퍼지 및 지능시스템학회 논문지 2004, Vol. 14, No. 2, pp. 789-792

Support Vector Regression을 이용한 희소 데이터의 전처리

A Sparse Data Preprocessing Using Support Vector Regression

*전성해, **박정은, **오경환

*Sung-Hae Jun, **Jung-Eun Park, **Kyung-Whan Oh

*청주대학교 통계학과

**서강대학교 컴퓨터학과

요 약

웹 마이닝, 바이오정보학, 통계적 자료 분석 등 여러 분야에서 매우 다양한 형태의 결측치가 발생하여 학습 데이터를 희소하게 만든다. 결측치는 주로 전처리 과정에서 가장 기본적인 평균과 최빈수뿐만 아니라 조건부 평균, 나무 모형, 그리고 마코프체인 몬테칼로 기법과 같은 결측치 대체 기법들을 적용하여 추정된 값에 의해 대체된다. 그런데 주어진 데이터의 결측치 비율이 크게 되면 기존의 결측치 대체 방법들의 예측의 정확도는 낮아지는 특성을 보인다. 또한 데이터의 결측치 비율이 증가할수록 사용가능한 결측치 대체 방법들의 수는 제한된다. 이러한 문제점을 해결하기 위하여 본 논문에서는 통계적 학습 이론 중에서 Vapnik의 Support Vector Regression을 데이터 전처리 과정에 알맞게 변형하여 적용하였다. 제안 방법을 이용하여 결측치 비율이 큰 희소 데이터의 전처리도 가능할 수 있도록 하였다. UCI machine learning repository로부터 얻어진 데이터를 이용하여 제안 방법의 성능을 확인하였다.

Abstract

In various fields as web mining, bioinformatics, statistical data analysis, and so forth, very diversely missing values are found. These values make training data to be sparse. Largely, the missing values are replaced by predicted values using mean and mode. We can used the advanced missing value imputation methods as conditional mean, tree method, and Markov Chain Monte Carlo algorithm. But general imputation models have the property that their predictive accuracy is decreased according to increase the ratio of missing in training data. Moreover the number of available imputations is limited by increasing missing ratio. To settle this problem, we proposed statistical learning theory to preprocess for missing values. Our statistical learning theory is the support vector regression by Vapnik. The proposed method can be applied to sparsely training data. We verified the performance of our model using the data sets from UCI machine learning repository.

Key words : Sparse data, Preprocessing, Support vector machine.

1. 서 론

데이터베이스 기술의 발전과 인터넷 사용의 증가에 따라 지식 추출(knowledge discovery)을 위하여 분석되어지는 데이터의 양은 무한히 증가하고 있다. 하지만 데이터 증가에 따라 분석을 위한 데이터의 품질 향상을 위한 도구들(tools)의 개발은 함께 발전하지 못하는 실정이다. 데이터가 증가하고 복잡해질수록 데이터 정제(data preprocessing)의 문제점도 함께 발생한다. 데이터 정제가 필요하게 되는 이유는 방대한 양의 데이터가 구축되면서 데이터의 결측치(missing value), 잡음(noise), 또는 불일치(inconsistency) 등도 아울러 증가하기 때문이다[3]. 본 논문에서는 데이터 품질에 대한 여러 문제점들 중에서, 특히 결측치 문제를 해결하기 위한 효과적인 모형을 제안하였다. 통계적 학습 모형(statistical learning theory) 중에서도 주로 예측 모형(predictive model)에 사용되는 Vapnik의 Support Vector Regression(SVR)을 이용하였다[2]. 제안 모형의 성능 평가를 위하여 UCI machine Learning Repository의 기계 학습 데이터를 이용하여 기존의 결측치 대체 모형(missing value imputation)들과 비교 실험을 수행하였다. 본 논문의 실험을 통하여 결측치의 비율이 큰 희소 데이터(sparse data)의 전처리에서도 제안 모형은 성능이 좋은 결과를 보여 주고 있음을 알 수 있었다.

접수일자 : 2004년 3월 31일
완료일자 : 2004년 9월 13일

2. Support Vector Regression을 이용한 결측치 대체

2.1 Support Vector Regression

Vapnik은 VC 차원(Vapnik Chervonenkis dimension)에 의한 support vector를 이용하여 주어진 데이터들을 이분법적으로 나눌 수 있는 이상적인 선형평면(hyper linear plane)을 구하는 방법을 소개하였다[9][10]. 평면방정식이 주어졌을 때 분류(classification) 문제를 해결하는 함수식은 다음과 같다.

