EE 559 Project

Classify Forest Cover

05/10/2014

Shankhoneer Chakrovarty

chakrova@usc.edu

Tools

Prtools description <Expand later>

1. **Baseline Performance Measure:**

Procedure:

To get the baseline performance measure, *qdc* classifier was used which is provided by prtools. 5 times 5-fold cross validation was performed to estimate the performance on the training set.

Baseline performance data:

Classifier	**Average Error Rate**	**Standard Deviation**
qdc	84.51	0.11

So we see that the baseline performance error rate is 84.51±0.59%. So the accuracy is 15.49±0.59%

Confusion Matrix is presented below:

True Labels	**Estimated Labels**							**Totals**
	1	**2**	**3**	**4**	**5**	**6**	**7**	
1	0	0	204	324	0	0	12	540
2	0	0	113	421	0	0	6	540
3	0	0	3	537	0	0	0	540
4	0	0	1	539	0	0	0	540
5	0	0	27	513	0	0	0	540
6	0	0	0	540	0	0	0	540
7	0	0	410	89	0	0	41	540
Totals	**0**	**0**	**758**	**2963**	**0**	**0**	**59**	**3780**

Comment on the above results:

`qdc' classifier gives really poor result on the given dataset. This could be due to following reasons:

1. Data is not normalized.
2. 54 dimension may be too `complex' for the classifier to classify fairly accurately, we need to reduce the dimension to see which dimension gives the best performance.

Appropriate Features:

a) **Feature Extraction:**

No remapping of the given feature set was done. Instead, I relied on normalization and dimension reduction procedures to find the optimum performance.

b) **Normalization/Scaling:**

`normm' method provided by prtools was used for normalizing the given data. `normm' method normalizes the distances of all features in the dataset such that their Minkowski-P distances to the origin equal one. Minkowski-P distance between two points

P = (x_1,x_2,…,x_n) and Q = (y_1,y_2,…,y_n) $\in \mathbb{R}^n$

is defined as:

$$\left(\sum_{i=1}^{n} |x^i - y^i|^p\right)^{1/p}$$

P=2 was selected for all the classifiers henceforth in this project. Also, normalization was done for each row as well as for each column which improved the accuracy.

Classification performance of the normalized data using 5-fold cross validation on training data:

Case 1: When all the rows of training data was normalized:

Classifier	**Average Error Rate**	**Standard Deviation**
qdc	63.74	0.17

So we see that the baseline performance error rate is 63.74±0.17%. So the accuracy is 36.26±0.17%

Confusion Matrix is presented below:

True Labels	**Estimated Labels**							**Totals**
	1	**2**	**3**	**4**	**5**	**6**	**7**	
1	47	0	161	0	90	3	239	540
2	14	0	164	6	196	4	156	540
3	0	0	253	282	5	0	0	540
4	0	0	4	536	0	0	0	540
5	4	0	241	0	277	2	16	540
6	0	0	231	275	7	27	0	540
7	0	0	320	0	9	0	211	540
Totals	**65**	**0**	**1374**	**1099**	**584**	**36**	**622**	**3780**

Case 2: When all the columns of training data is normalized:

Classifier	**Average Error Rate**	**Standard Deviation**
qdc	84.13	0.008

So we see that the baseline performance error rate is 84.13±0.008%. So the accuracy is 15.87±0.008%

Confusion Matrix is presented below:

True Labels	Estimated Labels							Totals
	1	**2**	**3**	**4**	**5**	**6**	**7**	
1	66	0	473	0	0	0	1	540
2	73	0	467	0	0	0	0	540
3	0	0	536	0	3	0	1	540
4	0	0	511	0	26	0	3	540
5	80	0	456	0	0	0	4	540
6	15	0	513	0	4	0	8	540
7	15	0	525	0	0	0	0	540
Totals	**249**	**0**	**3481**	**0**	**33**	**0**	**17**	**3780**

Case 3: When all the rows as well as columns are normalized:

Classifier	Average Error Rate	Standard Deviation
qdc	57.72	0.19

So we see that the baseline performance error rate is 57.72±0.19%. So the accuracy is 57.72±0.19%

Confusion Matrix is presented below:

True Labels	Estimated Labels							Totals
	1	**2**	**3**	**4**	**5**	**6**	**7**	
1	452	1	5	0	81	0	1	540
2	283	36	28	0	192	0	1	540
3	0	0	338	0	200	2	0	540
4	0	0	499	0	0	41	0	540
5	120	1	7	0	412	0	0	540
6	16	0	309	0	203	12	0	540
7	479	0	33	0	4	0	24	540
Totals	**249**	**0**	**3481**	**0**	**33**	**0**	**17**	**3780**

Clearly from the observed classifier performances for all the case, it can be concluded that qdc classifier performance on training data normalized by Minkowski-2 distances improve if we normalize all the rows as well as the columns of the training data.

