AN EFFICIENT SVM FRAMEWORK TUNED FOR THYROID CLASSIFICATION

Maddi Kamal Manisha, Pedapudi Venkata Sai Vijay Kumar, Monisha Prasad, Akshaya Balamurugan

National University of Singapore

ABSTRACT

Professor Lee Kok Onn, the head of the division of endocrinology at National University Hospital said Grave's disease, which is an autoimmune disorder, produces abnormal antibodies which stimulate the thyroid gland, causing an overproduction of hormones. It is estimated to affect about 1.28 per cent of people in Singapore which costs a lot to the government. A timely decision made can save the costs incurring to the government.

Each time a patient visits a hospital, the patient may get different opinions from different doctors about the same problem. There is no data-driven or evidential decision-making process in the sphere of health. Hence, a novel approach is proposed to help the doctors arrive at a proper conclusion about the patient's condition using support vector machines for classifying the patient condition.

1. INTRODUCTION

The thyroid gland, which sits in front of the trachea or windpipe in the throat, is designed to secrete thyroid hormones. Thyroid hormones are involved in regulating just about every bodily function, including heart rate, blood circulation, body temperature, brain, skin, bowels and fluid balance which gives the importance of thyroid in human body.

The following statistics attempts to extrapolate the above prevalence rate for Thyroid disorders to the populations of various countries and regions. Extrapolated Prevalence of Thyroid disorders in Singapore 320,139 and Extrapolated Statistics was found to be 4,353,893. These prevalence extrapolations for Thyroid disorders are only estimates, based on applying the prevalence rates.

A study published in Hormone and Metabolic Research in 2011 found that children and siblings of autoimmune thyroid diseases sufferers have a 16-fold and 15-fold- increased risk of developing the conditions, respectively. Studies have also shown women are five times more likely to develop thyroid disorders, especially those aged between 20 and 50 which is alarming.

Top 10 cancers affecting Singapore women

Site	Ranking	No.
Breast	1	9,634
Colo-rectum	2	4,424
Lung	3	2,489
Corpus Uteri	4	2,271
Ovary	5	1,797
Lymphoma	6	1,470
Skin (Incl. Melanoma)	7	1,404
Thyroid	8	1,269
Stomach	9	1,117
Cervix Uteri	10	1,037

The disease burden of Singapore ministry of health also evidences this fact in above shown picture, by Singapore Cancer Registry, Annual Registry Report 2015.

The dataset is taken from UCI machine learning repository which has 22 variables with 7201 instances about patients examined for thyroid disease collected by Garavan Institute in Sydney, Australia.

The data contains few Boolean and few continuous values. All the continuous values are scaled between 0 to 1. Also, there are no missing values in the dataset. Hence, it does not require any preprocessing.

The attributes age, sex, pregnant, sick, lithium, psych etc can be understood from their names. The attributes which are ambiguous are explained as below:

I131_treatment - radioactive iodine therapy for hyperthyroid
Goitre - patient has goitre or not
Tumor - patient has tumor or not
Hypopituitary - pituitary gland condition
TSH - Thyroid Stimulating Hormone Test

T3 - Triiodothyronine
TT4 - Total T4/ Total Thyroxine
T4U - Thyroxine utilization rate
FTI - Thyroid Function Tests

Among the 21 predictor variables, 6 are continuous variables and the rest are logical as shown below:

1 age: continuous	2 sex: {M, F} logical	3 on thyroxine: logical
4 maybe on thyroxine: logical	5 on antithyroid medication: logical	6 sick - patient reports malaise: logical
7 pregnant: logical	8 thyroid surgery: logical	9 I131 treatment: logical
10 test hypothyroid: logical	11 test hyperthyroid: logical	12 on lithium: logical
13 has goitre: logical	14 has tumor: logical	15 hypopituitary: logical
16 psychological symptoms: logical	17 TSH: continuous	18 T3: continuous
19 TT4: continuous	20 T4U: continuous	21 FTI: continuous

The target variable 'Functioning of thyroid glands' has 3 classes as follows:

Class	**Description**
1	Normal
2	Hyperthyroid - Hyperfunctioning
3	Hypothyroid - Subnormal functioning

2. BASELINE APPROACH

The multiclass classification is done using Support Vector Machines to classify the variants of thyroid gland functioning. This baseline approach sets a very basic benchmark above which many improvements can be done. Thus, if classification made at an early point is aided by a recommendation system, the risk of getting severely affected can be reduced and the patient can be treated at an early stage without complications.

