文章编号:0253-2778(2020)01-0029-10

基于模糊度的半监督自步协同下的微信流业务识别

刘玮康,秦晓卫,卫 国

(中国科学院无线光电通信重点实验室,中国科学技术大学,安徽合肥 230026)

摘要:网络数据流的精准业务识别是实现差异化服务的先决条件,常用的监督学习在构建训练数据集时因需要大量人力标注因而难以实施,基于少量标注数据的半监督学习成为研究的热点之一.自步协同训练(self-paced co-training)的半监督框架在处理未标记数据时采用了从易到难、多视角协同的方法,但该方法仅以置信度为选取依据给样本标记伪标签,容易导致多视角的差异性在训练过程中逐步下降,从而引起协同增益下降、模型性能受限等问题.为此面向微信数据流识别问题,提出了一种基于模糊度的自步协同训练模型(fuzziness based self-paced co-training, FBSpaCo),在标注伪标签时进一步引入模糊度评估机制.实验表明,该模型在保证置信度的前提下有效地避免了训练过程中两视角差异性下降,较已有方法较大地提升了识别准确度.

关键词:数据流识别;半监督学习;自步协同训练;模糊度

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2020.01.004

引用格式:刘玮康,秦晓卫,卫国. 基于模糊度的半监督自步协同下的微信流业务识别[J]. 中国科学技术大学学报,2020,50(1):29-38.

LIU Weikang, QIN Xiaowei, WEI Guo. Service identification of WeChat traffic based on fuzziness and semi-supervised self-paced co-training[J]. Journal of University of Science and Technology of China, 2020,50(1):29-38.

Service identification of WeChat traffic based on fuzziness and semi-supervised self-paced co-training

LIU Weikang, QIN Xiaowei, WEI Guo

(*CAS Key Laboratory of Wireless-Optical Communications, University of Science and Technology of China, Hefei* 230026, *China*)

Abstract: Accurate service identification of network data streams is a prerequisite for providing differentiated services. The commonly used supervised learning is difficult to implement when constructing training data sets due to the need for a large number of human annotations. Semi-supervised learning based on a small amount of annotated data has become one of the research hotspots. Semi-supervised framework of Self-paced Co-training adopts the method of collaboration that processes the easier pieces first using multiple perspectives when dealing with unlabeled data. However, this method only uses confidence as the criterion to select pseudo labels for samples, which can easily lead to the gradual decline of multi-perspective differences in the training process, resulting in the decline of synergy gain and the limitation of model performance. Therefore, for the recognition of WeChat data streams, a self-paced co-training model

收稿日期:2019-03-28;**修回日期**:2019-07-17

基金项目:国家重点研发计划(2018YFA0701603)资助.

作者简介:刘玮康,男,1994年生,硕士生.研究方向:用户行为感知、机器学习.E-mail:1210301638@qq.com

通讯作者:秦晓卫,博士/副教授,E-mail:qinxw@ustc.edu.cn

based on fuzziness (FBSpaCo) is proposed. When labeling pseudo labels, the fuzziness evaluation mechanism is introduced. Experiments show that the model can effectively avoid the decline of the difference between two perspectives in the training process. Compared with the existing methods, the recognition accuracy is greatly improved.

Key words: network data identification; semi-supervised learning; self-paced co-training; fuzziness

0 引言

移动互联网的迅速发展促使各类业务爆炸式增长,为用户提供优质的服务体验以增加用户黏度是移动运营商关注的核心点,深入分析各类业务的行为、了解用户习惯,为各类业务提供差异化服务成为网络智能运维的重点.其中,网络数据流的精准业务识别作为差异化服务的先决条件,逐渐成为业界和学术研究的热点.以微信数据流为代表的社交网络数据流,在移动网络中占有很大比重,由于其混合类业务特性,如微信包含了文字、图片、短视频、小游戏等业务,获得相关研究者的广泛关注.

传统网络侧数据流的业务识别大多基于协议规则,如基于端口号[1]或基于深度包解析(deep package inspection, DPI)[2].考虑到保护个人隐私,服务提供商逐渐对其应用业务数据采取加密处理,传统的基于协议规则的方法很多情况下不再适用,取而代之的是基于机器学习或深度学习解决网络数据流业务识别的主流方法[3],其中代表性的方法整体上可归纳两大类:有监督与半监督.有监督模式通常是基于大量有标签的训练数据集采用机器学习或深度学习模型,如朴素贝叶斯[4]、贝叶斯神经网络[5]、C4.5决策树[6]、多层自编码器(autoencoder)与卷积神经网络(convolutional neural network, CNN)[7]等.由于有监督模型需要大量具有人工标注的训练数据集,因人力成本过高导致实际实施中困难重重.为了减少人力成本,基于少量标签数据的半监督模式在数据流识别中受到了广泛的欢迎和研究,其中代表性的方法包括有约束的聚类算法[8]、概率图模型的随机漫步算法[9-10]等,但是这些方法较为简单,提取的信息不足,往往学习数据在特征空间的相似性,难以深层挖掘不同业务数据的关联,因此不适用于复杂的微信数据业务识别场景.2017年,Ma等[11]提出了基于自步学习的协同训练模型,相对于传统的半监督方法,模型从简单的有标签数据开始,逐步通过标注伪标签的方式,扩大训练数据集并构建复杂的半监督模型.该方法由于在标记伪标签时默认损失函数最小(即置信度)的原则,导致模型在迭代训练过程中两视角的差异性越来越小,无法保障多视角协同增益原则,即在保证置信度的前提下各视角提供其他视角所不具备的信息[12],因此该模型的识别性能在迭代过程中会因为差异性条件无法满足而受限.

为了解决自步协同训练模型在训练过程中因视角差异性减小而导致的模型性能受限的问题,本文提出了一种基于模糊度[13]评估的自步协同训练模型.在置信度选择框架下进一步引入模糊度选择机制,每一轮训练综合考虑置信度高和模糊度差异大(即差异性大)的样本以满足协同训练结构有效性的两个先决条件,在增加数据多样性的同时,提升模型的容错性能.微信数据业务识别的相关实验结果表明,本文提出的模型在保证置信度的前提下,有效避免了训练过程中两视角差异性下降,模型识别精度到达0.89,F_1-score到达0.87,相比已有的方法较大提升了识别准确度.

