L2 penalized regression models.” (2014).
[4] Hosmer D W, Lemesbow S. Goodness of fit tests for the multiple logistic regression model. Communications in Statistics-Theory and Methods, 1980, 9(10): 1043-1069.
[5] Hosmer Jr, David W., and Stanley Lemeshow. Applied logistic regression. John Wiley & Sons, 2004.
[6] Kurgan, Lukasz, and Krzysztof J. Cios. “CAIM discretization algorithm.” Knowledge and Data Engineering, IEEE Transactions on Knowledge and Data Engineering, 145-153. (2004):
[7] Menard, Scott. Applied logistic regression analysis. Vol. 106. Sage, 2002.
[8] Zhang, Lijun, et al. “Efficient Online Learning for Large-Scale Sparse Kernel Logistic Regression.” AAAI. 2012.
[9] Zhu, Ji, and Trevor Hastie. “Kernel logistic regression and the import vector machine.” Journal of Computational and Graphical Statistics (2005).
[10] https://archive.ics.uci.edu/ml/datasets.html

저 자 소 개

이 창 환(정회원)
1982년 서울대학교 계산통계학과 학사 졸업.
1988년 서울대학교 계산통계학과 석사 졸업.
1994년 University of Connecticut Computer Science 박사 졸업.
1994～1996년 AT&T Bell Laboratories, NJ, USA
1996～현재 동국대학교 정보통신학과 교수
<주관심분야: 기계학습, 데이터 마이닝, 인공지능>