一种基于海量数据挖掘的设备状态预测算法

唐 胜[1] 胡 洁[2] 赵京虎[1]
(江苏瑞中数据股份有限公司 南京 210003)[1] (南京大学数学系 南京 210008)[2]

摘 要 提出了一种基于海量数据挖掘的设备状态预警算法。工业设备有大量的历史运行数据,并且实时采样的数据维度多,数据量大,算法首先对设备良好运行状态下的大量历史数据进行自适应聚类分析,建立设备的数学模型,并根据此类模型和设备运行的实时状态值对设备的运行状态进行预测。该算法充分考虑工业应用的实际需求,自动确定聚类的数目,解决了传统聚类算法处理海量历史数据时的开销大和效率低的问题,并且保证了回归预测过程的高效性。仿真实验表明,该算法能够有效地处理海量数据,并且能够实时得到预测值,实现对设备的实时监控预测。

关键词 设备状态,海量数据,数据挖掘,预警
中图法分类号 TP393 **文献标识码** A

Equipment Condition Monitoring Algorithm Based on Massive Data Mining

TANG Sheng[1] HU Jie[2] ZHAO Jing-hu[1]
(China Realtime Database Co. Ltd., Sgepri, Nanjing 210003, China)[1]
(Department of Mathematics, Nanjing University, Nanjing 210008, China)[2]

Abstract An equipment condition monitoring algorithm based on massive data mining was proposed. Industrial equipment has massive historical data and has a lot of multidimensional real-time running data. The proposed algorithm makes adaptive cluster analysis with massive historical healthy data to establish the mathematical models of equipment. The algorithm combines these models and real-time running data to achieve predication data. This algorithm can automatically determine the count of clusters by fully considering actual requests from industrial applications, which solves the problem that traditional clustering algorithms have much spending and low efficiency, and it also guarantees the efficiency in the procedure of regression. Simulation results show that the algorithm can effectively deal with massive data and get real-time predicted values, which realizes equipment condition monitoring.

Keywords Equipment condition, Massive data, Data mining, Pre-Alarm

1 引言

现代大型企业的设备稳定、持续的运行与企业的利益息息相关,它们的故障甚至异常停机将给企业带来难以想象的重大损失。因此在其运行过程中,提前发现可能的故障并加以预防和排除非常重要。然而,有一些传统的方法,如定期的人工巡视、设备停工例行检查等,存在以下几个方面的问题:1)定期的检测需要耗费大量的人力、物力,效率很低;2)对一些不必要的设备也进行了检测,造成资源的浪费;3)停机检测可能会带来巨大的经济上的损失。在这样的背景下,企业对设备状态预警方面的需求日益突出。

本文提出的设备状态预警(Equipment Condition Monitoring)技术是利用现代传感技术和计算机技术对运行中的设备进行监测,获取反映运行状态的各种数据值,并对其进行分析处理,预测运行状况,在必要时提供报警和故障诊断信息,避免因故障的进一步扩大而导致事故的发生,为状态检修提供实时数据。状态预警技术在很多领域,如电力、医学、航空、核工业等都有着深刻而广阔的应用前景。

工业设备的状态预警技术,其处理的对象是海量的实时数据。采用的实现算法必须具有处理海量数据的能力,并且能够适应实时处理数据的要求。基于这两点,本文提出一种基于海量数据挖掘的设备状态监测预警方法。首先对设备良好运行状态下的大量的历史数据进行挖掘,得到对应于设备在不同的操作模式下正常运行的各种参数指标值;下一步,结合得到的结果与实时获得的设备参数测量值,给出当前设备运行状态的一个健康状况预测值;最后,将这个预测值与实际值作比较,分析其中的差别,进而得知设备的健康状态,以及潜在的故障相应可能发生在哪里,以指导实际的人工设备检修。

