基于禁忌搜索的管道状况集成检测方法*

王永雄[1,2] 苏剑波[1]

[1](上海交通大学 自动化系 教育部系统控制与信息处理重点实验室 上海 200240)
[2](井冈山大学 电子与信息工程学院 吉安 343009)

摘 要 为提高管道状况异常检测的识别率和实时性,提出基于禁忌搜索的半监督 *K*-means 聚类和 C4.5 决策树的集成检测方法. 在禁忌搜索中引入代价敏感函数,选择具有最佳分类性能的特征组合和最佳组合权值,提高了不平衡数据分布中少数类的识别率. 半监督 *K*-means 方法首先把样本特征聚类为 k 类,再利用 C4.5 方法精确每一类的边界,级联式集成方法缓解不平衡数据分布问题,提高管道检测的准确度. 并提出 3 种集成原则:加权叠加、最近一致和最邻近原则. 实验结果验证了算法的有效性,在管道状况的异常检测中具有较高的分类准确度.

关键词 异常检测, 集成分类, 不平衡数据, 半监督 *K*-means 和 C4.5, 禁忌搜索
中图法分类号 TP 24

An Ensemble Detection Method of Pipeline Condition Based on Tabu Search

WANG Yong-Xiong [1,2], SU Jian-Bo[1]
[1](*Key Laboratory of System Control and Information Processing of Ministry of Education, Department of Automation, Shanghai Jiao Tong University, Shanghai* 200240)
[2](*School of Electronics and Information Engineering, Jinggangshan University, Ji'an* 343009)

ABSTRACT

To improve the recognition rate of pipe anomaly detection and real-time performance, an ensemble classification method based on Tabu search is proposed which combines semi-supervise *K*-means clustering and C4.5 decision tree. The cost-sensitive function is introduced in Tabu search to select the most discriminating feature subset and the best ensemble weights. Thus, the classification performance of the minority class in imbalance data is improved. The semi-supervise *K*-means approach partitions the features of samples into k clusters firstly. Then, a supervised C4.5 decision tree in each *K*-means cluster is trained to refine the decision boundaries by learning the subgroups within the cluster. The ensemble classification by cascading *K*-means and C4.5 alleviates the problems of imbalance data and improves the classification accuracy of imbalance data. The final decisions of the *K*-means and C4.5 methods are integrated based on the weighted sum rule, the nearest-neighbor rule, and the nearest consensus rule respectively. The experimental results show that the proposed system is effective in classifying imbalance data and has high performance in detecting the anomaly of pipeline.

* 国家自然科学基金重点资助项目(No. 60935001)
收稿日期:2011-11-03;修回日期:2012-04-01
作者简介 **王永雄**,男,1970 年生,博士研究生,副教授,主要研究方向为智能机器人及视觉. E-mail:wyxiong@ sjtu. edu. cn.
苏剑波(通讯作者),男,1969 年生,博士,教授,主要研究方向为智能机器人理论与技术、机器学习与人机交互等. E-mail:jbsu@ sjtu. edu. cn.

Key Words Anomaly Detection, Ensemble Classification, Imbalance Data, Semi-Supervise *K*-Means and C4.5, Tabu Search

1 引 言

机器视觉被广泛应用于各种自动异常检测或异常监测系统,例如视频监视系统的异常检测、工业产品表面异常检测和下水管道的缺陷检测等领域[1-3].使用人工神经网络、决策树、SVM 和 Boosting 等[1,3-5]机器学习方法进行异常检测具有较高的检测精度和较低的拒真率(False Positive Rate),但在实际应用中还存在以下一些问题. 1)需要对视觉图像的高维特征进行降维,提高实时性,同时提高分类精度.2)异常检测中的数据分布不平衡性常被忽视.如果在类与类之间有重叠或训练样本不足的难分类问题中,较小的不平衡问题会引起较大分类误差[6].另外,基于机器视觉的自动管道检测具有以下独特性: 1)对管道状况的评价缺乏客观的统一标准.对缺陷和污染程度的划分没有明确的界限,甚至类别之间有重叠; 2)样本的特征分布呈多样性,使得类内还可能存在子类,难以精确划定类与类之间的边界; 3)部分少数类训练样本严重不足(详见实验数据集),难以使用需要很多训练样本的分类方法.因此,文献[1]、[5]中管道缺陷检测的识别率都不高.

