基于 GWO-DBSCAN 算法的电商用户价值分类模型设计与实现

赵 煜[1] 卢胜男[1]
ZHAO Yu LU Shengnan

摘 要 基于对电商平台用户画像中用户价值标签的现状了解，分析了以往电商平台常用的 K-means 聚类方法的不足之处，并在此基础上选取多个聚类方法进行横向对比，确定了 GWO-DBSCAN 聚类方法来处理电商用户行为数据。采用基于密度划分的 DBSCAN 聚类算法，针对 DBSCAN 算法聚类效果受扫描半径 eps 和最小包含点 minpts 影响较大的问题，利用灰狼优化算法的全局寻优特性对最佳扫描半径 eps 和最小包含点 minpts 求解，实现对电商用户群体更合理的聚类。通过实践检验发现，采取 GWO-DBSCAN 算法聚类的结果与使用其他聚类方法得到的结果相比，在用户分类的合理性方面有较明显的提升。

关键词 用户价值分析；聚类算法；RFM 模型；灰狼优化算法；DBSCAN 算法

doi：10.3969/j.issn.1672-9528.2024.07.014

0 引言

伴随电商平台的繁荣发展，行业竞争也日趋白热化，为了在激烈的市场竞争中脱颖而出，科学评价客户价值并精准划分客户群体，进而指导制定营销策略，是赢得客户获取利润的重要手段。RFM 客户价值模型作为一种定量分析模型，在电商用户价值领域得到广泛应用，从业者通过对电商用户价值进行深入的研究，做了不同的改进。N. Manjushree 等人通过改进的 K-means，基于在线分析对产品的动态价格进行预测 [1]；袁绮蕊 [2] 通过 RFM 模型筛选典型的用户，从事实、模型和预测三个维度构建用户画像模型，并使用 K-means 对用户数据进行聚类分析；施文幸等人 [3] 基于萤火虫 K-means 聚类的电力用户画像构建和应用；陈东清等人 [4] 基于熵权法改进 RFM 模型，通过权重计算得分评价电商用户价值；徐翔斌等人 [5] 基于电子商务行业，将总利润 P 引入 RFM 模型，得到了 RFP 模型，并使用 K-means 聚类算法对其顾客进行聚类，为电子商务行业提出了针对性营销策略；谢鹏寿等人 [6] 根据汽车 4S 店客户的消费特征，提出了适用于汽车销售行业的 TFM 模型，并采用 K-means 算法对客户进行细分。

可以看出，原先的研究大多使用 K-means 聚类方法对模型进行聚类，但 K-means 存在对初始聚类中心选择敏感，不同初始选择会导致不同的聚类结果、需要预先确定簇数 K，且 K 值的选取不易把握、对非凸形状的类簇识别效果差、易受噪声、边缘点、孤立点影响等缺点，导致算法聚类效率不高。目前许多研究都基于密度峰值的聚类基础上进行改进 [7]，基于密度划分的 DBSACN 算法可以解决上述的部分问题，但 DBSCAN 需要根据给出的数据集选择合适的参数：扫描半径 (eps) 和最小包含点 (minpts)，只有选择合适的参数，才能得到更好的聚类效果。灰狼优化算法作为一种新型的智能优化算法，受到灰狼捕食行为的启发，这类算法在优化问题中具有较高的求解精度和寻优能力，已被广泛应用在聚类算法中。孟涛等人 [8] 将改进遗传算法和 DBSCAN 聚类相结合，刘成汉等人 [9] 使用改进交叉算子后的自适应人工蜂群黏菌算法做研究。

针对以上问题，本文提出一种基于灰狼算法优化的 DBSCAN 聚类算法对用户价值分类。通过引入灰狼算法，优化 DBSCAN 聚类算法中的两个参数，并通过电商用户数据集来验证算法分类的有效性。

1 理论基础

1.1 DBSCAN 算法

DBSCAN（density-based special clustering of application with noise）是一种基于密度的聚类算法，该算法不需要预先定义簇的个数，而是将高密度的区域划分为簇，并将其他区域视为噪声，这样就可以在有噪声的数据中发现任意形状的簇，对于有噪声（即孤立点或异常值）的数据集有很好的鲁棒性。DBSCAN 的基本概念包括两个算法参数，即扫描半径 (eps) 和最小包含点 (minpts)；三种类别的点，即核心点、边界点与噪声点；四种点的关系，即密度直达、密度可达、密

