二、PCA+分类

用PCA提取人脸图像的特征，分别用BP,RBF,SVM进行人脸识别，进行分析讨论，同时讨论对高斯白噪声的抑制能力。

1、仿真方法

1.1 PCA提取人脸图像的特征:

因为第一题已经提过人脸图像的PCA，因此基础知识和PCA如何提取在此不再赘述。在此仅说明用于分类的数据。在这里为了保证PCA结果的稳定性和一致性，我在这里使用COVPCA即用协方差矩阵求取PCA结果，并且样本用一整幅图，即112×92pixel大小的向量进行PCA求取，由于本题的目标是提取特征，不是恢复图像，因此这样求取PCA可以更容易提取每幅图的特征。PCA结果取系数矩阵作为PCA提取的特征。

样本取法为将每幅

1.2 加性高斯白噪声

与上次作业一样，此处加性高斯白噪声将加在整个样本上，使用matlab的awgn函数对样本加上高斯白噪声。

1.3 分类结果的评价

分类器用trainingloss与testingloss来评价，两者计数均为分错类样本的个数：

记样本i的教师为t_i，结果为y_i

$$\text{loss} = \sum_{i \in \{i | t_i \neq y_i\}} 1$$

也可以用失误率即lossrate来评价：

记一共有N个样本：

$$\text{lossrate} = \frac{\text{loss}}{N}$$

1.4 结果展示

仿真结果主要以数据形式进行，由于分类结果会以序号的形式输出。

2. 分类方法

2.1 BP网络分类

BP网络的基础知识在上一次已经提到过了，在此就不再赘述了。在这里，使用matlab的神经网络工具箱进行仿真。

2.1.1 BP网络结构与设置

设置：

设置类型	值
迭代方法	traingdx
最大迭代次数	40000
误差限	1e-5
学习率	0.1

网络结构：

层数	网络类型	神经元数量
第一层	purelin	70
第二层	tansig	40
输出层	softmax	-

这里将最后一层改为softmax而不是sigmoid输出，原因是因为实测sigmoid输出分40类正确率只有5%左右，效果极低。因此在查阅相关文献以后将最后一层网络改为softmax。对应的教师用Onehot编码：

例如当x^i为第一类，则对应的教师$y^i = [1,0,0,0,\dots,0]^T, y^i \in R^{40}$。即只有对应位置所在的值为1，其余为0。

Softmax简介

Softmax回归模型，是logistic回归模型在多分类问题上的推广。对于训练集$\{(x^{(i)}, y^{(i)}\}, i = 1,2,\dots,m$，$y^{(i)} \in \{1,2,\dots,k\}$。则对于给定的输入，我们的估计输出为：

$$h_\theta(x^{(i)}) = \begin{bmatrix} p(y^{(i)}=1|x^{(i)};\theta) \\ p(y^{(i)}=2|x^{(i)};\theta) \\ \vdots \\ p(y^{(i)}=k|x^{(i)};\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x(i)}} \begin{bmatrix} e^{\theta_1^T x(i)} \\ e^{\theta_2^T x(i)} \\ \vdots \\ e^{\theta_k^T x(i)} \end{bmatrix}$$

θ_j 为模型的参数。上面的概率分布是归一化的，所有概率之和为1。

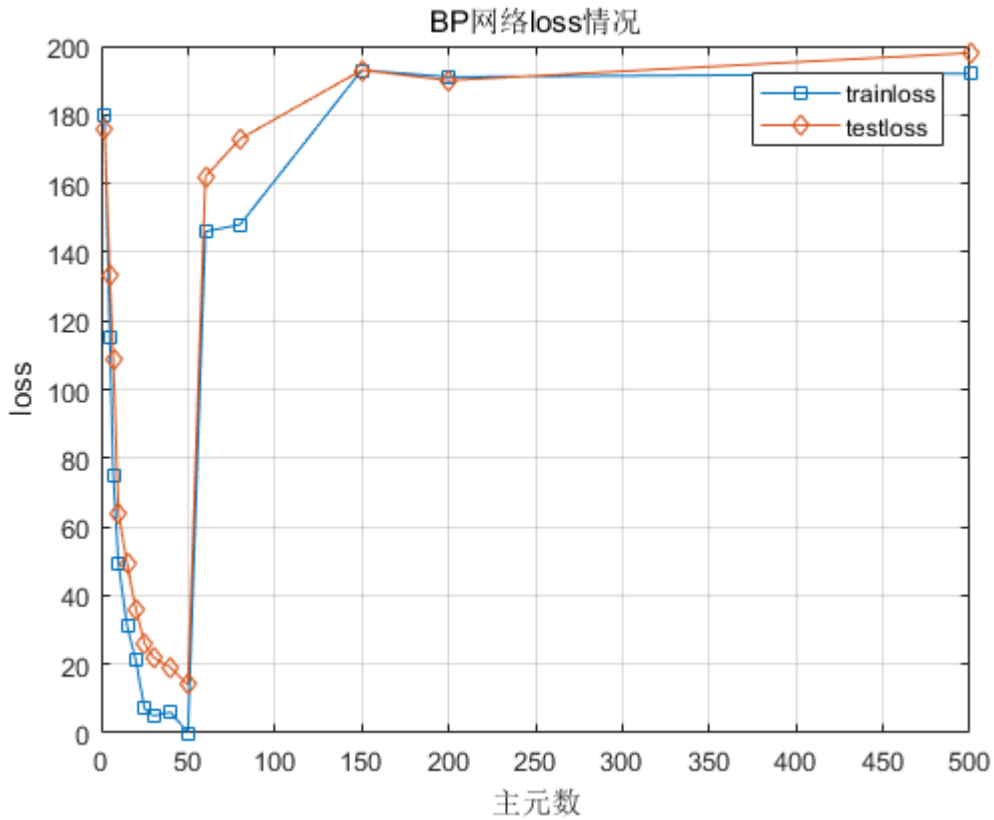
所以可以取估计输出的最大值为估计结果，即最有可能的类别。

2.1.2 分类结果

1. PCA主元数的影响：

   这里，训练集与测试集各一半，即用200张训练，200张测试。分类结果如下表：

PCA主元数	Trainingloss	Testingloss
2	180	176
5	115	134
7	75	108
10	49	64
15	31	49
20	21	36
25	7	26
30	5	22
40	1	13
50	0	14
60	146	162
80	148	173
150	193	193
200	191	190
500	192	198



