武漢大學
WUHAN UNIVERSITY

Deep Forest: Towards an Alternative to Deep Neural Networks

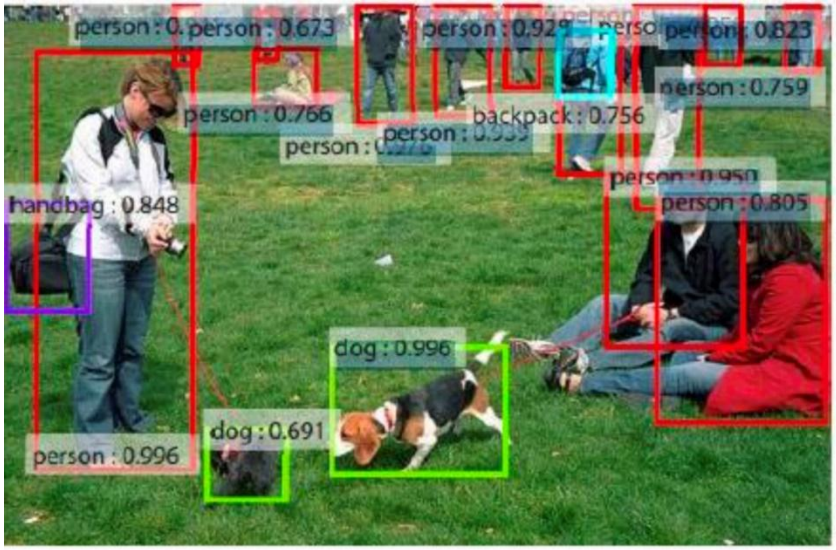
Liu Fang

2018.11.15

CONTENTS

简介 - Deep Learning

深度学习的巨大成就



每个人，闭上眼睛的时候，才能真正面对光明

他们在吱呀作响的船舷上，静静看着世界，没有痛苦的声音，碎裂的海洋里摇晃出阵阵沉默，吞噬过来。他们的躯体，一点，一点，逐渐暗淡在

你们虔诚的看着远方，我抬起头，不经意间，目光划过你们的面庞，上面淡淡的倔强印，那么坚强

尘世凡间
沉睡亿万光年的年轻战士
萦绕不散的寂寞烟云中
静候在末世岛屿之上
守候，女王何时归来
你的目光延向她迟归的方向
缓缓推进的海浪
这最后一夜
荡漾

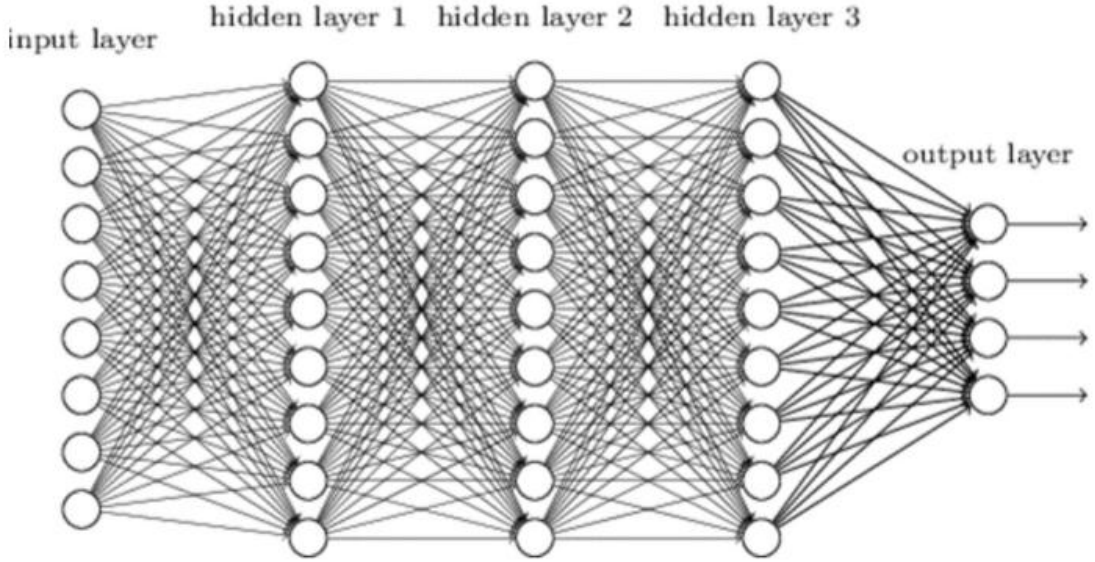
简介 - Deep Learning

何为深度 Deep?

增加模型的复杂性

→提高学习能力

- 添加隐藏的单元 (model width)
- 添加隐藏层 (model depth)



