University of Victoria

Department of Software Engineering

Analysis of Multiclass

Seed Kernel Dataset

By

Malcolm Newson

Class:

Data Mining (Seng 474)

Introduction:

A data set containing 210 sampled seed kernels were analyzed using K-Nearest Neighbour (knn), Naive Bayes (nb), Perceptron (percep), and Decision Tree (tree) classifier algorithms. Each sampled seed had 7 numerical attributes explored in every combination from pairs up to quintets to investigate which classifier-attributes groupings produced the highest accuracy during prediction. The results demonstrated that K-Nearest Neighbour, Naive Bayes, and Decision Tree models trained with all 7 attributes produced reasonably good accuracy (~90%), but specific subgroups of algorithm-attribute groupings can achieve ~94%. Perceptron did poorly in all cases.

Data Set:

The data set investigated was from the Machine Learning Repository at the Center for Machine Learning and Intelligent Systems (http://archive.ics.uci.edu/ml/datasets/seeds). It consists of 210 records of seed kernels that have been imaged on 13x18 cm X-ray KODAK plates, which is a non-destructive soft X-ray technique that is considerably cheaper than the alternatives like scanning microscopes or laser technologies. Three varieties of wheat: Kama, Rosa, and Canadian, where harvisted from experimental fields at the Institute of Agrophysics of the Polish Academy of Science in Lubin. 70 kernels were randomly selected from each variety and 7 numerical attributes were derived from the X-ray images, Area (A),

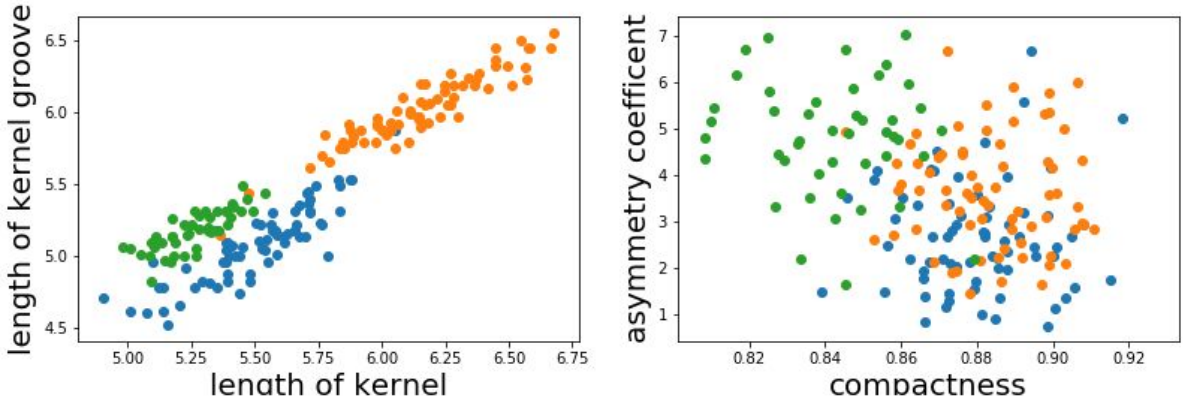


Figure 1

Perimeter (P), Compackness ($C = \frac{4\pi A}{P^2}$), Length of Kernel, Width of Kernel, Asymmetry Coefficient, and Length of Kernel Groove. Each record includes the classification of the seed, 0 for Kama, 1 for Rosa, and 2 for Canadian. Some minor data cleaning was required to remove extra tab characters in the downloaded data file.

This full set of 7 attributes would require a 7-D scatter plot to visualize. Instead, each pair of attributes was plotted in a 2-D scatter plot in Appendix A and two of note are presented in Figure 1. The left figure, Length of Kernel Groove vs Length of Kernel, nicely shows clusters of each kernel variety mostly separated. This would intuitively indicate that this attribute pair would effectively distinguish the class or kernel variety. Contrasted to this is the right figure, Asymmetry Coefficient vs Compactness, which shows significant overlap that would make classification challenging.

