IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 10, No. 1, March 2021, pp. 253~256
ISSN: 2252-8938, DOI: 10.11591/ijai.v10.i1.pp253-256

Linear discriminant analysis and support vector machines for classifying breast cancer

Zuherman Rustam, Yasirly Amalia, Sri Hartini, Glori Stephani Saragih
Department of Mathematics, University of Indonesia, Indonesia

Article Info

Article history:

Received Feb 22, 2020
Revised Dec 17, 2020
Accepted Feb 26, 2021

Keywords:

Breast cancer
Classification
Linear discriminant analysis
Machine learning
Support vector machines

ABSTRACT

Breast cancer is an abnormal cell growth in the breast that keeps changed uncontrolled and it forms a tumor. The tumor can be benign or malignant. Benign could not be dangerous to health and cancerous, but malignant could be has a probability dangerous to health and be cancerous. A specialist doctor will diagnose the patient and give treatment based on the diagnosis which is benign or malignant. Machine learning offer times efficiency to determine a cancer cell. The machine will learn the pattern based on the information from the dataset. Support vector machines and linear discriminant analysis are common methods that can be used in the classification of cancer. In this study, both of linear discriminant analysis and support vector machines are compared by looking from accuracy, sensitivity, specificity, and F1-score. We will know which methods are better in classifying breast cancer dataset. The result shows that the support vector machine has better performance than the linear discriminant analysis. It can be seen from the accuracy is 98.77%.

Corresponding Author:

Zuherman Rustam
Department of Mathematics
University of Indonesia
Depok 16424, Indonesia
Email: rustam@ui.ac.id

1. INTRODUCTION

Breast cancer is one of the common types of cancer causes death. About half a million people especially women die every year around the world. According to WHO in 2018, it is approximately 15% of all cancer deaths among women [1]. Cancer cell will grow and spread on breast tissue such as in the duct that brings milk to nipple, some in lobular which that makes breast milk, and some in other support tissue in breast [2]. Treatment will be different in every patient according to the status of the classes. Specialist doctors would determine the classes of the cancer. Detecting breast cancer often using machine learning techniques. Machine learning techniques provide time efficiency and a more accurate diagnosis to help doctors diagnose patients. There are some machine learning methods that used in classification of breast cancer such as normed kernel function-based fuzzy possibilistic C-means algorithm [3], sparse learning based fuzzy c-means [4], deep learning approach [5], combination of K-means, fuzzy C-means algorithm, and kernel function [6], convolutional neural network [7], using hybrid deep neural network [8], using SVM and hough transform [9]. In this research, a breast cancer dataset is performed by using linear discriminant analysis and support vector machines. Both methods have good performance for disease diagnosis and classification.

2. RESEARCH METHOD

2.1. Dataset

This research uses the wisconsin diagnostic breast cancer (WDBC) dataset from UCI machine learning repository [10]. This dataset has 32 attributes and 569 of instances without missing values. From the number of instances, 357 belongs to benign and 212 belongs to malignant. Benign and malignant are as a diagnosis in. All features are processed into numerical from a digitized image of a fine needle aspirate (FNA) of a breast mass and recoded with four significant digits. The dataset described characteristics of the cell nuclei present in the image.

2.2. Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA) is one of discriminant analysis method which can be used in classification and dimension reduction [11-13]. The main purpose of linear discriminant analysis is to predict the best categorize for multi-class labels [14]. Apply following equation:

$$Z = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d \quad (1)$$

$$S(\beta) = \frac{\beta^T \mu_1 - \beta^T \mu_2}{\beta^T C \beta} \quad (2)$$

Refers to the score function

$$S(\beta) = \frac{Z_1 - Z_2}{Z\ variance\ in\ the\ group} \quad (3)$$

The score function is maximized by the estimation linear coefficients. It is calculated by the following formula.