简介 - Deep Learning

深度学习的本质 - 表征学习

Previous

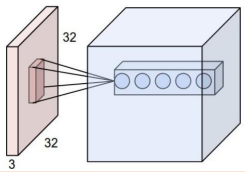
Manual feature selection

Classifier learning

With deep learning

Representation learning

Automatical feature learning



Classifier learning

简介 - Deep Learning

深度学习出现的问题

- 增加模型的复杂性

  → 增加过度拟合的风险

  → 增加训练难度

- 过拟合: 大量的训练数据，预测模型
- 训练: 强大的计算设施，如GPU

简介 - Deep Learning

最重要的是深度模型

- 逐层处理Layer-by-layer
- 功能转换
  - ✓ CNN
  - ✓ RNN

优点

- 计算的复杂性
- 需要手动调整参数
  - hidden layers
  - hidden units
  - Type of hidden units

缺点

简介 – 集成学习 Ensemble Learning

如何解决？

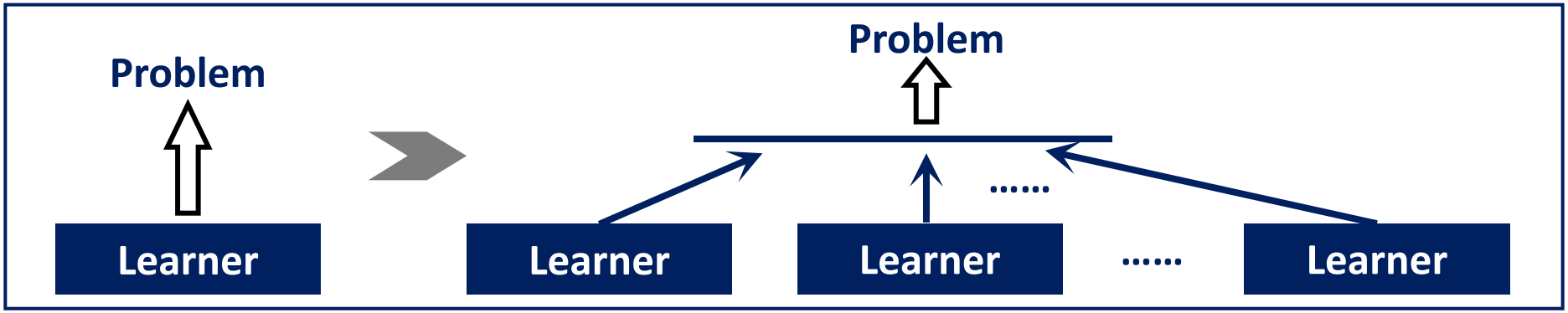
使用集成学习

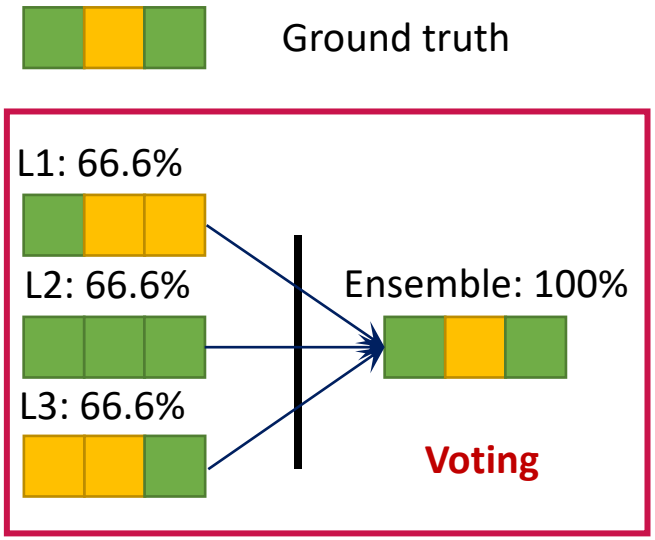
在过去的十年里，在KDDCup, Netflix等比赛中，集成学习获得了多次奖项。

- KDDCup 07: 1st place for "…Decision Forest and …"
- KDDCup 15: 1st place for "Three-Stage Ensemble and Feature Engineering for MOOC Dropout Prediction
- KDDCup 16: 1st place for "Gradient Boosting Decision Tree"

