基于海量标签的用户画像系统

作者：李晨、司靖辉、夏雯雯、辛德泰、张安澜

【摘要】近年来，随着用户数据的大量累积，越来越多的产品和企业开始重视利用现有的数据，分析数据之间的内在关联，从而更好地了解用户的需求和偏好，赢得市场。就校园内的学生数据而言，其数据的多样性更蕴含了丰富的潜在信息，对其加以充分利用可以最大限度地优化学校的资源和管理，为学生提供更优质的服务。然而当今现存的智慧校园相关应用对数据的利用尚不充分，仅仅进行了数据的表层分析而没有挖掘其深层的内在联系，因而其功能也是收效甚微。而在我们的工作中，我们依据数据的特征和功能的需要使用了多种机器学习算法，通过数据之间的关联关系生成了大量多维度的人物标签，从而构造了清晰完善的人物画像，进而应用到各种不同类别的模型中来支撑所需的功能，取得了令人满意的效果。

目录：

一、 背景

随着数据挖掘技术的不断成熟，越来越多的数据被人们搜集和利用，教育信息化也迎来了新的发展机遇。传统校园将经由电子校园、数字化校园阶段,逐步迈向智慧校园阶段。智慧校园主要是通过综合信息服务平台,依托社交网络、数据挖掘等关键技术支持,集成了校园的分布式信息系统资源,为广大师生提供了全面、协同的智能化感知环境,为教学、科研、管理和生活提供智能化、个性化、便捷化的信息服务。

反观目前的校园平台的服务，一方面是覆盖面窄，往往是一个应用侧重于某一个或几个方面，而没有实现全面系统的功能功效，使得学生经常同时使用多个 APP 才能满足学习生活的大多数需要；另一方面是功能性差，现有的校园服务对数据的分析往往流于表面而未能挖掘其内在的关联，从而使得某些推荐与预测服务不尽如人意。

而在我们的工作中，为了更为充分地利用学生的数据，我们利用了多种机器学习的相关算法，结合统计学的相关研究，对数据的内在关系进行了较为深层的挖掘，发现了很多数据之间的意想不到的关联。我们对处理过的数据进行分析，生成了与之对应的一系列标签，从而全方位的刻画我们的学生用户，利用推荐模型、异常检测模型、预测模型与评估模型为广大师生提供更为优质的服务，使学生可以根据自己的自身情况更好地调节自己的学习和生活状态，使教育工作者更好地制定和执行政策和方针，使招聘企业可以更好地了解学生的状态和能力。

二、 相关工作

随着信息与通讯技术的不断发展，大量的学生数据逐渐积累成了一个蕴含有大量价值的信息宝库，构建一个基于海量数据分析的一体化校园生活系统已然成为一个受到多方面重视的重大课题。截止到今日，不仅是这一课题的理论研究还是实践的雏形都已有了长足发展。下面我们将具体阐述当前的研究现状。

就理论研究方面而言，去年在会议上发表的《Discovering Different Kinds of Smartphone Users Through Their Application Usage Behaviors》一文可以说是这一课题研究现状的高度总结和典型代表。文中以智能手机的用户为研究对象，通过对手机 APP 使用、安装和卸载数据的清洗、聚类和分析，较为全面的得到了不同用户群体的具有代表性的特征，并未这些特征贴上了标签以用于手机的生产与 APP 的预装。然而其模型中不足的一点是，它的标签维度只有一维，并未涉及多维度标签的分析和利用，这对于文章中手机用户的分析也许足够，但将该模型置于多而繁杂的校园数据的应用中则略有缺陷，可以说是深度有余而广度不足。

就实践的雏形而言，我们以浙江大学的校园服务软件作为实例进行分析。浙江大学智慧校园规划了四个部分：智慧的应用、智慧的平台、云计算和通信网。其中智慧的应用包括智慧的校园管理和智慧的校园设施两个层次，涵盖了学校科研、教学、生活、交通、建筑等所有领域。然而细观其模型细节和实现方法我们可以发现，这些功能仅仅依靠多方位数据采集的便利而进行的一种信息整合与表层关联，相当于扩大了当前已知的信息网络让每个人都可以通过一个小点而到达信息网络的各个角落，但对于数据隐含信息的价值挖掘不足，造成了一种数据资源的浪费，可以说是广度有余而深度不足。

