The Seybold Report

ISSN 1533-9211

AN EFFICIENT SUPPORT VECTOR MACHINE BASED BREAST CANCER DETECTION MODEL

TSEHAY ADMASSU ASSEGIE[1], SUSHMA S J[2], BHAVYA B.G[3], PADMASHREE S[4]

[1] Lecturer, Department of Computing Science, Faculty of Computing Technology, Aksum Institute of Technology, Aksum University, Axum, Ethiopia
[2] Associate Professor, Department of Electronics and Communication Engineering, GSSS Institute of Engineering and Technology for Women, MYSURU, India
[3] Assistant Professor, Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysore, India
[4] Professor, Department of Electronics and Communication Engineering, GSSS Institute of Engineering and Technology for Women, MYSURU, India

DOI: 10.5281/zenodo.6553289

Abstract. In this research, a grid search approach is employed to develop an improved support vector machine (SVM) based breast cancer detection model. The grid search is used to find the best combinations of parameters that could maximize the accuracy of the SVM algorithm on breast cancer detection. Moreover, this study explores the effect of parameter tuning on the performance of SVM algorithm on breast cancer detection. The findings of this research shows that parameter tuning has a significant effect on the performance of the SVM algorithm on breast cancer detection. The effect of parameter tuning on the performance of SVM algorithm is experimentally tested on the Wisconsin breast cancer data collected from kaggle data repository. We have compared the performance of the SVM algorithm with the tuned hyper-parameters in the training and with default parameters. The result analysis on the performance of the SVM algorithm for breast cancer detection on the test dataset shows that the accuracy of the proposed enhanced model is 94.07% and the performance of the algorithm with the default hyper-parameters is 89.4%.

Keywords: parameter tuning, breast cancer, breast cancer detection, model optimization, parameter tuning.

1. Introduction

In recent years, breast cancer has become one of the most common types of cancer causing death in the world [1]. The diagnosis of breast cancer requires identification of the presence of the cancer. Medical experts are required in the decision making process that involves analysis of X-ray images and using ultrasound. The manual identification process is time-consuming task and errors are produced due to stress and physician expert experience during detection of the breast cancer. A breast cancer identification can be automated using machine learning algorithms. In this case, a learning algorithm is trained on a set of breast cancer observations and the machine predicts a new observation and determines whether the new observations is a cancerous or normal.

The Seybold Report
ISSN 1533-9211

Although machine-learning algorithms are widely applied in medical research for identification of breast, the accuracy of the learning algorithms desired to be of maximum possible level so that breast cancer identified correctly and with better precision. However, without the use of optimization approaches such as the hyper-parameter tuning, the default parameters used by inherent machine learning algorithms such as support vector machine (SVM) cannot be efficient to predict disease [3-6] with acceptable level of accuracy. Determining the hyper-parameter value that yields highest possible accuracy requires an automated approach. This research propose a novel SVM based model with grid search approach for parameter tuning when training the model. Primarily, this research is devoted to develop an efficient support vector machine based breast cancer detection model. Moreover, this research focuses on exploring the answers to the following research questions:

1) How to optimize the performance of SVM algorithm for breast cancer detection?
2) What is the best combinations of hyper-parameter that could produce maximum possible accuracy on breast cancer detection using SVM algorithm?
3) What is the effect of parameter tuning on the performance of SVM algorithm on breast cancer detection?
4) What is the accuracy of SVM for breast cancer detection?

2. Literature review

Numerous research works have been conducted on the identification of breast cancer using machine-learning algorithms [1-20]. But, most of the researches does not explored the effect of hyper-parameter tuning for the purpose of optimizing the accuracy of the machine learning algorithms in breast cancer detection.

In [2], a support vector machine based breast cancer prediction model is proposed. The proposed model is evaluated and evaluation results shows that the model has 81.80% accuracy on breast cancer prediction.

