doi:10.3772/j.issn.1002-0470.2021.03.006

基于特征加权 ML-kNN 的网页浏览业务 KQI 预测①

谢 苏② 刘子巍 李 克③

(北京联合大学智慧城市学院 北京 100101)

摘 要 传统以网络为中心的移动网络运维往往是在接到用户投诉时才采取相应补救措施,随着移动互联网(OTT)业务的高速发展,这一问题愈发突出。如何在监测用户业务感知的基础上对用户业务质量进行预测预警并及时干预,是提高移动业务保障能力和网络运维智能化水平的重要手段。本文利用从普通用户终端上采集的海量业务感知数据,重点针对网页浏览业务,研究了 ML-ReliefF 算法在业务感知采样数据降维中的应用。在此基础上,将特征选择结果与多标记 k 近邻(kNN)算法相结合,提出了基于特征加权的多标记 k 近邻算法应用于业务关键质量指标(KQI)预测。实验结果表明,该方法可有效提高 KQI 预测质量。

关键词 特征选择;智能网络运维(AIOps);关键质量指标(KQI);k 近邻(kNN);移动互联网(OTT);移动众包感知(MCS)

0 引 言

随着长期演进(long term evolution, LTE)网络成熟商用以及智能终端的普及,移动互联网(over-the-top,OTT)业务得到了高速发展并逐步成为移动网络承载的业务主体。一方面,电信运营商的传统业务包括语音、短彩信、增值业务等被大量替代,另一方面,新兴业务打破了移动网络的边界,促进了移动网和互联网的融合,业务性能和体验不再只是由移动网本身质量所决定。

移动互联网业务高速发展,传统以网络为中心的网络运维模式难以为继,在接到用户投诉后再去解决的传统方式难以应对快速多变的网络和业务环境。亟待采取前置手段,在用户体验劣化之前发现业务质量问题并指导运维人员采取手段优化提升,保障用户业务体验和黏性。这对提升网络运维的智能化水平有重要意义。Gartner 将这种基于海量网络数据和机器学习进行网络运维的新范式定义为智能网络运维(AI for IO operations,AIOps)[1]。

随着智能终端中大量传感器的应用,终端已逐步替代传统路测,成为采集用户业务感知的重要节点,为开展智能网络运维提供了数据基础。基于海量业务感知数据,采取适当的机器学习算法,是提升移动网络运维的智能化水平的必经之路。

本文针对网页浏览业务,尝试基于终端侧众包采集的业务感知数据,利用改进的特征选择和多标记学习算法开展业务关键质量指标(key quality indicator,KQI)预测。

1 相关工作

下面对终端侧业务感知数据的采集和评价、相关机器学习算法做简要介绍。

1.1 移动众包感知与 OTT 业务感知监测

随着智能终端以及可穿戴设备的兴起,基于众

① 国家自然科学基金(61841601,61972040)和北京联合大学人才强校优选计划(BPHR2018CZ05)资助项目。
② 女,1992 年生,硕士生;研究方向:移动网络智能运维技术;E-mail:13661365299@163.com
③ 通信作者,E-mail: like@buu.edu.cn
(收稿日期:2020-03-24)

包的终端侧网络测量得到了学术和产业界的重视，因其采集点更贴近用户因而能获取更真实的业务感知信息，为大数据在电信业务运营中的应用提供了一种崭新的视角[2-3]。文献[3]中将这类数据采集和分析范式命名为移动众包感知(mobile crowdsensing，MCS)。MCS 可应用于很多领域，包括移动网络运维[4-5]、城市交通监测和日常生活[6]等。

本文所涉及的 OTT 业务感知数据采集过程包括：用户终端上安装的采集前端以后台运行方式监测用户业务行为，采集业务感知信息并回传云端进行后续分析。为保护用户隐私，不采集包括手机号码、短信文本等在内的用户敏感信息，所采集数据对终端标识仅用于区分用户，并进行脱敏处理。

