ben.guo



滴滴科技 & ACM SIGSPATIAL • \$25,000 • 1171 支队伍 • 1295 名参寒者

SIGSPATIAL 2021 GISCUP



组队截止时间 2021-08-02

预估到达时间 (Estimated Time of Arrival) 在很多应用场景中扮演者至关重要的角色。

例如,在网约车订单分配、价格预估、行程中间预估、路线决策等各个场景。精准的到达时间预估可以帮助平台提升效率,提高用户体验,方便平台和用户做更优的决策。

然而,到达时间本身会收到路线长度、路段通行能力、道路等级、实时路况、红绿灯个数相位等因素的影响。

数据概览-原始数据

	0	1	2
0	5966280 531 2923.0000 495 75457 153	157298:4.5000,1.0000,1,0 511622:2.8552,1.0000,	47128_204460:23 624101_153565:30 30464_397883:
1	6099942 590 3212.2780 559 19210 157	177913:4.3311,0.1702,1,0 401603:25.9826,1.0000	243785_383862:30 224592_402375:23 402375_23283
2	2792816 559 7868.4474 626 38822 36	4089:8.5500,1.0000,0,0 621731:2.8174,1.0000,0,	412945_587168:30
3	5780075 677 4976.3791 446 51627 76	461148:7.9258,0.6199,1,0 270332:2.0571,1.0000,	434768_339809:32
4	7808558 968 4985.9186 872 39965 52	402771:0.5739,0.0816,0,0 403356:8.6000,1.0000,	279601_378428:30 607127_383862:9 224592_402375

数据概览-预处理后的数据

要预测的目标

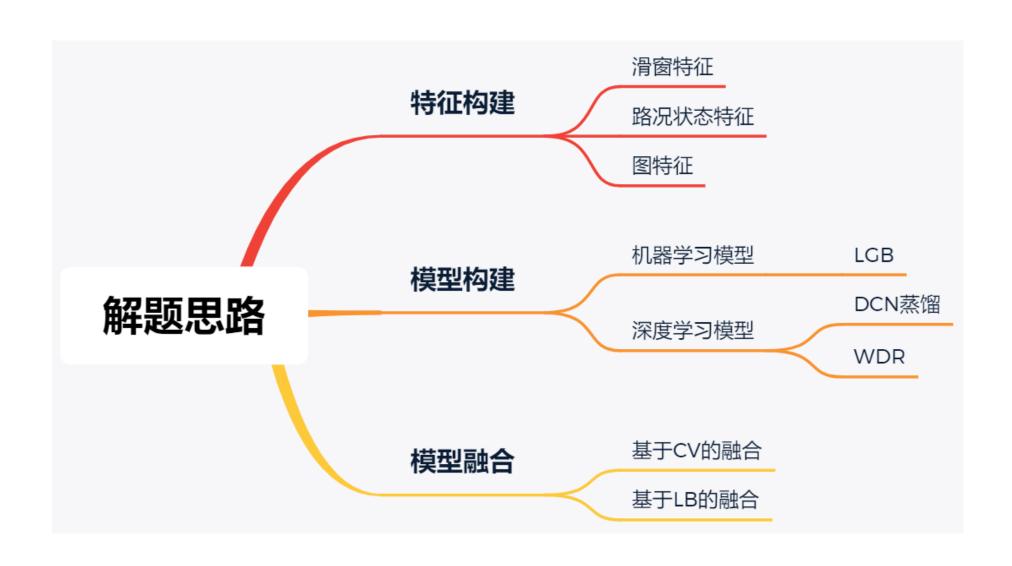
	order_id	ata	distance	simple_	eta driver_	_id slice_id	index	date_time				
0	5966280	531	2923.0000) 4	195 754	57 153	0	20200801		llood	公里 四月3世4 士 次丁	
1	6099942	590	3212.2780) 5	559 192	10 157	1	20200801				
2	2792816	559	7868.4474	1 6	388	22 36	2	20200801		Head—	—端到端特征	
3	5780075	677	4976.3791	1 4	146 516	27 76	3	20200801				
4	7808558	968	4985.9186	8	399	65 52	4	20200801				
ord	ler_id s	slice_id	link_id	link_time	link_ratio	link_current	t_status	link_arriva	al_status	date_time		
596	66280	153	157298	4.5000	1.0		1		0	20200801		
596	66280	153	511622	2.8552	1.0		1		1	20200801	Link	□ ∓∪∔±∠⊤
596	66280	153	524542	2.2345	1.0		1		0	20200801	├ Link—	—序列特征
596	66280	153	97293	3.6000	1.0		1		1	20200801		
596	66280	153	273845	3.2276	1.0		1		1	20200801		
ord	er_id	cr	ross id c	ross time	date_time							
596	6280	47128_2	204460	23.0	20200801		— Cross——序列特·					
596	6280 6	24101_1	153565	30.0	20200801				山州土公丁	k±⟨┬		
596	6280	30464_3	397883	26.0	20200801				川行仙	行 征		
596	6280 1	60457_5	527119	18.0	20200801							
596	6280 5	97673_	507113	30.0	20200801							

