Support Vector Regression을 이용한 희소 데이터의 전처리

A Sparse Data Preprocessing Using Support Vector Regression

*전성해, **박정은, **오경환
*청주대학교 통계학과
** 서강대학교 컴퓨터학과

*Sung-Hae Jun, **Jung-Eun Park, **Kyung-Whan Oh
*Dept. of Statistics, Cheongju University
**Dept. of Computer Science, Sogang University
E-mail : shjun@cju.ac.kr

요 약

웹 로그, 바이오정보학 등 여러 분야에서 다양한 형태의 결측치가 발생하여 학습 데이터를 희소하게 만든다. 결측치는 주로 전처리 과정에서 조건부 평균이나 나무 모형과 같은 기본적인 Imputation 방법을 이용하여 추정된 값에 의해 대체되기도 하고 일부는 제거되기도 한다. 특히, 결측치 비율이 매우 크게 되면 기존의 결측치 대체 방법의 정확도는 떨어진다. 또한 데이터의 결측치 비율이 증가할수록 사용가능한 Imputation 방법들의 수는 극히 제한된다. 이러한 문제점을 해결하기 위하여 본 논문에서는 Vapnik의 Support Vector Regression을 데이터 전처리 과정에 알맞게 변형한 Support Vector Regression을 제안하여 이러한 문제점들을 해결하였다. 제안 방법을 통하여 결측치의 비율이 상당히 큰 희소 데이터의 전처리도 가능하게 되었다. UCI machine learning repository로부터 얻어진 데이터를 이용하여 제안 방법의 성능을 확인하였다.

1. 서론

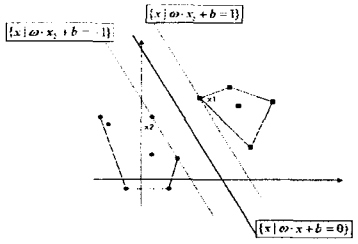
데이터베이스 기술의 발전과 인터넷 사용의 증가에 따라 의사 결정을 위하여 분석이 필요한 데이터의 양은 무한히 증가하고 있다. 하지만 데이터 크기의 증가에 비례하여 분석을 위한 데이터의 품질도 함께 증가하지는 못하는 실정이다. 데이터가 많아질수록 데이터 정제의 문제점도 함께 발생한다. 데이터 정제가 필요한 이유는 방대한 크기의 데이터가 구축되면서 데이터의 결측치(missing value), 잡음(noise), 또는 불일치(inconsistency)등이 발생하게 된다[3]. 본 논문에서는 이러한 문제점들 중에서, 특히 결측치 문제를 해결하기 위하여 효과적인 모형을 제안하였다. 통계적 학습 모형(statistical learning theory)인 Vapnik의 Support Vector Machine(SVM)과 Support Vector Regression(SVR)을 이용하여 데이터의 전처리(preprocessing)를 수행하였다. 특히 결측치의 비율이 매우 큰 희소한 데이터(sparse data)의 전처리를 위한 전략을 제시하였다. 제안 모형의 성능은 UCI machine Learning Repository의 기계 학습 데이터를 이용하였다

2. Support Vector Regression

Vapnik은 주어진 데이터들을 이분법적으로 나눌 수 있는 이상적인 선형평면을 구하는 방법을 SVM을 이용하여 제시하였다[9]. 평면방정식이 주어졌을 때 분류문제를 해결하는 함수식은 다음과 같다.

$$f(x) = sign\ (w \cdot x + b) \qquad (1)$$

식(1)의 함수식 부호에 의해 분류 모형이 결정된다. (그림 1)은 실제 문제 공간에서 이 평면과 방정식이 어떻게 표현되고 적용될 수 있는지 보여준다.



