新浪微博互动预测

小组成员：王学博，陈晓珍，谌丹璐，余睿哲

目录

新浪微博互动预测

摘要：新浪微博作为中国最大的社交媒体平台，旨在帮助用户发布的公开内容提供快速传播互动的通道，提升内容和用户的影响力。新浪微博互动预测的目标是发现能够最快找到有价值微博的方法，然后应用于平台的内容分发控制策略，对于有价值的内容可以增加曝光量，提高内容的传播互动量。

1.背景.

当今社会已经步入大数据时代，大量信息已经成为信息社会最重要的特征，如何更好的利用信息，如何从海量数据中发现知识，创造价值是人类面对的一个重要课题。

而随着社会的不断进步，科技水平的不断提高，人们进行社交活动的方式也日益繁多，新浪微博作为一种代表新时代社交方式的平台，脱颖而出，它旨在帮助用户发布的公开内容提供快速传播互动的通道，提升内容和用户的影响力。新浪微博互动预测的目标是发现能够最快找到有价值微博的方法，然后应用于平台的内容分发控制策略，对于有价值的内容可以增加曝光量，提高内容的传播互动量。

我们基于“阿里云天池大赛-新浪微博互动预测挑战赛”的数据集，首先观察数据，分析数据，挖掘有用信息，选择合适的预测模型，基于数据集进行训练，生成最终的预测模型。

2.问题描述：

对于一条原创博文而言,转发、评论、赞等互动行为能够体现出用户对于博文内容的兴趣程度，也是对博文进行分发控制的重要参考指标。因此本选题的任务就是根据抽样用户的原创博文在发表一天后的转发、评论、赞总数，建立博文的互动模型，并预测用户后续博文在发表一天后的互动情况。

3.方法：

3.1 数据获取：

我们下载了阿里云天池大数据竞赛所提供的微博用户数据集。天池官方提供两个数据集：测试集和预测集。

- 训练集数据是用户在 2015-02-01 至 2015-07-31 时间段发表的微博数据。包括用户标识、博文标识、发表时间、发表一周后的转发数、发表一周后的评论数、发表一周后的赞数、博文内容。博文的全部信息都映射为一行数据，其中对用户做了一定抽样，获取了抽样用户半年的原创博文，对用户标记和博文标记做了加密 发博时间精确到天级别。对数据集简单分析可得到：训练数据共有 122,5088 条（仅三条内容有缺失值），涉及 37251 个用户，每个用户至少发一条博文，博文最多数量前十在 4909~31015 条。
- 测试集数据是用户 2015-08-01 至 2015-08-31 时间段内发表的微博数据，包括用户标识、博文标识、发表时间、博文内容，共有 17,7923 条。我们将会预测测试集中的微博发表一周后转发、评论、点赞的具体数值。