2.1. Simple SVM Classifier

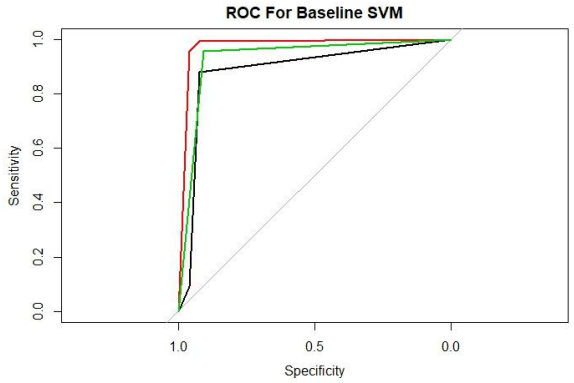
The basic classification is performed on the dataset using SVM function in R. Here we do not specify any parameters and no tuning is done. The accuracy is as shown below:

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3
         1   24    1    1
         2    4   26    3
         3    9   51 1321

Overall Statistics

               Accuracy : 0.9521
                 95% CI : (0.9397, 0.9625)
    No Information Rate : 0.9201
    P-Value [Acc > NIR] : 1.106e-06

                  Kappa : 0.5864
 Mcnemar's Test P-Value : 5.223e-11
```

Class one indicates that the condition of the patient is normal. It is well justified in the data as well since the patients who suspect to have thyroid gland problems are ought to take tests which obviously turned out to be negative. Also, class two indicates Hyperthyroidism and its prevalence is relatively less compared to class three Hypothyroidism category.

```
Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity           0.64865  0.33333   0.9970
Specificity           0.99857  0.99486   0.4783
Pos Pred Value        0.92308  0.78788   0.9566
Neg Pred Value        0.99081  0.96304   0.9322
Prevalence            0.02569  0.05417   0.9201
Detection Rate        0.01667  0.01806   0.9174
Detection Prevalence  0.01806  0.02292   0.9590
Balanced Accuracy     0.82361  0.66410   0.7376
```

The prevalence and balanced accuracy of class one justifies its sensitivity.

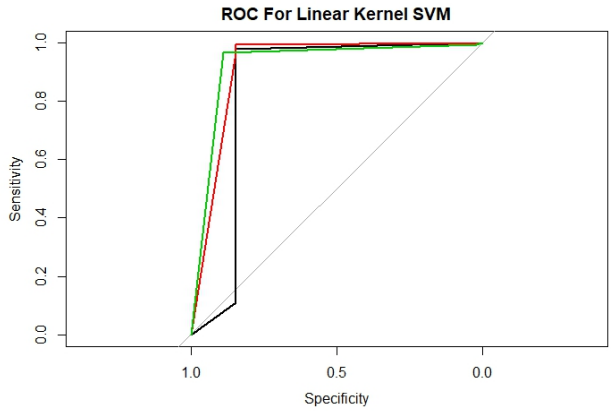


The Multi class Area Under the Curve for the baseline approach of SVM is 0.932 and it has a good accuracy of 0.9521. But, this classification is about a patient's health and thus risk factor should be reduced. There by, a fine-tuned model of the above approach with a better accuracy is required.

2.2. Model 1 - Linear Kernel

The model one is fitted with tuning parameters like a linear kernel and cost = 1. The results are as follows:

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3
         1   28    0    5
         2    1   39    5
         3    8   39 1315

Overall Statistics

               Accuracy : 0.9597
                 95% CI : (0.9482, 0.9693)
    No Information Rate : 0.9201
    P-Value [Acc > NIR] : 9.326e-10

                  Kappa : 0.6839
 Mcnemar's Test P-Value : 3.694e-06

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity           0.75676  0.50000   0.9925
Specificity           0.99644  0.99559   0.5913
Pos Pred Value        0.84848  0.86667   0.9655
Neg Pred Value        0.99360  0.97204   0.8718
Prevalence            0.02569  0.05417   0.9201
Detection Rate        0.01944  0.02708   0.9132
Detection Prevalence  0.02292  0.03125   0.9458
Balanced Accuracy     0.87660  0.74780   0.7919
```