Analysis:

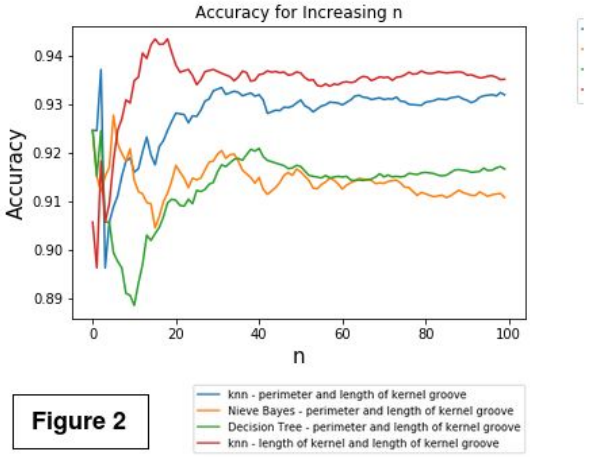


Figure 2

The entire data set was randomly split into a training set of 157 records (~75%) and a test set of 53 records (~25%). The training set is used to train a model based on an algorithm-attribute grouping that then predicts the class of the test records. The first measure used to assess the quality of each model is the accuracy, defined as the ratio of successful classifications during testing divided by the total number of training records. Because the data set was relatively small, the accuracy of the model varied depending on which records were placed in the training or test category. To address this, each model was repeatedly trained on n different random groupings of test and training records, and the accuracy calculated for all iteration was averaged. As n increases it is clear in Figure 2 that the calculated mean of accuracy settles. The full results from all pairs of attributes for each algorithm used is shown in Figure 3.

