工程与管理科学
Engineering and Management Science
第 5 卷◆第 03 期◆版本 1.0◆2023 年
文章类型：论文|刊号（ISSN）：2705-0637(P) / 2705-0645(O)

基于机器学习模型下智能手机 APP 使用情况聚类分析研究

王哲* 杨渠钏 卢灏 曾凡兆 梁兰青
广东海洋大学
DOI：10.12238/ems.v5i3.6280

[摘 要] 随着智能手机普及和移动互联网发展，智能手机 APP 成为人们日常生活中不可或缺一部分。每天，人们使用各种不同 APP 来满足各种需求，如社交、娱乐、购物、工作等。分析不同群体用户特征以及通过用户 APP 记录预测未来是否使用 APP 及使用时长尤其重要。本文通过构建三种聚类模型：kmeans 聚类、DBSCAN 聚类、层次聚类，并通过轮廓系数、DB 指数 Calinski-Harabasz 指数分析其聚类效果。其中 kmeans 效果最好，轮廓系数为 0.3437，DB 指数为 0.811884，Calinski-Harabasz 指数为 58552.461。
[关键词] 均值聚类；DBSCAN 聚类；分层聚类

Research on Cluster Analysis of Smartphone APP Usage Based on Machine Learning Models

Zhe Wang *、Quchuan Yang 、Hao Lu 、Fanzhao Zeng、Lanqing Liang
Guangdong Ocean University

[Abstract] With the popularization of smartphones and the development of mobile Internet, smartphone APP has become an indispensable part of people's daily life. Every day, people use a variety of different APPs to fulfill various needs, such as socializing, entertainment, shopping, and work. It is especially important to analyze the characteristics of different groups of users as well as to predict whether they will use APPs in the future and the duration of use through their APP records. In this paper, we construct three kinds of clustering models: kmeans clustering, DBSCAN clustering, hierarchical clustering, and analyze the clustering effect by profile coefficient and DB index Calinski-Harabasz index through the preprocessed dataset. Among them kmeans is the most effective with contour coefficient of 0.3437, DB index of 0.811884 and Calinski-Harabasz index of 58552.461.
[Keywords] kmeans clustering; DBSCAN clustering; hierarchical clustering

1引言

对于智能手机 APP 使用情况进行深入分析和理解，有助于解用户行为习惯、需求和喜好，为 APP 优化、个性化推荐和精准营销提供重要参考。在过去研究中，传统统计方法和简单数据汇总被广泛用于分析用户 APP 使用情况。然而，随着机器学习技术发展，越来越多研究开始探索基于机器学习模型智能手机 APP 使用情况聚类分析。通过聚类分析，将用户划分为若干个群体，每个群体代表具有相似 APP 使用行为用户。这将有助于解不同用户群体特点和需求，为 APP 推荐和营销提供个性化策略。通过对用户 APP 使用情况聚类分析，可以发现用户群体之间行为模式和差异，进一步洞察用户行为习惯和偏好。聚类分析结果可以为 APP 开发者提供有关如

何改进和优化 APP 功能、界面和用户体验重要指导，以满足不同用户群体需求。为达成这些目标，本研究将收集大量智能手机 APP 使用数据，并运用机器学习中聚类算法，如 K 均值聚类、层次聚类等，对用户进行群体划分和行为模式分析。最后，将通过实验和结果分析验证机器学习模型在智能手机 APP 使用情况聚类分析中有效性和实用性。

