调和聚类-分类方法在电力负荷预测中的应用

窦全胜[1),2),3)] 史忠植[1)] 姜 平[2)] 马君华[3)]

[1)](中国科学院计算技术研究所智能信息处理实验室 北京 100190)

[2)](山东工商学院计算机科学与技术学院 山东 烟台 264005)

[3)](烟台东方电子信息产业集团公司 山东 烟台 264001)

摘 要 分类和聚类是数据挖掘中两个重要的研究领域,分类需要相关的先验知识,而聚类往往依据某种相似性测度,从数据本身来寻找其内在特征.在电力系统负荷预测过程中,依靠先验知识得到的分类结果与聚类结果之间并不协调.针对这一问题,文中给出了调和矩阵的定义,并在此基础上,提出调和聚类-分类算法,将该方法应用于电力系统负荷预测的样本分类中,实际结果表明,通过文中方法得到的分类结果更加客观和科学,预测结果的可靠性得到了保证.

关键词 聚类;分类;负荷预测;调和矩阵

中图法分类号 TP391 **DOI 号**: 10.3724/SP.J.1016.2010.02644

Application of Associated Clustering and Classification Method in Electric Power Load Forecasting

DOU Quan-Sheng[1),2),3)] SHI Zhong-Zhi[1)] JIANG Ping[2)] MA Jun-Hua[3)]

[1)](*Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[2)](*School of Computer Science and Technology, Shandong Institute of Business and Technology, Yantai, Shandong* 264005)

[3)](*Yantai Dongfang Electronics Information Industry Group Co, Ltd, Yantai, Shandong* 264001)

Abstract Clustering and classification are two important research areas of data mining. Classification needs related prior-knowledge, while clustering normally finds its own inherent characteristics from the data based on similarity measure. In the process of power load forecasting, the results of classification and clustering are inconsistent. For this problem, this paper propose the definition of associated matrix and on that basis propose associated clustering-classification algorithm. This algorithm is applied to data sample classification for power load prediction, the experiment show that the classification results obtained by our method are more reliable.

Keywords clustering; classification; load forecasting; associated matrix

1 引 言

分类(classification)和聚类(clustering)是数据挖掘中两个重要的研究领域,分类是指依靠领域经验或某种模型,把数据映射到给定类别中的某一个类中.聚类是指根据"物以类聚"的原理,将本身没有类别的样本聚集成不同的簇,它的目的是使属于同

收稿日期:2009-01-11;最终修改稿收到日期:2009-12-14.本课题得到国家自然科学基金(60970088,60775035)、国家"九七三"重点基础研究发展规划项目基金(2007CB311004)、国家科技支撑计划项目基金(2006BAC08B06)、山东省中青年科学家奖励基金(2009BSD01383)资助.**窦全胜**,男,1971 年生,博士,副教授,研究方向为智能科学、数据挖掘及相关方法在电力系统中的应用. E-mail: douqs@ics.ict.ac.cn. **史忠植**,男,1941 年生,研究员,博士生导师,研究领域为人工智能、机器学习等. E-mail: shizz@ics.ict.ac.cn. **姜 平**,男,1979 年生,硕士,讲师,研究方向为人工智能、生物信息计算等. **马君华**,女,1965 年生,硕士,高级工程师,研究方向为智能电力技术.

一簇的样本之间彼此具有较好的相似性，而属于不同簇的样本间应具有足够的差异性. 当前，关于这方面的研究很多，文献[1-2]用粒度计算来解决分类问题，随着粒度计算理论本身的不断完善，这些方法也将得到更大的发展. 文献[3-4]用演化方法来进行分类规则挖掘，这些算法是数据挖掘与计算智能相结合的产物. 文献[5]提出了一种将支持向量机与无监督聚类相结合的新分类算法，并应用于网页分类问题，取得了较好的效果. 关于聚类方法，文献[6]做了系统地综述，对当前的聚类方法作了高度概括. 所有这些研究都代表了这一领域的最新进展，在此不逐一列举.

众所周知，分类需要相关先验知识，最理想的先验知识自然是以下的情形：在特征空间中，异类样本之间有明显的界限，相似性测度较小；而同类样本则聚集成一团，相似性测度较大. 从聚类的角度说，先验知识规定的属于一个类别的样本，依照选定的特征空间和相似性测度，也应当聚成一类. 然而不幸的是，在绝大多数情况下，都不会达到这种理想境界，常常遇到的情况是领域专家认为应该归为一类的点，往往在特征空间中距离特别远，或者说相似性测度特别小，而那些被认为应分别属于不同类别的样本，却相似性测度较大. 换句话说，先验知识极有可能和特征以及相似性测度函数不协调. 文献[7]应用粒度理论对这个问题进行了深入的分析，并给出了基于信息粒度原理的分类算法，有较强的理论和实际意义.