(그림 1) 이상 평면의 표현

그림 1에서 중앙의 굵은 직선을 구해내는 것이 SVM의 최종 목표이다. 이러한 이상 평면은 각 인스턴스들과의 폭을 최대로 하고 분류기로서의 몇 가지 조건들을 만족한다. 따라서 주어진 인스턴스들로부터 간격(margin) 폭을 최대화하고 몇 가지 조건식을 만족하는 평면의 방정식을 구해야 한다. 이상평면은 다음의 식을 만족해야 한다.

$$y_i(\langle w \cdot x_i \rangle + b) \geq 1, \quad i = 1,...,l \qquad (2)$$

식 (2)에서 점 x 와 평면과의 거리는 다음과 같이 정의된다.

$$d(w,b,x) = \frac{|\langle w, x_i \rangle + b|}{\|w\|} \qquad (3)$$

이상 평면(optimal hyperplane)은 위의 식 (3)을 만족하고 점과 평면 사이의 간격을 최대화 하는 w와 b를 구한다. SVM은 손실함수(loss function)를 이상 평면 방정식에 포함시킴으로서 회귀(regression) 문제에 적용 될 수 있다. 손실함수란 기대값과 측정값의 오차를 정의하는 함수식이다. 본 논문에서는 희소데이터에 대해 우수한 성능을 갖는 ε -Insensitive 손실함수를 사용한다.

$$D = \{(x^1, y^1), \cdots, (x^l, y^l)\}, \qquad x \in R^N, y \in R \qquad (4)$$

$$f(x) = \langle w \cdot x_i \rangle + b. \qquad (5)$$

식 (4)와 같은 데이터 구조를 식 (5)의 직선식으로 근사하는 최적의 회귀 함수는 SVM에서 다음 문제로 표현된다.

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2, \\ & y_i - \langle w \cdot x_i \rangle - b \leq \varepsilon, \\ \text{subject to} \quad & \langle w \cdot x_i \rangle + b - y_i \leq \varepsilon \end{aligned} \qquad (6)$$

ε -Insensitive 손실함수와 Lagrange 함수를 이용하여 식 (6)의 문제를 풀면 다음과 같은 해를 얻게 된다[2].

$$\begin{aligned} w &= \sum_{i=1}^{l} (\alpha_i, -\alpha_i^*) x_i \\ b &= -\frac{1}{2}\langle w, (x_r + x_s) \rangle \end{aligned} \qquad (7)$$

최종적으로 (7)식을 이용하여 결측치 대체에 의한 데이터 전처리가 이루어진다.

3. 제안 결측치 대체 기법

웹 로그 파일을 비롯한 많은 데이터는 (그림 2)와 같은 테이블(table) 구조를 갖는다. 본 논문에서는 이러한 테이블 구조의 데이터 중에서 결측치의 비율이 큰 희소 데이터의 전처리를 위하여 결측치 대체 전략을 취하였고 이러한 결측치 대체 전략으로써 SVR을 이용하였다.

	Page1	Page2	Page3	...	PageN
User1		8	17	...	
User2	6			...	
User3		5		...	
User4	11			...	3
User5			21	...	
⋮	⋮	⋮	⋮	⋮	⋮
UserM		7		...	

(a)

	Page1	Page2	Page3	...	PageN
User1	*8*	8	17	...	*3*
User2	6	*9*	*13*	...	*2*
User3	*10*	5	*11*	...	*1*
User4	11	*6*	*10*	...	3
User5	*9*	*4*	21	...	*4*
⋮	⋮	⋮	⋮	⋮	⋮
UserM	*6*	7	*12*	...	*3*

(b)

(그림 2) 희소한 데이터 구조

(그림 2)에서 열(column)로 나타나 있는 Page

들은 각 속성을 나타내는 애트리뷰트(attribute)이고 User들로 표현된 행(row)은 각각의 개체(object)이다. SVR은 각 page 마다 예측 모형을 구축하여 자신을 제외한 다른 Page들을 입력 변수로 하여 결측된 자신의 Page의 해당 셀(cell)을 채워 넣는다.

$$t_{page(i)} = F_{SVR}(t_{page1}, \cdots, t_{page(i-1)}, t_{page(i+1)}, \cdots, t_{pageN}) \quad (8)$$

(8)식은 i번째 Page가 (N-1)개의 Page들에 의해 예측되어지는 식을 나타내고 있다.