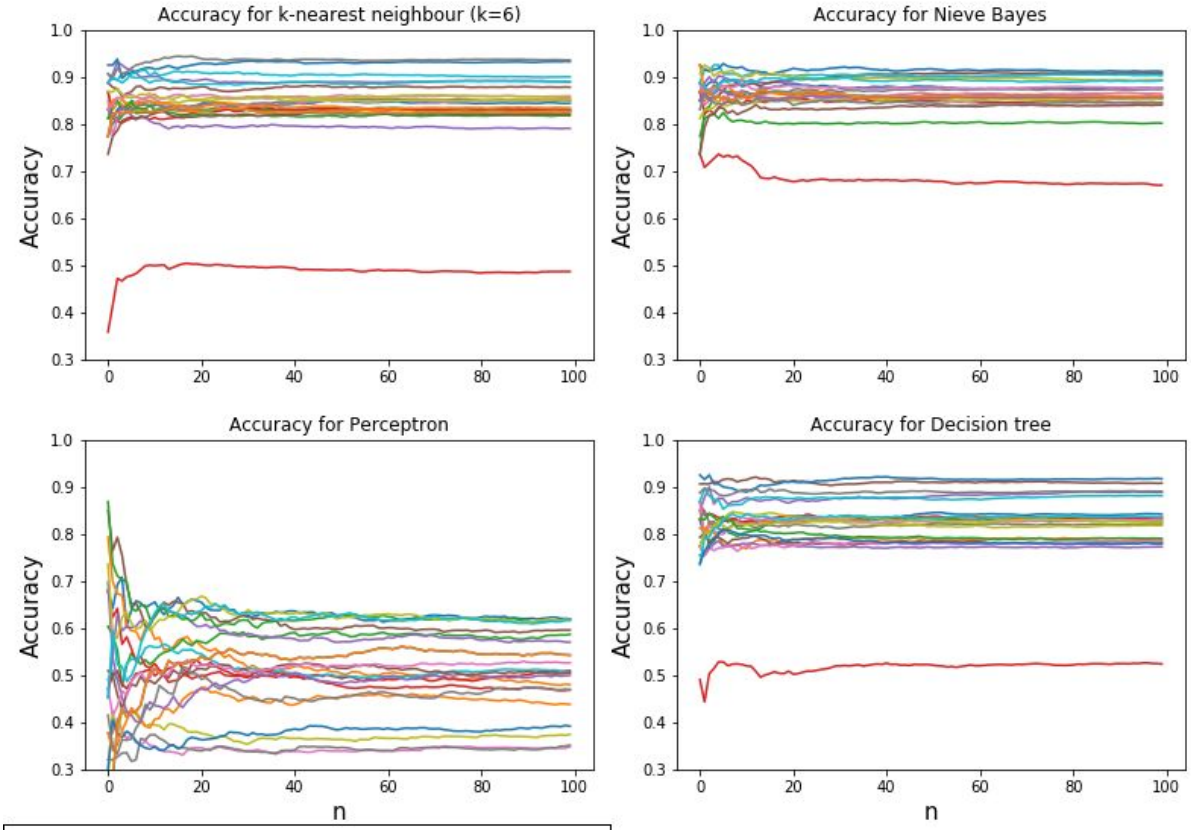


Figure 3: All above are for pairs of attributes:

Algorithms:

Four different algorithms are explored to see which can provide the greatest accuracy when predicting previously unseen records. These are K-Nearest Neighbour (knn), Naive Bayes (nb), Perceptron (percep), and Decision Trees (tree). Each was applied to the full 7 attributes as well as every combination of attributes: 21 pairs, 35 triplets, 25 quartets, and 21 quintents. K-Nearest Neighbour can have any positive integer k selected, and Figure 4 explores up to k=20 for attribute pairs to see which provides the highest accuracy (including n=100 iterations to average accuracy). It is clear from the right plot that the structure of the attribute pair accuracy is preserved regardless of k, pair 13 (Compactness vs Asymmetry Coefficient) remains obviously poor at predicting class. For most pairs of attributes two rules stand out.The first is to

avoid k=1 and the second is that accuracy drops off or stays relatively constant with increasing k after ~13. Two pairs compete with similarly high accuracy around 0.93. It is interesting to note that the blue line (Perimeter vs Length of Kernel Groove, pair id = 10) peaks at k=2,3 and then drops off for k>3. Contrasted to the gray line (Length of Kernel vs Length of Kernel Groove, pair id = 17) which varies around a high average for k between ~4 and ~17.

