运营商大数据轨迹聚类优化算法及其在疫情防控中的应用

Optimization Algorithm of Trajectory Clustering and Its Application for Epidemic Prevention Based on Telecom Big Data

成　晨，程新洲，晁　昆，张　涛，曹丽娟，徐乐西，韩玉辉，张晴晴（中国联通研究院，北京 100048）
Cheng Chen，Cheng Xinzhou，Chao Kun，Zhang Tao，Cao Lijuan，Xu Lexi，Han Yuhui，Zhang Qingqing（China Unicom Research Institute，Beijing 100048，China）

摘　要：
由于新冠病毒存在14天以上的潜伏期且在潜伏期具有传染性，密切接触者的排查至关重要，而运营商大数据以其独特的优势在寻找隐性接触中发挥着重要作用。在传统k-means聚类算法的基础上，优化损失函数并提出基于多目标函数的簇头选择算法，形成多目标轨迹聚类优化算法。在此基础上，构建基于运营商大数据的新冠肺炎疫情防控的密切接触者排查方法体系，将该算法用于隐性密切接触者的排查。

关键词：
多目标优化；鸟群觅食算法；数据挖掘；k-means；轨迹聚类
doi：10.12045/j.issn.1007-3043.2021.11.005
文章编号：1007-3043（2021）11-0023-05
中图分类号：TN915
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
As the incubation period for COVID-19，the time between exposure to the virus and the onset of symptoms，can be up to 14 days，during which these pre-symptomatic patients can be contagious，the investigation of close contacts is crucial，where telecom big data is playing an important role in finding close contacts with its unique advantages. In this paper，we propose a novel loss function and a cluster-head selection algorithm based on multi-objective optimization based on k-means algorithm. On the basis of the study，the multi-objective optimization algorithm of trajectory clustering has been brought out. By applying the algorithm in the investigation of close contacts，the system of precise and differentiated epidemic control measures based on telecom big data has been constructed.

Keywords：
Multi-objective optimization；Swarm intelligence；Particle Swarm Optimization；data mining；K-means；Trajectory Clustering

引用格式：成晨，程新洲，晁昆，等. 运营商大数据轨迹聚类优化算法及其在疫情防控中的应用[J]. 邮电设计技术，2021（11）：23-27.

1　概述

2019年末，首例新型冠状病毒肺炎在湖北省武汉市出现，并随着2020年春运期间的大规模人群迁徙迅速传播。经过艰苦卓绝的努力，我国疫情防控阻击战取得重大战略成果，目前已进入常态化的防疫阶段。

针对抗疫阻击战，习近平总书记多次作出重要批示指示，强调要运用大数据等手段，加强疫情溯源和监测。2020年，工业和信息化部多次召开疫情防控大数据专家会商会，传达国务院应对新型冠状病毒感染的肺炎疫情联防联控机制会议精神，研究部署大数据支撑服务疫情防控相关工作。

在疫情防控中，与感染者直接居住生活在一起、共同乘坐交通工具、乘电梯以及通过其他方式直接接触的人员被称为密切接触者。与显性密切接触者（共同居住生活或工作的人）相比，隐性密切接触者（无法

基金项目：工业和信息化部大数据产业发展试点示范项目（融合异构数据及深度学习的民生大数据创新应用试点示范）
收稿日期：2021-09-16

通过现有实名制数据直接追溯到的接触者）难以追溯和排查却依然存在感染风险。例如2020年1月19日，重庆市一名公交车乘客因为与一名患者相隔16秒登上同辆公交车，而被确诊为新冠肺炎患者；2020年1月22日，湖南某城市一个感染者乘坐公交车同时传染了13个人。

随着疫情防控工作逐渐常态化，对隐性密切接触者排查的精准化需求逐渐提升。运用传统的排查方法难以定位隐性接触者，而运营商大数据以其独特的优势在寻找隐性接触者时可发挥重要作用。运营商是天然的大数据集中地，拥有百万级的基站资源、亿级出账用户数、PB级日均数据生成及采集量，运营商大数据具备用户规模巨大、覆盖空间广、时间连续性强的优势，可以全面立体地刻画用户特征，为找到隐性接触者提供一定支撑。寻找乘坐共同的公共交通工具的隐性接触者，可以抽象为轨迹聚类问题，现有轨迹聚类算法的核心思想是采用欧式距离作为损失函数，基于k-means或基于密度的聚类算法进行轨迹聚类，而没有充分地考虑各类噪声数据对聚类结果的影响。另一方面，现有聚类方法多侧重于数据清洗后的聚类算法实施过程，而没有针对运营商OSS域大数据从预处理到模型训练的完整过程。

