Project Presentation

in

Pattern Recognition

Under the guidance of Prof. Sumeet Agarwal

By- ABHISHEK KUMAR (2009CS10175)
GAURAV AGARWAL (2009EE10390)

Aim

Classification of customers based on their attributes (eg. Age, country, etc) as a risky or non-risky customer.

Description

As the first part, we have performed credit risk analysis on German customers .

We have applied the SVM approach for classification and than combined the F-score approach that can simultaneously perform feature seletion task and model parameters optimization

In the second part, we have done comparative study with the Breiman Forest algorithm from the Mahout library with the dataset implemented in Hadoop Distributed File System.

Dataset

The credit dataset have been fetch from UCI repository

Function

SVM function with kernel chosen as Radial Basis Function (RBF)

SVM objective function

Min $0.5 * w'w + C (\sum error)$

subject to some constraints

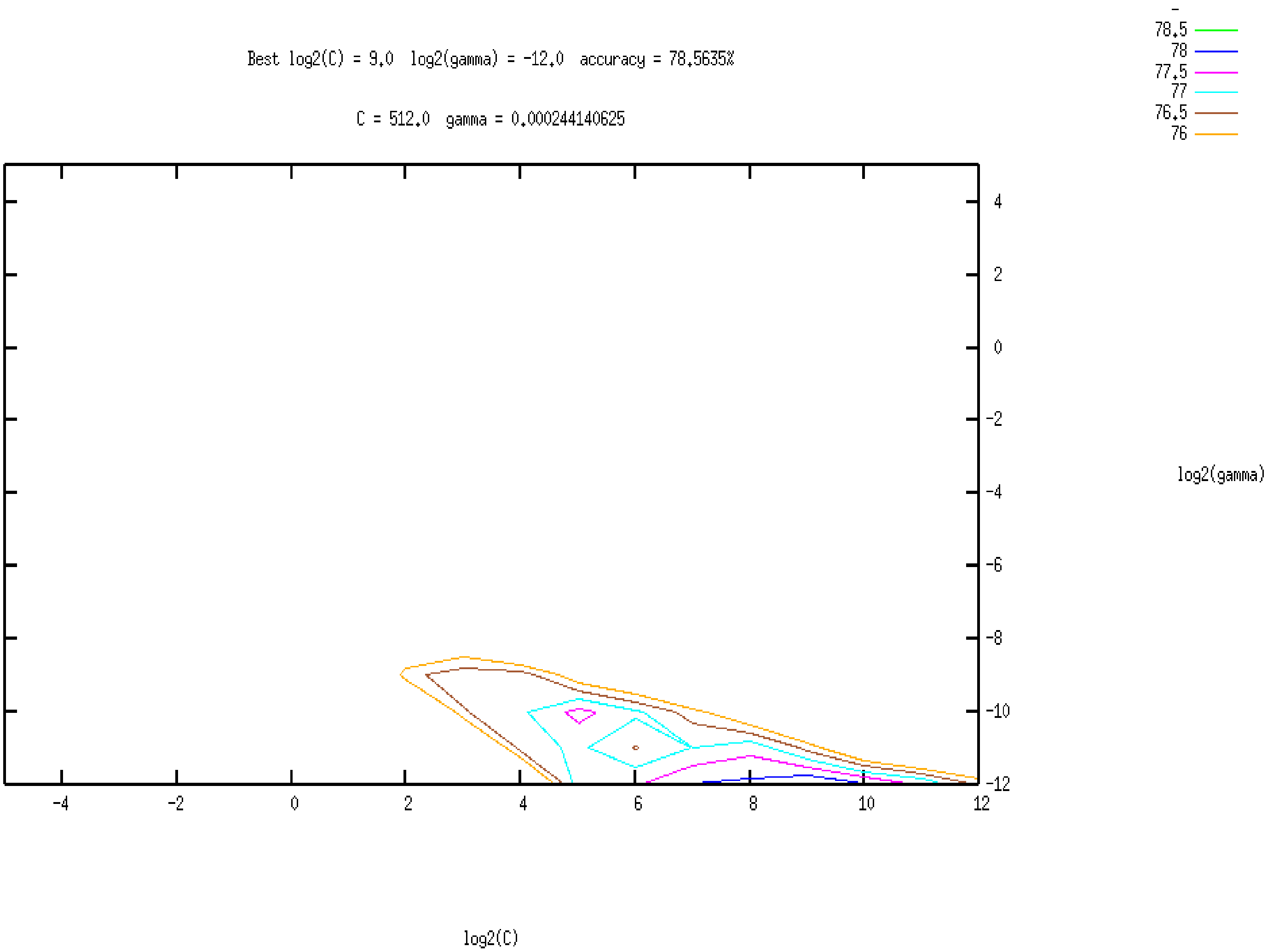
Kernel : $k(x,x')= \exp(- \gamma ||x-x'||^2)$

