(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 108573264 A
(43)申请公布日 2018.09.25

(21)申请号 201710130878.0

(22)申请日 2017.03.07

(71)申请人 中国科学院沈阳自动化研究所
地址 110016 辽宁省沈阳市东陵区南塔街114号

(72)发明人 朱云龙 吕赐兴 张浩 张丁一

(74)专利代理机构 沈阳科苑专利商标代理有限公司 21002
代理人 王倩

(51)Int.Cl.
G06K 9/62(2006.01)
G06N 3/00(2006.01)

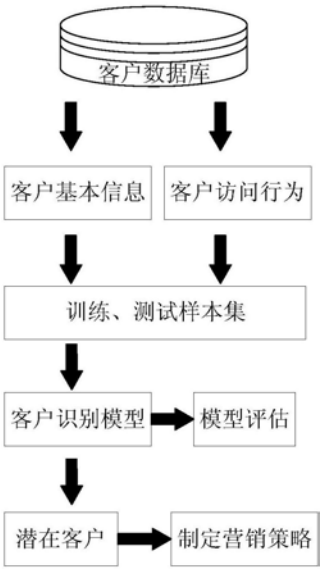
权利要求书2页 说明书4页 附图1页

(54)发明名称

一种基于新型蜂群聚类算法的家居行业潜在客户识别方法

(57)摘要

本发明涉及一种基于新型蜂群聚类算法的家居行业潜在客户识别方法，包括以下步骤：从客户集合中选择聚类中心并进行编码；随机给定所有人工蜜峰的初始位置；跟据人工蜜蜂的适应度对所有人工蜜蜂排序，从中选取前HN个位置作为食物源；根据人工蜜蜂当前位置进行聚类运算，并更新人工蜜蜂位置；更新食物源。本发明实现简便，不过分依赖参数的选择，具有较强的全局搜索能力，收敛速度快，识别精度高等优点，对于家居行业潜在客户识别这种复杂的聚类问题，有非常明显的优化识别效果。



CN 108573264 A

1.一种基于新型蜂群聚类算法的家居行业潜在客户识别方法，其特征在于包括以下步骤：

1)从客户集合X＝{x_1,x_2,...,x_N}中任意选择c个点作为聚类中心y_1,y_2,...,y_c，并进行编码：$Y_{sg}^t=(y_1,y_2,...,y_c)_{sg}^t$；

其中，Y_{sg}^t表示任意一个人工蜜蜂的编码；g表示人工蜜蜂在种群中的角色；s表示人工蜜蜂在角色g子群中的标号，t表示当前的迭代步骤；

2)随机给定所有人工蜜峰的初始位置；跟据人工蜜蜂的适应度对所有人工蜜蜂排序，从中选取前HN个位置作为食物源；

3)根据人工蜜蜂当前位置进行聚类运算，并根据聚类结果得到每个人工蜜蜂的适应度；通过以下步骤更新人工蜜蜂位置：

3-1)对雇佣蜂或跟随蜂进行位置更新：$v_{sq}=y_{sq}+\delta_{sq}(y_{sq}-y_{hq})$

其中，s＝1,2,…,n；n表示种群规模；h是随机确定的，不与s相同；v_{sq}代表更新后雇佣蜂或跟随蜂的位置；y_{sq}代表当前雇佣蜂或跟随蜂的位置；y_{hq}代表随机选取雇佣蜂或跟随蜂的位置；δ_{sq}为参数，在[-1,1]范围内随机产生；

3-2)对侦查蜂进行位置更新：$y_s^q=y_{\min}^q+\sigma(y_{\max}^q-y_{\min}^q)$

其中，σ是在[-1,1]范围内的随机数；y_s^q代表侦查蜂更新后的位置；$y_{\min}^q$代表侦查蜂当前位置向量中最小的一个维度；$y_{\max}^q$代表侦查蜂当前位置向量中最大的一个维度；

当$f_e^t>f_e^{best}$时，这只侦察蜂变为雇佣蜂；f_e^t表示第t次迭代的适应度，f_e^{best}表示上一次迭代中值最大的适应度；

4)更新食物源：计算所有人工蜜蜂个体当前位置的适应度，从中选择m个大于原有食物源适应度的位置，替换原有食物源中适应度最小的m个位置；

5)如果当前的迭代次数达到了预先设定的最大次数T_{max}，或最终结果小于收敛精度ξ，则停止迭代，输出当前的人工蜜蜂位置作为最终的聚类结果；否则返回步骤3)。

2.根据权利要求1所述的一种基于新型蜂群聚类算法的家居行业潜在客户识别方法，其特征在于所述适应度：

$$f_e=\frac{\sum_{xn=1}^{N}r_{xn}}{N}\text{其中，}\begin{cases}r_{xn}=0 & \text{正确聚类}\\ r_{xn}=1 & \text{错误聚类}\end{cases}$$