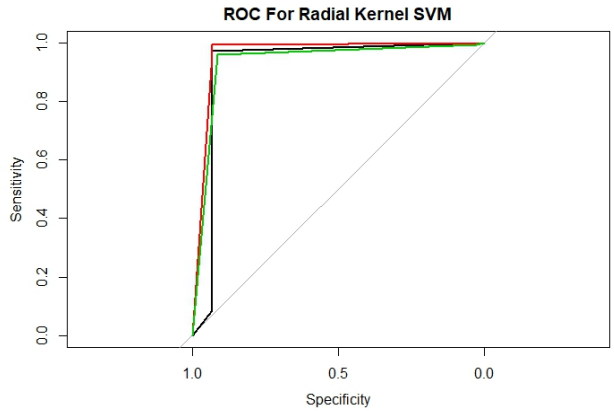
The Multiclass Area Under the Curve for this approach of SVM is 0.8972 and it has an accuracy of 0.9597

2.3. Model 2 - Radial Kernel

The model two is fitted with radial kernel with cost = 2 and sigma = 0.3. The results are as follows:

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3
         1   28    0    2
         2    1   31    3
         3    8   47 1320

Overall Statistics

               Accuracy : 0.9576
                 95% CI : (0.9459, 0.9674)
    No Information Rate : 0.9201
    P-Value [Acc > NIR] : 7.813e-09

                  Kappa : 0.6456
 Mcnemar's Test P-Value : 2.105e-09


Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity           0.75676  0.39744   0.9962
Specificity           0.99857  0.99706   0.5217
Pos Pred Value        0.93333  0.88571   0.9600
Neg Pred Value        0.99362  0.96655   0.9231
Prevalence            0.02569  0.05417   0.9201
Detection Rate        0.01944  0.02153   0.9167
Detection Prevalence  0.02083  0.02431   0.9549
Balanced Accuracy     0.87767  0.69725   0.7590
```

The Multiclass Area Under the Curve for this approach of SVM is 0.9402 and it has an accuracy of 0.9576



Looking at the values of both linear kernel and radial kernel approach it is worthy to note the closeness in terms of accuracies and slight variations in Multiclass Area Under the Curve. This led to the proposed scheme covering the other optimal solutions.

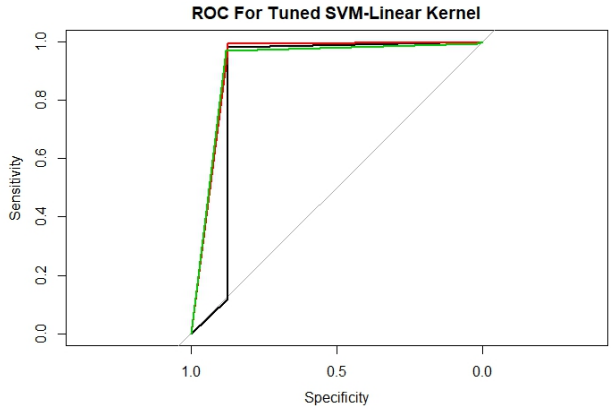
3. PROPOSED APPROACH

A fine-tuned model always gives better performance and it is a suggestable approach, especially for a sphere like health where every minor parameter is critical in evaluating the patient condition. After undergoing multiple iterations of validations, the model is presented below briefly.

3.1. Optimizing Linear Kernel Model

Here the linear kernel is chosen for the possible fine tuning with tuneGrid. Cross validation is also performed on the data. The tuneGrid has chosen the best value as C = 2

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3
         1   28    0    4
         2    1   44    6
         3    8   34 1315

Overall Statistics

               Accuracy : 0.9632
                 95% CI : (0.9521, 0.9723)
    No Information Rate : 0.9201
    P-Value [Acc > NIR] : 1.891e-11

                  Kappa : 0.7178
 Mcnemar's Test P-Value : 6.735e-05

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity           0.75676  0.56410   0.9925
Specificity           0.99715  0.99486   0.6348
Pos Pred Value        0.87500  0.86275   0.9690
Neg Pred Value        0.99361  0.97552   0.8795
Prevalence            0.02569  0.05417   0.9201
Detection Rate        0.01944  0.03056   0.9132
Detection Prevalence  0.02222  0.03542   0.9424
Balanced Accuracy     0.87695  0.77948   0.8136
```