2 数据预处理与模型构建

2.1 OSS域数据预处理

运营商大数据主要分为两大类：BSS域数据和OSS域数据。BSS数据来自于业务支撑系统，主要涉及计费、营业情况、账务和客户服务资料。OSS数据来自于运营支撑系统，涉及核心网络电路域、分组域、无线网络基础数据。本文所述的隐性接触者的定位方法主要采用运营商OSS域大数据的XDR（X Detail Record）数据源的信令面数据，关联工参获得基站位置，即经纬度信息，最终得到如下数据字段：imsi、开始时间、结束时间、基站经度、基站纬度。

数据预处理流程如下。

a）确认某一感染者的移动轨迹，获取其近日所乘坐的公交车线路。

b）采集相同时间区间内的全市OSS域XDR信令面数据，并关联工参获取位置信息。

c）数据去冗余，对于同一日的每个imsi，每5 min保留1条数据，即保留00:00:00，00:05:00，00:10:00，……，23:55:00的数据，作为time_id。最终得到的数据包含如下字段：imsi、time_id、基站经度、基站纬度。

2.2 用户筛选与向量化

将特定时间范围内、特定地理范围的用户转换为二维向量，方法为：

a）设感染者为I，在t_1~t_n的时间内在某一公交车上。则按照第2.1节所述，进行预处理，将其转化为二维向量：I_msisdn=$\{[t_1,s_1],[t_2,s_2],\cdots,[t_n,s_n]\}$，其中$t_i$为time_id，$s_i$为一个二维数组[经度，纬度]。

b）筛选(t_1,t_n)时间，在$s_1,s_2,\cdots,s_n$位置点中的任意2个或2个以上位置点出现过的所有用户的msisdn。

c）将b）中所筛选得到的所有用户的位置进行向量化，对于每一个用户，将时空分布转化为：p_msisdn=$\{[t_1,s_1],[t_2,s_2],[t_3,s_3],[t_4,s_4],[t_5,s_5],[t_6,s_6]\cdots\}$。

d）找到每一个p_msisdn的$[t,s]$第1次和最后一次和I_msisdn中的$[t,s]$重合点，分别为$[t_f,s_f]$和$[t_l,s_l]$，将p_msisdn更新：p_msisdn=$\{[t_f,s_f],[t_{f+1},s_{f+1}],[t_{f+2},s_{f+2}],\cdots\cdots,[t_{l-1},s_{l-1}],[t_l,s_l]\}$。例如，假设$p$_msisdn的$[t,s]$第1次和最后一次和I_msisdn中的$[t,s]$重合点分别为$[t_3,s_3]$和$[t_{(n-2)},s_{(n-2)}]$，则更新$p$_msisdn为$\{[t_3,s_3],[t_4,s_4],[t_5,s_5],[t_6,s_6],\cdots\cdots,[t_{(n-2)},s_{(n-2)}]\}$。

3 多目标轨迹聚类算法

3.1 优化损失函数

现有的k-means聚类算法，采用欧式距离作为损失函数。但是移动网络实际所产生的XDR数据，可能存在有大量的离群点或噪声点的问题。例如，由于工参上报错误，导致用户所在小区的经纬度上报错误，使得用户所在的小区的实际位置和关联到的小区经纬度不一致，产生了大量的离群点。又如，由于乒乓效应，用户在相邻小区间反复切换，使用户经纬度反复跳跃；或者由于越区覆盖，不同用户在同一位置，关联得到的经纬度却不同，由此产生大量的噪声点。

