

纵横交叉算法优化 FCM 在电力客户分类中的应用

孟安波¹, 卢海明¹, 李海亮², 谭火超², 郭壮志¹

(1. 广东工业大学, 广东 广州 510006; 2. 广东电网公司茂名供电局, 广东 茂名 525000)

摘要: 电力客户分类是供电企业客户关系管理的基石, 为了提高聚类算法的稳定性和精确性, 提出了一种纵横交叉算法(CSO)与模糊 C 均值算法(FCM)有机结合的新聚类算法(CSO-FCM), 并用新算法进行客户分类。新方法有效弥补了单一算法的不足, 拥有模糊理论处理不确定信息的能力以及纵横交叉算法全局收敛性强的特点。利用新算法对电力客户数据进行客观、科学的挖掘分析, 实现了对电力大客户较全面和准确的精细化分类, 为供电企业制定有针对性的营销策略提供了依据。

关键词: 客户细分; 纵横交叉算法; 模糊聚类; 增值服务

Electricity customer classification based on optimized FCM clustering by hybrid CSO

MENG Anbo¹, LU Haiming¹, LI Hailiang², TAN Huochao², GUO Zhuangzhi¹

(1. Guangdong University of Technology, Guangzhou 510006, China; 2. Maoming Power Supply Bureau of Guangdong Power Grid Corporation, Maoming 525000, China)

Abstract: Electrical consumers segmentation is the cornerstone of consumers relation management of electrical supply enterprises, in order to improve stability and exactness of clustering algorithm, this paper proposes a novel clustering algorithm to conduct consumers segmentation, which is organic combination by crisscross optimization algorithm and FCM. This method effectively compensates the demerits of single intelligent algorithm, which not only has the ability to dispose unstable information of fuzzy theory, but also has an advantage of global convergence of CSO. The new algorithm is used to objectively and scientifically analyze the electrical consumers data, achieving comprehensive and accurate segmentation, which can offer a pointed marketing strategies for enterprises.

Key words: consumers segmentation; crisscross optimization; fuzzy clustering; value-added services

中图分类号: F274

文章编号: 1674-3415(2015)20-0150-05

0 引言

电力体制改革的不断深化打破了国家对电力企业的垄断, 并将国内供电企业推向了市场。电力交易市场的建设使得客户在用电选择上有了更多的空间, 供电企业不能再依靠垄断的优势来留住和开发电力客户。因此, 要求供电企业必须增强市场意识, 注重企业效益, 重视客户资源, 对用电客户进行正确的评价, 实现客户细分。而经济学中的二八原则表明, 企业大部分利润的实现主要由少数的客户, 即大客户创造^[1-3]。可见, 发展和稳定电力大客户, 对供电企业提升核心竞争力具有十分重要的战略意义。

传统的电力客户细分方法依据一些定性的指标, 或者单一指标的定量分类方法来划分^[4-5]。随着管理信息系统和远程抄表技术的发展, 电力企业积累了大量的历史用电数据。目前, 对客户的分类主要通过客户数据的挖掘分析, 如文献[6]提出了模糊 C 均值算法作为聚类器, 应用于客户分类中。由于 FCM 聚类算法具有简单、快速的优点, 被广泛应用于数据分析中。但该算法对初始质心敏感, 并且基于梯度下降进行搜索常常使算法陷入局部最优。为了解决这一问题, 文献[7]将遗传算法引入聚类算法中, 虽然提高了算法的全局搜索能力, 但最优结果不稳定。文献[8]对 PSO 的粒子编码进行简化, 消除了样本向量维数对粒子群聚类算法的影响, 但操作步骤复杂, 控制参数较多。文献[9]提出了 GA-PSO-FCM 算法进行聚类分析, 该算法融合了遗

基金项目: 广东省自然科学基金(博士启动)项目(2013040013776); 广东省电网公司科技项目(K-GD2013-0789)

