Jurnal Mantik
Volume 4, Number 2, Aug 2020, pp. 1424-1429 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik

Classification of Breast Cancer Using Support Vector Machine and Forward Selection

Leliana Harahap[1], Erwin Setiawan Panjaitan[2], Muhammad Fermi Pasha[3]

[123]STMIK Mikroskil, Jl. M.H Thamrin No.140 Kel, Pusat Ps., Kec. Medan Kota, Kota Medan, Sumatera Utara 20212, Indonesia.

E-mail: leliharahap05@gmail.com[1], erwin@mikroskil.ac.id[2], fermi@ieee.org[3]

ARTICLE INFO

Article history:
Received: 12/01/2020
Revised: 22/07/2020
Accepted: 01/08/2020

ABSTRACT

Breast cancer represents about 12% of all new cancer cases and 25% of all cancers in women. Early detection and classification of cancer is essential to save a person's life. The causes of breast cancer are multi-factorial and involve family history, obesity, hormones, radiation therapy, and even reproductive factors. Each year, one million new women are diagnosed with breast cancer, according to a World Health Organization report, half of them will die, because it is usually too late when doctors detect cancer. After the selected variable is then evaluated based on certain criteria. If the first selected variable meets the criteria for inclusion, the selection continues. The procedure stops, if no other variables meet the entry criteria and adds the variables one by one. The accuracy of the Support Vector Machine is influenced by several factors, including the comparison of the amount of training data and test data adjusted for k-fold validation. In the comparison of training data and test data the resulting accuracy reaches 97.68% with a total composition of 345 training data (50%) and 345 test data (50%). In the tests carried out, the accuracy of Support Vector Machine and Forward Selection was obtained at 97.68%.

Keywords: Breast Cancer, Classification, Support Vector Machine, Forward Selection

1. Introduction

The second leading cause of female death is breast cancer (after lung cancer) 1,246,660 new cases of female invasive breast cancer are estimated to be diagnosed in the US during 2016 and an estimated 40,450 female deaths. Breast cancer represents about 12% of all new cancer cases and 25% of all cancers in women (Asria et al, 2016). Early detection and classification of cancer is essential to save a person's life. One of the terrible diseases that affect women is breast cancer and it is a major concern in the medical field. Breast cancer arises from the tissue of breast cells. Similar to other parts of the human body, the breast is made up of many microscopic cells. In the case of breast cancer, cell multiplication occurs rapidly in the breast and spreads to other parts of the human body (Prabhakar and Rajaguru, 2018).

The causes of breast cancer are multi factorial and involve family history, obesity, hormones, radiation therapy, and even reproductive factors. Each year, one million new women are diagnosed with breast cancer, according to a report by the World Health Organization, half of them will die, because it is usually too late when doctors detect cancer (Altonen et al., 1998). Breast cancer is caused by mutations in single cells, which can be closed by the system or cause reckless cell division. If the problem is not corrected after a few months, it will become a malignant tumor.

Malignant tumors develop into neighboring cells, which can cause metastases or reach other parts, while benign tumors cannot develop into other tissues, expansion is then limited to benign tumors (Chakraborty, 2009). Detection of SM may be difficult at the beginning of the disease, due to the absence of symptoms, after several clinical tests, an accurate diagnosis must have the ability to distinguish benign and malignant tumors. Good detection provides a low false positive rate (FP) and a negative false rate (FN) (Guyon et al., 2002).

Machine Learning (ML) is a set of tools used for the creation and evaluation of algorithms that facilitate prediction, pattern recognition and classification. Machine Learning is based on four steps: Collecting data,

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

choosing a model, training the model, testing the model (Gokhale, 2009). Machine Learning is a branch of science that implements mathematical algorithms into computer programming to identify data patterns and improve performance iteratively. Machine learning applications have solved many problems such as cancer patient prediction and corporate bankruptcy prediction (Lynch1, 2017). In recent years, various machine learning and soft computing techniques have been used to classify various medical problems including breast cancer. They classify breast cancer with the help of two techniques such as the Softmax Discriminant Classifier (SDC) and Linear Discriminant Analysis (LDA) (Prabhakar and Rajaguru, 2018).

