
阿里移动推荐算法大赛答辩
2015天池大数据竞赛
TIANCHI天池

SecRet;WeaPon

孝陵卫南京理工大学数据挖掘探险队

Nothing

中兴图灵杯人工智能一等奖
标签推荐数据挖掘方向
特征工程

Furong

数学建模一等奖
数据挖掘推荐系统方向
框架设计

Implus

南理校ACM队员，银奖第一
中兴图灵杯人工智能一等奖
Deep Learning方向
算法及融合模型

后续

任务描述

框架设计

特征工程

算法及
模型融合

任务描述

移动推荐任务

0～30天脱敏用户商品交互数据

31天

用户购买商品？

0～30天交互数据重点

- 用户4种操作：点击浏览、收藏、加入购物车、购买
- 商品的类别归属等

后续

任务描述

框架设计

特征工程

算法及
模型融合

框架设计

二分类

0～30 天用户商品交互数据

31天 用户购买商品？

考察日

用户商品全集

用户A 在 考察日 购买了商品 x

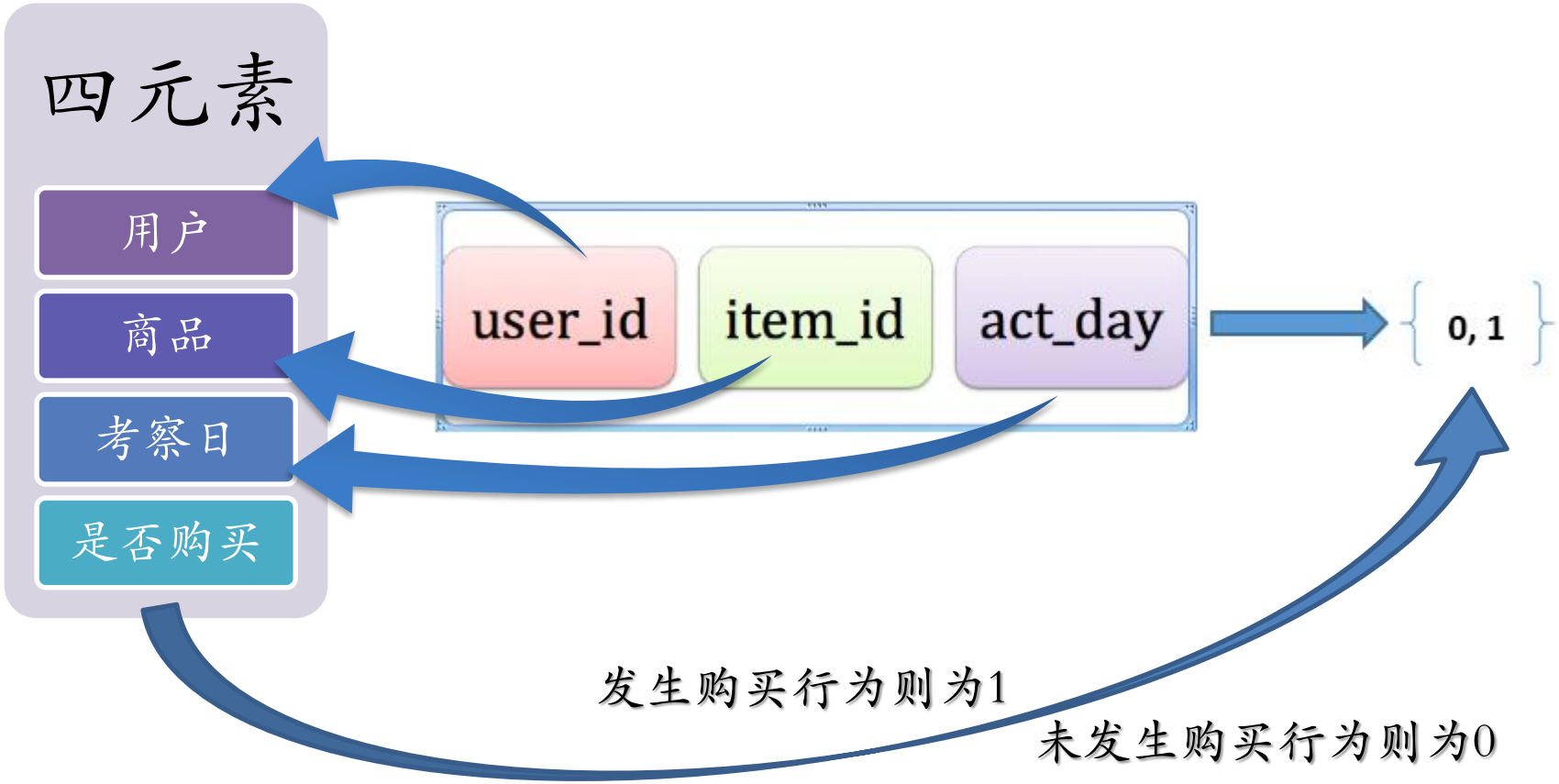
+

用户B 在 考察日 未购买商品 y

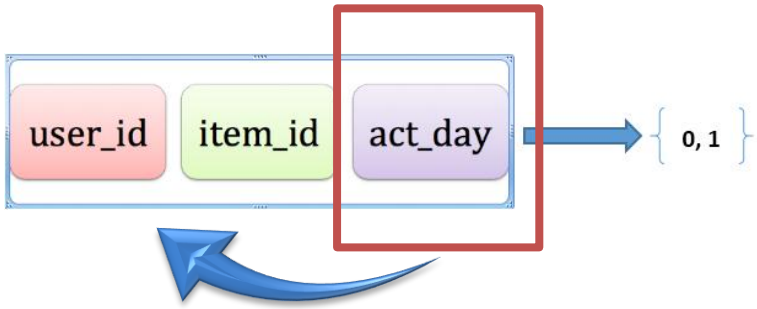
四元素

- 用户
- 商品
- 考察日
- 是否购买

样本结构


四元素
