微博虚假用户检测

——《社交网络挖掘》实践项目报告

吴茜茜 17300200036@fudan.edu.cn

何晓昕 17307130334@fudan.edu.cn

1 简介

本次课程项目我们研究新浪微博的虚假用户检测。

我们随机选择在蔡徐坤微博下评论的两千多位用户，爬取了他们个人信息和微博内容，人工标注类别，构成了本次实验数据集。在特征工程阶段，我们在原始数据集中提取了用户的社会特征、行为特征和微博内容特征，用 12 维的向量表征用户信息。在多个机器学习分类器上进行了实验，模型在 AUC、精度、召回率、F1 得分都表现良好，超过 90%。

为比较正常用户社交网络和虚假用户社交网络的差异，我们在原始数据集中随机选择了一位正常用户和一位虚假用户，爬取他们三阶粉丝，构成正常用户和虚假用户的社交网络。我们从网络的度分布、连通性、同质性等角度对两个网络的差异进行了充分比较。

2 数据集

本次课程项目我们构建了两个数据集 D_1 和 D_2。其中 D_1 完成虚假用户检测任务，D_2 比较两个网络的差异。

D_1 从蔡徐坤微博下评论的用户中随机选了两千位用户，爬取了他们的详细个人信息和微博内容（对每位用户，我们爬取的微博条数为 $min\,(weibo\_num, 1000)$），其中个人信息包括用户 id、昵称、地区、粉丝数、关注数、微博数，每一条微博内容包括微博 id、微博内容、发布时间、是否转发、是否包含 url、是否原创、长度、点赞数、评论数、转发数。对每一位用户，我们人工标注类别，1-虚假用户，0-正常用户。最终我们的数据集包括 2,234 位用户和 391,196 条微博内容，正负样本的比例接近 4:1。

D_2 是从 D_1 中随机选择了一位正常用户（id 为 6529218955，粉丝数 145，关注数 227）和一位虚假用户（id 为 6099859539，粉丝数 20，关注数 60），爬取了他们三阶粉丝网络构成。由于用户“新浪新手指南”几乎是所有微博用户的粉丝，给数据带来一定噪声，我们在预处理阶段两个网络中都剔除了该用户。最终 D_2 包含三部分：

- 网络中用户信息，包括用户 id、昵称、所在地区、关注数、粉丝数、微博数
- 网络中有向边信息，以 $(source\_id, target\_id)$ 的方式构成
- 网络中各个用户的微博内容，我们爬取了每个用户最近的 20 条微博

由此，我们构建了正常用户和虚假用户的社交网络，这两个网络的规模都在 900 左右。

3 网络可视化与分析

3.1 网络可视化

这一部分，我们分别对网络中的**关注关系**、**转发关系**和**地域间关注关系**进行可视化，并比较了两个网络的差异。

- **网络中关注关系**
  将两个网络中每个用户视为节点，粉丝节点向关注节点间有边的方式形成关注关系可视化。下图是两个网络可视化的结果，可以看到正常用户社交网络连接稠密，在外围形成小的聚集圈；而虚假用户的网络则是由一个个小的聚集圈构成，“小世界”现象显著，整个网络相对稀疏。