针对特征选择问题,只能通过穷举搜索才能保证获得最优解,然而穷举搜索一般只适合于低维问题的特征选择.禁忌搜索(Tabu Search, TS)和遗传搜索算法等元启发式(Meta-Heuristics)方法在处理多噪音并具有多个局部最优的高维特征选择问题具有较好效果[7-9].但遗传算法的搜索时间较长,可能陷入局部最优,而禁忌搜索通过调整编码方式、禁忌表、特赦准则、强化搜索和分散多样化搜索等策略避免循环搜索,从而提高搜索效率,通过局部极小突跳策略避免陷入局部最优.文献[8]、[9]利用禁忌搜索实现最优的或临近最优的组合特征选择,花费的搜索时间小于遗传算法和其它搜索算法.

为了解决异常检测中的不平衡数据问题,主要的方法如下. 1)数据层处理法[10].通过降采样多数类或过采样少数类来"再平衡"数据.或对训练集的重新组合划分,构成数据平衡的子训练集.2)代价敏感方法[4],在标准分类器中引入差异化的代价函数,提高少数类错误识别代价,从而分类器更倾向于提高少数类的识别率.3)特征选择和多分类器集成方法[3].一般单独使用上述某种方法的效果有限,因此本文采用上述多种方法的组合来处理不平衡数据问题,提高少数类的识别率.

针对特征分布的多样性问题,常用的方法是采用聚类对特征进行预处理[3,6],例如 *K*-means、自组织映射(Self-Organizing Map, SOM)和模糊 *C*-mean 等方法,无监督的聚类方法可以缓解人为标准不统一的问题.然而,样本分布的严重不平衡性,采用聚类方法就会出现两个问题[3]:样本过于集中某几个类时引起聚类的类优势(Class Dominance)问题和分类数太少引起的强迫分配(Forced Assignment)问题.文献[3]采用级联 *K*-means 和 ID3 的方法缓解 *K*-means 聚类的两个问题,应用到计算机网络异常点检测等领域,取得了很好的检测效果,但没有考虑特征选择问题和类的不平衡分布问题.

为实现异常检测中的特征选择、减少类的不平衡影响和提高实时性,本文提出一种基于禁忌搜索的组合半监督 *K*-means 和 C4.5 的异常检测方法.首先半监督 *K*-means 利用部分已标记的样本特征产生 l 类的初始聚类中心,实现基于特征的聚类,然后在每个类内用 C4.5 决策树修正或确认 *K*-means 聚类结果,最后通过集成方法确定最终检测结果.在分类器训练阶段,通过禁忌搜索实现特征和组合权值的选择.在测试阶段,采用最邻近、最近一致和加权叠加等 3 种原则组合 *K*-means 和 C4.5 的最终分类结果.算法中采用数据层的降采样、重采样、在特征选择中引入代价敏感函数和集成分类等多种策略来减少类的不平衡影响.

2 基于禁忌搜索的特征选择和最优权值选择

禁忌搜索在处理多噪音、局部最优的高维特征选择问题具有较好效果,考虑到特征提取占用异常检测的主要计算时间[1,5],而管道异常检测对实时性要求较高,所以本文在文献[8]、[9]的基础上权衡实时性和识别精度,在指定最大特征数情况下寻找最优特征子空间.虽然可能得到次优解,但可大幅减少搜索代价和特征提取时间,即保证检测性能,又满足检测的实时性.

本文通过组合 *K*-means 和 C4.5 分类结果来提

高检测精度,加权叠加是常用的组合方法之一.但是最优权值往往难以确定.因此这里把最优权值添加到解编码中,通过禁忌搜索自动寻优.算法流程如图1所示.禁忌搜索中的主要环节如下.

1)初始解和邻域解.为了搜索集成分类的最优加权值,在解的编码中增加集成分类的权值项 α,如图2所示,其中,n 表示每一个解 S 中的特征数,0表示第 n 个特征 F_n 没有包含在解中,1表示包含在解中.在迭代过程中,每次任意替换一个特征或权值更新一次产生一个邻域解,依据最小化代价函数

$$S^* = \arg\min_S Cost(S),$$

从邻域解中选择最优解 S^* 作为下一步迭代的当前解,直到找到预定特征数的最优解.

