一种加权的深度森林算法

宫振华[1] 王嘉宁[2] 苏 翀[2*]

[1](南京机电职业技术学院 江苏 南京 211135)

[2](江苏科技大学电气与信息工程学院 江苏 苏州 215600)

摘 要 深度森林 DF(Deep Forest)由多粒度扫描和级联森林两个部分组成。其中:多粒度扫描通过滑动窗口技术获取多个特征子集,以增强级联森林的差异性;级联森林则是将决策树组成的森林通过级联方式实现表征学习。因此,深度森林克服深度学习参数依赖性强、训练开销大以及仅适用于大数据集等不足之处。然而,深度森林中各个子树的预测精度是各不相同的,简单算术平均会导致子树的错误预测对整个森林的预测产生影响,进而随着级数增加,有可能使错误被进一步放大。为此,提出一种根据森林中每棵子树的预测精度进行加权的深度森林。在高维和低维数据集上进行实验,结果表明:加权的深度森林在高维和低维数据集上性能都获得一定提升,特别在高维数据集上优势较为明显。

关键词 深度森林 多粒度扫描 级联森林 加权

中图分类号 TP3 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.02.049

A WEIGHTED DEEP FOREST ALGORITHM

Gong Zhenhua[1] Wang Jianing[2] Su Chong[2*]

[1](*Nanjing Institute of Mechatronic Technology*, *Nanjing 211135*, *Jiangsu*, *China*)

[2](*School of Electrical and Information Engineering*, *Jiangsu University of Science and Technology*, *Suzhou 215600*, *Jiangsu*, *China*)

Abstract Deep forest can be categorized into two parts: multi-grained scanning and cascade forest. Multi-grained scanning obtains multiple feature subsets through sliding window technology to enhance the diversity of cascade forests. And cascade forest employs a cascade structure to realize representation learning. Thus, deep forest overcomes weaknesses such as strong dependence of deep learning parameters, high training cost and requirement for the big data. However, the prediction accuracy of each subtree in the deep forest is different. Simple arithmetic averages can lead to false predictions of subtrees and affect the entire forest. As the series increases, the error may be magnified further. Therefore, we proposed a weighted deep forest in which we calculate the weight according to the prediction accuracy of each subtree. Through experiments on high-dimensional and low-dimensional datasets, we find that the performance of weighted deep forests has been improved in both high-dimensional and low-dimensional datasets, especially in high-dimensional datasets.

Keywords Deep forest Multi-grained scanning Cascade forest Weighting

0 引 言

现时,深度学习网络模型取得了长足的进步和发展,这一切要得益于大数据时代的到来。虽然深度学习网络模型为数据分析提供了有力的工具,但其所具有的大量调解参数和复杂的网络结构都在一定程度上决定了模型的学习效果。此外,由于深度学习网络模型在训练时需要大量的训练样本,因此,整个过程既耗时又对设备有较高的要求。而深度森林 DF 是一种基于深度模型提出的级联随机森林方法,它具有较少的调节参数,允许使用者可以根据设备的资源决定训练

收稿日期:2018-09-06。中国博士后科学基金项目(2016M600430)。**宫振华**,高工,主研领域:数据挖掘。**王嘉宁**,本科生。**苏翀**,研究员。

的耗费,并能自适应地调节训练模型层数。文献[1]的实验结果充分表明,相较于深度学习网络模型,深度森林取得了更好的分类性能。在结构层面,深度森林由多粒度扫描和级联森林两个部分组成[1]。其中,多粒度扫描通过滑动窗口技术获取多个特征子集,以增强级联森林的差异性。级联森林则是将决策树组成的森林通过级联方式实现表征学习。可以说,深度森林沿用了深度学习对样本特征属性的逐层处理机制,利用多级结构实现表征学习。与深度学习不同之处主要表现在以下几个方面:

(1) 深度森林的级数是随着训练的不断不深入自动调节的;

(2) 深度森林具有很少的超参数且对超参数不敏感;

