计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(1):146−153 [doi: 10.15888/j.cnki.csa.007737]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于高斯混合聚类的风电出力场景划分①

张发才[1,2], 李喜旺[1], 樊国旗[3]

[1](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[2](中国科学院大学, 北京 100049)
[3](国网金华供电公司, 金华 321001)
通讯作者: 张发才, E-mail: zhangfacai_ucas@163.com

摘　要: 目前基于相似度的聚类方法对风电出力场景进行聚类划分, 而相似度又大多采用欧式距离长短作为衡量依据, 其结果反映时间序列曲线的幅度大小差异, 未能反映出曲线的形态特征及变化趋势的不同. 本文提出一种基于高斯混合聚类的风电出力场景划分的方法, 即通过属于某一类的概率大小来判断最终的归属类别. 首先根据 BIC 准则, 肘部法则和轮廓系数分别确定 GMM 聚类和 K-means 聚类的最佳数量, 然后以某地区实际风电为研究对象, 提取该地区 3 年春季风电出力典型场景, 并对这两种聚类结果进行对比分析, 验证本文方法的有效性. 最后通过 GMM 聚类模型提取该地区各个季节风电出力典型场景.

关键词: 聚类划分; 最佳聚类数; GMM; 典型场景

引用格式: 张发才,李喜旺,樊国旗.基于高斯混合聚类的风电出力场景划分.计算机系统应用,2021,30(1):146–153. http://www.c-s-a.org.cn/1003-3254/7737.html

Wind Power Output Scene Division Based on Gaussian Hybrid Clustering

ZHANG Fa-Cai[1,2], LI Xi-Wang[1], FAN Guo-Qi[3]

[1](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](State Grid Jinhua Power Supply Company, Jinhua 321001, China)

Abstract: At present, the clustering method based on similarity is used to classify the wind power output scene, and the similarity is mostly measured by the Euclidean distance. Hence, the results reflect the difference of the amplitude of the time series curve, not the difference of the morphological characteristics and changing trend of the curve. This study proposes a method of wind power output scene division based on Gaussian mixture clustering, that is, the final attribution category is judged by the probability of belonging to a certain category. Firstly, the optimal numbers of GMM clustering and K-means clustering are determined according to BIC criterion, elbow rule and contour coefficient, respectively. Then, taking the actual wind power in a certain area as the research object, the typical scenes of wind power output in spring in this area are extracted, and the two clustering results are compared and analyzed to verify the effectiveness of this method. Finally, the typical scenes of wind power output in each season in this region are extracted by GMM clustering model.

Key words: clustering; optimal number of clusters; GMM; typical scene

① 基金项目: 国家科技重大专项 (2017ZX01030-201)
Foundation item: National Science and Technology Major Program (2017ZX01030-201)
收稿时间: 2020-05-26; 修改时间: 2020-06-23; 采用时间: 2020-06-28; csa 在线出版时间: 2020-12-31

近年来, 中国风力发电发展速度快, 风电场的规模以及风电并网比例不断增大. 与传统的发电方式相比, 风力发电最根本的不同点在于其有功出力的随机性、间歇性和不可控性[1]. 由于地理地貌和季风变化影响着风电资源分布, 风电出力的随机性变化具有一定的季节周期性[2], 用典型场景集反映周期内风电出力的变化特征, 对含有风电电力系统的规划和调度具有重要意义.

目前, 风电出力典型场景的选取主要采用聚类划分方法. 文献 [3] 介绍了聚类算法大致可以分为层次聚类算法, 划分式聚类, 基于密度和网格的聚类算法和其他算法. 文献 [4] 提出基于改进 K-means 聚类的风电功率典型场景. 文献 [1] 采用改进的模糊 C 均值聚类算法和分层聚类算法, 实现对风电出力场景的选取. 文献 [5,6] 采用 K-means 算法对风电出力样本进行聚类划分, 得到具有代表性的典型风电出力场景. 文献 [7] 提出基于 Wasserstein 距离和改进 K-medoids 聚类算法, 构建覆盖调度空间的典型场景. 文献 [8] 提出主成分分析法和分层聚类算法相结合的方法, 计算出年度典型风电出力场景. 文献 [9] 采用模糊 C 均值聚类法, 完成对所研究区域风电功率典型场景的提取.