在电力系统负荷预测的过程中，电力专家先将负荷变化特征分成若干类，然后把预测当日的情况归于某一类，相同类别采用相同的预测模式. 这种分类方式依靠于这样的一种假设，即相同类型的负荷曲线的波形是类似的，但是由于造成不同负荷特征的原因非常复杂，领域专家的分类相对笼统，只能大体地反映负荷数据本身所具有的特征. 在实际负荷预测过程中，存在这样的情况，即部分领域专家认为属于同一类的负荷曲线却表现出了不同的特征. 另一方面，同类影响因素引起的负荷变化特征具有较强的相似性，通常能够聚成一类，然而这里存在的问题是，客观上聚成一类的样本集，领域专家却不能给出它们特征之所以相近的原因，从而无法为具体预测提供前提，使得聚类结果难以直接用于预测. 针对上述问题，本文提出调和聚类-分类算法，来保证分类和聚类的一致性，将该方法应用于电力系统负荷预报的样本分类中，取得了较好效果. 以下就调和聚类-分类方法加以详细论述.

2 调和聚类-分类方法

设 $U=\{x_1,x_2,\cdots,x_k\}$ 为样本集合，δ 为聚类操作，在 δ 作用下形成聚类谱系 G，在某一粒度[7]下，切割谱系 G，将 U 分割成相互独立的子集，即 $\delta(U)=\{G_1,\cdots,G_m\}$，依靠先验知识对 U 分类，得到分类 $C=\{C_1,\cdots,C_n\}$，对于 $\forall C_i\in C(i=1,2,\cdots,n)$，都被 $\delta(U)$ 分割成了 m 个子集，$C_i=\bigcup_{j=1}^{m}C_{ij}$，其中 m 是 $\delta(U)$ 中子集的个数，$C_{ij}=C_i\cap G_j$，为 C_i 和 G_j 的交集.

定义 1. 设 U 为样本空间，δ 为聚类操作，在某一粒度下对该聚类谱系进行切割得到 $G=\delta(U)=\{G_1,\cdots,G_m\}$，$C=\{C_1,\cdots,C_n\}$ 为 U 依赖先验知识得到的分类. 称矩阵 $\boldsymbol{\Lambda}=\{\lambda_{ij}\}_{n\times m}$ 为分类 C 基于 G 的调和矩阵. 其中，$\lambda_{ij}=\dfrac{|C_{ij}|}{|C_i|}$，$C_{ij}=C_i\cap G_j$，$|C_{ij}|$、$|C_i|$ 分别表示集合 C_{ij} 和 C_i 中元素个数. 称矩阵 $\boldsymbol{\Lambda}$ 中的每一行 $\boldsymbol{R}_i(i=1,2,\cdots,n)$ 为调和向量.

定义 2. 在定义 1 基础上设 $\boldsymbol{R}_i=(r_{i1},r_{i2},\cdots,r_{im})(i=1,2,\cdots,n)$ 为一调和向量，若存在 s 个分量使得 $r_{ij}\geq\dfrac{1}{m}(j=1,2,\cdots,s)$，则称向量 $\boldsymbol{R}_i$ 为 s 项占优或多项占优，若向量 $\boldsymbol{R}_i$ 中只有唯一的 $r_{ij}>\dfrac{1}{m}$，则称 $\boldsymbol{R}_i$ 单项占优. 若 r_{ij} 使 $\boldsymbol{R}_i$ 单项占优，且 $r_{ij}>0.8$，则称 $\boldsymbol{R}_i$ 是单项充分占优.

理想的情况下，我们期望依靠先验知识得到的分类 C 与聚类谱系在某一粒度下的切割所得到的聚簇集合大体保持一致，此时调和矩阵 $\boldsymbol{\Lambda}$ 为方阵，且每一行 $\boldsymbol{R}_i$ 都是单项充分占优，此时存在置换矩阵 $\boldsymbol{\theta}$，使得 $\boldsymbol{\Lambda\theta}$ 为对角充分占优矩阵，如式(1)所示.

$$\boldsymbol{\Lambda\theta}=\begin{bmatrix}\lambda_{11} & & \\ & \cdots & \\ & & \lambda_{nn}\end{bmatrix}_{n\times n} \tag{1}$$

而最糟糕的情况是，分类与聚类完全无关，每一个分类 C_i 中的样本平均分布在 $\delta(U)$ 中，事实上，这两种极端情况都是不容易出现的.

