

基于特征选择的集成分类器抗噪性能分析

韦艳艳¹, 李陶深²

(1. 广西民族大学 信息科学与工程学院 广西混杂计算与集成电路设计

分析重点实验室, 广西 南宁 530006;

2. 广西大学 计算机与电子信息学院, 广西 南宁 530004)

摘 要:特征选择有助于增强集成分类器成员间的随机差异性,从而提高泛化精度。研究了随机子空间法(Random Subspace)和旋转森林法(Rotation Forest)两种基于特征选择的集成分类器构造算法,分析讨论了两算法特征选择的方式与随机差异程度之间的关系。通过对UCI数据集引入噪声,比较两者在噪声环境下的分类精度。实验结果表明:当噪声增加及特征关联度下降时,基本学习算法及噪声程度对集成效果均有影响,当噪声增强到一定程度后,集成效果和单分类器的性能趋于一致。

关键词:集成分类器;特征选择;随机子空间;旋转森林;抗噪性能

中图分类号:TP181

文献标识码:A

文章编号:1673-629X(2012)11-0161-04

Anti-noise Performance Analysis of Classifiers Ensembles Based on Feature Selection

WEI Yan-yan¹, LI Tao-shen²

(1. Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, School of Information Science

& Engineering, Guangxi University for Nationalities, Nanning 530006, China;

2. School of Computer & Electronic Information, Guangxi University, Nanning 530004, China)

Abstract: Feature selection encourages random differentiation of the members of the ensembles to improve generation accuracy. In this paper, random subspace and rotation forest, two algorithms based on feature selection for constructing classifiers ensembles were researched and their relationship between ways of selecting features and its affection on diversity was discussed. By introducing noise into UCI data sets, compared anti-noise performance with different noisy level of two algorithms. Experimental results indicate that both base learning algorithms and noisy level affect the accuracy of an ensemble while noise increases and feature correlation decreases. In situation with higher classification noise, both ensembles and single classifier exhibit quite similar performance.

Key words: classifiers ensembles; feature selection; random subspace; rotation forest; anti-noise performance

0 引言

集成分类器将多个分类器按一定的方式组合而成,通常可以获得比成员分类器更好的分类性能,在模式识别、机器学习、数据挖掘等领域具有广泛的应用前景。Dietterich^[1]解释了集成分类器之所以有效的原因在于集成学习方式能较好地克服训练集信息量有限、假设空间H表达能力不足等问题,因而可以有效地提高泛化效果。在集成学习中,相关研究表明^[2,3],成员

分类器间的差异性(diversity)有助于提高集成分类器的泛化性能。增强差异性的有效方法之一就是通过特征选择方式得到特征子集来构造成员分类器^[4],即以随机方式得到若干特征子集来训练分类器,由此增强其相互间的差异程度,如随机子空间法(Random Subspace)^[5]、旋转森林法(Rotation Forest)^[6]。

集成分类器另一个重要优势在于增强抗噪声的能力。噪声数据在现实世界中普遍存在,如数据或标签错误、属性值不完整等,导致学习算法对假设空间搜索以找到能最好地拟合训练样本的假设时出现偏差,进而影响分类器的精度,使得分类器的方差(variance)增大。通过对多个分类器进行集成,则可以减小方差,降低学习算法对结果估计的扰动程度,从而提高算法的鲁棒性。据此,文中首先分析了集成环境下成员分类

收稿日期:2012-04-01;修回日期:2012-07-03

基金项目:广西自然科学基金项目(2010GXNSFA013127);广西教育项目(201106LX131)

作者简介:韦艳艳(1974-),女,广西贵港人,讲师,硕士,CCF会员,研究方向为机器学习、数据挖掘;李陶深,教授,博士,CCF高级会员,研究方向为网络路由及分布式计算。

器相互间的差异性对误差的影响,然后介绍了随机子空间法和旋转森林法的实现过程,讨论这两种集成分类器构造算法对特征选择的处理方式,最后通过在不同噪声环境下的实验测试集成精度以及与单分类器的对比,分析比较两算法的抗噪声性能。

