Comparison Four Kernels of SVR to Predict Consumer Price Index

M F Rohmah\textsuperscript{1}, I K G D Putra\textsuperscript{2}, R S Hartati\textsuperscript{3}, L Ardiantoro\textsuperscript{4}

\textsuperscript{1,4}Informatics Engineering, Universitas Islam Majapahit, Indonesia
\textsuperscript{2}Informatics Engineering, Universitas Udayana, Indonesia
\textsuperscript{3}Electrical Engineering, Universitas Udayana, Indonesia

Email: \textsuperscript{1}miminfr@gmail.com; \textsuperscript{2}ikgdarmaputra@unud.ac.id; \textsuperscript{3}rukmisari@unud.ac.id; \textsuperscript{4}ipan.ardianto@gmail.com

Abstract. The economy of a region is affected by the stability of food supplies. If the market price of the food supply is stable, the purchasing power level will increase. The price stability of food supplies can be anticipated by using the Support Vector Regression Method, to predict the Consumer Price Index, known as CPI. In the Consumer Price Index assessment, using data based on recording, measurement and calculation of the goods and services average price which consumed by households in a certain period of time. Goods and services that are deemed to represent household expenses are then averaged. The CPI in this study is a type of food supply issued by the Indonesian Central Statistics, and the input variable is taken from the prices of staple commodities in the city of Surabaya, Malang and Kediri based on data from the Siskaperbapo website. To get the supported vector data, the hyperplane maximized by the SVR concept. This concept is able to overcome the overfitting, in order to obtain more accurate prediction results. In predicting the Consumer Price Index, reference data is divided as training data 2016-2019 and testing data 2017-2020. All four kernels were used in the test, namely Spline kernel, Gaussian-RBF kernel, Linear kernel and Polynomial kernel. All four kernels are compared to see their MAPE, this can be shown by the Mean Absolute Percentage Error (MAPE) of less than three, if by using Gaussian RBF kernel. The smallest MAPE value showed by Malang CPI value, which is 1.8242 with $C = 50$, followed by Kediri with the MAPE value of 2.251 with $C = 50$ and MAPE value of Surabaya which is 2.5279 with $C = 50$.

Keywords: four kernel; hyperplane; overfitting; CPI value; SVR.

1. Introduction

A series of processes to explore the added value of a data set in the form of manual unknown knowledge or automatic analysis of complex big data with the aim of discovering patterns or trends that are usually not aware of its existence is called data mining. The data mining process often involves statistical methods, mathematics, and utilizing artificial intelligence technology. Data mining is part of data science by using an algorithm known as machine learning that functions to explore knowledge in datasets. It usually takes a long time to evaluate accuracy and select the best algorithm as the final model. Furthermore, communicating the final result through visualization that can be interpreted by related parties such as the user [1]. Science data is growing rapidly for many reasons, including: Fast growing data sets; Data is stored in a data warehouse; With web and intranet data access increases; Pressure of business competition increases market share in the global economy; Always increasing data; Storage media is growing rapidly [2]. SVR is also known as Support Vector Machine-Regression. The difference between SVM and SVR lies in the output and application of the system [3], [4]. In addition important information from large data, data mining is also used for predictions such as predict the CPI value with expectations of keeping food commodity prices stable. The purpose of using the kernel in research is to implement a model in a higher dimensional space (feature space) without having to define the mapping function from the input space to the feature space, so that for non-linear separable cases the input space is expected to be linear separable in the...
feature space. The hyperplane can be used as a decision boundary efficiently. A test of four kernels, RBF-Gaussian has a good MAPE value with shown values less than three [5], [6].

2. Research Methodology
There are 34 types of input data (t1, t2, t3,..., t34) which are taken from food prices from the siskaperbapo websites in 3 different cities for 4 years from 2016 - 2019. The amount of data as input variables are 1632 for each city. The target variable consists of one attribute (CPI) with 1680 data.

