* 1

一种新的动态 SVM 选择集成算法

廖　勇[1,2]，　王晓丹[1]，　齐俊杰[2]
(1. 空军工程大学　导弹学院，陕西　三原　713800；2. 95824 部队，北京　100195)

摘　要：针对动态选择集成算法存在当局部分类器无法对待测样本正确分类时避免错分的问题，提出基于差异聚类的动态 SVM 选择集成算法。算法首先对训练样本实施聚类，对于每个聚类，算法根据精度及差异度选择合适的分类器进行集成，并根据这些分类器集成结果为每个聚类标定错分样本区，同时额外为之设计一组分类器集合。在测试过程中，根据待测样本所属子聚类及在子聚类中离错分样本区的远近，选择合适的分类器集合为之分类，尽最大可能的减少由上一问题所带来的盲区。在 UCI 数据集上与 Bagging－SVM 算法及文献[10]所提算法比较，使用该算法在保证测试速度的同时，能有效提高分类精度。

关键词：差异聚类；支持向量机；动态集成
DOI：10.3969/j.issn.1009－3516.2010.05.006
中图分类号：TP391.4　　**文献标识码**：A　　**文章编号**：1009－3516(2010)05－0026－05

动态选择集成是选择性集成的重要手段之一，它根据不同的待测样本选择适合的集成，它通常基于一个特殊“函数”：对每个待测样本，这个函数都选择最有可能对其正确分类的分类器集合。很多学者对动态选择集成算法做了深入研究[1－7]，并提出了多种动态选择集成策略，其中最流行的选择策略有先验局部精度选择(A Priori Selection)、后验局部精度选择(A Posteriori Selection)、总体局部精度(Overall Local Accuracy)和局部类精度(Local Class Accuracy)。然而，在这些选择策略中，都存在同一个问题，即当局部分类器无法对待测样本正确分类时，这种错分就无法避免[5]。

针对这一问题，本文提出了基于差异聚类的动态支持向量机(Support Vector Machines, SVM)选择集成算法。该算法的关键点在于根据分类器集成对由训练样本生成的若干聚类进行评价，为每个聚类标定错分样本区，并额外为之设计一组分类器集合，当待测样本离此区域很近或属于此区域时，使用此分类器集合为之分类，尽最大可能地减少由上一问题所带来的盲区。

1　动态选择集成思想

动态选择集成思想涉及全局精度和局部精度 2 个概念。全局精度是指分类器在整个样本集合上的分类准确率，局部精度是指分类器在样本空间中某个局部区域上的分类精度。

把样本空间 R^d 分成 K 个区域即 $R_1, R_2, \cdots, R_K, K>0$，从分类器集合 $H=\{h_1, h_2, \cdots, h_L\}$ 中为每个区域 R_j 指定若干个分类器。设 h^* 是 R^d 中具有最高全局精度的分类器，$h^* \in \mathrm{H}$，$P(h_i \mid R_j)$ 为 h_i 分类器在区域 R_j 上正确分类的概率，即局部精度。$\{h^j_{i_1}, h^j_{i_2}, \cdots, h^j_{i_E}\} \subset H$ 为对区域 R_j 分类的分类器子集合，则整体正确分类的概率 P_c 为：

$$P_c = \sum_{j=1}^{K}\left(P(R_j)\sum_{e=1}^{E}P(h^j_{i_e} \mid R_j)\right) \tag{1}$$

* **收稿日期**：2010－06－22
基金项目：国家自然科学基金资助项目(60975026)
作者简介：廖　勇(1976－)，男，河南信阳人，副教授，博士生，主要从事智能信息处理、模式识别等研究；
E－mail：qige_svm@126.com
王晓丹(1966－)，女，陕西汉中人，教授，博士生导师，博士(后)，主要从事智能信息处理、模式识别等研究.