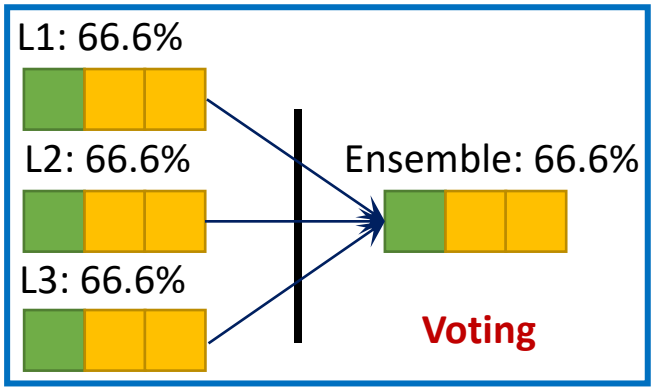
简介 - 集成学习Ensemble Learning

使用多个学习者解决问题

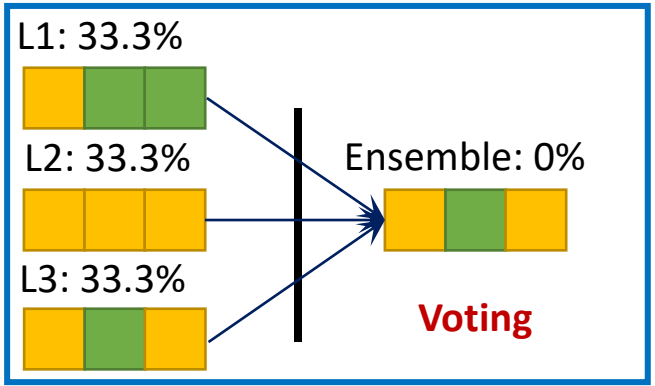




Ensemble really helps



Individuals must be different



Individuals must be not bad

简介 - 集成学习 Ensemble Learning