其中，N为客户的数量，f_e为适应度，r_{xn}为参数。

3.根据权利要求1所述的一种基于新型蜂群聚类算法的家居行业潜在客户识别方法，其特征在于所述根据人工蜜蜂当前位置进行聚类运算包括以下步骤：

1.1)将人工蜜蜂当前位置作为v_1,v_2,...,v_c，即c个聚类中心；

1.2)以v_1,v_2,...,v_c为中心点对客户集合X进行集合划分：

如果$||x_k-v_i||^2\leqslant||x_k-v_j||^2$,$x_k\in X$,i＝1,2,...,c,j＝1,2,...,c,i≠j，则将x_k划分到聚类客户集合A_i中，其中$A_i\subset X$；

1.3)计算新的聚类中心：

$$v_i^{'}=\frac{\sum_{k=1}^{N_i}x_k}{N_i},\ x_k\in A_i,\ i=1,2,...,c$$

其中,N_i表示第i个集合A_i中向量的数量;

1.4)令$v_i=v'_i$,计算类间距离:$J=\sum_{i=1}^{c}\sum_{x_k\in A_i}\|x_k-v_i\|^2$;将类间距离小于阈值的$x_k$分为点$v'_i$的类中;

1.5)当本次聚类中心与上次聚类中心相比无变化时,计算结束,此时的聚类中心为聚类结果;否则,返回步骤1.2)。

4.根据权利要求1所述的一种基于新型蜂群聚类算法的家居行业潜在客户识别方法,其特征在于根据最终的聚类结果对客户进行分类,完成识别。

一种基于新型蜂群聚类算法的家居行业潜在客户识别方法

技术领域

[0001] 本发明涉及一种基于新型蜂群聚类算法的家居行业潜在客户识别方法，属于家居行业的电子商务领域，同时涉及群体智能算法和聚类算法领域。

背景技术

[0002] 随着科学理念和技术的进步以及企业市场稳健发展的需要，客户资产作为企业的一项重要的无形资产，其重要性已经受到了广泛的关注，成为衡量企业市值的关键要素之一。在“以客户为中心”的市场环境下，能否很好的理解客户行为及市场的真实需求已成为决定企业竞争力的关键。企业的成功与否很大程度取决于企业是否能快速并准确地响应客户的需求，以及应对竞争对手的商业策略的变化。因而，如何高效利用海量、分散的客户资源，支持中小企业群的智能营销决策是一项有待解决的关键技术。

[0003] 针对家居行业潜在客户识别这种复杂的聚类问题，如果采用传统的数据挖掘方法解决，则无法在客户的识别精度和效率两个方面同时达到理想的结果。近年来，模拟蜜蜂觅食行为的人工蜂群智能优化算法得到了学者们的广泛关注，结合聚类算法用其来解决此类问题取得了较好的结果，显示出了基于生物启发计算的聚类算法在解决目标识别问题中的独特优势。但是较早提出的蚁群算法或遗传算法等实现复杂，稳定性较差，优化的聚类结果随机性大；粒子群算法虽然运行速度较快，但是经常陷入解空间的局部，无法得到全局最优解，特别是针对高维优化问题更加无能为力；经典的k-means等纯聚类算法，受参数的影响较大，结果随机性较大，无法实现对客户的精准识别。这些已存在的算法在解决相对复杂的优化问题时，其性能还不能达到满意的精度和稳定性要求。

发明内容

[0004] 针对现有计算在家居行业潜在客户识别这种复杂的聚类问题时所暴露出来的缺陷，本发明提出了一种借鉴大自然中多个蜜蜂协作觅食行为的基于新型蜂群聚类算法的家居行业潜在客户识别方法。

[0005] 本发明解决其技术问题所采用的技术方案是：一种基于新型蜂群聚类算法的家居行业潜在客户识别方法，包括以下步骤：

[0006] 1)从客户集合X＝{x_1,x_2,...,x_N}中任意选择c个点作为聚类中心y_1,y_2,...,y_c，并进行编码：$Y_{sg}^{t}=(y_1,y_2,...,y_c)_{sg}^{t}$；

