Twin support vector machine using kernel function for colorectal cancer detection

Zuherman Rustam, Fildzah Zhafarina, Jane Eva Aurelia, Yasirly Amalia
Department of Mathematics, University of Indonesia, Indonesia

Article Info

Article history:

Received Jul 22, 2020
Revised Jun 14, 2021
Accepted Oct 12, 2021

Keywords:

Colorectal cancer
Detection
Kernel
Machine learning
Twin support vector machine

ABSTRACT

Nowadays, machine learning technology is needed in the medical field. therefore, this research is useful for solving problems in the medical field by using machine learning. Many cases of colorectal cancer are diagnosed late. When colorectal cancer is detected, the cancer is usually well developed. Machine learning is an approach that is part of artificial intelligence and can detect colorectal cancer early. This study discusses colorectal cancer detection using twin support vector machine (SVM) method and kernel function i.e. linear kernels, polynomial kernels, RBF kernels, and gaussian kernels. By comparing the accuracy and running time, then we will know which method is better in classifying the colorectal cancer dataset that we get from Al-Islam Hospital, Bandung, Indonesia. The results showed that polynomial kernels has better accuracy and running time. It can be seen with a maximum accuracy of twin SVM using polynomial kernels 86% and 0.502 seconds running time.

Corresponding Author:

Zuherman Rustam
Department of Mathematics
University of Indonesia
Jl. Prof DR. Sudjono D. Pusponegoro, Pondok Cina, Depok, Jawa Barat 16424, Indonesia
Email: rustam@ui.ac.id

1. INTRODUCTION

One of the diseases that cause death in the world is cancer. Cancer is the second leading cause of death globally [1]. Detecting these diseases when still at an early stage is associated with markedly improved survival prospects [2], [3]. Early-stage of the cancer is more likely to treat [4]. Colorectal cancer is cancer with the third death rate. responsible for around 600,000 per year worldwide [5]-[8]. Information technology has an important role in the field of medicine. Cancer is a disease that can be detected by machine learning. Data is very useful in the medical field. It can be seen from the development of data mining in medical science is increasing rapidly. This increase can be seen from the high prediction results, can reduce treatment costs, increase the chances of recovery of patients, and decisions to save lives [9], [10].

Machine learning is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed [11]. One method that is popular because the learning performance is very good is the twin support vector machine (SVM) [12]. Kernel method is a method that uses functions when the algorithm operates in feature space with a higher dimension. This process uses product operations between images, all feature pairs. This method is used directly or indirectly by a SVM and twin SVM to classify data [13]. The kernel functions commonly used for SVM methods are linear kernel, polynomial kernel, RBF kernel, and gaussian kernel. This paper proposes the twin SVM method as a novel approach for the early detection of colorectal cancer. The kernel functions used are the linear kernel, polynomial kernel, RBF kernel, and gaussian kernel. This paper compares the performance of the twin SVM with each kernel to get the best kernel for the detection of colorectal cancer.

2. RESEARCH METHOD

2.1. Twin support vector machine

SVM is a method used to find a single hyperplane to classify samples [14] proposed twin SVM is found where samples are given to classes with two hyperplanes according to their distance from their hyperplanes. Equations of the two hyperplanes are as:

$$w_1^T x_s + b_1 = 0$$

$$w_2^T x_s + b_1 = 0$$

i-th hyperline parameters shown by w_i and b_i. Each hyperline is closest to its class sample, non-parallel in nature, and farthest from the opposite class sample. Assume a binary classification task with classes +1 and −1, and A ∈ $\mathbb{R}^{n_1 xd}$ and B ∈ $\mathbb{R}^{n_2 xd}$ indicate each matrix has a sample with each class +1 and -1 [15]. Based on the appropriate class, one sample is shown with each matrix row. The two hyperplanes of twin SVM obtained from (1) and (2):

$$\min \frac{1}{2}(\boldsymbol{A}\boldsymbol{w_1} + \boldsymbol{e}b_1)^T(\boldsymbol{A}\boldsymbol{w_1} + \boldsymbol{e}b_1) + p_1 e^T \xi$$

$$s.t\ -(\boldsymbol{B}\boldsymbol{w_1} + \boldsymbol{e}b_1) + \xi \geq e, \xi \geq 0 \tag{1}$$

$$\min \frac{1}{2}(\boldsymbol{B}\boldsymbol{w_2} + \boldsymbol{e}b_2)^T(\boldsymbol{B}\boldsymbol{w_2} + \boldsymbol{e}b_2) + p_2 e^T \xi$$

$$s.t - (\boldsymbol{A}\boldsymbol{w_2} + \boldsymbol{e}b_2) + \xi \geq e, \xi \geq 0 \tag{2}$$

