Machine Learning in the Area of Image Analysis and Pattern Recognition

Edward Tolson

Advanced Undergraduate Project

Spring 2001

Table of Contents

Introduction

This project investigates the use of machine learning for image analysis and pattern recognition. Examples are shown using such a system in image content analysis and in making diagnoses and prognoses in the field of healthcare. Given a data set of images with known classifications, a system can predict the classification of new images. As an example, in the field of healthcare, given a data set of fine needle aspirate (FNA) images of breast masses that are each classified as benign or malignant, a new FNA of a breast mass can be classified as benign or malignant.

There are at least two parts to any such system. The first part is an algorithm for creating a feature vector (also known as a data point) given an image. A feature vector consists of several numbers that are measured or calculated from the image. These features are then used by the second part of the system, a machine learning algorithm, to classify unknown feature vectors given a large database of feature vectors whose classifications are known.

These two parts of such a system are not entirely independent – that is the design of the machine learning algorithm may benefit by knowing how the features are extracted from the image, and the feature extracting may be more successful if the type of machine learning algorithm to be used is known. However, in order to limit the scope of this project, only the second part of such a system is explored. That is, this project focuses on developing a system that uses machine learning to classify unknown images given a database of images and classifications, all of which have already been broken down into feature vectors by an image processing algorithm.

The Machine Learning Algorithm

After researching several machine learning algorithms including Bayesian Nets, Decision Trees, Genetic Algorithms, Nearest Neighbors and Neural Nets, I decided to use a form of K-Nearest-Neighbors. The advantages of this type of algorithm are its transparency, quick training time, and the simplicity of outside parameters that must be adjusted for each data set. Furthermore, other researchers have had good results of classifying images with this algorithm, so this type of algorithm seemed like a good choice.

K-Nearest-Neighbors

The algorithm implemented in this project is a variant of the standard K-Nearest-Neighbor algorithm. To classify an unknown data point, K-Nearest Neighbor finds the k *closest* points in the set of training data. The definition of *closest* is discussed below. Of the k closest points, the algorithm returns the majority classification as the predicted classification of the unknown point. If there is a tie for the majority classification of the k closest points, what classification the algorithm returns is unspecified – it may return any of the majority classifications. (Cover and Heart, 1967; Stanfill and Waltz, 1986; Aha, Kibler, and Albert)

Variations on the K-Nearest-Neighbor Algorithm

The Distance Metric

As opposed to the standard Euclidean distance metric, a weighted distance metric is used to determine the distance between two data points. This is done in an attempt to give all features equal influence on the decision. For example, one feature may range in values from 0 to 1000 while another ranges in values from 0 to 1. If straight Euclidean distance were used as the distance metric, the value of the first feature would have a much more significant effect on the distance between points, and hence classification than the second feature. In order to improve on this, the variance of each feature is calculated during training. The Euclidean distance formula is then adjusted using the variance of each feature according to the following formula:

$$\text{Weighted distance} = \text{Square-root}[\ \sum_{f=1}^{F} (\ (\text{point1}(f) - \text{point2}(f))^2 / (\text{variance}(f))\ )\ ]$$

where F is the number of features that each point has, point1(f) is the value of the f^{th} feature of the first point, point2(f) is the value of the f^{th} feature of the second point, and variance(f) is the variance of the f^{th} feature over the set of training points. Hence, this is the Euclidean distance with the weight of each feature normalized according to the variance of that feature in the training data.

Sensitizing the Algorithm to Certain Classifications

The K-Nearest-Neighbors algorithm implemented in this project is further generalized to allow sensitizing to certain classifications over others. In cases where there are only two possible classification types, the algorithm can be adjusted to favor one classification more or less as needed. This can be useful when misclassifications have different costs. For example, in the case of diagnosing a disease, a false negative may be much more harmful than a false positive. Therefore, in this case there may be a desire to allow the algorithm to favor a positive diagnosis.

A parameter j, which works in conjunction with the parameter k discussed above, is added to facilitate this. The value of j must be between 1 and the value of k, inclusive. The system then specifies to the algorithm which classification is "standard" – for example in the disease diagnosis case, the standard classification would be negative, since presumably the majority of the population does not have the disease. Then, when classifying an unknown point, the algorithm collects the k *closest* points to the unknown point as discussed above. If j or more of these k points are not the standard classification, i.e. in this example if j of the k closest points are of the positive classification, then the classification of those j points is returned as the predicted classification of the unknown point. Hence, if j is greater than (k+1)/2, the algorithm will favor the standard classification, and if j is less than (k+1)/2, the algorithm will favor the non-standard classification.

Measuring the Algorithm's Performance

There are several possible ways of measuring the success of a classification algorithm. The first and most straightforward is simply the percentage of correct classifications the algorithm makes. However, this method can be deceiving. Consider, for instance, the classification of a rare disease that is found in one ten-thousandth of the population. Good classifying algorithms for diseases may return a success rate on the order of 95%. Consider, however, a blind classifier that always asserts that the patient does not have the disease. It will classify correctly 99.99% of the time, since the majority of the population does not have the disease. Intuitively, however, it does not seem that a blind classifier of this type is really as useful as this percentage seems to indicate.

Hence, a more reasonable measure of performance is examining both the correct positive classification rate and the correct negative classification rate. In the blind classifier example above, the correct positive classification rate is 0% while the correct negative classification rate is 100%. The fact that the correct positive classification rate of this blind system is 0% seems to better make evident the weakness of that system.

A third measure of an algorithm's ability to correctly classify unknown data is the resulting Receiver Operating Characteristic (ROC) curve. This type of analysis is used when there are only two classifications, for example, normal and abnormal. In this case, the percentage rate at which the algorithm correctly predicts that an abnormal data point is abnormal is known as the detect rate. Similarly, the false alarm rate is the percentage rate at which the algorithm incorrectly classifies normal points as abnormal. An ROC curve is generated by plotting the detect rate versus the false alarm rate while varying the sensitivity towards the two classifications. As discussed above, varying the parameter j varies the sensitivity of the algorithm used in this project. Hence by holding parameter k constant and varying parameter j from 1 to k-1, the sensitivity of the algorithm is varied, and several (depending on the value of k) points of an ROC curve are generated. Note that this differs from a standard ROC curve in that the sensitivity cannot be varied in a continuous way, but only over discrete values of the parameter j.

Using Cross-Validation to Measure Performance

Cross-validation was used to estimate the percentage rates of correct and incorrect classifications that are needed for the analysis mentioned above. Cross-validation is a way of estimating how the algorithm will perform on new, unknown data given a set of data with known classifications. The data set is randomly divided into two subsets – a training set and a testing set. The learning algorithm is then trained using the training set. After the algorithm has been trained, it is then used to predict the classifications of the test data set. Since the data in the test set has known classifications, the known classifications are compared with the predictions made by the algorithm, and the percent classified correctly and incorrectly can be obtained.

In this project, for fixed values of parameters k and j, the percentage rates of correct and incorrect classifications on a data set were obtained through the following process: The data was randomly divided into a training data set that contained 95% of the original data set, and a test data set that contained the remaining 5% of the data points. The algorithm is then trained using the training data, and used to classify the test data. The number of correct and incorrect classifications made by the algorithm per classification type are recorded. This process of dividing and testing is then repeated five hundred times, and the correct and incorrect classification numbers are averaged.

