International Journal of Engineering & Technology, 7 (2.33) (2018) 855-860

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A hybrid SVML method for survival of patient post breast cancer operation prediction by using SVM and logistic regression

Rakshanda Agarwal [1], Rajeshkannan Regunathan [2] *

[1] Bachelors of Technology, Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India
[2] Assistant Professor Senior, Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India
**Corresponding author E-mail: rajeshkannan.r@vit.ac.in*

Abstract

A Support Vector Machine is a supervised linear maximum margin classifier and used in many classification applications. While on the other hand Logistic Regression is a regression model which has a categorical dependent variable. Breast cancer operation is a critical one and the survival of the patient is not sure. For a person to be operated, we must know her survival chances after the cancer operation has been performed. Here, in this paper, we propose a hybrid model of a support vector Machine with Logistic Regression namely, SVML (Support Vector Machine-Logistic) which will help us predict the survival chance of the patient post operation. With this model, we can improve the performance of the SVM classifier in terms of its accuracy. Using our model and dataset, we have increased the accuracy to 85.24% for which SVM gave an accuracy of 78.03% and Logistic Regression gave an accuracy of 72.40%.

Keywords: *Breast Cancer; Classification; Logistic Regression; Support Vector Machine (SVM).*

1. Introduction

Breast cancer is the most common form of cancer that is diagnosed among women worldwide. It is a disease of public importance and concern. The number of risk factors for breast cancer exceeds than that of any other cancer type resulting in a complicated and rich aspect of research. The only cure for breast cancer is an operation, the victory of which is not assured or even if it is, patient's survival after a certain number of years is not assured. In the recent years, our understanding about the factors that cause breast cancer has increased at an overwhelming rate as discussed by Ferlay et al. [12]. This paper provides a widespread review and critical valuation of the survival rate or success rate of any patient after breast cancer operation using various machine learning algorithms such as Support Vector Machine (SVM) and Logistic Regression.

We know that Prediction, Classification, and Regression all work hand in hand. To know the result of a certain process we need to know its past result which we get through the training dataset. Now, the test dataset can be used to predict the outcome based on how your data has been trained previously i.e. by knowing or understanding the response faced previously. Training and testing is a vital part of machine learning. In this paper, our aim is to find out the best prediction mechanism by combining the benefits of both Support Vector Machine and Logistic Regression. We have used Haberman's Survival Data Set for our reference [11] purpose, but any other data set containing similar attributes can also be used to implement this algorithm. This dataset is used to predict the survival of the person within a five year period of breast cancer operation.

Cancer is a disease that involves an abnormal growth of cells which spreads in the other parts of the body. Similarly, breast cancer is a cancer that is established from the breast tissues. This can be healed by the removal of the breast. Thus, in our paper, we aim to predict the possibility of a person's survival after five years of breast cancer operation. The combined hybrid model of Support Vector Machine (SVM) and Logistic Regression has been used for accurate results in our data which we call as SVML.

Since this method is a combination of the good qualities present in SVM and Logistic Regression. It will be proved to be better than both the methods individually. This method will always have accuracy percentage greater than each individual method or in the worst case equal to the accuracy of any one method which has a higher accuracy for that data set.

1.1. Support vector machine (SVM)

A Support Vector Machine (SVM) is a linear marginal classifier, which can also be extended to nonlinear cases. Support Vector Machines are robust and commanding tools to categorize data according to Fung et al [1]. It generally categorizes data into two parts dividing it by a hyperactive plane. It is an intricate model that uses multidimensional surfaces to outline the relationship between features and outcomes. In spite of it being complex, it can be applied to real world problems where classification and prediction are essential [2], [18] and [24].

Let us imagine a Support Vector Machine to be a surface that can define boundaries between various points of data in a multidimensional space. Now, the vital goal of SVM is to create a boundary known as the hyper plane which then leads to the homogeneous partitioning of data on either side of this newly created hyper plane. SVM is learning is a combination of linear regression modeling and an instance based nearest neighbor learning. SVMs can be used for any type of learning task, including both classification and prediction [2-3].