如果采用欧式聚类作为损失函数，则会导致聚类效果受到噪声点和离群点的影响，效果不理想。因此引入Minkowski距离作为损失函数：

$$d=\left(\sum_{i=1}^{n}\left|x_i-y_i\right|^p\right)^{1/p} \tag{1}$$

其中，p越小，则对抗离群点的效果越好；p越大，则对抗噪声点的效果越好。

由于各个地域数据移动网络建设水平不同、数据采集厂商不同，数据中的离群点以及噪声点的占比无

法明确。为了在节约计算资源的情况下，尽可能准确地找到和感染者共同乘坐公共交通工具的用户，设计p值的边界作为损失函数的系数的最大值和最小值。

a）p_{min}=0.01，用于对抗离群点，例如工参错误带来的误差。

b）p_{max}=100，用于对抗噪声点，例如乒乓效应带来的误差。

3.2 基于多目标函数的簇头选择算法

k-means算法的核心思想是以样本之间的距离衡量相似性，通过将样本划分为若干提前设定好数量的簇，使得簇内各个样本间距离最小，而簇间距离最大。k-means算法选择簇头的方法的第1步是随机选择k个节点作为初始化的簇头选择方案，然后再进行更新迭代簇头位置。因此，若初始化簇头选择方案不合理，将会影响聚类效果。簇头选择问题是一个np-hard问题，传统线性算法难以找到最优解，因此引入群体智能算法解决该问题。另一方面，如3.1小节所述，移动网络实际所产生的XDR数据，可能存在有大量的离群点或噪声点的问题，因此需要引入Minkowski距离作为损失函数，且应根据现网数据特征选取不同的p值，从而形成不同的损失函数。因此，本文提出了一种群体智能算法——基于多目标的鸟群觅食算法，并将其应用于k-means算法的簇头选择过程中。

3.2.1 鸟群觅食算法

鸟群觅食算法（Particle Swarm Optimization，PSO）是群体智能算法中的一种，源于群居性生物通过自组织性的个体协作表现出的群体智能性，包括遗传算法、蛙跳算法等。本文提出的鸟群觅食算法旨在通过模拟鸟群寻找食物的过程寻找最优解，其过程为：通过每一次迭代，每只鸟通过飞行速度调整前进的距离和方向。其中，每只鸟的速度由它本身的运动过程中的最好适应度的位置以及整个鸟群的最好适应度的位置决定。算法步骤为：

设鸟群中有S只鸟，进行t次迭代，第i只鸟当前的位置为$X_i(t)=\left[X_{i1}(t),X_{i2}(t),...,X_{iM}(t)\right]$。速度为$V_i(t)=\left[V_{i1}(t),V_{i2}(t),...,V_{iM}(t)\right]$。每次迭代中，基于适应度值更新每只鸟的最好位置以及整个鸟群的最好位置。每只鸟的个体最好位置定义为局部最优解，即它在历次迭代中所经历的适应度最大的位置，表示为$P_i(t)=\left[P_{i1}(t),P_{i2}(t),...,P_{iM}(t)\right]$。整个鸟群的最好位置定义为全局最优解，即目前整个鸟群搜索到的最好位置，表示为$G(t)=\left[P_{g1}(t),P_{g2}(t),...,P_{gM}(t)\right]$，$1\leqslant g\leqslant M$。

每次迭代后，局部最优解的更新方式为：

$$P_i(t)=\begin{cases}P_i(t-1), & f\left[X_i(t)\right]<f\left[P_i(t-1)\right]\\ X_i(t), & f\left[X_i(t)\right]\geqslant f\left[P_i(t-1)\right]\end{cases}\tag{2}$$

每只鸟的速度的更新方式为：

$$V_{ij}(t+1)=V_{ij}(t+1)+c_1\times\text{rand}\times\left[P_{ij}(t)-X_{ij}(t)\right]+c_2\times\text{rand}\times\left[G_j(t)-X_{ij}(t)\right]\tag{3}$$

位置的更新方式为：

$$X_{ij}(t+1)=V_{ij}(t+1)+X_{ij}(t)\tag{4}$$

其中$1\leqslant i\leqslant S$，$1\leqslant j\leqslant M$，t为当前进化次数，T为预设的最大进化次数，c_1为其向局部最优解运动的步长，c_2为其向全局最优解运动的步长，rand服从[0，1]均匀分布，限制每只鸟的速度的变换范围为$V_{ij}(t)\in[-V_j^{max},V_j^{max}]$。当$t<T$时，不断重复上述过程，最终获得全局最优解$P(T)$。本场景中$V^{max}$=1，每次进化后，若$X_{i1}(t)>0$，则表示选择节点作为簇头；反之，则不选择该节点作为簇头。

3.2.2 多目标鸟群觅食算法

首先引入多目标鸟群觅食算法中的相关概念。

a）非支配解。假设多目标场景的目标为函数最大值，x和y是2个解，S是解集，若$f_i(x)\geqslant f_i(y),(i=1,2,\cdots m)$恒成立且$f_i(x)=f_i(y),(i=1,2,\cdots m)$不恒成立，则定义$x$为非支配解，$y$被$x$支配。

b）支配解。若$f_i(x)\leqslant f_i(y),(i=1,2,\cdots m)$恒成立且$f_i(x)=f_i(y),(i=1,2,\cdots m)$不恒成立，则$y$支配$x$，$y$称为非支配解。

c）Pareto最优解和最优解集。若存在解z不被任何解支配，则称其为Pareto最优解，所有的Pareto最优解组成了Pareto最优解集，该解集是算法进化的最终的目标。

