

用于水声目标识别的近邻无监督特征选择算法

陈含露, 杨宏晖, 申 昇

(西北工业大学航海学院, 陕西西安 710072)

摘要: 针对水声目标数据的特征冗余问题, 提出一种新的近邻无监督特征选择算法。首先利用顺序向后特征搜索算法生成原始特征集的子集, 然后利用基于代表近邻选取方法的特征评价机制评价特征子集的优越性。使用实测水声目标数据集和声呐数据集进行特征选择和分类实验, 在保持支持向量机平均分类正确率几乎不变的情况下, 特征数目分别降低了 90%和 75%。结果表明, 该算法选择出的特征子集, 在去除冗余特征后有效地提高了后续学习算法的效率。

关键词: 水声目标识别; 无监督; 特征选择; 代表近邻

中图分类号: TB533

文献标识码: A

文章编号: 1000-3630(2016)-03-0204-04

DOI 编码: 10.16300/j.cnki.1000-3630.2016.03.003

Neighbor based unsupervised feature selection algorithm for underwater acoustic target recognition

CHEN Han-lu, YANG Hong-hui, SHEN Sheng

(Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China)

Abstract: The problem of feature redundancy in underwater target recognition has been studying by plenty of researchers. In this paper, a new neighbor based unsupervised feature selection algorithm is proposed. Primarily, the subsets of the original feature set extracted from the dataset are produced by using backward feature searching strategy. Subsequently, these feature subsets are evaluated with the assessment mechanism based on the representative neighbors choosing method. Results of classification experiments with actual measured underwater acoustic target dataset and sonar dataset after feature selection show that the accuracies of SVM classifiers remain almost the same when the numbers of features are decreased by 90% and 75%, which indicates that the proposed method improves the efficiency of subsequent learning algorithm with the redundant features removed.

Key words: underwater acoustic target recognition; unsupervised; feature selection; representative neighbors

0 引言

为提高水声目标识别的正确率, 研究人员往往用多种方法提取水声目标辐射噪声的多域特征。然而, 水声目标样本获取的代价却很大。因此, 要在水声目标样本数目保持不变的前提下达到分类正确率损失尽可能小的目的, 进行特征选择以去除不相关和冗余特征, 在水声目标识别任务中具有重要的意义。

根据用于特征选择的数据有无类标, 可将特征选择方法分为有监督方法和无监督方法。在水声目标识别领域, 相比于有监督特征选择方法^[1]的趋于

成熟, 无监督特征选择方法^[2]仍有待深入研究。有监督特征选择方法通常通过类标的指导来评价特征与识别任务的相关程度。而无监督特征选择由于缺少类标指导, 往往倾向于选出能够保留样本内在聚类属性的特征。文献中大多现有无监督特征选择算法依赖于距离矩阵来寻找最优特征子集^[3-5], 近年来从样本近邻方面考虑的非参数方法为特征选择提供了新的思路^[6-8]。但这些方法往往面临两方面的问题: (1) 通常选择近邻是为了观察某个特征子集判别样本是否属于同一聚类的能力, 因此要求所选近邻必须对不同的特征具敏感性——保证样本与最近近邻属于同一聚类, 与最远近邻属于不同聚类; (2) 通常要求根据先验知识设定近邻数目, 若要求设定的近邻数目逼近样本数目且样本数目较大, 则将导致算法的计算量过大。本文研究了一种基于特征顺序搜索算法和代表近邻选取方法^[9]的非参数无监督特征选择方法——近邻无监督特征选

收稿日期: 2015-05-21; 修回日期: 2015-09-17

基金项目: 水声对抗技术重点实验室开放基金

作者简介: 陈含露(1991—), 女, 浙江丽水人, 硕士研究生, 研究方向为模式识别、声信号处理。

通讯作者: 杨宏晖, E-mail: hhyang@nwpu.edu.cn

择(Neighbor Based Unsupervised Feature Selection, NBUFS)算法, 其中代表近邻的选取机制可克服上述两个问题, 并利用实测水声目标数据集和声呐数据集的多域特征进行了特征选择和分类实验, 结果证明本文算法能够较好地解决水声目标特征选择问题。