Applications in Image Content Analysis

Visual analysis and pattern recognition can be used to estimate the content of images. The possible uses of such technology range from autonomous robots to car autopilot systems. All of these need a system that can analyze the content of digital images. The example chosen for this project is classifying outdoor images as grass, sky, foliage, cement, a window, a path, or a brick face. The data set is taken from the University of California at Irvine machine learning repository. Details and statistics of the data set are available in Appendix D.

Previous Results of using Machine Learning in Image Content Analysis

Machine leaning techniques have been used in an attempt to automatically detect rooftops in aerial images. In a comparison of detecting rooftops using a nearest neighbors algorithm, a naive Bayesian network algorithm, and a Budds classifier algorithm. In using these three methods on six images, naive Bayes performed slightly better then Nearest Neighbors, which performed significantly better than the Budds classifier method.

Brodley and Utgoff used Decision Tree algorithms for classifying images based on content. Using cross-validation on the same data set used in this project, they claim to have achieved 96.5% accuracy. This level of accuracy is extremely successful in image content analysis. (Brodley and Utgoff, 1992)

Results of this System on Applications in Image Content Analysis

This system seems to perform fairly well at classifying outdoor images from this data set. The results of the cross-validation testing are listed in Appendix A. Note that since there are more than two classification types, the parameter j is not used in the algorithm and ROC curves are not used for analysis. As the parameter k varies through all the odd numbers from one to fifteen, the overall percentage of correct classification varies from 85.20% to 89.20%, which is a strong overall performance. Furthermore, the system does reasonably well at classifying each type of image in particular; there are no types which the algorithm does a poor job of classifying with the exception of "window" as parameter k gets large. On the other hand there are several types at which the system does an outstanding job of classifying. "Brick face", "grass", "sky", and "path" in particular are almost always classified correctly. Overall, the system seems to perform well for most types of images, and could be of use in applications involving image content recognition.

Applications in Health Care

The field of healthcare also has the potential to benefit from visual analysis and pattern recognition using machine learning. Machine learning techniques can be used to analyze MRI's, X-ray's, etc. to aid in diagnosing and making a prognosis. In this project, the machine learning algorithm was used on two sets of data in the area of healthcare, both of which come from images of fine needle aspirates (FNA) of breast masses. The first data set contains points that are made of features extracted from an FNA and are classified with a diagnosis of whether the breast mass was benign or malignant. Hence, a learning algorithm using this data can predict from a new FNA of a breast mass whether the mass is benign or malignant. The second data set is taken from FNA's of malignant breast masses and is classified as to whether the cancer was recurrent or non-recurrent. Therefore a learning algorithm using this data can predict from an FNA whether a malignant tumor will recur. Both of these data sets are taken from the University of

California at Irvine machine learning repository. Details and statistics of the data sets can be found in Appendix D.

Previous Results of using FNA Images to make Diagnoses and Prognoses

A system called Xcyt uses the same techniques of machine learning analyzing FNA images to diagnose a breast tumor as benign or malignant. This Xcyt system uses linear programming to look for simple planar dividing classifiers. Used on the same data used in this project, Xcyt found a single plane separator in the three-dimensional space of three of the features (worst nucleus area, worst nucleus smoothness, and mean texture). Using cross validation, the predictive accuracy of the Xcyt system was 97.5%, with a detect rate of 96.7% and specificity rate of 98.0%. (Street, 2000)

Making a prognosis as to whether or not a malignant tumor will recur is by nature a much more difficult task. Research has been done into a variety of different machine learning algorithms for this task, including decision trees, median-based separation, and artificial neural networks (Wolberg, 1994; Street, 2000; Ripley, 1998). An artificial neural network system has been used on a different set of data to achieve the same purpose. The best predictive percentage achieved by the artificial neural net system was 79% accuracy. (Ripley, 1998)

Results of this System on Applications in Healthcare

The system did an extremely good job at classifying the FNA images as either benign or malignant. The cross-validation results and ROC plots are recorded in Appendix B. The algorithm used in this project with parameter k set to 15 and parameter j set to 7 yielded an estimated correct classification percentage of 97.34%, with a detect rate of 95.16% and a specificity rate of 98.82%. These results are very competitive with the results of the Xcyt system as described in the previous section. Moreover, over a wide range of values of k, the correct classification percentages are very high, and the ROC curves are close to ideal. This shows that such a system could be very helpful in identifying breast masses as benign or malignant from FNA images.

The system's performance on the task of deciding whether malignant tumors are recurrent or non-recurrent is less successful. The results are completely detailed in Appendix C. The system struggled in this area, though for a handful of parameters the true negative rate and the true positive rate are both greater than 50%, showing that the system's performance is certainly better than random guessing. Perhaps the best-balanced choice of parameters for this set of data is parameter k set to 2 and parameter j set to 1. With these settings, the system correctly classifies 59.55% of non-recurrent tumors and 54.92% of recurrent tumors. However, this system is not as accurate as the artificial neural network system described in the previous section.

Conclusions

The results of this project show that feature extraction and machine learning are a viable approach to visual analysis. The variant of the Nearest-Neighbor algorithm used in this project seems to work extremely well in breast mass diagnosing, and the parameter j provides a helpful way of adjusting sensitivity. Those in the healthcare field know the associated costs of misdiagnosis in both directions, and hence could adjust the algorithm's sensitivity accordingly. Though the example of determining whether a malignant tumor would recur or not was not very successful, the results were better than random guessing and it is possible that it is simply very difficult to make such a determination ahead of time. The recognition of the content of outdoor images was not as successful as the tumor diagnosing, but successful enough to warrant more attention and research. Moreover, a different machine learning algorithm or feature extraction algorithm may improve results.

There are many fields and uses in which systems that analyze images and recognize patterns could have much benefit. From high-tech uses to healthcare, such a system can benefit the community and improve quality of life. It is clear that feature extraction and machine learning algorithms provide a viable approach to creating such a system.

Appendix A – The Results of Cross-Validation in the Case of Identifying Outdoor Picture Contents

Results with the Parameter K Set to 1

Overall Correct Classification Rate	Brick Face	Cement	Foliage	Grass	Path	Sky	Window
86.67%	94.20%	71.26%	62.42%	100.00%	99.47%	100.00%	68.34%

Table 1: Results of identifying picture contents by type with parameter k set to 1

Results with the Parameter K Set to 3

Overall Correct Classification Rate	Brick Face	Cement	Foliage	Grass	Path	Sky	Window
87.56%	94.20%	71.33%	66.80%	100.00%	99.58%	100.00%	69.59%

Table 2: Results of identifying picture contents by type with parameter k set to 3

Results with the Parameter K Set to 5

Overall Correct Classification Rate	Brick Face	Cement	Foliage	Grass	Path	Sky	Window
88.31%	96.78%	71.15%	71.24%	100.00%	98.70%	100.00%	69.62%

Table 3: Results of identifying picture contents by type with parameter k set to 5

Results with the Parameter K Set to 7

Overall Correct Classification Rate	Brick Face	Cement	Foliage	Grass	Path	Sky	Window
89.20%	99.45%	70.03%	75.60%	100.00%	99.84%	100.00%	67.31%

Table 4: Results of identifying picture contents by type with parameter k set to 7

Results with the Parameter K Set to 9

Overall Correct Classification Rate	Brick Face	Cement	Foliage	Grass	Path	Sky	Window
88.89%	100.00%	72.57%	73.23%	100.00%	100.00%	100.00%	68.31%

