ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.12, No.35, December 2016
E-mail: wltx@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

KDDcup2015数据集研究

宋国琴,何春,章三妹

(西华师范大学 教育信息技术中心,四川 南充 637000)

摘要:KDDcup2015数据集提供了学堂在线半年内39门课程的部分时段学习行为信息,主要用于学生翘课行为预测研究。翘课行为反映了幕课的质量问题,也是在线教育的核心问题之一。该文通过对数据集的详细分析,解读了KDDcup2015数据集的格式和内容,介绍了数据分析的工具和平台,并通过实例展示如何将原始数据转化为有机的字典数据,以利于进一步的特征建立和机器学习。总结了数据集的不足和可能的影响,为同类数据集的建立和应用提供了依据。

关键词:KDDcup2015;幕课;翘课;Python

中图分类号:TP181 文献标识码:A 文章编号:1009-3044(2016)35-0005-03

DOI:10.14004/j.cnki.ckt.2016.4900

KDD是数据挖掘与知识发现(Data Mining and Knowledge Discovery)的简称,KDD CUP是由ACM(Association for Computing Machiner)的SIGKDD(Special Interest Group on Knowledge Discovery and Data Mining)组织的年度竞赛。

学生的高辍学率成为MOOC平台最核心的问题,也是在线教育的核心问题之一[1-3]。对辍学的了解和预测可以很好的维护和促进学生的学习活动。因此,KDDCup2015的题目为:对中国最大的MOOC平台之一学堂在线的辍学行为进行预测。通过对数据集的解读,根据用户之前的行为,对他在接下来的10天内是否会翘课进行预判。

由于在线教育的盈利要求,目前极少有完整的系统的在线教育公开数据,在线教育公开数据极其稀缺,KDDcup2015[4]数据集有极高的研究和应用价值。通过对KDDcup2015数据集的分析和研究,提出了数据的分析方法和手段,为进一步的数据挖掘[5]或机器学习[6]过程做铺垫。

1 数据集基本情况

数据资源有五个文件,均为CSV格式,如表1所示。CSV是一种通用的、相对简单的文件格式,被用户、商业和科学广泛应用。最广泛的应用是在程序之间转移表格数据,而这些程序本身是在不兼容的格式上进行操作的(往往是私有的和/或无规范的格式)。因为大量程序都支持某种CSV变体,因此在实践中,CSV文件还是非常方便的。

表1 KDDCup2015文件信息

文件名	大小(KB)	说明
enrollment_train.csv	8646	注册号(训练集)
log_train.csv	603782	学习日志(训练集)
truth_train	995	翘课标签(训练集)
enrollment_test.csv	5765	注册号(测试集)
log_test.csv	398863	学习日志(测试集)
Object.csv	3062	课程及模块信息
date.csv	3	课程的日志数据最早和最迟时间
sampleSubmission.csv	663	上交成绩模板

下面分别解释关键表中字段的含义:

1)object.csv -在这个文件中的每一行描述了一个课程中的模块,包括它的类别,它的子模块,以及发布时间。这些模块可能代表的课程的不同部分,例如章,节,在线视频材料、习题等。模块被组织成树型结构,每个课程包含几个章节;每章包含几个部分,每个部分包含几个对象(视频、习题等)。

- course_id -课程号
- module_id -模块号
- category - 模块种类
- children -模块的子模块
- start -模块向学生开放的时间

2)enrollment_train.csv - 每一行表明某用户参加了某课程。

- enrollment_id - 注册号
- username -学号
- course_id -课程号

3)log_train.csv - 每一行都是一个“事件”的日志行为记录。每条记录包含以下信息:

- enrollment_id -注册号
- time - Time of the event.(事件)操作发生的时间
- source - Event source (server or browser).(事件)操作的资源
- event -在事件类型方面,定义了7种不同的事件类型:

problem -做作业
video -看视频
access -读取课程的除了视频和作业外的其它对象
wiki -读取课程的维基百科
discussion -论坛讨论
navigate -浏览课程其它部分
page_close - 关闭网页