1 相关内容简介

1.1 分类器集成环境下的差异性分析

设 $F(x)$ 为集成分类器, $G(x)$ 为目标分类器, $f_i(x)$ 为成员分类器,若采用简单平均的集成规则,那么集成分类器可表示为:

$$F(x) = \frac{1}{N} \sum_{i=1}^N f_i(x) \quad (1)$$

其中 N 为成员分类器的个数。则集成分类器的均方误差可表示为^[7]:

$$\begin{aligned} \text{MSE}_e &= E\{F(x) - G(x)\}^2 \\ &= E(F(x) - E\{f_i(x)\})^2 + (E\{F(x)\} - G(x))^2 \\ &= E\left\{\left(\frac{1}{N} \sum_{i=1}^N f_i(x) - E\left\{\frac{1}{N} \sum_{i=1}^N f_i(x)\right\}\right)^2\right\} + \text{Bias}_e^2 \\ &= \frac{1}{N^2} E\left\{\sum_{i=1}^N \sum_{j=1, j \neq i}^N (f_i(x) - E\{f_i(x)\})(f_j - E\{f_j(x)\})\right\} \\ &\quad + \frac{1}{N^2} E\left\{\sum_{i=1}^N (f_i(x) - E\{f_i(x)\})^2\right\} + \text{Bias}_e^2 \\ &= \frac{1}{N^2} \text{Cov}_{i,j} + \frac{1}{N^2} \text{Var}_i + \text{Bias}_e^2 \quad (2) \end{aligned}$$

其中: $\text{Cov}_{i,j}$ 代表成员分类器之间的协方差, Var_i 代表单个成员分类器的方差, Bias_e 是集成分类器的偏差。

由式(2)可见,集成分类器的均方误差值与成员间的协方差密切相关。这时协方差可看作成员间的错误关联程度,成员分类器之间的关联越小,协方差也越小,表明成员分类器的差异性越强。因此,要获得理想的集成泛化效果,成员分类器之间应尽可能少地出现判决一致性或完全相等的计算结果(如预测概率等)、尽可能多地有不同的输出结果,这样才能生成更为精准的集成分类器。

增强成员分类器间的随机差异性的方法包括:采用不同学习算法生成异构分类器、调整学习算法的训练参数、改变训练样本或改变训练集特征等,这些方法都可以使得成员分类器在预测偏好、分类边界等方面存在不同程度的差别,从而达到降低集成分类器误差的目的。

1.2 特征选择

对于分类器的构造,理想的训练集特征应当具备的特点是:每一个特征与类别标记是强相关的但特征之间是弱相关或不相关的^[8]。因而,特征选择的目标

是消除那些对目标函数相关性低的特征或冗余的特征。由此可见,通过改变训练特征集的方式来构建具有差异性的成员分类器,要使得选取出来的特征与原类别标记具有比较强的相关性,否则,虽然可以得到差异性较大的成员分类器,却会出现分类器偏离原分类问题的现象^[9]。常见的特征选择方法有以粗糙集、信息熵等度量作为特征评价的有监督方法,以特征相关程度作为筛选评价的无监督方法,以及融合无标签样本和有标签样本进行学习的半监督方法等^[10]。

2 基于特征选择的集成分类器构造算法

2.1 随机子空间法

随机子空间法(以下简称 RS)的算法思路是通过划分特征集的方式降低成员分类器间的相关程度,以提高随机差异性,最终的分类判别根据成员分类器的投票结果得到。文献[5]证明,RS 对线性分类器在训练集样本量少时具有递减的学习曲线。

RS 算法描述如下:

输入:(1)成员分类器数量 T ;

(2)训练集 $D = (X, Y)$;

$X = \{x_1, x_2, \dots, x_N\}$ 为 $N \times M$ 的矩阵;

$Y = \{y_1, y_2, \dots, y_N\}$ 为 $N \times 1$ 的矩阵,其中 $y_i \in \{\omega_1, \omega_2, \dots, \omega_k\}$

(3)特征集 $F = \{f_1, f_2, \dots, f_M\}$;

(4)学习算法 L ;

(5)子空间维数 m , 其中 $m < M$ 。

输出:集成分类器 C 。

计算步骤:

Do For $i = 1, 2, \dots, T$

随机生成 m 维特征子集 F_i , 得到 D 在 F_i 上的映射集 D_i ;

用 L 在 D_i 上训练成员分类器 C_i ;

