HASIL CEK_SGY_TIA_1_Rough_Set_English

by Sgy_tia_1 Rough_set_english Rough_set_english

Submission date: 25-Oct-2021 09:18AM (UTC+0700)
Submission ID: 1683015691
File name: SGY_TIA_1_Rough_Set_English.pdf (461.9K)
Word count: 5086
Character count: 24111

*Corresponding author. Sugiyarto, Department of Mathematics, Ahmad Dahlan University, Yogyakarta, Indonesia. E-mail: sugiyarto@math.uad.ac.id.

New Fuzzy Support Vector Machine Classification With Dimensional Reduction Based on Rough set

Sugiyarto Surono[1*],Tia Nursofiyani[2], Rohmatul Fajriyah[3]

[1,2]Department of Mathematics, Ahmad Dahlan University, Yogyakarta, Indonesia
[3]Department of Statistics, Universitas Islam Indonesia, Yogyakarta, Indonesia

ABSTRACT

Support vector machine (SVM) is one of the most promising data classification engines because of its high generalization performance and wide application for classification, but one of the problems with the SVM method is when it is applied to high-dimensional data. In overcoming this obstacle, the dimension reduction using rough set will be used. In addition, SVM is very sensitive to noise and outliers which causes a decrease in accuracy in classifying data. To deal with this, one technique that can be applied is to combine SVM with fuzzy logic or often called fuzzy support vector machine (FSVM). The study proposes a new approach, namely building a membership function in FSVM by comparing the distance of positive and negative samples to the center of the class using the function of Chebyshev and Manhattan distances, then the fuzzy membership degree value is calculated by the Gaussian formula. The proposed FSVM is applied to data that has been dimensionally reduced using the rough set. The two proposed FSVM algorithms (Chebyshev and Manhattan FSVMs) show that dimensional reduction using the rough set can improve the accuracy of the Chebyshev and Manhattan FSVMs. Based on the analysis of the selected data training and testing scenarios (60:40, 70:30, 80:20, 90:10), Chebysev FSVM algorithm can effectively eliminate the impact of noise and improve classification accuracy. Chebyshev FSVM provides the best accuracy of 98.19% on the ratio of data training and testing 90:10.

Keyword: Classification; Fuzzy support vector machine; Support vector machine, Membership function, Rough set

1. Introduction

Classification is the process of grouping data, with the data used having a class or label (supervised data). Support vector machine (SVM) is one of the classification algorithms that has recently received attention because of its high generalization performance and wide application for classification.

The SVM method, which is introduced by Vapnik in 1995 based on structural risk minimization [6], is one of the superior methods of the supervised algorithm that works to find a classification model or a set of optimal separators (hyperplane) from training data so as to be able to separate datasets into binary or multi-class. It can classify linear and nonlinear data by building an optimal hyperplane. Building the optimal hyperplane means finding the hyperplane that has the maximum margin, the perpendicular distance from the support vector to the hyperplane.

The latest research related to SVM performance was carried out by Silva in 2020[1], in this study the SVM, ANN, KNN, fuzzy logic and RF (random forest) methods were used for the classification of driving models. The results of this study indicate that SVM has the highest accuracy value of 96%. SVM has been widely applied in real-world problems in many areas such as text categorization [2], image categories [3], pest detection [4], and identification of agricultural systems [5].

In some literatures it is stated that SVM has several weaknesses, as described by Chen (2011) where when using data with high dimensions (features) is one of the obstacles in the application of the SVM method [7]. To deal with this reducing features becomes very important. Dimensional reduction is a

*Corresponding author. Sugiyarto, Department of Mathematics, Ahmad Dahlan University, Yogyakarta, Indonesia. E-mail: sugiyarto@math.uad.ac.id.

technique of identifying important variables and eliminating irrelevant variables to build a good learning model. By reducing the dimension reduction on the contrary, there are many advantages including avoiding over-fitting, reducing the complexity of data analysis and improving data analysis performance [8].

Rough set is a dimension reduction method that can identify significant variables and eliminate irrelevant variables to produce a good learning model, so as to reduce the dimensions of the data without lacking the information contained in the data set. Research related to SVM and rough set was carried out by Meng et al (2018).

In addition to complex problems, the SVM method is experiencing a lot of interference due to noise and outliers, this causes a decrease in the generalization performance of SVM [10]. To solve these problems, there are many ways that can be applied, such as combining the SVM method with fuzzy logic [10-12]. Previous research applies fuzzy logic with two events, namely using fuzzy rules [14], [15] and fuzzy membership functions (fuzzy membership values) [16]–[20] on each sample so that the new sample input makes a different contribution to eliminating the effect of noise and outliers, and is expected to improve the performance of the generalized SVM classification.

