Optimization of Fuzzy Support Vector Machine (FSVM) Performance by Distance-Based Similarity Measure Classification

Sugiyarto Surono [1*], Tia Nursofiyani [1], Annisa E. Haryati [1]

[1] Department of Mathematics, Ahmad Dahlan Univesity, Yogyakarta, Indonesia.

Abstract

This research aims to determine the maximum or minimum value of a Fuzzy Support Vector Machine (FSVM) Algorithm using the optimization function. SVM is considered as an effective method of data classification, as opposed to FSVM, which is less effective on large and complex data because of its sensitivity to outliers and noise. One of the techniques used to overcome this inefficiency is fuzzy logic with its ability to select the right membership function, which significantly affects the effectiveness of the FSVM algorithm performance. This research was carried out using the Gaussian membership function and the Distance-Based Similarity Measurement consisting of the Euclidean, Manhattan, Chebyshev, and Minkowsky distance methods. Subsequently, the optimization of the FSVM classification process was determined using four proposed FSVM models and normal SVM as comparison references. The results showed that the method tends to eliminate the impact of noise and enhance classification accuracy effectively. FSVM provides the best and highest accuracy value of 94% at a penalty parameter value of 1000 using the Chebyshev distance matrix. Furthermore, the model proposed will be compared to the performance evaluation model in preliminary studies (Xiao Kang et al., 2018). The result further showed that using FSVM with Chebyshev distance matrix and a Gaussian membership function provides a better performance evaluation value.

Keywords: FSVM; Membership Function Fuzzy; Classification; Distance-based Similarity Measure.

1. Introduction

Classification is a grouping method based on the characteristics possessed by objects. The Support Vector Machine (SVM) is one of the classification methods that have been a subject of debate over the last decade due to its high generalization performance and wide application. Research related to SVM performance, such as SVM, ANN, KNN, fuzzy logic, and RF (random forest) methods [1] to classify driving models, indicates that SVM has the highest accuracy value of 96%. Furthermore, in [2] and [3], SVM is proven to have a high level of accuracy and generalization performance from other classification methods. In the real world, this method is applied in many areas such as text categorization [4-6], speech recognition [7], bioinformatics [8-11] , and network security [11].

Vapnik introduced SVM in 1995 [12, 13] based on structural risk minimization theory. It is one of the superior methods trained with an algorithm and used to separate a dataset into two or more classes. Stave Gunn stated that the SVM method is used to determine the optimal global solution and works by mapping the training data into a high-dimensional space while looking for a classification capable of maximizing the margin between the two classes [14].

Margin is the distance between the support vector and the hyperplane, while the support vector is the pattern of each class with the closest distance to the hyperplane.

The SVM method suffers a lot in complex problems with many parameters due to noise and outliers, which causes a decrease in the generalization performance [15]. Therefore, one of the methods used to solve this problem is by combining the SVM method with fuzzy logic [16-18]. Preliminary studies applied fuzzy logic with two events, namely using fuzzy rules [19, 20], and its membership functions [21-26]. In each sample, new input was used to provide a different contribution to eliminate the noise and outliers effect and to improve the generalization performance of SVM classification.

Fuzzy logic is a mathematical way of describing obscurity, and it was introduced by Lotfi A. Zadeh in 1965 [27]. 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 [28]. Several methods such as function approach, intuition, rank-ordering, inductive reasoning, neural networks, and genetic algorithm were used to build a membership function in fuzzy.

Euclidean distance is a general criterion chosen to determine the similarity of the data used to construct the membership function. Xiao-Kang Ding (2018) proposed the FSVM method based on the Euclidean distance using 3 methods, namely FSVM-1, FSVM-2, and FSVM-3, by comparing the distances of positive and negative samples. However, in the FSVM-3 method, point samples are mapped into a high-dimensional space and calculated using the FSVM-2 method. This research indicates that the best accuracy is given by FSVM-3 followed by FSVM-2 and FSVM-1 [29].

