Spark日志整合与FCM-DNN的网络流量分析算法

山东大学

报告人 李腾

目录

01 背景

02 系统模型

03 算法设计

04 实验过程

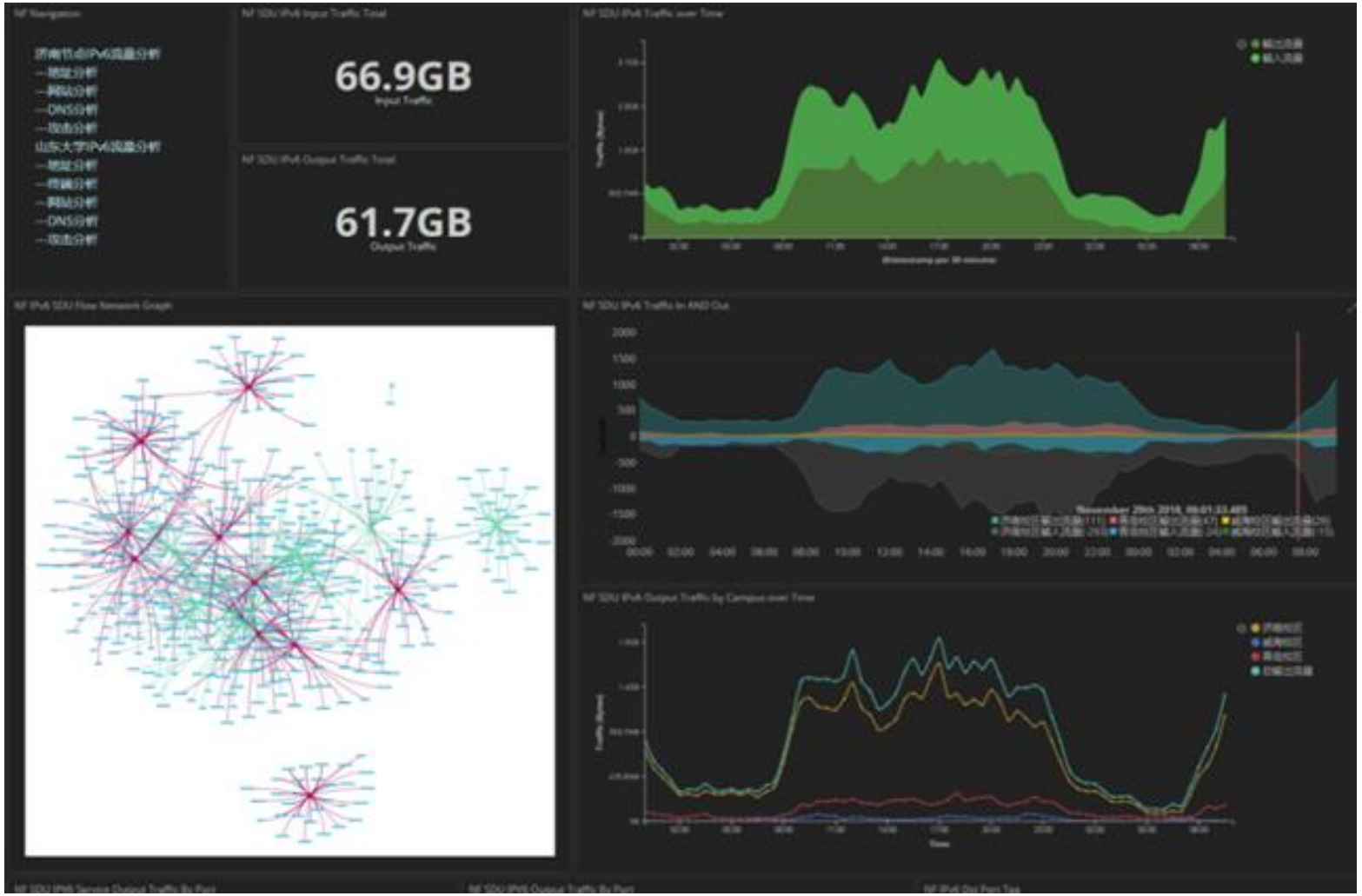
05 校园应用

01

背景

01 背景

山东大学是国内较早开展IPv6升级工作的高校之一，2017年《推进互联网协议第六版（IPv6）规模部署行动计划》发布以来，山东大学进一步建立健全了相关组织管理与工作机制，通过统筹规划，合理组织，加强协同，促进创新，全面实现三地一体（济南、威海、青岛）IPv6规模部署与应用创新。