The SVM method with a combination of fuzzy logic is called the Fuzzy Support Vector Machine (FSVM) where the membership function is a crucial step in classification using FSVM [21]. In 2018, Xiao-Kang Ding conducted a research entitled "Optimization and Analysis on Fuzzy SVM for Objects Classification". The research shows that building a membership function based on the comparison of the distance of the sample point to the positive and negative class center (Euclidean's distance) in general has a better performance than the conventional membership calculation method.

Based on these previous studies, then in this paper the dimension reduction of using rough set is implemented first before the binary classification by using FSVM. The new membership function algorithms based on the Chebyshev and Manhattan distances to compare the distance of the sample point to the positive and negative class center is calculated. Moreover in calculating the degree of membership, these two algorithms use the Gaussian curve formula. These two FSVM models are evaluated using Indian pima diabetes dataset.

The paper describes introduction in section 1 and it is followed by some explanation about rough set concept related in section 2. The SVM is described in section 3. Section 4 explains about Fuzzy Support Vector machine (FSVM). Section 4 is result and discussion and finally section 5 is conclusion and remarks.

2. Rough set

Rough set is an intelligent mathematical tool developed by Prof. Pawlak in 1982 to deal with uncertainty and incompleteness [23]. It is based on the concept of upper and lower approximation of a set, model and set space. Performing data analysis using rough set requires several basics knowledges as in the following subsections.

2.1. Distance measure

Distance measures is a method that works to measure the level of similarity of two objects in terms of the geometric distance of the variables included in the two objects. Some of the distance functions that can be used are Euclidean distance, Chebyshev distance, Manhattan distance, Minkowski distance, Hamming distance, Mahanalobis distance, etc. Each of these distance functions has advantages and disadvantages. According to Syaripudin (2013) Chebyshev distance has the advantage that this distance measurement is very sensitive to objects that have outliers. In addition to the Chebyshev distance, Manhattan distance is often used because of its ability to better detect special circumstances such as the presence of outliers [22].

2.2. Information System

The set in the rough set is reflected as a table, with the rows in the table reflecting the data (objects) and the columns in the table reflecting the variables from the data [24]. Broadly speaking, the information system is denoted as IS.

Information system is defined as follows:

$$IS = (U, A, V, f) \qquad (1)$$

Where U is Universal and A is a finite set of variables $\{a_1, a_2, \dots, a_n\}$. Each variable $a \in A, f_a : U \to V_a$ where V_a (the domain of the variable) is the set of variable values.

*Corresponding author. Sugiyarto, Department of Mathematics, Ahmad Dahlan University, Yogyakarta, Indonesia. E-mail: sugiyarto@math.uad.ac.id.

2.3. Indiscernibility relation **tdk ada referensi sumber**

Indiscernibility relation which is generally denoted IND is a data relation that can have the same value for a conditional variable. $IS = (U, A, V, f)$ is an information system such that the Indiscernibility relation data according to the variable D ($D - indiscernibility$) is defined

$$IND_{IS}(D) = \{(x, x') \in U^2 \mid \forall a \in B\, a(x) = a(x')\} \quad (2)$$

If $(x, x') \in IND_{IS}(D)$ then x and x' are data that cannot be distinguished from each other by the variable D. Equivalent classes are the classes which is equivalent to $IND_{IS}(D)$ and denoted as $[x]_D$.

2. 4. Reduct **tdk ada referensi sumber**

In rough set method there are redundant variables that will not affect the results of the classification if omitted. Therefore, this variable can be omitted without losing its true value. Reduct is a set of variables that can produce the same classification as if all variables were used. A variable that is not used in the reduct is a variable that is dispensable (not needed) in the classification process.
Suppose IS=(U,A) where B⊆A and a∈B means that a is dispensable (not needed) in variable B if IND_IS=IND_IS (B-{a}) and vice versa. If it is indispensable (required) it means that a is indispensable in variable B.

2.5. Core **tdk ada referensi sumber**

Dimensional reduction can remove redundant variables to provide simpler informed decisions. Core C is the set of variables that belong to the intersection of all reductions of C. It is defined as follows:

$$COR(C) = \cap_D \in_{RED(C)} l \quad (3)$$

3. Support Vector Machine (SVM) **tdk ada referensi sumber**

Support vector machine (SVM) is a supervised learning method that works by finding the best hyperplane that can separate two datasets from two different classes. Finding the best hyperplane is equivalent to measuring the margin of the hyperplane and finding its maximum point. Margin is the distance between the hyperplane and the closest pattern of each class. This closest pattern is called a support vector.