$$\beta = C^{-1}(\mu_1 - \mu_2) \quad (4)$$

$$C = \frac{1}{n_1 + n_2}(n_1 C_1 + n_2 C_2) \quad (5)$$

Where β menas the linear model coefficients, C means the covariance matrix and μ means the average vector.
To calculate the best discriminant between the two groups, use the Mahalanobis equation:

$$\Delta^2 = \beta^T(\mu_1 - \mu_2) \quad (6)$$

$$\beta^T \left(x - \left(\frac{\mu_1 + \mu_2}{2} \right) \right) > log \frac{P(C_1)}{P(C_2)} \quad (7)$$

The equations Δ represents the mahalanobis difference between the two groups, x represent data vector, and P represents class probabilities.
For the final step, if the condition in equation is satisfied, a new feature is classified [15-16].

2.3. Support vector machines (SVM)

Support vector machine is supervised machine learning technique for classification and regression problems which was proposed by Vapnik *et al*. in 1992. SVM is a computational algorithm that learns to assign labels to object from experience and examples. SVM can be applied to medical diagnosis [17-19] weather prediction, finance [20], stock market analysis [21-22] and image processing [23]. SVM has the fundamental feature of separating binary labeled data centered on a line that achieves the labeled data's maximum distance [24]. To help labeled data separate, SVM uses a hyperplane which divides plane into classes and measuring a maximum margin where in class lies on the either side. Given a dataset $\{x_i, y_i\}_{i=1}^N$ where $x_i$is an element of R^D, y_i is the class label, where $y_i \in \{-1,1\}$ for binary classification, and N is number of samples [25]. Since the goal of SVM is to find the best hyperplane, it follows:

$$w.x + b = 0 \quad (8)$$

The decision function can be expressed as:

$$f(x) = \text{sign}(w.x + b) \quad (9)$$

From (9) here $w = \sum_{i=1}^{N} \alpha_i y_i x_i$ and $b = \frac{1}{N}\sum(y_i - \sum \alpha_m y_m x_m)$

3. RESULTS AND ANALYSIS

This research used RStudio software for running the program of both methods which are support vector machines and linear discriminant analysis. From Figure 1, the result og linear discriminant analysis, the red graph is a group of samples diagnosed with benign breast cancer and the blue graph is for malignant. It can be said that the linear discriminant analysis successfully classifies base on the dataset. According to Tables 1-2, there are 355 of samples that has benign breast cancer correctly and 2 of healthy samples that were incorrectly identified breast cancer. Samples that has malignant correctly are 194 from Table 1 and 207 from Table 2. By testing the accuracy, sensitivity, specificity and F1-score with 80% of data training and 20% of data testing. The result is in following table.

From Table 3, support vector machines (SVM) has better performance than linear discriminant analysis (LDA) according to the percentage of the result. Accuracy measure how accurate of the model performance that perform the data. Accuracy from support vector machines is 98.77% it is representing the accurate of the model in support vector machines and its more accurate than linear discriminant analysis that has 96.49%. Sensitivity is the probability that patients with cancer are diagnosed with our model. Sensitivity is 99.44%, both of linear discriminant analysis and support vector machines are same. Specificity is the probability that patients without cancer are not diagnosed. Specificity from support vector machines is 97.64% and from linear discriminant analysis is 91.51%. F1-score measured the realistic accuracy the model performances. F1-score from support vector machines is 99% and 97,26% for linear discriminant analysis.