Conventionally, SVMs were used for binary classification and hence are most suitable for the same even today. In SVM, a Maximum Margin Hyper plane (MMH) by Lantz [2] is the one that is used to find the best hyper plane which will generate the greatest separation between classes. A best hyper plane is nominated so that future output we get is also best. Even the slightest variation in the position of the line might cause one point to fall into other class. Support vectors are the points which are closest to the Maximum Margin Hyper plane (MMH). Each class has at least one support vector. Using these support vectors it is easy to define a Maximum Margin Hyper plane (MMH).
To clearly explain a Support Vector Machine, we need to explain what exactly a hyper plane is. Let us consider the following equation in an n-dimensional space.

$$\vec{w}.\vec{x} + b = 0 \quad (1)$$

Here, w is a vector of n weights (w_1, w_2, w_3 . . . w_n) i.e. set of points in our case, x is the directional vector and b being the bias value.
Now, we can specify a hyper plane by grouping the set of points which have a value greater than or equal 1 on one side of the hyper plane and the points that have a value less than or equal to -1 on the other side of the hyper plane, as shown in Eq. (2) and Eq. (3). By this, we mean that all points of the first class will fall above the hyper plane and the points that belong to the second class will fall below the hyper plane we have just defined.

$$\vec{w}.\vec{x} + b \geq +1 \quad (2)$$

$$\vec{w}.\vec{x} + b \leq -1 \quad (3)$$

In the case of nonlinear data, we can train our SVM model by using a slack variable which allows some points to be misclassified. Nonlinear data can also be classified by a method of kernel trick. In this method proposed by Furey et al., a nonlinear association may appear to be linear one [3].
Though a widely accepted mechanism, Support Vector Machines also have multiple drawbacks. Choosing the perfect kernel for classification is a big issue and also for multiclass classifier research needs to be done are some of the major drawbacks of SVM.

1.2. Logistic regression

Linear regression is used to provide a commanding data analysis model, but there are certain shortcomings in it. In linear regression, if the dependent variables have only two or three response categories, several assumptions are likely to go unmet [4], [16]. These are overcome by implementation of a new method known as Logistic Regression. When we have a proportion of the response, the dependent variables and explanatory variables are linked together by logit or logistic transform. The transform is as follows:

$$Logit(P) = \log[\frac{p}{1-p}] \quad (4)$$

Here, the term inside log represents the odds that the event occurs, where p is the probability of the event occurring. Using logit or logistic the scale is transformed to plus and minus infinity and also while transforming logit back to normal probability measures we will get the result in the range of 0 to 1 only as obligatory as per Tranmer et al [5]. In Logistic Regression, the dependent variable is categorical unlike in linear regression. Binary Logistic Regression is the most common form of Logistic Regression which is used to forecast the probability of a binary reaction i.e. 0 and 1 or true and false.
Just like SVM, Logistic Regression too has many drawbacks like identification of independent variables, overfitting of the model and requirement for independent observations. Due to all these reasons, there is a need for some new algorithm.

2. Related work

Support Vector Machine and Logistic Regression have various uses in machine learning environment and also in the real world. Support Vector Machine is widely used in classification and also in regression. Its common uses include: text categorization, classification, face detection, forecasting etc. on the other hand Logistic Regression is used in cases where we do not get good results with linear regression. It is used in areas such as Image Segmentation and Categorization, Geographic Image Processing, Handwriting recognition, Healthcare: Examining a group of over million people for myocardial infarction within an epoch of 10 years is a solicitation zone of Logistic Regression and predicting whether a person is depressed or not based on bag of words from the corpus seems to be conveniently solvable using Logistic Regression and SVM. Hence, SVM and Logistic Regression can together work great in providing and predicting better results.
Support Vector Machine has been proved to be grander to neural networks in bankruptcy estimate. To improve the conventional SVM, add the integrated binary discriminant rule (IBDR) especially for corporate financial distress prediction. Here, the output of SVM is modified based on the output of Logistic Regression analysis. If Logistic Regression can provision the output given by SVM with a high probability, the IBDR will admit the output else it will modify the output of SVM. The study has proven that IBDR is largely better than the conventional SVM model in spite of having some limitations in updating the SVM output, such as the technique used to modify the yield of the SVM classifier has a significant impact on the presentation of the consequential approach. There are unconventional means of altering the output of SVM which are remarkable. Another concern according to Hua et al. [6] for forthcoming research communicates to a structured method of selecting an ideal value of constraints in IBDR and SVM for the top prediction performance.
Yilmaz et al. [7] made a comparison of conditional probability (CP), Support Vector Machine (SVM), Logistic Regression (LR) and artificial neural networks (ANNs) for landslide vulnerability mapping. The graph acquired and analyzed with the area under the curve manner shows that the ANN method looks more truthful than the others, but the results show that Conditional Probability (CP) is more simple and accurate in case of landslide susceptibility. Though all the models have a very close accuracy level, there is no specific method that has proved to be better than all other methods for all situations. Though all the models have a very close accuracy level, there is no specific method that has proved to be better than all other methods for all situations; hence there is a need for a new methodology.
Support Vector Machine is commonly used with an arbitrarily generated training dataset classified in advance. An innovative method for Support Vector Machine according to Tong et al. [19] is with active learning i.e. a method to select the instances and to request the next instance. This active learning method diminishes the need for labeled training instances for both transductive and inductive settings.
Using Support Vector Machine for text classification is dignified as the best measure in the paper by E. Leopold et al. [20]. It states that the term frequency transformations have more impact on SVM performance than the kernel. Lemmatization can be avoided in Support Vector Machine which makes it perform better since a large amount of time is hoarded. This also proves that SVM is better than other neural net algorithms and k-nearest neighbor algorithm. This algorithm has not been made generic yet and works only on one language.
Hua et al. [8] adapted Support Vector Machines (SVMs) to estimate the happening of the demand for spare parts. A hybrid forecast approach has been developed that can synthetically evaluate autocorrelation of demand time series and also the relationship that the explanatory variables have with the demand of the spare part. The presentation of exponential smoothening, Croston's method, IFM method, Markov's bootstrapping method and SVM method are compared with this new Logistic Regression and Sup-