The measuring methods commonly used to determine similarity measurements are Euclidean, Manhattan, Chebyshev, Minkowski, Hamming, Mahanalobis, and Minkowski Chebyshev distances. There are various advantages and disadvantages associated with the use of these methods. According to Mohammed and Abdulazeez (2018), euclidean distance is the most commonly used method for calculating distances in numerical data. It works efficiently by calculating the similarity in the grouping and has the ability to separate the data adequately [29]. Manhattan distance is often used due to its ability to detect special circumstances such as the presence of outliers [30]. This is in addition to the sensitivity of the Chebyshev distance in detecting objects with outliers.

Based on the description of several preliminary studies, the combination of SVM and fuzzy logic (FSVM) methods tend to optimize classification by selecting the right membership function. Therefore, this research aims to apply the FSVM with a Gaussian membership function to determine distance measures. Furthermore, comparative research is carried out using several methods such as Euclidean, Manhattan, Chebyshev, Minkowski, Hamming, and Minkowski Chebyshev distances.

2. Materials and Methods

The fuzzy system proposed by Lotfi A. Zadeh was built based on fuzzy set theory and fuzzy logic. This method is useful for dealing with complex real-world problems such as uncertainty and imprecision. Set theory and classification techniques are very useful in dealing with uncertainty and ambiguity, which leads to an increase in the generalizability of the classifier. FSVM is an extension method proposed by Lin and Wang in 2002 to reduce the sensitivity of SVM to outliers or noise. It works by assigning a low weight to each sample by determining a fuzzy membership function based on the similarity of data (distance). The application of fuzzy membership function (s_i) where $0 < s_i \leq 1$ is on the training dataset x_i with class $y_i \in [1, -1]$. Therefore, the fuzzy version dataset is determined as follows: $\left\{ \left(x_1, y_1, s_1\right), \left(x_2, y_2, s_2\right), ..., \left(x_n, y_n, s_n\right) \right\}$. Meanwhile, the optimal FSVM hyperplane is obtained by entering the membership function value in the standard SVM formula.

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

$$y_i \left(w^T x_i + b\right) \geq 1 - \xi_i ; \xi_i \geq 0; \qquad 1 \leq i \leq n \qquad (2)$$

where s_i denotes a fuzzy membership function with a value between 0 and 1 $\left(0 < s_i \leq 1\right)$. In 2016, Xiao-Kang proposed a research to determine the degree of membership of each data by adopting the calculation of the membership function. This research was carried out by comparing the distance from each positive and negative sample to the center of each class using the formula for calculating Euclidean, Manhattan, Chebyshev distance, and the Minkowsky distances. The calculation of the membership function is stated as follows:

$$s_i = \begin{cases} f(d_i^+), if \, \|x_i^+ - x_{center}^+\| \geq \|x_i^+ - x_{center}^-\| \\ 1, if j \, \|x_i^+ - x_{center}^+\| < \|x_i^+ - x_{center}^-\| \\ 1, if \, \|x_i^- - x_{center}^+\| > \|x_i^- - x_{center}^-\| \\ f(d_i^-), if \, \|x_i^- - x_{center}^+\| \leq \|x_i^- - x_{center}^-\| \end{cases} \tag{3}$$

where;

$$d_i \begin{cases} \|x_i^+ - x_{center}^+\| \\ \|x_i^- - x_{center}^+\| \end{cases} \tag{4}$$

When the distance of the positive class is less than the negative, it is considered a "useful point," and its membership is set as 1. However, supposing the distance of the positive class is greater than the negative, it is considered a "noisy point" and calculated according to the Gaussian Membership Function formula.

2.1. Gaussian Membership Function

The Gauss curve Membership Function formula is written as follows:

$$G(x_i; \sigma, c) = e^{\frac{-(x_i - c)}{2\sigma^2}} \tag{5}$$

2.2. Kernel Radial Basic Function (TBF)

The RBF kernel formula is as follows:

$$K(x, y) = exp\left(\frac{\|x - y\|^2}{2\sigma^2}\right) \tag{6}$$

2.3. Distance Based Similarity Measure

The similarity measure is an important part that needs to be considered in pattern matching and to carry out various types of classification. Distance-Based Similarity Measure works to measure the level of similarity of two objects in terms of the geometric distance from the variables included in both objects. These include the following.

