基于选择性集成学习的高速列车故障识别研究*

饶 川, 苟先太, 金炜东
(西南交通大学 电气工程学院, 成都 610031)

摘 要: 在 SVM 分类识别中,分类器模型一经训练得到,对所有测试样本进行无差别的识别。针对高速列车故障中样本的分类识别是存在区域分类精度的情况,提出了一种基于选择性集成学习的 SVM 多分类器融合算法。该方法选取测试样本最邻近的 k 个训练样本;然后选择对其分类效果好的 SVM 分类器进行融合,以提高分类准确率;最后使用高速列车故障数据进行了实验,并与 AdaBoost、KNN、Bayes、SVM 分类方法进行了比较。实验结果表明,该算法提高了分类识别准确率。

关键词: 选择性集成学习; 支持向量机; 多分类器融合; 区域分类精度; 高速列车故障分类
中图分类号: TP391 **文献标志码:** A **文章编号:** 1001-3695(2018)05-1365-03
doi:10.3969/j.issn.1001-3695.2018.05.018

Study on recognition of high speed rail malfunction based on selective ensemble learning

Rao Chuan, Gou Xiantai, Jin Weidong
(*School of Electrical Engineering, Southwest Jiaotong University, Chengdu* 610031, *China*)

Abstract: In the classification and recognition of SVM, classifier model will identify all the testing samples without difference after it's formed. Regarding that accuracy grade of region classification exists in classification and recognition of high-speed rail malfunction samples, this paper proposed a multi-classifier fusion algorithm based on selective ensemble learning. The process of this algorithm first selected the nearest training samples by testing samples of k numbers and then fused them with effective SVM classifiers to increase accuracy rate. And it conducted an experiment with high-speed rail malfunction data, compared this algorithm with AdaBoost, KNN, Bayes and SVM classifications. The experimental results show that the algorithm improves accuracy rate of classification and recognition.

Key words: selective ensemble learning; support vector machine(SVM); multi-classifier fusion; regional classification accuracy; fault classification of high-speed rail

高速列车故障识别主要是通过车体上的实时监测和采集系统收集到的故障振动信息对其进行分析,查找出具体的故障类型。目前,西南交通大学研究团队学者在故障数据特征和特征评价方面提出了一系列的方法来识别故障类型,但故障分类方法的分析和研究相对较少,大多采用支持向量机[1,2](support vector machine,SVM)对故障进行分类[3~5]。虽然 SVM 具有结构简单、泛化能力强和全局最优化等优点[6],但是在实际应用中也存在一些明显的缺点:a)SVM 为降低求解优化问题的时间和空间复杂度,运算过程中需要采用近似算法,导致其泛化性能低于理论期望水平[6];b)SVM 的性能很大程度上取决于模型参数和核函数的选择,目前还没有一个特别有效的方法可以准确找到最优参数,这也会导致支持向量机的训练结果不是最优的[7]。SVM 与选择性集成学习[8]的结合研究为弥补上述缺点提供了一条有效途径。

自从 Schapire[9]、Breiman[10]分别提出了 Boosting、Bagging 等集成学习方法以及周志华提出选择性集成学习方法后,许多学者就对集成学习与支持向量机的结合进行了研究[11~13]。大多数研究是利用 Boosting 和 Bagging 算法对训练集进行重新抽样形成多个训练子集,然后用作为基学习机的 SVM 进行投票组合。由于高速列车故障识别采用多通道传感器收集的监测数据,不同通道的数据集存在着差异性和互补性,不需要利用算法对数据集进行抽样形成训练子集。

本文基于选择性集成学习思想提出了一种基于最近有效领域的 SVM 集成学习方法,并将其用于故障识别。

1 高速列车转向架故障分析

高速列车转向架故障识别方法是在测量系统状态信息的基础上来判定该系统目前所处的运行状态,并在此基础上对故障的类型进行识别。本文研究数据来源于西南交通大学牵引动力国家重点实验室。监测数据是由高速列车转向架安装的多方位传感器采集而来,传感器把采集的位移和加速度转换为电信号,并上传到上位机,用于列车的故障识别[14]。图 1 是转向架监测模型简化图。图中 1、2、3、4、5 各部位都安装有横向或垂直传感器,用于采集各个零部件的振动位移和加速度信号[14]。