用户
商品
考察日
是否购买
user_id
item_id
act_day
0, 1
发生购买行为则为1
未发生购买行为则为0

样本选择



方案一：所有用户 × 所有商品

方案二：考察日前所有有交互的用户 × 交互商品

方案三：考察日前特定天数特定交互的用户 × 交互商品

框架设计

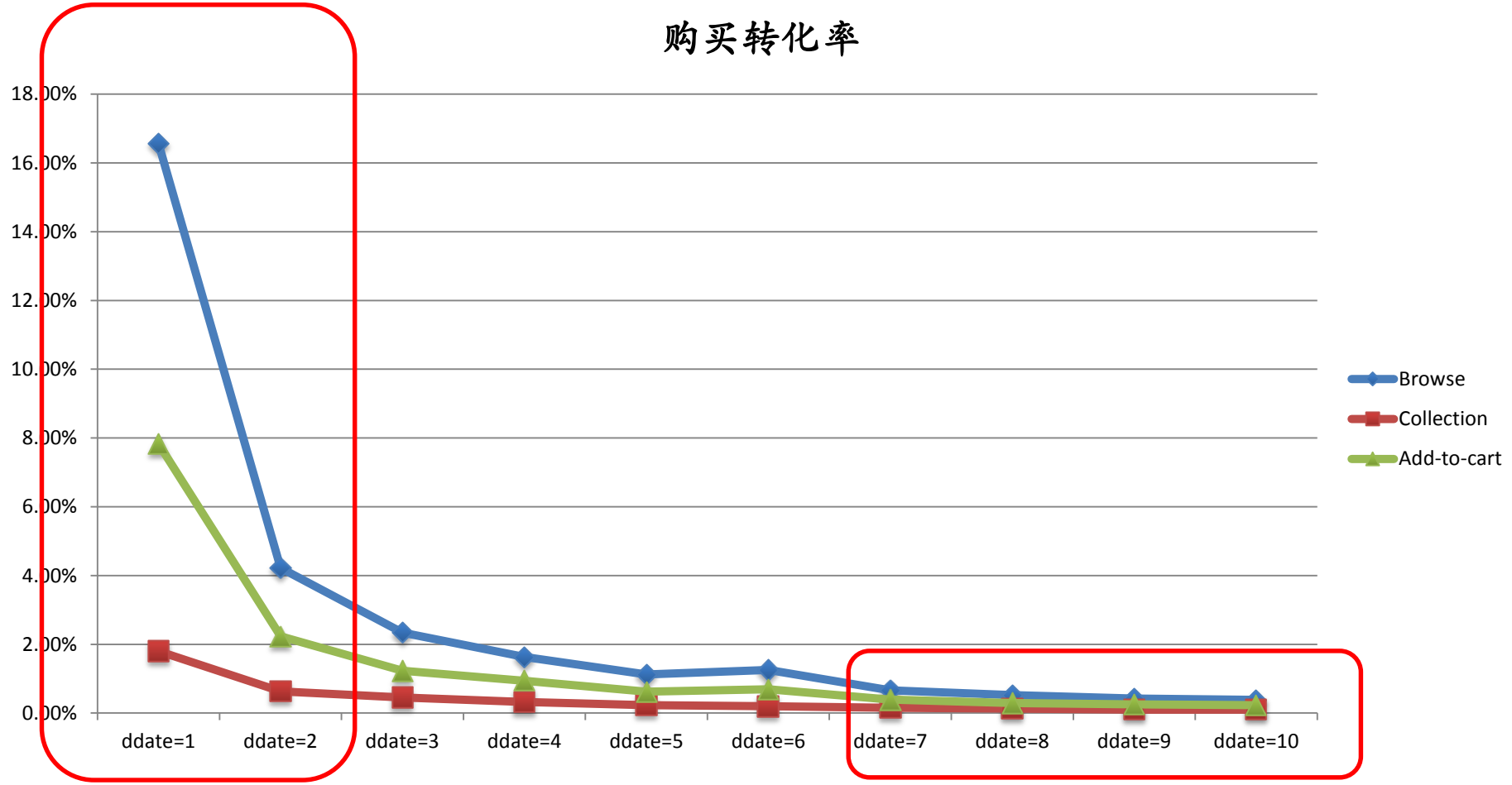
方案一：所有用户 × 所有商品

总样本约：20000000000000000 个

用我们的特征体系计算所有样本的时间是：273972 年

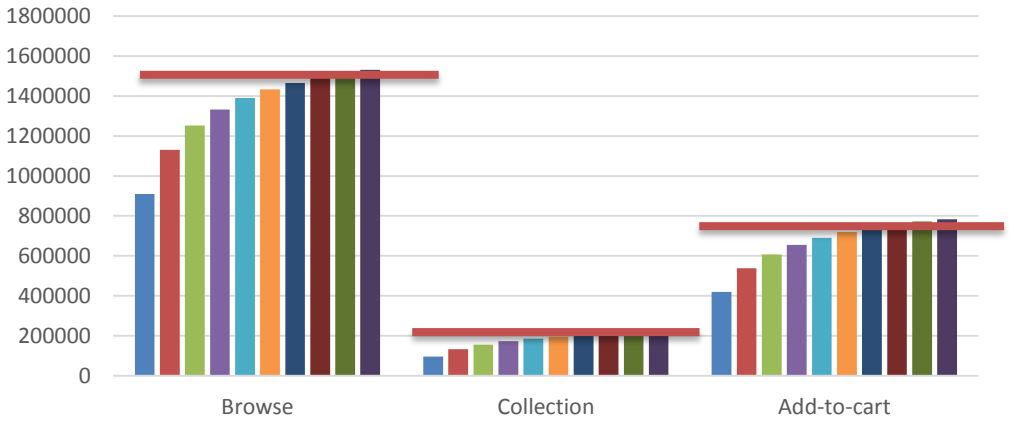
未交互样本量巨大 且缺乏大量有效信息，考虑过滤。

方案二：考察日前所有有交互的用户 × 交互商品



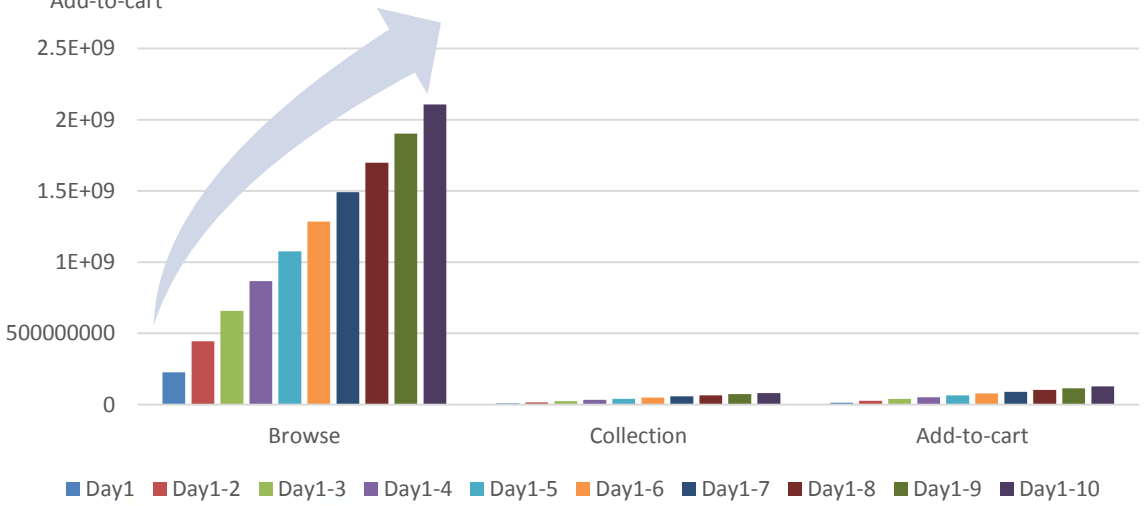
方案二：考察日前所有有交互的用户 × 交互商品

前n天交互对象为 考察日正样本 分布（平缓）



- 7天正样本 ~= 10+天正样本

前n天交互对象为 考察日负样本 分布



- 浏览负样本陡增，每往前推一天，＋2亿负样本

框架设计

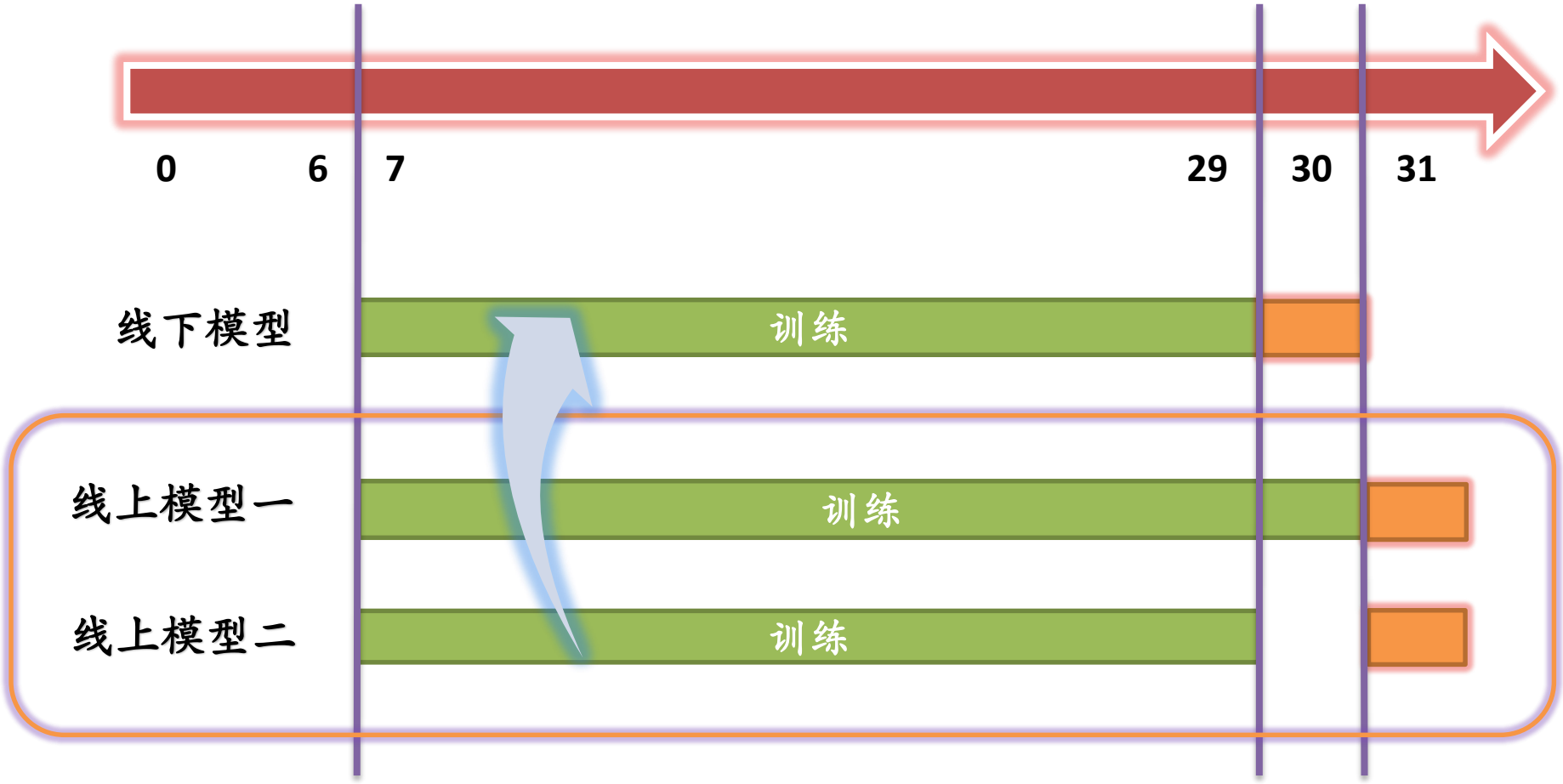
