大数据计算机基础——Python

基于 Python 的图像降维与分类

作　　者：　Yang Le

目　　录

一、 概述 1
二、 数据说明 1
三、 特征提取 2
（一） 核主成分分析（Kernel PCA） 2
（二） 参数优化与模型选择 2
（三） 特征提取与图像重构 5
四、 图片分类 5
（一） 分类方法 5
（二） 分类模型评估指标 7
（三） 分类性能评估 7
五、 总结 9

基于 Python 的图像降维与分类

一、 概述

图像识别在多个领域中具有重要的应用价值，如最为常见的人脸识别、生物医学影像识别、智能交通领域的路况识别等。图像识别在本质上是分类问题，但由于图像数据的高维特征，在建立分类模型之前往往需要对数据进行特征提取，以减少数据中的噪声和模型参数个数，从而提高模型性能。本文首先采用核主成分分析（Kernel PCA）对图像进行降维，比较了不同核函数的降维效果，发现采用多项式核时重构误差最小；进一步基于降维后的数据建立分类模型，比较了 K 近邻、支持向量机和朴素贝叶斯三种分类方法的在图像分类上的性能，发现采用 RBF 核的支持向量机的分类效果最佳。

二、 数据说明

本文采用的数据是 CIFAR-100 数据集，它包含有 20 个大类的图像，该 20 类图像被细分成 100 个小类，每个小类各有 500 个训练图像和 100 个测试图像。每一张图像的大小相同，均为 32×32 个像素点，每个像素点包含 RGB 三个通道的颜色信息，即每张图片包含 3072（32×32×3）个数字信息。本文选用了其中三个小类的图像，在数据集中的数值型标签分别为 2、3、70，类别名称分别为“baby”、“bear”和“rose”，图 1 展示了每一类中的前 6 张训练图像。

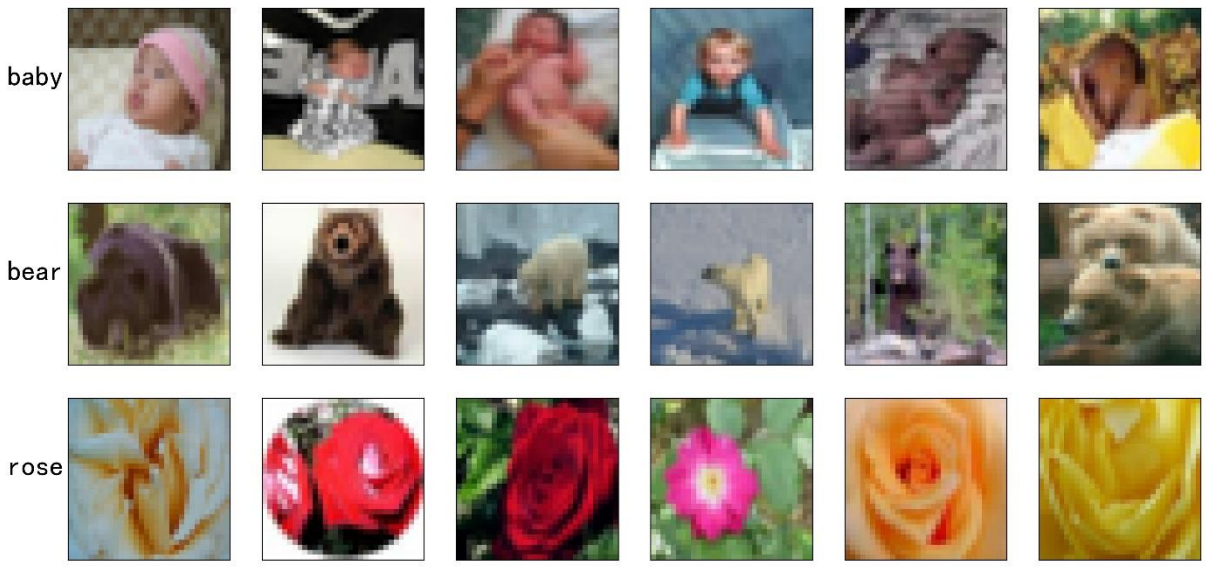

图 1. 本文所选的三类图像示例

三、 特征提取

（一） 核主成分分析（Kernel PCA）

本文采用核主成分分析（Kernel PCA）对图像进行特征提取。Kernel PCA 是对一般的主成分分析（PCA）的非线性扩展，在 PCA 的基础上增加了将数据通过非线性映射函数映射到一个高维空间的过程，在高维空间中使用 PCA 将其映射到另一个低维空间中。而非线性映射将带来较为高昂的计算成本，因而采用核函数计算两个高维空间中向量的相似度。Kernel PCA 相较 PCA 能够更好的提取数据中的非线性信息，而图像数据中的信息往往是非线性的，因而在理论上，我们更有理由选择 Kernel PCA 而不是 PCA。