ξ is a non-negative vector component, therefore ξ ≥ 0. Vector of the size slack variable n represented by e. letting the margin of decision make a few mistakes is the standard approach. a standard approach is taken if the sampling service cannot be separated linearly. (for example, some points are in or on the wrong margin). the cost for a wrong-classified sample that is proportional to the distance between the sample and the decision margin is determined by each zero-zero element of the slack variable vector. Based on these equations, ρ_1 and ρ_2 are penalty parameters. Twin SVM is in great demand in various fields with various versions of the proposed algorithm [16]. Recently, several fuzzy formulations from twin SVM have also been proposed [17]

2.2. Kernel function

Kernel method is a method that uses kernel functions to operate algorithms in feature spaces that have higher dimensions. This method uses product operations between images of all image pairs in the feature space [18]. Accuracy for classifying objects in the right cluster is difficult to obtain in high dimensional data sets, measuring euclidean distances on k-means, c-means, or fuzzy c-medoids. Distribution data can be represented to validate the truly central cluster. This difficulty can be overcome by using the kernel method [19]. Let X^n be an input space; F is a feature space and ϕ : X_n →F. In (3) defines kernel functions [20], [21]:

$$K(x_1, x_2) = \varphi(x_1)\varphi(x_2) \tag{3}$$

where $x_1, x_2 \in X^n$.

Kernel functions that are often used are linear kernel, polynomial kernel, RBF kernel, and gaussian kernel. Table 1 lists the formulas for kernel functions [22]-[23]:

Table 1. The formula of kernel function

Kernel Function	Formula
Linear Kernel	$K(x_1, x_2) = x_1^T x_2 + c$
Polynomial Kernel	$K(x_1, x_2) = (\gamma x_1^T x_2 + c)^d; \gamma > 0$
RBF Kernel	$K(x_1, x_2) = e^{-\gamma \lvert\lvert x_1 - x_2 \rvert\rvert^2}; \gamma > 0$
Gaussian Kernel	$K(x_1, x_2) = e^{-\frac{\lvert\lvert x_1 - x_2 \rvert\rvert^2}{2\sigma^2}}$

2.3. k-Fold cross validation

The dataset is divided into two, i.e training data and testing data. This is done so that the resulting model can be evaluated and obtained. Colorectal cancer data patterns are studied and recognized by machines with training data. Testing data are data used to evaluate models obtained after a machine learns data patterns [24]. By using the k-fold cross validation method, the dataset is divided into training data and testing data [25]. Training data samples were selected by the k-fold cross validation method. This method works by dividing the dataset with k-parts of the same size. Models and repetition of processes k times tested for each subsample taken as validation data.

2.4. Proposed method

Several stages are proposed in this study, including data divided into training and testing data. then the data is tested with k-fold cross validation. The k-value chosen was 10 and 45 for the random state. This means that the dataset was divided into 10 samples of the same size. In the second stage, the training data were used by the twin SVM method based on linear kernel, polynomial kernel, RBF kernel, and gaussian kernel to study data patterns and build classification models. The next step is to classify the models obtained and evaluated based on the parameters of accuracy and running time. To find the best kernel, the evaluation parameters produced by each kernel are compared.

3. RESULTS AND ANALYSIS

This research using Jupyter Notebook as software for running the program of twin SVM using linear kernel, polynomial kernel, RBF kernel, and gaussian kernel. The stages carried out in this paper using the Python 3 programming language.