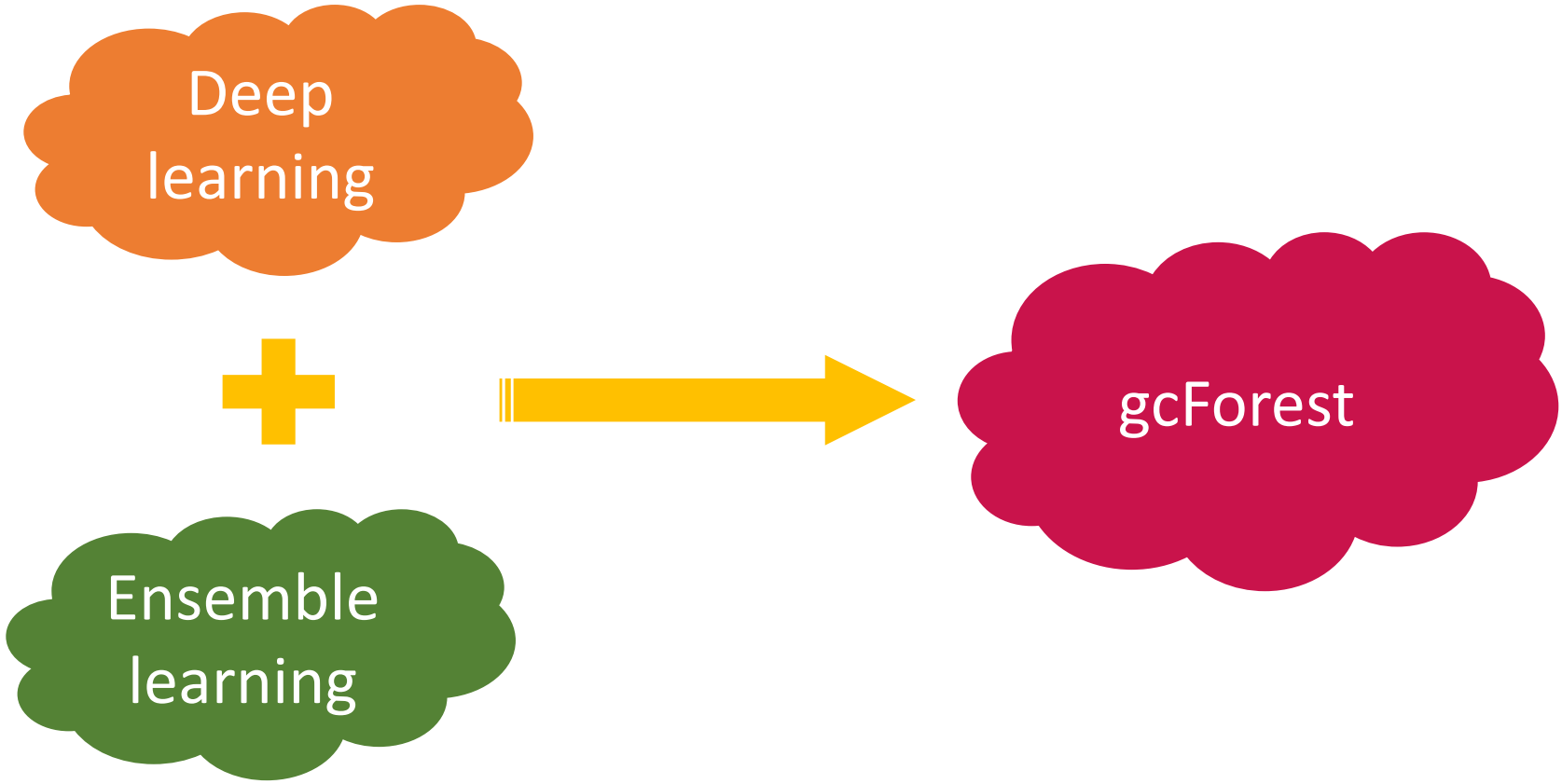
最重要的是集合模型

整体误差 = 个人的错误 − 个体的多样性

- individual
  - 个体学习者越准确，整体模型就越好
- Diversity of individual
  - 个体学习者越多样化，整体模式就越好

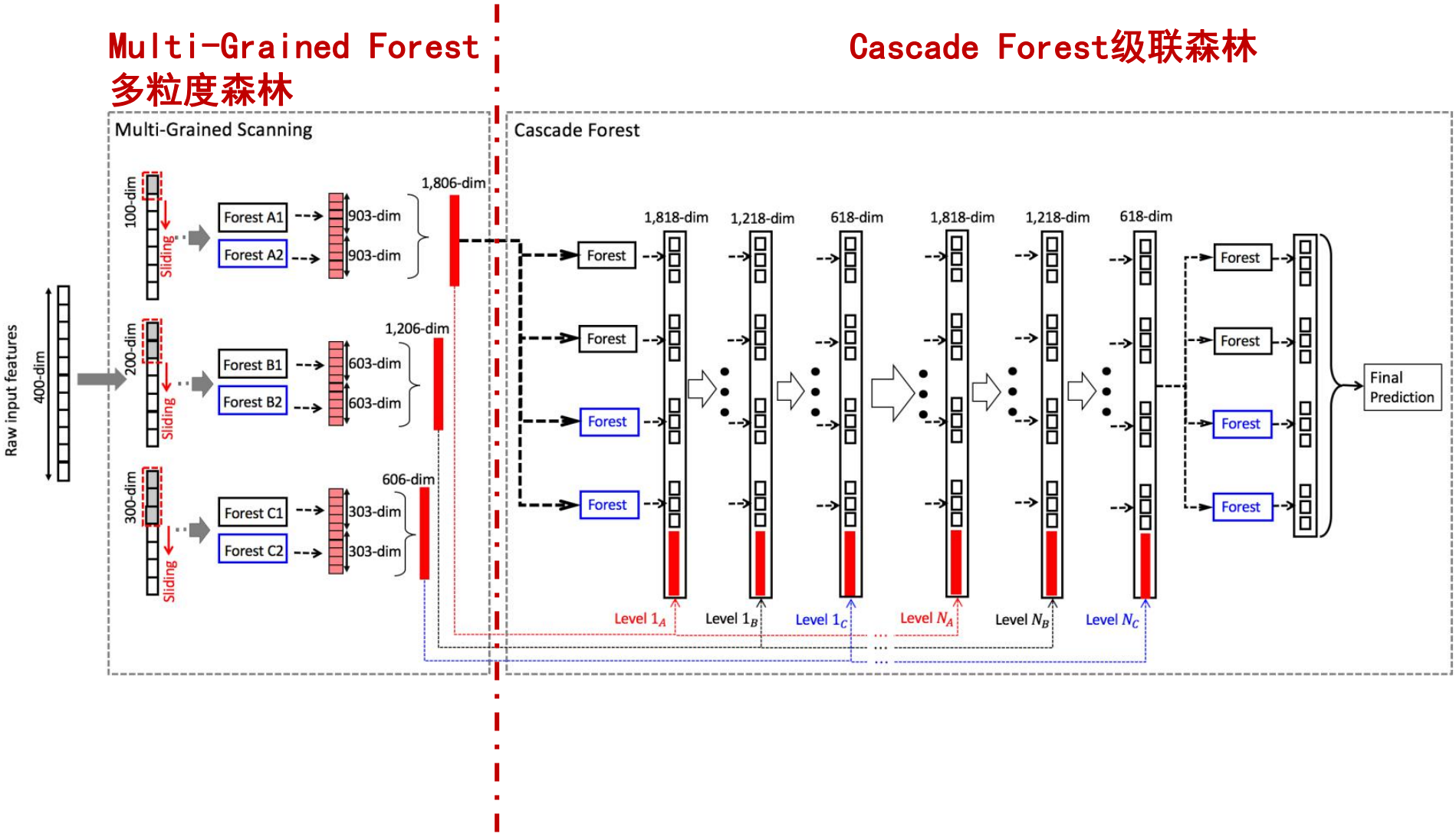
简介 - gcForest（Multi-Grained Cascade Forest）