传算法全局搜索能力强和 PSO 收敛速度快的优点, 多数情况下比遗传算法更快地收敛于最优解, 而且可以避免完全随机寻优的退化现象。

针对现有聚类算法的不足, 本文提出一种纵横交叉算法优化 FCM 的客户分类新方法。纵横交叉算法(Crisscross optimization, CSO)是一种全新的群智能算法。经测试函数实验证明, CSO 在处理复杂的非线性、高维度问题时, 其收敛速度和收敛精度方面具有巨大优势, 并且 CSO 算法结构简单, 容易实现, 控制参数只有一个^[10]。新方法有效克服了 FCM 聚类结果不稳定和容易陷入局部最优的缺点, 且拥有模糊理论处理不确定信息的能力以及纵横交叉算法全局收敛性强的特点。实验分析表明, 在电力客户数据分析中, 取得了较好的聚类效果, 为供电企业识别不同的客户群体、实现差异性服务提供了依据。

1 纵横交叉算法

纵横交叉算法受儒家中庸思想和遗传算法中交叉操作的启发, 其搜索行为由横向交叉和纵向交叉两种方式组成, 每次迭代这两种行为将交替进行, 交叉后得出的解称为中庸解(MS_{hc} , MS_{vc})。通过结合一个简单的竞争机制, 使得横向交叉和纵向交叉完美的结合起来: 交叉后产生的子代, 与其父代进行竞争, 只有比父代适应度更好的子代才会被保留下来, 竞争后得出的解称为占优解(DS_{hc} , DS_{vc}), 类似于 PSO 中的个体极值 P_{best} 。

1.1 横向交叉操作

横向交叉是在种群中两个粒子相同维之间的算数交叉。首先对所有粒子进行随机两两配对, 假设父代粒子 $X(i)$ 和 $X(j)$ 的第 d 维进行横向交叉, 则它们根据式(1)、式(2)产生子代。

$$MS_{hc}(i, d) = r_1 \times X(i, d) + (1 - r_1) \times X(j, d) + c_1 \times (X(i, d) - X(j, d)) \quad (1)$$

$$MS_{hc}(j, d) = r_2 \times X(j, d) + (1 - r_2) \times X(i, d) + c_2 \times (X(j, d) - X(i, d)) \quad (2)$$

$$i, j \in N(1, M), \quad d \in N(1, D)$$

式中: r_1 、 r_2 是 $[0, 1]$ 之间的随机数; c_1 、 c_2 是 $[-1, 1]$ 之间的随机数; M 是粒子规模; D 是变量维数; $X(i, d)$ 、 $X(j, d)$ 分别是父代粒子 $X(i)$ 和 $X(j)$ 的第 d 维; $MS_{hc}(i, d)$ 和 $MS_{hc}(j, d)$ 分别是 $X(i, d)$ 和 $X(j, d)$ 通过横向交叉产生的第 d 维子代。

横向交叉中的两两配对将多维搜索空间划分成了种群数目一半的子空间, 配对的粒子以较大概率 (p_h , 通常取 1) 在以它们自己为对角顶点的超立方

体内产生后代 $MS_{hc}(i)$ 和 $MS_{hc}(j)$ 。同时, 为减少搜索盲点, 如式(1)、式(2)中的第 3 项所示, 横向交叉以较小概率在超立方体外缘进行搜索, 这种跨界搜索机制区别于遗传算法的交叉操作, 有效地增强了算法的全局搜索能力。交叉后进入竞争机制, 与其父代进行比较, 适应度更好的粒子保留在 DS_{hc} 中。

1.2 纵向交叉操作

纵向交叉是所有粒子不同维之间进行的一种算数交叉。由于不同维之间的取值范围可能不同, 因而交叉前须对各维进行归一化处理。每次纵向交叉操作只产生一个子代, 目的是使陷入停滞的那一维跳出局部最优而又不破坏其他正常维。假定粒子 $X(i)$ 的第 d_1 维和第 d_2 维进行纵向交叉, 则根据公式(3)产生子代。

