摘　要

基于机器学习的电商在线消费者购买行为预测研究

近年来人们的日常消费方式发生了翻天覆地的变化。由于网络的普及，人们开始使用PC以及移动设备进行网上购物，该种方式突破了时间和地域的限制。线上商品种类繁多齐全且价格更为低廉，能够更好的满足消费者多样的需求。但海量的商品呈现给消费者的同时，也使得消费者需要花费大量的精力来挑选商品。越来越多商家为了更好的满足消费者的消费需要而对消费需求进行细化，研发满足消费者细化需求的网络购物平台并推向市场，使得网络零售市场竞争更为激烈。如何能够准确的了解消费者的消费需求，并对其提供更有针对性的垂直服务，是电商接下来的发展过程中，不得不思虑的重要一环。随着数据科技的不断进步，大数据成为近几年的新兴话题，在大数据存储计算水平上有了较大的提升，进而衍生出了区块链技术。消费者每一笔交易数据都被记录在了服务器中，进而可以通过机器学习以及各种智能计算方法分析消费者的在线行为以及交易数据等预测消费者未来的消费行为。

本文采用阿里云天池大数据平台提供的来自淘宝购物平台的已经进行脱敏处理的真实数据，对消费者的行为进行统计挖掘，预测消费者会购买哪种商品。对消费者购买行为预测模型的提出分为四步：第一步是数据异常值的处理。对原始数据去除噪声、去除缺省值，并对消费者行为进行初步统计得出基本的分布情况，为进一步的特征选择和提取以及机器学习方法的选择进行准备工作。第二步是特征选取。从商品的维度构造出消费者特征、商品特征以及消费者-商品交互的行为特征三大特征群。将消费者行为按照发生的时间顺序进行连接作为交互行为序列，并通过各种变换来找到更符合数据特点的其它不同的特征组合，将其加入特征集合。而后，以正样本集大小作为参考，对负样本进行不放回的随机抽样；由于正样本在整体数据集中占比过低，将正样本全部入样。第三步对于统计过的行为数据进行筛选处理，原始数据中存在大部分操作行为过少的记录，在训练中将会影响模型的精度。该问题通过对消费者行为的定性分析来对数据进行筛选处理，删除有嫌疑冲动消费以及行为次数过少的记录，并对数据按照不同行为序列长度对数据进行分层处理。第四步是模型训练和预测。本文尝试应用循环神经网络算法（RNN）对消费者行为序列进行研究，利用N vs 1结构RNN来对行为序列

行为倾向进行二分类，得出消费者行为倾向得分。而后将得分作为新的特征，将新的数据集利用朴素贝叶斯算法进行进一步的预测。将其结果和利用单一朴素贝叶斯算法建立模型所得结果进行比较。利用训练集对模型进行测试后的实验结果表明，使用RNN和贝叶斯融合后的模型预测效果更稳定，能够降低时序序列长度对预测准确度的影响；预测准确度相对单一朴素贝叶斯模型也有一定的提高，模型结果AUC值最优能够达到0.92。

最后，本文提出了模型在电子商务实际交易场景中应用方向以及思路，并分析模型自身不足，对该课题进一步研究方向进行更为详细的讨论。

关键词：

朴素贝叶斯，循环神经网络（RNN），行为分析，电子商务

4.1 数据来源与数据预处理

4.1.1 数据来源

经过十几年的发展，阿里巴巴旗下的淘宝以及天猫电子商务平台逐渐成为中国电子商务领域龙头企业，每天都有数以百万计的消费者登录，积累了大量的消费者数据。2009 年阿里巴巴集团成立阿里云计算，专注利用科技提升消费者购物体验，提供电子商务相关技术服务。自 2015 年阿里云举办第一届天池大数据竞赛起，开始为全国的学者提供脱敏大数据，供其进行研究。其中就包括网购平台的电子商务交易数据，该数据来源于真实的购物网站-“天猫”以及“淘宝”电商平台。

为了更好的对所建立模型进行实证演练，论文使用来源于阿里巴巴天池数据实验室的官方数据集，数据集包含了淘宝自 2017 年 11 月 25 日至 2017 年 12 月 3 日之间，原始数据集包含消费者数量 987994，商品数量 4162024、商品类目数 9439、所有行为数量一亿以上，其中，购买行为占比不到 1%。涉及到有行为的约一百万名消费者，行为包括点击（pv）、购买（buy）、加入购物车（cart）、喜欢（fav）。其它数据字段包含消费者 ID、商品 ID、商品类目 ID、行为类型以及行为发生的时间戳。消数据示例如图 4.1 所示。

表 4.1　原始数据示例

userID	goodID	goodclassID	Search	Time
1	1531036	2920476	pv	2017-11-25 09:21:25
1	3830808	4181361	pv	2017-11-25 15:04:53
1	4365585	2520377	pv	2017-11-25 21:28:01
1	4606018	2735466	pv	2017-11-26 05:22:22

4.1.2 预测目标

在本文的研究中所要解决的主要问题，是如何通过消费者在电商平台所产生的行为数据来对其未来可能发生的购买行为进行预测。近几年来，线上零售额占据了中国零售市场的较大比重，人们越来越倾向于网上购物。现有的网上购物平台包括阿里巴巴的天猫、淘宝以及京东商城积累了大量的消费者数据。如何对海量的消费者隐式反馈数据进行分析，对消费者的行为目标预测，更好的为消费者