思路

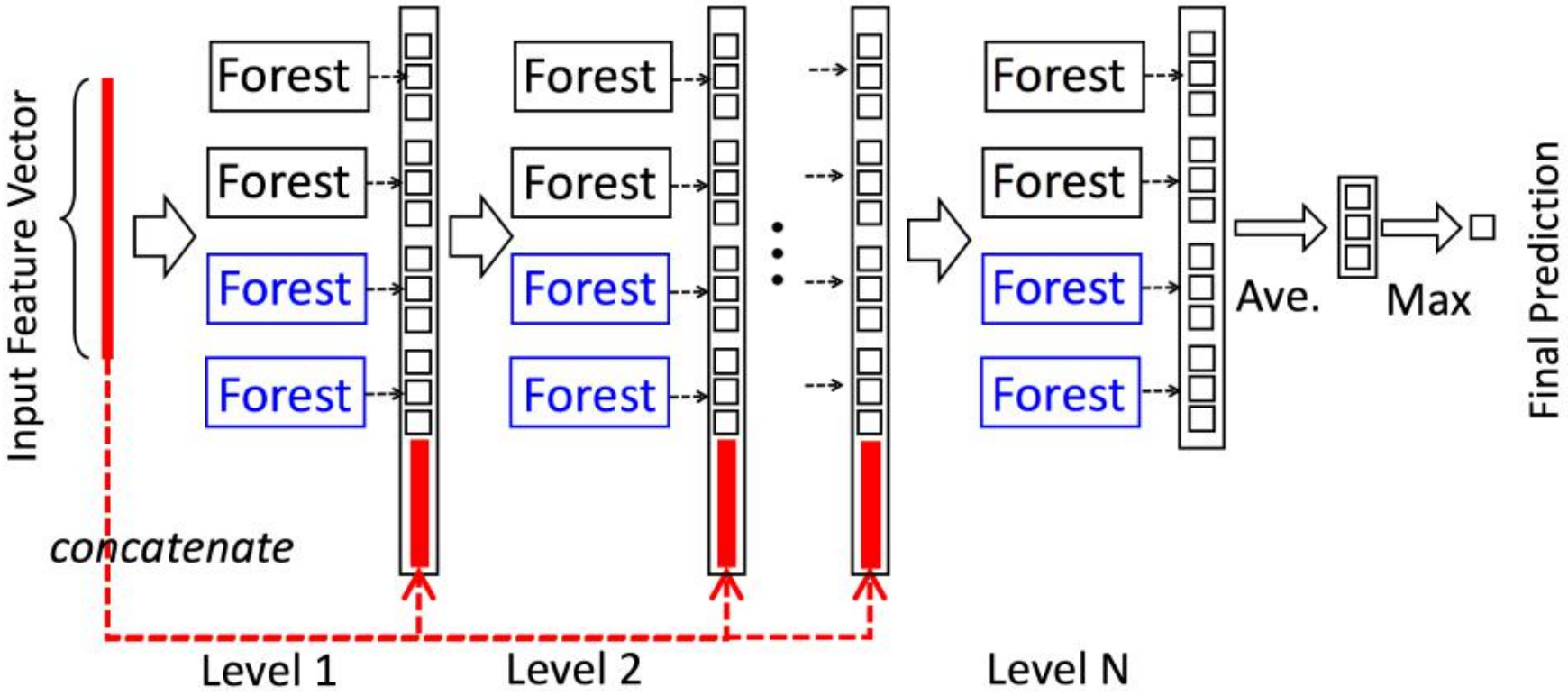


方法

gcForest的基本原理



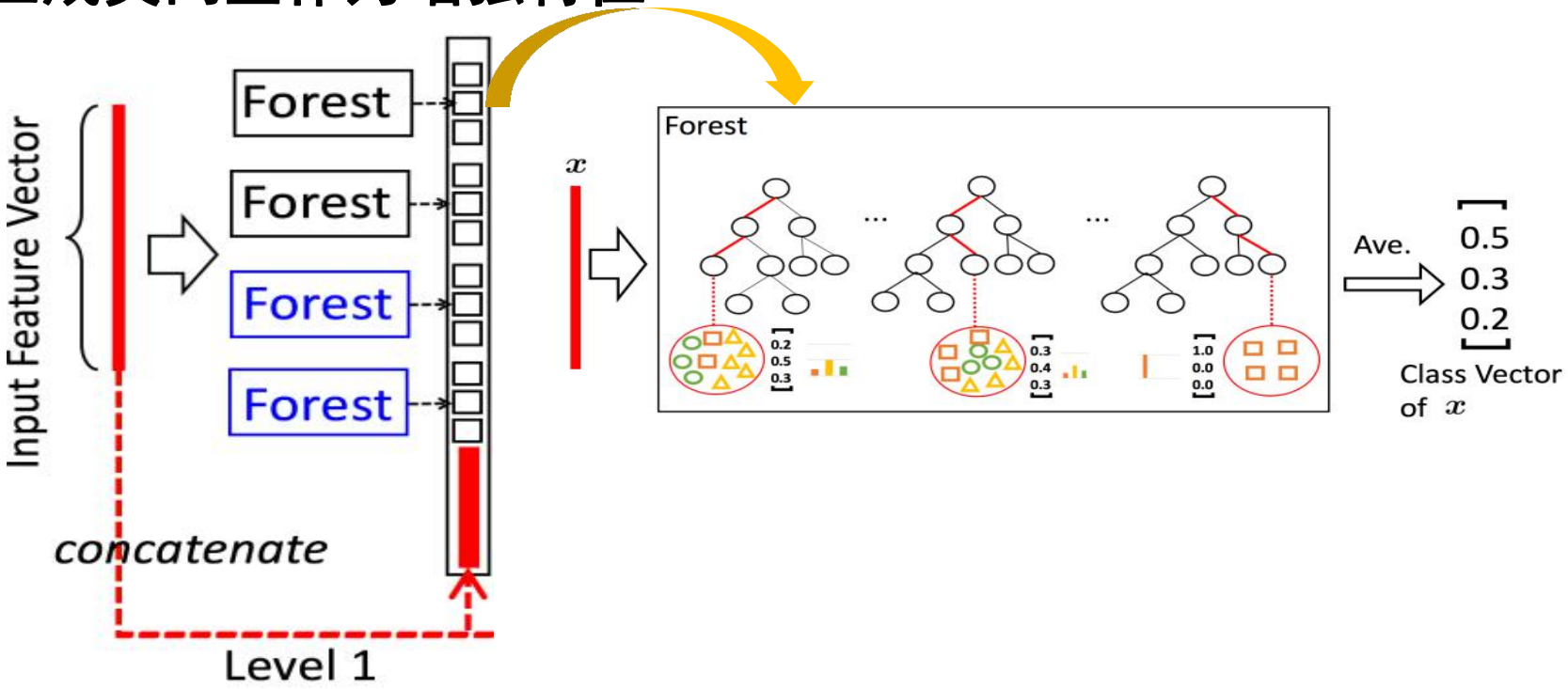
方法-cascade forest 结构图示



1. 每一个水平都是一系列决策森林的ensemble
2. 每一层上为了多样性的考虑，选择了两种不同的森林，分别是completely-random trees和一般的rt，图中有2个completely-random trees和2个一般的rt，每一个completely-random trees或者一般的rt都包含500棵树。

方法 - Cascade Forest

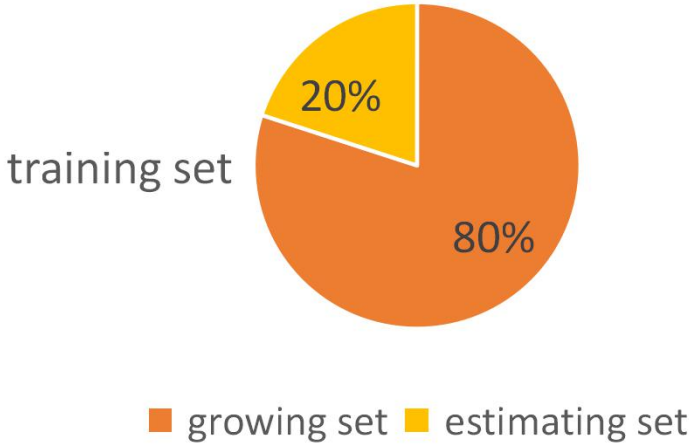
生成类向量作为增强特征



1. 每一个训练样本（此处的训练样本是被sliding window加工后得到的扩充样本）输入到森林中后，进入森林中的每一棵决策树。
2. 在每一棵决策树的结点上得出一个3维向量，分别代表三个类别的概率。然后再求取每棵树上的平均
3. 求取整个森林上的平均，并将所有301个向量（当sliding window大小为100-dim时）进行concatenate之后输出。因此最后每个森林的输出维度为301x3=903维。

