摘　要

随着公有云服务技术的发展，越来越多的互联网用户选择在公有云上访问虚拟服务器。同时，不少用户使用一段时间后便不再使用，成为“流失用户”。于是，如何通过用户监控指标的监控值建立用户流失预警模型并预测出潜在流失用户，成为了云服务提供商亟需解决的问题。本题基于公有云服务技术的应用背景，分析附件中流失用户与非流失用户的数据特征，并构造预测模型。

针对问题一，首先对数据做规范化处理。随后，综合指标的时间覆盖率、用户覆盖率、平均方差三种评价方式评价某一指标的数据质量，得到 20 个初筛指标。最后，基于数据典型性，即所有流失用户中该指标展现出明显流失趋势的用户的比例进行指标筛选，得到 8 个与用户流失相关的指标。

针对问题二，基于上述得到的 8 个关键指标，分别采用用户指标在记录时间内的均值、用户指标在单位时间内下降的平均幅度与用户指标值出现大幅下降的程度作为用户特征，标准化处理后通过 K-Medians 方法聚类，得到四种风险流失类型。

针对问题三，分别采用两种方法求解。一方面，将问题一得到的 8 个关键指标的用户数据作动态主成分分析降维，使用支持向量机完成二分类训练。另一方面，将通过长短期记忆网络的隐性时序信号与原始的显性时序信号叠加构成残差网络，输入时序卷积神经网络，将时间维度的信息转换为特征维度的信息，并在全连接层将上述特征融合为二分类的二维特征，提供预测结果。

关键词　用户流失分析　聚类　时序卷积神经网络　动态主成分分析

目录

一、 问题的背景与重述

1.1 问题背景

公有云服务能够为用户提供可通过互联网访问的虚拟服务器及其配套资源，近年来被广泛推广，在网上办公应用、机器学习的训练与存储、游戏开发等场景都得到应用。

随着市场的日趋饱和，云服务厂商面临着“用户流失”等问题，影响其用户活跃度与营收能力。本题希望能够构建用户流失预警模型，有效识别用户流失风险，以便厂商对潜在流失用户问询或挽留，提高云服务质量。

1.2 问题重述

问题一：根据 250 名流失用户监控指标的监控值，筛选出与用户流失紧密相关的重要指标。

问题二：根据流失用户和正常用户的资源利用情况，给出用户流失风险的分级并得到每级用户的对应特征。

问题三：基于流失用户与正常用户的流失指标监控值，给出流失用户具体判别方法，并构建模型来预测用户流失风险。依据流失用户监控指标的监控值，计算该模型的精确率、召回率及 F1-score 以评价该模型准确性，并分析模型对相关因素的依赖性。

二、 问题分析

2.1 问题一的分析

问题一要求根据附件 1 中的流失用户监控指标的监控值，建立筛选指标模型，筛选出与用户流失相关的重要指标，并说明选取的指标数量及原因。

在对附件一的时序数据预处理后，拟通过用户流失的特征在各指标时序数据的变动趋势体现来建立模型。本题难点在于用户多、多指标的时序数据维数极大，冗余较多，传统的 GRA 算法关联度分析较难解决；而时序数据的常用特征，包括均值、变异度等，在模型中也作用有限。

因此，我们考虑人为构建评价指标，综合时间覆盖率、用户覆盖率、平均方差三种评价方式评价某一指标的数据质量。考虑到代表性是重要指标最需要被考虑的特性，拟对通过上述方法筛选出的指标进行基于数据典型性的二次排序，得到最终指标。

2.2 问题二的分析

问题二要求根据附件 1 和附件 3 中的用户资源利用情况，建立用户流失风险分级模型，并给出每一流失风险等级用户特征的数学描述。

本题一方面有除了问题一中提到的用户多、多指标的时序数据维数极大，冗余较多的特点，另一方面也会受到数据的离群点的影响。因此，不仅要在提取特征时要消除用户间由记录信息不完全带来的差异，还要在使用聚类算法时要考虑到离群点的影响。

