Classification of rainfall data using support vector machine

H Sain1*, H Kuswanto2, S W Purnami2 and S P Rahayu2

1Department of Mathematics, Universitas Tadulako, Indonesia
2Department of Statistics, Institut Teknologi Sepuluh Nopember, Indonesia

*Email: hartayuni.sain@untad.ac.id

Abstract. Classification problems can be based on cross-section or time-series data. In general, some of the characteristics found in time series data are data that are very susceptible to containing noise and outliers. In this paper, time series data-based classification was carried out using the support vector machine (SVM) method. The SVM method is the most popular binary classification technique in machine learning. The advantage of this method is that it can find a global optimum solution and always achieve the same solution for every running. Another advantage is that it can solve the over-fitting problem by minimizing the upper limit of generalization errors. SVM classification model performance can be seen from classification accuracy and sensitivity and specificity tests. The results showed that based on the level of classification accuracy used the SVM method resulted in an accuracy rate of the prediction results with training data of 96.3% and the accuracy of the prediction results with testing data of 90.08%. Therefore, the classification of rainfall data using the SVM method has a very good performance.

1. Introduction

One of the studies of Machine Learning is classification. Classification is the process of finding a function that describes and distinguishes data classes with the aim that the function can be used in predicting the class of an object or data whose class is unknown. Classification is characterized as a method for determining a collection of models or functions to describe and distinguish concepts or data classes in order to predict classes of specific objects or to define labels of unknown objects [1].

Several machine learning-based modeling techniques were developed to assist in classification problems including Artificial neural networks (McCulloch and Pitts, 1943), Classification Adaptive Regression Tree (Breiman, Friedman, Olshen and Stone, 1984), Multivariate Adaptive Regression Spline (Friedman, 1991), k-Nearest Neighbor (Fix and Hodges, 1951), Support Vector Machine (Cortes and Vapnik, 1995) and so on.

The support vector machine (SVM) is one of those modelling techniques focused on machine learning for data classification. Vapnik [2] first implemented the SVM method as a prediction tool for classification and regression problems and was very successful. As this approach can be performed analytically (mathematically), the SVM method can effectively function for pattern recognition and become very common in machine learning. This can be done as SVM can specify the optimum global solution and always have the same solution for all runs [3]. As in addition, SVM has performed as an effective tool for dealing with overfitting problems by minimizing the upper limit of generalization of error, based on structural risk minimization theory. However, it should be noted that SVM was once theoretically and very efficiently built for the problem of classification in the resolution of two-class
classification [4]. Simply put, the principles of SVM function to decide the optimum hyperplane separating two groups [5].

The heartbeat time series classification[6], benchmark UCR time-series datasets[7], motif-based classification[8], Early Classification Model (ECM) and patient-specific classification of multivariate time series[9], brain activity classification[10], and multivariate time series classification[11] are many studies related to the use of the SVM approach in time series data classification. SVMs have been used more recently in many biomedical applications: age-related gait degeneration [12], EEG signal classification [13], brain-computer interfacing (BCI) [14, 15], scoliosis analysis and prediction [16, 17], and electromyogram analysis [18].

Originally, SVM was used for binary data classification, but real-world issues are not only issues with binary classification, but rather multi-class issues. In practice, decomposition is made into a set of binary issues in order to solve the problem of multi-class data classification so that standard SVM can be implemented directly. The main aim of this study is how the support vector machine method can be used to classify time series data.

2. Methods
2.1. Support vector machine
The Support Vector Machine (SVM) method was introduced for the first time by Vapnik et al in 1995 [2] and has been very successful in making predictions, both in classification and regression cases. For every running, the SVM method seeks the optimal global solution and always reaches the same solution. By mapping training information to a high-dimensional space, SVM works. In a high dimensional space, a classifier that is able to maximize the margin between two data classes will be sought. The use of this method tries to find the optimal separator function that can separate two data sets from two different classes (-1, +1) or also known as the best hyperplane among infinite functions.

A linear classifier, namely classification cases that can be separated linearly, is the basic concept of SVM. SVM is classified into 2 groups in the linear classification, namely separable and non-separable.

Given a set of \( X = \{x_1, x_2, \ldots, x_n\} \), with \( x_i \in \mathbb{R}^n \), \( i = 1, \ldots, n \). Knowing that \( X \) is a specific pattern, that if \( x_i \) is a member of a certain class then \( x_i \) given label (target) \( y_i = +1 \), otherwise \( y_i = -1 \). Data will therefore be given as a pair \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\) which is a training vector set of two groups that will be categorized by SVM [19-20],

\[
(x_i, y_i), \quad x_i \in \mathbb{R}^n, \quad y_i \in \{-1, 1\}, \quad i = 1, \ldots, n
\]  

(1)

from a normal vector parameter called \( w \), a separating hyperplane is defined, and parameter defines relative plane position towards coordinate center called \( b \), the two parameters are shown as follows

\[
(w^T.x) + b = 0
\]

