

支持向量机集成和特征选择联合算法

杨宏晖, 孙进才, 牛奕龙, 赵 妮

(西北工业大学航海学院环境工程系, 西安 710072)

摘要: 提出了两种基于支持向量机集成和特征选择联合算法。联合算法的核心思想是在构建基础分类器的同时选择有效特征。通过对实测舰船数据和公共数据的识别实验, 证明了两种算法都可以用于舰船目标识别。算法一更适用于冗余特征较多的情况。算法二在对舰船目标识别时, 选择的特征数目降低为原来特征数目的 30%, 正确分类率比单个支持向量机高近 10%。

关键词: 支持向量机; 特征选择; 分类器集成; 舰船辐射噪声

中图分类号: TN911.7

文献标识码: A

文章编号: 1000-3630(2006)-04-0337-04

Hybrid optimization algorithms of support vector machine ensembles and feature selection

YANG Hong-hui, SUN Jin-cai, NIU Yi-long, ZHAO Ni

(College of Marine Engineering, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: Two hybrid optimization algorithms of ensembles of support vector machine (SVM) and feature selection are proposed in this paper. The main idea of these two algorithms is that searching the effective feature subset during the training of base classifier. Experiments are performed on real-world datasets, and the results demonstrate that the proposed algorithms are well suited for ship radiated noise recognition. Algorithm 1 is more suitable for targets whose features are highly redundant. The number of classification features selected by algorithm 2 drops to 30% of the original feature set, meanwhile the classification accuracy increases by about 10% compared to the accuracy of a single SVM classifier.

Key words: support vector machine; feature selection; classifier ensembles; ship-radiated noise

1 引 言

由于舰船目标的复杂性和相关条件的限制, 对舰船目标的识别一直是国内外研究工作者的研究重点和难点^[1]。

舰船噪声样本特征对舰船目标识别系统的性能是至关重要的。为了提高舰船目标的分类识别率, 越来越多类型的特征被提取出来并加以组合^[2,3]。但

是样本特征空间的维数决定了识别系统的结构复杂度和运算时间的长短。如果增加特征向量的维数, 而训练样本数目没有增加, 类别间的可分离性会明显下降^[4]; 为了保持识别性能, 在特征数目增加的情况下, 就要增加样本数目。舰船目标样本的获取有很多限制条件, 样本数目偏少, 因此, 特征选择是提高舰船目标识别系统正确识别率的有效方法之一^[5]。

支持向量机^[6] (support vector machine, SVM) 是 Vapnik 博士提出基于统计学习理论的针对解决小样本模式识别问题的有效方法, 在很多实际识别问题中都取得了很好的效果。但是 SVM 在学习过程中因为时间和空间运算的高度复杂度, 在运算过程

中采取了近似计算^[7]。因此,有时候 SVM 对实测数据的分类结果与理论期望值有很大的差距。分类器集成本质上是一个基础分类器集合,其中基础分类器的决策被以某种方式加以组合得到综合分类器,用以对新的实例进行分类。基于 SVM 的分类器集成算法,是提高 SVM 对水声信号的分类正确率的有效途径之一。

分类器集成和特征选择是模式识别领域的两个方面,通常分类器集成和特征选择是分别独立进行的,这样做存在着一定的缺陷。首先是无法确定先进行特征选择还是先进行分类器集成。其次,如果先进行其中一个过程,那么另一个就要随机确定。因此,为了获得性能更好的学习机,本文打破传统的分类器集成和特征选择分别进行的方法,提出和两种基于 SVM 分类器集成和特征选择联合算法,并与单个 SVM 的分类性能进行比较。

