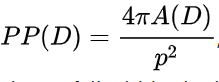
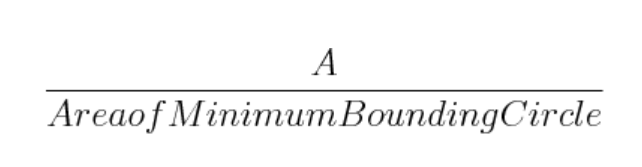
**Feature used to classify:**

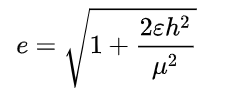
1. Contour Area
2. Contour Perimeter
3. Minimum Bounding rectangle width
4. Polsby-Popper Test for hull area compactness



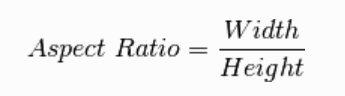
1. Minimum bounding circle radium
2. Reock Score



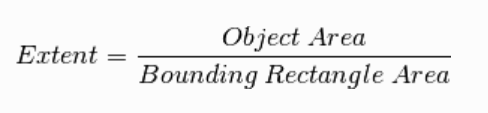
1. Eccentricity Closing ellipse maximum axis and minimum axis



1. Contour Aspect Ratio(Rectangularity)



1. Extent(Rectangularity)



1. Solidity



Above is contour shape feature.

From 11 to 19 I consist of color feature in selected contour:

11-13: splited mean R,G,B intensity value

14-16: splited mean H,S,V intensity value

17-19: splited mean L,A,B intensity value

**Feature threshold classifier manually:**

|  |  |  |  |
| --- | --- | --- | --- |
| Frame Label | Num of Frame | Correctly detected | Accuracy |
| Normal | 3645 | 2960 | 81.2% |
| Water | 148 | 129 | 87.2% |
| Reflecxtion | 236 | 216 | 91.53% |
| Total | 4029 | 3305 | 82.03% |

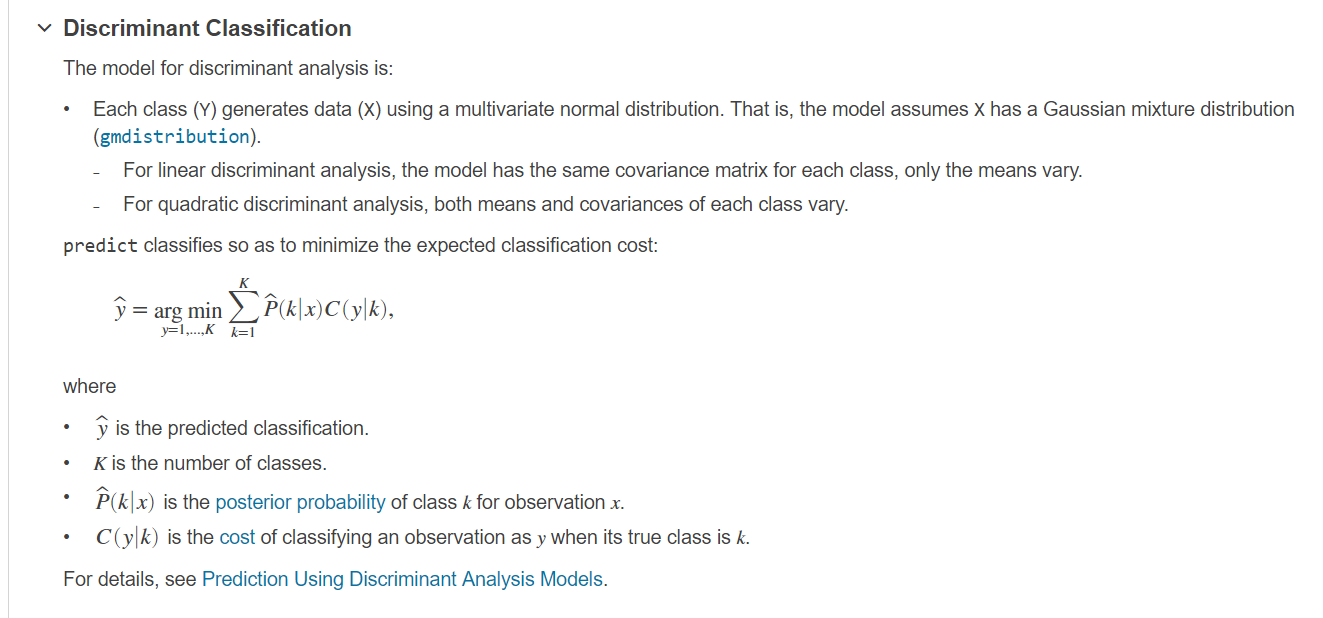
**Classifier comparison based on different machine learning models:**

