Case V:
Project Crystal Voice for Huawei, Model Reuse and something else

主讲教师：詹德川

目录

项目背景

项目概述：
无线业务场景下某种子特性打开之后会给通信质量带来多少增益（EVQI_Gain）

项目诉求：

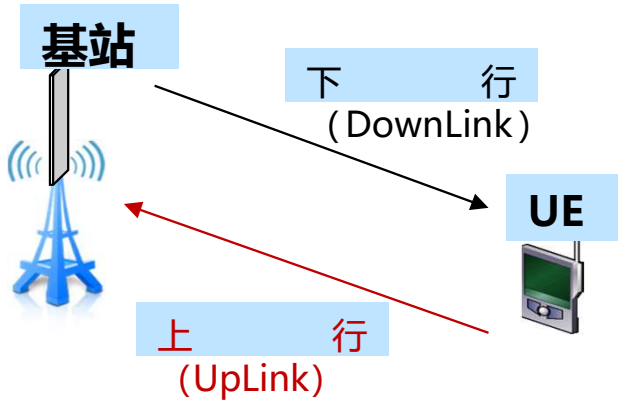
a) 业务方面：根据无线业务部门提供的特征数据，准确识别潜在高增益局点，并预测具体增益值

b) 技术方面：数据标签不准确，新局点无标记数据，数据分布变化等等

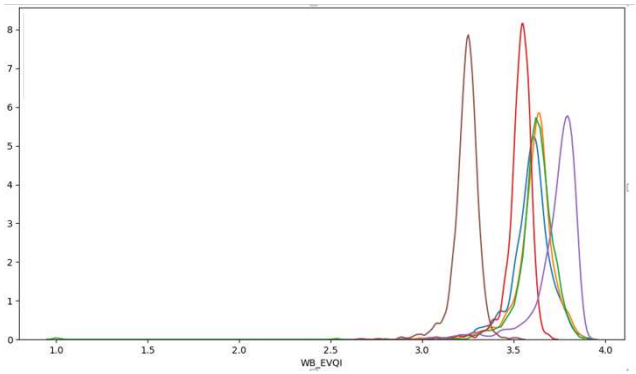
项目目标：

a) 项目指标：提升无线部门具体的业务性能

b) 通用解决方案：以具体的业务Case为基础，获得一整套通用性模型复用算法



华为无线业务基站通信质量、上行下行流量示意图



水晶语音项目样本数据分布差异性示例图（目标值分布p(y)）

项目背景

项目具体描述：

a) 水晶语音项目是以3G信号为应用场景收集相关数据

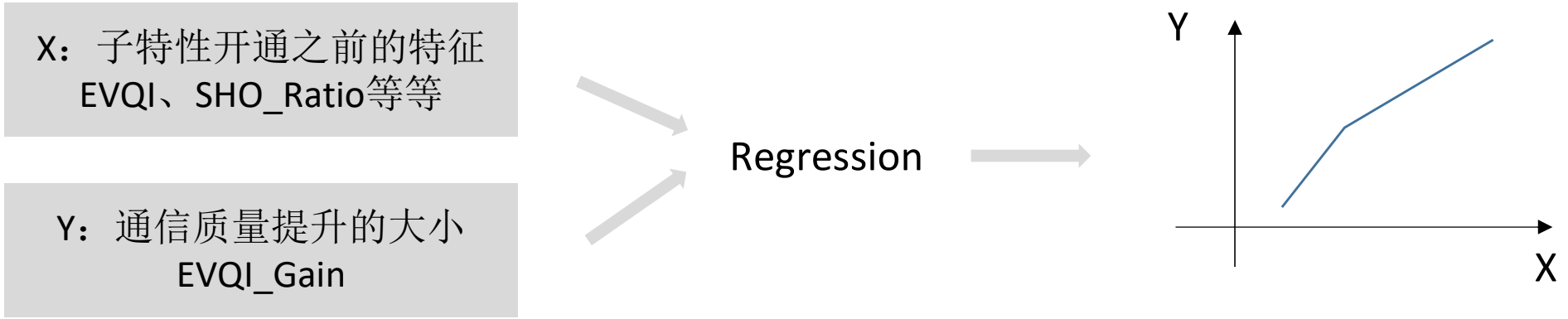
b) 子特性主要是以窄带通信的深度覆盖和无缝两项服务为主

c) 项目数据包括8个局点，每个局点约有1000个小区

d) 项目目标就是预测新局点下的小区开通子特性之后的通信质量提升

项目任务抽象：

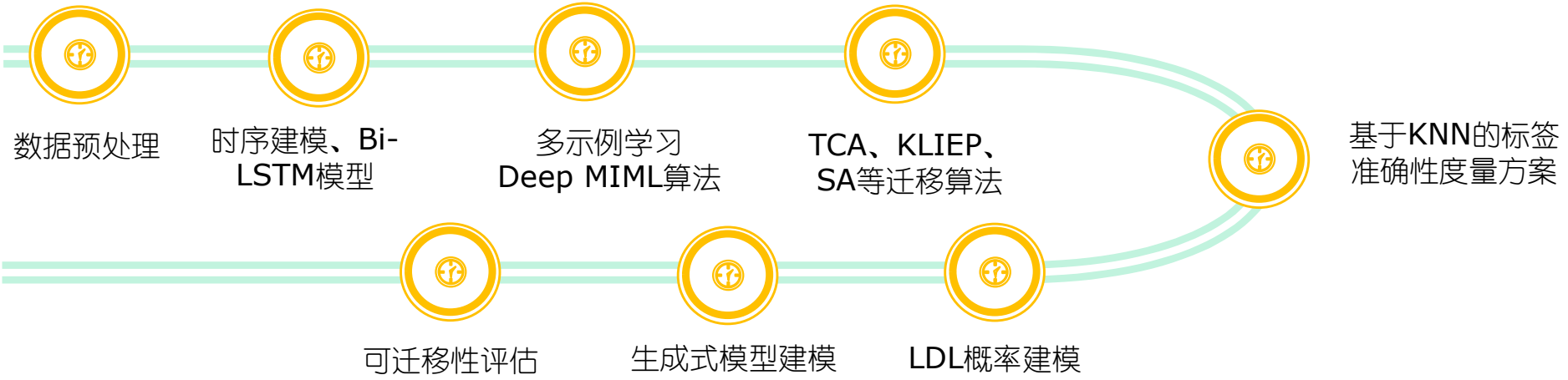
将预测增益的问题建模为机器学习里面的回归任务，预测的增益值超过某个阈值就判定为潜在高增益局点



