群顶科技
TrendyTech

KPI异常检测

群顶科技-烧脑特工队

目录

CONTENTS

团队介绍

群顶科技-烧脑特工队

隶属于南京群顶科技有限公司人工智能研究室。该研究室成立近2年的时间，主要针对电信运营商数据中心的软硬件多维度数据作探索，寻求用机器学习方法解决故障预测、异常发现、事件归因分析等运维难题，是专门从事AIOps研究的部门。

烧脑特工队是专门从事数据分析和算法建模的团队，由多名经验丰富和富有活力的数据及算法工程师组成，已实现如磁盘故障预测、基于请求数基线的业务异常发现等多种场景的AIOps应用，帮助电信运营商在灾害预防、节能减排方面取得创新性突破。

张永磊

2013年毕业于苏州大学自然语言处理实验室后一直从事社交、法律、医保、电信等数据挖掘相关工作。

徐磊

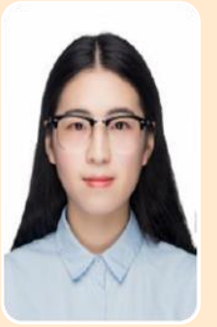
2011年吉林大学软件工程专业硕士毕业，专注于海量数据处理、分析以及建模。

周宇

南京信息工程大学应用统计学，硕士研究生，主要负责数据分析相关工作。

高鹏举

东南大学硕士研究生在读，主要负责磁盘SMART数据分析与故障预测建模。



张琳琳

东南大学硕士研究生在读，目前主要负责磁盘故障以及寿命预测的优化工作。

目录

CONTENTS

技术方案流程

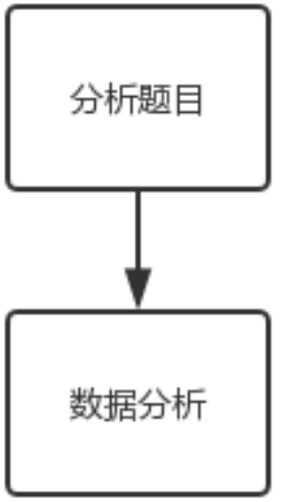
1、分类问题?