2 支持向量机

把已知一组 N 个 d 维的独立同分布的训练样本 $X=\{(x_i, y_i) \mid x_i \in R^d, y_i \in \{-1, 1\}, i=1, 2, \dots, N\}$ 通过非线性变换 $h(\cdot)$ 映射到一个高维特征空间 f 。在此高维特征空间中,构造最优线性决策函数 $y(x) = \text{sgn}(\omega \cdot h(x) + b)$ 。 ω 是分类超平面的系数向量, b 是分类阈值。应用 Lagrange 乘子法, ω 可以表示为

$$\omega = \sum_{i=1}^N a_i y_i h(x_i) \quad (1)$$

a_i 是 Lagrange 乘子。因此,在高维空间中,支持向量机的训练过程和检验过程都只是取决于 f 空间的样本的点积运算,即 $h(x_i) \cdot h(x_j)$ 。

引入原样本空间核函数

$$k(x_i, y_j) = h(x_i) \cdot h(x_j) \quad (2)$$

则将高维空间的点积运算转变成原空间的核函数,避免了指定非线性变换形式和真正实施复杂的非线性映射。在本文中所用的核函数是径向基核函数

$$k(x_i, y_j) = \exp(-\gamma \|x_i - y_j\|^2), \gamma > 0 \quad (3)$$

并用 10 折交叉确认法(Cross Validation)确定适合本问题的径向基函数的参数。

SVM 是针对两类问题设计的,不能直接用于多类分类问题。本文采用多数投票准则,把多类分类问题转化为两类分类问题的方法。 M 类数据分类问题,对每个可能的类别对训练一个识别器,那么 M

类分类问题需要训练 $M \cdot (M-1)/2$ 个识别器。在识别过程中,统计每个识别器的分类结果,输入样本被划分为在识别结果中占多数的类别。

3 SVMs 集成和特征选择联合算法

SVMs 集成和特征选择联合算法的核心思想是在构建基础分类器的过程中加入对有效特征的搜索,那么最终的综合分类器的分类性能和所选择的有效特征子集包含的分类信息是密切相关的。Freund 和 Schapire 提出的 Adaboost 算法可以显著提高单个分类器的正确识别率^[8]。本文在 Adaboost 分类器集成算法的启发下,提出两种 SVMs 集成和特征选择联合算法,如图 1 所示。两个算法的输入相同,算法总体结构相同,但是算法核心部分——特征选择过程以及相应构建的基础分类器是不同的。

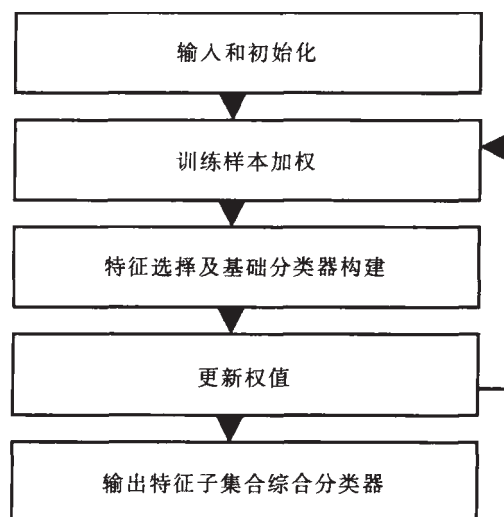


图 1 算法结构图

Fig.1 Structure of algorithm

两种算法的学习对象是一组 N 个 d 维的独立同分布的训练样本对

$X=\{(x_i, y_i) \mid x_i \in R^d, y_i \in \{1, 2, \dots, M\}, i=1, 2, \dots, N\}$, 这里 y_i 是类别标记,样本是有限的 M 类目标。如图 1 所示,两种联合算法都重复循环运行 T 次,在第 t 次循环($t=1, 2, \dots, T$),训练样本的分布参数为 D_t (分布参数的初始值设置为 $D_1(i)=1/N, i=1, 2, \dots, N$)。循环计算步骤如下:

(1) 训练样本加权实际是根据分布参数 D_t 对训练样本进行重抽样,重抽样的样本个数是 N 个,得到新的训练样本集 X_{t_i} ;