The designation of the classification problem is an example of a training dataset $M = \{(x_i, y_i)\}_{i=1}^{n}$ with n in m dimension. For all $x_i \in R^m$ there is exist a class of $y_i \in [1, -1]$. It is notated as $\{(x_1, y_1,), (x_2, y_2), \ldots, (x_n, y_n)\}$. Hyperplane can be represented as $w^T x_i b = 0$. Thus, the objective function to obtain the optimal hyperplane can be expressed as follows:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=0}^{n} \xi_i \quad (4)$$

$$y_i(w^T x_i + b) \geq 1 - \xi_i\,;\, \xi_i \geq 0; \quad 1 \leq i \leq n \quad (5)$$

Generally, classification cases that occur in everyday life are non-linear problems. For separate non-linear data, kernel functions $K(x_i, x_j)$ are applied.

Some common choices of kernel functions are given as follows:

1. Linear Kernel SVM

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T, \mathbf{x}_j \quad (6)$$

2. Polynomial Kernel (degree *d*)

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + 1 \quad (7)$$

3. Radial Basis Function (RBF) Kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\right. \quad (8)$$

4. Fuzzy Support Vector Machine (FSVM) **tdk ada referensi sumber**

FSVM is an extension method of SVM proposed by Lin and Wang in 2002. It serves to reduce

*Corresponding author. Sugiyarto, Department of Mathematics, Ahmad Dahlan University, Yogyakarta, Indonesia. E-mail: sugiyarto@math.uad.ac.id.

the sensitivity of SVM to outliers or noise. A training dataset $M = \{(x_i, y_i)\}_{i=1}^{n}$ with n in m dimension. For all $x_i \in R^m$ there is exist a class of $y_i \in [1, -1]$. It is notated as $\{(x_1, y_1,), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\}$. After applying fuzzy logic $(membership\, function\, fuzzy(s_i))$ with $0 < s_i \leq 1$, the training dataset is denoted as $\{(x_1, y_1, s_1), (x_2, y_2, s_2), (x_3, y_3, s_3), \ldots, (x_n, y_n, s_n)\}$. Determining the optimal hyperplane in this problem is given as follows:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=0}^{n} s_i \xi_i \quad (9)$$

$$y_i(w^T x_i + b) \geq 1 - \xi_i ; \xi_i \geq 0; \quad 1 \leq i \leq n \quad (10)$$

where C is the penalty constant. To solve the optimization problem, Langrangian is used and the problem can be changed to

$$\max W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j x_i x_j \quad (11)$$

$$subject\ to \sum_{i=1}^{n} y_i \alpha_i = 0 \quad 0 \leq \alpha_i \leq s_i C, \ i = 1,2,\ldots,n \quad (12)$$

With s_i is a fuzzy membership function with a value $0 < s_i \leq 1$, $x_i x_j$ is $K(x_i x_j)$ if the data is nonlinear.

4.1. New Fuzzy Membership Function for FSVM tdk ada referensi sumber

We study two FSVM models with membership functions are based on the comparison of each data sample point to the positive and negative class centers using the Chebyshev and Manhattan distances. The data is divided into two classes (binary) such that the class center can be defined as vector average of the variables in the data. The calculation to find the class center is as follows:

$$x^+_{center} = \frac{1}{n^+}\sum_{i=1}^{n^+} x_i^+ \quad (13)$$

$$x^-_{center} = \frac{1}{n^-}\sum_{i=1}^{n^-} x_i^- \quad (14)$$

where x^+_{center} and x^-_{center} is the average of the positive and negative classes. The n+ is number of data points in the positive class and the n_ is number of data points in the negative class. The two algorithms are denoted by Chebyshev and Manhattan FSVMs. The value of the degree of membership is computed by the Gaussian curve formula. This formula is

$$G(x_i; \sigma, c) = e^{\frac{-(x_i - c)}{2\sigma^2}} \quad (15)$$

4.1.1. FSVM Chebyshev tdk ada referensi sumber

In Chebyshev FSVM, the first step is to calculate the distance of each sample point to the center of the positive and negative classes using the Chebyshev distance formula. The membership function can be built using the following equation:

$$s_i = \begin{cases} f(d_i^+),\ if\ \left(max\left|x_i^+ - x^+_{center}\right|\right) \geq \left(max\left|x_i^+ - x^-_{center}\right|\right) \\ 1, if\ \left(max\left|x_i^+ - x^+_{center}\right|\right) < \left(max\left|x_i^+ - x^-_{center}\right|\right) \\ 1, if\ \left(max\left|x_i^- - x^+_{center}\right|\right) > \left(max\left|x_i^- - x^-_{center}\right|\right) \\ f(d_i^-), if\ \left(max\left|x_i^- - x^+_{center}\right|\right) \leq \left(max\left|x_i^- - x^-_{center}\right|\right) \end{cases} \quad (16)$$

Where

$$d_i \begin{cases} \left(max\left|x_i^+ - x^+_{center}\right|\right) \\ \left(max\left|x_i^- - x^-_{center}\right|\right) \end{cases} \quad (17)$$

When the distance of the positive sample to the center of the positive class is less than the distance of the

*Corresponding author. Sugiyarto, Department of Mathematics, Ahmad Dahlan University, Yogyakarta, Indonesia. E-mail: sugiyarto@math.uad.ac.id.

sample to the center of the negative class, then it is considered as a "useful point", and its membership is set as 1. When the distance of the positive sample to the center of the positive class is greater than the distance to the center of the negative class, then the point is considered as a “noisy point”, and the value of the degree of membership is calculated according to the membership function formula of the Gaussian curve; and vice versa.