Table 5: Results of identifying picture contents by type with parameter k set to 9

Results with the Parameter K Set to 11

Overall Correct Classification Rate	Brick Face	Cement	Foliage	Grass	Path	Sky	Window
87.84%	100.00%	65.42%	73.96%	100.00%	99.92%	100.00%	59.36%

Table 6: Results of identifying picture contents by type with parameter k set to 11

Results with the Parameter K Set to 13

Overall Correct Classification Rate	Brick Face	Cement	Foliage	Grass	Path	Sky	Window
86.34%	100.00%	67.46%	68.81%	100.00%	100.00%	100.00%	46.73%

Table 7: Results of identifying picture contents by type with parameter k set to 13

Results with the Parameter K Set to 15

Overall Correct Classification Rate	Brick Face	Cement	Foliage	Grass	Path	Sky	Window
85.20%	99.74%	68.65%	64.16%	100.00%	100.00%	100.00%	39.49%

Table 8: Results of identifying picture contents by type with parameter k set to 15

Appendix B – The Results of Cross-Validation in the Case of Classifying a Breast Mass as Benign or Malignant

Results with the Parameter K Set to 1

J	Overall Correct Classification Rate	True Negative Rate	Detect Rate
1	95.40 %	93.41 %	96.79 %

Table 9: Results of classifying breast mass FNA as malignant with parameter k set to 1

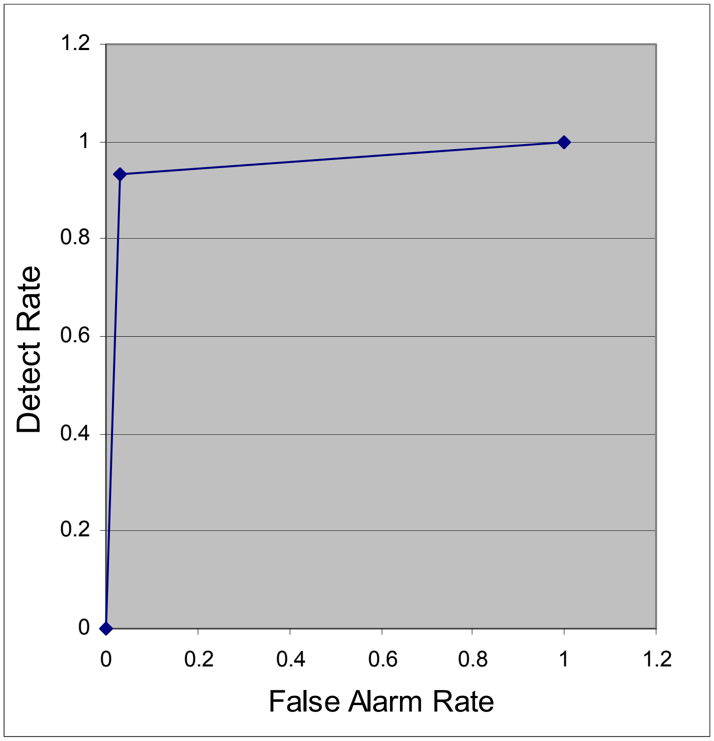

Figure 1: ROC curve of diagnosing a breast mass with parameter k set to 1

Results with the Parameter K Set to 2

J	Overall Correct Classification Rate	True Negative Rate	Detect Rate
1	94.80 %	94.33 %	95.50 %
2	95.79 %	99.76 %	89.89 %

Table 10: Results of classifying breast mass FNA as malignant with parameter k set to 2

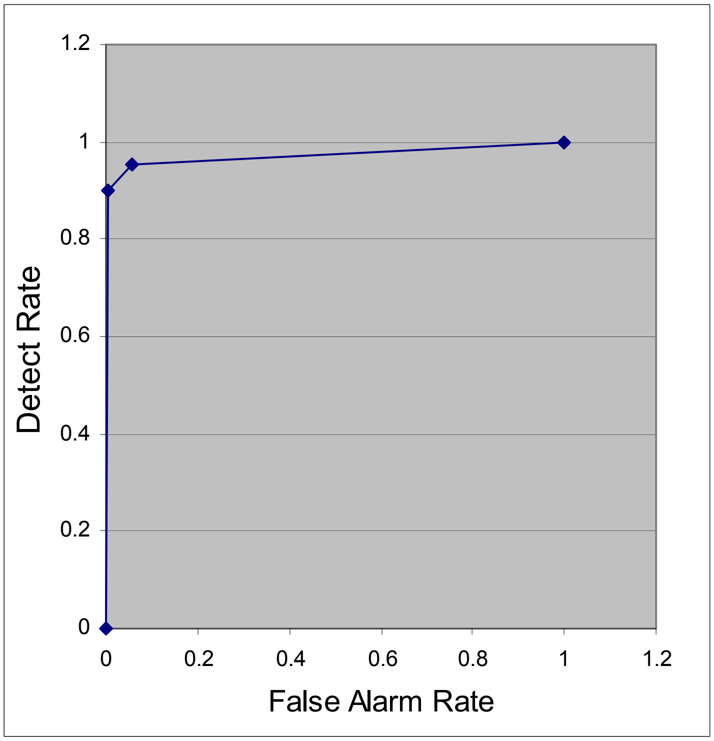

Figure 2: ROC curve of diagnosing a breast mass with parameter k set to 2

Results with the Parameter K Set to 3

J	Overall Correct Classification Rate	True Negative Rate	Detect Rate
1	94.06 %	91.76 %	97.54 %
2	96.70 %	98.90 %	93.50 %
3	94.63 %	100.00 %	86.95 %

Table 11: Results of classifying breast mass FNA as malignant with parameter k set to 3

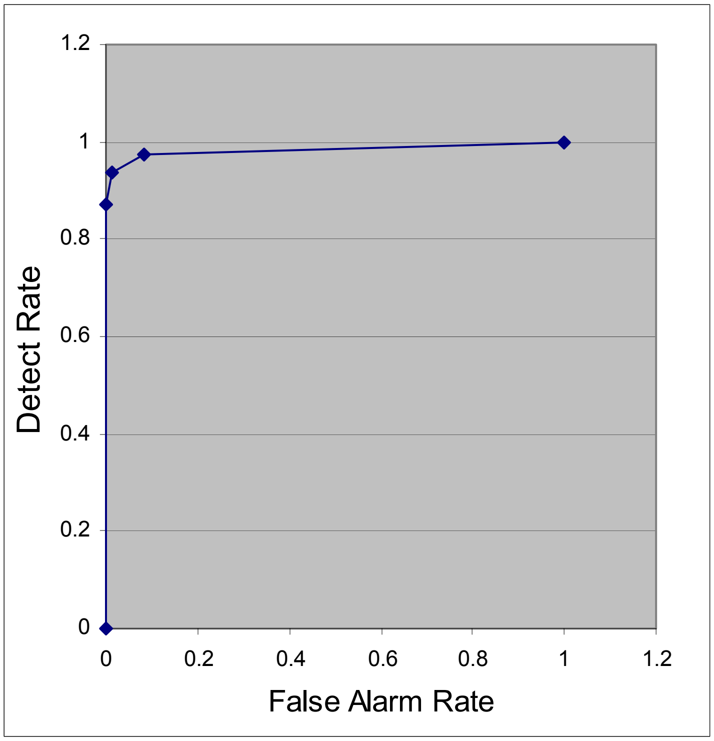

Figure 3: ROC curve of diagnosing a breast mass with parameter k set to 3

Results with the Parameter K Set to 4

J	Overall Correct Classification Rate	True Negative Rate	Detect Rate
1	93.39 %	90.17 %	98.10 %
2	97.04 %	98.10 %	95.48 %
3	97.10 %	99.74 %	93.14 %
4	93.22 %	99.99 %	83.23 %