In another study [3], on breast cancer detection the K-Nearest Neighbor (KNN), SVM and Naive Bayes algorithm are compared against their prediction accuracy on breast cancer identification. The comparative analysis on the performance of the three algorithms shows that SVM algorithm performed better than the KNN and naïve Bayes algorithm.

In [4], four supervised machine-learning algorithms, namely, SVM, logistic regression, random forest and naïve Bayes algorithm are compared on the breast cancer prediction. The comparative result of the predictive accuracy on breast cancer detection shows that random forest has better performance on breast cancer detection as compared to the Naïve Bayes, SVM and logistic regression.

In another study [5], a support vector machine based breast cancer diagnosis model is proposed. In the study, the authors highlighted that determining the best kernel value as a future work. The results analysis shows that the proposed model achieves an accuracy of the 93.06%.

The Seybold Report
ISSN 1533-9211

In [6], the authors compared the performance of SVM, Knearest Neighbor (KNN), random forest and decision tree on breast cancer detection. The authors used the Wisconsin breast cancer data repository for testing and training the model. The comparative result of the performance of these algorithms shows that the KNN algorithm has better accuracy as compared to the random forest, the SVM, decision tree and random forest algorithm.

In [7], a convolutional neural network (CNN), support vector machine (SVM) and random forest (RF) algorithms are used and a model is proposed for breast cancer detection. The result of performance analysis shows that better classification accuracy is achieved with CNN algorithm as compared to the RF and SVM and random forest is better than SVM.

In [8], a comparative study is conducted on support vector machine and random forest algorithms. The performance of support vector machine and random forest algorithm is compared for breast cancer detection. The result of performance comparison shows that the SVM algorithm has better performance as compared to the random forest algorithm on breast cancer detection.

In [9], random forest and support vector machine based breast cancer detection model is proposed. The authors used the Wisconsin breast cancer data repository for testing and training the proposed model.

In [10], the authors surveyed the importance of machine learning algorithms in breast cancer detection. Machine learning algorithms are widely used for detection and forecasting disease cases.

In [11], support vector machine based breast cancer prediction model is proposed using the Wisconsin breast cancer data repository. The author compared the random forest and naïve Bayes algorithm on breast cancer detection. Experimental result shows that the random forest is the most effective algorithm on breast cancer prediction with better accuracy as compared to the support vector machine and the Naive Bayes algorithm.

In [12], a K-Nearest Neighbor (KNN) and support vector machine based breast cancer detection model is proposed. The performance of the algorithms are compared using the accuracy metric as performance measure. The comparative result shows that the support vector machine has better accuracy as compared t the KNN algorithm on breast cancer detection.

In [13], three machine learning algorithms namely the random forest, support vector machine and the Naïve Bayes algorithm based breast cancer detection model is proposed. Experimental result analysis on the performance of the three algorithms on breast cancer detections shows that the random forest algorithm has better accuracy compared to the support vector machine and the Naïve Bayes algorithm.

In [14], the authors proposed a machine-learning model for breast cancer prediction using three machine learning algorithms namely, linear regression, decision tree and random forest. The most popular breast cancer data repository, namely the Wisconsin data repository is used for

The Seybold Report
ISSN 1533-9211

training and testing the model. The performance of the proposed model is analyzed and result shows 88.14% accuracy.

In [15], a Radom forest based breast cancer diagnosis model is proposed. The Wisconsin breast cancer data repository is used to train and test the proposed model and performance test result shows that the model has acceptable level of accuracy even though, the performance can be improved to a better level.

3. Research methodology

In this research, a breast cancer data collected form the kaggle breast cancer data repository is used for training and test support vector machine to develop the proposed model on breast cancer detection.

The dataset consists of 569 observation. To determine the best hyper-parameter combination that could maximize a support vector machine algorithm on breast cancer detection, a grid search method is employed. The proposed model is implemented using python programming language.