而我们进行的工作，是将上述两项工作的优点集中起来，既注重挖掘深层的数据关联又重视对各项数据的全方位利用，从而形成一个基于海量标签的人物画像系统，用来支撑校园服务平台的各项功能。

三、 需求分析

根据题目描述与数据预览，我们主要面向学生、学校和招聘企业三类群体做出了如下需求分析：

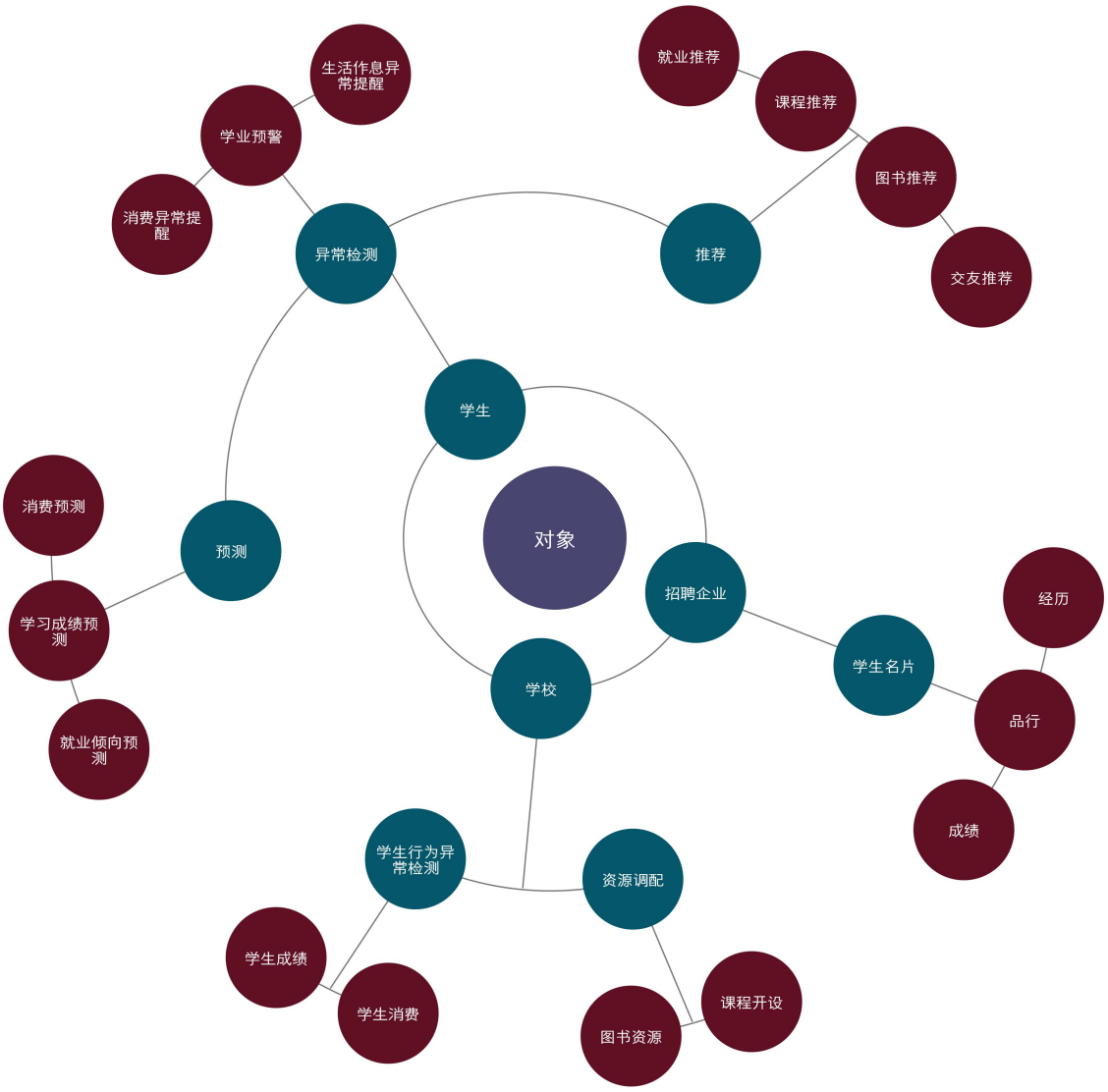

图 1 需求分析图

四、 具体方案实施

1、数据预处理

我们通过对现有的数据进行了初步的分析后发现，其存在格式不一、条件模糊、数据不完整等等一系列问题。为了使后续的分析更加顺利，我们首先对数据进行了清洗以提高数据的质量。具体步骤如下:

图 2 数据预处理流程图

(1)数据清理

在“消费记录.csv”文件中，我们删除了消费类型为空、消费地点为空、消费时间为空或者消费金额为空的记录；在“学生成绩.csv”中，我们从课程和学生两个维度进行了分析：我们先把每个课程的所有学生成绩从小到大进行排列，计算其四分位数，将落在第三个四分位数(Q_3)之上或第一个四分位数(Q_1)之下至少 1.5xIQR(IQR = Q_3 - Q_1)处的值作为离群点挑选出来得到 D1，同理，对学生对象的各科成绩也进行同样的处理得到 D2。将 D1 与 D2 的并集从“学生成绩.csv”文件中提取出来单独存储，不进行删除是因为我们根据现有数据尚未分清其中哪些是非正常数值哪些为异常数值，而且离群点数据对于异常行为分析也十分重要。

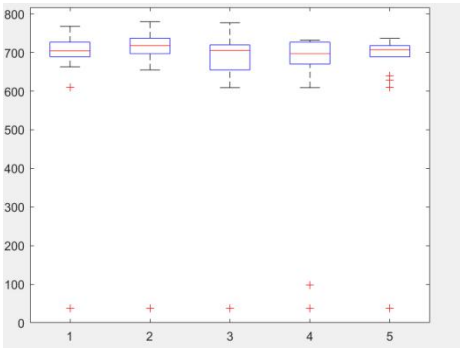

图 3 学生个人成绩盒图

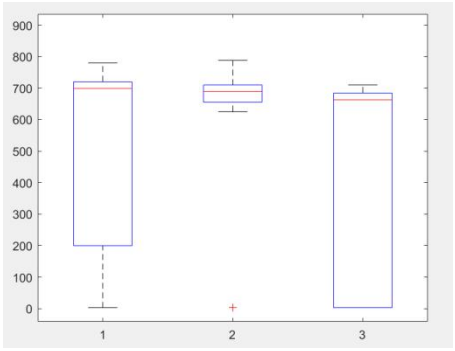

图 4 课程成绩盒图

注：图中“+”为离群点

(2)数据背景分析

为了方便将现有数据结合外部的微博、气象等数据进行联合分析，我们首先需要知道数据的背景和来源。

我们通过对消费类型属性的取值分析，发现了“喜付下发”和“支付宝充值”等包含隐藏信息的消费类型。之后我们调查了使用“喜付”的成都高校，结合提供的数据中的学生人数和车载消费信息（很多为班车消费），初步认定“成都某高校”应该为电子科技大学。之后我们对消费时间进行了分析，发现了在“e02-29”这一天的数据，同时结合支付宝发行和大范围推广使用的时间，我们认定消费记录数据中 e 代表的是 2016，那么其他时间也依次进行了推导替换。而在借阅记录数据中，我们对时间进行分析也发现了“e-02-29”，而“b”、“c”、“d”对应的年份没有“02-29”，同时“a”年份数据是从 9 月份开始的，从而结合上述分析我们推断这是 2012-2016 年的图书借阅记录。对其他数据文件的时间我们也进行了类似的分析处理，具体的对应关系见表 1。

加密字母	a	b	c	d	e	f
对应年份	2012	2013	2014	2015	2016	2017

表 1 字母年份对照表

(3)数据规范化

对于奖学金数据，我们根据电子科技大学的奖学金评审规则，我们把学校的奖学金类型与题目提供给的 11 种奖学金数据一一对应(具体请见表 2)；对于薪资数据，我们根据所在地区的普遍薪资水平和大小关系，用 1-9 的数字替代薪资中的字母。为了分析学生成绩、奖学金和薪资之间的关系，我们对学生成绩、奖学金和薪资数据进行了最小-最大规范化处理，然后单独存储。之后我们又根据各个省级行政区域的薪资水平，将就业地区进行了概念分层，分别用 H、M、L 代表地区薪资的“高”、“中”、“低”三个层次，例如：“上海-市辖区-浦东新区”被替换为“H”，“湖南-长沙市-天心区”别替换为“M”；同时根据消费类型的资金流向分析，我们还将消费记录中的消费类型转换为“-1”，“0”，“1”，其中“-1”代表收入，“0”代表未进行消费，“1”代表支出，例如：“POS 消费”用“1”来替代，“支付宝充值”用“-1”来替代而“卡冻结”用“0”来替代。

x912	z052	x616	y663	z512	x492/y524	y076	z918	z735	y786	x986
11000	10000	9000	8000	6500	6000	5000	3000	1500	1000	500