Table 12: Results of classifying breast mass FNA as malignant with parameter k set to 4

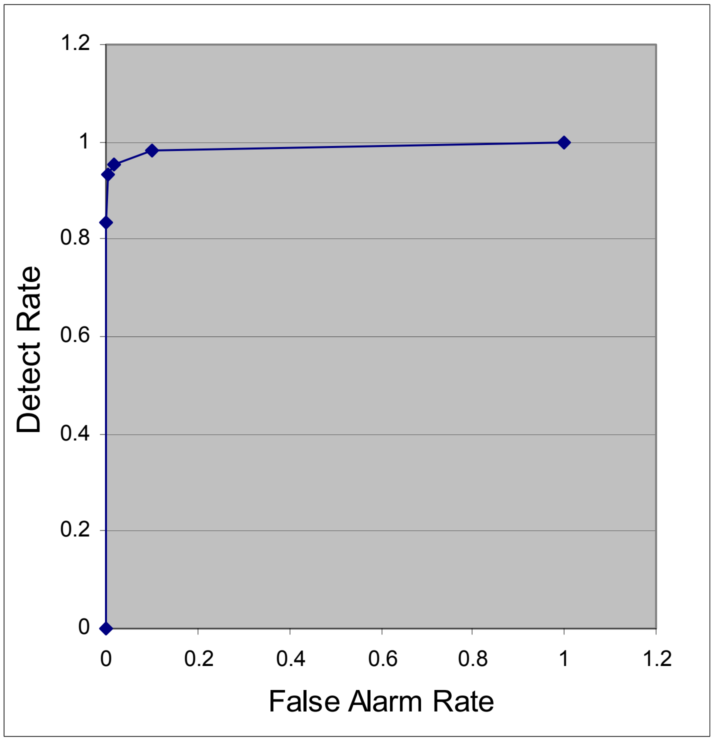

Figure 4: ROC curve of diagnosing a breast mass with parameter k set to 4

Results with the Parameter K Set to 5

J	Overall Correct Classification Rate	True Negative Rate	Detect Rate
1	92.23 %	87.96 %	98.53 %
2	96.51 %	96.60 %	96.37 %
3	96.94 %	99.15 %	93.77 %
4	95.17 %	99.64 %	88.73 %
5	91.63 %	100.00 %	79.46 %

Table 13: Results of classifying breast mass FNA as malignant with parameter k set to 5

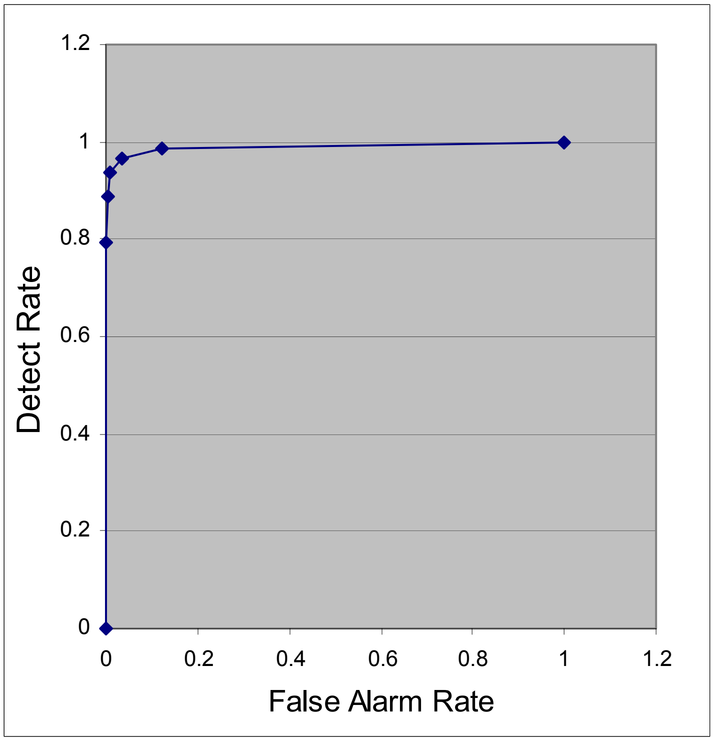

Figure 5: ROC curve of diagnosing a breast mass with parameter k set to 5

Results with the Parameter K Set to 7

J	Overall Correct Classification Rate	True Negative Rate	Detect Rate
1	89.90 %	83.91 %	98.57 %
2	95.67 %	94.96 %	96.69 %
3	96.75 %	97.52 %	95.63 %
4	96.57 %	99.23 %	92.76 %
5	95.16 %	99.70 %	88.48 %
6	93.03 %	99.68 %	83.21 %
7	90.04 %	99.99 %	75.41 %

Table 14: Results of classifying breast mass FNA as malignant with parameter k set to 7

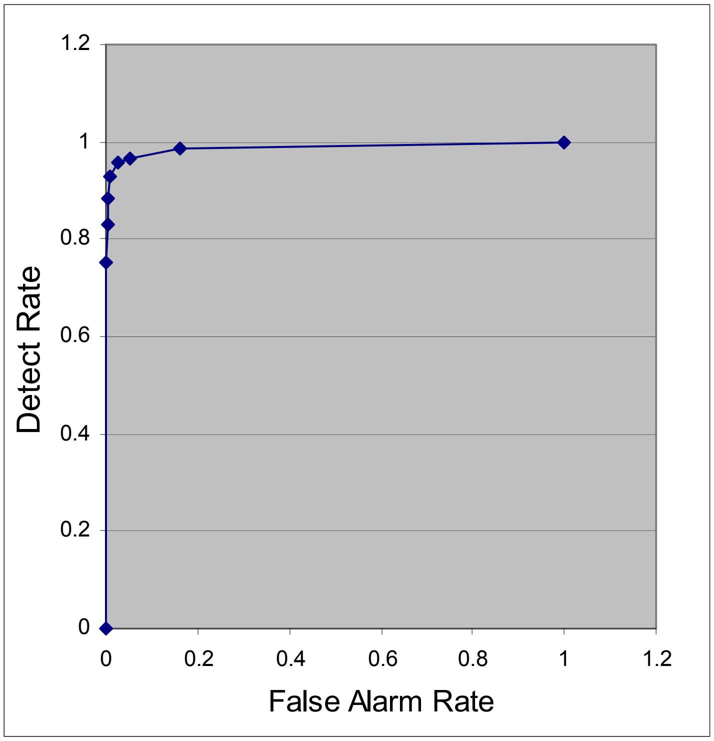

Figure 6: ROC curve of diagnosing a breast mass with parameter k set to 7

Results with the Parameter K Set to 10

J	Overall Correct Classification Rate	True Negative Rate	Detect Rate
1	87.24 %	78.75 %	99.23 %
2	93.50 %	91.00 %	97.15 %
3	95.56 %	95.12 %	96.21 %
4	96.50 %	97.61 %	94.86 %
5	97.04 %	99.07 %	94.10 %
6	96.81 %	99.44 %	92.97 %
7	94.42 %	99.60 %	86.89 %
8	93.34 %	99.84 %	84.05 %
9	91.00 %	100.00 %	78.03 %
10	88.42 %	100.00 %	71.47 %