可以看到，当PCA主元数取60以上的时候loss突然变高，这应该是由于随着PCA的主元数增加，网络对于输入样本来说太浅了，学习能力跟不上，因此产生了欠拟合。加深网络应该可以得到更好的效果。但是对于PCA主元少的时候，过深的网络会出现过拟合问题，所以实际上，网络深度要跟据当前的主元数进行设计。

比如将网络结构改为下面的结构时：

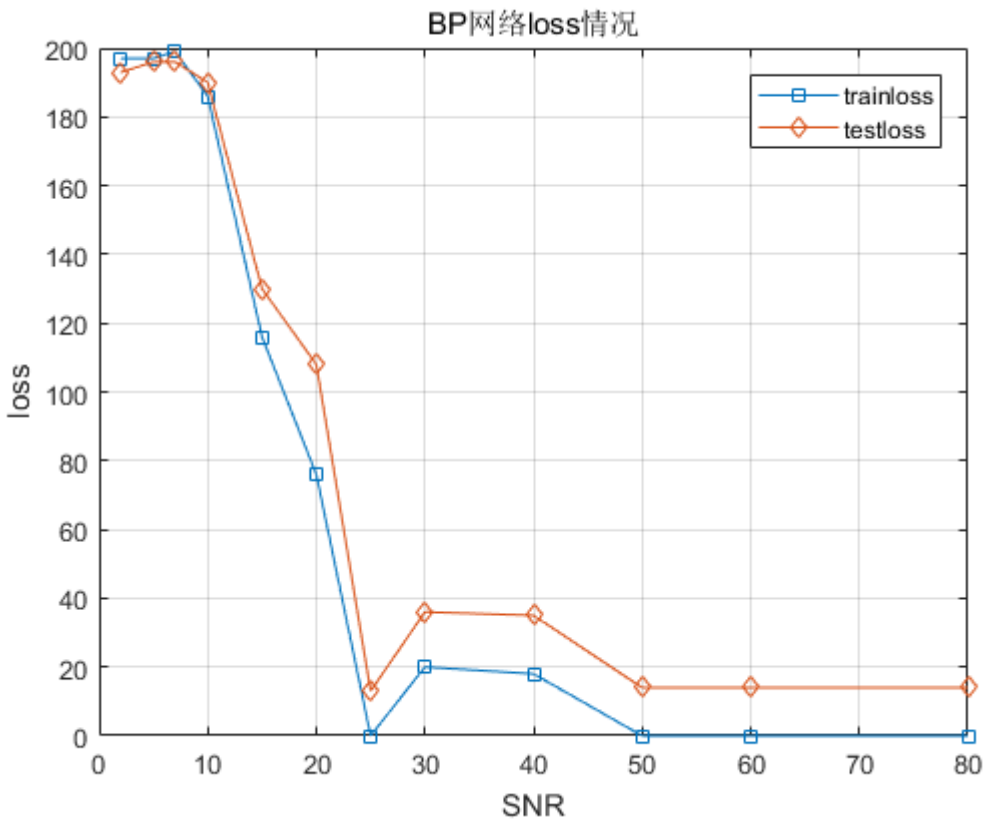
层数	**网络类型**	**神经元数量**
第一层	purelin	90
第二层	tansig	50
第三层	tansig	40
输出层	softmax	-

trainingloss为10，testingloss为95。当然此时还存在有过拟合，但是此时已经比之前要好很多。

2. 高斯白噪声的影响

这里，训练集与测试集各一半，即用200张训练，200张测试，PCA主元取50个。分类结果如下表：

高斯白噪声(SNR)	Trainingloss	Testingloss
2	197	193
5	197	196
7	199	196
10	186	190
15	116	130
20	76	108
25	0	13
30	20	36
40	18	35
50	0	14
60	0	14
80	0	14