2 算法描述

2.1 算法综述

图1展示了算法的执行场景,系统采用C/S架构,设备报警信息可以发送给各类通讯设备。算法分为两大步骤,一是利用设备运行的历史数据建立起设备运行状态模型,这一步通过聚类算法实现;二是利用经过聚类得到的设备状态模

唐 胜(1987—),男,工程师,主要研究方向为基于海量数据的数据挖掘,E-mail:talestory1314@gmail.com;胡 洁(1987—),女,工程师,主要研究方向为最优化理论与算法;赵京虎(1972—),男,高级工程师,主要研究方向为电网自动化调度。

型，结合设备运行的实时状态数据对当前运行状态进行回归预测。

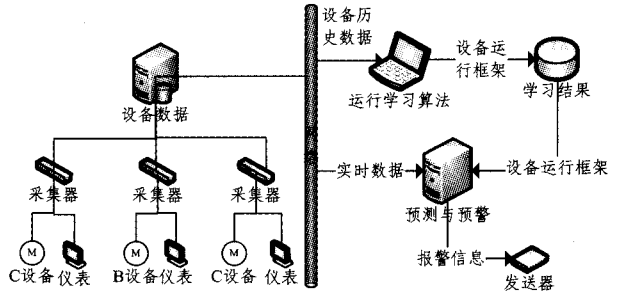

图 1　算法模拟实际应用模型

2.2　步骤一：改进的层次聚类算法

传统的层次聚类算法为树聚类算法，它以一种层次的架构方式，通过对数据进行反复的聚合或分裂，形成一个层次序列的聚类问题解[1]。图 2 描述了常见的凝聚(agglomerative)层次聚类算法以及分裂(divisive)层次聚类算法对数据集 $D=\{x_1,x_2,x_3,x_4,x_5\}$ 的处理思想，其中，

$\forall i=\{1,2,\cdots,5\}, x_i=(v_{i1},\cdots,v_{ik},\cdots v_{in})$

式中，$n\geqslant 1, 1\leqslant k\leqslant n$，$v_{ik}$ 为数据集 D 中向量 x_i 的第 k 维参数的值。

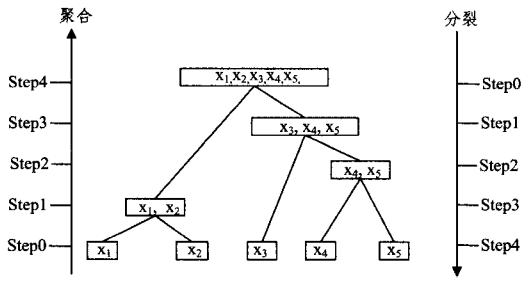

图 2　常见的凝聚层次聚类和分裂层次聚类算法思想

凝聚层次聚类算法采用的是自底而上的方法，分裂层次聚类算法采用的是自顶向下的方法，两者的策略是相反的。但它们的一个共同特点是，都需要检查和估算大量的对象或类[2]，并且进行了反复的迭代，因此传统层次聚类算法的效率比较低，为 $O(tn^2)$，其中 t 为迭代次数，n 为数据集中的数据样本个数。在工业设备状态预警的应用中，反映设备运行历史状态的数据样本通常具有海量的样本量，这时需要学习和处理的历史数据量将非常大，传统的层次聚类方法将无法满足实时处理海量数据的要求。因此，在对海量的工业设备历史状态数据进行挖掘时，本文提出了一种改进的层次聚类算法算法 LISDC(Large Industrial Sampling Data Clustering，海量工业采样数据聚类)，该算法能够在比较小的时间和空间的开销下对海量数据实现聚类，为进一步的回归预测做好准备。

2.2.1　算法思想

LISDC 算法将反映设备历史运行状态的数据样本作为训练数据集，依次读入训练集中的数据向量(Data Vector)，根据训练集的最大值、最小值向量将其标准化，然后确定其所在的类(或者属于某个当前已有类，或者自成一个新的类)，直到所有数据向量被扫描一遍，聚类过程结束。这样的处理过程避免了需要把所有数据一次性全部读入内存而后才能进行聚类的弊端，是凝聚型层次聚类算法的一种改进方法。图 3 描述了 LISDC 算法对数据集 $D=\{y_1,y_2,y_3,y_4,y_5\}$ 的处理思想，其中，