Table 15: Results of classifying breast mass FNA as malignant with parameter k set to 10

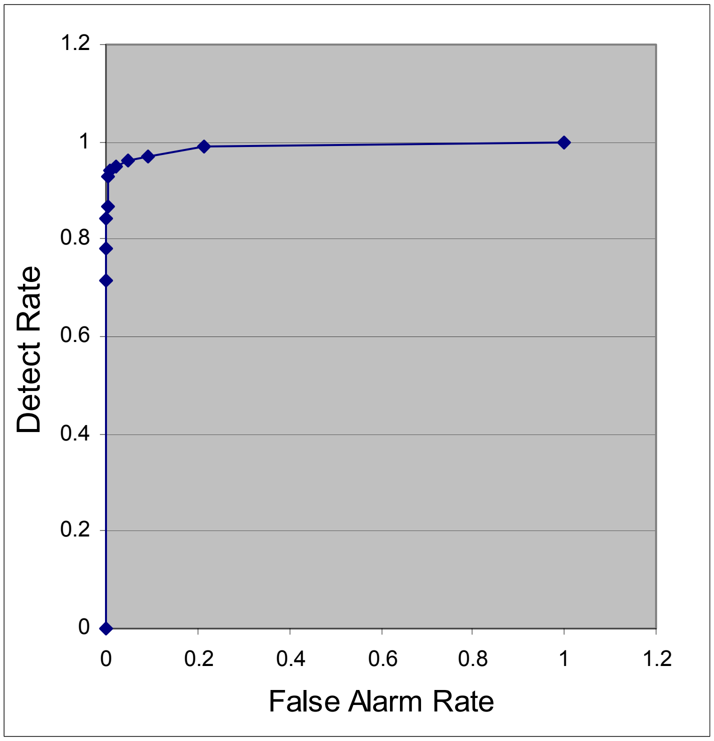

Figure 7: ROC curve of diagnosing a breast mass with parameter k set to 10

Results with the Parameter K Set to 15

J	Overall Correct Classification Rate	True Negative Rate	Detect Rate
1	83.36 %	72.23 %	99.56 %
2	90.90 %	85.25 %	99.22 %
3	93.46 %	90.44 %	97.88 %
4	95.06 %	93.89 %	96.77 %
5	96.22 %	96.21 %	96.24 %
6	96.59 %	97.70 %	94.94 %
7	97.34 %	98.82 %	95.16 %
8	96.37 %	99.38 %	91.99 %
9	95.96 %	99.56 %	90.65 %
10	95.72 %	99.80 %	89.96 %
11	93.66 %	99.79 %	84.59 %
12	91.64 %	99.95 %	79.34 %
13	90.57 %	100.00 %	77.06 %
14	88.03 %	100.00 %	70.84 %
15	85.31 %	100.00 %	63.12 %

Table 16: Results of classifying breast mass FNA as malignant with parameter k set to 15

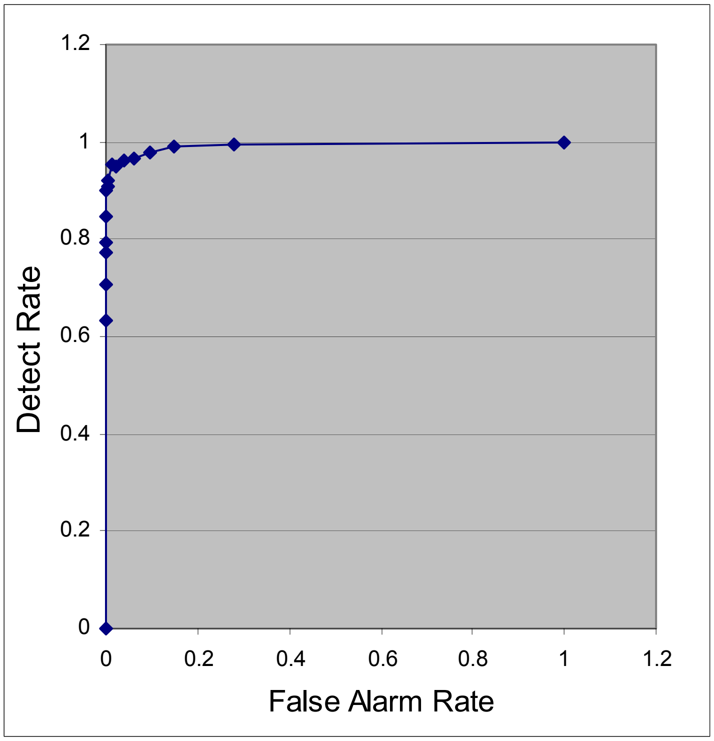

Figure 8: ROC curve of diagnosing a breast mass with parameter k set to 15

Appendix C – The Results of Cross-Validation in the Case of Classifying a Malignant Breast Mass as Recurrent or Non-Recurrent

Results with the Parameter K Set to 1

J	Overall Correct Classification Rate	True Negative Rate	Detect Rate
1	73.16 %	80.83 %	49.09 %

Table 17: Results of classifying an FNA as recurrent with parameter k set to 1

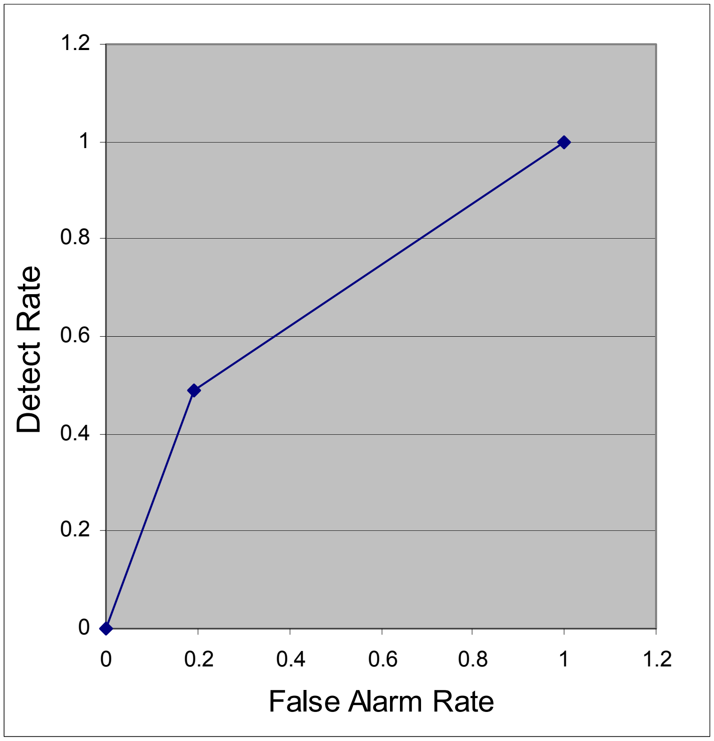

Figure 9: ROC curve of predicting the recurrence of a malignant breast mass with parameter k set to 1

Results with the Parameter K Set to 2

J	Overall Correct Classification Rate	True Negative Rate	Detect Rate
1	58.38 %	59.55 %	54.92 %
2	78.04 %	96.03 %	24.02 %