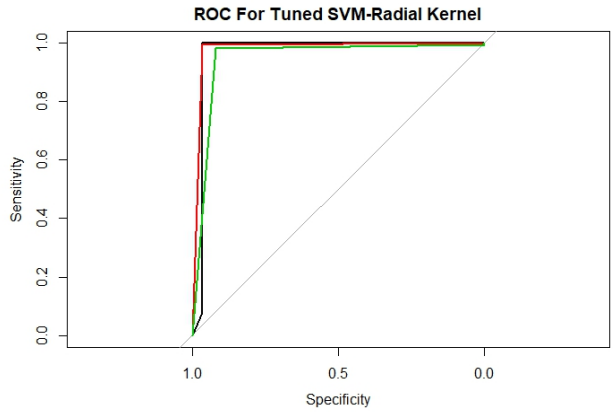


The Multiclass Area Under the Curve for this approach of SVM is 0.9101 and it has an accuracy of 0.9632

3.2. Optimizing Radial Kernel Model

Here the radial kernel is chosen for the possible fine tuning with tuneGrid. Cross validation is also performed on the data. The tuneGrid has chosen the best values as Sigma = 0.03125 and C = 32

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3
         1   29    0    1
         2    0   59    5
         3    8   19 1319

Overall Statistics

               Accuracy : 0.9771
                 95% CI : (0.968, 0.9842)
    No Information Rate : 0.9201
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8327
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity           0.78378  0.75641   0.9955
Specificity           0.99929  0.99633   0.7652
Pos Pred Value        0.96667  0.92187   0.9799
Neg Pred Value        0.99433  0.98619   0.9362
Prevalence            0.02569  0.05417   0.9201
Detection Rate        0.02014  0.04097   0.9160
Detection Prevalence  0.02083  0.04444   0.9347
Balanced Accuracy     0.89154  0.87637   0.8803
```



The Multi class Area Under Curve for this approach of SVM is 0.9654 and it has an accuracy of 0.9771

4. EXPERIMENTAL RESULTS

The comparison of various models taken as baseline approach and proposed approach is depicted below.

4.1. Validation Quantities Results

Model	Name of the approach	Accuracy	AUC
1	Simple SVM Classifier	0.9521	0.932
2	Linear Kernel SVM Model	0.9597	0.8972
3	Radial Kernel SVM Model	0.9576	0.9402
4	Optimized Linear Kernel SVM	0.9632	0.9101
5	Optimized Radial Kernel SVM	0.9771	0.9654

The table shows that the accuracy has improved significantly after fine tuning the hyperparameters in every SVM model.

5. CONCLUSION

The decision made at a right time can always a save a life. It is highly recommended to classify the different variants of thyroid gland functioning with high precision to provide the right treatment. Having said that SVM being a robust classifier is taken for devising a model for this case. Different models are trained to arrive at the proposed solution. Many trial runs are thus done for coming up with the final optimal approach using radial kernel which gives an accuracy of 97.71%.

REFERENCES

[1] Singapore Cancer Registry, Annual Registry Report 2015 https://www.nccs.com.sg/patientcare/whatiscancer/cancerStatistics/Pages/Home.aspx

[2] UCI Machine Learning Repository: Thyroid Disease DataSet https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease

[3] Thyroid Diagnosis based Technique on Rough Sets with Modified Similarity Relation https://pdfs.semanticscholar.org/535b/c1bf2e21ad1bb542cc0bf72ec036eeeb579e.pdf

[4] Efficient optimization of Support Vector Machines: https://www.sciencedirect.com/science/article/pii/S0377042705005856

[5] Parameter tuning of functions using grid search: http://ugrad.stat.ubc.ca/R/library/e1071/html/tune.html

[6] Nguyen H. Son and Skowron A.: Quantization of real value attributes. | Proc. Int. Workshop Rough Sets and Soft Computing at 2nd Joint Conf. Information Sciences (JCIS'95), Durham, NC, pp.34-37, (1995).

[7] Li-Na Li & Ji-Hong Ouyang & Hui-Ling Chen & Da-You Liu," A Computer Aided Diagnosis System for Thyroid Disease Using Extreme Learning Machine", J Med Syst, vol. 36 no. 5, pp. 3327-3337, 2012.

[8] Nesma Ibrahim, Taher Hamza, Elsayd Radwan, "An Evolutionary Machine Learning Algorithm for Classifying Thyroid Diseases Diagnoses, "Egyptian Computer Science Journal, vol.35, no. 1, pp.73- 86,Jan 2011.

[9] Ankit Gupta, Kishan G. Mehrotra,Chilukuri Mohan,"A Clustering-Based Discretization for Supervised Learning, "Statistics and Probability Letters, vol.80, pp.816-824, May 2010.

[10] Keles, A., "ESTDD: Expert System For Thyroid Diseases Diagnosis", Expert Syst. Appl. vol.34 no. 1,pp. 242–246, 2008.