

基于选择性聚类集成的图像目标分类方法

储岳中, 刘 恒, 张学锋, 潘 祥

(安徽工业大学 计算机科学与技术学院, 安徽 马鞍山 243002)

摘 要: 传统集成学习算法是对所有个体分类器进行组合决策, 由于无法反映个体分类器的差异性, 不能有效提高集成分类器的识别率. 为此, 提出基于互信息(Normalized Mutual Information, NMI)的个体分类器差异性度量方法, 利用匈牙利算法对个体分类器的标记向量进行匹配, 在此基础上提出基于成分数据的 AP (Affinity Propagation) 聚类集成算法作为选择性集成策略. 在遥感图像上分别与经典算法做比较实验, 结果表明此算法在分类性能上具有一定的优越性.

关键词: 聚类集成; 匈牙利算法; 近邻传播; 图像目标分类

中图分类号: TP391.4

文献标识码: A

文章编号: 1000-7180(2017)11-0058-05

Image Target Classification Method Based on Selective Cluster Aggregation

CHU Yue-zhong, LIU Heng, ZHANG Xue-feng, PANG Xiang

(School of Computer Science, Anhui University of Technology, Ma'anshan 243002, China)

Abstract: The traditional ensemble learning algorithm was a combination of the individual classifiers, which couldn't reflect the differences of individual classifiers and effectively improve the recognition rate of ensemble classifier. A measurement method based on Normalized Mutual Information was proposed, and the Hungarian algorithm was used to match the labeled vectors of individual classifiers. On the basis of that, a AP (Affinity Propagation) clustering integrated strategy was proposed based on component data. Compared with the classical algorithm on the remote sensing image, the results showed that the proposed algorithm has certain advantages in classification performance.

Key words: cluster aggregation; Hungarian algorithm; affinity propagation; image target classification

1 引言

研究表明, 解决实际分类问题时很难找到一个普适的分类器. 集成学习是近几年来被提出的一种分类器构建策略, 它将多个个体分类器融合起来, 以提高分类器的泛化性能和分类精度. 但随着个体分类器数目的增多, 对系统的存储要求越来越高, 计算开销也随之加大. 针对如何构建具有一定差异性的个体分类器和如何设计选择策略, 国内外学者提出了很多代表性的方法^[1-10]. 韩敏^[5]等采用核极端学

习机作为基学习器, 引入相关性和冗余性描述准确性和差异性, 提出了基于互信息的选择性集成学习算法. 余晓东^[6]等采用样本扰动和参数扰动的二重扰动策略产生子学习机, 并利用多数投票法融合识别结果, 提出了基于直觉模糊核匹配追踪集成的分类器学习算法. 孟军^[8]等针对高维数据, 采用反序聚合技术对数据特征进行过滤, 采用多数投票法进行融合集成, 提出了一种基于近邻传播聚类的集成特征选择方法. Koen^[10]分别以 Bagging 方法、随机子空间方法和两种方法的组合来构造基分类器, 在此

收稿日期: 2017-01-12; **修回日期:** 2017-03-01

基金项目: 国家自然科学基金(61602007); 安徽高校自然科学研究重点项目(KJ2016A085); 安徽高校自然科学研究重大项目(KJ2015ZD09)

基础上设计出不同的基于扩展加性模型(Generalized additive models, GAMs)的集成分类算法。

本文利用聚类方法构造多个个体分类器,基于互信息测度进行个体分类器差异性度量,在此基础上进行选择性的聚类集成,所构造集成分类器用于遥感图像分类,与现有相关算法相比,本文算法在分类成功率上具有一定的优越性。

2 聚类集成

2.1 聚类集成基本结构

聚类集成是在集成学习的基础上,将数据集的多个不同聚类结果进行重新组合形成最终聚类结果的无监督分类方法,适合解决一些无法事先确定数据集类别的问题。从现有文献看,大多学者采用的方法是将集成学习的学习机构造和组合器构造均用聚类的方法实现。聚类集成过程可表达为^[11]:假设有 n 个数据集 $X = \{x_1, x_2, \dots, x_n\}$, $x_i \in R^d$, 对数据集 X 利用 H 次聚类算法得到 H 个聚类结果 $\{\pi_1, \pi_2, \dots, \pi_H\}$ (我们称之为聚类成员), 其中 π_i ($i = 1, 2, \dots, H$) 为第 i 次聚类算法得到的聚类结果, 最后设计一个共识函数对这 H 个聚类结果进行合并, 得到最终的聚类结果。常见聚类集成结构如图 1 所示。