(3) 深度森林具有较低的训练开销,既适用于大规模数据集,也适用于小规模数据集;

(4) 其结构适用于并行处理。

深度森林在结构上由多级组成,每级分别由随机森林[2]和完全随机森林[3]两种森林组成。就每个样本而言,每个森林将其各个子树预测的类概率向量进行算术平均后,作为该森林的预测结果,并与样本的原始特征向量拼接,作为下一级的输入。由于森林中各个子树的预测精度是各不相同的,算术平均会导致子树的错误预测对整个森林的预测产生影响,进而随着级数增加,有可能使错误被进一步放大。为了避免上述影响,本文提出了一种加权的深度森林 WDF (Weighted Deep Forest)。主要思想是根据森林中每棵子树的预测精度计算其相应权重,再对各个子树的预测概率向量进行加权求和,以提高深度森林的预测精度,降低级联级数。

1　深度森林

深度森林与深度神经网络都是通过多级结构进行表征学习[4],但深度森林以其简单的训练模型以及不依赖于大量数据进行训练的特点弥补了深度神经网络的缺点,并逐渐被应用于工程实践中[5-6]。

深度森林由多粒度扫描和级联森林两个部分组成。多粒度扫描主要处理高维数据和图像数据。假设长度为 n 的一维特征向量,若使用长度为 m 的窗口进行滑动且每次滑动一个单位长度,将产生 $n-m+1$ 个具有 m 维特征向量的数据子集;类似地,对于一个 $n\times n$ 的二维图像数据,若使用 $m\times m$ 的窗口进行滑动,每次滑动一个单位长度,将产生 $(n-m+1)^2$ 个具有 $m\times m$ 特征向量的数据子集。这些数据集将分别输入到 1 个完全随机森林和 1 个随机森林。对于 c 个类别的分类问题,经过两个不同的随机森林分类后,长度为 n 的一维特征向量将产生长度为 $2c(n-m+1)$ 的类向量;类似地,对于一个 $n\times n$ 的二维图像数据,将产生长度为 $2c(n-m+1)^2$ 的类向量。随后,这些类向量将被拼接到原始的样本的特征空间里,作为后面级联森林的输入特征。整个多粒度扫描结构如图 1 所示。

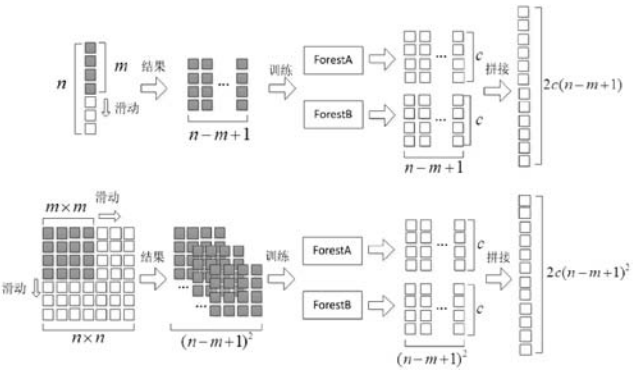

图 1　多粒度扫描

级联森林主要由随机森林和完全随机树森林两种森林组成。森林之间通过层级方式形成级联结构。对于每一级森林,首先,训练样本通过 k 折交叉验证,训练其中各棵子树,与此同时,每棵子树对每个训练样本给出一个预测的类概率向量;其次,将测试样本输入训练得到的森林,每棵子树对每个测试样本也给出一个预测的类概率向量;再次,森林对所有子树预测的类概率向量按训练样本和测试样本分别计算平均类概率向量;最后,该级的所有森林将所有样本上的平均类概率向量与样本的原始特征向量拼接后作为下一级森林的输入特征。每级结束预测后,会在验证集上对预测结果进行评估,以决定是否扩展下一级。如果不再扩展,则在已扩展的级中,找出最优评估结果所对应的级,将所有森林在测试样本上的平均类概率向量算术平均后,取概率最高的类向量作为整个深度森林的预测结果。其中,级联森林结构和单个森林的结构分别如图 2 和图 3 所示。