2.3.1. Euclidean Distance

Euclidean distance is often used in measuring data similarities, as shown in Equation 7:

$$d_{euclidean} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \qquad i = 1, 2, 3, \dots, n \tag{7}$$

2.3.2. Manhattan Distance (Minkowski Distance)

Manhattan distance is used to calculate the absolute difference between the coordinates of a pair of objects, as shown in Equation 8:

$$d_{Manhattan} = \sum_{i=1}^{n} \|x_i - y_i\|, \qquad i = 1, 2, 3, \dots, n \tag{8}$$

2.3.3. Chebysev Distance

The Chebyshev distance is measured using the following formula:

$$d_{Chebyshev} = \left(max \, |x_i - y_i|\right), \qquad i = 1, 2, 3, \dots, n \tag{9}$$

2.3.4. Minkowski Distance

The Minkowsky distance is a generalization of the Euclidean and the Manhattan, whereby the power (p) acts as the determining parameter. When p equals 1 and 2, the Minkowsky distance space becomes equivalent to Manhattan and Euclidean, respectively. The following formula is used to calculate the Minkowski distance:

$$d_{minkowski} = \left(\sum_{i=1}^{n} \left| x_i - y_i \right|^P \right)^{\frac{1}{p}}, \qquad i = 1,2,3,\dots,n \tag{10}$$

3. Results and Discussion

The Fuzzy Support Vector Machine algorithm proposed in this research is applied to 3 datasets considered as a representative that effectively verifies the proposed FSVM model. Furthermore, the data processing was carried out using Jupyter Notebook Software with Python programming language. The details of the 3 datasets are shown in Table 1.

Table 1. Dataset information

Dataset	Sample variable	Number of samples	Positive sample	Negative sample
Herberman	3	306	225	81
Wine	13	130	59	71
Quality	11	3365	1457	2198

The basic concept of the Fuzzy Support Vector Machine is first used to determine the degree of fuzzy membership of the data used for FSVM calculation, which comprises positive and negative classes. Therefore the center of the class can be defined as the average vector of the attributes determined using Equation 11.

$$
\begin{aligned}
x_{center}^{+} &= \frac{1}{n_+} \sum_{i=1}^{n_+} x_i \\
x_{center}^{-} &= \frac{1}{n_-} \sum_{i=1}^{n_-} x_i
\end{aligned}
\tag{11}
$$

where x_{center}^{+} and x_{center}^{-} denote the means of the positive and negative classes, while n_+ and n_- are the number of data points in the positive and negative classes, respectively.

The calculation of the distance matrix from the data points of each class to the center uses Equations 7 to 10. The results are used to determine the value of the degree of fuzzy membership based on Equation 3 as follows:

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

This research uses a different penalty parameter (C), including the values of $C = 2, C = 10, C = 50, C = 200, C = 500, C = 1000$, and the RBF kernel. In 2013, Bekker et al. proposed the AUC approximation method in the binary case, as shown in the Equation 12.

$$Accuracy = \frac{1}{2}(SE + SP) \tag{12}$$

where SE (sensitivity) and SP (specivity) are the ratios of the completeness or accuracy of the correct prediction of positive and negative data, respectively. The results of the FSVM classification accuracy based on the AUC (Area Under Curve) approach are shown in Figure 1.