port Vector Machine (LRSVM) method. Statistical results show that LRSVM performs best for almost all the times.

A hybrid of Logistic Regression and Support Vector Machine is used for β-turn prediction, which is a secondary protein structure. Here, non β-turn classes are under-sampled multiple times and then combined with β-turn classes till a balanced number of sets is created. Now, each balanced set trains one SVM at a time and its result is passed and aggregated through Logistic Regression model. This method proposed by Elbashir et al. [9] helps to diminish the problem of imbalanced class and as well as the calculation time.

The extension of Support Vector Machine for multiclass classification as proposed by Zhu et al. [17] is still a research centered topic. A new method for classification known as the Import Vector Machine (IVM) has been anticipated which is based on Kernel Logistic Regression (KLR). IVM can not only perform as fit as SVM in two classes but can also be used in multiclass classification. It also provides an estimate of the probability and uses much less training data as compared to the Support Vector Machine (SVM). The only problem in my opinion with this method is its high computation cost of O (N^2q^2) for binary class and O (MN^2q^2) for multi-classes.

The local SVM classifier is used as the groundwork, and then it is unified using Logistic Regression model to perform a more strong operation. Here, Logistic Regression helps us take advantage of the statistical theory of model selection methods such as the added variable plot, step wise regression etc. This method by Chang et al. [10] can also be applied to multicast problems in case of heterogeneity and boost the performance of an SVM classifier.

The two most likely indicators of breast cancer according to Carter et al. [13] are the extent of axillary lymph node participation and tumor size. The data that had been recorded in SEER (Surveillance, Epidemiology and End Results) Program was applied in the evaluation of the breast cancer survival of women in the United States. Five year survival was calculated using the auxiliary lymph node data and the tumor diameter. The lymph nodes status and the tumor diameter were found to be sovereign but important indicators. As the tumor size amplified, survival declined irrespective of the lymph node status. Also, as the lymph node participation amplified, the survival status thus diminished irrespective of the tumor size.

Early Breast Cancer Trialists' Collaborative Group proposed [14] a 10-year and 15-year consequence of breast cancer has been confirmed by a combined meta-analysis of 194 not confounded random trials of hormonal therapy and chemotherapy. The annual death rate of breast cancer when calculated was 38% for women younger than 50 years when detected by breast cancer. While for those who were of age 50-69 years after diagnosis, the death rate was 20% irrespective of tumor characteristics or lymph nodes. There could be an enhancement in the long-term survival of women operated with breast cancer by presenting newer drugs which should be better than the older ones.

Rare events such as wars, vetoes, epidemiological infections, etc. are difficult to predict and explain. This problem has two main sources which are; Logistic Regression which underestimates the probability of prediction of such sporadic actions and common data collections strategies which are inefficient for such rare events. A method with a lower mean square error has been recommended which increases the probability of an event, and thus makes a huge difference to the initial result. According to King et al. [16] the effect of this method will be maximum when the numbers of observations are few or small, for a large number of observations this method is not that efficient.