评价体系

提交的预测结果会和行程的真实的时间比较,并根据MAPE (Mean absolute percentage error) 评分:

$$MAPE = \frac{1}{n} \sum_{i=0}^{n} \frac{|eta-ata|}{ata}$$

思路分析: 序列+图+回归



特征构建:统计特征

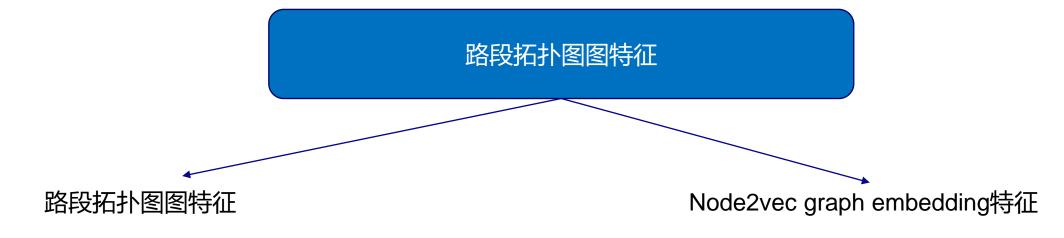


- 订单路段数均值、总和等统 计特征
- 订单通行时间均值、总和、 变化等统计特征
- 订单路段状态统计特征
- 订单current status 词袋模型, 4维向量特征
- Cross特征同上

- 订单路段倒序排列,统计滑窗长度 3,5,....,100...中:
- 路段状态趋势特征
- 路段通行时间趋势特征,趋势变化特征
- 路段通行时间间隔趋势特征
- Cross特征同上

- 历史两周路段、时间片统 计路段状态特征, 路段占 比百分比
- 历史两周路段、星期、时 间片统计路段状态特征, 路段占比百分比

特征构建: 图特征



(a) 用户行为序列

(b) 构建物品关系图

(c) 随机游走

- 度中心性,紧密度中心性等
- 通过度中心性筛选城市拓扑图中心节点
- 每个订单中心节点统计特征单独统计

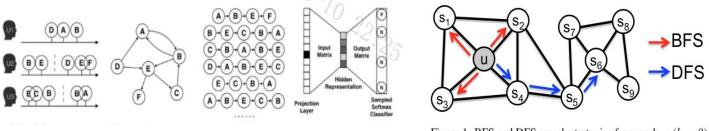


Figure 1: BFS and DFS search strategies from node $u\ (k=3)$.

