Optical Character Recognition of printed text

Mynhardt, S. & Reisle, C. R.

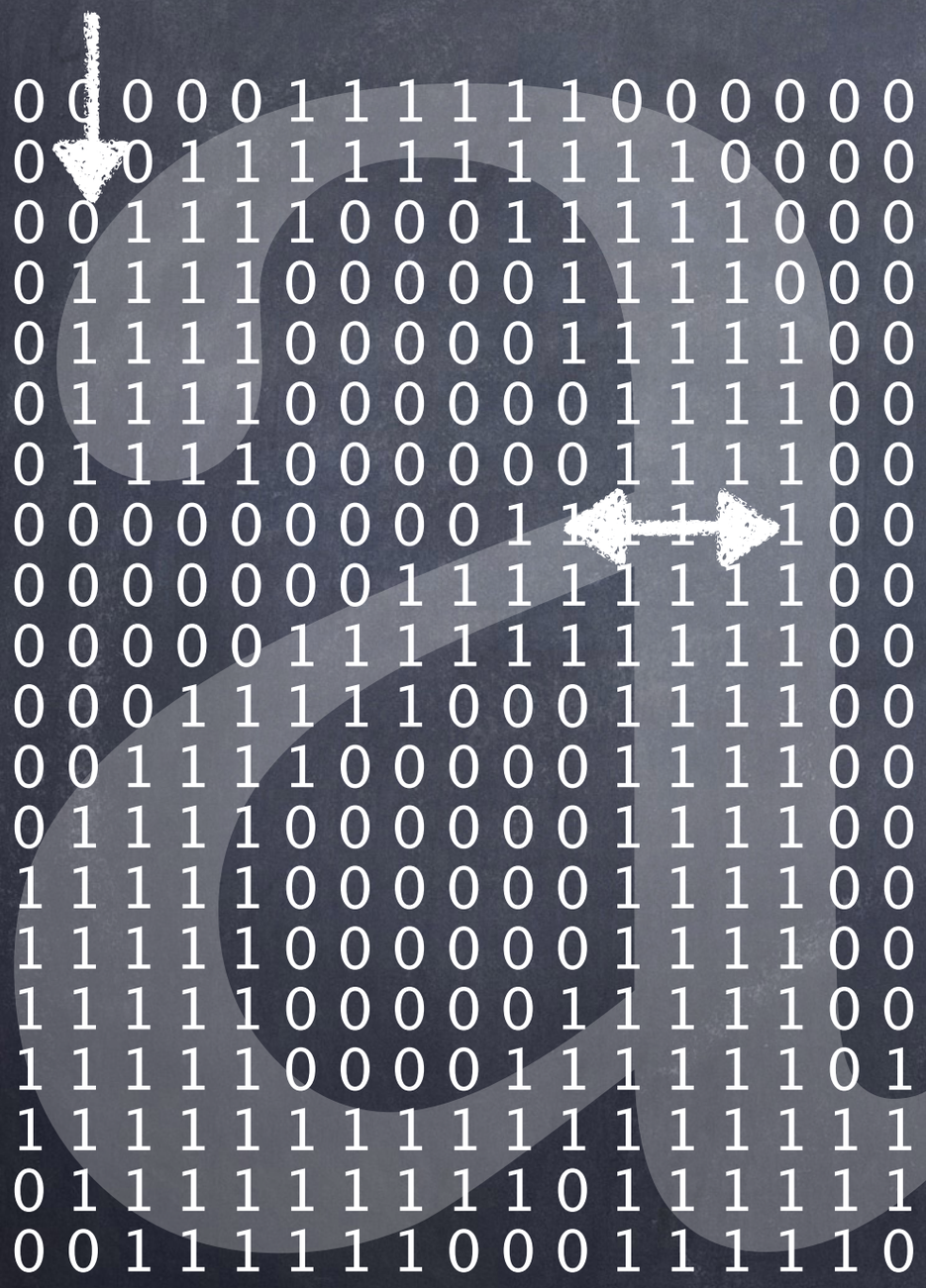
Outline

Project Goal: take jpg file input, extract text, output to plain text file

- Pre-Processing
- Classification
- Attributes
- Model
- Test data
- Results
- Next steps

Pre-processing

Simple Clustering



- makes a pass over the columns of binary matrix, first labelling adjacent pixels with the same set number

- on the second pass over the matrix, the sets of adjacent pixels are merged

- on the third pass, the regions are extracted based on the pixels belonging to each set/component