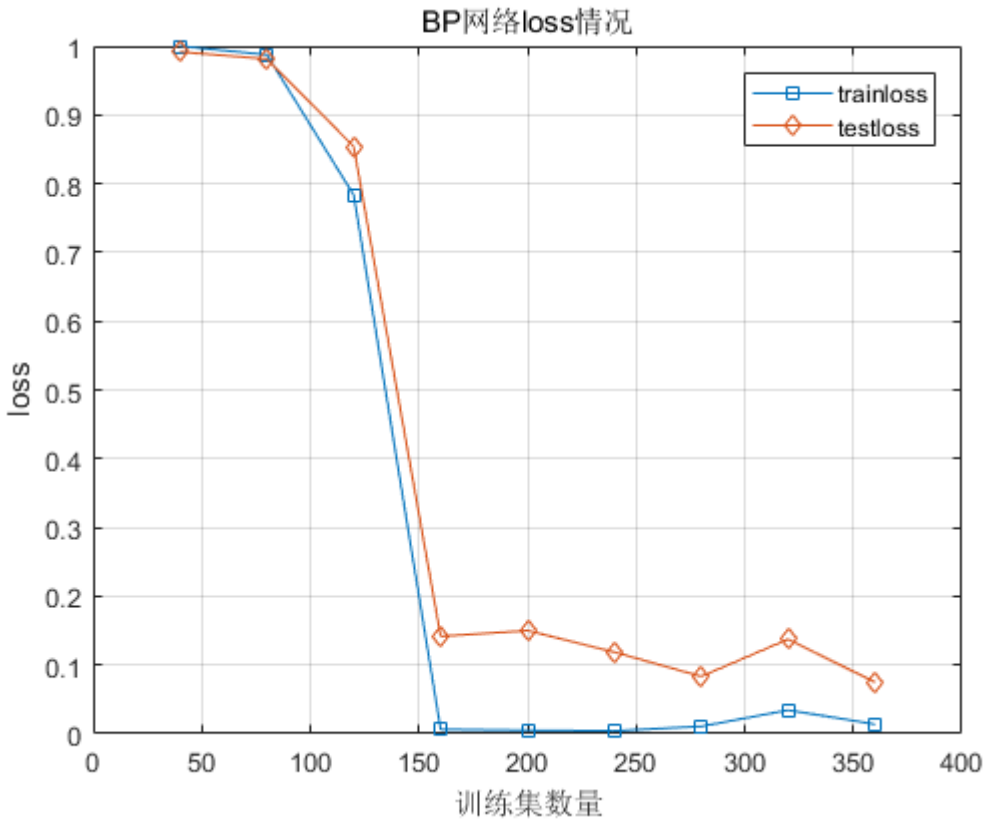


从图中可以看到，大约在25SNR时，BP的Loss就回归正常了，但是当SNR低的时候，Loss非常的高。

3. 训练集数量的影响

这里依旧取50个主元，无高斯白噪声。

训练集数量	Traininglossrate	Testinglossrate
40	1	0.9917
80	0.9875	0.9813
120	0.7833	0.8536
160	0.0063	0.1417
200	0.0050	0.1500
240	0.0042	0.1188
280	0.0107	0.0833
320	0.0343	0.1375
360	0.0139	0.0750



可以看到，当训练集数量到160时存在一个突变，在此之前此网络的loss很高，之后loss很低。因此可以认为此网络最少需要160个训练样本。

4. 总结

可以看到，这个BP网络大约只能支持PCA主元数为50，大于50的主元数会带来欠拟合问题。当保持PCA主元数为50时，最多只能容忍约SNR为25db的高斯白噪声，最少需要160个训练样本。