$\forall i=\{1,2,\cdots,5\}, y_i=(y_{i1},\cdots,y_{ik},\cdots y_{in})$

式中，$n\geqslant 1, 1\leqslant k\leqslant n$，$v_{ik}$ 为数据集 D 中向量 y_i 的第 k 维参数的值。

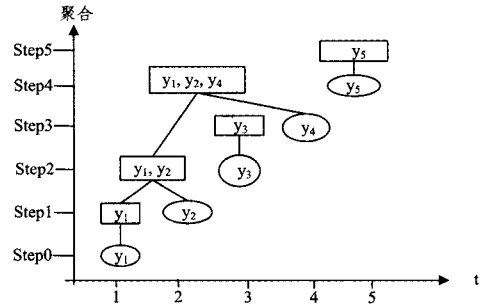

图 3　LISDC 算法思想

图 3 表明了算法的执行过程。$t=1$ 时，历史数据向量 y_1 被读入内存，此时没有任何已存在的类，y_1 自成一个初始的类；$t=2$ 时，y_2 被读入内存，算法根据设定的参数确定 y_2 在哪一个类中，如图 3 所示，将 y_2 加入到之前 y_1 形成的类中，组合生成一个新类；$t=3$ 时，y_3 被读入内存，它超出了算法参数设定的阈值，自成一个新类……如此循环，直到训练集中所有数据向量都被处理过。与传统层次聚类的以“树”的形式呈现结果的方式并不完全相同，LISDC 算法没有聚类形成最终的树状模型，而是以“森林”的方式呈现，主要原因是 LISDC 算法每次只处理当前内存中读入的一个数据向量，仅遍历了一次训练集就生成了最终的聚类模型。

2.2.2　算法实现

定义 1(标准化因子，scale factor)　进行数据标准化时采用的参数，决定数据标准化后的范围，一般取为 1。

定义 2(最大值向量，max vector)　历史数据向量训练集中各个维度的最大值组成的向量，数据标准化时使用。

定义 3(最小值向量，min vector)　历史数据向量训练集中各个维度的最小值组成的向量，数据标准化时使用。在算法中以 min vector 表示。

定义 4(初始半径，initial radius)　当某个数据向量需要自己生成一个新类时，将该向量的各个维度向上扩展 initial radius 距离，向下扩展 initial radius 距离，这样该初始聚类就有了初始的大小范围。

定义 5(扩展容许度，expansion tolerance)　在数据的训练过程中，某个类吸收新读入的训练数据向量时，可以超过其上限或低于其下限的容许值，该值决定了聚类可以超过其上下限而进行膨胀的程度。

定义 6(信号扰动容许度，signal disturbance tolerance)　为一个百分比值。信号在采集传输过程中可能有扰动，这会导致所测数据与实际数据之间产生一个差异，直接用测得值进行聚类可能会影响结果的准确性。于是采取如下策略：当某个类需要扩展去吸收某个训练数据向量时，如果去除其扰动容许度的比例后能聚合到该类中，那么该数据就可以被吸收进来。该值是对训练数据向量即信号的不确定性的度量。