Figure 2 shows that when the penalty parameter is 2 (C=2) while using Manhattan and Chebyshev distances, the best accuracy generator is FSVM. Furthermore, at parameters C=50 and 500, FSVM using the Manhattan penalty distance produces the best accuracy from other methods. The best accuracy result for the penalty parameter C=10 is obtained by the FSVM method using Euclidean distance. Meanwhile, the highest accuracy in the penalty parameters C=200 and C=1000 is generated by FSVM with Chebyshev distance. The results of the accuracy of each method are shown in Table 2.
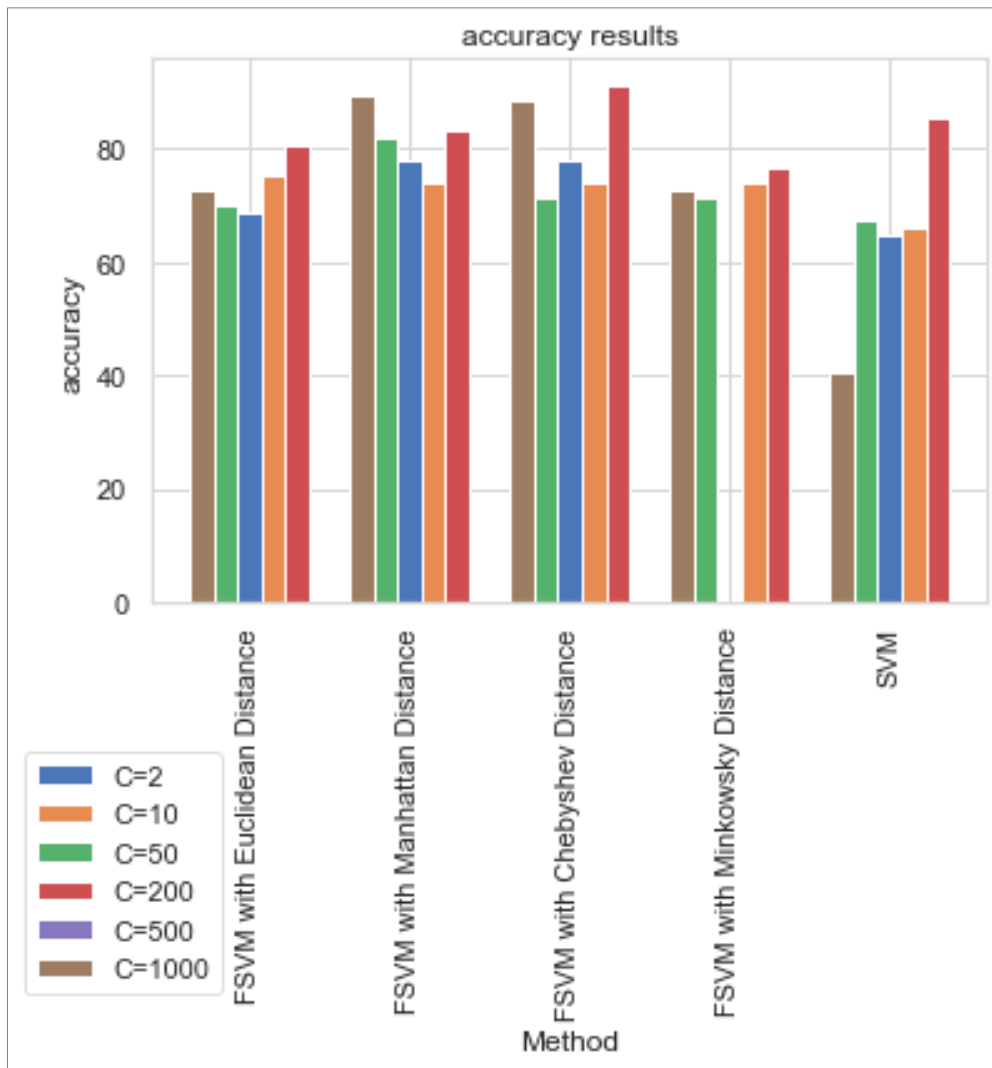
Figure 1. FSVM and SVM accuracy results

Table 2. Classification accuracy results

	FSVM 1	FSVM 2	FSVM 3	FSVM 4	SVM
$C = 2$	68.83%	77.92%	77.92%	0.0%	64.93%
$C = 10$	75.32%	74.02%	74.03%	74.03%	66.23%
$C = 50$	70.13%	81.81%	71.43%	71.43%	67.53%
$C = 200$	80.52%	83.12%	91.30%	76.62%	85.36%
$C = 500$	72.72%	89.58%	88.37%	72.72%	40.32%
$C = 1000$	74.03%	79.22%	94%	71.42%	80.52%