本文定义的调和矩阵和一些文献中[8]所述的混淆矩阵(confusion matrix)密切相关，所谓混淆矩阵是描绘样本数据的真实属性与识别结果之间的关系、评价分类器性能的一种方法. 具体定义如下：设

给定数据集 $D=\{T_1,T_2,\cdots,T_n\}$ 包含 N 类样本,分类器 C 的混淆矩阵为

$$\boldsymbol{CM}(C,D)=(cm_{ij})_{N\times N},$$

其中,cm_{ij} 表示第 i 类样本被分类器 C 判断成第 j 类样本的数据占第 i 类样本总数的百分率,混淆矩阵中元素的行下标对应目标的真实属性,列下标对应分类器产生的识别属性.对角线元素表示各模式能够被分类器 C 正确识别的百分率,而非对角线元素则表示发生错误判断的百分率.在电力系统负荷分类问题中,完全正确的分类是不容易得到的,通过聚类得到的结果往往更能够客观地反应负荷变化特征,我们试图通过聚类来修订主观分类,调和矩阵中元素的行下标对应的是依据先验知识所产生的属性,列下标对应聚类产生的识别属性.调和矩阵从本质上可以看成混淆矩阵的一种近似形态.

算法 1.　调和聚类-分类算法.

1. 依据领域经验对 U 分类,而得到初始分类,$C=\{C_1,C_2,\cdots,C_n\}$;

2. 对 U 按距离实施聚类操作 δ,得到聚类谱系图 G,在聚类谱系图 G 的最顶部处进行切割,得到两个分支,每一分支都构成一类.由此得到调和矩阵:

$$\boldsymbol{\Lambda}=\begin{pmatrix}\lambda_{1G1} & \lambda_{1G2}\\ \vdots & \vdots\\ \lambda_{nG1} & \lambda_{nG2}\end{pmatrix} \tag{2}$$

3. 设在某一时刻,调和矩阵

$$\boldsymbol{\Lambda}=\begin{pmatrix}\lambda_{1G1} & \lambda_{1G2} & \cdots & \lambda_{1Gj} & \cdots & \lambda_{1Gl}\\ \lambda_{2G1} & \lambda_{2G2} & \cdots & \lambda_{2Gj} & \cdots & \lambda_{2Gl}\\ \vdots & \vdots & & \vdots & & \vdots\\ \lambda_{jG1} & \lambda_{jG2} & \cdots & \lambda_{jGj} & & \lambda_{jGl}\\ \vdots & \vdots & & \vdots & & \vdots\\ \lambda_{nG1} & \lambda_{nG2} & \cdots & \lambda_{nGj} & \cdots & \lambda_{nGl}\end{pmatrix} \tag{3}$$

考察 $\boldsymbol{\Lambda}$ 的每一列 $(\lambda_{1Gj},\lambda_{2Gj},\cdots,\lambda_{nGj})^{\mathrm{T}}$,若存在两个或两个以上分量 λ 大于 μ,且 $|G_j|>\tau|U|$,其中,$0<\tau<\mu<1$ 为设定的阈值参数,$|G_j|$ 和 $|U|$ 分别表示 G_j 和 U 中的元素个数,则在聚类谱系图 G_j 的最顶部进行切割,形成新的分支,并修正调和矩阵 $\boldsymbol{\Lambda}$.反复执行步 3,直至 $\boldsymbol{\Lambda}$ 中每列只存在唯一的 $\lambda_{iGk}>\mu$ 或 $|G_j|<\tau|U|$;

4. 考察矩阵 $\boldsymbol{\Lambda}$ 中的每一列 $(\lambda_{1Gj},\lambda_{2Gj},\cdots,\lambda_{nGj})^{\mathrm{T}}$,若存在两个或两个以上分量 λ 在其所在行内充分占优,不防设这些分量为 $\lambda_{hGj},\lambda_{h+1Gj},\cdots,\lambda_{pGj}$,则将 $C_h,C_{h+1},\cdots,C_p$ 合并成一类.修正矩阵 $\boldsymbol{\Lambda}$;

5. 考察矩阵 $\boldsymbol{\Lambda}$ 中每一行 $\boldsymbol{Row}_i=(\lambda_{iG1},\lambda_{iG2},\cdots,\lambda_{iGm})$,若 $\boldsymbol{Row}_i$ 单项充分占优,则将 C_i 单独作为一类,否则,设定阈值 κ,假定 $\boldsymbol{Row}_i$ 中大于 κ 的分量有 l 个,分别为 $\lambda_{is},\lambda_{is+1},\cdots\lambda_{is+l-1}$ 则根据 $G_s,\cdots,G_{s+l-1}$ 将 C_i 分成 l 类,对于 $\forall x\in C_i-\bigcup_{r=0}^{l-1}G_{s+r}$,按照与集合中心距离最短原则添加到 C_i 的某个分类中.