以上聚类算法都是以欧氏距离作为样本相似度判断, 欧氏距离能反映样本曲线间的远近程度, 不能反映曲线形态的相似程度. 文献 [10] 提出基于考虑序列互相关性的“形态距离”的聚类算法, 并提取春, 夏, 秋, 冬的风电出力典型场景, 避免了基于欧式距离聚类的缺点. 文献 [11] 比较得出 GMM (Gaussian Mixture Model) 聚类质量优于层次聚类, K-means, K-medoids, SOM 聚类. 文献 [12] 提出基于高斯混合模型的公交出行特征分析. 文献 [13] 通过应用高斯混合模型对伊朗西南部某水域进行分区, 取得很好的效果. 文献 [14] 提出基于 EM 和 GMM 的朴素贝叶斯岩性识别, 结果表明高斯混合模型有很好的拟合效果. 文献 [15] 采用 GMM 聚类进行汉语数字识别. 此外, GMM 聚类不仅具有灵活的类簇形状, 还能够很好的捕获属性之间的相关性和依赖性[16].

本文提出了一种基于概率分布的高斯混合聚类模型 GMM, 通过样本属于某一类的概率大小来判断其归属类别, 本文选取某地区的风电出力情况, 与传统的基于欧式距离的聚类算法的划分结果对比分析, 验证本文提出的风电出力场景划分方法的有效性.

1 基于高斯混合聚类对风电出力场景的划分方法

1.1 高斯混合聚类模型

高斯混合模型是由有限个独立的多元高斯分布模型线性组合而成, 每一个多元高斯分布成为混合高斯模型的成分, 而多元高斯分布则是一元高斯分布在高纬度空间中的扩展[17].

假设一天内每个小时的风电功率为$x_i(i=1,2,\cdots,24)$, 则高斯混合模型可以表示为:

$$p(x)=\sum_{k=1}^{K}\alpha_k N(x_k|\mu_k,S_k) \tag{1}$$

高斯混合模型有 3 个参数需要估计, 分别为μ, α和S, 其中, μ表示模型的期望, α表示各个分布的权重, S 表示模型的方差.

上式可化为

$$p(x|\alpha,\mu,S)=\sum_{k=1}^{K}\alpha_k N(x|\mu_k,S_k) \tag{2}$$

下面采用最大似然法 (EM) 进行参数估计.

算法步骤如下:

(1) 指定μ, α和S的初始值.

(2) 计算后验概率$\gamma(z_{nk})$:

$$\gamma(z_{nk})=\frac{\alpha_k N(x|\mu_k,S_k)}{\sum_{j=1}^{K}\alpha_j N(x|\mu_j,S_j)} \tag{3}$$

(3) 求解μ_k的最大似然函数:

$$\mu_k=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n \tag{4}$$

(4) 求S_k的最大似然值

$$S_k=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k)(x_n-\mu_k)^{\mathrm{T}} \tag{5}$$

(5) 求解α_k的最大似然函数

$$\alpha_k=N_k/N \tag{6}$$

(6) 循环重复计算步骤 (2)~(5), 直至算法收敛.

1.2 最佳聚类数目确定方法

对于最佳聚类个数确定, GMM 聚类往往是采用 BIC 准则[18]. 贝叶斯信息准则 (Bayesian Information Criterion, BIC), 1978 年由 Schwarz 提出, 用于实际中选择最优的模型, 如式 (7):

$$BIC = k\ln(n) - 2\ln(L) \tag{7}$$

其中, k 为模型参数个数, n 为样本数量, L 为似然函数, $k\ln(n)$ 惩罚项在维数过大且训练样本数据相对较少的情况下, 可以有效避免出现维度灾难现象.

对于 K-means 聚类, 采用肘部法则和轮廓系数相结合的方法确定最佳聚类数目. 肘部法则的核心指标是误差平方和 (Sum of the Squared Errors, *SSE*), 如式 (8):

$$SSE = \sum_{i=1}^{k}\sum_{p\in C_i}|p - m_i|^2 \tag{8}$$

其中, C_i 是第 i 个簇, p 是 C_i 的样本点, m_i 是 C_i 的质心, SSE 是所有样本的聚类误差, 代表了聚类效果的好坏.

当 k 小于真实聚类数时, 由于 k 的增大会大幅增加每个簇的聚合程度, 故 *SSE* 的下降幅度会很大, 而当 k 到达真实聚类数时, 再增加 k 所得到的聚合程度会迅速变小, 所以 *SSE* 的下降幅度会骤减, 然后随着 k 值的继续增大而趋于平缓, *SSE* 和 k 的关系图是一个手肘的形状, 而这个肘部对应的 k 值就是数据的真实聚类数

轮廓系数是类的密集与分散程度的评价指标, 如式 (9):

$$s = (b - a)/\max(a, b) \tag{9}$$

其中, a 表示样本到彼此间距离的均值, b 表示样本到除自身所在簇外的最近簇的样本的均值, s 取值在 [−1, 1] 之间, 如果 s 越接近 1, 代表所在簇合理, 如果 s 越接近−1, s 应该分到其他簇中. 对于使用轮廓系数确定聚类的数量, 应该选取较大的轮廓系数.

2 实验和结果分析

风电出力通常具有明显的季节分布特性, 与单风电场相比, 一个地区的风电功率具有更明显的季节性规律.

