基于AMI大数据和自组织映射网络的电力市场用户特征研究

胡 鑫[1],周 睿[1]，黄靖茵[1]，包 俊[2]，杨宇忠[2]，谭 讯[2]

（1.广东电力交易中心有限责任公司，广东 广州 510600；2.北京清大科越股份有限公司，北京 100084）

摘要：随着智能电网的发展和AMI系统的应用，电网数据量大大增加，海量电网大数据一方面有助于电网运行、调度策略的制定，一方面也大大增加了数据处理的难度。为此采用自组织映射神经网络（Self-Organizing Map，SOM）模型，基于实际电力用户的消费数据进行了用户消费行为模式分析和识别，同时利用主成分（PCA）模型结合k均值聚类的方法与SOM模型进行对比。结果表明SOM模型在数据降维、用户行为模式识别方面具有一定优势，实际算例结果表明SOM模型效果优于PCA模型，研究结果可为电力大数据处理方法和用户行为模式识别提供参考。

关键词：智能电网；神经网络；数据降维；大数据；模式识别

Research on user behavior pattern recognition of smart grid big data dimensionality reduction based on AMI and SOM

HU Xin[1], ZHOU Rui[1], HUANG Jingyin[1], BAO Jun[2], YANG Yuzhong[2], TAN Xun[2]

(1. Guangdong Power Exchange Center Co,. Ltd, Guangzhou 510600, China; 2. Beijing QU Creative Technology Co., Ltd., Beijing 100084, China)

Abstract: With the development of smart grid and the application of AMI system, the amount of power grid data increases greatly. On the one hand, massive power grid big data is conducive to the formulation of power grid operation and dispatching strategy, on the other hand, it also greatly increases the difficulty of data processing. In this paper, the self-organizing map model is used to analyze and identify the consumer behavior patterns based on the actual consumption data of power users. At the same time, PCA model combined with k-means clustering method is used to compare with SOM model. The results show that SOM model has some advantages in data dimensionality reduction and user behavior pattern recognition. The results of practical examples show that SOM model is better than PCA model. The research results can provide reference for power big data processing methods and user behavior pattern recognition.

Key words: smart grid; neural network; data dimensionality reduction; big data, pattern recognition

0 引言

近年来随着电力系统和电力市场的发展，电力系统已经从传统的垂直集成和集中结构过渡到了竞争型状态结构。这种结构条件下，电网中所有参与者（电厂、用电户和输配电企业等）都处于有效竞争状态。实现这种电

广东电力交易中心有限责任公司科技项目资助（GDKJXM20185842）

网构架需要在电力生产和消费之间交换足够的数据和信息，为此电力部门引入了自动计量读数（AMR）来实现电力数据的计量和监测，此后又引入了AMI来提高电网通信基础结构，以解决AMR在数据通信方面的不足。利用AMI可以汇总电网用户、电厂和输配电企业在电力市场需求、运营和计划等方面的信息，已经逐渐成为智能电网中一项不可或缺的技术。

可见AMI的重点功能是实现电网通信需求和数据交换，随着AMI的应用，电网中的数据交换量和信息解析量都大大增加，虽然信息量的增加有利于电网制定更有效的运行决策，但是从原始数据中提取有效信息的过程也变得更加复杂。因此，管理电网中的大量数据，需要数据挖掘和模式识别等适当的算法来实现。

用电户参与电力市场后，电网运营商首先需要解决用电户行为特征识别的问题，精确地识别和表征用电户的消费行为模式，对于制定合理可靠的电网运行策略、保持电网稳定运行至关重要。为此研究人员采用了自组织映射网络（SOM）来分析电网数据，也有研究人员利用SOM检测异常电网数据中的异常数据。在电网用户行为表征方面，文献[1]使用数据挖掘技术分析用电户的消费概况，文献[2]提出了一个用电户电力消费的分析方法，但有关电力市场大数据管理和数据降维仍然是分析用电户消费模式的难题。研究表明用电户的消费特征可用于分析用户行为模式，如电费、负载预测等，通常分析和研究用户消费行为模式，都是基于每小时消费数据进行分析的。