方案三：考察日前特定天数特定交互的用户 × 交互商品

转化率　样本量　模型性能效率

7：非浏览　6：非浏览　5：非浏览　4：非浏览　3：非浏览　2：非浏览　1：所有交互　考察日 = D

框架设计

样本分配-训练测试



后续

任务描述

框架设计

特征工程

算法及
模型融合

特征工程

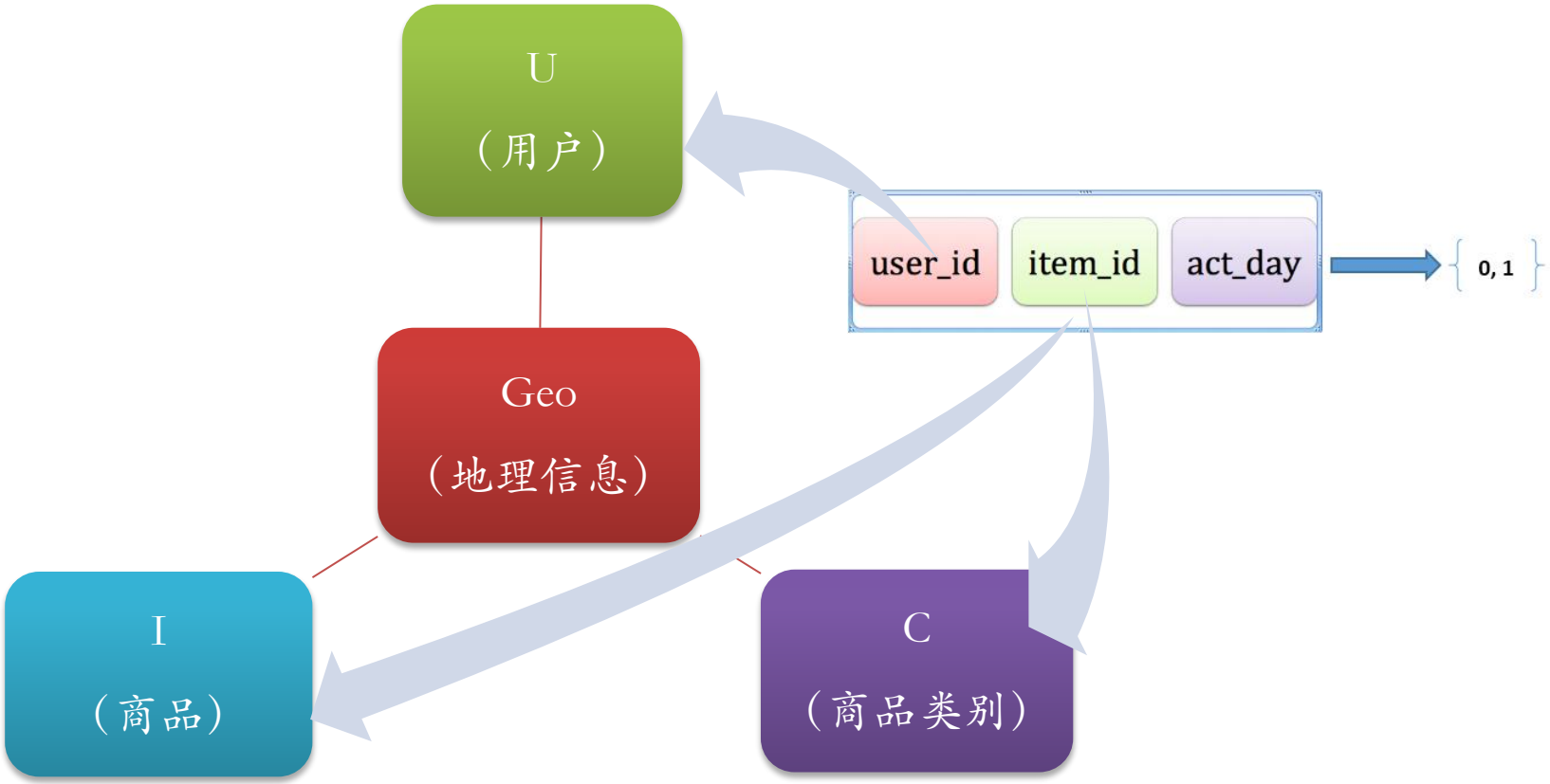
特征创新

引入丰富的特征群（即按照特征属性分为10类）

精心设计了大量二次组合统计特征

特征工程

特征群-基础群



特征工程

特征群-基础群

U特征群

- 计数特征
- 加和特征
- 加权特征
- 转化率特征
- 活跃度特征

I特征群

- 计数特征
- 加和特征
- 商品热度特征
- 交互时间特征
- 交互人数特征
- 星期分布特征

C特征群

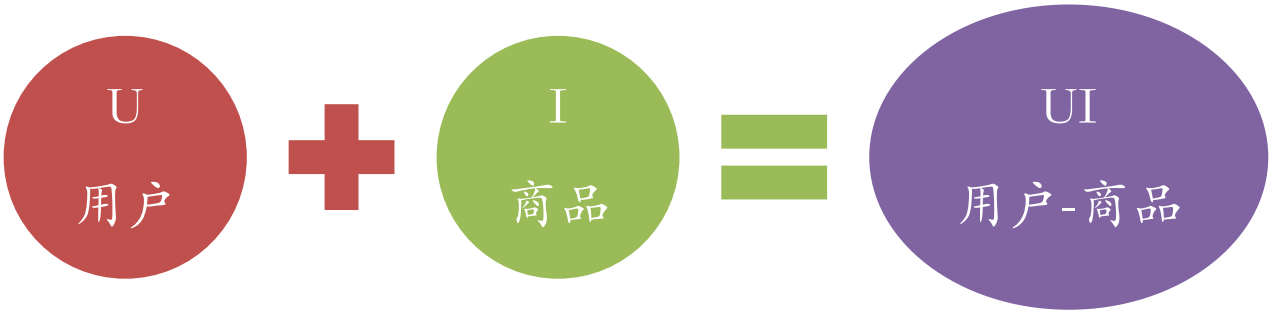
- 计数特征
- 加和特征
- 类别热度特征
- 回头客特征

Geo群

- 用户商品最近距离特征
- 商品是否具有地理信息特征

特征工程

特征群-衍生群



	U特征	I特征	C特征
U特征	-	UI特征群	UC特征群
I特征	UI特征群	-	IC特征群
C特征	UC特征群	IC特征群	-

特征工程

特征群-衍生群

