Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 1 : 2024
ISSN : **1906-9685**

ANALYSIS ON CREATING MINIMUM CLASSIFICATION GUIDELINES FOR BREAST CANCER DIAGNOSIS THROUGH THE APPLICATION OF SUPPORT VECTOR MACHINE ALGORITHMS

Musa Om Prakash, Kalpana Allani, Ramya Sakilam, Assistant Professor, Dept. of Computer Science Engineering, Brilliant Institute of Engineering and Technology, Hyderabad, Telangana, India
Uppalapati Vijaya Durga, Student Dept. of Computer Science Engineering, Brilliant Institute of Engineering and Technology, Hyderabad, Telangana, India

ABSTRACT
The Support Vector Machine, or SVM, is a linear model that can be used to solve classification and regression issues. It can solve both linear and nonlinear problems and is useful for a wide range of applications. SVM is a basic concept: The method divides the data into classes by drawing a line or hyperplane. The goal of this study is to see if putting past knowledge into an SVM classifier for Breast Cancer Diagnosis can increase its accuracy. In the diagnosis of breast cancer, a rule-based classifier is extensively utilized. In recent decades, a classifier with good illness classification performance has been developed, and it is in high demand. Because classification rules are derived from previous diagnostics with a large number of features, it is difficult to create a minimum number of high performance rules while retaining all diagnostic information. In this paper, we proposed a SVM model in order to provide the best performance classifier. Based on the experimental results of the SVM model, we can achieve an accuracy of 1.0.
Keywords—Support Vector Machine, Breast Cancer Diagnosis, Machine learning.

1. INTRODUCTION
The support vector machine (SVM) and its extensions are one class of the most successful machine learning methods [1], and have been widely adopted in various application fields [2], [3]. Actually, SVM aims to seek the optimal hyperplane with the maximum margin principle, but the generalization error of SVM actually is a function of the ratio of the radius and margin, i.e., radius-margin error bound [4]. When feature mapping is given, the radius is fixed and can be ignored, and thus SVM can safely minimize the generalization error by maximizing the margin. However, for joint learning of feature transformation and the classifier, the radius information is valuable and cannot be ignored.
Breast cancer is the second leading cancer among the women worldwide. The occurrence of breast cancer is increasing every year by year, due to heredity, increase life expectancy, different lifestyles and food habits. The genuine motivation of this research is tobuild the classification model to classify the breast cancer and to provide the accurate diagnosis to physicians to provide effective treatment to save a life. Thus, efficient classification model increases the mortality of the women. Currently, we have different techniques like X-ray Mammogram, Ultrasound, Magnetic resonance imaging (MRI), Biopsy, Positron Emission Tomography (PET), etc. to evaluate cancer in humans. Though we have different techniques; diagnosis is made by the experienced physicians. When compared to a physician, machine learning diagnosis is more correct, and it is approximated with an accuracy of 91.1% [5]. Thus, usage of machine learning classifier systems in medical diagnosis is increased. The classifier algorithms help experienced/ inexperienced physicians to diagnosis accurately by minimizing possible errors. The most common classifier algorithm used to classify medical data is J48 decision tree.

1.1 Background and Motivations

http://doi.org/10.36893/JNAO.2024.V15I1.028-033

Many people are affected from breastcancer at the present time. Causing of this disease depends on man factors and cannot be simply determined. In addition, the identificationmethod that determines whether or not the cancer is benign or malignant additionally needs an excellent deal of effort from a doctors and physicians. Once many tests are concernedwithin the identification of breast cancer, like clump thickness, uniformity of cell size, uniformity of cell form, etc., the ultimate result could also be troublesome to get, even for doctors. This has given an increase within the previous few years to the utilization of machine learning and computing generally as diagnostic tools. The diseases that take numerous lives, diagnostic computer-based applications are usedwide.