常用的核函数有线性核、多项式核、RBF 核以及双曲正切核，它们的函数形式如表 1 所示[①]。本文后续将通过实验的方法比较不同的核函数及其参数取值下的降维效果。

表 1. 常用的核函数

核函数	函数形式
线性核 (Linear kernel)	$k(x,y)=x^{\top}y$
多项式核 (Linear kernel)	$k(x,y)=(\gamma x^{\top}y+c_0)^d$
高斯核 (RBF kernel)	$k(x,y)=\exp(-\gamma \parallel x-y \parallel^2)$
双曲正切核 (Sigmoid kernel)	$k(x,y)=\tanh(\gamma x^{\top}y+c_0)$
余弦相似度 (Cosine similarity)	$k(x,y)=\frac{xy^{\top}}{\parallel x \parallel \parallel y \parallel}$

（二） 参数优化与模型选择

应用 Kernel PCA 方法需要选定一个合适的核函数，并对带有参数的核函数进行参数调优，最终需要确定提取的主成分个数。

1. 核函数的选择及其参数调优

参数优化的目标是使得测试集上的重构效果达到最优，即在不同参数取值下训练模型，利用得到的模型对测试集进行降维再重构，计算重构数据与原始数据之间的均方误差，将使得误差最小的参数值作为最优的参数取值。由于 Sigmoid 核函数无法重构测试集图像，只考虑其余几个核函数，其中 Poly、RBF 核函数带有参 γ，通过实验寻找它们各自最优的 γ 取值。虽然 Sigmoid 核函数无法重构测试集，但不能说明降维后的分类效

[①] https://scikit-learn.org/stable/modules/metrics.html#metrics

果不好，因此后续可以考虑利用分类误差来进行参数优化和模型选择，该部分未在本文进行讨论。

实验过程中，控制主成分个数为 200 个，后续将讨论主成分个数是否对参数的最优取值有影响；主要的计算过程基于 multiprocessing 库进行并行计算。

结果如图 2 和表 2 所示。图 2 分别给出了使用 Poly 和 RBF 核函数时，训练样本上的重构误差（拟合误差）和测试样本上的重构误差（测试误差）随 γ 的变化情况。在各自给定的 γ 取值范围内，使用两种核函数降维的拟合误差均随 γ 增大而下降；对于测试误差，使用 Poly 时，存在一个 γ 的取值使得测试误差最小；而使用 RBF 时，测试误差随 γ 增大而增大。表 2 汇总了核函数参数优化后的降维效果以及不带参数的核函数的降维效果。使用不带有参数的核函数 Linear 和 Cosine 时，拟合误差和测试误差都远大于其他核函数。基于上述结果，可以认为采用 Poly 核的 Kernel PCA 降维效果较其他核更好，且当其 γ 参数取 1.15×10^{-6} 时，测试集上的重构误差最小，能够相对较好的提取原始数据中的信息。