算法的第 1 步和第 2 步对样本集进行初始分类和聚类.算法第 3 步可以保证,在 C 中只有一个分类 C_i 的绝大多数样本在某一聚类 G_j 中,如果 C 中存在两个以上分类绝大部分元素在同一聚类 G_j,则在 G_j 对应的聚类谱系的最顶层分割 G_j,在最终时刻,若依然存在两个以上 C 中分类绝大部分元素在同一聚类 G_j 中,必有 G_j 中的样本数量小于一定规模,步 4 将这些分类并为一类.在步 5,若调和矩阵中的某一行 $\boldsymbol{Row}_i$ 单项充分占优,则将 C_i 单独置成一类,步 3 和步 4 已经确保了不可能存在两个以上分类 C_i 的绝大多数样本在某一聚类 G_j 中.如果 $\boldsymbol{Row}_i$ 不是单项充分占优向量,且 C_i 中的大部分样本分布在 $G_s,\cdots,G_{s+l-1}$ 中,则依据 $G_s,\cdots,G_{s+l-1}$ 将 C_i 分成 l 类,C_i 中不在 $G_s,\cdots,G_{s+l-1}$ 中的样本按距离最小原则,添加到 C_i 的某个分类中.

若将算法的步 5 中的相关调整近似地看成是聚类的结果,则关于算法收敛性有如下结论:存在合适的参数 μ、τ 使得算法必在某一迭代步骤上停止,停止时刻调和矩阵 $\boldsymbol{\Lambda}$ 有 k 行为单项充分占优,$s-k$ 行为多项占优,其中,$k\geq 0$,$s\leq n$,n 为依靠先验知识所得到的分类数,且 $\boldsymbol{\Lambda}$ 中不存在使所有行都充分占优的列.如式(4)所示:

$$\boldsymbol{\Lambda}=\begin{pmatrix}1 & 0 & \cdots & & & \\ 0 & 1 & \cdots & & & \\ \vdots & \vdots & & & & \\ \hdashline \lambda_{i1} & 0 & & \lambda_{ij} & \cdots & 0\\ 0 & \cdots & \lambda_{i+1,j} & \lambda_{i+1,m} & \cdots & \\ \vdots & \vdots & & & & \end{pmatrix}\begin{matrix}k\\ k+1\\ \\ \\ s\end{matrix} \tag{4}$$

证明.　不妨设向量 $\boldsymbol{v}=(\lambda_{1Gj},\lambda_{2Gj},\cdots,\lambda_{nGj})^{\mathrm{T}}$ 为调和矩阵 $\boldsymbol{\Lambda}$ 的任意一列,只需取 $0<\tau<\mu<0.8$,由于样本集中的数据是有限的,由算法步 3 不难看出,算法必在某一迭代步骤上停止,且算法停止时,$\boldsymbol{\Lambda}$ 中的每一列 $\boldsymbol{v}$ 必满足下述条件之一:

(1) $\boldsymbol{v}$ 中只存在唯一的 $\lambda_{iGk}>\mu$;

(2) 若 $\boldsymbol{v}$ 中只存在两个以上的 $\lambda_{iGk}>\mu$,则必有 $|G_j|<\tau|U|$;

(3) 对于 $\forall\lambda_{iGm}\in\boldsymbol{v}$,均有 $\lambda_{iGm}<\mu$.

首先证明在算法停止时刻,调和矩阵 $\boldsymbol{\Lambda}$ 中不存在使所有行都充分占优的列.用反证法,假设 $\boldsymbol{v}=(\lambda_{1Gj},\lambda_{2Gj},\cdots,\lambda_{nGj})^{\mathrm{T}}$ 是 $\boldsymbol{\Lambda}$ 中的一列,若该列使 $\boldsymbol{\Lambda}$ 所有行都充分占优,即对于 $\forall\lambda_{iGm}\in\boldsymbol{v}$,$\lambda_{iGm}>0.8$,则 $\boldsymbol{v}$ 必满足以上条件(2),则有

$$|G_m|=\sum_{i=1}^{n}|C_i|\lambda_{iGm}>0.8\sum_{i=1}^{n}|C_i|=0.8|U|.$$

与中止条件 $|G_m|<\tau|U|$ 矛盾.