Robotics is taking part in an awfully necessary role in operational rooms. Also, the skilled systems are conferred within the intensive treatment rooms. In turn, using another side of Artificial intelligence for breast cancer designation isn't unworthy. It's reported that breast cancer illness is that the second commonest cancer that affects girls, and was the rife cancer within the world by the year of 2002. This cancer may be a quite common sort of cancer among girls and therefore the second highest reason behind cancer death.

Within the United State, regarding one in eight girls over their time period includes a risk of developing breast cancer. With the uncontrolled division of one cell inside the breast leads to beginning to the breast cancer which results in a visible mass, called a tumour. The tumour can be either benign or malignant. The correct designation in determinant whether or not the tumour is benign or malignant may result in saving lives. Therefore, the necessityfor precise classification within the clinic maybe an explanation for nice concern for specialistsand doctors. This importance of Artificial intelligence has been actuated for the last twenty five years, once scientists began to understand the quality of taking bound selections to treat specific diseases.

The employment of machine learning and data processing as tools in diagnosing becomes terribly effective and one amongst the crucial diseases in medicines wherever theclassification task plays a really essential role is that the diagnosis of breast cancer. Therefore, machine learning techniques will facilitate doctors to create an correct identification for breast cancer and make the proper classification of being benign or malignant tumor. There is little question that analysis of information taken from the patient and selections of doctors and specialists are the foremost necessary factors within the identification, however knowledgeable systems and artificial intelligence techniques like machine learning forclassification tasks, conjointly facilitate doctors and specialists in a great deal.

2. LITERATURE SURVEY

Several risk factors are determined for the cause of breast cancer. Though all women have the chance to be affected by the breastcancer, these factors influence more. Some of these risk factors can be reduced and some cannot be changed. Parker and Folsom [6] discovered that Intentional weight loss of women increase the chance of getting breast cancer. It stated that stoutness and nourishment habits influence the menace of breast cancer in postmenopausal females. Kullberg et al [7] studied both white color and blue color job women dataset and concluded that white color job women are at high risk. It is because of life style and reproduction style, and surveyed all risk factors which cause breast cancer and concluded that developed countries are more inconvenient than developing country.

Akerstedt et al [8] focused the women who are working in night shift and concluded that they are in very risk zone for breast cancer, and studied the influence of genetic disorders and genetically issues accompanying with breast cancer. This work discovered many genetically related proteins which was reason for causing breast cancer. Majoor et al. [9] analyzed the age risk factor in every country. The breast cancer is directly proportional to the age risk factor. The color of the women also participates in causing breast cancer in women. And also proposed the influence of Family history, and stated that the women have high risk chance of getting breast cancer when anyone of her close blood relationhaving this disease.

Kamińska [10] proposed various risk factors which causes breast cancer. This research discovered that the women have high risk ofbreast cancer when they are with condensed breast tissue. If there is sudden change in the size of breast tissue, it may lead to breast cancer. Lobular carcinoma

http://doi.org/10.36893/JNAO.2024.V15I1.028-033

situation arise when the cells that appear similar to cancer cells are in the Breast tissue, on the other hand they may not mature over the wall of the lobules. This condition is not a hunky-dory cancer or pre- cancer. Eventually the probabilities of possibility increases with having LCIS.

3. EXISTING SYSTEM

A. Principal Component Analysis(PCA)

PCA is a mathematical method used for data analysis. It is one of the most significant features extraction techniques [11]. Normally, PCA transforms a set of dependent variables into a set of independents which handles with uncorrelated variables called Principal Component (PC). Most of the largest possible variances will be retained in the first PC andthen the next PCs will decrease the possible variances [12]. The objectives of PCA are toreduce the dimension of the data and select new variables that relevant to the best outcome. There are two approaches using in PCA .i.e. Eigenvalues and Eigenvectors. An eigenvector represents the direction of the line (horizontal, vertical, etc.) and an eigenvalue is a number of variances in the data's direction of eigenvector. The basic process of reducing data dimension byPCA which can reduce the rules of the model. It can be explained as below:

Algorithm 1PCA algorithm

1. Re-center the original dataset to the origin at means zero
2. Compute the sample variance- covariance
3. Compute the eigenvalues andeigenvectors.
4. Decision which principal components should be retained based on theeigenvalues in order to select highest to lowest eigenvalues. It could achieve 95% confidence interval.
5. Find the transformation matrix based on selection of PCs

B. Decision Tree J48

J48 is an algorithm used to create a decision tree for decision making. It is an implementation of C4.5 algorithm by using Java application in the Weka Data mining tool [11]. Many problems have been solved by decision tree classification approach based on dividing and conquering strategy. It can be used to predict an unseen data set based on the various attribute value of the available data. The decision tree is represented by a rule based (if- then rules) which described by nominal and numeric properties. The construction of decisiontree is built from a root node at the top of thetree to any leaf node that defined the feature. A branch feature may stop into a leaf node when searching for subset instance in the same class ormay further create the leaf node when the nodes are not the same class. Every branch from the root node to leaf ode is represented as a rule. It uses Gain Ratio as a splitting condition to separate the data set for normalizing the data into the form of information gain. The highest value of information gain ratio is selected as a root node and then splitting process is continued until reaching the leaf node.

4. PROPOSED SYSTEM

Support Vector Machine Algorithm

Support Vector Machines are a type of supervised machine learning algorithm that provides analysis of data for classification and regression analysis. While they can be used for regression, SVM is mostly used for classification. We carry out plotting in the n- dimensional space. The value of each feature is also the value of the specified coordinate. Then, we find the ideal hyperplane that differentiates between the two classes. These support vectors are the coordinate representations of individual observation. It is a frontier method for segregating the two classes.

http://doi.org/10.36893/JNAO.2024.V15I1.028-033

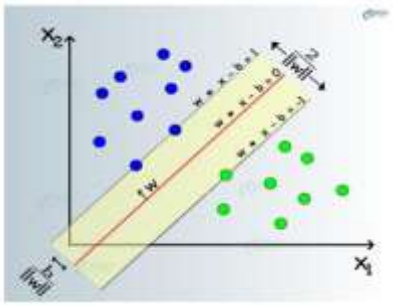

Fig. 1 SVM-Frontier-model

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.

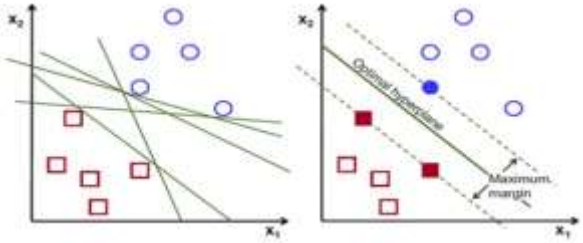

Fig. 2 Possible Hyperplanes

To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e. the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Hyperplanes and Support Vectors

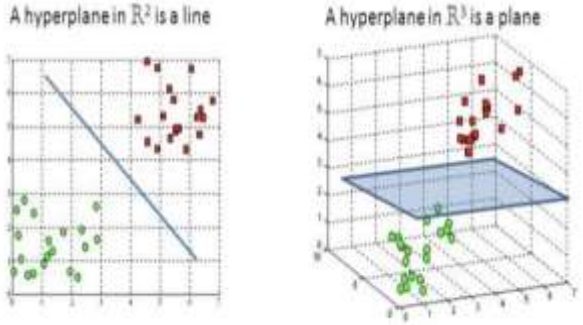

Fig.3 Hyperplanes in 2D and 3D featurespace

Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds 3.

