基于聚类集成的高铁故障诊断分析

陈云风 王红军 杨 燕
(西南交通大学信息科学与技术学院 成都 610031)

摘 要 聚类集成是对若干独立基聚类器的结果进行组合,从而得到一个对原始数据最优的聚类结果。聚类集成能够减小噪声和孤立点对结果的影响,同时增强聚类结果的鲁棒性和稳定性。从3方面阐述了基于聚类集成的高铁故障诊断分析:1)将原始高铁仿真数据通过傅里叶变化把信号从时域转换到频域,再用不同的特征选择算法进行数据预处理分析;2)分别采用Affinity Propagation(AP)、模糊C均值(FCM)、高斯混合模型(EmGauussian)、Kmeans 4种不同的聚类算法对预处理后的数据进行分析比较;3)引入HGPA、MCLA、CSPA 3种不同聚类集成模型,将得到的基聚类结果分别进行集成。首次把聚类集成算法运用于高铁故障分析中,对比实验结果表明,该方法相比于单个的聚类算法能够更准确有效地进行高铁故障诊断。

关键词 故障诊断,特征选择,聚类分析,聚类集成
中图法分类号 TP277 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.6.049

Fault Diagnosis of High-speed Rail Based on Clustering Ensemble

CHEN Yun-feng WANG Hong-jun YANG Yan
(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)

Abstract Clustering ensemble is the combination of some independent cluster's results, so as to get an optimal clustering result to the original data. Clustering ensemble can reduce the influence of noise and outlier on the clustering result, and at the same time it can also improve the robustness and stability of the clustering results. This paper divided three aspects to describe the fault diagnosis of high-speed rail analysis based on clustering ensemble. In the first aspect, we switched the original simulation data from time domain to frequency domain through discrete Fourier transform, and used different feature selection algorithms for data preprocessing. In the second aspect, we used AP, FCM, EmGauussian and Kmeans, four different clustering algorithms, to analyze it. In the last aspect, we used HGPA, MCLA and CSPA, three different Cluster Ensemble models, to integration the results of clustering algorithms. This paper applied clustering ensemble algorithm in fault diagnosis of high-speed rail for the first time. The experimental results show that this method has better performance than a single clustering algorithm, and can be more accurate and effective for fault diagnosis of high-speed rail.

Keywords Fault diagnosis, Feature selection, Cluster analysis, Clustering ensemble

1 引言

在高速铁路的飞速发展给人们带来交通便利的同时,高铁的安全隐患问题已经成为一个不可忽视的研究项目。减振器是列车运行中用来减少或消除外部有害振动及冲击影响的重要部件,它能使车辆行驶时不仅具有良好的平稳性和舒适性,而且具有良好的安全性。本文实验中用到的数据主要来自于以下几种重要的减震器:抗蛇行减振器、横向减振器和空气弹簧。抗蛇行和横向减振器是车辆系统阻尼元件,抗蛇行减震器可以抑制转向架的蛇行运动,提高列车的安全性;横向减振器主要对车体横向振动起抑制作用,可以改善列车运行时的平稳性并提高列车蛇行运动时的稳定性;空气弹簧位于列车车体与转向架之间,起着承载车体和二系隔振作用,可以获得较高的动扰度,提高乘坐时的舒适感[1]。因此,及时分析减振器数据,定位相应的故障位置并进行分析,是高速列车安全运行的前提。传统的故障诊断方法有基于解析模型的方法、基于信号处理的方法和基于知识的方法,这些方法有的需要准确的数学模型;有的虽然不需要准确的数学模型,但计算量却较大;有的需要丰富的专家经验知识。然而,这些条件相对于高铁故障数据分析来说是难以获取的。聚类集成是一个无监督的学习过程,其对原始数据集的多个基聚类结果进行学习和集成,得到一个能够较好反映数据集内在结构的划分。聚类集成可以明显提高系统的泛化能力,较好地检测和处理孤立点,提高聚类结果质量。相比于传统的故障诊断方法,已经有学者把机器学习的一些聚类算法应用到了故障诊断的研究中,但目前并没有学者把聚类集成技术应用到高铁数据的

到稿日期:2014-06-27 返修日期:2014-10-10 本文受国家自然科学基金项目:藏文Web信息的社会网络动态演化机理研究(61262058),西南交通大学牵引动力国家重点实验室自主研究课题:基于云计算的海量高铁数据处理关键技术研究(2012TPL_T15)资助。
陈云风(1989—),男,硕士生,主要研究方向为机器学习;王红军(1977—),男,副研究员,主要研究方向为机器学习、数据挖掘、计算智能、聚类集成等;杨 燕(1964—),女,教授,主要研究方向为数据挖掘、计算智能、集成学习等。