定义 7(聚类上限，cluster upper limit)　对聚类结果中的每个类取类中所有数据向量各个维度的上限组成的向量，每一维度都有上限。

定义 8(聚类下限，cluster lower limit)　对聚类结果中的

每个类取类中所有数据向量各个维度的下限组成的向量，每一维度都有下限。

图 4 表明了一个初始聚类 C_i 的第 k 维分量的可扩展吸收范围，其他分量的吸收范围类似地依据其初始值及各参数值生成。

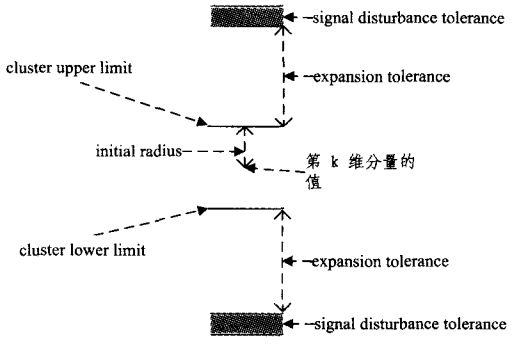

图 4　某个初始聚类第 k 维分量的可吸收范围示意图

经过一阶段的聚类，C_i 已经经历了扩展，其上下限都会有变化，图 5 描述了这一结果。

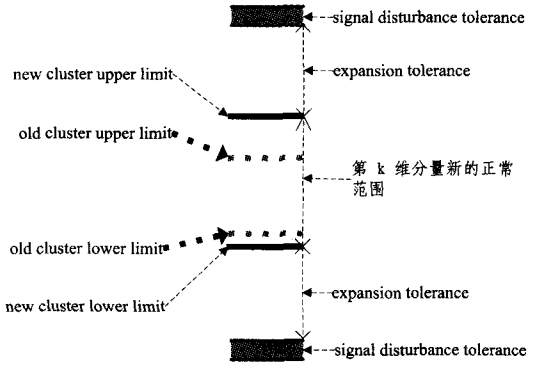

图 5　变化后的第 k 维分量的可吸收范围示意图

算法实现的过程如下：假定数据向量训练集经标准化后为 $D=(o_1,o_2,\cdots,o_m)$，聚类结果的类集合为 C，其中，

$\forall i=\{1,2,\cdots,m\}, o_i=(o_{i1},\cdots,o_{ik},\cdots,o_{in})$

式中，$n\geqslant 1, 1\leqslant k\leqslant n$，$v_{ik}$ 为数据集 D 中向量 o_i 的第 k 维参数的值。n 为每个数据向量的维度。

LISDC1[初始化]　$C=\phi$，读入第一个训练数据向量并对其标准化，记录结果为 o_1，生成初始聚类 C_1，并记录 upper limit 和 lower limit，$C=\{C_1\}$；

LISDC2[处理新数据向量]　读入一个新的训练数据向量并对其标准化，记录结果为 o_2，如果 $\exists C_j\in C$，使得 $distance(o_i,C_j)=0$，那么转至步骤 LISDC3，否则转至步骤 LISDC4；

/ * $distance(o_i,C_j)$ 表示向量 o 与某个类 C 各个维度的欧式距离之和，如果 o 的每个维度分量值都在类 C 对应的分量值范围内（包括扩展和扰动范围），那么 $distance(o_i,C_j)=0$ * /

LISDC3[o_i 并入 C_j]　$C_j=C_j\cup o_i$，同时更新 C_j 的参数，现有的类总个数不变；

LISDC4[o_i 自成一类]　由 $C_{n+1}=\{o_i\}$，$C=C\cup C_{n+1}$ 其中 n 为聚类结果集中聚类的个数；

/ * 此时，C 中类的总个数要加 1 * /

LISDC5　若所有历史训练数据向量都已经处理完，则算法终止；否则，返回步骤 LISDC2。

2.3　步骤二：基于相似性的回归预测算法

本文提出一种基于相似性的回归预测算法 SBR（Similarity Based Regression，基于相似性的回归），该算法利用 LISDC 算法聚类所得的模型进行回归预测。预测过程中利用每个类的上下限作为各种状态的边界值，直接讨论实时测得的数据是否在边界值内，以及差值的大小。避免了逐个对类中数据点进行处理，取得算法正确性和算法性能之间的平衡点。