目录

研究内容

项目技术路线：

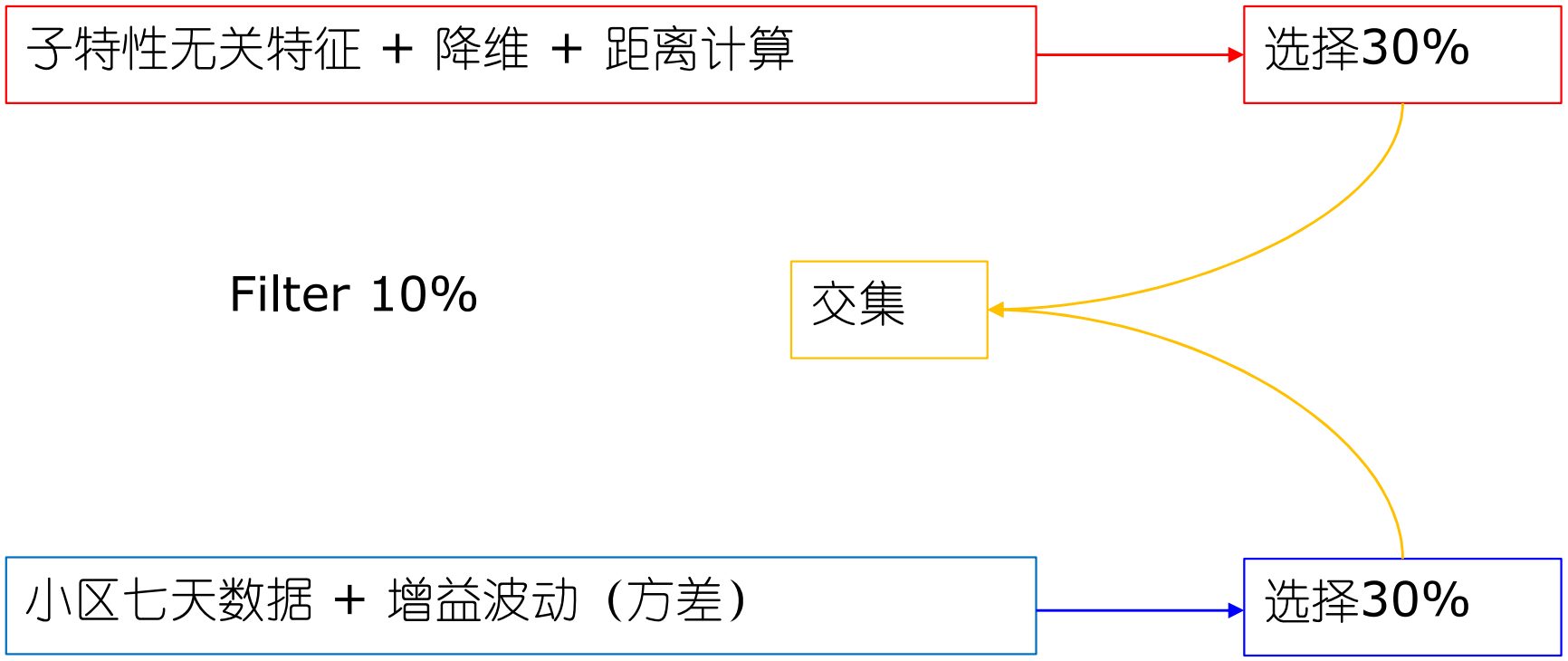


项目迁移学习SDK：

研究内容

数据预处理：

Noisy Label去除



研究内容：特征选择、降维是什么

研究内容

浅层模型建模：

算法：
Ridge

评估指标：
潜在高增益局点识别：F1
增益预测：P30

循环验证：使用除了目标领域的所有局点训练模型（三折交叉验证选最优模型参数），在目标领域上预测

局点	A	B	C	D	E1	E2	E3	T
F1	0.947644	0.500699	0.77681	0.466667	0.95339	0.94434	0.965079	0.938493
P30	0.699055	0.535865	0.664399	0.253165	0.765556	0.675497	0.637459	0.600277

* P30是指的预测值和真实值误差在30%以内的小区比例

研究内容：一些关于**Ridge Regression**的问题

研究内容

深度模型建模：

MLP：使用小区平均数据

BiLSTM：使用一周时序数据

DeepMIL：使用小时级别数据

数据粒度：综合使用了全连接神经网络、长短时记忆网络和深度多示例网络来处理各种粒度的数据

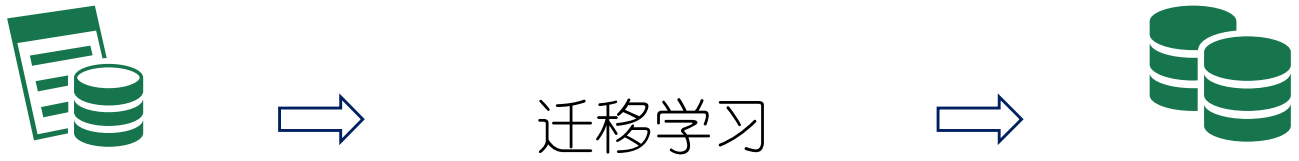
局点	A	B	C	D	E1	E2	E3	T
F1	0.957983	0.401515	0.772991	0.421769	0.952381	0.943161	0.962025	0.965287
P30	0.703104	0.324895	0.686319	0.177215	0.732222	0.651656	0.613363	0.503458