With a value of C=1000, the accuracy is determined with different training and testing data scenarios as follows:

Table 3. Evaluation of Model Performance with Different Training and Testing Data Partitions

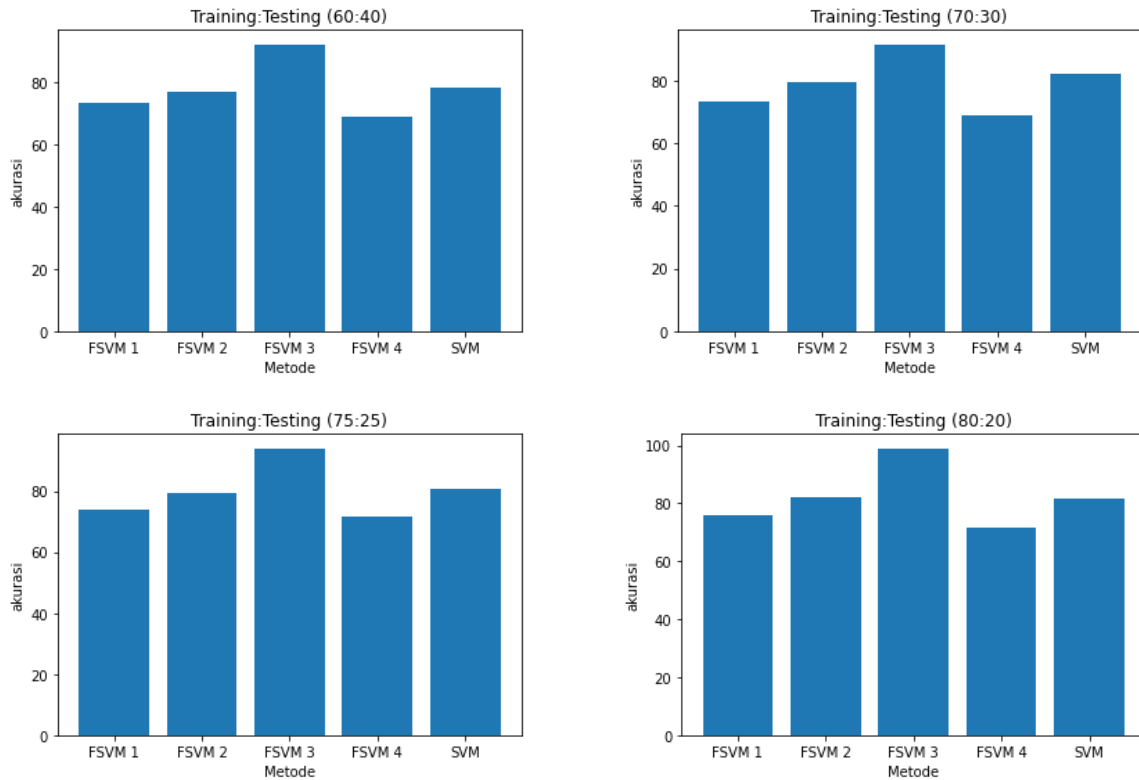| Training data and testing data	FSVM 1	FSVM 2	FSVM 3	FSVM 4	SVM
60:40	73.45%	77.23%	92.34%	69.07%	78.2%
70:30	73.33%	79.58%	91.62%	68.91%	82.33%
75:25	74.03%	79.22%	94%	71.42%	80.52%
80:20	76.09%	82.15%	99%	71.48%	81.67%

Figure 2. Graph of Model Performance Evaluation with Different Partitions of Training and Testing Data

Table 3 and Figure 2 show that the best performance of FSVM using Chebyshev distance and RBF kernel with C=1000 is obtained from the proposed methods. Furthermore, the results of the G-means (GM), SE, SP from FSVM are compared with the FSVM proposed by Xiao Kang Ding et al. (2018) as shown in Equation 3 using the membership function linearly ascending and exponential FSVM with the Euclidean function. The results of the performance evaluation are shown in Table 3.

