Classification tide levels in Semarang City use support vector machine

Sugito¹, D Safitri¹, Mustafid¹, D Ispriyanti¹, A Prahutama¹

¹Department of Statistics, Diponegoro University
Jl. Prof. Soedharto, SH, Tembalang, Semarang 50275, Indonesia
E-mail: sugitostat@gmail.com

Abstract. As the capital of Central Java, Semarang is a coastal town located on the north coast of Java. This geographical feature exposes the city to environmental problems such as tidal flood (locally known as rob). Therefore, it is high at stake for the city to classify a certain model to determine the high of tide. One of the expert classification methods of machine-based learning is support vector machine (SVM), it is classified as the nonparametric machine which does not require any assumptions. The classification using SVM requires a kernel as a weight to determine support vector data to classify. Therefore, this study uses Kernel of polynomial and radial base function. As for variables of tidal classification, this study used wind speed and rainfall. On the basis of the analysis, the maximum tide level classification accuracy was carried out on the distribution of 80: 20 training and testing data resulting in a classification accuracy of 69.42%. Classification accuracy was determined by the distribution of training and testing data.

1. Introduction

As the capital of Central Java Province, Semarang is located on the north coast of Java. It is known as one of the flourishing economic centers in Central Java. However, its geographical location nearby the coastline of Java Island exposes its northern area to tidal flood known as rob. This is a disconcerting fact because of the northern area of the city links between the west and the east Pantura roads or vice versa. Moreover, in the north area of the city of Semarang located some industrial and economic centers such as the Kaligawe and Johar market. Kaligawe and Johar market are often exposed to tidal floods caused by high tides. The height of the rob flood varies in accordance with the height of the tide. Thus, more information is required to find out the factors triggering the tides such as wave height, rainfall and wind speed. It highlights the significance of classifying the tidal levels based on these factors.

Statistics is known as a science of great function for modeling and classifying. Currently, data mining is seen as an important part of statistics carried out based on observations and complex variables used as sources of information. One area in data mining is the classification method which is included Classification Regression Tree (CART), C 45, Quick Unbiased Estimator (QUEST), and ID3. Nowadays data mining comes up with development in machine learning. The machine learning method in data mining used for classification is Support Vector Machine (SVM). The SVM method is considered more accurate and faster to use than other data mining classification methods [1, 2]. Its concept relies on the best hyperplane to determine the data that will support vector data for classification. To determine the best hyperplane, the researcher used the Lagrange multiplier [3]. In this study, the maximum height of sea water in the city of Semarang was classified based on rainfall
and wind speed. To see the accuracy of the method, the researcher tried some random training and testing data sharing through several scenarios, while to see the accuracy based on time, the training and testing data was distributed based on time division.

2. Support Vector Machine (SVM)

Basically SVM is originally designed to work in linear classifier principle, which is developed further to work in non-linear cases using the kernel concept in a high-dimensional [4]. Figure 1 is an example of the SVM concept, of the set \( X = \{x_1, x_2, ..., x_m\} \), where the available data is denoted as \( x_i \in \mathbb{R}^n, \ i=1,2, m \).

![Hyperplane concept in SVM](image)

Figure 1. Hyperplane concept in SVM

Figure 1 shows that some patterns are members of two classes +1 and -1. The pattern incorporated in class -1 is symbolized by a red square, while the pattern in class +1 is symbolized by a yellow circle. Meanwhile, to find the best separator hyperplane of the two classes can be done by way of measuring the margin of the hyperplane and finding its maximum point. The best hyperplane is derived by maximizing the margin value to pass the mean between the two classes. The sample hyperplane that is closest to the hyperplane is called vector support. The learning process in SVM aims to look for vector support to obtain the best hyperplane [5].