[0007] 其中，Y_{sg}^{t}表示任意一个人工蜜蜂的编码；g表示人工蜜蜂在种群中的角色；s表示人工蜜蜂在角色g子群中的标号，t表示当前的迭代步骤；

[0008] 2)随机给定所有人工蜜峰的初始位置；跟据人工蜜蜂的适应度对所有人工蜜蜂排序，从中选取前HN个位置作为食物源；

[0009] 3)根据人工蜜蜂当前位置进行聚类运算，并根据聚类结果得到每个人工蜜蜂的适应度；通过以下步骤更新人工蜜蜂位置：

[0010] 3-1)对雇佣蜂或跟随蜂进行位置更新：$v_{sq}=y_{sq}+\delta_{sq}(y_{sq}-y_{hq})$

[0011] 其中，s＝1,2,…,n;n表示种群规模;h是随机确定的，不与s相同;v_{sq}代表更新后雇佣蜂或跟随蜂的位置;y_{sq}代表当前雇佣蜂或跟随蜂的位置;y_{hq}代表随机选取雇佣蜂或跟随蜂的位置;δ_{sq}为参数，在[-1,1]范围内随机产生;

[0012] 3-2)对侦查蜂进行位置更新：$y_s^q = y_{\min}^q + \sigma(y_{\max}^q - y_{\min}^q)$

[0013] 其中，σ是在[-1,1]范围内的随机数;y_s^q代表侦查蜂更新后的位置;$y_{\min}^q$代表侦查蜂当前位置向量中最小的一个维度;$y_{\max}^q$代表侦查蜂当前位置向量中最大的一个维度;

[0014] 当$f_e^t > f_e^{best}$时，这只侦察蜂变为雇佣蜂;f_e^t表示第t次迭代的适应度，f_e^{best}表示上一次迭代中值最大的适应度;

[0015] 4)更新食物源:计算所有人工蜜蜂个体当前位置的适应度，从中选择m个大于原有食物源适应度的位置，替换原有食物源中适应度最小的m个位置;

[0016] 5)如果当前的迭代次数达到了预先设定的最大次数T_{max}，或最终结果小于收敛精度ξ，则停止迭代，输出当前的人工蜜蜂位置作为最终的聚类结果;否则返回步骤3)。

[0017] 所述适应度：

[0018] $f_e = \frac{\sum_{xn=1}^{N} r_{xn}}{N}$ 其中，$\begin{cases} r_{xn} = 0 & \text{正确聚类} \\ r_{xn} = 1 & \text{错误聚类} \end{cases}$

[0019] 其中，N为客户的数量，f_e为适应度，r_{xn}为参数。

[0020] 所述根据人工蜜蜂当前位置进行聚类运算包括以下步骤：

[0021] 1.1)将人工蜜蜂当前位置作为$v_1, v_2, \ldots, v_c$，即c个聚类中心;

[0022] 1.2)以$v_1, v_2, \ldots, v_c$为中心点对客户集合X进行集合划分：

[0023] 如果$||x_k - v_i||^2 \leqslant ||x_k - v_j||^2$，$x_k \in X$，i＝1,2,...,c,j＝1,2,...,c,i≠j，则将x_k划分到聚类客户集合A_i中，其中$A_i \subset X$;

[0024] 1.3)计算新的聚类中心：

[0025] $v_i^{'} = \frac{\sum_{k=1}^{N_i} x_k}{N_i},\ x_k \in A_i,\ i = 1,2,...,c$

[0026] 其中，N_i表示第i个集合A_i中向量的数量;

[0027] 1.4)令$v_i = v'_i$，计算类间距离：$J = \sum_{i=1}^{c} \sum_{x_k \in A_i} \| x_k - v_i \|^2$;将类间距离小于阈值的$x_k$分为点$v'_i$的类中;

[0028] 1.5)当本次聚类中心与上次聚类中心相比无变化时，计算结束，此时的聚类中心为聚类结果;否则，返回步骤1.2)。

[0029] 根据最终的聚类结果对客户进行分类，完成识别。

[0030] 本发明具有以下有益效果及优点：

[0031] 1.为了更好解决这些问题，模拟蜜蜂觅食行为，结合经典的聚类算法并融入模隶属度概念，发明了基于新型蜂群聚类算法的家居行业潜在客户识别方法，该算法实现了自适应聚类，具有全局搜索能力强，收敛速度快，识别精度高等优点。

[0032] 2.算法实现简便，不过分依赖参数的选择，具有较强的全局搜索能力，收敛速度快，识别精度高等优点，对于家居行业潜在客户识别这种复杂的聚类问题，有非常明显的优