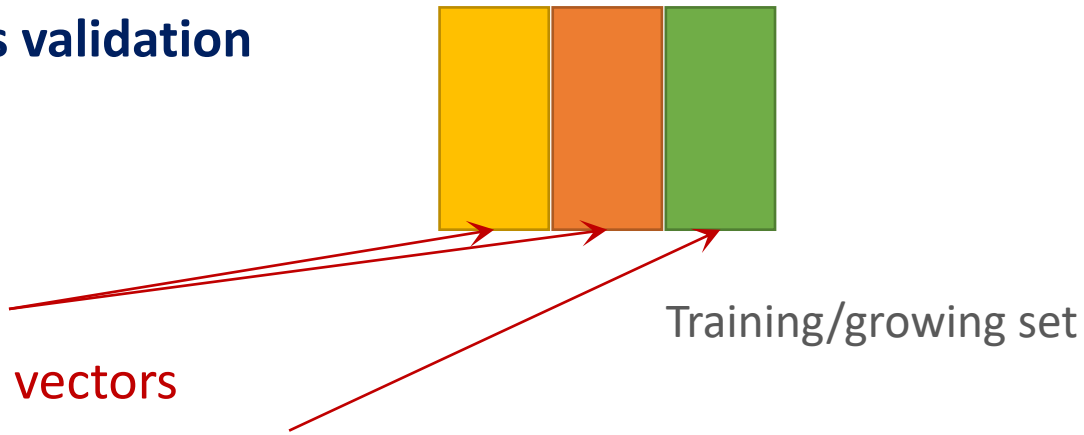
方法 - Cascade Forest

- **如何确定级联级别的数目？**
  - 首先分开训练集
    - →增长集Growing set
    - →预测集Estimating set

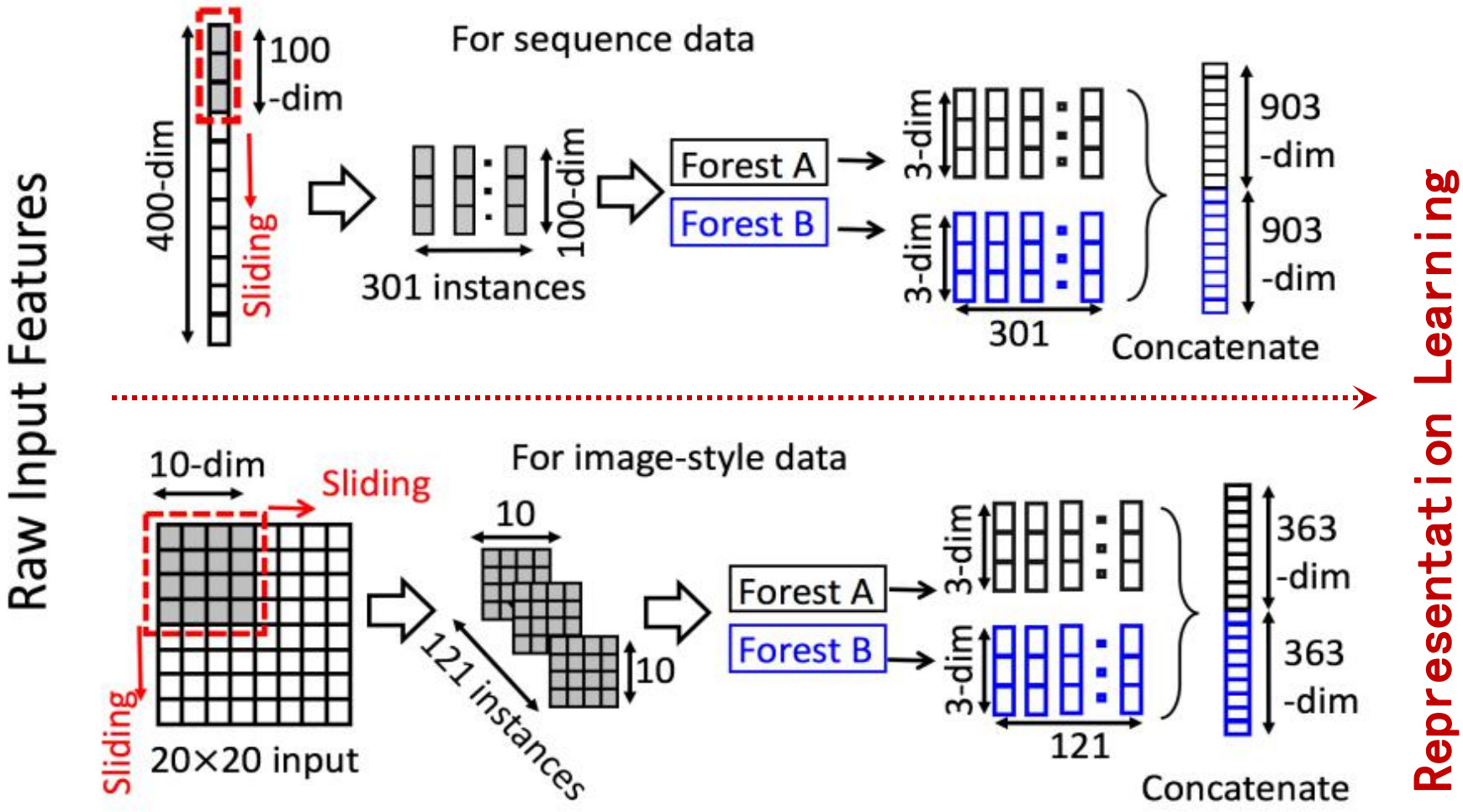


- **如何得到类向量和减轻过拟合?**
  - K-交叉验证 **k-fold cross validation**

    划分训练集/增长集
    - →Subset for training
    - →Subset for generating vectors



方法 - 滑动窗口扫描Multi-Grained Scanning