Principally, SVM is a method for classifying a set of training vectors from two classes \((x_1,y_1), (x_2,y_2), \ldots (x_l,y_l)\), with \( x \in \mathbb{R}^n, \ y \in \{-1,1\} \) and a hyperplane \( (w^T \cdot x) + b = 0 \)

The vector set is said to be perfectly separated if it is separated without causing any error with a close distance between the vector and the maximum hyperplane. There are several redundancies in equation, without reducing its generality to consider canonical hyperplane, where the parameters \( w \) and \( b \) are limited by [6]:

\[
\min_{x_i} |w^T x + b| = 1
\]

In separated hyperplane with canonical form, must be required some of constraint

\[
y_i ((w \cdot x_i) + b) \geq 1, i = 1, \ldots, l,
\]

while for the distance from \( x \) to hyperplane \((w,b)\) as follows [5]:

\[
d(w,b;x) = \frac{|w^T \cdot x + b|}{||w||}
\]

For optimal hyperplane is given by \( \text{margin } \rho(w,b) \) as [7]:

\[
\rho(w,b) = \min_{\{x_i, y_i=1\}} d(w,b; x_i) + \min_{\{x_i, y_i=-1\}} d(w,b; x_i)
\]
\[
L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{l} \alpha_i \left( (x_i \cdot w) \cdot b \right) y_i - 1, \]

where \(\alpha_i\) is called Lagrange multiplier. The solution of this Lagrange function can be obtained by minimizing \(L\) against primal variables. When the solution is obtained (optimal point), the gradient \(L = 0\), so that the optimal value of the Lagrange function is given by:

\[
\begin{align*}
\frac{\partial L}{\partial b} = 0 & \implies \sum_{i=1}^{l} \alpha_i y_i = 0 \\
\frac{\partial L}{\partial w} = 0 & \implies w = \sum_{i=1}^{l} \alpha_i x_i, y_i,
\end{align*}
\]

So the dual problems as follows:

\[
\max_{\alpha} W(\alpha) = \max_{\alpha} \left\{ \min_{\alpha} L(w, b, \alpha) \right\}, \text{ and the solution of this equation as follows as:}
\]

\[
\hat{\alpha} = \arg \min_{\alpha} \frac{1}{2} \sum_{i=1}^{l} \alpha_i \sum_{j=1}^{l} \alpha_i \alpha_j (y_i y_j) \langle x_i, x_j \rangle - \sum_{i=1}^{l} \alpha_i,
\]

where \(\alpha_i \geq 0, i = 1, 2, \ldots, l\)

\[
\sum_{i=1}^{l} \alpha_i y_i = 0
\]

The solution can be used to determine the Lagrange function multiplier and the optimal divider hyperplane [3]:

\[
w = \sum_{i=1}^{l} \alpha_i x_i, y_i \quad \text{and} \quad \hat{b} = -\frac{1}{2} w \cdot (x_i + x_j),
\]

where \(x_i\) and \(x_j\) are support vector for each class with the requirement as:

\[
\hat{\alpha}_i, \hat{\alpha}_j > 0, \quad y_i = 1, \quad y_j = -1.
\]

So that the classification equation with soft margins uses the equation:

\[
f(x) = \text{sign} \left( w^T \cdot x + \hat{b} \right)
\]

Whereas for soft margins using linear interpolation is:

\[
f(x) = h \left( w^T \cdot x + \hat{b} \right) \quad \text{where} \quad h(x) = \begin{cases} 
-1 & x < -1 \\
1 & -1 \leq x < 1 \\
0 & x = 1 
\end{cases}
\]

In general, data that has been linearly separated will fulfill the following equation:

\[
\|w\|^2 = \sum_{i=1}^{l} \hat{\alpha}_i = \sum_{i=1}^{l} \sum_{j=1}^{l} \hat{\alpha}_i \hat{\alpha}_j \langle x_i, x_j \rangle y_i y_j
\]

In reality, there is a high probability of misclassification. To overcome this, the formulation that has been done previously will be expanded so that non-separable data can be used. Previous optimization problems both on objective and constraint functions are modified by following the slack variable \(\xi > 0\) which is a measure of misclassification. The following is a modified constraint for non-separable cases [1]:

\[
y_i \left( w^T \cdot x_i + \hat{b} \right) \geq 1 - \xi, i = 1, 2, \ldots, l
\]

Hyperplane or optimal separator is determined by vector \(w\), that is by minimizing the function:
\[ \Phi(w, \xi) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} \xi_i \]

Where \( C \) is a regulatory parameter used to control the relationship between variable slack and \( \|w\|^2 \).