1.2 网页浏览业务感知 KQI

对于 OTT 浏览业务，通常根据业务特征以及与用户体验的相关性定义若干 KQI 指标[7]，包括以下两点。

(1)首包时延(D_k)：用户发起网页请求到收到服务器第一个 HTTP 200 OK 报文包之间的时长，即：

$$D_k = T_{200} - T_{req} \tag{1}$$

其中，T_{req} 为用户发起网页浏览请求的时间点，T_{200} 为收到第一个 HTTP 200 OK 报文的时间点。

一次网页浏览过程包括若干环节，即 DNS 解析、TCP 连接建立和 HTTP 交互，因此首包时延可分解为 3 个分段时延指标之和，即：

$$\begin{cases} D_{dns} = T_{dns} - T_{req} \\ D_{tcp} = T_{tcp} - T_{dns} \\ D_{get} = T_{200} - T_{tcp} \end{cases} \tag{2}$$

(2)页面打开时延(D_p)：指用户发起浏览请求到整个 HTTP 页面下载完毕并渲染完成的时长。页面时延是在首包时延基础上增加接收响应时延，即：

$$D_p = D_k + D_{res} \tag{3}$$

其中，$D_{res} = T_{res} - T_{200}$ 为接收响应时延，指从收到第一个响应到终端发出[FIN，ACK]的时间差。

为了进行业务质量的问题定位，除上述信息之外，通常还会同步采集网络环境信息和定位信息等。

1.3 过滤式特征选择

特征选择是从原始特征集中按照某一种评价准则选择出一组具有良好区分特性的特征子集。这样一方面可以避免特征数量较多的应用场景中存在的“维数灾难”问题，还可以通过剔除冗余或不相关特征提高机器学习算法的性能。

其中过滤式选择最具代表性的是 Relief 算法。Kononenko[8]在此基础上提出了 ReliefF 算法，使其可以解决多分类和回归问题。有人进一步将其推广到多标记分类应用场景[9-11]。文献[9]提出了一种多标记 ReliefF(ML-ReliefF)算法。该算法基于各标记的共现性假设以及每个类标记对样本贡献值相等的假设，改进了特征权值更新公式。

特征权重表征了特征对标记的影响程度。算法任务就是找到理想的权重向量，$\boldsymbol{W}[A] = 1 \sim p$，$p$ 为总特征数。子集重要性由子集中各特征权重之和决定。

ML-ReliefF 算法的伪码描述如算法 1 所示。

算法 1 ML-Relief F 算法

W = ML-ReliefF(D, m, k)

输入：训练数据集 $D = \{X_1 \sim X_m\}$，迭代次数 m，近邻数 k。

输出：预测的特征权值向量 $\boldsymbol{W}$。

1. 初始化特征权值向量 $\boldsymbol{W}[A] = 0.0$，$A = 1, 2, \cdots, p$
2. for $i = 1$ to m do
3. 从 D 中随机选一个样本 R_i，class(R_i) = (h_1，h_2，$\cdots$，h_t)
4. for $t = 1$ to T do
5. 寻找与 $\boldsymbol{R}_i$ 的标记 h_t 相同的 k 个近邻 $\boldsymbol{H}_j(\boldsymbol{R}_i)$，$j = 1 \sim k$
6. 对标记 $C \neq h_t$，找出与 $\boldsymbol{R}_i$ 标记不同的 k 个近邻 $\boldsymbol{M}_j(C)$
7. for A = 1 to p do
8. 按照式(4)更新各特征权值：

$$\boldsymbol{W}[A] := \boldsymbol{W}[A] + \frac{w_i}{mk}\left\{ -\sum_{j=1}^{k} diff(\boldsymbol{A}, \boldsymbol{R}_i, \boldsymbol{H}_j) + \sum_{C \neq h_t}\left[\frac{P(C)}{1 - P(h_t)}\sum_{j=1}^{k} diff(\boldsymbol{A}, \boldsymbol{R}_i, \boldsymbol{M}_j(C))\right]\right\} \tag{4}$$

9. end for
10. end for
11. end for

1.4 ML-kNN 算法

文献[12，13]在 k 近邻(k-nearest neighbor，kNN)