推荐其感兴趣的商品，提升服务水平促成交易，是电商平台提升服务水平的关键。

本文希望利用阿里天池提供的 98 万消费者在不同类别商品以及不同时刻产生的行为历史数据，建立基于机器学习的消费者购买行为预测模型。以商品为最小单位，通过研究消费者在某一类商品下所产生的行为序列数据，结合消费者和商品的其它特征指标预测消费者在产生该种行为序列后，是否会对该商品产生购买行为。

通过对消费者的行为数据判断消费者是否会购买商品可以转化为机器学习中的二分类。分类目标分为两种，具体数据中使用 0，1 进行标记：购买、不会购买。若消费者在该类行为序列下产生了购买行为，则该条记录标记为 1。

4.1.3 数据分析

为了更好对所选用数据进行建模以及对原始数据进行处理，论文首先对原始数据进行统计，并通过统计结果进行描述分析，根据数据特点以及统计情况来采取适当的预处理方式对其进行筛选和可视化统计。在原始数据中，消费者行为包括四种，分别是点击行为、喜欢行为、加入购物车行为和购买行为；其次，数据中还标识了消费者产生行为的时间点，以及所对应的商品种类以及商品 ID。对网站来说，时间是判断消费者兴趣的重要参考，且消费者在一天中的可用空闲时间不定，也就是说每个消费者在进行网上操作的时间点有所差别。所以，首先对各个行为所发生的时间点来进行统计。

在对消费者 24 小时点击行为统计中可知，消费者在 17:00 到 22:00 间的活跃度高，也就是消费者在下班或者放学后对网站的登录浏览情况最为活跃，夜晚睡前达到最高，符合人们日常作息规律。其次，日间高峰期是 10:00 到 14:00 左右，消费者在接近午休的时间较为活跃，点击行为较为平稳阶段为 10:00-16:00。由此可见，消费者在线浏览的时间点也与消费者所产生的行为有一定的关联，如图 4.2。

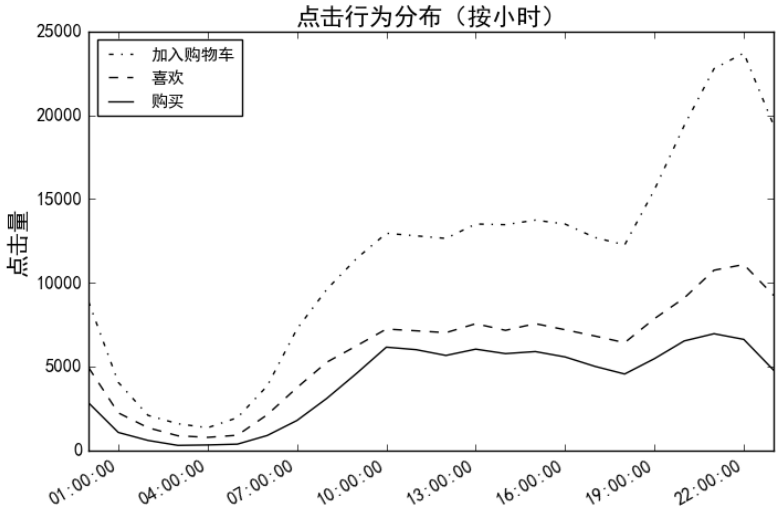

图 4.2　消费者点击行为统计

其次，从定性的角度来分析，消费者在网络购物平台产生购物行为之前都会产生一定的点击、喜欢或者加入购物车行为。而从商品维度对产生购买行为的消费者的行为数据进行分析发现，消费者在产生购买行为之前，大多数消费者都产生 0~5 次左右的点击行为来浏览了解商品详情等信息 ，其中绝大多数产生 2 次浏览行为；在消费者进行浏览后消费者会产生 0~3 次喜欢行为，其中绝大多数为 0 次；加入购物车行为产生 0~2 次，而大部分人都没有进行加入购物车行为，如图 4.2。这与日常对于消费者行为的理解有一定偏差，事实上绝大多数消费者在购买前都不会产生喜欢和加入购物车行为，而是浏览 2 次左右商品信息就进行购买。说明浏览行为在很大程度上预示着消费者下一步的行动目标。综上，消费者在电子商务平台活动时，往往在产生较短操作序列后，便会对商品进行购买。