Parameters : (C, γ)

Glipmse of the algorithm followed

- We did consider a grid space of (C, γ) with log C Є {-5, -4, …, 12} ang log γ Є {-12, …, 5}
- For, each parameter (C, γ) , we did conduct n-fold cross validation on the training set with 'n' Є (1,10).
- Choose the parameter (C, γ) that lead to lowest CV error classification rate.
- We then used the best parameter to create the model as a predictor.

GNU Plot of the Grid



Results

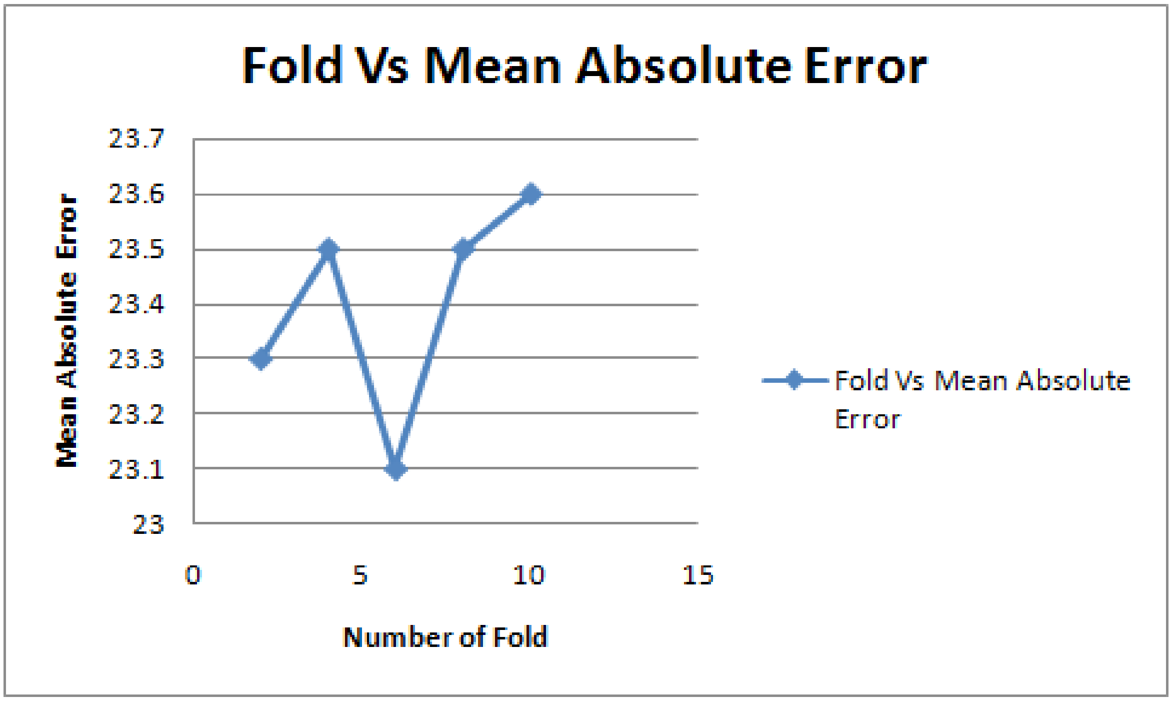
C=512, γ=0.000244 , **Folds=2** RBF kernel: K(x,y) = e^-(0.01* <x-y,x-y>^2)
No. of support vectors= 568
Correctly classified instances= 767 (76.7%)
Incorrectly classified instances=233 (23.3%)
Mean absolute error= 23.3%
Time taken to build= 0.84 seconds