$$MS_{vc}(i, d_1) = r \cdot X(i, d_1) + (1 - r) \cdot X(i, d_2) \quad (3)$$

$$i \in N(1, M), \quad d_1, d_2 \in N(1, D)$$

式中: $r \in [0, 1]$; $MS_{vc}(i, d_1)$ 是父代粒子 $X(i)$ 的第 d_1 维和第 d_2 维通过纵向交叉产生的子代。

在实际应用中, 绝大多数随机搜索算法的早熟问题往往是因为种群的部分维处于停滞不前, 称之为维局部最优。因此, 根据概率 p_v (通常取 0.2~0.8) 进行纵向交叉不仅能使陷入局部最优的维有机会摆脱出来, 进而使整个种群摆脱局部最优, 同时它的变异方式能较好地维持种群的多样性。交叉后同样进入竞争机制, 与其父代粒子进行比较, 择优保留在 DS_{vc} 中。

两种交叉方式的有机结合不仅加速了种群的收敛速度, 同时大大提高了收敛精度。一旦粒子陷入停滞不前的某维在纵向交叉操作下摆脱局部最优, 会通过横向交叉操作迅速传播至整个种群, 而更新后的维也会使其余陷入局部最优的维有更多机会通过纵向交叉跳出局部最优, 两种交叉方式的结果会呈链式反应在整个种群中蔓延。相比其他群智能优化算法, 这种纵横交叉机制使得 CSO 在全局收敛能力和收敛速度方面具有明显的优势。

2 CSO-FCM 算法简介

2.1 模糊 C 均值聚类(FCM)

FCM 是通过求取所有样本点与聚类中心距离加权的目标函数极值点而得到聚类中心迭代公式, 其权值是隶属度函数的指数^[11]。

设样本集 $Y = (y_1, y_2, \dots, y_i, \dots, y_n)$ 划分为 c 类, y_i 为任意维向量, 则聚类中心的集合为 $V = (v_1, v_2, \dots, v_c)$ 。各样本以一定的隶属度 u_{ij} 隶属于各个类, u_{ij}

表示第 j 个样本对第 i 个聚类中心的隶属度。 u_{ij} 满足式(4)条件。

$$\sum_{i=1}^c u_{ij} = 1, u_{ij} \in [0,1] \quad (4)$$

因此, FCM 可以表示为目标函数求极值的问题。

$$\min J_m = \sum_{j=1}^n \sum_{i=1}^c (u_{ij})^m (d_{ij})^2 \quad (5)$$

式中: n 为样本数; d_{ij} 为样本 x_j 与第 i 个聚类中心 v_i 之间的距离; m 为模糊系数, 是一个控制算法柔性的参数, 当 m 过大时, 聚类精度较低; 反之当 m 过小, 则算法接近传统的均值算法。

2.2 CSO 算法优化 FCM

CSO-FCM 聚类算法采用基于聚类中心的编码方式, 也就是每个粒子是由 c 个聚类中心组成。设样本向量维数为 d , 则粒子的维数是 $c \times d$, 粒子 x 的编码结构为: $x = [x_1^1, x_1^2, \dots, x_1^d, \dots, x_c^1, \dots, x_c^d]$ 。迭代过程中每个粒子的适应度按照式(6)确定^[9]。适应度函数为

$$fit = \frac{1}{J_m + 1} \quad (6)$$

如果聚类效果越好, 则 J_m 越小, 粒子适应度 fit 就越大。CSO-FCM 的终止条件为: 满足最大进化代数或 J_m 小于预先设定的 ε 或模糊矩阵不变。模糊矩阵不变可定义为

$$|u'_{ij} - u_{ij}^{t-1}| \leq \varepsilon \quad (7)$$

具体算法流程如下:

1) 对数据样本进行归一化处理, 设定聚类数目 c , 模糊系数 m , 种群大小 N , 控制变量数 D (即维数 $c \times d$), 最大迭代次数 $Maxgen$, 纵向交叉率 p_v , 横向交叉率 p_h 。

2) 初始化种群, 把每个粒子拆分成 c 个聚类中心, 按照隶属度大小把所有样本指派到不同簇中。

3) 按照式(6)计算每个粒子的适应度 fit 。

4) 根据式(1)、式(2)执行横向交叉, 产生的中庸解保存在矩阵 MS_{hc} 中, 然后计算 MS_{hc} 中每个解的适应值, 与其父代种群 X (即 DS_{vc} , 第一代除外) 进行比较, 只有适应度更好的粒子才能保留在 DS_{hc} 中。