相比较于Ridge，P30性能有所下降，和数据标记噪音有关系，因此后文只考虑使用线性模型Ridge来预测

研究内容

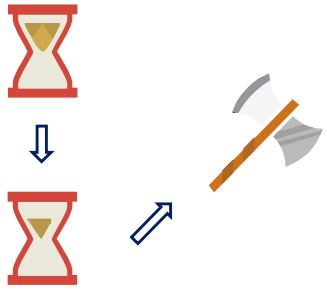
迁移学习：

迁移学习主要目的是将已有领域（源域）的知识迁移到新的场景（目标域），辅助目标域快速有效地部署好的模型

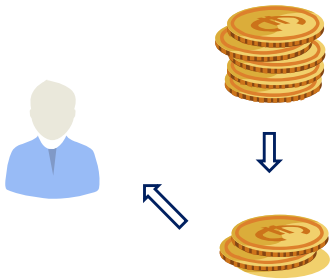


源域　　目标域

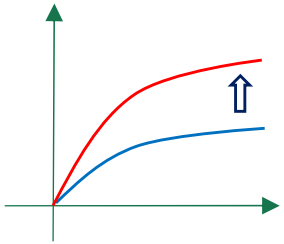
迁移学习具有节省时间成本、节省标注成本、提升模型性能等优点，可以解决目标域因缺乏算力、缺乏有效标记数据等难点



减少新场景下模型训练的时间



新场景标记样本不足，降低标注成本



利用任务之间的相关性提升模型性能

研究内容

KLIEP、SA等算法：

KLIEP (Kullback Leibler Importance Estimation Procedure)通过KL距离来衡量两个分布之间的差异，并通过样本加权方式去学习一组和样本相关的权重

问题抽象 $\arg\min_{M} \|X_S M - X_T\|_F^2$

$$KL\left(P_{te}(x)||\beta_x P_{tr}(x)\right) = \int_{\mathcal{D}} P_{te}(x) \log \frac{P_{te}(x)}{\beta_x P_{tr}(x)} dx = -\int_{\mathcal{D}} P_{te}(x) \log \beta_x dx + const$$

样本权重 ⟹ $\beta_x = \sum_{l=1}^{b} \alpha_l \varphi(x)$ $\int_{\mathcal{D}} \beta_x P_{tr}(x) dx = 1$ ⟹ 优化求解

$$\max_{\{\alpha_l\}_{l=1}^{b}} \sum_{j=1}^{n} \log\left(\sum_{l=1}^{b} \alpha_l \varphi_l(x_j)\right)$$

$$s.t. \quad \sum_{i=1}^{m}\sum_{l=1}^{b} \alpha_l \varphi(x_i') = m, \quad \{\alpha\}_{l=1}^{b} \geq 0$$

SA (Subspace Alignment)对数据的子空间进行对齐，采用简单的线性映射将源域的子空间X_S和目标域的子空间X_T对齐

Data: Source data S, Target data T, Source labels L_S, Subspace dimension d
Result: Predicted target labels L_T
$X_S \leftarrow PCA(S, d)$;
$X_T \leftarrow PCA(T, d)$;
$X_a \leftarrow X_S X_S' X_T$;
$S_a = S X_a$;
$T_T = T X_T$;
$L_T \leftarrow Classifier(S_a, T_T, L_S)$;
Algorithm 1: Subspace alignment DA algorithm

研究内容

TCA、KLIEP、SA等迁移算法：

高低增益局点识别：通过一个局点的增益预测值的平均值反映，红色是潜在高增益局点

局点	A	B	C	D	E1	E2	E3	T
Real	0.054	0.017	0.036	0.006	0.065	0.063	0.074	0.062
Predict	0.059	0.018	0.043	0.011	0.067	0.074	0.071	0.054

高增益局点**P30**指标：

- 使用特殊的特征工程方法使得性能提升约**10%**
- 使用KL和SA迁移综合使用可以将性能提升约**5%**
- 局点A、E1、E3、T基本可以满足要求，P30可接近0.7达到落地标准

P30	A	C	E1	E2	E3	T	Mean
原始数据	0.500	0.480	/	/	/	0.390	0.480
特征工程	0.788	0.487	0.673	0.562	0.603	0.417	0.588
KL迁移	0.772	0.488	0.658	0.580	0.665	0.550	0.619
SA迁移	0.813	0.484	0.664	0.575	0.648	0.675	0.643

目录

关键技术点

标签不准确性度量：

在无线业务场景，标签是通过特定的数据（比如子特性开通之后才能收集到的数据）拟合出来的，因此存在标签本身就不准确的问题

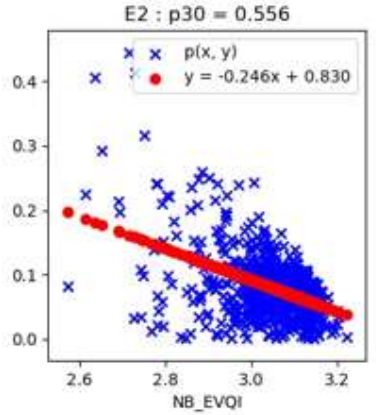
Algorithm：度量标签不准的程度

Step1：对于所有样本有监督地降维（线性、非线性）

Step2：对于每一个样本x，在隐层空间使用KNN寻找近邻，比如10个近邻样本

Step3：逐一计算10个近邻样本的标签的方差

Step4：计算所有方差的平均值当作样本标签的不准确性，即p(y | x)的方差



上图：标签不准确的定性分析，同一个x可能对应多个目标值，造成了回归任务变得很困难

局点	A	B	C	D	E1	E2	E3	T
3p30	0.822	0.364	0.518	0.28	0.67	0.584	0.673	0.73
ratio	**0.239**	0.626	**0.426**	0.672	0.317	**0.431**	0.34	**0.28**