4.1.2 FSVM Manhattan tdk ada referensi sumber

Similar to Chebyshev, the Manhattan FSVMs membership function is determined by calculating the Manhattan distance from each sample point of the sample to the positive class center with the following equation:

$$s_i = \begin{cases} f(d_i^+),\ if \left(\sum_{i=1}^{n^+} \left| x_i^+ - x_{center}^+ \right| \right) \geq \left(\sum_{i=1}^{n^+} \left| x_i^+ - x_{center}^- \right| \right) \\ 1, if \left(\sum_{i=1}^{n^+} \left| x_i^+ - x_{center}^+ \right| \right) < \left(\sum_{i=1}^{n^+} \left| x_i^+ - x_{center}^- \right| \right) \\ 1, if \left(\sum_{i=1}^{n^-} \left| x_i^- - x_{center}^+ \right| \right) > \left(\sum_{i=1}^{n^-} \left| x_i^- - x_{center}^- \right| \right) \\ f(d_i^-), if \left(\sum_{i=1}^{n^-} \left| x_i^- - x_{center}^+ \right| \right) \leq \left(\sum_{i=1}^{n^-} \left| x_i^- - x_{center}^- \right| \right) \end{cases} \quad (18)$$

Where

$$d_i \begin{cases} \left(\sum_{i=1}^{n^+} \left| x_i^+ - x_{center}^+ \right| \right) \\ \left(\sum_{i=1}^{n^-} \left| x_i^- - x_{center}^- \right| \right) \end{cases} \quad (19)$$

When the distance of the positive sample to the center of the positive class is less than the distance of the sample to the center of the negative class then it is considered a "useful point", and the membership degree value is 1. When the distance of the positive sample to the center of the positive class is greater than the distance to the center of the negative class, then the point is considered as a “noisy point”, and the value of the degree of membership is calculated according to the membership function formula of the Gaussian curve; and vice versa.

3. Classification Model Performance Evaluation tdk ada referensi sumber

Classification performance is the ability of the classification model to predict the class of data. The results of the classification can be arranged in a confusion matrix. The confusion matrix is also known as the error matrix. In the confusion matrix there are several terms that are commonly used, namely:

a) *True Positives* (TP) = Number of correct predictions in the positive class
b) *True negatives* (TN) = Number of correct predictions in the negative class
c) *False positives* (FP) = Number of wrong predictions in the positive class
d) *False negative* (FN) = Number of wrong predictions in the negative class

Several methods can be used fro evaluate the performance, namely

1. *Sensitivity* (SE) tdk ada referensi sumber

Sensitivity is the ratio of the completeness or accuracy of the correct prediction of positive data. The sensitivity value describes how many samples of the positive class are correctly labeled. The sensitivity value can be calculated by the formula:

$$SE = \frac{TP}{TP + FN} \quad (20)$$

2. *Specificity* (SP) tdk ada referensi sumber

Specificity is the ratio of the completeness or accuracy of the correct prediction of negative data. The specificity value describes how many samples of the negative class are correctly labeled. The specificity value can be calculated by the formula:

$$SP = \frac{TN}{TN + FP} \quad (21)$$

3. G-means (GM) tdk ada referensi sumber

G-means or geometric means is the multiplication of accuracy in the positive class (sensitivity) and accuracy in the negative class (specificity). The G-means value can be calculated by the formula:

$$GM = \sqrt{SE \times SP} \quad (22)$$

4. Result and Discussion

In this study we apply the Chebyshev and Manhattan FSVM algorithms on the “Pima Indian Diabetes” dataset in which eight input variables and one output variable are obtained from UCI. Parts of the data set are shown in Table 1 The output variables in

*Corresponding author. Sugiyarto, Department of Mathematics, Ahmad Dahlan University, Yogyakarta, Indonesia. E-mail: sugiyarto@math.uad.ac.id.

this data set fall into two categories: "1" and "-1". The "1" indicates that the patient has diabetes and "-1" indicates that the patient does not have diabetes.