的基础上提出多标记 k 近邻算法(multi-label k-nearest neighbor, ML-kNN)以解决多标记分类问题。它是将 kNN 与贝叶斯算法相结合而构造的分类器,可对多标记数据进行有效分类。对每个测试样本,在训练集中找到它的 k 个近邻。然后,基于近邻样本的统计信息和最大后验概率(maximum A posteriori, MAP)原则计算测试样本的标记集合。主要步骤如下。

首先,在训练样本集 $\{x_i, i = 1 \sim m\}$ 中寻找各样本矢量的 k 近邻并构造其 k 近邻样本集 $N(x_i)$。对各标记项 $y_j, j = 1 \sim q$,计算先验概率 $\mathbb{P}(H_j)$ 和 $\mathbb{P}(\overline{H_J})$。$H_j$ 和 $H_{\bar{j}}$ 表示未知样本 x 具有和不具有该标记。

然后计算归一化频数矩阵,即训练样本的近邻中具有和不具有 y_j 的样本数。对未知样本 x 同样构造其近邻集 $N(x)$,并计算近邻中同标记样本数 $\{C_j\}$。

最后在计算 x 的似然概率 $\mathbb{P}(C_j \mid H_j)$ 和 $\mathbb{P}(C_j \mid \tilde{H}_j)$ 的基础上,根据式(7)计算出其标记估计结果。

$$y = \left[\frac{\mathbb{P}(H_j)\,\mathbb{P}(C_1 \mid H_1)}{\mathbb{P}(\tilde{H}_1)\,\mathbb{P}(C_1 \mid \tilde{H}_1)} > 1\right] \tag{5}$$

需要注意的是,ML-kNN 在特征空间中近邻搜索时并没有考虑各特征权重的影响。

2　基于特征加权 ML-kNN 的 KQI 预测

本节研究将 ML-kNN 算法引入网页浏览业务的 KQI 指标预测,并结合特征选择对算法进行改进,提出了一种基于特征加权 ML-kNN 的 KQI 预测算法。

用户在智能终端上使用业务时,其业务体验的优劣往往受到网络环境、业务平台和终端质量等各环节因素的影响。当用户处于特定环境下,就可以基于其当前所处的网络和业务环境信息对其所用请求的业务的质量做出预测,这是典型的分类或回归问题。为了简化起见,这里仅对其网页浏览业务的关键 KQI 指标做出优劣的预测,即多标记二分类问题。

ML-kNN 是多标记分类算法中性能较优的一类算法,但算法中近邻搜索时,样本在特征空间中的距离计算采用的是各特征项等权重的欧氏距离方法。而以 ReliefF 为代表的过滤式特征选择虽然是与后续的学习任务分离的,但因为特征选择结果包含的权重大小反映了各特征项对标记的影响程度,因此可根据各特征项的权重来施加不同影响以搜索最有效近邻。考虑到 KQI 预测所涉及特征和标记项都较少,因此这里特征选择的目的并非减少参与学习的特征数,而是利用特征权重提高 KQI 预测算法的性能。

2.1　数据预处理

首先需要针对业务感知样本的特征进行预处理,包括特征值的归一化、标记项的转换等。

(1)初始特征和标记空间的构造

在移动网络(本文以 LTE 网络为例)下,当用户在智能终端上使用网页浏览类业务 App 浏览特定网页时,终端上的采集前端所采集的业务感知样本就构成了初始的众包业务感知数据集,主要包含以下字段:日期,时间,大区编号(TAC),小区编号(CellID),经度,纬度,场强(RSRP),信号质量(RSRQ),网站名称,网站 IP,DNS IP,终端标识,终端型号,D_{dns}, D_{tcp}, D_{req}, D_{res}。

选择前 13 个字段作为初始特征集 $\bar{\boldsymbol{x}} = \{\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_d\}$, $d = 13$。其中字段{日期,时间,经度,纬度,场强,信号质量}为数值型数据,其余字段为名目型数据。将 KQI 指标字段 $\{D_{dns}, D_{tcp}, D_{req}, D_{res}\}$ 作为初始标记集,$\bar{\boldsymbol{Y}} = \{\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_d\}$, $d = 4$。

