Finding Minimal Feature Set for Cancer Data using RFE technique and SVM

Venkatesham Chintapandu, Vaishnavi Mahipathi

Department of Computer Science, Georgia State University
Atlanta, GA

vchintapandu1@student.gsu.edu

vmahipathi1@student.gsu.edu

Abstract— **Cancer is one of the leading causes of death in the world, particularly in developing countries. Hence, its early screening and accurate diagnosing becomes very necessary. In this work, we have developed and evaluated a model that uses Cancer data for screening and predicting the presence of cancerous cells using feature selection techniques. The data set was obtained from the UCI Machine Learning Repository. The model was trained and executed on a certain example set. We use a Feature Selection technique called Recursive feature elimination (RFE) to extract the principal feature items. Our method uses *eight* sensitive feature items for training where as the previous work uses ten feature items. In training step, the data set was classified using Support Vector Machines (SVM). Our results proved to be better than the experimental results obtained from previous works. We use *eight* feature items to cover major aspects of cancerous condition. Thus the practitioners can screen and confirm the presence of malignant or benign tumours in a person.**

Keywords— **Feature Selection, Classification, RFE, SVM, Malignant, Benign**

I. Introduction

Cancer is a group of conditions in which cells in the body grow, change, and multiply out of control. Cancer remains the second most common cause of death in the US, accounting for nearly 1 of every 4 deaths. [1]. A group of rapidly dividing cells may form a lump or mass of extra tissue. These masses are called tumours. Tumours can either be cancerous (malignant) or non-cancerous (benign). Malignant tumours penetrate and destroy healthy body tissue.

In the late stages of cancer, cells break through normal texture and metastasize or spread to new region in the body. The leading cause of death among women between 40 and 55 years of age is Breast cancer and the second overall cause of death among women (exceeded only by lung cancer). Luckily, amount of fatality from breast cancer has decreased in recent years with an increased insistence on early detection and more impressive treatments [5].

Correct decision taken by the physician avoids surgery but due to incorrect decision, patients suffer unnecessary surgery [6]. Proposed method uses RFE-SVM classifier that avoids errors during diagnosis and assists the doctors for breast cancer diagnosis. This technique is used to diagnosis in shorter time that avoids the cancer death.

This paper is organised as follows. Section 2 provides information regarding the prior and existing methods for Cancer diagnosis using various feature selection techniques. Section 3 gives the details of our proposed method. Section 4 summarizes the results and discussions of the experiments conducted. The conclusion and future work is discussed in Section 5.

II. Previous and Related Work

From Experimental study and surveys, it can be inferred that Data Mining techniques can be used to solve these problems efficiently with accurate results. In [2], a pattern recognition system was introduced, which helps the pathologist to identify the condition more accurately. In [3], a learning algorithm that combined logarithmic simulated annealing with the perceptron algorithm was used. 10-fold cross-validation with C4.5 decision tree method was used in [4].

Ster & Dobnikar, (1996) used LDA method using neural networks to predict the presence of malignant tumours in a body tissue [7]. In [8], the authors use Support Vector Machines for Feature Selection and Classification of breast cancer data.

In [14], the author uses MRMR technique and decision trees to screen and predict depression in adults.

The present application demonstrates that SVMs are also very effective for discovering informative features or attributes. Our techniques outperform other methods in classification performance in cancer diagnosis while selecting cells that have plausible relevance to cancer diagnosis

III. Proposed Work

In this paper, we develop a model that uses breast cancer data to screen and predict whether the body tissue contains cancerous cells. The first step is to identify the set of principal feature items using Recursive Feature Elimination. The second step is to classify the feature items using a classifier called SVM.