(2)

It must fulfill this restriction when determining canonical-shaped hyperplane separation [19-20],

\[ y_i [(w^T.x_i) + b] \geq 1, \quad i = 1, 2, \ldots, n \]

(3)

to get the ideal hyper-plane is done by maximizing the margin \( \frac{2}{||w||} \) or minimizing the following functions:

\[
\Phi(w) = \frac{1}{2} ||w||^2
\]

(4)

The optimization issue can then be solved by the Lagrange function, where \( \alpha_i \) is the multiplier of Lagrange. The function of Lagrange is primal space, so it must be converted into dual space so that the function can be solved simpler and more effectively. Therefore, the dual space solution can be obtained as follows:
\[ \alpha^* = \arg \min_{\alpha} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \cdot \mathbf{x}_j) - \sum_{i=1}^{n} \alpha_i \]  
subject to,

\[ \alpha_i \geq 0, \quad i = 1, 2, \cdots, n \]  
and

\[ \sum_{i=1}^{n} \alpha_i y_i = 0 \]  
thus the classification using the following formula [20],

\[ f(x) = h(\hat{w}^T \cdot x + \hat{b}) \]  
where,

\[ h(x) = \begin{cases} 
-1 & x < -1 \\
1 & -1 \leq x \leq 1 \\
1 & \text{otherwise} \end{cases} \]  

\[ \hat{w} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \mathbf{x}_i \]  
\[ \hat{b} = -\frac{1}{2} \hat{w}(\mathbf{x}_r + \mathbf{x}_s) \]  

In fact, not all data can be considered linearly separable, so defining a particular linear hyperplane is very hard. By converting the data into a higher dimensional feature space, this issue can be solved so that the data can be segregated linearly in the new feature space. On nonlinear data, SVM also works. Basically, the following optimization problem[20] is solved by nonlinear classification.

\[ \hat{\alpha} = \arg \min_{\alpha} \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{l} \alpha_i \]  

\( K(\mathbf{x}_i, \mathbf{x}_j) \) It is a kernel function that displays the nonlinear mapping of the space of the feature. This equation provides a hard hyperplane classifier that separates the feature space

\[ f(x) = \text{sign} \left( \sum_{i \in S} \hat{\alpha}_i y_i K(\mathbf{x}_i, \mathbf{x}) + \hat{b} \right) \]  
where

\[ \hat{\omega} \cdot \mathbf{x} = \sum_{i \in S} \hat{\alpha}_i y_i K(\mathbf{x}_i, \mathbf{x}) \]  

\[ \hat{b} = -\frac{1}{2} \sum_{i \in S} \hat{\alpha}_i y_i [K(\mathbf{x}_r, \mathbf{x}_i) + K(\mathbf{x}_s, \mathbf{x}_i)] \]  

The linear kernel, the polynomial kernel, and the Gaussian radial basic function (RBF) are some of the kernel functions which are widely used.

2.2. Evaluation of model performance
The overall performance of the classification model is demonstrated by classification accuracy, where the higher the classification accuracy, the better the performance of the classification model.
Average Classification Accuracy = \[ \frac{1}{G} \left( \sum_{i=1}^{G} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i} \right) \] (13)

To obtain an optimal and more specific classification, sensitivity, and specificity can be tested.

\[
Sensitivity = \frac{\sum_{i=1}^{G} tp_i}{\sum_{i=1}^{G} (tp_i + fn_i)}
\] (14)

\[
Specificity = \frac{\sum_{i=1}^{G} tp_i}{\sum_{i=1}^{G} (tp_i + fp_i)}
\] (15)

3. Methodology
3.1. Data source and description
The data source used in this study is secondary data in the form of monthly rainfall data for the period 1999 to 2015 obtained from the National Center for Environmental Prediction - NOAA on the website http://www.esrl.noaa.gov. The data used only select 1 observation station in East Nusa Tenggara Province, namely the Komodo station with data descriptions in Table 1.

In 1993, McKee developed a drought index calculation using the SPI method for the first time. The help is to detect and minimize dryness. Meteorological drought analysis using the SPI method can be carried out over a period of one month, three months, six months, twelve months, and so on according to the purpose of the analysis. In this study, the SPI method was used for a period of 3 months. To find out the nature of SPI can be seen in Table 2.

**Table 1. Data description**

| Data     | Y               | Total |
|----------|-----------------|-------|
| Komodo   | Near Normal (2) | 134   |
| Severe   | Wet (1)         | 32    |

**Table 2. SPI Class and Weather Nature**

| Drought index | Whether nature     | Drought index | Whether nature |
|---------------|--------------------|---------------|----------------|
| SPI≥2.0       | Extremely wet      | SPI≥1.0       | Very Wet       |
| 1.5≤SPI≤1.99  | Very wet           | -0.99≤SPI≤0.99| Near Normal    |
| 1.0≤SPI≤1.49  | Moderately wet     | -1.49≤SPI≤-1.0| Moderately Dry |
| 0.99≤SPI≤0.99 | Near Normal        | -1.99≤SPI≤-1.5| Severe Dry     |
| SPI≤-2.0      | Extremely Dry      | SPI≤-1.0      | Severe Dry     |
4. Results and Discussion

Figure 1 shows the time series plot for rainfall data usage. It shows a pattern formed from plots of monthly rainfall data over a 16 year period. As for the rainfall data plot based on the SPI classification, it can be described in Figure 2. The data that will be used consist of monthly rainfall data \(X_1\), SPI value \(X_2\) and Weather nature \(Y\).

