Weather data estimation by sensitive features selection

Tin Nilar Lin† Hiroshi Watanabe†

Abstract— **Weather data estimation is one of the most important issues for disaster management. As the size of the data, period of observation time and the number of influence factors become large and the model will be complex. In this paper, we focus on selecting the optimal feature for reducing data, time and computational complexity without suffering the accuracy. Experimental result shows that our proposed approach can select the valid features and have a good accuracy with smaller dataset and reduce the computational time.**

Keywords— support vector regression, machine learning, feature selection, weather data estimation

1. INTRODUCTION

Sensitive feature selection is one of the main tasks of machine learning and data mining. Depending on the different tasks, there are many different methods applied to find the optimal features that balance the speed and quality of feature selection. In weather datasets, there are many influence factors that affect the estimation of weather and climate data.

In general, using more features should result in more effective prediction. However, some parameters may not important to the expected outcomes. Using this irrelevant information will make the model complex. Thus, the selection of the most sensitive parameters using the reasonable methods without suffering the accuracy is the important. There are several different applications that have been developed with different objectives and tasks using support vectors network [1], [2], and [3]. However, they have their own constraints and have uncertainties in prediction.

In this study, we apply kernel based support vector regression approach for valid features selection to obtain the reasonable accuracy, decrease its computational time and model complexity.

2. METHOD AND DATA USED

2.1 Data Used

The proposed approach has been applied for the sensitive feature selection on the weather dataset from the well-known Kaggle data science repository [4]. The task of the weather dataset is to predict how sales of weather-sensitive products are affected by stormy weather. We rearranged this weather dataset and created as our own dataset for prediction of sea level pressure. Our proposed approach is conducted with 18 features and 1800 instances.

2.2 Support Vector Regression

The basic idea of Support Vector Regression to find the linear regression function $f(x)$ in a high dimensional feature space. Suppose our training dataset T is composed of m samples and n features.

$$T = \{(x_1, y_2), (x_2, y_2), \ldots\ldots\ldots., (x_m, y_m)\}$$

where $x_i \epsilon R^n$ is the input vector of m observed samples and $y_i \epsilon R^1$ is the corresponding target values. The goal of ε-SV regression is to perform the liner regression to find the function f(x) that has at most the predefined deviation ε from the obtained targets, y_i for all training data T [5]. In other words, we neglect the error smaller than predefined threshold of ε and if the error higher than the threshold, we will penalize it. The generic SVR takes the form

$$f(x) = \omega^T \emptyset(x) + b \qquad (1)$$

where $\omega \epsilon R^n$ and $b \epsilon R, \emptyset(x)$ denotes the non-linear transformation function to map the data point into higher dimensional feature space where linear regression is performed. The idea is to determine the optimal function $f(x)$ that can estimate future values accurately by minimizing the loss function.

$$\begin{aligned} &\text{minimize} \quad \tfrac{1}{2}\left||\omega|\right|^2 \\ &\text{subject to} \quad |y_i - (\omega^T x_i + \mathrm{b})| \le \varepsilon \qquad (2) \end{aligned}$$

Such assumption in (2), function $f(x)$ exists that approximates all pairs (x_i, y_i) with ε precision. But sometimes. there is a case to allow some errors. Analogously to loss function, the soft margin loss function is proposed. According to this function, we can define slack variables ξ and ξ^* in loss function to penalize the points beyond the predefined variable ε through the cost parameter C. The SVR solution of convex quadratic optimization problem with minimizing the loss function is as follows:

$$\begin{aligned} &\text{minimizing} \quad \tfrac{1}{2}\left||\omega|\right|^2 + C \textstyle\sum_{i=1}^{m}(\xi + \xi^*) \\ &\text{subject to} \quad y_i - (\omega x_i + b) \le \varepsilon + \xi \qquad (3) \\ &\qquad (\omega x_i + b) - y_i \le \varepsilon + \xi^* \\ &\qquad \xi, \xi^* \ge 0 \end{aligned}$$

where the constant C is regularization parameter and C > 0 determines penalties to estimate the error. Thus, the optimization problem can be expressed in its primal object function by dual set of variables [5].

$$\begin{aligned} L = \tfrac{1}{2}\left||\omega|\right|^2 &+ C \textstyle\sum_{i=1}^{m}(\xi + \xi^*) - \sum_{i=1}^{m}(\eta_i\xi_i + \eta_i^*\xi_i^*) \\ &- \textstyle\sum_{i=1}^{m} \alpha_i(\varepsilon + \xi_I - y_i + \omega^T x_i + b) \\ &- \textstyle\sum_{i=1}^{m} \alpha_i^*(\varepsilon + \xi_i^* - y_i - \omega^T x_i - b) \qquad (4) \end{aligned}$$

where L is the Lagrangian and η_i , η_i^*, ξ_i, ξ_i^* are Lagrange multiplier.