(2)特征值和标记值转换

对初始数据集中各数值型特征进行归一化,即:

$$x_i = \frac{g(\bar{x}_i) - LB(\bar{x}_i)}{UB(\bar{x}_i) - LB(\bar{x}_i)} \quad 0 \leqslant i \leqslant m \tag{6}$$

其中 $\bar{x}_i$ 表示特征 i 的原始值。$g(\bar{x}_i)$ 为相对于上下边界 $UB(\bar{x}_i)$ 和 $LB(\bar{x}_i)$ 的截断函数,即

$$g(\bar{x}_i) = \begin{cases} UB(\bar{x}_i) & \bar{x}_i > UB(\bar{x}_i) \\ LB(\bar{x}_i) & \bar{x}_i < LB(\bar{x}_i) \\ \bar{x}_i & \text{其他} \end{cases} \tag{7}$$

在实际数据集中由于采样误差和终端个体差异等原因会导致过小或过大采样值的存在。为避免该因素对归一化的影响,这里并不直接采用最小和最大值,而采用其箱形图分布的下外限和上外限,即:

$$\begin{cases} LB(\bar{x}_i) = \max(Q1(\bar{x}_i) - 3IQR(\bar{x}_i), \min(\bar{x}_i)) \\ UB(\bar{x}_i) = \min(Q3(\bar{x}_i) + 3IQR(\bar{x}_i), \max(\bar{x}_i)) \end{cases} \tag{8}$$

对于原始训练样本中的各数值型标记字段，根据预设门限$\{T_1 \sim T_q\}$按下式转换成布尔型，即：

$$y_j = [\bar{y}_J < T_j] \quad 1 \leqslant j \leqslant q \tag{9}$$

其中$[c]$表示当条件 c 成立时返回 1，否则返回 0。由此可得到实验数据集 $D = \{(x_i, Y_i) \mid 1 \leqslant i \leqslant m\}$。

2.2 ML-ReliefF 的应用和改进

分析 ML-ReliefF 算法发现，贡献值参数的设计初衷是对多标记数据给予重视，又不能把多标记数据的权重设得过大。算法给出了 3 种权重取值，即 1 范权重法、单位权重法和 2 范权重法[9]。但因该参数对所有迭代样本是固定常数项，归一化后的结果是等价的，对最终结果没有实际意义，因此在本算法中将该参数去除。此外，考虑到本文应用场景是多标记二分类，且为了简便起见取 $k=1$，则式(4)改为

$$\boldsymbol{W}[A] := \boldsymbol{W}[A] + \frac{1}{m}\{-diff(\boldsymbol{A}, \boldsymbol{R}_i, \boldsymbol{H}_j) + diff(\boldsymbol{A}, \boldsymbol{R}_i, \boldsymbol{M}_j)\} \tag{10}$$

其中 m 是对迭代样本数做归一化的因子。

此外，在迭代特征权重过程中，各权重反映了其在近邻搜索时更好地表征样本间的相似度，因此可在每次迭代时采用当前特征权重对近邻搜索的距离进行加权以提高迭代收敛速度。具体地，可在寻找样本的同标记和异标记最近邻 $\boldsymbol{H}_j$ 和 $\boldsymbol{M}_j$ 时，将欧氏距离修改为归一化加权欧氏距离。此外，需将各权重初始值设为 1，以避免 0 初始值在迭代时导致的无法归一化问题，最终结果中再减掉初始值即可。

2.3 算法描述

KQI 预测算法的具体过程如下。

输入：原始训练样本集 $\bar{D}$；原始未知样本 $\bar{x}_0$。

(1)数据预处理得到 D 和 x_0。

(2)采用 ML-ReliefF 获得各特征权重 $\boldsymbol{W}[A]$，并选择前 $N(N \leqslant p)$ 个最大权重对应的特征子集。