在实际应用层面，用户风险等级划分要考虑到划分标准的实际意义，因此考虑人为构建指标，选取平均流量、下降比例与大幅下降值作为评价指标，同时消除时间维度，便于不同用户间的横向比较，最后采用 K-Medians 聚类分析。

2.3 问题三的分析

问题三要求基于问题（1）筛选出的重要监控指标，根据附件 1 与附件 3 中的用户监控指标的监控值，构建用户流失预测模型，说明流失用户的具体判别标准。并根据评价指标评估模型的正确性。考虑到数据维数多且含有时间序列，拟采取两种方法求解。

其一，可以对时序特征使用动态主成分分析算法降维数据，采用支持向量机的方法分类。

其二，采用自行搭建的残差长短期记忆卷积预测模型。考虑到数据量大、时间序列关系强，存在复杂的隐式时序特征，可以通过长短期记忆网络的隐式特征与原信号的显式特征相叠加，得到残差网络，再通过卷积神经网络将残差信号的时间维度转为特征维度信息，最后通过全连接层压缩信息，得到预测结果。

三、 符号说明

序号	符号	含义
1	n_j	拥有指标 j 的用户个数
2	N_i	第 i 个用户拥有的指标个数
3	t_{ij}	第 i 个用户的指标 j 覆盖的时间

四、 模型假设

- 假设云服务厂商对同一监测指标在不同用户中的监测方式都是相同的。
- 假设每个用户不同检测指标的检测时间不同、每个用户拥有检测指标的数量不同是云服务厂商记录平台导致而与用户自身无关。

五、 模型建立与求解

5.1 数据预处理

为了提高用于后续研究的数据集的质量，找出并纠正数据集中不完整或者错误的数据，我们首先要清洗数据。

5.1.1 缺失值的处理

观察发现，部分用户监控指标的最大值或最小值出现缺损。考虑到自行补充数据集会造成数据污染，且用户的最大最小值波动程度较大难以体现总体趋势，而监控指标的均值没有缺失情况且较为平稳，因此除特殊说明外后续求解均用监控指标的均值。

5.1.2 异常值处理

1. 错误值的处理
   统计发现，数据中一共有 10.5% 的数据值存在着均值大于最大值或者均值小于最小值的不合理情况，若全部删除会对数据的完整性带来影响。因此，我们对异常值采用线性填充的方法，即用前后两个数据值的平均值覆盖原有的异常值，以便在剔除异常数据时最大程度的保留原有的趋势。

2. 离群值的处理
   数据中同时出现了脱离正常波动与跳变范围内的极大或极小值。采用基于箱型图的识别方法处理，可以避免受极端值影响。若某一数据值小于 QL-1.5IQR 或大于 QU+1.5IQR（QL 为下四分位数，QU 为上四分位数，IQR 为四分位距），则采取同上的线性填充法替换。

5.1.3 数据平滑处理

观察数据发现，数据含有轻微噪声，同时存在大范围突变波动波动。据题意，由于用户的实际使用和自身业务的特点，监控指标发生突变较为正常，当监控指标发生长期的趋势性变化时，才能判定为“流失用户”，因此考虑采用滑动平均滤波算法对数据处理。

具体操作上，用宽度为 3 的滑动窗口通过某用户某指标的时间序列，取窗内数据的平均值作为新序列输出值。这样可以在保留数据原本的波动性的基础上增加其平滑性。

5.2 问题一：选取重要指标

5.2.1 数据规范化

为解决不同用户不同监控指标的数据间量纲不一致、数值间差别较大的问题, 对所有指标分别除以其极差，将数据映射到 $[0,1]$ 区间内：

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

其中，X 为处理前的数据，X_{norm} 为规范化处理后的数据，X_{max} 与 X_{min} 分别是处理前同一指标数据中的最大最小值。

5.2.2 基于数据质量的指标选取

我们采用如下三种评价方式评价某一指标的数据质量：

- **指标的时间覆盖率**

  在我们通过如下公式计算指标 j 的时间覆盖率 T_j：

$$T_j = \frac{1}{n_j}\sum_{i}^{n}\frac{t_{ij}}{t_{imax}}, i \in \{1,2,...,n_j\} \tag{2}$$