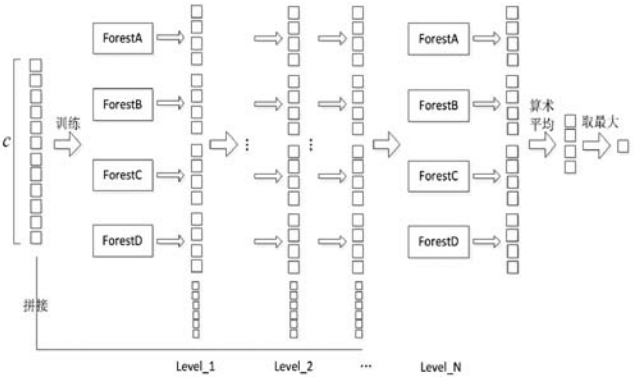

图 2　级联森林结构

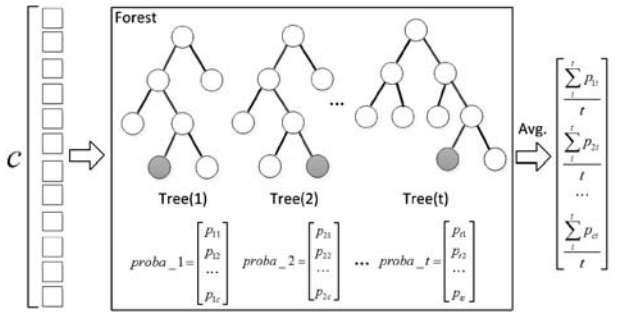

图 3 单个森林结构

2 加权的深度森林

深度森林中各棵子树对应不同的预测精度,简单的算术平均法,忽略了各棵子树之间的预测差异,使预测错误率较高的子树对整个森林的预测结果产生较大影响。为此,本文对森林中各棵子树的结合策略进行改进,提出一种加权的深度森林 WDF,其中每个森林的结构如图 4 所示。

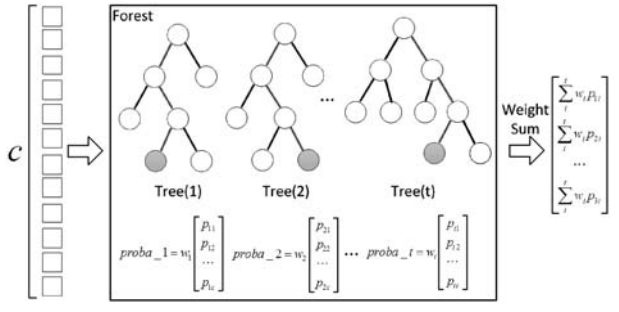

图 4 单个加权森林结构

假设训练样本集 T 和测试样本集 S 的示例个数分别为 m 和 n,类别标记的个数为 c,记为 $L=\{l_1,l_2,\cdots,l_c\}$。设森林 F 包含 t 棵子树,记第 k 棵子树为 $T_k(k\in[1,t])$,第 i 个训练样本被子树 T_k 预测为类 l_j 的概率为 p_{ij},则在训练样本集 T 上,子树 T_k 预测的类概率矩阵如下所示:

$$Proba(T_k,T)=\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1c} \\ p_{21} & p_{22} & \cdots & p_{2c} \\ \vdots & \vdots & & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mc} \end{pmatrix} \tag{1}$$

定义函数 $Max(\boldsymbol{X})$ 表示获取二维矩阵中行向量 $\boldsymbol{X}$ 的最大值元素所对应的列下标,当有多个相同的最大值时,取最小的列下标。令列下标从 1 开始,则子树 T_k 对训练样本集 T 的预测结果向量如下所示:

$$Predict(T_k,T)=\begin{pmatrix} Max(p_{11} & p_{12} & \cdots & p_{1c}) \\ Max(p_{21} & p_{22} & \cdots & p_{2c}) \\ \vdots & & & \\ Max(p_{m1} & p_{m2} & \cdots & p_{mc}) \end{pmatrix}=\begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{pmatrix} \tag{2}$$