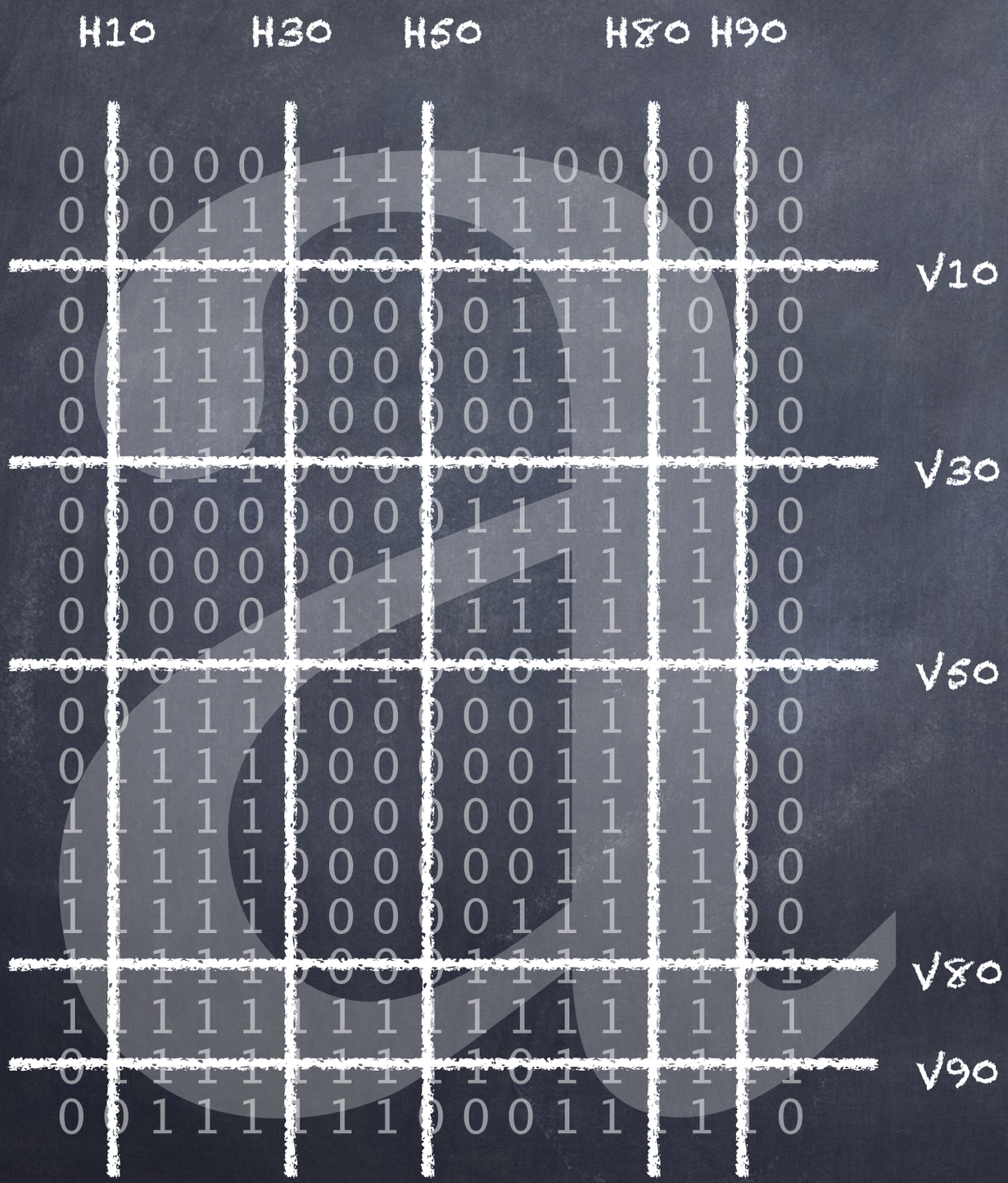
Classification

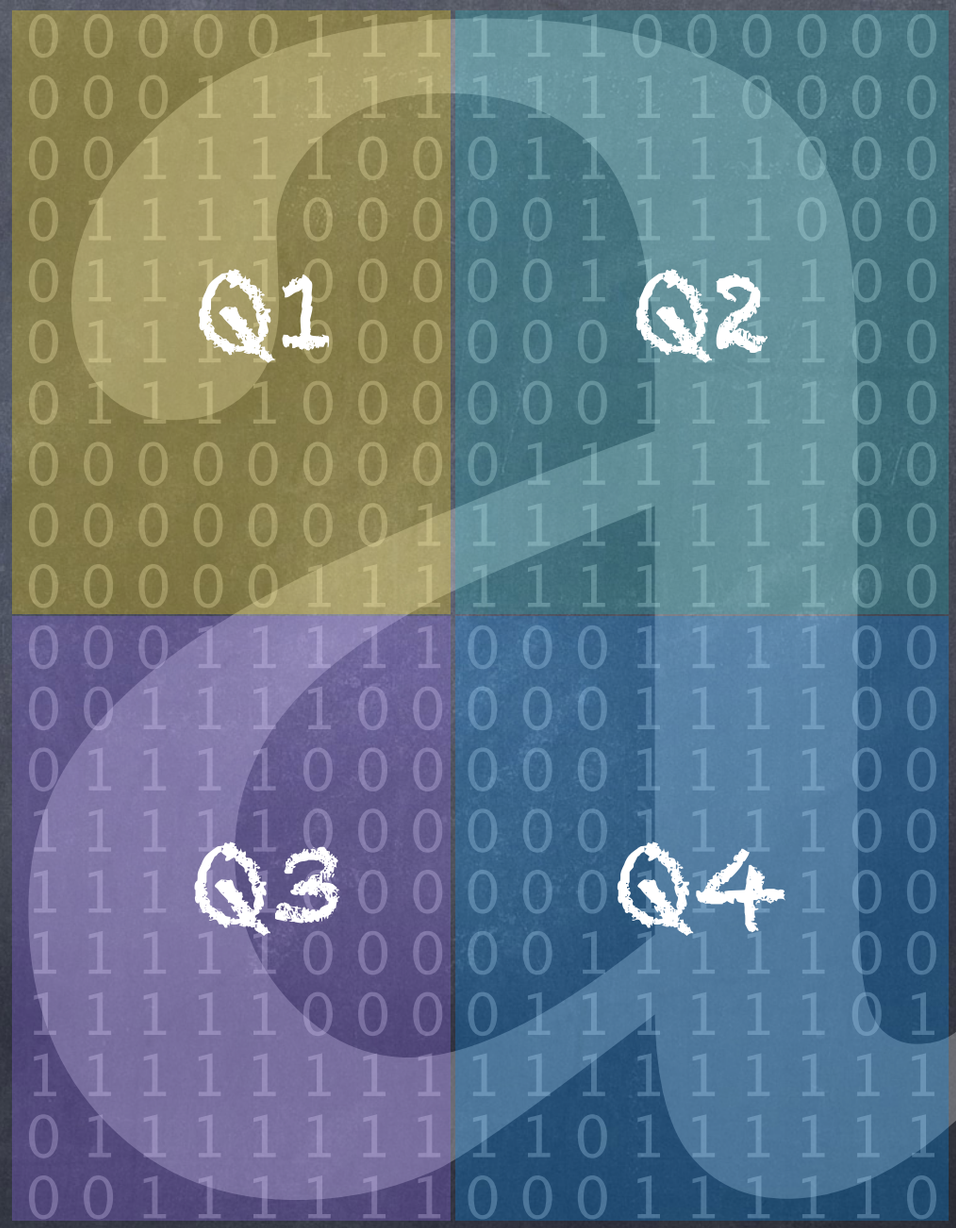
Computing Attribute Vectors

Feature	Function
H{10, 30, 50, 80, 90}	Sum of black pixels in the row/col of pixels at X% of the total height/width
V{10, 30, 50, 80, 90}	
H-symm	Measure of the horizontal symmetry
V-symm	Measure of the vertical symmetry
CC	number of closed components
H/W ratio	height to width ratio of the component
C	number of merged components
Q{1, 2, 3, 4}	ratio of black pixels that fall into the top left, top right, bottom left and bottom right quadrants, respectively, of the image