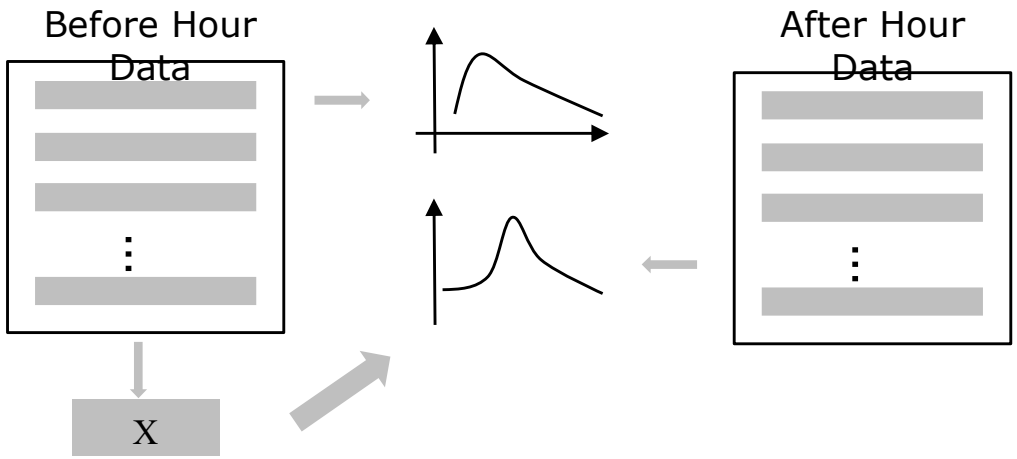
Ratio指的是使用上述度量方法计算的标签的方差和标签真实的均值的比例。

从统计学原理，当ratio小于0.3时，p30才可能达到0.7以上，比如A、T局点；并且**ratio越大，标签越不准确，p30性能越差**。

关键技术点

解决标签不准问题：

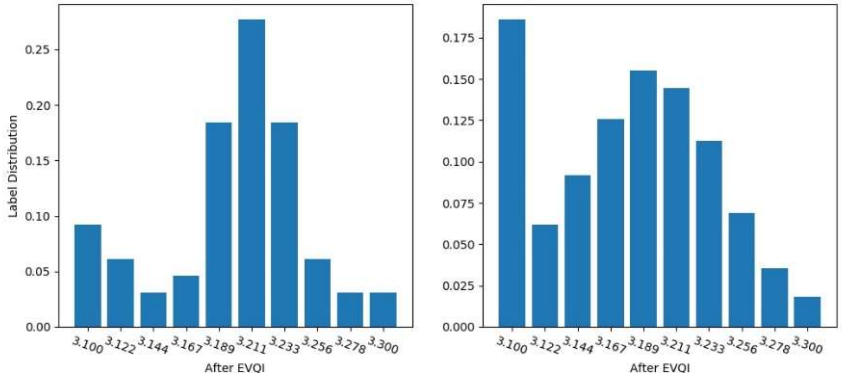
引入概率建模， 借助LDL方法，其做法是将标签建模成一个概率分布，而不是一个单独的标签值，特别适合标签本身就不准确的业务场景



LDL在水晶语音case中的示例图

局点	A	B	C	D
Self	0.827	0.737	0.799	0.777
Transfer	0.824	0.644	0.761	0.708
局点	E1	E2	E3	T
Self	0.821	0.804	0.801	0.699
Transfer	0.805	0.762	0.775	0.600

LDL预测Intersection指标，使用迁移的性能可以逼近自身局点有标记数据集情况下训练的性能



LDL预测示例图，左图是真实的标签分布，右图是预测的标签分布，评价指标是两个分布的Intersection指标

关键技术点

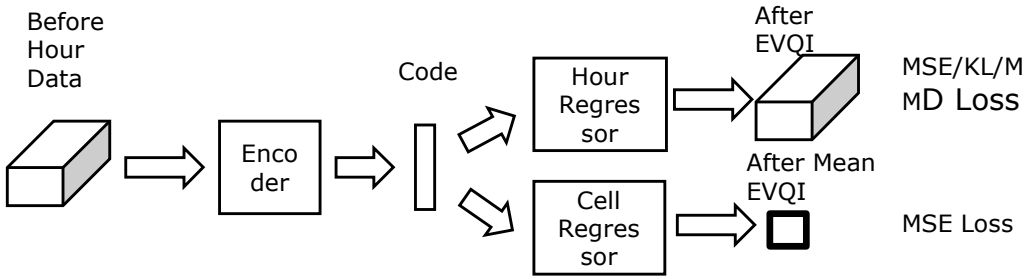
解决标签不准问题：

利用生成式模型，生成式模型将小时级数据和标签问题综合考虑，使用生成网络拟合子特性开通之后的数据/通信质量，无需打标签

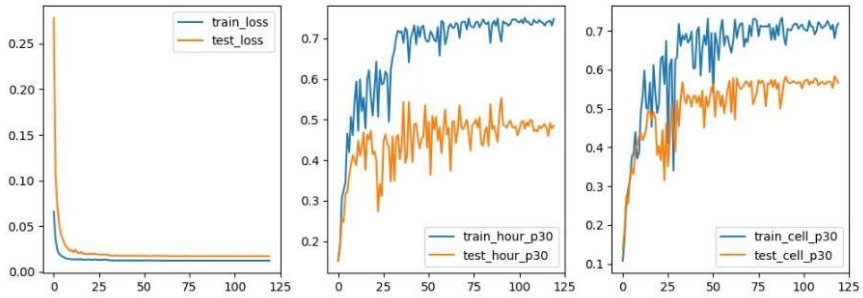
Model	MLP	LSTM	Conv
P30	**0.57**	0.52	0.49

