文章编号：2095-2163(2020)12-0185-04 中图分类号：TP183 文献标志码：A

四种机器学习算法在 MNIST 数据集上的对比研究

肖 驰
（韩山师范学院 计算机与信息工程学院，广东 潮州 521041）

摘 要：MNIST 数据集是检验机器学习算法性能常用的数据集。本文以 MNIST 数据集为例，研究四种机器学习方法的性能。首先，介绍支撑向量机、随机森林、BP 神经网络和卷积神经网络；其次，将四种学习方法在 MNIST 数据集上训练学习；最后，对四种学习模型的性能做对比分析。就实验结果而言，卷积神经网络在性能上优于其它三种学习算法。

关键词：支撑向量机；随机森林；BP 神经网络；卷积神经网络；MNTST 数据集

Comparative study of four machine learning algorithms on the MNIST dataset

XIAO Chi
（School of computer and information engineering, Hanshan Normal University, Chaozhou Guangdong 521041, China）

【Abstract】 The MNIST dataset is a commonly used data set for testing the performance of machine learning algorithms. This paper uses the MNIST dataset as an example to study the performance of four machine learning methods. First, support vector machines, random forests, BP neural networks and convolutional neural networks are introduced. Then the four learning methods are trained and learned on the MNIST data set. Finally, the performance of the four learning models is compared and analyzed. As far as the experimental results are concerned, the performance of the convolutional neural network is better than the other three learning algorithms.

【Key words】 Support vector machine; Random forest; BP neural network; Convolution neural network; MNIST dataset

0 引 言

万物互联的时代，人工智能在人们的生产和生活中扮演重要的角色，机器学习是人工智能中最前沿的研究领域之一[1]。机器学习算法是从海量数据中学习，并给出相应的决策或者预测[2]。

MNIST 数据集是由手写数字的图片和相应的标签组成，一共有 10 类，分别对应数字从 0～9[3]。训练图片一共有 60 000 张，可采用学习方法训练出相应的模型。测试图片一共有 10 000 张，可用于评估训练模型的性能。

本文拟应用支撑向量机、随机森林、BP 神经网络和卷积神经网络四种机器学习方法在 MNIST 数据集上训练相应的模型，并评估不同模型的性能。

1 机器学习方法

大数据时代，需要借助机器学习方法挖掘潜在知识信息，并给出所需的决策或者预测。如谷歌的“人机大战”，AlphaGo 的胜利本质上是信息技术综合运用的胜利，是大数据的胜利，也是机器学习的胜利[4]。本文从四个方面来探讨 MNIST 数据集上的数据建模，并分类预测。

1.1 支持向量集

支持向量机（SVM）是一种基于统计学习理论的二分类模型，在特征空间上使用间隔最大化进行分类。对于线性可分问题，SVM 利用二次规划来实现最大分类间隔；对于非线性分类问题，SVM 先通过合适的核函数将输入空间映射到高维空间，再在高维空间实现线性可分。SVM 非线性映射的目的是寻求一个最优的超平面使分类间隔最大，同时，在样本数据的学习泛化能力和映射的复杂性方面达到最佳[5]。

假设 x 表示二分类的数据点，用 y 表示类别，那么线性分类器的学习目标是在 n 维的数据空间寻找到一个超平面，这个超平面的方程如下：

$$w^{\mathrm{T}}x + b = 0. \tag{1}$$

其中，w^{T} 中的 T 代表转置；w 和 b 分别表示权重和偏置；在超平面 $w^{\mathrm{T}}x + b = 0$ 确定的情况下，$|w^{\mathrm{T}}x + b|$ 表示点 x 到距离超平面的远近。如图 1 所示，五角星和圆点分别表示两种类别，分布在超平面的两侧。

超平面离数据点的“间隔”越大，分类的确信度也越大。为了提高分类的确信度，需要让所选择的

基金项目：韩山师范学院一般项目（LY201801）；潮州市科技局项目（2018GY20）。

作者简介：肖 驰（1971-），男，学士，讲师，主要研究方向：信号处理。

收稿日期：2020-05-20

超平面能够最大化这个“间隔”。目标函数为：

$$\min \frac{1}{2} \| w \|^2 \quad \text{s.t.} \quad y_i(w^T\varphi(x_i) + b) \geqslant 1, i = 1, \cdots, n. \tag{2}$$