故障诊断研究中。本文首次采用聚类集成技术对上述几种故障利用采集到的数据进行故障识别，整体过程分为3个阶段：数据预处理、聚类分析、集成分析。

在数据预处理阶段，特征选择是故障诊断中的重要阶段。通过特征选择算法处理能够消除数据的一些噪声，降低数据维度和算法的时间复杂度，从而提高故障诊断效率。被选择的特征的质量也会直接影响故障状态识别的准确率。本文分别采用基尼系数(Gini Coefficient)[4]、统计假说(statistical hypothesis)[5]、非参数检验(Kruskal-Wallis)[6]和fisher score[7]4种方法进行特征选择；然后分别采用基于划分的Kmeans[8]聚类算法、基于模糊均值的FCM[9]聚类算法、基于近邻传播的AP算法[10]和基于极大似然估计的高斯混合模型算法[11]对预处理后的数据进行单个聚类算法的对比分析；最后本文分别选择了HGPA[12]、文献[13]中的MCLA和CSPA 3种聚类集成模型，对单个基聚类结果进行融合、分析，得到最后的故障状态识别结果。

本文第2节介绍了传统故障诊断方法、特征选择算法、基聚类算法和聚类集成模型的相关知识；第3节详细阐述基于聚类集成模型的高铁故障诊断；第4节分析实验及其结果；最后对本文进行总结。

2 相关工作

2.1 传统故障诊断方法

高速运行的机车车辆状态直接关系到列车的行车安全。列车安全、高速和平稳的运行，对列车的转向架、减震器提出了更高的要求。列车的故障诊断系统能够及时探测高速运行时转向架的工作状况、振动加速器和空气噪音的状态值，进而保障行车安全。传统的故障诊断方法有基于解析模型的参数估计方法[14]、基于信号处理的小波变换方法[15]、基于知识的神经网络方法[16]、基于单个聚类算法的故障诊断系统。本文在传统单聚类算法的故障诊断系统上首次将聚类集成了算法用在高铁故障诊断上，克服了传统故障诊断方法计算量大、需要大量先验知识的缺陷，集成多种基聚类算法的优势，使故障识别率达到最优。

2.2 特征选择算法阐述

特征选择也叫特征子集选择(Feature Subset Selection, FSS)，是指从已有的M个特征(Feature)中选择N个特征使得系统的特定指标最优化[17]。Fisher score线性判别的基本原理如下：

$$y(x)=a_0+a_1x_1+\cdots+a_dx_d=a^{\mathrm{T}}x+a_0 \tag{1}$$

可以将d维矢量$a=(a_1,a_2,\cdots,a_d)^{\mathrm{T}}$视作特征空间$x^d$中的以$a_1,a_2,\cdots,a_d$为分量的一个矢量，$a^{\mathrm{T}}x$表示矢量$x$在以$a$为方向的轴上投影的$\|a\|$倍。所求的$a$使投影后同一簇内样本距离小，不同簇内样本距离大。该算法就是求解满足类间离散度和总的类内离散度之比最大的投影方向，然后在一维空间中确定判别规则。基尼系数(Gini Coefficient)是基尼根据伦茨曲线计算出反映收入分配平等程度的指标，直接计算公式如下：

$$G=\frac{1}{2n(n-1)\mu}\sum_{i=1}^{n}\sum_{j=1}^{n}|Y_i-Y_j| \tag{2}$$

式中，n表示样本数量，μ为收入均值，Y为样本的收入值。对应到高铁数据集中，n表示特征维数，μ为特征向量均值，Y表示特征向量。统计假说(statistical hypothesis)是指有关某一总体参数的假设。T分布是一组对称密度函数曲线，具有一个单独参数v以确定某一特定分布，v是自由度。T分布的密度函数为：

$$f_v(t)=\frac{[(v-1)/2]!}{\sqrt{\pi v}[(v-2)/2]!}\left(1+\frac{t^2}{v}\right)^{-\frac{v+1}{2}},-\infty<t<+\infty \tag{3}$$

非参数检验(Kruskal-Wallis)实质是两个独立样本的曼-惠特尼U检验[18]在多个样本下的推广，可用于检验多个总体的分布是否存在显著差异。

2.3 聚类算法分析

聚类分析是指将物理或抽象对象的集合划分成为由类似的对象组成的多个簇的分析过程，使得同一个簇中的对象有很大的相似性而不同簇间的对象有很大的相异性[19]。聚类是搜索簇的无监督学习过程。与分类不同，聚类不依赖预先定义带类标记的训练数据，需要由聚类算法自动确定标记，而分类学习的实例或数据对象有类别标记。

Kmeans算法是最为经典的基于划分的聚类算法，其基本思想是以空间中K个点为中心进行聚类，对最靠近它们的样本归类。通过迭代的方法，逐次更新各聚类中心的值，直至聚类中心的值不再变化。算法描述如下：