化识别效果。

附图说明

[0033] 图1是客户识别过程的总体结构图；

[0034] 图2是蜂群聚类算法对客户属性集的聚类结果图。

具体实施方式

[0035] 下面结合实施例对本发明做进一步的详细说明。

[0036] 如图1展示了客户识别过程的总体结构。首先，从客户数据库数据获取，如客户的基本信息，客户的偏好行为等，形成训练和测试样本集合，获取数据质量的高低在很大程度上影响了最终获得结果的好坏；然后，为了实现客户的聚类，建立潜在客户识别模型，训练出的模型需要经过评估后，才可用于潜在客户识别；最后，根据构建的潜在客户识别模型，对新访问的客户进行识别，发现潜在客户，进行目标营销。

[0037] 步骤1：建立基于k-means聚类方法的客户识别模型

[0038] 1.1)给定聚类个数c；

[0039] 1.2)从客户集合X＝{$x_1,x_2,...,x_N$}中任意选择c个点$v_1,v_2,...,v_c$分别作为c个聚类集合的聚类中心；其中，x_k表示客户对家居产品的偏好，包括颜色，材质等信息，例如某些客户偏爱红色和玻璃材质的家居产品等；k＝1,…,N；

[0040] 1.3)以$v_1,v_2,...,v_c$为中心点对X进行集合划分，划分的原则是：如果$||x_k-v_i||^2 \leq ||x_k-v_j||^2$，$x_k \in X$，i＝1,2,...,c，j＝1,2,...,c，i≠j，则将x_k划分到聚类客户集合A_i中，其中$A_i \subset X$；

[0041] 1.4)根据聚类客户集合$A_1,A_2,...,A_c$中的点计算新的中心点：

[0042] $$v_i^{'}=\frac{\sum_{k=1}^{N_i}x_k}{N_i},\ x_k \in A_i,\ i=1,2,...,c$$

[0043] 其中N_i表示集合A_i中向量的数量；

[0044] 1.5)令$v_i＝v'_i$，根据下式计算类间距离：

[0045] $$J=\sum_{i=1}^{c}\sum_{x_k \in A_i}||x_k-v_i||^2$$

[0046] 1.6)当聚类中心不再变化计算结束，否则，返回步骤1.3)。

[0047] 步骤2：编码

[0048] 聚类算法的核心是聚类中心的确定，所以对(重新更新数据即待测数据的客户集合X＝{$x_1,x_2,...,x_N$}中任意选择c个点)聚类中心$y_1,y_2,...,y_c$进行编码。任一人工蜜蜂可以编码如下：

[0049] $$Y_{sg}^{t}=(y_1,y_2,...,y_c)_{sg}^{t}$$；

[0050] 其中，Y_{sg}^{t}表示任意一个人工蜜蜂的编码(位置)；g表示人工蜜蜂在种群中的角色；s表示人工蜜蜂在角色g子群中的标号，t表示当前的迭代步骤；y_l表示第l个聚类中心，其中l＝1,2,...,c；

[0051] 步骤3:初始化各类参数

[0052] 给定聚类个数c;种群规模n,即人工蜂群的个数;随机给定所有人工峰个体的初始位置;跟据适应度从大到小对所有个体排序,从中选取前HN个位置作为食物源;最大迭代次数T_{max},收敛精度ξ;

[0053] 步骤4:数据更新

[0054] 人工蜂共有三种类型:雇佣蜂、跟随蜂和侦查蜂。其中,雇佣蜂的数量占到总数的一半,其余的是跟随蜂和侦察蜂。侦察蜂进行的是空间的探索,而雇佣蜂和跟随蜂分不同阶段在搜索空间中执行开发过程。

[0055] 将步骤1的聚类结果,根据下面公式计算人工蜂的适应度(函数值),目的是使聚类误差率最小:

[0056] $$f_e=\frac{\sum_{xn=1}^{N} r_{xn}}{N}\text{其中,}\begin{cases} r_{xn}=0 & \text{正确聚类} \\ r_{xn}=1 & \text{错误聚类} \end{cases}$$

[0057] 其中,N为需要识别客户的数量,f_e为误差率。这里是根据历史客户分类数据来判断聚类划分结果。

[0058] 设置当前迭代次数t=0,完成初始化阶段后,对每个人工蜜蜂执行下面两个迭代过程:

[0059] 3.1)利用下面公式对雇佣蜂和跟随蜂进行位置更新:

[0060] $v_{sq}=y_{sq}+\delta_{sq}(y_{sq}-y_{hq})$

[0061] 其中,s=1,2,…,n;q=1,2,…,c是随机选取的索引,h是随机确定的,但是不能与s相同;v_{sq}代表更新后雇佣蜂和跟随蜂的位置;y_{sq}代表更新前雇佣蜂和跟随蜂的位置;y_{hq}代表随机选取雇佣蜂或跟随蜂的位置;δ_{sq}在[-1,1]范围内随机产生,这个参数控制着在y_{sq}周围的新蜜源的产生。