A. Recursive Feature Elimination

Many feature selection routines used a wrapper approach to find appropriate variables such that an algorithm that searches the feature space repeatedly fits the model with different predictor sets. The best predictor set is determined by some measure of performance (i.e. R^2, classification accuracy, etc). Recursive Feature Elimination [9] is type of search routine algorithm. .

```
Algorithm 1: Recursive Feature Elimination
1.1Train the model on the training set using all predictors.
1.2 Calculate the model performance.
1.3. Calculate the variable rankings
1.4. for each subset size $S_i$, i=1…..S do
    1.4.1 Keep the $S_i$ most important variables
    1.4.2 pre process the data if required.
    1.4.3 Train the model on the training set using $S_i$ predictors
    1.4.4 Calculate model performance
    1.4.5 Recalculate the rankings of each predictor if required
1.5 End
1.6 Calculate the performance profile over the $S_i$
1.7 Determine the appropriate number of predictors
1.8 Use the model corresponding to the optimal $S_i$
```

First, the algorithm fits the model to all predictors. Each predictor is ranked using its importance to the model. Let S be a sequence of ordered numbers which are candidate values for the number of predictors to retain (S1 > S2 ...). At each iteration of feature selection, the Si top ranked predictors are retained, the model is refit and performance is assessed. The value of Si with the best performance is determined and the top Si predictors are used to fit the final model.

B. Support Vector Machines

Support vector machines are used to analyze data and recognize patterns for classification is based on statistical machine learning theory. It constructs a d-dimensional hyper plane [10]. SVM are powerful classifiers with good performance in the field of classification. Support vector machines (SVMs) are primarily two-class classifiers and are basically binary classification algorithms. Their main purpose is their mathematical duty and geometric analysis. It investigates to find the trade off between maximizing the margin and minimizing the training set error to perform the best generalization ability and remains stable to over fitting

IV. RESULTS AND DISCUSSIONS

This section describes in details the steps carried out in our experiment and the results are presented at the end of this section.

A. Data Set Description

We have downloaded the Wisconsin breast cancer data (WBCD) that taken from the UCI machine learning repository which was obtained from the University of Wisconsin Hospitals. The dataset consisted of 569 instances with no missing values and 32 attributes including both, the sample's ID and the class label. Each sample was obtained from fine needle aspirates taken from patient's breast texture. The remaining 30 attributes were re-coded in a domain between 1 and 10 that represent nuclear features namely radius, texture, compactness, concavity, symmetry, etc.

Since the data available was raw data, we segregated the data into two distinct sets. Namely the Training data set and the training label set. The sample's ID was removed from the training data set. The class label set consisted of letters M and B denoting Malignant (cancerous cells) and benign (non cancerous cells) respectively. We replaced M and B with 1 and 2 respectively. Therefore, class label 1 corresponds to presence of cancer cells and 2 otherwise.

B. Assessment metrics

In this paper, the performance of the classifier is evaluated by *overall accuracy*, *G-Mean* and *F-Measures* [11].
The confusion matrix as shown in Table I represents the contingency table for evaluating the performance of machine learning algorithm on the classifying problems. Let *{p, n}* be the positive and negative testing examples and *{Y, N}* be the classification results given by a learning algorithm for positive and negative predictions.

In this paper, the malignant class was set as the positive class and benign class was set as the negative class

TABLE I
CONFUSION MATRIX

	Predicted Positive(Y)	Predicted Negative (N)
Actual Positive (p)	TP (True Positives)	FN (False Negatives)
Actual Negative (n)	FP (False Positives)	TN (True Negatives)

Based on Table I, the evaluation metrics used to assess classifying the data sets are defined as follows

Overall Accuracy (OA):

$$OA = \frac{TP + TN}{TP + FP + FN + TN}$$

TP Rate:

$$TPRate = \frac{TP}{TP + FN}$$

FP Rate:

$$FP\ Rate = \frac{FP}{F\,P + TN}$$

Precision:

$$Precision = \frac{TP}{TP + FP}$$

Recall:

$$Recall = \frac{TP}{TP + FN}$$

F-Measure:

$$F\text{-}Measure = \frac{(1 + \beta) \times Recall \times Precision}{\beta^2 \times Recall + Precision}$$

Where β is a coefficient to adjust the relative importance of precision versus the recall (usually β = 1). F-Measure is high when both the recall and precision are high. It indicates that the F-Measure is the measure of goodness of a learning algorithm on the interest class.

$$G\text{-}Mean = \sqrt{\frac{TP}{TP+FN} X \frac{TN}{TN+FP}}$$

G-Mean is used to evaluate a performance of classifier on the skew data. The term $\frac{TP}{TP+FN}$ is called the *Positive Accuracy* and the term $\frac{TN}{TN+FP}$ is called the *Negative Accuracy.* The idea of G-Mean is to maximise the accuracy on both, the Cancer affected and non affected classes.