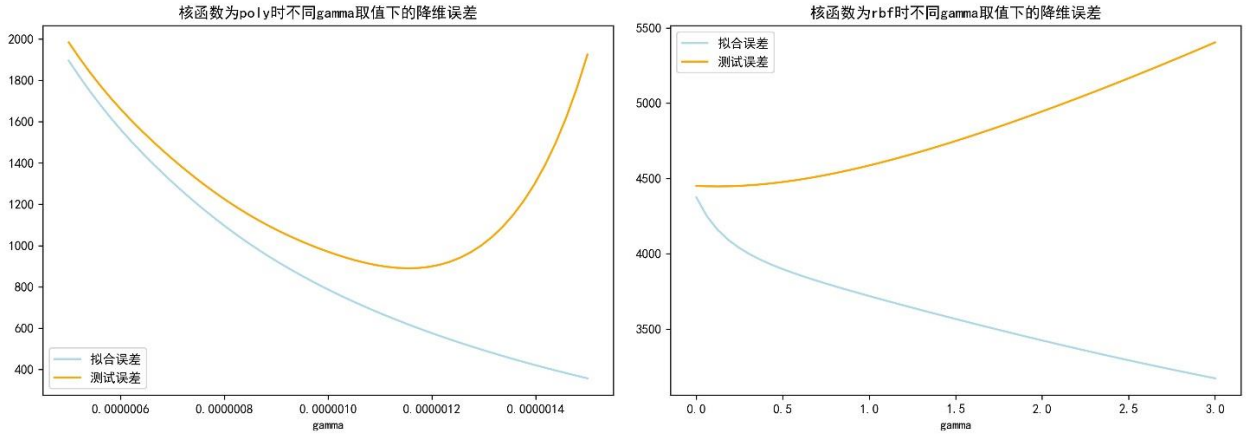

图 2. 核函数为 Poly 和 RBF 时降维误差随 γ 的变化

表 2. 各核函数的降维效果

核函数	γ	拟合误差	测试误差
Cosine	-	11840.22	12550.24
Linear	-	13411.89	13531.05
Poly	1.15E-06	619.51	889.91
RBF	0.12	4159.87	4447.32

2. 主成分个数的确定

从理论上来说，主成分个数越多，原始数据中的信息也就保留得更多，重构误差更小；但同时主成分中也可能包含了更多的噪声，且维数较高，后续模型的参数个数较多，

不利于模型训练。因此，需要确定一个合适的主成分个数。

图 3 给出了主成分个数分别取 100、200 和 300 时测试误差随 γ 的变化情况。可以看出，无论 γ 的取值如何，主成分个数更多时，重构误差都更小。同时，最优的 gamma 取值对主成分个数不敏感，即不同主成分个数下，最优的 γ 取值都大致不变。因此，可基于前面参数调优的结果选用 Poly 核及其最优的参数取值，来进一步选择主成分个数，而无需重新进行参数调优。

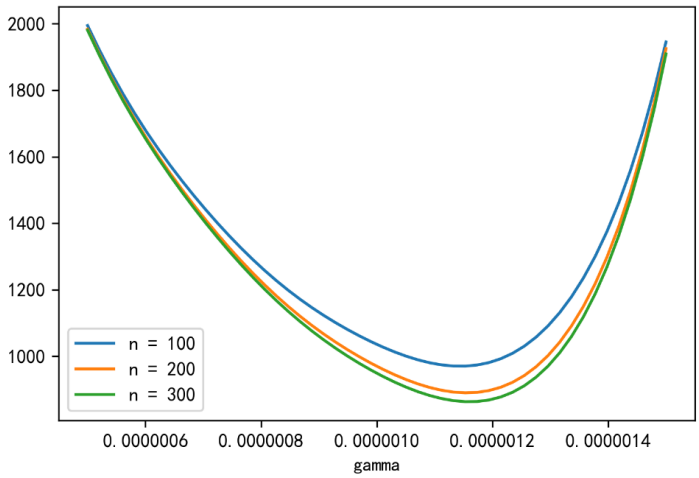

图 3. 不同主成分个数下测试误差随 γ 的变化

主成分个数选择的依据是：当测试误差随主成分个数增加的变化量（margin）的绝对值小于某个阈值时，则认为增加主成分个数不再明显的较小测试误差，将此时的主成分个数作为模型适用的主成分个数。本文设置该阈值为 0.1。从图 4 中可以看出，当主成分个数大于 300 时，margin 的绝对值趋于稳定且小于 0.1，因此本文选用 300 个主成分。