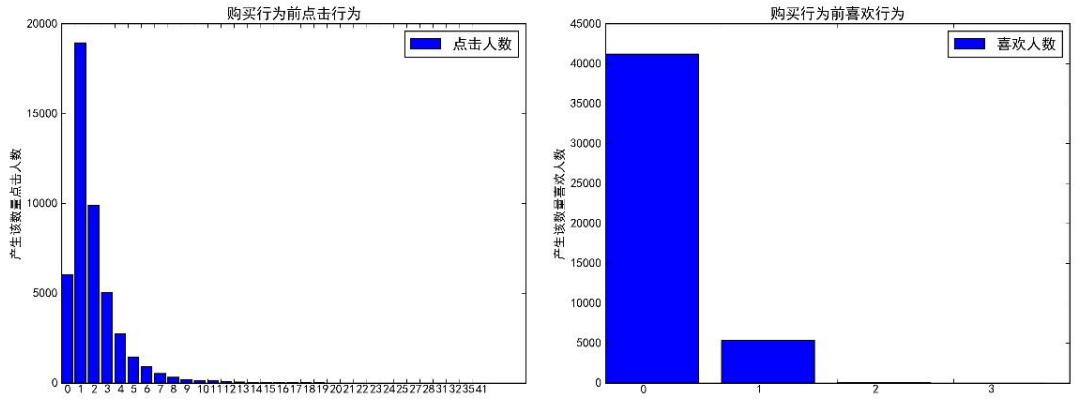

图 4.3　购买前点击行为统计

在对消费者行为的分析中，消费者的转化率是不可忽视的部分。如图 4.4 所示，对原始数据集中 300 万名消费者所产生的行为与行为产生所在商品种类别数量应用散点图进行分析。通过图中所示点的分布状况我们可以分析出，消费者在进行购买前产生的行为中，绝大多数是点击行为，且点击行为所涉及到的商品类别数量不到其产生点击行为数量的二分之一；在加入购物车、喜欢、以及购买行为上，消费者行为数量依旧大于其行为所产生的商品类别数。由此可知，消费者在网上进行购物时，往往会对所选品类进行多次操作，对同类商品反复产生点击、喜欢以及加入购物车的行为后进行购买，其中点击行为占绝大多数。由于本文所应用的数据为五天之内的行为数据，结合图中数据分布可知，消费者在短期内可能会对同类商品进行重复购买，反映了消费者在该时段对于该类商品有极高的兴趣程度。消费者会对同类多数商品产生喜欢的行为，加入购物车行为相比较少一些。

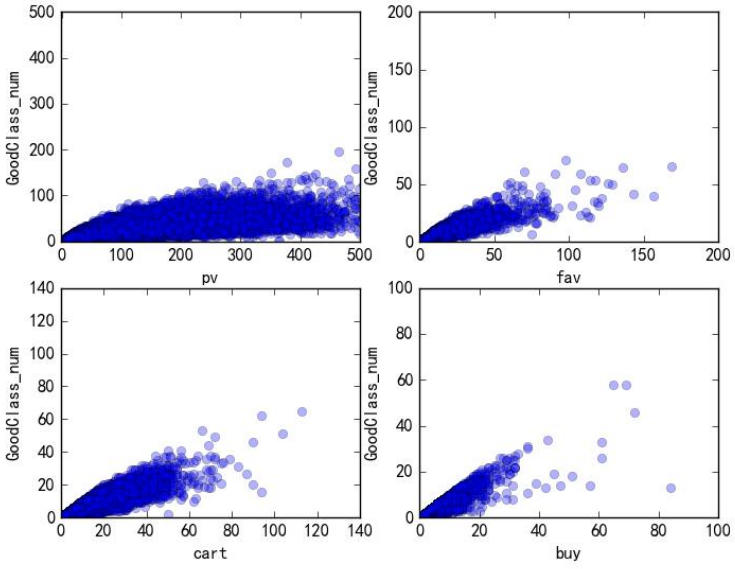

图 4.4　行为与商品类别关联

4.1.4 数据预处理

1.数据清洗

首先，数据清洗是为了去除数据中的噪音，包括重复值、缺失值等。是对数据训练前必不可少的一步。由于本文是针对消费者行为展开研究，而原始数据是

按照每个消费者对于每个商品单个行为进行统计。所以，在数据去噪、去重过后，需要对消费者行为序列统计之后再对其进行筛选。

2.数据统计筛选

为了更加方便的描述消费者对产品的单次消费选择情况，本文首先对数据按照消费者 ID 以及商品类别 ID 和商品编号对数据进行分组统计，按照时间顺序连接消费者行为序列数据作为 behavior_list 字段，消费者对该类别商品按照时间顺序操作次数统计作为 behavior_cout 字段，也就是记录消费者是第几次查看该类别商品，消费者产生行为的时间序列作为 time_list 字段，与行为序列相对应。消费者是否对该产品产生了购买行为使用 buy 字段。例如表 4.2 的第八行，表示消费者 165889 在对 1102540 所属类别下的 829406 产品操作序列为 pv，pv，pv，pv，buy，在产生三次点击浏览行为过后，没有产生喜欢和加入购物车行为的情况下购买了两次。这里需要注意的是，表 4.2 忽略了消费者在不同类别商品点击次序，仅就类别下的单一商品进行了行为统计，即本文主要关注于挖掘统一产品操作中的消费者序列模式。

从所研究的行为序列来考虑，消费者常常会对一些商品产生极少量的操作，兴趣程度不高，浏览行为在其中占绝大多数，此类行为序列将会大大影响模型判断的准确程度。通过对原始数据的查看发现，点击行为（pv）在整体行为数据中占比高达 89%以上，同时对统计后的行为序列进行勘察发现，产生购买行为的消费者仅占消费者总数的 0.153，正负样本比例严重失衡。在机器学习模型中，正负样本比例不均衡会导致训练结果过拟合，也就是说预测结果可能在数据较多的哪一类准确率更高。如此大的数据量级差异将会导致该属性在实验过程中占据主导地位，同时导致迭代收敛速度减慢，严重影响模型的训练。因此，为了加快预测速率和预测的精度，根据预测的目的性以及需求性，需要删除满足一下 3 个条件的数据：

（1）冲动消费或人为恶意刷单嫌疑数据

从定性的角度分析发展，数据中存在一些影响模型精度的无效数据。一些消费者在对产品没产生点击行为时就进行购买，在不考虑刷单行为的前提下，此类消费者行为属于冲动消费行为。冲动消费的发生与商品类别、消费者经济条件、环境因素以及商家促销活动相关；其一，当消费者所要购买的商品是其日常使用