C. Feature Ranking

For evaluation, the attributes are ranked as per the priority using RFE. Out of 30 features, we eliminated 29 features to find that attribute with highest priority using Recursive feature elimination. To find the top two features, we eliminated 28 features using RFE. This process was repeated in order to rank the features in the order of their priority. The rank wise priority corresponding to the feature set is listed in Table III.

D. Analysis of Results:

.

The training data set has 30 features. Our aim was to find the minimal feature set with maximum prediction accuracy. We started with all 30 features to assess the classification accuracy. We ended it with one feature to assess the classification accuracy. To calculate the prediction accuracy, we followed 10-fold cross validation. The training data set is divided into 10 sets. Each time one set was used for testing and the remaining nine sets were used for training. This is repeated for all the ten sets.

We have used RFE to select the top 29 features. Once the 29 features were selected, we applied 10-fold cross validation method to predict the class labels using SVM classifier. After the prediction was completed, we calculated all the assessment metrics mentioned in section B. Next, we selected 28 features, applied 10-fold cross validation method to predict the class labels using SVM classifier and then all the assessment metrics were calculated. This process is repeated for 27, 26…1 feature(s).

Recursive feature elimination (with SVM) algorithm was written in python programming language. We have written MATLAB programs to predict the class labels using SVM and calculate corresponding assessment metrics mentioned in section B.

Table II shows the results for the values obtained after calculating the assessment metrics for the features. Firstly, we calculate the F-Measure, G-Mean and Overall Accuracy rate considering all the 30 features. We can see that the values in percentage for F-Measure, G-Mean and Overall Accuracy corresponding to 30 feature items are 96.6%, 96.27% and 97.54% respectively. We can see that these values are maximum when we consider all the 30m features. Similarly we calculate these metrics by decreasing the number of features step wise. For 28 features, the values are 96.45%, 97.13% and 97.36%. It shows that the accuracy reduces as the number of feature sets decreases. We continue to calculate the metrics for all the feature sets considered. When the number of features considered is 10, we can see that the F-Measure obtained is 92.79%, G-Mean is 93.93% and overall accuracy is 94.73%.

Our aim is to minimise the number of features considered and yet maximise the above mentioned metrics. G-Mean evaluates the performance of both the affected classes and the non affected classes. It can be seen that these values are maximum for 30 features. It can be observed that selecting 8 features also gives an optimal result where the F-Measure, G-Mean and Overall Accuracy values are as follows: 93.3%, 94.42% and 95.08%.

Table III describes the Number of features considered and the corresponding features selected according to the priority. The RFE algorithm ranks the model considered according to the importance. After each iteration, the top ranked predictor is selected and the feature with least importance is omitted. To find the feature with highest priority, we have applied RFE to select one feature. It is observed that the top ranked feature is feature number 28. When the number of features selected is 2, the first highest ranked feature number is 28 and the second highest ranked feature number is 8. We repeated this procedure for all the 30 features. When the number of features selected is 8, the features considered are 28, 8, 17,16,27,29,6,11 in the order of their priority, 28 being the highest and 11 being the lowest.

Fig.1 is a simple line graph that shows the overall accuracy values for the number of features selected. It is observed that the Overall accuracy is maximum of 97.54% for 30 features and minimum of 92.44% for 1 feature selected. For 8 features, we obtain the overall accuracy rate of 95.08%.

Similarly, Fig.2 shows the G-Mean values for the number of feature items. G-Mean for 8 feature items is 94.42%. This gives accuracy for both affected and non affected classes.

Fig.3 is a line graph showing the F-Measure and features. The F-Measure value for 8 features is observed to be 93.3%.

