KE5206 CI1, CA2 programming assignment submission

PREPARED BY:

CHAN SU-WEN, PHILEMON A0110588E
LIANG SHIZE A0178178M
SANCHIT MITTAL A0178507W
TERESA CHENG SIEW LOON A0178510H

ABSTRACT

Dielectric spectroscopy is a method of measuring the dielectric properties of a substance while varying frequency. It is principled on the interaction of an external field with the electric dipole moment of the sample and is quantified and termed as permittivity. [1]

This study seeks to determine if classification of electrical impedance spectroscopy measurements of breast tissue is feasible for testing cancerous tissue.

Different support vector machine (SVM) algorithms utilizing kernel functions such as linear, radial basis function and polynomial were used for pattern classification. Based on the confusion matrix, the SVM with polynomial kernel function and radial basis function achieves the best classification accuracy of 0.9375.

1. INTRODUCTION

Various breast cancer prediction models have been known to make use of machine learning and statistical techniques. We have taken the data set from https://archive.ics.uci.edu/ml/datasets/Breast+Tissue#.

To diagnose breast cancer, a tissue biopsy is obtained from the patient. The suspected tissue sample that is removed from the patient is sent for histological and chemical analysis. The disadvantage of this procedure is that the waiting time is too long, and the patient must wait 1 to 2 days for the results. Also, they will often need to go for a second biopsy as the results from their first biopsy can be inconclusive. A faster method is to do a tissue analysis to determine if the patient requires further screening. In this way, the medical expenses incurred from screening and the anxiety of the patient waiting for analysis results can be abated.

We will be using the data set to predict 6 classes of breast tissue, they are as follow:

Short Name	Full Name		# of cases
Car	Carcinoma		21
Fad	Fibro-adenoma		15
Mas	Mastopathy		18
Gla	Glandular		16
Con	Connective		14
Adi	Adipose		22
			106

To building the SVM classifier, we must decide on the kernel function as different kernel functions can result in different predictions and model performance.

Fibroadenoma is a term that medical practitioners use to describe a broad range of solid, benign breast lesions that commonly effect premenopausal women.

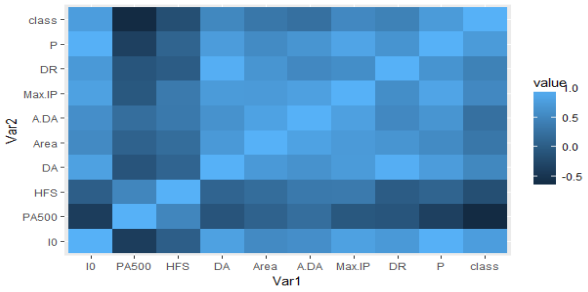
2. BASELINE APPROACH

2.1. Data Exploration

9 Independent Variables:

Short Name	Description
I0	Impedivity (ohm) at zero frequency
PA500	phase angle at 500 KHz
HFS	high-frequency slope of phase angle
DA	impedance distance between spectral ends
AREA	area under spectrum
A/DA	area normalized by DA
MAX IP	maximum of the spectrum
DR	distance between I0 and real part of the maximum frequency point
P	length of the spectral curve

The correlation heatmap shows high correlation between DA & DR and I0 (using R).



Analysis of Covariance (Using R)

I0, PA500, A.DA, DR and P are statistically significant variables.

```
summary(fit)
             Df Sum Sq Mean Sq F value Pr(>F)
I0            1  82.04   82.04 636.449 <2e-16 ***
PA500         1  13.07   13.07 101.432 <2e-16 ***
HFS           1   0.20    0.20   1.582 0.2116
DA            1   0.39    0.39   3.058 0.0835 .
Area          1   0.44    0.44   3.430 0.0671 .
A.DA          1   0.53    0.53   4.129 0.0449 *
Max.IP        1   0.01    0.01   0.086 0.7703
DR            1   0.21    0.21   1.616 0.2067
P             1   0.81    0.81   6.283 0.0139 *
Residuals    96  12.37    0.13
```

1.2 Data transformation and Feature Engineering

Target variables were transformed into numeric form. Fad, Mas and Gla were merged into a singular class based on domain knowledge as their discrimination is not important.

Short Name	Full Name	Transformed Variable	# of cases
Car	Carcinoma	0	21
Fad	Fibro-adenoma	1	15
Mas	Mastopathy	1	18
Gla	Glandular	1	16
Con	Connective	2	14
Adi	Adipose	3	22
			106

70% of the data set is used for training and 30% of the data is used for testing.