频繁且对其较为了解或者价格较低，不需要过多考虑时会直接购买；其二，若消费者具有一定水平的经济条件，对生活以及商品有了更高的要求，购买风险意识降低；其三，消费者空闲时间较少且较为需要某类商品时，浏览的频率会降低，甚至也可能直接购买；最后，在类似“双十一”或“双十二”的活动期间，商品进行促销或者开展大牌低价抢购活动时，消费者会产生冲动消费。与此同时，网上也存在恶意刷单的可能性，一些商家利用小额红包招揽“伪顾客”来为商铺刷高销售额以及好评数量，在对常规消费者的行为分析中属于干扰数据，此类数据将影响模型整体的有效性。

（2）低频单一操作单个商品

消费者对一类别商品仅产生过一次点击行为，也就是消费者仅仅查看了该类商品的其中一种商品，且产生的操作次数小于 2，仅限于浏览行为，并没有其它加入购物车、喜欢行为来展现消费者对于该商品感兴趣，行为数量过少，难以通过该类行为判断消费者行为习惯。且此类行为序列特征过少、维度低，对于机器学习来说，此类问题将会导致模型准确率降低，不具有与研究意义。

（3）低频单一操作多个商品

对不同类别下多个商品进行查看，行为数目均小于 2，且行为类型仅限于点击行为，未产生具有辨识性的商品喜欢和商品加入购物车行为。该类行为类似一些伪交易行为，伪交易行为指的是一些商家利用红包、返现等方式，雇佣一些伪消费者查看其它几个相关商品后对其店铺商品进行购买模拟了消费者真实浏览商品后购买的行为。消费者每次操作不超过 1，行为长度过短，同样会对模型最终训练所得结果的精度具有一定影响。

处理后的消费者行为序列部分样式如下表所示：

表 4.2　行为序列数据处理情况示例

userID	goodclassID	goodID	time_list	behavior_list
25313	4643350	602041	0，7，7，7，23	cart，pv，pv，pv，pv
246654	381850	4802139	20，20，23，7	pv，pv，pv，pv
159913	4789432	4685183	10，10，8，8	cart，pv，pv，pv
165889	1102540	829406	15，15，22，22，22	pv，pv，pv，pv，buy
286669	4672807	2931238	21，21，15，10	fav，buy，pv，pv
249004	815501	1083709	16，16，16，16，18	pv，pv，buy，pv，pv

3.数据分层

在经过数据统计过后发现消费者活动过程中所产生的行为数目各不相同，在不同的操作次数下消费者行为可能存在差别。对所有序列截取 buy 字段之前的部分序列，按照不同行为次数统计数据集。经过筛选统计后发现消费者在购买行为前所产生的行为数在[3，8]区间内占比最高。为了更细致的对不同数目的行为序列进行研究，将数据按照行为数目不同分别统计为 6 个数据集，模型训练时分为 6 个行为层次来进行研究。在数据集中，同一层次数据均为行为序列的原始长度相等的记录，若该层次中存在一些行为序列不存在购买行为，则序列从最后一个字段开始截取，保证数据集中所有记录的实际长度相等，处理过程实例如图 4.4。

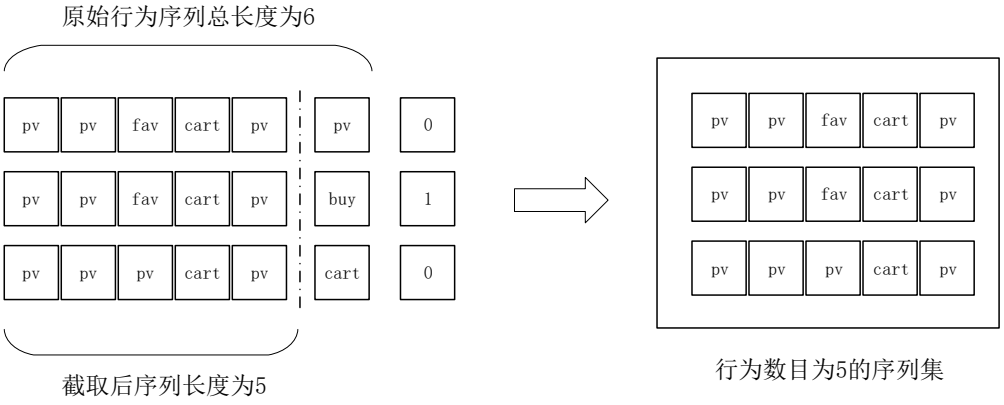

图 4.5　数据处理演示

数据中对于消费者行为的记录使用的是字符串，在 RNN 中不方便训练。因此，在统计完消费者行为序列后，按照点击行为、喜欢行为以及加入购物车行为的顺序，也就是通过分析消费者不同行为所表现出对商品的兴趣程度进行排序，分别将三种行为定义为整型 1、2、3。每种行为可能出现的时间点有 24 种，将行为所出现的时间点处赋值为之前定义的数值，其余 23 个时间点分别填充 0。若消费者行为序列长度为l，则行为序列处理过后长度为$24 \times l$。

4.2 预测模型训练

4.2.1 模型实验环境

本文所用数据利用 Mysql5.6 来进行处理和筛选过程，数据存储为 csv 文件。实验采用 Python3.5 编程和作图分析以及展示，使用开源的 Anaconda4.2.0 来更