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

식(1)의 함수식 부호에 의해 분류 모형이 결정된다. x는 입력 벡터(input vector)이고 w는 가중치 벡터(weight vector), 그리고 b는 편이(bias)를 나타낸다. 다음의 그림은 실제

문제 공간에서 이 평면과 방정식이 어떻게 표현되고 적용될 수 있는지 보여준다.

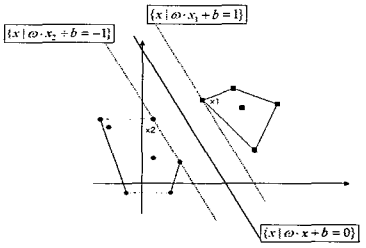

그림 1. 이상 평면의 표현
Fig. 1. Optimal hyper-plane

위의 그림에서 중앙의 굵은 직선을 구해내는 것이 Vapnik의 분류모형의 최종 목표이다. 이러한 이상 평면은 각 개체들(instances))과의 폭을 최대로 하는 분류기(classifier)로서의 조건들을 만족한다. 따라서 주어진 개체들로부터 간격(margin) 폭을 최대화하고 몇 가지 조건식을 만족하는 평면의 방정식을 구해야 한다. 이상평면은 다음의 식을 만족한다.

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

식 (2)에서 점 x 와 평면과의 거리는 다음과 같이 정의된다.

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

위의 식에서 $|w|$는 w의 노름(norm)을 나타낸다[101]. 이상 평면(optimal hyperplane)은 위의 식을 만족하고 점과 평면 사이의 간격을 최대화 하는 w와 b를 구하는 것이다. 이러한 분류 모형은 손실함수(loss function)를 이상 평면 방정식에 포함시킴으로서 회귀(regression) 문제에 적용 될 수 있다. 손실함수란 기대값과 측정값의 오차를 정의하는 함수식이다. 본 논문에서는 희소 데이터에 대해 우수한 성능을 보이는 ε-insensitive 손실함수를 사용한다[10]. Vapnik의 분류 모형에 손실함수를 변형하여 회귀모형을 구축할 수 있는 통계적 학습 모형이 SVR이다[10]. SVR은 다음과 같은 데이터 구조식을 갖는다.

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

위 식은 다음식과 같은 함수 구조로 근사된다.

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

식 (4)와 같은 데이터 구조를 식 (5)의 선형식으로 근사하는 최적의 회귀 함수는 다음의 문제로 표현된다.

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

ε-insensitive 손실함수와 Lagrange 승수(multipliers)를 이용하여 식 (6)의 문제를 풀면 다음과 같은 해를 얻게 된다[2].

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

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

2.2 데이터 전처리와 결측치 대체

기가 바이트(giga byte)가 넘는 데이터를 다루어야 하는 오늘날 데이터의 품질을 높이고 그로 인한 지식 추출의 과정도 효과적으로 이루기 위하여 데이터 전처리는 매우 중요한 이슈가 되었다[3]. 데이터의 정제(cleaning), 통합(integration), 변환(transformation), 축소(reduction)등의 수많은 데이터 정제 방법들이 있다. 데이터 전처리를 하는 가장 큰 이유는 불완전한 데이터 품질을 최대한 높이는 것이다. 본 논문에서는 데이터 전처리의 여러 방법들 중에서 데이터 정제에 대한 기법들 중에서 결측치 대체에 대한 연구를 수행하였다. 데이터 정제에서 결측치가 발생하면 해당 개체를 제거하기도 한다. 이렇게 되면 결국 해당 다른 애트리뷰트(attribute)의 값은 가지고 있지만 결측된 해당 애트리뷰트 때문에 개체가 제거되어 최종적으로는 정보의 손실을 감수해야 하는 문제가 발생한다. 따라서 본 논문에서는 결측치를 포함하고 있는 해당 개체를 제거하는 대신에 해당 애트리뷰트의 값을 예측하여 채워 넣은 결측치 대체 전략을 취하였다.

데이터 집합에서 발생되는 결측 데이터 패턴은 다음의 그림에서 잘 나타내고 있다[6].

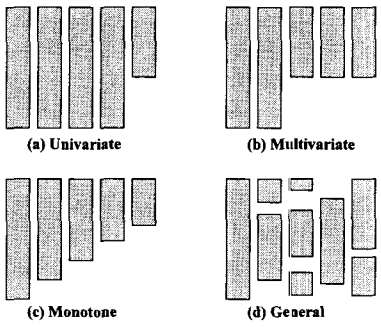

그림 2. 결측 데이터 패턴
Fig. 2. Missing data patterns

위의 그림을 통하여 일반적으로 발생하는 4가지 유형의 데이터 패턴을 볼 수 있다. (a)와 (b) 같은 유형은 일반적인 결측치 대체 모형을 통하여 해결할 수 있다. 하지만 (c)와 같이 조금 복잡한 결측 유형이 발생되면 좀 더 정교한 모형인 마코프 체인 몬테 칼로(Markov Chain Monte Carlo)와 같은 다중 결측치 대체 모형(multiple imputation method)을 사용해 한다[7][8]. 그런데 (d)패턴은 기존의 결측치 대체 모형으로는 해결이 쉽지 않다. 즉 예측된 결측치에 대한 정확도(accuracy)가 떨어진다. 하지만 대부분의 희소 데이터의 결측치 구조는 (d)의 형태를 띠고 있다.