定义函数 $Acc(\boldsymbol{X}_1,\boldsymbol{X}_2)$ 表示两个同维向量 $\boldsymbol{X}_1$、$\boldsymbol{X}_2$ 中相同位置元素相等的个数所占的比例。例如:

$$Acc\left(\begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix},\begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}\right)=\frac{2}{3}$$

令 $Y(T)$ 是由训练样本集 T 的真实类映射到类集合中的下标所组成的向量,则子树 T_k 对训练样本集 T 的预测准确率 a_k 为:

$$a_k=Acc(Predict(T_k,T),Y(T)) \tag{3}$$

至此,第 k 棵子树的权重[7-8] W_k 为:

$$W_k=\frac{\log_2(a_k/(1-a_k))}{\sum_{i=1}^{t}\log_2(a_i/(1-a_i))}$$

$$W_k\propto\log_2(a_k/(1-a_k)) \tag{4}$$

最后,森林 F 在训练样本集 T 和测试样本集 S 上预测的类概率矩阵分别为:

$$Proba(F,T)=\sum_{i=1}^{t}Proba(T_i,T)\times W_i \tag{5}$$

$$Proba(F,S)=\sum_{i=1}^{t}Proba(T_i,S)\times W_i \tag{6}$$

更进一步,假设级联森林中每一级又包含 h 个森林,那么第 e 级森林组合 C_e 在训练样本集 T 和测试样本集 S 上预测的类概率矩阵分别为:

$$Proba(C_e,T)=\frac{1}{h}\sum_{i=1}^{h}Proba(F_i,T) \tag{7}$$

$$Proba(C_e,S)=\frac{1}{h}\sum_{i=1}^{h}Proba(F_i,S) \tag{8}$$

类似地,如果把上述矩阵改写成行向量组的形式,可分别得到第 e 级森林组合 C_e 对训练样本集 T 和测试样本集 S 的预测结果向量,具体如下:

$$Predict(C_e,T)=\begin{pmatrix} Max(\tilde{r}_{T1}) \\ Max(\tilde{r}_{T2}) \\ \vdots \\ Max(\tilde{r}_{Tm}) \end{pmatrix} \tag{9}$$

$$Predict(C_e,S)=\begin{pmatrix} Max(\tilde{r}_{S1}) \\ Max(\tilde{r}_{S2}) \\ \vdots \\ Max(\tilde{r}_{Sn}) \end{pmatrix} \tag{10}$$

同理,令 $Y(T)$、$Y(S)$ 分别表示由训练样本集 T、测试样本集 S 的真实类映射到类集合中的下标所组成的向量,则第 e 级森林组合分别在训练样本集 T 和测试样本集 S 上预测准确率 A_{et} 和 A_{es} 分别为:

$$A_{et} = Acc(Predict(C_e, T), Y(T)) \quad (11)$$

$$A_{es} = Acc(Predict(C_e, S), Y(S)) \quad (12)$$

当级联森林不再扩展时,则在已扩展的级中,找出在训练样本集 T 上预测准确率最高值所对应的级,将该级森林组合在测试样本集 S 上预测结果向量和预测准确率作为整个加权深度森林的预测结果。加权的深度森林(WDF)如算法 1 所示。

算法 1　加权的深度森林(WDF)

输入　训练集 T,测试集 S,森林中子树的数目 N,每一级森林的数目 M

```
if T 是高维数据集 then
    T = 多粒度扫描(T);
end if
for i = 1 to M
    for j = 1 to N
        使用 T 训练子树;
        根据式(1) - 式(3)计算子树的准确率;
        根据式(4)计算当前子树的权重;
        输入测试集 T 到当前子树;
    end for
    根据式(5)、式(6)分别计算当前森林在训练集 T 和测试集 S 上的预测类概率矩阵;
end for
根据式(7)、式(8)分别计算当前级联森林在训练集 T 和测试集 S 上的预测类概率矩阵 P;
if 评估后继续扩展下一级 then
    将概率矩阵 P 拼接到原始特征空间,形成新的训练集 T* 和测试集 S*;
    返回 step 1 继续执行;
else
    在所有扩展的级中找出训练集上预测准确率最高的那一级 opt_lay_id,并输出该级在测试集 S 预测的结果;
end if
```