The fine ambient particulate matter has extensively been linked with numerous health effects. Alleviation pivots on comprehending which sources add to its toxicity. Black Carbon (BC), which is an indicator of elements that are produced from traffic sources, has been connected with a number of health effects, but due to its great spatial inconsistency, it is difficult to guesstimate its concentration. A model was first fit to estimate the concentration of BC but was built using inadequate monitoring data and hence was unable to arrest the complex spatiotemporal forms of ambient BC. In order to expand the projecting ability, a data for 24,301 measurements was obtained from 368 monitors over a period of 12 years in Massachusetts, Rhode Island and New Hampshire. Also, Nu-Support Vector Regression (nu-SVR) – a machine learning method that integrates higher order interactions and nonlinear terms, with suitable regularization of parameter estimates has been used. A model by Awad et al. [22] has been used to correctly evaluate short and long-term exposures to BC and for readings focusing on various health outcomes in MA, RI and Southern NH have been developed in [22].

The development of medical training and knowledge has returned a cumulative number of clinical actions and trials which are used to assess a patient's development. The surplus of accessible tests may be troublesome to clinicians in the lack of confirmation that demonstrates the usefulness of a given method. Hence, there exist serious needs to recognize a distinct number of metrics to arrest during clinical valuation for the effective and concise care of patient care. The data in this paper indicates an extreme of 5 variables can be used to create a risk profile that is grounded on the projected five-factor maximum model. This model provides a policy for future research in progress of injury prediction. In this method by Hewett et al. [23] it is validated how a model that previously exists for the anticipation of primary ACL injury can direct the expansion of the second ACL injury risk analysis, and also in what way the five-factor maximum model might be functional across the injury range for improvement of the injury risk analysis.

3. Methodology

Just like various Support Vector Machine and Logistic Regression models explained above, there is a need for a hybrid algorithm which is better and efficient than both of them specifically in terms of its accuracy. There has been continuous research in these fields as discussed above. Most of them have drawbacks which we need to overcome, one major being the precision and accuracy of the model. The perfect area to test our model is in the field of breast cancer as it is one of the most trending and research topics these days. The operation of breast cancer is challenging and so are the survival chances.

There is no point in an operation if survival for a longer duration is not guaranteed as the operation will be expensive and also troubling to the patient.

Using our model, hospitals can predict survival chances of the patient in a much accurate manner by just knowing the age of the patient and the number of auxiliary nodes. This will help doctors to make a decision of whether or not the patient should be operated.

Table 1: Sample Haberman's Dataset

Age of patient (X30)	Year of Operation (X64)	Number of positive axillary nodes (X1)	Survival Status (X1.1)
30	64	1	1
33	58	10	1
34	59	0	2
34	66	9	2
37	60	15	1
37	63	0	1
38	69	21	2
39	67	0	1
39	59	2	1
42	69	1	2
57	64	0	1

Haberman's survival dataset [11] has been used to predict the survival of the patient after a breast cancer operation. This dataset comprises of [4] attributes which are as follows, also the sample attributes have been shown in Table 1.

- Patient's age when the operation was performed (X30)
- Year in which the operation was performed (X64)
- The Number of positive axillary nodes that were detected (X1)

- Survival status (X1.1 class attribute) 1 = the patient survived for a period of 5 years or more 2 = the patient died within a period of 5 years of operation

In the above dataset [11], Lymph nodes or the number of positive auxiliary nodes i.e. X1 are small clumps of immune cells which act as sifters for the lymphatic system. This lymphatic system runs throughout the entire body and is responsible for transporting cells and fluids. Breast cancer is probable to spread first in the lymph nodes in the underarm i.e. the axillary lymph nodes and then to the entire body. Positive axillary lymph node specifies that the node contains cancer. The removal of these nodes is a fundamental part of the operation [21].

The above mentioned parameters are used to foretell if a patient will survive for more than 5 years after a breast cancer operation. The attributes that help us in this prediction are the age of the patient at the time of operation (X30) and the Number of positive axillary nodes that were perceived (X1). The project is coded in R using e1071 package [15], [18] and [24].

The architectural design of our SVML model is shown in Fig. 1. It consists of an SVM classifier, a Logistic Regression classification model, which comprises of a Logistic Regression system and a classification system, and then it contains an accuracy calculator and comparator which will calculate accuracy using Eq. (5) and finally, a final classifier that classifies the final result and hence completes the SVML model. This model can be trained and then used for additional testing. The working of each block of the architecture diagram has been explained further in the paper.