End For

对于测试样本 (x', y') , 首先根据特征子集 F_1, F_2, \dots, F_T 得到 x'_1, x'_2, \dots, x'_T 个子样本, 然后计算

$$C(x') = \arg\max \left(\sum_{i=1}^T C_i(x'_i) \right) \quad (3)$$

分类结果由成员分类器按多数投票法得出。

2.2 旋转森林法

旋转森林法(以下简称 RF)的思路是将原始特征投影到若干个新的特征空间,从而使生成的成员分类器之间具有随机差异性。具体方法是,将原始特征集随机划分成 K 个特征子集,相应地得到 K 个训练集,使用 PCA 主成分分析方法从 K 个训练集中抽取出 n 个主成分,将原始数据线性投影到新的特征空间,然后从新的样本空间生成决策树分类器。

RF 算法描述如下:

输入: (1) 成员分类器数量 T ;

(2) 训练集 $D = (X, Y)$;

$X = \{x_1, x_2, \dots, x_N\}$ 为 $N \times M$ 的矩阵;

$Y = \{y_1, y_2, \dots, y_N\}$ 为 $N \times 1$ 的矩阵, 其中

$y_i \in \{\omega_1, \omega_2, \dots, \omega_k\}$

(3) 特征集 $F = \{f_1, f_2, \dots, f_M\}$ 以及特征子集个数 K ;

(4) 学习算法 L 。

输出: 集成分类器 C 。

计算步骤:

Do For $i = 1, 2, \dots, T$

将特征集 F 划分成 K 个子集 $F_{i1}, F_{i2}, \dots, F_{iK}$;

Do For $j = 1, 2, \dots, K$

映射 D 到特征子集 F_{ij} 得到 D_{ij} ;

随机从 D_{ij} 中选出一个非空的类标签子集 D_{ij}' ;

对 D_{ij}' 进行 PCA 变换, 得到矩阵 M_{ij} , 矩阵的第 i 列表示第 i 个主成分的因子。

End For

按 F 的特征序列重新排列 $M_{ij} (j = 1, 2, \dots, K)$, 得到 M_{ij}' , 放入 R_i^α 中。

$$R_i^\alpha = \begin{bmatrix} M_{i1}' & [0] & \dots & [0] \\ [0] & M_{i2}' & \dots & [0] \\ \dots & \dots & \dots & \dots \\ [0] & [0] & \dots & M_{iK}' \end{bmatrix}$$

由矩阵 $XR_i^\alpha Y$ 构造成员分类器 C_i 。

End For

对于测试样本 (x', y') ,

$$C(x') = \operatorname{argmax}_{y \in \Phi} \sum_{i=1}^T I(C_i(XR_i^\alpha) = y) \quad (4)$$

即由成员分类器根据 XR_i^α 计算出 x' 属于类别 ω_i 的概率均值, 均值最大的即为 x 所属的类别。在 PCA 计算过程中, 所有主成分均被保留, 以描述原始数据的全部信息。

2.3 算法讨论

根据上述算法描述, RS 集成的效果受成员分类器个体精度及成员间判别差异的共同影响, 优化其中一项并不一定会得到更好的准确率。Ho^[5] 认为, RS 的子空间维数等于原特征集维数的一半较适合, 而文献[11]则用随机搜索来确定子空间维数。

RF 在特征选择过程中增加了对特征子集的 PCA 变换。通过随机选择得到特征子集, 其原始特征相互间可能存在强关联性, 而经过 PCA 变换, 使得这些特征在新的映射空间中表现出不相关或变成弱相关, 这无疑表明变换后的特征子集较原先更接近于“理想”

特征集。因此, 构造成员分类器时, 学习算法能够在新特征空间中较好地不同类别的数据点区分开来。文献[12]指出, 参数 $K=3$ 时, 用 PCA 进行特征变换获得的集成分类效果最好。

3 实验与分析

文中用 UCI^[13] 中 16 个二类数据集在 WEKA^[14] 平台上进行了一系列对比实验, 以了解 RS 及 RF 在无噪声与有噪声环境下的分类情况。

3.1 数据集

表 1 为数据集描述。

表 1 数据集描述

Datasets	Inst.	Attr.	Dis.	Con.
breast-c	286	9	9	0
breast-w	699	9	0	9
colic	368	22	15	7
credit-a	695	15	9	6
credit-g	1000	20	13	7
diabetes	768	8	0	8
echoc	132	8	1	7
heart-s	270	13	7	6
hepatitis	155	19	13	6
ionosphere	355	34	0	34
kr-vs-kp	3196	36	36	0
labor	57	16	8	8
promoters	106	57	57	0
sonar	208	60	0	60
tic-tac-toe	958	9	9	0
vote	435	16	16	0