5) 根据式(3)执行纵向交叉并将产生的解保存在矩阵 MS_{vc} 中, 然后计算 MS_{vc} 中每个解的适应值, 与其父代种群 X (即 DS_{hc}) 进行比较, 择优保留在 DS_{vc} 中。

6) 判断终止条件是否满足。若满足, 则迭代终止并将 DS_{vc} 中适应值最好的一组解拆分成 c 个聚

类中心, 按照隶属度大小把所有样本指派不同簇中, 并输入聚类结果。否则, 转到步骤 4) 进行新一轮迭代。

2.3 算法检验

为了测试本文提出的新方法对不同维数数据的分类效果, 采用来自权威的 UCI 机器学习数据库的 Iris 和 Wine 数据集样本^[12]。Iris 数据集样本的 4 个特征分别表示 Iris 数据的萼片长度、萼片宽度、花瓣长度和花瓣宽度, 由 3 类共 150 个样本组成, 类别分布各为 50 个; Wine 数据集由 3 类共 178 个样本组成, 每个样本有 13 个属性, 3 个类不存在交叠, 类别分布为 59, 71, 48。表 1 给出了 k-means 算法、基于 PSO 的 k-means 算法和 CSO 聚类算法的聚类精度, 其中第三组实验结果来自文献[8]。

CSO 优化过程的参数设置如下: 最大迭代次数 200 次, 种群数目为 30, 纵向交叉概率 p_v 设为 0.5, 平均聚类度是 20 次实验的平均值。

表 1 不同算法聚类精度比较

Table 1 Comparison of the accuracy of different clustering

| algorithms | | | | |
|-----------------------|------|--------|--------|--------|
| 算法 | 数据集 | 最高聚类精度 | 最低聚类精度 | 平均聚类度 |
| FCM | Iris | 89.33% | 82.00% | 87.30% |
| | Wine | 74.16% | 70.22% | 71.01% |
| PSO-FCM | Iris | 89.33% | 89.33% | 89.33% |
| | Wine | 70.22% | 70.22% | 70.22% |
| 简化编码 | Iris | 92.20% | - | - |
| PSO 聚类 ^[8] | Wine | 93.20% | - | - |
| CSO-FCM | Iris | 90.00% | 90.00% | 90.00% |
| | Wine | 94.94% | 94.94% | 94.94% |

由表 1 可见, 对于低维度的 Iris 的数据集, CSO-FCM 聚类的最高精度为 90.00%, 略比传统 FCM 和 PSO-FCM 高, 比改进的 PSO 聚类低; 对于 Wine 数据集, 本文方法的聚类精度明显高于另外三种算法, 最高聚类精度达 94.94%, 这是因为对 Wine 数据集进行聚类分析时粒子变量数 $D = 39$, 虽然 PSO-FCM 有可能收敛于全局最优点, 但它们容易早熟, 这种弊端在特征维数较高的情况下尤为明显。而 CSO 算法由于纵向交叉操作的加入, 能使陷入局部最优的维有机会摆脱出来, 进而使整个种群摆脱局部最优, 同时它的变异方式能较好地维持种群的多样性, 充分说明本文算法对于复杂的多特征、多类别数据集聚类效果具有明显优势。

3 实验分析

本文从某电业局的客户数据中根据以下三条标

准划分出电力大客户^[13]:

- (1) 供电电压不小于 10 kV;
- (2) 装接容量不小于 800 kVA;
- (3) 年用电量不小于 250 000 kWh。

该电业局供电客户 17.8 万户, 其中, 大工业客户营业户数只占总营业户数的 0.4%, 但用电量却占总用电量的 77.58%, 而用电量最多的前十位总用电量占整个地区用电量的 55%^[13]。

经过筛选, 符合条件的电力大客户共有 99 个, 每个客户有 4 个特征指标, 分别表示年用电量、用电增长率、平均电价和变异系数。经过 CSO-FCM

聚类后的客户分成了 4 类, 分别是重要保持客户、重要发展客户、无价值客户和一般客户, 具体指标数据如表 2 所示。CSO-FCM 聚类算法的客户细分流程如图 1 所示。