2.1. SVR method
The SVR method purpose is to improve generalized performance by selecting the appropriate use of kernel functions. Therefore, the kernel selection is very important for a particular application [7]. SVR was first introduced and developed from the concept of SVM theory [8]. SVM is a technique for making predictions, both in classification and regression cases [9]. SVM can approach the regression function by using the ε-insensitive loss function concept. The concept of ε-insensitive loss function is used to evaluate how well the regression function is used. The application of SVM in regression cases is called SVR-Regression. SVR is a method that can produce good performance, because it can overcome the problem of overfitting, which is a condition in where a model does not describe the relationship between input and output variables well, but instead describes random error or noise, which will result in poor predictive results [9, [10]. The basic objective of the SVR is to find the function \( f(x) \) which has the most ε deviation from the actual target obtained from all training data, and at the same time the function must be as flat as possible. In other words, the error does not matter, as long as the error is less than epsilon ε. In SVR it is known as support vector, support vector is training data used in testing.
\[ f(x) = w^T \Phi(x) + b \]  \hspace{1cm} (1)

\( \Phi(x) \) is the mapping result of the T function in the input space, \( w \) is weighting vector dimension 1and \( b \) is bias or deviate. The \( w \) and \( b \) coefficients are estimated by minimizing the risk function defined in the equation (1).

The coefficients \( w, b \) minimize the risk function of the following equation:

\[ R = \min \frac{1}{2} ||w||^2 + \frac{1}{l} \left( \sum_{i=1}^{l} L_\varepsilon(y_i, f(x_i)) \right) \]  \hspace{1cm} (2)

SVR will find a function \( f(x_i) \) which has the greatest deviation \( \varepsilon \) from the actual target \( y_i \) for all training data. then with SVR, when \( \varepsilon \) is equal to 0, a perfect regression will be obtained. Conversely, a high \( \varepsilon \) value is associated with a small slack variable value and low accuracy. The addition of this slack variable is to solve the problem of infeasible margin limiter in the optimization problem [9].

2.2. Function Kernel

According to the kernel function is a function \( k \) where all the input vectors \( x, z \) will fulfill [3] the following conditions:

\[ k(x, z) = \phi(x)^T \phi(z) \]  \hspace{1cm} (3)

Regression problems that have a nonlinear pattern (nonlinearity) can be solved by using kernel functions [11]. The four kernel functions used in the consumer price index forecasting test are as follows:
2.3. CPI Calculation

The calculation of the CPI in Indonesia is carried out by Indonesian Statistics. Started from January 2014 until now, CPI data have been obtained from surveys conducted in 82 cities throughout Indonesia. In East Java, the CPI calculated by eight representative cities / districts. This study used three cities, namely Surabaya, Malang and Kediri. The CPI calculation based on a 2012 Cost of Living (SBH) survey. This data used as the basis for determining the commodity package, weight, city coverage and base year for processing the CPI. Every five years, an improvement of the weighing list held, so that the Consumer Price Index value is on the same scale, which is 100. This CPI calculation carried out by Statistics Indonesia every month using the Modified Laspeyres method [12] as follow:

\[
In = \frac{\sum Pn \times Qo}{\sum Po \times Qo} \times 100\%
\]  

\( In = n \)-th index
\( Pn = n \)-th year price
\( Po = \) base year price
\( Qo = \) quantity on base year
Table 2. Example of Consumer Price Index and Change of Percentage in Surabaya 2019 [12]

| NO | Name of Staple | Unit | Yesterday’s Price (Rp) | Price Now (Rp) | Change (Rp) | Change (%) |
|----|----------------|------|------------------------|----------------|-------------|------------|
|    | RICE           |      |                        |                |             |            |
|    | Bunga Rice     | kg   | 12.300                 | 12.300         | 0           | 0.00%      |
|    | Merik Rice     | kg   | 12.167                 | 12.167         | 0           | 0.00%      |
|    | IR 64          | kg   | 9.313                  | 9.313          | 0           | 0.00%      |
|    | SUGAR          |      |                        |                |             |            |
|    | Domestic Sugar | kg   | 12.040                 | 12.040         | 0           | 0.00%      |
|    | COOKING OIL    |      |                        |                |             |            |
|    | Bimoli 620     | ml   | -                      | -              | -           | -          |
|    | Bimoli 2       | L    | 12.275                 | 25.275         | 0           | 0.00%      |
|    | Olive Oil 1    | kg   | 10.200                 | 10.200         | 0           | 0.00%      |
|    | Bimoli 2       | L    | 15.000                 | 15.000         | 0           | 0.00%      |