(3)采用 ML-kNN 分类，其中训练样本和未知样本的 k 近邻搜索时均采用归一化特征加权欧氏距离。

(4)对未知样本进行分类，得到其标记集 Y_0 预测值。

3 实验结果与分析

3.1 实验数据集

实验所用数据为上海 LTE 网络下采集，所涉及的目标网址包括新浪、搜狐、微博、淘宝等 10 个主流网站。因为原始采集数据的标记字段 $\{D_{dns}, D_{tcp}, D_{req}, D_{res}\}$ 均为实数，需按照设定门限转换为正负标记。考虑到实际采集样本中正例样本占比往往较高的情况，为了减少样本不平衡对预测性能的影响，判决门限的设定采用 Q3 箱形图法，将正负例的比例控制在 3:1 附近。此外，为避免门限选取受局部数据分布特征的影响，在全样本空间内取各标记字段的 Q3 作为判决门限，超过门限记为负例，即为{55, 57, 358, 459}(ms)。

经过清洗，包括剔除 Q3 + 1.5IQR 和 Q1-1.5IQR 之外的异常值以及字段值溢出或缺失的样本后，根据经纬度保留中心城区区域内样本共计 12.6 万条作为实验数据集。

3.2 性能评价准则

多标记学习算法的主要评价指标可以分为两类，即基于预测的指标簇和基于排序的指标簇[14]。前者主要评价预测结果的正确性，包括 Accuracy、F1-measure、Hamming Loss 等，而后者基于评分函数评价标记的排序质量，包括 One-error、Coverage、Ranking Loss 等。一般来说，基于预测的指标应用更为普遍。考虑到本实验中标记项仅 4 个，基于排序的指标簇的指示意义不大，因此采用上述基于预测的指标簇。

本实验采用 Python 语言在 Pycharm 平台完成。

3.3 特征选择实验分析

首先采用简化后的 ML-ReliefF 对业务感知样本进行特征选择，结果如图 1(不加权迭代)所示。

图 2 为选取不同迭代次数 m 时的各特征项归一化权重的收敛趋势，可见大部分的特征项权重在 1000 次左右基本收敛。为了对比，本文按第 2.2 节

所述在每次迭代时用瞬时特征权重进行加权距离的近邻搜索并同步更新特征权重(见图 1)。可见,加权迭代与原算法结果基本相同。

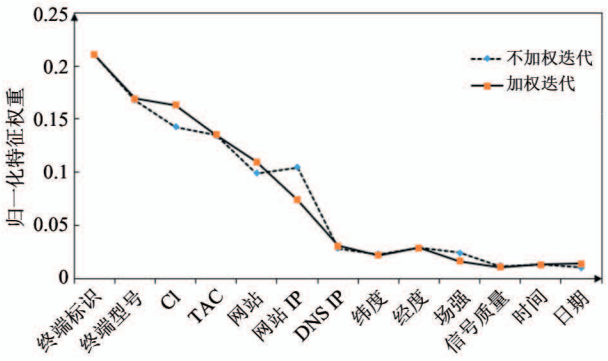

图 1　ML-ReliefF 实验的归一化特征权值结果

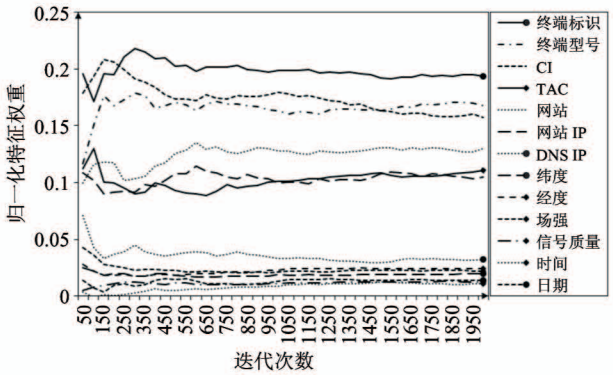

图 2　ML-ReliefF 特征权重收敛趋势

3.4　KQI 预测实验分析

这里采用 10×10 折交叉验证进行样本集分割。

(1) 特征选择对非特征加权 KQI 预测的性能影响

首先考察不同的特征选择结果对 KQI 预测性能的影响。分别保留特征权重最大的前 1、4、7、10、13 个特征参与 KQI 预测,近邻搜索时的近邻数 k 分别取 5、10、15 并保留最佳结果即 $k=15$,表 1 为实验结果。可见,不同程度的保留特征数对性能有较明显的影响,其中特征数为 7 或 10 时性能最佳,太多或太少特征参与预测性能均不佳。