1. 西安石油大学 陕西西安 710065

度相连与非密度相连。

1.2 灰狼优化算法（GWO）

灰狼算法（grey wolf optimizer）是 Seyedali Mirjalili 等人[10]在 2014 年提出的一种基于自然界灰狼行为的启发式优化算法，算法模拟了灰狼群体中不同等级灰狼间的社会等级制度和捕食行为，通过不断搜索最优解来解决复杂问题。在算法中每一个灰狼个体的位置代表了解向量中的一个解，种群中适应度最优的解、次优的解和第三优的解分别对应 α 狼、β 狼和 δ 狼，而其余的解被看作 ω 狼：α 狼具有最高领导地位；β 狼服从 α 狼的领导并引导其他的狼；δ 狼服从 α 狼和 β 狼的领导并负责保卫和侦察；ω 狼服从其他狼的领导。灰狼算法主要分为三个步骤：包围猎物、狩猎和攻击猎物。

（1）包围猎物，狼群在搜索猎物时逐渐接近并包围，具体数学模型为：

$$\vec{D}=\left|\vec{C}\bullet\vec{X_p}(t)-\vec{X}(t)\right| \tag{1}$$

$$\vec{X}(t+1)=\vec{X_p}(t)-\vec{A}\bullet\vec{D} \tag{2}$$

式中：$\vec{X_p}$为猎物的位置向量，$\vec{X}$为狼群的位置向量，$\vec{D}$为狼群距离猎物的长度，t 为当前算法迭代次数，协同系数向量$\vec{A}$、$\vec{C}$定义如下：

$$\vec{A}=2\vec{a}\bullet\vec{r_1}-\vec{a} \tag{3}$$

$$\vec{C}=2\bullet\vec{r_2} \tag{4}$$

$\vec{A}$用来模拟灰狼的攻击行为，受收敛因子$\vec{a}$的影响。$\vec{a}$在整个迭代过程中，由 2 减小到 0 。

（2）狩猎行为：

$$\begin{cases}\vec{D_\alpha}=\left|\vec{C_1}\square\vec{X_\alpha}-\vec{X}\right|\\ \vec{D_\beta}=\left|\vec{C_2}\square\vec{X_\beta}-\vec{X}\right|\\ \vec{D_\delta}=\left|\vec{C_3}\square\vec{X_\delta}-\vec{X}\right|\end{cases} \tag{5}$$

上述是灰狼个体跟踪猎物位置的数学模型，$\vec{D_\alpha}$、$\vec{D_\beta}$ 和 $\vec{D_\delta}$分别为是 α 狼、β 狼和 δ 狼与其他个体之间的距离；$\vec{X_\alpha}$、$\vec{X_\beta}$ 和 $\vec{X_\delta}$ 分别是 α 狼、β 狼和 δ 狼的当前最优位置。当$|A|$<1 时，狼群集体更新位置：

$$\begin{cases}\vec{X_1}=\vec{X_\alpha}-\vec{A_1}\bullet\vec{D_\alpha}\\ \vec{X_2}=\vec{X_\beta}-\vec{A_2}\bullet\vec{D_\beta}\\ \vec{X_3}=\vec{X_\delta}-\vec{A_3}\bullet\vec{D_\delta}\end{cases} \tag{6}$$

$$\vec{X}(t+1)=\frac{\vec{X_1}+\vec{X_2}+\vec{X_3}}{3} \tag{7}$$

式（6）为 ω 狼在 α 狼、β 狼和 δ 狼的分别领导下更新后的位置，其中的$\vec{A_1}$、$\vec{A_2}$和$\vec{A_3}$是根据式（3）计算得到的系数向量，式（7）为 ω 狼最后确定的位置。

（3）攻击猎物

$$a=2\left(1-\frac{t}{T}\right) \tag{8}$$

式中：t 表示当前迭代次数，T 为人为设定的最大迭代次数，在迭代过程中，a 值线性逐渐减小，其对应$\vec{A}$的值也会在区间 [-a,a] 变化；a 值越小表示灰狼越靠近猎物，当 $|\vec{A}|<1$ 算法完成收敛，获取猎物位置。本文选用灰狼算法来选取最佳的扫描半径 eps 与最小包含点 minpts。