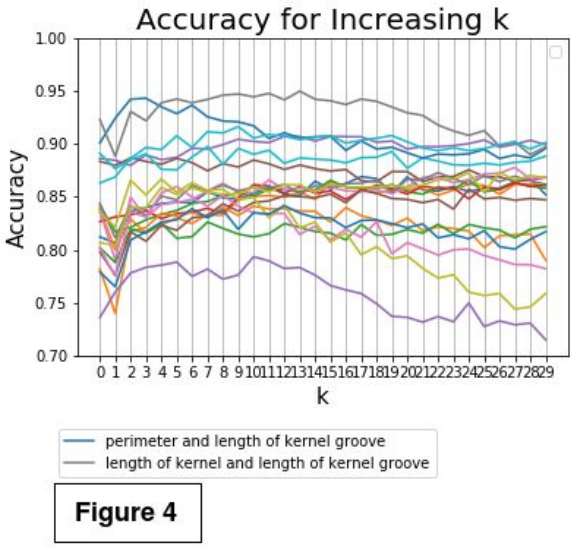


Figure 4

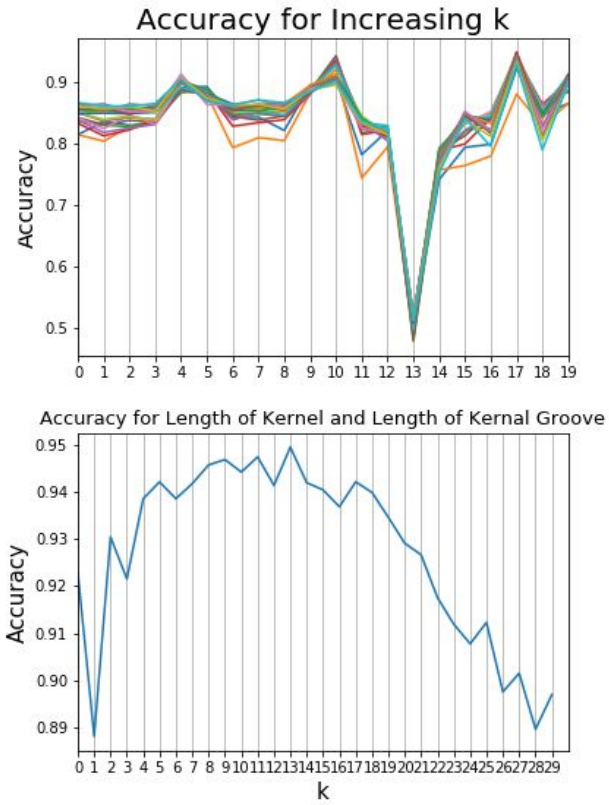
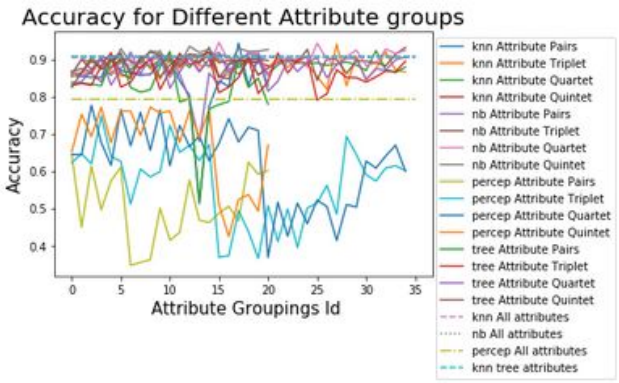
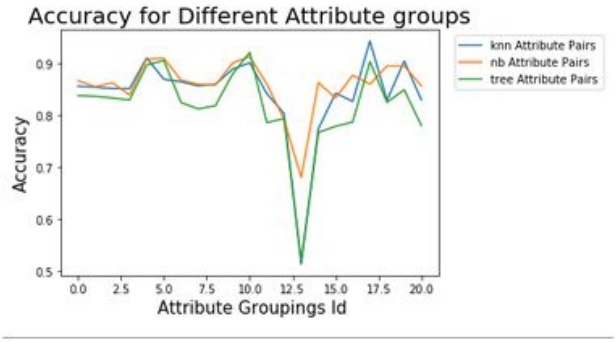


Figure 5 displays the accuracy calculated for each attribute combination for all 4 classifiers (n=100) in the top left. Perceptron does exceptionally poorly; the algorithm applied to all 7 attributes does better than any subset but is still well below the other 4 algorithms. The top right figure displays attribute pairs for all K-Nearest Neighbour, Naive Bayes, and Decision trees. It is interesting to note that all three display the same significant dip on pair id 13 (Compactness and Asymmetry Coefficient). The bottom figure displays algorithm-attribute groupings that have a peak





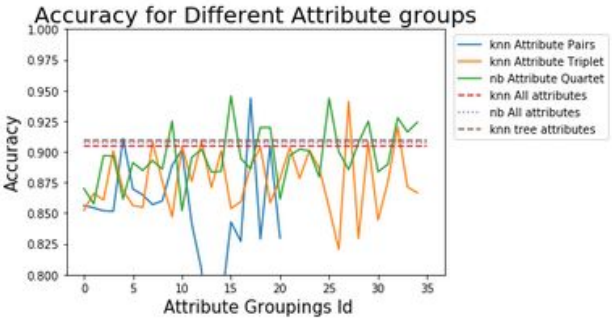


Figure 5: k=15

above 0.93 (Naive Bayes quartet id 15, K-Nearest Neighbour pair 17, Naive Bayes quartet id 25 and K-Nearest Neighbour triplet id 27). These specific attribute combinations do better than the same algorithms applied to all 7 attributes, which produced accuracies of knn = 90%, nb = 90%, percep = 78%, and tree = 91%.