1 近邻无监督特征选择算法

1.1 算法原理

NBUFS 算法的原理如图 1 所示。对于数据集 X , 首先计算两两样本间的欧式距离, 得到相异度矩阵。接着将相异度矩阵输入聚类倾向视觉评估算法(Visual Assessment of cluster Tendency, VAT)^[10-11], 由输出的重组图像评估聚类数目。然后利用 K-means 算法对数据集进行聚类, 并选择出基于聚类结果的代表近邻。最后采用封装模型进行特征选择, 得到结果子集, 其中封装模型的特征评价函数是基于代表近邻设计的。

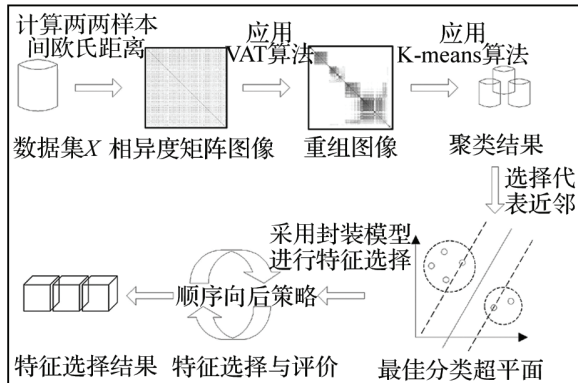


图1 NBUFS 算法原理

Fig.1 The procedure of NBUFS algorithm

1.2 代表近邻的选取方法

代表近邻主要用于后续封装模型特征评价函数的设计, 其选取步骤如下:

(1) 假设有输入数据集 $X = \{x_i\}_{i=1}^m, x_i \in R^n$, 其中 m 为数据集的样本数目, n 为数据集的特征数目。又有 $C = \{c_1, \dots, c_k, \dots, c_M\}$, 其中 c_k 表示利用 1.1 节中提到的 VAT 算法和 K-means 算法获得的第 k 类样本。对于样本 $x_i \in c_k$, 选择两个最近邻: 一个与 x_i 处于同一类中, 另一个与 x_i 处于不同类中。现定义如下:

$$WN(x_i) = \underset{x_r \in c_k}{\operatorname{argmin}} D(x_i, x_r) \quad (1)$$

$$BN(x_i) = \underset{x_r \notin c_k}{\operatorname{argmin}} D(x_i, x_r) \quad (2)$$

其中: $r=1, \dots, m$, $D(\cdot, \cdot)$ 用于计算向量间的欧氏距离。称 $WN(x_i)$ 为 x_i 的类内最近邻, 它实际上是除 x_i 外与 x_i 处于同一类中且与 x_i 距离最近的一个样本; 称 $BN(x_i)$ 为 x_i 的类外最近邻, 它实际上是和 x_i 处于不同类中且与 x_i 距离最近的一个样本。如此, 则可得两个近邻集: $\{WN(x_i)\}_{i=1}^m$ 和 $\{BN(x_i)\}_{i=1}^m$ 。

(2) 将 $\varepsilon(x_i, WN(x_i))$ 的类标设置为 -1 , 将 $\varepsilon(x_i, BN(x_i))$ 的类标设置为 $+1$, 用于训练支持向量机(Support Vector Machine, SVM), 其中 $i=1, \dots, m$, $\varepsilon(\cdot, \cdot)$ 用于计算向量各元素间的曼哈顿距离绝对值。例如 $\varepsilon([0.3, 0.6]^T, [0.9, 0.7]^T) = [0.6, 0.1]^T$ 。

(3) 由步骤(2)中的 SVM 训练结果得到分类超平面的特征向量 w :

$$w^T \cdot \varepsilon(x_i, WN(x_i)) + w_0 \leq -1 \quad (3)$$

$$w^T \cdot \varepsilon(x_i, BN(x_i)) + w_0 \geq +1 \quad (4)$$

(4) 定义代表样本 $x_{\text{repr}} \in \{x_1, \dots, x_m\}$ 为满足如下条件的样本:

$$\underset{i \in \{1, \dots, n\}}{\operatorname{argmin}} (w^T \cdot \varepsilon(x_i, BN(x_i)) - w^T \cdot \varepsilon(x_i, WN(x_i))) \quad (5)$$