Ih{30, 50, 80}	number of times the pixels transition from black to white and white to black at the specified row or column position
Iv{30, 50, 80}	

Attributes...

Attributes...





Classification

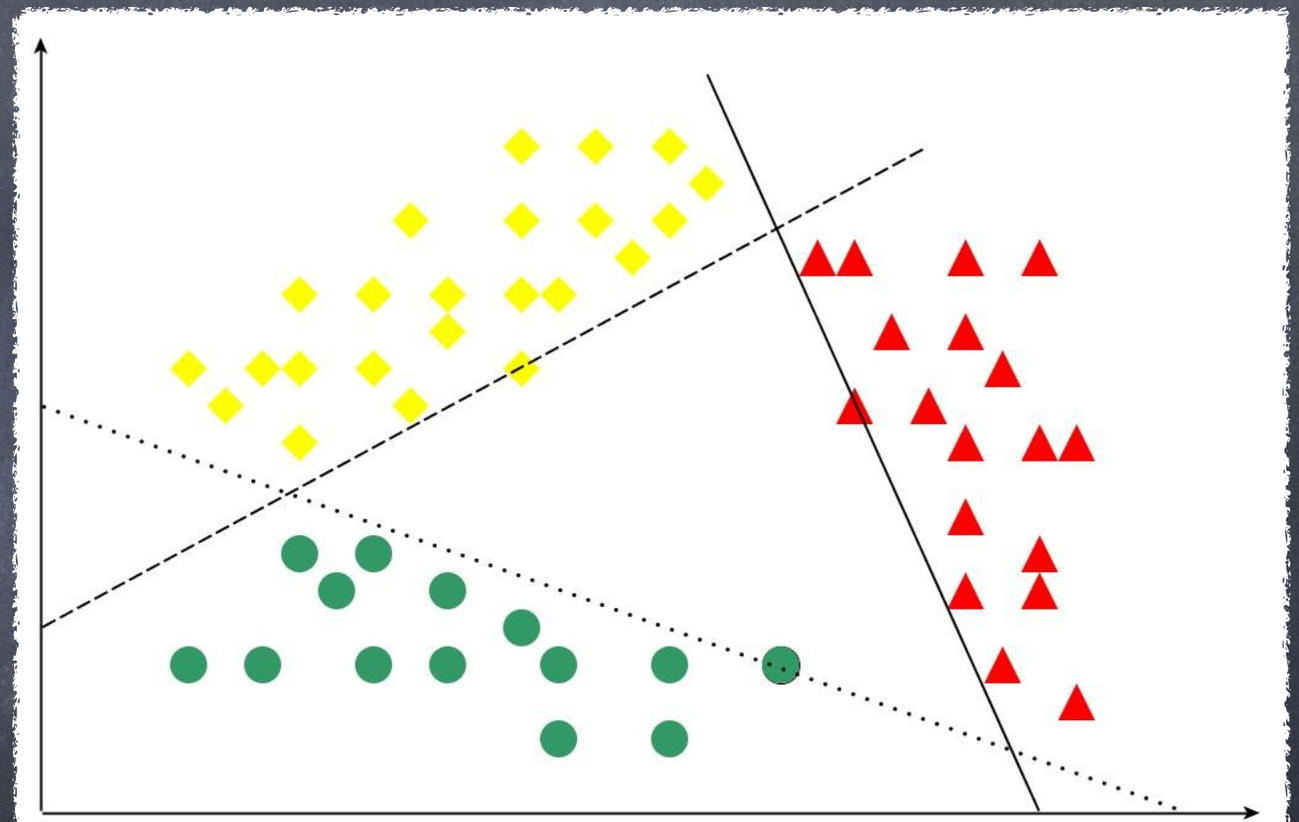
Mathematical Models

Support Vector Machine

- We have multiple (82) classes of characters

- We need 82 SVM classifiers or hyperplanes to separate characters out from each other

- Example: Classifier for 0, call it SVM0, will separate all the samples of 0 from the other characters