1 相关知识

1.1 多视角学习

所谓多视角学习有别于传统的单视角学习,通过多个视角观察相同对象并协同学习获取增益,如多个摄像头捕捉同一目标的动作信息构成多视角.为保障多视角学习的性能增益,应满足以下两个条件[14]:

(Ⅰ)视角的一致性:假设两个视角下的模型函数为 f_1 和 f_2,需要尽可能地保证同一个样本在这两个视角函数下有相同的表示.下面的公式表明,两个视角分类不一致的概率的下界,为两个视角分类错误率最大值.

$$p(f_1 \neq f_2) \geqslant \max\{p_{\mathrm{err}}(f_1), p_{\mathrm{err}}(f_2)\} \quad (1)$$

(Ⅱ)视角差异性:即某个视角包含其他视角所不具备的信息,视角之间存在差异性.差异性是多视角能够全面准确的描述数据的前提.在涉及多视角机器学习问题中,可以利用多个视角的互补信息来改善学习的性能.

1.2　自步协同训练

协同训练是一种较为经典的半监督模型. 首先通过将原始数据分成两个数据子集，称之为视角(view)；其次针对两个视角下分别训练模型，并交叉地给另一视角的无标签数据标记伪标签[15]，利用两个视角的协同增益，即视角的一致性与差异性，提升模型性能. 自步协同训练则是在协同训练框架中进一步引入了自步学习方式来标注伪标签，如图 1 所示. 自步学习的核心思想是对模型的迭代，每次倾向于选择所有样本中具有较小的训练误差、置信度高的样本[11].

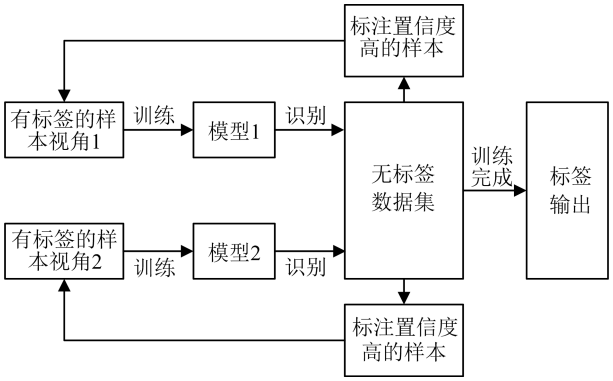

图 1　自步协同训练模型

Fig. 1　Self-paced co-training model

对于一个自步学习模型，其优化目标为[11]

$$\min E(\boldsymbol{w},\boldsymbol{v};\lambda)=\sum_{i=1}^{n}(v_iL(y_i,f(\boldsymbol{x}_i,\boldsymbol{w}))+r(\boldsymbol{w})+p(v_i,\lambda)) \tag{2}$$

式中，矢量 $\boldsymbol{x}$ 表示输入的样本，y 表示样本对应的标签，L 表示模型的损失函数，r 表示模型参数矢量 $\boldsymbol{w}$ 的正则化项，f 是表示模型，$\boldsymbol{v}$ 是一个二分变量，用于表示样本是否被选择. p 函数表示自步学习的正则化，其中 λ 参数控制每一轮迭代选择多少样本标记伪标签加入训练集. 自步学习的优化问题是一种双凸优化问题(biconvex optimization problem). 双凸优化问题是指相对待优化参数集合 z 而言，参数集合可以被划分为两个互斥的子集 z_1 和 z_2. 如果任意一个参数集合 z_i 固定时，另一组参数的优化问题看作凸优化问题，那么整体就可以看作双凸优化问题. 例如，在优化目标中需要优化的两组参数 $\boldsymbol{w}$ 和 $\boldsymbol{v}$，我们可以先固定 $\boldsymbol{w}$ 求解 $\boldsymbol{v}$ 最优解，再固定 $\boldsymbol{v}$ 求解 $\boldsymbol{w}$ 的最优解，该优化方法可以称为 ACS(alternative convex search)方法. 通过联合优化模型参数 $\boldsymbol{w}$ 与权重 $\boldsymbol{v}$，并逐渐增加参数 λ 的值，模型可以从“简单”的样本开始，逐渐扩充训练数据集以解决更加复杂的问题.

以自步学习的方式标注伪标签，每一轮迭代中重新计算权重 v，所有的伪标签样本将会从训练数据集中剔除，模型重新评估无标签样本的置信度，标注伪标签后形成新的训练数据集. 一方面，保证当前数据集对于当前模型是最为置信的；另一方面，可以避免协同训练模型中一旦引入错误的标签，将无法从训练数据集中剔除的问题.

2　整体模型框架

网络数据流是由应用业务产生的数据包的集合[16]，数据包作为传输通信的基本单元，分为包头和包体两个部分. 其中包头包含时间戳、包长、5 元组(源 IP 地址、目的 IP 地址、源端口号、目的端口号、协议类型)以及数据链路层/互联层/传输层/应用层等 4 层的协议头信息；包体对应数据包传输的有效载荷. 假设仅使用时间戳和包长，数据流可表示为 $\mathrm{TF}=(\{t_i,P_i\})_{i=1}^{I}$，其中 I 表示集合中包含的数据包个数，t_i 和 P_i 分别表示第 i 个数据包的时间戳与包长. 使用 5 元组，可以将 TF 分为单向数据流与双向数据流，其中单向数据流的所有数据包的 5 元组完全相同，双向数据流中数据包的源和目的 IP 地址/端口号是可互换的. 依据文献[7,17]，本文选择包含信息更多的双向数据流作为识别对象，研究高可靠的业务标注方法.

由于运营商无法直接获取数据业务类型，一般需要专业技术人员人工标注，有标签的样本非常少，所以本文使用半监督的方法识别微信数据的业务类型. 在微信数据流中，每一条数据流对应一种业务，由于微信本身包含的业务较多，从网络上获取的微信数据中，不同业务的数据流混在一起，直接识别较为困难，所以本文首先把不同业务的数据过滤出来，将问题转化为半监督的分类问题，模型的整体结构如图 2 所示.