3.1. Data

In this study, the data consisted of 210 samples and seven features. these seven features consist of CEA, hemoglobin, leukocytes, hematocrit, platelets, age. diagnosis features become a target feature in detecting colorectal cancer. The data are colorectal cancer data obtained from Al-Islam Hospital, Bandung, Indonesia with cancer diagnoses (1), and no cancer (0). Table 2 represented part of the data:

Table 2. Part of colorectal cancer data

Age	CEA	Hemoglobin	Leukocyte	Hematocrit	Platelets	Diagnosis
74	3.26	11.8	19400	37.3	341000	0
84	29.12	8	12400	26.6	465000	1
81	4.5	8.8	19900	26.2	468000	0
56	0.96	13.9	9400	41.5	260000	0
75	3.24	7.7	13500	22.5	377000	0
58	0.71	11	18200	34	259000	0
63	1.65	10.1	19900	32.1	151000	0
73	36.49	11.1	9700	33.4	267000	1

3.2. Confusion matrix

In this paper, a confusion matrix was used to assist in calculating the evaluation parameters of the classification model. Table 2 shows the confusion matrix used to evaluate the twin SVM classification model based on the kernel for the diagnosis of colorectal cancer. Table 3 shown confusion matrix.

Table 3. Confusion Matrix

		Predict	
		Cancer (Y)	Non-Cancer (N)
Actual	Cancer (Y)	TP	FN
	Non Cancer (N)	FP	TN

Explanation:
TP (true positive): many cases of colorectal cancer are predicted to be correct
TN (true negative): many cases of not colorectal cancer are predicted to be correct
FP (false positive): many cases of not colorectal cancer are predicted to be wrong (predicted as colorectal cancer)
FN (false negative): many colorectal cancer cases are predicted to be wrong (predicted as not pancreatic cancer)

3.3. Evaluation parameters

The parameters to evaluate the performance of the twin SVM classification model were accuracy and required running time. In 4 shows the formula for accuracy:

$$Accuracy = \frac{(TN+TP)}{(FN+TP+FP+TN)} x\ 100\% \qquad (4)$$

Accuracy is used to compare the number of cases of colorectal cancer and not colorectal cancer that identified correctly with the total number of cases.

3.4. Results

In this section, we discuss the performance evaluation of the twin SVM classification model with linear kernel, polynomial kernel, RBF kernel, and gaussian kernel. The twin SVM classification model based on kernel detects colorectal cancer using a twin SVM with a linear kernel, polynomial kernel, RBF kernel, and gaussian kernel. In this research, the highest accuracy is from the polynomial kernel. This indicates that the polynomial kernel is the appropriate kernel in detecting colorectal using a twin support vector machine. In this paper, we have built the twin SVM classification model with linear kernels, polynomial kernels, radial basis function kernels, and gaussian kernels in detecting colorectal cancer. Table 4 presents a comparison of twin SVM performance linear kernel, polynomial kernel, RBF kernel, and gaussian kernel. All kernel parameter is 1. The performance evaluation parameters compared are accuracy and running time. Table 4 shows the result of the accuracy and running time twin SVM classification model based on kernel.

Table 4. Results of the twin SVM classification model based on kernel

Classification Model	Accuracy (%)	Running Time (seconds)
Linear Kernel	81%	0.565
Polynomial Kernel	86%	0.502
RBF Kernel	76%	1.605
Gaussian Kernel	76%	1.612

Based on Tabel 4, that can be seen that for accuracy, twin SVM models the highest accuracy of 86% was recorded when using the polynomial kernel at 0.502 seconds. While the lowest accuracy at 76% was recorded when RBF and Gaussian kernel with a running time of 1.605 seconds for RBF kernel and 1.612 for the gaussian kernel. For consideration of running time, the twin SVM model with polynomial kernel has the fastest running time compared to linear, RBF, and gaussian kernels, which is around 0.502 s. The twin SVM model with the gaussian kernel actually produces the longest running time which is around 1.612 s. Based on the results obtained, the polynomial kernel gets the best results in terms of accuracy and running time. Thus, the polynomial kernel is the best kernel for the twin SVM in detecting colorectal cancer dataset.