Confusion matrix:

In this case of 3 classes {0,1,2} the confusion matrix is 3x3, whereas in a simple binary {0,1} classification problem the confusion matrix is 2x2. Accuracy is defined as the number of successful classifications of the model divided by the total number of classifications. This is visualized in Figure 6 by two grids where the sum of the dark squares in the numerator is divided by the sum of the dark squares in the denominator. Other measures of a model's

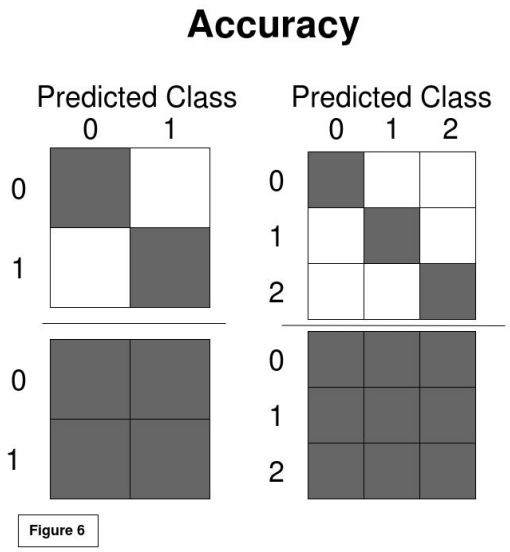


Figure 6

predictions include Sensitivity, Specificity, Positive Predictive Value (PPV), and Negative Predictive value (NPV) are also visualized in Figure 7. However these measures are always with respect to one class. PPV with respect to class 0 will be different than with respect to class 1 or 2.

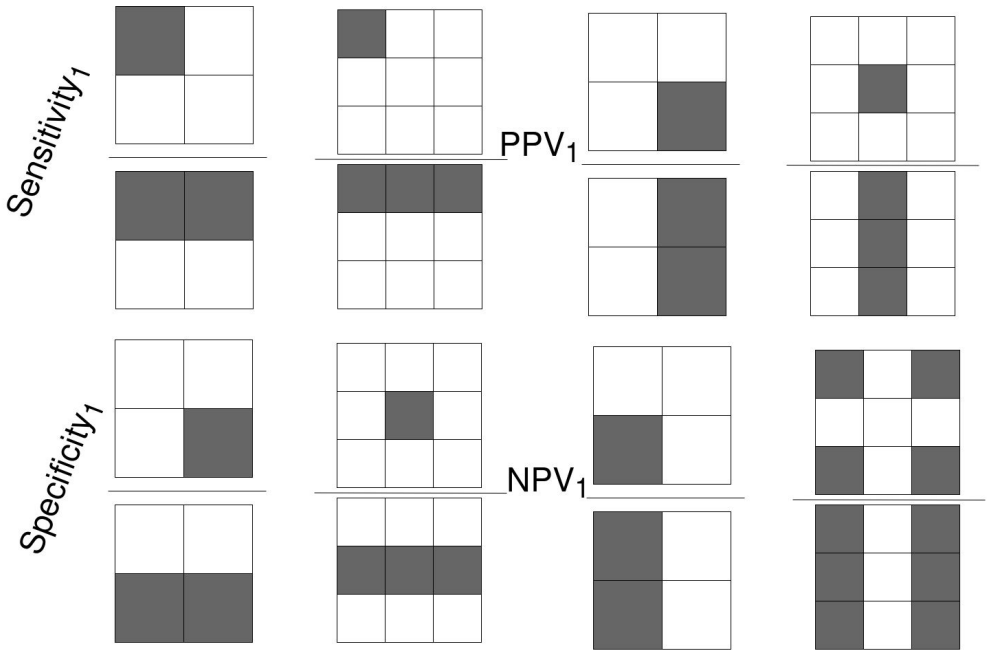


Figure 7

Figure 7 shows the squares used for each calculation with respect to class 1, where the top square of each pair is the numerator and the bottom square is the denominator. The english definitions are as follows:

Sensitivity$_c$ - the number of cases correctly classified as c divided by the number of cases truly belonging to c.

Specificity$_c$ - the number of cases correctly classified as *not* c divided by the number of cases truly belonging to *not* c.