Tabel 1. DataSet Pima Indian Diabetes

No	X1	X2	X3	X4	X5	X6	X7	X8	Y
1	6	148	72	35	0	33.6	0.6270	50	1
2	1	85	66	29	0	26.6	0.3510	31	-1
3	8	183	64	0	0	23.3	0.6720	32	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
767	1	126	60	0	0	30.1	0.3490	47	1
768	1	93	70	31	0	30.4	0.3150	23	-1

Prior to classification using the two algorithms, the Pima Indian Diabetes dataset was dimensionally reduced using roughest. Dimensional reduction using Rough gives the result that with the original 8 input variables, it produces 5 input variables which can be seen in Table 2.

Table 2. The Pima Indian Diabetes DataSet section after rough set dimension reduction

No	X1	X2	X3	X4	X6	y
1	6	148	72	35	33.6	1
2	1	85	66	29	26.6	-1
3	8	183	64	0	23.3	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
767	1	126	60	0	30.1	1
768	1	93	70	31	30.4	-1

The 'new' five input variables will be used as input in the classification process using the Chebyshev's and Manhattan FSVMs.

The principle of the Fuzzy Support Vector Machine Algorithm is to minimize the objective function such that the optimal hyperplane is obtained to classify the data. The basic concept of the Fuzzy Support Vector Machine is first to determine the value of the fuzzy membership degree of the data that will be used as a new variable in the input data. In order to get that value, the center of the positive and negative classes must be determined first. It is obtained by using equations (14) and (15). After that, Determining the value of the degree of membership is based on the comparison of the distance from the data sample point to the center of the positive class using equations (16) and (18). Once the value of the degree of fuzzy membership is obtained, then the FSVM input dataset can be written as

$$\left\{(x_1, y_1, s_1), (x_2, y_2, s_2), \ldots, (x_n, y_n, s_n)\right\}.$$

Determining the optimal hyperplane in this FSVM problem is equivalent to solving optimization problems in equations (10) and (11) by changing the form to langrange according to equations (12) and (13). The input data for Chebyshev and Manhattan FSVMs were processed by modifying the amount of training data and testing data. The scenarios for training data and testing data used are 60%:40, 70%:30%, 75%:25%, and 80%:20%.

To see the performance evaluation, the geometric mean of sensitivity was used to evaluate the Chebyshev and Manhattan FSVMs. It shows the comprehensive classification performance. Sensitivity (SE) represents the accuracy of the positive class, specificity (SP) is the accuracy of the negative class, and GM shows the overall accuracy. The results of the performance evaluation of the Chebyshev and Manhattan FSVMs are shown in Table 3 and Figure 1.

Table 3. Accuracy Results of Chebyshev and Manhattan FSVM with Second Degree Polynomial Kernel

Training and Testing Scenarios	Performance Evaluation	FSVM Chebyshev	FSVM Manhattan
60% : 40%	GM	92.34%	94.01%
	SE	97.44%	94.5%
	SP	87.5%	93.51%
70% : 30%	GM	94.76%	93.84%
	SE	97.95%	98.63%
	SP	91.66%	89.28%
80% : 20%	GM	93.47%	92.55%
	SE	99.02%	97.08%
	SP	88.23%	88.23%
90% : 10%	GM	98.19%	97.19%
	SE	100%	97.95%
	SP	96.42%	96.42%

*Corresponding author. Sugiyarto, Department of Mathematics, Ahmad Dahlan University, Yogyakarta, Indonesia. E-mail: sugiyarto@math.uad.ac.id.

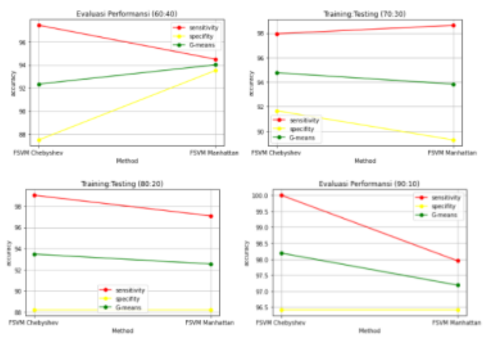

Figure 1. Comparison of the Accuracy Levels of Chebyshev and Manhattan FSVMs with a Second Degree Polynomial Kernel

From Table 3 it can be seen that in the training and testing scenario of 90%:10%, the accuracy for the positive class (sensitivity (SE)) reaches 100%. This indicates that the hyperplane capability produced by Chebyshev FSVM classifies samples from class '1' (showing that patients with diabetes) was 100% accurate with 0% misclassification.

Based on Figure 1, from the overall scenario of training data and data testing used in this study, Chebyshev's FSVM gives the best overall accuracy (G-means(GM)) results where the highest value is given by the training and testing data partition 90:10. In this partition Chebyshev FSVM gave a score of 98.19% which indicates that the hyperplane ability of this method shows the ability of the separator function to distinguish class 1 (indicating that the patient has more diabetes) or -1 (indicating that the patient does not have diabetes) is very good. To see the effect of implementing dimension reduction using roughest on Chebyshev and Manhattan FSVMs, the results from Table 3 will be compared with Table 4 where in Table 4 are the results of Chebyshev and Manhattan FSVM s without dimension reduction using the rough set.