①选择K个初始中心点，例如$c[0],\cdots,c[K-1]$；

②对于$data[0],\cdots,data[n]$，分别与$c[0],\cdots,c[K-1]$比较，假定与$c[i]$差值最少的标记为i；

③对于所有标记为i的点，重新计算$c[i]=\{$所有标记为i的$data[j]$之和$\}$/标记为i的个数；

④重复②、③，直到所有$c[i]$值的变化小于给定阈值。

模糊C均值聚类算法(FCM)也是一种基于划分的聚类算法，与普通的硬性划分算法不同的是，它是一种模糊的软聚类划分，是用隶属度确定每个数据点属于某个聚类的程度的一种聚类算法。算法描述如下：

①利用值在0,1间的随机数初始化隶属矩U；

②计算c个聚类中心$c[i],i=1,\cdots,c$；

③计算目标函数，如果它小于某个确定的阈值，或相对上次目标函数值的改变量小于某个阈值，则算法停止；

④计算新的U矩阵，返回步骤②。

Affinity Propagation(AP)聚类算法的原理是根据N个数据点之间的相似度进行聚类，AP算法的具体过程如下：

①计算数据点之间的相似度值，构建相似矩阵S，再计算矩阵S对角线上每个点的preference值；

②设置一个最大迭代次数，迭代过程开始后，计算每一次的Responsiility值和Availability值，根据$R(i,k)+A(i,k)$值来判断k是否为聚类中心；

③当迭代次数超过最大值或者当聚类中心连续多少次迭代不发生改变时终止计算。

混合模型是指随机变量x的概率密度为：

$$p(x|\Theta)=\sum_{k=1}^{M}\alpha_k p_k(x|\theta_k),\sum_k\alpha_k=1 \tag{4}$$

此处的$\Theta=(\alpha_1,\cdots,\alpha_M,\theta_1,\cdots,\theta_M)$，即混合模型是由$M$个分量组成，每个分量的权重系数为$\alpha_k$。当每一个分量都能满足高斯分布时，则称其为高斯混合模型。其计算步骤如下：

①求完整数据对数似然函数的期望；

②求满足期望最大的参数Θ'；

③借助拉格朗日乘子求α_k。

2.4 聚类集成模型

聚类集成的工作原理如图1所示。聚类集成主要分为两个阶段：第一阶段是得到基聚类器的结果数据，这个阶段主要

是用不同的基聚类算法对原始数据集进行多次重复实验来得到结果数据；第二阶段是数据集成，根据聚类集成算法对前一阶段采集的基聚类结果进行集成，从而得到一个最好的对原始数据划分的聚类结果。

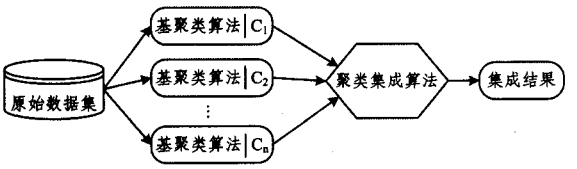

图1 聚类集成工作原理

Alexander Strshl 最早在机器学习研究中明确提出聚类集成问题以及解决方案。通过计算 m 个基聚类结果的相关信息和它们之间的信息熵，来最大程度地分享 m 个基聚类结果，从而得到最好的结果[13]。目前主要的聚类集成算法有3类：第1类是基于图形分割[20]的方法，该方法的一般过程是先把基聚类算法的结果转换成图或超图的点顶和边，然后基于最小权重或者最小切的方法切割图形，最后切割成顶点和边不交叉的几个子图，每一个子图表示一个类别；第2类是基于矩阵相似计算[21]的方法，这类方法是先把基聚类结果转化为矩阵，然后再进行矩阵计算得到数据之间的相似度，并按此相似度聚类；第3类是基于概率统计[22-24]的方法，这类方法主要是先求出基聚类结果的统计特征，基聚类器的权重与其置信度成正比[25]。本文采用的集成方法是由文献[13]提出的3种算法：①CSPA 将数据点当作图的点顶，两个数据点之间的相似度当作边的权重，利用图划分算法 METIS 来得到聚类结果；②HGPA 将数据点当作超图的顶点，每个聚类成员个体的每个类作为一条超边，利用超图划分算法 HMETIS 来得到聚类结果；③MCLA 将每个聚类成员个体的每个类作为图的顶点，两个类之间共有相同数据点的比例作为边的权重，利用图划分算法 METIS 压缩超边集得到元簇，最后根据数据点的比例来决定聚类结果。