(2) 特征加权 KQI 预测的性能

进一步,当采用特征选择结果进行加权 KQI 预测时,结果如表 2 所示。其中全部特征均参与预测时的性能最优,因而表明预测结果的错误率和方差均下降,结果更稳定。

表 1　KQI 预测实验结果(特征选择,不加权,$k=15$)

特征数	Hamming Loss	Macro-F1	Micro-F1	Instance-F1
1	0.316±0.011	0.785±0.011	0.802±0.008	0.778±0.010
4	0.197±0.009	0.863±0.009	0.866±0.007	0.849±0.009
7	0.193±0.007	0.862±0.005	0.868±0.005	0.853±0.006
10	0.192±0.012	0.863±0.007	0.868±0.009	0.852±0.009
13	0.202±0.009	0.859±0.006	0.864±0.006	0.847±0.005

表 2　特征加权 KQI 预测的实验结果($k=15$)

特征数	Hamming Loss	Macro-F1	Micro-F1	Instance-F1
1	0.309±0.017	0.809±0.013	0.816±0.012	0.792±0.013
4	0.194±0.010	0.862±0.006	0.867±0.006	0.852±0.009
7	0.190±0.004	0.865±0.004	0.870±0.003	0.855±0.004
10	0.187±0.011	0.866±0.009	0.871±0.007	0.856±0.008
13	0.186±0.005	0.871±0.005	0.874±0.004	0.859±0.004

图 3 对不同特征个数情况下不加权和加权 KQI 预测结果的趋势进行了对比(仅对比 Hamming Loss,其他指标趋势相同)。由图 3 可知,特征加权 KQI 预测由于较好地利用了每个特征项,各特征都按照自己的权重参与了预测并做出相应的贡献,因此参与预测的特征项越多性能越好,而不加权情况

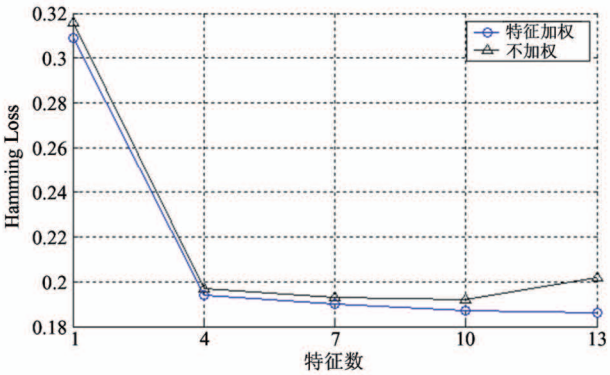

图 3 KQI 算法 Hamming Loss 性能对比

下则需要谨慎选择实际使用的特征数。

4 结 论

本文从目前移动网络运维中对于用户业务体验问题往往不能及时发现和干预、主要依赖于人工巡检和客户投诉的现状入手,研究了将特征选择和多标记学习用于业务质量预测的问题。基于众包业务感知数据,以网页浏览 OTT 业务为对象,研究了 ML-ReliefF 算法在业务感知数据的特征选择上的应用,并将特征选择结果与 ML-kNN 算法相结合,提出了基于特征加权的 ML-kNN 算法用于业务 KQI 预测。实验结果验证了算法在对特定场景下用户的业务体验质量预测应用上的有效性。

本文的主要创新在于:(1)将多标记特征选择引入众包业务感知数据的处理,有效降低了非关键特征对 KQI 预测的影响。(2)与传统过滤式特征选择仅作为后续学习的预处理不同,本文将特征选择结果嵌入后续的学习过程,充分发挥所有特征的信息,提升了预测的准确率。