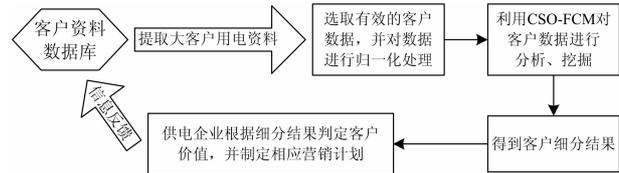


图 1 基于 CSO-FCM 聚类的客户细分流程图

Fig. 1 Customer segmentation flowchart based on CSO-FCM

表 2 通过 CSO-FCM 聚类分析后产生的客户类别

Table 2 Customer category produced by using CSO-FCM clustering analysis

| 客户类别 | 客户数量 | 平均用电量 (E) | 平均增长率 (G) | 平均电价 | 平均变异 系数(T) | 比较结果 | 客户级别 |
|------|------|---------------|--------------|---------|---------------|--------|--------|
| 1 | 6 | 206 309 055 | 63.338% | 0.479 1 | 0.203 | E↑G↑T↓ | 重要保持客户 |
| 2 | 8 | 13 362 323.38 | 56.646% | 0.542 | 0.382 | E↓G↑T↓ | 重要发展客户 |
| 3 | 11 | 27 191 557.56 | -16.297% | 0.478 | 0.482 | E↓G↓T↑ | 无价值客户 |
| 4 | 74 | 18 561 937.27 | 6.441% | 0.616 | 0.150 | E↓G↑T↓ | 一般客户 |

由表 2 可见, 重点保持客户数量仅为 5.05%, 但用电总量却达到了 37.93%, 为四个级别中最高, 保持高增长率同时用电波动小。针对此类能带来较大利润的客户, 供电企业应除基本服务外, 针对客户需求给予个性化的增值服务, 加强与客户的交流, 帮助其降低用电成本。重点发展客户虽然当前用电量不大, 但增长率达到 56.65%, 而且用电波动小, 具有较大的潜在价值。针对此类客户, 服务的重点是提供完备的技术支持和深入了解客户的发展需求。一般客户数占整个客户群总体的 77.78%, 表现为用电增长缓慢, 但用电波动小, 属于企业的优质客户, 针对此类客户营销重点在于提供更专业、更便捷的服务。对于无价值客户应重点了解和掌握这类客户出现用电负增长及用电波动大的原因, 引导他们规范用电。

4 结论

本文以电力客户用电信息为基础, 建立了以 CSO-FCM 为主的客户细分模型并予以论证。CSO-FCM 聚类汲取模糊聚类的处理不确定信息的能力与纵横交叉算法全局收敛的特点, 有效克服了传统 FCM 聚类结果不稳定和容易陷入局部最优等缺点。

实例分析中, 以某电业局电力大客户为研究对象, 使用本文提出的新方法对客户进行识别和细分, 由每一类的客户人数和对应的用电信息可以看出对客户分类的实际结果也符合“二八定律”。对大

客户的细分为供电企业识别不同客户群体的基本特征和需求差异, 进而为制定有针对性的差异化营销服务策略提供了依据, 也为构建以“客户为导向”的电力营销理念打下了基础。

参考文献

- [1] 吴杨, 刘俊勇. 基于云测度的电力营销状态评估及预警模型研究[J]. 电力系统保护与控制, 2013, 41(1): 61-69.
WU Yang, LIU Junyong. Research on assessment and forewarning model of power marketing condition based on cloud measure theory[J]. Power System Protection and Control, 2013, 41(1): 61-69.
- [2] 彭显刚, 赖家文, 陈奕. 基于聚类分析的客户用电模式智能识别方法[J]. 电力系统保护与控制, 2014, 42(19): 68-73.
PENG Xiangang, LAI Jiawen, CHEN Yi. Application of clustering analysis in typical power consumption profile analysis[J]. Power System Protection and Control, 2014, 42(19): 68-73.
- [3] FLINT D J, BLOCKER C P, BOUTIN JR P J. Customer value anticipation, customer satisfaction and loyalty: an empirical examination[J]. Industrial Marketing Management, 2011, 40(2): 219-230.
- [4] 李正明, 张纪华, 陈敏洁. 基于层次分析法的企业有序用电模糊综合评估[J]. 电力系统保护与控制, 2013, 41(7): 136-141.
LI Zhengming, ZHANG Jihua, CHEN Minjie. Fuzzy