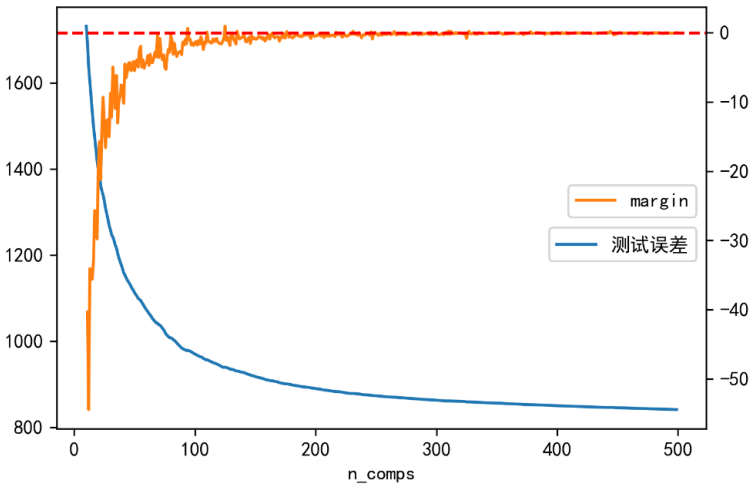

图 4. 测试误差及其变化量随主成分个数的变化

（三） 特征提取与图像重构

基于参数优化与模型选择的结果，建立最终的 Kernel PCA 模型并对图像进行特征提取。利用训练好的模型对训练图像进行降维，部分原始图像与重构图像如图 5 所示，可以看出重构图像在色彩分布上与原始图像比较类似，但损失了较多细节信息，整体上较为模糊。利用模型对测试图像进行降维并绘制图像如图 6，重构效果明显比训练图像差，在色彩上和细节上有较多失真。

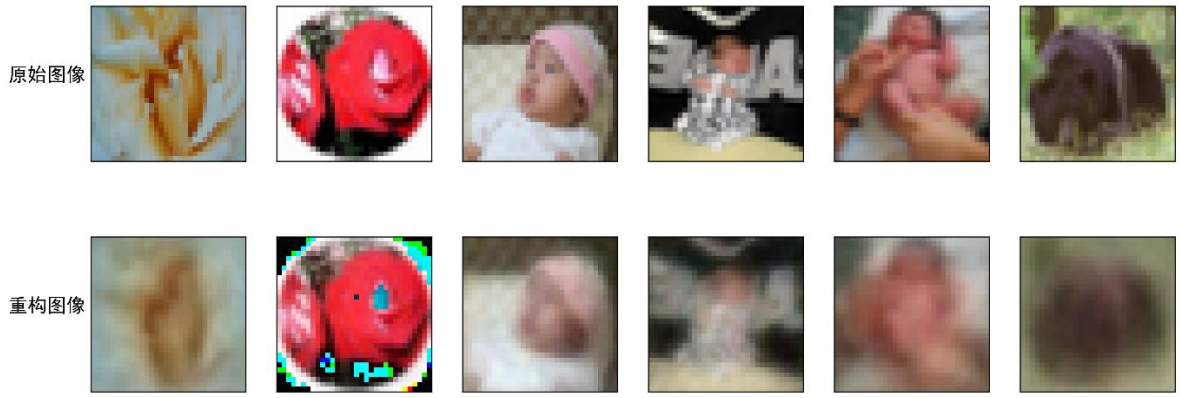

图 5. 训练图像的原始图像与重构图像

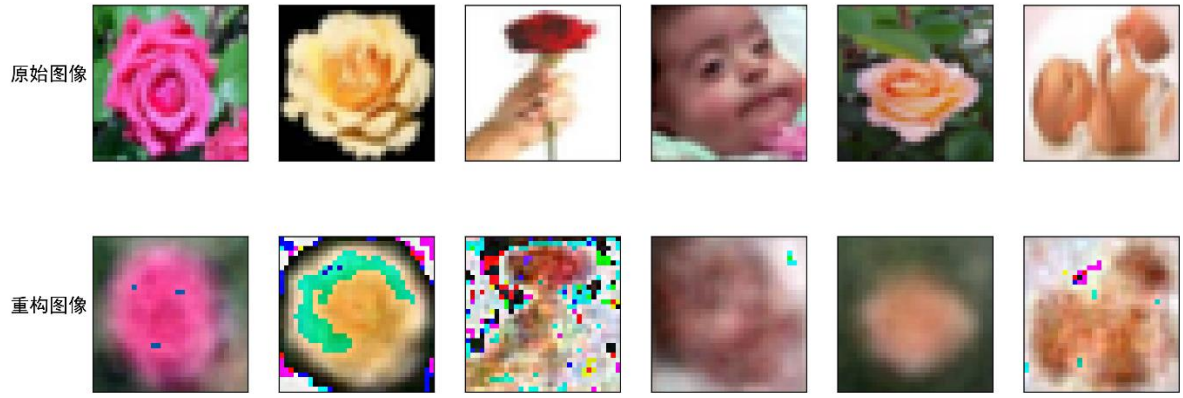

图 6. 测试图像的原始图像与重构图像

四、 图片分类

（一） 分类方法

1. k 近邻法

k 近邻法(k-Nearest Neighbor, k-NN)是 1967 年由 Cover T 和 Hart P 提出的一种基本分类与回归方法。对于某个需要分类的新样本，首先计算它与其他有类别标签的样本之间