TABLE II RESULTS OF PREDICTING CANCER CELLS:
Number of Attributes, TP Rate, FP Rate, Recall, F-Measure, G-Mean, Overall Accuracy

Number of Attributes	TP Rate	FP Rate	Precision	Recall	F-Measure	G-Mean	Overall Accuracy
30	0.9623	0.0168	0.9714	0.9623	0.9668	0.9727	0.9754
28	0.9623	0.0916	0.9668	0.9623	0.9645	0.9713	0.9736
25	0.9623	0.0916	0.9668	0.9623	0.9645	0.9713	0.9736
21	0.9481	0.0224	0.9617	0.9481	0.9549	0.9627	0.9666
18	0.9292	0.0168	0.9704	0.9292	0.9494	0.9558	0.9631
14	0.9151	0.028	0.951	0.9151	0.9327	0.9431	0.9508
10	0.9104	0.0308	0.9461	0.9104	0.9279	0.9393	0.9473
8	0.9198	0.0308	0.9446	0.9198	0.933	0.9442	0.9508
5	0.8868	0.0532	0.9082	0.8868	0.8974	0.9163	0.9244

TABLE III RANKING OF FEATURES

#of Attributes Considered	Ranking of the Features
1	28
2	28,8
3	28, 8,17
4	28, 8,17,16
5	28, 8,17,16,27
6	28, 8,17,16,27,29
7	28, 8,17,16,27,29,6
8	28, 8,17,16,27,29,6,11
9	28, 8,17,16,27,29,6,11,9
10	28, 8,17,16,27,29,6,11,9,7
11	28, 8,17,16,27,29,6,11,9,7,25
12	28, 8,17,16,27,29,6,11,9,7,25,5
13	28, 8,17,16,27,29,6,11,9,7,25,5,4
14	28, 8,17,16,27,29,6,11,9,7,25,5,4,30

