

纵横交叉算法在电力大客户细分中的应用

电力客户分类应用领域广泛涉及数据挖掘与特征提取问题,为了提高聚类算法的稳定性和准确性,提出一种基于纵横交叉(Crisscross Optimization, CSO)算法的聚类方法,能有效克服k均值聚类算法对初始质心敏感,容易陷入局部极值的缺点。CSO算法采用一种双交叉搜索机制,其中横向交叉引入扩展因子增强全局搜索能力,纵向交叉引入维交叉概念,从而避免维局部最优问题。两种交叉算子交替产生中庸解,通过与父代竞争产生的占优解在种群中相互催化,从而避免早熟问题的同时能够迅速收敛到全局最优。利用新方法对电力大客户数据进行客观、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。

李超 谭火超/广东电网公司茂名供电局
孟安波 卢海明 郭壮志/广东工业大学自动化学院



李超/工程师

随着电力体制改革的不断深化,作为直接面向市场的供电企业,如何优化营销策略,实现多元化和差异化服务,满足不同客户的多元化需求,提高核心竞争力,扩大电能在社会消费终端中的占有率,已成为供电单位的一项迫切任务^[1-5]。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^[6]。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^[7]。

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,如地域、电压等级和受电容量等^[5-6]。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对客户数据的挖掘分析,由于K均值聚类算法具有简单、快速的优点,被广泛应用于数据分析中。但K均值算法对初始质心敏感,没有良好的稳定性,并且基于梯度下降进行搜索常常使算法陷入局部最优^[8-9]。为了解决这一问题,文献[10]将遗传算法引入聚类算法中,但是实验结果表明当聚类问题的规模较大时,该算法会出现早熟现象,且在进化过程中可能产生退化现象^[11]。文献[8]利用PSO算法进行聚类分析,与遗传算法相比PSO算法简单易行,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。但缺点是PSO算法面对大规模问题同样容易陷入局部最优。文献[12]对PSO算法的粒子编码进行简化,消除了样本相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。

针对以上聚类算法的不足,本文提出一种基于纵横交叉(Crisscross Optimization, CSO)算法的聚类算法。纵横交叉算法是一种全新群智能算法。经测试函数实验证明^[13],CSO算法在处理复杂的非线性、高维度问题时,其收敛速度和收敛准确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。实验分析表明,在电力大客户数据分析中,取得了较好的聚类效果,为供电企业识别不同的客户群体,现实差异性服务提供了依据。

关键词/Keywords

客户细分·
纵横交叉算法·
聚类分析·
增值服务·

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)。

CSO算法

CSO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地结合起来: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. 横向交叉

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,则其根据式(1)、式(2)产生子代,即

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

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

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

式中, r_1 、 r_2 为 $[0,1]$ 之间的随机数; c_1 、 c_2 为 $[-1,1]$ 之间的随机数; M 为粒子规模; D 为变量维数; $X(i, d)$ 、 $X(j, d)$ 分别为父代粒子 $X(i)$ 和 $X(j)$ 的第 d 维; $MS_{hc}(i, d)$ 和 $MS_{hc}(j, d)$ 分别为 $X(i, d)$ 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 $1/2$ 的子空间,配对的粒子以较大概率(p_h ,通常取1)在以其本身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。

2. 纵向交叉

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,则根据式(3)产生子代,即

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

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

式中, $r \in [0,1]$; $MS_{vc}(i, d_1)$ 为父代粒子 $X(i)$ 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 $0.2 \sim 0.8$)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,同时其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。

CSO算法的整体流程如图1所示。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准确度。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算法在全局

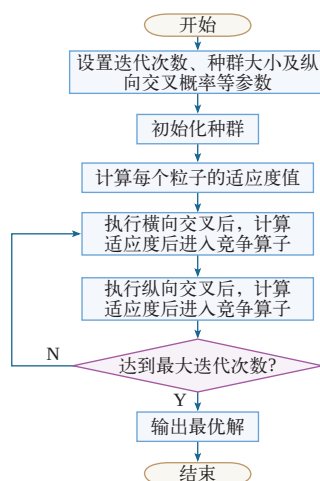


图1 基于CSO聚类的客户细分流程图

收敛能力和收敛速度方面具有明显的优势。

基于CSO的聚类算法

基于CSO的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 m 个聚类中心组成(m 为数据集的类别数)^[14]。设样本相量维数为 d ,则粒子的维数是 $m \times d$,粒子 x 的编码结构为 $x = [c_1^1, c_1^2, \dots, c_1^d, \dots, c_m^1, \dots, c_m^d]$ 。迭代过程中的适应度函数按照最近邻法则确定,即每个样本被划分到与其最近的聚类中心代表的簇。适应度函数如下

$$\text{fit} = \sum_{i=1}^n \sum_{j=1}^m d(x_j^i, c_j) \quad (4)$$

式中, n 为簇的样本数量; c_j 为第 j 个聚类中心; x_j^i 为属于第 j 簇的第 i 个样本; $d(x_j^i, c_j)$ 为 x_j^i 与 c_j 之间的欧氏距离。