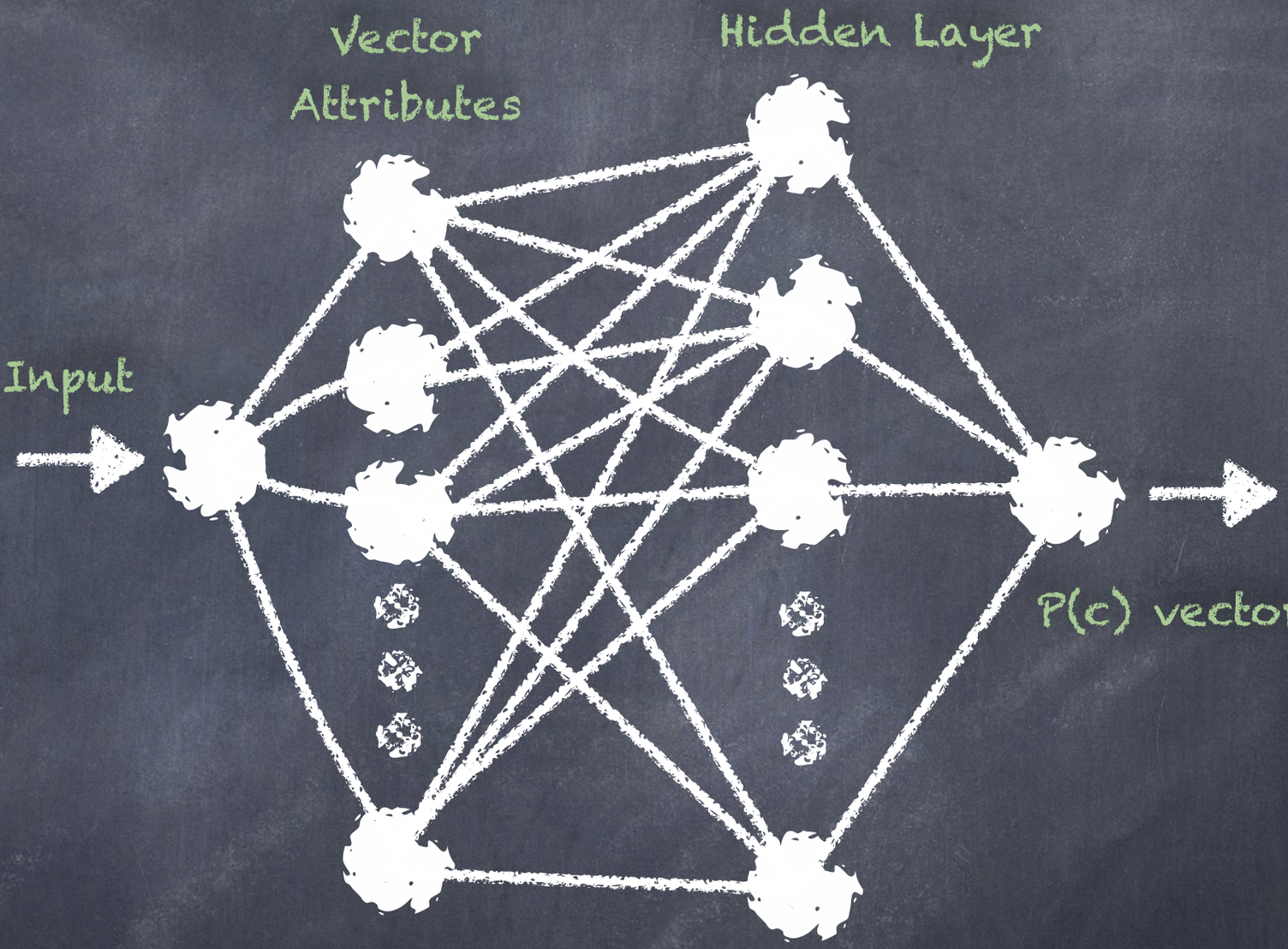


Method for a three class problem

Artificial Neural Network

- An ANN is an abstract copy of biological neural networks

- An ANN is a function that maps input vectors from our input space (a binary matrix) to an output space (class probability vector)



- The function has several layers consisting of neurons (called perceptrons)

- For each input our particular ANN outputs a vector with the probability of an ouput belonging to a given class

Classification

Model Evaluation

=== Evaluation on training set ===

	ANN	SVM
Correctly Classified Instances	411 (100 %)	410 (97.7567 %)
Incorrectly Classified Instances	0 (0 %)	1 (0.2433 %)
Kappa statistic	1	0.9975
Mean absolute error	0.0033	0.0021
Root mean squared error	0.0175	0.0137
Relative absolute error	13.7823 %	8.8617 %
Root relative squared error	15.9544 %	12.5025 %
Total Number of Instances	411	411

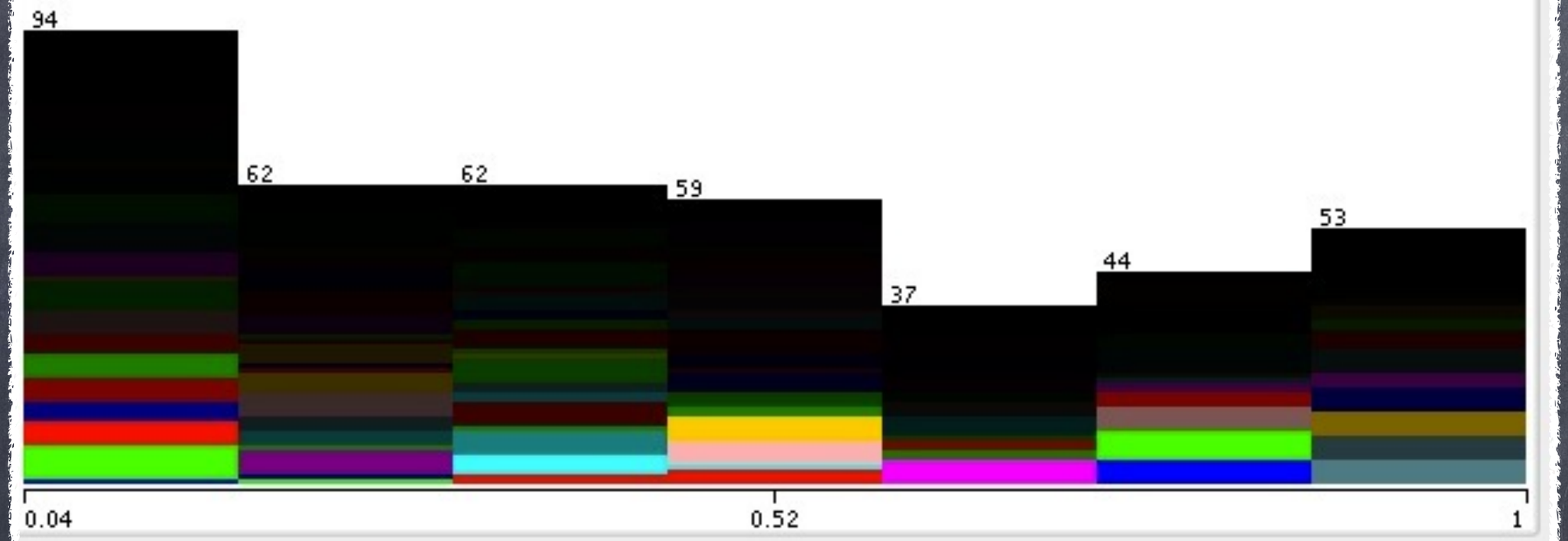
=== Stratified 10-fold cross-validation ===