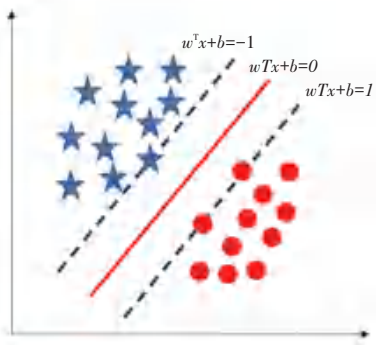


图 1 SVM 的分类超面

Fig. 1 Classification hyperplane of SVM

其中，$\varphi(x)$ 是非线性映射，将数据从低维空间映射高维空间。目标函数是二次的，约束条件是线性的，(2)是一个凸二次规划问题。

由于数据集中存在噪声，引入一个松弛变量 ξ，来允许一些数据可以处于超平面错误的一侧，如：

$$\begin{aligned} &\min_{w,b,\xi_i} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{m} \xi_i \\ &\text{s.t. } y_i(w^{\mathrm{T}}\varphi(x_i) + b) \geqslant 1 - \xi_i \\ &\xi_i \geqslant 0, \ i = 1, 2, \cdots, m \end{aligned} \tag{3}$$

其中，C 是惩罚参数，表示对分类错误的惩罚度。

通过拉格朗日乘子法将公式(3)转换为原问题等价的对偶问题，从而得到原始问题的最优解。目标函数变为：

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \theta(x_i, x_j) \\ &\text{s.t. } \sum_{i=1}^{m} \alpha_i y_i = 0, \\ &0 \leqslant \alpha_i \leqslant C. \end{aligned} \tag{4}$$

其中，$\theta(x_i, x_j)$ 为非线性核函数，α_i 为拉格朗日乘子。

1.2 随机森林

决策树(DT)和 SVM 都是单个分类器，都有性能提升的瓶颈以及过拟合的问题。因此，集成多个分类器来提升预测性能的方法应运而生。随机森林(RF)是一种 Bagging 集成学习方法，利用多个 DT 对样本进行训练、分类并预测的一种算法[6]。

1.2.1 DT

DT 是基于树结构的决策，建树的过程主要包括自根至叶的递归过程和在每个中间结点寻找一个“划分”属性。不同 DT 采用不同的准则来寻找划分属性，如 ID3 是基于信息熵，C4.5 基于信息增益率，CART 是基于 Gini 不纯度。但是，没有一种算法能在所有数据集上得到最好的分类精度。

假设样本数据集 D，第 k 类样本所占的比例为 P_k，$|y|$ 为类别个数，则 D 的信息熵：

$$Ent(D) = -\sum_{k=1}^{|y|} P_k \log_2 P_k. \tag{5}$$

假设属性 α 有 V 个可能的取值，如果 a 来对样本集 D 进行划分，D^v 表示第 v 分支结点包含了 D 中所在属性 α 上取值为 α^v 的样本。属性 α 对样本 D 划分的信息增益：

$$Gain(D,a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v). \tag{6}$$

由于信息增益偏重对取值数目较多的属性，为了减少这种影响，提出了有信息增益率：

$$Gain\_ratio(D,a) = \frac{Gain(D,a)}{IV(a)}. \tag{7}$$

$$IV(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|}$$

属性 α 的可能取值数目越多，则 $IV(\alpha)$ 的值通常就越大。信息增益率本质是在信息增益的基础之上乘上一个惩罚参数。特征个数较多时，惩罚参数较小；特征个数较少时，惩罚参数较大。基尼指数：

$$Gini(D) = \sum_{k=1}^{|y|} \sum_{k' \neq k} P_k P_{k'} = 1 - \sum_{k=1}^{|y|} P_k^2. \tag{8}$$

表示在样本集合 D 中一个随机选中的样本被分错的概率。$Gini(D)$ 越小，数据集 D 的纯度越高。

DT 对给定训练样本学习分类规则，不需要先验知识。DT 可能存在过度分割训练样本，使得树结构复杂，导致过拟合。尽管可以通过剪枝避免过拟合问题，但会增加 DT 的算法复杂性。

1.2.2 Bagging 思想

Bagging 是一种并行集成学习方法。首先，如果训练集大小为 N，对于每棵树而言，随机且有放回地从训练集中的抽取 N 个训练样本，作为该树的训练集；其次，训练并建立相应的 DT；最后，通过投票决定分类结果。这种策略具有很好的抗噪性，从而提高准确度。

RF 算法是在给定数据集上产生 m 个 DT 为基分类器，进行集成学习得到的一个组合分类器，其输出是由每个 DT 投票决定的，算法流程如图 2 所示。RF 将 Bagging 集成学习方法和随机子空间相结合。