1.3 GWO-DBSCAN 聚类算法

本文结合 GWO 灰狼优化算法对 DBSCAN 算法进行改进，以 silhouette 指标为目标函数，通过 GWO 算法对 DBSCAN 中的扫描半径与最小包含点进行优化，通过多次迭代寻优，找到使 silhouette 取值最大时的 eps 和 minpts。改进后算法具体流程如下。

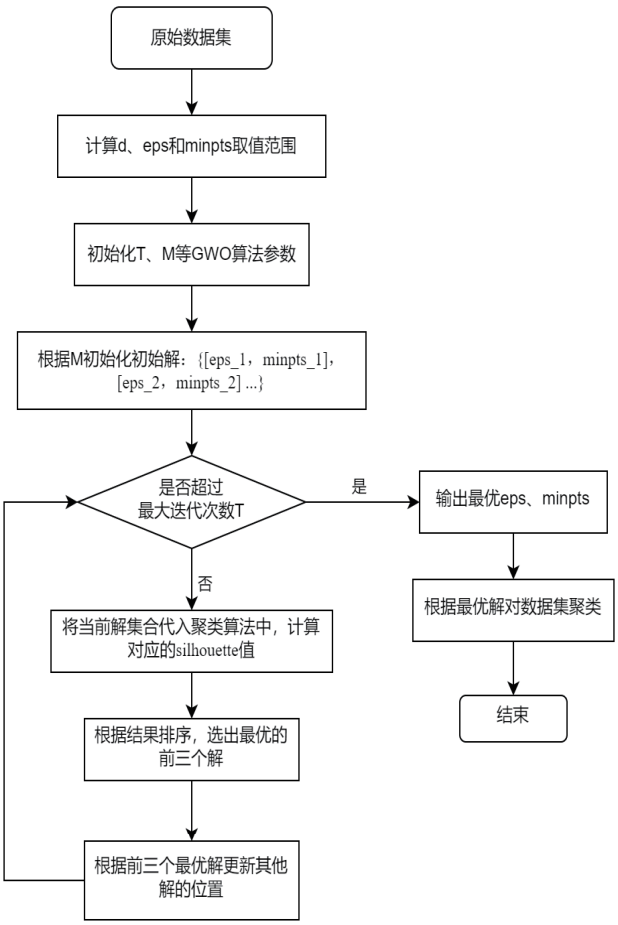

图 1 GWO-DBSCAN 聚类算法流程

步骤 1：计算用户数据样本之间的欧氏距离，对比每个样本与其最近 / 最近邻居之间的平均距离 d，选取合适的 eps 和 minpts 的取值范围：[eps_min，minpts_min] 和 [eps_max，minpts_max]。

步骤 2：设计 GWO 算法的最大迭代次数 T，种群规模 M，初始向量值，即 eps 与 minpts 组成的集合。

步骤 3：根据种群规模 M 和 eps、minpts 的取值范围，随机生成初始解集合 {[eps_1, minpts_1]，[eps_2, minpts_2] …}。

步骤 4：将初始解集合中的元素依次代入 DBSCAN 算法中，对原始用户数据集进行聚类，计算并存储每次聚类结果的 silhouette 指标。

步骤 5：对初始解集合与对应的 silhouette 值按照升序排序，得到初始解集合中的最优的前三个解和最优的前三个 silhouette 值。

步骤 6：计算系数向量的参数 a，利用式（6）对初始解集合中的每个元素进行位置更新，并重新代入 DBSCAN 算法中对数据集进行聚类，计算 silhouette 值。

步骤 7：判断是否到达 GWO 算法的最大迭代次数 T，若满足，结束迭代并执行步骤 8；若不满足，执行步骤 5 继续寻优过程。

步骤 8：将得到的最优解 eps 和 minpts 代入 DBSCAN 算法中，得到最终聚类结果。

2 实验结果与分析

2.1 数据集选取及数据预处理

为验证改进算法在聚类电商用户价值的合理、可靠性，本课题采用 RFM 模型中的三个指标作为实验输入数据。RFM 模型是常用于衡量客户价值的数学模型。该模型主要是通过分析最近一次购买时间（recency）、消费频率（frequency）和消费金额（monetary）3 个角度，对客户进行细分，进而来确定客户价值的。因此，RFM 模型在电商平台得到了比较广泛的应用。

选取的数据集为 kaggle 上的某电商脱敏后的用户行为数据集。该数据集包括用户 ID、性别、住址、订单日期、购买商品品类、购买商品数量等信息，通过对数据集预处理，得到 RFM 模型中近一次购买时间、消费频率和消费金额这三个数据。

