聚类的动态分类器集成选择①

王宁燕[1], 韩晓霞[2]

[1](西安通信学院, 西安 710106)
[2](第二炮兵工程大学 502 教研室, 西安 710025)

摘 要: 动态分类器集成选择(DCES)是当前集成学习领域中一个非常重要的研究方向. 然而, 当前大部分 DCES 算法的计算复杂度较高. 为了解决该问题和进一步提高算法的性能, 本文提出了基于聚类的动态分类器集成选择(CDCES), 该方法通过对测试样本聚类, 极大地减少了动态选择分类器的次数, 因而降低了算法的计算复杂度. 同时, CDCES 是一种更加通用的算法, 传统的静态选择性集成和动态分类器集成为本算法的特殊情况, 因而本算法是一种鲁棒性更强的算法. 通过对 UCI 数据集进行测试, 以及与其他算法作比较, 说明本算法是一种有效的、计算复杂度较低的方法.

关键词: 动态分类器集成选择; 集成学习; 聚类

Cluster-Based Dynamic Classifier Ensemble Selection

WANG Ning-Yan[1], HAN Xiao-Xia[2]

[1](Xi'an Communication Institute, Xi'an 710106, China)
[2](502 Unit, Second Artillery Institute, Xi'an 710025, China)

Abstract: Dynamic classifier ensemble selection (DCES) is an important field in the machine learning. However, computational complexity of the current methods is very high. In order to solve the problem and improve the performance further, cluster based dynamic classifier ensemble selection (CDCES) is proposed in this paper. Using the proposed method to cluster the testing sample, the degree of DCES is reduced enormously and the computation complexity is decreased. At the same time, CDCES is a more general method and the traditional static ensemble selection and dynamic classifier is a specific case of the proposed method, so CDCES is more robust. Compared with the other algorithms on UCI data set, it is demonstrated that the proposed method is a more effective and lower computational complexity method.

Key words: dynamic classifier ensemble selection; ensemble learning; cluster

集成学习在过去的二十多年中已经成为机器学习领域的研究热点之一. 该方法已经广泛应用于手写体识别、人脸识别、生物特征识别和计算机辅助诊断等众多领域[1-5], 因而对集成学习的研究具有非常重大的意义.

集成学习最初由 Hansen 和 Salamon 提出, 他们通过训练多个神经网络并且将结果按照一定的规则进行组合, 发现可以显著提高整个学习系统的泛化性能[6]. 由于集成学习主要利用多个子分类器来获得比单一分类器更好的性能, 因而大部分方法均是通过使用大量的子分类器来获得更好的性能[13]. 最有代表性的工作是 Bagging, Boosting 等方法[7,8]. 然而, 这些做法会带来一些负面影响, 首先使用更多的子分类器会导致更大的计算和存储开销; 其次, 随着子分类器个数的增加, 各子分类器之间的差异性将越来越难获得. 为了解决该问题, 周志华等人提出了选择性集成, 并且说明了选择性集成优于传统的集成方法[9]. 由于该方法能够进一步提高集成学习的有效性和泛化性, 因而受

① 基金项目:国家自然科学基金(61004069)
收稿时间:2014-08-16;收到修改稿时间:2014-10-08

到了广泛的关注. 传统的选择性集成是一种静态的选择性集成, 即对所有的测试样本, 选择出的子分类器均是一样的. 然而, 该方法并没有考虑每一个测试样本的差异性. 为了针对每一个测试样本的差异性, 有学者提出了动态分类器集成选择(DCES)[9]. 该方法针对每一个测试样本的特征, 动态的选择出一组子分类器对该测试样本进行分类. 其中比较有代表性的工作有 KNORA[10], 以及 overproduce-and-choose 策略[11]. 然而, 针对大规模数据, 由于当前的动态分类器集成选择对每一个测试样本均选择出一组子分类器, 因而计算复杂度较高. 为了解决以上问题, 本文提出了基于聚类的动态分类器集成选择(CDCES). 该方法不仅能够进一步提高动态集成的分类性能, 而且能够降低算法的计算复杂度. 同时, 传统的静态选择、动态选择是本文方法的两种特殊情况, 因而 CDCES 是一种更为通用的方法. 利用 UCI 数据集测试本算法的性能, 并且与其他方法作比较, 表明 CDCES 是一种更加有效的方法.

本文的结构安排是: 第二部分介绍了当前 DCES 的相关工作, 第三部分为 CDCES 的算法描述, 第四部分为实验结果和分析, 最后给出了本文的结论和展望.