2.3.1　算法思想

算法的依据是 LISDC 算法得出的聚类模型已经覆盖了设备正常运行下的各种状态。假如算法的输入是某个设备正常运行的状态数据，按照 LISDC 算法的聚类结果，每个类代表了设备运行过程中的一种正常状态，当异常的实时数据输入时，将无法把它合并到任何一个正常类中。基于历史训练数据集的完整性、算法结果的正确性，正常的设备状态必定落在 LISDC 算法聚类结果中的某个类中。如果测得的实时数据向量正常，根据其相似的聚类模型得到的预测值必定与其自身相差不大，而若实时数据向量异常，那么预测值与其相差会很明显。本文正是利用这种原理给出设备运行状态的实时情况。

图 6 描述了基于图 3 所示的聚类结果进行回归预测的思想，其中向量 y_6 为实时获得的需要进行回归预测的当前所测数据向量。已知聚类结果集为 $C=\{C_1,C_2,C_3\}$，其中 $C_1=\{y_1,y_2,y_4\}$，$C_2=\{y_3\}$，$C_3=\{y_5\}$。

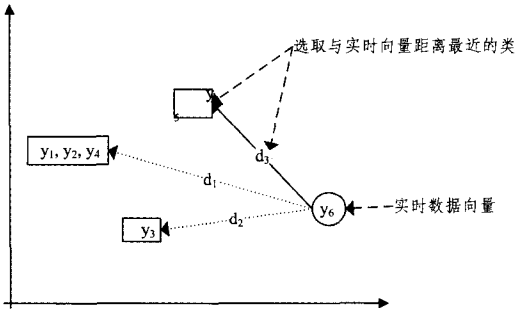

图 6　SBR 算法思想

图 6 描绘了 SBR 算法为一个实时数据向量寻找相似类模型的过程：当 y_6 到来时，SBR 算法遍历聚类结果集。计算该向量与类模型 C_1,C_2,C_3 的距离，记录为 d_1,d_2,d_3，比较 d_1,d_2,d_3 的大小，d_3 最小，选取类 C_3 作为 y_6 的最相似类，认为 y_6 代表设备在此种运行状态下。最后将 y_6 的各个维度与类 C_3 相应的各个维度的上限和下限分别进行比较，某个维度如果在类对应维度的上限和下限之间，那么该维度的预测值就为其本身，否则选取与其距离最近的上限或下限作为预测值。通过这种方式得到各个维度的预测值，即可得到当前设备运行状态的预测值。

2.3.2　算法实现

SBR 算法的基础是 LISDC 算法已经得到了聚类结果集 C，并且使用了 LISDC 算法中的一些变量作为算法的输入。记聚类结果集 $C=\{C_1,C_2,\cdots,C_m\}$，其中 m 为聚类模型的总个数。

SBR 算法有两个过程，一是寻找实时输入的数据与各个类模型的距离的最小值，然后通过最小距离所对应的类模型确定该输入向量最终所在模型；二是根据找到的最相似模型得出输入数据的预测值。

当系统接收到一个实时数据向量 o 时，计算其预测向量的过程描述如下：

SBR1[初始化] $i=0$；

SBR2[计算距离] 计算向量 o 与聚类结果集 $C=\{C_1, C_2, \cdots, C_m\}$ 中聚类的距离，若 $d_i=0$，转至步骤 SBR4；若 $i=m$，转至步骤 SBR3；否则，令 $i=i+1$，转至步骤 SBR2；

SBR3[寻找最近距离] 从距离集合 $d=\{d_1, d_2, \cdots, d_m\}$ 中找到一个 d_i，使得 d_i 为集合 d 中的最小值；

SBR4[确定相似模型] 根据 d_i 确定聚类模型在聚类结果集中的序号为 i，即相似的聚类模型为 C_i；

SBR5[计算预测向量] 根据聚类模型 C_i 得到向量 o 的预测向量 o'。

3 数值实验

3.1 数据集

实验数据来自于 Champaign IL 的艾伯特发电厂的一个水蒸气发电机系统[3]。一共有 8 个监测点，采样频率为 3s 一次，每次采样的 8 个数据就组成了一个数据向量。原始数据向量一共有 9600 个。本文根据需要对这些数据向量进行划分，取不同数量的向量组成训练集，用作预警的实时输入也从这些数据向量中选取。