3.2 实验设置

对表 1 数据集引入了从 5% 至 50% 不同比例水平的类噪声, 降低其特征关联度, 然后分别用决策树 C4.5, 朴素贝叶斯 NB 以及 K 近邻 IBK ($K=3$) 三种不同的学习算法构造成员分类器, 成员分类器个数为 10, 集成分类器用 10 次 10-折交叉验证计算其分类准确率。除基本学习算法外, RF 与 RS 的其它参数均采用 WEKA 的默认设置。

3.3 评估度量

实验采用几何平均错误率 (GM error ratio) 来对比 RS 与 RF 的分类性能。几何平均错误率 G 定义为:

$$G = \sqrt[n]{\prod_{i=1}^n \frac{E_{iA}}{E_{iB}}} \quad (5)$$

其中 E_{iA} 和 E_{iB} 分别表示算法 A 和算法 B 在第 i 个数据集上的错误率。若 G 小于 1, 则说明算法 A 优于算法 B; 若 G 大于 1, 结果反之; 而 G 等于 1 时, 表示算法 A

与 B 的分类性能相同。几何平均错误率在计算时考虑到了算法在所有数据集中的表现,因此可以比较客观地评价两种算法的优劣程度。

3.4 结果分析

(1) 无噪声时的对比。

在无噪声情况下,如表 2 所示,用 C4.5 训练成员分类器时,RF 与 RS 间几何平均错误率为 0.6996,RF 的性能要明显优于 RS;而使用贝叶斯和最近邻法时,RF 与 RS 间的几何平均错误率均大于 1,RS 的性能则优于 RF。

表 2 无噪声的分类性能对比

Datasets	RF / RS		
	C4.5	Naïve Bayes	IBK
平均准确率	86.07/83.35	78.22/81.35	83.29 /84.12
几何平均错误率	0.6996	1.1437	1.0412

(2) 噪声环境下的对比。

在类噪声比例为 0% ~ 15% 时,使用决策树作为基本学习算法时,RF 要优于 RS,而 RS 用贝叶斯方法和 IBK 时均表现好于 RF,但从 20% 起,不管采用何种学习算法,RF 与 RS 的分类效果差别不大。噪声比例越大,特征关联程度也随之降低,几何平均错误率的值越接近,如图 1 所示。

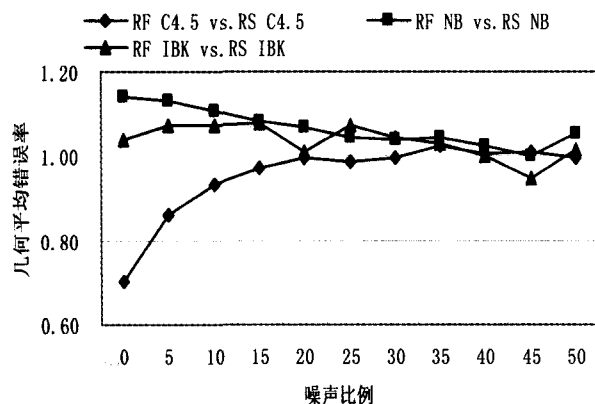


图 1 RF 与 RS 在不同噪声环境下的对比

(3) 与单分类器的对比。

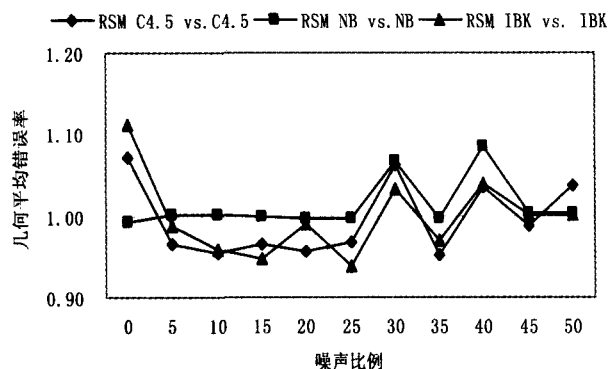


图 2 RS 与单分类器的对比

从图 2 来看,RS 与单分类器之间出现了较明显的波动,RS 与单分类器之间在噪声 25% ~ 45% 时波动较明显。相比之下,RF 与单分类器的对比则要平缓得多,如图 3 所示。在低噪声时,虽然 RF 对贝叶斯分类器和最近邻分类器的集成效果不如单分类器,但是当噪声逐渐增加,它们之间的差异变小。