的距离，将与之距离最小的 k 个样本中出现最多的类别作为该样本的类别。通常 k 是不大于 20 的整数，本文通过交叉验证的方法确定最优的 k 取值为 6。另外，可选的距离度量方式很多，诸如欧氏距离、马氏距离、曼哈顿距离以及明氏距离等，本文采用 KNeighborsClassifier 函数默认的明氏距离。

2. 支持向量机

支持向量机（Support Vector Machines，SVM）是一种二分类模型，它的目的是寻找一个超平面来对样本进行分割，分割的原则是间隔最大化，最终转化为一个凸二次规划问题来求解。而本文的图像分类属于多分类问题，对于一个 n 分类的问题，可通过 n-1 个 SVM 的堆叠来解决。

对于非线性可分的问题，可以将训练样本从原始空间映射到一个更高维的空间，使得样本在这个空间中线性可分，如果原始空间维数是有限的，即属性是有限的，那么一定存在一个高维特征空间使样本可分。这样的非线性映射在实现上与 Kernel PCA 类似，通过核函数计算得到向量在特征空间中的内积。因此这里同样涉及到模型选择与参数调优。与 Kernel PCA 类似的，SVM 中常用的核函数有线性核（即直接在原始空间中进行分割，以解决线性可分问题）、Poly 多项式核、RBF 高斯核以及 Sigmoid 核，其中 Poly、RBF 和 Sigmoid 带有 γ 参数，需要进行参数调优。

SVM 假设训练样本在样本空间或者特征空间中是线性可分的，但在现实任务中往往很难确定合适的核函数使训练集在特征空间中线性可分，为解决这一问题，可以对每个样本点引入一个松弛变量，使得 SVM 的目标为间隔加上松弛变量大于等于 1。对于松弛变量，进一步引入一个罚参数 C，C 越大，则松弛变量接近 0，即对误分类的惩罚增大，趋向于对训练集全分对的情况，这样对训练集测试时准确率高，但泛化能力弱。C 值小，对误分类的惩罚减小，将被误分类的样本作为噪声点，模型泛化能力较强。本文在不同的核函数下，采用网格搜寻和交叉验证的方法，选择合适的 C 和 γ 的取值，参数调优的结果如表 3 所示。

表 3. SVM 各核函数的最优参数取值与

核函数	C	γ	Score
RBF	10	8.00E-07	0.78
Poly	20	5.00E-06	0.76
Sigmoid	650	5.00E-08	0.78

3. 朴素贝叶斯

朴素贝叶斯分类（NBC）是以贝叶斯定理为基础并且假设特征条件之间相互独立的

方法，先通过已给定的训练集，以特征之间独立作为前提假设，学习从输入到输出的联合概率分布，再基于学习到的模型，输入待预测样本的观测值求出使得后验概率最大的类别 Y。

（二） 分类模型评估指标

本文采用查准率、召回率和 F1 score 对分类性能进行评估。将每一类样本的分类都视为一个独立的二分类问题，对每类样本计算上述三个指标，用各指标的平均值来衡量模型综合分类性能。

1. 查准率(Precision)

查准率用以衡量正样本的分类准确率，即被预测为正样本的样本中真的正样本的占比。

$$Precision = \frac{TP}{TP + FP}$$

2. 召回率(Recall)

召回率表示在所有正样本中，被识别为正样本的占比。

$$Recall = \frac{TP}{TP + FN}$$

3. F1 score

F1 score 是查准率和召回率的调和平均。

$$\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R} \Rightarrow F_1 = \frac{2PR}{P + R} = \frac{2TP}{2TP + FP + TN}$$