1.3 SVM

The formulation of SVM adheres to the Structural Risk Minimization (SRM) principle and promised to be better [2] than traditional Empirical Risk Minimization (ERM) principle, as utilized in traditional neural networks. In the case of SRM, upper bound of the expected risk is minimized. On the other hand, ERM reduces and minimizes the training data error. This key difference enables SVM to better generalize the dataset. SVMs were strategically developed to solve classification problems. Currently, SVMs have been modified to tackle regression problems [3].

How a linear kernel based SVM works is that it maps the nonlinear input space to a new linearly separable space. More specifically, the set of vectors residing on one side of the hyperplane are labelled as -1, and all vectors residing on the other side of the hyperplane are labelled as +1. The training instances that appear nearest to the hyperplane in the transformed space are called support vectors. The proportion of support vectors is small compared to the training set size and they dictate the hyperplane margin, and thus the decision surface.

Polynomial and Gaussian Radial Basis Function (RBF) kernel functions are the commonly used functions in such problems and their formulas are shown subsequently.

1.4 TRAINING SVM MODEL USING R

Th Caret package provides train() method for training our models based on the training data. We must input different parameter values for different algorithms. Prior to using the train method, we will first use trainControl method. It controls the computational subtleties of the train method.

We are setting 3 parameters being fed into trainControl method. The "method" parameter holds the details about resampling method. We can use many methods such as "boot", "boot632", "cv", "repeatedcv", "LOOCV" and "LGOCV". In this assignment, we are using repeated cross-validation. We repeated the k-fold Cross Validation process by splitting the data into k-folds repeatedly. Finally, the resultant model accuracy is calculated as the mean from the number of repeats. We set number of folds at 10 and number of repetitions at 3.

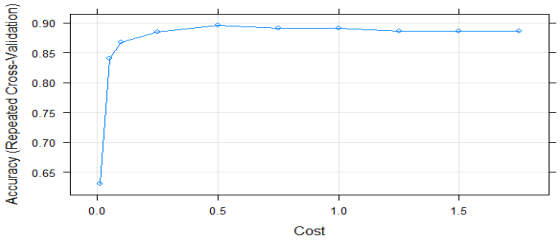
Furthermore, we passed 2 values into our pre-process parameter "center" & "scale". These two parameters will help to center and scale the data. After transformation, our training data will have a mean value at approximately "0" and standard deviation of "1". The "tuneLength" parameter is an integer value and is used for tuning our algorithm. See results in appendix (table 1)

1.5 Linear SVM

We achieved an accuracy rate of 0.9375 on the training data (refer to Confusion Matrix 1 in the appendix)

Assumption:
We are assuming linear separability.

After tuning of soft Margin, we found the best value for the model is c at 0.5 (refer to Figure 1)



3. PROPOSED APPROACH

3.1 Kernel Function

Kernels functions which allow the SVM to classify features that are nonlinear functions of the training vector attributes. In this section, we will be using polynomial and radical basis kernel function in our proposed approach.

Polynomial: A polynomial mapping is a popular method for non-linear modelling. The second kernel is usually preferable has it addresses the problem of a vanishing hessian.

$K(x,x') = <x, x'>^d$

$K(x,x') = (<x, x'> + 1)^d$

Gaussian Radial Basis Function: Radial basis functions most commonly with a Gaussian form

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

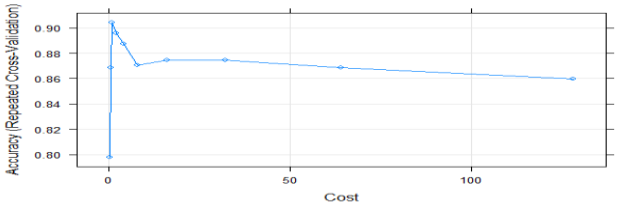
As we increase the order of the polynomial kernel, the size of the function class increases. An n-th order polynomial kernel gives us all analytic functions whose derivatives of order (n+1) are constant, and hence all derivatives of and

above order (n+2) are zero. The squared exponential kernel gives you access to all analytic functions (that is all infinitely differentiable functions).
The squared exponential kernel is nonparametric, while the polynomial kernels model is parametric. A nonparametric model works well on a complex data model. In contrast a parametric model's size is fixed, so it works on a simple data model.
However, a squared exponential kernel is generally more flexible than the linear or polynomial kernels which allows more functions with its function space.