The relationship between breast cancer and machine learning is not very new as it has been widely used for decades to classify tumors and other malignancies, predict gene sequences responsible for cancer and determine prognostics (Tang et al., 2009). The purpose of classification is to place each observation in the category it belongs to. In this research, we will use a machine learning classifier, namely Support vector machine. The goal is to determine whether the patient has a benign or malignant tumor. In this study using the Wisconsin breast cancer database. The aim of the study was to develop an effective machine learning approach to cancer classification using classifiers in a data set. The performance of the classification results will be evaluated in terms of accuracy, training process and testing process using the Feature Selection technique.

Feature selection is one of the most important techniques and is often used in pre-processing. This technique reduces the number of features involved in specifying a target class value, reduces irrelevant, redundant features and data that causes misunderstanding of the target class which has an immediate effect on the application. The main purpose of feature selection is to select the best feature from a data feature set.

Feature selection is a process of removing redundant and irrelevant features from the actual dataset. So that the time used to execute the classifier that processes data is reduced, and it can also increase accuracy because irrelevant features can worsen the data negatively affecting classification accuracy (S. Doraisami and S. Golzari, 2008). With feature selection, it can increase understanding and reduce data handling costs (A. Arauzo-Azofra, 2011). The Feature selection algorithm is divided into three groups: filters, wrappers, and embedded selectors. Filters evaluate each feature independently of the classifier, ranking the features after evaluating and taking the best ones (Guyon Isabelle and A. Elisseef, Journal of Machine learning Research). Wrappers take a subset of the feature set, evaluate the classification performance of this subset, and then the other subsets are evaluated by classifiers. The subset that has the maximum performance in the classification will be selected. So the wrappers depend on the classifier chosen. Even wrappers are more reliable because the classification algorithm affects the level of accuracy (J. Novakovic, 2010).

Forward selection is a type of incremental regression that starts with a blank model. In forward selection, the first variable selected for entry into the built model is the one with the greatest correlation with the dependent variable. After the variables have been selected, they are evaluated based on certain criteria. If the first selected variable meets the criteria for inclusion, then forward selection continues. The procedure stops, when no other variables are left that meet the entry criteria and add the variables one by one.

2. Theoretical basis

A. Breast cancer

Cancer is a disease caused by abnormal cell growth. These cells exist because of changes in gene expression, so the cancer will develop into a cell population that can attack certain tissues (Ruddon R, 2007). Changes in the appearance of genes that extend into the cells can cause functional shifts of these cells. This is very dangerous because it can cause death. Based on the Global Cancer (GLOBOCAN) statistical data from the International Research on Cancer (IARC) in 2018, there were 18.1 million cases of cancer in the world and 9.6 million of them had died. In 18.1 million cases of cancer, the most common cancer cases experienced by men were prostate cancer cases, while the most common cancer cases experienced by women were breast cancer cases (IARC, 2018).

B. Data Mining

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data. The idea is to build a computer program that filters through the database automatically, looking for regularities or patterns. Strong patterns, if found, will likely generalize to make accurate predictions of future data. Anything that is found will be incorrect but there will be exceptions to every rule and cases that are not covered by any rule. Algorithms must be robust enough to deal with imperfect data and extract imprecise but useful regularities (Ian & Eibe 2005).

C. Classification

Classification is one of the Data Mining techniques which is used to analyze a given dataset and take each instance of it and assign this instance to a particular class so that there will be minimal misclassification.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

This classification is used to extract a model that accurately defines important data classes in a given dataset. There are 2 stages in the classification process, the first step the model is created by applying a classification algorithm to the training data set and in the second step the extracted model is tested against a predetermined test dataset to measure the performance and accuracy of the model trained by the model. So classification is the process of assigning class labels from datasets whose class labels are unknown (Nikam, Orient. J. Comp. Sci. & Technol, 2015). There are many types of algorithms used to classify data, including: ID3 Algorithm, C4.5 Algorithm, K Nearest Neighbors Algorithm, Naïve Bayes Algorithm, ANN Algorithm, SVM Algorithm.