山东大学IPv6应用规模同样处于国内高校前列。学校主页（域名：www.sdu.edu.cn）和所有二级单位网站，OA系统、人事系统等关键信息化应用系统已全部完成IPv6升级，并通过自主开发的校内IPv6部署动态检测平台，及时发现并解决校内信息系统的IPv6部署盲点，推动IPv6的全方位部署。

学校积极开展支持IPv6的智慧校园应用创新，青岛校区建设了支持IPv6的物联网应用试点，实现了园区环境监控、大型仪器定位管理和学生宿舍无线门锁等创新应用。

通过Portal方式实施双栈和纯IPv6用户认证，实现了IPv6身份认证管理。升级了网管系统，全面实现对IPv6网络设备、终端和地址等进行管理。建设了流量分析系统，实现对IPv6流量的分析与展示。

01 背景

随着IPv6的快速发展，不仅要对出入口流量的变化趋势进行分析，还需对用户行为进行分析。例如，对高校学生的上网行为进行分析，可识别出沉迷于网络游戏、网络电视的“网瘾学生”，尽早对其进行干预，以对他们的学业进行帮助，在带宽紧张时，分析占用带宽较高的上网行为，以便高校对网络是否需要扩容进行决策。

Session数据可记录学生上网过程中的网站地址、子域名、域名(上网使用协议)、域名所属大类(上网类型)、上网起止日期、使用流量等信息，是分析上网行为的重要数据。目前使用Session日志进行上网行为分析存在两个难题。一是高校学生基数较大，网络日志较多，而高校存储设备有限，需要对日志按照域名所属大类进行合并存储。二是校园网出口流量控制设备虽然会对大部分域名进行归类，但由于流量控制设备的目的是对网络安全进行保护和对异常流量进行监控，使得域名所属大类标记过于宽泛，很多网站都被标记为HTTPS 协议或者 HTTP 协议。通过统计归类于HTTP 大类的域名占比高达30%, 这使得学生真正的上网类型无法被挖掘。

上网行为分析

- 识别沉迷网络游戏、网络电视的“网瘾学生”
- 识别上网时间过长的学生
- 对网络是否扩容做出决策

01 目前存在的挑战

山东大学 SHANDONG UNIVERSITY

1. 每日产生300G网络日志，网络流量特征提取困难
2. 域名的特征维度少，不利于通过深度学习进行分类
3. 单一Session连接特征具有相似性和不平衡性的特点
4. 通过神经网络进行分类时需通过优化提升识别的精准度