得到代表样本后, 定义类内代表近邻为 $WN(x_{\text{repr}})$, 类外代表近邻为 $BN(x_{\text{repr}})$ 。

1.3 顺序向后特征搜索算法

顺序向后特征搜索算法由特征全集开始, 在其后的每一次迭代中去掉一个特征(每次迭代中去掉该特征时得到的特征子集评价价值比去掉其它任何一个特征时得到的评价价值都高), 直到特征数目减少到规定数目为止, 算法流程见图 2。

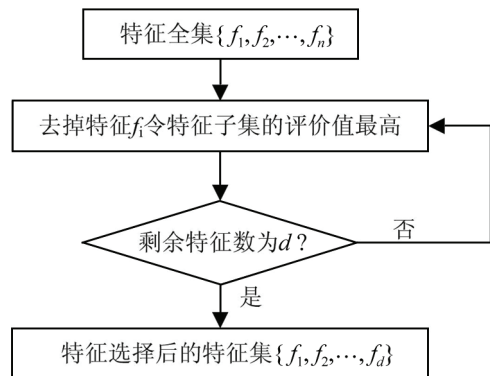


图2 顺序向后特征搜索算法流程图

Fig.2 The procedure of backward feature search strategy

1.4 封装式特征选择模型

采用封装模型进行特征选择是本文算法的核心部分, 其具体步骤如下:

(1) 开始: 运行顺序向后特征搜索算法。

(2) 评价初始化: 对于用待评价特征子集描述

的数据集 X , 和近邻集 $\{WN(x_i)\}_{i=1}^m, \{BN(x_i)\}_{i=1}^m$, 将它们以相同的方式分为 k 部分, 每部分所含样本数目相同。将其中的 $k-1$ 个部分作为 X_{train} , 剩下的 1 个部分作为 X_{valid} 。

(3) 训练: 按 1.2 节中的方法训练 X_{train} 得到超平面的特征向量 w , 并利用方程(5)得到 $WN(x_{\text{repr}})$ 和 $BN(x_{\text{repr}})$ 。此时将方程(3)和方程(4)改写为:

$$w^T \cdot \varepsilon(x_i \in X_{\text{train}}, WN(x_i) \in X_{\text{train}}) + w_0 \leq -1 \quad (6)$$

$$w^T \cdot \varepsilon(x_i \in X_{\text{train}}, BN(x_i) \in X_{\text{train}}) + w_0 \geq +1 \quad (7)$$

(3) 测试: 将特征子集的评价值记为 Acc , 它可由下式计算得到:

$$\frac{1}{|X_{\text{valid}}|} \sum_{i=1}^m I[(w^T \cdot \varepsilon(x_i \in X_{\text{valid}}, WN(x_i) \in X_{\text{valid}}) \leq w^T \cdot \varepsilon(x_{\text{repr}}, WN(x_{\text{repr}}))) \quad (8)$$

$$\text{and}(w^T \cdot \varepsilon(x_i \in X_{\text{valid}}, BN(x_i) \in X_{\text{valid}}) \geq w^T \cdot \varepsilon(x_{\text{repr}}, BN(x_{\text{repr}}))))]$$

其中, $I[\]$ 的输出分两种: 当 $[\]$ 内的条件满足时为 1, 不满足时为 0。 $|X_{\text{valid}}|$ 表示 X_{valid} 的样本数目。

(5) 循环: 重复步骤(3)~(4), 直到步骤(2)中所述的 k 部分数据都曾作为测试集。

(6) 评价完成: 将以上步骤得到的 k 个评价值的算术平均作为特征子集的最终评价值。

(7) 循环: 重复步骤(2)~(6), 直到顺序向后特征搜索算法运行结束。

2 实验结果及分析

2.1 实验数据

利用实测水声目标数据集和加州大学用于机器学习的 UCI(University of California Irvine, UCI) 数据库中的声呐数据集对本文算法的性能进行验证实验, 数据说明如表 1 所示。

表 1 数据说明
Table 1 Data specifications

数据	特征数目	类别数目	各类样本数目
实测水声目标数据集	71	4	(120,120,120,120)
声呐数据集	60	2	(72,66)