好的管理 Python 包。TensorFlow 是谷歌开源的深度学习框架之一，由于 TensorFlow 具有灵活的架构，方便在移动设施、服务器以及计算机一个或多个 CPU 等多种平台上开展计算，其具有高度的灵活性、可移植性、多语言支持。因此，本文的循环神经网络采用 TensorFlow 框架，版本为 1.2.1。

4.2.2 模型成立条件

（1）以行为发生的时刻为时间刻度

在近几年来对消费者行为探索进行购买预测的研究中，大部分学者对于消费者行为的处理是按照日来计算，通过前一天的行为来预测消费者在之后会不会产生购买行为。而在实际生活中，具体的时间点真正影响着消费者浏览商品的频率。消费者往往是通过在休息时间的点击行为对商品产生兴趣。所以，本文在对行为序列进行统计时，缩小了日期的影响，以小时为最小刻度对消费者行为序列进行统计。

（2）不考虑消费者在不同商品间的穿插行为序列

在本文的研究中，以商品 ID 为最小的统计单位对消费者的行为序列进行统计，在不考虑消费者在不同商品类别的商品间产生的穿插行为的前提下，对消费者可能发生的购买行为进行预测。也就是说，当消费者在查看一类商品的过程中，某一次行为发生在了其它类别的商品当中，不会将它加入序列作为当前行为序列的一部分。

（3）不考虑消费者个人私密信息

大多数电子商务平台均能够提供平台数据，但其中与消费者相关的隐私信息往往会加密或者缺失，在大多数情况下这些信息难以获取或容易造成侵权。在本文的研究中没有考虑消费者性别、年龄等私密信息，仅对消费者在平台中所产生的行为进行分析。

4.2.3 模型训练过程

为了验证 BNN-NB 算法的有效性，本文拟将融合算法与传统朴素贝叶斯算法的准确性进行比较。将进行处理过的数据集 300 万条数据进行划分，截取消费者行为序列中 buy 行为前的数据字段。由于消费者在电子商务平台中操作数量有

所不同，也就是消费者对商品的浏览行为或者兴趣程度稍有差别。为了分别研究不同行为数量的人群的网络购物习惯，将数据按照行为数目来分表统计进行预测，利用行为数量大于 2 的所有记录进行训练，按照不同的行为长度进行分层实验，数据处理后正负样本比例为 1:8。抽取其训练集以及测试集以 3:1 的比例进行划分。并分别输入朴素贝叶斯分类器单一模型和 RNN-NB 模型分类器进行学习和测试。

大致实现过程为：将序列输入到结构为 N vs 1 的循环神经网络中，得到 RNN 的序列分类结果，并将所得分类概率值作为新的特征项加入到的其它特征的数据集中。将此特征集作为朴素贝叶斯的输入，最终得出分类结果，结果使用 0，1，表示。

4.2.4 模型评价指标

在对 RNN-NB 模型进行构建之后，必须通过一定的方法和计算指标对模型效果进行评估，通过评估结果来对模型的特征以及参数等方面进行调制直至得到模型最佳分类效果。对于二分类模型的评价来说，常用的方法包括混淆矩阵、PR 曲线（查重率（Precision）-查全率（Recall）曲线）、接受者操作特征曲线（ROC 曲线）和 AUC、KS 曲线（又叫洛伦兹曲线）。PR 曲线在可视化分类器结果中，可以通过调整分类阈值得到不同的 P-R 值从而绘出曲线，其方向与 ROC 相反；KS 曲线图中将绘出两条曲线，其横轴是模型分类所定义的阈值，纵轴是真阳性率与假阳性率，所得 KS 值依旧是两条曲线相距最远的距离值。其中，在正负样本量足够大的情况下，ROC 曲线在对结果的表达中更稳定、简洁且易于理解。

由于本文所研究的问题属于二分类问题，为了更好的描述模型的分类效果，对分类结果的分析将使用混淆矩阵以及 ROC 曲线来对其进行描述，用以评价模型的准确性等分类效果。

1.混淆矩阵

混淆矩阵是衡量分类模型准确率中最直观的、计算最为简便的方法，同时也为 ROC 曲线的绘制提供了数据准备，并用以计算模型分类结果的准确度、召回率、F1 Score 等指标。针对二分类问题的混淆矩阵结构如下表：

表 4.3 混淆矩阵结构

预测 / 真实	Positive	Negative
True	True Positive(TP)	True Negative(TN)
False	False Positive(FP)	False Negative(FN)

其中，列标（Positive/Negative）表示模型分类所得出的结果结果，行标（True/False）表示模型的预测结果与真实值相比是否预测准确。上述指标仅为初步的统计结果，在得到以上数据之后，需要计算混淆矩阵的二级指标，包括准确率、精确率、灵敏度以及特异度且可以进一步通过计算三级指标 F1 Score 来评估模型的整体分类效果。

（1）正确度（Accuracy）：判定正确的数量占总样本的比例。反映模型的判定能力以及准确程度。计算的是整个模型的准确率，也就是模型整体识别样本能力。

$$\mathrm{ACC}=\frac{TP+TN}{TP+FP+TN+FN} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots（4.1）$$

（2）精确度（Precision）：是指判定为正样本的样本数量与被判定样本中的总正样本数的比率，用以评价模型识别正样本的能力。

$$\mathrm{PPV}=\frac{TP}{TP+FP} \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots（4.2）$$

（3）灵敏度（Sensitivity）：是指真实值为正样本的结果中，模型预测结果也为正样本的比重，也就是召回率（Recall）。该指标直接影响着模型的准确率。

$$\mathrm{TPR}=\frac{TP}{TP+FN} \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots（4.3）$$