(2) SVMs 学习训练样本 X_t , 得到特征子集 x^* 以及相应的分类超平面 $h_t(x^*)$;

(3) 计算 $h_t(x^*)$ 的分类错误率: $\varepsilon_t = \sum_{i: h_t(x_i^*) \neq y_i} D_t(i)$, 如果 $\varepsilon_t \geq 0.5$ 或 $\varepsilon_t = 1$, 则设 $T = t - 1$ 并结束循环;

(4) 计算加权权值 $\beta_t = \varepsilon_t / (1 - \varepsilon_t)$; (4)

(5) 更新分布参数 D_t : 如果 $h_t(x_i^*) = y_i$, $D_{t+1} = \frac{D_t(i)}{Z_t}$ β_t ; 否则 $D_{t+1} = \frac{D_t(i)}{Z_t}$ 。 Z_t 是归一化常量。

最后利用加权多数投票法获得综合分类器:

$$h_{fin} = \operatorname{argmax}_{y_i \in \{1, 2, \dots, M\} : h_t(x) = y_i} \sum \lg \frac{1}{\beta_t}$$

(5)

两种 SVM 集成和特征选择联合算法在上述步骤(2)是不同的, 下面分别介绍。

第一种算法的特征选择和构建基础分类器的方法是: 给训练样本集 X_t 的每个特征训练一个分类器, 选择分类错误率最小的特征子集 x^* 及相应的分类超平面 $h_t(x^*)$ 。在这个算法中, 每次循环选择的特征子集只包含一个特征。

第二种算法的特征选择和构建基础分类器的方法是: 给训练样本集 X_t 的每个特征训练一个分类器, 利用排序法选择 l 个特征构成特征子集 x^* , 特征的个数 l 需要预先设置。对特征进行排序的评价函数是每个特征相应的分类器的分类正确率。然后, SVM 学习 l 得到分类超平面 $h_t(x^*)$ 。

综上所述, 在每次循环中, 错分样本获得的分布参数大, 正确分类的样本的分布参数小。因此, 算法不仅将分类算法的重点放在分类效果差的样本上, 而且同时也选择了对分类困难样本分类效果好的特征子集, 并根据本次循环得到的分类器分类性能影响下一个循环的分类器的分类性能以及特征子集的选取。这样可以很好的解决分类器过学习问题, 增

强综合分类器的推广能力。

4 实验和结果

4.1 实验所用数据

本文用实际舰船辐射噪声数据集和公共数据集来验证两种 SVM 分类器集成和特征选择联合算法的有效性。舰船目标数据集共分为 A、B、C 三类目标, 共 450 个样本 (150 × 3)。每个样本由 12 维波形特征、6 维包络特征、9 维小波分解特征共 27 维特征描述。公共数据是从 http://www.ics.uci.edu/ml/learn/ML_Repository.html 得到。实验中所用数据集的结构如表 1 所示。

表 1 实验中所用数据集

Table 1 Datasets for the experiments

数据集	特征数	类别数	训练样本数	测试样本数
舰船目标	27	3	300	450
Wine	13	3	100	178
Sonar	60	2	150	208
iris	4	3	100	150
diabetes	8	2	400	768

4.2 实验和结果

在实验中, 把两种联合算法应用于舰船目标的识别, 并与单个 SVM 的分类性能进行比较。每个算法重复运行 50 次, 实验结果为 50 次的平均值。实验结果列在表 2 中。

从对舰船目标的识别实验结果可以看到, 联合算法二不仅可以大幅度减少特征数目 (选择的特征数目是原特征数目的 30%), 而且比单个的 SVM 的分类正确率高 10%。算法一对特征的选择更加集中, 特征数目下降到原来特征数目的 20%, 虽然分类正确率比单个 SVM 有所下降, 但分类正确率仍然