1 相关研究工作

一般而言, 传统的动态分类器集成分为三个基本的步骤[12]: (1)利用训练样本产生基分类器, 组成基本分类器集(BCP); (2)利用训练样本产生一组验证样本, 用于测试样本选择子分类器; (3)利用相应的方法, 为每一个测试样本选择一组子分类器组成集成分类器. 之所以称其为动态, 是因为步骤(2)和(3)是在测试阶段执行. 接下来对几种常见的 DCES 策略做一个简单的描述[10,12].

1.1 基于 K 近邻的动态分类器集成选择

K-nearest-oracles(KNORA)把集成的精确性作为选择的标准[10]. 对每一个测试样本, KNORA 首先寻找到其 K 个近邻作为验证样本, 然后选择出能够正确分类该 K 个样本的子分类器作为针对该测试样本的子分类器. 最后利用投票等融合方法得到该测试样本的分类结果.

1.2 基于动态分类器选择(DCS)的动态分类器集成选择

在集成选择时, 集成的差异性与精确性同等重要. 为了同时考虑以上两种特性, 有学者提出了基于 DCS 的 DCES[12]方法. 同时提出两种策略: 聚类和选择策略(C-V), 以及 K 近邻和选择策略(K-V). 以上两种策略的主要区别在于: 在 C-V 中, 所有的排序和选择均是在训练阶段执行, 而在 K-V 中, 这些步骤在测试阶段执行.

1.3 动态 overproduce-and-choose 策略

动态 overproduce-and-choose 策略(DOCS)由传统的 overproduction 和选择两步组成[10]. 而选择这一步又分为两步: 优化和动态选择. 第一步通过利用单目标或者多目标优化方法选择一个具有较高精确率的候选种群. 第二步利用信度测量方法选择出信度最高的一组集成分类器作为当前测试样本的集成分类器. 实验结果表明多目标优化通常能够得到比单目标优化更好的性能.

2 基于聚类的动态分类器集成选择

传统的 DCES 针对每个测试样本动态地选择一次集成分类器. 因而, 随着数据规模的增加, DCES 的计算复杂度会极大地增加. 基于此, 为了降低算法的计算复杂度和进一步提高性能, 本文提出了基于聚类的动态分类器集成选择(CDCES). 由于该方法利用聚类将测试样本归类, 因而只需要对每一类测试样本选择子分类器, 而不需要对每一个测试样本选择子分类器. 因此降低了算法的计算复杂度.

CDCES 分为以下四步: 首先利用相应的方法, 比如 Bagging 等产生 BCP. 第二, 利用 K 均值聚类等方法对测试样本聚类, 并且找到聚类中心. 第三, 为每个聚类中心寻找其 N 个近邻, 并将 N 个近邻作为验证样本集合, 验证 BCP 中的每个子分类器是否作为该类测试样本的子分类器. 第四, 利用选择出的子分类器组成该类测试样本的集成分类器, 并采用投票等融合方法输出分类结果.

该方法与传统方法的区别在于, 利用 K 均值聚类等方法将测试样本聚类, 并且找到每一类的聚类中心. 对该聚类中心在验证样本集合中寻找 N 个近邻, 然后利用已经训练好的子分类器测试该 N 个近邻的输出, 并将能够正确分类的子分类器保留下来, 作为该测试样本的集成子分类器. 由于本方法只需要动态选择 K 组集成分类器, 因而极大地降低了计算复杂度. 图 1 为 CDCES 的算法框架. 算法 1 给出了本文算法的实现

流程.

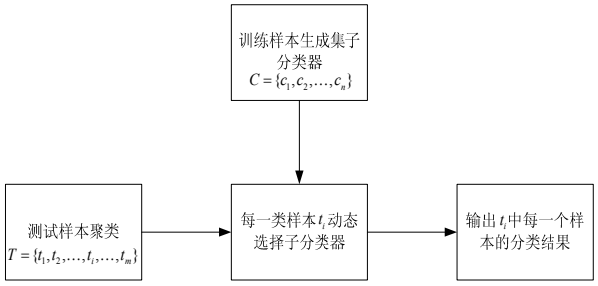

图 1 CDCES 算法框架

算法 1 基于聚类的动态分类器集成选择

步骤 1: 设训练样本为 $S=\{s_1,s_2,\ldots,s_p\}$, 利用 Bagging 等方法产生基本分类器集 (BCP), $C=\{c_1,c_2,\ldots,c_n\}$.

步骤 2: 对测试样本 $T=\{t_1,t_2,\ldots,t_m\}$, 采用 K 均值聚类将测试样本聚为 K 类, 每一类的聚类中心为 $C=\{c_1,c_2,\ldots,c_K\}$.