UI特征群

- 计数特征
- 加和特征
- 权值特征
- 交互时间特征
- **习惯偏差特征**

UC特征群

- 计数特征
- 加和特征
- 权值特征
- 交互时间特征
- **习惯偏差特征**
- 星期分布特征

IC特征群

- 比例特征
- **排序特征**

特征工程

同理-衍生群

UI&UC特征群

- 竞争特征
- 排名特征

U&UI特征群

- 基本比率特征
- 二次购买特征
- 交互时间比特征
- 交互排名特征

U&UC特征群

- 基本比率特征
- 二次购买特征
- 竞争特征
- 交互时间比特征
- 交互排名特征

特征工程

二次统计特征

特征群	特征名	特征含义	优势及作用
UI&UC	uiuc_row_ln_weight_day_1_7	该用户在考察日前7天对该商品4种操作加权值在用户对该类下所有商品加权值中的排序	防止预测一个用户购买同类商品下的大量不同物品
U&UI	uiu_row_ln_weight_day_1_7	该用户在考察日前7天对该商品4种操作加权值在用户对所有商品加权值中的排序	可以预测出用户最想购买的商品
U&UC	ucu_row_ln_weight_day_1_7	该用户在考察日前7天对该类别商品4种操作加权值在用户对所有类别加权值中的排序	可以预测出用户最想购买的类别