Looking at Table 3 and Table 4, it can be seen that dimension reduction with rough set increases the accuracy of Chebyshev's and Manhattan's FSVM

Table 4. Results of Chebyshev and Manhataan FSVM Accuracy with Second Degree Polynomial Kernel Without Dimensional Reduction

Training and Testing Scenarios	Performance Evaluation	Chebyshev FSVM	Manhattan FSVM
60% : 40%	GM	77.07%	87.05%
	SE	83.16%	83.92%
	SP	71.42 %	90.30%
70% : 30%	GM	91.92%	93.37%
	SE	95.91%	92.76%
	SP	88.09 %	94.80%
80% : 20%	GM	95.99%	94.59%
	SE	100%	97.08%
	SP	92.15 %	92.15%
90% : 10%	GM	95.37%	94,37%
	SE	92.85%	95.91%
	SP	97.95 %	92.85%

5. Conclusion

In this paper, we introduce the Chebyshev and Manhattan FSVMs, where a new membership function is proposed to improve classification accuracy. In FSVM, the membership function is built based on the comparison of the distance from the data sample point to the position of the positive and negative class centers calculated using the Chebyshev and Manhattan distance functions respectively, then the fuzzy membership degree value is calculated using the Gaussian formula. Chebyshev's FSVM will be applied to data that has been reduced in dimensions using a rough set.

Based on the results and discussion in the previous section, it is concluded that the rough set calculation produces 4 variables as the main variables from the initial 8 variables. The four variables are the number of times pregnant, plasma glucose concentration 2 hours in the glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, and Body Mass Index. The best classification accuracy of dimensional reduction data with rough set is given by Chebyshev FSVM, with ratio of training and testing data is 90%:10%, that is 98.19%. This means that the

*Corresponding author. Sugiyarto, Department of Mathematics, Ahmad Dahlan University, Yogyakarta, Indonesia. E-mail: sugiyarto@math.uad.ac.id.

proposed FSVM Chebyshev performs the best to reduce the noise and outliers effects such that the classification is optimum. In addition, looking at the comparison of research results, it can be concluded that dimension reduction with rough set can increase the accuracy of Chebyshev and Manhattan FSVMs.

Daftar Pustaka

[1] I. Silva and J. E. Naranjo, "A Systematic Methodology to Evaluate Prediction Models for Driving Style Classification," *Sensors (Switzerland)*, vol. 20, no. 6, pp. 1–21, 2020.

[2] S. Chatterjee, P. George Jose, and D. Datta, "Text Classification Using SVM Enhanced by Multithreading and CUDA," *Int. J. Mod. Educ. Comput. Sci.*, vol. 11, no. 1, pp. 11–23, 2019.

[3] M. A. Chandra and S. S. Bedi, "Survey on SVM and their application in image classification," *Int. J. Inf. Technol.*, 2020.

[4] M. A. Ebrahimi, M. H. Khoshtaghaza, S. Minaei, and B. Jamshidi, "Vision-based pest detection based on SVM classification method," *Comput. Electron. Agric.*, vol. 137, pp. 52–58, 2017.

[5] Y. Lebrini *et al.*, "Identifying Agricultural Systems Using SVM Classification Approach Based on Phenological Metrics in a Semi-arid Region of Morocco," *Earth Syst. Environ.*, vol. 3, no. 2, pp. 277–288, 2019.

[6] V. Vanpik and C. Cortes, "Support Vector Network," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.

[7] H. L. Chen, B. Yang, J. Liu, and D. Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 9014–9022, 2011.

[8] E. S. Wahyuni, "Penerapan Metode Seleksi Fitur Untuk Meningkatkan Hasil Diagnosis Kanker Payudara," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 7, no. 1, p. 283, 2016.

[9] Z. Meng, Z. Jian, L. Guoxi, and Z. Kai, "Performance prediction of a modular product variant with RS-SVM," *Procedia CIRP*, vol. 79, pp. 495–499, 2019.

[10] Q. Wu and W. Wang, "Piecewise-Smooth Support Vector Machine for Classification," *Math. Probl. Eng.*, vol. 2013, 2013.

[11] C. F. Lin and S. De Wang, "Fuzzy Support Vector Machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 464–471, 2002.

[12] W. An and M. Liang, "Fuzzy Support Vector Machine Based on within-Class Scatter for Classification Problems with Outliers or Noises," *Neurocomputing*, 2013.

[13] T. Guernine and K. Zeroual, "New Fuzzy Multi-Class Method to Train SVM Classifier," *3rd Int. Conf. Adv. Databases, Knowledge, Data Appl.*, no. c, pp. 77–82, 2011.