[0062] 3.2)利用下面公式对侦查蜂进行位置更新:

[0063] $$y_s^q=y_{\min}^q+\sigma(y_{\max}^q-y_{\min}^q)$$

[0064] 其中,σ是在[-1,1]范围内的随机数;y_s^q代表侦查蜂更新后的位置;$y_{\min}^q$代表侦查蜂当前位置向量中最小的一个维度;$y_{\max}^q$代表侦查蜂当前位置向量中最大的一个维度。当$f_e^t>f_e^{best}$时,即侦察蜂发现蜂蜜含量更加丰富的蜜源后,这只侦察蜂将变成一只雇佣蜂。f_e^t表示该侦察蜂第t次迭代的适应度,f_e^{best}表示上一次迭代中所有人工蜜蜂适应度的最大值;

[0065] 步骤5:更新食物源:计算所有蜜蜂个体当前位置的适应度,从中选择好于原有食物源适应度的位置替换食物源中较差的位置;

[0066] 步骤6:如果当前的迭代次数达到了预先设定的最大次数T_{max},或最终结果小于预定收敛精度ξ要求,则停止迭代,输出聚类结果,如图2所示,否则,t=t+1,转到步骤4。

[0067] 聚类结果的聚类中心表示客户对家居产品的偏好,针对聚类中心对客户进行分组,完成客户识别。

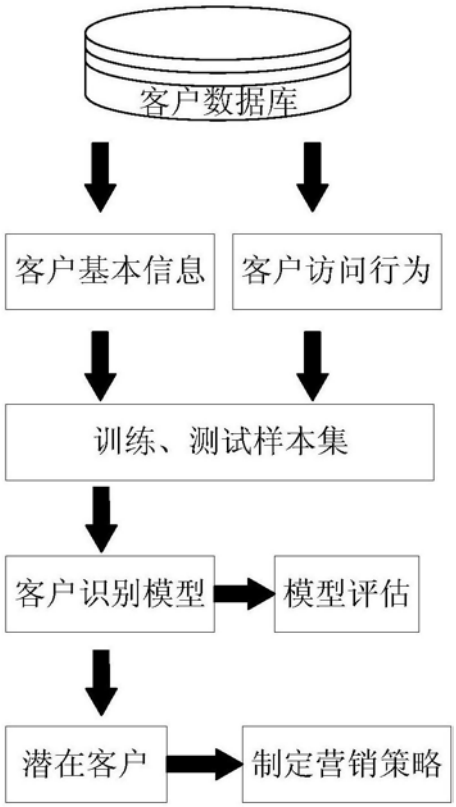

客户数据库
客户基本信息
客户访问行为
训练、测试样本集
客户识别模型
模型评估
潜在客户
制定营销策略

图1

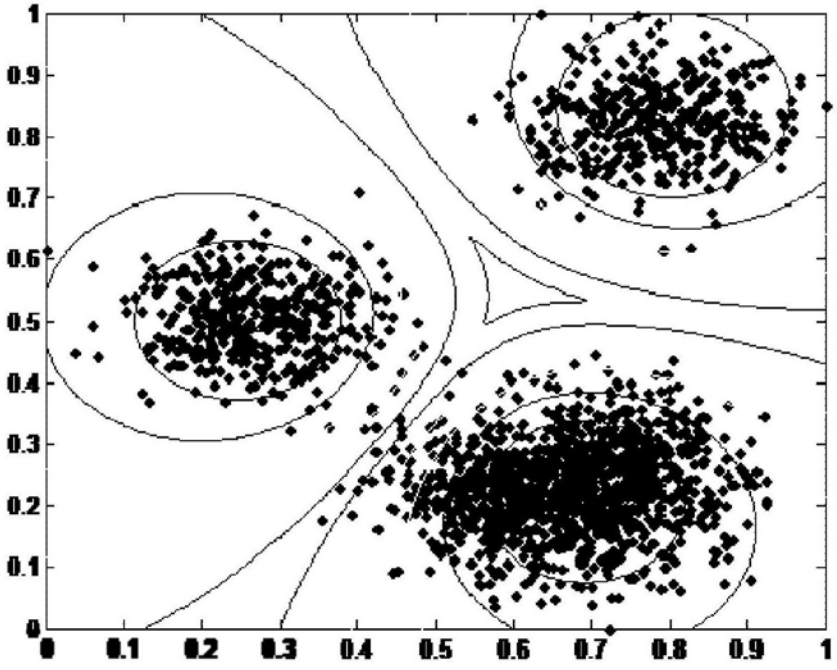

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

图2