引入多目标的鸟群觅食算法的步骤如下。

步骤1：初始化鸟群S，将每个向量$X_i(t)$的每一位随机赋值为-1或1，并初始化速度向量$V_i(t)$每一位为0，根据$X_i(t)$选择簇头初始化，然后按照k-means算法的更新过程进行轨迹聚类。分别令损失函数中的参数p赋值为0.01以及100，计算2种损失函数所对应的S集合的每个$X_i(t)$的适应度值。

步骤2：得到S中每个$X_i(t)$的非支配解等级并将其排序，即：首先遍历S中的每个解$y1$，得到每个$y1$支配的解集S_{y1}以及被$y1$支配的解的数量n_{y1}。若n_{y1}=0，

则没有任何一个解支配$y1$，则令$y1$的非支配等级为1。随后，对于所有非支配等级为1的解$y1$，遍历S_{y1}中的每个解$y2$以及被$y2$支配的解的数量n_{y2}，并令$n_{y2}=n_{y2}-1$，若$n_{y2}=0$，则把解$y2$放入集合S_{y2}，令S_{y2}中所有解的非支配等级为2。接着，对于所有非支配等级为2的解$y2$，遍历S_{y2}中的每个解$y3$以及被$y3$支配的解的数量n_{y3}，并令$n_{y3}=n_{y3}-1$，若$n_{y3}=0$，则把解$y3$放入集合S_{y3}，令S_{y3}中所有解的非支配等级为3。按同样的方法操作集合S_{y3}，得到非支配等级为4的集合，最终得到所有的解的非支配等级。

步骤3：计算每个解的拥挤度，针对非支配等级相同的解，按照拥挤度由小到大排列，把非支配等级为1的前M个解保存在集合Z中。

步骤4：用鸟群觅食算法进行进化，得到新的簇头初始化方案，然后按照k-means算法的更新过程进行轨迹聚类，针对每个损失函数，得到新的解集S_{new}。

步骤5：混合S和S_{new}，并按照非支配等级对里面的解进行排序，对于非支配等级相同的解，按照拥挤度从小到大排列，把前M个解保存在集合Z中。

步骤6：若未达到最大进化次数，另$S=S_{new}$，返回第5步；否则，进化完成，Z集合中的每个解都是一种簇头初始化方法，并得到对应的轨迹聚类结果。

3.3 多目标鸟群觅食算法性能测试

评价多目标优化问题，一般基于真实解与计算得出的解的距离（Generation Distance，GD）和分散度（Spacing，SP）2个指标。下面使用通用测试函数ZDT4函数评估算法性能，该函数可以有效评价算法是否容易陷入局部最优解：

$$\begin{cases}\min f_1(x)=x_1\\ \min f_2(x)=g(x)\times h(x)\end{cases} \tag{5}$$

其中，$g(x)=11+x_2{}^2-10\cos(4\pi x_2)$，

$$h(x)=\begin{cases}1-\sqrt{\dfrac{f_1(x)}{g(x)}} & ,f_1(x)\leqslant g(x)\\ 0 & ,\text{其他}\end{cases}。$$

仿真参数设定为：迭代次数T=1 000，鸟群规模为100，$c_1=c_2$=0.01，仿真结果如图1~图3所示。

由图1可知，多目标鸟群觅食算法求得的支配解集与真实Pareto最优解集基本相同，可见该算法可以有效找到最优解集。算法求出的非支配解集和真实Pareto最优解集见图2和图3。计算得到GD=4.425×10^{-4}，SP=0.004 105。2个值均较小，算法性能较好。

