

分类器模拟算法及其应用

杨利英, 覃 征, 张选平

(西安交通大学电子与信息工程学院, 710049, 西安)

摘要: 针对标准数据集在评估多分类器系统的组合方法时存在的不足, 设计了一种新的分类器模拟算法. 该算法利用分类器的识别率建立混淆矩阵, 由混淆矩阵生成基分类器的决策, 进而结合分类器之间的相关性度量生成所有的模拟数据. 通过实验评估表明, 该算法能够模拟任意多个分类器和任意多个模式类别的数据, 且能够表达出分类器之间的关联性. 又应用生成的模拟数据集对多数投票和堆叠泛化这 2 种组合方法进行了实验, 结果表明分类器之间的负相关有助于提高系统的性能, 特别是当单个分类器识别率取 0.8、关联度从 0.829 5 降至 -0.484 7 时, 多数投票和堆叠泛化的性能分别提高了 14.98% 和 41.99%.

关键词: 多分类器系统; 分类器模拟算法; 相关性

中图分类号: TP181 **文献标识码:** A **文章编号:** 0253-987X(2005)12-1311-04

Classifier Simulation Algorithm and Its Applications

Yang Liying, Qin Zheng, Zhang Xuanping

(School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Aiming at the deficiency of evaluating classifier combination methods with standard data sets, a new classifier simulation algorithm was proposed. The confusion matrix was established by the classifier's recognition rate, by which the decision of the base classifier was made. Then all simulating data were generated by combining the correlation measure between classifiers. Through experimental investigation it is shown that the algorithm can simulate any number of classifier and any number of data with any kind of pattern, and it can also express the dependency between classifiers. With simulated datasets, experiments were carried out on both of majority vote and stacking combination method. The results indicate that negative correlation can improve the classification performance. The accuracy of the two methods increases by 14.98% and 41.99% respectively when decreasing the correlation from 0.829 5 to -0.484 7, particularly when the recognition rate of individual classifier is set to 0.8.

Keywords: multiple classifier system; classifier simulator; correlation

在多分类器系统中, 探讨分类器之间相关性对组合效果的影响是一个重要的问题. 由于不同的组合方法会得到不同的效果, 应用时需谨慎选取, 这就要求能对各种组合方法进行评估. 评估通常是在标准数据集上进行的, 但标准数据集的特性是固定的, 不便于调节和控制, 因此使用模拟数据评估组合方法成为一个新的研究方向^[1]. Lecce^[2]用相似性索引衡量分类器之间的关联, 产生模拟数据; Zouari^[3]提

出了一种分类器模拟算法, 但没有考虑分类器之间的关联; Kuncheva^[4]推导出 2 个分类器在指定识别率和关联程度时输出的计算公式, 用它可以精确地产生模拟数据, 但其模拟输出的是二进制向量, 而非样本类别.

本文提出了一种新的分类器模拟算法. 该算法首先产生一个分类器的模拟输出, 然后根据 Q 统计量等参数产生其他分类器的输出. 与以往研究所不

同的是,它既能根据混淆矩阵生成用样本类别表示的模拟数据,又能表达多个分类器之间的关联性。

1 算法描述

1.1 相关性度量

Q 统计量是描述 2 个分类器之间关联程度的一种度量,当有多个分类器时,用 Q 统计量的均值描述分类器集合的关联^[5]。本文采用 Q 统计量作为分类器关联的度量,因为它对于独立关系及正负相关都有清晰的解释,且易于计算。分类器 R_i 和 R_j 之间的 Q 统计量定义为

$$Q_{ij} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \quad (1)$$

N^{ab} 是指测试样本集中 R_i 判决为 a 、 R_j 判决为 b 的样本数目。式(1)中 a 和 b 取值为 1 表示分类器判决正确,取值为 0 则表示判决错误。

1.2 输入参数

按照参与组合的单个分类器的输出异同,多分类器组合可在 3 个层次水平上进行:抽象级、排序级、度量级。本文提出的分类器模拟算法产生的是抽象级输出,如果需要输出排序级或者度量级类型,可以通过混淆矩阵进行转化。模拟算法的一个输出可用二元组(Original_Class, Simulating_Class)定义,其中 Original_Class 指模拟的样本真实类别,Simulating_Class 指模拟算法的判决类别。分类器模拟算法的输入参数包括类别数 M ,每类样本的数目 N ,所模拟的 l 个分类器 $\{R_1, R_2, \dots, R_l\}$ 及各自的识别率 $\{L_1^T, L_2^T, \dots, L_l^T\}$,任意 2 个分类器之间的关联度 $Q_{ij}, i, j \in \{1, 2, \dots, l\}$ 。