首先选取某地区 2017 年至 2019 年 3 年春季 3 个月的每 1 小时实测地区风电出力数据进行分析, 验证该方法的有效性, 然后再对该地区其他季节风电出力特性进行分析.

2.1 最佳聚类数目确定

使用 *BIC* 对高斯混合模型进行选择, 既涉及协方差的类型, 也涉及模型中聚类的数量. 如图 1 所示, 其中 spherical, tied, diag, full 分别对应球面协方差矩阵, 相同的完全协方差矩阵, 对角协方差矩阵, 完全协方差矩阵, GMM 应选择聚类数目为 4 的和相同的完全协方差矩阵.

针对 K-means 聚类, 综合考虑 *SSE* 和轮廓系数, 如图 2 所示, 蓝色曲线表示 *SSE* 随着 k 变化的曲线, 红色曲线表示轮廓系数随着 k 变化的曲线. 一般来说, 平均轮廓系数越高, 聚类的质量也相对较好. 在这, 最优聚类数应该是 2, 这时平均轮廓系数的值最高. 但是, 聚类结果 (k=2) 的 *SSE* 值太大了, 根据肘部法则, 当 k=4 时, *SSE* 的值会低很多, 但此时平均轮廓系数的值较高. 因此, k=4 是最佳的选择.