为此，针对电力大数据背景下数据分析、用户行为模式识别的需求，本文采用SOM模型分析和研究电力用户的行为模式，且采用PCA模式结合k均值聚类的方法与SOM模型进行对比，研究结果可为电力大数据分析和处理提供参考。

1 数学模型

1.1 自组织（特征）映射模型

自组织映射是一种通过竞争学习算法形成竞争层的神经网络模型。与传统竞争方法相反，该算法中不仅获胜者神经元的权重会在出现新输入值后更新，而且相邻神经元的权重也会随之更新，更新范围根据相邻神经元与获胜者神经元的相对距离来决定。通过这种模式，可用神经元覆盖几乎所有的输入空间。因此可以灵活地用各种形状的覆盖方式来覆盖输入空间，对于输入空间中没有样本分布的情况，这一覆盖方式很有效。

实际运用中，通常是通过SOM将N维输入数据集投影到输出空间，由此将数据集的维度降低到二维，如图1所示，其中输入空间称为样本空间，输出空间称为特征空间。输入空间由具有多个维度的初始样本组成，假设有I个样本，每个样本都有N个维度，表明输入空间的维度为N。在神经网络的竞争层中，每个神经元具有与样本相同的维度，通过将每个样本输入该网络，即可开始网络的训练。假设获胜者的神经元单位为w_j，与输入样本的最小距离为x_j，由此可得

$$J^{j,\mathrm{win}} = \arg\min \| X_i - W_j \|$$
$$\| X - W \| = \sqrt{\sum_{n=1}^{N} (x_n - w_n)^2} \tag{1}$$

式中，每个样本都有N个维度，式（1）计算所得的结果是获胜者单位，获胜者单位对周边单位的影响函数则可表示为

$$m_{ij}(t) = \exp(-\frac{1}{2} \times \frac{d_{ij}}{\sigma(t)^2}) \tag{2}$$

$$\sigma(t) = \sigma_0 \exp(-\frac{t}{\lambda}) \tag{3}$$

式中，$m_{ij}(t)$为影响函数；$\sigma(t)$为标准偏差；d_{ij}为距离函数。

网络的学习法则通过Kohonen公式实现，即

$$w_j^{\mathrm{new}} = w_j^{\mathrm{old}} + \eta m_{ij}(t)(x - w_j^{\mathrm{old}}) \tag{4}$$

式（4）实际上反映了学习速率为η时，神经元权重的更新过程，通过这一模型就可以建立起SOM神经网络。

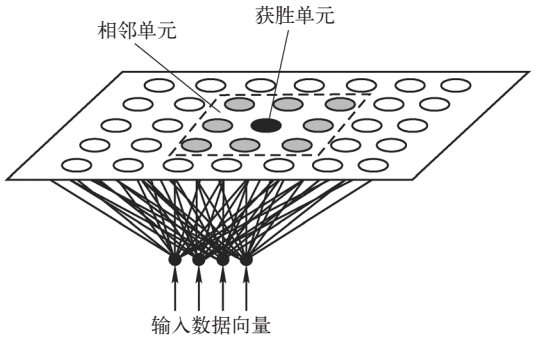

图1 SOM模型数据降维示意图

Fig. 1 Diagram of SOM model data dimensionality reduction

1.2 主成分（PCA）模型

信号处理中的一种主流方法是数据降维，其本质是将初始N维数据集投影到数据维度更小的空间中。该数据空间从初始空间构建而成，数据维度为K维（$K<N$），在降低数据维度过程中应保留初始空间的实质特征。因此按如下步骤进行数据降维：

1）准备初始数据集。

2）计算每个数据的差异和数据集的平均值。

3）计算协方差矩阵。

4）计算协方差矩阵的特征值和特征向量。

5）确定数据组中的主要成分。

PCA算法中有k个正交向量，这些向量中的每一个都是数据集的主要成分，并且这些主成分按照重要性的不同将以降序方式排列。图2为二维数据集(X_1，X_2)及其两个主要成分(Y_1，Y_2)的示意图。