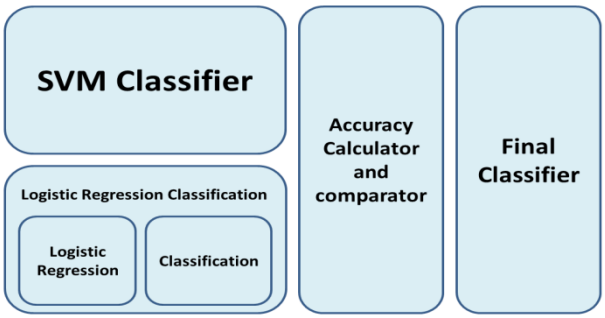

Fig. 1: Architecture Diagram of SVML Model.

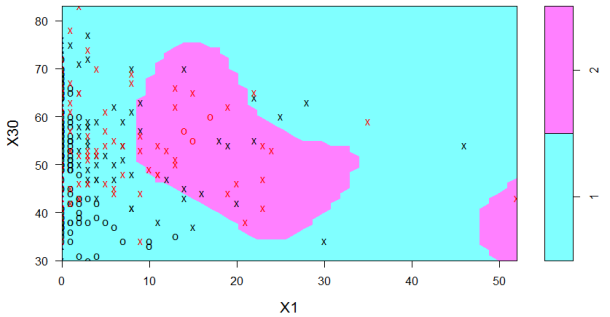

Fig. 2: SVM Classification Plot.

Firstly, we apply SVM classification to our training dataset and classify the data as it can be seen in Fig. 2. We also predict the accuracy of this classification using Eq. (5). The pink area represents patients who died within 5 years of operation i.e. patients belonging to class 2, while the blue area is classified as class 1 meaning that the patient had survived for 5 years or longer.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Where,
TP represents True Positive values' count
TN represents True Negative values' count
FP represents False Positive values' count
FN represents False Negative values' count

Similarly, we now apply Logistic Regression to the same training dataset i.e. Haberman's dataset, classify it, test it and then find out its accuracy using Eq. (5). The output graph obtained using Logistic Regression method for prediction is shown in Fig. 3, this comprises of four graphs residual versus fitted, normal distribution, residual versus leverage and scale location.

Finally, the next step is to predict the output from both Support Vector Machine (SVM) and Logistic Regression combined i.e. the output of our SVML model. For each training data consider the classifier that gives us a better output, i.e. at points where SVM gives us a better result, SVM is considered while on the other where Logistic Regression gives us the better output we consider the output of Logistic Regression. Scenarios where both give us the same output are perfect.