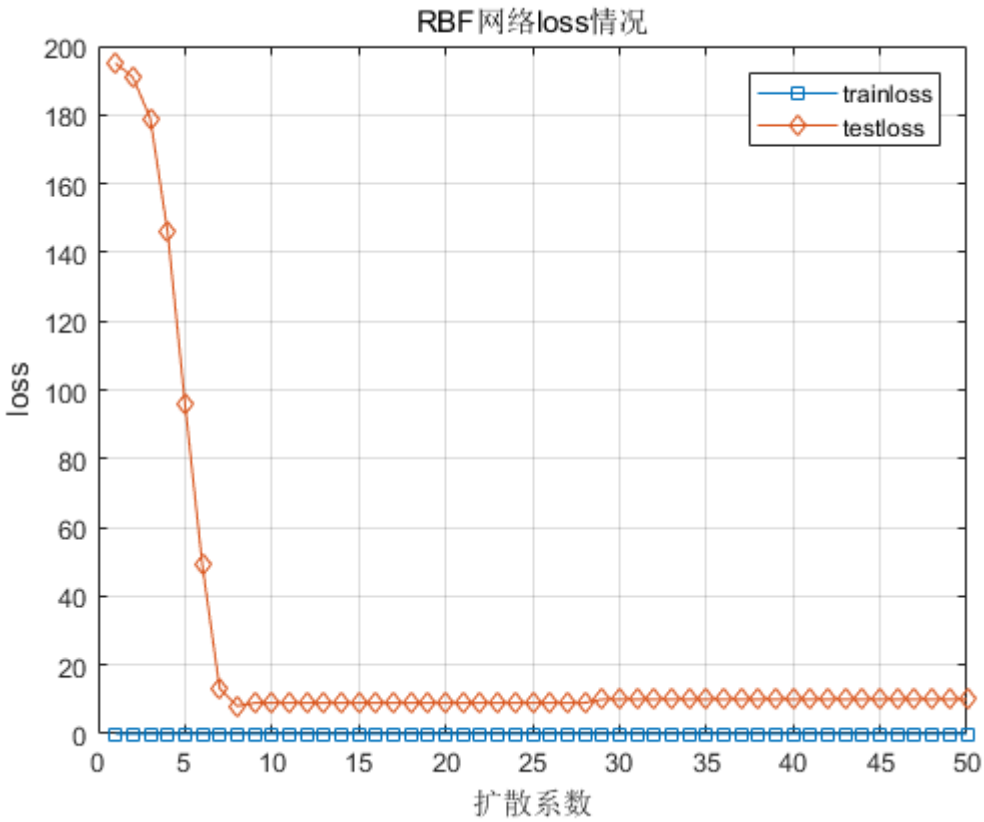
2.2 RBF网络分类

RBF网络的基础知识在上一次已经提到过了，在此就不再赘述了。在这里，使用matlab的神经网络工具箱进行仿真。Label与BP一样采用Onehot编码。

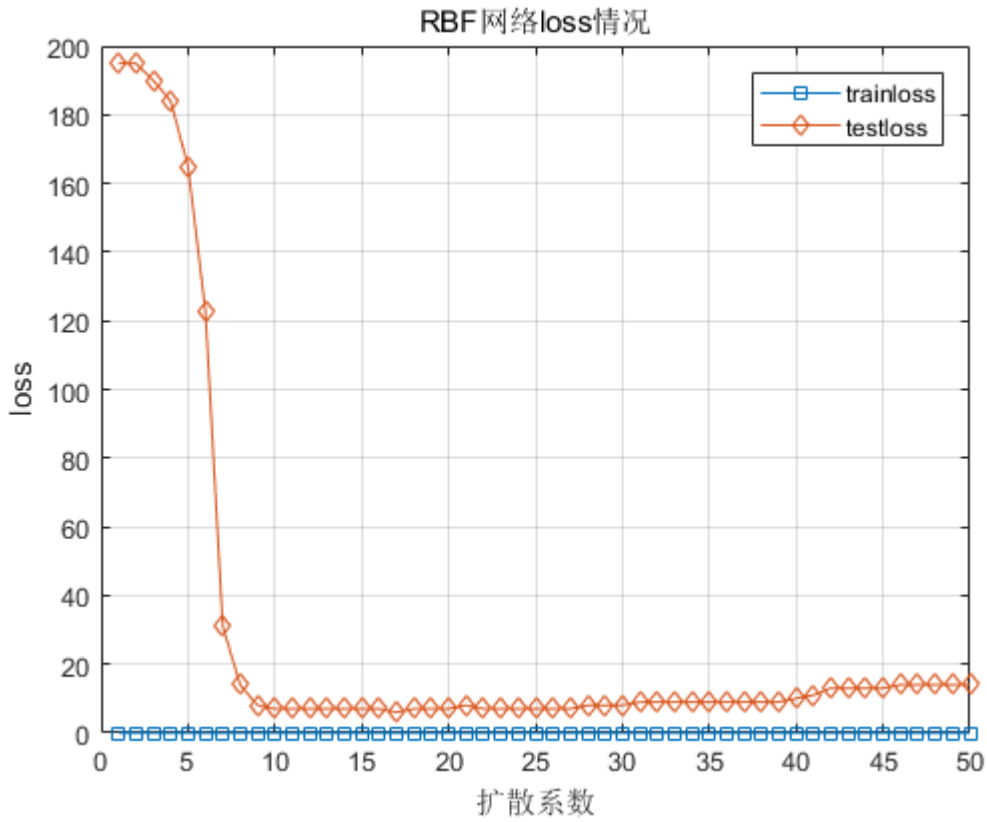
2.2.1 网络参数设置：

RBF分类网络使用newbfe函数进行设计，要对newbfe的扩散系数进行设置，如用默认的1，则分类准确率低于5%。

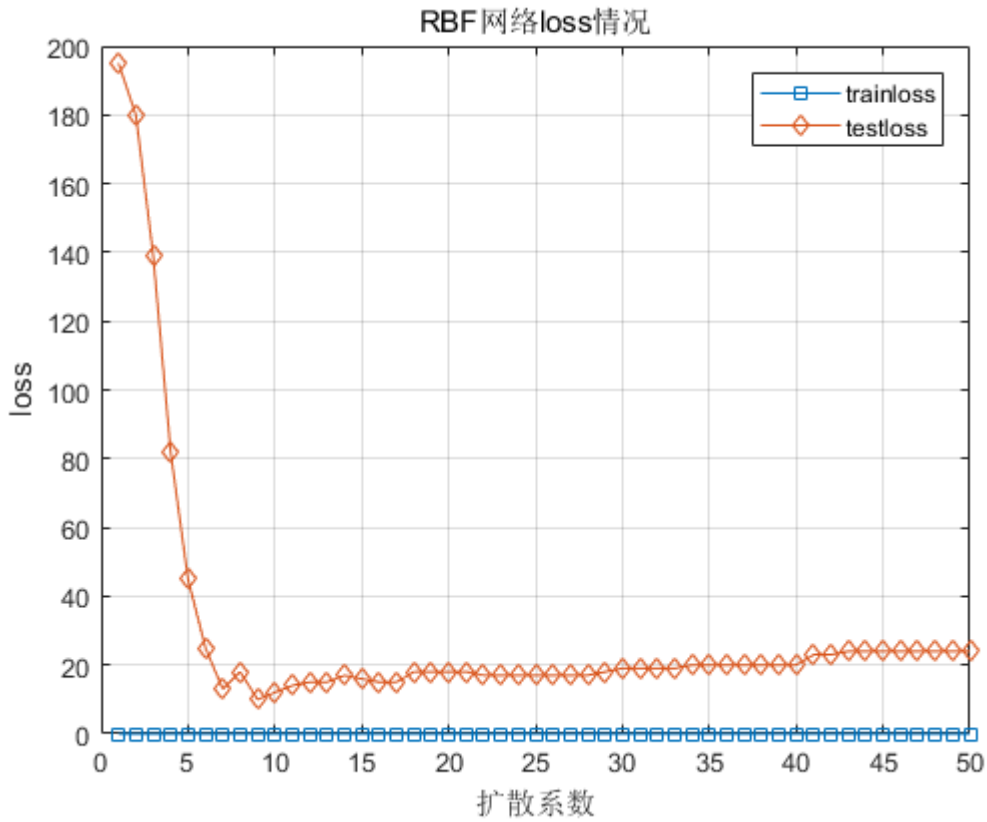
取PCA50个主元，训练集和测试集各用一半进行测试扩散系数对结果的影响，发现取扩散系数为8时，loss最低：