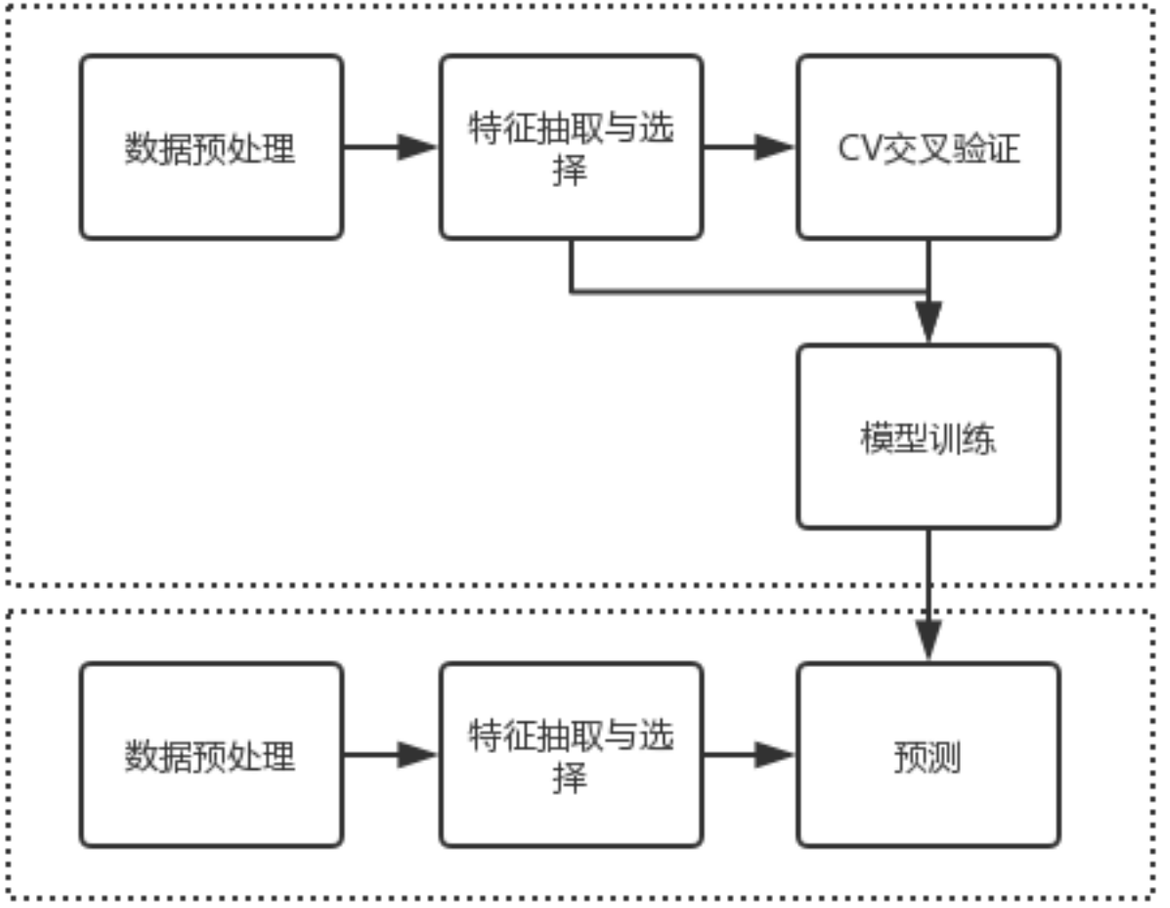
正负样本分布不平衡

2、回归问题?

异常点没有好的处理方式

3、无监督学习方法?





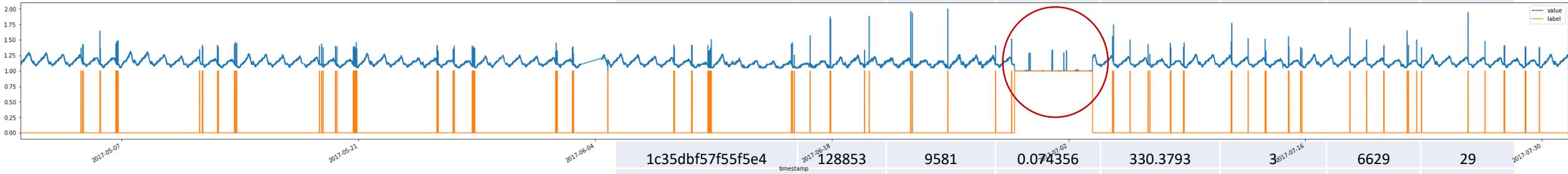
技术方案-数据分析

结论：随着时间变化时序数据在发生变化，但曲线类型以及周期等不会发生变化

处理方式：数据标准化处理将根据最近2w条数据进行处理

KPI ID	train_mean	test_mean	train_std	test_std
02e99bd4f6cfb33f	1.950093238	1.456195361	0.653738948	0.615630579
046ec29ddf80d62e	-0.021680707	0.021680706	0.960396762	1.03775024
07927a9a18fa19ae	0.155113421	-0.198748731	1.05112791	0.910105794
09513ae3e75778a3	1025131262	1014029766	155216825	117773086
18fbb1d5a5dc099d	4.296634656	5.351093186	8.34129986	9.98829972
1c35dbf57f55f5e4	1258.931531	1357.56323	124.563839	177.279185
40e25005ff8992bd	1348.611168	1382.5878	541.900571	554.957669
54e8a140f6237526	-0.13189758	0.13188159	0.935478932	1.04418134
71595dd7171f4540	1519.416008	1290.81875	648.767664	497.988293
769894baefea4e9e	0.088519127	0.094763535	0.046166213	0.050231145
76f4550c43334374	0.093644829	0.089637832	0.049568352	0.04699931
7c189dd36f048a6c	1792.331223	1721.737707	674.276096	635.069822
88cf3a776ba00e7c	0.167033412	0.200166594	0.079311622	0.089916292
8a20c229e9860d0c	0.104737148	0.078545513	0.052330522	0.039905666
8bef9af9a922e0b3	1697.345307	1872.208669	659.291449	700.260405
8c892e5525f3e491	6498.535892	6288.837147	2364.48853	2195.99635
9bd90500bfd11edb	31.61091766	24.1068212	11.8581694	7.48273119
9ee5879409dccef9	0.117959105	0.154540807	0.061467499	0.078139488
a40b1df87e3f1c87	1779.908211	1737.1547	675.179619	641.963981
a5bf5d65261d859a	0.010538757	0.014037476	0.137453242	0.12640971
affb01ca2b4f0b45	1766.516299	1790.843013	691.173819	673.966299
b3b2e6d1a791d63a	-0.124606245	0.124591138	0.885393916	1.08868837
c58bfcbacb2822d1	1051384034	1028128022	92924775.3	100623868
cff6d3c01e6a6bfa	1507.293926	1292.809489	645.285543	498.540785
da403e4e3f87c9e0	2.465291355	2.805912529	1.15419221	1.13069512
e0770391decc44ce	6593.521003	6370.531934	2.38E+03	2.20E+03