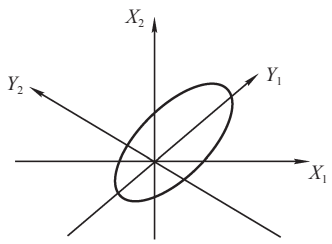

图2 PCA数据向量降维示意图

Fig. 2 PCA data vector dimensionality reduction diagram

2 计算方法

SOM模型计算流程如图3所示。对初始数据集进行预处理之后，首先将获取的实际用电负载曲线利用降维算法进行降维。本文在此处采用了PCA和SOM两个单独的算法以并行方式实现降维。其中SOM具有降维和聚类的内在功能，但与之相比，PCA仅具有降维的功能，因此在利用PCA对数据进行降维之后，再使用k均值算法对数据进行聚类。

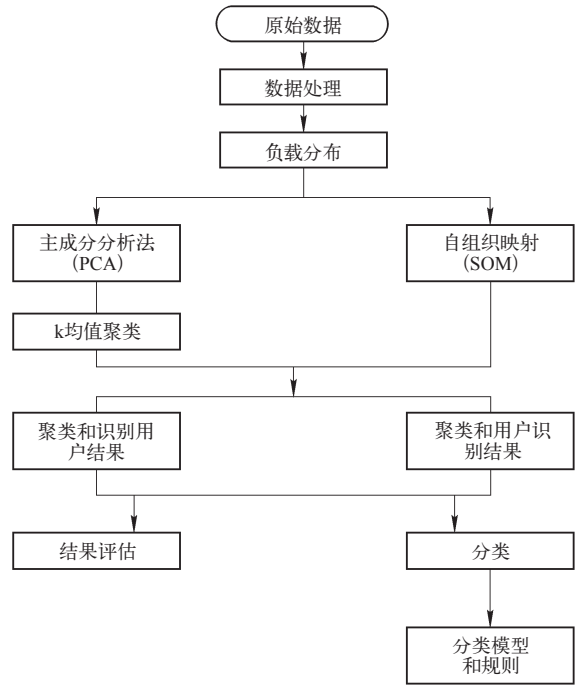

图3 SOM模型计算流程示意图

Fig. 3 Calculation flow diagram of SOM model

本文采用某省电网实际用电数据进行分析。该数据集包含1 000户家庭在一年时间内，每30 min记录一次的消费数据，如图4所示。因此每24 h内的每个样本（家庭）有48个消费数据。

2.1 基于SOM的大数据和消费行为模式识别

在模式识别中，聚类起着相当重要的作用，因此本文采用聚类算法，根据用电户消费行为的相似性对单个集群中的用户进行分类。从消费行为的角度来看，聚类后每个集群的用电户彼此最相似，并且与其他集群的用电户也具有最大的相似性，同时认为每个顾客都是具有特征向量的一个样本，其特征向量是每30 min

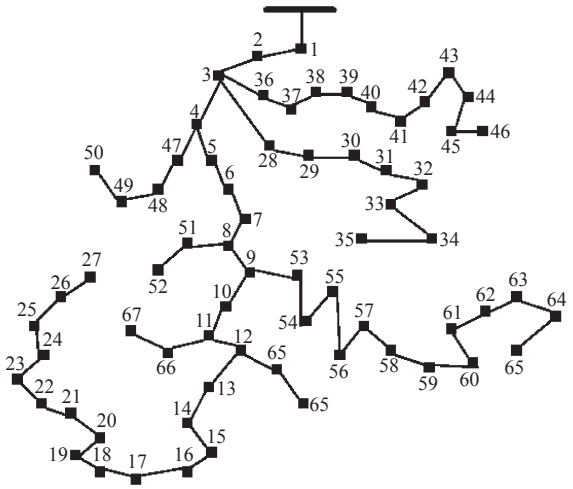

图4 算例采用的电网结构示意图

Fig. 4 Schematic diagram of power grid structure used in the calculation example

间隔一次的用电量记录。根据上述描述，可知本文采用的初始数据集是1 000个用电户一年时间内的消费数据记录，这些用电户中的每个样本都有48个特征向量，通过这些特征向量可以进行相似性分析。使用SOM模型进行分析时，输出值是描绘神经元在网络中具有预定义大小位置的映射图。通过比较每个聚类操作中竞争层的映射图就能够识别样本消费行为的变化。SOM算法与其他聚类算法（如k均值）相比，优势在于SOM算法能够自动确定聚类中心并检测用电户的消费行为变化。

