应用机器学习来刻画社交网络的智能体基于模型

Haoyuan Li[0009−0003−7608−1396], Lidia Conde Matos[0009−0001−8240−154X], Eduardo César Galobardes[0000−0002−9729−8557], and Anna Sikora[0000−0003−0090−4109]

Universitat Autònoma de Barcelona, Cerdanyola 08193, Spain

摘要 如今，社交媒体网络在我们的生活中变得越来越重要，研究社交媒体网络的必要性也日益增强。随着平台上的数十亿用户和持续不断的更新，建模社交网络的复杂性极大。基于代理的建模（ABM）被广泛用于研究社交网络社区，使我们能够定义个体行为并模拟系统级演化。它可以成为测试算法如何影响用户行为的强大工具。为了充分利用基于代理的模型，优秀的数据处理和存储能力是必不可少的。高性能计算（HPC）提供了一个理想的解决方案，擅长管理复杂的计算和分析，特别是对于大量或迭代密集型任务而言。我们利用机器学习（ML）方法来分析社交媒体用户，因为它们能够高效地处理大量的数据，并提炼出有助于理解用户行为、偏好和趋势的见解。因此，我们的提案涉及使用 ML 来描述用户属性并开发一种通用的用户模型，用于在 HPC 系统上进行社交网络的 ABM 模拟。

Keywords: 社交网络 · 机器学习 · ABMS · 高性能计算

1 介绍

我们的长期目标是定义一个基于代理的社会媒体网络模型，并研究它们如何用于更好地理解这些网络的动力学。鉴于社交网络影响力的不断扩大，理解其动力学变得至关重要。ABM 允许研究人员模拟网络中个体的行为，并研究他们的互动如何产生整个网络的涌现特性。

我们旨在将 ABM 在 HPC 平台上实现，以充分发挥 ABMS 的潜力。由于代理规模和复杂性，需要在 HPC 上实施并行社会网络模拟器。我们打算首先定义模型，然后探索如何使用 ABM 来模拟用户行为、信息传播，并分析其动态。通过我们的研究，希望为社交媒体网络的基本机制提供洞见。

为了实现我们的目标，我们首先需要对社交网络用户进行建模。然而，由于缺乏标注数据集，这具有挑战性。我们也旨在理解每个类别中用户属性

的统计分布以创建合成数据集。在这项工作中，我们介绍了使用机器学习技术进行用户分类的方法以及通过使用 Twitter 数据集获得的初步结果。

2 研究方法论

我们从收集社交网络用户数据开始，并创建了自己的包含可以分析以表征用户的属性的数据集。然后，利用收集到的所有信息，我们专注于通过主成分分析（PCA）[1] 评估各种属性。随后，我们应用一种机器学习方法根据这些识别出的属性对用户进行分类，然后获得每个组中每种属性的分布模式以生成用户模型。这项工作的总体步骤如图 1所示。

图 1. 提出的 方法论 的一般步骤

2.1 数据准备

为了构建有效的数据集，充分的数据收集和处理是必不可少的。许多社交媒体平台提供了 API，允许访问用户生成的文本、图像和视频以及相应的元数据。我们选择了最具特色的属性，并将原始数据转换为仅包含这些选定属性。

一个数据集通常包含数值型和二进制属性，捕捉多样化的用户特征和行为。由于这种多样性，评估每个属性在用户分类中的有用性至关重要。此评估旨在识别最具信息量的属性以实现有效的分类。

我们采用了主成分分析（PCA）进行属性评估。因为它可以减少具有众多相互关联变量的数据集的维度，同时尽可能保留变化量。这是通过转换到一组新的无相关性变量来实现的，即主成分（PC），它们按顺序排列，使得前几个主成分保留了原始所有变量中的大部分变化 [2]。为了确定在 PCA 中要保留的主成分数量，存在几种常用方法，包括特征值准则、碎石图以及领域知识考虑。在这项研究中，我们将通过借鉴先前的研究见解来决定组件的数量。解释属性的相关性通常意味着查看 PC 中的负载，并据此决定哪些变量对该组成部分重要 [3]。