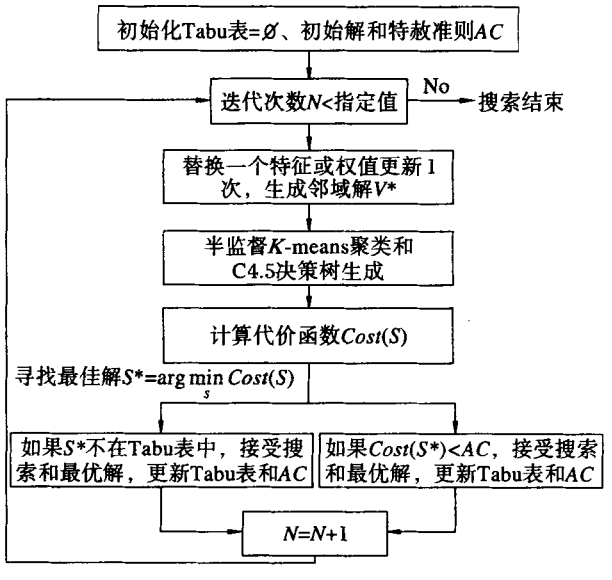

图1　禁忌搜索算法流程图

Fig. 1　Flowchart of Tabu search algorithm

F_1	F_2	…	F_n	α
1	1	…	0	0.5

图2　禁忌搜索的编码方法和一个初始解的例子

Fig. 2　Encoding scheme used in Tabu search and an example of initial solution

2)禁忌搜索的目标函数:为提高少数类的分类精度,目标代价函数定义为

$$Cost = \sum_{i=1}^{k} K_i \frac{N_{ei}}{N_i}, \tag{1}$$

其中,k 是分类数(3个大类:干净、污染和工程缺陷类),N_{ei} 是第 i 类中被错误分类的样本数,N_i 是第 i 类总样本数,K_i 是第 i 类样本的错误分类代价,一般通过先验知识确定,在实验中,多数类的错分代价 $K_1=1$,少数类的错分代价 $K_i=2$,$i\neq 1$.与文献[10]中的代价函数

$$Cost = \sum_{i=1}^{k} K_i N_{ei}$$

不同的是,这里不仅考虑类之间的错误分类代价差别,而且考虑类的不平衡性.多数类的总样本数远大于少数类的总样本数,因此错分少数类样本的代价远大于错分多数类的样本,体现对少数类的重视,使得选择的特征更利于提高少数类的分类精度.通常情况下,多数类(正常样本)的识别率都很高,我们正是通过平衡"每个类中被误分样本的比例"来提高少数类的识别率.

3)禁忌搜索参数选择.禁忌表可避免循环搜索,从而提高搜索效率.Tabu表长度 T 和算法中其它参数相关,如邻域解的数量或迭代次数增加时,都应适当增加 T 的值,T 值的增加也增加搜索的广泛性.我们根据 $T=\sqrt{M}$(见文献[9]、[10])预选 T 的初始值,M 是总特征数.Tabu表长度 T 是很关键的参数,当 T 太小时,容易产生循环搜索,难以跳出局部最优;当 T 太大计算量大,搜索质量差,因此 T 需要通过实验的方法再调整,以获得最佳的搜索效果.

3　基于特征的组合半监督 *K*-means 和 C4.5 检测

3.1　基于特征的半监督 *K*-means 检测

K-means 是完全由数据驱动的算法,并能保证至少是局部最优,但是对噪音和不相关特征敏感.从管道图像提取的大量特征中存在可分性差或包含噪音成分较多的特征,甚至对分类的效果产生负面影响,降低分类精度,同时增加计算量.在很多情况下,由于样本知识的不完备性和特征的多样性,采用半监督的聚类方法可利用部分已知特征生成初始聚类中心(即种子),正确的种子可使得特征聚于一个较好的区域[11],产生一个类似用户指定的聚类,因此产生较好的聚类效果.这里把选中的特征组合成一维特征列向量,然后进行基于特征的聚类.在半监督 *K*-means 聚类中,种子仅在训练阶段的初始化使用[11].半监督 *K*-means 分类的算法步骤如下.

step 1　用第 i 类种子集的均值初始化类"C_i"的中心 r_i,$i=1,2,\cdots,l$,假设每一类中至少有一个种子(根据特征的多样性,采用半监督的方法把管道的3大类再细分为8个子类,即 $l=8$,具体类别详见

实验 4.1 节).

step 2 根据最短距离分配样本 x 到类"C_i",这里

$$i = \arg\min_i \| x - r_i \|^2.$$

step 3 计算各类中数据的均值,更新所有类的中心 r_i.