* 4种操作加权值是指对4种操作数目加权统一成一个数值

特征工程

特征细节

- 总维度有2064，核心维度有780+
- 统计特征窗口为 1/2/3/4/5/6/7/10/15/21/30
- ln(1 + x) 平滑
  - 化比率除法为减法，同时避免除0错误
  - 数据平滑标准化，减弱异常数据的影响
- 统计特征窗口＞7，用平均值解决数据截断问题
  - 比如考察日为12，那么对于前20天的浏览量的统计特征就会出现**数据截断丢失问题**
  - 使用前20天的**平均浏览量**能够有效避免因为数据截断而导致的数据分布不一致

后续

任务描述

框架设计

特征工程

算法及
模型融合

算法创新

从第一赛季开始引入深度学习（卷积神经网络CNN），

自主设计了适应推荐系统的二维特征模式，

并在尝试中发现了由drop-out正则化带来自融合方式；

深度学习

IMAGENET Large Scale Visual Recognition Challenge 2012 rank

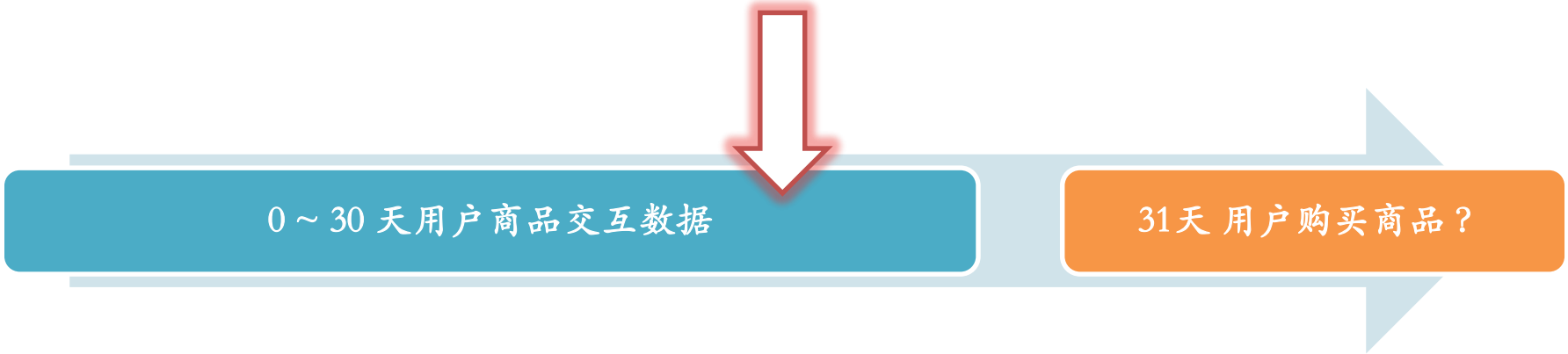
Method	Top-5 error
Deep CNN [1]	**15.315%**
SIFT+FV	26.172%
High-Level SVM	26.979%