4. CONCLUSION

Colorectal cancer detection quickly is very important. it is useful for handling cancer quickly before being infected to all organs of the body. However, this is difficult because colorectal cancer has no specific symptoms. The twin SVM method can help detect colorectal cancer based on blood tests and age. The most appropriate kernel for the twin SVM method in detecting colorectal cancer is the polynomial kernel which produces an accuracy of 86% and the required running time is 0.502 seconds.

ACKNOWLEDGEMENTS

Zuherman Rustam supported financially by University of Indonesia with a Hibah FMIPA 2021 grant schema. Fildzah Zhafarina, Jane Eva Aurella, Yasirly Amalia supported financially by PUTI 2020 grant schema.

REFERENCES

[1] T. Nadira and Z. Rustam, “Classification of cancer data using support vector machines with features selection method based on global artificial bee colony,” *AIP Conference Proceedings*, vol. 2023, no. 1, pp. 1-8, 2018, doi: 10.1063/1.5064202.

[2] N. Bannister and J. Broggio, “Cancer survival by stage at diagnosis for England (experimental statistics): adults diagnosed 2012, 2013 and 2014 and followed up to 2015,” *Produced in collaboration with Public Health England*, 2015.

[3] P. Muller, S. Walters, M. P. Coleman and L. Woods, “Which indicators of early cancer diagnosis from population-based data sources are associated with short-term mortality and survival?,” *Cancer epidemiology*, vol. 56, pp. 161-170, 2018, doi: 10.1016/j.canep.2018.07.010.

[4] R. Siegel, C. Desantis and A. Jemal, “Colorectal cancer statistics, 2014,” *CA: a cancer journal for clinicians*, vol. 64, no. 2, pp. 104-117, 2014, doi: 10.3322/caac.21220.

[5] W. C. Shangkuan, “Risk analysis of colorectal cancer incidence by gene expression analysis,” *PeerJ*, 5, e3003, 2017, doi: 10.7717/peerj.3003.

[6] M. S. Kim, D. Kim and J. -R. Kim, “Stage-Dependent Gene Expression Profiling in Colorectal Cancer,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 5, pp. 1685-1692, 2019, doi: 10.1109/TCBB.2018.2814043.

[7] A. Calon, *et al.*, “Stromal gene expression defines poor-prognosis subtypes in colorectal cancer,” *Nature genetics*, vol. 47, no. 4, pp. 320-329, 2015, doi: https://doi.org/10.1038/ng.3225.

[8] P. F. Simmonds, *et al*., “Surgery for colorectal cancer in elderly patients: a systematic review,” *The Lancet*, vol. 356, no. 9234, pp. 968-974, 2000, doi: 10.1016/s0140-6736(00)02707-0.

[9] H. Asri, H. Mousannif, H. A. Moatassime and T. Noel, “Using machine learning algorithms for breast cancer risk prediction and diagnosis,” *Procedia Computer Science*, vol. 83, pp. 1064-1069, 2016, https://doi.org/10.1016/j.procs.2016.04.224.

[10] Z. Rustam, V. A. W. Hapsari and M. R. Solihin, “Optimal cervical cancer classification using Gauss-Newton representation based algorithm,” *AIP Conference Proceedings*, vol. 2168, no. 1, pp. 020045 1-6, 2019, doi: 10.1063/1.5132472.

[11] Z. Rustam and N. P. A, Ariantari, “Support Vector Machines for Classifying Policyholders Satisfactorily in Automobile Insurance,” *Journal of Physics: Conference Series*, 2018, vol. 1028, no. 1, pp. 1-9, doi :10.1088/1742-6596/1028/1/012005.

[12] H. Huajuan, W. Xiuxi and Z. Yongquan, “Twin support vector machines: A survey,” *Neurocomputing*, vol. 300, pp. 34-43, 2018, doi: 10.1016/j.neucom.2018.01.093.

[13] H. J. S. Taylor, J. S. Taylor and N. Cristiani, “Kernel methods for pattern analysis,” *Cambridge university press*, 2004.

[14] Jayadeva, R. Khemchandani and S. Chandra, “Twin Support Vector Machines for Pattern Classification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 905-910, 2007, doi: 10.1109/TPAMI.2007.1068.