2.2 评价指标

本文的实验结果评价指标为：轮廓系数（silhouette coefficient）、Calinski-Harabasz index（CH）、戴维森堡丁指数（Davies-bouldin index，DBI）。轮廓系数和 CH 指标的计算方法为：

$$s=\frac{b-a}{Max(a,b)} \tag{9}$$

$$\mathrm{SC}=\frac{\sum_{i=1}^{N}s_i}{N} \tag{10}$$

$$\mathrm{CH}=\frac{tr(B_k)(N-K)}{tr(W_k)(K-1)} \tag{11}$$

$$\begin{cases} B_k=\sum_{q=1}^{k}n_q(c_q-c_e)(c_q-c_e)^T \\ W_k=\sum_{q=1}^{k}\sum_{x\in C_q}(x-c_q)(x-c_q)^T \end{cases} \tag{12}$$

式（9）中，s 表示某个样本的轮廓系数，a 表示某个样本与其所在簇内其他样本的平均距离，b 表示某个样本与其他簇样本的平均距离；式（10）中 SC 为聚类总的轮廓系数；式（12）中 $\boldsymbol{B}_k$ 为类间的协方差矩阵，$\boldsymbol{W}_k$ 为类内数据的协方差矩阵，c_q 表示类 q 的中心点，c_e 表示数据集的中心点，n_q 表示类 q 中数据的数目，C_q 表示类 q 的数据集合。

2.3 实验结果分析

为验证改进后算法的有效性，分别使用 K-means 算法、FCM 算法、OPTICS 算法、DBSCAN 算法和 GWO-DBSCAN 算法对给出的电商用户行为数据集进行聚类，并对结果作对比。

表 1 是通过 3 种内部聚类指标，对 5 种不同聚类算法的结果作比较。可以看出，GWO-DBSCAN 算法在轮廓系数上对比其他的聚类算法，分别提高了 0.014 5、0.096、0.075 8、0.014 5；在 DBI 值上对比其他聚类算法，分别降低了 0.024 6、0.026 5、0.121 7、0.024 6；在 CH 值上对比其他的聚类算法最高。这说明本文提出的改进后的算法聚类更合理，簇间距离更大，簇内样本更紧密，更少出现交叉重叠现象。

表 1 对比实验结果

	Kmeans	FCM	OPTICS	DBSCAN	GWO-DBSCAN
聚类中心数	4	3	4	4	4
轮廓系数	0.816 0	0.734 5	0.754 6	0.816 0	0.830 5
CH 值	7 223.637 0	1 376.943 1	2 261.449 3	7 223.637 0	12 844.375 2
DBI 值	0.257 9	0.259 8	0.355 0	0.257 9	0.233 3

表 2 为采用 GWO-DBSCAN 聚类算法得到的聚类结果，并展示了每个类别中各自的 RFM 指标均值和该类别的用户质量；根据聚类结果来看，该数据集的用户分为 4 类。

表 2 GWO-DBSCAN 聚类结果

	R 均值	F 均值	M 均值	占比	用户质量
类别 1	31.6	3	473.4	0.33	低
类别 2	45.5	4	690.0	0.17	中
类别 3	24.6	5	1 165.2	0.17	高
类别 4	15.3	4	805.4	0.17	中高

类别 3 为高价值用户群体，这类用户的特点是消费金额最多，购物频率最高，最后一次上线的时间也较短，这类用户是电商用户中需要维持和发展的客户。

类别 4 的用户为中高价值用户群体，消费金额仅次于高质量用户群体，最后一次购买时间最近，消费频率也很高，

此类用户是企业重点保持但不必大力发展的用户。

类别 2 为中等质量用户群体，潜在价值高，此类用户的特点是最后一次购买时间久，但消费金额和消费频率不算太低，是企业大力发展且需要不断维护防止流失的用户。

类别 1 是低价值客户群体，其购买频率和消费金额都是最低的，总的来说是电商用户中具有一定价值但需企业取舍的客户。

图 2 为采用 GWO-DBSCAN 聚类算法聚类，并排除异常用户数据后各类别用户所占比重，图 3 为各类用户总花费占比，其中高质量的用户占比最少，但消费金额最多，分别为 19.9% 和 32.5%；中等价值的用户和中高价值用户占比接近，且花费相当，差不多为 20%，低价值用户虽然消费金额不是最低的，但人数上占比最多，多达 39.7%。可以看出，改进后聚类算法的聚类结果很符合帕累托法则（Pareto's principle）：高价值的用户占总人数的 20%。这表明改进后算法在聚类电商用户的合理性表现优异。