	ANN	SVM
Correctly Classified Instances	399 (97.0803 %)	399 (97.0803 %)
Incorrectly Classified Instances	12 (2.9197 %)	12 (2.9197 %)
Kappa statistic	0.9704	0.9704
Mean absolute error	0.0031	0.0028
Root mean squared error	0.025	0.0275
Relative absolute error	12.7235 %	11.7249 %
Root relative squared error	22.7414 %	25.0182 %
Total Number of Instances	411	411

Classification

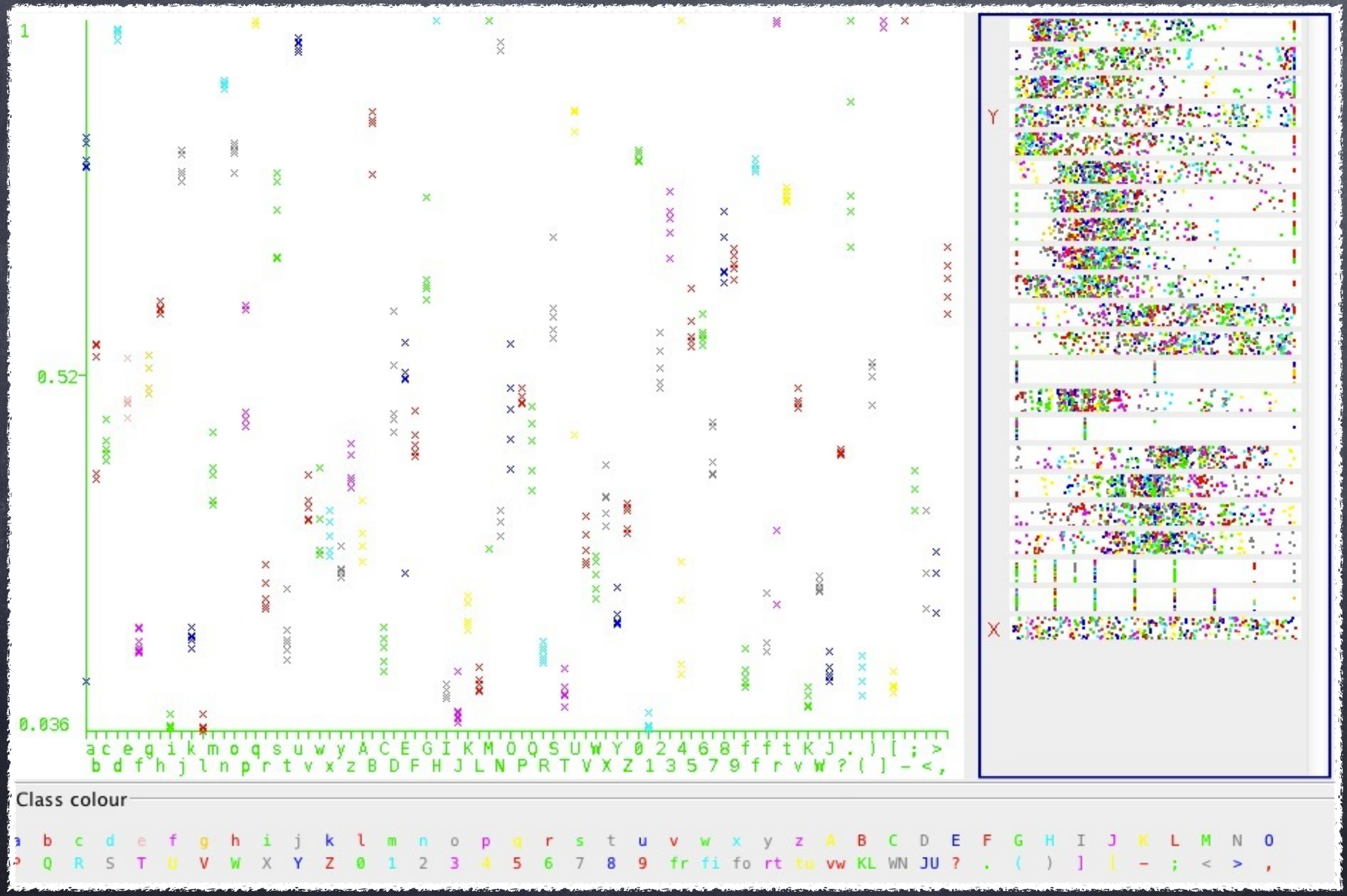
Attribute Evaluation

Attribute: V80

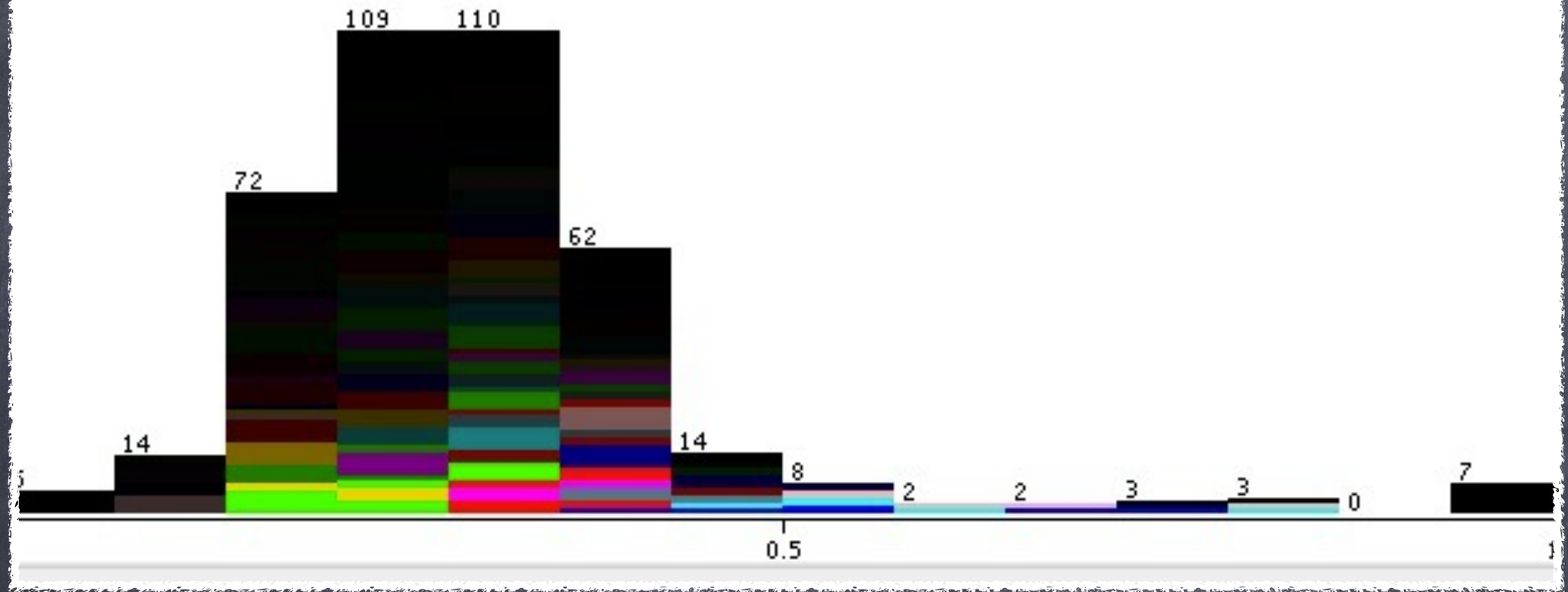


- The V80 values are fairly evenly distributed
- This indicates that this attribute on its own is good at differentiating between different characters