2文献综述

利用聚类分析方法对大量用户智能手机 APP 使用情况进行探索性研究，研究通过采集和分析用户 APP 使用数据，将用户划分为若干个群体，并揭示不同群体之间 APP 使用行为模式和差异[1]。研究结果为 APP 开发者和营销人员提供重要参考和优化建议。将用户根据其智能手机 APP 使用情况和用户个人信息划分为不同群体。通过对不同群体特征分析，研究发现用户群体之间 APP 使用行为模式和用户特征之间关联。这为 APP 推荐和个性化营销提供有益洞察。采用多种聚类算法，包括 K 均值聚类和层次聚类，对智能手机 APP 使用情况进行分析[2]。通过比较不同算法效果，研究发现不同用户群体 APP 使用特征，为 APP 优化和用户行为分析提供有益信息[3]。对现有基于机器学习模型智能手机 APP 使用情况聚类分析研究进行综合调查和总结。归纳不同研究中使用聚类算法和特征选择方法，并对研究结果和应用价值进行讨论[4]。

3基于多种机器学习模型建立

3.1 T-SNE 降维

T-SNE 降维是用于降维一种机器学习算法，是由 Laurens van der Maaten 等在 08 年提出来。此外，T-SNE 是一种非线性降维算法，非常适用于高维数据降维到 2 维或者 3 维，进行可视化。该算法可以将对于较大相似度点，t 分布在低维空间中距离需要稍小一点，而对于低相似度点，t 分布在低维空间中距离需要更远其优点在于对于不相似点，用一个较小距离会产生较大梯度来让这些点排斥开来，这种排斥又不会无限大(梯度中分母)[5]，避免不相似点距离太远，本文决定利用 T-SNE 降维成二维数据。

3.2 kmeans 聚类模型

K-Means 算法问是一种典型基于划分聚类算法，也是一种无监督学习。K-Means 算法思想很简单，对给定样本集，用欧氏距离作为衡量数据对象间相似度指标，相似度与数据对象间距离成反比，相似度越大，距离越小。预先指定初始聚类数以及个初始聚类中心，按照样本之间距离大小，把样本集划分为个簇根据数据对象与聚类中心之间相似度，不断更新聚类中心位置，不断降低类簇误差平方和(Sum of Squared Error，SSE)，当 SSE 不再变化或目标函数收敛时，聚类结束，得到最终结果。

K-Means 算法核心思想：首先从数据集中随机选取 k 个初始聚类中心c_i（iS1sk），计算其余数据对象与与聚类中心C_i欧氏距离，找出离目标数据对象最近聚类中心c_i，并将数据对象分配到聚类中心c_i所对应簇中。然后计算每个簇中数据对象平均值作为新聚类中心，进行下一次迭代，直到聚类中心不再变化或达到最大迭代次数时停止。空间中数据对象与聚类中心间欧氏距离计算公式为：

$$d(X,C_i)=\sqrt{\sum_{j=1}^{m}(X_j-C_{ij})^2} \quad (1)$$

其中，X为数据对象;C为第i个聚类中心；m为数据对象维度;X，C为X和C第j个属性值。

整个数据集误差平方和 SSE 计算公式为：

$$SSE=\sum_{i=1}^{k}\sum_{X\in C_i}|d(X,C_i)|^2 \quad (2)$$

其中，SSE 大小表示聚类结果好坏;k 为簇个数。最后得出最佳 k 值为 8，为后续聚类提供可靠参数。

3.3 DBSCAN 聚类模型

DBSCAN 具有噪声基于密度聚类方法)是一种典型基于密度空间聚类算法。显著优点是聚类速度快且能够有效处理噪声点和发现任意形状空间聚类。但是当空间聚类密度不均匀、聚类间距差相差很大时，聚类质量较差。不需要事先指定簇数量，而是通过密度连通区域来确定簇数量和形状。可以通过调整两个重要参数来影响簇数量和形状，这两个参数是半径大小(eps) 和邻居数量值，分析并找出最佳 eps 大小。随着 eps 半径增加，聚类簇个数显著降低，噪声点个数趋于下降趋势。

本文将噪声点个数占整体数据集比例作为评判标准。噪声点个数占整体数据集比例随着 eps 半径大小增加而减少，因此选择最佳 eps 值 0.5，此时聚类簇个数为 1202，噪声点个数为 1724，噪声点数量占整个数据集比例为：2.47%，效果最佳。