  其中，n_j 为拥有指标 j 的用户个数，t_{ij} 为第 i 个用户的指标 j 所覆盖的时间，t_{imax} 是第 i 个用户所有指标中最长的覆盖时间。同一用户的不同指标数据，记录的起始时间相近，中间均保持 3 天一次的记录频率，但记录的终止时间差异较大，因此时间覆盖程度不同主要体现终止时间的不同。时间覆盖率越高的指标，表达的用户波动信息更全，反之，时间覆盖率低的指标容易出现后期波动信息的缺失。

- **指标的用户覆盖率**

  指标 j 的用覆盖率 U_j 计算公式如下：

$$U_j = \frac{n_j}{n} \tag{3}$$

  其中，n_j 为拥有指标 j 的用户个数，n 为总用户个数。用户覆盖率越高的指标，越具有普遍性与可信性。

- **指标的平均方差**

  指标 j 的平均方差 V_j 计算公式如下：

$$V_j = \frac{1}{n_j}\sum_{i}^{n} v_{ij}, i \in \{1,2,...,n_j\} \tag{4}$$

  其中，n_j 为拥有指标 j 的用户个数，v_{ij} 为第 i 个用户的指标的方差。方差越大的数据有较大的波动性，反应的用户变换趋势信息也越多。

对于三种评价方式，我们选取其交集：只有在三种评价方式下均超过平均值的指标才被纳入候选指标中。经过第一轮筛选后，我们得到如下指标：{'13,1,11', '13,1,6', '1,1,2', '4,2,5', '4,2,1', '4,2,2', '1,1,1', '13,1,8', '4,1,2', '13,1,7', '13,1,4', '4,1,5', '1,1,4', '4,1,4', '4,1,3', '4,2,3', '4,1,1', '13,1,9', '4,2,4', '1,1,3'}。

5.2.3 基于数据典型性的指标选取

我们评价某一指标具有典型性的标准为：在所有流失用户中该指标展现出明显流失趋势的用户的比例。该比例越高，表明该指标的变化趋势越能反映用户流失的整体趋势。

1. **数据二值化处理**

   为了更好地展现用户的流失趋势，我们模仿图像的二值化方法将指标数据值 0-1 化。具体公式如下：

$$a_{ijt} = \begin{cases} 0 & x_{ijt} \leq 0.1 * \mu_{ij} \\ 1 & x_{ijt} \geq 0.1 * \mu_{ij} \end{cases} \tag{5}$$

   其中，x_{ijt} 为第 i 个用户在时间 t 上指标 j 的数值，a_{ijt} 为 0-1 化的后的数值，μ_{ij} 为第 i 个用户在全记录时间内指标 j 的平均值。

2. **时间的预处理**

   同一用户不同指标记录的起始时间基本相同，而我们关心的是每个用户各自独立的变化趋势，因此我们把所有用户的时间起始点平移到同一时间点，记该点为时间零点，以 3 天为时间间隔（原始数据的记录间隔）记录数据。由于不同用户同一监控指标的持续时间不同，我们用时间较短的监控指标按照最后一个时间点时的监控值延拓数据至相同长度。