Table 4. Model performance evaluation

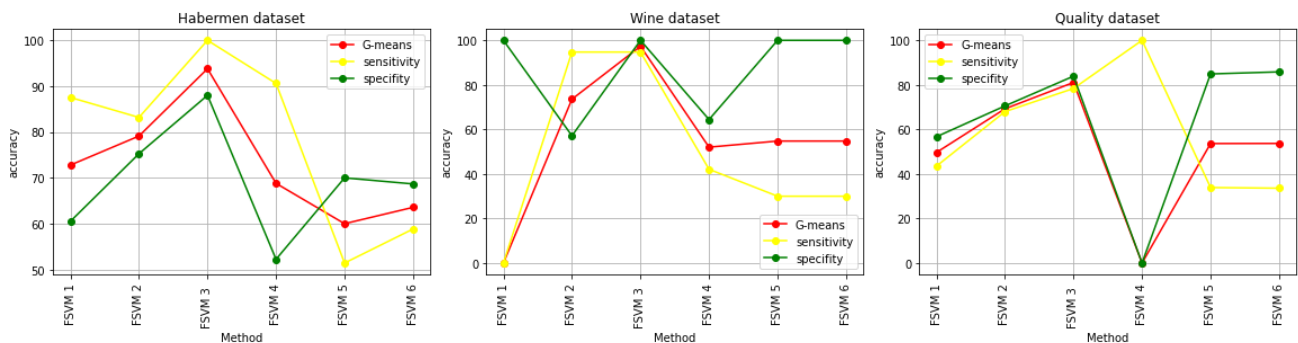| Dataset	Performance evaluation	FSVM 1	FSVM 2	FSVM 3	FSVM 4	FSVM 5	FSVM 6
Habermen	GM	72.79%	79.12%	93.80%	68.8%	60%	63.57%
	SE	87.54%	83.18%	100%	90.6%	51.42%	58.85%
	SP	60.52%	75.26%	88%	52.24%	70%	68.67%
Wine	GM	0%	73.57%	97.33%	52.03%	54.77%	54.77%
	SE	0%	94.73%	94.73%	42.1%	30%	30%
	SP	100%	57.14%	100%	64.43%	100%	100%
Quality	GM	49.633	69.25%	81.08%	0%	53.65%	53.7%
	SE	43.5%	67.98%	78.37	100%	33.9%	33.6%
	SP	56.70%	70.54%	83.9%	0%	84.91%	85.83%

Figure 3. Model Performance Evaluation Graph for Wine and Quality Datasets

Description:

FSVM 1: FSVM with Euclidean distance matrix and Gaussian membership function

FSVM 2: FSVM with Manhattan distance matrix and Gaussian membership function

FSVM 3: FSVM with Chebyshev distance matrix and Gaussian membership function

FSVM 4: FSVM with Minkowsky distance matrix and Gaussian membership function

FSVM 5: FSVM proposed by Xiao Kang Ding et al. (2018) with an increased Linear membership function

FSVM 6: FSVM proposed by Xiao Kang Ding et al. (2018) with exponential membership function

Table 4 and Figure 3 show that FSVM 2 and FSVM 3 provide better classification performance evaluations than previous studies. Meanwhile, FSVM 5 and FSVM 6 provided better classification performance evaluations than FSVM 1. Out of all the models discussed in this research, FSVM 3 provides the best classification evaluation.

4. Conclusion

This research introduced a fuzzy membership function based on Distance-Based Similarity Measure with the Euclidean, Manhattan, Chebyshev, and Minkowsky distance methods to determine the best method capable of optimizing the Fuzzy Support Vector Machine classification process. This application is a new method because there are no previous studies on the analysis of FSVM based on Distance-Based Similarity Measures. Therefore, based on the results and discussion of this study, it is concluded that the FSVM using the Chebyshev distance and the Gaussian membership function has the best performance in reducing the effects of noise and outliers.