4. EXPERIMENTAL RESULTS

4.1 Radical Basis Function

Next, we will build a model using Non-Linear Kernel like Radial Basis Function and Polynomial kernel function. For using RBF kernel, we change our train method parameter to "svmRadial". In Radial kernel, we set the value of Cost to "C" parameter and "sigma" .
It's showing that final sigma parameter's value is 0.3947888 & C parameter's value as 1 (refer to training data set 2 for more information). Our model's accuracy on our test set as follow:



Confusion Matrix

Accuracy : 0.9375
95% CI : (0.7919, 0.9923)
No Information Rate : 0.5625
P-Value [Acc > NIR] : 3.289e-06

Kappa : 0.8991
Mcnemar's Test P-Value : NA

Statistics by Class:

	Class: 0	Class: 1	Class: 2	Class: 3
Sensitivity	0.8000	0.9444	1.000	1.0000
Specificity	0.9630	0.9286	1.000	1.0000
Pos Pred Value	0.8000	0.9444	1.000	1.0000
Neg Pred Value	0.9630	0.9286	1.000	1.0000
Prevalence	0.1562	0.5625	0.125	0.1562
Detection Rate	0.1250	0.5312	0.125	0.1562
Detection Prevalence	0.1562	0.5625	0.125	0.1562
Balanced Accuracy	0.8815	0.9365	1.000	1.0000

We achieved an accuracy of about 94% using the radical basis function.

Grid-Search Optimization

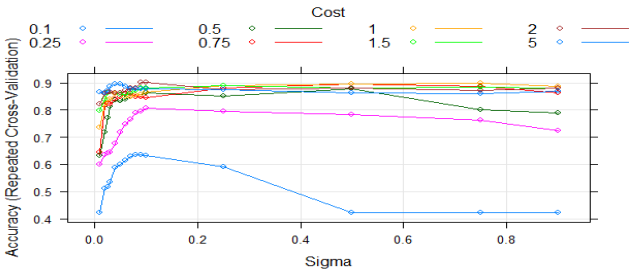
We used grid-search methods for optimizing parameters in each kernel function. First, the experiments were completed with package e1071 in R studio software.

Following is our tuning parameters:

Setting sigma at 0.01, 0.02, 0.025, 0.03, 0.04,0.05, 0.06, 0.07,0.08, 0.09, 0.1, 0.25, 0.5, 0.75,0.9 and cost function, c at 0.1, 0.25, 0.5, 0.75,1, 1.5, 2,5. The optimal model is when sigma = 0.1 and C = 2.

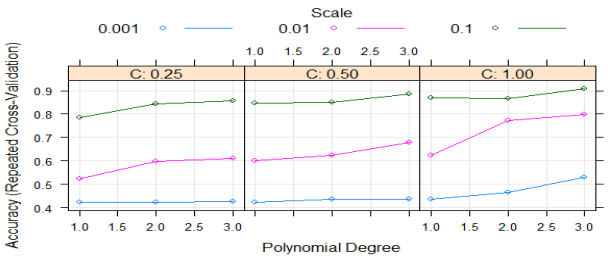
Refer to table 2 in appendix for more information.

Accuracy was utilised to select the optimal model using the largest value.
The final values used for the model were sigma = 0.1 and C = 2.



4.2 Polynomial Kernel

Accuracy was used to select the optimal model using the largest value. The final values used for the model were degree = 3, scale = 0.1 and C = 1 (refer to figure 4 for more details).



Confusion Matrix
Accuracy : 0.9375
95% CI : (0.7919, 0.9923)
No Information Rate : 0.5625
P-Value [Acc > NIR] : 3.289e-06

Kappa : 0.903
Mcnemar's Test P-Value : NA

Statistics by Class:

	Class: 0	Class: 1	Class: 2	Class: 3
Sensitivity	1.0000	0.8889	1.000	1.0000
Specificity	0.9259	1.0000	1.000	1.0000
Pos Pred Value	0.7143	1.0000	1.000	1.0000
Neg Pred Value	1.0000	0.8750	1.000	1.0000
Prevalence	0.1562	0.5625	0.125	0.1562
Detection Rate	0.1562	0.5000	0.125	0.1562
Detection Prevalence	0.2188	0.5000	0.125	0.1562
Balanced Accuracy	0.9630	0.9444	1.000	1.0000

We achieved near to 94% accuracy using polynomial kernel function.
We didn't apply any 'ensemble' methods in this assignment since both RBF and polynomial functions are us the same accuracy rate.