C=512, γ=0.000244 , **Folds=4** RBF kernel: K(x,y) = e^-(0.01* <x-y,x-y>^2)
No. of support vectors= 568
Correctly classified instances= 765 (76.5%)
Incorrectly classified instances=235 (23.5%)
Mean absolute error= 23.5%
Time taken to build= 0.77 seconds

C=512, γ=0.000244 , **Folds=6** RBF kernel: K(x,y) = e^-(0.01* <x-y,x-y>^2)
No. of support vectors= 568
Correctly classified instances= 769 (76.9%)
Incorrectly classified instances=231 (23.1%)
Mean absolute error= 23.1%
Time taken to build= 1.21 seconds

C=512, γ=0.000244 , **Folds=8** RBF kernel: K(x,y) = e^-(0.01* <x-y,x-y>^2)
No. of support vectors= 568
Correctly classified instances= 763 (76.3%)
Incorrectly classified instances=237 (23.7%)
Mean absolute error= 23.7%
Time taken to build= 0.79 seconds

C=512, γ=0.000244 , **Folds=10** RBF kernel: K(x,y) = e^-(0.01* <x-y,x-y>^2)
No. of support vectors= 568
Correctly classified instances= 761 (76.1%)
Incorrectly classified instances=239 (23.9%)
Mean absolute error= 23.9%
Time taken to build= 0.79 seconds



Plot of Number of folds vs Absolute Error

NEXT STEP

Glipmse of the algorithm followed

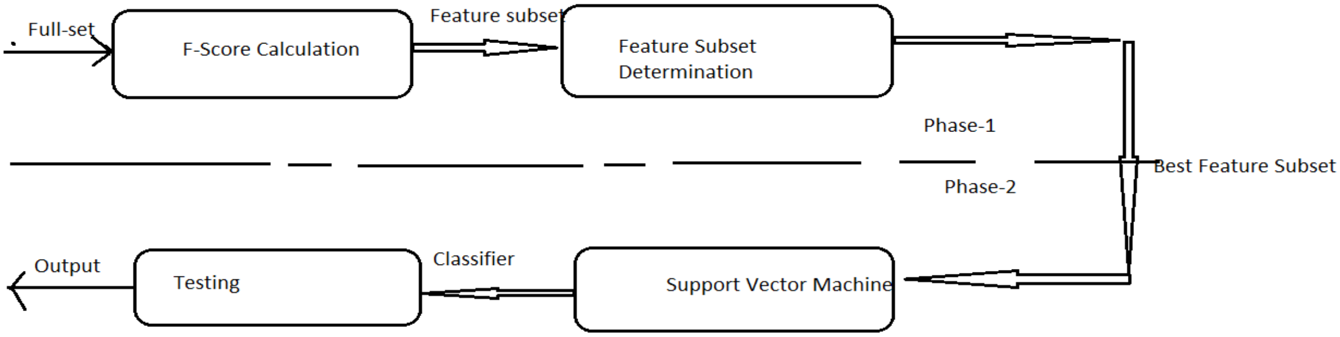
- Calculate F-score for every feature
- Sort F-score and set possible number of features from 15 to 24 (total number of attributes)
- For each 'f' threshold

  -Keep the first f-features according to F-score
  -Apply the n-fold cross-validation
  -Calculate the average validation error of the n-fold cross-validation
- Choose the f(threshold) with the lowest average validation error
- Drop the features with F-score below the selected threshold. Rerun the SVM training and measure the classification accuracy.