（4）特异度（Specificity）：在真实值是负样本的结果中，模型预测也为负样本的比重，用以评估模型识别负样本的能力。

$$\mathrm{TNR}=\frac{TN}{TN+FP} \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots（4.4）$$

（5）F1-Score 该指标综合了精确率与召回率也就是灵敏度的产出结果，解决了在日常分类中难以兼顾两者的问题，对于该问题以往学者的解决方法是利用 F-Measure（又称为 F Score），用以计算两者的加权平均值。其取值范围在 0 到 1 之间，所得值大小和模型分类效果正相关。其中 P 代表精确率，R 代表召回率。当参数α取值为 1 时，所得结果也就是常见的 F1 Score。

$$F = \frac{(\alpha^2+1)P*R}{\alpha^2(P+R)} \qquad F1\ Score = \frac{2PR}{P+R} \cdots\cdots\cdots\cdots (4.5)$$

2.ROC 曲线

ROC 揭示敏感度与特异度的相互关系，且引入了另外两个关于计算识别能力的指标，即真阳性率（True Positive Rate，TPR）以及假阳性率（False Positive Rate，FPR）。其中 TPR 也就是混淆矩阵中介绍的敏感度，而 FPR 也就是特异性。图像以假阳性率（FPR）为横坐标，真阳性率（TPR）为纵坐标绘制成曲线。在一些二分类的分类器中，会预先提出相应预测所得概率的阈值，就会导致 FPR 会随着 TPR 的提高而提高。为了更形象的描述模型识别能力，本文将利用 ROC 曲线来表示。AUC 被定义为 ROC 曲线与坐标轴围成的面积，面积越大表示模型预测效果越好，曲线越接近左上角表示模型预测有效性以及准确度越高。ROC 曲线下面积区间在[0.5，1]，也就是 AUC 值在该区间范围内模型具有一定分类效果。当 AUC 值在 0.5~0.7 时表明模型分类效果较差；在 0.7~0.9 时表明模型预测效果良好；当小于 0.5 时表明模型失效。如图 4.5 所示，模型分类效果越好，TPR 值将会更大，最大不大于 1，FPR 更小，最小不小于 0，也就是曲线将更接近左上角，标识 A 处也就是 AUC，面积越大模型分类效果越好。

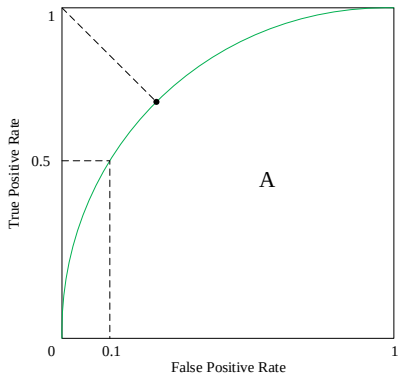

图 4.6 ROC 演示图

4.2.5 实验结果对比

1.朴素贝叶斯模型

通过对预处理过后的特征数据集进行分析发现，数据经筛选过后剩余数据量为 303 万。其中，行为数在[3，8]区间占总体数据量的 80%左右。为了更好的对模型进行训练，提取其中行为数目在所研究区间的数据，并分为六个行为层次分

别输入模型对其进行预测。在朴素贝叶斯模型中不考虑行为序列与时间的关联性，仅将行为序列以及时间序列分别拆分作为预测模型的特征输入项。

从每个行为层次中提取 3 万数据，按照正负样本 1:8 的比例不放回抽样作为训练集；并在余下的数据集中提取 2000 作为测试集，正负样本比例同训练集一致。模型最终要得到的是二分类的结果，每条预测结果使用 0，1 值表示。模型训练结果相关指标如表 4.4，通过该表可以看出，模型在 6 不同行为长度的被测试数据集上都表现出了较好的分类效果。

表 4.4 预测结果比较

	ACC	PPV	TPR	TNR	F1 Score
3_behavior	0.777654	0.32699	0.915	0.856723	0.481475
4_behavior	0.78099	0.3387	0.915	0.85641	0.494395
5_behavior	0.78	0.332192	0.905	0.85525	0.485993
6_behavior	0.79889	0.348315	0.832	0.8825	0.4910521
7_behavior	0.80444	0.354962	0.83	0.88475	0.4972622
8_behavior	0.813333	0.365079	0.82	0.89	0.5052233

预测样本总数量为 52380 条，其中包含 5820 条正样本和 46560 条负样本。整体预测的精确率为 79.25%，召回率为 86.95%。F1 Score 得分随着消费者行为序列变长而增加，说明模型的整体准确度随着消费者行为序列长度的增加而增加，在行为长度增加到 8 个时召回率达到 50%以上。模型对于正样本的识别能力比对于负样本的识别能力强，当序列长度达到 8 时，对正样本的识别准确率最低。所以，该模型在对较短行为序列进行研究时较为有效，序列越长，模型对于数据中正样本识别能力越差，模型整体效果中等，AUC 最大值为 0.9126，如图 4.6。