[14] J. L. Castro, L. D. Flores-Hidalgo, C. J. Mantas, and J. M. Puche, "Extraction of fuzzy rules from support vector machines," *Fuzzy Sets Syst.*, vol. 158, no. 18, pp. 2057–2077, 2007.

[15] K. Wu, M. Zhou, X. Sean Lu, and L. Huang, "Fuzzy logic based text classification method for social media data," *2017 IEEE Int. Conf. Syst. Man, Cybern. SMC 2017*, vol. 2017-Janua, pp. 1942–1947, 2017.

[16] M. Mohammadi and M. Sarmad, "Robustified distance based fuzzy membership function for support vector machine classification," *Iran. J. Fuzzy Syst.*, vol. 16, no. 6, pp. 191–204, 2019.

[17] D. Xiaokang, Y. Lei, Y. Jianping, and Z. Zhaozhong, "Optimization and Analysis on Fuzzy SVM for targets Classification in Forest," *Open Cybern. Syst. J.*, vol. 10, pp. 155–162, 2016.

[18] S. G. Chen and X. J. Wu, "A New Fuzzy Twin Support Vector Machine for Pattern Classification," *Int. J. Mach. Learn. Cybern.*, vol. 9, no. 9, pp. 1553–1564, 2018.

[19] Congqin-Yi, R. Zhou, and K. Hu, "Fuzzy Support Vector Machine for Breast Cancer Gene Classification Congqin-Yi," pp. 676–679, 2017.

[20] D. Xiaokang, Y. Lei, Y. Jianping, and Z. Zhaozhong, "Optimization and Analysis on Fuzzy SVM for Objects Classification," *Open Cybern. Syst. J.*, vol. 9, pp. 1421–1429, 2018.

[21] W. Liu, L. L. Ci, and L. P. Liu, "A New Method of Fuzzy Support Vector Machine Algorithm for Intrusion Detection," *Appl. Sci.*, vol. 10, no. 3, 2020.

[22] U. Syaripudin, I. Badruzaman, E. Yani, D. K, and M. Ramdhani, "Studi Komparatif

*Corresponding author. Sugiyarto, Department of Mathematics, Ahmad Dahlan University, Yogyakarta, Indonesia. E-mail: sugiyarto@math.uad.ac.id.

Penerapan Metode Hierarchical, K-Means Dan Self Organizing Maps (SOM) Clustering Pada Basis Data," *Istek*, vol. VII, no. 1, pp. 132–149, 2013.

[23] M. A. Rahman, "Penerapan Metode Rough Set Dalam Memprediksi Penjualan Perumahan (Studi Kasus Di PT. Anugerah Pasenda Pekanbaru)," *Univ. Dharmawangsa*, vol. 14, no. April, pp. 5–48, 2020.

[24] A. Prajana, F. Sains, T. Universitas, I. Negeri, A. Raniry, and B. Aceh, "Penerapan Teory Rough Set Untuk Memprediksi Tingkat Kelulusan Siswa Dalam Ujian Nasional Pada Sma Negeri 5 Kota Banda Aceh," *J. Islam. Sci. Technol.*, vol. 2, no. 1, pp. 75–88, 2016.

HASIL CEK_SGY_TIA_1_Rough_Set_English

ORIGINALITY REPORT

16%	11%	11%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	journal.uad.ac.id Internet Source	2%
2	docshare.tips Internet Source	2%
3	Muhammad Kamal, Nur Mohammad Farda, Ilham Jamaluddin, Artha Parela, Ketut Wikantika, Lilik Budi Prasetyo, Bambang Irawan. "A preliminary study on machine learning and google earth engine for mangrove mapping", IOP Conference Series: Earth and Environmental Science, 2020 Publication	1%
4	Sugiyarto Surono, Rizki Desia Arindra Putri. "Optimization of Fuzzy C-Means Clustering Algorithm with Combination of Minkowski and Chebyshev Distance Using Principal Component Analysis", International Journal of Fuzzy Systems, 2020 Publication	1%
5	hdl.handle.net Internet Source	1%

6	www.docstoc.com Internet Source	1 %
7	www.scribd.com Internet Source	<1 %
8	pt.scribd.com Internet Source	<1 %
9	www.open-open.com Internet Source	<1 %
10	Zheng, J.. "A support vector machine classifier with automatic confidence and its application to gender classification", Neurocomputing, 201105 Publication	<1 %
11	journals.ums.ac.id Internet Source	<1 %
12	Chen, H.L.. "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis", Expert Systems With Applications, 201107 Publication	<1 %
13	open.library.ubc.ca Internet Source	<1 %
14	"Advances in Computing and Data Sciences", Springer Science and Business Media LLC, 2018 Publication	<1 %