技术方案-数据分析



	数据量	异常点数	异常比例	平均连续点	最小连续异常点	最大连续异常点	异常区间数
1c35dbf57f55f5e4	128853	9581	0.074356	330.3793	3	6629	29
40e25005ff8992bd	100254	654	0.006523	11.27586	2	95	58
54e8a140f6237526	8248	4	0.000485	4	4	4	1
71595dd7171f4540	147668	1113	0.007537	16.13044	1	202	69
769894baefea4e9e	8784	9	0.001025	3	3	3	3
76f4550c43334374	8784	87	0.009904	43.5	4	83	2
7c189dd36f048a6c	147689	426	0.002884	6.553846	2	95	65
88cf3a776ba00e7c	65436	3106	0.047466	65.2766	1	625	47
8a20c229e9860d0c	8784	4	0.000455	4	4	4	1
8bef9af9a922e0b3	129453	527	0.004071	9.581818	2	136	55
8c892e5525f3e491	147009	3083	0.020972	256.9167	4	2678	12
9bd90500bfd11edb	128613	143	0.001112	71.5	4	139	2
9ee5879409dccef9	65449	2945	0.044997	60.10204	1	632	49
a40b1df87e3f1c87	137925	381	0.002762	5.013158	2	59	76
a5bf5d65261d859a	128679	29	0.000225	4.833333	3	10	6
affb01ca2b4f0b45	147680	584	0.003954	8.848485	1	154	66
b3b2e6d1a791d63a	8247	12	0.001455	6	6	6	2
c58bfcbacb2822d1	128667	124	0.000964	5.166667	3	10	24
cff6d3c01e6a6bfa	147629	1087	0.007363	14.68919	1	202	74
da403e4e3f87c9e0	129035	7666	0.05941	68.44643	4	6643	112
e0770391decc44ce	147024	3074	0.020908	279.4545	3	2698	11