取PCA500个主元，训练集和测试集各用一半进行测试扩散系数对结果的影响，发现取扩散系数为17时，loss最低：



取PCA20个主元，训练集和测试集各用一半进行测试扩散系数对结果的影响，发现取扩散系数为9时，loss最低：



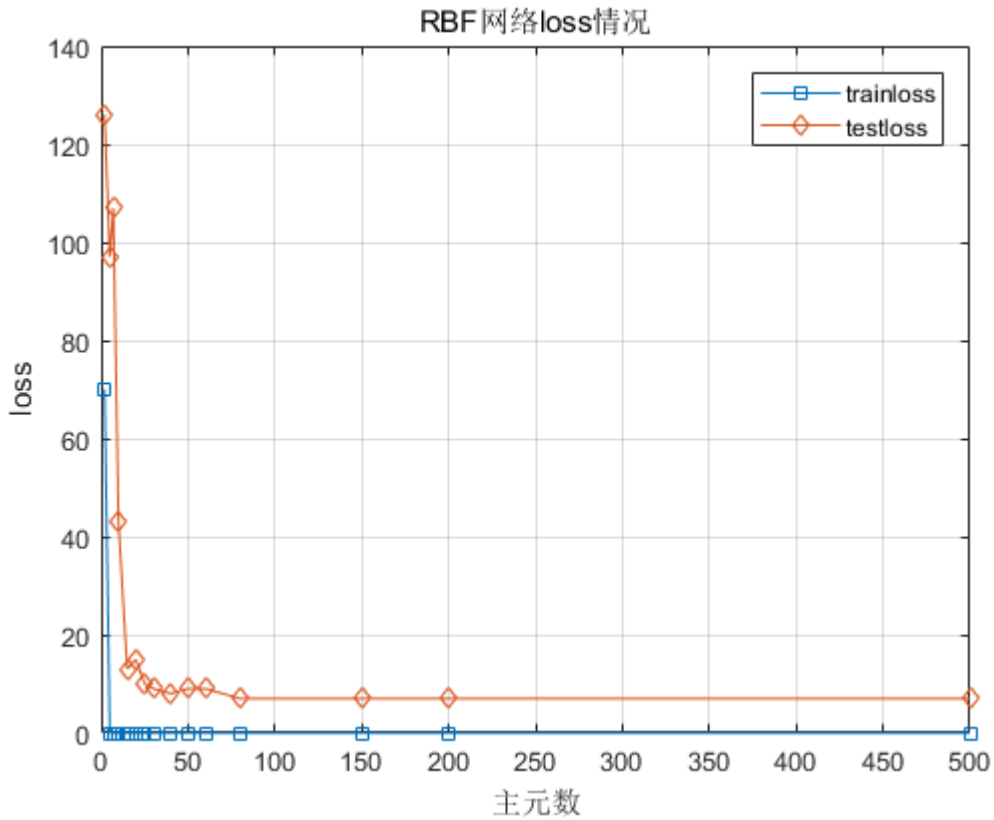
所以，综合上述测试，为了平衡在各种情况下的测试结果，下面测试全部取扩散系数为12。

2.2.2 分类结果

1. PCA主元数的影响：

   这里，训练集与测试集各一半，即用200张训练，200张测试。分类结果如下表：

PCA主元数	Trainingloss	Testingloss
2	70	126
5	0	97
7	0	107
10	0	43
15	0	13
20	0	15
25	0	10
30	0	9
40	0	8
50	0	9
60	0	9
80	0	7
150	0	7
200	0	7
500	0	7

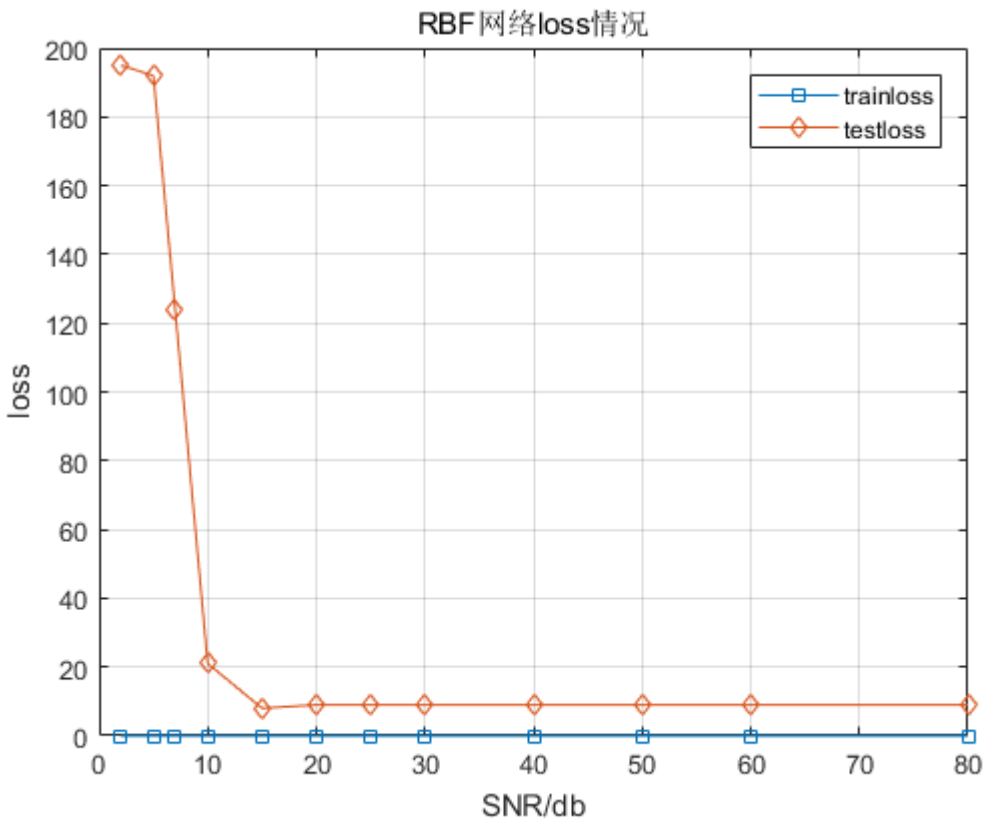


由于RBF网络工具箱会随训练样本的不同调整权置和神经元数目使训练集的误差最低，因此，相比BP网络有更好的表现。这并不是说明BP网络本身性能低于RBF网络，而是因为BP网络的参数多于RBF网络，因此自适应调整有难度，如上面的结果，当BP网络调整得好的时候，结果也是很好的，但需要每次为输入样本调整参数。RBF网络loss的情况来看，在主元数大于15以后，结果就很好，过拟合没有了，而且testloss很低。而且随着主元数的增加，testloss也基本没有起伏。