2.3 희소 데이터의 결측치 대체

웹 로그 파일을 비롯한 대부분의 데이터는 (그림 3)와 같

은 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. (그림 3)에서 Page와 User는 각각 애트리뷰트와 개체를 나타낸다. (a)는 결측치를 많이 포함하고 있는 불완전한 데이터 구조이고 본 논문의 SVR을 이용하여 결측치가 대체되어 데이터 품질을 향상시킨 결과로서 나타나는 데이터 구조가 (b)에 나타나 있다. 본 논문의 SVR은 각 애트리뷰트 마다 예측 모형을 구축하여 자신을 제외한 다른 애트리뷰트들을 입력 변수들로 하여 결측된 자신의 애트리뷰트의 해당 결측셀(missing cell)을 채워 넣는다. 다음은 결측치 대체를 위한 이러한 SVR 모형이다.

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

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

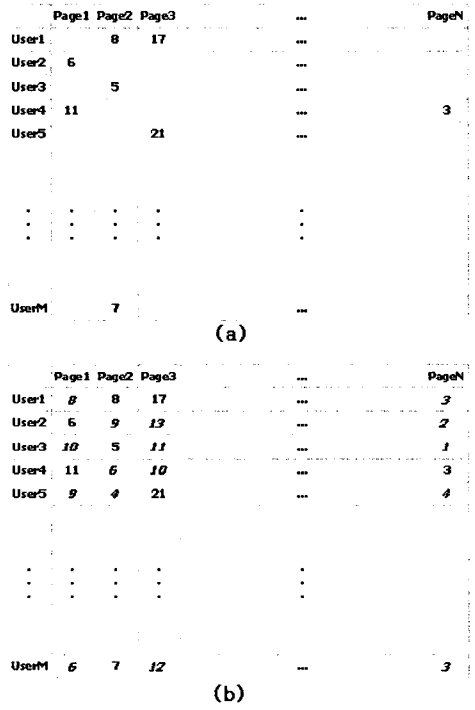

그림 3. 희소한 데이터 구조
Fig. 3. The structure of sparse data

결론적으로 본 논문에서는 기존의 연구들에서는 예측 모형에만 적용되어 오던 통계적 학습 모형인 Vapnik의 SVR을 데이터 전처리 도구로서 적용할 수 있게 하였다.

3. 실험 및 결과

본 논문의 실험을 위하여 UCI Machine Learning Repository의 Glass Identification 데이터와 Iris Plant 데이터를 이용하였다[11]. 첫 번째 실험을 위한 Glass 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) 데이터 구조를 띠고 있다. 본 논문의 실험을 위하여 목표 변수는 사용하지 않고 9개의 입력 변수들만을 사용하여 임의로 5%, 10%, 20%, 30%, 40%, 그리고 50%의 결측 비율(missing rate)을 갖는 데이터 집합을 만들어 실험에 이용하였다. <표 1>에서 SVR과 비교되는 3개의 결측치 대체 모형과의 성능 평가에 대한 결과를 나타내고 있다. 성능 비교는 실제값과 예측값의 제곱 평균을 나타내는 평균제곱오차(MSE, mean squared error)을 이용하였다[1]. 다음은 MSE에 대한 정의이다.

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(y_j - y_j^*)^2 \quad (9)$$

위 식에서 y_j는 알고 있는 j번째의 실제값이고, y_j^*는 j번째의 예측값이다. 이 값이 작을수록 실제값과 예측값의 차이가 작은 것이기 때문에 모형의 예측 정확도가 높다고 결론내릴 수 있다.

표 1. 모형들 간의 MSE (Glass)
Table 1. MSE among comparative models (Glass)

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 2. 모형들 간의 MSE (Iris)
Table 2. MSE among comparative models (Iris)

mssing rate	tree.	dist.	huber	SVR
5%	0.018 (0.155)	0.020 (0.231)	0.031 (0.203)	0.012 (0.043)
10%	0.029 (0.236)	0.031 (0.369)	0.058 (0.308)	0.018 (0.058)
20%	0.048 (0.298)	0.063 (0.431)	0.082 (0.455)	0.025 (0.090)
30%	0.069 (0.339)	0.109 (0.498)	0.155 (0.534)	0.038 (0.088)
40%	0.121 (0.501)	0.188 (0.668)	0.254 (0.703)	0.055 (0.099)
50%	0.215 (0.785)	0.298 (0.831)	0.388 (0.990)	0.068 (0.111)

위의 실험에서 기존에 결측치 대체 기법으로 많이 쓰이고 있는 tree imputation(tree.)[4], distribution based imputation(dist.)[5], 그리고 Huber의 single value imputation(huber)[6]과 본 논문의 SVR을 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 우수해 짐을 알 수 있었다.