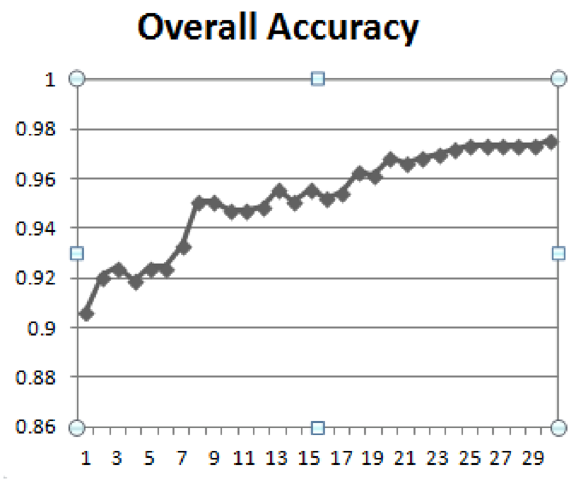

Fig.1 Number of features Vs Overall Accuracy

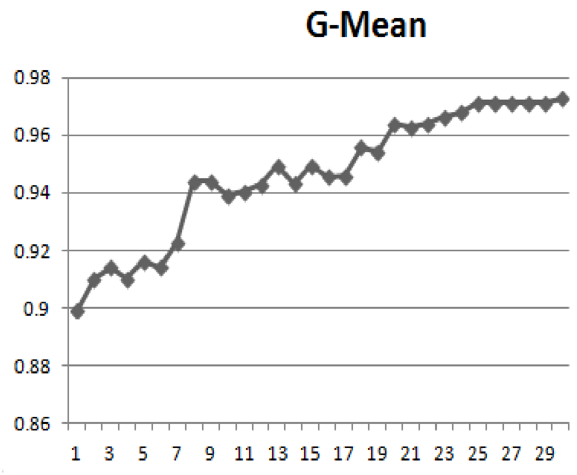

Fig.2 Number of features Vs G-Mean

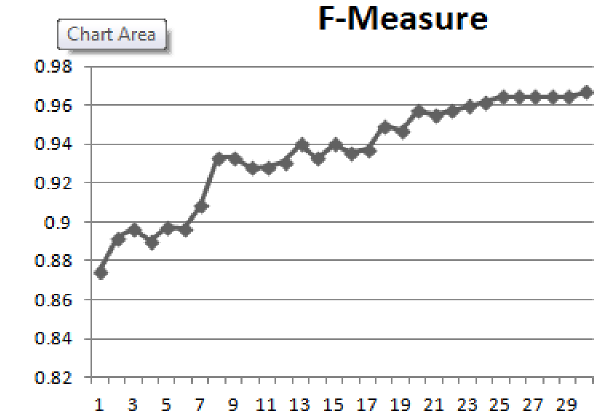

Fig.2 Number of features Vs F-Measure

ACKNOWLEDGMENT

We would like to thank UCI Repository site for the data. We would also like to thank our professor Dr. Yanqing Zhang, Georgia State University and his Ph.D student Ma, Long, Georgia State University for helping us in carrying out this experiment.

V. CONCLUSIONS AND FUTURE WORK

This paper developed a hybrid method for screening and predicting the presence of cancerous cells in body tissue using the feature selection technique called Recursive Feature Elimination and classifier called Support Vector Machine. RFE and SVM are used together to select the sensitive features. Using this method, we identified that selecting 8 features from the original set of 30 features gives an optimal result. The overall accuracy is seen to be 95.08%. Our method has proved to be very effective for discovering informative features or attributes. Our techniques outperform other methods in classification performance in cancer diagnosis while selecting cells that have plausible relevance to cancer diagnosis. Using SVM along with RFE has increased the accuracy results. Using RFE and SVM not only achieved higher accuracy for affected cases but also proved equally accurate for non affected cases as well.

Furthermore, the proposed results are particularly useful for screening Cancer in the patients and help the practitioners observe symptoms clearly. Especially, this method can reduce time of diagnosis.

However, in this method, one cannot predict the severity of Cancer. Modifying the existing Recursive Feature Algorithm may yield in better prediction accuracy. We will be working on modifying this algorithm and combining it with other feature extraction techniques to get better accuracy in future.

REFERENCES

[1] All Cancer Facts and Figures [Online] : Available at : http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2014/index

[2] Albrecht, A. A., Lappas, G., Vinterbo, S. A., Wong, C. K., & OhnoMachado, L. (2002). Two applications of the LSA machine. In Proceedings of the 9th international conference on neural information processing (pp. 184–189).

[3] Hamiton, H. J., Shan, N., & Cercone, N. (1996). RIAC: A rule induction algorithm based on approximate classification. Technical Report CS 96-06, University of Regina.

[4] V. E. Ekong, U. G. Inyang, and E. A. Onibere, "Intelligent decision support system for depression diagnosis based on neuro-fuzzy-cbr *hybrid," Modern Applied Science, vol. 6, no. 7, pp. 79–88, 2012*. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in *Proc. ECOC'00*, 2000, paper 11.3.4, p. 109.

[5] "Feature Subset Selection and Parameters Optimization for Support Vector Machine in Breast Cancer Diagnosis" , Elnaz Olfati, Hassan Zarabadipour, Mahdi Aliyari Shoorehdeli

[6] Mehmet F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis", Elsevier, Expert Systems with Applications, vol. 36, no. 2, pp. 3240- 3247, 2009*FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.

[7] Ster and Dobnikar. "Neural networks in medical diagnosis,Comparison with other methods". In proceedings of the international conference on engineering applications of neural networks (pp. 427–430), 1996.
.

[8] Santi Wulan Purnami and S.P. Rahayu, "Feature selection and classification of breast cancer diagnosis based on support vector machines "..

[9] Isabelle Guyon, Jason Weston, " Gene Selection for Cancer Classification using Support Vector Machines ," 2002.

[10] Vapnik, V. Statistical Learning Theory. 3rd ed. New York, Wiley Interscience publication, 1989..

[11] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, pp. 1263–1284,2009

[13] S. Chen, H. He, and E. A. Garcia, "RAMO Boost: Ranked minority oversampling in boosting," *IEEE Transactions on Neural Networks* vol. 21,pp.1624–1642,2010.

[14] Putthiporn Thanathamathee, "Boosting with Feature Selection Technique for Screening and Predicting Adolescents Depression",2013