除了k均值以外，在SOM模型中每个新样本输入后都需要对映射图进行再次训练，并重新调整映射图的布局。图5为所有样本经过6次聚类后神经元之间间距的示意图。由图5可以清晰地看出用电户消费行为的变化，图中深色区域表示神经元之间的距离较长，而浅色表示神经元之间间距较短，可知相互之间具有联系性的神经元用黄色表示，随着相互联系的神经元之间间距的增加颜色加深。这种颜色分布可以反映SOM模型对数据进行聚类的作用，实际上这是SOM的固有属性，它可以自动确定聚类的数量，并以SOM分布图的形式提供特征空间中的样本分布。

通过将样本数据集从N维空间投影到二维

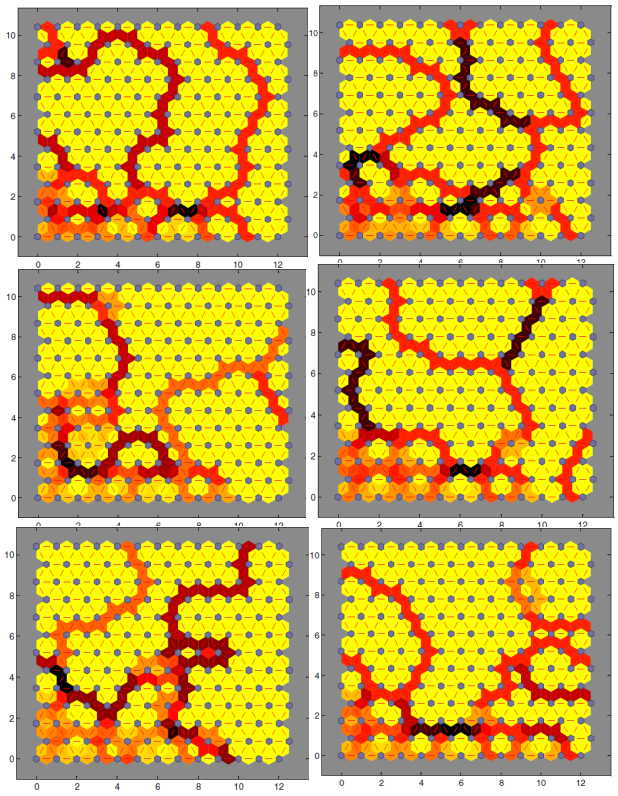

图5 SOM映射图中神经元距离变化示意图

Fig. 5 Schematic diagram of neuron distance change in SOM map

空间，使数据分析变得更加容易，尽管采用降维技术丢弃了部分数据特征，但是这一简化具有更多的优点。随着AMI的应用，在进行大量数据管理时不可避免地要使用大数据管理算法，而数据降维可以有效提高数据分析效率及降低数据分析难度，从本文采用的用电户数据来看，SOM算法能够有效反映用电户的消费行为。

2.2 模型评价指标

文献[3]中引入了两个指标，以评估聚类算法的性能，本文通过应用SOM和PCA两套算法得到了两组单独的输出数据。通过计算两组输出数据的性能指标，就可以对比PCA和SOM的性能。

假设用电户共有N户，每一户的用电限量由一个数据组构成$d=\{d_z,z=1,\cdots,Z\}$，其中Z为数据组的长度，假设数组$D=\{d^{(n)},n=1,\cdots,N\}$是$N$个用电户各自的用电数据，数据的分析结果为$C=\{c^{(n)},n=1,\cdots,J\}$，采用SOM和PCA两种方法对数据组进行分析，按如下公式就可以