01 解决挑战的工具

山东大学 SHANDONG UNIVERSITY

- 数据汇聚工具
- Spark大数据平台
- FCM（模糊C均值聚类）
- DNN（全连接神经网络）

网络出口设备每日产生约300G，2亿条数据。需通过数据汇聚工具将Session日志抽取为HDFS文件。再通过Spark对日志进行计算整合。

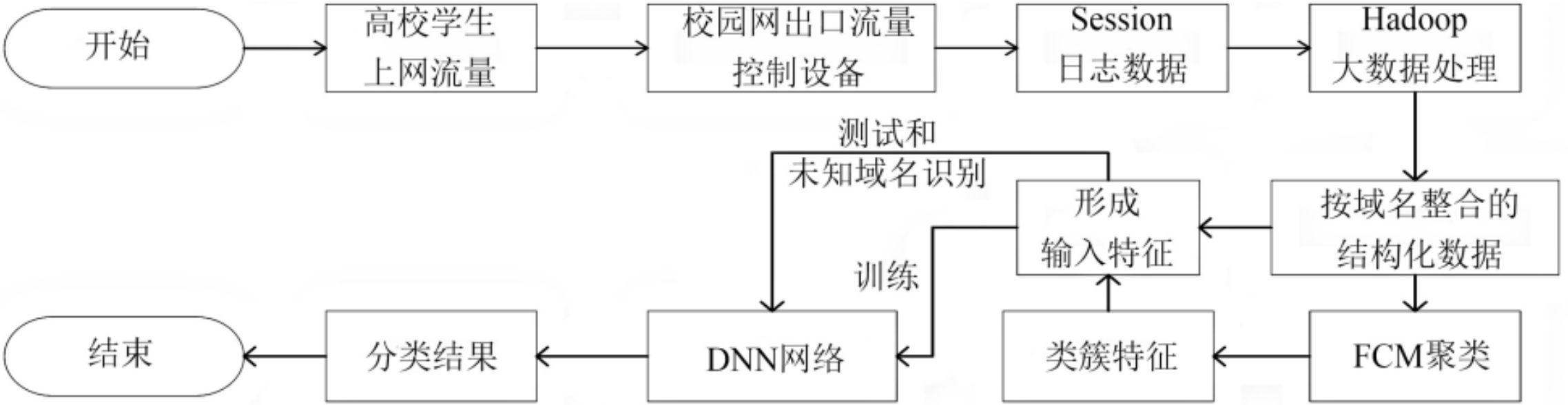
域名的特征维度较少，且单一Session连接特征具有相似性和不平衡性的特点。相似性是指不同类型网站的单一Session连接具有相似的特征。例如，各类网站主页的Session连接，无论时长还是流量都是近似的。不平衡性是指不同类型的网站流量和上网时长分布不同，但其平均值可能相近。通过FCM聚类，可提取类簇的特征，解决上述问题。