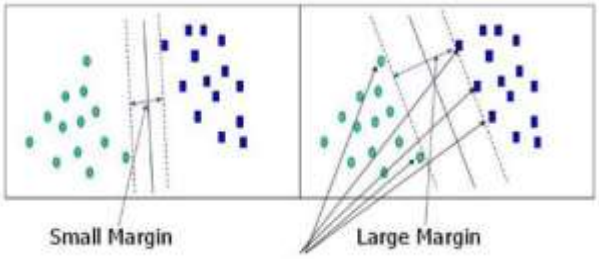

Fig.4 Support Vectors

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM.

Large Margin Intuition

http://doi.org/10.36893/JNAO.2024.V15I1.028-033

In logistic regression, we take the output of the linear function and squash the value within the range of [0,1] using the sigmoid function. If the squashed value is greater than a threshold value (0.5) we assign it a label 1, else we assign it a label 0. In SVM, we take the output of the linear function and if that output is greater than 1, we identify it with one class and if the output is -1, we identify is with another class. Since the threshold values are changed to 1 and -1 in SVM, we obtain this reinforcement range of values ([- 1,1]) which acts as margin.

Cost Function and Gradient Updates

In the SVM algorithm, we are looking to maximize the margin between the data points and the hyperplane. The loss function that helps maximize the margin is hinge loss.

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases}$$

$$c(x, y, f(x)) = (1 - y * f(x))_+$$

The cost is 0 if the predicted value and the actual value are of the same sign. If they are not, we then calculate the loss value. We also add a regularization parameter the cost function. The objective of the regularization parameter is to balance the margin maximization and loss. After adding the regularization parameter, the cost functions looks as below.

$$min_w \lambda \parallel w \parallel^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+$$

Now that we have the loss function, we take partial derivatives with respect to the weights to find the gradients. Using the gradients, we can update our weights.

$$\frac{\delta}{\delta w_k} \lambda \parallel w \parallel^2 = 2\lambda w_k$$

$$\frac{\delta}{\delta w_k} (1 - y_i \langle x_i, w \rangle)_+ = \begin{cases} 0, & \text{if } y_i \langle x_i, w \rangle \geq 1 \\ -y_i x_{ik}, & \text{else} \end{cases}$$

When there is no misclassification, i.e. our model correctly predicts the class of our data point, we only have to update the gradient from the regularization parameter.

$$w = w - \alpha \cdot (2\lambda w)$$

When there is a misclassification, i.e. our model make a mistake on the prediction of the class of our data point, we include the loss along with the regularization parameter to perform gradient update.

$$w = w + \alpha \cdot (y_i \cdot x_i - 2\lambda w)$$

5. SVM IMPLEMENTATION IN PYTHON

The dataset we will be using to implement our SVM algorithm is the Iris dataset. Since the Iris dataset has three classes, we will remove one of the classes. This leaves us with a binary class classification problem.

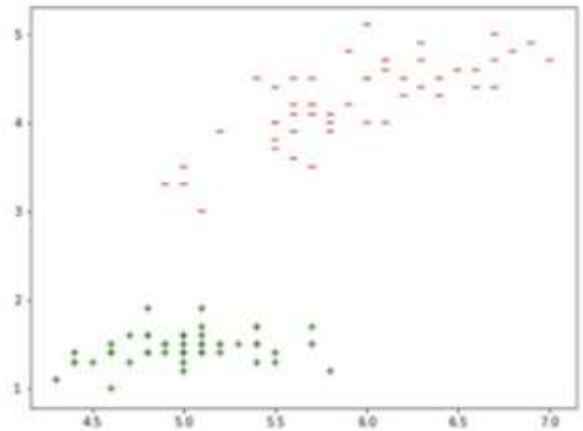

Fig. 5 visualizing data points

Also, there are four features available for us to use. We will be using only two features, i.e Sepal length and Petal length. We take these two features and plot them to visualize. From the above graph,

http://doi.org/10.36893/JNAO.2024.V15I1.028-033

you can infer that a linear line can be used to separate the data points. We extract the required features and split it into training and testing data. 90% of the data is used for training and the rest 10% is used for testing. α(0.0001) is the learning rate and the regularization parameter λ is set to 1/epochs. Therefore, the regularizing value reduces the number of epoch's increases. We now clip the weights as the test data containsonly 10 data points. We extract the features fromthe test data and predict the values. We obtain the predictions and compare it with the actual values and print the accuracy of our model.