表 2 实验结果

Table 2 Results of experiments

数据集	算法一选 择特征数	算法二选 择特征数	算法一分类正确率		算法二分类正确率		单个 SVM 分类正确率	
			训练集	测试集	训练集	测试集	训练集	测试集
舰船目标	5 ±1	9 ±1.3	0.820	0.811	0.981	0.967	0.886	0.869
Wine	7	3 ±1.2	0.952	0.922	0.995	0.951	0.98	0.972
Sonar	10 ±2	30 ±7	0.83	0.810	0.951	0.922	0.754	0.734
Iris	1	2 ±1	0.879	0.870	0.902	0.902	0.891	0.900
diabetes	5	2 ±1	0.752	0.754	0.714	0.723	0.700	0.708

保持在 80%以上。由于算法一选择的舰船特征数目比算法二更少,那么所得到的综合分类器的结构更简单,在线测试的时候,算法一可以比算法二节省更多时间。在实际的识别应用中,可以根据对时间和分类精度的实际要求,选择使用合适的算法。

从对公共数据的识别分类结果可以看到,算法二与单个 SVM 相比对 4 个公共数据集的分类正确率都有很大的提高,对测试集的正确分类率最少提高 1.5%(diabetes),最大提高 18.8%(sonar)。在构成正确分类率高的综合分类器的同时,算法二也选择了有效特征子集,各个数据集的特征数目大致减少为原来特征数目的 30%。算法一在显著降低特征数目的同时,正确分类率与单个 SVM 相比,wine 数据、iris 数据分别下降了 5%和 3%,sonar 数据、diabetes 数据分别提高了 7.6%和 4.6%。比较算法一和算法二可以知道,算法一更适用于特征高度冗余的情况。

5 结 论

舰船目标的特征选择和分类器设计对舰船目标识别是至关重要的。本文提出两种 SVM 分类器集成和特征选择联合算法,并和单个 SVM 的分类性能进行比较。对实测三类舰船目标和公共数据的识别实验结果证明了,两种算法都可以应用于水声目标识别问题。算法一比算法二更适合用于特征高度冗余的情况,算法二不仅可以选择有效特征,还可以得到分类精度高、推广性能好的综合分类器。

参 考 文 献

[1] 丁玉薇. 被动声纳目标识别技术的现状与发展[J]. 声

学技术, 2004, 23(4): 253-257.

DING Y W. Review on passive sonar target recognition [J]. Technical Acoustics, 2004, 23(4): 253-257.

[2] 樊养余, 孙进才, 李平安, 等. 基于高阶谱的舰船辐射噪声特征提取[J]. 声学学报, 1999, 24(6): 611-616.

FAN Y Y, SUN J C, LI P A. et al. Feature extraction of ship-radiated noise using higher-order spectrum [J]. Acta acoustic, 1999, 24(6): 611-616.

[3] 蔡悦斌, 张明之, 史习智, 林良骥. 舰船噪声波形结构特征提取及分类研究[J]. 电子学报, 1999, 27(6): 129-130.

CAI Y B, ZHANG M Z, SHI X Z et al. Extraction and classification of ocean acoustic signals based on wave structure[J]. Acta electronica sinica, 1999, 27(6): 129-130.

[4] Elshaikl T S, Wacker A G. Effect of dimensionality and estimation on the performance of Gaussian classifiers[J]. IEEE Trans PAMI, 1980, 2(12): 187-198.

[5] YANG H H, SUN C J. Feature selection for multi-class underwater acoustic targets via SVMs and genetic algorithms. Proceeding of international symposium on intelligence computation and its applications[C]. Wuhan: China university of geosciences of press, 2005. 397-401.

[6] Vapnik V N. The nature of statistical learning theory[M]. John Wiley, New York, 1998.

[7] Burges C. A tutorial on support vector machines for pattern recognition[J]. Data mining and knowledge discovery, 1998, 2(2): 121-167.

[8] Yoav Freund, Robert E S. Experiments with a new boosting algorithm[C]. In Proc. of the 13th international conference on machine learning. Morgan Kaufmann, 1996, 148-156.