执行 PCA 后，我们可以评估收集的所有属性，并确定保留哪些属性进入下一步。

2.2 使用机器学习进行聚类

为了进行聚类，我们采用了机器学习（ML），它提供了处理高维度和异构数据的能力，确保了稳健的分类。ML 方法可以处理各种类型的数据，包括数值型、类别型和混合属性类型。在 ML 中，分类主要涉及监督和无监督的方法。监督方法依赖于带有标签的数据集进行训练，而我们的数据集中缺乏标签。因此，我们选择了无监督方法来对数据进行聚类。对于无监督方法，有许多选项可用，例如 K 均值 [4]、K-原型 [7]、层次聚类 [5] 和 BIRCH[6]。我们选择了 K 均值，因为它在计算上更高效，易于实现，并且能够产生良好的结果。

K-均值是一种流行的无监督学习算法，用于聚类，它侧重于在没有标签或先前答案的情况下发现数据中的隐藏模式或结构 [8]。该算法包括两个独立的阶段。首先，它随机选择 k 个中心点，其中 k 的值是预先固定的。其次，将每个数据对象分配到最近的中心点 [9]。通常情况下，聚类的数量通过使用肘部法定义，然而，在本研究中，我们根据分布和先前对社交网络用户典型分组的分析选择了固定数量。当所有数据对象都包含在某些簇中时，聚类过程结束。

2.3 特征分析

最后一步是分析每个组内属性的统计分布。为此，我们需要识别这些分布。该方法涉及对通过 K-means 算法获得的每个聚类内的属性分布进行检查。我们最初分析了直方图以了解与每个属性相关的可能分布。接下来，我们进行了拟合优度测试，使用残差平方和 (RSS) 指标将观察到的频率与理论分布预期的频率进行比较。最后，我们通过概率图评估了拟合优度。这使我们能够生成反映真实用户特征分布的新用户，并通过比较这些分布来验证它们。

3 初步结果：用例

在本节中，我们将该方法应用于使用真实的 Twitter 推文数据集来刻画用户。

3.1 步骤 1：数据准备

本工作中使用的数据由一组来自 2022 年 11 月的一般 Twitter 流的 JSON 文件组成，该数据包含超过 130000 名用户的推文 [13]。我们处理了这些数据以获得用户属性。表 1的第一列显示了我们选择的所有属性。

然后，为了基于这些属性有效分类用户我们使用了 PCA。表 1显示了第一主成分中的属性载荷。我们可以看到大多数数值型属性的载荷高于二进制属性，表明它们对主成分的影响更大。二进制属性不是分类用户的最重要部分。最终我们选择了 friends_count、followers_count、listed_count、favourites_count、statuses_count 用于聚类。

表 1. 所有属性的主成分负载量

Attributes	PC1	PC2	PC3
location	5.313587e-09	1.467293e-07	-5.952920e-07
protected	-0.000000e+00	-1.110223e-16	-1.110223e-16
verified	3.946670e-08	3.000574e-08	5.879996e-08
followers_count	**9.999922e-01**	-3.192501e-03	-1.511976e-03
friends_count	8.453385e-04	4.239099e-03	-1.111189e-02
listed_count	1.544425e-03	3.381150e-04	2.707960e-04
favorites_count	-1.069482e-03	**1.370760e-01**	**-9.904922e-01**
status_count	3.366800e-03	**9.905463e-01**	**1.371111e-01**
created_at	-3.165097e-07	-3.738327e-06	1.040501e-05
geo_enabled	1.439318e-08	2.101619e-07	-6.423289e-07
default_profile	-2.844280e-08	-2.706425e-07	6.173020e-07
default_profile_image	0.000000e+00	-0.000000e+00	-0.000000e+00

3.2 步骤 2：使用机器学习方法进行聚类

在准备好所有数据之后，我们可以应用 K-means 对数据集进行聚类。在这种情况下，我们首先使用肘部法来获得最佳的簇数，这给出了 K=7。然而，定义得到的 7 个簇中的每一个都具有挑战性，因为某些簇表现出属性分布不明显的情况。因此，基于工作 [10] 并结合对我们案例的分析，我们决定自己确定 K 值。在比较了分布之后，我们假设 K=4。图 2显示了每个簇中用户的分布，表 2显示了每个簇中用户的百分比。