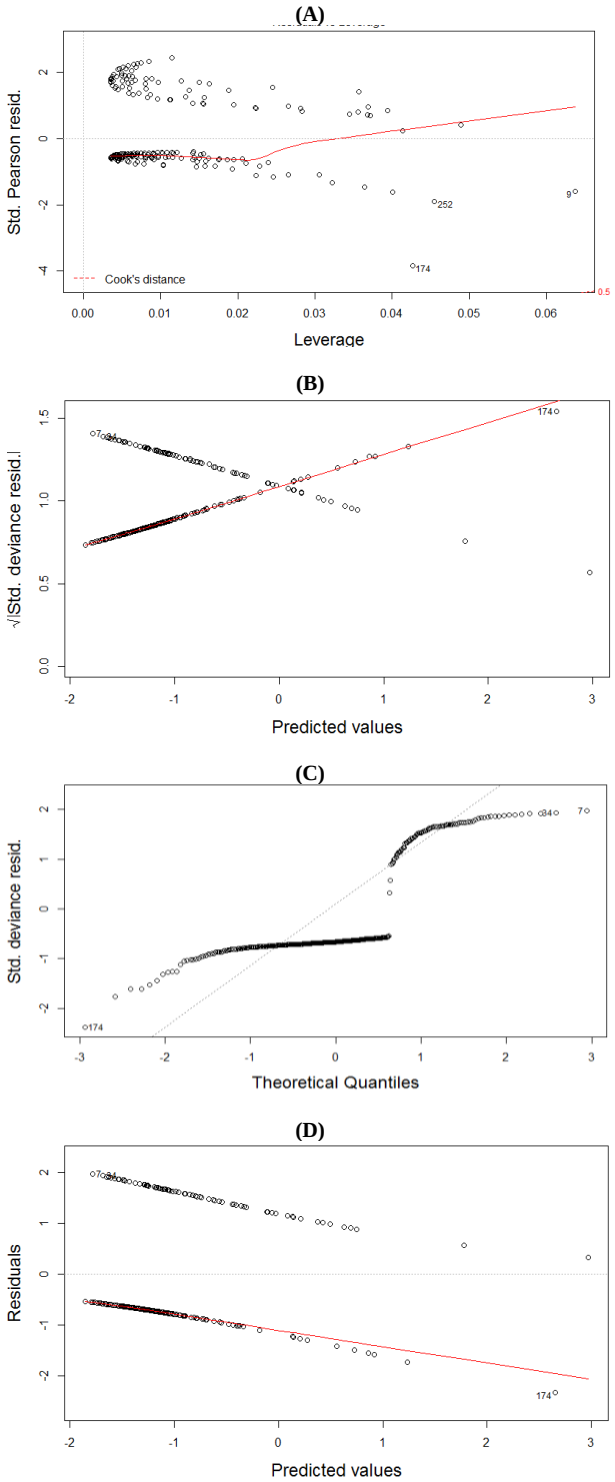

Fig. 3: Output Graph of Logistic Regression: (a) Residual v/s Leverage (b) Scale location (c) Normal Q-Q (d) Residual v/s Fitted.

Now, using this trained data predict the output for the testing dataset and again calculate the accuracy using Eq. (5). Construct a

confusion matrix for all the three outputs; SVM, Logistic regression and the SVML model's final output we just predicted. The confusion matrix for all the three outputs is shown in the Tables 2, 3 and 4 respectively.

Table 2: Confusion Matrix Obtained because of SVM

SVM	Class 1	Class 2
Class 1	214	57
Class 2	10	24

Table 3: Confusion Matrix Obtained because of Logistic Regression

Logistic Regression	Class 1	Class 2
Class 1	175	35
Class 2	49	46

Table 4: Confusion Matrix Obtained because of the SVML Algorithm

SVML	Class 1	Class 2
Class 1	214	35
Class 2	10	46

Table 5: Comparison of the Accuracy Levels Calculated for SVM, Logistic Regression and SVML Model along with their TP, TN, FP and FN Values

Method	True Positive (TP)	True Negative (TN)	False Positive (FP)	False Negative (FN)	Accuracy (%)
SVM	214	24	10	57	78.03
Logistic Regression	175	46	49	35	72.40
SVML	214	46	10	35	85.24

Based on these confusion matrices, calculate the accuracy percentage of each method using Eq. (5). We observe and prove that the SVML algorithm provides better accuracy in predicting the survival of a person after breast cancer. For your clear understanding, the accuracy level, true positive, true negative, false positive and false negative values have been summarized in Table 5. The above mentioned steps have been summarized in the data flow diagram elucidated in Fig. 4. The entire SVML model algorithm has also been specified below.

Though in this paper, our model considers only Haberman's data set, it can be easily implemented for any other data set provided we know the following information: patient's age, year of operation and the number of lymph nodes detected. These three attributes can help us predict the fourth one that is the survival status of the patient.

Also in cases where there is some errant value that deviates from the normal form, we need not worry since our SVM classifier will reject those values during classification. Hence we have used the advantages of both SVM and Logistic Regression to construct our model (SVML) which in almost all cases will give us a perfect or near to perfect result.

3.1. Algorithm

SVML ()

1) Load the dataset.
2) Train it in the SVM classifier algorithm svm_model<- svm_classifier(X1.1~.,data=x) //classify age and number of nodes based on X1.1 i.e. class, the output is shown in Fig. 1.
3) Predict the class of the testing datasets using the svm_model constructed in the above step pred <- predict(svm_model, z) //z is the dataset without class label.
4) Construct a confusion matrix of the above output as shown in Table 2 and calculate its accuracy level using Eq. (5).
5) Train the dataset in step1 using Logistic Regression glm_model <- glm (X1.1~X30+X1, data) //call to Logistic Regression.
6) Predict the output of this glm_model with the given test dataset predict1 <- predict (glm_model)
7) Calculate the mean of the predicted value in step 6.
8) Calculate the class of the predicted data in step 6 For i=0 to number of elements, n if(predict1[i]>mean) ans[i]=2 Else ans[i]=1 End for.
9) Repeat step 4 for Table 3, i.e. output of Logistic Regression.
10) Compare the accuracy in step 4 and step 9 using Eq. (5) and modify the output accordingly for the SVML final output.
11) Repeat step 4 for Table 4, i.e. SVML output and calculate accuracy.