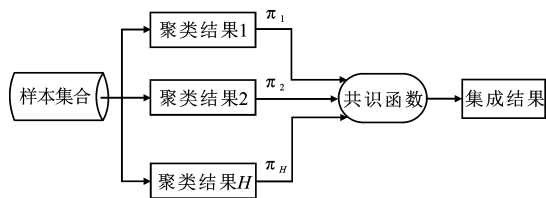


图 1 聚类集成结构

与单一聚类算法相比,聚类集成的划分结果的平均性能表现更为优越,能够部分克服某些由聚类算法本身所造成的参数敏感问题,并且能够从集成分布中评估聚类的不确定性。通过聚类集成设计无监督分类器要解决两个关键问题:(1)如何生成多个具有一定差异性的聚类成员,当前主要方法有不同算法、相同算法的不同输入参数、不同数据子集、不同特征子集或子空间投影等;(2)如何设计有效的决策函数(共识函数)对聚类成员进行合并,当前主要共识函数设计方法有关联矩阵法、投票法、信息论法、超图法和各种混合模型法等。

2.2 聚类个体差异性度量

众多研究表明,聚类集成学习要求输入集成学习的聚类成员要有一定的差异性,同时又要有一定的精确度。传统的聚类集成算法往往是利用得到的

所有聚类个体通过相关算法获得最终的集成结果,而选择性聚类集成是从中选择质量和差异性折中的个体聚类,以期获得更好的集成结果。为了评价聚类个体间的差异性,本文提出基于互信息(Normalized Mutual Information, NMI)的方法^[12],并将差异性和精确性结合起来构造聚类个体选择策略,为聚类个体的选择提供依据。

给定 π_a, π_b 两个聚类个体,两个划分之间的互信息定义为^[12]

$$NMI(\pi_a, \pi_b) = \frac{\sum_{h=1}^{k_a} \sum_{l=1}^{k_b} n_{hl} \log_2 \left[\frac{N \cdot n_{hl}}{n_h \cdot n_l} \right]}{\sqrt{\left[\sum_{h=1}^{k_a} n_h \log_2 \frac{n_h}{N} \right] \left[\sum_{l=1}^{k_b} n_l \log_2 \frac{n_l}{N} \right]}} \quad (1)$$

式中, k_a 和 k_b 分别为聚类 π_a 和 π_b 的类别数; n_{hl} 为同时在 π_a 中的第 h 个类以及 π_b 中的第 l 个类的数据点的个数; n_h 为聚类 π_a 中第 h 类中的数据点个数; n_l 为聚类 π_b 中第 l 类中的数据点个数; N 为数据集的样本数。互信息是信息论中的一个重要概念,用来衡量两个分布间的统计相关性,它为成对聚类结果之间的互相关性提供了一个合理的度量。互信息的取值范围为 $[0, 1]$ 。显然,当 π_a 和 π_b 等价时, $NMI(\pi_a, \pi_b) = 1$; 当 π_a 不包含任何 π_b 的信息或 π_b 不包含任何 π_a 的信息时, $NMI(\pi_a, \pi_b) = 0$ 。若要考虑测量一个聚类集体之间的差异性,可以使用所有聚类对之间的平均 NMI 值来表示,如下式所示。

$$NMI_{\text{平均}} = \frac{2}{L(L-1)} \left(\sum_{i=1}^{L-1} \sum_{j=i+1}^L NMI(\pi_i, \pi_j) \right) \quad (2)$$

式中, L 为聚类集体中成员个数。由于 $NMI_{\text{平均}}$ 的值越大差异性越小,为此一个基于互信息的差异性测度 Div 定义如下:

$$Div = 1 - NMI_{\text{平均}} \quad (3)$$

这样, Div 越大,则两个聚类个体之间的差异性越大,当 Div 为 0 时,两个聚类是等价的。

3 基于成分数据的 AP 聚类集成算法

3.1 数据转换

对于 N 个聚类个体,我们任意选择一个为参考向量 C_r , 对其余的所有非参考向量,通过匈牙利算法与参考向量做匹配 (C_r, C_i) ($i = 1, 2, \dots, N, i \neq r$), 使得所有聚类个体的标记一致。