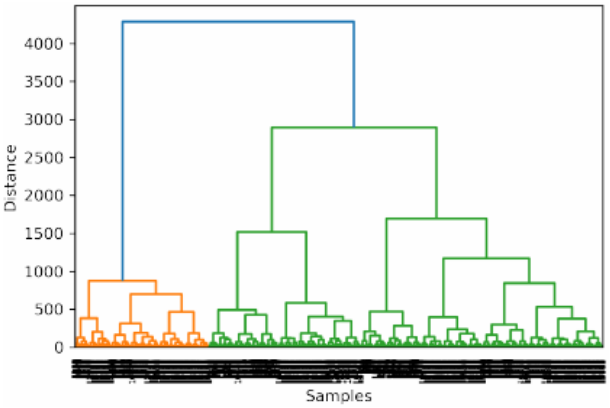

图 1　树状图

3.4 层次聚类模型

层次聚类在不同层次对数据集进行划分从而形成树形聚类结构，数据集划分可采用“自底向上(合并)”聚合策略，也可采用“自顶向下(拆分)”分析策略。依据采用策略可以将层次聚类方法分为：聚合式聚类、分拆式聚类，两种方法均是启发式策略没有去优化一个明确目标函数来实现聚类，很难严格评价聚类效果。

层次聚类是一种无参数聚类方法，不需要提前设置簇数量因此不需要确定最佳 k 值。它在聚类过程中可以自动形成不同大小簇，并且可以通过选择簇高度或距离来控制聚类数量。可以使用树状图来帮助确定聚类数量。

由上图可知，树状图中最长垂直线与水平线相交，在该点上画一条水平线，选择该水平线与垂直线交点作为最佳距离。可以看出最佳距离为 800.

4结论

通过三个聚类模型，分析三种内部指标轮廓系数、DB 指数 Calinski-Harabasz 指数分析其聚类效果，得出初步模型性能结果。综合三个模型对比得知，DBSCAN 聚类效果最差，kmeans 聚类效果最好。三种模型效果对比如下表所示:

表 1　模型效果对比表

	kmeans 聚类	DBSCAN 聚类	层次聚类
轮廓系数	0.3437	0.1009	0.3357
DB 指数	0.811884	1.578564	0797575
Calinski-Harabasz 指数	58552 461	1875.918	21600.569

因此选择 kmeans 作为聚类模型进行分析,通过机器学习中聚类算法，成功地将智能手机用户划分为不同群体。每个群体代表具有相似 APP 使用行为和特征用户群体。这种用户群体划分有助于更好地理解不同用户之间差异和共性，为 APP 推荐和个性化营销提供重要参考。聚类分析揭示不同用户群体之间 APP 使用行为模式。发现在不同群体中，用户 APP 使用偏好和使用频率存在显著差异。这为 APP 开发者和营销人员提供洞察和决策支持,可以更好地满足不同用户群体需求。

[参考文献]

[1]Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3), 160.

[2]Liu, K., Xu, S., Xu, G., Zhang, M., Sun, D., & Liu, H. (2020). A review of android malware detection approaches based on machine learning. IEEE Access, 8, 124579-124607.

[3]Kim, T., Kang, B., Rho, M., Sezer, S., & Im, E. G. (2018). A multimodal deep learning method for android malware detection using various features. IEEE Transactions on Information Forensics and Security, 14(3), 773-788.

[4]Ramesh, A., Nikam, D., Balachandran, V. N., Guo, L., Wang, R., Hu, L., ... & Jia, Y. (2022). Cloud-based collaborative road-damage monitoring with deep learning and smartphones. Sustainability, 14(14), 8682.

[5]Li, J., Sun, L., Yan, Q., Li, Z., Srisa-An, W., & Ye, H. (2018). Significant permission identification for machine-learning-based android malware detection. IEEE Transactions on Industrial Informatics, 14(7), 3216-3225.