2 构建混淆矩阵

如果不考虑拒绝输出,分类器 R_k 的混淆矩阵是一个 $M \times M$ 矩阵,记为 M_k^C 。其中,第 i 行、第 j 列的元素 $M_k^C(i, j)$ 表示分类器 R_k 将真实类别为 i 的样本判决为类别 j 的概率,记为混淆率 C_k^j 。当 $i=j$ 时, $M_k^C(i, j)$ 表示的是分类器 R_k 对第 i 类样本的识别率。不失一般性,假设每一分类器对于各类别样本的识别率都相同,则 M_k^C 对角线元素都为 L_k^T 。在 M_k^C 的第 i 行中,除第 i 列的元素外,其余列之和为分类器 R_k 相对于类别 i 的总混淆率,各个混淆率可随机确定,但需满足

$$C_k^i = \sum_{j=1, j \neq i}^M C_k^j = 100\% - L_k^T \quad (2)$$

3 产生相关分类器的输出

3.1 2 个分类器输出的生成算法

给定 2 个分类器 R_i, R_j 的识别率 L_i^T, L_j^T 及其关联度 Q_{ij} ,生成算法首先产生分类器 R_i 的输出,然后根据此输出得到分类器 R_j 的输出。对于每一个模拟输出,当 R_i 判决出错时,按概率 $P_{F \rightarrow T}(i, j)$ 将其变换成正确的类别以产生 R_j 的相应输出,而当 R_i 判决正确时,按概率 $P_{T \rightarrow F}(i, j)$ 将其转变为错误的判决以产生 R_j 的相应输出,且由混淆概率矩阵 M_j^C 确定变换为何种混淆类别。 $P_{F \rightarrow T}(i, j)$ 和 $P_{T \rightarrow F}(i, j)$ 通过已知量得到^[4]。

当 $Q_{ij} \neq 0$,即 R_i, R_j 相关时,可得

$$P_{F \rightarrow T}(i, j) = \frac{-(1 - Q_{ij} + 2Q_{ij}(L_i^T - L_j^T)) \pm \Delta^{1/2}}{4Q_{ij}(1 - L_i^T)} \quad (3)$$

其中

$$\begin{aligned} \Delta &= (1 - Q_{ij} + 2Q_{ij}(L_i^T - L_j^T))^2 - \\ &8Q_{ij}(1 - L_i^T)L_j^T(Q_{ij} - 1) \\ P_{T \rightarrow F}(i, j) &= 1 - P_{F \rightarrow T} + \frac{P_{F \rightarrow T} - L_j^T}{L_i^T} \end{aligned} \quad (4)$$

当 $Q_{ij} = 0$,即 R_i, R_j 相互独立时,可得

$$P_{F \rightarrow T}(i, j) = L_j^T \quad (5)$$

$$P_{T \rightarrow F}(i, j) = 1 - L_j^T \quad (6)$$

具体描述如下。

步骤 1: 输入参数 L_i^T, L_j^T 和 Q_{ij} 。

步骤 2: 生成混淆矩阵 M_i^C, M_j^C 。

步骤 3: 计算 $P_{F \rightarrow T}(i, j)$ 和 $P_{T \rightarrow F}(i, j)$ 。

步骤 4: 由 M_i^C 的各行构建累积频率矩阵 $F_{M \times M}^C$,

即

For $ii=1$ to M

$$F^C(ii, 1) = M_i^C(ii, 1)$$

End ii

For $ii=1$ to M

For $jj=2$ to M

$$F^C(ii, jj) = F^C(ii, jj-1) + M_i^C(ii, jj)$$

End jj

End ii

步骤 5: 产生 R_i 的 $M \times N$ 个输出,并把它们存储于矩阵 P_i^O ,该矩阵具有 $M \times N$ 行和 2 列,即

For $ii=1$ to M

For $jj=1$ to N

取随机数 $Z, 0 \leq Z < 1$,扫描 F^C 的行 ii 以寻找 $F^C(ii, kk)$ 的第一个满足 $F^C(ii, kk) \geq Z$

的值(最小的索引), kk 是取自于类 ii 样本的模拟分类结果. 这样得到一个输出 (ii, kk) , 其中 $P_i^O(N * (ii-1) + jj, 1) = ii$ 和 $P_i^O(N * (ii-1) + jj, 2) = kk$
 End jj