随机是 RF 的核心，一是随机在原始训练数据中有放回的选取等量的数据作为训练样本；二是在建立 DT 时，随机的选特征中选取一部分特征建立 DT。通过随机的选择样本和属性特征，降低了 DT 之间的相关性。RF 解决 DT 性能瓶颈的问题，对高维数据分类具有良好的可扩展性和并行性。

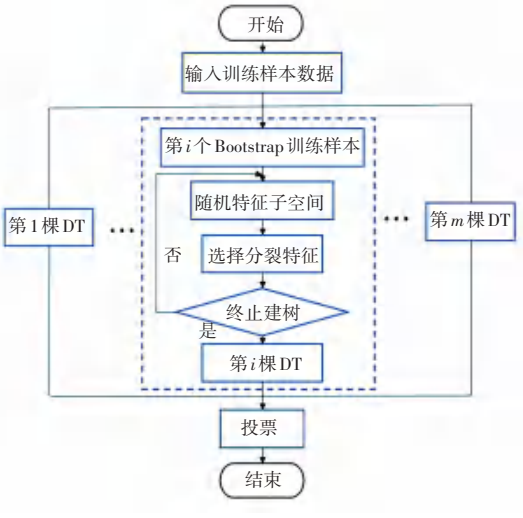


图 2　RF 流程图

Fig. 2　RF flow chart

1.3　BP 神经网络

BP 神经网络包括输入层、隐藏层和输出层，通过神经元相互连接而成[7]。每个神经元都有输入和输出，神经元模型如图 3 所示。

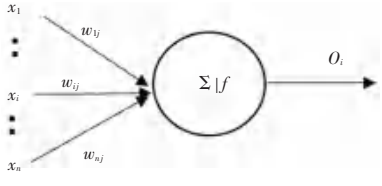


图 3　神经元

Fig. 3　Neuron

其中，$x_1, x_2, \ldots, x_n$ 表示某个神经元的 n 个输入；w_{ij} 表示第 i 个神经元与第 j 个神经元的权重；b_j 是偏置参数；神经元模型将这些输入加权后经激励函数 f 输出：

$$O_j = f\left(\sum_{j=1}^{n} x_i w_{ij} - b_j\right). \tag{9}$$

在 BP 神经网络中，前向传播的信息流为输入层->隐藏层->输出层，反向传播残差信号是输出层->隐藏层。根据残差不断从后向前调整网络权重和偏置，使残差减小以满足目标。通过网络学习提取样本数据的内在特征并输出，此方法有很强的泛化能力和非线性映射能力。

三层 BP 算法的具体步骤如下：

（1）初始化网络中的权重和偏置项，$w_{ih}^{(0)}, b_{ih}^{(0)}, w_{ho}^{(0)}, b_{ho}^{(0)}$。

（2）激活向前传播，得到各层输出和损失函数的期望：

$$E(\theta) = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{10}$$

其中，y_i 是真实值，$\hat{y}_i$ 是预测值。

（3）根据损失函数，计算输出单元的残差项和隐藏单元的残差项，并更新权重和偏置项。

（4）判断损失函数小于给定的阈值或者迭代次数，若是则结束，否则转（2）。

1.4　卷积神经网络

卷积神经网络（CNN）是由输入层、卷积层、激活函数、池化层、全连接层组成[8]。CNN 保留了空间信息，特征图上的一个点对应输入图上的区域，从一个局部区域学习到的信息，应用到图像的其它地方。因此，可以更好地适应图像分类。不同的特征通过多个不同的卷积核实现。不同位置共享相同权值，可实现数据的不同位置检测相同的模式。

1.4.1　卷积

卷积层由一组滤波器组成，在图像的某个位置上覆盖滤波器；将滤波器中的值与图像中的对应像素的值相乘；把这些乘积累加起来，得到的和是输出图像中目标像素的值；对图像的所有位置重复此操作，卷积过程如图 4 所示。

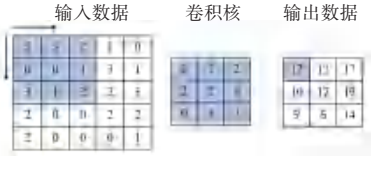


图 4　卷积过程

Fig. 4　Convolution

1.4.2　池化

图像中相邻像素倾向于具有相似的值，卷积层输出中包含的大部分信息都是冗余的。池化层操作可以对输入的特征图进行压缩，提取主要特征，简化网络计算复杂度。池化操作如图 5 所示，一种是均值池化，一种是最大池化。