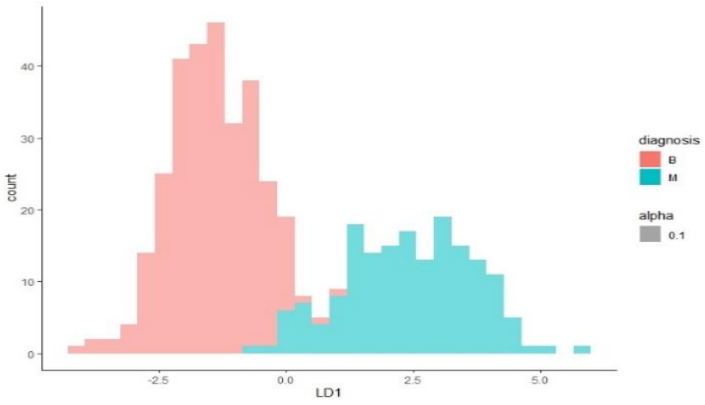

Figure 1. Linear discrimnant analysis classification result

Table 1. Confusion matrix of linear discriminant analysis

Prediction	Reference	
	B	M
B	355	18
M	2	194

Table 2. Confusion matrix of support vector machines

Prediction	Reference	
	B	M
B	355	5
M	2	207

Table 3. Result from linear discriminant analysis and support vector machines

	LDA	SVM
Accuracy (%)	96.49	98.77
Sensitivity (%)	99.44	99.44
Specificity (%)	91.51	97.64
F1-Score (%)	97.26	99

4. CONCLUSION

According to the result, both support vector machine and linear discriminant analysis has a good performance based on accuracy, sensitivity, specificity and F1-score. By comparing two methods based on the number of results, it can be concluded that support vector machine better than linear discriminant analysis. Support vector machines has been widely used by researchers especially on breast cancer classification because it has a good performance. Support vector machines is suggested to help the doctor to predict and classify a disease or a dataset that similar.

ACKNOWLEDGEMENT

This research supported financially by University of Indonesia, with a DRPM PUTI Q2 2020 research grant scheme.

REFERENCES

[1] World Health Organization, "Breast cancer: prevention and control," available: https://www.who.int/cancer/detection/breastcancer/en/ (accessed January 10, 2020).

[2] American Cancer Society, "Breast Cancer," available: cancer.org//cancer/breast-cancer(accessed January 10, 2020).

[3] Lestari, A. W., & Rustam, Z., "Normed kernel function-based fuzzy possibilistic C-means (NKFPCM) algorithm for high-dimensional breast cancer database classification with feature selection is based on Laplacian Score," In *AIP Conference Proceedings*, vol. 1862, no. 1, AIP Publishing LLC, 2017, doi:10.1063/1.4991247.

[4] Fijri, A. L., & Rustam, Z., "Comparison between fuzzy kernel C-means and sparse learning fuzzy C-means for breast cancer clustering," In *2018 International Conference on Applied Information Technology and Innovation (ICAITI)*, pp. 158-161, 2018, doi: 10.1109/ICAITI.2018.8686707.

[5] M. Toğaçar and B. Ergen, "Deep learning approach for classification of breast cancer," In *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, pp. 1-5, 2018, doi: 10.1109/IDAP.2018.8620802.

[6] Rustam, Z., & Hartini, S., "Classification of Breast Cancer using Fast Fuzzy Clustering based on Kernel," In *IOP Conference Series: Materials Science and Engineering*, vol. 546, no. 5, 2019, DOI: 10.1088/1757-899X/546/5/052067.

[7] Ting, F. F., Tan, Y. J., & Sim, K. S. Convolutional neural network improvement for breast cancer classification. *Expert Systems with Applications*, vol. 120, pp. 103-115, 2019, DOI: 10.1016/j.eswa.2018.11.008.

[8] Yan, R., Ren, F., Wang, Z., Wang, L., Zhang, T., Liu, Y., & Zhang, F. Breast cancer histopathological image classification using a hybrid deep neural network. *Methods*, vol. 173, pp. 52-60, 2020, doi: 10.1016/j.ymeth.2019.06.014.

[9] Vijayarajeswari, R., Parthasarathy, P., Vivekanandan, S., & Basha, A. A. "Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform," *Measurement*, vol. 146, pp. 800-805, 2019, doi:10.1016/j.measurement.2019.05.083.

[10] Wolberg, W. H., Street, W. N., & Mangasarian, O. L., "Breast cancer Wisconsin (diagnostic) data set," *UCI Machine Learning Repository*, 1992, Available:http://archive.ics.uci.edu/ml/.

[11] Wang, S., Lu, J., Gu, X., Du, H., & Yang, J., "Semi-supervised linear discriminant analysis for dimension reduction and classification," *Pattern Recognition*, vol. 57, pp. 179-189, 2016, doi:10.1016/j.patcog.2016.02.019.