通过构建DNN网络，对已知分类的网站进行训练，并通过归一化数据、设置聚类时的分组长度、使用交叉熵损失函数等优化策略，提高网站域名识别的精准度。

02

系统模型

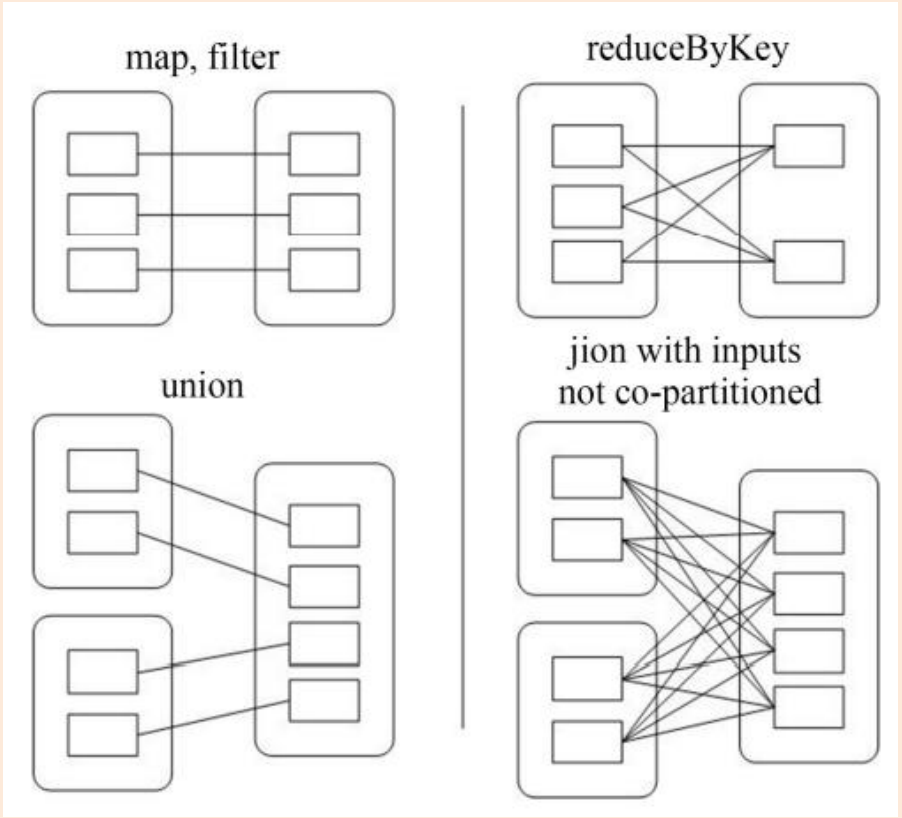
02 系统模型



模型将网站域名分为常用协议、网络电视、网络游戏、社交、流媒体、下载、游戏共 7 个大类。大数据处理工具将流量控制设备生成的日志按照域名整合为可分析的结构化数据，每个大类选取已知域名进行聚类，计算各大类最大类簇数(z)和簇内属性，再结合域名原始特征得到综合特征。对综合特征进行统一化处理，包括类簇属性按流量排序、聚类分组长度统一化，从而生成最终的神经网络输入特征。 使用部分已知类型的域名数据进行训练， 选取其他已知类型的域名数据进行测试，得到分类的准确率。

Spark数据处理

使用Spark批处理，对日志进行整合。Spark是一种基于内存的大数据并行计算框架，常被用于批量计算，其计算速度是MapReduce的100倍以上，适用于多次读取大批量数据的场景。此外，Spark可以集成Hadoop，读取Hadoop上的任意数据。不同于传统Hive调用MapReduce的过程，本研究使用Spark中的RDD进行计算，计算过程如图2所示。RDD通过丰富的算子进行转换，从而提升日志整合速度。



01 HDFS

02 HIVE

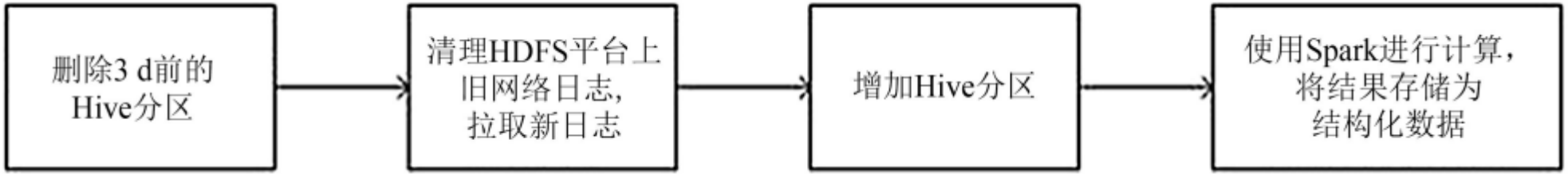
03 RDD

04 Spark

05 整合完成

02 数据处理过程

Spark处理整合日志文件的流程为：通过数据集成工具将日志按照规范抽取到 HDFS 平台后，自动关联到Hive分区，使用Spark工具，按照相关规则对日志进行计算，将结果存储到结构化数据库，定期清理计算完成的日志和Hive分区，以实现快速自动化的计算过程，最终将海量的非结构化数据转化为可分析的结构化数据。



02 FCM聚类

使用FCM对Spark计算好的各域名的属性进行聚类。模糊c均值聚类(fuzzy c-means，FCM)是基于模糊理论的聚类算法，主要用于数据的聚类分析。其主要思想是使被划分到同一类的对象之间的相似度达到最大，而不同类对象之间的相似度达到最小。聚类结果是每个数据点对聚类中心的隶属度，可用数值表示

FCM算法首先对参数初始化，然后求解各类的聚类中心，并迭代更新聚类中心和隶属度矩阵，从而使目标函数达到最优。

$$J(U,V)=\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}^{m}(d_{ik}(x_k,v_i))^2$$