step 4 重复 step 2 ~ step 4,直到类中心稳定.

step 5 对于测试样本 Z.

step 5.1 计算样本 Z 到类中心 r_i 的距离 $D(r_i, Z)$, $i = 1,2,\cdots,l$,由最近距离原则确定类别.

step 5.2 使用阈值原则或贝叶斯决策原则分类样本 Z. 阈值原则:$Z \to 1$,如果

$$P(\omega_i \in 1 \mid Z \in C_i) > \tau,$$

这里 0 和 1 分别表示正常和异常(缺陷),$\omega_i \in 1$ 表示样本 ω_i 为异常样本,条件概率

$$P(\omega_i \in 1 \mid Z \in C_i)$$

表示样本 Z 被分类为 C_i 类的条件下 C_i 类中异常样本的概率,τ 是阈值;否则,$Z \to 0$.

贝叶斯决策原则:$Z \to 1$,如果

$$P(\omega_i \in 1 \mid Z \in C_i) > P(\omega_i \in 0 \mid Z \in C_i),$$

$\omega_i \in 0$ 表示样本 ω_i 为正常样本;否则,$Z \to 0$,

$$P(\omega_i \in 0 \mid Z \in C_i)$$

表示样本 Z 被分类为 C_i 类的条件下,C_i 类中正常样本的概率.

在实验中,只有当样本聚类到异常样本占多数的类中,才把此样本分类为异常样本,因此采用阈值原则确定样本类别,人为设定阈值 τ 为 0.5.

3.2 C4.5 决策树检测

C4.5 算法是 ID3 上的一个改进算法. 用信息增益率来选择属性,克服了用信息增益选择属性时偏向选择取值多属性的不足,并且能处理连续型数据和不完整数据. 信息增益率定义如下:

$$GainRatio(\Omega,A) = \frac{Gain(\Omega,A)}{SplitI(\Omega,A)}, \tag{2}$$

其中,分裂信息 $SplitI(\Omega,A)$ 代表按照属性 A 分类样本集 Ω 的广度和均匀性,

$$SplitI(\Omega,A) = -\sum_{i=1}^{C} \frac{|\Omega_i|}{|\Omega|} \log_2 \frac{|\Omega_i|}{|\Omega|}, \tag{3}$$

其中,Ω_i 是总输入空间 Ω 分类后的子集,C 是分类数,这里作为二分类问题 C 为 2,信息增益 $Gain(\Omega, A)$ 与 ID3 算法中的信息增益相同:

$$Gain(\Omega,A) = Entropy(\Omega) - \sum_{i=1}^{C} \frac{|\Omega_i|}{|\Omega|} Entropy(\Omega_i), \tag{4}$$

信息熵

$$Entropy(\Omega) = \sum_{i=1}^{C} -p_i \log_2 p_i,$$

其中 p_i 是第 i 类的概率.

C4.5 算法是在 K-means 聚类之后,在每一类内(即 C_i 类内),采用 2 叉树分类方法将类内样本划分为"正常"和"异常"两个子集,进一步判断类内样本是否异常. 具体步骤是:根据属性的信息增益率将训练样本划分为这两个子集,即首先按属性的值排序,找出信息增益率最大的分割点(这里和 C4.5 处理连续数据的方法相同),然后计算两个子集的概率:

$$\mu_i^1 = \frac{\sum_{j=1}^{m} h(j)}{m}, \qquad h(j) = \begin{cases} 1, & \text{样本 } j \text{ 是异常样本} \\ 0, & \text{样本 } j \text{ 是正常样本} \end{cases} \tag{5}$$

其中,m 为 C_i 类中样本的总数,μ_i^1 表示在 C_i 类中样本为异常的概率. 在测试时,若测试样本 Z 判断为异常样本时,则输出异常的概率 μ_i^1,否则输出正常的概率 $\mu_i^0 = 1 - \mu_i^1$.

3.3 组合 K-means 和 C4.5 异常检测

在禁忌搜索选择特征和组合权值的基础之上,通过级联方式组合两种常见的分类方法. 组合 K-means 和 C4.5 异常检测总体框架如图 3 所示.