Ridge Baseline	Train	Test
无特征工程	0.57	**0.56**

尝试使用了MLP，LSTM和CNN等模型构建生成式模型，使用生成式模型无须打标签（无监督训练方式），达到无特征工程下的Ridge的基线（有监督训练）



生成式模型训练框架图，根据自特性开关打开前的数据预测打开后的流量数据，然后计算信息质量增益



生成式模型训练过程训练测试损失、训练测试的P30值变化情况

关键技术点

可迁移性度量：

在实际过程中要解决局点之间模型是否可以迁移的问题，从学术界理论研究进行分析，然后提出相应的解决工业界近似方案

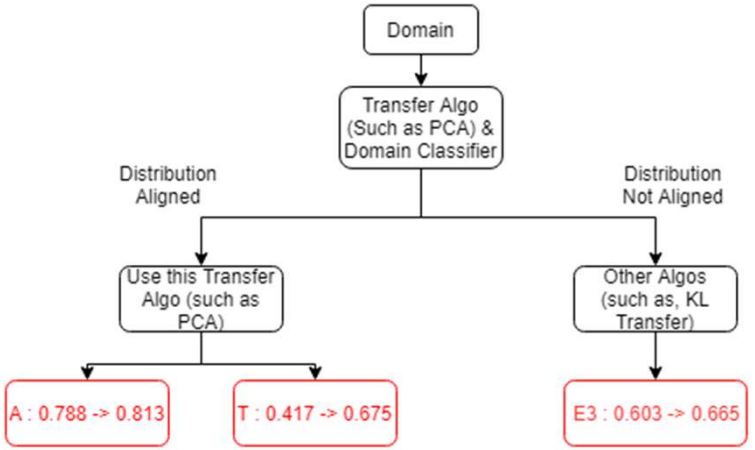
学术界理论研究：

$$\epsilon_T(h) \leq \epsilon_S(h) + \mathbf{d_1}(\boldsymbol{D_S}, \boldsymbol{D_T}) + \min\{\mathrm{E}_{D_S}[|f_S(x) - f_T(x)|], \mathrm{E}_{D_T}[|f_S(x) - f_T(x)|]\}$$

工业界经验近似：提出基于Domain Classifier的可迁移性判别技术

局点	A	B	C	D
可迁移性	0.366	0.591	0.275	0.281
P30性能提升	0.028	0.144	-0.006	-0.007
局点	E1	E2	E3	T
可迁移性	0.195	0.18	0.295	0.926
P30性能提升	-0.018	0.026	0.056	0.248