3. **比例值计算**

   比例值 r_{jt} 衡量指标 j 在 t 时间点的流失用户数占 t 时间点拥有指标 j 的人数的比例，其计算公式为：

$$r_{jt} = 1 - \frac{1}{n_{jt}} \sum_{i}^{n_{jt}} a_{ijt} \tag{6}$$

   其中，n_{jt} 为 t 时间点拥有指标 j 的总用户数，a_{ijt} 为上文二值化后的数值。

4. **指标选取**

   选取指标的标准是

   - 最后一天的比例值大于 0.6
   - 每种指标最多三项，最小一项

最终，我们选取了以下八个指标：’1,1,1’, ’4,1,1’, ’4,1,3’, ’4,1,4’, ’13,1,6’, ’13,1,9’, ’13,1,7’, ’1,1,3’，其比例值趋势图如图1。若没有特殊说明，下文将用这八个指标做分析。

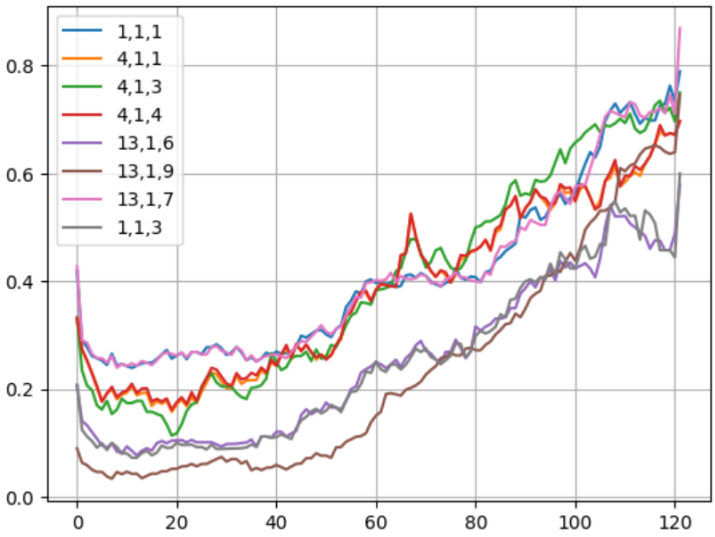

图 1: 比例值曲线

5.3 问题二：用户的流失信息描述

5.3.1 问题分析

本题目要求，利用数据建模刻画用户画像，对用户的流失风险分级。我们**先提取用户数据特征，降维数据，再聚类分析**。

5.3.2 特征提取

根据观察，每个用户拥有的指标数目不同，数据记录的时间不同，同一种指标不同用户的数据大小也有数量级的差别。在提取特征时，我们主要提取“均值”、“比例”等指标，尽量缩小因为记录数据的完整性不同带来的差异。

我们将对每个用户提取以下三个指标：

- **平均流量**：即用户的全部 8 个指标在记录时间内的均值。平均流量越高，说明用户对云服务平台使用量越大、依赖度越高，流失风险越低。
- **下降比例**：第 i 个用户的比例值 α_i 计算公式如下：

$$\alpha_i = \sum_{j}^{8} \frac{x'_{ij} - x''_{ij}}{t_{ij} N_i mean_{ij}} \tag{7}$$

其中，x'_{ij}, x''_{ij} 分别为第 i 个用户的指标 j 最初三天与最末三天的均值，t_{ij} 为第 i 个用户的指标 j 总共记录的时间天数，N_i 第 i 个用户拥有的总指标数，$mean_{ij}$ 为第

i 个用户指标 j 的均值。下降比例值表现了用户单位时间内每个指标的平均下降的幅度，下降比例越大，流失风险越高。

- **大幅下降值**：第 i 个用户的大幅下降值 α_i 计算公式如下：

$$r_i = \sum_{j}^{8} \frac{\sum_k drop_{ijk}}{N_i mean_{ij}} \tag{8}$$

其中，$\sum_k drop_{ijk}$ 为第 i 个用户的指标 j 所有大幅下降点的下降值的加和，大幅下降的判断标准为：下降幅度大于该用户该指标所有下降时间点的下降幅度的均值。根据附件 2，用户的阶段流失预警时间点是根据用户监控指标大幅下降的时间点所求得的，因此有充足理由相信，大幅下降值对用户流失风险有一定影响。