表 2 奖学金对照表

2、生成数据标签

在对数据进行了初步的清洗之后，我们要进行的是聚类和贴标签工作。聚类是为了更好地把握某一个或者某几个属性的数据分布，挖掘其内在联系，而贴标签则是为了用一种更容易为人所理解的方式来展现数据的特征。最后我们对聚类进行了评估，对标签进行了解释。具体的分析流程如下：

图 5 生成数据标签流程图

(1)数据的整理重组

为了让数据的分析展现出更好的效果，挖掘出更多的特征，我们首先对初步清洗过后的数据进行了整理重组。这一过程主要分为新属性构造和数据向量重构两大部分。其中新属性的构造主要由消费类型衍生而出。由于消费类型的取值范围广，信息量大，我们便从中提取了一些有用的信息来进行属性重构从而使我们的分类和标签更加准确和多样化。例如，我们将卡丢失等类似信息单独列为一个数值属性来统计每个学生的丢卡次数以辅助分析学生的性格；将每个学生充电的金额、时间单独提取出来单独存放以探究同宿舍人群和宿舍的用电情况等等。而在数据向量的重构方面则是对现有属性和新生属性的合理整合，我们将同类或者不同类但彼此之间有相关关系的属性聚合到一起生成多种向量形式，为下面的聚类和之后的模型处理做铺垫。例如，我们以学生为对象，分别对图书和课程进行了向量生成，形成了类似于（图书 A，图书 B，图书 C……）和（课程 A，课程 B，课程 C……）的向量形式，为之后的推荐模型做准备。

(2)组合聚类与标签生成

在这一处理阶段，我们使用整理重组的数据向量对其中多个属性值进行有选择的组合聚

类，然后在对聚类结果进行评估之后对各个类别进行人工贴标签工作。在这里简要介绍一部分我们所聚类和贴标签的方向。

根据图书借阅时间和图书借阅次数对图书编号进行聚类，我们可以得到专业相关类书籍和非专业相关类书籍；根据学生 ID、课程编号和专业书籍类图书编号之间的关联，我们可以区分出学生的专业类别、课程的专业类别以及图书的专业类别；根据课程编号、开课学期、专业类书籍类图书编号以及借阅时间之间的关系，我们可以大致推断出课程类与图书类之间的对应关系；根据学生成绩、平均奖学金金额和图书借阅量聚类，我们可以对学生的专业能力和学习能力打上标签；根据消费类型中 POS 消费、消费时间对消费地点进行聚类分析，可以推断出消费地点是食堂、超市还是其他地方；根据消费类型中 POS 消费、消费时间和消费地点标签对学生进行聚类，可以推断学生之间的好友关系……类似的关系还有很多，就不在此一一列举了。

至于在聚类分析中用到的方法，我们对不同的数据组成使用了不同的聚类算法。例如：在区分专业相关和非专业相关类书籍时，我们便依据图书借阅量--时间构成的柱形图进行了数据切割，将借书高峰期出现的各个书籍借阅计数情况与平常进行比较，有明显不同（0、1 的变化也算入在内）的书籍分类为专业相关书籍；在分析学生、课程编号、书籍编号之间的关联的时候，由于数据是高维系数矩阵，所以我们一方面使用 PCA 进行降维后再进行聚类分析，另一方面也尝试使用 Ng-Jordan-Weiss 算法进行谱聚类分析。在其他的数据集上我们用到过还有 K-means、CLARANS、BIRCH、DBSCAN、EM、DENCLUE 和 CLIQUE 等聚类算法，其具体的应用由数据集的特性来决定。

