文章编号:1000-582X(2007)08-0044-05

基于结构自适应神经网络用电量时间特征的聚类分析

袁忠军[1],陈 刚[2]
(1.广西水利电力职业技术学院 电力工程系,广西 南宁 530023;
2.重庆大学 电气工程学院 高电压与电工新技术教育部重点实验室,重庆 400030)

摘 要:鉴于聚类分析在数据挖掘中具有重要的作用,针对聚类分析中聚类数确定难的问题,深入研究了聚类准则的选择和曲线特性,提出了一种基于SOFM神经网络的结构自适应聚类神经网络,其特点是能够自动确定最佳的聚类数。基于实际营销数据,采用结构自适应聚类神经网络技术实现了用户用电量时间特征分析,所得结论对于电价的针对性的调整以及合理地安排电力生产具有重要的参考价值。

关键词:聚类分析;最优聚类数;人工神经网络;用电量时间特征
中图分类号:TM715 **文献标志码**:A

聚类(clustering)是数据挖掘领域最为重要和基本的技术之一,用于发现在数据库中未知的知识分类。从数据库知识发现的角度来讲,对聚类问题的研究是要从大量的数据集中智能地、自动地抽取出有价值的聚类知识。

人们先后提出了聚类分析领域的大量算法和方法,总结起来可以归纳为五大类[1-5]:划分方法、层次方法、密度方法、网格方法以及模型方法。模型方法由于具有设计简单、解决问题范围广且最终可以归结为优化问题等优点,成为聚类分析的主流方法[6-7]。但该方法采用数值迭代算法,在大数据量的聚类中的运算速度较慢。到了20世纪90年代初,随着神经网络(ANN)理论及其应用研究的重新兴起和快速发展,ANN以其分布式存储、并行协同处理以及自学习等特性被用于聚类分析领域。目前常用于聚类分析的ANN有Kohonen自组织神经网络、自组织共振神经网络(ART)以及竞争神经网络(CNN)等。对于各种ANN聚类算法来说,均需要事先给定聚类数或者相关的参数和假设条件,这些参数很难给出或无法恰当地指定,而这些参数对于聚类效果和性能具有重要的影响。其中聚类数确定问题,一直都是聚类分析中的一个难点。笔者将结合自组织特征神经网络和聚类准则研究一种能自动确定聚类数的自适应神经网络,并将其应用在电力营销的用户用电特征分析中,以支持电力营销的辅助决策。

1 自组织特征神经网络的聚类原理

1981年Kohonen提出了自组织特征映射SOFM(Self-Organizing feature Map)的模型。图1给出了由输入层和竞争层组成的无隐层SOFM网络模型示意图,该网络利用自组织特点,将n个一维输入模式,通过欧氏距离函数进行相似性判断,映射到二维神经元阵列上,以实现信息的特征提取或称聚类,这种自组织聚类过程是系统自主、无导师指导的过程下完成的。

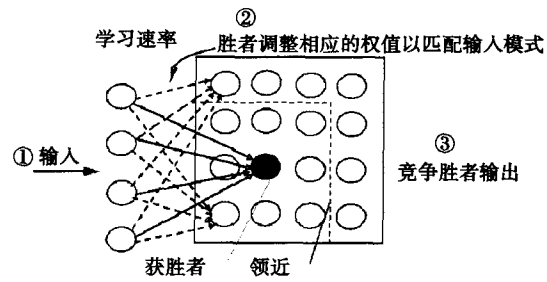

图1 Kohonen自组织特征图

收稿日期:2007-05-08
基金项目:重庆市科委自然科学基金资助项目(CSTC,2006BB6209)
作者简介:袁忠军(1968-),女,广西水利电力职业技术学院硕士研究生,主要从事电力系统安全、经济运行、电力营销等研究,(E-mail)yunzjqq@163.com。

标准的 SOFM 网络的学习算法步骤如下：

1）为网络中每一个神经元的权重赋较小的随机值。

2）对于每个输出神经元 i，对应输入向量的神经元权重向量为 $w_{ij}(j=1,2,\cdots,n)$，分别计算欧氏距离（作为相似性判断依据）：

$$d_i = \sqrt{\sum_{j=1}^{n}(x_j - w_{ij})^2}(i=1,2,\cdots,k)。\quad (1)$$

3）$d_c = \min\{d_1, d_2\cdots, d_n\}$，则以 c 所对应的神经元为竞争得胜神经元。

4）对以 c 为中心，$Nc(t)$ 范围内的神经元，按照式(2)调节权重，在 Nc 范围外的神经元权重不调整：

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(t)(x_j - w_{ij}(t)),\quad (2)$$