5. Declarations

5.1. Author Contributions

Conceptualization, S.S. and T.N.; methodology, S.S.; software, S.S.; validation, S.S., T.N. and A.E.H.; formal analysis, S.S.; investigation, S.S.; resources, S.S.; data curation, S.S.; writing—original draft preparation, S.S.; writing—review and editing, S.S.; visualization, S.S.; supervision, S.S.; project administration, S.S.; funding acquisition, T.N. All authors have read and agreed to the published version of the manuscript.

5.2. Data Availability Statement

The data presented in this study are available in article.

5.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

5.4. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

6. References

[1] Silva, I., & Naranjo, J. E. (2020). A systematic methodology to evaluate prediction models for driving style classification. Sensors (Switzerland), 20(6), 1–21. doi:10.3390/s20061692.

[2] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Computer Science, 83(Fams), 1064–1069. doi:10.1016/j.procs.2016.04.224.

[3] Fauzi, F. (2017). K-Nearset Neighbor (K-NN) dan Support Vector Machine (SVM) untuk Klasifikasi Indeks Pembangunan Manusia Provinsi Jawa Tengah. Jurnal Mipa, 40(2), 118–124.

[4] Han, K. X., Chien, W., Chiu, C. C., & Cheng, Y. T. (2020). Application of support vector machine (SVM) in the sentiment analysis of twitter dataset. Applied Sciences (Switzerland), 10(3). doi:10.3390/app10031125.

[5] Lin, Y., Yu, H., Wan, F., & Xu, T. (2017). Research on Classification of Chinese Text Data Based on SVM. IOP Conference Series: Materials Science and Engineering, 231(1), 0–5. doi:10.1088/1757-899X/231/1/012067.

[6] Widiastuti, N. I., Rainarli, E., & Dewi, K. E. (2017). Peringkasan dan Support Vector Machine pada Klasifikasi Dokumen. Jurnal Infotel, 9(4), 416. doi:10.20895/infotel.v9i4.312.

[7] Pavithra, S., & Janakiraman, S. (2020). Enhanced polynomial kernel (EPK)–based support vector machine (SVM) (EPK-SVM) classification technique for speech recognition in hearing-impaired listeners. Concurrency and Computation: Practice and Experience, 1–12,. doi:10.1002/cpe.5210.

[8] Orlando, G., Raimondi, D., Khan, T., Lenaerts, T., & Vranken, W. F. (2017). SVM-dependent pairwise HMM: An application to protein pairwise alignments. Bioinformatics, 33(24), 3902–3908. doi:10.1093/bioinformatics/btx391.

[9] Sohail, A., & Arif, F. (2020). Supervised and unsupervised algorithms for bioinformatics and data science. Progress in Biophysics and Molecular Biology, 151, 14–22. doi:10.1016/j.pbiomolbio.2019.11.012.

[10] Lin, X., Li, C., Zhang, Y., Su, B., Fan, M., & Wei, H. (2018). Selecting feature subsets based on SVM-RFE and the overlapping ratio with applications in bioinformatics. Molecules, 23(1). doi:10.3390/molecules23010052.

[11] Akbani, R., & Korkmaz, T. (2010). Applications of Support Vector Machines in Bioinformatics and Network Security. Application of Machine Learning, February. doi:10.5772/8618.

[12] Vanpik, V., & Cortes, C. (1995). Support Vector Network. Mach. Learn, 20, 273–297.

[13] Pushpita Anna Octaviani, Yuciana Wilandari, D. I. (2014). Penerapan Metode SVM Pada Data Akreditasi Sekolah Dasar Di Kabupaten Magelang. Jurnal Gaussian, 3(8), 811–820.

[14] Brereton, R. G., & Lloyd, G. R. (2010). Support Vector Machines for classification and regression. Analyst, 135(2), 230–267. doi:10.1039/b918972f.

[15] Wu, Q., & Wang, W. (2013). Piecewise-smooth support vector machine for classification. Mathematical Problems in Engineering, 2013. doi:10.1155/2013/135149.