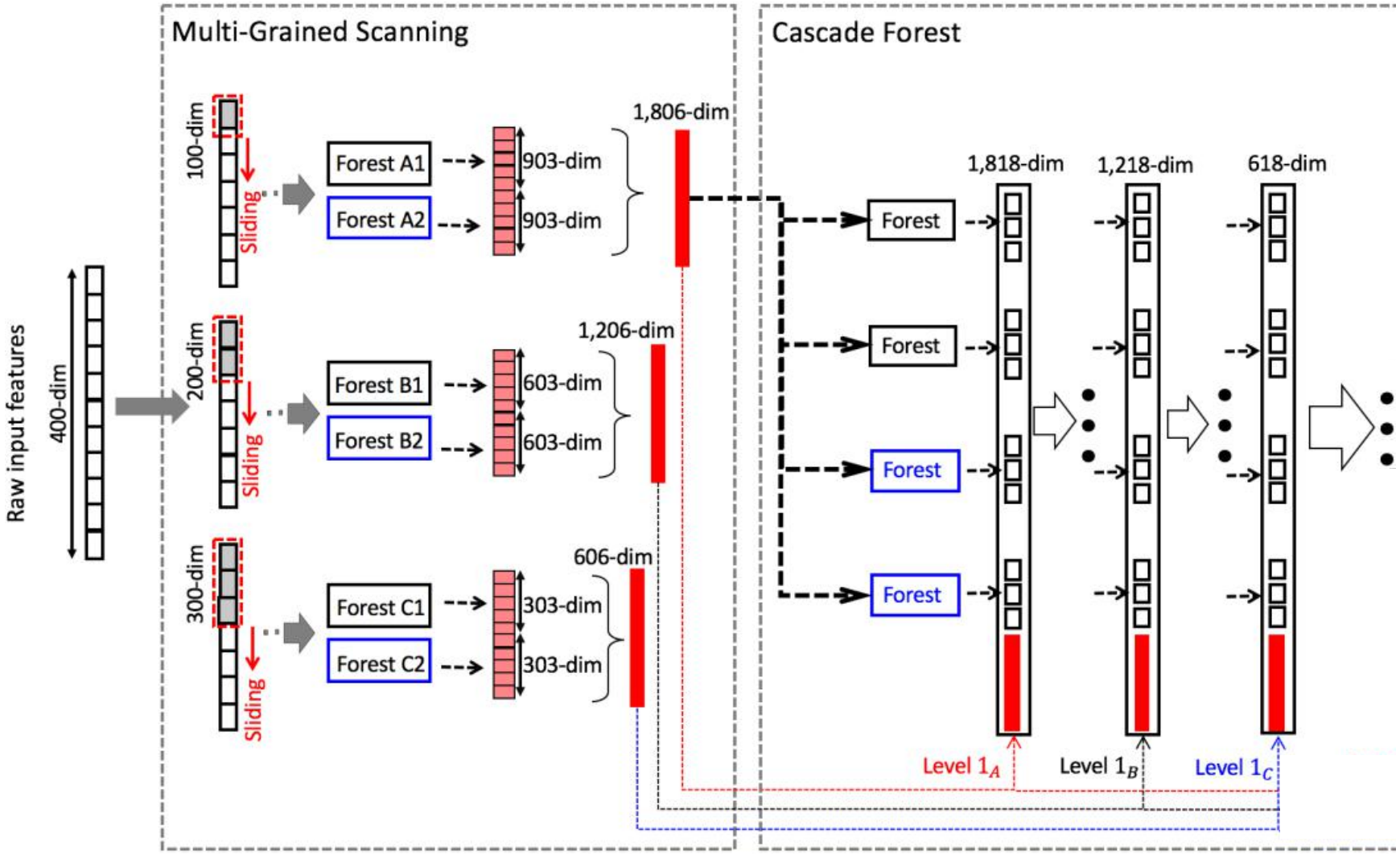


滑动窗口是用来对原始特征进行扫描的。

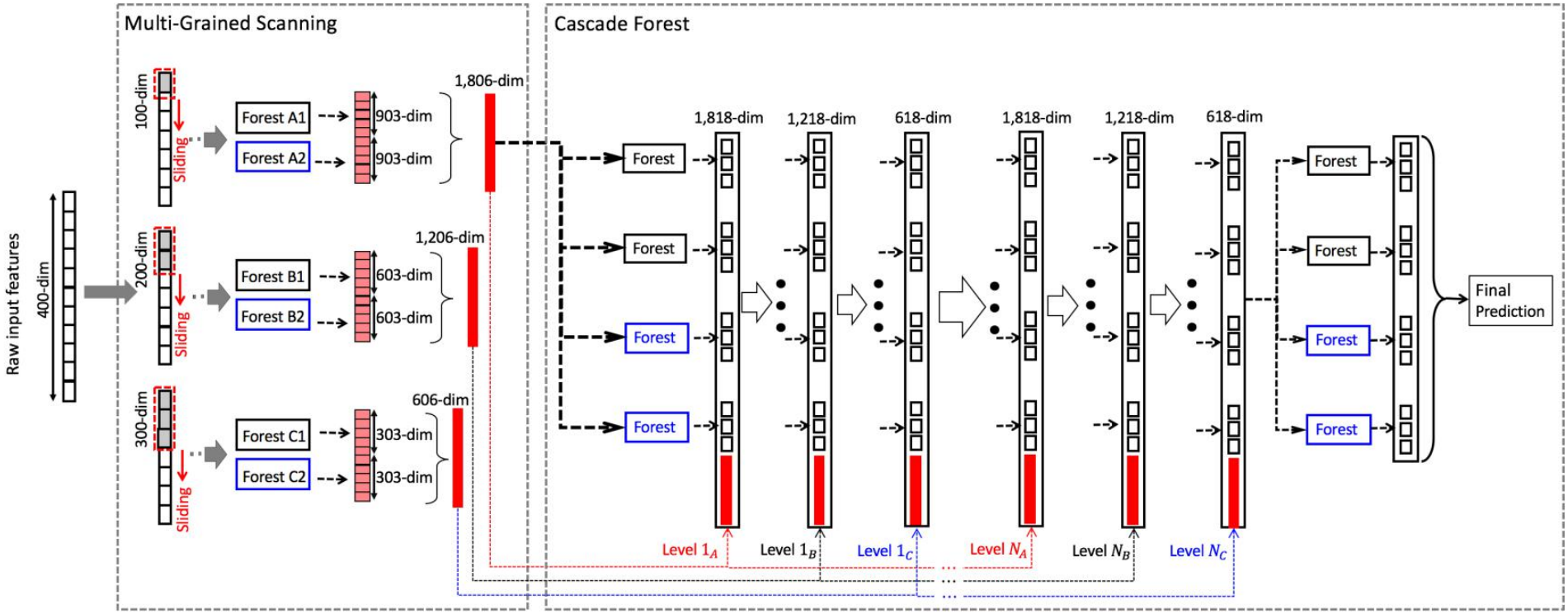
假设原始特征的维数为400-dim，滑窗大小为100-dim。对于序列数据，滑动窗口滑动一次会产生一个100-dim的向量，因此总共会产生301个特征向量。

方法 - Multi-Grained Scanning

多粒度 → 每层有多个级别



方法 - Multi-Grained Scanning



1. 合并成一个3618-dim的原始数据，表示原始的一个数据样本，第一级的输出是12+3618=3630，
2. 直到最后第N级，只有12个输出，然后在每一类别上做avg，然后输出max那一类的label，那就是最终的预测类别。