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | All feature used | Color feature used | Shape feature used |
| Tree | 99.4% | 99.3% | 96.6% |
| Linear Discriminant | 99.5% | 99.5% | 94.7% |
| Quadratic Discriminant | 62.4% | 99.5% | 98.0% |
| Logistic Regression | 99.5% | 99.4% | 94.1% |
| SVM | 99.4% | 99.6% | 99.1% |
| KNN | 99.5% | 99.5% | 98.2% |

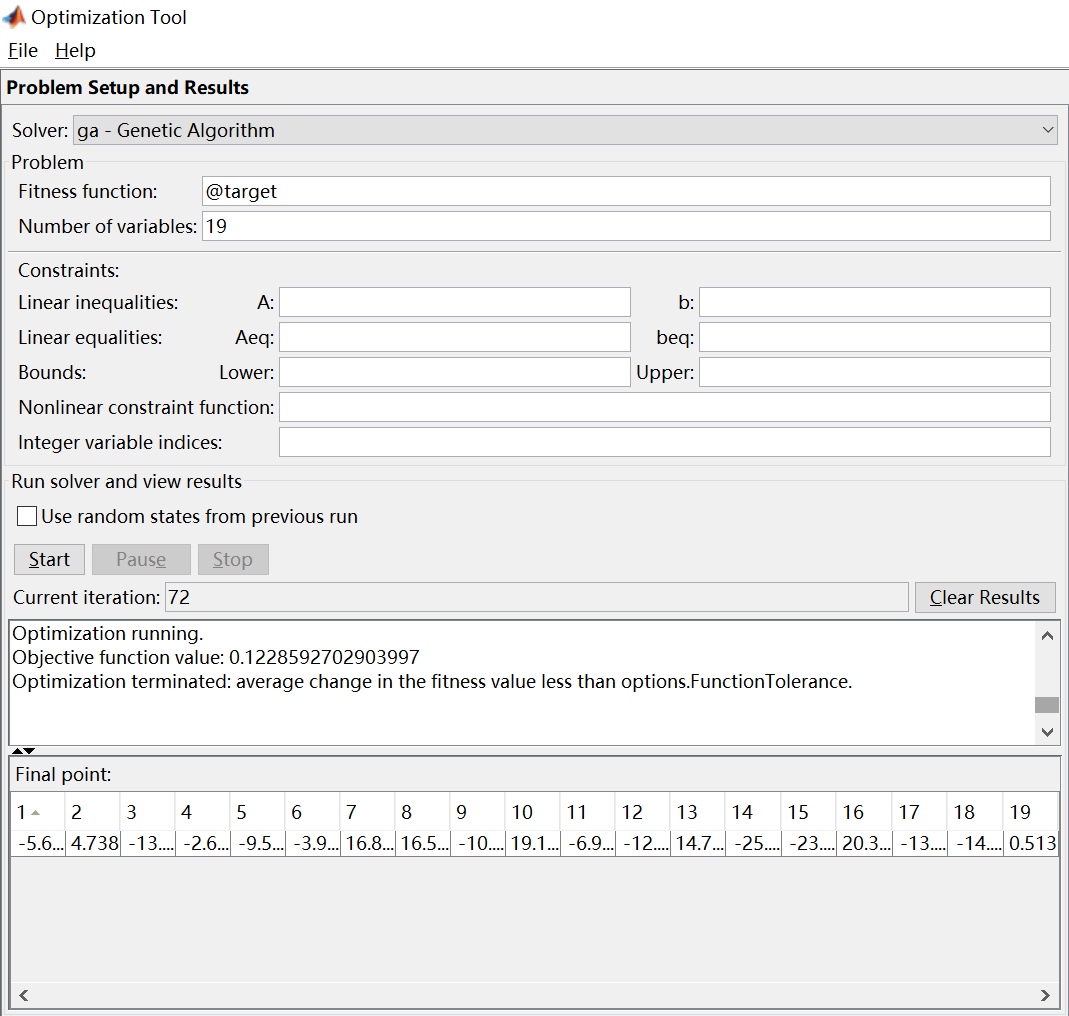
The strange thing is when using Quadratic Discriminant analysis to do classification, result is good when using only color or shape feature but accuracy drops much more and really low when using all feature. So next step using genetic algorithm to optimize(GA algorithm).