2. 高斯白噪声的影响：

训练集与测试集各一半，即用200张训练，200张测试，PCA主元取50个。分类结果如下表：

高斯白噪声(SNR)	Trainingloss	Testingloss
2	0	195
5	0	192
7	0	124
10	0	21
15	0	8
20	0	9
25	0	9
30	0	9
40	0	9
50	0	9
60	0	9
80	0	9

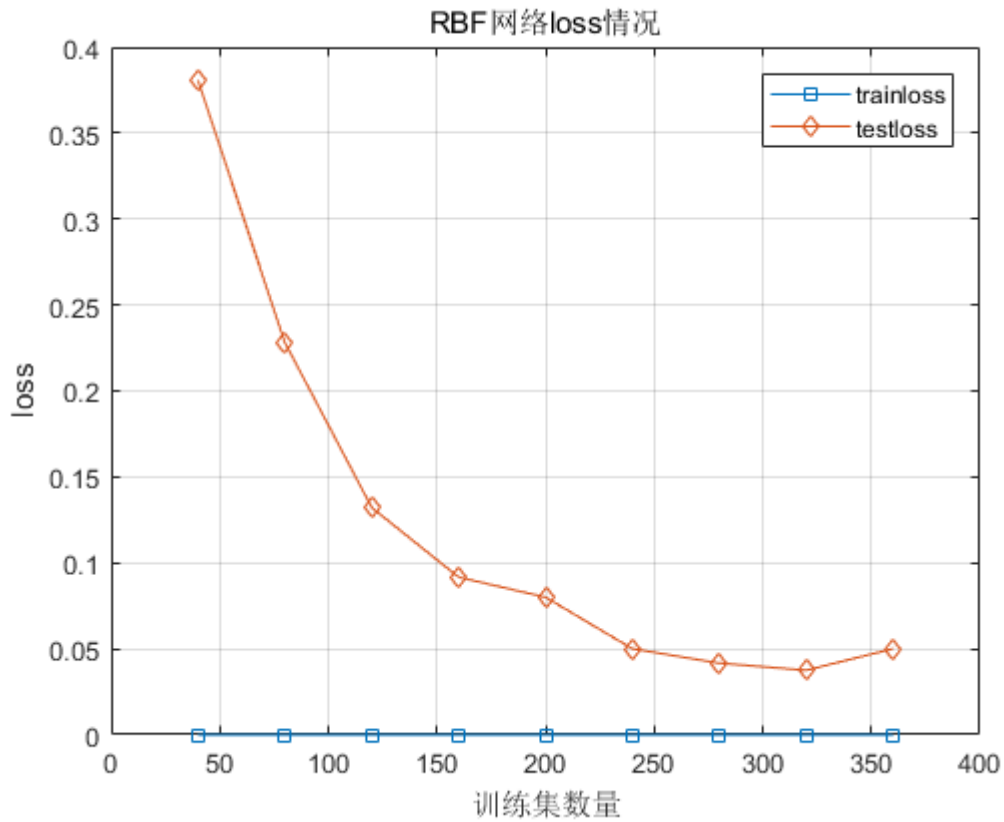


可以看到在50个主元的情况下，在实验中RBF网络都可以保证对训练集完全分类，但是只有当$\mathbf{SNR} \geq 15$时，testloss才开始正常。可以认为在这个情况下，RBF网络可以容忍$\mathbf{SNR} \geq 15$的高斯白噪声。

3. 训练集数量的影响：

这里依旧取50个主元，无高斯白噪声。

训练集数量	Traininglossrate	Testinglossrate
40	0	0.3806
80	0	0.2281
120	0	0.1321
160	0	0.0917
200	0	0.0800
240	0	0.0500
280	0	0.0417
320	0	0.0375
360	0	0.0500



在这里我们看到，当训练样本为160个的时候，RBF网络的testloss下降率的变化率发生了变化，同时，此时的测试集正确率大于90%，因此认为在训练集大于160个的时候此网络误差就在可容许的范围内。

2.3 SVM分类

2.2.1 网络参数设置：

由于matlab本身自带的工具箱中的SVM并不能直接用于多分类系统，虽然也可以手动改为用softmax回归或者投票机制的多分类器，但是最终实验下来效果都不够理想，准确率都不能超过30%。因此最终使用了第三方的svm工具箱，libsvm。同时，由于libsvm中的SVM函数svmtrain与svmpredict都与matlab自带工具箱重名，因此将libsvm中的函数改为svmtrain2与svmpredict2。