c) Dimensionality Reduction

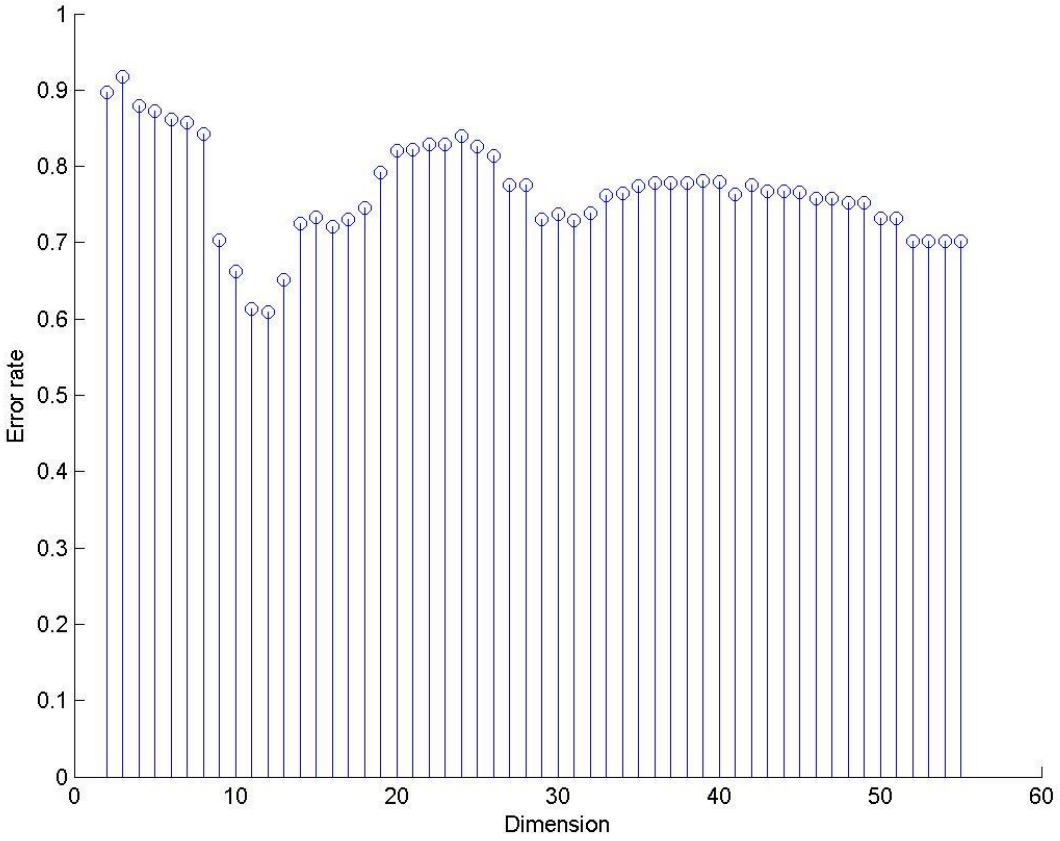
`qdc' classifier was used to find out the optimum dimension and PCA was used to reduce dimension.

Training data reduced to 11 dimension gave optimal accuracy. Following are the statistics of the qdc classifier on ` testing_set_int_labeled'

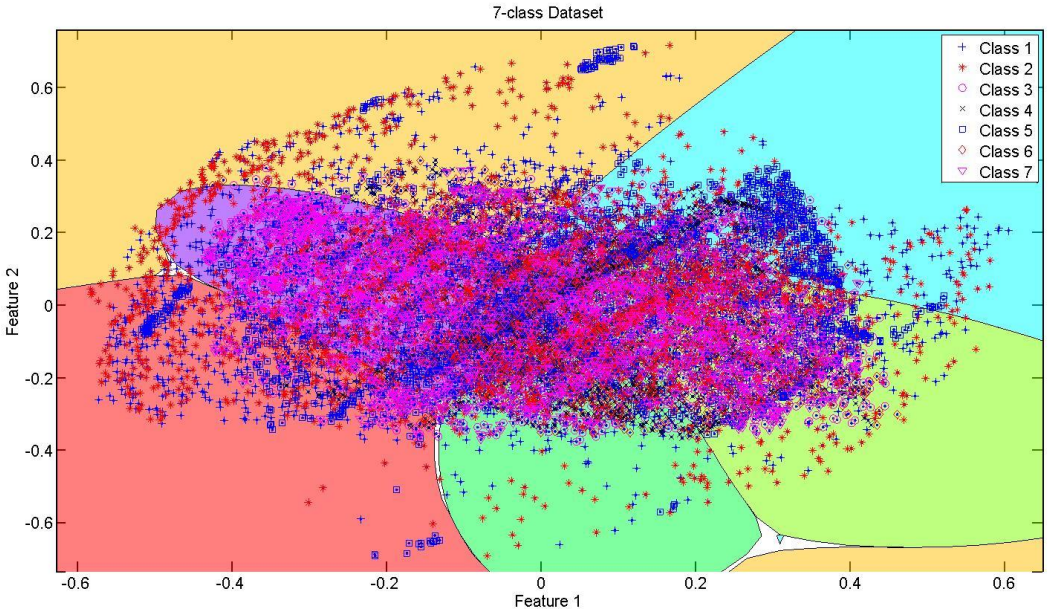
Classifier	Average Error Rate
qdc	60.98

So we see that the baseline performance error rate is 60.98%. So the accuracy is 39.02%

Classification performance (accuracy percentage) vs. dimensionality of feature on training data



Scatter plot in feature space for the 2-dimensional case, showing decision boundaries, training data, and testing data



Comparison with baseline performance:

<Expand>