设算法中止时，$\boldsymbol{\Lambda}$ 中中止于条件(1)的列有 w 列，中止于条件(2)的列为 h 列，设$(\lambda_{1Gj},\lambda_{2Gj},\cdots,\lambda_{nGj})^{\mathrm{T}}$是 h 列中的任意一列，该列中分量为 λ_{hGj}，$\lambda_{h+1Gj},\cdots,\lambda_{pGj}$使其所在行充分占优，由步 4 将 C_h，$C_{h+1},\cdots,C_p$合并成一类. 显然，新合成的分类依然是充分占优，设合并操作后产生充分占优的行数为 l，则 $\boldsymbol{\Lambda}$ 中所有充分占优的行数为 $k=w+l$，不妨设经过步 4 合并操作后，调和矩阵 $\boldsymbol{\Lambda}$ 的行数为 s，则 $\boldsymbol{\Lambda}$ 中的其它 $s-k$ 行为多项占优，步 5 将 $\boldsymbol{\Lambda}$ 中的每一行内 λ 值较小的类按距离最近原则，分配到其它类中，由于 $\boldsymbol{\Lambda}$ 中的行与次序无关，经过适当的行次序调整，即为式(4)所示状态. 证毕

在这里应该强调的是：分类采用的先验知识与聚类使用的测度函数所刻画问题的本质是相同的，否则进行调和就没有任何实际意义. 同时，本文所述方法在一定程度上依赖于聚类谱系的划分，在实际应用中我们总是在相关聚类谱系的最顶端进行切割，这种切割方式已经在电力负荷预测中取得了不错的效果，对于不同的问题，是否存在更合理的切割方式，还需要进一步研究. 在这里涉及到两种标准，一个是领域专家的先验知识，这种先验知识往往来源于日常生活中的一些规律，由于影响电力负荷变化的因素多且复杂，一些造成负荷特殊变化的原因还不清晰，电力专家依据的先验知识往往只能大体地反映负荷变化的规律，存在一定的不确切性. 另一方面，通过聚类方法能够从客观上较好地识别负荷变化的不同特征，但是样本之所以聚成一类的原因还无法确定，这样就使得聚类方法无法直接用于预测. 故而我们退而求其次，采取一种相对折衷的办法. 调和聚类-分类算法一定程度上保证了主观经验和客观规律的统一. 以下就本算法在电力负荷预测中的应用过程加以阐述.

3 负荷预测问题描述

负荷预测是电力系统的一个传统研究问题，是指从已知的电力系统、经济、社会、气象等情况出发，通过对历史数据的分析和研究，探索事物之间的内在联系和发展变化规律，对负荷发展做出预先估计和推测. 负荷预测方法众多[9-11]，按照预测期限的不同，电力系统负荷预测分为长期预测、中期预测、短期预测、超短期预测. 其中超短期预测是研究最多的，也是最困难的.

在电力系统负荷预测过程中，通常需要按照专家经验将样本分成不同类别，然后提取不同类别数据的特征，找出特征和影响因素之间的关系，根据这些影响因素的变化来预测未来某日的电力负荷情况. 在这个过程中，合理分类是有效预测的基础，通常情况下，领域专家依靠经验或习惯对样本进行分类，这种方式存在着上面所述的问题，即专家认为应该归为一类的点，往往在特征空间中距离特别远，而不同类别的样本，从特征上看又相当接近.

本文对某电力公司近几年的数据样本分别按经验和上节所述调和聚类-分类方法进行分类，在不同分类的基础上，根据气象因素预测负荷情况，并根据实际负荷数据对基于不同分类产生的预测结果加以评估. 在这里，不同分类方法采用的预测模型是相同的，下面就本文采用的预测模型加以简单描述.

首先，对样本进行分类，应用小波变换针对不同类别对 96 点数据进行特征的提取. 本例中，采用 Daubechies 小波提取负荷数据的特征.

设$\{p(t)\}$，$t=1,2,\cdots,96$ 为某天 96 点负荷值，令 $C_0(t)=p(t)$，小波分解如下：

$$\left\{\begin{array}{c} C_0=p(t) \\ C_j[k]=C_{j-1}\bar{h}[2k] \\ D_j[k]=C_{j-1}\bar{g}[2k] \end{array}\right\},\ j=1,2,\cdots,L \qquad (5)$$

式中，$\bar{h}[-k]=h[k]$，$\bar{g}[k]=g[-k]$，$g[k]=(-1)^{k-1}h[k-1]$，$h[k]$为低通滤波器，$g[k]$为高通滤波器，L 为分解层数 $C_j[k]$，$D_j[k]$，$j=1,2,\cdots,L$ 分别为第 j 层小波变换的低通信号(特征)和高通信号(噪声). 称$\{C_L,D_L,\cdots,D_2,D_1\}$为在尺度 L 下的小波变换序列. 通过小波变换，将 96 点时间序列分解成特征和噪声两部分. 在本例中，采用三层小波分解，具体如图 1 所示.