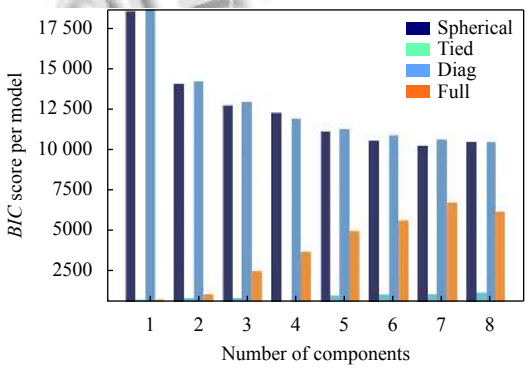


图 1 混合高斯模型参数选择

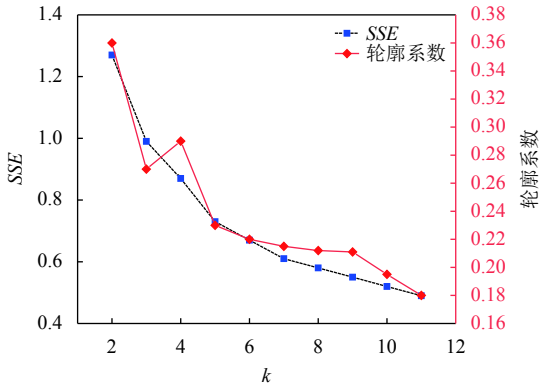


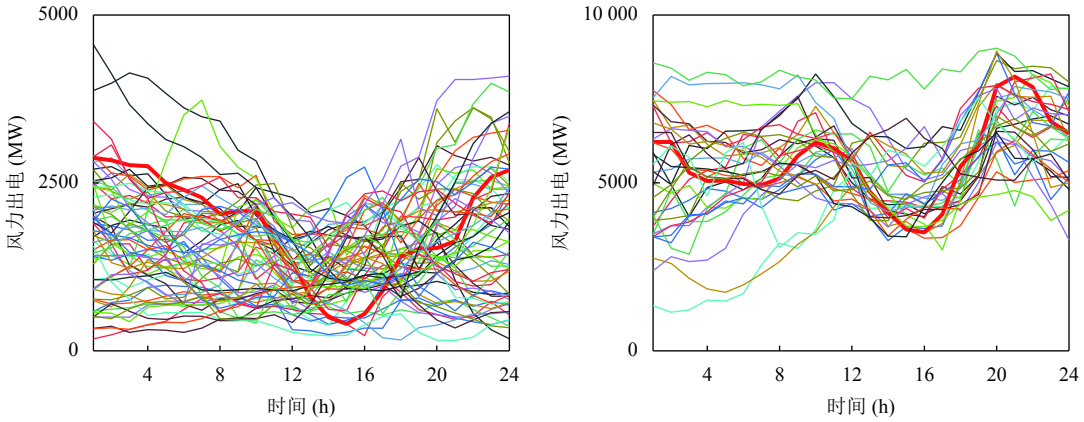
图 2 K-means 聚类模型的 k 选择

2.2 基于概率的聚类划分

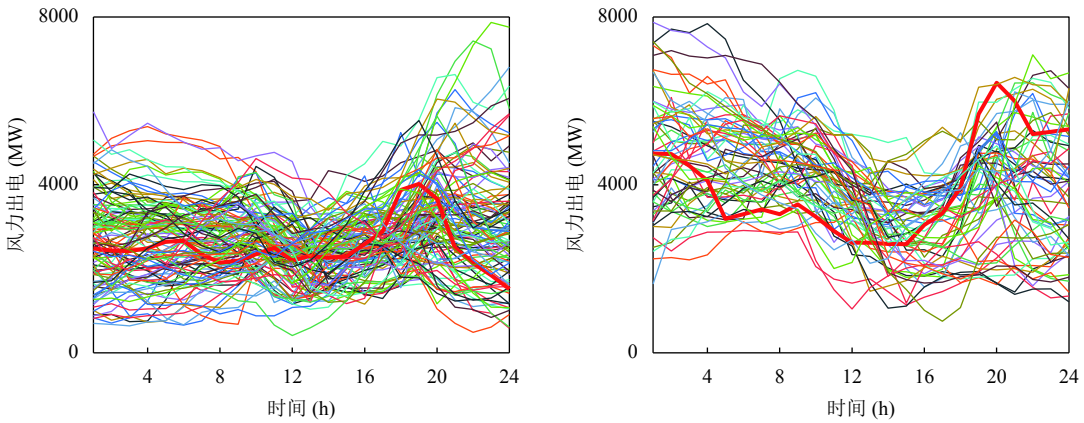
图 3 中, 红色曲线为聚类中心, 代表该地区风电风力的典型场景. 在 4 种形态的样本簇中, 每一簇的风力出力功率范围明显不同, 大部分的类内样本都与聚类中心相似, 只有少数曲线的形状与中心曲线的形态不同.

2.3 基于欧式距离的聚类划分

采用 K-means 聚类算法对同一组数据进行聚类划分, 得到的风电出力曲线簇如图 4 所示.



(a) 第1类出力曲线包含65个样本 (b) 第2类出力曲线包含33个样本



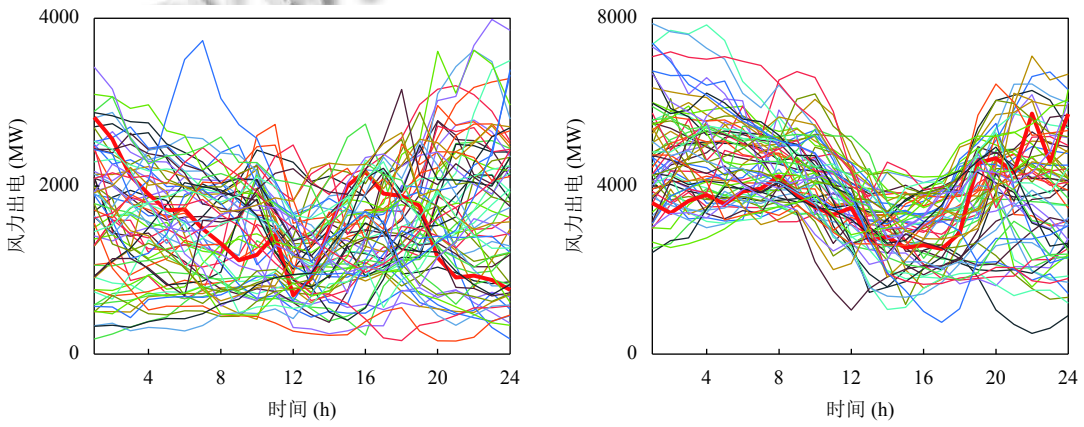
(c) 第3类出力曲线包含112个样本 (d) 第4类出力曲线包含60个样本

图 3　4 类风电出力曲线簇 (GMM)

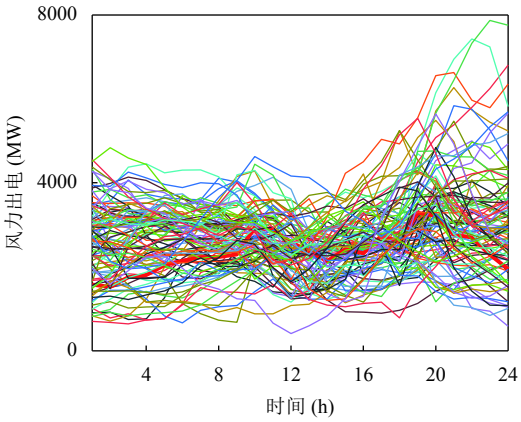
红色曲线为聚类中心, 代表该地区风电风力的典型场景. 如图 4 所示, 类内包含多种形态的出力曲线, 很多曲线形态与聚类中心曲线形态不一致, 仅能反映出风电出力的幅度大小.

为进一步比较这两种聚类方法, 分别提取其聚类中心曲线.