其中，$\alpha(t)$ 是学习增益系数。

5）递减邻域半径 $Nc(t)$ 以及学习率 $\alpha(t)$

$$\alpha(t+1) = \alpha(t)(1 - t/T),\quad (3)$$

$$N_c(t+1) = \text{int}[N_c(t)(1 - t/T)]。\quad (4)$$

6）令 $t=t+1$，如 $t<T$，取下一个输入向量返回执行步骤 2；否则学习结束。

SOFM 网络训练完后，可以直接在输出神经元获得聚类结果。另外还可采用距离图法对输出结果进行合并处理，以获得相对较优的聚类结果。但是，该方法的前提是输出神经网络个数要大于实际聚类数，如果小于又该如何呢？如何确定最佳的聚类数？

从前面分析可以看出，SOFM 网络采用输出神经元的权值向量作为聚类中心，通过学习算法不断调整权值向量靠近输入向量的聚类中心，这与统计学的k－均值聚类算法基于距离比较的原理是一致的。因此针对聚类数的确定问题，须引入聚类准则来进行分析和研究。

2　聚类准则的研究

2.1　聚类准则的设计

假设将 n 个样本 $x_j \in \mathscr{R}^n(j=1,2,\cdots,n)$ 分成 k 类 $\{C_1, C_2, \cdots, C_K\}$，对 $i=1,2,\cdots,k$ 和 $j=1,2,\cdots,n$，定义：

$$\mu_{ij} = \begin{cases} 1 & if \quad x_j \in C_i \\ 0 & if \quad x_j \notin C_i \end{cases},\quad (5)$$

设 $m_i \in \mathscr{R}^n$ 表示第 i 类的中心，n_i 表示第 i 类所包含的样本个数，则：

$$m_i = \frac{\sum_{j=1}^{n}\mu_{ij}x_j}{\sum_{j=1}^{n}\mu_{ij}} = \frac{1}{n_i}\sum_{j=1}^{n}\mu_{ij}x_j,\quad (6)$$

整体类内差异为：

$$S(u) = \sum_{i=1}^{k}\sum_{j=1}^{n}\mu_{ij}\|x_j - m_i\|^2。\quad (7)$$

式(7)中的 $\|\ \|$ 表示欧氏范数，$S(u)$ 就是经典的类内平方误差和（WGSS，within－group sum of squared error）准则函数[3]。k－均值算法的目的是针对一个预先给定的聚类数 K 寻找最优的 μ^* 使得 $S(u)$ 取得极小值。

对于聚类数 K 未知的情况下，如 $K=n$ 时上式的聚类则取得最小值 0，因此必须修改。有的文献将聚类准则改为“类内差异最小，而类间差异最大”，也即求解下面的优化问题：

$$\begin{cases} \min\limits_{k} LN = \sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n}\mu_{ij}\|x_j - m_i\|^2; \\ \max\limits_{k} LJ = \sum\limits_{i=1}^{k}\sum\limits_{j<i}\|m_i - m_j\|^2。 \end{cases}\quad (8)$$

进一步，可以将式(8)中的第 2 个式子改为用类间最小连接距离代替，同时将该多目标的优化问题用转化为一个单目标的优化问题，即：

$$\max_{k,\mu} JZ = \max_{k,\mu}\left(\frac{\min\limits_{1\leqslant j<i\leqslant k}\|m_i - m_j\|^2}{\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n}\mu_{ij}\|x_j - m_i\|^2}\right)。\quad (9)$$

2.2　聚类准则的实验分析

观察类间距离 LJ 的变化趋势可发现：当聚类数为最大 n 时，LJ 较小；根据聚类思想，不同类之间的相似性尽量小，随着聚类数的减少，LJ 应该将逐渐增大。在文献[8]中证明：LJ 是聚类数 K 的严格单调递减函数，LN 为 K 的严格单调递减函数。

笔者在 Matlab 软件平台进行了聚类准则的曲线特性的实验。针对二维平面上正态分布的数据集，以不同的中心点产生符合高斯分布随机点，即分别取中心点为 4 个和 9 个，每个中心点附近分别用 2 种方差值设置 100 个点，组合成 4 种情况，然后通过 SOFM 神经网络计算聚类数从 2 到 20 的 LJ、LN、LJ/LN 的值以及变化情况。

实验结果说明：1）总体上 LJ 随聚类数 K 的增大而减小，但并非严格的单调递减函数；2）LN 并不是 K 的严格递减函数，随 K 的增加，整体趋势是下降的，但存在若干局部极小点。在这些极小点中包含了最优的聚类数点，而且往往在最优聚类数点的 LN 为全局极小。这种情况的出现，可以这样理解：