使用的svm分类器使用默认参数。即是使用高斯核，C-SVC，gamma取1。因为默认参数的效果已经足够好了，因此不再调试参数，直接仿真。

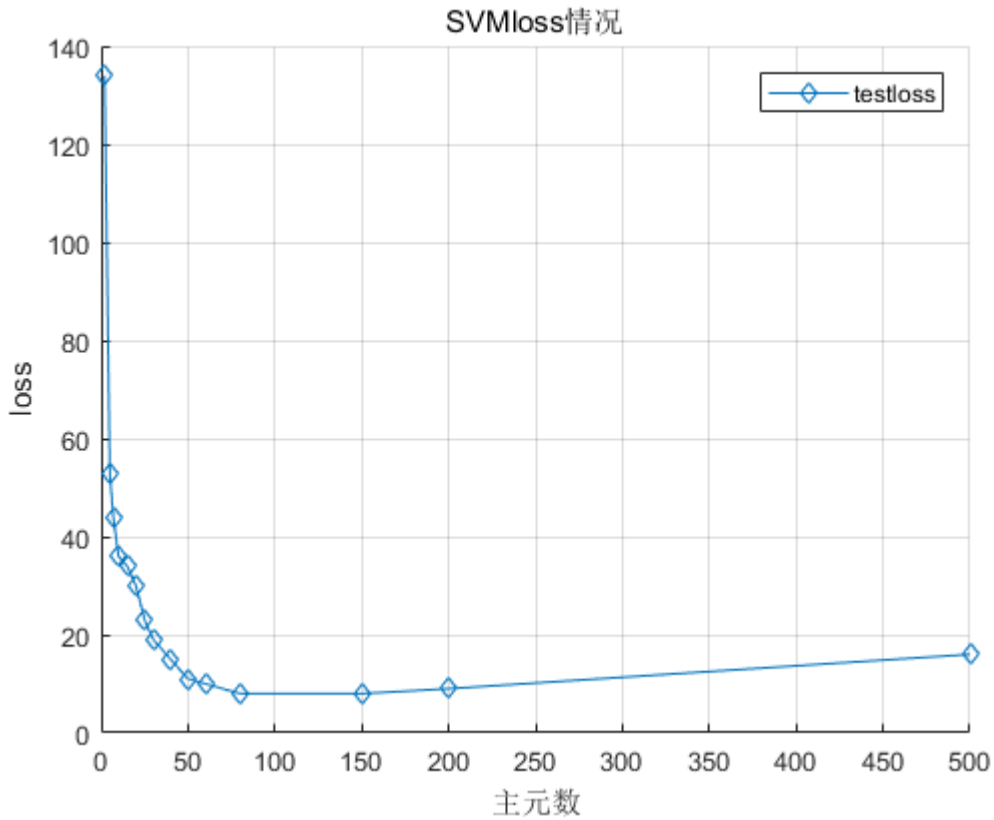
2.2.2 分类结果

由于这里使用的是高斯核函数，分类时总是可以达到对训练集的完美分类，因此，这里不列出trainingloss。

1. PCA主元数的影响：

   这里，训练集与测试集各一半，即用200张训练，200张测试。分类结果如下表：

PCA主元数	**Testingloss**
2	134
5	53
7	44
10	36
15	34
20	30
25	23
30	19
40	15
50	11
60	10
80	8
150	8
200	9
500	16

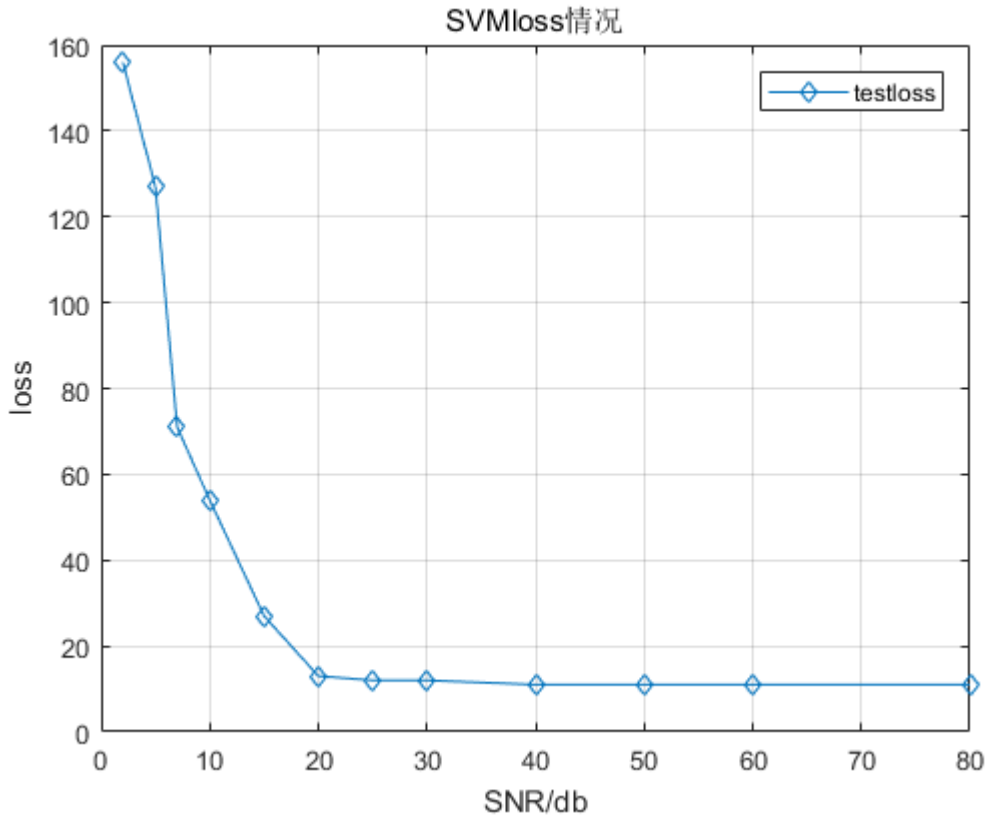


可以看到，SVM在25个主元时，testloss降低到了20左右，在并在80个主元时达到最低，但总体来说，SVM在主元更少的时候的表现要好于RBF网络与BP网络。但当主元数逐渐增加，似乎又出现了过拟合的情况。