具体算法流程如下。

- 1) 对数据样本进行归一化处理,设定算法参数。
- 2) 初始化种群,把每个粒子拆分成 m 个聚类中心,按照最近邻法则把所有样本指派到不同簇中。
- 3) 按照式(4)计算每个粒子的适应度 $\text{fit}(x)$ 。
- 4) 根据式(1)和式(2)执行横向交叉,产生的中庸解保存在矩阵 $\mathbf{MS}_{\text{hc}}$ 中,然后计算 $\mathbf{MS}_{\text{hc}}$ 中每个解的适应值,与其父代种群 $\mathbf{X}$ (即 $\mathbf{DS}_{\text{vc}}$,第一代除外)进行比较,只有适应度更好的粒子才能保留在 $\mathbf{DS}_{\text{hc}}$ 中。
- 5) 根据式(3)执行纵向交叉并将产生的解保存在矩阵 $\mathbf{MS}_{\text{vc}}$ 中,然后计算 $\mathbf{MS}_{\text{vc}}$ 中每个解的适应值,与其父代种群 $\mathbf{X}$ (即 $\mathbf{DS}_{\text{hc}}$)进行比较,择优保留在 $\mathbf{DS}_{\text{vc}}$ 中。
- 6) 判断终止条件是否满足。如果迭代次数大于设定的最大值时,则迭代终止并将 $\mathbf{DS}_{\text{vc}}$ 中适应值最好的一组解拆分成 m 个聚类中心,按照最近邻法则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。

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。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和48。表1为 k -means算法、基于PSO的 k -means算法和CSO聚类算法的聚类准确度比较,其中前两种算法实验结果来自文献[8]。

表1 不同算法聚类准确度比较

(单位: %)

算法	数据集	最高聚类准确度	最低聚类准确度	平均聚类度
k-means ^[8]	Iris	89.33	82.00	87.30
	Wine	74.16	70.22	71.01
PSO-k-means ^[8]	Iris	89.33	89.33	89.33
	Wine	70.22	70.22	70.22
CSO聚类	Iris	89.33	88.67	88.95
	Wine	94.38	94.38	94.38

CSO算法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。

由表1可见,对于低维度的Iris的数据集,CSO聚类的最高准确度为89.33%,与传统 k -means和PSO-k-means一样,最低聚类准确度略比PSO-k-means低;对于Wine数据集,本文方法的聚类准确度明显高于另外两种算法,最高聚类准确度达94.38%,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-k-means有可能收敛于全局最优点,但容易早熟,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。

实验分析

本文从某电业局的客户数据中根据以下3条标准划分出电力大客户^[15]。

- 1) 供电电压不小于10 kV。
- 2) 装接容量不小于800 kV·A。

3) 年用电量不小于250 000 kW·h。
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%^[15]。

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聚类算法的客户细分流程如图2所示。

表2 通过CSO聚类分析后产生的客户类别

客户类别	客户数量	平均用电量E/kW·h	平均增长率G (%)	平均电价/元	平均变异系数T	比较结果	客户级别
1	5	217 665 488.6	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	9	27 191 557.56	- 16.297	0.478	0.482	E ↓ G ↓ T ↑	无价值客户
4	77	18 561 937.27	6.441	0.616	0.150	E ↓ G ↑ T ↓	一般客户

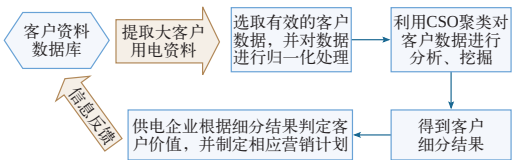


图2 基于CSO聚类的客户细分流程图

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4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，供电企业应提供除基本服务外，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。重点发展客户虽然用电总量不大，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，引导其规范用电。

结束语

针对k-means聚类算法存在的问题，本文提出一种基于CSO的聚类算法，测试数据及实验证明新方法有效克服了k-means聚类算法对初始质心敏感，容易陷入局部极值的缺点。尤其在多特征、多类别的数据集分类中优势更加明显，证明

了该算法在信息挖掘和分类方面的优越性。

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，进而制定有针对性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] 董世安. 基于客户价值评估的供电企业客户关系管理研究 [D]. 北京: 华北电力大学, 2010.