高速列车的动力学仿真模型是在仿真软件 SIMPACK 环境下搭建。该模型包括一个车体、两个构架、四个轮对、二系弹簧、二系减震器、抗蛇行减震器、一系弹簧、一系减震器等多个力元,62 个自由度。仿真环境为 LMA 型车轮踏面、CN60 钢轨、轮对内侧局 1 353mm,仿真采样频率为 243 Hz。

高速列车实验工况包含原车正常状态、空簧失气、横向减

收稿日期: 2017-01-05; **修回日期:** 2017-03-06 **基金项目:** 国家自然科学基金重点资助项目(61134002);国家自然科学基金资助项目(61075104);中央高校基本科研业务费专项资金资助项目(SWJTU11BR039,SWJTU11ZT06);四川省科技计划项目—重点研发项目(2017GZ0159)

作者简介: 饶川(1991-),男,湖北监利人,硕士研究生,主要研究方向为智能信息处理、模式识别(raochuan@my.swjtu.edu.cn);苟先太(1971-),男,副教授,博士,主要研究方向为数据通信、agent 技术;金炜东(1959-),男,教授,博导,主要研究方向为智能信息处理、模式识别、图像处理.

振器失效和抗蛇行减振器失效。牵引动力国家重点实验室测试的单一工况故障数据是在保证其他减振器完全正常的情况下,某一种类型的减振器完全拆除来模拟。采集到的监测数据包括系统对正常状态、两种位置的空气弹簧失效、四种位置的二系横向减振器失效、八种位置的抗蛇行减振器失效、四种工况类型的共 15 种不同类型的振动信号仿真数据。

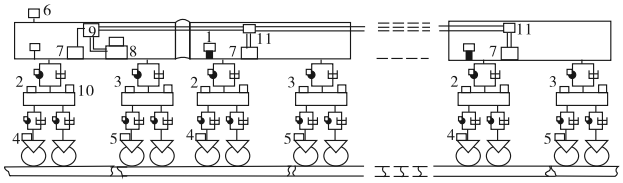

图1 转向架监测模型简化图

2 选择性集成学习的 SVM 多分类器融合算法 SVMg

2.1 选择性集成学习

选择性集成学习的过程如图 2 所示。其主要分为基学习机的生成和基学习机的融合[15]两步。目前,基学习机的生成方法主要包括两大类:a)将不同的学习算法应用到同一数据集上,得到的基学习机被称为异质类型;b)将同一学习算法应用于不同的训练集,得到的基学习机被称为同质类型。基学习机的生成算法主要包括 Bagging、Boosting、旋转森林、Random Forest 和 AdaBoost。其中 Bagging 通过可重复抽样并独立的训练支持向量机作为基学习机[16],而 Boosting 中各基学习机的训练集决定于在其之前产生的基学习机的表现,被已有基学习机错误判断的示例将以较大的概率出现在新基学习机的训练集中。AdaBoost(adaptive boosting,自适应增强)算法是 Freund 等人[17]提出的 Boosting 算法的一种改进。该算法通过迭代生成多个训练集,会更关注上一次迭代被错分的样本,赋予错分样本更大的抽样权重。基学习机的融合方法主要包括投票法、加权投票法、菩萨贝叶斯法、行为知识空间法等[1]。