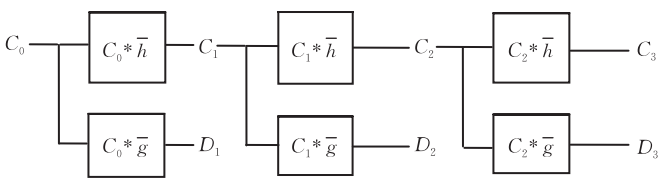


图 1 96 点数据 $p(t)$的三层小波分解

在每一次对 C_j 的分解中，所得到的低通信号 C_{j+1} 和高通信号 D_{j+1} 的维度是 C_j 维度的一半. 初始 C_0 为 96 点负荷数据，经过三层小波分解后，C_3，D_3，D_2，D_1 的维度分别是 12，12，24，48，因此，$\{C_3, D_3, D_2, D_1\}$ 的维度依然是 96，其中，前 12 个分量 C_3 包含着 $\{p(t)\}$ 总体波动信息，即特征分量，而 D_3，D_2，D_1 则是 $\{p(t)\}$ 在不同尺度空间上的高频信息，即噪声分量. 可以通过向量 $\{C_3, D_3, D_2, D_1\}$ 的重构得到 $\{p(t)\}$.

考察气象因素与特征分量和噪声分量的关系，尽管人们通常用气温来反映环境的冷热，但人体对外界冷热的舒适感并不能通过温度一项因素来评价，因此，本文采用实感温度作为影响负荷的主要因素. 实感温度由以下公式计算：

$$\begin{aligned} T_e &= 37 - Q_1/Q_2 - T_r, \\ Q_1 &= \Delta T(1.76 + 1.4V^{0.75}), \\ Q_2 &= (0.68 - 0.41R^h)(1.76 + 1.4V^{0.75}) + 1, \\ T_r &= 0.29T'(1 - R^h), \\ \Delta T &= 37 - T_\alpha \end{aligned} \tag{6}$$

其中，T_e 为实感温度，T_α 为测量温度，V 为风速，R^h 为相对湿度. 从定义上可以看出：风速可以使实感温度降低，但随着气温的增加，对实感温度的影响越来越小，当温度超过 37℃ 时，风速增大会使实感温度增加，当风速一定，气温较高时，湿度的增加使实感温度增加，而气温较低时，湿度增大使实感温度降低. 计算实感温度的气象因素都可以从气象台获得.

考察实感温度和特征分量之间关系，特征分量值随着温度的变化，呈现出一定规律，对这一规律进行回归，从而得到实感温度和特征分量之间的关系多项式. 从而根据实感温度的变化可以对特征分量进行预报.

噪声分量关于实感温度呈散点分布，无法采用回归的方式确定温度和噪声之间的关系. 如上文所述，噪声分量是由 96 点数据经过三层小波分解得到的不同尺度空间上的高频信息构成，向量长度为 84，我们采用如下方式确定噪声分量：

设 $D_i = \{d_{i1}, d_{i2}, \cdots, d_{i84}\}$，$i = 1, 2, \cdots, q$ 为某天 96 点数据的噪声分量，令 $f(d) = \sum_{i=1}^{q}\sum_{j=1}^{84} |d - d_{ij}|$，其中，$q$ 为该分类样本总数. 通过求解优化问题 $\min f(d)$ 来确定噪声分量 $\bar{d}_j$，$j = 1, 2, \cdots, 84$.

由上所述，通过实感温度可以对未来某天的负荷特征进行预测，预测结果与 $\bar{d}_j$ 重构即为当日的电量负荷的预测值.