Attribute: Class vs V&O

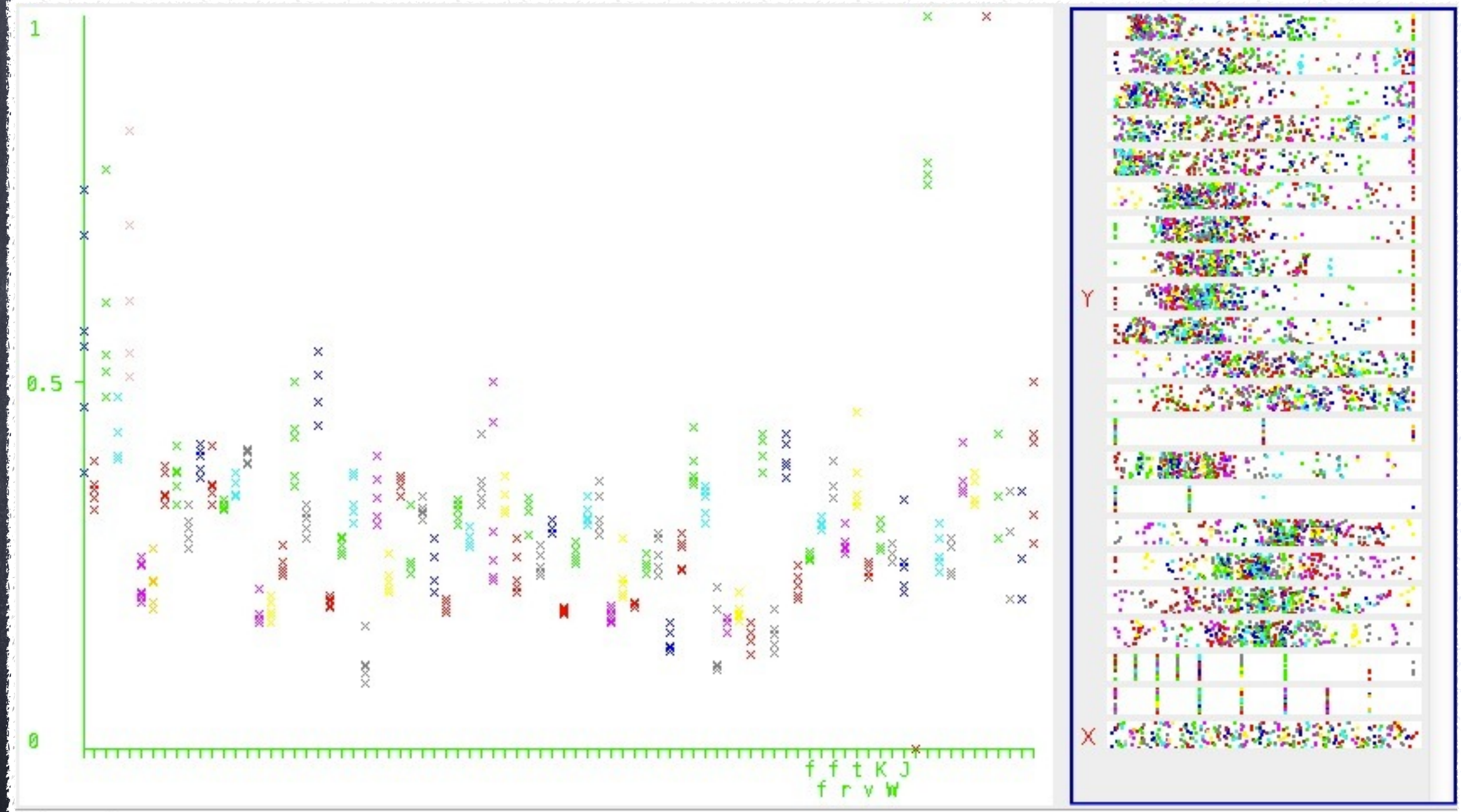


Attribute: H80



- The majority of the values tend to be around 0.25
- This indicates that this attribute on its own, is not particularly good at differentiating characters