对比图 2 和图 4 可以发现,加权的深度森林可以利用权重值修正森林的类概率矩阵。当修正的概率矩阵作为下一级的输入时,会使下一级森林在训练过程中不断优化并提高其预测精度,在一定程度上,不仅能提高最终预测精度,还可以减少扩展级数。

3　实　验

本文在高维和低维数据集上分别对深度森林(DF)和加权的深度森林(WDF)进行实验比较。实验平台配置如下:160 GB 内存、24 核 CPU、64 位 Ubuntu16.04 操作系统、Anaconda2(Python2.7)、类库包括 Numpy、Scikit-learn、Tensorflow 等[9-10]。

为了公平比较,这里采用与文献[1]一致的实验参数,即每个森林包括 500 棵子树,随机森林每次随机选择的特征数是$\sqrt{d}$(d 表示特征总数);在多粒度扫描结构中,随机森林和完全随机森林各一个,滑动窗口的大小分别取$\lfloor \frac{d}{16} \rfloor$、$\lfloor \frac{d}{8} \rfloor$和$\lfloor \frac{d}{4} \rfloor$;在级联森林里,每级均包括 4 个随机森林和 4 个完全随机树森林,采用 3 折交叉验证方式。

3.1　实验数据集

采用文献[1]中所使用的实验数据集,与原文相同,每个数据集的 80% 用于训练,20% 用于验证。参与实验的高维数据集有:GTZAN[11]、SEMG[12]、MNIST[13]以及 IMDB[14];低维数据集有:ADULT[15]、YEAST[16]、和 LETTER[17],其中,低维数据集无需进行多粒度扫描。参与实验的数据集描述如表 1 所示。

表 1　实验数据集

	序号	数据集	特征数	样本数量
低维	1	ADULT	14	48 842
	2	YEAST	8	1 484
	3	LETTER	16	20 000
高维	4	GTZAN	3 640	100 000
	5	SEMG	500	1 800
	6	MNIST	784	70 000
	7	IMDB	1 024	50 000

3.2　实验结果及分析

实验分别采用测试集上的准确率和扩展的级数作为评价指标。具体实验结果如表 2 所示。

表 2　实验结果

	序号	数据集	准确度/%		级数	
			DF	WDF	DF	WDF
低维	1	ADULT	85.76	86.17	13	19
	2	YEAST	61.66	61.83	6	5
	3	LETTER	97.22	97.42	6	6
		Avg	**81.55**	**81.81**	**8.33**	**10**
高维	4	GTZAN	67.00	68.67	13	10
	5	SEMG	72.41	75.56	5	5
	6	MNIST	99.25	99.40	17	12
	7	IMDB	89.14	91.08	16	17
		Avg	**81.95**	**83.68**	**12.75**	**11**

从表 2 可以看出,在低维数据集上,加权的深度森林预测准确率要略高于深度森林,但扩展级数多于深度森林。与此相反的是,在高维数据集上,无论准确率还是扩展级数,加权的深度森林都要优于深度森林。出现这一现象,主要有以下几点原因:

(1) 低维数据集包含的特征数较少,造成了森林中训练得到的子树之间的差异较少,则每棵子树的预测准确率较为接近。从式(4)可以看出,每棵子树的权重也较为接近。因此,其性能提高有限。

(2) 高维数据集往往包含较多的特征数,再经过多粒度扫描处理后,非常有利于增加后续级联森林中训练所得子树之间的差异。由于每棵子树的预测准确率波动较大,最终导致差异较大的权重分布。