The dual form of the Lagrange problem becomes:

\[
\max_{\alpha} w(\alpha) = \max -\frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i, x_j) + \sum_{i=1}^{l} \alpha_i
\]

The solution as

\[
\arg \min_{\alpha} \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i, x_j) + \sum_{i=1}^{l} \alpha_i
\]

With constrain as:

\[ 0 \leq \alpha_i \leq C, \quad i = 1, 2, \ldots l \]

\[ \sum_{i=1}^{l} \alpha_i y_i = 0 \]

And the solutions to this problem as follow as:

\[ w = \sum_{i=1}^{l} \hat{\alpha}_i y_i x_i \quad \text{and} \quad \hat{b} = -\frac{1}{2} w(x_i + x_j) \]

In the explanation of the above section, the data has been explained using linear data. Furthermore, the SVM method will be discussed for non-linear data, where this data can only be separated using non-linear surface. If all data is transmitted to feature space, then the hyperplane that is in the feature space is then re-transformed into the input space. Basically non-linear data classification has optimization in \( \alpha \)

\[ \alpha = \arg \min_{\alpha} \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) + \sum_{i=1}^{l} \alpha_i \]

The value of \( K(x, y) \) is a kernel function that shows non-linear mapping in feature space. This equation gives the hard classifier the separation of the hyperplane in feature space, with the equation

\[ f(x) = \text{sign} \left( \sum_{\alpha \in S_{\nu}} \alpha_i y_i K(x, x_i) + \hat{b} \right) \]

With values as:

\[ w^T \cdot x = \sum_{S_{\nu}} \hat{\alpha}_i y_i K(x, x_i) \quad \text{and} \]

\[ b = -\frac{1}{2} \sum_{\alpha \in S_{\nu}} \alpha_i y_i [K(x, x_i) + K(x_i, x_i)] \]

The kernels function in SVM as follow as [5]:

1. Linear: \( x^T x \)
2. Polynomial: \( (x^T x_i + 1)^p \)
3. Radial basic function (RBF): \( \exp (-\frac{1}{2\alpha^2} \|x - x_i\|^2) \)

3. Methodology of Research

This study used the classification of a maximum height of 1.25 m (1) and a maximum height of 4 m (2). The independent variables used for classification were wind speed and rainfall. Data were derived from the Meteorology, Climatology and Geophysics Agency of Semarang city for the period of January 2017 to August 2017 based on daily data. For classification, the researcher used SVM with several divisions of training and testing, namely 90:10; 80: 20 and 70: 30. The sharing of training and testing were based on time and holdout. Holdout is a method of randomly retrieving training and testing data in a group of data. In addition, the training and testing size was tested based on the following scenario. The first scenario is January 1st 2016 – July 31st 2017 as training data and August...
1st to August 31st 2017 as testing data. The second scenario is January 1st, 2016- June 30th 2017 as training data and July 1st to August 31st 2017 as testing data. The third scenario is January 1st 2016 - May 31st 2017 as training data and June 1st to August 31st 2017 as testing data. The fourth scenario is January 1st 2016 – April 30th 2017 as training data, while those from May 1st - August 31st 2017 were used as data testing. The SVM classification method relied on 2 types of weighting, namely polynomial and Radial Basis Function (RBF). The polynomial order values used in the study were 1 and 2, while the sigma value on the RBF function used was 1, 2, and 3.

4. Results and Discussion

On the basis of descriptive statistics of the research data, it is prominent that a wave height reaching a maximum of 1.25 m amounted to 352, while the wave height reaching a maximum of 4 m amounted to 256 data. Meanwhile, the average wind speed from January 1st 2016 to August 31st 2017 was 5.2 km/hour and the average rainfall reached 6.47 mm. In this study classification was carried out using the Support Vector Machine (SVM) method with kernel Radial Base Function (RBF) function and with a value of $\sigma = 2$. Distribution of training and testing were conducted by the holdout method.