[15] M. Tzelepi and A. Tefas, “Improving the performance of lightweight CNNs for binary classification using Quadratic Mutual Information regularization,” *Pattern Recognition*, vol. 106, no. 107407, 2020, doi: 10.1016/j.patcog.2020.107407.

[16] S. Ding, Y. An, X. Zhang, F. Wu and Y. Xue, “Wavelet twin support vector machines based on glowworm swarm optimization,” *Neurocomputing*, vol. 225, pp. 157-163, 2017, doi: 10.1016/j.neucom.2016.11.026.

[17] D. Gupta, B. Richhariya and P. Borah, “A fuzzy twin support vector machine based on information entropy for class imbalance learning,” *Neural Computing and Applications*, vol. 31, no. 11, pp. 7153-7164, 2019, doi: 10.1007/s00521-018-3551-9.

[18] W. Sadewo, Z. Rustam, H. Hamidah and A. R. Chusmarsyah, “Pancreatic Cancer Early Detection Using Twin Support Vector Machine Based on Kernel,” *Symmetry*, vol. 12, no. 667, pp. 1-8, 2020, doi: 10.3390/sym12040667.

[19] Z. Rustam and A. S. Talita, “Fuzzy kernel K-medoids algorithm for anomaly detection problems,” *AIP Conference Proceedings*, vol. 1862, no. 030154, pp. 1-8, 2016, doi: 10.1063/1.4991258.

[20] Z. Rustam and R. Faradina, “Face recognition to identify look-alike faces using support vector machine,” *Journal of Physics: Conference Series*, vol. 1108, no. 012071, pp. 1-7, 2018, doi: 10.1088/1742-6596/1108/1/012071.

[21] C. Bishop, “Pattern recognition and machine learning,” *Springer*, 2006.

[22] A. Zheng, “Evaluating Machine Learning Models: A Beginner’s Guide to Key Concepts and Pitfalls,” Sebastopol, CA: O’Reilly Media, Inc, 2015.

[23] Arfiani, Z. Rustam, J. Pandelaki and A. Siahan, “Kernel spherical k-means and support vector machine for acute sinusitis classification,” *IOP Conference Series: Materials Science and Engineering*, vol. 546, no. 052011, pp. 1-10, 2019, doi: 10.1088/1757-899X/546/5/052011.

[24] H. Glanz, L. Calvanho, D. S. Menashe and M. A. Frield, “A parametric model for classifying land cover and evaluating training data based on multi-temporal remote sensing data,” *ISPRS journal of photogrammetry and remote sensing*, vol. 97, pp. 219-228, 2014, doi: 10.1016/j.isprsjprs.2014.09.004.

[25] S. Saud, B. Jamil, Y. Upadhyay and K. Irshad, “Performance improvement of empirical models for estimation of global solar radiation in India: A k-fold cross-validation approach,” *Sustainable Energy Technologies and Assessments*, vol. 40, no. 100768, 2020, doi: 10.1016/j.seta.2020.100768.

BIOGRAPHIES OF AUTHORS

Zuherman Rustam is an Associate Professor and a lecturer of the intelligence computationat the Department of Mathematics, University of Indonesia. He obtained his Master of Science in 1989 in informatics, Paris Diderot University, French, and completed his Ph.D. in 2006 fromcomputer science, University of Indonesia. Assoc. Prof. Dr. Rustam is a member of IEEE who is actively researching machine learning, pattern recognition, neural network, artificial intelligence.

Fildzah Zhafarina is a final year student from Departement of Mathematics, University of Indonesia. Ms. Zhafarina is passionately researching machine learning, computer vision, neural networks and deep learning in various fields.

Jane Eva Aurelia was born in Jakarta, 19 June 1998. She is a final year student in the Departement of Mathematics, University of Indonesia. She is currently working on her thesis, which is firmly about applied mathematics using machine learning. Also, Ms. Jane's specialties in research are mostly about machine learning, mathematical modeling, and data mining.

Yasirly Amalia is a final semester student in Department of Mathematics, University of Indonesia. Ms. Yasirly is passionately in machine learning, mathematical modelling, and data mining.