图像处理　语音识别　自然语言处理

[1] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

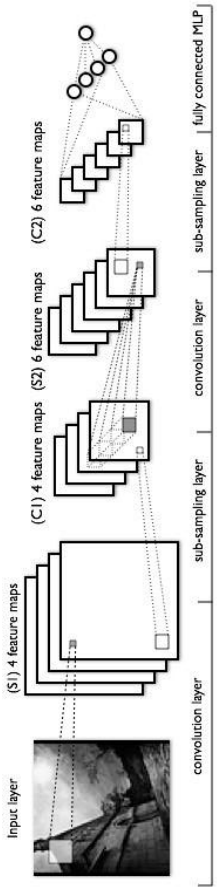
推荐系统？

深度卷积神经网络



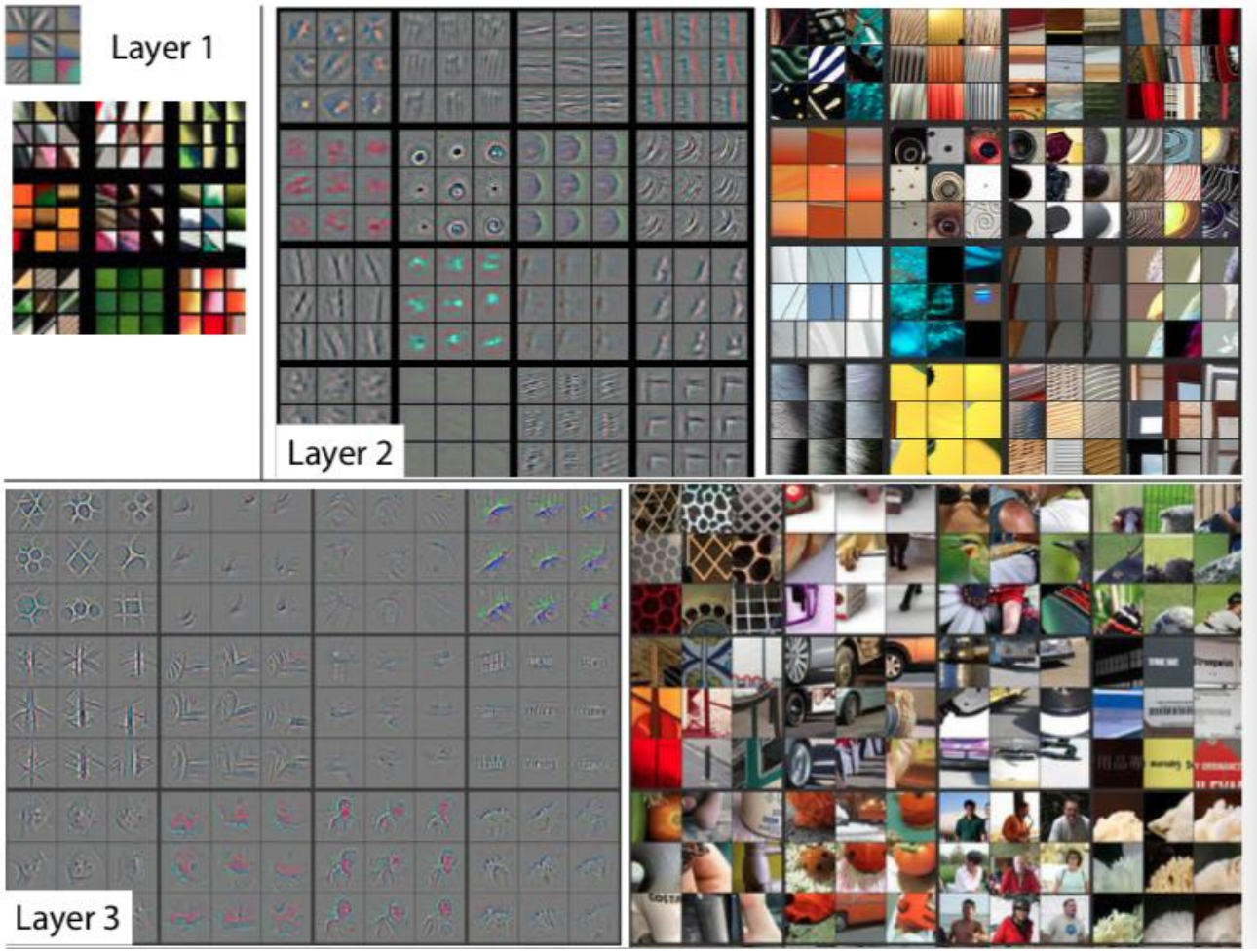
1. 综合各大领域，卷积神经网络带来的革命性最大
2. **卷积核**具有很强的从**局部到全局**抽取鉴别特征的能力

借鉴


fully connected MLP
sub-sampling layer
convolution layer
sub-sampling layer
convolution layer
(C2) 6 feature maps
(S2) 6 feature maps
(C1) 4 feature maps
(S1) 4 feature maps
Input layer