图 5　池化过程

Fig. 5　Pooling

2 实验仿真结果

本文采用 SVM、RF、BP、CNN 对 MNIST 数据集分类,仿真硬件环境 Macbook Pro,软件环境 Python 3.6 。

SVM 中采用的核函数为 rbf,惩罚参数为 1 000,提高训练集的准确率。RF 采用以 Gini 划分为 100 个 CART 树,内部节点需划分最小样本数为 2。BP 神经网络采用 200 次迭代,一次训练样本数为 128,采用随机梯度优化算法,在最后一层使用 *Softmax* 激活函数。CNN 神经网络框架如图 6 所示,卷积层 1 为 5×5 卷积核,卷积层 2 为 3×3 卷积核,池化层采用最大池化层。

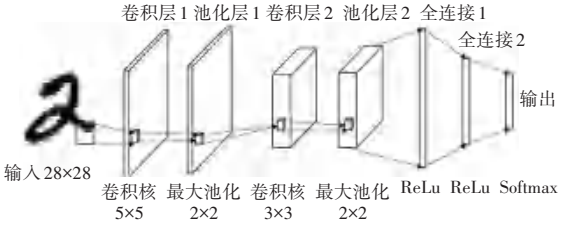


图 6 CNN 框架

Fig. 6 CNN framework

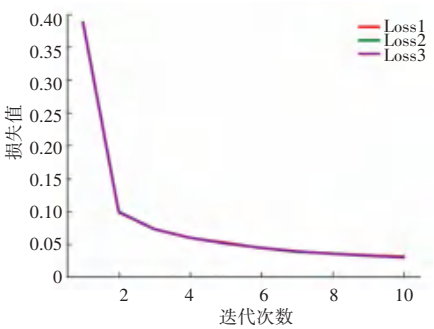
SVM、RF、BP 和 CNN 在 MNIST 的训练集与测试集上的准确率见表 1。其中 RF 在训练集最高 100%,但测试集则 96.91%,出现过拟合现象;BP 神经网络在训练集和测试集上都是最低,但其运行速度最快;SVM 性能比 BP 神经网络高,但 SVM 训练时间过长;卷积神经网络表现优秀,没出现 RF 的过拟合,在测试集上的准确率最高。

表 1 不同学习方法的准确率

Tab. 1 Accuracy of different learning methods

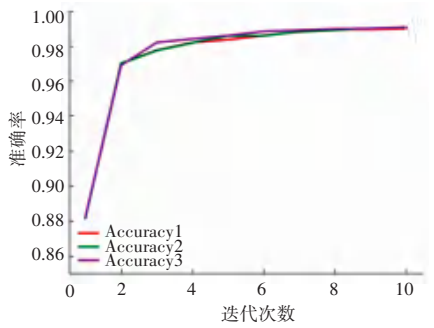
	SVM	RF	BP	CNN
训练集	94.71	100	92.34	99.48
测试集	94.46	96.91	92.27	99.11

CNN 在运行不同次数下迭代次数与分类精确和损失函数值的关系如图 7 所示。CNN 迭代 10 次,其损失随着迭代次数迅速收敛,准确率也越来越高,并趋于迅速稳定。



(a) 损失

(a) Loss



(b) 准确率

(b) Accuracy

图 7 CNN 的损失和准确率

Fig. 7 Loss and accuracy of CNN

3 结束语

通过四种机器学习方法对 MNIST 数据库分类,就训练集和测试集的分类识别准确率来说,CNN 无疑是最优的,数据量越大,准确率越高。CNN 代表深度学习的一种算法,而深度学习代表人工智能的前沿,未来需要优化深度学习框架,以轻量级的框架来提取特征,减少训练时间,提高图像分类的精度。

参考文献

[1] 周志华. 机器学习[M]. 北京:清华大学出版社, 2016.
[2] 张润,王永滨.机器学习及其算法和发展研究[J]. 中国传媒大学学报(自然科学版),2016,23(2):10-18,24.
[3] http://yann.lecun.com/exdb/mnist/.
[4] http://news.sciencenet.cn/sbhtmlnews/2016/3/310264.shtm.
[5] 张松兰.支持向量机的算法及应用综述[J]. 江苏理工学院学报,2016,22(2):14-17+21.
[6] 王奕森,夏树涛.集成学习之随机森林算法综述[J]. 信息通信技术,2018,12(1):49-55.
[7] 王宏涛,孙剑伟.基于 BP 神经网络和 SVM 的分类方法研究[J]. 软件,2015,36(11):96-99.
[8] 巴桂.基于卷积神经网络的图像分类算法[J]. 电脑与信息技术,2020,28(1):1-3.