2.4 MAPE
Mean Absolute Percentage Error (MAPE) is a measure of relative error. MAPE is more accurate because MAPE states the percentage error in the results of estimates or forecasts against the actual results during a certain period. MAPE, the average absolute error over a certain period multiplied by 100% and is also a measure of the relative precision used to determine the percentage of deviation in the estimation results. This approach is useful when the forecast variable size is important in evaluating the accuracy of the forecast. In addition, MAPE indicates how much error in estimating is compared to the real value [14]. In this paper, MAPE is used to measure the performance of four kernel of SVR to predict the Consumer Price Index. The MAPE equation is shown bellow.

\[
\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{X_i - F_i}{X_i} \right| \times 100\%
\]

\(X_i\) = period true value data i
\(F_i\) = data predicts of period i,
\(n\) = predicts time period

If the MAPE value is still below 10%, it can be said that the MAPE value is very good [5].

Table 3. Example of Average Surabaya City Prices at the Consumer Level 2019-11-08 [13]
3. Results and Discussion

3.1. Testing Result

In the first study, it was tried to use four kernels alternately, namely using the Spline, Gaussian-RBF, Linear, and Polynomial but the smallest MAPE value was less than 3 using the Gaussian-RBF kernel. After that followed by the Polynomial kernel and then the Spline kernel.

| City       | C value | MAPE Spline | MAPE Gaussian RBF | MAPE Linear | MAPE Polynomial |
|------------|---------|-------------|-------------------|-------------|----------------|
| Surabaya   | 50      | 6.9425e+91  | 2.5279            | 2.4476e+69  | 3.0268e+43     |
|            | 500     | 1.1275e+91  | 1.9666            | 1.0452e+83  | 2.3946e+73     |
|            | 1000    | 5.8588e+90  | 1.9173            | 2.8185e+89  | 1.8768e+84     |
| Malang     | 50      | 1.8556      | 1.8242            | 1.8559      | 1.8557         |
|            | 500     | 2.2058      | 1.8411            | 2.7973e+06  | 1.2678e+16     |
|            | 1000    | 1.8556      | 1.8415            | 2.1066e+11  | 6.5523e+23     |
| Kediri     | 50      | 1.7592      | 2.2511            | 3.1157      | 3.7857e+03     |
|            | 500     | 1.7594      | 1.8809            | 1.7790      | 5.5784e+16     |
|            | 1000    | 2.2415e+06  | 1.7897            | 1.0292e+13  | 4.9410e+23     |

- Consumer Price Index Curve for the city of Surabaya, Malang city and Kediri city:
  - **Surabaya** (blue curve : prediction, green curve : reference)

![Consumer Price Index Curve for Surabaya, Malang, and Kediri cities](image)

**Figure 3.** CPI Graph in Surabaya (a) City C=50, MAPE = 2.5279 (b) C=1000, MAPE = 1.9173

For the city of Surabaya with C = 1000, the predicted CPI value (in blue) is almost close to the reference...
CPI. And months 8-10, the predictive value is almost the same as the reference value. For the city of Surabaya with \( C = 1000 \), the predicted CPI value (in blue) is almost close to the reference CPI. And months 8-10, the predictive value is almost the same as the reference value.

- **Malang** (blue curve : prediction, green curve : reference)

![CPI Graph in Malang City](image)

Figure 4. CPI Graph in Malang City  
(a) \( C = 50 \), MAPE = 1.8242  
(b) \( C = 1000 \), MAPE = 1.8415

For the city of Malang with \( C = 1000 \), the predicted CPI value (in blue) is almost close to the reference CPI (in green)

- **Kediri** (blue curve : prediction, green curve : reference)

![CPI Graph in Kediri City](image)

Figure 5. CPI Graph in Kediri City  
(a) \( C = 50 \), MAPE = 2.2511  
(b) \( C = 1000 \), MAPE = 1.7897
For the city of Kediri with $C = 1000$, the predicted CPI value (in blue) is almost close to the reference CPI (in green).