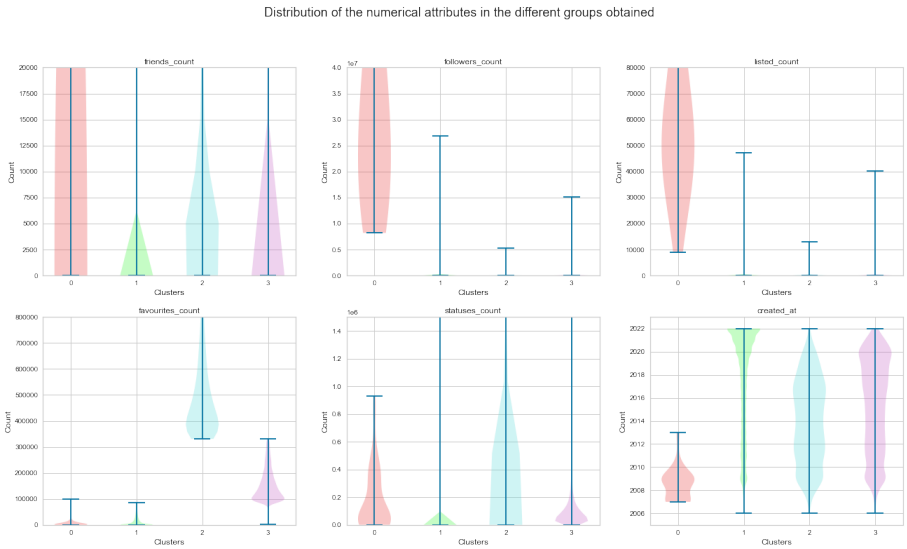

图 2. 不同聚类中数值属性的分布

表 2. 通过 K-均值算法获得的每个聚类中的用户分布。

cluster	users	distribution
0	61	0.000
1	115096	0.878
2	14158	0.108
3	1774	0.014

结合该表与这 4 个聚类的属性分布，我们可以按其特征定义每个聚类如下：

聚类 0：影响者

这是最小且最具特色的聚类。属于这个聚类的用户拥有大量的朋友和关注者，并被包含在许多列表中。相比之下，他们拥有的收藏数和状态数较少。此外，他们的账户创建年份倾向于集中在平台早期的几年内，所有账户均不晚于 2013 年创建。另外，该聚类内的所有用户均已验证，这一特征在其他聚类中明显不太普遍，并且已修改了其用户资料的主题或背景。

集群 1：标准用户这是最大的一个集群。这些用户往往拥有较少的朋友、关注者、收藏和状态，且他们不包含在许多列表中。此外，他们的账户创建

年份倾向于集中在平台后期的几年。该集群内有相当比例的用户没有自定义其个人资料的主题或背景，并且指定位置或激活地理定位的用户数量最少。

集群 2：共享者这是第二大集群。这些用户的特点是拥有大量收藏和状态更新，反映出他们在平台上的高活跃度，并表明他们与他人互动或表达思想和意见的频率很高。

聚类 3：潜伏者这是第二小的群组。这些用户的特点是拥有中等数量的收藏，而他们的状态值较低。我们可以得出结论，他们在社交媒体平台上花费时间，但并不积极贡献内容。他们可能会享受浏览或参与吸引他们兴趣的帖子，但不会主动寻求他人的关注。

3.3 步骤 3：特征分析

聚类后，我们使用概率图评估了每个分布的拟合优度。点与线之间的紧密对齐表明数据非常接近理论分布。虽然大多数分布表现出很强的拟合性，但某些属性并未很好地符合任何特定分布。作为潜在的改进方向，我们设想探索将这些属性拟合到复合分布模型中的可能性，旨在提高未来分析的准确性。图 3的第一行显示了我们获得的不同聚类的不同属性示例，其中数据点被绘制在理论分位数线上。