（三） 分类性能评估

1. 不同模型分类效果的比较

基于参数调优的结果建立模型，并对模型的分类性能进行测试。各模型在训练集和测试集上的分类性能如表 4 所示。

从训练集上的分类性能来看，除朴素贝叶斯之外的其他模型的拟合效果都较好，查准率、召回率以及 F1 score 均在 0.8 以上。线性 SVM 的各评估指标值都达到 1，则可认为降维后的训练集图像数据是线性可分的；多项式核 SVM 的拟合效果也较好，各指标均为 0.96；朴素贝叶斯的训练集分类效果最差，各指标均在 0.5 上下。

从应用角度来说，我们更关注模型在测试集上的表现。综合三个指标来看，测试集上的分类效果最佳的是 RNF 核 SVM 和 Sigmoid 核 SVM，各项指标达到 0.78；多项式核 SVM 表现稍逊，但差别不明显。K 近邻和线性核 SVM 的各项指标值在 0.75 上下，虽然线性 SVM 在训练集上的准确率达到 100%，但在测试集上的表现并不突出，说明模型存在过拟合。朴素贝叶斯的分类效果最差，F1 score 仅 0.38，对于本文的三分类问题，其

分类效果比较接近随机分类的效果。

表 4. 各模型在训练集和测试集上的分类性能

分类方法	训练集			测试集		
	查准率	召回率	F1 score	查准率	召回率	F1 score
KNN	0.81	0.8	0.8	0.76	0.75	0.75
SVM (Linear kernel)	1	1	1	0.75	0.74	0.74
SVM (RBF kernel)	0.88	0.88	0.88	0.78	0.78	0.78
SVM (Poly kernel)	0.96	0.96	0.96	0.78	0.77	0.77
SVM (Sigmoid kernel)	0.88	0.87	0.87	0.78	0.78	0.78
Naive Bayes	0.56	0.46	0.43	0.55	0.43	0.38

2. 最优模型的分类效果

根据模型比较的结果，本文选出最优的模型为 RBF 核 SVM 和 Sigmoid 核 SVM。表 5 和表 6 给出了它们在各类样本上的分类表现及混淆矩阵。两个模型在不同样本类别上的分类效果都有类似的差别，即对“bear”类的图片分类准确率最高，对“baby”类的分类准确率最低。从混淆矩阵来看，“baby”类更易被误分类为“rose”，而“bear”和“rose”都更易误分类为“baby”；“bear”误分类为“baby”相较“rose”表现得更明显。由此我们可以认为，模型的分类效果与训练样本类别的选择有关，若不同类别的样本之间的比较相似（比如都属于动物），模型在训练时更不容易提取出各类别的各自的特征，进而在测试集上更可能出现误分类。

表 5. RBF 核 SVM 和 Sigmoid 核 SVM 在各类别样本上的分类性能

类别标签	RBF 核 SVM			Sigmoid 核 SVM		
	查准率	召回率	F1 score	查准率	召回率	F1 score
baby	0.72	0.73	0.72	0.73	0.76	0.75
bear	0.84	0.81	0.82	0.84	0.81	0.83
rose	0.8	0.81	0.81	0.78	0.78	0.78
平均值	0.78	0.78	0.78	0.78	0.78	0.78