5.3.3 聚类分析

将上述三个指标标准化，并用 **K-Medians** 方法聚类。因为特征值中有少量离群点，而 K-Medians 是用数据集的中位数而不是均值来计算数据的中心点，受离群点影响小。我们将用户分成四类：高风险流失用户，中高风险流失用户，中低风险流失用户，低风险流失用户，分类结果如图2所示。下降比例与大幅下降值均与用户流失风险成正比，可视化过程中，将其并为一维。

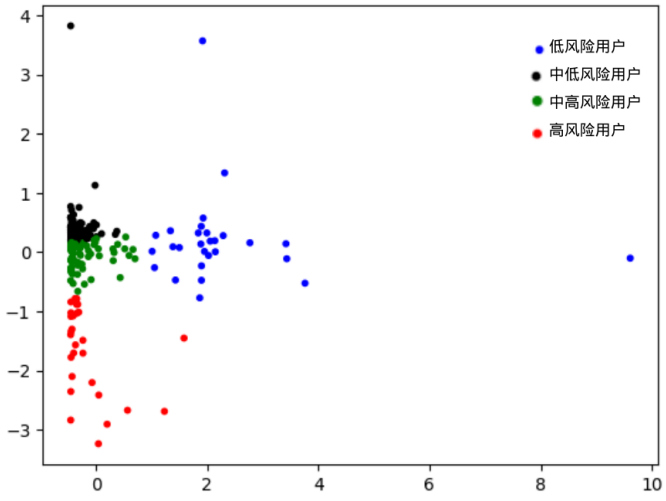

图 2: 流失风险分类图

5.4 问题三：用户的流失预测的网络模型

5.4.1 方法一：动态 PCA 与 SVM

1. **预处理**

由问题一，我们可以得到所有用户的八个指标随时间变化的特征数据。由于不同

用户拥有的数据长度不一，现以所有用户中所有指标的右边界值作为时序数据的时间右边界值，所有用户的所有指标均往前取 384 天，即 128 个数据单元，作为模型的输入；其中，从实际意义的角度和机器学习的角度，若某用户某指标的数据长度不足 128，则在数据最后填充右边界值，直至长度为 128。此外，为解决不同特征数据绝对大小不一而影响指标权重的问题，将各数据归一化。最后得到 432*8*128 的三维数组，三个维度分别代表用户、指标类型、时间点。

2. **动态 PCA**

预处理后得到的数据较长、特征较多，作为特征易导致过拟合的情况出现，因此考虑降维。

我们采用动态主成分分析算法提取数据特征:

(a) 将矩阵转为 432*1024 的矩阵，其中每 8 列为 8 个指标在同一时间点下的值，共 128 个时间点，即 8*128=1024 列。

(b) 将该新矩阵带入主成分分析算法，分析得到前四列特征向量的方差贡献率，分别为 0.547、0.272、0.130、0.021，选择前三列（累计方差贡献率为 94%）作为主成分。

具体实现是，

3. **SVM 分类**

由动态 PCA，得到 432*3 的特征矩阵，样本较小，而标签只有 0（未流失）与 1（流失）两项，属于二分类问题，考虑采用支持向量机（Support Vector Machine）求解。

根据实际情况，我们将数据集随机划分为训练集与测试集，选取正则化参数 C=2，RBF 核参数 gamma=0.001，得到训练集准确率为 78.4%，测试集准确率为 71.3%，略微过拟合。

5.4.2 方法二：RLC 网络——残差长短期记忆卷积预测模型

根据上述筛选出来的 8 种指标，我们延长其时间方向长度为 128 个时间单位，即每个用户有 8×128 个数据描述。由于数据量大、时间序列关系强，存在复杂的隐式时序特征，我们使用长短期记忆网络和时序卷积神经网络构造预测器。