Attribute: Class vs H&O



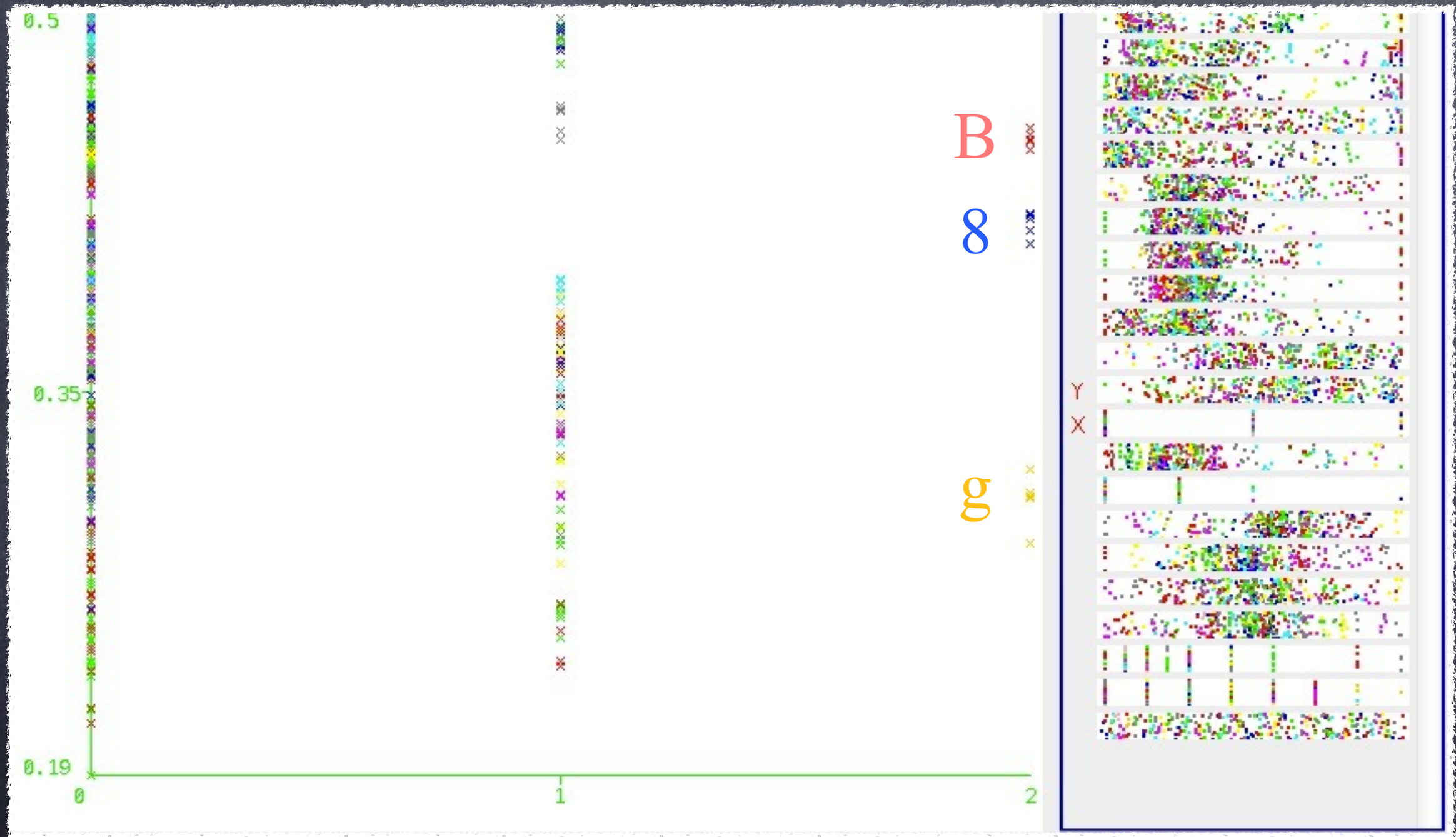
Class colour

a b c d e f g h i j k l m n o p q r s t u v w x y z A B C D E F G H I J K L M N O
P Q R S T U V W X Y Z 0 1 2 3 4 5 6 7 8 9 fr fi fo rt tu vw KL WN JU ? . ( )] [- ; < > ,

Attributes ...