2. 高斯白噪声的影响：

训练集与测试集各一半，即用200张训练，200张测试，PCA主元取50个。分类结果如下表：

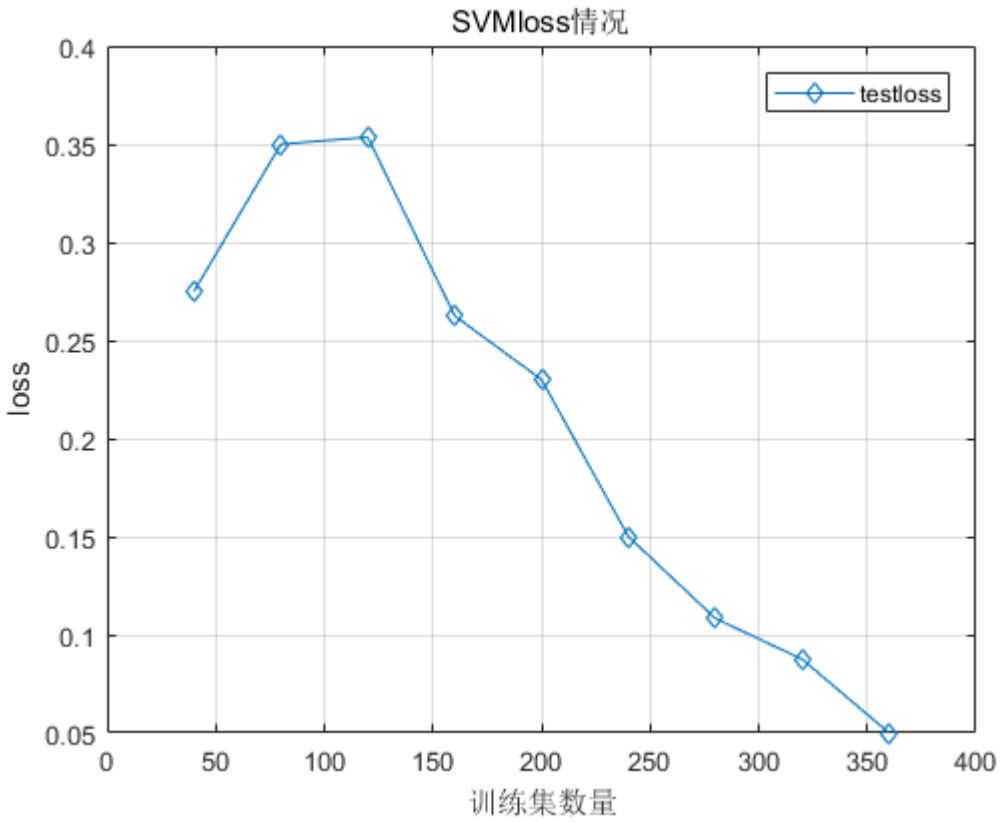
SNR(db)	**Testingloss**
2	156
5	127
7	71
10	54
15	27
20	13
25	12
30	12
40	11
50	11
60	11
80	11



可以看到，$\mathbf{SNR} \geq \mathbf{20}$时，SVM的loss回归正常水平，大约在15db时，loss也还可以接受。对噪声的抑制要好于BP网络，但略逊色于RBF网络。

3. 训练集数量的影响：

训练集数	**Testinglossrate**
40	0.2750
80	0.3500
120	0.3536
160	0.2625
200	0.2300
240	0.1500
280	0.1083
320	0.0875
360	0.0500



相比较前两者，SVM少量训练样本时的表现比BP网络和RBF网络更好。

3. 总结

可以看到这三种分类器在合适的参数下，用PCA进行特征提取以后，都可以很好的完成分类目标，下面就从三方面分别讨这三者的优劣。

3.1 PCA主元数影响

以错误降到20个左右进行划分（此时相当于90%正确率。）列出下表

BP	RBF	SVM
20	15	25

由此来看，RBF表现最好，BP其次，但是，实际上，根据全表可以看到，SVM在主元数少的时候，loss要小于另两者，而BP由于网络并未根据情况进行调整，而是一直用同一网络，因此，其结果也不具有太大的代表性，但总的来说，就方法的普适性和运行速度来说，RBF和SVM要优于BP。但是就从性能上来说，可以看到，这三者在分类效果最优时，性能差矩不大。

总的来说，SVM在主元少的时候就可以有优秀的表现，但当前的参数设置随着主元数升高，正确率提高不如另两者快。RBF可以最快达到90%正确率，但是这是由于其网络结构会根据样本分布和数量自适应调整的结果，这个优秀的结果是好的参数+方法得到的，并不能单纯的说RBF就适合于分类。而BP则是最不稳定的，在实测中，也常常有BP网络训练时陷入某个局部极小而使得loss突然增大，但当参数合适时，也可达到另两者的性能。

3.2 高斯白噪声的影响

以错误降到20个左右进行划分，列出下表：

BP	RBF	SVM
25	10	15

可以看到，抗噪能力上，RBF的表现也是最好的，而SVM也差不多。但是从BP的趋势图可以看到，在SNR25db处有突变即突然分类效果就回归正常了，而RBF随SNR增加，loss减少的最快，SVM的loss减少的也很快，但是不如RBF快。因此，在这个问题上，可以认为，RBF网络的抗噪能力是最好的。但是，同时比较噪声极大的时候，SVM的正确率也是高于另两者的，说明SVM在极大噪声情况下，表现要好于另两者。

3.3 训练集数量的影响

以错误率降到0.1左右进行划分，列出下表：

BP	RBF	SVM
200	120	240

可以看到，RBF所需要的训练样本数更少，BP其次，SVM最多。但是SVM在10%训练集时就可以达到很好的效果，这一点是比另两者要优秀的。

3.4 总结

综上，三种方法在参数足够好的情况下都可以达到很好的分类效果。RBF由于其网络结构会根据样本分布和数量自适应调整以达到最小误差，所以有最优秀的表现。BP由于其参数过多，本身难以调整，再加上迭代和局部最小等问题，其调整难度是最大的，而且每次训练的结果不稳定。SVM则相比更中庸但稳定，同时，SVM在少样本，大噪声，少主元等极端情况下的表现最好。