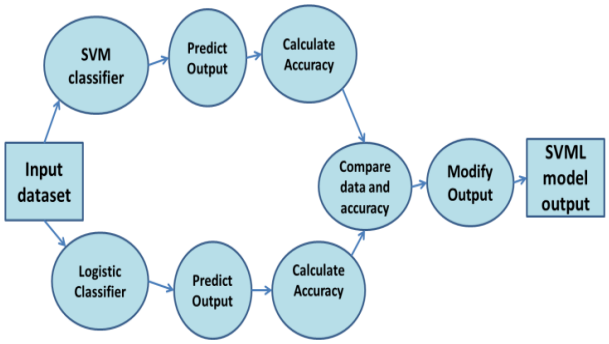

Fig. 4: Data Flow Diagram of SVML Model.

The above algorithm has been used to train the dataset and successfully and then test it by predicting the survival status of a breast cancer operated patient after a duration of five years. From data summarized in Table 5, we observe that our SVML model gives us the highest accuracy of 85%, while SVM gives an accuracy of 78% and Logistic Regression of 72%. The classified output of the SVML model has been plot in the graph shown in Fig. 5 and the sample output snapshot has also been shown in Fig. 6.

The accuracy of our model is also clearly visible in Fig. 5 as the bottom part represents data that has been accurately classified and the top part represents incorrectly predicted the output.

Although this model has currently been proposed only using R code we can assure that this model can be developed in any language using similar method and/or algorithm. According to our analysis, this proves to be the best method so far to test data and predict information in terms of accuracy. We would strongly recommend the use of this model in real world applications in the future.

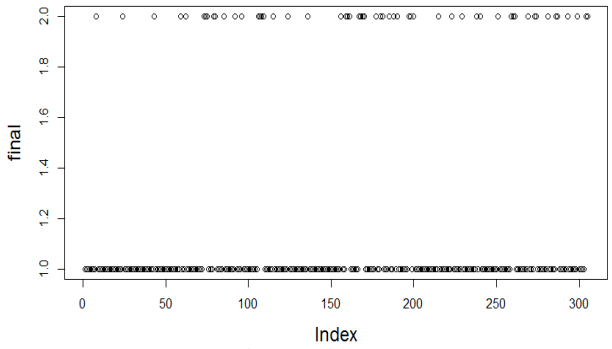

Fig. 5: Classification of the SVML Model.

```
> final_table <- table(final, X1.1)
> ans_table
   X1.1
ans   1   2
  1 175  35
  2  49  46
> pred_table
     X1.1
pred   1   2
   1 214  57
   2  10  24
> final_table
     X1.1
final   1   2
    1 214  35
    2  10  46
> pred_accuracy <- sum(diag(pred_table))/sum(pred_table)
> ans_accuracy <- sum(diag(ans_table))/sum(ans_table)
> final_accuracy <- sum(diag(final_table))/sum(final_table)
> pred_accuracy
[1] 0.7803279
> ans_accuracy
[1] 0.7245902
> final_accuracy
[1] 0.852459
```

Fig. 6: Sample Code Snippet.

4. Conclusion

Breast cancer being the most painstaking issue in the present world, we have used it to analyze our model. We have successfully applied SVM and Logistic Regression to our algorithm and then found a hybrid model of the same. We have named this model as Support Vector Machine-Logistic (SVML) and as the name suggests it will have the positive features of both Support Vector Machine and Logistic Regression. In addition, the main feature of this algorithm is its simplified nature. This makes it easy to be used by everyone and in every field.

We can infer that SVML algorithm is better than both SVM and Logistic Regression working separately in predicting the survival rate of a person after a breast cancer operation. In terms of accuracy, our model is 85% accurate for our dataset, while SVM and Logistic Regression are 78% and 72% accurate for the same dataset respectively. With SVML algorithm, we can easily predict whether a person will survive after five years of operation when her age and number of lymph nodes are known. Not only can we use this algorithm in the field of breast cancer, we can also use it for classification and prediction of data for different scenarios. Thus, whenever we want to predict the output of a certain dataset, we can say that our SVML algorithm will give a better output as it has a better accuracy level, surpassing all the outputs predicted by SVM and Logistic Regression separately. We in future also aim and wish to test this algorithm in real life situations by collecting the required data from medical professionals. This will help our model to be more accurate, justified and will help to recognize flaws if any.

This algorithm is better than the other two algorithms we have used for our comparison due to many reasons. Firstly, it gives us a better accuracy than both of them applied individually do. Secondly, it considers and takes into account the positives aspects of both the algorithms and hence makes it a better choice for classification.

References

[1] Fung, M. Glenn, and O. L. Mangasarian. "Multicategory proximal support vector machine classifiers." Machine learning Vol. 59, No. 1-2, pp.77-97, 2005.