**Feature Selection based on GA algorithm:**

* Fitness Function Design



Set weight as 19\*1 vector ,use loss function in Quadratic Discriminant classifier as fitness function, so our object is to find global minimum in this loss function. The target function wanna to optimize is weight for every feature, if absolute value is weight on specific feature it will get more influence on classification result. The fitness function can be represented as:

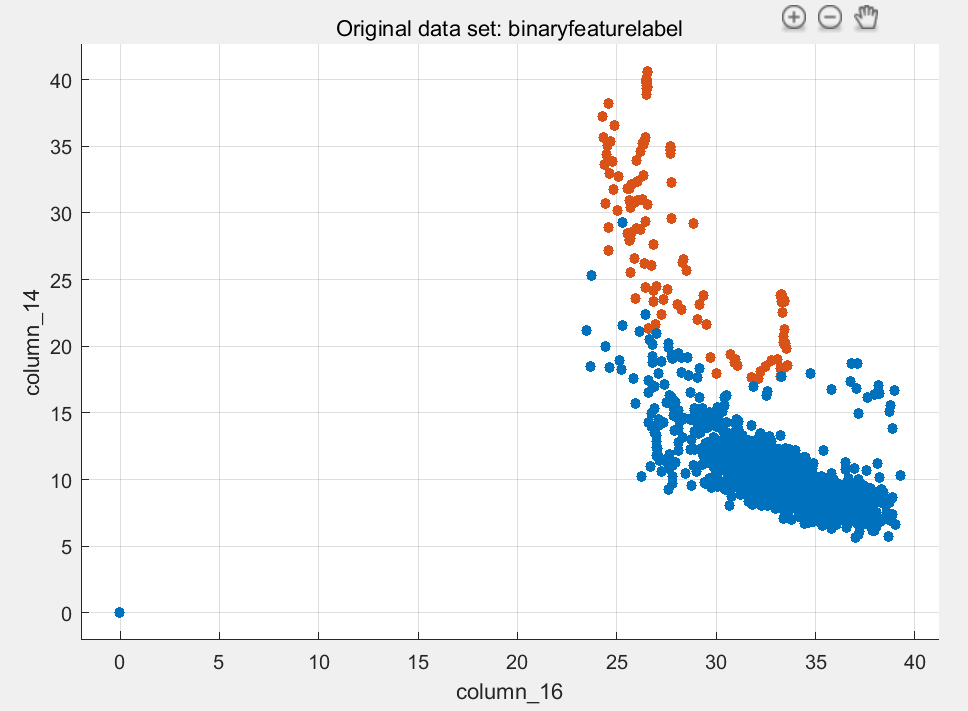


**The accuracy is about 98.8% after optimization by GA algorithm using all features, while no selection on model accuracy is about 62.4%!**

So from weight we can also learn about which feature is important and which one is not, so I only use two high-valuable feature to apply on all machine learning classification models, here is result:

|  |  |  |
| --- | --- | --- |
| Classifier | All feature used | 2 Selected feature used |
| Tree | 99.4% | 99.4% |
| Linear Discriminant | 99.5% | 98.5% |
| Quadratic Discriminant | 62.4% | 98.8% |
| Logistic Regression | 99.5% | 98.5% |
| SVM | 99.4% | 99.4% |
| KNN | 99.5% | 99.3% |

Even two feature instead of all 19 feature can get pretty same results but dimension and time consume is really low. I plot these two features distribution they are apparently linear separable.



**More Testing Result:**

***Reflecxtion2:*** Manually threshold selection classify:

|  |  |  |  |
| --- | --- | --- | --- |
| Frame Label | Num of Frame | Correctly detected | Accuracy |
| Normal | 2400 | 2224 | 92.7% |
| Water | 188 | 182 | 96.8% |
| Reflecxtion | 412 | 233 | 56.6% |
| Total | 3000 | 2639 | 88.0% |

Machine learning model classify:

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | All feature used | Color feature used | Shape feature used |
| Tree | 96.4% | 96.3% | 94.1% |
| Linear Discriminant | 95.9% | 95.2% | 90.5% |
| Quadratic Discriminant | 95.7% | 94.9% | 89.2% |
| Logistic Regression | 95.8% | 95.5% | 90.5% |
| SVM | 96.6% | 96.5% | 94.0% |
| KNN | 39.9% | 39.8% | 37.2% |