$$d_{ik}(x_k,v_i)=\| x_k-v_i \|$$

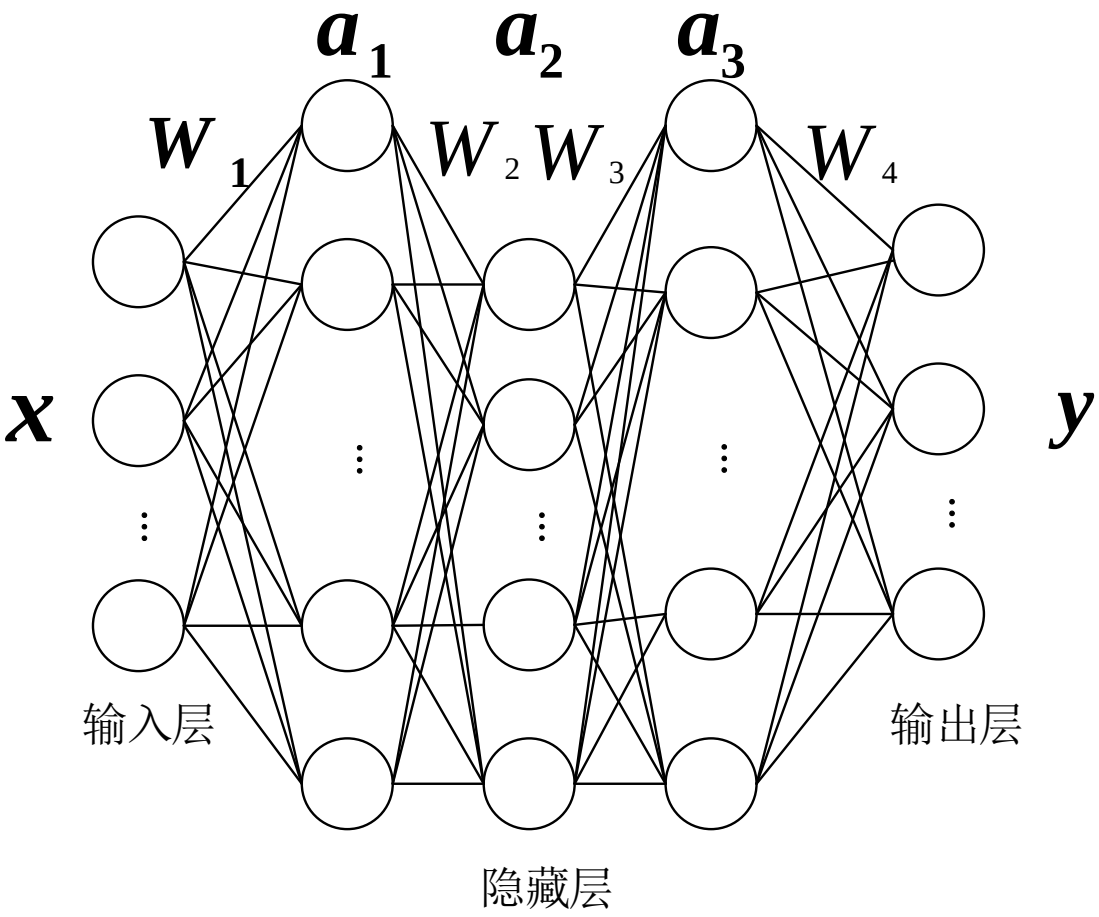
$$u_{ik}=(\sum_{j=1}^{c}(\frac{d(x_k,v_i)}{d(x_k,v_j)})^{\frac{2}{m-1}})^{-1}$$

$$v_i=\frac{\sum_{k=1}^{n}(u_{ik})^{m}\ x_k}{\sum_{k=1}^{n}(u_{ik})^{m}}$$

$$a_1 = f(W_1^{\mathrm{T}} x) \qquad a_2 = f(W_2^{\mathrm{T}} a_1)$$

$$a_3 = f(W_3^{\mathrm{T}} a_2) \qquad y = f(W_4^{\mathrm{T}} a_3)$$

全连接神经网络共5层，包括1个输入层、1个输出层和3个隐藏层。其中，输入层的节点数与FCM聚类结果相关。如果所有大类的聚类结果中最大的类簇数为z，则输入层节点数为3z+1，隐藏层中每层设置50个节点。由于日志类型共分为7类，所以将输出层设置为7个节点。



激活函数 $$y = \text{sigmoid}(\boldsymbol{W}^{\mathrm{T}} \cdot \boldsymbol{x}) = \frac{1}{1+\mathrm{e}^{-\boldsymbol{W}^{\mathrm{T}} \cdot \boldsymbol{x}}}$$

损失函数 $$H(p,q) = -\sum_{x} p(x) \ln q(x)$$

损失值 $$\text{Loss} = -\frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=1}^{M} y_{ij} \ln p_{ij}$$