步骤 3: 对每一类测试样本的聚类中心 $c_i,i=1,\ldots,K$, 分别寻找其 N 个近邻作为验证样本集合, $V=\{v_{11},v_{12},\ldots,v_{1j_1};\ldots v_{K1},v_{K2},\ldots,v_{Kj_K}\}$. 对BCP中的基分类器进行验证, 能够正确分类 N 个近邻的基分类器作为该类样本的子分类器. 并组成集成分类器.

步骤 4: 如果所有的基分类器均不能正确的分类 N 个近邻, 令 $N=N-1$, 返回步骤 3.

步骤 5: 利用为每一类聚类好的测试样本所产生的集成分类器, 利用投票法输出测试样本的分类结果.

3 实验结果及分析

为了说明 CDCES 的有效性, 本文利用 UCI 数据集测试算法的有效性. 这里选取 6 组 UCI 数据集作为测试数据. 表 1 为所使用的数据集. 利用 Bagging 产生 BCP, 每次产生 10 个子分类器, 采用五倍交叉验证. 其中, 训练样本的 70%用于训练子分类器, 30%用于验证每个子分类器对当前测试样本的有效性. 为了充分说明算法的有效性, 本文采用人工神经网络(ANN), Fisher 判别分析(FDA)以及朴素贝叶斯(NB)等弱分类器作为子分类器. 对于 CDCES 和 KNORA, 所选取近邻的范围为 $1\le K\le 30$, 对 CDCES, 聚类的类别数 $5\le k\le 10$.

为了进一步说明本算法的有效性, 本文将传统的动态集成(KNORA), 传统的静态集成(ALL), 以及单个分类器(Single best)的性能与本算法作比较. 表二为子分类器为朴素贝叶斯分类器的实验结果, 表三为子分类器为 Fisher 判别分析的实验结果, 表四为子分类器为人工神经网络的实验结果. 表 2-4 中所示的结果为五倍交叉平均后的最好结果. 从表中可以看出, 在大部分测试数据集中, 本文所提算法的识别率优于其他方法(黑体加底纹部分). 实验结果说明本文方法能够取得较好的性能, 并且具有更好的鲁棒性.

表 1 UCI 数据集

数据集	样本数量	特征数量	类别数量
WINE	178	13	3
IRIS	150	4	3
LIVER	345	6	2
SCALE	625	4	3
CAR	1728	6	4
MAGIC	19020	10	2

表 2 NB 为子分类器的实验结果

数据集	CDCES (%)	KNORA (%)	ALL (%)	Single best (%)
WINE	**98.32**	97.75	97.19	96.05
IRIS	**96.00**	95.33	95.33	95.33
LIVER	**61.45**	60.29	59.42	60.58
SCALE	88.80	86.56	**89.76**	88.96
CAR	**84.32**	76.88	80.20	80.18
MAGIC	72.11	**73.45**	72.70	72.83

表 3 FDA 为子分类器的实验结果

数据集	CDCES (%)	KNORA (%)	ALL (%)	Single best (%)
WINE	**99.44**	**99.44**	98.89	97.17
IRIS	**98.00**	**98.00**	97.33	97.33
LIVER	64.93	63.48	**64.93**	64.64
SCALE	**83.84**	83.04	80.00	76.16
CAR	**79.75**	79.44	77.09	76.04
MAGIC	78.53	**79.75**	79.41	79.41

表 4 ANN 为子分类器的实验结果

数据集	CDCES (%)	KNORA (%)	ALL (%)	Single best (%)
WINE	97.73	96.63	**98.32**	94.37
IRIS	**96.67**	96.00	**96.67**	**96.67**
LIVER	**66.80**	65.50	67.25	66.09
SCALE	**91.04**	90.24	89.12	89.28
CAR	**92.46**	91.72	90.80	89.06
MAGIC	**86.23**	85.46	86.15	86.17

为了说明本算法的计算复杂度, 本文给出 CDCES 和 KNORA 在 WINE、IRIS、LIVER 以及 SCALE 四个数据集上的运行时间. 算法运行于 PC 机上, 操作系统为 Windows XP, CPU 为 Intel Core 双核, 主频为 2.99 GHz, 内存为 2G. 运行平台为 Matlab2009.

图 2 和图 3 分别给出了子分类器分别为 NB 和 FDA 时, KNORA 和 CDCES 两算法的运行时间. 从图中可以明显看出, CDCES 的运行时间要远远低于 KNORA. 同时, 随着数据规模的增加, CDCES 的运行时间并没有较大的增加, 而 KNORA 有较大的增加. 这说明相比于传统的 DCES, 本文所提算法的计算复杂度较低.