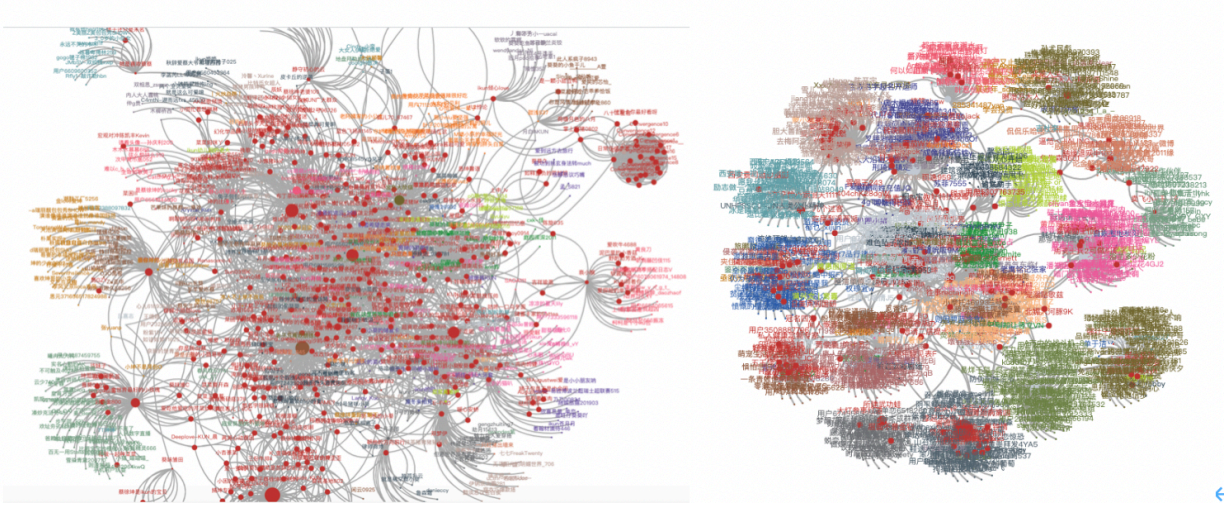

图 1: 关注关系可视化, 左图为正常用户网络，右图为虚假用户网络

- **转发关系可视化**
  新浪微博中用户间存在微博转发关系。前面已介绍，我们爬取了每个用户最近的二十条微博内容，以此查看两个网络间的转发关系。正常用户网络中转发关系非常可观，比如用户“一点繁星 kk”的微博内容被 4 位好友转发过。这样微博内容间的转发关系印证了正常用户网络间用户的频繁互动与社交。但是虚假用户网络并不存在用户间的微博转发关系，即网络中用户的互动趋于沉寂。

- **地域间关注关系可视化**

  我们在两个网络中各选择了一个节点，比较他们的位置与好友的位置 (位置信息由用户个人信息中的“地区”得到，这里我们仅具体到省份)。可以看到，我们选择的正常用户在江苏，他的好友 (包括关注、粉丝) 在江浙沪地区形成聚集。我们选择的虚假用户在沈阳，好友却分别在北上广集中。因此，我们得出：正常用户与好友在地理位置上更接近，而虚假用户与好友在地理位置上更远。