The Table can be analyzed that from three cities (Surabaya, Malang and Kediri) with a trial of four kernels (Spline, Gaussian-RBF, Linear, and Polynomial). It can be concluded that the Gaussian-RBF kernel is more appropriate in determining the predicted value, where by trying all $C$ values, MAPE value Gaussian RBF tends to be more stable with MAPE values less than three [5]. With the same iteration $C = 50$, the MAPE value for Malang city is 1.8242, smaller than the MAPE value of Surabaya city and Kediri city. Followed by Kediri with 2.2511 and Surabaya with 2.5279. Malang City has a stable MAPE value which is close to 1.8 tested using four kernels at $C = 50$, it can be concluded that Malang City is relatively stable in the prediction of the Consumer Price Index.

4. Conclusion

The SVR method with 4 kernels has succeeded in predicting the Consumer Price Index in 3 cities, namely Malang, Surabaya, and Kediri. The prediction value of the Consumer Price Index for the three cities; Malang has the smallest MAPE value, which is 1.8242 with $C = 50$, followed by Kediri with the MAPE value of 2.251, $C = 50$ and Surabaya with the highest MAPE value of 2.5279, $C = 50$. From the four kernels modelling methods, the Gaussian-RBF kernel shows the most stable performance among other kernels namely Spline, Linear, and Polynomial kernels. The Gaussian-RBF kernel is the best method because it produces a stable MAPE value.

References

[1] I. H. Witten et.all. 2011. Data Mining: Practical Machine Learning Tools and Techniques. Elsevier Inc.
[2] T.L, Daniel. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons.
[3] Cristianini N., Taylor J.S. 2005. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge Press University, 2000.
[4] G,Bin Gui et all. 2014. Financial Time Series Forecasting Using Support Vector Machine. Tenth International Conference on Computational Intelligence and Security.
[5] Chang et all. 2007. The Development of a Weighted Evolving Fuzzy Neural Network for PCB Sales Forecasting. Elsevier, 32(Expert Systems with Applications), pp. 86-96.
[6] MF, Rohmah et al. 2020. Predicting Consumer Price Index Cities and Districts in East Java with The Gaussian-Radial Basis Function Kernel. J. Phys.: Conf. Ser. 1456 012026.
[7] S,Abe. 2005. Support Vector Machine for Pattern Classification. Springer - Verlag. London Limited.
[8] M,Hossein et.all . 2016. Performance Evaluation of Support Vector Regression Models for Survival Analysis: A Simulation Study. International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6.
[9] Santosa, B., Umam, A. 2018. Data Mining dan Big Data Analysis. Penebar Media Pustaka, Yogyakarta. ISBN : 978-602-5414-76-3.
[10] B Scholkopf , A Smola. 2002. Learning With Kernel. MIT Press.
[11] S,Ruping. 2001. SVM Kernels For Time Series Analysis. 43, Tech. Rep. 43, CS Departement, AI Unit, 44221 Dortmunt.
[12] https://sirusa.bps.go.id/sirusa/index.php/indicator/ index-Indeks Harga Konsumen (IHK).
[13] http://siskaperbapo.com/harga/tabel. Sistim Informasi Ketersediaan dan Perkembangan Harga Bahan Pokok di Jawa Timur.
[14] Heizer, Jay dan Render, Barry. 2015. Manajemen Operasi. Jakarta: Salemba Empat. p.129.
Acknowledgments

The researchers are grateful to the Directorate of Research and Community Service (DPRM-Dikti), LLDIKTI Region VII, the Research Institute at the Islamic University of Majapahit namely LP4MP, which disbursed research funds for the second year of 2020, PKPT scheme. This research was also carried out in collaboration between researchers in the Faculty of Engineering at Udayana University - Bali as TPM (Partner Research Team). The researchers also grateful to the staffs from Statistics Indonesia, East Java and also the staffs of the East Java Province Industry and Trade Office (Disperindag-Jatim), both of whom have taken the time to explain the data about the process of the Consumer Price Index Index.