3 基于聚类集成的高铁故障诊断

基于聚类集成模型的高铁故障诊断方法的总体框图如图2所示。

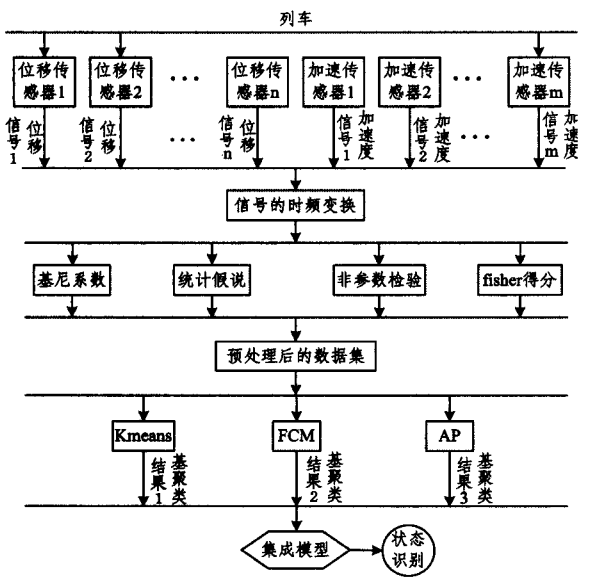

图2 基于聚类集成模型的高铁故障诊断方法总体框图

基于聚类集成的高铁故障诊断算法步骤如下：

输入：{原始数据集}

输出：{数据的聚类集成标签}

①对原始数据进行 FFT 变换，得到频域数据集 $DataSet_n$；

②以100维为单位对连续的一类信号数据 $DataSet_n$ 进行切割，得到数据集 $DataSet_m$；

③将数据集 $DataSet_m$ 代入特征选择算法，并以5维为递增单位，从而得到数据集 $\{DataSet_{m\_i}, i=5,10,\cdots,100\}$；

④将第③步得到的数据集分别代入 Kmeans、FCM、AP 3种聚类算法，每种算法对每个数据集循环20次，并记录平均值和标准差。

⑤将第④步3个基聚类算法对每一个数据集的输出结果进行合并后作为集成算法的输入数据集；

⑥根据集成算法输出的标签进行状态识别。

3.1 信号采集

高铁故障诊断通过分析多个传感器传回的振动信号进行故障判别。通过在列车不同部位安装的位移传感器和加速度传感器采集列车不同位置或同一位置不同速度且不同方向的位移信号，这些信号从不同角度反映了列车的性能。直接从通道采集的列车振动信号是不能直接使用的，一是因为这些信号中包含有噪声，二是传感器直接采集到的时域信号不利于分析，因此需要对采集到的数据进行去噪处理和时频变换。

3.2 特征选择

当列车出现故障时，振动信号将会呈现非平稳、非线性的特性。故障信号是由列车不同位置的传感器的不同通道传递过来的，由于传感器的位置不一样，因此其信号对故障的反映能力也有强弱。如果对所有信号都逐一分析不仅会降低故障诊断的效率，而且一些冗余的信号还会影响到诊断的准确率，因此应选择最能反映故障信息的信号作为分析信号。从通道传送的故障信号维度过高，因此还需要对其进行特征选择来降维。文中分别采用 Fisher score、Gini Coefficient、statistical hypothesis、Kruskal-Wallis 4种算法对数据进行特征选择。首先算出每个维度的权重，然后对其进行排序，我们认为排在前面的维度更具有分析价值。

3.3 状态识别

先将通过特征选择之后的数据集分别作为 Kmeans、FCM、AP、EmGauussian 4种基聚类器算法的输入数据进行聚类分析，然后将4种算法的输出结果作为集成模型的输入数据。聚类集成能够得到一个对原始数据最好的聚类结果，本文分别选取了3种不同的集成模型进行对比实验，最后根据状态识别率来判定算法的有效性。

4 实验结果与分析

4.1 实验数据和评价标准

本文实验数据来源于某高铁研究项目组针对横向减振器故障、抗蛇行减振器故障、空气弹簧故障对某型号动车进行的模拟实验。该模拟实验共设置了4种工况，即实验中聚类时的4种类别，如表1所列。

表1 实验中设置的4种工况

	工况名称	描述
1	抗蛇形减震器全拆	只拆除列车的抗蛇形减震器
2	横向减震器全拆	只拆除列车的横向减震器
3	空簧失气	只使列车的空气弹簧完全失气
4	原车	正常的列车

每一种工况下又对应不同速度的数据，本文实验所用数据的采集速度分别是在 $V=200$km/h、$V=160$km/h、$V=140$km/h、$V=120$km/h、$V=80$km/h 和 $V=40$km/h 下进行的。每个数据集是包含相等数量的 4 种工况的组合，一共有 6 种不同的速度，因此有 6 个不同的数据集进行实验。经过一个傅立叶变换，原始数据达到一个时频变换的效果用于后面的分析。由于实验中每个通道传过来的数据是一个连续的波形信号，如果不对其进行切割处理，一个信号数据集将只有一条连续记录。本文将每个信号数据以 100 维为一个单位截取为一条记录，每种工况下选择 52 条记录组合为实验数据集。因为有 4 种工况，所以最后每个数据集都是一个 208 行 100 列的矩阵。