Link id -> 128维embedding特征(选择表达结构性和同质性2种游走方式,共256维embedding特征) cross id同上

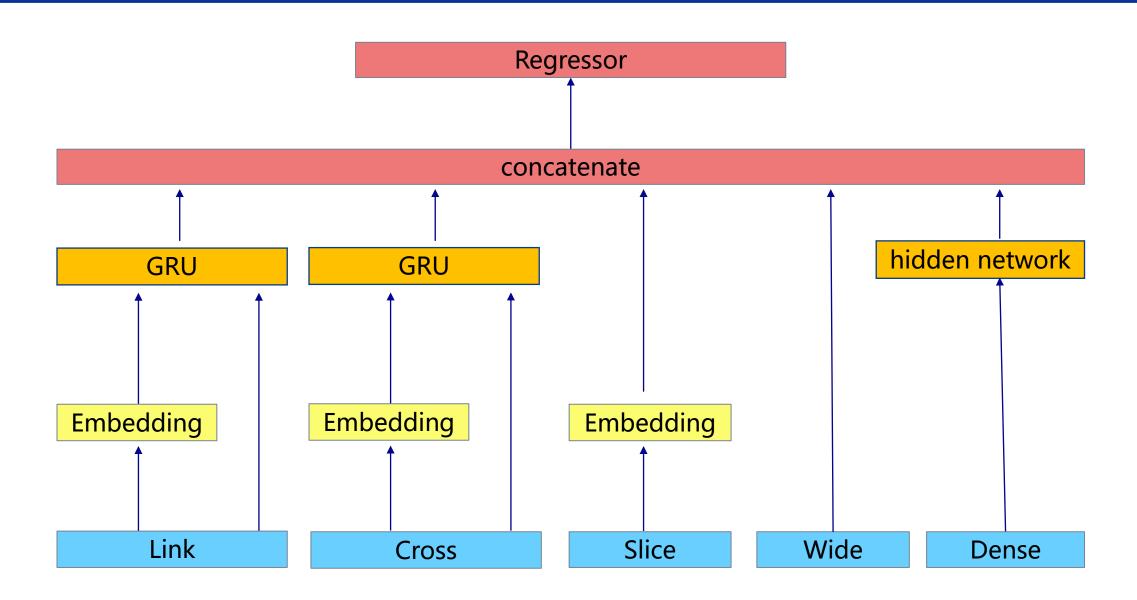
产生Embedding

模型构建:蒸馏模型

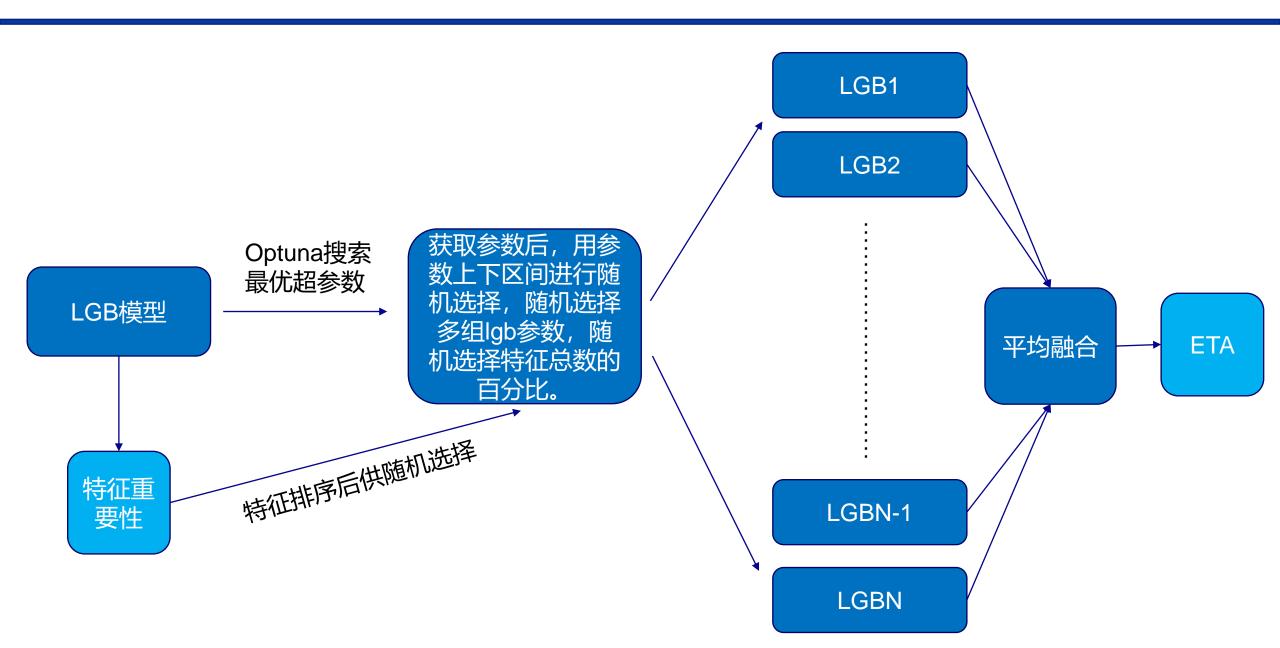
https://www.kaggle.com/mathormad/knowledgedistillation-with-nn-rankgauss **RESULT** MAPE Teacher Model Student Model MAPE+T MAPE MAPE Batch Norm Laver **Batch Norm Layer** Batch Norm Layer **Batch Norm Layer** Mean & Max Mean & Max Mean & Max Mean & Max Encode Encode **Pooling Pooling Pooling Pooling** Cross product Cross product +MLP +MLP **BI-GRU BI-GRU BI-GRU BI-GRU** Embedding & Embedding & Embedding Embedding **Embedding** Embedding Stacking Stacking Sparse Arrival Dense Link Cross Sparse Dense Link Cross

Arrival包含当前order_id中,4种路况状态的百分比; Beta值为0.8,即mape(t mape,s mape)占总体loss的权重;

模型构建: WDR模型



模型构建: LGB模型



模型结果

将各种模型融合后处理,各模型的表现如下:

模型	public LB	private LB
LGB模型	0.155x	
图特征+LGB模型	0.135x	
WDR模型	0.133x	
DCN模型	0.133x	
DCN-蒸馏模型	0.131x	
增加更多特征蒸馏	0.129x	
到达路况状态特征	0.128x	
三折交叉深度模型	0.126x	0.126x
三模型融合	0.125x	0.125x

模型融合后的结果即可得到Top1%的成绩。