5. CONCLUSIONS

In this report, we utilized the SVM to recognize the breast tissue classification patterns which showed promising performance. Furthermore, the experimental results showed that the SVM with polynomial kernel function and RBF achieved the best classification accuracy of 0.9375.
The benefits of immediate tissue classification are that this could reduce the cost of screening and the amount of time a patient needs to wait before getting their results. Before this technique can be adopted clinically, the classification rate must be drastically improved to prevent the occurrence of false positives.

REFERENCES

Source of data: http://archive.ics.uci.edu/ml/datasets/breast+tissue

[1] Grenier, K., Dubuc, D., Chen, T., Artis, F., Chretiennot, T., Poupot, M., & Fournie, J. (2013). Recent Advances in Microwave-Based Dielectric Spectroscopy at the Cellular Level for Cancer Investigations. *IEEE Transactions on Microwave Theory and Techniques*,*61*(5), 2023-2030. doi:10.1109/tmtt.2013.2255885
[2] Burges C., "A tutorial on support vector machines for pattern recognition", In "Data Mining and Knowledge Discovery". Kluwer Academic Publishers, Boston, 1998, (Volume 2).
[3] V. Vapnik, S. Golowich, and A. Smola. Support vector method for function approximation, regression estimation, and signal processing. In M. Mozer, M. Jordan, and T. Petsche, editors, Advances in Neural Information Processing Systems 9, pages 281–287, Cambridge, MA, 1997. MIT Press.

APPENDIX

Table 1
Pre-processing: centered (9), scaled (9)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 66, 66, 68, 66, 67, 66, ...
Resampling results:

Accuracy	Kappa
0.8906746	0.8414096

Tuning parameter 'C' was held constant at a value of 1

Confusion Matrix 1

Statistics by Class:

	Class: 0	Class: 1	Class: 2	Class: 3
Sensitivity	1.0000	0.8889	1.000	1.0000
Specificity	0.9259	1.0000	1.000	1.0000
Pos Pred Value	0.7143	1.0000	1.000	1.0000
Neg Pred Value	1.0000	0.8750	1.000	1.0000
Prevalence	0.1562	0.5625	0.125	0.1562
Detection Rate	0.1562	0.5000	0.125	0.1562
Detection Prevalence	0.2188	0.5000	0.125	0.1562
Balanced Accuracy	0.9630	0.9444	1.000	1.0000

Overall Statistics

Accuracy : 0.9375
95% CI : (0.7919, 0.9923)
No Information Rate : 0.5625
P-Value [Acc > NIR] : 3.289e-06

Kappa : 0.903
Mcnemar's Test P-Value : NA

Table 2

Pre-processing: centered (9), scaled (9)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 66, 66, 68, 66, 67, 66, ...
Resampling results across tuning parameters:

sigma	C	Accuracy	Kappa
0.010	0.10	0.4208333	0.0000000
0.010	0.25	0.6007937	0.3488567
0.010	0.50	0.6325397	0.4159956
0.010	0.75	0.6456349	0.4439615
0.010	1.00	0.7363095	0.5999746
0.010	1.50	0.7972222	0.6973452
0.010	2.00	0.8228175	0.7364614
0.010	5.00	0.8670635	0.8028195
0.020	0.10	0.5121032	0.1789162
0.020	0.25	0.6367063	0.4234660
0.020	0.50	0.7190476	0.5699140
0.020	0.75	0.8144841	0.7232658
0.020	1.00	0.8228175	0.7364614
0.020	1.50	0.8406746	0.7631124
0.020	2.00	0.8587302	0.7901345
0.020	5.00	0.8628968	0.7964825
0.025	0.10	0.5162698	0.1858848
0.025	0.25	0.6408730	0.4323880
0.025	0.50	0.7726190	0.6571291
0.025	0.75	0.8242063	0.7390689
0.025	1.00	0.8353175	0.7554965
0.025	1.50	0.8628968	0.7964770
0.025	2.00	0.8670635	0.8028195
0.025	5.00	0.8712302	0.8089593
0.030	0.10	0.5347222	0.2209796
0.030	0.25	0.6450397	0.4407707
0.030	0.50	0.8138889	0.7225983
0.030	0.75	0.8228175	0.7364614
0.030	1.00	0.8400794	0.7627882
0.030	1.50	0.8670635	0.8028195
0.030	2.00	0.8670635	0.8028195
0.030	5.00	0.8878968	0.8351072
0.040	0.10	0.5876984	0.3227838
0.040	0.25	0.6773810	0.5027233
0.040	0.50	0.8359127	0.7557492
0.040	0.75	0.8400794	0.7627882
0.040	1.00	0.8587302	0.7905565
0.040	1.50	0.8623016	0.7950574
0.040	2.00	0.8628968	0.7968990
0.040	5.00	0.8954365	0.8482118
0.050	0.10	0.6007937	0.3533694
0.050	0.25	0.7184524	0.5707112
0.050	0.50	0.8339286	0.7530836
0.050	0.75	0.8490079	0.7757974
0.050	1.00	0.8628968	0.7968990
0.050	1.50	0.8581349	0.7891369
0.050	2.00	0.8623016	0.7950629
0.050	5.00	0.8954365	0.8490584
0.060	0.10	0.6146825	0.3820984
0.060	0.25	0.7472222	0.6182312
0.060	0.50	0.8394841	0.7614170
0.060	0.75	0.8587302	0.7912308
0.060	1.00	0.8484127	0.7743778
0.060	1.50	0.8581349	0.7891369
0.060	2.00	0.8706349	0.8073188
0.060	5.00	0.8865079	0.8354836
0.070	0.10	0.6283730	0.4081767
0.070	0.25	0.7644841	0.6460535
0.070	0.50	0.8484127	0.7750511
0.070	0.75	0.8587302	0.7912308
0.070	1.00	0.8484127	0.7738508
0.070	1.50	0.8623016	0.7950629
0.070	2.00	0.8795635	0.8221394
0.070	5.00	0.8767857	0.8214820
0.080	0.10	0.6367063	0.4239678
0.080	0.25	0.7900794	0.6861259
0.080	0.50	0.8531746	0.7817178
0.080	0.75	0.8484127	0.7743778
0.080	1.00	0.8539683	0.7829417
0.080	1.50	0.8706349	0.8073188
0.080	2.00	0.8817460	0.8262536
0.080	5.00	0.8767857	0.8220602
0.090	0.10	0.6367063	0.4244108
0.090	0.25	0.7948413	0.6934046
0.090	0.50	0.8573413	0.7880603
0.090	0.75	0.8484127	0.7743778
0.090	1.00	0.8581349	0.7891369
0.090	1.50	0.8795635	0.8221394
0.090	2.00	0.9009921	0.8566799
0.090	5.00	0.8767857	0.8220602
0.100	0.10	0.6325397	0.4170937
0.100	0.25	0.8079365	0.7140663
0.100	0.50	0.8628968	0.7971512
0.100	0.75	0.8442460	0.7681825
0.100	1.00	0.8664683	0.8012582
0.100	1.50	0.8817460	0.8262536
0.100	2.00	0.9009921	0.8565452
0.100	5.00	0.8767857	0.8220602
0.250	0.10	0.5920635	0.3391866
0.250	0.25	0.7942460	0.6952492
0.250	0.50	0.8513889	0.7805038
0.250	0.75	0.8817460	0.8272100
0.250	1.00	0.8900794	0.8401083
0.250	1.50	0.8900794	0.8401083
0.250	2.00	0.8761905	0.8191538
0.250	5.00	0.8761905	0.8191538
0.500	0.10	0.4208333	0.0000000
0.500	0.25	0.7847222	0.6824013
0.500	0.50	0.8775794	0.8212648
0.500	0.75	0.8950397	0.8498605
0.500	1.00	0.8950397	0.8498605
0.500	1.50	0.8811508	0.8296636
0.500	2.00	0.8811508	0.8296636
0.500	5.00	0.8630952	0.8019418
0.750	0.10	0.4208333	0.0000000
0.750	0.25	0.7632937	0.6523243
0.750	0.50	0.8019841	0.7112946
0.750	0.75	0.8873016	0.8371692
0.750	1.00	0.8998016	0.8567233
0.750	1.50	0.8853175	0.8345743
0.750	2.00	0.8714286	0.8130427
0.750	5.00	0.8589286	0.7962736
0.900	0.10	0.4208333	0.0000000
0.900	0.25	0.7248016	0.5947098
0.900	0.50	0.7888889	0.6937406
0.900	0.75	0.8638889	0.8020691
0.900	1.00	0.8859127	0.8342089
0.900	1.50	0.8805556	0.8277225
0.900	2.00	0.8811508	0.8285701
0.900	5.00	0.8686508	0.8110327

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.1 and C = 2.

breast tissue.R