由于网站分类是one-hot向量问题，如果把结果当作是概率分布来看，标签表示的就是数据真实的概率分布，由激活函数产生的结果其实是对于数据的预测分布，预测分布和真实分布差值叫做KL散度或者是相对熵。我们希望的是预测值尽量靠近真实分布，也就是希望相对熵可以越来越小。相对熵又等于交叉熵减去数据真实分布的熵，后者是确定的，所以最小化相对熵就等价于最小化交叉熵。

因此在DNN网络中使用交叉熵损失函数来衡量的是预测值和真实标签之间的差异性，训练的目的是不断减少Loss，也就是让预测值不断靠近真实值。

03

算法设计

算法 1 得到最大类簇数 z 和存储训练数据的 D_{map}.

1）算法输入日志整合后的结构化数据 D_{log} 和各大类的典型网站域名集 M_{domain}. D_{log} 表示为 (f, t, l, a, r_q). 其中，f 为流量，t 为时长，l 为连接数，a 为域名，r_q 为日期. M_{domain} 表示为 $(M_1, \cdots, M_7)$. 其中，M_i 表示第 i 类网站的域名集合.

2）定义数据的时间区间 $(t_s, t_e]$，定义时长阈值 t_{min}. 令 $z=0$，创建存储训练数据的 D_{map}.

3）循环 M_{domain}. 对于第 i 类网站的域名集合 M_i，获取满足条件的日志，条件表示为

$$(a \in M_i) \wedge (t \geqslant t_{min}) \wedge (r_q \in (t_s, t_e]) \tag{14}$$

4）生成协议类簇. 将 3）得到的特征集中的流量、时长属性，组成矩阵 $\boldsymbol{X}$；为减少聚类时间，将 $\boldsymbol{X}$ 分为 m 个子矩阵 $\boldsymbol{X}'$. 令 $\boldsymbol{y}'=\mathrm{FCM}(\boldsymbol{X}')$，得到结果向量 $\boldsymbol{y}'$. c 为聚类的维度，如果 $c>z$，则令 $z=c$.

5）获取训练集 D_{train}. $D_{train}=\mathrm{cal}(\boldsymbol{X}', \boldsymbol{y}', c)$，$\mathrm{cal}(\boldsymbol{X}', \boldsymbol{y}', c)$ 的计算过程见算法 2，将 D_{train} 保存到 D_{map} 中.

6）对 D_{map} 中的训练集进行补 0 操作.

算法 2 通过矩阵 $\boldsymbol{X}'$、结果向量 $\boldsymbol{y}'$和聚类数 c 获取训练数据.

1）定义协议类簇属性集($f_{\text{sum},1}$，…，$f_{\text{sum},c}$；$t_{\text{sum},1}$，…，$t_{\text{sum},c}$；$r_{\text{sum},1}$，…，$r_{\text{sum},c}$)．令 $i=1, 2, \cdots, c$，则 $f_{\text{sum},i}$代表第 i 类数据的流量之和，$t_{\text{sum},i}$代表第 i 类数据的时长之和，$r_{\text{sum},i}$代表第 i 类数据的个数.

2）循环结果向量 $\boldsymbol{y}'$. 循环聚类数 c，得到 $f_{\text{sum},i}=\sum_{j=1}^{n}\sum_{i=1}^{c}\boldsymbol{X}'[i][1]$、$t_{\text{sum},i}=\sum_{j=1}^{n}\sum_{i=1}^{c}\boldsymbol{X}'[i][2]$、$r_{\text{sum},i}=\sum_{j=1}^{n}\sum_{i=1}^{c}1$. 式中：$n$ 为向量 $\boldsymbol{y}'$ 的长度；i 和 j 需满足条件 $\boldsymbol{y}'$；$\text{get}(j)=i$.

3）计算类簇内平均流量、平均时长和聚为 j 类的数量比例. 平均流量 $\overline{f_i}$ 为 $f_{\text{sum},i}/r_{\text{sum},i}$，平均时长 $\overline{t_i}$ 为 $t_{\text{sum},i}/r_{\text{sum},i}$，个数比例 r_i 为 $r_{\text{sum},i}/n$.