① 整体曲线趋势的下降。随着聚类数 K 的增加，分中心点逐渐增多，则样本点到分中心点之间的距离和 LN 将会减小，当聚类数达到全部样本数 n 时，每个样本单独成类，其 $LN=0$，但此时的 LJ 也达到极小。这是极端的情形，没有任何实际应用价值。

② 多个极小点。由于样本数据本身的分类聚集特性,当聚类数符合样本数据的实际分类特性时,此时必将出现 LN 较小的情形,这就是局部的极小点。由于聚类多样性,这种相对较小的聚类数并不是只有一个,但在最佳聚类数时 LN 最小。

因此,LJ/LN 并非为单峰函数,而是一个多峰函数,在最优聚类数处的峰值是全局最大极值。这正好符合“类间距离大,类内距离小”的聚类思想。根据这个思想,就可以确定最优的聚类数。

3 结构自适应的聚类神经网络模型

多峰函数的优化问题可以采用多种优化算法求得。如遗传算法具有全局搜索和性能稳定等优点,但由于聚类数是离散的且变化范围不是很大,采用遗传算法的计算工作量反而较大。因此,针对这个特殊的聚类数的优化问题,可直接采用枚举法实现,算法的具体步骤要点如下:

1)根据问题领域的实际情况,确定聚类数的大致范围,一般情况下最大聚类数不超过样本总数的1/2。

2)设置最大的训练次数以及权值差收敛阈值。权值的收敛阈值应考虑较小,为最大范围的1%左右。

3)从最大聚类数开始,聚类数依次递减,用SOFM网络分别训练样本数据进行聚类分析,直到某聚类数对应的神经网络收敛为止。

4)从收敛的聚类数的神经网络开始,按式(10)合并最小距离的神经元,实现聚类数的快速递减,并且分别计算 LJ/LN,直到最小聚类数为止。

$$W_h = \frac{1}{N_p + N_q}(N_p W_p + N_q W_q). \qquad (10)$$

式(10)中,2个神经元 p 和 q 对应的权值矢量分别为 W_p 和 W_q,聚类所含样本个数分别为 N_p 和 N_q。

5)比较各个 LJ/LN 的值,其中最大值对应的 K 即为最优的聚类数。

6)用最优聚类数作为输出神经元重新构建SOFM神经网络,采用较小的权值阈值(1%左右)和较大的最大训练次数重新训练网络,输出正确的聚类结果。

从上面算法步骤可以看出:聚类数及神经网络结构(主要是输出神经元个数)事先勿需确定,要用训练数据反复训练神经网络之后,最后通过聚类准则来确定最佳聚类数的大小和聚类神经网络的结构,因此可称之为结构自适应的聚类神经网络。

4 用户用电量时间特征的聚类分析

用户用电量时间特征的聚类分析即:按地区和行业等类别分析其分时电量比例结构以及分段电量的比例构成情况,并进行时间特征的聚类分析,这对于供电部门了解本地区各个行业的用电时间特征以及采取何种措施增供促销等方面都具有重要的参考价值。

算例对某地区1999年用电数据库的原始数据进行聚类挖掘分析[9-11],挖掘采用的编程工具为Matlab神经网络工具箱,数据库系统是SQLServer2000,数据分析采用Excel电子表格工具。

4.1 数据选择

原始用电数据存储在各个相互关联的表中,涉及的表及其关系如图2所示。

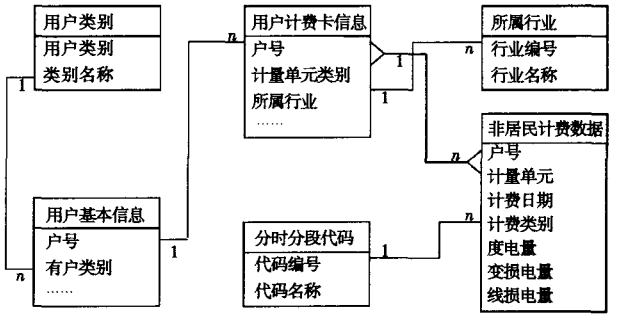

图2 电量电费 E-R 关系图

作为聚类分析,各个属性字段必须具有一定的可比性。需要了解不同行业在丰枯段以及峰谷段如何安排用电,各个时段的用电比例如何,各个行业的分段及分时用电的构成比例是否具有相似性,通过比例结构的相似性进行一定划分,这样就可以掌握不同行业用户在时间上(每年丰枯段以及每天的峰谷段)的用电趋势。针对实际问题,需要对涉及的原始表进行复杂的SQL统计查询以生成所需要的新表,然后考察各类电量在总电量的百分比情况,最后形成的字段包括:年、用户类别、所属行业、丰期峰(平、谷)段电量百分数、平期峰(平、谷)段电量百分数、枯期峰(平、谷)段电量百分数。