F-Score

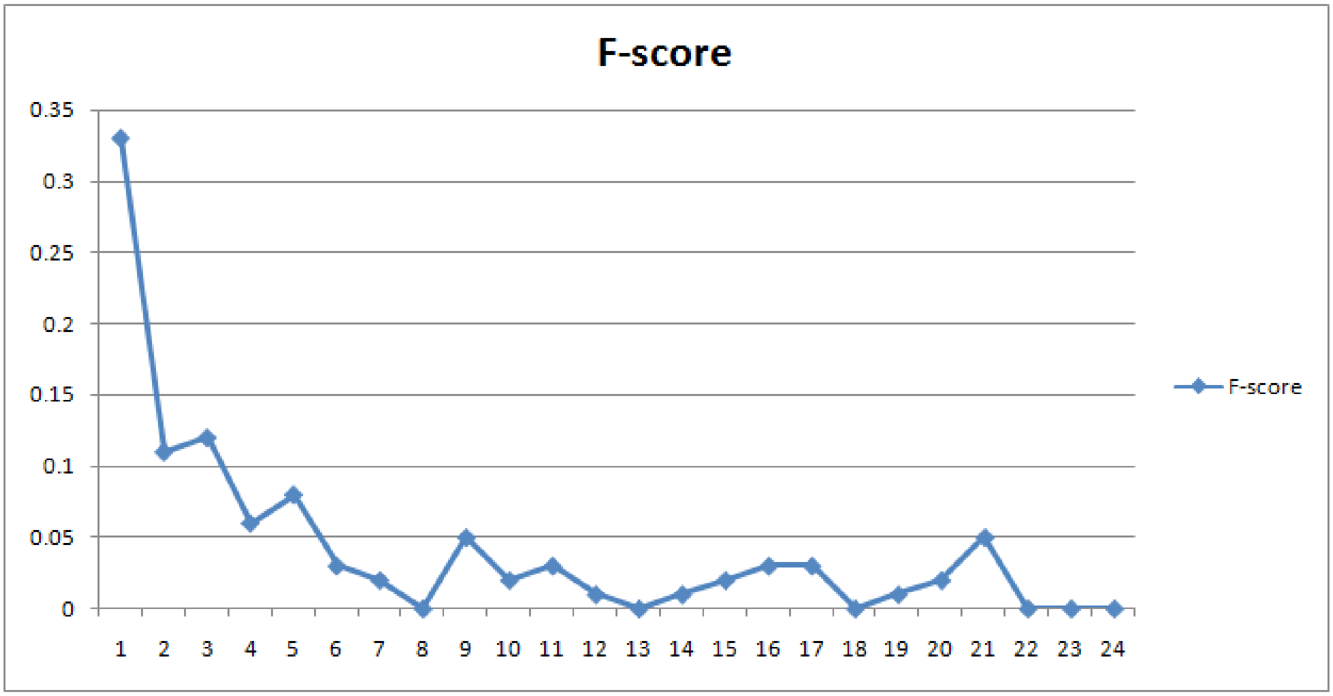
$$F(i) \equiv \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2}$$

Algorithm description



Results

F-Score of parameters



Parameters: 20 parameters sorted according to F-score

C=512, γ=0.000244 , **Folds=10** RBF kernel: K(x,y) = e^-(0.01* <x-y,x-y>^2)
No. of support vectors= 569
Correctly classified instances= 763 (76.3%)
Incorrectly classified instances=237 (23.7%)
Mean absolute error= 23.7%
Time taken to build= 0.83 seconds