End ii

步骤 6: 生成 R_j 的输出矩阵 P_j^O , 即

For $ii=1$ to $M \times N$

$P_j^O(ii, 1) = P_i^O(ii, 1)$

If $P_i^O(ii, 1) = P_i^O(ii, 2)$

$P_j^O(ii, 2)$ 以概率 $P_{T \rightarrow F}(i, j)$ 被改变为一个混淆类, 从而得到 $P_j^O(ii, 2)$, 混淆类由 M_j^C 而定

Elseif

$P_j^O(ii, 2)$ 以概率 $P_{F \rightarrow T}(i, j)$ 被改变为正确的分类 $P_i^O(ii, 1)$, 得到 $P_j^O(ii, 2) = P_i^O(ii, 1)$

Endif

End ii

3.2 多分类器输出生成算法

在 2 个分类器的生成算法中, 无论采用哪个分类器作为基分类器, 都不会给模拟结果带来影响. 对多个分类器而言, 如果仍固定地以一个分类器为基础, 则随后生成的分类器之间的关联度就无法得以保证. 为了克服此问题, 可以针对每一输出随机生成 $\{R_1, R_2, \dots, R_l\}$ 的一个置换 $\{R_{w_1}, R_{w_2}, \dots, R_{w_l}\}$, 按

照该置换给定的顺序产生数据, 即以 R_{w_1} 为基分类器生成 R_{w_2} 的输出, 再以 R_{w_2} 为基础生成 R_{w_3} 的输出, 直至 l 个分类器的输出全部生成. 当输出的数目足够大时, 任意 2 个分类器都有足够的相邻机会, 从而可以近似地表达出二者之间的相关关系.

4 实验

设模式类别数目 $M=10$, 每类样本的数目 $N=1\ 000$. 每组实验运行 100 次, 然后取其平均值作为实验结果. 为了讨论方便, 设各分类器的识别率和其间的关联度都相同.

实验 1 分类器模拟算法.

(1) 2 个分类器的生成算法. 设 2 个分类器 R_i, R_j 识别率的取值范围为 $\{0.5, 0.6, 0.7, 0.8, 0.9\}$, 关联度 Q_{ij} 的取值范围为 $\{-1, -0.9, \dots, -0.1, 0, 0.1, \dots, 0.9, 1\}$. 实验得到的各模拟值的最大误差见表 1.

(2) 多个分类器的生成算法. 不失一般性, 以 3 个分类器为例给出实验结果及分析. 实验中 3 个分类器的识别率取值范围为 $\{0.6, 0.7, 0.8, 0.9\}$, 关联度 Q_{ij} 的取值范围为 $\{-0.8, -0.5, -0.2, 0.2, 0.5, 0.8\}$, 这样共有 24 种组合方式, 每种方式下得到的 3 个模拟识别率均值 \bar{L}^T 和 3 个模拟 Q 统计量的均值 \bar{Q} 如表 2 所示.

由表 1、表 2 可以看出: 2 个分类器的模拟算法

表 1 2 个分类器情况下的模拟值最大误差

模拟参数	模拟值最大误差				
	$L=0.5$	$L=0.6$	$L=0.7$	$L=0.8$	$L=0.9$
L_i^T	$\pm 0.001\ 1$	$\pm 0.000\ 9$	$\pm 0.001\ 1$	$\pm 0.001\ 0$	$\pm 0.000\ 6$
L_j^T	$\pm 0.001\ 1$	$\pm 0.001\ 2$	$\pm 0.000\ 7$	$\pm 0.000\ 9$	$\pm 0.000\ 8$
Q_{ij}	$\pm 0.005\ 8$	$\pm 0.002\ 9$	$\pm 0.002\ 5$	$\pm 0.003\ 7$	$\pm 0.013\ 5$

表 2 3 个分类器情况下模拟识别率均值和模拟 Q 统计量均值

Q	\bar{L}^T				\bar{Q}			
	$L=0.6$	$L=0.7$	$L=0.8$	$L=0.9$	$L=0.6$	$L=0.7$	$L=0.8$	$L=0.9$
-0.8	0.600 2	0.699 9	0.800 2	0.899 9	-0.466 5	-0.456 0	-0.443 5	-0.428 9
-0.5	0.600 1	0.699 9	0.800 1	0.900 1	-0.302 8	-0.298 4	-0.295 7	-0.291 3
-0.2	0.599 8	0.700 3	0.800 0	0.899 9	-0.128 6	-0.127 1	-0.125 9	-0.127 8
0.2	0.599 7	0.699 9	0.800 1	0.899 8	0.141 8	0.141 2	0.139 1	0.140 7
0.5	0.600 0	0.700 0	0.799 9	0.899 9	0.389 5	0.393 0	0.390 1	0.393 5
0.8	0.600 4	0.700 6	0.800 1	0.900 1	0.709 5	0.711 8	0.711 6	0.716 1

能精确地给出所有结果;多个分类器的模拟算法能精确模拟识别率.对于关联度,算法自身的特性决定了模拟值偏低,这可通过事先设置较高的关联度来产生需要的数据.