Accuracy of our SVM model

```
Accuracy: 1.0
```

There is another simple way to implement the SVM algorithm. We can use the Scikit learn library and just call the related functions to implement the SVM model. Thenumber of lines of code reduces significantly toofew lines. Support vector machine is an elegant and powerful algorithm.

CONCLUSION

A Support Vector Machine is a linear model for classification and regression problems;and for breast cancer diagnosis has been a powerful tool supporting doctor diagnosis. Such a system requires classification rules derived from historical diagnosis. The desirable rules should be minimal in their number and give a good performance. This paper is to obtain such rules from the Wisconsin Breast Cancer data set. It performed experiments on the data set to determine the best classifier among Logistic regression, Naive Bayes Classifier, Nearest Neighbor. It found that SVM classifier giving thebest accuracy.

REFERENCES

1. V. N. Vapnik, Statistical Learning Theory. New York, NY, USA: Wiley, 1998.
2. H. Do, A. Kalousis, and M. Hilario,―Feature weighting using margin and radius based error bound optimization inSVMs,‖ in Proc. ECML PKDD, 2009, pp. 315–329.
3. J. Wu and H. Yang, ―Linear regression- based efficient SVM learning for large- scale classification,‖ IEEE Trans.Neural Netw. Learn. Syst., vol. 26, no. 10, pp. 2357–2369, Oct. 2015.
4. V. Vapnik and O. Chapelle, ―Bounds on error expectation for support vector machines,‖ Neural Comput., vol. 12, no. 9, pp. 2013–2036, Sep. 2000
5. 1.RWBrause. Medical analysis and diagnosis by neural networks. Lecture Computer Sci 2001;2199:1-13.
6. E. Parker and A. Folsom, "Intentional weight loss and incidence of obesity- related cancers: the Iowa Women's Health Study," International journal of obesity, vol. 27, p. 1447, 2003.

C. Kullberg, J. Selander, M. Albin, S. Borgquist, J. Manjer, and P. Gustavsson, "Female white-collar workers remain at higher risk of breast cancer after adjustments for individual risk factors related to reproduction and lifestyle," Occup Environ Med, pp. oemed-2016- 104043, 2017.

7. T. Åkerstedt, A. Knutsson, J. Narusyte,
8. P. Svedberg, G. Kecklund, and K. Alexanderson, "Night work and breast cancer in women: a Swedish cohort study," BMJ open, vol. 5, p. e008127, 2015.
9. B. C. Majoor, A. M. Boyce, J. V. Bovée,
10. V. T. Smit, M. T. Collins, A. M. Cleton-Jansen, et al., "Increased risk of breast cancer at a young age in women with fibrous dysplasia," Journal of Bone and Mineral Research, vol. 33, pp. 84-90, 2018.
11. M. Kamińska, T. Ciszewski, K. Łopacka-Szatan, P. Miotła, and E. Starosławska, "Breast cancer risk factors," Przegladmenopauzalny=Menopause review, vol. 14, p. 196, 2015.
12. J.-W.Liu, Y.H.Chen, and C.H.Cheng, ―Owa based information fusion method with PCA preprocessing for data classification,‖ in International Conference on Machine Learning and Cybernetics, 2012, pp. 3322– 3327.
13. T. R. Patil and S. S. Sherekar, ―Performance analysis of naive bayes and j48 classification algorithm for data classification,‖ International Journal Of Computer Science And Applications, vol. 6, no. 2, 2013.

http://doi.org/10.36893/JNAO.2024.V15I1.028-033