3.2 数据分析与统计：

观察训练数据中，字段并不多，包括：用户标记，博文标记，发博时间，转发数，评论，赞数，博文内容。

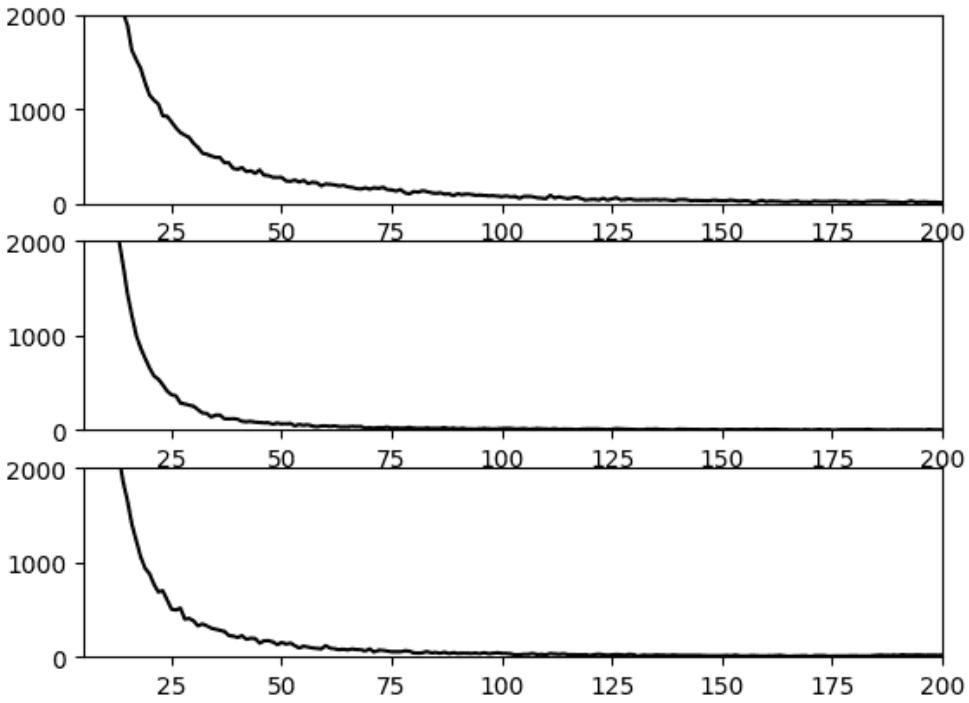
- 用户标记：大多数用户发文不止一条，可通过转发数，评论数，赞数预测该用户的粉丝，以及粉丝的习惯。
- 博文标记：是微博的 id，可看作索引。
- 发博时间：可分解出工作日，节假日，时间段等属性。
- 转发数，评论数，赞数：是预测的目标，也可用于计算用户的特征和分析其相关性。
- 博文内容：可解析出更多特征，如分词聚类，情绪分析，是否包含链接，是否包含表情，是否包含视频，是否自动生成，是否为广告（含：天猫，淘宝，超便宜），长度，是否@谁，是否为转发#，文章分类（新闻，技术，笑话，心情情…）