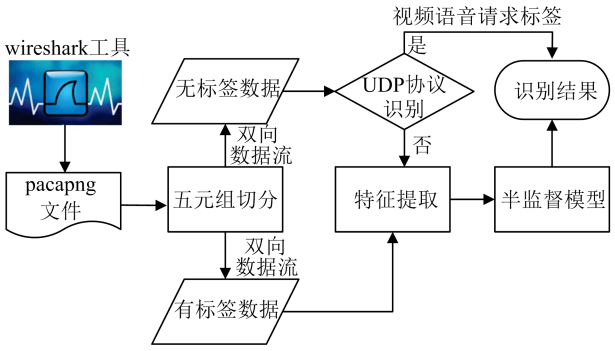

图 2　整体模型框架

Fig. 2　Overall model framework

其具体步骤如下：

(Ⅰ)数据预处理：使用 WireShark 软件从代理服务器上抓取微信原始数据，使用 5 元组将数据切分成双向数据流，每一条流中包含的数据包属于同一类型业务.

(Ⅱ)协议识别：通过户数据报协议(user datagram protocol，UDP)，识别出无标签数据中的视频请求和语音请求两类，然后将无标签的数据与已经通过人工标注的包含 4 种类型(视频，图片，位置，其他 4 类)的有标签数据整合.

(Ⅲ)特征提取：对于整合的数据，提取其时间间隔、数据包的包长以及其组合作特征、有标签的数据构成训练数据、无标签的数据构成测试数据.

(4)构建半监督模型：训练基于模糊度的自步协同训练模型，并给无标签的数据打标签.

3 基于模糊度的自步协同训练

3.1 特征提取

微信数据流半监督识别是一个分类问题，首先介绍从微信原始数据中提取的特征，作为后续半监督模型的训练与测试数据. 微信的双向数据流可以分成：①由数据包时间间隔构成的时间序列；②由数据包包长构成的时间序列. 为了解决数据流的加密的问题，提升模型在其他应用业务数据识别问题上的鲁棒性，本文仅选取数据包时间间隔，数据包的包长以及其组合作为样本数据的特征，包括基本统计特征、前向与后向方差、最长单调子序列长度、跳跃点特征、K 阶子段占比、Top-N 频繁序列频次[18]. 除此之外，依据文献[19-20]，本文还提取了动态时间规整特征，特征描述如表 1 所示.

表 1 特征描述

Tab. 1 Feature description

特征名称	说明
基本统计特征	序列的一阶与二阶描述性统计特征，包括均值、方差、中值、最小值、最大值、最大值个数、平均方差、序列偏度与散度.
前向与后向方差	从 3 个四分位数获得方差，即 $\mathrm{Var}(\mathrm{TF}'_{I_1})$、$\mathrm{Var}(\mathrm{TF}''_{I_1})$、$\mathrm{Var}(\mathrm{TF}'_{I_2})$、$\mathrm{Var}(\mathrm{TF}''_{I_2})$、$\mathrm{Var}(\mathrm{TF}'_{I_3})$、$\mathrm{Var}(\mathrm{TF}''_{I_3})$ 6 对方差，每一对包含时间间隔序列与包长序列两个方差. 其中 $\mathrm{TF}'_{I_1}=(\{t_i,P_i\})_{i\subset[0,I_1]}$、$\mathrm{TF}''_{I_1}=(\{t_i,P_i\})_{i\subset(I_1,I]}$、$\mathrm{TF}'_{I_2}=(\{t_i,P_i\})_{i\subset[0,I_2]}$、$\mathrm{TF}''_{I_2}=(\{t_i,P_i\})_{i\subset(I_2,I]}$、$\mathrm{TF}'_{I_3}=(\{t_i,P_i\})_{i\subset[0,I_3]}$、$\mathrm{TF}''_{I_3}=(\{t_i,P_i\})_{i\subset(I_3,I]}$、$I_1$，$I_2$，$I_3$ 表示序列 $\mathrm{TF}=(\{t_i,P_i\})_{i=1}^{I}$ 的 3 个四分位索引，I 表示序列的长度.
最长单调子序列长度	包长与时间间隔序列中单调递增和单调递减的子序列的长度.
跳跃点特征	该特征描述了数据流量的稳定性，例如相对于视频流数据或语音数据，文本消息包含了更多的随机大小的包与时间间隔，本文中，计算序列中包长或时间间隔的值是下一个位置的值 2.5 倍的索引个数.
K 阶子段占比	将区间 $[t_{\min},t_{\max}]$ 与 $[P_{\min},P_{\max}]$ 等间隔的划分为 K 段，其中 $t_{\min}$、$t_{\max}$、$P_{\min}$、$P_{\max}$ 分别表示某时间间隔序列与对应的包长序列的中的最小值与最大值，对于 $\mathrm{TF}=(\{t_i,P_i\})_{i=1}^{I}$，统计每一个划分的子区间所包含的数据包或时间间隔占整体的比重，本文中 $K=7$.
Top-N 频繁序列频次	与 K 阶子段类似，将区间 $[t_{\min},t_{\max}]$ 与 $[P_{\min},P_{\max}]$ 等间隔的划分为 K 段，命名为 $t^1,t^2,\cdots,t^{K-1}$，t^K 与 $P^1,P^2,\cdots,P^{K-1},P^K$，将包长序列与时间间隔序列中的每一个值用其所属的区间名称代替，得到新的两条序列. 对于新得到的字符序列，统计其长度在 3～20 内的连续子序列的出现次数，并用最频繁的 N 个子序列的个数作为特征.
动态时间规整特征	使用每一类有标签样本的包长序列的频繁序列作为每一类的标准序列，对于所有样本，计算其包长序列到每一类标准序列的 DTW 距离 $\mathrm{Dist}(\lvert X\rvert,\lvert Y\rvert)$ 作为特征，其中 X，Y 表示两个序列，计算方式为：$D(i,j)=\mathrm{Dist}(i,j)+\min\{D(i-1,j),D(i,j-1),D(i-1,j-1)\}$，$\mathrm{Dist}(i,j)$ 表示序列 X 第 i 个点和序列 Y 中第 j 个点的距离，即包长之差，$D(i,j)$ 衡量了序列 X 的前 i 个点与序列 Y 前 j 个点的相似性.

3.2 FBSpaCo 算法

传统的自步协同训练模型每轮迭代中仅对置信度高的无标签样本标注伪标签，随着训练迭代两个视角下的模型越来越相似，导致差异性条件无法满足、模型性能受限[21-22]；此外，协同训练还要求两个视角下的模型具有相同的输出结果，即满足多视角一致性，因此寻求差异性与一致性的平衡是保障协同训练增益的重点[12]. 本文将“模糊度”概念引入自

步协同训练架构，通过模糊度来解决视角差异性下降的问题.