表 6-1. RBF 核 SVM 混淆矩阵

类别标签		预测类别		
		baby	bear	rose
真实类别	baby	73	8	19
	bear	18	81	1
	rose	11	8	81

表 6-2. Sigmoid 核 SVM 混淆矩阵

类别标签		预测类别		
		baby	bear	rose
真实类别	baby	76	6	18
	bear	15	81	4
	rose	13	9	78

图 7 给出了 RBF 核 SVM 模型在测试集上的部分分类结果，绿色标签表示被正确分类，红色标签表示误分类。

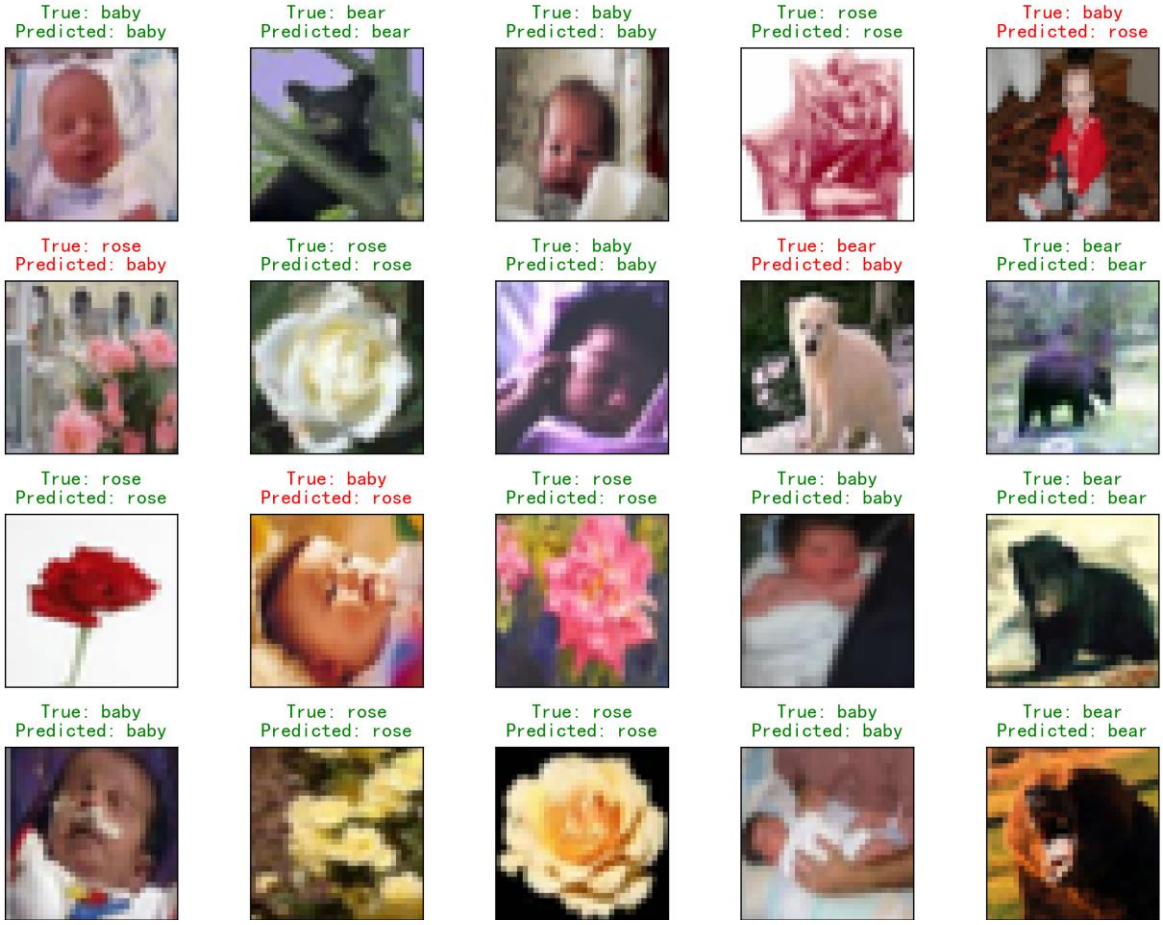

图 7. RBF 核 SVM 部分测试集分类结果

五、 总结

本文以 Python 为主要实现工具，利用 Kernel PCA 对图像进行降维，并基于降维后的数据进行图像分类，对比了 KNN、SVM 以及朴素贝叶斯三种分类模型的效果。基于前文的实验与结果，以下几点总结：

1. 模型带有超参数时，需要通过实验进行参数优化。参数优化的常用方法是交叉验证，目标通常是使得模型获得较好的泛化能力而不是对训练数据的拟合能力。参数调优范围的选择需要结合优化问题的性质以及多次实验的结果，以避免获得局部最优。
2. 本文所采用的图像数据量较小，但采用并行计算仍旧比较耗时。实际应用中的训练集数据量往往比较庞大，对数据处理、模型训练的效率有更高的要求，除了计算硬件上的支持，算法上的优化也十分重要。
3. 图像分类模型训练样本的选择对模型性能有较大的影响，特征类似但所属类别不同的样本更容易被误分类。对于这样的分类任务，则要求模型能够更多地捕捉到数据中的细节信息，因此相比本文所使用的分类方法，深度学习的方法可能能够更好地处理图像识别任务。