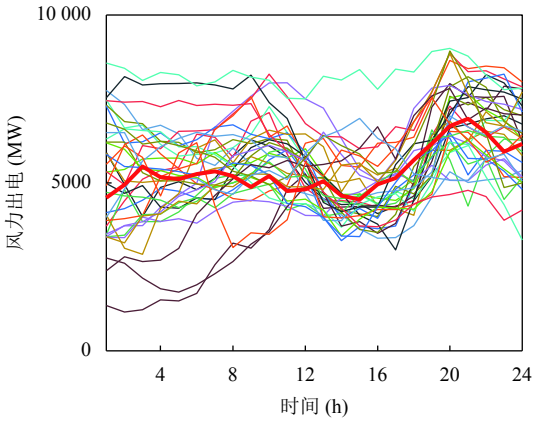
在图 5 中, 从峰谷差分布范围来看, 基于 K-means 算法风电功率峰谷差分布范围集中在 1700–3300 MW 之间, 不能反映出风电峰谷差特点, 对调度安排实用价值较小. 基于 GMM 聚类算法风电功率分布范围从 2400–4600 MW 之间, 较能反应该地区风电峰谷差波动范围.



(a) 第1类出力曲线包含67个样本 (b) 第2类出力曲线包含66个样本

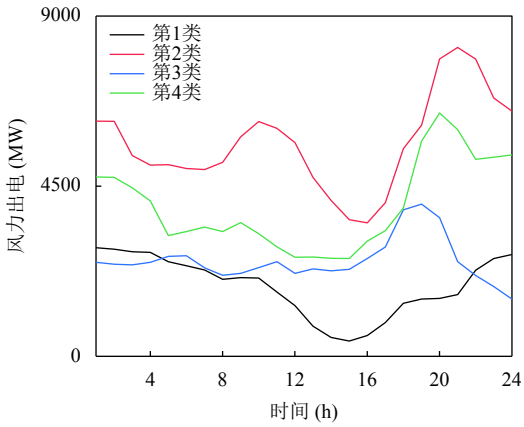


(c) 第3类出力曲线包含101个样本

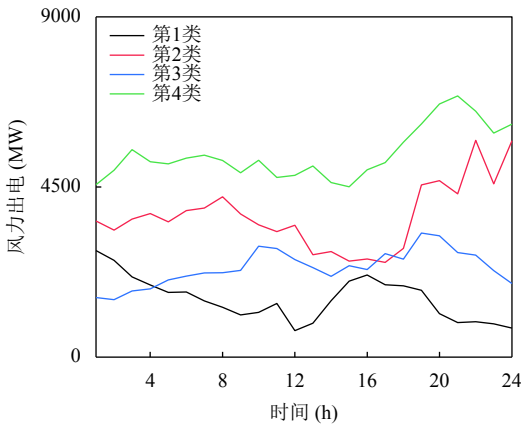


(d) 第4类出力曲线包含36个样本

图 4　4 类风电出力曲线簇 (K-means)



(a) 本文方法聚类中心曲线



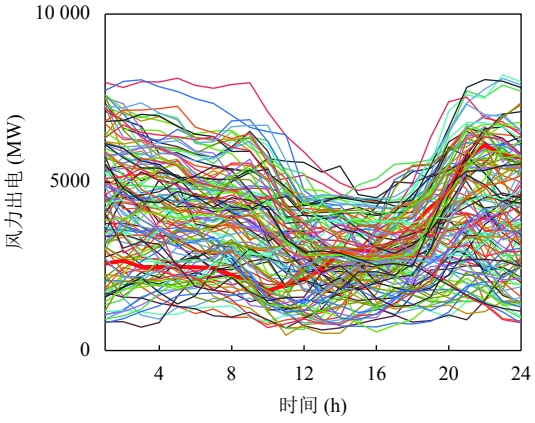
(b) K-means 方法聚类中心曲线

图 5　GMM 与 K-means 聚类中心曲线对比

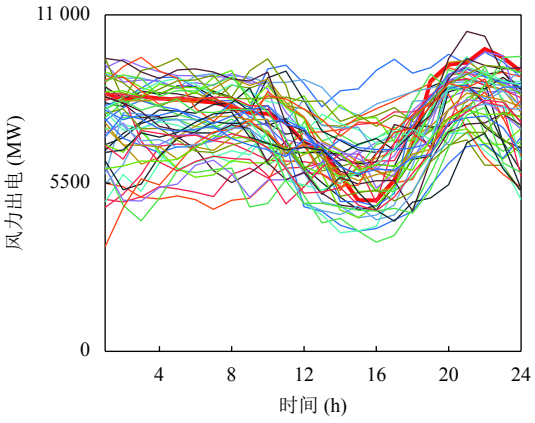
从功率波动范围来看, K-means 波动范围较小, 不能反映某些情况下风电的大范围波动特点、多峰谷特点 (如 GMM 第 2 类出力) 以及正反调峰特点 (K-means 风电波动特征选取较差).