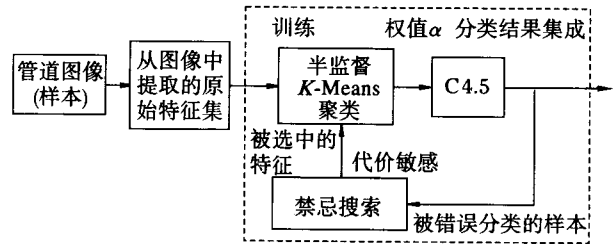

图 3 基于禁忌搜索的半监督 K-means/C4.5 算法原理框图

Fig. 3 Flowchart of semi-supervised K-means/C4.5 classification algorithm based on Tabu search

在训练阶段,首先对被选中的特征执行半监督 K-means 聚类,把训练空间分割成 l 个互不相连的类别 $C_1, C_2, \cdots, C_l$,并得到 l 类的中心 $r_1, r_2, \cdots, r_l$. K-means 方法确保每个训练样本唯一分配到一个类. 但是,如果类与类之间有重叠或者类内存在一些子类,K-means 就可能产生分类错误. 然后在聚类后的每个 C_i 类内,通过监督的方法分别生成一个 C4.5 决策树,C4.5 决策树是对 K-means 分类结果的进一步确认和修正,因此可认为 C4.5 是进一步精确划分特征空间的分类边界.

在测试阶段,首先执行 K-means 聚类和 C4.5 分类,然后通过 3 种集成规则:加权叠加原则、最邻近原则和最近一致原则组合 K-means 和 C4.5 的分类结果,得到最终的检测结果. 在测试时,给定一个测试图像样本 Z,从中抽取特征,应用半监督 K-means 聚类方法进行分类,假设样本 Z 分类为 C_i 类,则得到异常的概率 $P(\omega_i \in 1 \mid Z \in C_i)$ 以及 C4.5 的分类输出概率 μ_i^1. 定义 K-means 结果的异常概率 P_i[3]:

$$P_i = D_i P(\omega_i \in 1 \mid Z \in C_i),\ i = 1,2,\cdots,f, \quad (6)$$

其中

$$D_i = 1 - \frac{d_i}{\sum_{l=1}^{k} d_l}, \quad (7)$$

D_i 是基于欧氏距离的比例因子,$P(\omega_i \in 1 \mid Z \in C_i)$ 表示样本属于 C_i 类异常的概率,f 是样本 Z 可能的候选类数,在实验中 $f \leqslant 3$. $d_1, d_2, \cdots, d_l$ 是样本 Z 到对应的 l 个聚类中心 $r_1, r_2, \cdots, r_l$ 的欧氏距离.

为了得到 K-means 和 C4.5 最终输出概率,提出 3 种集成原则进行组合. 后两个集成原则类似文献[3]的方法,具体如下.

1) 加权叠加原则. 计算样本异常概率 M_i:

$$M_i = \alpha\mu_i^1 + (1-\alpha)P_i,\ i = 1,2,\cdots,f, \quad (8)$$

μ_i^1 是 C4.5 分类器的输出概率,见式(5),P_i 是 K-means 聚类输出概率,见式(6),$\alpha \in [0,1]$ 是加权系数,通过禁忌搜索获得最优解. 这里通过 K-means 和 C4.5 输出概率的加权求和得到样本异常的最终概率 M_i,最优加权值 α 使得组合方法得到最好的分类结果,即通过 C4.5 决策树进一步确认或修正 K-means 的聚类结果. 最后参照 K-means 方法的步骤 step 5.2,根据阈值原则,确定样本的最终检测结果.

2) 最邻近原则. 首先根据阈值原则,分别确定样本 K-means 和 C4.5 的二进制分类结果(见图 4 中括号内的值,0 和 1 分别表示正常和异常). 假设样本 Z 到 C_1 类距离最短,即 d_1 最小,那么样本 Z 被 K-means 分类为 C_1 类. 在图 4 第 1 行中,升序排列样本 Z 到 C_i 类的距离 d_i(即候选类的可能性从大到小排序),K-means 方法把样本 Z 分类到 C_1 类),$d_1 < d_2 < \cdots < d_f$,最后 1 行是对应候选类的 C4.5 分类结果,最后选择第 1 列的 C4.5 分类结果为最终分类结果,即最邻近 K-means 分类结果. 在图 4 的示例中,把 d_1 列 C4.5 的分类结果 0 作为最终分类结果,即判定样本 Z 是正常.

3) 最近一致原则. 从第 1 列(d_1 列)开始,当 K-means 和 C4.5 的二进制分类结果一致时为最终分类结果. 如图 4 中 d_2 列,判定样本 Z 为异常.