D. Support Vector Machine(SVM)

Support Vector Machine (SVM) is a learning system for classifying data into two groups of data using a hypothetical space in the form of linear functions in a high-dimensional feature space. SVM has properties that are not shared by machine learning in general, namely in the process of finding the best hyperplane so that the maximum size of the margin between nonlinear input space and feature space is obtained using kernel rules (Cortes & Vapnik 1995). The margin is twice the distance between the hyperplane and the support vector. The point closest to the hyperplane is called the support vector.

In what is assumed it is given a set S of points x_i ϵ R ^ n IRn with i = 1, 2, ..., N. Each point xi belongs to one of the two classes and is thus labeled Yi ϵ (1, - 1}. Its purpose is to define a hyperplane equation that divides S leaving all points of the same class on the same side while maximizing the minimum distance between one of the two classes and the hyperplane. For this purpose several preliminary definitions are required (Pontil & Verri 1997).

First, the set S can be separated linearly if there are w ϵ R ^ n and b ϵ R in such a way

$$x_i.w + b \geq +1 \; if \; y_i = +1 \quad (1)$$

$$x_i.w + b \leq -1 \; if \; y_i = -1 \quad (2)$$

With

w = weight vector perpendicular to the hyperplane (normal plane)

b = position of the plane relative to the coordinate center

In simpler notation, the two inequalities above can be rewritten

$$Y_i(w.x_i + b) \geq 1, \quad (3)$$

Untuk$i$=1,2,…., N. Pasangan (w, b) menunjukkan hyperplane persamaan

$$w.x + b = 0 \quad (4)$$

Dinamakan dengan *separating hyperplane.* Jika dilambangkan dengan ω yang berarti w, jarak yang ditandai $d_i$ditikikx_i dari hyperplane pemisah (w, b) diberikan oleh

$$d_i = \frac{w.x_i+b}{\omega} \quad (5)$$

Kombinasi dari pertidaksamaan dan persamaan diatas untuk seluruh $x_i \in S$, maka

$$y_i d_i \geq \frac{1}{\omega} \quad (6)$$

E. K-Fold Cross Validation

Cross validation is a statistical technique that is generally used to examine and evaluate algorithms or learning models by partitioning data into learning sets to train the model and test sets to evaluate them. The training and test sets in cross-validation are randomly divided into partitions (60% data in the training set and 40% data in the test set) and go through successive crossover loops so that each instance is tested. K-fold cross validation is the basic form of one of the K partitions used as a validation set (A. Lavecchi, 2005).

F. Feature selection

Feature selection is a process of removing redundant and irrelevant features from the actual dataset. So that the time used to execute the classifier that processes data is reduced, and it can increase accuracy also because irrelevant features can worsen the data negatively affecting classification accuracy (S. Doraisami and S. Golzari, 2008). With feature selection, it can increase understanding and reduce data handling costs (A. Arauzo-Azofra, 2011).

G. Confusion Matrix

A confusion matrix is a method that is usually used to calculate the accuracy of the data mining concept. In performance measurement using confusion matrix, there are 4 (four) terms as a representation of the results of the classification process. The four terms are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). True Negative (TN) value is the number of negative data detected correctly, while False Positive (FP) is negative data but detected as positive data. Meanwhile, True Positive (TP) is positive data that is detected correctly. False Negative (FN) is the opposite of True Positive, so the data is positive, but is detected as negative data.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

3. Results and Conclusions

A. Dataset

At this stage, preprocessing data sharing has been carried out as described in the previous chapter with the amount of data of 690 and 9 attributes and 2 classes. dataset information used:

Table 1
Dataset Information

No	Atribut	Type
1	Clump Thickness	Numeric
2	Uniformity of Cell Size	Numeric
3	Uniformity of Cell Shape	Numeric
4	Marginal Adhesion	Numeric
5	Single Epithelial Cell Size	Numeric
6	Bare Nuclei	Numeric
7	Bland Chromatin	Numeric
8	Normal Nucleoli	Numeric
9	Mitoses	Numeric
10	Class	Numeric

B. Data Normalization

The attribute variable with a large value has a greater influence in making classification predictions than the variable with a small value. To solve this problem, normalization techniques are used so that all variables are in the same range and no variable has a dominant influence on other variables. To calculate data normalization, the formula is used:

$$Data_{normalisasi} = \frac{Datai - \min\ Data}{MaxData - MinData} \quad (1)$$

Where the mindata value is the minimum value of the dataset on the ith data attribute, maxdata is the maximum value of the dataset on the i th data attribute, ai is the result of normalization and vi is the i th data attribute data. Data before normalization can be seen in table 4.2.