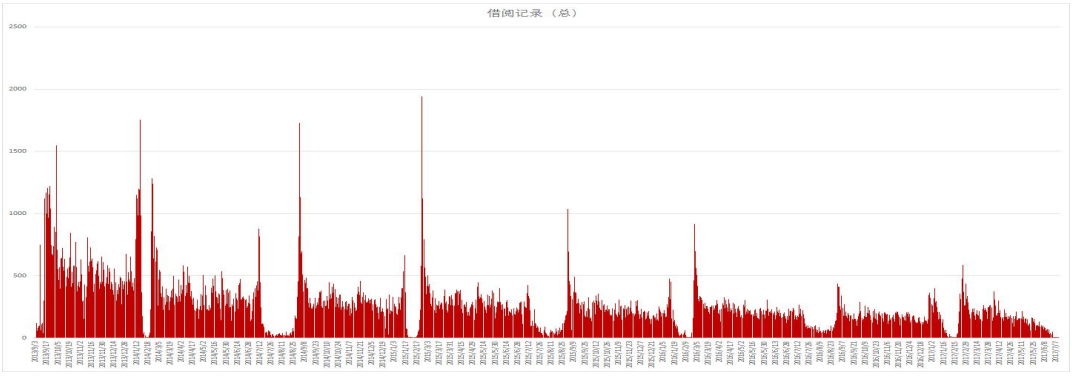

图 4　借阅记录与时间关系

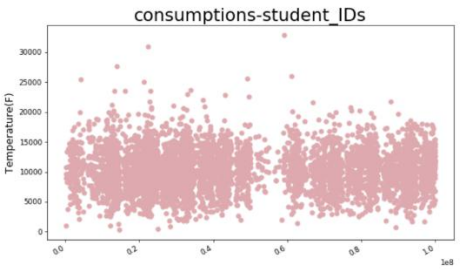

图 5 消费与学号关系

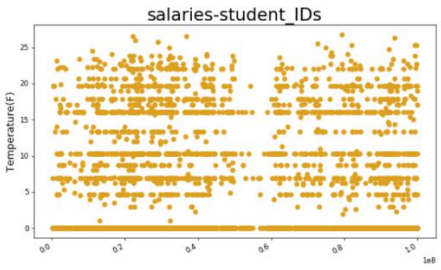

图 6 就业薪资与学号关系

(3)聚类评估与标签解释

在用一个或几个聚类方法对某个数据集进行聚类之后，我们对聚类结果进行了评估。在这方面我们主要是利用了轮廓系数这一个度量，我们通过使用数据集中所有对象的轮廓系数的平均值来比较不同聚类算法的聚类质量。

而对于由聚类结果生成的标签，我们从其解释性的角度分为可解释和不可解释两种。其中可解释标签指的是像消费水平高或低这种具有实际意义的标签，而不可解释标签则是像图

书编号这种只能用字母或者数字进行抽象表示类别而无法获知其实际含义的标签。而从其应用的角度划分，标签主要分为能力类、生活类、书籍类、课程类和工作类标签，例如消费水平的高低属于生活类标签；图书的专业类别属于书籍类标签。在对人物进行分析时，我们就可以通过其学习能力、阅读能力、生活作息、消费水平以及偏好的书籍和课程这些标签来进行人物模型的刻画。