在这 3 种组合原则中,加权叠加是综合考虑 2 种分类方法的结果,最近一致原则是考虑 2 种方法的一致性,最邻近原则是以聚类方式确定特征的分类能力,再以 C4.5 的分类结果作为最终结果.

		$C_1(d_1)$	$C_2(d_2)$	$C_3(d_3)$	…	$C_f(d_f)$
		d_1 <	d_2 <	d_3 <…<		d_f
K-Means	P_i	0.89(1)	0.78(1)	0.17(0)	…	0.82(1)
C4.5	μ_i^1	0.12(0)	0.82(1)	0.25(0)	…	0.55(0)

↑
最近一致

图 4　最近一致原则和最邻近原则集成示例

Fig. 4　Example of combining nearest-consensus rule and nearest-neighbor rule

4　实验和结果分析

本文将基于禁忌搜索的组合半监督 K-means/C4.5 方法应用到自动管道检测系统,此系统由移动管道机器人携带数字摄像头和光源进入空调管道,将获取的数字录像或数字图像传送到电脑.

4.1　空调管道异常检测实验

从原始的管道图像中,共选取 328 张管道图,由于样本数据存在严重的不平衡分布,有污染或缺陷的图片比例很小,所以对干净的管道图像进行降采样,同时采用重采样技术,从每张有污染或缺陷的图片抽取多张子图,组成新的子图库,总共 817 张子图,其中干净且无缺陷子图 263 张,具有金属纹理的干净管道子图 280 张,污染的子图 105 张,包含异物 18 张,锈斑 52 张,通孔 5 张、裂缝 36 张和接缝 58 张. 采用形态学方法对管道图像进行图像分割,从这些子图中提取 98 个用于预选的特征,其中基于小波变换的小波特征 16 个[12](分解层数为 3 层),几何特征 5 个,包括面积、长度、紧致度、离心率和凸包[13],表面粗糙度[14] 和灰度直方图等一级统计特征 5 个,LBP 特征 56 个,基于灰度共生矩阵特征 16 个[15]. 在提取基于灰度共生矩阵的特征时,为减少计算量,降低矩阵维数,对图像进行灰度直方图规范化,规范化后的图像灰度级为 8,相邻距离选择为 1,角度分别为 0°,45°,90° 和 135°. 实验采用 5 折交叉验证.

根据清扫和修复目的,管道分为:干净的(正常部分)、有异物或污染的(需清洗部分)、有裂缝、通孔或锈斑等工程缺陷(需更换部分)3 大类,但是由于特征的多样性,在上述 3 大类中还可再细分为多

个子类,这里再细分为 8 个子类:1) 正常类,包含干净、具有金属纹理的 2 个子类;2) 污染类,包含重灰尘、轻度灰尘和异物 3 个子类;3) 工程缺陷类,包含锈斑、通孔和裂缝 3 个子类,即 K-means 中设定类别数 l 为 8.

4.2 实验结果

为评价不平衡多分类问题的性能指标,验证少数类的分类准确度,定义每一大类的分类正确率:

$$Cr(C_i)=\frac{N_C(C_i)}{N(C_i)}\times 100\%, \tag{9}$$

其中,$N(C_i)$ 是 C_i 类的总样本数,$N_C(C_i)$ 是 C_i 类中正确分类的样本数. 接受操作特性曲线(Receiver Operating Characteristic, ROC) 下的面积(Area under Curve, AUC) 是评价分类器整体性能的常用指标,对于多分类系统,参考文献[16]. 这里首先把 C_i 类的所有样本作为正样本,其它样本作为负样本,计算 C_i 类的独立 $AUC(C_i)$,然后定义总 AUC_T 是独立 $AUC(C_i)$ 的加权和:

$$AUC_T=\sum_{C_i}AUC(C_i)p(C_i), \tag{10}$$

其中 $p(C_i)$ 是 C_i 类的先验概率.