对两种模型的输出结果进行评估，即

$$\mathrm{dis}(c^{(j)},D^{(j)})=\sqrt{\frac{1}{k^{(j)}}\sum_{n=1}^{k^{(j)}}\mathrm{dis}^2(c^{(j)},d^{(n)})} \tag{5}$$

$$\mathrm{MIA}=\sqrt{\frac{1}{J}\sum_{j=1}^{J}\mathrm{dis}^2(c^{(j)},d^{(j)})} \tag{6}$$

$$\mathrm{CDI}=\frac{\sqrt{\frac{1}{j}\sum_{j=1}^{j}\frac{1}{2k^{(j)}}\sum_{k=1}^{k^{(j)}}\mathrm{dis}^2(d^{(k)},D^{(j)})}}{\sqrt{\frac{1}{2J}\sum_{j=1}^{j}\mathrm{dis}^2(d^{(j)},C)}} \tag{7}$$

式中，CDI和MIA两个指标可以衡量聚类的准确性。

如前文所述，如果每个聚类的样本彼此之间最相似，并且与其他聚类的样本也最相似，则适合进行聚类。根据MIA和CDI两个参数的计算公式，可知如果MIA和CDI较小，则数据组可以有效地进行聚类。本文采用SOM和PCA两种算法对比研究了二者的计算效果，得到的CDI和MIA指标如图6、图7所示。可见与PCA相比，SOM模型的两个指标值都较小，表明SOM在降维和聚类方面的性能更好。

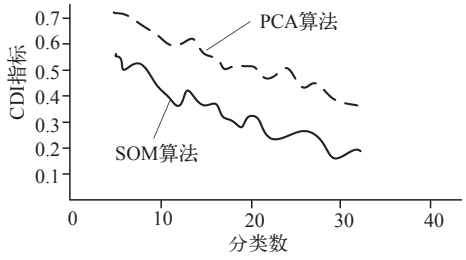

图6　SOM、PCA算法的CDI指标

Fig. 6 CDI index of SOM and PCA algorithm

3　结束语

本文研究了应用自组织映射模型分析用电户消费行为模式中的性能。结果表明SOM模式通过其竞争层神经网络能较好反映出用电户消费行为的变化。样本数据的变化导致神经元之间距离的变化，也有助于揭示用电户在消费行为方面的变化。

同时，SOM模型兼具数据降维和数据分类识别的功能，本文通过引入CDI、MIA两个评估指标，对比分析了SOM模型和PCA模型对同一组数据的分析能力，结果表明相比于PCA模型，SOM模型在数据降维和聚类方面都具有更好的性能。

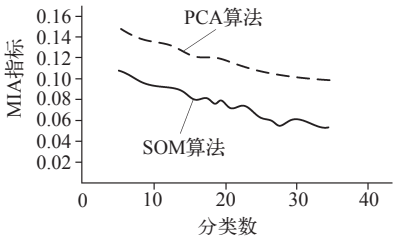

图7　SOM、PCA算法的MIA指标

Fig. 7 MIA index of SOM and PCA algorithm

参考文献

[1] KASKI Samuel，SINKKONEN Janne，PELTONEN Jaakko．Bankruptcy analysis with self-organizing maps in learning metrics[J]．IEEE Transactions on Neural Networks，2001，12(4)：936-947．

[2] 于惠鸣，张智晟，龚文杰，等．基于深度递归神经网络的电力系统短期负荷预测模型[J]．电力系统及其自动化学报，2019，31(1)：116-120．

[3] WANG Y F，DENG M H，BAO Y K，et al．Power system disaster-mitigating dispatch platform based on big data[C]．International Conference on Power System Technology，IEEE，2014．

作者简介

胡　鑫

胡鑫（1988－），女，硕士研究生，高级工程师，主要研究方向为信息安全。

周睿（1984－），男，硕士研究生，高级工程师，主要研究方向为电力调度和电力交易。

黄靖茵（1992－），女，本科，助理工程师，主要研究方向为电力交易和信息技术。

(收稿日期：2019－04－08)EA