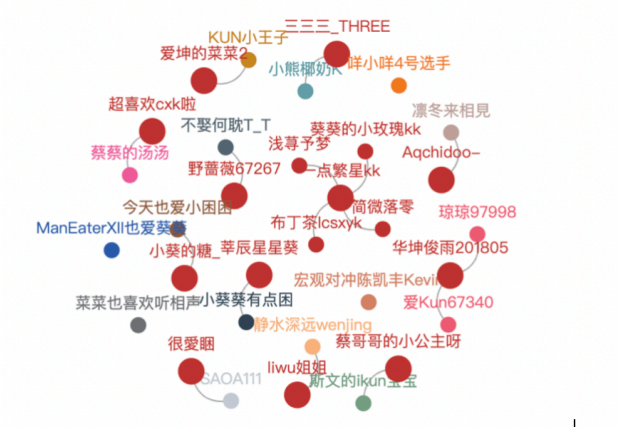

图 2: 正常用户网络间转发关系可视化

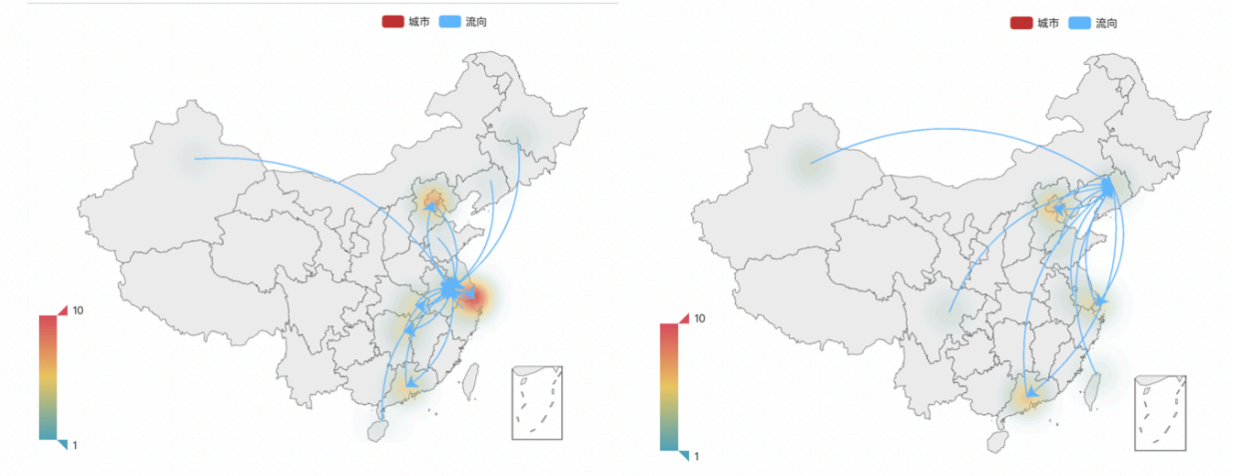

图 3: 地域间关注关系可视化，左图为正常用户，右图为虚假用户

3.2 网络属性比较

属性	NormNet	SpamNet	说明
节点数	909	891	-
边数	2659	1423	-
平均度数	2.925	1.597	NormNet 连接更稠密
强连通分量数	420	733	NormNet 连通性更好
平均路径长度	5.22	3.563	NormNet 接近六度空间理论，SpamNet 符合小世界现象
聚类系数	0.169	0.18	SpamNet 小世界现象显著

表 1: 网络属性比较,NormNet 正常用户网络，SpamNet 虚假用户网络

此外，我们还研究了网络的同质性，主要是用户的微博内容与好友微博内容的相似度。我们随机选择正常用户和虚假用户，比较他们的微博内容与好友微博内容。从词云图中我们直观地看到正常用户与好友微博内容相似度高、有多个公共词；但是虚假用户与好友微博内容相似度低。

3.3 用户活跃度差异

正常用户与虚假用户在活跃度上差异明显。就一天中活跃时间看，正常用户在中午和晚上有两个高峰，符合他们日常通勤的作息。但是虚假用户主要在夜间进行发微博等活动。就微博数量按月统计的结果来看，正常用户在一年中各月份的微博数量大致相当，分布均衡。但是异常用户发微博则呈现出短时突发性的特征，集中于某个月份或时间段。

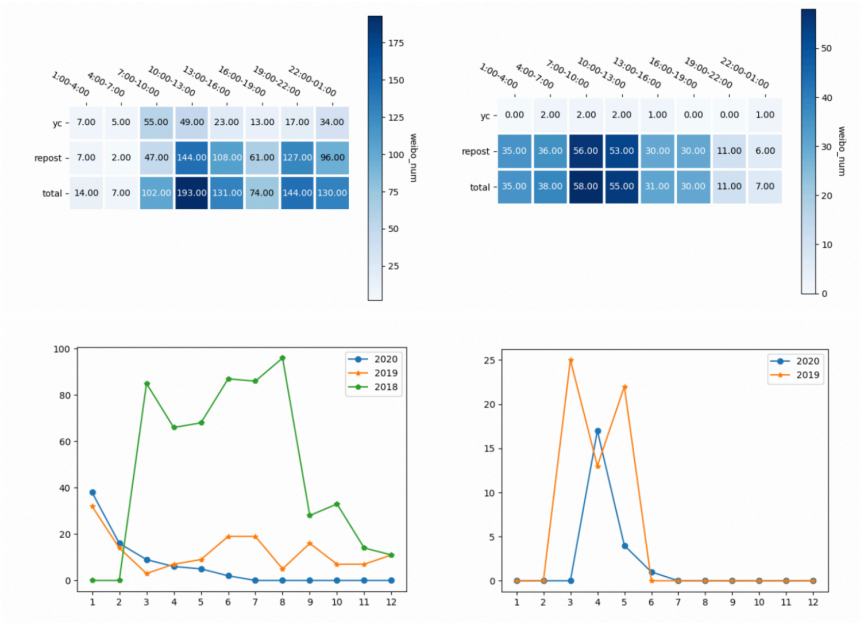

图 4: 活跃状态比较，左为正常用户，右图为虚假用户

4 模型、评价方法和实验结果

4.1 特征生成与分析

结合现有研究，从社会关系、用户行为和内容信息 3 方面选择 12 个区分度大的特征。

4.1.1 社会关系

关注数 (F1) 为当前微博用户关注其他微博用户的总数。虚假用户的关注数远少于正常用户，其原因可能是虚假用户仅用于进行虚假宣传、网络营销等具有明确目的性的活动，毋需关注其他用户。

粉丝数 (F2) 为当前微博用户的粉丝总数。虚假用户的粉丝数要明显少于正常用户，可能因为垃圾用户没有正常的社会关系，很少有人会关注它。

关注粉丝比 (F3) 为关注数与粉丝数的比值。

4.1.2 用户行为

月均微博 (F4) 为用户每月所发的微博数，可衡量用户所发微博的活跃频度。虚假用户的活跃度要低于正常用户的。原因可能是它会为某种利益目的进行短期非法活动，当活动结束就不再发送博文。如有些流量粉可能为某位明星“打榜”之后就停用。