模糊度是一种衡量概念或者对象与集合隶属关系的属性，在语言学中表现为不同词汇之间的不确定边界. 模糊度最早由 Zadeh[23] 在 1965 年提出，作者使用概率来度量对象的模糊度，并推荐使用信息论中信息熵度量不确定性的方式来构建模糊度计算公式. 本文对样本的模糊度给出定义：设 $V=\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_C\}$ 是一个样本的模糊集（概率输出），依据文献[24]，样本的模糊度为

$$\boldsymbol{F}(V)=-\frac{1}{C}\sum_{i=1}^{C}(\boldsymbol{\mu}_i\log\boldsymbol{\mu}_i+(1-\boldsymbol{\mu}_i)\log(1-\boldsymbol{\mu}_i)) \tag{3}$$

模糊度表示不同样本属于不同类别的不确定边界，样本在两个视角下不确定边界的差异可以表示其在两个视角下的差异. 边界差异越大，则表示样本在两个视角下的信息差异越大. 同时，给模糊的样本标记伪标签又会增加模型整体的不确定性，可能会引入错误的标签数据，但考虑到自步学习本身具有剔除有问题的样本的能力，这将减少了模糊样本的负面影响，因此本文提出了基于模糊度的自步协同训练模型来解决上述提到的问题.

本文提出的基于模糊度的自步协同训练模型的整体优化目标标识为

$$\min_{\substack{w^{(j)},y_k,v_k^{(j)}\in[0,1]\\ j=1,2;k=l+1,\cdots,l+u}} E(\boldsymbol{w}^{(j)},\nu_k^{(j)},y_k;\lambda^{(j)},\gamma,\beta)=K_{\text{sup}}+K_{\text{spco}}+K_{\text{fuzzy}} \tag{4}$$

优化目标包括 3 个部分，公式表示为

$$\left.\begin{aligned}K_{\text{sup}}&=\sum_{j=1}^{2}\sum_{i=1}^{l}L(y_i,f_j(x_i^{(j)};\boldsymbol{w}^{(j)}))+\frac{1}{2}\sum_{j=1}^{2}\|\boldsymbol{w}^{(j)}\|_2\\ K_{\text{spco}}&=\sum_{j=1}^{2}\sum_{k=l+1}^{l+u}(\nu_k^{(j)}L(y_k,f_j(x_k^{(j)};\boldsymbol{w}^{(j)}))-\lambda^{(j)}\nu_k^{(j)})\\ K_{\text{fuzzy}}&=-\gamma(\beta(\boldsymbol{F}(\boldsymbol{g}^{(2)})-\boldsymbol{F}(\boldsymbol{g}^{(1)}))+\nu^{(1)})^{\mathrm{T}}\cdot(\beta(\boldsymbol{F}(\boldsymbol{g}^{(1)})-\boldsymbol{F}(\boldsymbol{g}^{(2)}))+\boldsymbol{\nu}^{(2)})\end{aligned}\right\} \tag{5}$$

K_{sup} 表示在有监督数据集上模型的损失函数，K_{spco} 与 K_{fuzzy} 共同组成了模型在无监督数据集上的自步学习框架. 其中，在正则化项 K_{fuzzy} 中引入模糊度评估. 公式新增的符号中，$\boldsymbol{v}^{(i)}$ 表示二分变量 $\boldsymbol{v}$ 在第 i 个视角下的矢量，$\lambda^{(i)}$ 控制了每一轮自步学习的步长，γ 是调整正则化项对整体目标函数影响的参数，$\boldsymbol{g}^{(i)}$ 表示视角 i 下模型的概率输出矢量，β 表示样本模糊度差异相对于置信度的影响程度，β 值越大，说明标注样本伪标签时，模糊度差异的影响更大. $\boldsymbol{F}(\boldsymbol{g}^{(i)})$ 表示样本在第 i 个视角下的模糊度矢量.

本文提出的基于模糊度的自步协同训练模型的正则化项可以写作

$$\begin{aligned}K_{\text{fuzzy}}=&-\gamma(\boldsymbol{v}^{(1)})^{\mathrm{T}}\boldsymbol{v}^{(2)}-\gamma\beta^2(\boldsymbol{F}(\boldsymbol{g}^{(2)})-\boldsymbol{F}(\boldsymbol{g}^{(1)}))^{\mathrm{T}}(\boldsymbol{F}(\boldsymbol{g}^{(1)})-\boldsymbol{F}(\boldsymbol{g}^{(2)}))-\\&\gamma\beta(\boldsymbol{v}^{(1)})^{\mathrm{T}}(\boldsymbol{F}(\boldsymbol{g}^{(1)})-\boldsymbol{F}(\boldsymbol{g}^{(2)}))-\\&\gamma\beta(\boldsymbol{v}^{(2)})^{\mathrm{T}}(\boldsymbol{F}(\boldsymbol{g}^{(2)})-\boldsymbol{F}(\boldsymbol{g}^{(1)}))\end{aligned} \tag{6}$$

整体表示为 4 项求和，为了使整体的目标函数小，正则化项需要尽可能小. 第一项 $(\boldsymbol{v}^{(1)})^{\mathrm{T}}\boldsymbol{v}^{(2)}$ 表示两个视角下的模型对同一个样本尽可能有相同的置信度表达；后 3 项表示两个视角下的模糊度之差尽可能大（包括模糊度差的平方以及模糊度差与 $\boldsymbol{v}$ 的乘积）. 相比传统的自步协同训练模型的正则化项 $(\boldsymbol{v}^{(1)})^{\mathrm{T}}\boldsymbol{v}^{(2)}$，本文的正则化既保证了多视角的一致性特性，又保证了多视角的差异性.

由于算法仅修改了正则化项，保持了原本的自步学习结构，所以该优化问题仍然可以使用 ACS 的优化方法. 下面描述目标函数的优化求解过程.

初始化：首先初始化模型的参数，$\boldsymbol{v}^{(1)}$ 与 $\boldsymbol{v}^{(2)}$ 初始化为全零的矢量，$\lambda^{(1)}$ 与 $\lambda^{(2)}$ 设置为一个较小的值保证初始模型只选择少量样本标记伪标签，γ 与 β 设置为一个固定的参数，在整个迭代训练过程中不变.