4　不同分类方法的预报结果

依据先验知识按日期类型对样本进行分类，记作 $C = \{C_1, C_2, \cdots, C_n\}$，对于每一个 96 点历史负荷数据 $(p_0, p_1, \cdots, p_{95})$，令

$$\Delta p = \left(\frac{p_1 - p_0}{p_0}, \frac{p_2 - p_1}{p_1}, \cdots, \frac{p_{95} - p_{94}}{p_{94}}\right),$$

对 Δp 按欧式距离聚类，形成聚类谱系，采用调和聚类-分类方法对上述分类重新分类，在调和聚类-分类方法作用下，原分类 $\{C_1, \cdots, C_n\}$ 被重新划分，得到的分类为 $C' = \{C'_1, \cdots, C'_m\}$，考察最终的调和矩阵

$$\begin{pmatrix} 1 & 0 & \cdots & & \\ 0 & 1 & \cdots & & \\ \vdots & \vdots & & & \\ \hline \lambda_{i1} & 0 & \lambda_{ij} & \cdots & 0 \\ 0 & \cdots & \lambda_{i+1,j} & \lambda_{i+1,m} & \cdots \\ \vdots & \vdots & & & \end{pmatrix},$$

对于调和矩阵 $\boldsymbol{\Lambda}$ 中所有充分占优的行，依据先验知识的分类结果与客观聚类结果大体相同，预报方法与第 3 节所述方法一致.

对于调和矩阵 $\boldsymbol{\Lambda}$ 中所有多项占优的行，我们做如下处理，设 $(0, \cdots, \lambda_{ij}, \cdots 0 \cdots, \lambda_{im}, \cdots, 0)$ 为 $\boldsymbol{\Lambda}$ 中多项占优的行，不妨设 $\lambda'_{is}, \lambda'_{is+1}, \cdots, \lambda'_{is+l-1}$ 为其中占优项，则在新分类集合 C' 中必存在分类 $C'_{is}, C'_{is+1}, \cdots, C'_{is+l-1}$ 与之对应，在这些分类中分别提取特征和噪声，按上节所述方法回归出特征与平均温度之间的关系，并与相应的高频信号重构，得到预测器 $P'_{is}, P'_{is+1}, \cdots, P'_{is+l-1}$，由于 C_i 被分成 $C'_{is}, C'_{is+1}, \cdots, C'_{is+l-1}$ 的原因是不得而知的，具体预测时也只能从先验知识入手，该行所对应的一类情况采用如下方式进行预测：

$$P_i(t) = \sum_{j=s}^{s+l-1} \lambda_{ij} P_{ij} \tag{7}$$

为验证本文所述方法的有效性，分别采用不同的分类方式，以预测当年前两年的历史数据作为学习样本，对国内某电力公司 2007、2008 及 2009 年上半年的负荷情况进行预测，在负荷预测的过程中一些特殊节假日，如春节、元旦等数据规模较小的类，由于历史数据样本少，进行回归是没有意义的，只能按历史趋势进行预测.

表 1 列举了实验过程中的几组统计结果，其中统计类型 A 为数据点中误差小于 1% 数据点的百分比，B 为误差点大于 1% 而小于 3% 点的百分比，C

则是误差大于 3%点的百分比，D 为预测结果与实际数据的均方根误差的平均值.

表 1 基于不同分类的预测结果

年份	误差类型	依据先验知识分类的预测结果	调和聚类-分类的预测结果
2007	*A*	68%(23827 点)	76%(26630 点)
	B	17%(5957 点)	14%(4906 点)
	C	15%(5256 点)	10%(3504 点)
	D	3.02%	2.41%
2008	*A*	77%(27055 点)	81%(28460 点)
	B	15%(5270 点)	13%(4568 点)
	C	8%(2811 点)	6%(2108 点)
	D	2.16%	1.88%
2009 上半年	*A*	75%(13032 点)	80%(13901 点)
	B	16%(2780 点)	15%(2606 点)
	C	9%(1564 点)	5%(869 点)
	D	2.25%	1.82%

从表 1 的统计结果中可以看出，基于新分类的预测结果要明显好于原有的基于经验的分类，通过上述分析不难看出，在依据先验知识的分类中，负荷数据往往在客观上存在着不同的特征，将客观上特征并不趋同样本集当作一类进行回归，是造成回归曲线误差相对较大的主要原因，本文中的调和聚类分类方法一定程度上避免了这个问题，预测的准确率有了明显的提高，这也验证了本文提出的方法较好地解决了先验知识和相似性测度函数不协调的问题.

5 结 论

分类和聚类是数据挖掘中两个重要的研究领域，然而，在实际应用中，通常存在依靠先验知识得到分类结果与聚类结果之间不协调问题，针对这一问题和电力系统的实际应用背景，本文给出了调和矩阵的定义，并在此基础上，提出调和聚类-分类算法，使分类结果与聚类保持较高的一致性，并将该方法应用于电力系统负荷预测中，实验结果证实了本文方法的有效性.