对于数据集中的每个样本数据 X_i , 以 N 个聚类个体为依据,统计每个标记所占的比例,得到一个

成分向量记为 $\mathbf{V}_i = (V_{i1}, V_{i2}, \dots, V_{iL})$, L 为类别数,

$\mathbf{V}_i$ 满足 $\sum_{j=1}^L V_{ij} = 1$ 。数据具有不适定性, 不利于利用常规方法对成分向量所构成的向量集进行聚类, 为此, 对 L 维成分数据向量 $\mathbf{V}_i = (V_{i1}, V_{i2}, \dots, V_{iL})$ 采用中心化对数比做进一步转换, 将 $\mathbf{V}_i$ 映射为另一个空间中的数据点 Z_i , 即:

$$Z_{ij} = \ln \frac{V_{ij}}{g(\mathbf{V}_i)} (j = 1, 2, \dots, L) \quad (4)$$

式(4)中, $g(\mathbf{V}_i) = \exp\left(\frac{1}{L} \sum_{j=1}^L \ln V_{ij}\right)$ 。该变换是空间 $\mathbf{V}_i$ 到 R^L 空间的一一映射, 因此, 向量 $\mathbf{Z}$ 可在 L 维空间 R^L 中任意取值。这种中心化对数比变换使得聚类样本数据不再受成分数据不适定性的约束。对新得到的样本集 Z_i , 随后可以应用一些经典算法对其进行聚类, 为聚类集成提供差异性的划分结果。

3.2 算法步骤

综上所述, 本文提出基于成分数据的 AP 聚类集成算法 (AP Clustering Ensemble based on Composition Data, APECD) 步骤如下。

输入: 样本集合 X , 样本类别数 L , 参数 M 和 K 。

输出: 每个样本的类别属性, 分类成功率。

1. 重复运行 M 次 K -均值聚类算法, 产生 M 个聚类个体成员, 组成聚类个体集合, 记为 Sub 。每次聚类时随机产生 L 个聚类中心。

2. 利用 NMI 测度进行聚类个体间差异性度量, 两聚类个体 C_i 与 C_j 间的距离为 $d(C_i, C_j)$, 利用式(3)所描述的度量值来计算任意两个聚类个体之间的距离 d , 并计算每个聚类个体与其余所有聚类个体间的平均距离, 选择 H 个平均距离大的聚类个体组成集合 Sub' 。

3. 在 Sub' 集合中任意选择一个聚类个体为参考向量, 采用匈牙利算法将其余聚类个体与参考向量进行类别标签匹配;

4. 对 Sub' 集合所对应的样本集按式(4)进行中心化对数比转换, 形成新的样本集 Z ;

5. 应用近邻传播 (Affinity Propagation, AP) 算法对样本集 Z 进行聚类, 得到样本的聚类集成结果, 作为样本数据的分类器;

6. 对样本数据的实际分类结果和聚类集成结果利用匈牙利算法进行类别标签匹配, 使用 Micro-Precision 测度进行准确性评价。

注: Micro-Precision 测度定义如下^[13]:

$$P_{M-P} = \frac{1}{N} \sum_{i=1}^L a_i \quad (5)$$

式中, L 为聚类集成结果中类别的个数; N 为数据集总样本数; a_i 是第 i 个类中分类正确的数据点数。

图 2 给出了算法的流程图。

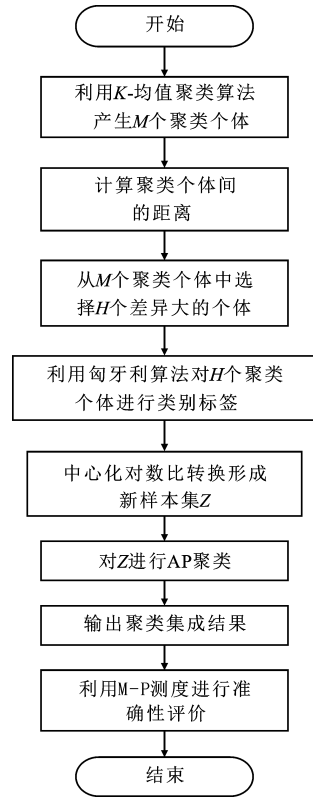


图 2 APECD 算法流程图

4 实验与分析

图 3 为华盛顿地区 SAR 图像中分割出来的 12 座桥梁目标。要识别 12 座桥梁目标, 需要补充样本, 为此, 对原始 SAR 图像每隔 6° 旋转一次, 采用双立方插值, 以此模拟现实中从不同角度成像得到的 SAR 图像^[14]。对每次旋转得到的 SAR 图像进行手工分割, 共获取 720 幅桥梁图像, 子图像大小为 50×50 , 样本集中包括每座桥图像各 60 幅, 再增加伪目标图像 60 幅, 样本集合计 780 幅图像。在目标识别前, 首先利用 Radon 变换对其进行特征提取, 然后再对 Radon 变换结果进行奇异值分解, 从大到小取前 25 个奇异值作为各图像样本的特征向量。

首先, 通过 K -均值聚类算法 (K 取 12), 随机产生聚类中心, 产生 100 个聚类个体数。然后采用互信息 (NMI) 的方法来度量聚类个体的差异性, 利用本文所提算法 (APECD) 进行选择集成实验, 被选择的聚类个体数范围为 $10 \sim 100$ 。训练结束后再在测试集上进行行测试。选用 Bagging (BAG)、Adaboost (ADA)、随机子空间方法 (RSM) 等经典算法与本文算法 (APECD) 做比较, 以测试正确率作为评价标准。

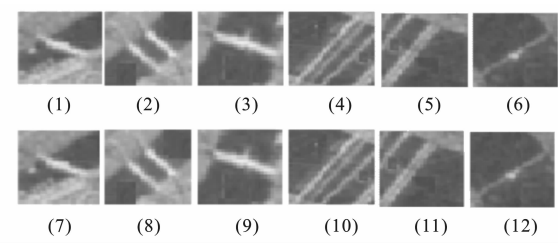


图 3 12 座桥梁目标图像

表 1 给出了全集成算法(BAG、ADA 和 RSM)和选择性聚类集成算法针对桥梁图像的测试正确率对比.图 4 给出了选择不同聚类个体数目时桥梁图像的测试正确率对比图,从图中可看出,相对全集成算法,选择性聚类集成算法的测试正确率有所提高,本文所提算法在两个数据集上的分类效果均要好于经典算法.

表 1 全集成算法和选择性聚类集成算法测试正确率对比

数据集	算法	聚类个体规模	测试正确率
桥梁 图像	BAG	100	0.819 5
	ADA	100	0.841 0
	RSM	100	0.837 7
	选择性 集成算法 (APECD)	10	0.823 5
		20	0.825 2
		30	0.828 7
		40	0.831 2
		50	0.832 4
		60	0.847 8
		70	0.853 9
		80	0.835 6
		90	0.828 0
		100	0.829 1

5 结束语

集成学习是近年来国内外学者的研究热点问题,全集成学习是对个体分类器的集体整合,由于各个个体分类器的差异性没有体现出来,集成效果往往达不到预期.为此,本文通过互信息进行个体分类器差异性度量,利用匈牙利算法对个体分类器的标记向量进行匹配,在此基础上提出基于成分数据的 AP 聚类集成算法.在遥感图像上分别与经典算法比较实验结果表明,本文算法在分类性能上具有一定的优越性.但本文算法的最好分类效果受聚类个体规模影响,目前只能通过搜索的方法来确定某一

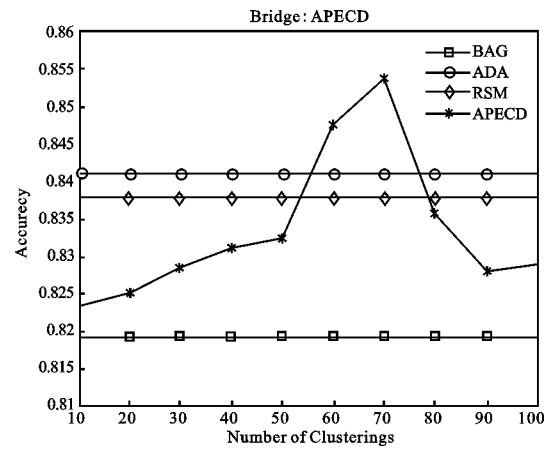


图 4 选择不同个体数目时桥梁图像测试正确率

数据集的最佳聚类个体规模,分类精度是以算法计算负担为代价的,当数据规模很大时,需对算法进一步优化,这将是下一步需研究的内容.