3　其余季节风电出力场景

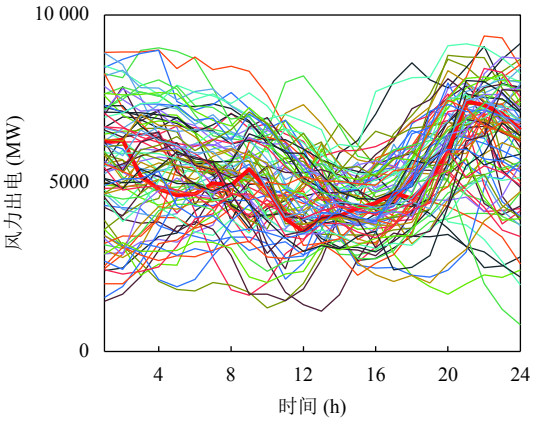
对于夏、秋、冬季节, 使用 *BIC* 对高斯混合模型进行选择, 得到最佳聚类数目应为 3. 提取这 3 个季节的风电出力场景, 图 6 到图 8 分别为夏、秋、冬季的场景曲线簇, 其中, 红色曲线代表该季节风电出力的典型场景.



(a) 第1类出力曲线包含134个样本

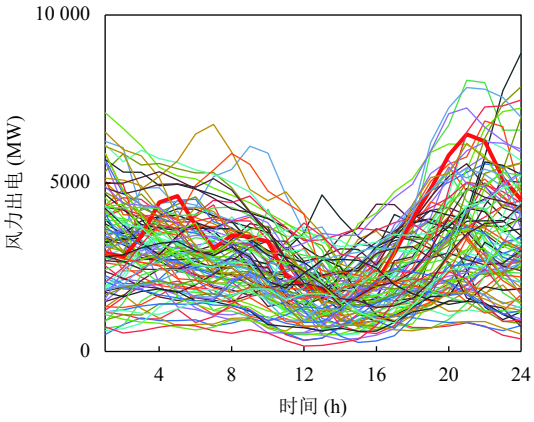


(b) 第2类出力曲线包含52个样本



(c) 第3类出力曲线包含87个样本

图 6 夏季地区风电出力场景曲线簇

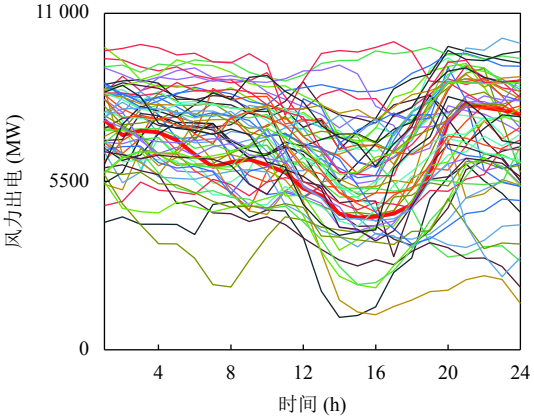


(c) 第3类出力曲线包含104个样本

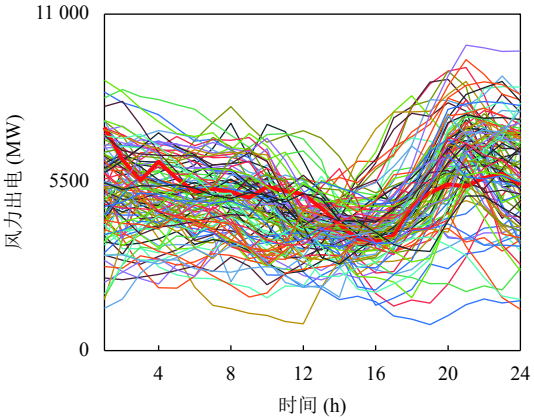
图 7 秋季地区风电出力场景曲线簇

由图 9 可知, 该地区夏秋季节风电出力功率较大、功率波动范围分布变化较小、功率波动范围较大, 夏季上半日相比秋季风电波动较小; 冬季风电波动与春季相似, 但呈现多峰谷特点更加明显.

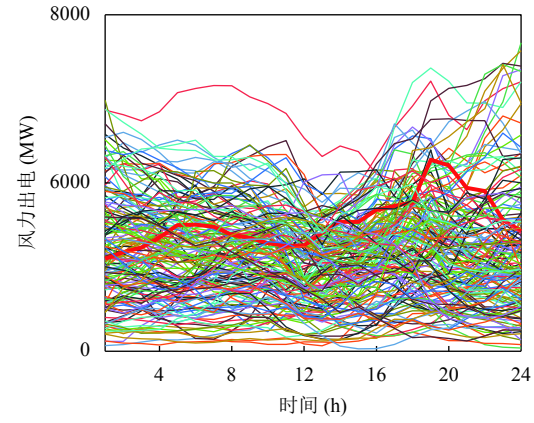
在调度计划中, 夏秋季节应安排调峰能力较强机组和其他调峰资源, 应对风电功率大范围波动, 且秋季上半日应多安排爬坡性能较高机组或灵活性调节资源, 应对风电功率频繁波动. 针对春冬季节风电多峰谷特性对峰谷电价合理优化, 通过负荷参与电网调度减少风电峰谷差.