3.1 Correlation model

We have employed a Pearson's correlation for analysis and visualization of the relationship among breast cancer dataset features to the target feature. Visualization of the relationship among each feature to the target variable (feature) helped us to investigate predictable relationship among the dataset features. Pearson's correlation for each feature to the target feature in breast cancer dataset is illustrated in figure 1.

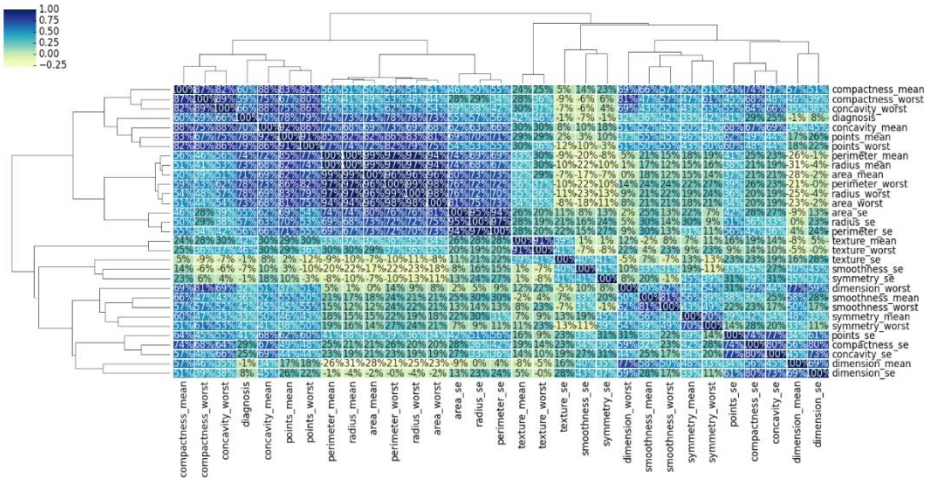

Fig. 1. Relationship among breast cancer dataset features.

Figure 1 shows Pearson correlation among the breast cancer dataset features to the target feature. As shown in figure 1, perimeter worst, radius and mean area has strongly correlated to

The Seybold Report
ISSN 1533-9211

heart failure and ejection fraction is negatively relationship to the target feature (class label namely the diagnosis).

4. Results and Discussions

In this section, experimental results are analyzed using different performance metrics such as the confusion matrix and predictive accuracy. The proposed model is evaluated with confusion matrix and accuracy as performance metrics.

4.1. Confusion Matrix

The confusion matrix for the proposed support vector machine based breast cancer detection model is shown in figure 2. We have employed confusion matrix to analyze the number of correct and incorrect classification by the proposed model, which demonstrated in figure 2. A comparative analysis on internet support vector machine with default hyper parameter and the optimized hyper-parameters with grid search appears to prove that, the optimized model is more efficient as compared to the inherent support vector machine.

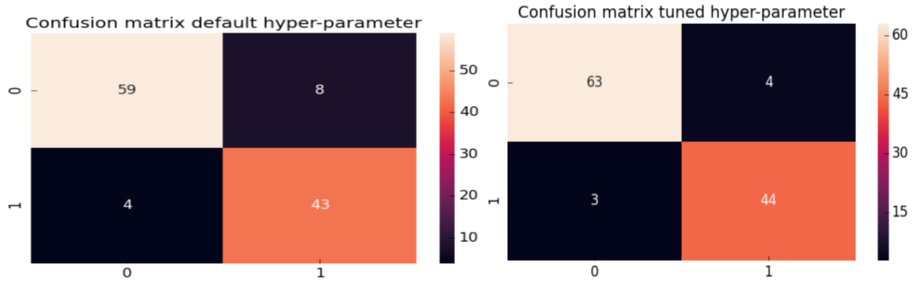

Fig. 2. Confusion matrix for proposed and default parametrized support vector machine on breast cancer detection.