3、提出模型

在得到了大量多类别的标签数据之后，我们就可以设计相关模型来解决我们问题的实际需要。在这里，我们主要提出了推荐、异常检测、预测和人物评价四种具体模型。

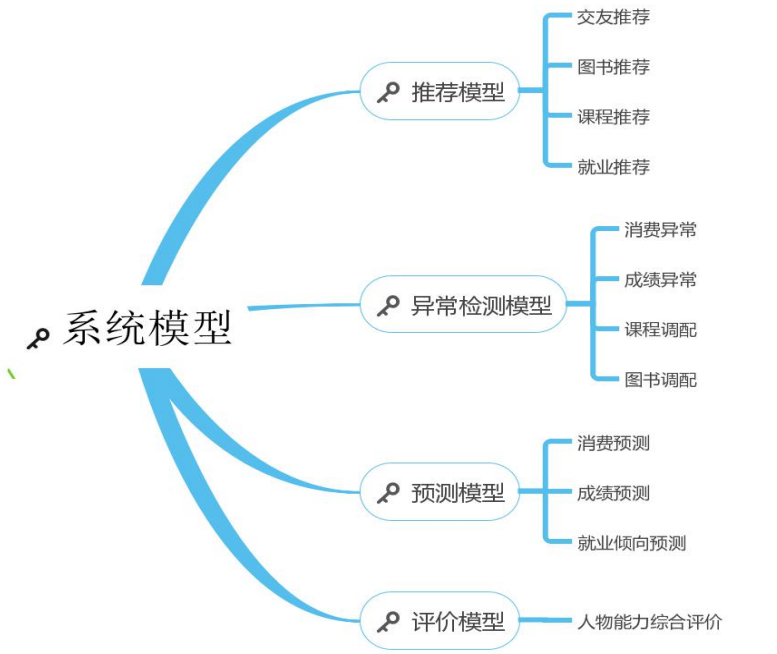

图 7 模型-功能关系图

(1)推荐模型

该模型主要服务于学生的需求，它所支持的功能有好友推荐、图书推荐、课程推荐和就业推荐。它主要以学生作为对象，学生标签作为属性进行处理分析。在这个过程中我们用到的推荐算法主要分为基于人口统计学的推荐、基于内容的推荐、基于用户的协同过滤推荐以及基于物品的协同过滤推荐三种，其具体使用根据数据和推荐的内容的不同而不同。

在好友推荐的服务板块，我们主要用到的是基于人口统计学的推荐和基于内容推荐。一方面我们认为一个人与他/她好友的好友成为朋友的可能性比较大，所以我们会将位于其好友的朋友圈中但又不属于他自己朋友圈中的用户推荐给他/她。在另一方面，我们认为标签相似度高的人很可能成为朋友，例如性格标签中性格相近，在图书借阅标签中偏好同一类书籍，在课程标签中课程吻合度高等等，于是我们基于某一类或者某几类标签数据间的欧式距离进行好友推荐。

在图书和课程的推荐板块，我们用到了上述中的三种算法。在人口统计学算法上，我们会根据用户好友群中的书籍和课程对用户进行合理的推荐，因为我们认为好友之间的喜好总会有相同的部分；在基于用户的协同过滤系统中，我们根据用户的对图书和课程的偏好标签而找到相似的人群，从相似的人群中再找合适的图书和课程进行推荐，即一种“物-人-物”的推荐模式；在基于物品的协同过滤系统中，我们根据所有用户对不同标签数据的偏好，发

现标签与标签之间的相似度，再根据该用户的标签属性，通过皮尔逊相关度进行标签推荐，即一种“人-物-人”的推荐模式。

而在就业推荐板块，由于就业数据相对于整体学生数据来说所占比例不高，我们就简单的使用了基于人口统计学的推荐模型，将与用户相似度高的就业人群中的就业地区类别推荐给用户本身。

图 8 推荐模型-算法-功能图

(2)异常检测模型

该模型主要服务于学生、校方和辅导员三种用户。对学生而言，该模型主要用于校园卡消费异常提醒、学业预警和生活作息异常提醒；对学校而言，该模型主要用于资源的调配例如图书资源调配和课程资源调配；对辅导员而言，该模型主要用于对整体学生的状况的把握，便于即时发现学生在学习生活上的异常。

该模型的分析算法主要分为横向和纵向两个方面，其数据由历史数据和实时更新的当下数据组成。在横向的角度，我们分析的是单个学生对象在整个学生群体的相对位置，主要使用聚类方法进行离群点检测，在之前对各项历史数据进行聚类的过程中已经标记出了离群点并在有必要时进行了单独存储，当新的数据到来时对新数据首先进行临时存储，当到达预定的阀值时再重新进行聚类更新信息，这一角度的检测是有一定的时间延迟的；在纵向的角度，我们分析的是当下数据和历史数据之间的关系，使用的是基于统计学的方法，当前数据到来时立即与历史数据的状况进行分析对比得出结果，具有实时性。

在具体的实现方面，不同的功能使用不同的方法。学生的异常信息提醒主要使用的是纵向方法，将历史数据用四分位图表示出来，将当前数据与历史数据进行比较，在这里我们拟定将落在第三个四分位数(Q_3)之上或第一个四分位数(Q_1)之下至少 2.5xIQR 的数据作为异常数据并对学生进行异动提醒，例如 POS 消费值异常可能是校园卡被盗等等；校方对资源的调配中，对图书的调配使用的是纵向方法，对课程的调配主要使用的是横向和纵向方法，当某一个图书或者某一个标签的图书的借阅量在某一段时期相对于历史数据突然升高时，校方应该考虑加大该图书或该类图书的投放量，当一个课程的成绩均值和方差相对于其自身历史数据及其他同难度系数课程的数据有较大出入时，我们有理由认为该课程存在一些问题，校方应该介入调查询问原因；辅导员对学生异常行为的检测需要使用两种方法，一方面导员需要了解每个学生在整个学生群体的相对位置分布，另一方面需要观察两个不同的时间段学生相对位置的变化情况，当相对变化超过一定的阀值时，导员则需要给出额外关注，对信息有所了解。