Table 18: Results of classifying an FNA as recurrent with parameter k set to 2

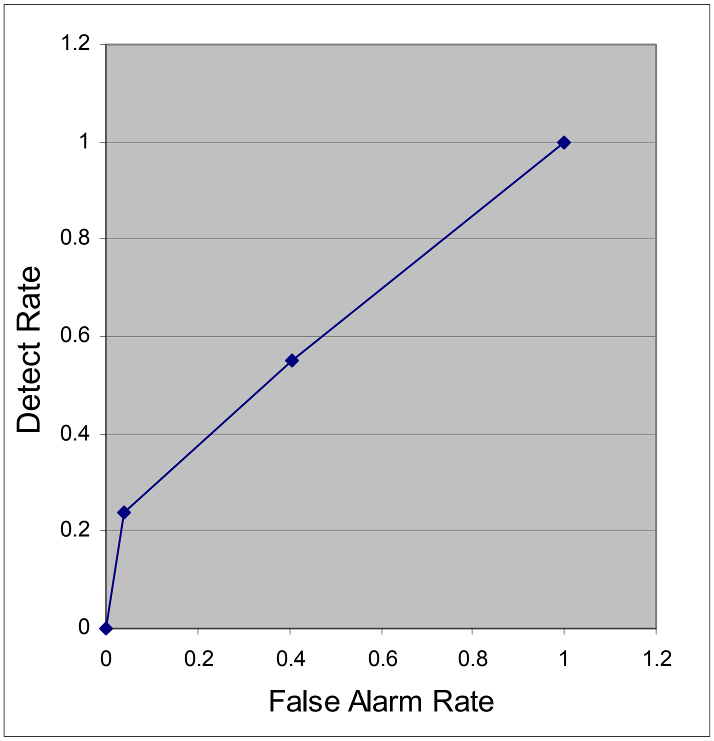

Figure 10: ROC curve of predicting the recurrence of a malignant breast mass with parameter k set to 2

Results with the Parameter K Set to 3

J	Overall Correct Classification Rate	True Negative Rate	Detect Rate
1	44.02 %	39.48 %	58.37 %
2	73.08 %	84.99 %	35.15 %
3	75.84 %	99.31 %	5.14 %

Table 19: Results of classifying an FNA as recurrent with parameter k set to 3

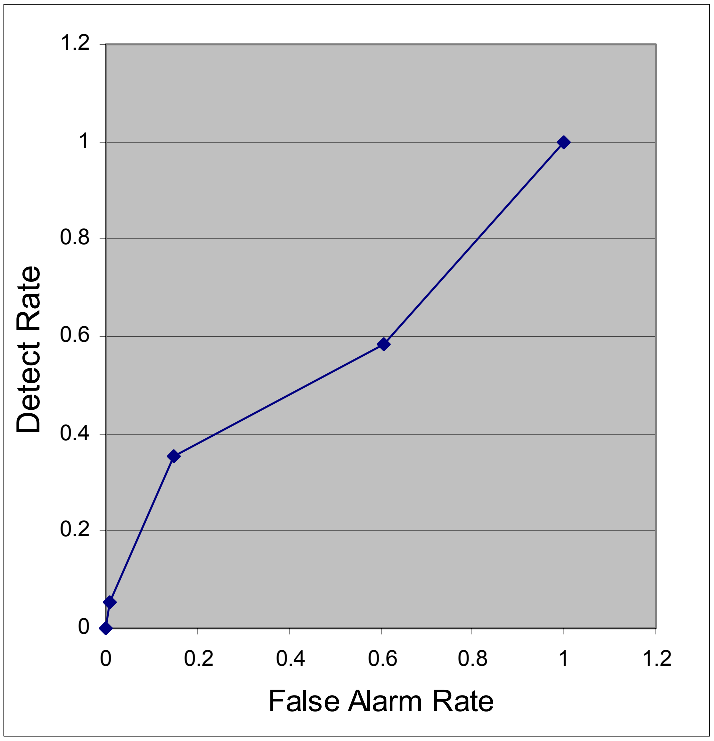

Figure 11: ROC curve of predicting the recurrence of a malignant breast mass with parameter k set to 3

Results with the Parameter K Set to 4

J	Overall Correct Classification Rate	True Negative Rate	Detect Rate
1	37.72 %	27.77 %	67.99 %
2	67.14 %	75.57 %	40.81 %
3	73.46 %	93.62 %	11.27 %
4	75.54 %	99.97 %	0.08 %

Table 20: Results of classifying an FNA as recurrent with parameter k set to 4

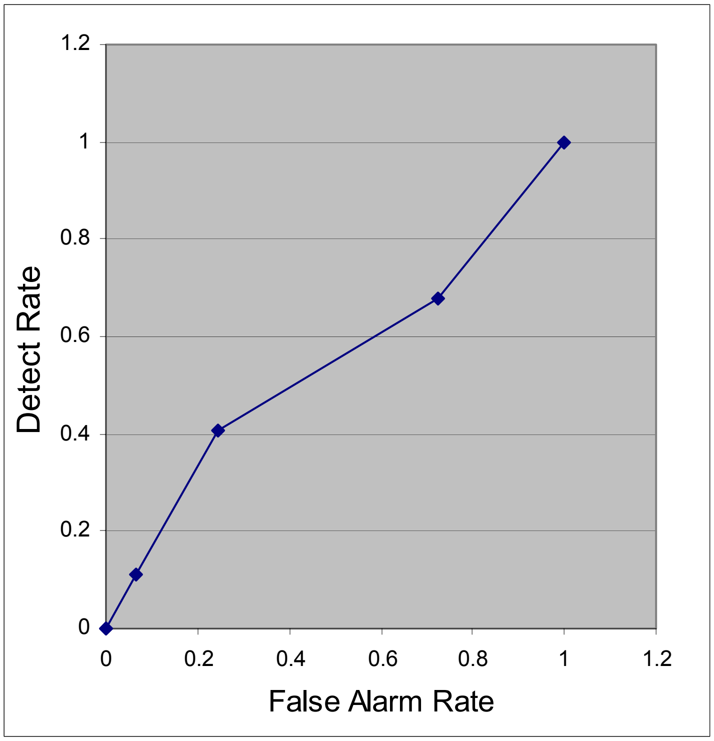

Figure 12: ROC curve of predicting the recurrence of a malignant breast mass with parameter k set to 4

Results with the Parameter K Set to 5

J	Overall Correct Classification Rate	True Negative Rate	Detect Rate
1	33.22 %	19.08 %	74.78 %
2	61.90 %	65.72 %	50.16 %
3	75.96 %	89.99 %	33.95 %
4	74.82 %	98.56 %	2.67 %
5	75.88 %	100.00 %	0.08 %