更新 $\nu_k^{(3-j)}$ 与 $\nu_k^{j}(j=1,2)$：这一步的物理含义为计算不同视角下样本的置信度，并选择合适的无标签样本标记伪标签加入到 j 视角下模型的训练数据集中. 对于 $v_k^{(3-j)}$ 更新，首先对目标函数求解关于 $v_k^{(3-j)}$ 的偏导数，即

$$\begin{aligned}\frac{\partial E}{\partial\nu_k^{(3-j)}}=&L_k^{(3-j)}-\lambda^{(3-j)}-\gamma\nu_k^{j}-\\&\gamma\beta(F_k(\boldsymbol{g}^{(3-j)})-F_k(\boldsymbol{g}^{(j)}))\end{aligned} \tag{7}$$

当偏导数小于 0，说明目标函数对于 $v_k^{(3-j)}$ 是单调递减的，为了使目标函数最小，$v_k^{(3-j)}$ 取最大值 1；同理，当偏导数大于 0，目标函数对于 $v_k^{(3-j)}$ 是单调递增的，$v_k^{(3-j)}$ 取最小值. 具体的 $v_k^{(3-j)}$ 更新公式为

$$\nu_k^{(3-j)^*}=\begin{cases}1, L_k^{(3-j)}-\gamma\beta(F_k(\boldsymbol{g}^{(3-j)})-F_k(\boldsymbol{g}^{(j)}))<\\ \quad\lambda^{(3-j)}+\gamma\nu_k^{j}\\ 0,\text{otherwise}\end{cases} \tag{8}$$

该更新公式表明，当损失函数 $L_k^{(3-j)}$ 与模糊度之差

$F_k(\boldsymbol{g}^{(3-j)}) - F_k(\boldsymbol{g}^{(j)})$ 的差小于参数 $\lambda^{(3-j)}$ 与第 j 个视角下 $\gamma\nu_k^j$ 的和时，$\nu_k^{(3-j)}$ 在本轮迭代中更新为 1，即在 $(3-j)$ 视角下给第 k 个样本标记伪标签，放入训练数据集. 这说明模型在训练中选择置信度尽可能高以及在两视角下差异性尽可能大的样本标记伪标签，平衡了差异性与一致性的需求.

更新 $\boldsymbol{w}^{(j)}$：使用伪标签与有标签的数据共同组成的训练数据集在 j 视角下重新训练模型，使用梯度下降或者其他优化算法更新模型参数 $\boldsymbol{w}^{(j)}$. 初始阶段，由于在两个视角下有标签的样本个数很少，为了保证模型在初始阶段训练的稳定性，减少伪标签错误的概率，本文选择在小样本上鲁棒性高的随机森林作为两个视角下的训练模型.

更新 y_k：当更新完模型参数 $\boldsymbol{w}^{(j)}$ 后，使用新模型更新样本的伪标签. 由于存在多个视角，所以选择使下面公式最小化的标签，即

$$y_k = \underset{y_k}{\operatorname{argmin}} \sum_{j=1}^{2} \nu_k^{(j)} L(y_k, f_j(\boldsymbol{x}_k^{(j)}; \boldsymbol{w}^{(j)})) \quad (9)$$

为了获得式(9)的全局最小值，可以直接将两个视角下模型的预测输出的加权求和作为伪标签 y_k. 针对微信业务数据识别问题，可直接从 4 种类型中选择使上式最小的一类作为伪标签.

更新 $\lambda^{(j)}$：一旦样本的伪标签更新后，增加 $\lambda^{(j)}$ 的值，增加自步学习的步长，保证下一轮迭代中选择更多的样本数据标记伪标签.

在不同视角下重复以上的优化步骤，直到所有无标签数据都打上伪标签或者迭代次数达到设置的上限. 整体的算法流程如算法 3.1 所示.

算法 3.1 基于模糊度评估的自步学习的协同训练模型

```
1 输入数据，人工提取特征
2 初始化 ν(1)，ν(2)，λ(1)，λ(2) 以及 γ
3 训练初始模型，更新 w(1)，w(2)
4 训练周期设置为 1
5    while 有无标签数据未标记伪标签或训练周期小于最大周期：
6        for j in (1,2)：
7            更新 ν(j)，使用 L_k^(j)，γν_k^(3−j) 获得当前视角下置信度高且模糊度差异大的样本，加入训练集
8            更新 w(j)，根据新得到的训练集重新训练模型
9            更新 y_k，依据误差最小的方式选择伪标签
10           更新 λ(1)，λ(2)
11       end for
12   end while
```

4 实验结果

4.1 实验数据集

表 3 是通过本文搭建的一台 Linux 代理服务器数据采集平台获取的微信数据，使用 6 部装有微信应用的安卓终端(安卓版本 6.0，内存 3G)连接代理服务器，在服务器上使用 WireShark 软件抓取由终端微信产生的数据包. 具体来说，首先，在智能手机上安装微信应用软件，删除其他应用程序；其次，志愿者在早中晚 3 个时间段分别测试微信的 6 种业务，并上报每种业务的开始时间、结束时间和标签；最后，将服务器上的 WireShark 生成的 pcapng 文件添加至数据库.

表 3 实验数据描述

Tab. 3 Description of experimental data

#	业务类型	双向流个数	数据大小
1	其他	2 390	17.5MB
2	图片	2 557	272MB
3	视频	2 361	957MB
4	位置	2 276	27.4MB
5	语音通话	2 140	356MB
6	视频通话	1 985	2 599MB

4.2 评价指标

模型评估是衡量机器学习或深度学习模型的重要环节，本文采用精度(Acc)、准确率(Pr)、召回率(Re)以及 F_1-Score 作为衡量模型的评价指标，公式为

$$\mathrm{Acc} = \frac{1}{V(S)} \sum_S \delta(y_s - y'_s) \quad (10)$$

$$\mathrm{Pr}_t = \frac{\mathrm{TP}_t}{\mathrm{TP}_t + \mathrm{FP}_t} \quad (11)$$

$$\mathrm{Re}_t = \frac{\mathrm{TP}_t}{\mathrm{TP}_t + \mathrm{FN}_t} \quad (12)$$

式中，S 表示所有无标签的双向数据流，$V(S)$ 表示其个数，y_s 与 y'_s 分别表示模型的预测输出与样本的真实标签，δ 是冲激函数，当且仅当 $y_s = y'_s$ 时值为 1，否则为 0. 对于数据业务的每一种类型，TP_t 表示预测类型为 t 的样本中实际类型为 t 的样本个数，FP_t 表示预测类型为 t 的样本中实际类型不是 t 的样本个数，FN_t 表示预测类型不为 t 的样本中实际类型不是 t 的样本个数，求解出 Pr_t 与 Re_t 后在所有类别上平均得到整体的精度与召回率. 由于实际使用中精度和召回率往往成反比，所以本文使用 F_1-

score 来衡量模型的整体表现性能，F_1-score 越高模型的性能越好.