3.2 聚类 LISDC 过程

包含两组实验。第一组实验用于考察程序的运行时间与历史数据量的关系。该组实验分为 10 小组，数据向量个数分别为 500，1000，2000，3000，4000，5000，6000，7000，8000，9000，参数统一设置如下：标准化因子为 1，初始半径为 0.04，扩展容许度为 0.04，信号扰动容许度为 0.02%，实验结果描点展示如图 7 所示。第二组实验用于考察不同初始参数值对聚类结果以及运行时间的影响。实验分为 5 组，相关参数设置可参见表 1。

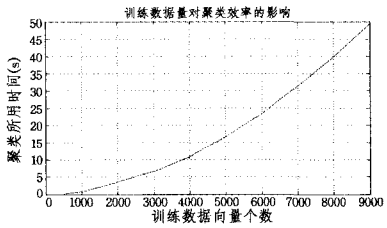

图 7 训练数据量对聚类效率的影响曲线示意图

表 1 参数初始值设定对结果的影响

标准化因子	初始半径	扩展容许度	信号扰动容许度	聚类所得类个数	运行时间(s)
1	0.01	0.02	0.02%	8977	70.647207
1	0.02	0.04	0.02%	8146	65.424975
1	0.03	0.02	0.02%	8679	68.945700
1	0.04	0.04	0.02%	6425	49.557523
1	0.05	0.05	0.02%	3892	25.791215

3.3 回归预测 SBR 过程

对训练数据集的数据量为 9000、回归预测数据集数据量为 600 的情况做了两次实验，聚类过程的两组参数标准化因子、初始半径、扩展容许度、信号扰动容许度分别设定为 1，0.01，0.02，0.02%和 1，0.05，0.05，0.02%，由 3.2 节知第一组参数的聚类结果中有 8977 个类，第二组参数的聚类结果中有 3892 个类。对其中 3 个分量的预测结果如图 8 以及图 9 所示。