두 번째 실험을 위한 Iris 데이터는 붓꽃의 외형을 결정하는 4개의 입력 변수들(SepalLength, SepalWidth, PetalLength, PetalWidth)와 3가지 꽃의 종류(setosa, versicolor, virginica)를 결정하는 1개의 목표변수로 이루어져 있다. Glass 데이터와 마찬가지로 5%에서부터 50%까지의 결측치를 임의로 만들어 성능 평가를 수행하였다. 그러나 Iris 데이터에 대한 실험에서는 Glass 데이터에서의 실험과는 달리 한 번의 실험으로 끝나지 않고 100번의 반복 실험을 수행하였다. 즉

전체 데이터에 대한 결측치 생성을 매번 다르게(random) 발생시켜 실험을 수행하였다.

표 2에서 각 셀의 첫 번째 값은 MSE에 대한 평균(mean)이고, 두 번째 괄호안의 값은 표준 편차(standard deviation)이다. Glass 데이터의 실험결과 마찬가지로 Iris 데이터의 실험 결과에서도 SVR에 의한 결측치 대체 결과의 정확도가 가장 좋게 나온 것을 알 수 있었다. 특히 SVR의 예측값에 대한 표준 편차가 다른 기법들에 비해 작게 나왔다. 이는 그만큼 SVR에 의한 전처리 결과가 다른 모형들에 비해 안정적임을 알 수 있었다.

4. 결론 및 향후 연구과제

본 논문에서는 예측을 위한 통계적 학습 이론인 SVR을 이용하여 결측치 대체에 의한 데이터 전처리를 수행할 수 있는 방법을 제안하였다. 특히, 결측 비율이 높은 희소한 데이터의 정제에서 SVR이 기존의 결측치 대체 기법들에 비해 성능이 우수한 것으로 나타났다. 희소성이 적은 데이터의 정제는 오히려 SVR보다는 기존의 Tree Imputation 등이 더 좋은 성능을 보여주기도 하였다. 따라서 본 논문에서는 웹로그 데이터와 같이 매우 희소한 데이터의 전처리 과정에서 SVR의 사용될 수 있기를 추천한다.

향후 연구과제로는 다양한 손실 함수를 이용한 변형된 통계적 학습 모형을 이용하여 더욱 성능이 우수한 결측치 대체 모형에 대한 구축이 가능하리라고 본다.

참 고 문 헌

[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, nderstanding 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. N. Vapnik, "The Nature of Statistical Learning Theory", Springer, 1995.

[10] V. N. Vapnik, "Statistical Learning Theory", Hohn Wiley & Sons, 1998.

[11] UCI Machine Learning Repository, www.ics.uci.edu/mlearn

저 자 소 개

전성해(Sung-Hae Jun)

1993년 : 인하대 통계학과 (학사)
1996년 : 인하대 통계학과 (이학석사)
2001년 : 인하대 통계학과 (이학박사)
2003년 : 서강대학교 컴퓨터학과 (공학박사 수료)
2003년 : 현재 청주대학교 통계학과 전임강사

관심분야 : 데이터마이닝, 기계학습, 데이터공학
Phone : 043-229-8205
Fax : 043-229-8432
E-mail : shjun@cju.ac.kr

박정은(Jung-Eun Park)

2001년 : 성공회대학교 전산정보학과 (학사)
2003년 : 서강대학교 컴퓨터학과 (공학석사)
2003년~현재 : 서강대학교 대학원 컴퓨터학과 박사과정

관심분야 : 데이터마이닝, 시맨틱웹, 기계학습, 에이전트
Phone : 02-703-7626
Fax : 02-704-8278
E-mail : fayemint@empal.com

오경환(Kyung-Whan Oh)

1978년 : 서강대학교 수학과 (학사)
1985년 : Florida State University, Computer Science(공학석사)
1988년 : Florida State University, Computer Science(공학박사)
1989년~현재 : 서강대학교 컴퓨터학과 교수

관심분야 : 퍼지로직, 인공지능, 다중에이전트
Phone : 02-703-7626
Fax : 02-704-8278
E-mail : kwoh@ccs.sogang.ac.kr