3.2.1 统计转发，评论，点赞个数（后统称反馈）

个数	0	1	2
转发	0.821	0.063	0.025
评论	0.793	0.068	0.042
赞	0.749	0.103	0.046

可见，如果把所有情况都预测成 0，也会是个效果不错的预测

3.2.2 反馈个数做图



上面列出了转发，评论，点赞的分布图，横坐标是反馈个数（如转发数），纵坐标是该反馈出现的次数，如 0 次转发出现了上百万次（由于影响显示，做图截取掉了）。

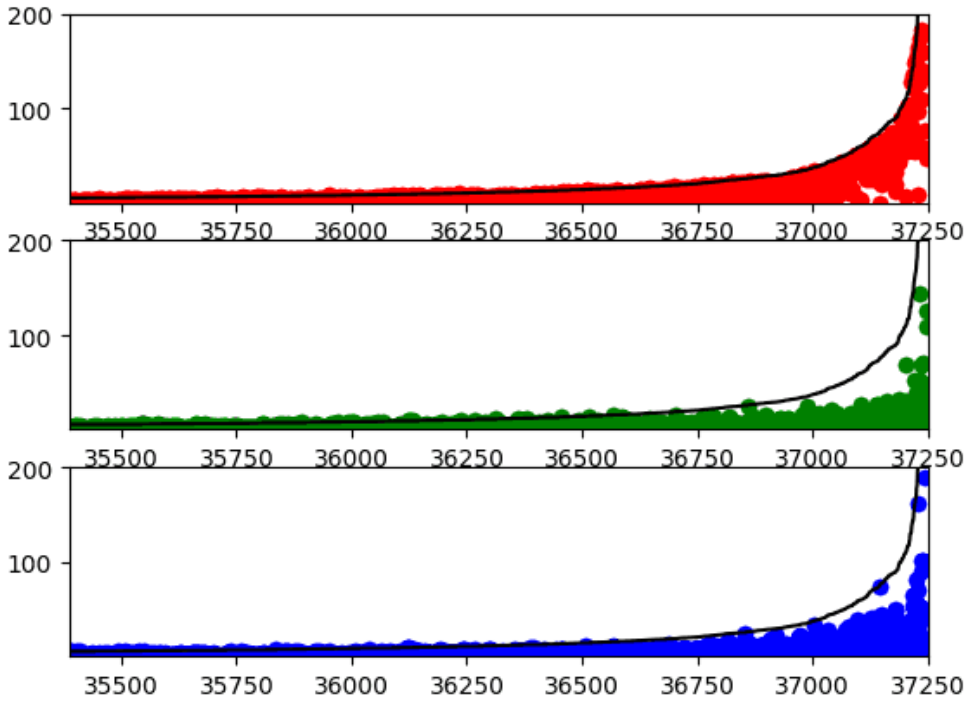
3.2.3 反馈均数

平均每篇获得反馈个数是，转发：3.54，评论：1.26，赞：2.22。可见，虽然大多数人没得到反馈，但被关注的少数人拉高了平均分。

3.2.4 统计用户

训练数据中共 37000 多个用户，人均发文 33 篇，首先用把每个用户得到的转发，评论，点赞的均值加在一起，可计算出关注度，即下图中的黑线，按关注度对用户排序，下图分别显示了关注度和各种反馈之间的关系，以及分布，从中也能看到在 30000 多人里只有几十个人平均每篇的反馈之和超过 100，且以转发为主。

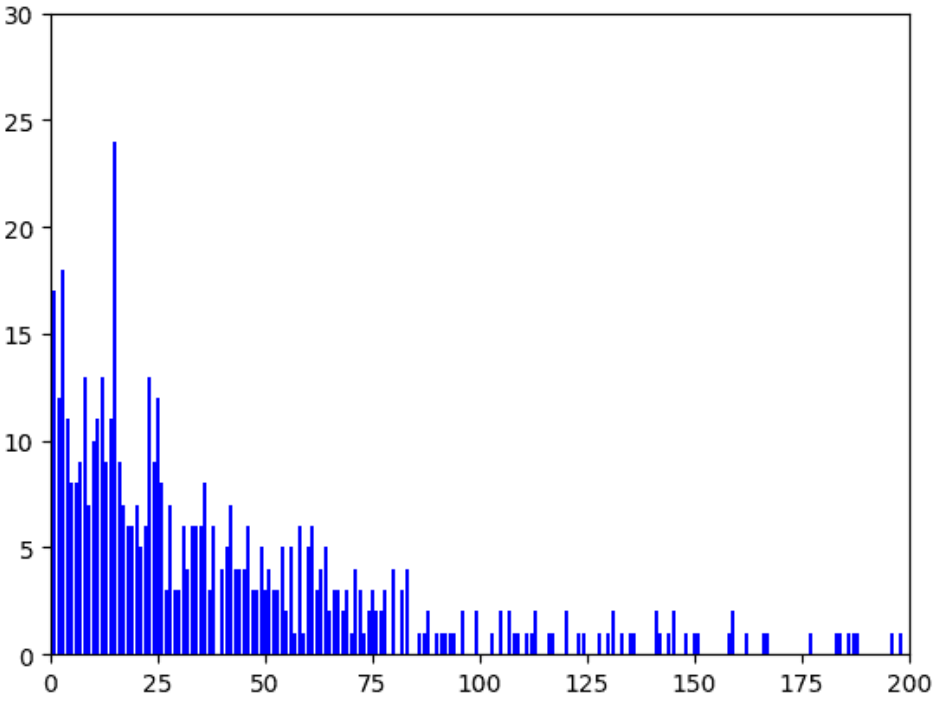
截掉了图的左侧，其中显示有 15000 多人，从未得到过任何反馈，占了全体用户数的 0.412，所以说没人理也很正常。估计可能因为不太使用微博，只发广告，自动生成消息，或者好友太少。



3.2.5 单个用户分析

下面是对某个用户的转发分析，他共发文 733 篇，其中最多的一篇被转发 8949 次，也因为影响显示被截掉了，其中有 167 篇文是 0 次转发，大多数分布在 0-100 次以内。从中还可以估计一下他的粉丝数，至少有 8949 人，方法是 max(f,l,c)。