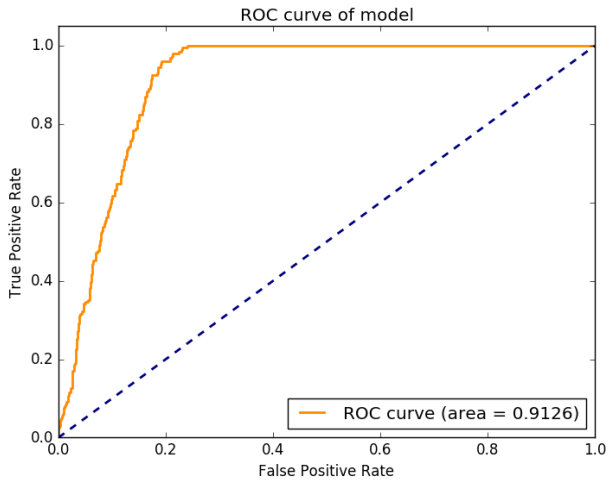

图 4.7 朴素贝叶斯模型最优 ROC 曲线

2.循环神经网络（RNN）-朴素贝叶斯（NB）融合模型

首先，分层提取已经处理好的行为序列数据以及时间序列数据作为行为集合，将行为序列按照不同操作次数分层输入循环神经网络进行训练。训练数据集共 10 万条记录，训练样本正负比例为 1:8。由于数据在通过该模型训练后结果还需要进一步检验，因此设置训练集、测试集比例为 3:1，其中测试集部分不少于 1 万条，以保证有充足的数据进行下一个模型训练。通过不断的变换模型中的参数，得到 RNN 的最佳训练结果，最后将 RNN 网络结构的具体参数设置为如下形式：将学习率设置为 0.006，代表模型训练时参数的更新速率；设置批大小（Batch_Size）为 128；本文建立的 N vs 1 循环神经网络是典型的三层网络，包括输入层，隐藏层以及输出层，其中输出层只输出一个值。为了得到二分类结果，本文采用 N vs 1 的结构对网络进行训练，其中隐藏层节点设置为 10 个。（张兆晨等，2018）由于不同序列长度对模型训练的准确率也有一定的影响，本文将消费者行为序列按照不同长度进行分层，研究不同长度的行为序列对模型结果的影响。

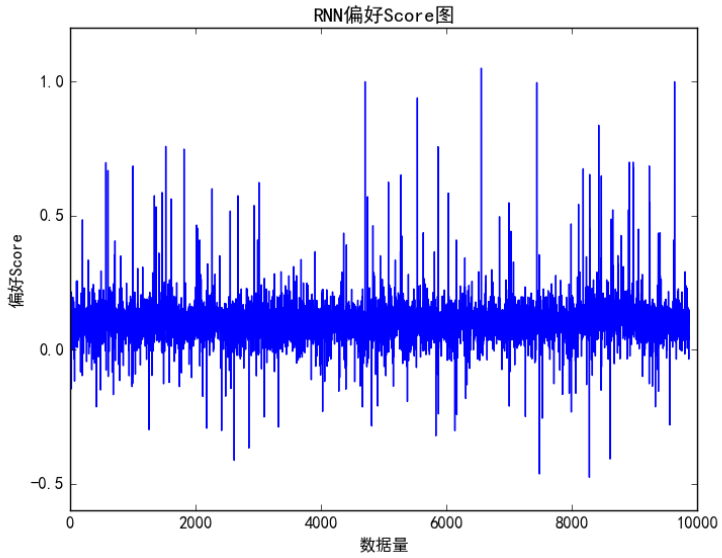

图 4.8　RNN 部分偏好 Score 演示图

图 4.7 展示了部分通过 RNN 演练过后的行为序列所得偏好得分，得分越接近 1，表示该行为序列下的消费者的购买意图越强烈，得分越接近 0 或者小于零，表示消费者的购买偏好较小，不大可能产生购买行为。RNN 仅对消费者的行为序列偏向进行判断，得分并不能客观的表现出消费者是否会产生购买行为，仅是对行为序列的评分，还需要结合其它特征综合对消费者行为进行进一

步的分析。得出偏好得分后，将该得分作为下一步朴素贝叶斯分类器的特征进一步训练。

表 4.5 预测结果比较

	ACC	PPV	TPR	TNR	F1 Score
3_behavior	0.833888	0.394456	0.925	0.8225	0.553064
4_behavior	0.83	0.388655	0.925	0.828125	0.54733
5_behavior	0.82444	0.386718	0.920	0.83375	0.545413
6_behavior	0.83777	0.375968	0.910	0.84875	0.532098
7_behavior	0.8875	0.35199	0.9175	0.8525	0.518788
8_behavior	0.88055	0.35613	0.9181	0.8531	0.513192

朴素贝叶斯预测样本总数量为 52380 条，其中包含 5820 条正样本和 46560 条负样本，样本中均包含 RNN 训练后所得偏好值。模型整体预测的精确率为 84.56%，召回率为 91.92%。F1 Score 得分随着消费者行为序列变长而降低，说明模型的整体准确度随着消费者行为序列长度的增加而减少，在各个行为长度的行为序列训练中，所得 F1 Score 均保持在 0.5 以上。所以，该模型在对较短行为序列进行研究时较为有效，序列越长，模型整体效果会相对减弱，但整体指标保持较为平稳。基于混合模型下的 ROC 曲线面积最大为 AUC=0.9251，整体模型预测效果较好，如图 4.8。