实验2 关联度对多个分类器系统性能的影响.

为了观察分类器之间的关联性对多分类器系统性能的影响,针对2种常用的组合方法进行了实验研究,即多数投票和堆叠泛化[6].用多分类器模拟算法产生了具有相同识别率和关联度的3个分类器的12组模拟数据,如表3所示,其中模式类别数目为10,每类样本数目为1000.

表3 模拟数据集

Q	L^T		Q	
	L=0.8	L=0.6	L=0.8	L=0.6
-0.9	0.8	0.6	-0.49	-0.52
-0.5	0.8	0.6	-0.29	-0.30
-0.2	0.6	0.8	-0.13	-0.12
0.2	0.8	0.6	0.14	0.14
0.5	0.8	0.6	0.39	0.39
0.9	0.8	0.6	0.84	0.84

对于每一组数据,取出2/3的样本作为训练集,剩余的1/3作为测试集.在堆叠泛化中,0级分类器使用整个训练集进行训练,1级分类器的训练数据则由0级分类器通过2分交叉验证方法(2-fold cross validation)获得.实验结果如图1、图2所示,从中可以看出,在2种组合方法中,无论参与组合的分类器性能如何,多分类器系统的性能总是随着关联程度的增加而降低,且这种影响在堆叠泛化中的作用要比在多数投票中更为显著.以单分类器的识别率取0.8为例,当关联度从0.8295降至-0.4847时,多数投票系统的识别率从0.8239增至0.9473,性能提高了14.98%,堆叠泛化系统的识别率从0.6647增至0.9438,性能提高了41.99%.

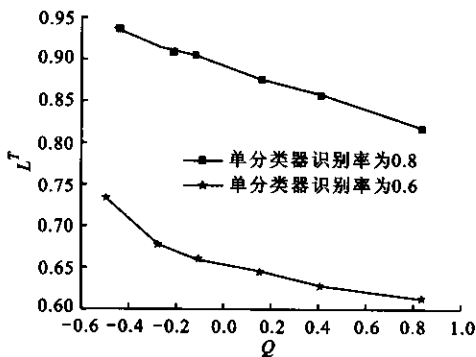


图1 关联度对多数投票的影响

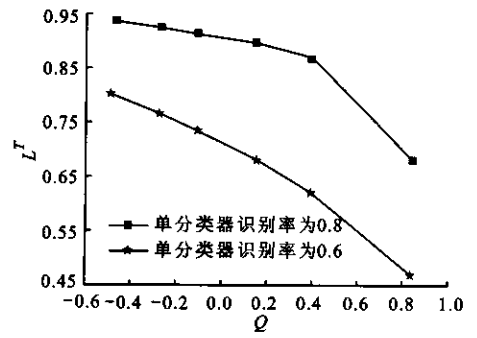


图2 关联度对堆叠泛化的影响

5 结论

为了更充分地评估多分类器组合方法,本文提出了一种能够表达分类器之间相关性的分类器模拟算法,通过建立一个分类器的混淆矩阵生成其模拟输出,结合分类器之间的关联生成所有模拟数据.应用此算法产生的模拟数据对多数投票和堆叠泛化2种组合方法进行了评价,结果表明分类器之间的负相关能够提高系统的性能.

参考文献:

- [1] Parker J R. Evaluating classifier combination using simulated classifiers [R]. Research Report, # 2000/659/11. Calgary, Canada: Department of Computer Science, University of Calgary, 2000.
- [2] Lecce V D, Dimauro G, Guerriero A, et al. Classifier combination: the role of a-priori knowledge [A]. Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition [C]. Amsterdam, Netherlands: International Unipen Foundation, 2000, 143-152.
- [3] Zouari H, Heutte L, Lecourtier Y, et al. A new classifier simulator for evaluating parallel combination methods [A]. Proceedings of the Seventh International Conference on Document Analysis and Recognition [C]. New York: IEEE Computer Society, 2003.
- [4] Kuncheva L I, Kuncheva R K. Generating classifier outputs of fixed accuracy and diversity [J]. Pattern Recognition Letters, 2002, 23(5): 593-600.
- [5] Kuncheva L I, Whitaker C J. Measures of diversity in classifier ensembles [J]. Machine Learning, 2003, 51(2): 181-207.
- [6] Wolpert D. Stacked generalization [J]. Neural Network, 1992, 5(2): 241-259.