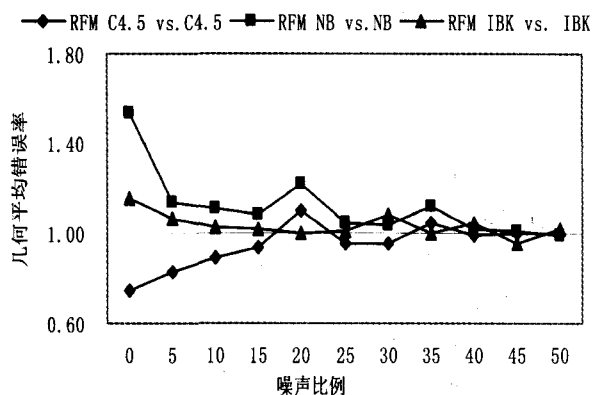


图 3 RF 与单分类器的对比

这说明,RF 对决策树算法有明显偏好,而采用其它学习算法时,RS 占优势。这可以解释为原始数据的特征空间经过 PCA 转换后,更利于 C4.5 算法而非贝叶斯和 IBK 进行泛化,由此影响了成员分类器的精度。

4 结束语

文中研究了两种基于特征选择技术的分类器集成算法 Random Subspace 与 Rotation Forest,分析比较了两者在噪声环境下的分类精度。实验结果表明:低噪声环境下,使用不同的基本学习算法对 RF、RS 泛化性能有较明显的影响,随着噪声的增加,特征关联度下降,两算法差异逐渐缩小,集成的效果趋于一致;其次,在与单分类器的对比中,集成分类器的精度受基本学习算法的影响,且当噪声增加到一定程度后,集成的优势已经基本消失,泛化性能与单分类器相当。

参考文献:

- [1] Dietterich T G. Ensemble methods in machine learning[M]//Multiple Classifier Systems. Cagliari, Italy: Springer-Verlag, 2000:1-15.
- [2] Brown G. Ensemble Learning[M]//Encyclopedia of Machine Learning. New York: Springer-Verlag, 2010:1-24.
- [3] Dietterich T G. An Experimental Comparison of Three Methods for Constructing Ensemble of Decision Trees: Bagging, Boosting and Randomization[J]. Machine Learning, 2000, 40(2): 139-157.
- [4] 叶云龙,杨明. 基于随机子空间的多分类器集成[J]. 南京师范大学学报(工程技术版), 2008, 8(4): 87-90.
- [5] Ho T K. The Random Subspace Method for Constructing Deci-

(下转第 168 页)

无线传感器网络是靠电池供电,网络的寿命是关键。

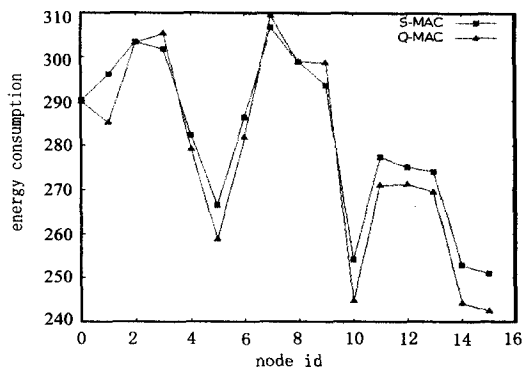


图 5 能量分析

根据各节点的能量消耗图,在实际各项应用中,比如军事、医疗等方面,可根据各节点的能量浪费,运用不同的电池供电,能量消耗多的节点可用容量较大的电池,能量消耗少的节点可用容量较小的电池,这样可以更有效地控制网络,延长网络的寿命。