- object -读取或浏览的对象

4)true_trian.csv - 每行包含一个注册号是否翘课的信息。
-第1列 - 注册号

收稿日期:2016-11-15
基金项目:四川省教育厅项目(编号:16ZA0171)
作者简介:宋国琴(1979—),女,四川南充人,副教授,硕士,主要研究方向:特征选择、神经网络、深度学习、数据挖掘。

-第2列 - 是否辍学(离散类型,0表示辍学事件,1表示持续学习)

这些文件中的主要对象及关系如图1所示。

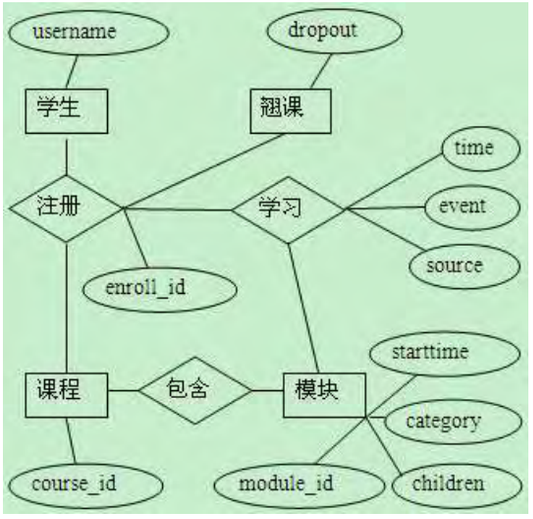

图1 KDDCUP2015数据E-R图

2 样本分析

KDD2015数据集一共包含的39门课程,每门课程的学习信息都是开课后的一个月,总的时间跨度为半年。图2和图3显示了训练集和测试集中不同时间点上的日志分布。通过比较可以看出,在春节期间的日志数量极少,说明放假后学生基本不参与学习。另外,训练集和测试集的数据分布高度相似,使训练集上的应用可以有效地应用于测试集。

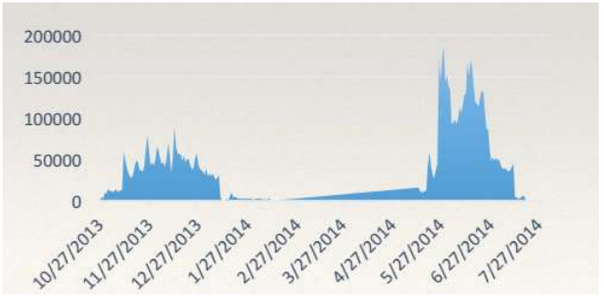

图2 训练集数据分布

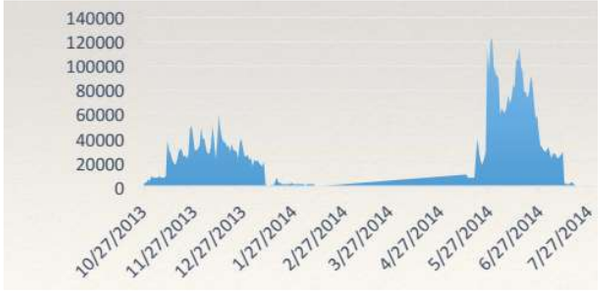

图3 测试集数据分布

3 数据分析及处理

3.1 工具和平台

Python[7-8]是一种面向对象的解释型计算机程序设计语言,是纯粹的自由软件,源代码和解释器CPython遵循GPL(GNU General Public License)协议。Python语法简洁清晰,特色之一是强制用空白符(white space)作为语句缩进。它使你能够专注于解决问题而不是去搞明白语言本身。

本文基于64位的windows7平台,使用基于Python2.7解释器的64位Anaconda2软件读取和分析数据。Python2.7是目前为止最稳定使用最广泛的python版本,而Anaconda2集成了完备的python科学计算的第三方库,安装方便,简化了数据分析和科学计算的很多程序设置工作。

3.2 程序实现

Python字典是另一种可变容器模型,且可存储任意类型对象,如字符串、数字、元组等其他容器模型。字典由键和对应值成对组成。字典也被称作关联数组或哈希表。字典设置实例如下:

```
dict = {'Alice': '2341', 'Beth': '9102', 'Cecil': '3258'}
```