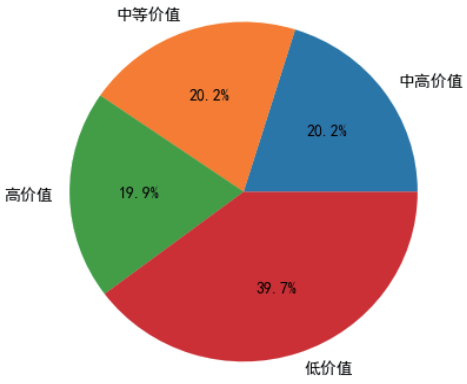

图 2 各用户价值人数占比图

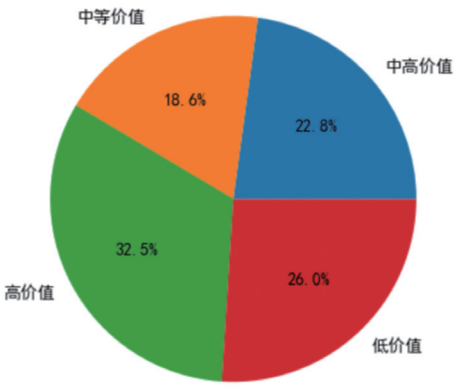

图 3 各类用户总花费占比

3 结语

针对传统 K-means 方法根据 RFM 模型对电商用户进行聚类时需要提前确定用户类别数、容易受到异常点影响等问题，本文提出一种基于灰狼算法优化的 GWO-DBSCAN 聚类算法。针对 DBSCAN 聚类算法的两个重要参数进行优化，使用聚类内部评价指标 silhouette 作为目标函数，利用灰狼优化算法寻找两个参数的全局最优解。通过在数据集上进行实验，验证了改进算法有效提高了聚类结果的质量，可以为后续开发人员制定电商用户画像标签提供可靠的依据和参考。

参考文献：

[1]MANJUSHREE N, BHAVANA N.Predicting dynamic product price by online analysis:modified k-means cluster[J].Advances in intelligent systems and computing,2020,1120:1-15.

[2] 袁绮蕊 . 基于 K-MEANS 的在线健康社区用户画像模型构建 [J]. 科技情报研究 ,2021,3(4):95-106.

[3] 施文幸 , 曹诗韵 . 基于萤火虫 K-means 聚类的电力用户画像构建和应用 [J]. 计算机系统应用 ,2021,30(8):281-287.

[4] 陈东清 , 叶翀 , 黄章树 . 基于熵权法改进 RFM 模型的电商客户价值细分研究 [J]. 西安电子科技大学学报 (社会科学版), 2020,30(2):39-45.

[5] 徐翔斌 , 王佳强 , 涂欢 , 等 . 基于改进 RFM 模型的电子商务客户细分 [J]. 计算机应用 , 2012,32(5):1439-1442.

[6] 谢鹏寿 , 张宽 , 范宏进 , 等 . 汽车 4S 店 TFM 客户细分模型及其方法研究 [J]. 小型微型计算机系统 , 2019, 40(10): 2165-2169.

[7]DU M, DING S, JIA H.Study on density peaks clustering based on k-nearest neighbors and principal component analysis[J]. Knowledge-based systems,2016,99(May 1):135-145.

[8] 孟涛 , 王晓勇 , 胡胜利 . 基于改进遗传算法和 DBSCAN 聚类的学习数据深度挖掘方法 [J]. 齐齐哈尔大学学报 (自然科学版), 2024(1):1-7.

[9] 刘成汉 , 何庆 . 改进交叉算子的自适应人工蜂群黏菌算法 [J]. 小型微型计算机系统 , 2023,44(2):263-268.

[10]MIRJALILI S, MIRJALILI S M, LEWIS A.Grey wolf optimizer[J].Advances in engineering software, 2014, 69: 46-61.

【作者简介】

赵煜（1997—），男，陕西西安人，硕士研究生，研究方向：机器学习、大数据技术与应用。

卢胜男（1982—），女，江苏徐州人，博士，副教授，硕士生导师，研究方向：图像处理及机器学习、大数据技术应用。

（收稿日期：2024-05-11）