PPV$_c$ - the number of cases correctly classified as c divided by the number of cases classified as c.

NPV$_c$ - the number of cases correctly classified as *not* belonging to c divided by the number of cases classified as *not* belonging to c.

The confusion matrices of three algorithm-attribute combinations are given below where each cell has been the average of n=100 iterations. These include two very high accuracy combinations; K-Nearest Neighbour pair id 17 (Length of Kernel and Length of Kernel Groove) and Naive Bayes quartet id 25 (Perimeter, Compactness, Asymmetry Coefficient and Length of Kernel Groove). Both knn pair 17 and nb quartet 25 have similar results because their accuracy is very similar. Some differences include that knn pair 17 has a low sensitivity with respect to class 0 and a low PPV for class 3. This means that the classifier is doing a better job for class 2 than for either class 0 or 3. Nb quartet 25 has more stable results with just a slightly lower sensitivity for class 0. The group knn pair id 13 is one of the low spikes that is noticeable in Figure 5 right, and has a low accuracy of 0.517. It does much worse at predicting class 1, with a very low sensitivity of 0.289 and low PPV and NPV.

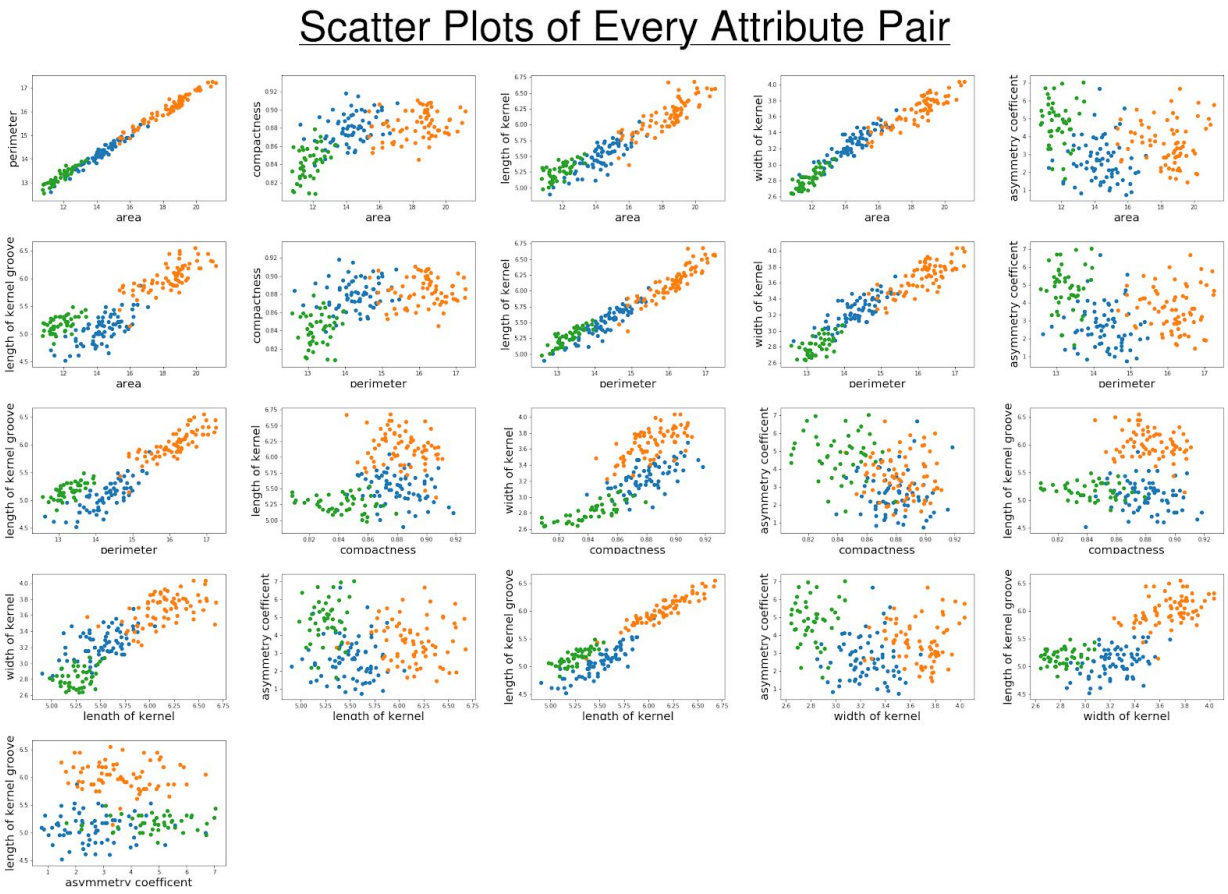
Knn pair 17 - Length of Kernel and Length of Kernel Groove					Nb quartet 25 - Perimeter, Compactness, Asymmetry Coefficient and Length of Kernel Groove				
Accuracy = 0.942					Accuracy = 0.942				
Confusion matrix:	**predicted class = 0**	**predicted class = 1**	**predicted class = 2**		Confusion matrix:	**predicted class = 0**	**predicted class = 1**	**predicted class = 2**	
true class = 0	15.48	0.4	1.75		**true class = 0**	15.45	0.51	0.94	
true class = 1	0.21	17.16	0.48		**true class = 1**	0.46	17.49	0	
true class = 2	0.17	0	17.35		**true class = 2**	1.16	0	16.99	
	Sensitivity	**Specificity**	**PPV**	**NPV**		**Sensitivity**	**Specificity**	**PPV**	**NPV**
true class = 0	0.878	0.989	0.976	0.942	**true class = 0**	0.914	0.955	0.905	0.96
true class = 1	0.961	0.989	0.977	0.981	**true class = 1**	0.974	0.985	0.972	0.987
true class = 2	0.99	0.937	0.886	0.995	**true class = 2**	0.936	0.973	0.948	0.967