[2] Lantz, Brett. Machine learning with R. Packt Publishing Ltd., 2013.

[3] Furey, S. Terrence, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer and D. Haussler. "Support vector machine classification and validation of cancer tissue samples using microarray expression data." Bioinformatics Vol. 16, No. 10, pp. 906-914, 2000.

[4] Menard, Scott. Applied logistic regression analysis. Vol. 106, Sage, 2002.

[5] Tranmer, Mark, and M. Elliot. "Binary logistic regression." Cathie Marsh for census and survey research, paper, Vol. 20, 2008.

[6] Hua, Zhongsheng, Y. Wang, X. Xu, B Zhang, and L. Liang. "Predicting corporate financial distress based on integration of support vector machine and logistic regression." Expert Systems with Applications, Vol. 33, No. 2, pp. 434-440, 2007.

[7] Yilmaz, Işık. "Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: conditional probability, logistic regression, artificial neural networks, and support vector machine." Environmental Earth Sciences Vol. 61, No. 4, pp. 821-836, 2010.

[8] Hua, Zhongsheng, and B. Zhang. "A hybrid support vector machines and logistic regression approach for forecasting intermittent demand of spare parts." Applied Mathematics and Computation, Vol. 181, No. 2, pp.1035-1048, and 2006.

[9] Elbashir, M. Khalafallah, W. Jianxin, and F. Wu. "A hybrid approach of support vector machines with logistic regression for β-turn prediction." In: Proc. of IEEE International Conference on In Bioinformatics and Biomedicine Workshops (BIBMW), IEEE, pp. 587-593, 2012.

[10] Chang, Y. Ivan. "Boosting SVM classifiers with logistic regression." See www.stat.sinica.edu.tw-/library/c_tec_rep/200303.pdf, 2003.

[11] Haberman's Survival data set, T. S. Lim, UCI Machine Learning Repository, http://archive.ics-.uci.edu/ml Irvine, CA, University of California, School, 1999.

[12] Ferlay, Jacques, C. Héry, P. Autier, and R. Sankaranarayanan. "Global burden of breast cancer." In Breast cancer epidemiology Springer NewYork, pp. 1-19, 2010.

[13] Carter, L. Christine, C. Allen, and D. E. Henson. "Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases." Cancer Vol. 63, No. 1, pp. 181-187, 1989.

[14] Early Breast Cancer Trialists' Collaborative Group. "Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials." The Lancet Vol. 365, No. 9472, pp.1687-1717, 2005.

[15] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, A. Weingessel, MF Leisch. Package 'e1071'. R Software package, available at http://cran.rproj-ect.org/web/packages/e1071/index.html Jan 6, 2009.

[16] King, Gary, and L. Zeng. "Logistic regression in rare events data." Political analysis Vol.9, No. 2, pp. 137-163, 2001.

[17] Zhu, Ji and T. Hastie. "Kernel logistic regression and the import vector machine." Journal of Computational and Graphical Statistics, Vol. 14, No. 1, pp. 185-205, 2005.

[18] D. Meyer, F. T. Wien. Support vector machines. The Interface to libsvm in package e1071. Aug 5, 2015.

[19] Tong, Simon, and D. Koller. "Support vector machine active learning with applications to text classification." Journal of machine learning research, Vol. 2, pp. 45-66, Nov 2001.

[20] E. Leopold and J. Kindermann. "Text categorization with support vector machines. How to represent texts in input space?" Machine Learning Vol. 46, No. 1-3, pp. 423-444, 2002.

[21] "Learn About Lymph Node Status and Breast Cancer at Susan G. Komen." Susan G.Komen®, ww5.komen.org/BreastCancer/LymphNodeStatus.html.

[22] Awad, Y. Abu, P. Koutrakis, B. A. Coull, and J. Schwartz. "A spatio-temporal prediction model based on support vector machine regression: Ambient Black Carbon in three New England States." Environmental Research, no. 159, pp. 427-434, 2017.

[23] Hewett, E. Timothy, K. E. Webster, and W. J. Hurd. "Systematic Selection of Key Logistic Regression Variables for Risk Prediction Analyses: A Five-Factor Maximum Model." Clinical Journal of Sport Medicine, 2017.

[24] D. Meyer. "Support Vector Machines—the Interface to libsvm in package e1071."Paper available at http://cran.rproject.org/-web/packa-ges/e1071/vignettes/svmdoc.pdf., 2014.