4.2 数据表示

用户类别分为9类(居民照明、非居民照明、……)。这种类型的变量取值不涉及排序问题,不存在关联关系,因此采用简单的映射成0~1.0间的数字即可。

行业类别按行业规定划分为39类(未含全部的行业类别,仅仅针对实际数据库中涉及的行业)。这种类型的变量取值同样不涉及排序问题,不存在关联关系,因此采用简单的映射成0~1.0间的数字即可。

其它字段均为连续的数字类型,范围为百分数0.00~100.00。这样,最后给出所选择的数据及其表示方法,见表1所示。

表 1　聚类分析的数据表示

序号	意义	类型	值范围	表示
1	用户类别	符号	共九大类	0,1,…,1.0
2	所属行业	符号	共 39 个行业	025,05,…,1.0
3	丰期峰段电量百分数	连续数字	0.00 ~ 100.00	0.0 ~ 1.0
4	丰期平段电量百分数	连续数字	0.00 ~ 100.00	0.0 ~ 1.0
5	丰期谷段电量百分数	连续数字	0.00 ~ 100.00	0.0 ~ 1.0
6	平期峰段电量百分数	连续数字	0.00 ~ 100.00	0.0 ~ 1.0
7	平期平段电量百分数	连续数字	0.00 ~ 100.00	0.0 ~ 1.0
8	平期谷段电量百分数	连续数字	0.00 ~ 100.00	0.0 ~ 1.0
9	枯期峰段电量百分数	连续数字	0.00 ~ 100.00	0.0 ~ 1.0
10	枯期平段电量百分数	连续数字	0.00 ~ 100.00	0.0 ~ 1.0
11	枯期谷段电量百分数	连续数字	0.00 ~ 100.00	0.0 ~ 1.0

4.3　最优聚类数的确定

综合某地区的 1999 年的用电记录,提取出共 87 个训练样本,样本编号为 1,2,…,87。设置聚类数的范围为 2 ~ 20,采用前述研究的结构自适应的聚类神经网络,计算结果如图 3 所示,可得出最优聚类数为 13。

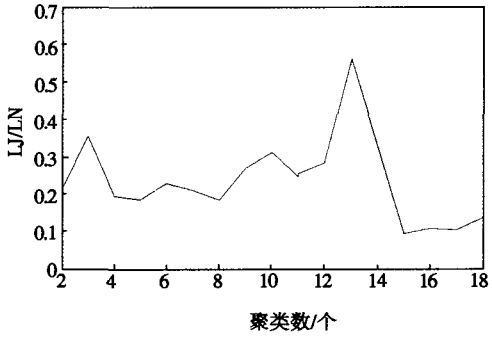

图 3　用电量时间特征聚类的类间距 *LJ/LN* 的变化曲线

4.4　结果及分析

聚类结果分析是数据挖掘中必不可少的重要环节。采用聚类数为 13 的聚类分析的结果如图 4 的聚类图所示。

为了更加清楚地说明聚类结果的意义,根据聚类图以及其它资料我们设计出如表 2 所示的聚类统计表格(部分),从中可详细察看聚类意义所在。

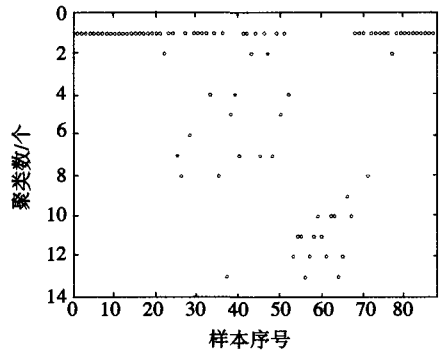

图 4　用电量时间特征的聚类图

表 2　用电量时间特征聚类分析结果表(部分)

分类	电量占用比例	占百分比 丰	平	枯	占百分比 谷	峰	平	行业说明	简要说明
1	6.72				100			涉及几乎各个行业	全部在平段用电
2	2.06				1	98	1	非居民城市居民用电、普非工业中的国内商业、其他教育事业、农村综合其他行业(金融、保险业)	几乎全集中在平段用电,比例超过 98%,峰段和谷段用电极少,在 2%左右
.....									
13	0.73	5	6	89	37	38	25	普通工业:锅炉制造业。大工业:水利发电业、金属结构制造业	几乎全在枯水用电量较大。峰平谷用电相当。