2.2 NBUFS 算法性能验证

采用如下两种方式验证 NBUFS 算法的有效性: (1) 支持向量机(Support Vector Machine, SVM) 的分类结果; (2) 样本的空间分布。

2.2.1 SVM 分类结果

分别用上述两种数据对 NBUFS 算法的特征选

择结果进行 SVM 分类实验, 采用 10 次 10 折交叉验证 SVM 运行结果的分类正确率平均值作为最终的分类正确率。结果得到选择出的特征数目与 SVM 分类正确率的关系如图 3 所示, 特征选择前后所需的分类时间如表 2 所示。

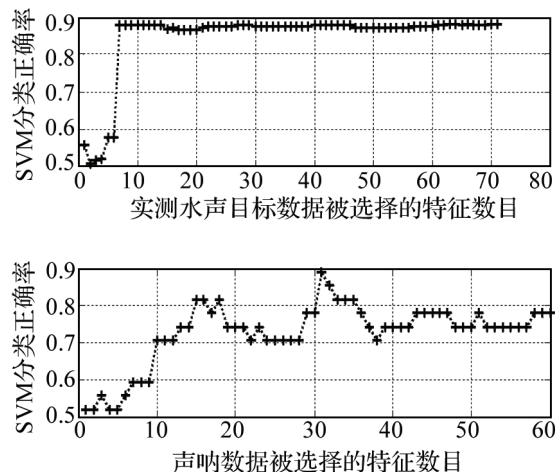


图 3 被选择的特征数目与 SVM 分类正确率的关系

Fig.3 Relationship between number of features selected and SVM classification accuracy

表 2 特征选择前后所需的分类时间

Table 2 Classification time needed before and after feature selection

数据	特征选择前 分类时间/s	特征选择后 分类时间/s
实测水声目标数据集	2.688	1.121
声呐数据集	0.015	0.009

由图 3 可以看出, 两种数据的特征数目与 SVM 分类正确率关系曲线的变化趋势相似: 开始时 SVM 分类正确率总体上随特征数目的增加而增加, 当特征达到一定数目后, 分类正确率趋于相对稳定。实测水声目标数据使用 7 个特征即可使分类正确率与使用完全数据集的 71 个特征时相当; 声呐数据使用 15 个特征即可使分类正确率与使用完全数据集的 60 个特征时相当。这说明使用 NBUFS 算法进行特征选择以后, 使用部分特征就可以表征完全数据集的全部分类信息。

由表 1 可以看出, 使用经特征选择得到的特征子集进行 SVM 分类实验, 可以在较大程度上减少分类时间。

因此, 使用本文方法进行特征选择后, 在不牺牲分类正确率的前提下, 有效提高了后续学习算法的速度, 提高了计算效率。

2.2.2 样本的空间分布

对于实测的水声目标数据集, 使用 NBUFS 算法选择出最佳二维特征(记为特征 1 和特征 2)和最

差两维特征(记为特征3和特征4),分别在二维平面绘制该数据在最佳两维特征和最差两维特征表示下的4类样本散布图,如图4和图5所示。另外由于声呐数据本身分类性能较差,使用任意两维特征,甚至三维特征均无法很好地区分不同类样本,而高于三维特征表示的样本散布图又无法用图形表示,因此本节仅以实测水声目标数据为例进行阐明。

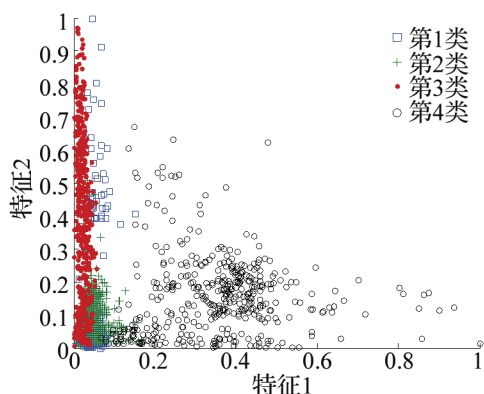


图4 实测水声目标数据最佳两维特征表示的样本散布图
Fig.4 Samples of actually measured underwater acoustic target dataset described by the best 2 features