聚类的评价标准有很多，本文选用 Micro-Precision[26] 标准，该标准是聚类准确率的一种形势，其计算公式如下：

$$P=\frac{1}{N}\sum_{i=1}^{i=k}a_i k \tag{5}$$

其中，N 表示数据集中数据点的数量，a_i 表示对数据在某一类聚类正确的数量，k 表示数据集中聚类的类别数。

4.2 实验结果与分析

本实验先对数据进行特征选择，通过不同的特征选择算法计算出每一维的权重，然后对权重进行由大到小的排序，并每次以 5 维属性作为一个步长递增形成一个特征选择后的子数据集，分别用 Kmeans、FCM、EmGauussian 计算其每个子数据集的准确率，每种算法迭代 20 次并记录下其平均值和标准差，直到属性维数到 100 为止。图 3－图 8 分别给出了速度在 $V=40$km/h、$V=80$km/h、$V=120$km/h、$V=140$km/h、$V=160$km/h 和 $V=200$km/h 下采集的数据集在 4 种不同特征选择算法下，分别采用 3 种聚类算法分析的实验结果。图中 X 轴表示数据维数，从 5 开始，以每 5 维为步长递增至 100 维，因此一共有 20 个结果；Y 轴表示状态识别的准确率。图中在曲线每点的上下振幅表示该算法在对应属性维数下计算的准确率的标准差。图 3－图 8 中 ▷ 标注的线条代表 FCM 算法、* 标注的线条代表 Kmeans 算法、△ 标注的线条代表 EmGauussian 算法。可以发现数据的维度与聚类的准确度并不是成正比，这就说明数据存在部分的噪声；还可以发现 FCM 算法的准确率波动不大，而 Kmeans 准确率的波动相对较大，这是因为 Kmeans 每次都随机地分配中心点，所以聚类结果的准确率在很大程度上取决于随机初始化的中心点。