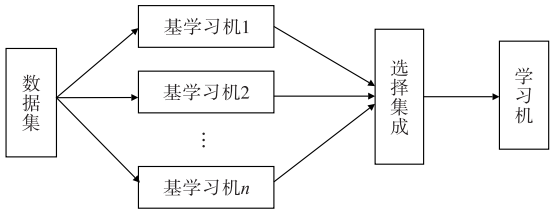

图2 选择性集成学习过程

2.2 算法中基学习机的生成

在高速列车的故障中,转向架安装的多方位传感器采集到不同类型的数据作为训练集,所以基学习机的生成方法主要是同质类型的。在选择学习算法方面,像 AdaBoost 这样经典集成学习算法,主要是通过在同一训练集中抽取不同的训练样本训练出基学习机。由于不同类型传感器采集到的数据带有很大的差异性和互补性,如何将其合并成同一训练基之后再训练出不同的基学习机是得不偿失的,这就需要对不同通道的传感器采集到的数据集进行单独训练。

SVM 是定义在特征空间上的间距最大、结构风险最小的线性分类器[11,12]。如图 3 所示,SVM 通过核函数将低维空间向量集映射到高维空间,然后再寻找一个合适的超平面将不同类型的点区分开,具有很高的预测精度。本算法的基学习机的生成采用 SVM 算法,即通过车体上不同传感器通道采集到的信息训练出不同的 SVM 模型作为基学习机。

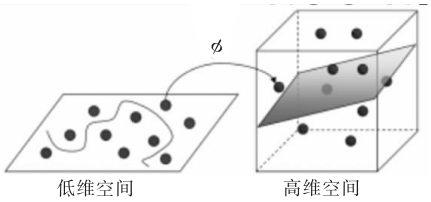

图3 SVM通过核函数映射过程

2.3 算法中基学习机的融合

在选择性集成学习中,成员分类器之间的差异性被视为分类器集成的一个关键因素,而反映在高速列车故障分类方面就是各分类器存在区域分类精度。融合的过程就是要体现成员分类器之间的差异性,即保留一些分类效果好的分类器,剔除一些分类效果不好的分类器。算法融合的基本思路是:首先在训练出的基学习机下找出训练样本中各样本被正确分类的分类器组合,并在每个分类器组合中根据验证集的总体识别率对这些分类器进行排序;然后根据当前输入的测试样本选择出距离最近的 k 个训练样本,再根据这 k 个训练样本的分类器组合对测试样本进行测试,并对测试结果进行多数投票表决。

2.4 基于最近有效邻域的选择性集成学习 SVM 算法 SVMg 的实现

假设选择性集成分类器由 L 个学习机组成,并通过差异性准则 P 和 Q 进行选择,最后通过投票预测输出故障类型。其结构如图 4 所示。

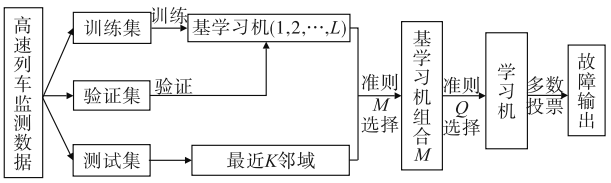

图4 基于选择性集成学习算法结构

算法具体步骤如下:

a)在训练集中,设 ω_j 为 m 个目标类标签,$j \in 1,2,\cdots,m$,$H=\{h_i, i=1,2,\cdots,L\}$ 为训练得到的 L 个不同的分类器 SVM 模型,分类器 h_i 对训练样本 x 的分类输出为

$$h_i(x)=(\theta_{i1},\theta_{i2},\cdots,\theta_{im}) \quad \theta_{ij}\in 0,1; j=1,2,\cdots,m \tag{1}$$