为了验证基于禁忌搜索的特征选择有效性,针对不同的组合特征集(人为组合) 分别进行实验,最终分类结果都采用加权叠加方式进行集成. 对图像样本随机分组,重复 10 次实验的平均结果如图 5 所示. 在禁忌搜索中,预定最大特征数为 10,即 $n=10$,权值 α 在初始解中预设为0.5. 在迭代过程中,每次任意替换一个特征或权值更新一次产生一个邻域解,权值项每次迭代步长为 0.1,最终选出的特征包含几何特征 1 个、小波特征 4 个、LBP 特征 2 个、灰度共生特征 1 个和一级统计特征 2 个(10 次实验中禁忌搜索选择的特征子集基本不变,分布也基本相同,具有较好的稳定性). 从图 5 可看出,组合特征集对分类性能的影响,人为组合的几何特征与灰度共生矩阵特征组合分类效果最差,几何和 LBP 特征对少数类的分类效果也很差,LBP 和小波特征组合有较好效果,随着组合特征的增加,分类的性能逐渐提高. 通过禁忌搜索剔除无关特征、对分类结果作用小和包含噪音成分较多的特征,提高分类准确度,这意味着无关特征"混淆" K-means 聚类,对整个分类结果有负面作用,同时大幅减少特征提取的计算量. 从图 5 中还可看出,在特征选择中使用加入样本分布信息的代价敏感函数,使得少数类的分类性能有明显提高.

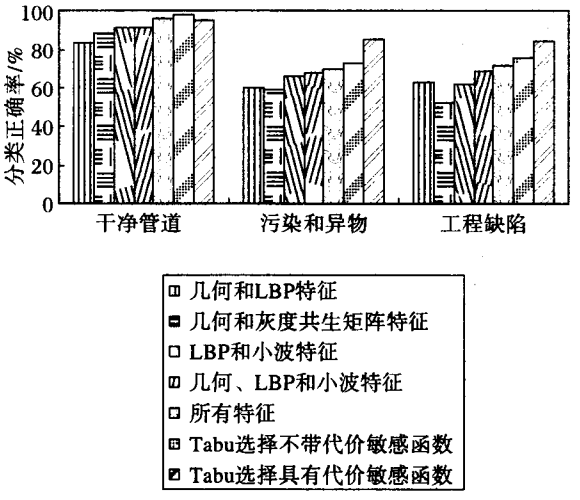

图 5 基于不同特征集的分类结果

Fig. 5 Results of classification based on different combined feature sets

为验证 K-means/C4.5 方法的有效性,分别采用半监督 K-means、BP 神经网络[1]、C4.5、SVM 和 3 种不同集成原则的 K-means/C4.5 方法的进行分类 10 个. 采用式(10)定义的总 AUC_T 作为分类器整体性能评价指标,结果如表 1 所示,粗体字表示获得的最好结果. 表 1 可清楚看出,组合 K-means/C4.5 方法性能明显得到提高,好于单独采用半监督 K-means 或 C4.5

表 1 不同分类方法和 3 种集成原则的分类性能对比

Table 1 Comparison of classification performances by different methods and 3 combining rules

分类方法	正确率/%			总 AUC_T/%	TS 选取的特征个数
	干净管道	污染和异物类	工程缺陷类		
TS+半监督 K-means	86.7	64.1	64.8	73.9	16
TS+BP 神经网络[1]	92.3	69.3	68.9	78.6	15
TS+C4.5	89.4	68.4	69.7	78.5	15
TS+SVM	94.8	**79.8**	73.8	81.6	10
TS+最邻近原则 K-means/C4.5	**96.1**	75.1	75.2	81.3	10
TS+最近一致原则 K-means/C4.5	95.1	74.3	76.1	79.6	10
TS+加权叠加原则 K-means/C4.5	95.2	79.6	**81.8**	**83.8**	10

方法,这说明在预定最大特征数的情况下,禁忌搜索选择最佳分类能力的特征组合,有利于提高少数类的检测识别率,同时由于提取的特征数少,保证检测的实时性. 训练样本不足时,基于分界面支撑点(Support Point)的支持向量机(SVM)方法具有较好的检测精度,但和组合 *K*-means 和 C4.5 方法相比还略逊一筹,这正是样本特征分布的多样性造成的. 从实验的效果看,这 3 种组合原则中加权叠加原则最灵活,分类效果也最好.

5　结 束 语

为提高异常检测识别率和实时性,本文提出基于禁忌搜索的组合 *K*-means 和 C4.5 管道检测方法,实现较高的异常检测效果. 该方法采用禁忌搜索从多噪音多局部最优的高维特征中选择最优或者邻近最优特征组合和最优权值,通过半监督 *K*-means 对图像特征进行预分类,然后利用加权叠加、最近一致原则和最邻近原则三种原则组合 C4.5 分类结果,提高异常检测的性能. 采用数据层的降采样、重采样、在特征选择中引入代价敏感函数和集成分类等多种策略来缓解异常检测中不平衡数据分布问题,保证少数类的分类精度和实时性. 将这个方法应用到基于视觉的自动管道异常检测等分类困难问题,取得较好的识别效果.