参 考 文 献

[1] Yao T, Yao Y Y. Granular computing approach to machine learning//Proceedings of the 1st International Conference on Fuzzy Systems and Knowledge Discovery: Computational Intelligence for the E-Age. Orchid Country Club, Singapore, 2002, 2: 732-736

[2] Yao Y Y, Yao J T. Induction of classification rules by granular computing//Proceedings of the 3rd International Conference on Rough Sets and Current Trends in Computing. Malvern, PA, USA, 2002: 331-338

[3] Parpinelli R S, Lopes H S, Freitas A A. Data mining with an ant colony optimization algorithm. IEEE Transactions on Evolutionary Computation, 2002, 6(4): 321-332

[4] Sousa T, Silva A, Neves A. A particle swarm data miner// Fernando Moura Pires, Salvador Abreu eds. EPIA 2003. LNAI 2902. Berlin: Springer, 2003: 43-53

[5] Li Xiao-li, Liu Xu-Min, Shi Zhong-Zhi. A Chinese web page classifier based on support vector machine and unsupervised clustering. Chinese Journal of Computers, 2001, 24(1): 62-68(in Chinese)
(李晓黎，刘继敏，史忠植. 基于支持向量机与无监督聚类相结合的中文网页分类器. 计算机学报，2001，24(1)：62-68)

[6] Xu Rui. Survey of clustering algorithms. IEEE Transactions on Neural Networks, 2005, 16(3): 645-678

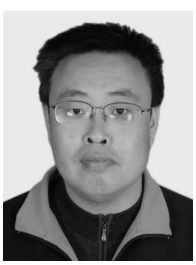
[7] Pu Dong-Po, Bai Shuo, Li Guo-Jie. Principle of granularity in clustering and classification. Chinese Journal of Computers, 2002, 25(8): 810-816(in Chinese)
(卜东波，白硕，李国杰. 聚类/分类中的粒度原理. 计算机学报，2002，25(8)：810-816)

[8] Zhang Jing, Song Rui, Yu Wen-Xian, Xia Sheng-Ping, Hu Wei-Dong. Construction of hierarchical classifiers based on the confusion matrix and Fisher's principle. Journal of Software, 2005, 16(9): 1560-1567(in Chinese)
(张静，宋锐，郁文贤，夏胜平，胡卫东. 基于混淆矩阵和Fisher准则构造层次化分类器. 软件学报，2005，16(9)：1560-1567)

[9] Alfares H K, Nazeeruddin M. Electric load forecasting literature survey and classification of methods. International Journal of Systems Science, 2002, 33(1): 23-34

[10] Metaxiotis K, Kagiannas A, Askounis D, Psarras J. Artificial intelligence in short term electric load forecasting: A state-of-the-art survey for the researcher. Energy Conversion and Management, 2003, 44(9): 1525-1534

[11] Kang Chongqing, Xia Qing, Zhang Boming. Review of power system load forecasting and its development. Automation of Electric Power Systems, 2004, 28(17): 1-11(in Chinese)
(康重庆，夏清，张伯明. 电力系统负荷预测研究综述与发展方向的探讨. 电力系统自动化，2004，28(17)：1-11)



DOU Quan-Sheng, born in 1971, Ph. D., associate professor. His research interests include intelligence computation and data mining.

SHI Zhong-Zhi, born in 1941, professor, Ph. D. supervisor. His research interests include artificial intelligence, machine learning and distributed artificial intelligence.

JIANG Ping, born in 1979, M. S., lecturer. His research interests include artificial intelligence and bioinformatics computation.

MA Jun-Hua, born in 1965, M. S., senior engineer. His research interests include artificial intelligence and application in power system.

Background

This work was partially supported by the National Natural Science Foundation of China (Nos. 60970088, 60775035), Doctor Foundation of Shandong province (grand No. 2009BSD01383), the National High-Tech Research and Development Plan of China (No. 2007AA01Z132). These projects focus on the fields of intelligence science and its applications, authors have done the research on the fields of data mining, intelligence computation and application of artificial intelligence methods in electric power system.

The main contents of this paper involved to electric power load forecasting, which is a traditional research problem on power system. In the process of power load forecasting, electricity experts always divide the forecasting situation into several categories, the same category uses the same forecasting model. There exists such a situation that some load curves which domain experts consider belonging to the same category have shown different characteristics, but some load curves which belong to different categories are very similar, and usually will be gathered into one category by clustering, i. e. the results of classification and clustering are always inconsistent. In this paper, we adopted the compromise solution to solve this problem, and developed the Associated Clustering-classification algorithm. The experimental results show that the classification results obtained by this method are more reliable.

This work was also supported by Yantai Dongfang Electronics Information Industry Group Co, Ltd. (Dongfang). Some ideas of this paper have been embedded into the software system developed by Dongfang.