异常连续集中且比例高：

（1）线下划分训练测试集

（2）数据偏向性

数据采样：在训练数据中，对连续性异常区间进行采样

技术方案-数据分析

在保证准确率为1时，通过极值过滤异常点，自动选择阈值

1、周期性

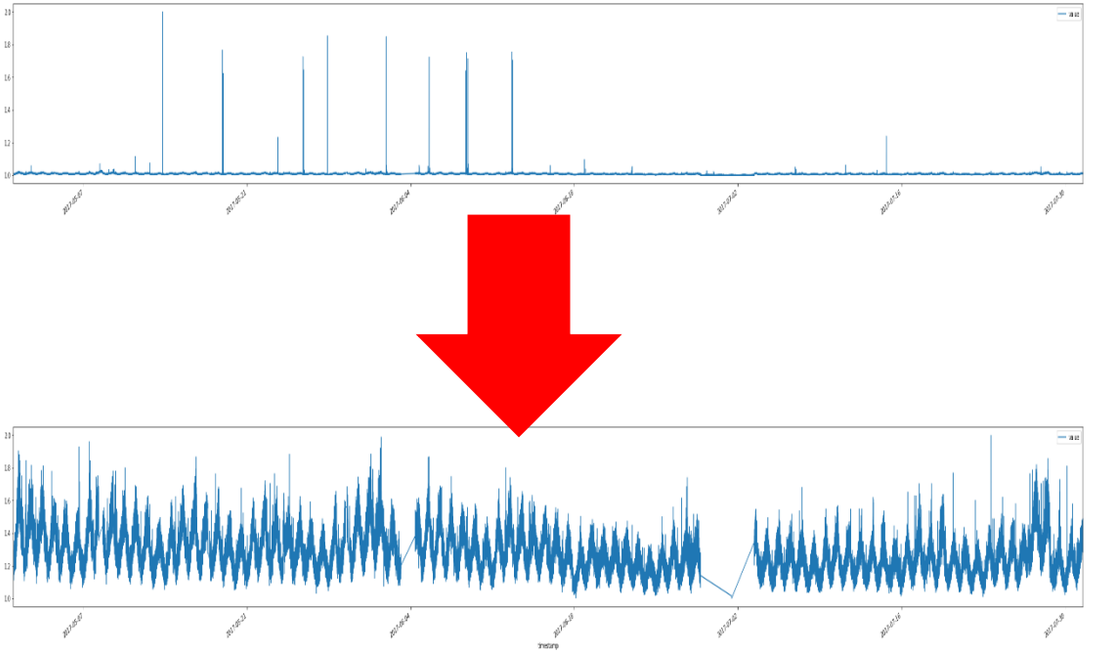
2、阈值也在发生变化

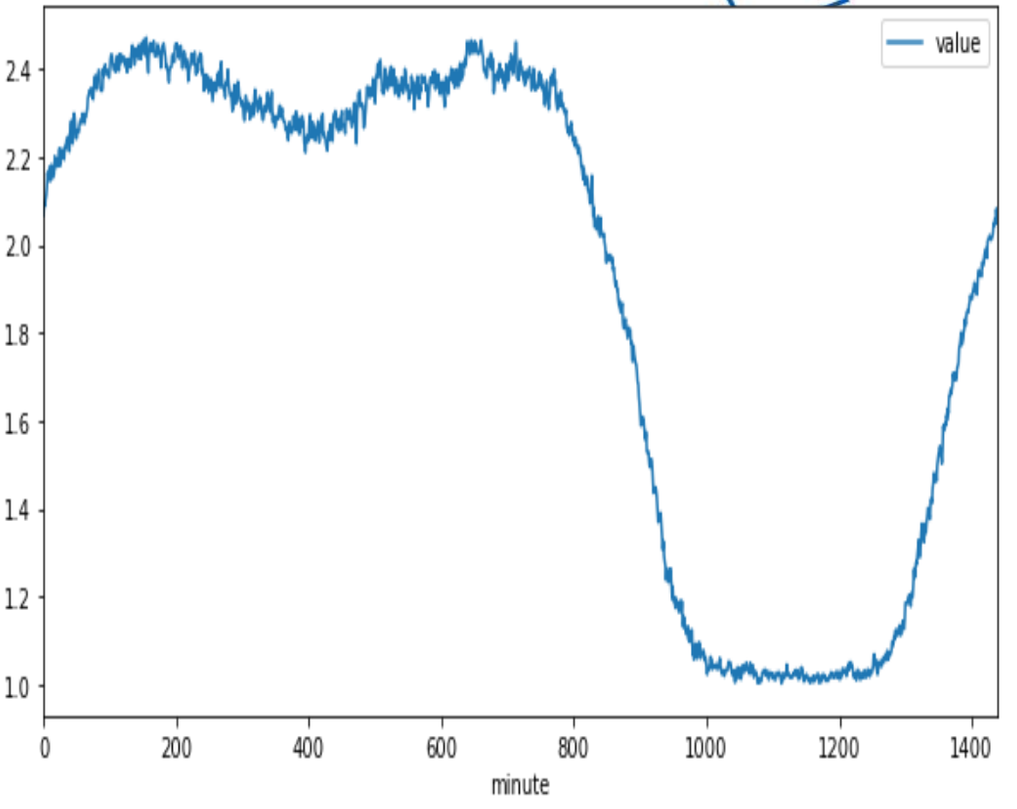
方案：

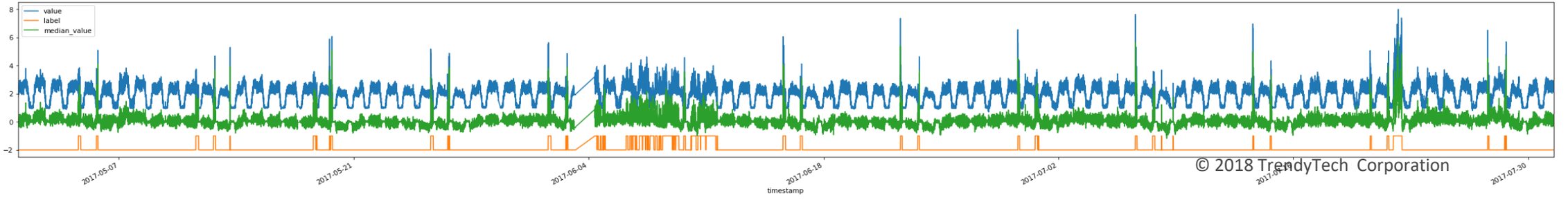
分别以：上阈值与最大值50%分位、下阈值与最小值50%分位截取

KPI ID	极值	异常点数	占比
02e99bd4f6cfb33f	max	589	0.055829
046ec29ddf80d62e	min	59	0.7375
07927a9a18fa19ae	min	68	0.523077
09513ae3e75778a3	max	1	0.004202
09513ae3e75778a3	min	2	0.008403
18fbb1d5a5dc099d	max	386	0.049091
18fbb1d5a5dc099d	min	6330	0.805036
1c35dbf57f55f5e4	max	257	0.026824
1c35dbf57f55f5e4	min	11	0.001148
40e25005ff8992bd	max	6	0.009174
40e25005ff8992bd	min	39	0.059633
71595dd7171f4540	max	16	0.014376
71595dd7171f4540	min	38	0.034142
76f4550c43334374	min	63	0.724138
7c189dd36f048a6c	max	4	0.00939
7c189dd36f048a6c	min	52	0.122066
88cf3a776ba00e7c	max	23	0.007405
88cf3a776ba00e7c	min	349	0.112363