下一步,将进一步分析标记项间相关性对特征选择的影响并对特征选择算法进行改进以提升业务质量预测的效果。此外,不平衡样本集尤其是正例占优情况下,负例预测的效果还有较大提升空间,这也是下一步的研究重点。

参考文献

[1] Siegfried G. Beginning AIOps: data science for IT operations[EB/OL]. https://www.gartner.com/doc/3893177?ref=mrktg-srch:Gartner, 2018

[2] Naboulsi D, Fiore M, Ribot S, et al. Large-scale mobile traffic analysis: a survey[J]. *IEEE Communications Surveys and Tutorials*, 2016, 18(1): 124-161

[3] Ganti R K, Ye F, Lei H. Mobile crowd sensing: current state and future challenges[J]. *IEEE Communications Magazine*, 2011, 49(11):32-39

[4] Casas P, Seufert M, Wamser F, et al. Next to you: monitoring quality of experience in cellular networks from the end-devices[J]. *IEEE Transactions on Network and Service Management*, 2016, 13(2): 181-196

[5] Li K, Wang H, Xu X L, et al. A crowdsensing based analytical framework for perceptional degradation of OTT web browsing[J]. *Sensors*, 2018, 18(5): 1566

[6] Bulut M, Demirbas M, Ferhatosmanoglu H. LineKing: coffee shop wait-time monitoring using smart phones[J]. *IEEE Transactions on Mobile Computing*, 2015, 14(10): 2045-2058

[7] 中国电信. 中国电信移动互联网业务感知测试 APP 功能规范[S]. 北京:中国电信,2015

[8] Kononenko I. Estimating attributes: analysis and extension of relief[C]// European Conference on Machine Learning, Berlin, Germany, 1994:171-182

[9] 黄莉莉, 汤进, 孙登第, 等. 基于多标签 ReliefF 的特征选择算法[J]. 计算机应用,2012, 32(10):2888-2890

[10] Huang Z, Yang C, Zhou X, et al. A hybrid feature selection method based on binary state transition algorithm and ReliefF[J]. *IEEE Journal of Biomedical and Health Informatics*, 2018, 23:1888-1898

[11] Spolaor N, Cherman E A, Monard M C, et al. ReliefF for multi-label feature selection[C]//2013 IEEE Brazilian Conference on Intelligent Systems, Fortaleza, Brazil, 2013:6-11

[12] Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning[J]. *Pattern Recognition*, 2007, 40(7):2038-2048

[13] Zhang M L, Zhou Z H. A review on multi-label learning algorithms[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2014, 26(8):1819-1837

[14] Wu X Z, Zhou Z H. A unified view of multi-label performance measures[C]// The 34th International Conference on Machine Learning, Sydney, Australia, 2017: 3780-3788

KQI prediction for web browsing based on feature weighted ML-kNN

Xie Su, Liu Ziwei, Li Ke

(College of Smart City, Beijing Union University, Beijing 100101)

Abstract

Traditional network-centric mobile network operation often takes corresponding remedial measures when receiving user complaints about service quality. With the rapid development of over-the-top (OTT) services, this problem has become increasingly prominent. How to predict and warn the user's service quality and timely intervene based on the service perception monitoring is an important means to improve the intelligence of network operation. In this paper, the service perception data crowdsensed from massive user terminals are utilized, focusing on the web browsing service, and the ML-ReliefF algorithm in the dimension reduction of service perception data is applied. On this basis, combined with the feature selection results with the multi-label k-nearest neighbor (ML-kNN) algorithm, a feature weighted key quality indicator (ML-kNN for KQI) prediction is proposed. Experimental results show that this method can effectively improve the quality of key quality indicator (KQI) prediction.

Key words: feature selection, AI for IT operations (AIOps), key quality indicator (KQI), k-nearest neighbor (kNN), over-the-top (OTT), mobile crowdsensing (MCS)