$$F_1\text{-score}=\frac{2\times \mathrm{Pr}\times \mathrm{Re}}{\mathrm{Pr}+\mathrm{Re}} \tag{13}$$

同时，本文分析了模型在训练过程中的差异性变化，该差异性通过计算两个视角下模型在无标签样本上的概率输出的平均 Kullback-Leibler(KL)散度来度量.

$$\mathrm{KL\text{-}div}=\frac{(D(\boldsymbol{\mu}^{(j)}\parallel\boldsymbol{\mu}^{(3-j)})+D(\boldsymbol{\mu}^{(3-j)}\parallel\boldsymbol{\mu}^{(j)}))}{2}$$

$$D(\boldsymbol{\mu}^{(j)}\parallel\boldsymbol{\mu}^{(3-j)})=\sum_K \boldsymbol{\mu}_k^{(j)}\log\frac{\boldsymbol{\mu}_k^{(j)}}{\boldsymbol{\mu}_k^{(3-j)}},(j=1,2) \tag{14}$$

公式(14)是 KL 散度的计算公式，其中，$\boldsymbol{\mu}^{(j)}$ 与 $\boldsymbol{\mu}^{(3-j)}$ 表示某样本在两模型下的概率输出矢量，矢量长度为 K，即类别的个数，$\boldsymbol{\mu}_k^{(j)}$ 则表示矢量中第 k 位的值. KL 散度用于两概率分布的差异性度量. 由于其非对称特性，本文使用两个方向的平均 KL 散度 KL-div.

4.3　实验结果分析

由于微信的视频请求和语音请求使用 UDP 协议可以直接区分，所以实验结果分析中使用的微信数据仅包含图片、视频、位置、其他 4 种业务类型.

首先分析动态时间规整特征的有效性. 图 3 是有标签序列到标准序列的 DTW 距离平均值的热力图，图中每一列方格的颜色(深度)代表某一类包长序列到 4 种不同标准序列(图中的列)的 DTW 距离的平均值，颜色越深则代表 DTW 距离越大. 从图 3 可以发现，对角线上的方格颜色在其所在的行和列最浅，表示该包长序列集合到其所属类的标准序列的平均 DTW 距离最短，说明 DTW 相似度可以作为序列识别的特征，包含充足的信息. 对于每一个无标签序列，计算序列到 4 个标准序列的 DTW 值，并与之前计算的统计特征一起输入到半监督模型中.

分析在迭代训练过程中两个视角下的模型的差异性变化，如图 4 所示. 实验中，随机森林的弱分类器树的个数为 500，树的最大深度为 3，自步协同训练模型中 $\lambda^{(1)}$ 和 $\lambda^{(2)}$ 初始化为 0.5，每一轮迭代完成，$\lambda^{(1)}$ 和 $\lambda^{(2)}$ 增加 0.8，模型的优化目标公式中参数 γ 为 0.3，β 的值为 3.33，即 $\beta\gamma=1$，迭代次数设定为 50 次. 为了尽可能模拟真实的半监督微信业务数据识别任务，有标签的样本仅占整体样本的 0.3%. 图中蓝色线(下方线条)表示无模糊度评估的模型差异性，红色线(上方线条)表示有模糊度评估的模型差异性.

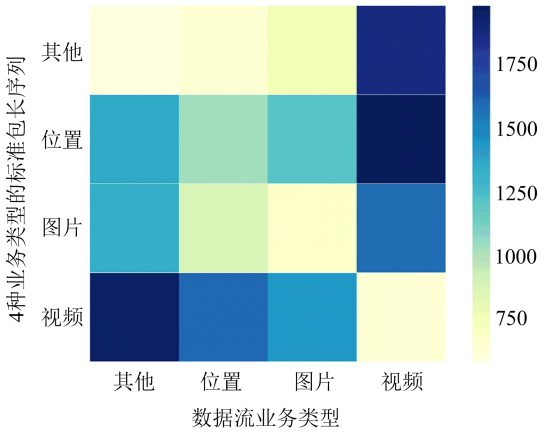

图 3　包长序列的 DTW 距离热力图

Fig. 3　DTW distance thermodynamic map of packet length sequence

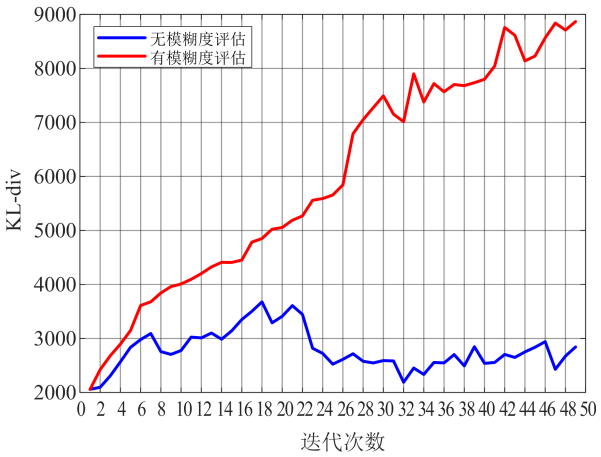

图 4　模型的视角差异性随着迭代次数的变化

Fig. 4　View difference of the model changes with the number of iterations

从图 4 可知，在初始阶段无论是否有模糊度评估，两个视角下模型的差异性都在提升，原因是初始阶段有标签的样本少，两视角下的初始模型 f_1^0 与 f_2^0 只学习到少量的信息，称为弱模型. 在后续的迭代训练中，随着训练数据的增加，两个模型学习到了各自视角下更多的信息，模型的差异性继续增加[21]. 随着继续训练，同协同训练一样，自步协同训练模型的两个视角下的训练数据趋于饱和，模型差异性下降，加入模糊度的自步协同训练模型的两视角的差异性在后续的迭代中仍然保持增长，说明模糊度的引入可以避免自步协同训练模型训练过程中两视角差异性下降的问题，并通过实验分析差异性的增加对于模型分类性能的影响.