预测的可迁移性和真实迁移之后的P30性能提升非吻合



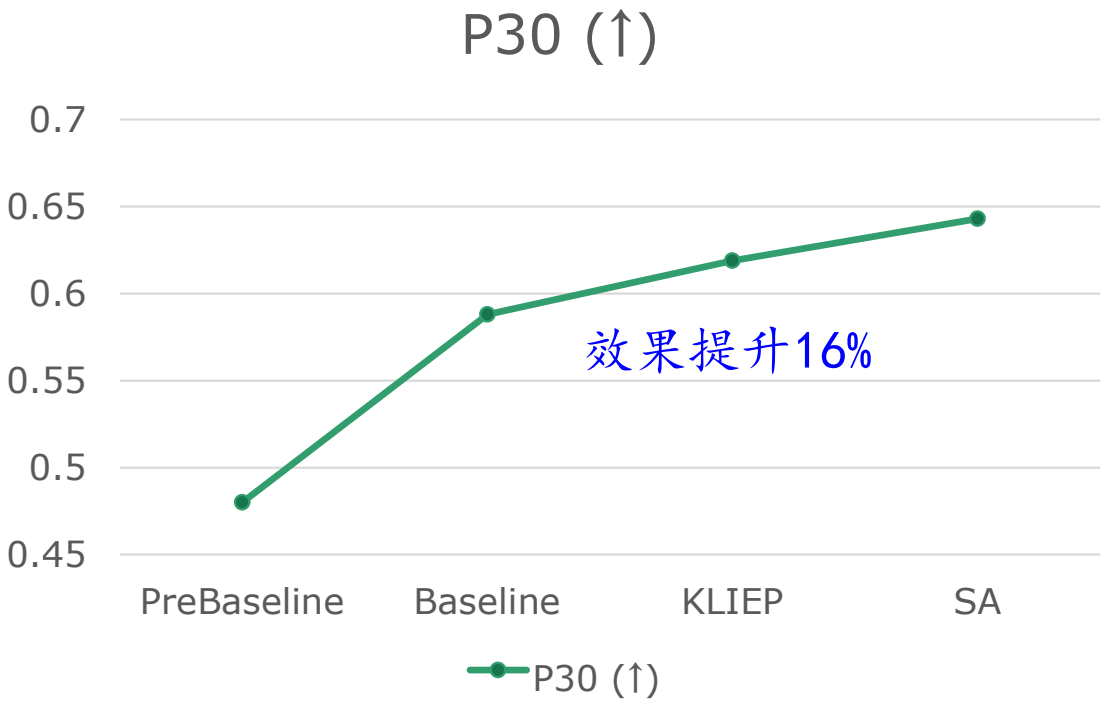
基于领域分类器的可迁移性判别流程

目录

研究成果

项目指标性能：

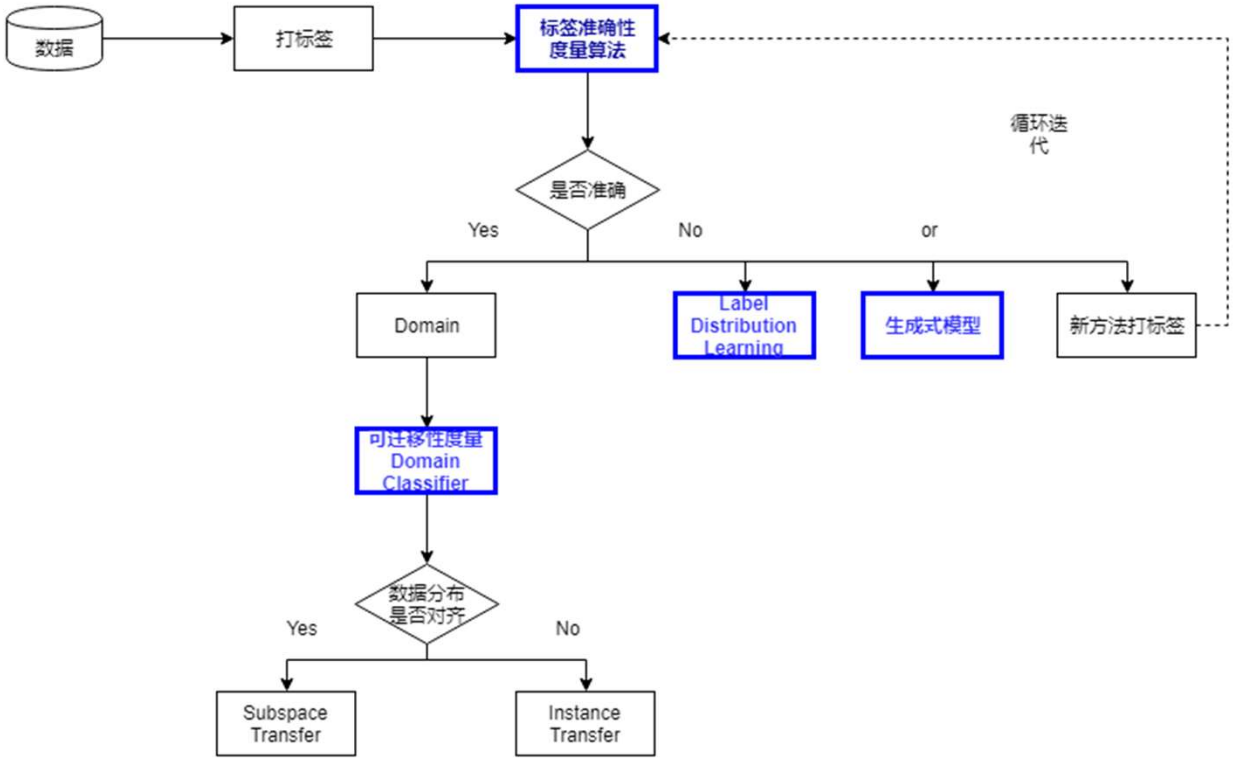
P30	Mean
原始数据	0.480
特征工程	0.588
KL迁移	0.619
SA迁移	0.643



在所提特征工程、**KL**迁移和**SA**迁移算法的支撑下，项目指标**P30**提升高达**16%**，很多局点上的性能已经达到实际落地标准

研究成果

无线业务场景述求通用解决方案：

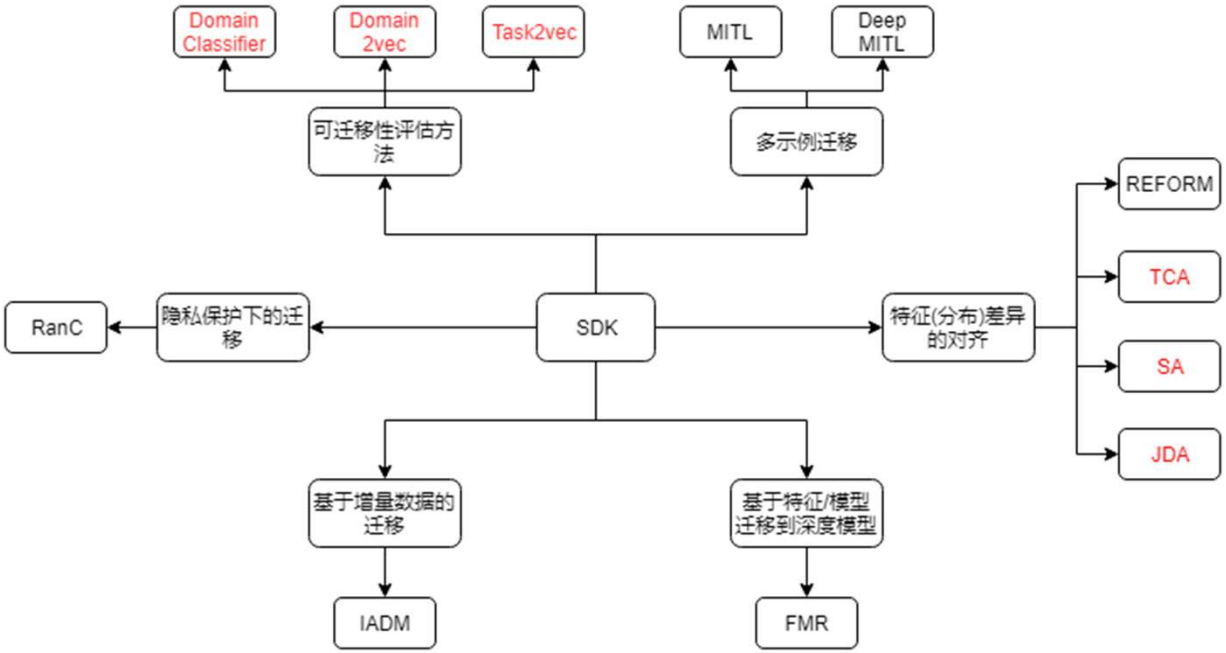


- 收集到数据之后，对数据打标签
- 使用之前的标签度量方法判断标签是否足够准确
- 如果标签准确，使用迁移技术度量方案评估选择基于子空间的迁移技术还是基于样本的迁移技术
- 如果标签不准确，可以使用Label Distribution Learning的方法或者生成式模型等方法打标签