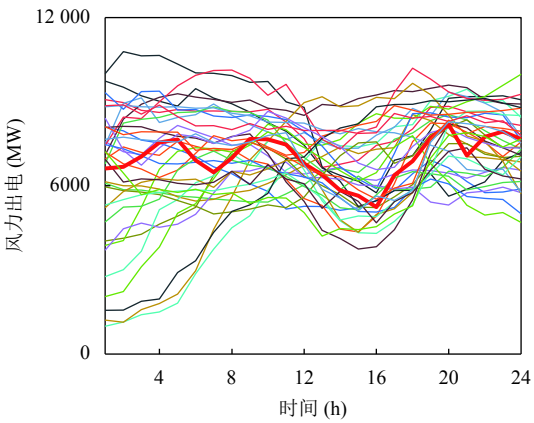
(a) 第1类出力曲线包含61个样本



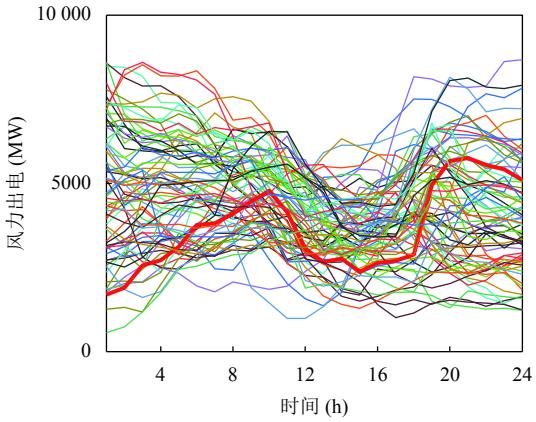
(b) 第2类出力曲线包含111个样本



(a) 第1类出力曲线包含164个样本



(b) 第2类出力曲线包含36个样本

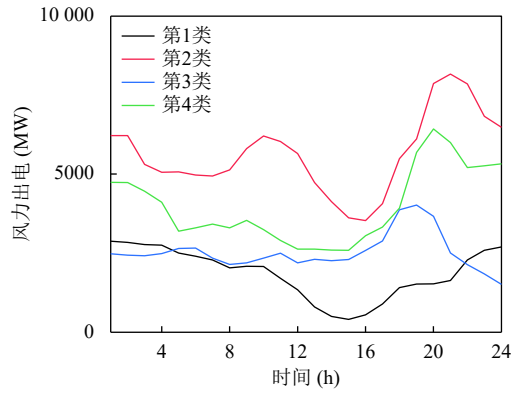


(c) 第3类出力曲线包含76个样本

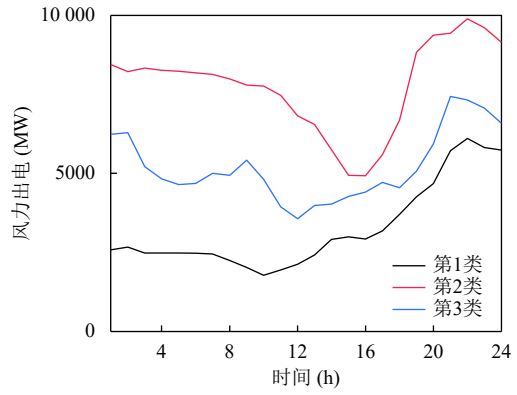
图 8 冬季地区风电出力场景曲线簇

4 结论

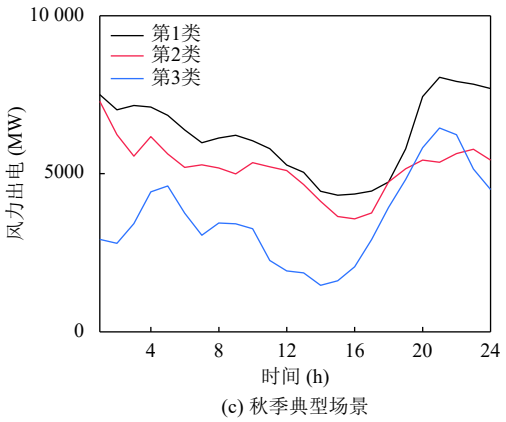
随着清洁能源在社会发展中扮演越来越重要的角色, 风能资源的利用也逐渐增多. 本文针对风电出力场景进行研究, 提出的高斯混合聚类模型, 能够提取典型风电出力场景, 并与 K-means 聚类方法对比, 该文提取的方法更能得到同类形态相近的曲线, 反映出风电功率变化的特征, 例如风电的正反调峰特性和波动特性, 对电网的调度具有重要意义.