4. 실험 및 결과

본 논문의 실험을 위하여 UCI Machine Learning Repository의 Glass Identification 데이터를 이용하였다[10]. 이 데이터는 유리의 종류를 결정하는 9개의 입력 변수들(Ri, Na, Mg, Al, Si, K, Ca, Ba, Fe)과 1개의 목표변수로 이루어져 있다. 또한 결측치가 없는 완전한(complete) 데이터이다. 실험을 위하여 목표 변수는 사용하지 않고 임의로, 5%, 10%, 20%, 30%, 40%, 그리고 50%의 결측 비율(missing rate)을 갖는 데이터 집합을 만들어 사용하였다. <표 1>에서 SVR과 비교되는 3개의 결측치 대체 모형과의 성능 평가에 대한 결과를 나타냈다. 성능 비교는 실제값과 예측값의 제곱 평균을 나타내는 평균제곱오차(MSE, mean squared error)을 이용하였다[1]. 이 값이 작을수록 모형의 예측 정확도는 높은 것이다.

위의 실험에서 기존에 결측치 대체 기법으로 많이 쓰이고 있는 tree imputation(tree.)[4], distribution based imputation(dist.)[5], 그리고 Huber의 single value imputation[6]을 본 논문의 SVR과 비교하였다. 데이터의 결측 비율이 작은 5% 환경에서는 오히려 SVR에 비해 tree imputation 모형이 더 좋은 결과를 보여 주고 있다. 하지만 결측 비율이 커질수록 다른 비교 모형들에 비해 SVR의 성능이 훨씬 우수해짐을 알 수 있다.

mssing rate	tree.	dist.	huber	SVR
5%	0.007	0.014	0.050	0.009
10%	0.013	0.015	0.072	0.012
20%	0.043	0.036	0.131	0.024
30%	0.042	0.078	0.152	0.019
40%	0.079	0.153	0.210	0.025
50%	0.142	0.252	0.321	0.031

<표 1> 비교 모형들간의 MSE

5. 결론 및 향후과제

본 논문에서는 통계적 학습 이론 중에서 예측 모형인 SVR을 이용하여 결측치 대체에 의한 데이터 전처리를 수행하였다. 특히, 결측 비율이 높은 희소한 데이터의 정제에서 SVR의 성능이 우수한 것으로 나타났다. 희소성이 적은 데이터의 정제는 오히려 SVR보다는 기존의 Tree Imputation 이 더 좋은 성능을 보여준다. 따라서 웹 로그 데이터와 같은 희소한 데이터의 전처리에 SVR의 사용을 추천한다.

향후 연구과제로는 MCMC(Markov Chain Monte Carlo) 등과 같은 다중 결측치 대체 기법(multiple imputation)[7],[8]을 이용한 희소한 데이터의 정제에 대한 연구도 가능하리라고 본다.

감사의 글

본 연구는 과학 기술부 주관 뇌신경정보학 사업에 의해 지원되었음.

참고문헌

[1] G. Casella, R. L. Berger, "Statistical Inference", Duxbury Press, (1990).

[2] C. Cortes, V. Vapnik, "Support Vector Networks", Machine Learning, vol. 20, 273-297, 1995.

[3] J. Han, K. Kamber, "Data Mining: concepts and Techniques", Morgan Kaufmann Publishers, 2000.

[4] D. C. Hoaglin, F. Mosteller, J. W. Tukey, "Understanding robust and exploratory data analysis", John Wiley & Sons Inc. 2000.

[5] R. J. A. Lavori, R. Dawson, D. Shera, "A Multiple Imputation Strategy for Clinical Trials with Truncation of Patent Data", Statistics in Medicine, vol. 14, 1913-1925, 1995.

[6] R. J. A. Little, D. B. Rubin, "Statistical Analysis with Missing Data", Wiley Interscience, 2002.

[7] D. B. Rubin, "Multiple Imputation for Nonresponse in Surveys", John Wiley & Sons, 1987.

[8] J. L. Schafer, "Analysis of Incomplete Multivariate Data", Chapman and Hall, 1997.

[9] V. Vapnik, "The Nature of Statistical Learning Theory", Springer, 1995.

[10] www.ics.uci.edu/mlearn