Table 2.
Data Before normalized

Data Id	1	2	3	4	5	6	7	8	9	Class
1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2
1035283	1	1	1	1	1	1	3	1	1	2

Data that has been normalized can be seen in table 3

Table 3
Data After being normalized

1	2	3	4	5	6	7	8	9	Class
0,57	0,00	0,00	0,00	0,17	0,00	0,25	0,00	0,00	0,00

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

1	2	3	4	5	6	7	8	9	Class
0,57	0,33	0,33	0,57	1,00	1,00	0,25	0,17	0,00	0,00
0,29	0,00	0,00	0,00	0,17	0,11	0,25	0,00	0,00	0,00
0,71	0,78	0,78	0,00	0,33	0,33	0,25	1,00	0,00	0,00
0,43	0,00	0,00	0,29	0,17	0,00	0,25	0,00	0,00	0,00
1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,00	1,00
0,00	0,00	0,00	0,00	0,17	1,00	0,25	0,00	0,00	0,00
0,14	0,00	0,11	0,00	0,17	0,00	0,25	0,00	0,00	0,00
0,14	0,00	0,00	0,00	0,17	0,00	0,00	0,00	1,00	0,00
0,43	0,11	0,00	0,00	0,17	0,00	0,13	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,25	0,00	0,00	0,00

C. Testing and Analysis Results

The test scenario carried out in this study is testing 690 breast cancer data with 2-fold validation and testing with variations in the value of k (neighbor). The data to be tested is 50% of the total data and 50% of it is training data. Tests were carried out to obtain the accuracy of breast cancer classification using a Support Vector Machine and forward selection.

Table 4

Table of accuracy results for testing with variations in the k-fold value.

K-Fold	Accuracy	Fitur
2	97,68	8 2 6 9 0 1 0 0 0
3	97,39	6 1 4 3 9 8 0 0 5
5	97,25	2 6 0 1 5 9 8 3 0
6	97,25	9 6 1 5 3 0 0 8 2
10	97,25	4 3 0 7 2 1 6 0 0
15	97,54	1 3 0 6 8 9 4 0 0

D. Classification Results

For testing with validation using variations in the k-fold value, the accuracy results for the greater k value, the greater the accuracy results. For all tests, the best accuracy was obtained at the value of the F-Fold k = 2. From the variation of testing, the best overall accuracy is obtained at 97.68%. The accuracy obtained from the Support Vector Machine classification with forward selection gets a better value than previous research using Support Vector Machine, which is 95.61% (Amrane et al, 2018). From these results it can be concluded that the accuracy of Support Vector Machine with forward selection is higher than using Support Vector Machine alone. The graph of the accuracy results for testing with variations in the k-fold value can be seen in Figure 4.1

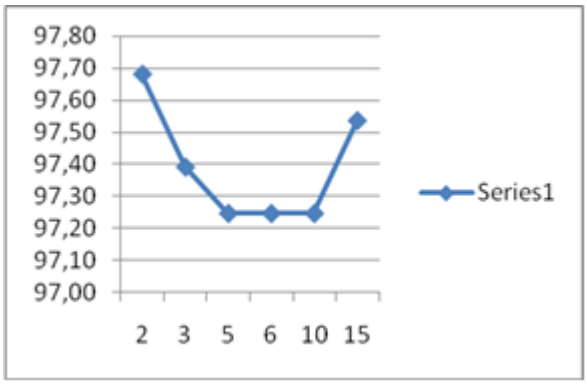

Fig 1. Graph of accuracy results for testing with variations in k-fold values

E. Conclusion

SVM can be applied to the classification of breast cancer data using the Forward Selection technique. There are two main processes involved in determining the classification of breast cancer, namely data

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

preprocessing and the classification process. The preprocessing process aims to normalize the data for each attribute so that each attribute can have a balanced effect on the resulting accuracy in the classification process. Furthermore, the classification process uses forward selection techniques to increase the accuracy value.