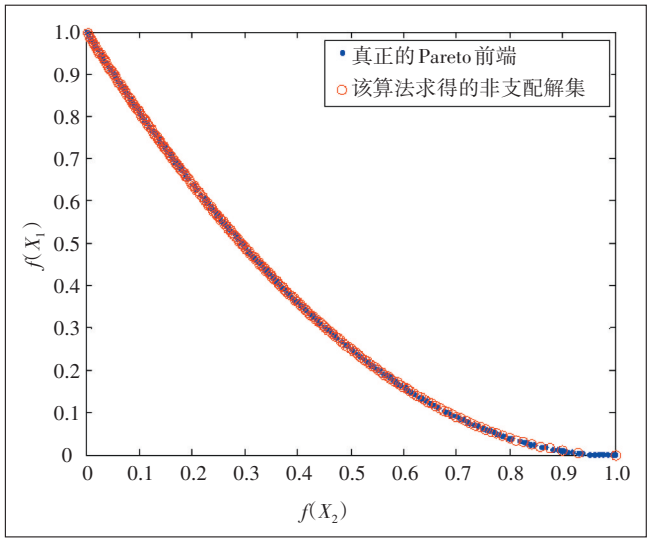

图1　多目标鸟群觅食算法求得的ZDT4最优解和真正的Pareto最优解对比图

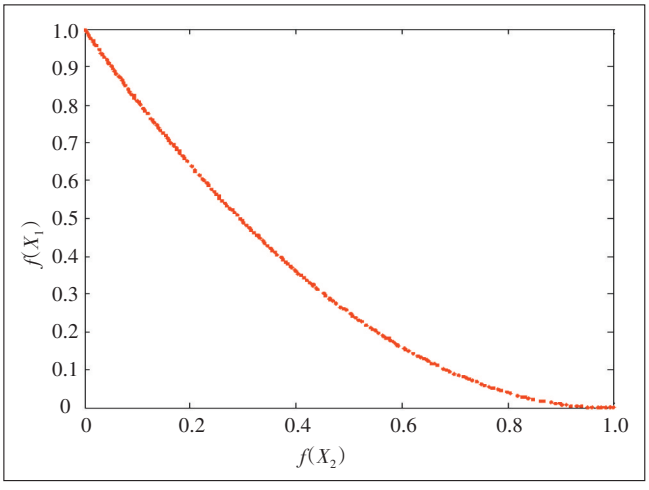

图2　多目标鸟群觅食算法求得的ZDT4函数最优解

4 用轨迹聚类优化算法获取隐性接触者

将多目标鸟群觅食算法引入k-means算法的簇头选择中，并将优化的k-means聚类算法用于轨迹聚类，方法如下。

a）将$I_msisdn=\{[t_1,s_1],[t_2,s_2],\cdots,[t_n,s_n]\}$进行分解，转化为如下向量集合：

①$I_msisdn_{1_n}=\{[t_1,s_1],[t_2,s_2],\cdots,[t_n,s_n]\}$

②$I_msisdn_{1_(n-1)}=\{[t_1,s_1],[t_2,s_2],\cdots,[t_{(n-1)},s_{(n-1)}]\}$

③$I_msisdn_{1_(n-2)}=\{[t_1,s_1],[t_2,s_2],\cdots,[t_{(n-2)},s_{(n-2)}]\}$

④$I_msisdn_{1_(n-3)}=\{[t_1,s_1],[t_2,s_2],\cdots,[t_{(n-3)},s_{(n-3)}]\}$

⑤…

⑥$I_msisdn_{2_n}=\{[t_2,s_2],[t_2,s_2],\cdots,[t_n,s_n]\}$

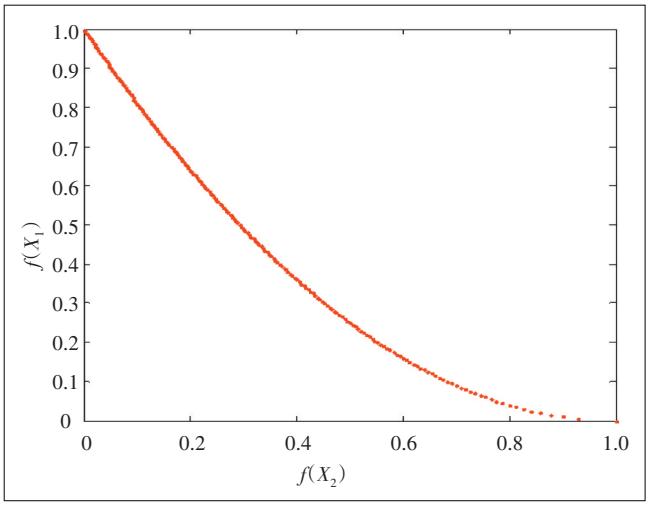

图3　ZDT4函数真正的Pareto最优解

⑦…

⑧$I_msisdn_{(n-1)_n}=\{[t_1,s_1],[t_2,s_2]\}$。

b）将a）中生成的所有I_msisdn按照向量的长度，由长到短排列。对于每一个相同长度的I_msisdn，找到所有相同长度的p_msisdn，分别令p=10以及p=0.1，作为损失函数。