局部特征

高阶特征


Layer 1
Layer 2
Layer 3

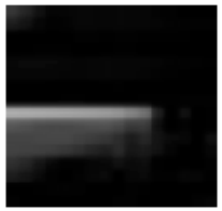
二维特征模式

时间轴

原始统计特征量

	……	前3天	前2天	前1天
UI特征群				
UC特征群				
U特征群				
I特征群				

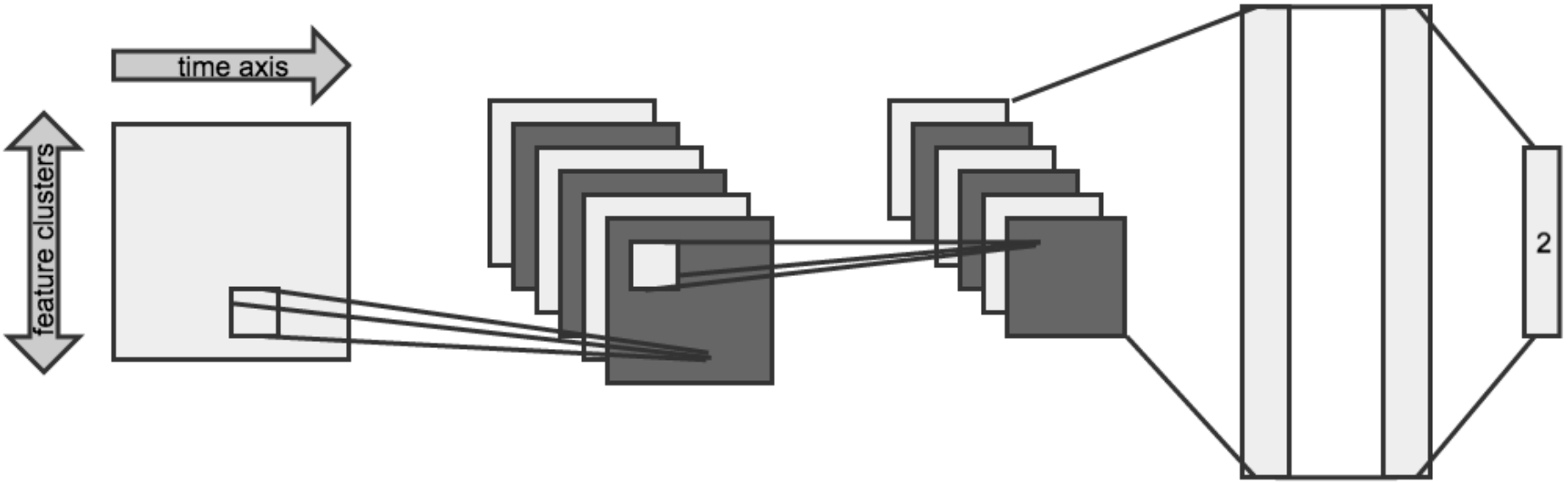
features



label 0

label 1

卷积网络结构



* 不采用任何 pooling
* 使用 drop-out regularization，有自融合功效

卷积网络

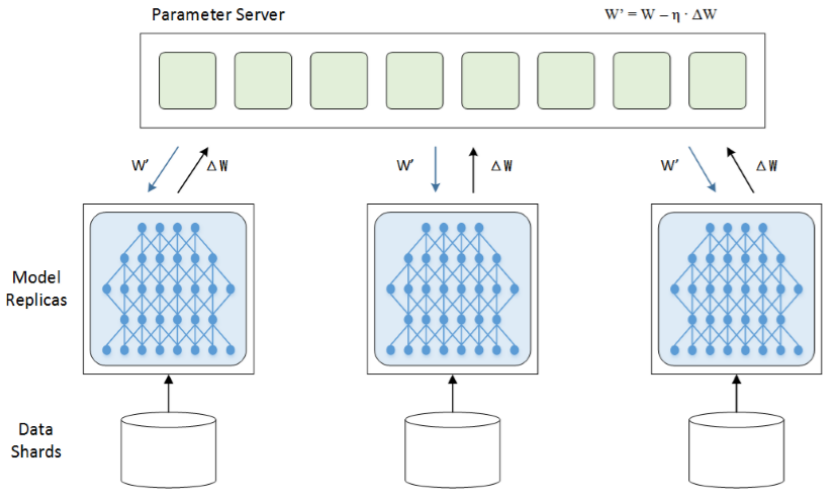
深度卷积神经网络（Deep CNN）在Season 1：

Method	*F1-Score Offline*	*F1-Score Online*
Gradient Boosted Decision Trees (GBDT)	6.85%	-
Random Forest (RF)	6.98%	-
Logistic Regression (LR)	7.36%	10.56%
Deep CNN (self-ensemble)	7.40%	10.84%
LR + RF	7.68%	11.67%
Deep CNN (self-ensemble) + LR	8.02%	12.20%
Deep CNN (self-ensemble)+ LR + RF	**8.12%**	**12.69%**

遗憾

- 深度卷积神经网络未能在Season 2用于最后结果
  - 暂时无GPU 支持, 无多线程支持
    - 训练相当耗时
  - Graph结构无法很好地支持参数服务器（分布式迭代的异步模式）
    - 同步（聚合）模式下

收敛非常缓慢，
没有很好的weight update
合并方案（无相关文献）



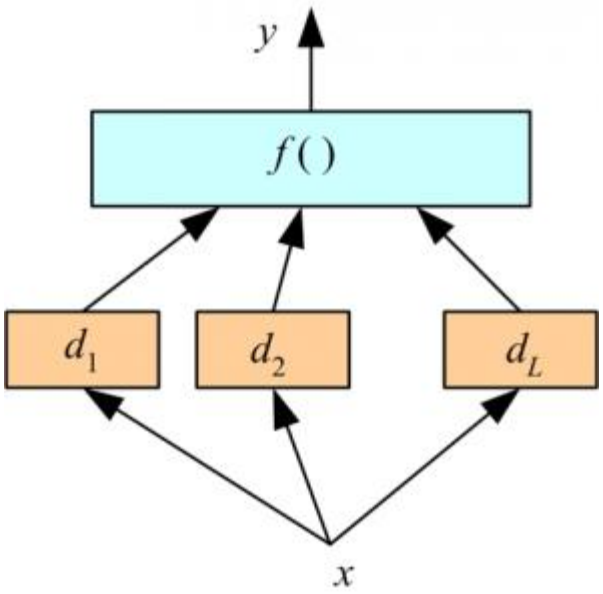
模型融合创新

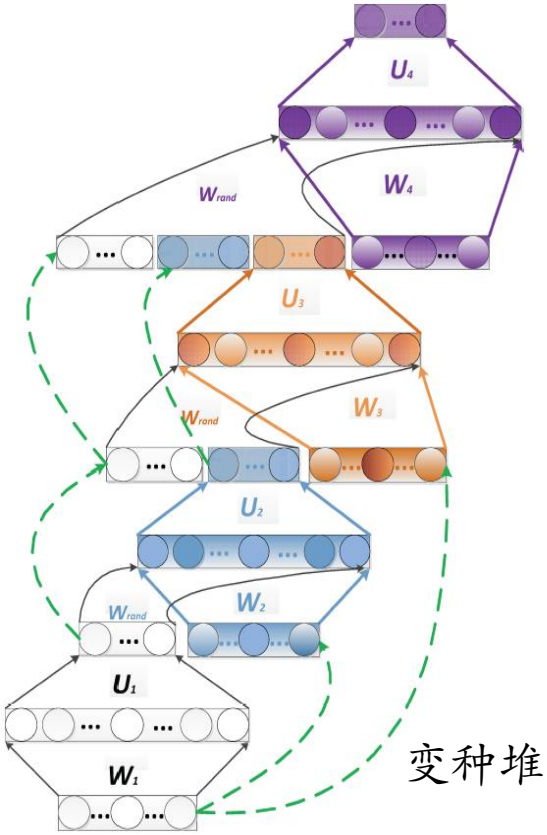
模型融合中借鉴并扩展了**堆融合**的技巧，

根据我们的解决方案体系设计了基于不同特征群的**多视角堆模型**融合结构；

堆模型

基础堆模型 [1] (Wolpert,1992)