无线业务场景的标签不准确问题、如何对标签引入概率因素问题都有了相应的解决方法

研究成果

迁移算法SDK：



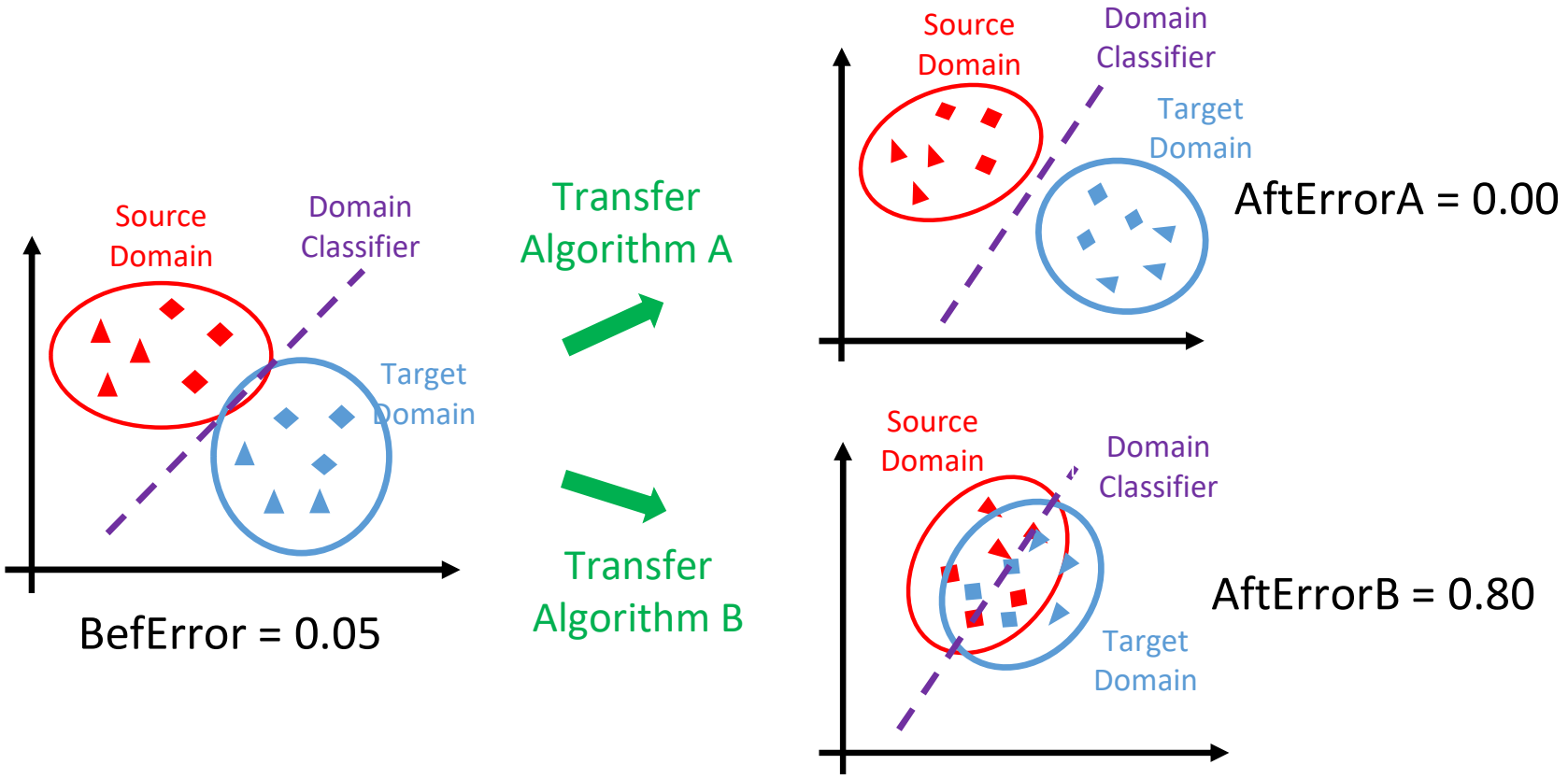
- 项目验证过程中开发实现的各种迁移算法，比如PCA、KMM、TCA等算法可以被封装进SDK，为以后的迁移项目节省开发成本
- 封装基于Domain Classifier的迁移性判别技术
- 封装多示例学习的相关算法，比如DeepMIML和AttentionMI，可以应用到后续适合使用多示例的项目中
- 封装Label Distribution Learning的相关算法

无线业务场景的标签不准确问题、如何对标签引入概率因素问题都有了相应的解决方法

研究成果

专利：基于领域分类器度量可迁移性度量

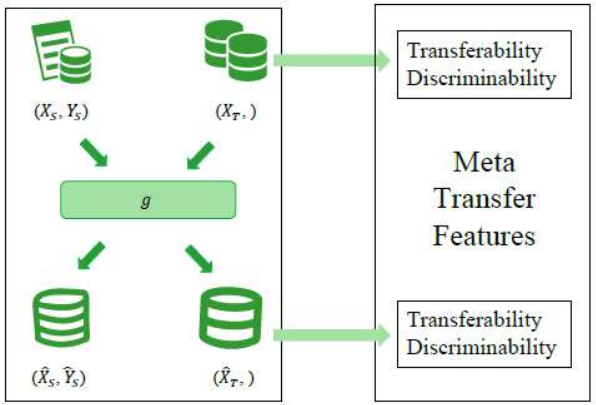
本发明重点解决源域和目标域数据在特征分布存在差异情况下的可迁移性度量，相比较于前人研究的定性分析和评估，本发明提出了一种基于领域分类器进行定量计算可迁移性的方法，通过训练一个二分类器来评估两个领域数据之间的可迁移性：



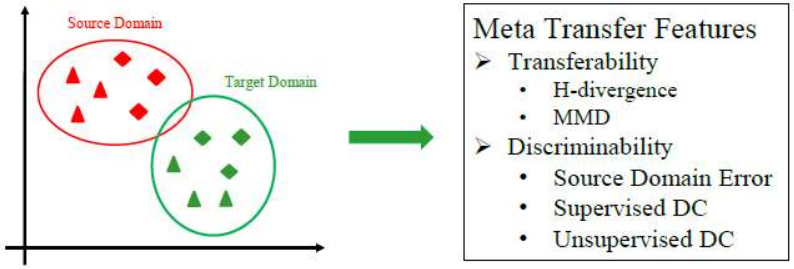
研究成果

论文： Towards Understanding Transfer Learning Algorithms Using Meta Transfer Features. PAKDD 2020.