Table 1. The accuracy of SVM method based on training testing by Holdout

| Proportion of Training:Testing | Polynomial P=1 | Polynomial P=2 | Polynomial P=3 | RBF $\sigma = 1$ | RBF $\sigma = 2$ | RBF $\sigma = 3$ |
|--------------------------------|----------------|----------------|----------------|-----------------|-----------------|-----------------|
| 90:10                          | 58.33%         | 58.33%         | 66.67%         | 65%             | 60%             |
| 80:20                          | 57.85%         | 66.12%         | 69.42%         | 66.12%          | 63.64%          |
| 70:30                          | 58.01%         | 61.88%         | 64.09%         | 64.64%          | 60.22%          |
| 60:40                          | 64.46%         | 61.98%         | 66.53%         | 64.05%          | 61.98%          |

Based on table 1, it is seen that the highest accuracy value is found in SVM grouping using 80:20 Training and Testing data, with a kernel function of polynomial with $p = 2$ and base function radial kernel function with value of $\sigma = 1$. In this data grouping, the accuracy rate amounted to 66.12% for the polynomial, while the rbf function reached 66.94%. On the other hand, the smallest level of accuracy was found for the polynomial kernel function with $p = 1$ with training: 80:20 testing with classification accuracy of only 57.85%, whereas the RBF kernel function came up with the smallest accuracy value of $\sigma = 3$ by 60%.

Table 2. Classification result for Testing Data in the best SVM method

| Group    | Prediction 1st Class | Prediction 2nd Class | Total |
|----------|----------------------|----------------------|-------|
| 1st Class| 61                   | 9                    | 70    |
| 2nd Class| 28                   | 23                   | 51    |
| Total    | 89                   | 32                   | 121   |

Based on table 2, it is observed that the number of testing data was 121. Subsequently, it was followed by training and testing based on time.

Based on table 3, it is revealed that the highest accuracy value for the distribution of training testing based on the time contained in SVM grouping using Training data from 1 January 2016-31 May 2017 and Testing on 1 June to 31 August 2017 with the used kernel function of radial base function reached the value of $\sigma = 1$. In this grouping the level of accuracy only amounted to 40.22%, whereas the level of accuracy using the polynomial kernel function for $p = 1$ or $p = 2$ was relatively the same with that of each training and testing data distribution.

Table 3. The accuracy value of SVM method in training and testing based on time
Meanwhile, the smallest accuracy value in the RBF kernel function was = 3 for the distribution of training data from January 1st 2016 - May 31st 2017. In the kernel polynomial function the smallest accuracy value was 32.26% for either p = 1 or p = 2 for training distribution from January 1st 2016 until July 31st 2017. On the use of the Polynomial kernel function, the accuracy value is relatively the same for all training and testing sharing scenarios. However, in RBF kernels, the accuracy value is more varied for each scenario. The best classification of the SVM model with the distribution of training and testing data on time is as follows:

**Table 4.** Classification of testing data the best SVM model based on time

| Group          | Prediction | 1st Class | 2nd Class | Total |
|----------------|------------|-----------|-----------|-------|
| 1st Class      | 11         | 4         | 15        |
| 2nd Class      | 51         | 26        | 77        |
| Total          | 62         | 30        | 92        |

Based on table 4, it can be seen that the number of testing data is 92. Compared to Table 2, data testing is based on a smaller amount of time than the percentage of training and testing data. Accuracy using the Holdout method produces better accuracy than the distribution of training and testing based on time.

**Conclusion.**

The classification of the maximum wave height of 1.25 m and 4 m using the SVM method produces an accuracy of 69.42% with the distribution of training and testing in Holdout. It is important to determine the amount of training and testing data to make sure the classification accuracy. Classification accuracy using SVM based on time division results in poorer accuracy than that of the holdout method.

**References**

[1] Chen H L, Yang B, Lie J and Liu D Y 2011 *Expert Systems with Applications* **5864** 9.
[2] Yasin H, Caraka H E, Tarno and Hoyyi A 2016 *A Global Journal of Pure and Applied Mathematics* **12** 4.
[3] Gun S 1998 *Support Vector Machines for Classification and Regression* (Southampton: University of Southampton).
[4] Lee Y J 2001 *Support Vector Machines in Data Mining* (Madison: University of Wisconsin).
[5] Vapnik V N 1999 *The Nature of Statistical Learning Theory* (New York: Springer).
[6] Addeh A and Ebrahimzadeh A 2012 *Journal of Medical Signal and Sensor* **2** 22.
[7] Huang C M, Lee Y J, Lin D K J and Huang S Y 2007 *Computational Statistics and Data Analysis* **52** 335.