其中:θ_{ij} 表示在分类器 h_i 作用下 x 属于 ω_j 的概率;0 表示被错误分类,1 表示被正确分类。

b)根据每个分类器对训练样本 x 的分类输出结果 $h_i(x)$,得到训练样本被正确输出的分类器组合为

$$M(x)=(h_1, h_2,\cdots, h_k) \quad k\in 1,2,\cdots,L \tag{2}$$

c)在验证集中,将 n 个验证样本输入 L 个不同分类器中,得到总体识别率由大到小的组合为

$$P=(h_{\lambda_1}, h_{\lambda_2},\cdots, h_{\lambda_L}) \tag{3}$$

其中:λ_i 代表在验证样本下,总体识别率第 i 高的分类器在原 L 个分类器中的序号。

d)对于测试样本 x,找出其最近 k 个邻近的训练样本。

e)根据式(2),找出这 k 个训练样本被正确分类的分类器组合 $M(1)$,$M(2)$,$\cdots$,$M(k)$,记为选择准则 M。

f)根据式(3),计算出每个分类器组合 M 中分类器识别率的降序排序 $Q(1)$,$Q(2)$,$\cdots$,$Q(k)$。

g)取每个分类器组合 Q 中前 μ 个组合,将其记为选择准则 Q,得到最终的学习机。设 $S(x_\lambda)$ 为所有选择性集成选择出的个体学习机在测试样本 x_λ 上的实际测试之和,其值为

$$S(x_\lambda)=\sum_{i=1}^{k\times u} f_i(x_\lambda) \tag{4}$$

则集成学习器在测试样本 x_λ 上的预测输出为

$$F(x_\lambda) = \text{sign}(S(x_\lambda)) \tag{5}$$

这里的 sign( )是符号函数,即

$$\text{sign}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} \tag{6}$$

其中:1 表示预测样本输出故障类型正确; -1 表示结果输出错误;而当结果为 0 时,表示故障类型较难判断。

3　实验与分析

为了验证本文提出的基于最近有效邻域的选择性集成学习的 SVM 多分类器融合算法 SVMg 的有效性,对高速列车故障数据进行了实验,并在同样处理的条件下与 AdaBoost、KNN、Bayes、SVM 分类方法进行了比较。

3.1　实验设计

在高速列车四种工况(正常、横向减震器失效、抗蛇行减震器失效、空气弹簧失效)的 15 种数据类型中,原车、横向减震器失效、空气弹簧失效之间的区分性较大,不经过选择性集成学习就可以达到 0.95 以上的识别率,较难区分的是八种抗蛇行减震器失效。本文实验对象选取的是构架 2 右 1 和构架 2 右 2 抗蛇行减震器失效,其时域信号如图 5 所示。

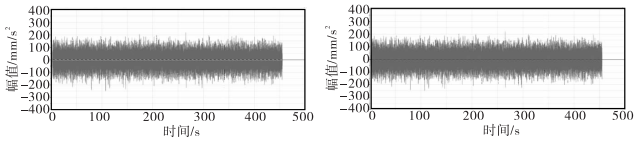

图5　两种工况下的时域信号

每种故障类型采集到的数据以 1 024 的采样点作为一个样本,对其进行三层小波包分解得到一个 8 维特征。每种类型数据有 108 个样本,选取 72 个作为训练样本,18 个作为验证样本,最后的 18 个样本作为测试样本。

3.2　实验结果分析

图 6 表示测试样本选取不同 k 个邻域,并在每个分类器组合中选取不同的 u 值下的分类识别率对比。表 1 为 SVMg 算法与 AdaBoost、KNN、Bayes、SVM 分类识别率的多次对比结果。

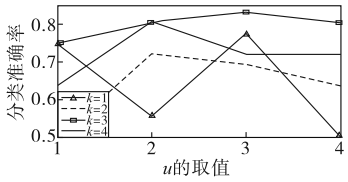

图6　不同邻域k和不同分类器组合u的分类准确率

表 1　各种不同分类方法的平均分类准确率

分类方法	AdaBoost	KNN	Bayes	SVM	SVMg
第一次	0.666 7	0.583 3	0.672 2	0.722 2	0.833 3
第二次	0.666 7	0.500 0	0.650 0	0.833 3	0.805 6
第三次	0.500	0.583 3	0.661 1	0.638 9	0.722 2
第四次	0.666 7	0.472 2	0.661 1	0.611 1	0.805 6
第五次	0.638 9	0.555 6	0.672 2	0.666 7	0.777 8
第六次	0.555 6	0.527 8	0.661 1	0.638 9	0.750 0
第七次	0.722 2	0.694 4	0.666 7	0.638 9	0.805 6
第八次	0.666 7	0.527 8	0.655 6	0.611 1	0.722 2
第九次	0.722 2	0.500 0	0.650 0	0.666 7	0.750 0
第十次	0.638 9	0.555 6	0.655 6	0.694 4	0.805 6
平均准确率	0.644 4	0.550 0	0.660 6	0.672 2	0.777 8