对于实行峰谷段分时电价政策的用户。原则上用户应积极进行错峰填谷,可减少企业的电费支出,又支持了供电企业的分时电价政策。可以按以下几类情况分别考察:谷段用电大于峰段用电的用户、峰段用电大于谷段用电的用户、主要用电量集中在平段的用户。

对于实行丰平枯分时电价政策的用户。在水电比重较大的电网,实行丰、枯季节电价。丰水的"弃水"期电价可比现行电价低 30% ~50%;枯水期电价可比现行电价高 30% ~50%。主要根据中国水电的季节发电的特征,每年根据水库来水情况分为 3 个季节,即丰水季节、平水季节和枯水季节。可考察以下 2 种情况:丰枯季节用电较大的用户、主要集中在平水季节用电的用户。

5　结　论

在 SOFM 神经网络的基础上,从聚类准则出发,通过实验对聚类准则的曲线特性进行了详细的分析和论证,设计出一种结构自适应的聚类神经网络,该网络能自动确定最佳的聚类数(输出神经元),并提出了一种

减少计算量的改进算法。最后应用该模型和算法对电力营销中的用电量时间特征进行了全过程的仿真计算分析,证明了该聚类神经网络的有效性,并对聚类结果进行了分析。用户用电量的特征聚类分析可以得出一些很有意义的结论,如分地区、行业、用电类别的用户的丰枯和峰谷的用电比例情况,这些结论对于电价的调整以及合理地安排发电计划等都具有参考价值。

参考文献:

[1] KARYPIS, E-H HAN, KUMAR V. Chameleon a hierarchical clustering algorithm using dynamic modeling[J]. IEEE Computer, Special issue on Data Analysis and Mining, 1999,32(8):68-75.

[2] GUHA R,RASTOGI, SHIM K. Rock: a robust clustering algorithm for categorical atributes[C]. In Proc. 1999 Int. Conf. Data Engineering (ICDE'99), Sydney, Australia, Mar. 1999: 512-521.

[3] ALSABTI K, RANKA S, SINGH V. An efficient K-means clustering algorithm[C]. Proc. the First Workshop on High Performance Data Mining, Orlando, Florida, 1998.

[4] AGRAWAL R, GEHRKE J, GUNOPULOS D,et al. Automatic subspace clustering of high dimensional data for data mining aplications[C]. In Proc. 1998 ACM-SIGMOD Int. Conf. Management ofData (SIGMOD'98), Seattle, WA, June 1998: 94-105.

[5] GOIL, SNAJAY, HARASHA NAGESH ,et al. MAFIC: efficient and scalable subspace clustering for very large data sets[C]. Technical Report Number CPDC-TR-9906-019, Center for Parallel and Distributed Computing, Northwestern University, 1999.

[6] 赵艳广. 数据挖掘中聚类算法研究与仿真[D]. 北京:北京邮电大学,2003.

[7] 湛燕, 杨芳, 王熙照. 基于遗传算法学习聚类算法的中心个数[J]. 计算机工程与应用, 2003,(16):86-87.

[8] 杨凌. 聚类分析中聚类数的确定问题[D]. 武汉:武汉科技大学,2001.

[9] 周波. 德阳电业局电力营销决策支持系统的设计与实现[D]. 重庆:重庆大学,2002.

[10] 陈刚, 王超, 周波. 电力营销决策支持系统的设计[J]. 电力需求侧管理, 2003, (4):26-29.

[11] 陈刚. 基于数据挖掘的电力营销决策支持系统的结构原理及算法研究[D]. 重庆:重庆大学,2004.

Clustering Analysis for Time Feature of User Power Consumption Based on Structural Self-adaptation ANN

YUAN Zhong-jun[1] , *CHEN Gang*[2]

(1. Guanxi Occupational Technology College of Water Conservancy and Electric Power, Guanxi 530023, China;
2. Key Laboratory of High Voltage Engineering and Electrical New Technology, Ministry of Education, Electrical Engineering College of Chongqing University, Chongqing 400030, China;)

Abstract: In view of the important effect of clustering analysis in data mining, the clustering rules and its curve are studied to solve the problem of determining clustering number. A kind of self-adaptation clustering ANN is presented based on SOFM ANN, which can automatically determine the clustering number. Based on practical sales data, the time feature analysis of power user consumption are carried out by using the self-adaptation clustering ANN, whose conclusion has the important referenced values for adjusting power price correspondingly and arranging power producing reasonably.

Key words: clustering analysis; optimum clustering number; artificial neural networks (ANN); time feature of user power consumption

(编辑 姚 飞)