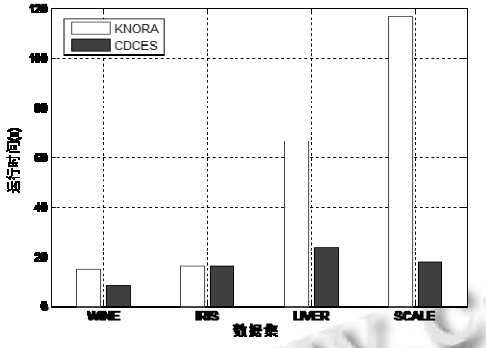

图 2　子分类器为 NB 时算法的运行时间

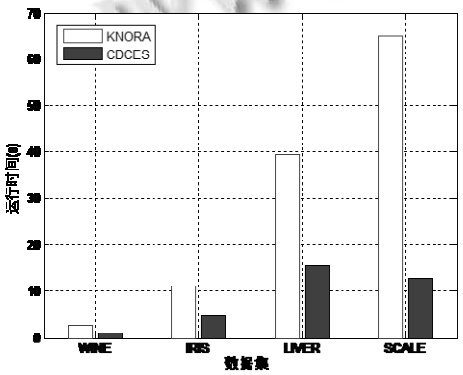

图 3 子分类器为 FDA 时算法的运行时间

另外, CDCES 是一种更加通用的算法. 当聚类的类别数与测试样本数量相当时, CDCES 就等价于 KNORA; 当聚类类别数为 1 时, CDCES 就等价于传统的静态选择性集成.

4　结论与展望

本文首先提出了基于聚类的动态分类器集成选择算法(CDCES). 由于 CDCES 对测试样本聚类, 因而需要动态选择集成分类器的次数仅为聚类的类别数量, 而传统的动态分类器集成选择需要为每个测试样本选择一次集成分类器. 因此, CDCES 极大的降低了计算复杂度. 通过利用 UCI 数据集测试本算法, 并且与传统的静态以及动态分类器选择算法相比, 本算法在大部分情况下取得了较好的性能, 说明 CDCES 具有较好的鲁棒性. 通过与传统 DCES 方法运行时间的比较, 说明本算法的计算复杂度较低. 另外, 本算法是一种更加通用的方法, 传统的静态选择以及动态选择均为本算法的特殊情况.

尽管 CDCES 在 UCI 数据集上展示了较好的性能, 但是在实际应用中能否够取得较好的性能, 有待于进一步验证. 本算法利用聚类对测试样本分组, 降低了算法复杂度, 但是聚类的类别参数需要调整, 因而在接下来的研究中需要发展一种自适应的方法来调整参数.

参考文献

1 Baruque B, Corchado E, Mata A, Corchado JM. A forecasting solution to the oil spill problem based on a hybrid intelligent system. Information Sciences, 2010, 180(10): 2029–2043.

2 Cho SB. Pattern recognition with neural networks combined by genetic algorithm. Fuzzy Sets and Systems, 1999, 103(2): 339–347.

3 Mallipeddi R, Mallipeddi S, Suganthan PN. Ensemble strategies with adaptive evolutionary programming. Information Sciences, 2010, 180(9): 1571–1581.

4 Yang G, Lin Y, Bhattacharya P. A driver fatigue recognition model based on information fusion and dynamic Bayesian network. Information Sciences, 2010, 180(10): 1942–1954.

5 Shi L, Xi L, Ma XM, Weng M, Hu XH. A novel ensemble algorithm for biomedical classification based on Ant Colony Optimization. 2011. (in press).

6 Hansen LK, Salamon P. Neural network ensembles. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1990, 12(10): 993–1001.

7 Sehapire RE. The strength of weak learnability. Machine Learning, 1990, 5(2): 197–227.

8 Breiman L. Bagging predicators. Machine Learning, 1996, 24(2): 123–140.

9 Zhou ZH, Wu J, Tang W. Ensembling neural networks: many could be better than all. Artificial Intelligence, 2002, 137(1-2): 239–263.

10 Ko AHR, Sabourin R, de Souza Britto Jr. A. From dynamic classifier selection to dynamic ensemble selection. Pattern Recognition, 2008, 41(5): 1718–1731.

11 Santos EMD, Sabourin R, Maupin P. A dynamic overproduce-and-choose strategy for the selection of classifier ensembles. Pattern Recognition, 2008, 41(10): 2993–3009.

12 Santana A, Soares RGF, Canuto AMP, Souto MCP. A dynamic classifier selection method to build ensembles using accuracy and diversity. Proc. of the Ninth Brazilian Symposium on Neural Networks (SBRN). 2006. 36–41.

13 李青,焦李成.利用集成支撑矢量机提高分类性能.西安电子科技大学学报,2007,34(1): 68–70.