[16] Lin, C. F., & Wang, S. De. (2002). Fuzzy support vector machines. IEEE Transactions on Neural Networks, 13(2), 464–471. doi:10.1109/72.991432.

[17] An, W., & Liang, M. (2013). Fuzzy support vector machine based on within-class scatter for classification problems with outliers or noises. Neurocomputing, 110, 101–110. doi:10.1016/j.neucom.2012.11.023.

[18] Guernine, T., & Zeroual, K. (2011). New fuzzy multi-class method to train SVM classifier. 3rd International Conference on Advances in Databases, Knowledge, and Data Applications, 77–82.

[19] Castro, J. L., Flores-Hidalgo, L. D., Mantas, C. J., & Puche, J. M. (2007). Extraction of fuzzy rules from support vector machines. Fuzzy Sets and Systems, 158(18), 2057–2077. doi:10.1016/j.fss.2007.04.014.

[20] Wu, K., Zhou, M., Sean Lu, X., & Huang, L. (2017). Fuzzy logic-based text classification method for social media data. 2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2017, 2017-January, 1942–1947. doi:10.1109/SMC.2017.8122902.

[21] Sridevi, P. (2019). Identification of suitable membership and kernel function for FCM based FSVM classifier model. Cluster Computing, 22, 11965–11974. doi:10.1007/s10586-017-1533-9.

[22] Gupta, D., Richhariya, B., & Borah, P. (2019). A fuzzy twin support vector machine based on information entropy for class imbalance learning. Neural Computing and Applications, 31(11), 7153–7164. doi:10.1007/s00521-018-3551-9.

[23] Ding, X. K., Yang, X. J., Jiang, J. Y., Deng, X. L., Cai, J. C., & Ji, Y. Y. (2018). Optimization and analysis on fuzzy SVM for objects classification. Journal of Information Hiding and Multimedia Signal Processing, 9(6), 1421–1429.

[24] Wang, R., Zhang, X., & Cao, W. (2016). Clifford Fuzzy Support Vector Machines for Classification. Advances in Applied Clifford Algebras, 26(2), 825–846. doi:10.1007/s00006-015-0616-z.

[25] Xiaokang, D., Lei, Y., Jianping, Y., & Zhaozhong, Z. (2016). Optimization and Analysis on Fuzzy SVM for Targets Classification in Forest. The Open Cybernetics & Systemics Journal, 10(1), 155–162. doi:10.2174/1874110x01610010155.

[26] Jiang, X., Yi, Z., & Lv, J. C. (2006). Fuzzy SVM with a new fuzzy membership function. Neural Computing and Applications, 15(3–4), 268–276. doi:10.1007/s00521-006-0028-z.

[27] Al-Mumtazah, N. S., & Surono, S. (2020). Quadratic Form Optimization with Fuzzy Number Parameters: Multiobjective Approaches. International Journal of Fuzzy Systems, 22(4), 1191–1197. doi:10.1007/s40815-020-00808-x.

[28] Liu, W., Ci, L. L., & Liu, L. P. (2020). A new method of fuzzy support vector machine algorithm for intrusion detection. Applied Sciences (Switzerland), 10(3). doi:10.3390/app10031065.

[29] Mohammed, N. N., & Abdulazeez, A. M. (2018). Evaluation of partitioning around medoids algorithm with various distances on microarray data. Proceedings - 2017 IEEE International Conference on Internet of Things, IEEE Green Computing and Communications, IEEE Cyber, Physical and Social Computing, IEEE Smart Data, IThings-GreenCom-CPSCom-SmartData 2017, 2018-January, 1011–1016. doi:10.1109/iThings-GreenCom-CPSCom-SmartData.2017.155.

[30] Syaripudin, U., Badruzaman, I., Yani, E., K, D., & Ramdhani, M. (2013). Studi Komparatif Penerapan Metode Hierarchical, K-Means Dan Self Organizing Maps (SOM) Clustering Pada Basis Data. Istek, VII(1), 132–149.