The level of accuracy on the Support Vector Machine is influenced by several factors, including the comparison between the amount of training data and test data adjusted for the k-fold validation. In the comparison of training data and test data, the resulting accuracy reaches 97.68% with the total composition of training data as much as 345 data (50%) and test data as much as 345 data (50%). In the tests carried out, the SVM and Forward Selection accuracy obtained was 97.68%.

References

[1] Amrane, M., Oukid, S., Gagaoua, I., & Ensar, T. (2018). Breast Cancer Classification Using Machine Learning. *IEEE* .

[2] Aslan, M. F., Celik, Y., Sabanci, K., & Durdu, A. (2018). Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Analysis Data. *International Journal of Intelligent Systems and Applications in Engineering* .

[3] Chougrad, H., Zouaki, H., & Alheyane, O. (2018). Deep Convolutional Neural Networks for Breast Cancer Screening. *Computer Methods and Programs in Biomedicine* .

[4] Devarriya, D., Gulati, C., Mansharamani, V., Sakalle, A., & Bhardwaj, A. (2019). Unbalanced Breast Cancer Data Classification Using Novel Fitness Functions in Genetic Programming. *Journal Pre-proof* .

[5] Ghaddar, B., & Sawaya, J. N. (2017). High Dimensional Data Classification and Feature Selection using Support Vector Machines. *European Journal of Operational Research* .

[6] Guenther, N., & Schonlau, M. (2016). Support vector machines. *The Stata Journal* , 917–937.

[7] Hamad, Y. A., Simonov, K., & Naeem, M. B. (2018). Breast Cancer Detection and classification Using Artificial Neural Networks . *IEEE* .

[8] Houthuys, L., Langone, R., & Suykens, J. A. (2017). Multi-View Least Squares Support Vector Machines Classification. *Neurocomputing* .

[9] Huang, J., Yu, Z. L., & Gu, Z. (2017). A Clustering Method based on Extreme Learning Machine. *Neurocomputing*.

[10] Jasmir, Nurmaini, S., Malik, R. F., Abidin, D. Z., Zarkasi, A., Kunang, Y. N., et al. (2018). Breast Cancer Classification Using Deep Learning . *International Conference On Electrical Engineering and Computing Science (ICECOS)* .

[11] Liu, N., Shen, J., Xu, M., Gan, D., Qi, E. s., & Gao, B. (2018). Improved Cost-Sensitive Support Vector Machine Classifier for Breast Cancer Diagnosis. *ResearchArticle* , 13.

[12] Nie, F., Wang, X., Jordan, M. I., & Huang, H. (2016). The Constrained Laplacian Rank Algorithm for Graph-BasedClustering. *Association for the Advancement of Artificial Intelligence* .

[13] Nilashi, M., Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). A Knowledge A Knowledge A Knowledge A Knowledge--Based System for Breast Cancer Classification Based System for Breast Cancer Classification Based System for Breast Cancer Classification Based System for Breast Cancer Classification Using F Using F Usi. *Telematics and Informatics* .

[14] Obaid, O. I., Mohammed, M. A., Ghani, M. K., Mostafa, S. A., & Dhief, F. T. (2018). Evaluating the Performance of Machine Learning Techniques in the Classification of Wisconsin Breast Cancer . *International Journal of Engineering & Technology* .

[15] Omondiagbe, D. A., Veeramani, S., Sidhu, A. S., & Sidhu, A. S. (2019). Machine Learning Classification Techniques for Breast Cancer Diagnosis . *Materials Science and Engineering* .

[16] Sahu, H., Shrma, S., & Gondhalakar, S. (n.d.). A Brief Overview on Data Mining Survey . *International Journal of Computer Technology and Electronics Engineering* .

[17] Turgut, S., Dagtekin, M., & Ensari, T. (2018). Microarray Breast Cancer Data Classification Using Machine Learning Methods . *IEEE* .

[18] Verma, A., Kumar, A., & Kumar, M. S. (2019). Breast Cancer Prediction Using Support Vector Machine . *International Research Journal of Engineering and Technology (IRJET)* .