参考文献:

[1] ZHOU Z H, WU J, TANG W. Ensembling neural networks; Many could be better than all[J]. Artificial Intelligence,2002, 137(1/2): 239-263.

[2] ROKACH Lior. Collective-agreement-based pruning of ensembles[J]. Computational Statistics and Data Analysis ,2009, 53(4): 1015-1026.

[3] 姚旭,王晓丹,张玉玺,等. 基于 AdaBoost 和匹配追踪的选择性集成算法[J]. 控制与决策,2014, 29(2): 208-214.

[4] 杨长盛,陶亮,曹振田,等. 基于成对差异性度量的选择性集成方法[J]. 模式识别与人工智能, 2010, 23(4): 565-571.

[5] 韩敏,吕飞. 基于互信息的选择性集成核极端学习机[J]. 控制与决策,2015, 30(11): 2089-2092.

[6] 余晓东,雷英杰,宋亚飞,等. 基于直觉模糊核匹配追踪集成的目标识别方法[J]. 通信学报, 2015, 36(10): 165-171.

[7] YU G X, ZHANG G J, YU Z W, et al. Semi-supervised ensemble classification in subspaces[J]. Applied Soft Computing,2012, 12(5): 1511-1522.

[8] 孟军,尉双云. 基于近邻传播聚类的集成特征选择方法[J]. 计算机科学,2015, 42(3): 241-245.

[9] ZHNAG L, ZHOU W D. Sparse ensembles using weighted combination methods based on linear programming [J]. Pattern Recognition ,2011, 44(1): 97-106.

[10] KOEN W, DE B, KRISTOF C, et al. Ensemble classification based on generalized additive models[J]. Computational Statistics and Data Analysis ,2010, 54(6):

1535-1546.

[11] STREHL A,GHOSH J. Cluster ensembles-a knowledge reuse framework for combining multiple partitions[J]. The Journal of Machine Learning Research, 2002,3(12):583-617.

[12] 罗会兰,孔繁胜,李一啸. 聚类集成中的差异性度量研究[J]. 计算机学报,2007,30(8):1315-1324.

[13] 毕凯,王晓丹,邢雅琼. 基于模糊测度和证据理论的模糊聚类集成方法[J]. 控制与决策,2015,30(5):823-830.

[14] 王磊,周乐因,姬红兵,等. 一种面向信号分类的匹配追踪新方法[J]. 电子与信息学报,2014,36(6):1299-1306.

作者简介:

储岳中 男,(1971-),博士,副教授. 研究方向为模式识别.
E-mail:mychu@126.com.

刘 恒 男,(1976-),博士,教授. 研究方向为图像处理.

张学锋 男,(1978-),博士,教授. 研究方向为计算机仿真.

潘 祥 男,(1983-),博士,讲师. 研究方向为图像处理.

(上接第 57 页)

[7] Doucet A, Godsill S, Andrieu C. On sequential Monte Carlo sampling methods for Bayesian filtering[J]. Statistics and computing, 2000, 10(3): 197-208.

[8] Zisserman A. Multiple view geometry in computer vision[J]. Kybernetes, 2001, 30(9/10):1865-1872.

[9] Ross D A, Lim J, Lin R S, et al. Incremental learning for robust visual tracking[J]. International Journal of Computer Vision, 2008, 77(1):125-141.

[10] 薛明. 基于稀疏表示的在线目标跟踪研究[D]. 上海: 上海交通大学, 2014.

作者简介:

陈 滨 男,(1988-),博士研究生. 研究方向为计算机视觉.
E-mail:cbsai1988@163.com.

赵建军 男,(1965-),博士,教授. 研究方向为武器装备与作战指挥一体化技术.

姚 刚 男,(1985-),博士,讲师. 研究方向为武器装备与作战指挥一体化技术.

杨利斌 男,(1978-),硕士,讲师. 研究方向为武器装备与作战指挥一体化技术.