8bef9af9a922e0b3	max	4	0.00759
8bef9af9a922e0b3	min	30	0.056926
8c892e5525f3e491	max	160	0.051898
8c892e5525f3e491	min	2474	0.802465
9bd90500bfd11edb	min	1	0.006993
9ee5879409dccef9	min	59	0.020034
a40b1df87e3f1c87	max	6	0.015748
a40b1df87e3f1c87	min	57	0.149606
a5bf5d65261d859a	max	11	0.37931
affb01ca2b4f0b45	max	7	0.011986
affb01ca2b4f0b45	min	45	0.077055
b3b2e6d1a791d63a	min	1	0.083333
cff6d3c01e6a6bfa	max	10	0.0092
cff6d3c01e6a6bfa	min	55	0.050598
da403e4e3f87c9e0	max	530	0.069136
da403e4e3f87c9e0	min	6578	0.858075
e0770391decc44ce	max	155	0.050423
e0770391decc44ce	min	2502	0.813923

技术方案-数据分析







技术方案-特征工程

- **历史窗口选择**

  历史窗口分别取（7-28），线下5折交叉验证，最后选取窗口11

- **差分特征**

  时序数据中连续相邻两项之差，体现了数据的变化

- **tsfresh时序特征抽取**

  开源时序特征抽取工具包，最终只是人工抽取了部分特征加入模型。

历史窗口	时序原始数据	时序差分	数据标准化处理	异常点采样	tsfresh特征	F-score
7	Y	Y				0.390493
7	Y	Y	Y			0.654987
11	Y	Y	Y			0.7051657
11	Y	Y	Y	Y		0.7588002
11	Y	Y	Y	Y	Y	未提交线上