[12] Babu, P. H., & Gopi, E. S. "Medical data classifications using genetic algorithm based generalized kernel linear discriminant analysis," *Procedia Computer Science*, vol. 57, pp. 868-875, 2015, doi:10.1016/j.procs.2015.07.498.

[13] Marchevsky, A. M., Tsou, J. A., & Laird-Offringa, I. A. Classification of individual lung cancer cell lines based on DNA methylation markers: use of linear discriminant analysis and artificial neural networks. *The Journal of Molecular Diagnostics*, vol. 6, no. 1, pp. 28-36, 2004, doi 10.1016/S1525-1578(10)60488-6.

[14] Toğaçar, M., Ergen, B., & Cömert, Z., "Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders," *Medical hypotheses*, vol. 135, 2020, doi:10.1016/j.mehy.2019.109503.

[15] Mandelkow, H., de Zwart, J. A., & Duyn, J. H., "Linear discriminant analysis achieves high classification accuracy for the BOLD fMRI response to naturalistic movie stimuli," *Frontiers in human neuroscience*, vol. 10, no. 37, 2016, DOI: 10.3389/fnhum.2016.00128.

[16] Bernstein, R., Osadchy, M., Keren, D., & Schuster, A., "LDA classifier monitoring in distributed streaming systems," *Journal of Parallel and Distributed Computing*, vol. 123, pp. 156-167, 2019, doi:10.1016/j.jpdc.2018.09.017.

[17] Rustam, Z., & Maghfirah, N., "Correlated based SVM-RFE as feature selection for cancer classification using microarray databases," In *AIP Conference Proceedings*, vol. 2023, no. 1, 2018, doi: 10.1063/1.5064232.

[18] Nadira, T., & Rustam, Z., "Classification of cancer data using support vector machines with features selection method based on global artificial bee colony," *Proceedings Of The 3rd International Symposium On Current Progress In Mathematics And Sciences 2017 (ISCPMS2017)*, vol. 2023, no. 1, 2018, DOI: 10.1063/1.5064202.

[19] Rampisela, T. V., & Rustam, Z., "Classification of schizophrenia data using support vector machine (SVM)," In *Journal of Physics: Conference Series*, vol. 1108, no. 1, 2018, DOI: 10.1088/1742-6596/1108/1/012044.

[20] Rustam, Z., & Ariantari, N. P. A. A., "Support Vector Machines for Classifying Policyholders Satisfactorily in Automobile Insurance. In *Journal of Physics: Conference Series*, vol. 1028, no. 1, 2018, DOI: 10.1088/1742-6596/1028/1/012005.

[21] Rustam, Z., Vibranti, D. F., & Widya, D., "Predicting the direction of Indonesian stock price movement using support vector machines and fuzzy kernel C-means," In *Proceedings Of The 3rd International Symposium On Current Progress In Mathematics And Sciences 2017 (ISCPMS2017)*, vol. 2023, no. 1, 2018.

[22] Puspitasari, D. A., & Rustam, Z., "Application of SVM-KNN using SVR as feature selection on stock analysis for Indonesia stock exchange," In *Proceedings Of The 3rd International Symposium On Current Progress In Mathematics And Sciences 2017 (ISCPMS2017)*, vol. 2023, no. 1, 2018, DOI: 10.1063/1.5064204.

[23] Rustam, Z., & Ruvita, A. A., "Application Support Vector Machine on Face Recognition for Gender Classification. In *Journal of Physics: Conference Series*, vol. 1108, no. 1, 2018, DOI: 10.1088/1742-6596/1108/1/012067.

[24] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D., "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, 906-914, 2000, DOI: 10.1093/bioinformatics/16.10.906.

[25] Cristianini, N., & Shawe-Taylor, J., "An introduction to support vector machines and other kernel-based learning methods," Cambridge university press, 2000, doi:10.1017/CBO9780511801389.