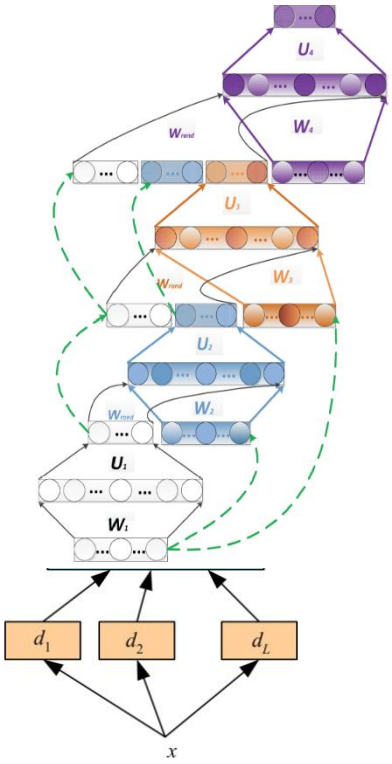


变种堆结构[2] (Deng Li)

[1] Wolpert, David H. "Stacked generalization." *Neural networks* 5.2 (1992): 241-259.

[2] Deng, Li, Dong Yu, and John Platt. "Scalable stacking and learning for building deep architectures." *Acoustics, Speech and Signal Processing (ICASSP), 2012*

多视角堆模型



多特征群

U feature cluster → GBDT → U stacking feature

I feature cluster → GBDT → I stacking feature

C feature cluster → GBDT → C stacking feature

UI&UC feature cluster → GBDT → UI&UC stacking feature

original feature clusters all

U stacking feature / I stacking feature / C stacking feature / UI&UC stacking feature / original feature clusters all → GBDT

多视角堆模型

多视角堆模型（Multi-View Stacking Ensemble）在Season 2：

Method	*F1-Score Offline*	*F1-Score Online*	*Our Team Rank*
GBDT (single model)	7.90%	8.65%	2
GBDTs (average ensemble models)	7.94%	8.71%	2
MVSE GBDT (single model)	7.95%	8.71%	2
MVSE GBDT + GBDTs (average ensemble models)	**8.01%**	**8.78%**	**1**

堆优势-提升幅度

融合方式	融合类型	线上F1值	提升幅度(%)
GBDT（单模型）	(Boosting)	8.65%	
3 GBDTs（平均融合）	Bagging	8.68%	0.03
4 GBDTs（平均融合）	Bagging	8.70%	0.02
5 GBDTs（平均融合）	Bagging	8.71%	0.01
GBDTs + MVSE（平均融合）	Stacking + Bagging	8.78%	0.07

* 我们尝试过加权融合，实际效果并没有平均融合好

后续

任务描述

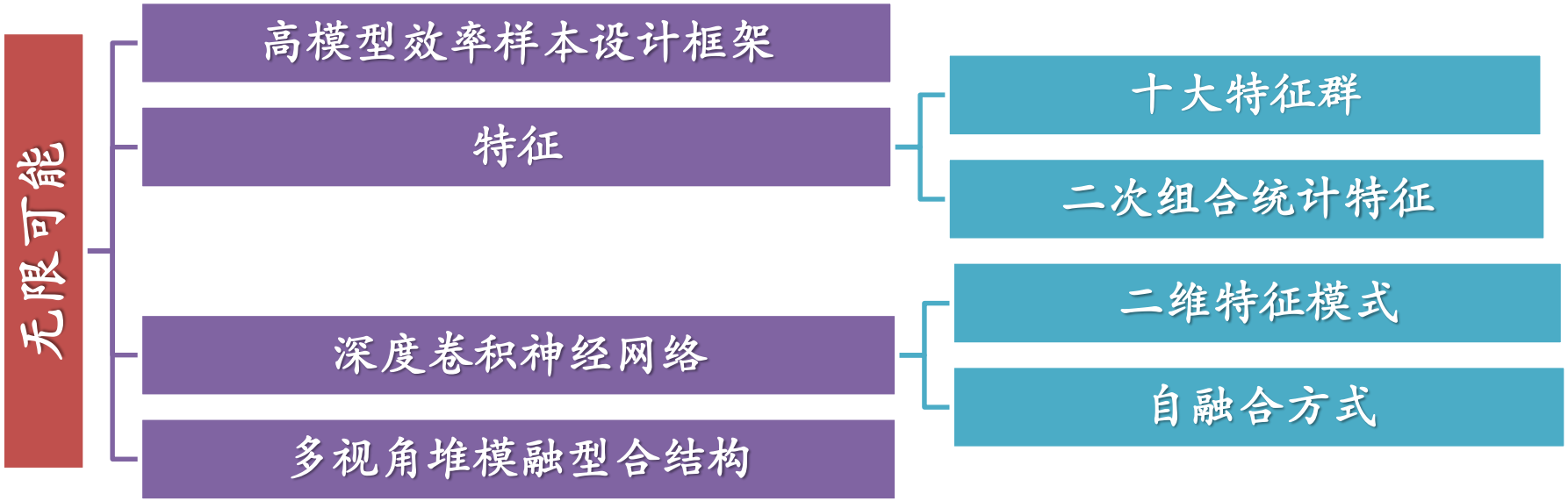
框架设计

特征工程

算法及
模型融合

后续

创新总结


无限可能
高模型效率样本设计框架
特征
十大特征群
二次组合统计特征
深度卷积神经网络
二维特征模式
自融合方式
多视角堆模融型合结构

后续

寄语

- 感谢 @岱月 @凝岚 @励辰 @泽熠 @煜霈 @无影 @贤木 @一婷 @天渡 @ 崇慧 @默默付出的阿里人 带我们愉快的“玩耍”
- 感谢阿里提供了一个这么好的平台能让我们这些普普通通的孩子找到 **研究和超越自我** 的乐趣
- 感谢所有在比赛过程中帮助我们的其他战队队员，师兄们，老师们，家人们；感谢今天在场的所有老师同学的悉心聆听
- 希望GPU，**多线程**，caffe，lasagne什么的快点支持起来吧

后续

生命不息，奋斗不止