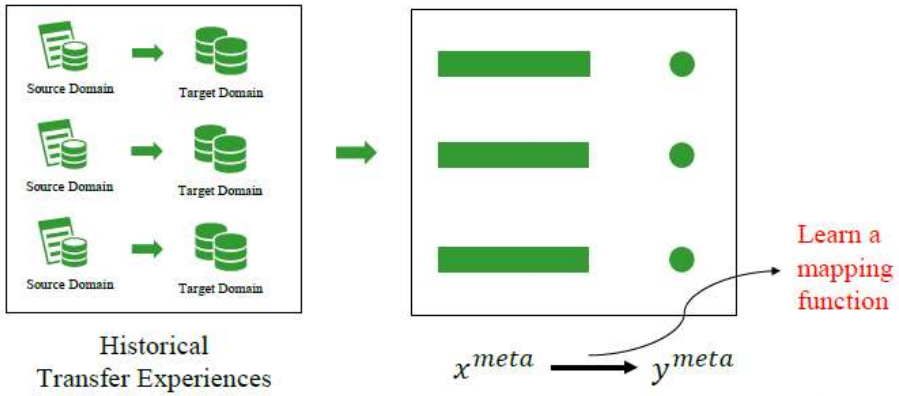
Meta Transfer Features



We focus on the unsupervised transfer learning (unsupervised domain adaptation): the source domain is labelled, while the target domain has no labelled data.



在两个领域数据上提取元迁移特征



Given a pair of source and target domain, with a transfer learning algorithm, we can obtain meta transfer features as x^{meta}, and the transfer improvement ratio as y^{meta}. MetaTrans learns a mapping from x^{meta} to y^{meta}.

构建元迁移特征到迁移性能提升的映射

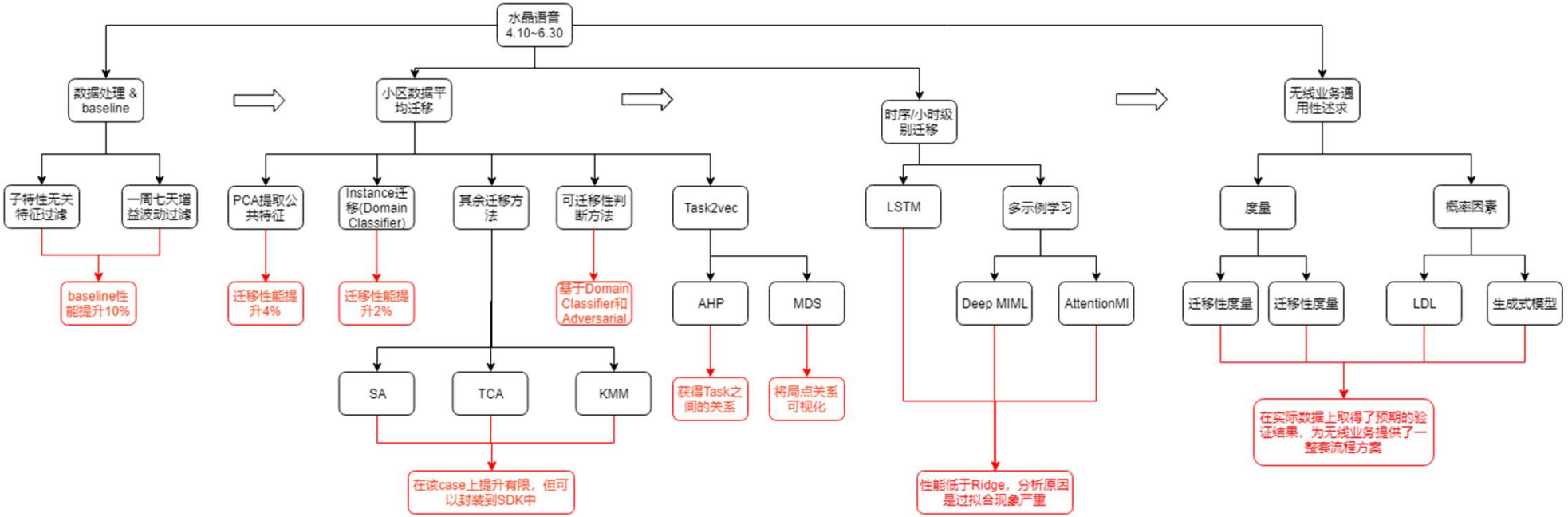
Train and Test Sets	Method	MSE	MAE
Train: $\mathbf{A} \to \mathbf{C}, \mathbf{A} \to \mathbf{D}, \cdots, \mathbf{W} \to \mathbf{D}$	Meta-Sin	0.0339	0.1573
	Meta-Inv	0.0418	0.1724
Test: $\mathbf{U} \to \mathbf{M}, \mathbf{M} \to \mathbf{U}$	Meta-MTL	**0.0314**	**0.1507**
Train: $\mathbf{A} \to \mathbf{C}$	Meta-Sin	0.0104	0.0821
	Meta-Inv	0.0162	0.1065
Test: $\mathbf{A} \to \mathbf{D}$	Meta-MTL	**0.0081**	**0.0729**

所提算法预测的迁移性能提升能够很好的拟合迁移任务的真实迁移性能

目录

研究成果

研究过程梳理：



可扩展方向：

- 跨区域通信流量预测
- 多型号硬盘故障检测
- 多APP下的智能业务感知