Table 21: Results of classifying an FNA as recurrent with parameter k set to 5

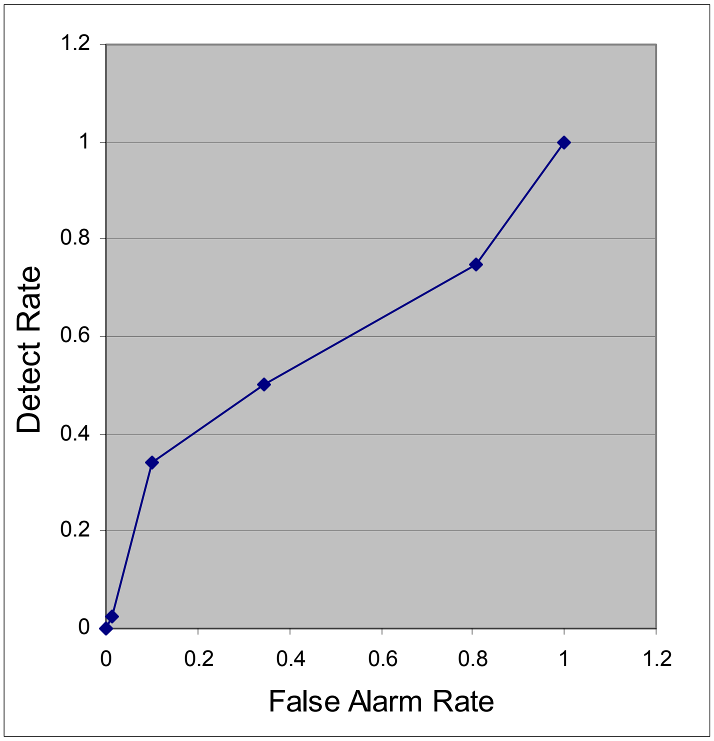

Figure 13: ROC curve of predicting the recurrence of a malignant breast mass with parameter k set to 5

Results with the Parameter K Set to 7

J	Overall Correct Classification Rate	True Negative Rate	Detect Rate
1	29.84 %	11.97 %	84.90 %
2	51.20 %	47.94 %	60.87 %
3	69.46 %	79.71 %	38.39 %
4	78.96 %	94.95 %	28.15 %
5	78.82 %	98.57 %	17.42 %
6	75.54 %	99.97 %	0.08 %
7	74.24 %	100.00 %	0.00 %

Table 22: Results of classifying an FNA as recurrent with parameter k set to 7

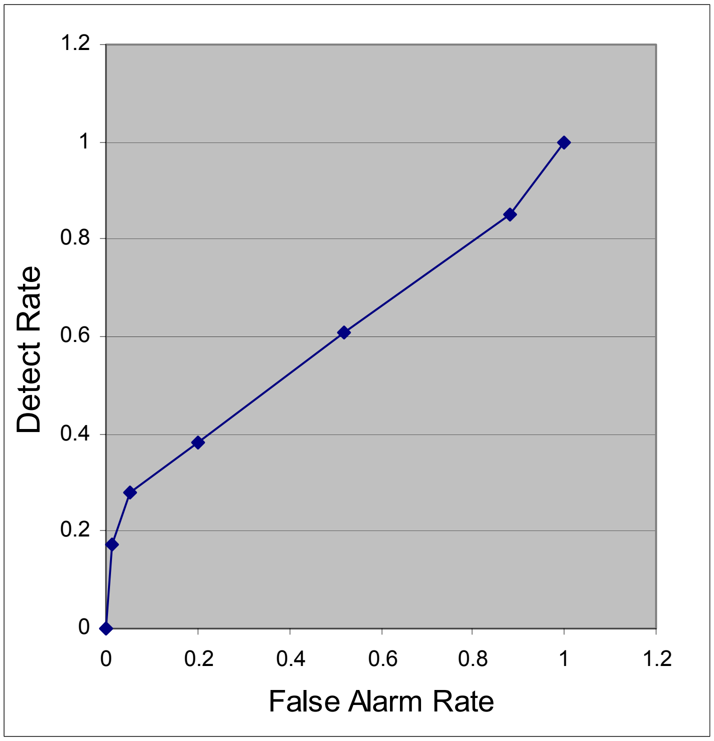

Figure 14: ROC curve of predicting the recurrence of a malignant breast mass with parameter k set to 7

Results with the Parameter K Set to 10

J	Overall Correct Classification Rate	True Negative Rate	Detect Rate
1	26.48 %	5.79 %	92.15 %
2	42.64 %	32.82 %	72.90 %
3	57.58 %	59.17 %	52.78 %
4	74.90 %	85.92 %	41.64 %
5	79.46 %	95.07 %	30.29 %
6	76.62 %	98.60 %	9.05 %
7	75.80 %	99.58 %	1.32 %
8	75.18 %	99.97 %	0.00 %
9	74.50 %	100.00 %	0.00 %
10	75.12 %	100.00 %	0.00 %

Table 23: Results of classifying an FNA as recurrent with parameter k set to 10

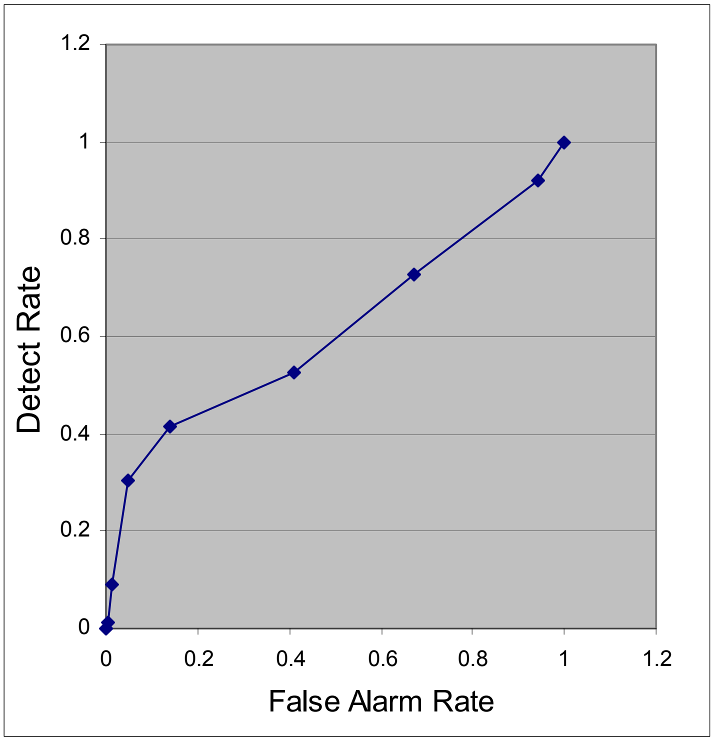

Figure 15: ROC curve of predicting the recurrence of a malignant breast mass with parameter k set to 10

Results with the Parameter K Set to 15

J	Overall Correct Classification Rate	True Negative Rate	Detect Rate
1	25.62 %	1.77 %	99.92 %
2	35.42 %	15.40 %	96.59 %
3	43.50 %	33.29 %	73.33 %
4	56.72 %	57.74 %	53.61 %
5	69.68 %	78.91 %	42.36 %
6	75.50 %	90.56 %	31.37 %
7	74.08 %	94.59 %	10.93 %
8	73.50 %	97.85 %	3.04 %
9	74.75 %	99.23 %	0.32 %
10	75.74 %	99.95 %	0.00 %
11	75.68 %	100.00 %	0.00 %
12	76.16 %	100.00 %	0.00 %
13	75.84 %	100.00 %	0.00 %
14	74.32 %	100.00 %	0.00 %
15	74.64 %	100.00 %	0.00 %