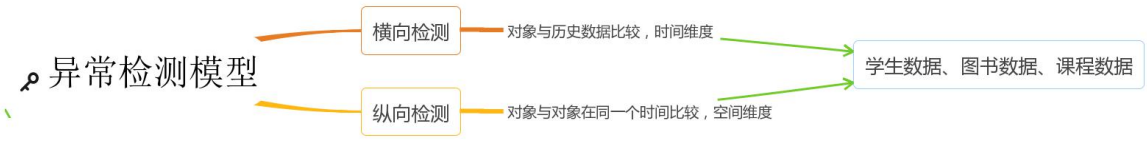

图 9 异常检测模型-方法-数据图

(3)预测模型与评价模型

由于相对于上述两种模型而言，我们在这两个模型上面所做的工作不足，处理方式也相对简单，所以在此合并起来分析。

预测模型主要用于学生的消费预测、学生成绩预测和就业倾向预测。其中消费预测和学生成绩预测主要是依据学生历史消费和成绩数据的均值和方差进行的线性预测；而就业倾向预测主要依赖于上述的推荐系统模型中的就业推荐，根据相似的学生群体来对用户的就业地点和薪资进行预测。

图 10 预测模型-算法-功能图

评价模型主要是刻画了学生的整体人物形象，方便校外企业加深对学生的理解。例如学生的学生能力和专业能力的强弱（由成绩和奖学金情况刻画）、关注时事的倾向（由论坛发帖数量和微博行为分析）、性格特质偏沉稳还是创新（由书籍和课程的跨标签情况、成绩的稳定度、消费的稳定度等得到）以及一些其他的信息。这些可以方便企业为自己的部门和岗位招揽到更合适的人才。

图 11 评价模型-方法-功能图

五、 总结与展望

在本项工作中，我们首先对提供的数据进行了清洗，然后根据数据的特征和应用利用 k-means、CLARANS、BIRCH、DBSCAN 等聚类算法进行聚类分析，之后对聚类的结果进行评估并贴上可解释或不可解释的标签，进而为每个用户生成了基于海量标签的用户画像，为我们推荐模型、异常检测模型、预测模型以及评价模型的构造打下数据基础，并为我们对各个方面功能的实现提供了有力的支撑。

然而由于时间的缘故，截止到目前，我们的系统还依旧停留在模拟实验的基础上，尚未能开发出以该系统为核心的 APP，而且该系统仅仅是以历史数据的分析建模为主体，未能加入对时序数据流处理的功能模块。同时在今后，我们拟利用在数据预处理中推测出来的时间地点信息，结合微博发帖、天气变化、社会新闻等外部数据进行更深层次的优化处理，从而真正的实现智能化、智慧化，为广大师生提供一个良好的服务平台。

六、 参考文献

【1】 基于大数据的智慧校园建设研究 张玲 温向明

【2】 智慧校园建设总体架构模型及典型应用分析 王燕

【3】 基于数据挖掘的学生成绩分析 严的兵

【4】 智慧校园规划设计

【5】 数据挖掘概念与技术 第 3 版

【6】 电子科技大学学生奖学金条例（试行） 电子科技大学信息公开网 http://www.xxgk.uestc.edu.cn/xxgk/read_xx.aspx?id=126

【7】 常用的推荐算法解析 http://blog.csdn.net/u014605728/article/details/51274814

【8】 推荐系统实践 项亮

【9】 Sha Zhao, Julian Ramos, Jianrong Tao, Ziwen Jiang, Shijian Li, Zhaohui Wu, Gang Pan, and Anind K.Dey. 2016. Discovering Different Kinds of Smartphone Users Through Their Application Usage Behaviors. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 498–509.

【10】 Chunqiu Zeng, Wubai Zhou, Tao Li, Larisa Shwartz, and Genady Ya. Grabarnik. 2017. Knowledge Guided Hierarchical Multi-Label Classification Over Ticket Data. *the Journal of Network and Service Management* 14(2): 246-260

【11】 Shan Wu, Shangfei Wang, and Qiang Ji. 2017. Capturing Dependencies among Labels and Features for Multiple Emotion Tagging of Multimedia Data. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence.* AAAI: 1026-1033