- Of course, this is why we do not have only one point of comparison
- There are a total of 22 attributes that are compared between each test character and the training characters
- One important observation is that although most of the letters have similar H80 values, any outliers, in this case "." and "-", are effectively isolated from the rest of the data by this attribute

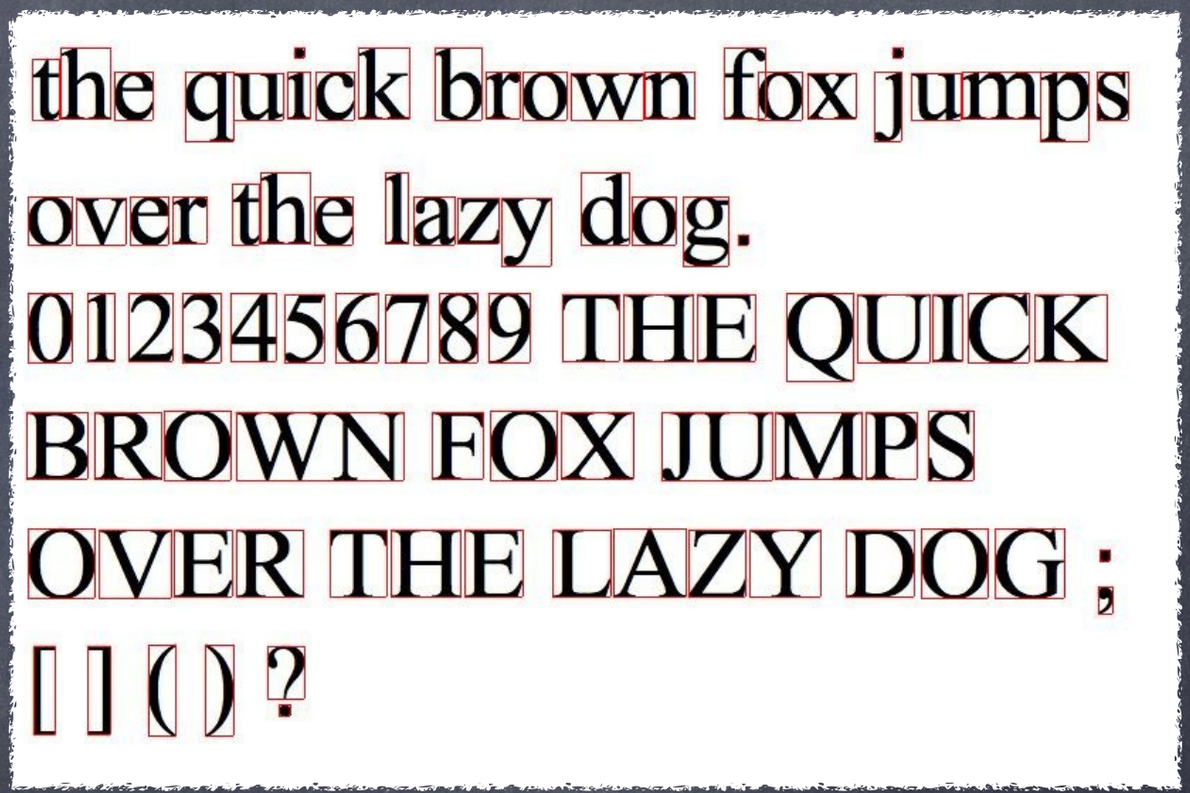
CC vs Vsymm coloured by class value



Examples

Example 1

the quick brown fox jumps over the lazy dog.
0123456789 THE QUICK BROWN FOX JUMPS OVER THE LAZY DOG ; [ ] ( ) ?



SVM

the quick brown fox jumps

over the lazy dog.

0123456789 THE QUICK

BROWN FOX JUMPS

OVER THE LAZY DOG ;

[ ] ( ) ?

ANN

the quick brown fox jumps

over the lazy dog.

0123456789 THE QUICK

BROWN FOX JUMPS

OVER THE LAZY DOG ;

[ ] ( ) ?

Example 2

c. Repeat the analysis for part (b) using the same
Compare the F-measure results for both classifie
results consistent with what you expect from the

c. Repeat the analysis for part (b) using the same
Compare the F-measure results for both classifie
results consistent with what you expect from the

SVM

c. fepeat the analysis for part (b) using the same

Compare the F-measure results for both classifie]

results consistent with what you expect from the

ANN

c. Repeat the analysis for part (b) using the same

Compare the F-measure results for both classifie]

results consistent with what you expect from the

Next Steps ...

- Pre-processing: Scanning algorithm used to isolated putative letters is not robust in terms of noise, would like to explore other options

- Pre-processing: image straightening step

- post-processing: determining spacing on edge cases (i.e. no separate words, or all separate characters)

- other: find frequent pairs of letters for training on pairs of letters that may end up clustered together

References

1. Neves, E. et al. (1997). IEEE. A Multi-Font Character Recognition Based on its Fundamental Features by Artificial Neural Networks

2. Shrivastava, V. & Sharma, N. (2012). SIPIJ 3(5). Artificial Neural Network Based Optical Character Recognition.

3. Hall, M. et al. (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

4. Tautu E.D. & Leon, F. (2012). Bul. Inst. Polit. Optical Character Recognition using support vector machines.

5. Sadri, J. et al. (2003). MVIP. Application of Support Vector Machines for Recognition of Handwritten Arabic/Persian Digits.

6. Rashnodi, O. (2011). International journal of computer applications 29(12). Persian Handwritten Digit Recognition using Support Vector Machines.