参 考 文 献

[1] Duran O, Althoefer K, Seneviratne L D. Automated Pipe Defect Detection and Categorization Using Camera/Laser-Based Profiler and Artificial Neural Network. IEEE Trans on Automation Science and Engineering, 2007, 4(2): 118-126

[2] Tao Xiang, Gong Shaogang. Video Behavior Profiling for Anomaly Detection. IEEE Trans on Pattern Analysis and Machine Intelligence, 2008, 30(5): 893-908

[3] Gaddam S R, Phoha V, Balagani K S. K-means+ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-means Clustering and ID3 Decision Tree Learning Methods. IEEE Trans on Knowledge and Data Engineering, 2007, 19(3): 345-354

[4] Suna Y, Kamela M S, Wong A K C, *et al*. Cost-Sensitive Boosting for Classification of Imbalanced Data. Pattern Recognition, 2007, 40(12): 3358-3378

[5] Yang M D, Su T C. Segmenting Ideal Morphologies of Sewer Pipe Defects on CCTV Images for Automated Diagnosis. Expert Systems with Application, 2009, 36(2): 3562-3573

[6] Japkowicz N, Stephen S. The Class Imbalance Problem: A Systematic Study. Intelligent Data Analysis Journal, 2002, 6(5): 429-450

[7] Tahir M A, Smith J E, Caleb-Solly P. A Novel Feature Selection Based Semi-Supervised Method for Image Classification // Proc of the 6th International Conference on Computer Vision Systems. Santorini, Greece, 2008: 484-493

[8] Zhang Hongbing, Sun Guanyu. Tabu Search Algorithm for Feature Selection. Acta Automatica Sinica, 1999, 25(4): 487-496 (in Chinese)
(张鸿宾,孙广煜. Tabu 搜索在特征选择中的应用. 自动化学报, 1999, 25(4): 487-496)

[9] Tahir M A, Bouridane A, Kurugollu F. Simultaneous Feature Selection and Feature Weighting Using Hybrid Tabu Search / K-Nearest Neighbor Classifier. Pattern Recognition Letters. 2007, 28(4): 438-446

[10] Theodoridis S, Koutroumbas K. Pattern Recognition. 4th Edition. London, UK: Elsevier, 2009

[11] Basu S, Banerjee A, Mooney R. Semi-Supervised Clustering by Seeding // Proc of the 19th International Conference on Machine Learning. Sydney, Australia, 2002: 19-26

[12] Gonzalez R C, Woods R E. Digital Image Processing. 2nd Edition. Upper Saddle River, USA: Prentice-Hall, 2002

[13] Sonka M, Hlavac V, Boyle R. Image Processing, Analysis, and Machine Vision. 2nd Edition. Pacific Grove, USA: Brooks/Cole, 2002

[14] Zaklit J, Wang Yongxiong, Shen Yantao, *et al*. Quantitatively Characterizing Automotive Interior Surfaces Using an Optical TIR-Based Texture Sensor // Proc of the IEEE International Conference on Robotics and Biomimetics. Guilin, China, 2009: 1721-1726

[15] Latif-Amet A, Ertüzün A, Eric A. An Efficient Method for Texture Defect Detection: Sub-Band Domain Co-Occurrence Matrices. Image and Vision Computing, 2000, 18(6/7): 543-553

[16] Ghanem S A, Venkatesh S, West G. Multi-Class Pattern Classification in Imbalanced Data // Proc of the IEEE International Conference on Pattern Recognition. Istanbul, Turkey, 2010: 2881-2884

基于禁忌搜索的管道状况集成检测方法

作者： 王永雄， 苏剑波， WANG Yong-Xiong， SU Jian-Bo

作者单位： 王永雄,WANG Yong-Xiong(上海交通大学自动化系教育部系统控制与信息处理重点实验室 上海200240;井冈山大学电子与信息工程学院 吉安343009)， 苏剑波,SU Jian-Bo(上海交通大学自动化系教育部系统控制与信息处理重点实验室 上海200240)

刊名： 模式识别与人工智能 ISTIC EI PKU

英文刊名： Pattern Recognition and Artificial Intelligence

年，卷(期)： 2013,26(1)