- comprehensive evaluation of enterprise's orderly power utility based on analytic hierarchy process[J]. Power System Protection and Control, 2013, 41(7): 136-141.
- [5] CAI Z, AGUILAR F X. Consumer stated purchasing preferences and corporate social responsibility in the wood products industry: a conjoint analysis in the U.S. and China[J]. Ecological Economics, 2013, 95: 118-127.
- [6] 吴剑飞, 姚建刚, 陈华林, 等. 电力客户能效状态模糊综合评估[J]. 电力系统保护与控制, 2010, 38(13): 94-98, 103.
WU Jianfei, YAO Jiangang, CHEN Hualin, et al. Fuzzy comprehensive evaluation of power customer energy efficiency assessment[J]. Power System Protection and Control, 2010, 38(13): 94-98, 103.
- [7] 杨兴春, 王刚, 张安妮. 改进的 GA-FCM 算法及其在交通事故挖掘中的应用[J]. 计算机系统应用, 2010, 19(9): 159-162.
YANG Xingchun, WANG Gang, ZHANG Anni. Improved GA-FCM and its application in traffic data mining[J]. Computer Systems & Applications, 2010, 19(9): 159-162.
- [8] 魏新红, 张凯. 一种改进的 PSO-Means 聚类优化算法[J]. 河南科技大学学报: 自然科学版, 2011, 32(2): 41-43.
WEI Xinhong, ZHANG Kai. An improved PSO-Means clustering algorithm based on kernel methods[J]. Journal of Henan University of Science & Technology: Natural Science, 2011, 32(2): 41-43.
- [9] 雷浩籍, 刘念, 崔东君, 等. 基于 GA 与 PSO 混合优化 FCM 聚类的变压器故障诊断[J]. 电力系统保护与控制, 2011, 39(22): 52-56.
LEI Haoxia, LIU Nian, CUI Dongjun, et al. Transformer fault diagnosis based on optimized FCM clustering by hybrid GA and PSO[J]. Power System Protection and Control, 2011, 39(22): 52-56.
- [10] MENG A, CHEN Y, YIN H, et al. Crisscross optimization algorithm and its application[J]. Knowledge-Based Systems, 2014, 67: 218-229.
- [11] MUKHOPADHYAY A, MAULIK U, BANDYOPADHYAY S. Multiobjective genetic algorithm-based fuzzy clustering of categorical attributes[J]. IEEE Transactions on Evolutionary Computation, 2009, 13(5): 991-1005.
- [12] 傅涛, 孙亚民. 基于 PSO 的 k-means 算法及其在网络入侵检测中的应用[J]. 计算机科学, 2011, 38(5): 54-55.
FU Tao, SUN Yamin. PSO-based k-means algorithm and its application in network intrusion detection system[J]. Computer Science, 2011, 38(5): 54-55.
- [13] 程志炯. 乐山电业局大客户细分及增值服务研究[D]. 北京: 华北电力大学, 2009.
CHENG Zhijiong. Study on value-added service for key accounts in high-quality customer engineering of Leshan Power Bureau[M]. Beijing: China Electric Power University, 2009.

收稿日期: 2015-01-08; 修回日期: 2015-04-08

作者简介:

孟安波(1971-), 男, 博士, 副教授, 主要研究方向为电力系统自动化, 系统分析与集成;

卢海明(1990-), 男, 通信作者, 硕士研究生, 主要研究方向为智能算法在电力系统中的应用; E-mail: 496113194@qq.com

李海亮(1974-), 男, 高级工程师, 主要研究方向为电力市场营销的管理和研究工作。

(编辑 葛艳娜)