从图 6 中可以看出,最佳邻域 k 和每组分类器的组合 u 的选取对分类准确性有着很大的影响。在 k 和 u 值增加过程中,分类准确率呈现先增加后下降的趋势。测试样本选择三个最近邻域和选取每组分类器组合的前三个时分类的效果最好。从表 1 中可以看出,前四种方法对比中 SVM 的分类效果最好;而 KNN 由于没有学习过程,分类准确率最差;SVMg 和 KNN 都是需要选取最佳邻域进行预测,但 SVMg 增加了 SVM 的强学习过程,平均分类准确率提高了 22.78%;而 SVMg 相比于 SVM,由于 SVMg 增加了选择性集成学习过程,平均分类准确率也提高了 10.56%。以上说明 SVMg 算法相比于 AdaBoost、KNN、Bayes、SVM 分类方法,能够明显地提高分类准确率。

4　结束语

本文从考虑样本区域分类精度出发,结合选择性集成学习和 SVM 的各自优点,提出了基于选择性集成学习的 SVM 多分类器融合分类方法。该方法通过选择待测样本的最佳邻域和最佳 SVM 分类器组合,然后将其预测结果进行动态投票输出。在 SVMg 与 AdaBoost、KNN、Bayes、SVM 等分类方法进行实验对比中,SVMg 方法相比于单一 SVM 和选择性集成学习算法 AdaBoost 能有效地提高高速列车故障分类的准确性。

参考文献:

[1] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报,2000,26(1):32-42.

[2] Vapnik Z H. The nature of statistical learning theory[M]. New York: Spring-Verlag,1995.

[3] 朱明,吴思东,付克昌. 基于熵特征的高速列车故障诊断方法[J]. 振动、测试与诊断,2015,35(2):381-405.

[4] 秦娜,金炜东,黄进,等. 高速列车转向架故障信号的小波熵特征分析[J]. 计算机应用研究,2013,30(12):3657-3659,3663.

[5] 吴志丹,秦娜,金炜东. 基于 EEMD 的高速列车横向减振器故障的排列熵特征分析[J]. 计算机科学,2016,43(5):304-307.

[6] 吴杰长,刘海松,陈国钧. 基于选择性 SVM 集成的模拟电路故障诊断方法[J]. 机械与电子,2011(11):26-29.

[7] 王金彪,周伟,王澍. 基于集成支持向量机的故障诊断方法研究[J]. 电光与控制,2012,19(2):87-91.

[8] Zhou Zhihui, Wu Jianxin, Tang Wei. Ensembling neural networks: many could be better than all[J]. Aritificial Intelligence,2002,137(1-2):239-263.

[9] Schapire R E. The strength of weak learn ability[J]. Machine Learning,1990,5(2):197-227.

[10] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2):123-140.

[11] 谷雨,郑锦辉,戴明伟,等. 基于 Bagging 支持向量机集成的入侵检测研究[J]. 微电子学与计算机,2005,22(5):17-19.

[12] 陈涛. 选择性支持向量机集成算法[J]. 计算机工程与设计,2011,32(5):1807-1809.

[13] 扈晓君. 基于选择性集成学习的支持向量机分类研究[D]. 济南:山东师范大学,2015.

[14] 石晶晶. 基于 FRFT 的高速列车安全性态评估数据特征分析[D]. 成都:西南交通大学,2013.

[15] 张春霞,张讲社. 选择性集成学习算法综述[J]. 计算机学报,2011,34(8):1399-1410.

[16] 魏玲,张文修. 基于支持向量机集成的分类[J]. 计算机工程,2004,30(13):1-2.

[17] Freund Y, Schap I R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences,1997,55(1):119-139.