由上述分析可知,深度森林中训练得到的子树之间的差异往往决定了最后的预测精度和级联数,而子树之间的差异性却受到数据集中特征数的影响。这与级联森林主要由随机森林和完全随机树森林两种森林组成有关。通过表 1 和表 2 可以看出,由于高维数据集上训练得到的子树之间的差异性增加,每棵子树的预测准确率差异也较大,加权的思想会赋予准确率高的子树较大的权重,以增加其在决策中的作用。因此,加权深度森林优势在高维数据集上具有明显的优势,在准确率提高的同时,森林的级联数也有所减少,降低了训练时间;而在低维数据集上,由于权重之间差异性较小,延缓了收敛速度,增加了森林的级联数,虽然增加了训练时间。但与原算法相比,还是获得了可比的性能。综上所述,加权的深度森林更适合利用多粒度扫描处理高维数据集。

4 结 语

深度森林沿用了深度学习对样本特征属性的逐层处理机制,但克服了深度学习参数依赖性强、训练开销大以及仅适用于大数据等不足之处。然而,深度森林中各个子树的预测精度是各不相同的,简单算术平均会导致子树的错误预测,对整个森林的预测产生影响,进而随着级数增加,有可能使错误被进一步放大。为此,本文提出了一种根据森林中每棵子树的预测精度进行加权的深度森林。通过在高维和低维数据集上进行实验,结果表明:加权的深度森林在高维和低维数据集上性能都获得了一定提升,特别在高维数据集上,这一优势较为明显。由于目前所使用的加权方式较为简单,下一步将进一步考虑更为全面的权重评估方式,比如综合集成分类器的多样性和分类性能的权重评估方式等。

参 考 文 献

[1] Zhou Z H, Feng J. Deep forest: Towards an alternative to deep neural networks[J]. eprint arXiv:1702.08835, 2017.

[2] 李建更,高志坤. 随机森林针对小样本数据类权重设置[J]. 计算机工程与应用,2009,45(26):131-134.

[3] 姚明煌. 随机森林及其在遥感图像分类中的应用[D]. 厦门:华侨大学,2014.

[4] 王剑云,李小霞. 一种基于深度学习的表情识别方法[J]. 计算机与现代化,2015(1):84-87.

[5] 刘广东,邱晓晖. 基于多模式 LBP 与深度森林的指静脉识别[J]. 计算机技术与发展, 2018, 28(7): 83-87.

[6] 朱晓妤,严云洋,刘以安,等. 基于深度森林模型的火焰检测[J]. 计算机工程, 2018, 44(7): 264-270.

[7] Kuncheva L I. Combining pattern classifiers: methods and algorithms[M]. New York: Wiley-Interscience,2004.

[8] 刘擎超. 多分类器加权集成技术研究[D]. 镇江:江苏大学,2011.

[9] Kramer O. Scikit-Learn[M]. Springer, 2016.

[10] Abadi M, Barham P, Chen J, et al. TensorFlow: a system for large-scale machine learning[C]//Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation. Berkeley :USENIX Association, 2016.

[11] Tzanetakis G, Member S, Cook P. Automatic musical genre classification of audio signals[J]. IEEE Transactions on Speech & Audio Processing, 2002, 10(5):293-302.

[12] Sapsanis C, Georgoulas G, Tzes A, et al. Improving EMG based classification of basic hand movements using EMD[C]//Engineering in Medicine and Biology Society. IEEE, 2013:5754-5757.

[13] Lécun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.

[14] Maas A L, Daly R E, Pham P T, et al. Learning word vectors for sentiment analysis[C]//Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011.

[15] Newadult. UCI adult income data set, adapted[Z].

[16] Feng J, Dai X, Qian X, et al. New insights into two distinct nucleosome distributions: comparison of cross-platform positioning datasets in the yeast genome[J]. Bmc Genomics, 2010, 11(1):33.

[17] Lichman M. UCI machine learning repository[Z]. 2013.