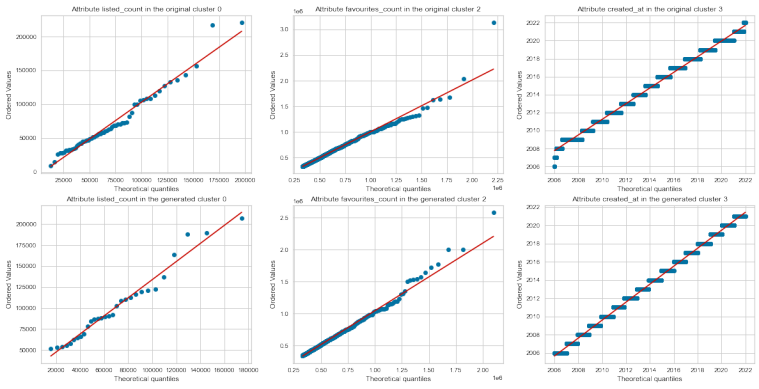

图 3. 原始和生成用户的概率图

获得分布后，我们构建了一个包含这些分布的模型，并使用该模型将分布与另一数据集的数据进行比较。新数据集采用与原始数据集相同的方法获取，但包含不同的用户和推文。新数据集通过 K-均值进行了聚类。将模型与新数据进行比较显示了显著接近的拟合度。

在此验证之后，我们基于该模型生成了新用户。生成的用户数量与原始数据集中的数量相匹配，并且每个聚类中用户的百分比相同。我们也比较了这些生成的用户，如图 3第二行所示的例子。我们提供了具有匹配属性和匹配聚类的生成用户的概率图，与原始用户所描述的一致。比较结果显示，当与我们原始数据集中的聚类进行对比时，相似性非常明显，这证实了识别出分布的准确性。

4 结论

在这项工作中，我们创建了一个社交媒体用户数据库，并使用机器学习技术对其进行聚类。然后我们在每个集群内确定属性分布以构建一个可以生成具有每个集群特征的新用户的模型。总结来说，我们的努力识别出了四种不同的 Twitter 用户类型，每种都有独特的属性。这使我们能够开发出一种能够生成具有特定特征的新用户的模型。这项工作的结果可以帮助进一步的研究，通过开发基于代理的模型来模拟用户的行为和互动。

Acknowledgments. 此项工作得到了西班牙科学与创新部 MCIN AEI/10.13039/501100011033 的资助，合同编号为 PID2020-113614RB-C21，并由加泰罗尼亚政府项目 2021 SGR 00574 资助。

参考文献

1. Shlens, J.: A tutorial on principal component analysis. (2014).
2. Jolliffe, I.: Principal Component Analysis. Springer-Verlag, New York (2002).
3. Al-Kandari, N.M., Jolliffe, I.T.: Variable selection and interpretation in correlation principal components. (2005). https://doi.org/https://doi.org/10.1002/env.728.
4. Ahmed, M., Seraj, R., Islam, S. M. S.: The k-means algorithm: A comprehensive survey and performance evaluation. (2020).
5. Nielsen, F., Nielsen, F.: Hierarchical clustering. Springer (2016).
6. Zhang, T., Ramakrishnan, R., Livny, M. : BIRCH: A new data clustering algorithm and its applications. Springer (1997).
7. Huang, Z.: Extensions to the k-means algorithm for clustering large data sets with categorical values. Springer (1998).
8. Chander, S., Vijaya, P.: Unsupervised learning methods for data clustering. Artificial Intelligence in Data Mining (2021).

9. Fahim, A.M., Salem, A.M., Torkey, F.A. et al.: An efficient enhanced k- means clustering algorithm. J. Zhejiang Univ. - Sci. A (2006). https://doi.org/https://doi.org/10.1631/jzus.2006.A1626.

10. Uddin, M., Imran, M., Sajjad, H.: Understanding Types of Users on Twitter. (2014).

11. Huang, Z.: Clustering large data sets with mixed numeric and categorical values. (1997).

12. Collier, N., North, M.: Parallel agent-based simulation with Repast for High Performance Computing. 89, 10, (2013). https://doi.org/10.1177/0037549712462620.

13. Twitter users data, https://archive.org/details/archiveteam-twitter-stream-2022-11, last accessed 2024/04/25