通过算法 1、2，将 Spark 整合得到的结构化数据，转化成全连接神经网络的训练数据. 算法 1 通过设置 $t_{\min}$，过滤打开网页后直接关闭，或者打开网页时连接超时的数据；设置聚类分组长度，将 $\boldsymbol{X}$ 分为 m 个子矩阵，提高聚类速度；通过调整 t_s 和 t_e 可持续获取训练数据；对流量和时长进行聚类，得到类簇内平均流量、平均时长和各类数量所占比例，连接数取子矩阵的平均值；对训练数据统一化，按照流量从大到小进行排序，保证输入数据顺序的一致性；得到最大类簇数 z，对小于 z 的数据补 0，保证输入节点个数的一致性.

04

实验过程

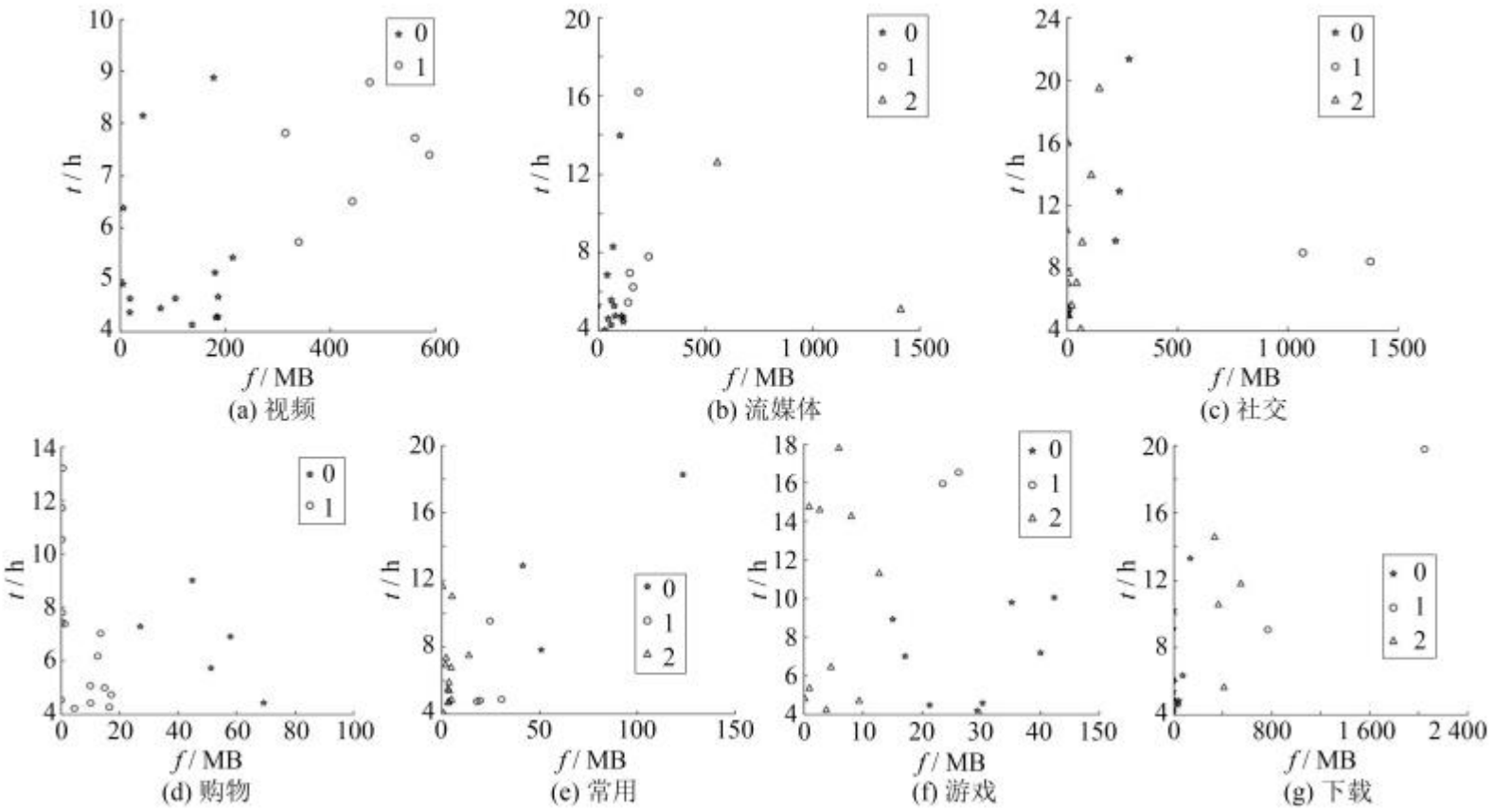
系统运行环境

表 1　系统运行环境

Tab.1　System operating environment

服务器	作用	配置
master1	大数据主节点	12 核 CPU，256 GB 内存，6 TB 硬盘
master2	大数据副节点	12 核 CPU，256 GB 内存，6 TB 硬盘
slave	大数据从节点	12 核 CPU，256 GB 内存，6 TB 硬盘
exchange	数据交换	8 核 CPU，128 GB 内存，500 GB 硬盘
train	聚类、训练	16 核 CPU，256 GB 内存，500 GB 硬盘

获取类簇属性



结果如图所示。可以发现视频类和流媒体类的平均流量和平均时长虽然相似， 但流媒体网站因为有下载操作，会出现高流量的点，而视频类流量较为平均， 下载和社交网站同样存在流量和时长不平衡的问题，导致如果直接求均值，会出现数据分布不同，但平均值相同，影响分类的精准度，因此无法直接将数据送入神经网络。而获取类簇中各类别点的属性均值和各类点的比例，可有效提高分类的精准度。