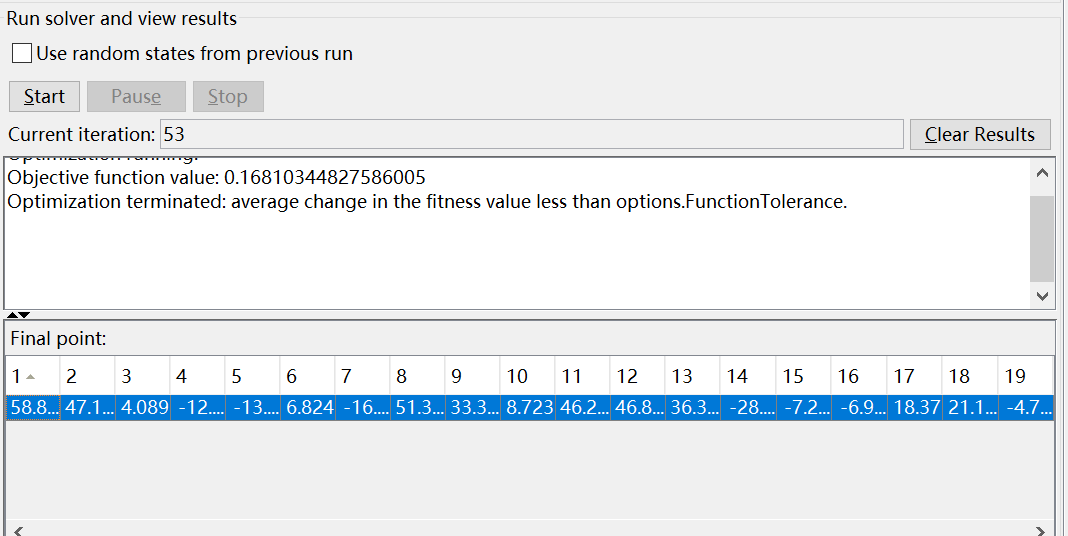
***Reflecxtion 3:*** Manually threshold selection classify:

|  |  |  |  |
| --- | --- | --- | --- |
| Frame Label | Num of Frame | Correctly detected | Accuracy |
| Normal | 282 | 250 | 88.7% |
| Reflecxtion | 182 | 61 | 33.52% |
| Total | 464 | 311 | 67.03% |

Machine learning model classify:

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | All feature used | Color feature used | Shape feature used |
| Tree | 97.6% | 97.6% | 95.9% |
| Linear Discriminant | 97.4% | 97.0% | 92.9% |
| Quadratic Discriminant | 97.6% | 97.4% | 95.7% |
| Logistic Regression | 97.4% | 97.6% | 95.3% |
| SVM | 97.4% | 97.6% | 96.6% |
| KNN | 82.8% | 93.3% | 82.5% |

**GA optimization:**



So selected only 1 2 8 9 11 12 13 14 18

|  |  |  |
| --- | --- | --- |
| Classifier | All feature used | 2 Selected feature used |
| Tree | 97.6% | 97.4% |
| Linear Discriminant | 97.4% | 97.4% |
| Quadratic Discriminant | 97.6% | 97.2% |
| Logistic Regression | 97.4% | 97.4% |
| SVM | 97.4% | 97.4% |
| KNN | 82.8% | 84.3% |

***Biopsy1：***

Manually threshold selection classify:

|  |  |  |  |
| --- | --- | --- | --- |
| Frame Label | Num of Frame | Correctly detected | Accuracy |
| Normal | 1024 | 825 | 80.6% |
| Biopsy | 838 | 758 | 90.45% |
| Total | 1862 | 1583 | 85.02% |

Machine learning model classify:

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | All feature used | Color feature used | Shape feature used |
| Tree | 95.1% | 93.5% | 93.5% |
| Linear Discriminant | 92.3% | 91.7% | 88.1% |
| Quadratic Discriminant | 91.8% | 90.2% | 89.7% |
| Logistic Regression | 93.7% | 91.7% | 90.3% |
| SVM | 94.3% | 92.4% | 92.9% |
| KNN | 93.9% | 91.2% | 92.4% |

***Biopsy2：***

Manually threshold selection classify:

|  |  |  |  |
| --- | --- | --- | --- |
| Frame Label | Num of Frame | Correctly detected | Accuracy |
| Normal | 2400 | 1838 | 76.58% |
| Biopsy | 686 | 562 | 81.9% |
| Total | 3086 | 2400 | 77.77% |

Machine learning model classify:

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | All feature used | Color feature used | Shape feature used |
| Tree | 94.1% | 91.1% | 93.6% |
| Linear Discriminant | 92.9% | 91% | 87% |
| Quadratic Discriminant | 91.7% | 91% | 87.7% |
| Logistic Regression | 93.5% | 91.4% | 89.0% |
| SVM | 94.2% | 91.5% | 92.5% |
| KNN | 47.4% | 44.5% | 46.8% |