Figure 2 (right) demonstrates the confusion matrix for optimized support vector machine using the grid search to determine the best hyper-parameter combination that produces better classification result. Figure 2(left) shows the confusion matrix for default parameterized support vector machine on breast cancer detection. As demonstrated in figure 2 (left and right), the proposed (optimized) support vector machine miss-classified 7 observations. Whereas the default parametrized support vector machine miss-classified 12 observations. Overall, the confusion matrix demonstrated in figure 2, appears to prove that the proposed support vector machine with tuned hyper-parameters with grid search is more efficient compared to the default parameterized support vector machine for breast cancer classification.

4.2. The effect of gamma and C on accuracy of SVM

The accuracy of the proposed model varies based on the parameters used in training the model. The effect of values of gamma and C is analyzed in this section. The performance of the proposed model for various gamma values is shown in figure 3.

The Seybold Report
ISSN 1533-9211

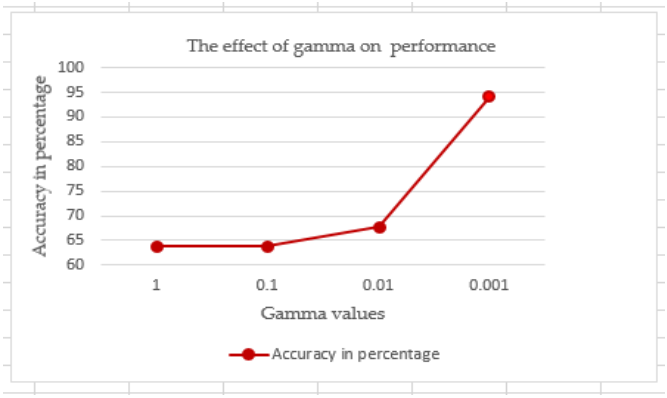

Fig. 3. The effect of gamma on predictive accuracy of SVM

As shown in figure 3, the breast cancer detection accuracy by the proposed SVM based model is the highest when the lower values of gamma is chosen. As demonstrated in figure 3, the highest accuracy (94.04%) is achieved when gamma is equal to 0.001 the result shows that the support vector machine is highly sensitive to gamma values and hence, the accuracy of the support vector machine algorithm is significantly affected by the value of gamma.

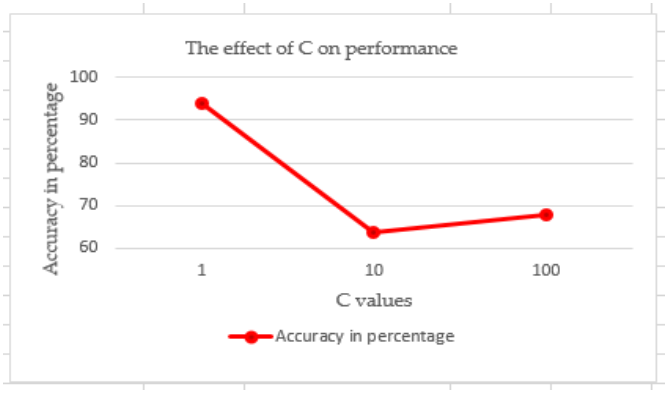

Fig. 4. The effect of C on predictive accuracy of SVM

The effect of C values on the performance of the support vector machine is demonstrated in figure 4. As shown in figure 4, the better performance is achieved when lower value of C is chosen. However, an experimental test result shows that the C value is also depends on the gamma value. We have analyzed that, the lower C and gamma value gives better performance of detection accuracy.

The Seybold Report
ISSN 1533-9211

5. Conclusion

In this, research, we have proposed an optimized support vector machine based model for breast cancer detection. We have employed grid search method for hyper-parameter tuning for selecting the best combination of hyper-parameters for training support vector machine. Moreover, the performance of the proposed model is analyzed with accuracy and confusion matrix as performance measures and result appears to prove that the proposed model has 94.07% accuracy for breast cancer prediction. Furthermore, experimental result reveals that the performance of support vector machine can be improved using grid search approach for parameter tuning..