时间间隔 (F5) 为用户最近一次发布微博距数据采集结束时刻的时间间隔（单位为天）。虚假用户的时间间隔比较大，原因同 F4。

转发比 (F6) 为转发的微博与微博总数之比。较正常用户，虚假用户常会转发其他用户的微博，以达到其扩散某些非法信息的目的。

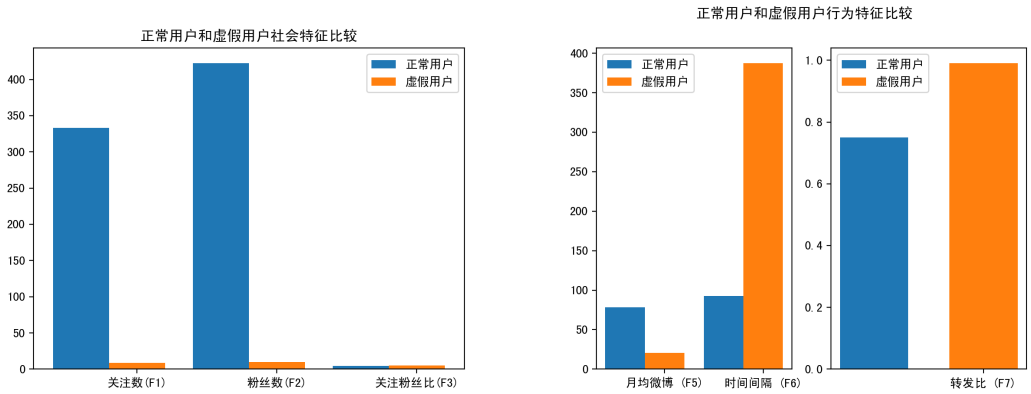

图 5: 社会特征和行为特征

4.1.3 内容特征

URL 链接比 (F7) 为含有 URL 的微博数量与微博总数之比。

微博评论比 (F8) 为收到的评论数与微博总数之比。正常用户会与好友就所发微博进行交流，而虚假用户由于其微博的信息价值低且没有真正的“好友”，故所发的微博一般不会有用户去评论。