可见，在粉丝多的情况下，反馈更多地取决于内容。



3.3 数据预处理：

3.3.1 重新对数据进行处理，分别对训练集和预测集进行处理。

- 删除“无用”的用户。因为数据集中存在大量的“无用”的用户，虽然他们看起来比较活跃，会存在微博发表的活跃度特别高，但是对于这部分用户互动（转发、评论、点赞）数平均值为 0，并没有与其他的用户产生交互。这类用户很可能属于“僵尸粉”用户，可以直接处理为最低的档位，不用参与模型的预测。对于“无用”的用户，我们可以设置了一些规则将他们过滤掉，比如发表微博数很大，但是总的互动数基本为 0 的用户。
- 找出测试集中用户没有历史数据的用户集合。找到需要预测的用户之后，我们从训练集中抽取出他们的历史数据，从而进一步发现有 1214 个用户在训练集中不存在历史数据。对于这部分用户我们有两个方面考虑，一个是直接将他们统一处理为最低的档位；另一方面是根据模型，这个模型是根据用户发表微博的特征维度来考虑，而不结合用户的特征。

3.3.2 将训练集分为训练集和验证集

第二步就是划分合理的训练集和测试集。 整个数据集的时间跨度为（2015-02-01 至 2015-07-31），我打算以三个月的数据为训练集，之后的一个月数据集为测试集。因此整个数据集可以划分为以下三个部分：

- 训练集一（674019）：2015-02-01 ~ 2015-04-30 - 测试集一（188029）：2015-05-01~2015-05-31

- 训练集二（621376）：2015-03-01 ~ 2015-05-31 - 测试集二（178823）：2015-06-01~2015-06-30

- 训练集三（573141）：2015-04-01 ~ 2015-06-30 - 测试集三（184214）：2015-07-01~2015-07-31

3.4 模型选择

3.4.1 传统搜索策略

第一种方法就是传统的搜索策略，搜索策略中，也进行两种实验对比：非固定值的预测和固定值的预测。

- 非固定值预测：对于每个 uid，我们首先得到它的（f_min，f_median，f_max），（c_min，c_median，c_max），（l_min，l_medain，l_max），然后：1.固定 c_median(评论数的中位数)和 l_medain（点赞数的中位数），搜索<f_min，f_max>（转发的最小值和最大值）之间的前向值，这会导致（f_medain，c_medain，l_medain）更高的分数，如果存在多个得分相同的结果，我们选择 f_medain 附近的结果。如果不存在获得比（f_medain，c_medain，l_medain）更高分数的任何结果，则比我们选择 forward = f_medain，通过相同的方法搜索评论数，通过相同的方法搜索点赞数
- 固定值预测：大约 80％的训练数据是：0 0 0（forward_count，comment_count，like_count），受此启发，我们为所有使用者尝试一些固定值，并计算他们在训练数据上的得分。

3.4.2 神经网络预测

第二种方法就是神经网络的方法。用来模拟统计数据（min, median, max, mean）和真实的数据（转发数，评论数，点赞数）之间的关系。将每个 Uid 的统计数据作为输入，就是有 f_min，f_median，f_max,f_mean,c_min，c_median，c_max,c_mean，l_min，l_medain，l_max，l_mean，12 个数据作为输入，输出端是 c_count,f_count,l_count。在拟合高次方程之间需要对数据进行预处理，由于 37251 条用户的统计数据中，大部分都是 0，并且有 21507 的用户的数据信息低于 10 条，所以选了删掉一部分统计量比较少的用户统计数据，减少这些小数据对于模型的影响。损失函数采用的最小均方误差。采用梯度下降的训练方法。

4.实验结果

4.1 固定值预测：

固定值预测结果在验证集上如下所示：

预测分数（%）
0 0 0 34.10%
1 0 0 29.23%
0 1 0 35.01%
0 0 1 32.20%
1 1 0 29.30%
1 0 1 29.39%
0 1 1 33.45%
1 1 1 13.46%
2 0 0 7.04%
0 2 0 29.93%
0 0 2 28.22%
0 1 2 12.85%

4.2 非固定值预测：

最小值 34.17%

最大值 8.08%

中位数 40.94%

4.3 神经网络预测结果：

基于神经网络的方法回归出每个用户的统计特性和真实的值之间的关系。网络设置 1 层隐藏层，隐藏层单元个数 256，输入为 12 个统计特征，输出为评论数，点赞数，转发数三个值，预测最终的分数为 31.25%，效果不尽人意。