oration

技术方案-模型选择

5折交叉验证，选择模型，最终选择xgb作为基础模型。

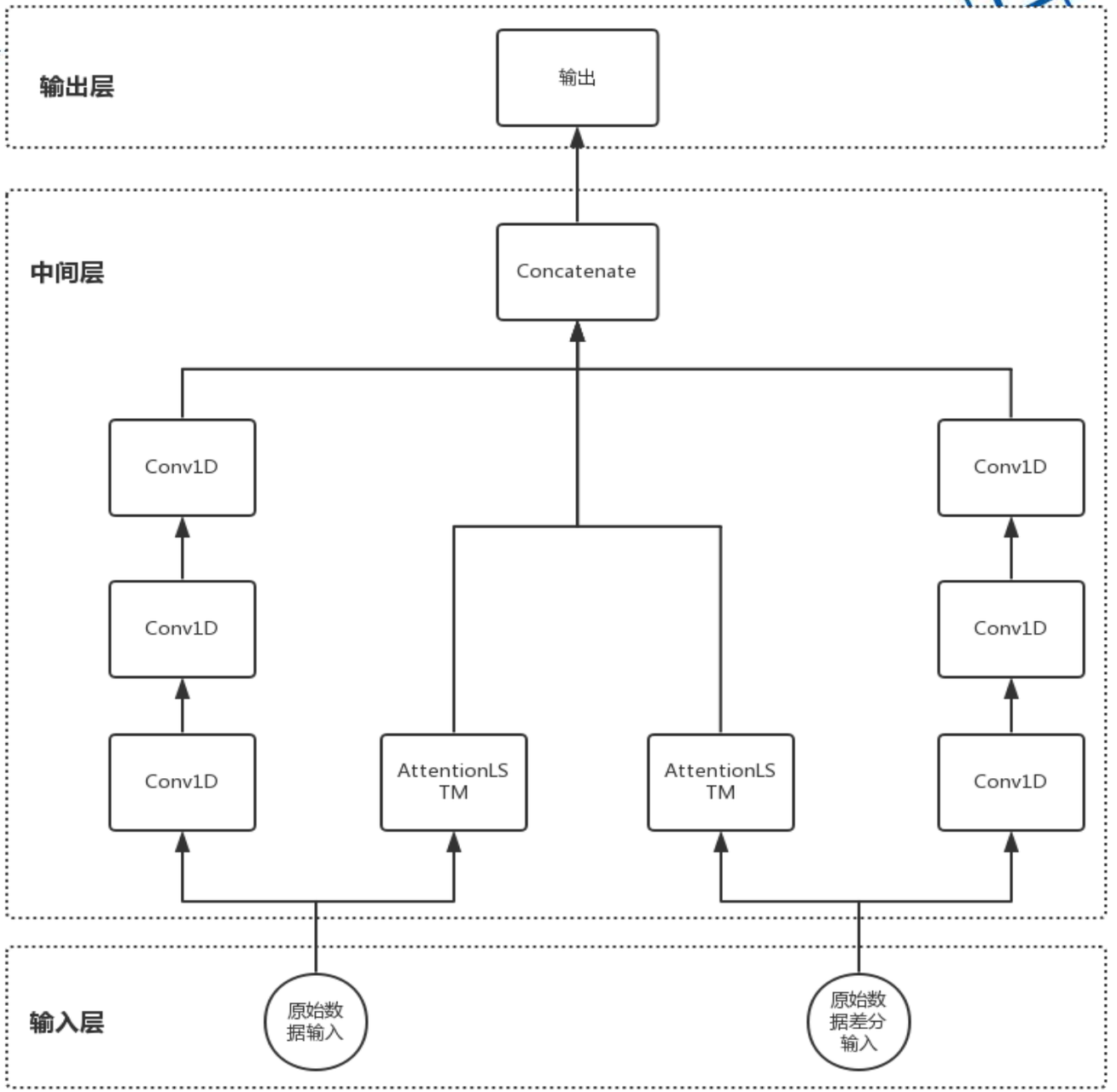
model	test_score-m	test_score-s	train_score-m	train_score-s	fit_time-m	fit_time-s	pred_time-m	pred_time-s
kbn	0.591199	0.136991	0.661241	0.01991	4.77759	1.471323	493.4611	12.6047
svm	0.626232	0.121125	0.664889	0.01786	730.569	179.5511	124.0037	42.2412
xgb	0.664286	0.050602	0.784104	0.00685	203.558	4.317997	1.140131	0.11369
en	0.620019	0.118301	0.66598	0.02328	711.496	167.0106	345.298	31.5414

注：en = [xgb, svm, kbn] + lr

技术方案-模型选择

- AttentionLSTM+CNN

- 初赛得分：0.620821733894



技术方案-模型集成

- 多模型投票（0.783470506235）

  使用不同采样方式训练多个模型，多个模型预测结果投票决定是否是异常点。

- Stacking方法（ml-ens开源工具包）

  将训练好的所有基模型对整个训练集进行预测，第j个基模型对第i个训练样本的预测值将作为新的训练集中第i个样本的第j个特征值，最后基于新的训练集进行训练。

  缺点：部分kpi异常数据非常少，正样本不足。

目录

CONTENTS

后续工作

- 进一步分析连续异常，提取连续异常特性
- 时序SAX表示，并对KPI数据采样聚类分组
- KPI分类之后，进行分组建模
- 针对不同类型的KPI选择不同的模型
- 增加KPI周期相关特征
- 调优深度模型网络结构及超参数

目录

CONTENTS

场景应用

磁盘故障预测

电信运营商正越来越多地使用服务器的机内磁盘作为存储介质，磁盘数量庞大，老化和损坏事件发生频率高且无序，给计划性维护带来挑战。通过对磁盘预估其剩余寿命，结合业务负载时序特征给出建议的更换时间窗口，不仅有效防范了系统性的数据丢失灾难，还能够实现预先采购，优化采购流程节约成本。

业务负载异动感知

电信运营商的私有云资源池的业务复杂性高，异常发现困难。通过监控1000多个业务的访问会话数量和资源负载情况，形成时序模型，比对实时负载，准确发现业务异常，实现按需的资源扩缩容。

群顶科技
TrendyTech

THANK YOU!