References

1. Anita Paneri, Mayank Patel, An Improved Model for Breast Cancer Classification Using Svm with Grid Search Method, International Journal of Innovative Technology and Exploring Engineering, 2019.
2. Ze-Wei Guo,, Matben Suchkov, Breast cancer detection via wavelet energy and support vector machine, Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication, Nanjing, China, August 27-31, 2018.
3. Anusha Bharat, Pooja N, R Anishka Reddy, Using Machine Learning algorithms for breast cancer risk prediction and diagnosis, IEEE Third International Conference on Circuits, Control, Communication and Computing, 2018.
4. Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S, Breast Cancer Prediction using Machine Learning, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019.
5. Nasser H. Sweilam, A.A. Tharwat, N.K. Abdel Moniem, Support vector machine for diagnosis cancer disease: A comparative study, Egyptian Informatics Journal, 2010.
6. Arpita Joshi and Dr. Ashish Mehta, Comparative Analysis of Various Machine Learning Techniques for Diagnosis of Breast Cancer, International Journal on Emerging Technologies (Special Issue NCETST-2017) 8(1): 522-526, 2017.
7. S.Vasundhara , B.V. Kiranmayee , Chalumuru Suresh, Machine Learning Approach for Breast Cancer Prediction, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1, May-2019.
8. Dana Bazazeh, Raed Shubair, Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis, IEEE, 2016.
9. Badal Soni, Angshuman Bora, Arpita Ghosh, and Anji Reddy, RFSVM: A Novel Classification Technique for Breast Cancer Diagnosis, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-12, October, 2019.

The Seybold Report
ISSN 1533-9211

10. K.Prasuna, K.V.S.N.Rama Rao, CH.M.H.Saibaba, Application of Machine Learning Techniques in Predicting Breast Cancer – A Survey, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8 June, 2019.
11. Assegie, T.A, Nair,P.S, The Performance of Different Machine Learning Models on Diabetes Prediction, International Journal of Scientific and technology research, January, 2020.
12. Ahmet Saygılı, Classification and Diagnostic Prediction of Breast Cancers via Different Classifiers, International Scientific and Vocational Journal, 2018.
13. Md. Milon Islam, Hasib Iqbal, Md. Rezwanul Haque, Md. Kamrul Hasan, Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors, IEEE Region 10 Humanitarian Technology Conference, 3027.
14. Madeeh Nayer Elgedaw, Prediction of Breast Cancer using Random Forest, Support Vector Machines and Naïve Bayes, International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 6 Issue 1 Jan. 2017.
15. S. Murugan, B. Muthu Kumar, S. Amudha, Classification and Prediction of Breast Cancer using Linear Regression, Decision Tree and Random Forest, International Conference on Current Trends in Computer, Electrical, Electronics and Communication, IEEE, 2017.
16. Bin Dai, Rung-Ching Chen, Shun-Zhi Zhu, Wei-Wei Zhang, Using Random Forest Algorithm for Breast Cancer Diagnosis, International Symposium on Computer, Consumer and Control, IEEE, 2018.
17. Assgie, T.A, ,Sushma S.J, Dr. Prasanna Kumar S.C, Weighted Decision Tree Model for Breast Cancer Detection, Technology Reports of Kansi University, Volume 62, Issue 03, January, 2020.
18. Assgie, T.A, Sushma S. J, A Support Vector Machine and Decision Tree Based Breast Cancer Prediction, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-3, February, 2020.
19. Assegie, T.A, An optimized K-Nearest Neighbor based breast cancer detection, Journal of Robotics and Control (JRC) Volume 2, Issue 3, May 2020.
20. Assegie,T.A, R. Lakshmi Tulasi, N. Komal Kumar, Breast cancer prediction model using decision tree and adaptive boosting algorithm, IAES Journal of Artificial Intelligence, Vol. 10, No.1, March, 2021.