方法 - Multi-Grained Scanning

- 下图展示了DNN和gcForest的一些超参数，gcForest中给出的是作者在实验中的一些默认值

Deep neural networks (e.g., convolutional neural networks)	gcForest
Type of activation functions: Sigmoid, ReLU, tanh, linear, etc. Architecture configurations: **No. Hidden layers**: ? **No. Nodes in hidden layer**: ? **No. Feature maps**: ? **Kernel size**: ? Optimization configurations: **Learning rate**: ? Dropout: {0.25/0.50} **Momentum**: ? **L1/L2 weight regularization penalty**: ? Weight initialization: Uniform, glorot_normal, glorot_uni, etc. Batch size: {32/64/128}	Type of forests: Completely-random tree forest, random forest, etc. Forest in multi-grained scanning: **No. Forests**: {2} **No. Trees in each forest**: {500} Tree growth: till pure leaf, or reach depth 100 **Sliding window size**: $\{\lfloor d/16 \rfloor, \lfloor d/8 \rfloor, \lfloor d/4 \rfloor\}$ Forest in cascade: **No. Forests**: {8} **No. Trees in each forest**: {500} Tree growth: till pure leaf

- gcForest可以为各种数据使用默认的超参数
- DNN必须为几乎所有类型的数据调整超参数

实验

图像分类 (MNIST)

gcForest	**99.26%**
LeNet-5	99.05%
Deep Belief Net	98.75% [Hinton *et al.*, 2006]
SVM (rbf kernel)	98.60%
Random Forest	96.80%

手部运动识别 (sEMG)

gcForest	**71.30%**
LSTM	45.37%
MLP	38.52%
Random Forest	29.62%
SVM (rbf kernel)	29.62%
Logistic Regression	23.33%

实验

面部识别(ORL)

Small-scale training data

	5 image	7 images	9 images
gcForest	**91.00%**	**96.67%**	**97.50%**
Random Forest	91.00%	93.33%	95.00%
CNN	86.50%	91.67%	95.00%
SVM (rbf kernel)	80.50%	82.50%	85.00%
kNN	76.00%	83.33%	92.50%

低维的数据(UCI)

	16 features	14 features	8 features
	LETTER	ADULT	YEAST
gcForest	**97.40%**	**86.40%**	**63.45%**
Random Forest	96.50%	85.49%	61.66%
MLP	95.70%	85.25%	55.60%

结论

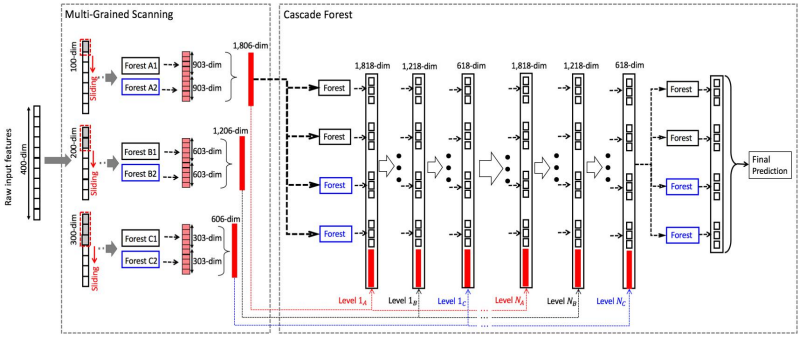
Deep Learning

优点

缺点

Ensemble Learning

gcForest



结论

gcForest

- 更少的参数
  - 更容易构建
  - 更容易调整
- 降低计算复杂度
  - 有效的训练
- 更健壮
  - 在不同任务中使用默认配置良好工作
  - 即使在非常小的训练数据集上也能很好地工作

参考文献

- Zhou Z H, Feng J. Deep Forest: Towards An Alternative to Deep Neural Networks[J]. 2017.

Deep Forest: Towards An

来自arXiv.org

收藏 引用 批量引用

作者 ZH Zhou , J Feng

摘要 In this paper, we propose gcForest, a decis neural networks. In contrast to deep neur...

被引量 28

cn.arXiv.org > cs > arXiv:1702.08835v2

Computer Science > Machine Learning

Deep Forest: Towards An Alternative to Deep Neural Networks

Zhi-Hua Zhou, Ji Feng

(Submitted on 28 Feb 2017 (v1), revised 31 May 2017 (this version, v2), latest version 14 May 2018 (v3))

In this paper, we propose gcForest, a decision tree ensemble approach with performance highly competitive to deep neural networks. In contrast to deep neural networks which require great effort in hyper-parameter tuning, gcForest is much easier to train. Actually, even when gcForest is applied to different data from different domains, excellent performance can be achieved by almost same settings of hyper-parameters. The training process of gcForest is efficient and scalable. In our experiments its training time running on a PC is comparable to that of deep neural networks running with GPU facilities, and the efficiency advantage may be more apparent because gcForest is naturally apt to parallel implementation. Furthermore, in contrast to deep neural networks which require large-scale training data, gcForest can work well even when there are only small-scale training data. Moreover, as a tree-based approach, gcForest should be easier for theoretical analysis than deep neural networks.

Comments: IJCAI 2017
Subjects: **Machine Learning (cs.LG)**; Machine Learning (stat.ML)
Cite as: **arXiv:1702.08835 [cs.LG]**
(or **arXiv:1702.08835v2 [cs.LG]** for this version)

Submission history

From: Zhi-Hua Zhou [view email]
[v1] Tue, 28 Feb 2017 16:10:31 GMT (1414kb,D)
[v2] Wed, 31 May 2017 17:15:34 GMT (1976kb,D)
[v3] Mon, 14 May 2018 18:06:13 GMT (2519kb,D)

武漢大學
WUHAN UNIVERSITY

THANKS