图 5 给出了模型的 F_1-score 随着迭代次数的

变化，实验的初始化设置同上. 除了分析模糊度的有效性之外，实验还分析了加入 DTW 特征对模型性能的影响. 由于文献[25]中，作者表示对于一条双向数据流，包含其业务特征的部分主要集中在流的前几个数据包，所以本文计算 DTW 距离时仅使用流的前 20 个数据包组成的包长序列，对于长度不足 20 的双向流，取双向流的所有数据包构成包长序列. 图 5 中，横坐标是模型训练中自步学习迭代的次数，纵坐标是模型在测试集上的 F_1-score，整体上具有模糊度评估并加入 DTW 相似度特征的自步协同训练模型的表现最好，没有模糊度评估且没有 DTW 相似度特征的模型性能最差. 从趋势上看，初始阶段所有的模型随着迭代训练性能都有提升，代表了初始阶段弱分类器逐渐学习的过程；后续阶段具有模糊度评估的模型在分类性能上仍然有提升，最后稳定在 0.87 左右，而不具有模糊度的模型仅能保持性能，甚至出现模型的恶化，其 F_1-score 初始时上升很快，到达约 0.862，之后性能迅速下降，最后稳定在 0.825 左右. 再比较有无 DTW 特征的模型的区别，在训练过程中，拥有 DTW 特征的数据训练的模型性能更好，在有模糊度与无模糊度条件下 F_1-score 分别提升 0.4 与 0.35，说明 DTW 特征能够提供其他特征所不具有的信息，可以提升模型的识别准确度.

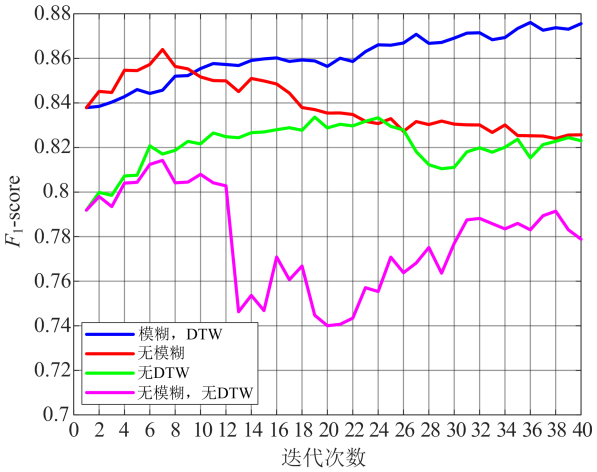

图 5 模型的 F_1-score 随着迭代次数的变化

Fig. 5 The F_1-score of the model changes with the number of iterations

为了表现模型整体的有效性，本文与其他用于网络数据业务识别的半监督模型进行比较，包括基于模糊度的神经网络(fuzziness NN)、半监督最小生成树聚类模型(semi-cluster)、随机漫步半监督模型(random walk). 由于文中引入 DTW 相似度作为特征，所以模型对比使用的数据集包括有 DTW 特征和无 DTW 特征的微信数据. 图 6 表示的是不同模型对比结果，4 张图分别表示不同模型精度、准确率、召回率以及 F_1-score 的对比.

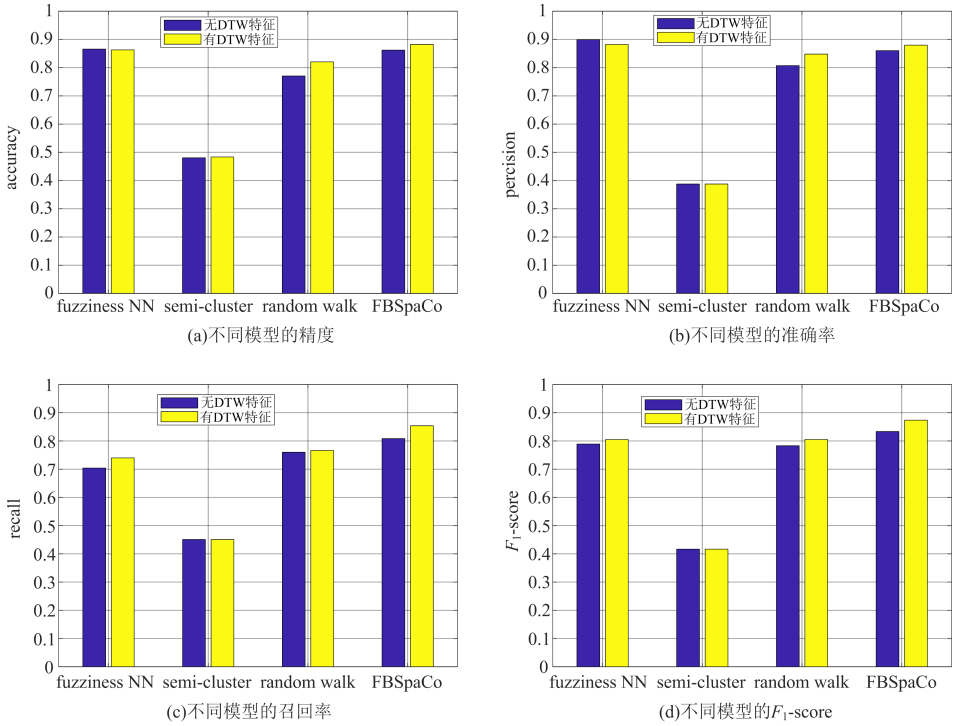

图 6 模型在微信数据上性能比较

Fig. 6 Performance comparison on WeChat data

从图 6 可以看出，在 F_1-score 上，本文提出的模型最高，其 F_1-score 在有 DTW 特征和无 DTW 特征下分别为 0.8336 与 0.8738，表明在微信业务数据集上，本文提出的基于模糊度评估的自步协同训练模型具有最好的表现. 同时，图 6 中也展示了 DTW 特征的引入对模型识别性能的影响，从模型整体的表现来看，有 DTW 特征引入的模型在 F_1-score 上均高于无 DTW 特征引入的模型，特别是本文提出的基于模糊度评估的自步协同训练模型，F_1-score 提升了 4.8%，说明该特征包含的包长序列波形特性是提升网络业务数据识别模型性能所需关注的重点.