![Figure 1. The plot of time series data](image1)

For 16 years, the average condition for Indonesia has been near normal, meaning that it is neither too dry nor too humid. From Figure 1, Indonesia experienced a severe drought, but not yet considered an extreme one, namely in August 2000, August 2004, and August 2006.

![Figure 2. The plot of rainfall data based on SPI classification](image2)
The SVM method of rainfall data classification is carried out using only radial kernels with parameter values of $C = 1$ and $\sigma = 0.33$. The data used in this study will be partitioned into training data and testing data with a ratio of 40%: 60%. Obtained the number of training data as many as 81 data and for testing data as many as 121 data. The results of the accuracy are presented in Table 3 as follows:

| Data     | Sensitivity | Specificity | Accuracy |
|----------|-------------|-------------|----------|
|          | Training    | Testing     | Training | Testing |
| Komodo   | 95.15%      | 87.81%      | 97.4%    | 96.3%    | 90.08% |

Table 3 shows that the results of the accuracy of time series data classification using the SVM method in this study are that the accuracy value for training data is higher than the accuracy value for testing data.

5. Conclusion

The results showed that the results of the time series data classification model performance evaluation, in this case, were monthly rainfall data using the SVM method. It can be concluded that the level of accuracy of the prediction results with training data is 96.3%, for the sensitivity value obtained is 95.15%, then the specificity value was 97.4%. As for the level of accuracy from the prediction results with testing data of 90.08%, for the sensitivity value obtained for 87.81%, then the Specificity value of 92.97%. Therefore, the classification of rainfall data using the SVM method has a very good performance.

References

[1] Han J, Kamber M 2001 *Data Mining Concepts & Techniques* USA: Academic Press
[2] Vapnik V 1995 *The Nature of Statistical Learning Theory* Springer-Verlag, NY
[3] Farquad M A H and Bose I 2012 *Decision Support System* 53 226-233
[4] Kumar M, Gromiha M Michael and Raghava G P S 2008 *Proteins* 71(1) 189-194
[5] Rahman F and Purnami W Santi 2012 *Journal Sains Dan Seni ITS* 1 (1) ISSN 2301928X
[6] Kompouraki A, Manis G and Nikou C 2008 *IEEE Engineering in Medicine and Biology Society* 13(4) 512-8
[7] Alweshah M, Rashaideh H, Hammouri A I, Tayyeb H and Ababneh M 2017 *International Journal of Data Analysis Techniques and Strategies* 9 (3)
[8] Buza K 2009 *Data Handling and Business Intelligence* pp 105-114
[9] Ghalwash M F, Ramjak D and Obradovic Z 2012 *IEEE International Conference on Bioinformatics and Biomedicine* 978-1-4673-2560-8
[10] Chaovalitwongswe W and Pardalos P M 2008 *Cybernetics and Systems Analysis* 44 125–138
[11] Orsegino C and Vercellis C 2010 *Pattern Recognition* 43 3787–3794.
[12] Begg R, Palaniswami M and Owen B 2005 *IEEE Trans. Biomed. Eng.* 52 (5) 828–838
[13] Guler I and Ubeysi E D 2007 *IEEE Trans. Inf. Technol. Biomed* 11 (2) 117–126
[14] Lal T N, Schroder N, Hinterberger T, Weston J, Bogdan M, Birbaumer N and Scholkopf B 2004 *IEEE Trans. Biomed. Eng* 51 (6) 1003–1010
[15] Kaper M, Meinicke P, Grossekathoefer U, Lingner T and Ritter H 2004 *IEEE Trans. Biomed. Eng* 51 (6) 1073–1076
[16] Ramirez, L, Durdle N, Hill D and Raso J 2005 *IEEE Trans. Inf. Technol. Biomed* 9 (2) 276–282
[17] Ramirez L, Durdle N, Raso J and Hill D 2006 *IEEE Trans. Inf. Technol. Biomed* 10 (1) 84–91.
[18] Liang H and Lin Z 2001 *IEEE Trans. Biomed. Eng* 48 (5) 601–604
[19] Gunn Steve 1998 *Support Vector Machines for Classification and Regression* ISIS Technical Report, Image Speech and Intelligent System Group, Southampton: University of Southampton
[20] Sain H and Purnami S Wulan 2015 *In Procedia Computer Science* **72** 59-66