均值数据

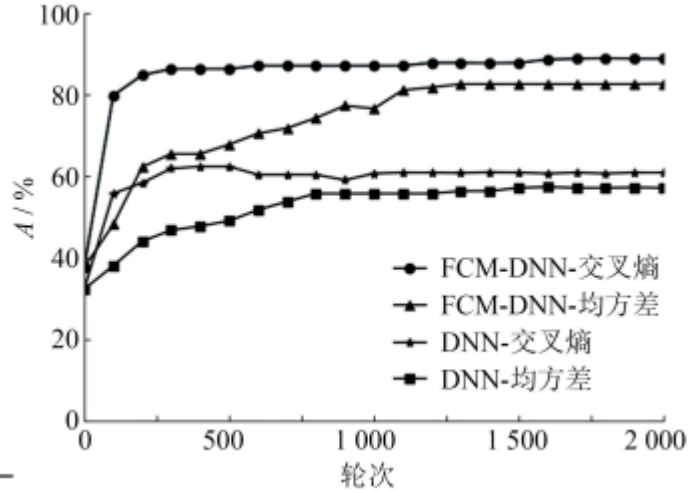
子域名	域名	f/MB	t/h	l
122.245.81.136	baiduyun	812.3	4.2	4 133
pro.m.jd.com	jingdong	12.4	14.1	1 248
act.wegame.com.cn	HTTP	30.2	0.1	1 120

FCM

获取类簇属性算法

类簇均值数据

$\overline{f}_1$/MB	$\overline{t}_1$/h	r_1	$\overline{f}_2$/MB	$\overline{t}_2$/h	r_2	$\overline{f}_3$/MB	$\overline{t}_3$/h	r_3	l
985.1	8.9	0.1	237.3	7.8	0.05	93	6.5	0.85	1 618
454.6	7.3	0.3	111.1	5.3	0.70	0	0	0	2 884



05

校园应用

05 校园应用

利用Spark对日志进行整合后，对已知同类型网站进行聚类，得到该类型网站的类簇和各类簇的特征集；将类簇进行统一化处理，将数据处理为相同长度；结合网站自身的特征，一并送入全连接神经网络进行训练，并使用测试集进行测试。该算法可弥补网络数据特征维度较少、特征具有相似性和不平衡性的问题，分类精度得到提高。同时，利用分类器分出的大类，将日志数据按照学生学号和大类进行压缩，把每日产生的22亿条数据，压缩为约30万条数据，且压缩的日志拥有分析学生上网画像的能力，并可以将每日数据进一步压缩为每周和每月的数据，解决高校存储资源不足的难题。

敬请批评指正！