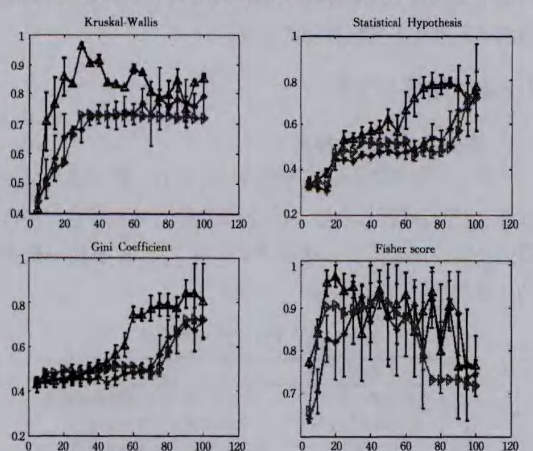

图 3　$V=40$km/h

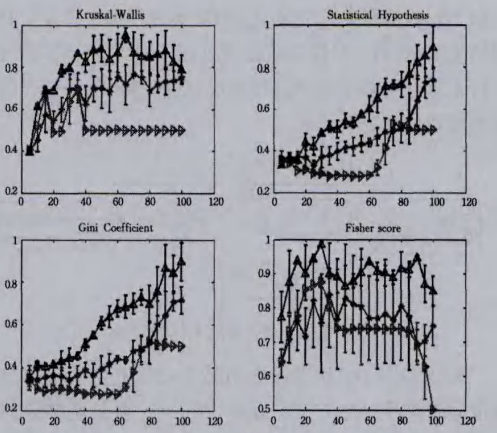

图 4　$V=80$km/h

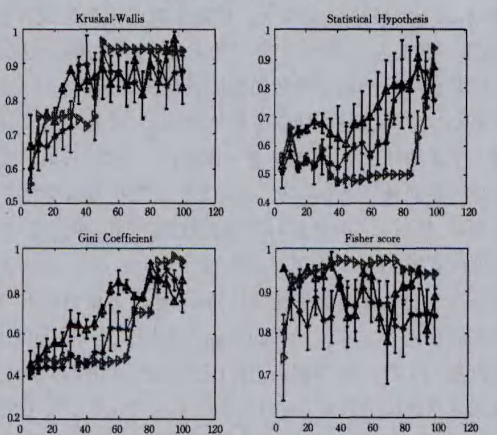

图 5　$V=120$km/h

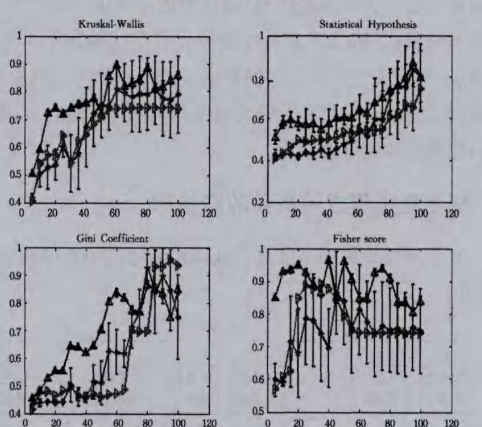

图 6　$V=140$km/h

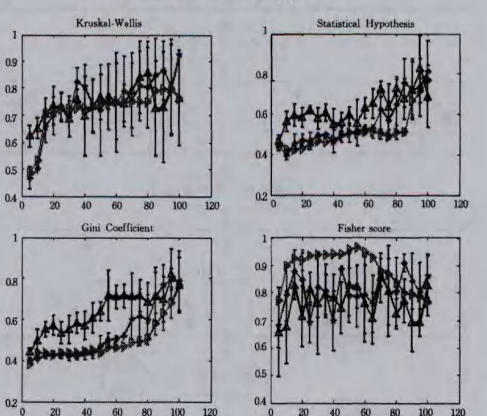

图 7　$V=160$km/h

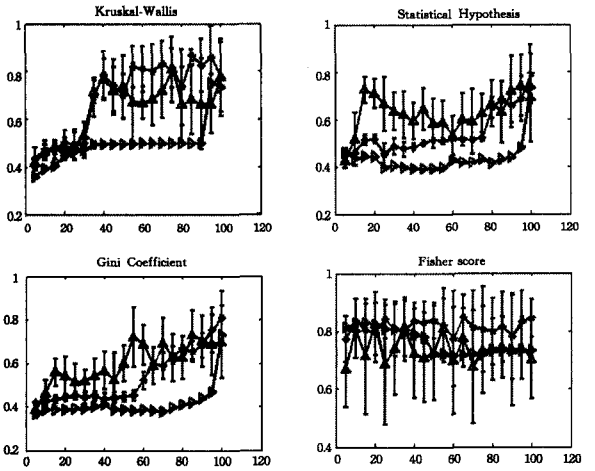

图 8 $V=200$km/h

本文先将传统的聚类算法与特征选择算法相结合用于高铁故障分析，然后用 3 种不同的聚类集成算法与之进行对比，实验结果如表 2 所列。表 2 中 4 种基聚类算法结果中的黑体数据表示该方法迭代 20 次时准确率的最高值，其他数据表示 20 次实验的平均值，最高值反映了该算法能够达到的一个峰值，而平均值则表示该算法对于相应数据的一个平稳结果。不难发现，传统的聚类算法在用于高铁故障诊断时状态识别率明显低于聚类集成算法。为了更直观地描述实验结果，图 9 给出基于集成模型的高铁故障诊断曲线图。从图 9 和表 2 中看出，MCLA 和 CSPA 的实验结果几乎都高于单个基聚类器的结果，说明相比于传统故障诊断方法。该方法的准确率更高、健壮性更好，并且在所有速度下，故障诊断准确率都在 84.6%以上，最高可达 99.5%，说明该方法在一定程度上能有效识别高铁故障。

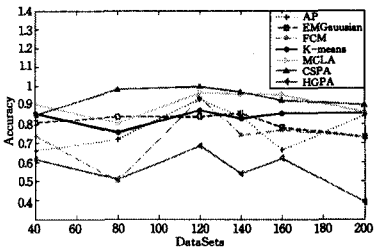

图 9 基于聚类集成的高铁故障诊断折线图

表 2 不同速度下基于集成模型的高铁故障诊断结果

Methds / Dataset	Kmeans		AP		FCM		EmGauussian		MCLA	CSPA	HGPA
Speed_40	**0.9231**	0.8548	0.6578	**0.6578**	0.7380	**0.7380**	0.8055	**0.9808**	0.9038	**0.8462**	0.6154
Speed_80	**0.7837**	0.7555	0.7199	**0.7212**	0.5000	**0.5000**	0.8373	**0.9375**	0.8029	**0.9808**	0.5096
Speed_120	**0.9663**	0.8712	0.9309	**0.9423**	0.9353	**0.9423**	0.8361	**0.9952**	0.9615	**0.9952**	0.6827
Speed_140	**0.9183**	0.8269	0.8365	**0.8365**	0.7401	**0.7404**	0.8550	**0.9856**	0.9567	**0.9663**	0.5385
Speed_160	**0.9279**	0.8550	0.6629	**0.6635**	0.7683	**0.7981**	0.7803	**0.9856**	0.9519	**0.9231**	0.6202
Speed_200	**0.8702**	0.8548	0.8431	**0.8462**	0.7308	**0.7500**	0.7313	**0.9712**	0.8606	**0.8990**	0.3894