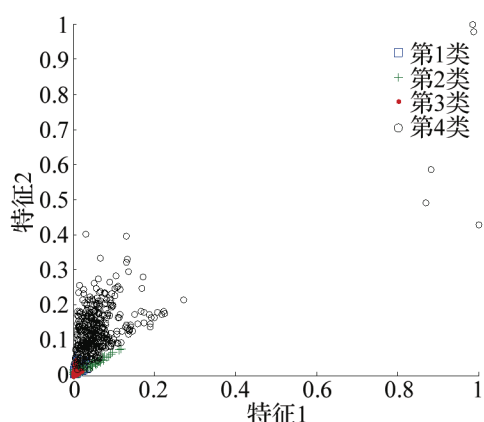


图5 实测水声目标数据最差两维特征表示的样本散布图
Fig.5 Samples of actually measured underwater acoustic target dataset described by the worst 2 features

通过对比图4和图5可以看出,最佳两维特征表示的4类样本在平面中较好地分布在不同区域,可分性较好;而最差两维特征的4类样本混在一起,可分性较差。这说明本文特征选择算法对特征分类性能的评价是可靠的。

综上所述,本文提出的NBUFS算法能有效地选择水声目标优化特征子集,在一定程度上解决水声目标识别的问题。

3 结论

本文提出的近邻无监督特征选择算法(NBUFS),利用顺序向后特征搜索算法生成原始特征集的子集,并利用基于代表近邻选取方法的特征评价机制评价特征子集的优越性。使用实测水声目标数据和声呐数据对其进行实验验证的结果表明:本文算法能够准确地选择出优秀的特征子集,在分类实验中获得较高的分类正确率,并减少分类时间,能较好地解决水声目标多域特征选择问题。

参考文献

- [1] 杨宏晖,王芸,孙进才,等.融合样本选择与特征选择的 AdaBoost 支持向量机集成算法[J]. 西安交通大学学报, 2014, 48(12): 63-68.
YANG Honghui, WANG Yun, SUN Jincan, et al. An Adaboost support vector machine ensemble method with integration of instance selection and feature selection[J]. Journal of Xi'an Jiaotong University, 2014, 48(12): 63-68.
- [2] 申昇,杨宏晖,袁帅.用于水声目标识别的互信息无监督特征选择[J]. 声学技术, 2013, 32(6): 30-33.
SHEN Sheng, YANG Honghui, YUAN Shuai. Mutual information unsupervised feature selection for underwater acoustic targets[J]. Technical Acoustics, 2013, 32(6): 30-33.
- [3] Zhao Z, Liu H(2007)Spectral feature selection for supervised and unsupervised learning[C]//Proceedings of the 24th international conference on machine learning(ICML), ACM, New York, 1151-1157.
- [4] Ng A, Jordan M, Weiss Y. (2001) On spectral clustering: analysis and an algorithm[C]//Proceedings of advances in neural information processing systems (NIPS), vol 14, MIT Press, Cambridge, 2001, 849-856.
- [5] Mitra P, Murthy C a, Pal S K. Unsupervised Feature Selection Using Feature Similarity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI, 2002, 24(3): 301-312.
- [6] Yang L, Jin R, Mummert L, et al. A boosting framework for visibility-preserving distance metric learning and its application to medical image retrieval[J]. IEEE Trans. Patt. Anal. Mach. Intell, 2010, 32(1): 30-44.
- [7] Yan R, Zhang J, Yang J, et al. A discriminative learning framework with pairwise constraints for video object classification[J]. IEEE Trans. Patt. Anal. Mach. Intell, 2006, 28(4): 578-593.
- [8] Chen C H. Feature selection for clustering using instance-based learning by exploring the nearest and farthest neighbors[J]. Information Sciences, 2015, 318(3): 14-27.
- [9] CHEN C H. A semi-supervised feature selection method using a non-parametric technique with side information[J]. Information Sciences, 2013, 39(3): 359-371.
- [10] Havens T C, Bezdek J C. An Efficient Formulation of the Improved Visual Assessment of Cluster Tendency (iVAT) Algorithm[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(5): 813-822.
- [11] Bezdek J C, Hathaway R J. VAT: A Tool for Visual Assessment of (Cluster)tendency[C]//Proceedings of the International Joint Conference on Neural Networks, 2002, 2225-2230.