原创微博评论比 (F9) 为收到的评论数与原创微博总数之比。

微博平均长度 (F10) 为博文的平均长度。

博文间余弦相似度 (F11, F12) 分别计算用户在相邻两天、一天内所发微博的余弦相似度。据观察，虚假用户通常在一天之内大量转发相似内容的博文。

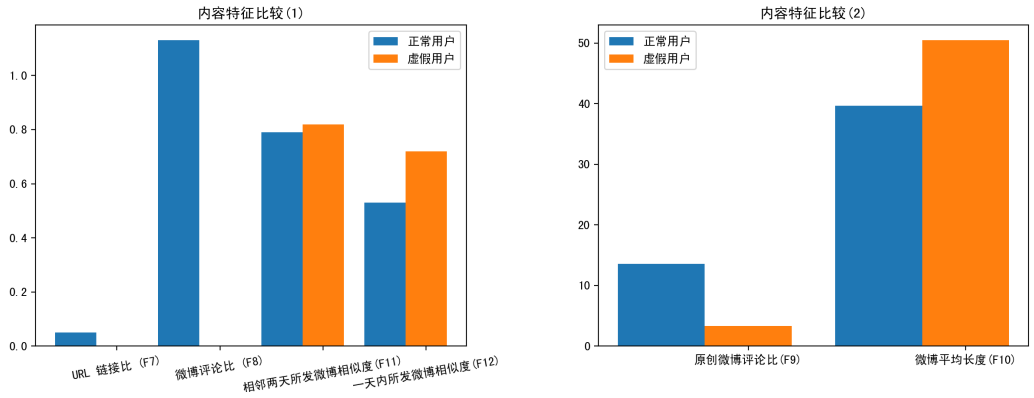

图 6: 内容特征

4.2 模型

由于分类器对样本数量的敏感度、特征之间相关度的敏感度等均不相同，故选择不同的分类器得到的分类效果往往不同。使用的 7 个分类器如表 2 所示。

名称	说明
Naive Bayes(NB)	基于贝叶斯定理与特征之间独立假设基础之上，根据某对象的先验概率利用贝叶斯公式计算出其后验概率，选择具有最大后验概率的类作为该对象所属的类
logistic regression(LR)	使用逻辑回归 sigmod 函数来计算后验概率，根据后验概率对所给对象进行分类识别
support vector machine (SVM)	建立在统计学理论中的结构风险最小化准则基础上，原理是将低维空间的点映射到高维空间，使它们成为线性可分，再使用线性划分的原理来判断分类边界
k-nearest neighbor (kNN)	一种基于实例学习的非参数估计的分类方法，计算新样本与训练样本之间的距离，找到距离最近的 k 个邻居，如果邻居的大多数属于某一个类别，则该样本也属于这个类别
AdaBoost.M1 (AdaBoost)	一种提高给定学习算法精度的方法，使用同一个训练集训练不同的弱分类器，然后把这些弱分类器集合起来，构成一个强的最终分类器
decision trees(DT)	一种简单且快速的非参数树状分类方法，利用信息增益率来选择特征，将信息增益率最大的特征作为决策树的分裂节点，每个分支均重复这一过程
random forest (RF)	以决策树为基本分类器的一个集成学习分类方法，它包含多个由 BA 集成学习技术训练得到的决策树，当输入待分类的样本时，最终的分类结果由单个决策树的输出结果投票决定

表 2: 分类器

4.3 评价方法与实验结果

实验结果如图 7 所示，每个模型各个指标都可以达到 90% 以上的结果，说明我们所选取的特征十分具有区分度。其中，效果最好的是随机森林，它的 AUC、准确率和 F1 得分都非常接近于 1。

5 前端

为更好地演示项目，我们开发了一个基于 django 的网页。主要功能为生成《微博关注检测报告》和对我们的所做的工作进行介绍。由于已在小组汇报中演示以及篇幅限制，不再赘述，可访问 此链接 了解更多。

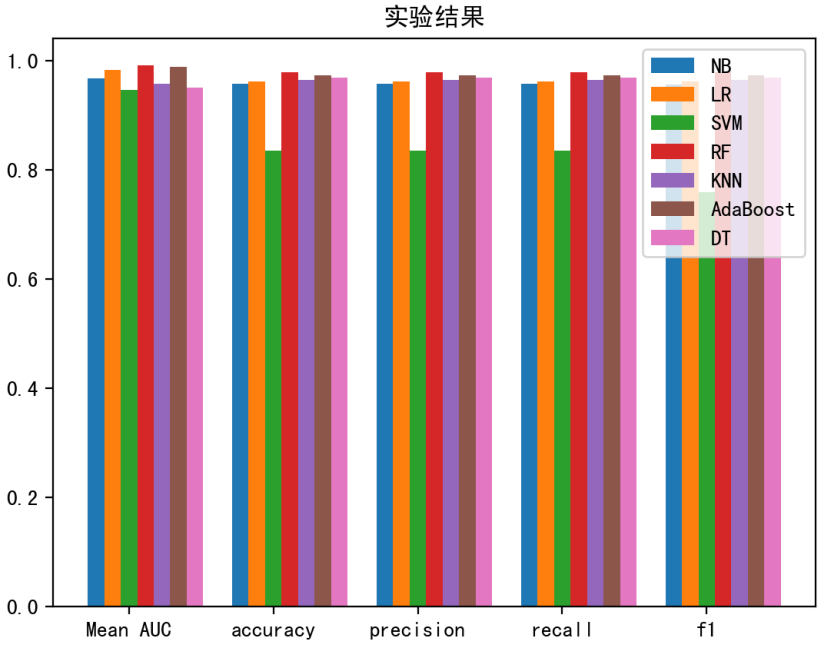

图 7: 实验结果

6 小组分工

- 何晓昕：数据处理、特征工程、分类模型、效果展示前端
- 吴茜茜：数据集构建、两类网络可视化与分析、效果展示后端

7 总结

总的来说，本次课程设计我们有如下贡献：

- 对微博社交网络中的关注关系、转发关系、地域间关注关系进行可视化。对正常用户网络和虚假用户网络的度分布、连通性、微博内容的同质性、用户活跃状态进行了充分的分析和比较。
- 从微博用户信息和微博内容中抽取了社会关系、用户行为和微博内容这 3 方面 12 个区分度大的特征，使用机器学习分类器训练模型，最优模型（随机森林）效果良好，在 AUC、准确率、F1 都非常接近 1.
- 实现了一个个性化微博报告的网页应用。用户输入自己的 id，可以得到自己的微博内容关键词词云图、好友在中国地图分布图、沉寂关注（半年没有发微博的关注）、异常粉丝（使用我们预训练的分类模型，检测出用户的虚假粉丝）等。