The first partial derivatives of L with respect to the primal variables $(\omega, b, \xi_i, \xi_i^*)$ can be substituting in (4), we obtain the dual formulation of non-linear SVR solution with ε loss function (8).

$$\partial_b L = 0 => \textstyle\sum_{i=1}^{m}(\alpha_i^* - \alpha_i) = 0 \qquad (5)$$

$$\partial_\omega L = 0 => \omega - \textstyle\sum_{i=1}^{m}(\alpha_i - \alpha_i^*)x_i = 0 \qquad (6)$$

†一 Waseda University, Graduate School of Fundamental Science and Engineering

$$\partial_{\xi}^{(*)}L = 0 \Rightarrow \mathrm{C} - \alpha^{(*)} - \eta^{(*)} = 0 \qquad (7)$$

maximizing $-\frac{1}{2}\sum_{i,j=1}^{m} (\alpha_i + \alpha_i^*)(\alpha_j - \alpha_j^*)\phi(x_{i,}-x_j)$

$$-\varepsilon \sum_{i=1}^{m}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{m} y_i(\alpha_i - \alpha_i^*) \qquad (8)$$

subject to $\sum_{i=1}^{m}(\alpha_i + \alpha_i^*) = 0$ and $\alpha_i, \alpha_j^* \epsilon (0, C)$, i, j =1,2…m. $\alpha^{(*)}$ refers to α_i^* and α_i.

In (8), the dual variables η_i , η_i^* through condition (7) by reformulating as $\eta^{(*)}$= C- α_i^* and rewriting the solution of (6) is

$$\omega = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)x_{i,} \text{thus } f(x) = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)(x_i, x) + b \quad (9).$$

In this study, we apply the radial basic function (RBF) kernel function: $K(x_i, x_j) = \exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right), \gamma > 0$. SVR uses the kernel (x_i, x_j) and its transformation function ϕ to map the original data in a higher dimension where γ is a scaling parameter.

It is possible to solve the regression problem in SVR by knowing the kernel function; penalty parameters C, which decides the penalties to estimation error and ε which determines the data under the predefined ε to be neglected in regression and free parameter γ.

3. MODEL EVALUATION AND EXPERIMENT RESULT

In this section, we present the performance of the prediction approach in terms of correlation coefficient (CE), mean absolute error (MAE) and root mean square error (RMSE) using 10 folds cross validation. First, we investigated the correlation coefficient (10) between each observed and predicted value to measure the degree of relationship between the observed and predicted values by applying the SVR approach.

$$\mathrm{CE} = 1 - \frac{\sum_{i=1}^{m}[O_i - P_i]^2}{\sum_{i=1}^{N}[O_i - \bar{O}_i]^2} \qquad (10)$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{m}[|O_i - P_i|] \qquad (11)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{m}[O_i - P_i]^2} \qquad (12)$$

In these equation, O_i and P_i is the observed and predicated values. If O_i and P_i are completely correlated, CE takes the values of 1 or -1.

TABLE I shows the comparison of the correlation coefficient for the selection of sensitive features. We arranged the relative sensitivity of each features from the highest rank to the lowest rank according to their correlation coefficient values of the predicted value.

TABLE I Features Used and Their Relative Sensitivity

Feature	Relative sensitivity (CE)
Departure from normal	0.2399 *
Average Temperature	0.1974 *
Minimum Temperature	0.1965 *
Wind speed	0.1911 *
Precipitation	0.1904 *
Heat	0.1867 *
Wet bulb temperature	0.1815 *
Dewpoint temperature	0.1777 *
Maximum temperature	0.1758 *
Resultant wind speed	0.0888 *
Codesum	-0.0718 *
Resultant wind direction	0.0690 *
Station number	0.0606 *
Cool	0.0513 *
Station pressure	-0.0412
Snowfall	0.0342
Date	-0.0159

Secondly, we selected the features that have high sensitivity value. Notation * in the TABLE I indicates the selected features and we filtered the three unselected features. Finally, we compared the model performance before and after features selection on our dataset as shown in TABLE II.

TABLE II Evaluation of Model Performance

Evaluation Index	Before feature reduced	After feature reduced
MAE	**0.2578**	**0.1426**
RMSE	0.3963	0.3501

MAE and RMSE are adopted to evaluate the performance of the proposed approach. In our study, we use the penalty parameters C ∈ (-3, -2, …, 3) and C ∈ (10, 20, ..., 100) using RBF kernel and ε is ranging from 0.01 to 0.1 and γ ∈ (0.01, 0.02, 0.03, …., 0.1) .The optimal parameters of C and ε are estimated as 1 and 0.01 ($\gamma = 0.01$). After the sensitive feature selection, the performance of the model becomes better with the smaller dataset and reduce the training time. Before the feature selection, the training time is 35.97 sec. and after feature selection, the time decrease to 2.69 sec. It is possible to reduce much time when the number of instances is increased.

4. Conclusion

In this paper, we analyze the SVR to find the optimal subset of features for estimation of the weather data. SVR parameters can be tuned to obtain the good solution.

References

[1] Z. A. Sunkad and Soujanya, "Feature Selection and Hyperparameter Optimization of SVM for Human Activity Recognition," 3rd International Conference on Soft Computing & Machine Intelligence, pp.104-109, Nov. 2016.

[2] H. Zhao and F. Magoulès, "Feature selection for support vector regression in the application of building energy prediction," IEEE 9th International Symposium on Applied Machine Intelligence and Informatics (SAMI), pp.219-223, 2011.

[3] F. Wang, G. Tan, C. Deng, and Z. Tian, "Real-time traffic flow forecasting model and parameter selection based on ε-SVR," 7th World Congress on Intelligent Control and Automation, pp.2870-2875, June 2008.

[4] "Kaggle," Kaggle Inc, [Online]. Available: https://www.kaggle.com /c/walmart-recruiting-sales-in-stormy-weather/data. [Accessed 15 August 2016].

[5] T. Nguyen, Q. N. Huu, and M. J. Li, " Forecasting Time Series Water Levels on Mekong River Using Machine Learning Models," in 7th international Conference on Knowledge and Systems Engineering, pp.292-297, 2015.