15 Salim Rezvani, Xizhao Wang, Farhad Pourpanah. "Intuitionistic Fuzzy Twin Support Vector Machines", IEEE Transactions on Fuzzy Systems, 2019 — <1 %
Publication

---

16 coek.info — <1 %
Internet Source

---

17 dokumen.pub — <1 %
Internet Source

---

18 studyres.com — <1 %
Internet Source

---

19 www.hindawi.com — <1 %
Internet Source

---

20 Juan Pineda-Jaramillo. "Travel time, trip frequency and motorised-vehicle ownership: A case study of travel behaviour of people with reduced mobility in Medellín", Journal of Transport & Health, 2021 — <1 %
Publication

---

21 export.arxiv.org — <1 %
Internet Source

---

22 M D Murjadi, Z Rustam. "Fuzzy Support Vector Machines based on Adaptive Particle Swarm Optimization for Credit Risk Analysis", Journal of Physics: Conference Series, 2018 — <1 %
Publication

23	Mozhgan Nasr Azadani, Azzedine Boukerche. "Toward Driver Intention Prediction for Intelligent Vehicles: A Deep Learning Approach", 2021 IEEE 46th Conference on Local Computer Networks (LCN), 2021 Publication	<1%
24	Submitted to National Chiao-Tung University Student Paper	<1%
25	Weiya Shi. "An Improved Kernel Principal Component Analysis for Large-Scale Data Set", Lecture Notes in Computer Science, 2010 Publication	<1%
26	M Irfan, A R Nurhidayat, A Wahana, D S Maylawati, M A Ramdhani. "Comparison of K-Nearest Neighbour and support vector machine for choosing senior high school", Journal of Physics: Conference Series, 2019 Publication	<1%
27	Shiqing Zhang. "Speech emotion recognition based on Fuzzy Least Squares Support Vector Machines", 2008 7th World Congress on Intelligent Control and Automation, 06/2008 Publication	<1%
28	repository.its.ac.id Internet Source	<1%

29	Sugen Chen, Junfeng Cao, Zhong Huang, Chuansheng Shen. "Entropy-Based Fuzzy Twin Bounded Support Vector Machine for Binary Classification", IEEE Access, 2019	Publication	<1 %
30	research-publication.com	Internet Source	<1 %
31	scitepress.org	Internet Source	<1 %
32	ejurnal.stmik-budidarma.ac.id	Internet Source	<1 %
33	epdf.pub	Internet Source	<1 %
34	www.freepatentsonline.com	Internet Source	<1 %
35	"Smart Monitoring of Rotating Machinery for Industry 4.0", Springer Science and Business Media LLC, 2022	Publication	<1 %
36	Abha Gejji, Shruti Shukla, Siddhee Pimparkar, Tamanna Pattharwala, Anand Bewoor. "Using a Support Vector Machine for building a Quality Prediction Model for Center-less Honing process", Procedia Manufacturing, 2020	Publication	<1 %

37	Hongjian Fan, K. Ramamohanarao. "A weighting scheme based on emerging patterns for weighted support vector machines", 2005 IEEE International Conference on Granular Computing, 2005 Publication	<1 %
38	Inbarani, H. Hannah, Ahmad Taher Azar, and G. Jothi. "Supervised Hybrid Feature Selection based on PSO and Rough sets for Medical Diagnosis", Computer Methods and Programs in Biomedicine, 2013. Publication	<1 %
39	Jim Samuel, Ali G. G. Md. Nawaz, Md Mokhlesur Rahman, Ek Esawi, Yana Samuel. "COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification", Cold Spring Harbor Laboratory, 2020 Publication	<1 %
40	Submitted to University of Florida Student Paper	<1 %
41	downloads.hindawi.com Internet Source	<1 %
42	dspace.lboro.ac.uk Internet Source	<1 %
43	ir.uiowa.edu Internet Source	<1 %

44	orion.math.uwaterloo.ca Internet Source	<1 %
45	B. Richhariya, M. Tanveer. "A robust fuzzy least squares twin support vector machine for class imbalance learning", Applied Soft Computing, 2018 Publication	<1 %
46	Muhammad Haqqi Ghufran Rifaldi, Erwin Budi Setiawan. "Competence Classification of Twitter Users Using Support Vector Machine (SVM) Method", 2019 7th International Conference on Information and Communication Technology (ICoICT), 2019 Publication	<1 %
47	Shifei Ding, Huajuan Huang, Junzhao Yu, Han Zhao. "Research on the hybrid models of granular computing and support vector machine", Artificial Intelligence Review, 2013 Publication	<1 %
48	Jilin Yang. "Filtering E-Mail Based on Fuzzy Support Vector Machines and Aggregation Operator", Lecture Notes in Computer Science, 2006 Publication	<1 %
49	ejournal.st3telkom.ac.id Internet Source	<1 %