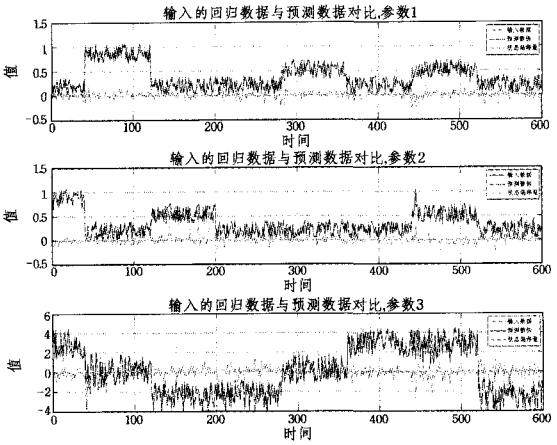

图 8 通过第一组聚类结果进行预测的输入和预测曲线对比

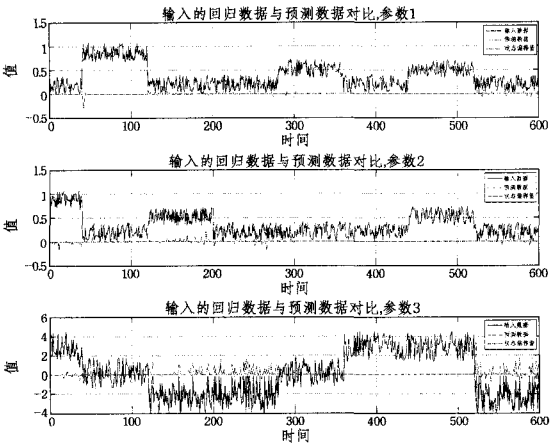

图 9 通过第二组聚类结果进行预测的输入和预测曲线对比

3.4 实验结果评价

由图 7 看出，本文的聚类算法的运行时间随着数据量的增长几乎呈二次增长，这和算法参数设置有关。由于聚类过程中根据每次新读入的数据向量去找其最相似的状态类，因此最坏的情况是每个向量自成一类，此时需要的时间复杂度为 $O(n^2)$。而最好的情形是所有数据向量都被加入到一个类中，此时复杂度为 $O(n)$。由此结合表 1，可以看出根据不同需求，人工设定不同的参数值，可以起到调节运行时间与聚类结果的作用。

在整个算法实现过程中，最终结果呈现在图 8 和图 9 中。这两个图给出了实时的输入值、预测值与两者之间的差值。从时间效率的角度来看，对于类数目为 8977 个类和 3892 个类的两组实验，前者用时 70.6s，后者用时 25.8s。从两个图的预测效果来看，图 9 中的差值在很长一段时间几乎都为 0，表明模型对实时状态的数据不敏感，很多时候无法达到很好的预测效果。相比之下图 8 更好地在某一时刻给出了实时状态的一个评估，当差值较大时，表明设备当前状态的健康程度可能偏低，预示着设备可能发生故障。

结束语 本文提出了一种能够处理工业生产过程中设备状态的海量实时采样数据的算法。算法主体分为两部分，聚类部分和回归预测部分。实验证明，该方法能较好地实时给出设备状态的预测与诊断结果，为企业实现大型设备的状态监测、预警诊断等起到良好的作用。把先进的传感器技术和智能信息处理技术应用于状态预警系统，将在提高设备可靠

（下转第 327 页）

种聚类问题，该算法都得到了比较满意的结果。

Horn 等人提出了一种新的量子聚类算法。该方法是对尺度空间向量聚类和支撑矢量机聚类固有思想的一种扩充。类似于支撑机聚类算法，该方法也与 Hilbert 空间中向量的每个点相关联；同时，还强调了它们的总和，这等于尺度空间概率函数。在这一点上与尺度空间聚类算法类似。新方法是研究 Hilbert 空间的一个算子，由 Schrodinger 等式表示，其概率函数是一个解。这个 Schrodinger 等式包括一个从概率函数中解析导出的势函数。本文将聚类中心与势能最小值联系在一起，最后验证了新方法在已知数据集合上的可行性，并通过限定 Schrodinger 势能对数据点位置的估价，将此方法应用到高维空间中的聚类问题。

4.2 核聚类算法

目前比较经典的聚类算法，如 K-means、模糊 K-means 聚类算法和 Kohonen 自组织神经网络等，只能对一些经典分布的样本奏效。它们没有对样本的特征进行优化，而是直接利用样本的特征进行聚类。因此这些方法的有效性在很大程度上取决于样本的分布情况。例如在一类样本散布较大，而另一类散布较小的情况下，这些方法的聚类效果就比较差。如果样本分布更加混乱，则聚类的结果反而会面目全非。

通过把核方法引入到聚类算法中，本文提出了一种核聚类方法。该方法增加了对样本特征的优化过程，通过利用 Mercer 核把输入空间的样本映射到高维特征空间，并在特征空间中进行聚类。核聚类方法是普适的，并在性能上优于经典的聚类算法，它通过非线性映射能够较好地分辨、提取并放大有用的特征，从而实现更为准确的聚类；同时，算法的收敛速度也较快。在经典聚类算法失效的情况下，核聚类算法仍能够得到正确的聚类。