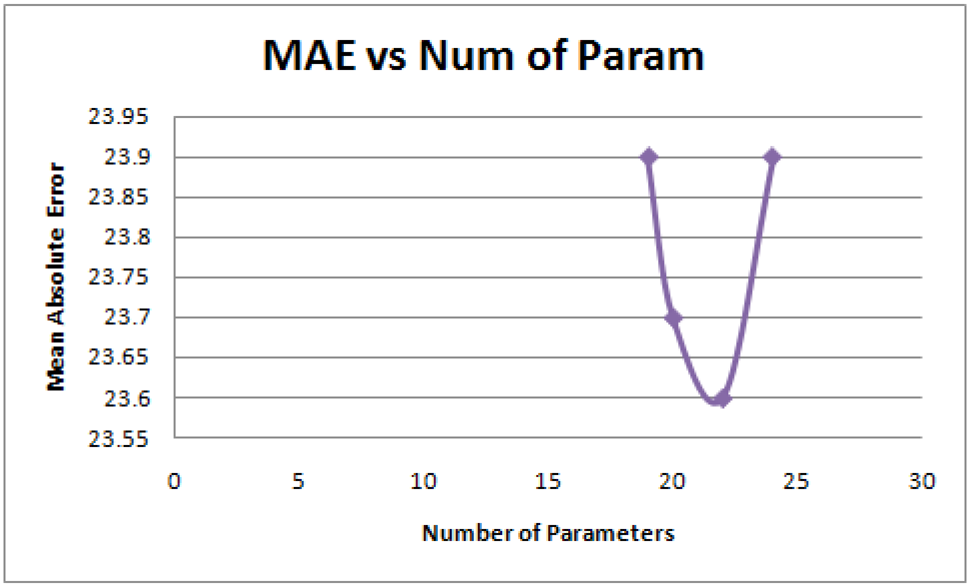
Parameters: 21 parameters sorted according to F-score

C=512, γ=0.000244 , **Folds=10** RBF kernel: K(x,y) = e^-(0.01* <x-y,x-y>^2)
No. of support vectors= 578
Correctly classified instances= 761 (76.1%)
Incorrectly classified instances=239 (23.9%)
Mean absolute error= 23.9%
Time taken to build= 0.96 seconds

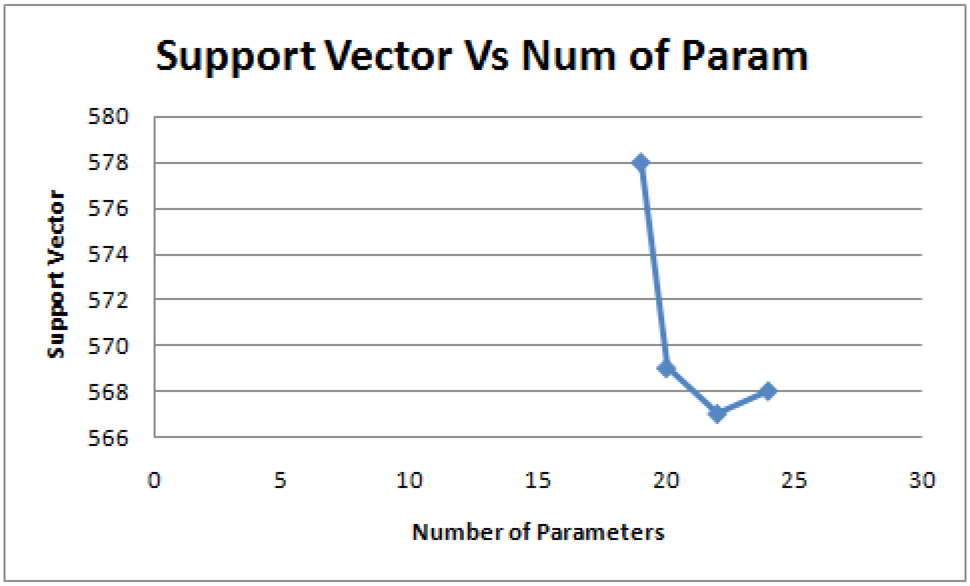
Parameters: 22 parameters sorted according to F-score

C=512, γ=0.000244 , **Folds=10** RBF kernel: K(x,y) = e^-(0.01* <x-y,x-y>^2)
No. of support vectors= 567
Correctly classified instances= 764 (76.4%)
Incorrectly classified instances=236 (23.6%)
Mean absolute error= 23.9%
Time taken to build= 0.85 seconds

Plots



Plot of mean absolute error versus the number of parameters



Plot of Support Vectors versus the No. of parameters

COMPARISION WITH BREIMAN FOREST ALGORITHM IN HADOOP

Error using Breiman Forest algorithm in Mahout on dataset implemented in Hadoop= **24.7%**

Error using the above SVM approch= **21.44 %**

Thanks...