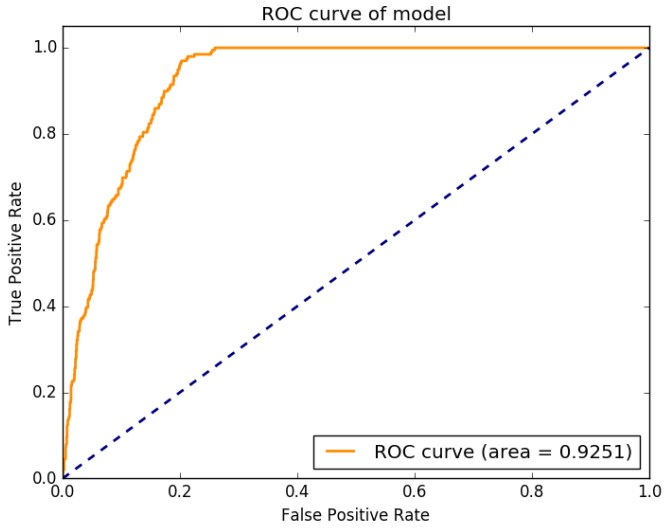

图 4.9 RNN-朴素贝叶斯模型最优 ROC 曲线

4.3 实验结果分析

本文在以消费者操作序列为重点，对消费者行为进行的研究中，为了更好的体现消费者行为序列在消费者网络购物中的影响以及所能显现出的消费者特性，在融合模型中，首先使用循环神经网络（RNN）对消费者行为序列进行分析学习，试图利用不同长度的消费者行为来对其下一步行为进行预测，也就是探讨消费者的点击、喜欢、加入购物车行为对于消费者购买可能性的影响，调节网络参数得到模型的最佳结果。其后，将 RNN 训练结果作为新的特征项加入到剩余特征集中，输入朴素贝叶斯模型得出最终的预测结果。最后实验结果表明，混合模型在准确度、特异性、F1 Score 等指标中较为平稳，且相比较于朴素贝叶斯模型准确度更高，模型最佳训练状态所得 AUC 值达到 0.925，结果表明 RNN 模型对于消费者行为序列的分析评价更能够体现行为的时序性，相比较朴素贝叶斯直接将行为拆分作为特征所得实验结果，更能够准确的判断消费者的行为倾向。

1.结果定量分析

通过上述实验结果可以看出，在对消费者行为的分析中，使用单一的朴素贝叶斯模型进行建模预测时，通过实证得出其准确度依赖消费者线上操作序列长度，消费者行为越长，在各种指标上的表现越好，预测整体准确率轻微上升。对序列长度为 3~6 的消费者数据分析时准确率小于 0.8，对于正样本的识别能力在序列最短时达到最佳效果，序列越长识别正样本能力越差，但依旧保持在 0.8 以上。对于负样本的识别能力一直保持在 0.8~0.9 之间。正负样本识别能力与行为序列长度负相关。整体预测效果得分 F1Score 值在序列最长时达到最佳。

融合模型对于不同长度消费者行为序列表现较为均衡，其准确率能够平稳保持在 0.8 左右，同朴素贝叶斯类似的是，预测准确率同样随着行为序列长度的增大准确率有所提高。融合模型准确率从行为数 6 开始出现明显的增长，当消费者行为序列达到 7 时模型预测效果达到峰值。在对正样本的识别能力上一直保持在 0.9~0.39 区间内，对于负样本的识别准确率波动在 0.8~0.85 之间，整体趋势较为平稳。BNN-NB 模型对于正负样本识别能力较为均衡，并未过多受到消费者行为序列长度的影响，其中对于短序列的识别能力稍强，能更好的处理不同长度的消费者行为信息，当消费者行为数目达到六时，模型显现出最佳效果，如图 4.9。

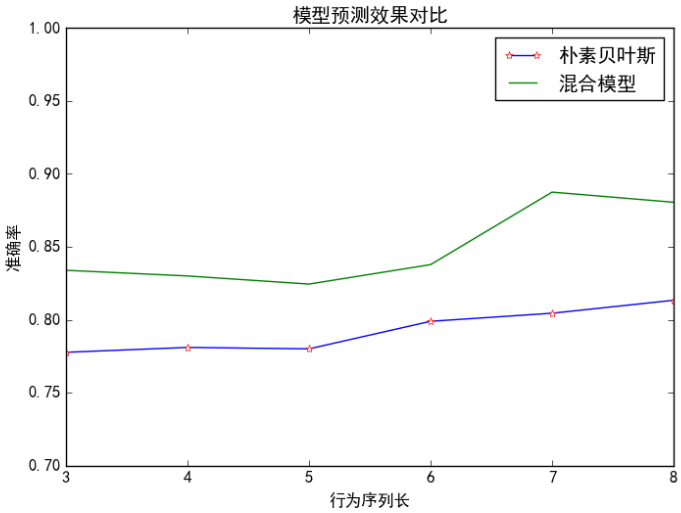

图 4.10 模型预测效果对比

2.结果定性分析

通常我们对消费者的行为规律进行定性分析的过程中，在假定消费者会购买该商品的情况下，消费者有可能会产生加入购物车或者喜欢的行为，从其行为的种类以及数目来判断消费者对于该商品所产生兴趣程度。但在对原始数据进行统计分析中，通过图表发现大多数消费者会在两次点击行为后产生购买行为，也就是说绝大多数消费者在购买商品前产生 2 次点击行为，没有进行加入购物车和喜欢行为，对于商品不会过多挑剔，往往短短的几次行为就会促使其购买该商品。这与我们的常识认定有所不同。所以在本文所研究的问题中，模型若能够对较短的消费者行为预测中表现良好，才能够从定性的角度满足购买行为预测的目的。从模型训练整体指标来看，融合模型在较短的行为序列的预测研究中分类效果较好，相对于朴素贝叶斯单一模型来说更有效。