结束语 本文提出一种基于聚类集成的高铁故障诊断方法，首先通过不同的特征选择算法对数据进行特征选择，然后使用多种不同的基聚类算法对数据进行单个的聚类分析，再将基聚类结果作为集成算法的输入，从而得到最终的状态识别结果。实验结果表明，不管是在速度较大还是速度较小的情况下，聚类集成算法的结果都优于单个聚类算法，其都能有效识别高铁故障。本文首次尝试将聚类集成算法运用在高铁故障诊断上，并取得了较好的状态识别效果，可见聚类集成算法在高铁故障数据分析研究上有着十分可观的应用前景。

参 考 文 献

[1] 赵晶晶，杨燕，李天瑞，等. 基于近似熵及 EMD 的高铁故障诊断[J]. 计算机科学，2014，41(1)：91-94
Zhao Jing-jing, Yang Yan, Li Tian-rui, et al. Fault Diagnosis of High-Speed Rail Based on Approximate Entropy and Empirical Mode Decomposition[J]. Computer Science, 2014, 41(1): 91-94

[2] 常伟刚，孙跃. 基于贝叶斯网络的 CPT 故障诊断专家系统[EB/OL]. http://www.paper.edu.cn/releasepaper/content/201201-120
Chang Wei-gang, Sun Yue. Fault Diagnosis Expert system of CPT Based on Bayesian Network[EB/OL]. http://www.paper.edu.cn/releasepaper/content/201201-120

[3] 秦娜，王开云，金炜东，等. 高速列车转向架故障的经验模态熵特征分析[J]. 交通运输工程学报，2014，14(1)：58-73
Qin Na, Wang Kai-yun, Jin Wei-dong, et al. Fault feature analysis of high-speed train bogie based on empirical mode decomposition entropy[J]. Journal of Traffic and Transportation Engineering, 2014, 14(1): 58-73

[4] Gini C. Variabilità e mutabilità[M]. Reprinted in Memorie di metodologica statistica, Pizetti E, Salvemini T, eds. Rome: Libreria Eredi Virgilio Veschi, 1912

[5] Veklerov E, Llacer J. Stopping rule for the MLE algorithm based on statistical hypothesis testing[J]. IEEE Transactions on Medical Imaging, 1987, 6(4): 313-319

[6] Breslow N. A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship[J]. Biometrika, 1970, 57(3): 579-594

[7] Gu Q, Li Z, Han J. Generalized fisher score for feature selection [J]. arXiv preprint arXiv: 1202.3725, 2012

[8] Hartigan J A, Wong M A. Algorithm AS 136: A k-means clustering algorithm[J]. Applied statistics, 1979, 28(1): 100-108

[9] Klawonn F, Kruse R, Runkler T. Fuzzy cluster analysis: methods for classification, data analysis and image recognition[M]. New York: John Wiley, 1999

[10] Frey B J, Dueck D. Clustering by passing messages between data points[J]. science, 2007, 315(5814): 972-976

[11] Bilmes J A. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models[J]. International Computer Science Institute, 1998, 4(510): 126

[12] Han E H, Karypis G, Kumar V, et al. Clustering based on association rule hypergraphs[M]. University of Minnesota, Department of Computer Science, 1997

[13] Strehl A, Ghosh J. Cluster ensembles--a knowledge reuse

framework for combining multiple partitions[J]. Journal of Machine Learning Research,2002(12):583-617

[14] 马继涌,高文. 基于最大交叉熵估计高斯混合模型参数的方法[J]. 软件学报,1999,10(9):974-978
Ma Ji-yong,Gao Wen. An Approach for Estimating Parameters in Gaussian Mixture Model Based on Maximum Cross Entropy [J]. Journal of Software,1999,10(9):974-978

[15] 洪飞,吴志美. 基于小波的 Hurst 指数自适应估计方法[J]. 软件学报,2005,16(9):1685-1689
Hong Fei,Wu Zhi-mei. Adaptive Hurst Index Estimator Based on Wavelet[J]. Journal of Software,2005,16(9):1685-1689

[16] 商琳,王金根,姚望舒,等. 一种基于多进化神经网络的分类方法[J]. 软件学报,2005,16(9):1577-1583
Shang Lin,Wang Jin-gen,Yao Wang-shu,et al. A Classification Approach Based on Evolutionary Neural Networks[J]. Journal of Software,2005,16(9):1577-1583

[17] Dy J G,Brodley C E. Feature subset selection and order identification for unsupervised learning[C]// ICML. 2000:247-254

[18] Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph[J]. Journal of mathematical psychology,1975,12(4):387-415

[19] 阳琳赟,周海京,卓晴,等. 基于属性重要性的加权聚类融合[J]. 计算机科学,2009,36(4):243-245
Yang Lin-yun,Zhou Hai-jing,Zhuo Qing,et al. Weighted Cluster Ensemble Based on Significance of Attribute[J]. Computer Science,2009,36(4):243-245