网络结构如下表所示：

Layer	Channels In	Channels Out	Kernel	Stride	Padding	Normalize	Dropout	Activation
LSTM	8	8	None	None	None	Layer Norm	0.5	tanh
Conv1	8	128	4	2	1	None	0.5	leakyReLU
Conv2	128	512	4	2	1	None	None	leakyReLU
Conv3	512	512	4	2	1	None	None	leakyReLU
Conv4	512	64	4	2	1	None	None	leakyReLU
FC1	512	64	None	None	None	Layer Norm	None	Sigmoid
FC2	64	2	None	None	None	None	None	Softmax

LSTM 用于获取网络的隐性时序信号，并与输入原始信号的显性时序信号相加，即构成残差网络。残差网络不仅可以提高后续网络层的训练速度，而且可以融合隐性时序信号和显性时序信号，提高网络的解析能力。

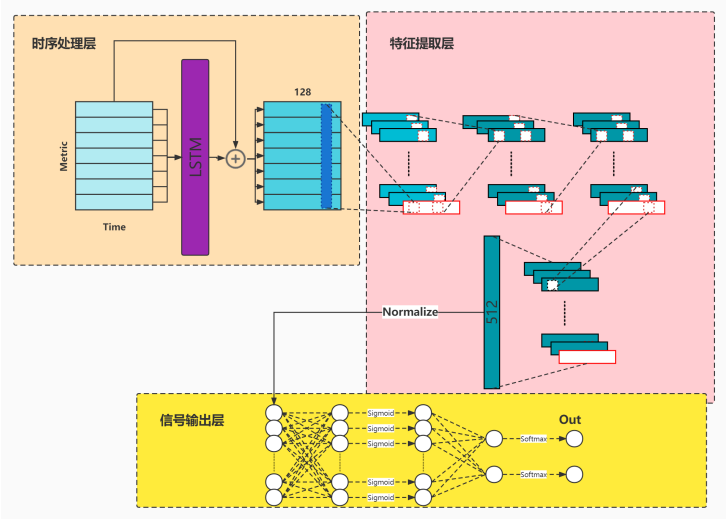

图 3: 网络结构示意图

卷积神经网络将时间维度的信息转化为特征维度的信息: 前三个卷积网络将时间维度的数据压缩到 16 个时间单位，但是在特征维度上有 512 个特征。最后一个卷积网络需要将数据简化压缩，即 **NeckLayer**，输出 64×8 的数字特征。

在全连接层，把上述特征融合成二分类的二维特征。第一个全连接层初步融合了特征，并用 Sigmoid 函数激活，避免梯度消失，提高训练速度；第二个全连接层输出二维向量，经 Softmax 激活后给出预测结果。

选取训练 Batch_size = 8, 优化器为 Adam, 损失函数为交叉熵损失，得到训练模型。

训练过程中训练集损失与测试集损失几乎同步变化，过拟合现象不严重。

方法	精确率	召回率	F1
DPCA+SVM	0.765	0.643	0.699
RLC	0.989	0.984	0.986

上表中记录了两种方法训练的模型的结果。神经网络的拟合能力明显高于 DPCA+SVM 方法。

六、 模型评价

6.1 模型的优点

1. 使用滑窗滤波平滑数据和横向对比法筛选数据。滑窗滤波可以减少局部噪声对模型性能的影响；通过横向对比可以得到数据充足、波动幅度大、具有广泛代表性

的数据指标，提高模型性能。

2. 利用特征工程将长时序信号降维至低维空间。我们根据先验知识，将平均流量、下降比例和大幅下降值从时序信号中提取出来，以此数学地表征用户的流失风险。

3. 使用残差长短期记忆卷积神经网络解析数据特征。残差网络可以提高模型的训练效率；长短期记忆可以提取隐性时序信息；卷积神经网络可以时移不变地将局部时域信息提取为特征。

6.2 模型的缺点

数据量小。题给数据量大约只包含 400 多个用户，50 多个指标，每个用户在每个指标上的数据长度不一。或可采用数据增强的办法提高数据量。