4 结束语

Q-MAC 协议是对 S-MAC 协议的改进,针对 S-MAC 协议退避窗口固定,不能适应信道变化,造成信道利用率下降的问题,在 S-MAC 协议的基础上,采用自适应的退避窗口,并结合队列长度预测流量的思想,经过 NS2 模拟实验表明,Q-MAC 协议在保持了与 S-MAC 协议相同能量消耗的情况下,有效地缩短了延迟,提高了吞吐量。

参考文献:

- [1] 储邵勋,胡艳军. 无线传感器网络技术[J]. 计算机技术与发展,2006,16(4):64-66.
- [2] 严谨,李平. 能量有效的无线传感器网络 MAC 协议[J]. 计算机工程,2010,36(23):98-100.
- [3] 苏俊,胡访宇. 无线传感器网络 SMAC 协议的节能改进[J]. 计算机工程,2009,35(5):106-121.
- [4] Chatterjee M, Das S K, Turgut D. WCA: A Weighted Clustering Algorithm for Mobile Ad Hoc Network [J]. Journal of Clustering Computing, 2002, 5(2): 193-204.
- [5] Cali F, Conti M, Gregori E. Dynamic Tuning of the IEEE 802.11 Protocol to Achieve a Theoretical Throughput Limit [J]. IEEE/ACM Trans. on Networking, 2000, 8(6): 785-799.
- [6] Bharghvan V. Performance evaluation of algorithms for wireless medium access [C]//IEEE International Computer Performance and Dependability Symposium. [s. l.]: [s. n.], 1998:142-149.
- [7] 杨武,史浩山,杨俊刚,等. 无线传感器网络中 SMAC 协议的改进与仿真[J]. 传感器与微系统,2010,29(7):47-52.
- [8] 徐雷鸣,庞博,赵耀,等. NS 与网络模拟[M]. 北京:人民邮电出版社,2003.
- [9] 赵志伟,张信明,刘道科,等. 基于竞争允许 TDMA 的无线传感器 MAC 层协议[J]. 计算机工程,2007,33(5):116-121.
- [10] 刘善平,林亚平,周四望,等. 一种低能耗低延时的无线传感器网络 MAC 协议[J]. 计算机应用,2006,26(2):287-291.
- [11] 刘俊,杨全胜. 一种基于定位应用的无线传感器网络 MAC 层方案[J]. 计算机技术与发展,2009,19(1):204-206.
- [12] 杨树森,周小佳,阎斌. 无线传感器网络在环境监测中的应用[J]. 计算机技术与发展,2008,18(9):170-172.
- [13] 蒋盛益,王边喜. 基于特征相关性的特征选择[J]. 计算机工程与应用,2010,46(20):153-156.
- [14] Skurichina M, Duin R P W. Bagging, boosting and the random subspace methods for linear classifier [J]. Pattern Analysis and Applications, 2002, 5(2): 121-135.
- [15] Kuncheva L I, Rodríguez J J. An Experimental Study on Rotation Forest Ensembles [C]//MCS 2007, Lecture Notes in Computer Science. Berlin: Springer, 2007: 459-468.
- [16] Frank A, Asuncion A. UCI Machine Learning Repository [D]. Irvine, CA: University of California, 2010.
- [17] Hall M, Frank E, Holmes G, et al. The WEKA Data Mining Software: An Update [J]. SIGKDD Explorations, 2009, 11(1): 10-18.
- [18] Mark A H. Correlation-based Feature Subset Selection for Discrete and Numeric Class Machine Learning [C]//Proceeding of the 17th International Conf on Machine Learning. [s. l.]: [s. n.], 2000: 359-366.
- [19] 张宏达,王晓丹,韩钧,等. 分类器集成差异性研究[J]. 系统工程与电子技术, 2009, 31(12): 3007-3012.

(上接第 164 页)

- [5] sion Forests [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 832-844.
- [6] Rodríguez J J, Kuncheva L I, Alonso C J. Rotation Forest: A New Classifier Ensemble Method [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(10): 1619-1630.
- [7] Arijun C, Xin Y. Evolving hybrid ensembles of learning machines for better generation [J]. Neurocomputing, 2006, 69(7-9): 689-700.
- [8] Mark A H. Correlation-based Feature Subset Selection for Discrete and Numeric Class Machine Learning [C]//Proceeding of the 17th International Conf on Machine Learning. [s. l.]: [s. n.], 2000: 359-366.
- [9] 张宏达,王晓丹,韩钧,等. 分类器集成差异性研究[J].