[20] Asur S, Parthasarathy S, Ucar D. An ensemble approach for clustering scale-free graphs[C]// Proc. of ACM KDD. 2006

[21] Li T,Ding C,Jordan M I. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization[C]// Seventh IEEE International Conference on Data Mining, 2007 (ICDM 2007). IEEE,2007:577-582

[22] Topchy A P,Jain A K,Punch W F. A Mixture Model for Clustering Ensembles[C]// SDM. 2004

[23] Berikov V. Weighted ensemble of algorithms for complex data clustering[J]. Pattern Recognition Letters,2014,38:99-106

[24] Minaei-Bidgoli B,Parvin H,Alinejad-Rokny H,et al. Effects of resampling method and adaptation on clustering ensemble efficacy[J]. Artificial Intelligence Review,2014,41(1):27-48

[25] 王红军,李志蜀,成飚,等. 基于隐含变量的聚类集成模型[J]. 软件学报,2009,20(4):825-833
Wang Hong-jun,Li Zhi-shu,Cheng Yang,et al. A Latent Variable Model for Cluster Ensemble[J]. Journal of Software,2009, 20(4):825-833

[26] Zhou Z H,Tang W. Clusterer ensemble. Knowledge-Based Systems,2006,19(1):77-83

(上接第 222 页)

[2] Dai Liu-ling,Liu Bin,Xia Yu-ning. Measuring Semantic Similarity between Words Using HowNet[C]// International Conference on Computer Science and Information Technology. 2008: 601-605

[3] 朱嫣岚,闵锦,周雅倩,等. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报,2009,20(1):14-20
Zhu Yan-lan,Min Jin,Zhou Ya-qian,et al. Semantic Orientation Computing Based on HowNet [J]. Journal of Chinese Information Processing,2009,20(1):14-20

[4] 赵妍妍,秦兵,刘挺. 文本情感分析[J]. Journal of Software, 2010,21(8):1834-1848
Zhao Yan-yan, Qin Bing, Liu Ting. Sentiment Analysis [J]. Journal of Software,2010,21(8):1834-1848

[5] Kamps J, Marx M, Mokken R J, et al. Using WordNet to Measure Semantic Orientations of Adjectives[C]// Proc of LREC-04, 4th Int Conf on Language Resources and Evaluation. Lisbon,2004:1115-1118

[6] Turney P D. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews[C]// Proceedings of the 40th Annual Meeting of the Association Computational Linguistics(ACL). 2002:417-424

[7] Turney P D, Littman, Michael L. Measuring praise and criticism: inference of semantic orientation from association [J]. ACM Transactions on Information Systems, 2003, 21(4): 315-346

[8] 宋晓雷,王素格,李红霞,等. 基于概率浅语义分析的词汇情感倾向判别[J]. 中文信息学报,2011,25(2):89-93
Song Xiao-lei,Wang Su-ge,Li Hong-xia,et al. Word Sentiment Orientation Discrimination Based on PLSA [J]. Journal of Chinese Information Processing,2011,25(2):89-93

[9] Takamura H, Inui T, Okumura M. Extracting Semantic Orientation of Words Using Spin Model[C]// Proc. Ann. Meeting of the Assoc. Computational Linguistics. 2005:133-140

[10] Takamura H,Inui T,Okumura M. Extracting Semantic Orientation of Phrases from Dictionary[C]// Proc. Conf. North Am. Ch. Assoc. for Computational Linguistics. 2007:292-299

[11] 杜伟夫,谭松波,云晓春,等. 一种新的情感词汇语义倾向计算方法[J]. 计算机研究与发展,2009,46(10):1713-1720
Du Wei-fu,Tan Song-bo,Yun Xiao-chun,et al. A New Method to Compute Semantic Orientation [J]. Journal of Computer Research and Development,2009,46(10):1713-1720

[12] Yoshida Y,Hirao T,Iwata T,et al. Transfer Learning for Multiple-Domain Sentiment Analysis-Identifying Domain Dependent/ Independent Word Polarity[C]// Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence. 2011: 1286-1291

[13] Mladenic D, Grobelnik M. Feature selection for classification based on text hierarchy[C]// Proceedings of the Conference on Automated Learning and Discovery(CONALD-98). 1998

[14] Bollegala D,Weir D,Carroll J. Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus[J]. IEEE Transactions on Knowledge and Data Engineering,2013,25(8):1719-1731

[15] Pantel P, Ravichandran D. Automatically Labeling Semantic Classes[C]// Proc. Conf. North Am. Ch. Assoc. for Computational Linguistics: Human Language Technologies. 2004: 321-328

[16] Blitzer J,Dredze M,Pereira F. Biographies,bollywood,boom-boxes and blenders:Domain adaptation for sentiment classification [C]// Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007:432-439