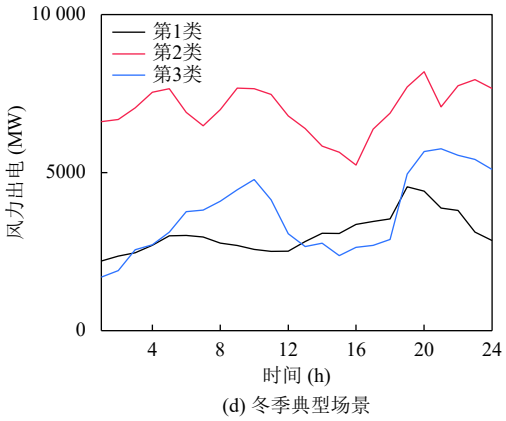
(a) 春季典型场景



(b) 夏季典型场景



(c) 秋季典型场景



(d) 冬季典型场景

图 9 四季曲线簇的典型场景

参考文献

1 潘珍华. 基于聚类算法的地区风电功率典型场景选取方法 [硕士学位论文]. 北京: 华北电力大学 (北京), 2018.

2 Wang T, Bi TS, Wang HF, *et al*. Decision tree based online stability assessment scheme for power systems with renewable generations. CSEE Journal of Power and Energy Systems, 2015, 1(2): 53–61. [doi: 10.17775/CSEEJPES.2015.00019]

3 孙吉贵, 刘杰, 赵连宇. 聚类算法研究. 软件学报, 2008, 19(1): 48–61.

4 廖攀峰, 齐军, 孙绥, 等. 基于改进 k-means 聚类的风电功率典型场景在日前调度中的应用. 电工材料, 2020, (1): 46–52.

5 熊强, 陈维荣, 张雪霞, 等. 考虑多风电场相关性的场景概率潮流计算. 电网技术, 2015, 39(8): 2154–2159.

6 邱宜彬, 欧阳誉波, 李奇, 等. 考虑多风电场相关性的场景概率潮流计算及无功优化. 电力系统保护与控制, 2017, 45(2): 61–68. [doi: 10.7667/PSPC160100]

7 王群, 董文略, 杨莉. 基于 Wasserstein 距离和改进 K-medoids 聚类的风电/光伏经典场景集生成算法. 中国电机工程学报, 2015, 35(11): 2654–2661.

8 王洪涛, 刘旭, 陈之栩, 等. 低碳背景下基于改进通用生成函数法的随机生产模拟. 电网技术, 2013, 37(3): 597–603.
9 姚剑峰, 凌静, 曲立楠, 等. 基于改进 FCM 聚类算法的清洁能源典型场景构建. 电网与清洁能源, 2019, 35(4): 76–82. [doi: 10.3969/j.issn.1674-3814.2019.04.013]
10 林俐, 肖舒, 费宏运, 等. 基于曲线形态特征的地区规模化风电出力场景划分. 电网与清洁能源, 2020, 36(3): 74–81, 88. [doi: 10.3969/j.issn.1674-3814.2020.03.012]
11 Li KH, Ma ZJ, Robinson D, *et al*. Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering. Applied Energy, 2018, 231: 331–342. [doi: 10.1016/j.apenergy.2018.09.050]
12 黄艳国, 韩亮, 张硕, 等. 基于高斯混合聚类模型的公交出行特征分析. 现代电子技术, 2019, 42(16): 174–178.
13 Ahani A, Nadoushani SSM, Moridi A. Regionalization of watersheds by finite mixture models. Journal of Hydrology, 2020, 583: 124620. [doi: 10.1016/j.jhydrol.2020.124620]
14 赵铭, 金大权, 张艳, 等. 基于 EM 和 GMM 的朴素贝叶斯岩性识别. 计算机系统应用, 2019, 28(6): 38–44. [doi: 10.15888/j.cnki.csa.006948]
15 高文曦, 于凤芹. 对 MFCC 进行 GMM 聚类的汉语数字识别方法. 计算机系统应用, 2011, 20(11): 167–170. [doi: 10.3969/j.issn.1003-3254.2011.11.041]
16 柳天虹. 工业大数据时间序列预测方法研究及应用 [博士学位论文]. 南京: 东南大学, 2018.
17 黄咏宁. 基于混合高斯模型的面板数据聚类研究 [硕士学位论文]. 广州: 华南理工大学, 2016.
18 张美霞, 李丽, 杨秀, 等. 基于高斯混合模型聚类和多维尺度分析的负荷分类方法. 电网技术, 2020, 44(11): 4283–4293.