4.3 谱聚法

最近一类有效的聚类方法开始受到广泛关注。该类方法建立在谱图理论基础之上，并利用数据的相似矩阵的特征向量进行聚类，因而统称为谱聚类方法。谱聚类算法是一种基于两点间相似关系的方法，这使得该方法适用于非测度空间。算法与数据点的维数无关，而仅与数据点的个数有关，可以避免由特征向量的过高维数所造成的奇异性问题。谱聚类算法是一个判别式算法，不用对数据的全局结构作假设，而是首先收集局部信息来表示两点属于同一类的可能性；然后根据某一聚类判据作全局决策，将所有数据点划分到不同的数据集合中。通常这样的判据可以在一个嵌入空间中得到解释，该嵌入空间是由数据矩阵的某几个特征向量张成的。谱方法成功的原因在于：通过特征分解，可以获得聚类判据在放松了的连续域中的全局最优解。

与其他方法相比，谱聚类方法具有明显的优势。该方法不仅思想简单、易于实现、不易陷入局部最优解，而且具有识别非凸分布的聚类能力，非常适合于许多实际应用问题。目前，谱聚类方法已应用于语音识别、视频分割、图像分割、VLSI 设计、网页划分、文本挖掘等领域。

谱聚类方法尽管取得了很好的效果，但目前仍处在发展的初期。算法本身仍存在许多值得深入研究的问题。

结束语 聚类分析作为数据挖掘中的重要组成部分，已经广泛应用于各个领域。在实际应用中，应根据具体问题具体分析，选择使用最佳的聚类方法。纵观数据挖掘中聚类分析方法的发展，可以看出聚类分析的新趋势：a)新方法不断涌现，如基于群的分类方法和基于粒度计算的分类方法。b)根据实际问题的需要，可以有针对性地综合众多领域的技术，以提高分类的性能。总之，数据挖掘中的聚类算法综合了机器学习、数据挖掘、模式识别、物理等领域的研究成果。相信随着这些领域中相关理论的发展、完善和相互渗透，聚类方法也将得到更进一步的发展。

参 考 文 献

[1] 马刚，李志刚. 数据仓库与数据挖掘的原理及应用[M]. 北京：高等教育出版社，2008：20-42

[2] 陈志泊. 数据仓库与数据挖掘[M]. 北京：清华大学出版社，2009：8-37

[3] 郭子龙，等. 免疫进化模糊聚类算法在边缘检测中的应用[M]. 西安：西安交通大学学报，2004：372-373

[4] Chen Y，Tu L. Density-Based Clustering for Real-Time Stream Data[J]. ACMKDD，San Jose，California，USA，2007：133-142

[5] 曲福恒，等. 浅议模糊网络结构聚类算法[J]. 吉林大学学报，2008：18-96

[6] 孙玉芬. 基于网格方法的聚类算法研究[J]. 华中科技大学，2006：62-104

[7] Han J，Kamber M. Data Mining：Concepts and Techniques [J]. Morgan Kaufmann Publishers，2001：33-82

[8] Chenm S，Han Jia-wei，Yup S. Datamining：an overview from a database perspective [J]. IEEE Trans on Knowledge and Data Eng，1996：886-883

[9] Han J，Kamber M. Data Mining：Concepts and Techniques [J]. Morgan Kaufmann Publishers，2001

[10] Wei Yong-qing，Yang Ren-hua，Liu Pei-yu. An improved Apriori algorithm for association rules of mining[C]//Proc of IEEE International Symposium on IT in Medicine &Education. Beijing：IEEE Press，2009：942-946

[11] 黄名选，严小卫，张师超. 基于矩阵加权关联规则挖掘的伪相关反馈查询扩展[J]. 软件学报，2009，20(7)：1854-1865

(上接第 321 页)

性、降低维修成本，优化维修策略方面起到重要作用[4]，对设备全寿命管理有着很好的辅助支撑作用。

参 考 文 献

[1] 孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. Journal of Software，2008，19(1)：48-61

[2] 蒋盛益. 基于聚类的入侵检测算法研究[M]. 北京：科学出版社

[3] ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/espinosa/datasets/powplant.dat

[4] 刘天安. 浅谈电力设备状态监测技术[Z]. China Science and Technology Review，115