Table 24: Results of classifying an FNA as recurrent with parameter k set to 15

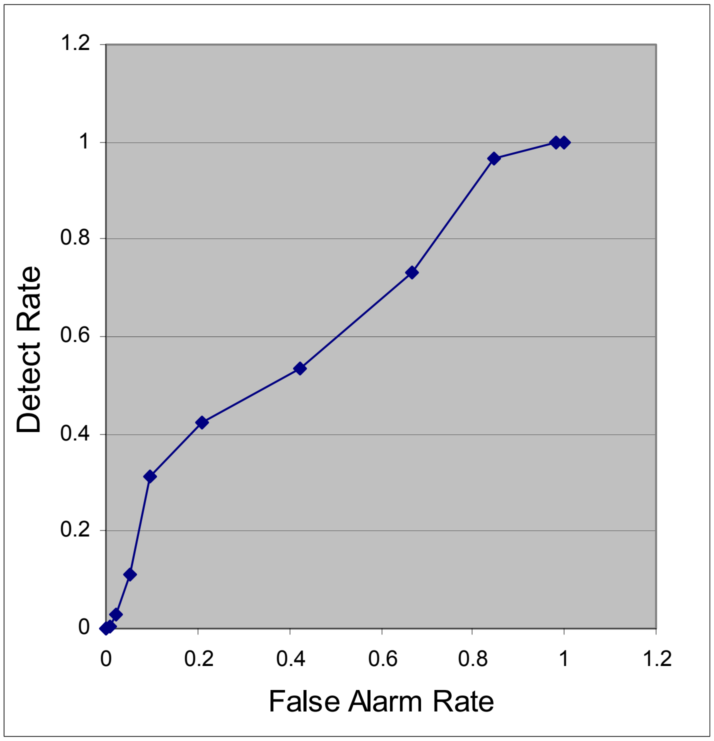

Figure 16: ROC curve of predicting the recurrence of a malignant breast mass with parameter k set to 15

Appendix D – The Data Sets

All three of the data sets used in this project are taken from the University of California at Irvine machine learning database repository, available online at http://www.ics.uci.edu/~mlearn/MLRepository.html. All the statistics and feature information listed in this appendix are taken from the information files that come with each data set.

Classifying the Content of Outdoor Images

This data set is referred to in the repository as the Image Segmentation Database training data file. The images used in the data set were chosen randomly, and then hand-segmented. This data set contains 210 entries, thirty entries in each of the following seven classes: brick face, cement, foliage, grass, path, sky, and window. Each entry contains the following nineteen features:

1) The column of the center pixel of the region
2) The row of the center pixel of the region
3) The number of pixels in the region (this feature is the same for all the images)
4) The results of a line extraction algorithm that counts how many lines of length 5 (any orientation) with low contrast, less than or equal to 5, go through the region
5) The results of a line extraction algorithm that counts how many lines of length 5 (any orientation) with high contrast, greater than 5, go through the region
6) Mean of the contrast of horizontally adjacent pixels in the region
7) Standard deviation of the contrast of horizontally adjacent pixels in the region
8) Mean of the contrast of vertically adjacent pixels
9) Standard deviation of the contrast of vertically adjacent pixels
10) The average intensity
11) The average red color value of each pixel
12) The average blue color value of each pixel
13) The average green color value of each pixel
14) Measure of the excess red (2*Red – (Green + Blue))
15) Measure of the excess blue (2*Blue – (Red + Green))
16) Measure of the excess green (2*Green – (Red + Blue))
17) 3-d nonlinear transformation of RGB (Algorithm can be found in Foley and VanDam, Fundamentals of Interactive Computer Graphics)
18) Saturation mean (Algorithm can be found in Foley and VanDam, Fundamentals of Interactive Computer Graphics)
19) Hue mean (Algorithm can be found in Foley and VanDam, Fundamentals of Interactive Computer Graphics)

Diagnosing a Breast Lump as Benign or Malignant

This data set is referred to in the repository as the Wisconsin Breast Cancer Diagnostic Database. Each entry in the data set contains thirty features that are extracted from a digitized FNA image along with a classification of benign or malignant. To create the thirty features, the following ten features are calculated for each cell nucleus in the mass:

1) The length of the radius (the radius is calculated as the mean distance from the center to points on the perimeter)
2) The texture (calculated as the standard deviation of the gray-scale values)
3) The length of the perimeter
4) The area
5) The smoothness (calculated using the local variation in radius length)
6) The compactness (calculated as perimeter^2/(area-1.0))
7) The concavity (calculated using the severity of the concave portions of the perimeter)
8) The number of concave points along the perimeter
9) The symmetry
10) Fractal dimension (calculated as the "coastline approximation" – 1.0)

For each of these ten features, the mean, standard error, and worst or largest single value each become a feature in the final feature vector. As an example, the average radius length, standard error of the radius length, and the largest single radius are each a feature in the final feature vector. In this way, the thirty features of each FNA are computed.

The data set contains 569 entries, of which 357 are benign and 212 are malignant.

Deciding Whether Malignant Tumor will be Recurrent

This data set is referred to in the repository as the Wisconsin Breast Cancer Diagnostic Database. Each entry in the data sets contains thirty-three features, thirty of which are extracted from a digitized FNA image, along with a classification of recurrent or non-recurrent. To create the thirty features that are taken from the FNA image, the following ten features are calculated for each cell nucleus in the mass:

1) The length of the radius (the radius is calculated as the mean distance from the center to points on the perimeter)
2) The texture (calculated as the standard deviation of the gray-scale values)
3) The length of the perimeter
4) The area
5) The smoothness (calculated using the local variation in radius length)
6) The compactness (calculated as perimeter^2/(area-1.0))
7) The concavity (calculated using the severity of the concave portions of the perimeter)
8) The number of concave points along the perimeter
9) The symmetry
10) Fractal dimension (calculated as the "coastline approximation" – 1.0)

For each of these ten features, the mean, standard error, and worst or largest single value each become a feature in the final feature vector. As an example, the average radius length, standard error of the radius length, and the largest single radius are each a feature in the final feature vector. In this way, the thirty features of each FNA are computed. Each entry also contains the following three features that do not come from the FNA:

1) The diameter of the excised tumor in centimeters
2) The number of positive auxiliary lymph nodes observed at the time of surgery
3) A time factor – if the tumor was recurrent, the length of time until it recurred; if the tumor has not recurred, the length of time since the removal surgery

Since this third feature, the time factor, is not known until after the time a prognosis would be useful, it was ignored and not treated as a feature in this project.

The data set contains 198 entries, of which 151 are non-recurrent and 47 are recurrent.

References

Aha, D. W., Kibler, D., and Albert, M. K. 1991. Instance-based learning algorithms. *Machine Learning, 6,* 37-66.

Brodley, C. E., and Utgoff, P. E. 1992. Multivariate Decision Trees. *Machine Learning, 19,* 45-77.

Cover, T. M., and Heart, P. E. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13,* 21-27.

Ripley, R.M. (1998), *Neural Networks for Breast Cancer Prognosis*, Ph.D. Thesis, Department of Engineering Science, University of Oxford.

Stanfill, C., and Waltz, D. 1986. Toward memory-based reasoning. *Communications of the ACM, 29,* 1213-1228.

Street, W. N. 2000. Xcyt: A system for remote cytological diagnosis and prognosis of breast cancer. *Soft Computing Techniques in Breast Cancer Prognosis and Diagnosis,* World Scientific Publishing.

Wolberg, W.H., Street, W.N. and Mangasarian, O.L. (1994), "Machine Learning Techniques to Diagnose Breast Cancer from Image-Processed Nuclear Features of Fine Needle Aspirates," *Cancer Letters*, vol. 77, pp. 163-171.