通过对字典的键的访问或遍历,可以很方便地实现数据表的访问、联结或存储。

以下程序使用字典方式处理enrollment_train.csv内容。

```
class Enrollment:
def __init__(self, filename):
    fin = open(filename)
    fin.next()
    self.enrollment_info = {}
    self.user_info = {}
    self.user_enrollment_id = {}
    self.course_info = {}
    self.ids = []
    for line in fin:
        enrollment_id,username,course_id = line.strip().split(",")
        if enrollment_id == "enrollment_id":
            continue
        self.ids.append(enrollment_id)
            self.enrollment_info[enrollment_id] = [username, course_id]
        if username not in self.user_info:
            self.user_info[username] = [course_id]
            self.user_enrollment_id[username] = [enrollment_id]
                else: self.user_info[username].append(course_id) self.user_enrollment_id[username].append(enrollment_id)
            if course_id not in self.course_info:
            self.course_info[course_id] = [username]
            else: self.course_info[course_id].append(username)
    print "load Enrollment info over!", len(self.course_info), len(self.enrollment_info)
    print self.enrollment_info.get("1")
if __name__ == "__main__":
    enrollment = Enrollment("../data/train1/enrollment_train.csv")
```

通过上面程序,enrollment_train.csv文件中的信息分别存入了course_info{}, enrollment_info{}, user_info{}, user_enrollment_id{}4个字典中,如表2所示。这4个字典中,分别使用不同关键字存储信息,可以通过不同键值查询信息,实现了信息

的归类与存储。

表2　字典信息

字典名	关键字	键值
e course_info{}	课程号	选修此课程的用户名
enrollment_info{}	注册号	注册号对应的用户名,课程号
user_info{}	用户名	此用户选修的课程号
user_enrollment_id{}	用户名	此用户的所有注册号

4 KDDcup2015存在的问题及应用研究方向

KDDcup2015的预测问题为翘课,它的主题任务是预测某个注册号未来是否翘课,界定原则为某个时间点后面10天是否有日志记录。实际上,如果注册号10天后回来继续学习,这是翘课行为;如果注册号10天后再也不继续学习,这是辍学行为。这两种行为的性质大不一样,而数据集并不没有有效区分。

KDDcup2015数据集样本具有不平衡性。从数据分析中可以看出,训练集中绝大部分是负样本,即大部分样本都有翘课行为,这会导致正样本代表性不足,使预测结果的泛化性降低。另一方面,数据集缺乏用户及课程特征数据,也缺乏一些细节信息,如观看视频的进度,使得整个数据集不够完整,也会使机器学习的结果不具有代表性。

KDDcup2015数据集包含了大量的实际行业数据,未来可以作进一步的数据挖掘研究,机器学习试验性研究,以及特征自动建立[9]的探索研究。

参考文献:

[1] JOSEP G, JULIA M. Rethinking dropout in online higher education: The case of the Universitat Oberta de Catalunya[J], The International Review of Research in Open and Distributed Learning, 2014, 15(1).

[2] DANIELl F O O, JANE S, RUSSELL B. Dropout rates of massive open online courses :behavioral patterns,6th International Conference on Education and New Learning Technologies [C]. EDULEARN14 Proceedings, IATED, 2014.

[3] TAN M, SHAO P. Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method[J]. International Journal of Emerging Technologies in Learning. 2015, 10(1).

[4] SIGKDD, KDD Cup 2015-Predicting dropouts in MOOC[EB/OL].(2015-8-4). http://www.KDDCup2015.com/information.html

[5] HAN J W, KAMBER M. 数据挖掘:概念与技术[M]. 范明,孟小峰译. 2版. 北京: 机械工业出版社, 2007.

[6] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[7] (美)BILL LUBANOVIC 著. 丁嘉瑞, 梁杰, 禹常隆译. PYTHON语言及其应用[M]. 北京: 人民邮电出版社, 2016.

[8] 张若愚. Python科学计算第2版[M]. 北京: 清华大学出版社, 2016.

[9] JAMES M K, KALYAN V. Deep Feature Synthesis: Towards Automating Data Science[C]. Proceedings of the The 3rd IEEE International Conference on Data Science and Advanced Analytics(DSAA). IEEE, 2015.