[2] 刘万林, 栗秋华, 周林, 等. 基于加速遗传算法和Shepard插值理论的大电力客户信用评估模型 [J]. 电气应用, 2010, 29 (9): 46-50.

[3] 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 (10): 118-127.

[4] Dzobo O, Alvehag K, Gaunt C T, et al. Multi-dimensional customer segmentation model for power system reliability-worth analysis [J]. International Journal of Electrical Power & Energy Systems, 2014, 62 (2): 532-539.

(下转第82页)

- [3] 谢俊文, 陆继明, 毛承雄, 等. 基于变平滑时间常数的电池储能系统优化控制 [J]. 电力系统自动化, 2013, 37 (1): 97-102.
- [4] 侯朝勇, 胡学浩, 惠东. 锂电池储能并网变换器的设计与控制 [J]. 电网技术, 2012, 36 (13): 247-251.
- [5] Haihua Z, Bhattacharya T, Duong Tran, et al. Composite energy storagesystem involving battery and ultracapacitor with dynamic energymanagement in microgrid applications [J]. IEEE Transactions on Power Electronics, 2011, 26 (3): 923-930.
- [6] 史林军, 陈中, 王海风, 等. 应用飞轮储能系统阻尼电力系统低频振荡 [J]. 电力系统自动化, 2010, 34 (8): 29-33.
- [7] 周念成, 郭燕, 王强钢. 飞轮储能系统建模与及提高电网暂态稳定性研究 [J]. 华东电力, 2011, 39 (5): 724-728.
- [8] 颜志敏, 王承民, 郑健, 等. 配电网中蓄电池储能系统价值评估 [J]. 电力自动化设备, 2013, 33 (2): 57-61.
- [9] Bludszuwei H, Domínguez-Navarro J A. A probabilistic method for energy storage sizing based on wind power forecast uncertainty [J]. IEEE Transactions on Power Systems, 2011, 26 (3): 1651-1658.
- [10] 曾鸣, 李娜, 马明娟, 等. 电力市场环境下的风储联合运行系统的经济效益评估模型 [J]. 华东电力, 2012, 40 (9): 1439-1443.
- [11] Fouad Abou Chacra, Patrick Bastard, Gilles Fleury, et al. Impact of energystorage costs on economical performance in a distribution substation [J]. IEEE Transactions on Power Systems, 2005, 20 (2): 684-691.
- [12] 严干贵, 冯晓东, 李军徽, 等. 用于松弛调峰瓶颈的储能系统容量配置方法 [J]. 中国电机工程学报, 2012, 32 (28): 27-35.
- [13] 肖晓伟, 肖迪, 林锦国, 等. 多目标问题的研究综述 [J]. 计算机应用研究, 2011, 3 (5): 15-19.

(收稿日期: 2014-09-10) EA

(上接第61页)

- [5] 杨淑霞, 李翔, 黄陈锋. 区间数信息下的用电客户信用评价 [J]. 电气应用, 2005, 24 (11): 24-27.
- [6] 张铮, 闫红漫, 陆融. 电力监控系统跨平台集成方案的研究与实现 [J]. 电气应用, 2008, 27 (18): 65-68.
- [7] 靳方明. 基于客户价值分析的电力客户服务管理研究 [D]. 北京: 华北电力大学, 2007.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 杨乐新, 吴宾宾, 万旭. 论历城电力的增供扩销策略与实践 [J]. 电气应用, 2013 (S1): 229-233.
- [12] 王纵虎, 刘志镜, 陈东辉. 一种改进的粒子群优化快速聚类算法 [J]. 西安电子科技大学学报, 2012 (5): 61-65.
- [13] Meng A, Chen Y, Yin H, et al. Crisscross optimization algorithm and its application [J]. Knowledge-Based Systems, 2014, 67 (3): 218-229.
- [14] 王春叶. 基于数据挖掘的电力客户细分研究 [D]. 保定: 华北电力大学, 2009.
- [15] 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.

(收稿日期: 2015-01-04) EA