c）用k-means聚类算法，找到所有与I_msisdn为同一聚类的用户，作为潜在的隐性接触者。

下面以某市某个感染者为例，假设有感染者1，其msisdn为1565176××××，某日他乘坐公共交通工具时，时空轨迹如图4所示。

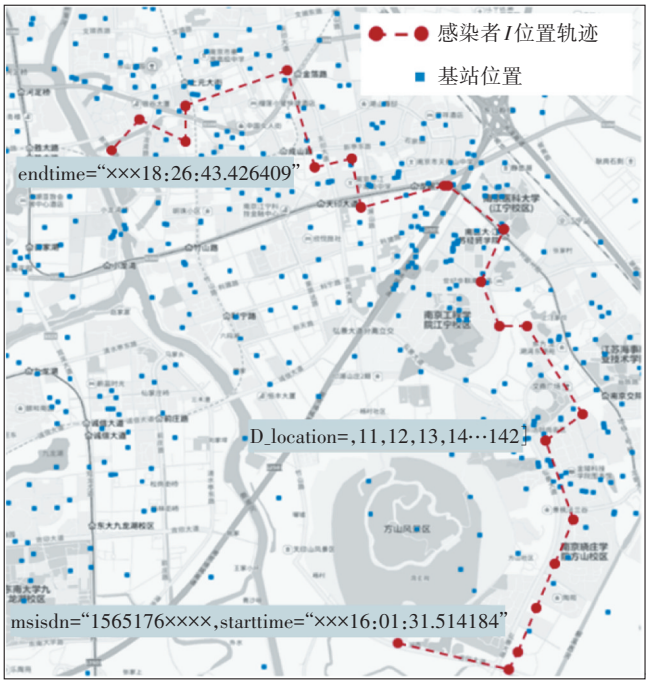

图4　感染者当日轨迹

按照第2.2节所述，将其时空轨迹向量化为I_msisdn，并将多目标鸟群觅食算法引入k-means算法的簇头选择中，并将优化的k-means聚类算法用于轨迹聚类，按照第3.3节所述方法进行轨迹聚类，共找到324名隐性密切接触者，可用于向防疫部门报送。

5　总结

目前我国已进入常态化疫情防控阶段，感染者的密切接触者排查对防止疫情扩散有着至关重要的作用。传统的密切接触者排查方法在一定程度上难以找到隐性密切接触者，给疫情防控工作带来了一定的挑战。而通过运营商大数据，可以获取用户的身份属性、时空轨迹、常驻区域、业务偏好、交际圈子等多维度信息，从而在疫情联防联控、人口流动洞察、疫情态势研判和决策等方面发挥重要的作用，特别是为隐性接触者的排查提供强有力的支撑。本文在传统k-means聚类算法的基础上，针对不同地区运营商数据的特征引入了Minkowski距离作为损失函数，提出基于多目标函数的簇头选择算法，形成了多目标轨迹聚类优化算法。在此基础上，构建基于运营商大数据和多目标函数的簇头选择算法的新冠肺炎疫情防控的密切接触者排查方法体系，助力隐性接触者的排查。

参考文献：

[1] PASSINO K M. Biomimicry of bacterial foraging for distributed optimization and control[J]. IEEE Control Systems Magazine, 2002, 22(3):52-67.

[2] CHENG C, CHENG X Z, YUAN M Q, et al. A novel cluster algorithm for telecom customer segmentation[C]//2016 16th International Symposium on Communications and Information Technologies (ISCIT). IEEE, 2016.

[3] GAO H Y, CAO J L. Non-dominated sorting quantum particle swarm optimization and its application in cognitive adio spectrum allocation [J]. Journal of Central South University, 2013(20):1878-1888.

作者简介：

成晨，工程师，主要从事通信大数据分析及挖掘等技术领域的研究工作；程新洲，教授级高级工程师，主要从事通信大数据分析及架构的研究工作；晁昆，高级工程师，主要从事通信大数据分析及产品等技术领域的研究工作；张涛，工程师，主要从事通信大数据分析及挖掘等技术领域的研究工作；曹丽娟，工程师，主要从事大数据算法研究及行业应用研究的工作；徐乐西，教授级高级工程师，博士，主要从事通信大数据行业应用等领域研究工作；韩玉辉，高级工程师，主要从事通信大数据行业应用、移动互联网DPI技术等领域的研究工作；张晴晴，工程师，主要从事通信大数据行业应用、移动互联网DPI技术等领域的研究工作。