5　结论

本文提出了一种使用基于模糊度评估的自步协同训练模型的微信数据业务识别方法，结合多视角学习的一致性与差异性两种特性，利用自步协同训练半监督模型从易到难地选择置信度高的样本标注伪标签的优势，在每一轮迭代中引入模糊度评估，保证两视角的差异性不会随着模型训练而下降，避免模型因视角相似度过高而性能受限. 同时，考虑到包长序列之间存在相似性，在特征提取中加入动态时间规整特征. 实验结果表明，基于模糊度评估的自步协同训练模型在微信数据业务识别问题上精度到达 0.89，F_1-score 到达 0.87，优于当前常用的半监督网络数据业务识别模型的表现.

参考文献(References)

[1] 吕述望，苏波展，王鹏，等. SM4 分组密码算法综述[J]. 信息安全研究，2016，2(11)：995-1007.
LYV Shuwang，SU Bozhan，WANG Peng，et al. Overview on SM4 algorithm[J]. Journal of Information Security Research，2016，2(11)：995-1007.

[2] PAPADOGIANNAKI E，HALEVIDIS C，AKRITIDIS P，et al. OTTer：A scalable high-resolution encrypted traffic identification engine[C]//International Symposium on Research in Attacks，Intrusions，and Defenses. Springer，Cham，2018：315-334.

[3] 赵双，陈曙晖. 基于机器学习的流量识别技术综述与展望[J]. 计算机工程与科学，2018 (10)：1746-1756.
ZHAO Shuang，CHEN Shu-hui. Review：Traffic identification based on machine learning[J]. Computer Engineering and Science，2018 (10)：1746-1756.

[4] MOORE A W，ZUEV D. Internet traffic classification using Bayesian analysis techniques[C]// SIGMETRICS Performance Evaluation Review. Alberta，Canada：ACM，2005，33(1)：50-60.

[5] AULD T，MOORE A W，GULL S F. Bayesian neural networks for internet traffic classification[J]. IEEE Transactions on Neural Networks，2007，18 (1)：223-239.

[6] WILLIAMS N，ZANDER S，ARMITAGE G. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification [J]. ACM SIGCOMM Computer Communication Review，2006，36(5)：5-16.

[7] LOTFOLLAHI M，ZADE R S H，SIAVOSHANI M J，et al. Deep packet：A novel approach for encrypted traffic classification using deep learning[J]. arXiv preprint，2017，arXiv:1709.02656.

[8] WANG Y，XIANG Y，ZHANG J，et al. A novel semi-supervised approach for network traffic clustering [C]// Proceedings of 5th International Conference on Network and System Security. Milan，Italy：IEEE，2011：169-175.

[9] LIN F，COHEN W W. Semi-supervised classification of network data using very few labels[C]//2010 International Conference on Advances in Social Networks Analysis and Mining. Odense，Denmark：IEEE，2010：192-199.

[10] MAHDAVI E，FANIAN A，HASSANNEJAD H. Classification of encrypted traffic for applications based on statistical features[J]. International Journal of Information Security，2018，10(1)：29-43.

[11] MA F，MENG D，XIE Q，et al. Self-paced co-training [C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org，2017：2275-2284.

[12] QIAO S，SHEN W，ZHANG Z，et al. Deep co-training for semi-supervised image recognition[C]// Proceedings of the European Conference on Computer Vision. Munich，Germany：Springer，2018：135-152.

[13] ASHFAQ，RANA AAMIR RAZA，et al. Fuzziness based semi-supervised learning approach for intrusion detection system[J]. Information Sciences，2017，378：484-497.

[14] XU C，TAO D，XU C. A survey on multi-view learning[J]. arXiv preprint，2013，arXiv:1304.5634.

[15] BLUM A，MITCHELL T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the Eleventh Annual Conference on Computational Learning Theory. Park Norwell，USA：ACM，1998：92-100.

[16] HENZINGER M R，RAGHAVAN P，RAJAGOPALAN S. Computing on data streams[J]. External Memory Algorithms，1998，50：107-118.

[17] WANG W, ZHU M, ZENG X, et al. Malware traffic classification using convolutional neural network for representation learning[C]//International Conference on Information Networking. Taipei, China: IEEE, 2017: 712-717.

[18] LIU J, FU Y, MING J, et al. Effective and real-time in-app activity analysis in encrypted internet traffic streams[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2017: 335-344.

[19] PETITJEAN F, FORESTIER G, WEBB G I, et al. Faster and more accurate classification of time series by exploiting a novel dynamic time warping averaging algorithm[J]. Knowledge and Information Systems, 2016, 47(1): 1-26.

[20] WEGNER MAUS V, CÂMARA G, APPEL M, et al. dtwSat: Time-weighted dynamic time warping for satellite image time series analysis in *R*[J]. Journal of Statistical Software, 2019, 88(5): 1-31.

[21] WANG W, ZHOU Z H. Theoretical foundation of co-training and disagreement-based algorithms[J]. arXiv preprint, 2017, arXiv:1708.04403.

[22] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint, 2014, arXiv:1412.6572.

[23] ZADEH L A. Probability measures of fuzzy events[J]. Journal of mathematical analysis and applications, 1968, 23(2): 421-427.[24] DE LUCA A, TERMINI S. A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory[J]. Information and control, 1972, 20(4): 301-312.

[25] LOPEZ-MARTIN M, CARRO B, SANCHEZ-ESGUEVILLAS A, et al. Network traffic classifier with convolutional and recurrent neural networks for Internet of things[J]. IEEE Access, 2017, 5: 18042-18050.

(上接第 17 页)

[12] 孙伟松,孙小兵,李斌,等.软件历史代码库词库自动构建技术及实现[J].中国科学技术大学学报,2017,47(1):80-86.
SUN Weisong, SUN Xiaobing, LI Bin, et al. On automatic construction of the word base for historical program repository[J]. Journal of University of Science and Technology of China, 2017, 47(7):80-86.

[13] 卜尧,吴斌,陈玉峰,等.BDAP——一个基于 Spark 的数据挖掘工具平台[J].中国科学技术大学学报,2017,47(4):358-368.
BO Yao, WU Xiaobin, CHEN Yufeng, et al. BDAP: A data mining platform based on Spark[J]. Journal of University of Science and Technology of China, 2017, 47(4):358-368.

[14] 王国燕,汤书昆.前沿科学成果的图像传播范式[J].中国科学技术大学学报,2014,44(9):754-760.
WANG Guoyan, TANG Shukun. A graphic communication paradigm for forefront scientific results[J]. Journal of University of Science and Technology of China, 2014, 44(9):754-760.