knn - Compactness and Asymmetry Coefficient				
Accuracy = 0.517				
Confusion matrix:	**predicted class = 0**	**predicted class = 1**	**predicted class = 2**	
true class = 0	11.5	4.08	2.02	
true class = 1	6.97	5	5.36	
true class = 2	2.63	4.49	10.95	
	Sensitivity	**Specificity**	**PPV**	**NPV**
true class = 0	0.653	0.729	0.545	0.809
true class = 1	0.289	0.76	0.368	0.687
true class = 2	0.606	0.789	0.597	0.795

Conclusion:

Of the four algorithms explored for classifying seed kernels, perceptron does the worst by far. Of the remaining three, Decision Trees has the highest accuracy when trained on all 7 attributes at 91%. However, specific sub selections of attributes used in K-nearest Neighbour and Naive Bayes can be 93%. There are subtle differences between high accuracy classifiers that can be seen in the confusion matrix derivatives, such as Sensitivity and PPV. A poor model was demonstrated with pair id 13 (Compactness and Asymmetry Coefficient), where the accuracy drops significantly from any other pair. Using groups of more than 2 attributes avoids the drop in accuracy that those two attributes trigger. Because of the small size of the data set, each training and testing of a model was repeated 100 times and the results averaged. This attempts to stabilize the results but does not guarantee that the resulting accuracy is the model's true accuracy. A more extensive data set is needed to provide definitive selections of the optimum model to predict unobserved seed kernels.

Appendix: A


Scatter Plots of Every Attribute Pair
perimeter
area
compactness
area
length of kernel
area
width of kernel
area
asymmetry coefficent
area
length of kernel groove
area
compactness
perimeter
length of kernel
perimeter
width of kernel
perimeter
asymmetry coefficent
perimeter
length of kernel groove
perimeter
length of kernel
compactness
width of kernel
compactness
asymmetry coefficent
compactness
length of kernel groove
compactness
width of kernel
length of kernel
asymmetry coefficent
length of kernel
length of kernel groove
length of kernel
asymmetry coefficent
width of kernel
length of kernel groove
width of kernel
length of kernel groove
asymmetry coefficent