式中 $P(R_j)$ 为测试样本 x 取自 R_j 的概率。为了最大化 P_c,指定 $\{h^j_{i_1}, h^j_{i_2}, \cdots, h^j_{i_E}\} \subset H$,使得:

$$\sum_{e=1}^{E} P(h^j_{i_e} \mid R_j) \geqslant P(h_l \mid R_j) \quad , \forall l = 1,2,\cdots,L \tag{2}$$

由式(1) - (2) 得:

$$P_c \geqslant \sum_{j=1}^{K} P(R_j) P(h^* \mid R_j) = P(h^*) \tag{3}$$

通过式(3) 可以看出,选择局部精度高的分类器集合 $\{h^j_{i_1}, h^j_{i_2}, \cdots, h^j_{i_E}\} \subset H$,其分类性能等于或好于具有最好全局精度的分类器 h^*。

选择局部精度高的分类器,即动态分类器选择思想(Dynamic Classifier Selection),相对于选择局部精度高的分类器集合可能在算法上要简单一些,但是其精度需依赖于我们对这些分类器泛化估计的信任程度[8],而选择局部精度高的分类器集合,不仅最终的分类精度没有降低,而且也可分散产生过拟合的风险。

2　基于差异聚类的动态 SVM 选择集成(DC - SVM)思想

2.1　"分类盲区"问题

动态选择思想是针对待测样本所属区域进行操作,对区域内样本的判别直接影响着对待测样本的最终判决。假设这一区域的某些样本被局部分类器(或局部分类器集合)错分,那么这些样本将被错分,且无法避免,其周围的样本可能被错分的概率也相对较大。见图 1,假设样本空间中待测样本所属区域 R_i,在此区域中三角形表示被局部分类器(或局部分类器集合)错分的样本,这些样本一旦被错分,那么在以后的识别过程中,无论这些样本被局部分类器识别多少次,这些样本将还会被错分,我们称此问题为"分类盲区"问题。

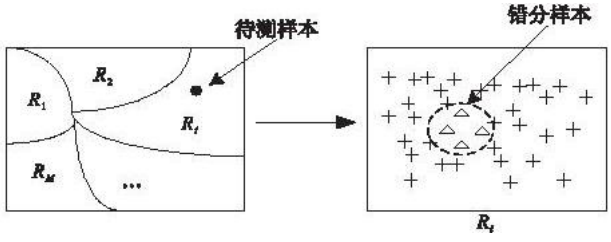


图 1　错分样本区

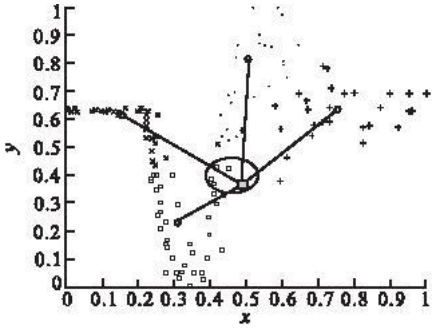
Fig. 1　Sample areas which are misclassified

2.2　基于差异聚类的动态 SVM 选择集成思想

基于差异聚类的动态 SVM 选择集成算法在"分类盲区"上重新训练一组分类器,有效缩减了"分类盲区"的范围,其算法思想是:

首先利用 boostrap 方法提取训练样本训练多个 SVM 分类器;对训练样本进行初始聚类,使性质相似的样本尽可能聚为一类。然后对于每个聚类,根据 SVM 分类器在聚类中的局部精度及相互之间的差异度,选择合适的分类器集成;把各聚类样本作为验证样本,用已生成的集成对其分类,确定错分样本,并依此选择对错分样本分类精度高且差异性大的分类器集合。

对于新样本 x,见图 2,圆形为各个聚类中心,首先根据新样本到各个聚类中心的距离确定其所属聚类,再确定其邻域是否存在错分样本,以选择合适的分类器集合对其进行分类。



2　基于聚类的动态 SVM 选择集成示意图

Fig. 2　Dynamic SVM selected ensemble based on clustering sketch map

3　基于差异聚类的动态 SVM 选择集成算法

输入:训练集 $\{(x_i, y_i)\}_{i=1}^{n} \in S$;测试集 T;SVM 基分类器参数 σ, C;K - mean 聚类参数 k;初始生成的基分类器个数 L。

输出:测试集的集成分类结果。

步骤 1　对训练集 S 使用 bootstrap 方式训练生成 L 个 SVM,$h_1, h_2, \cdots, h_L$。

步骤 2 不理会类标记,把整个训练集 S 当作验证集进行 K – mean 聚类,生成 $C_1, C_2, \cdots, C_K$,及类中心 $v_1, v_2, \cdots, v_K$。

步骤 3 用 L 个 SVM 基分类器对每个聚类中的样本进行评价,生成 K 个标记分类对错的 0 – 1 矩阵 $\boldsymbol{D}_1$, $\boldsymbol{D}_2, \cdots, \boldsymbol{D}_K$:

$$\boldsymbol{D}_i = \begin{bmatrix} d_{11} & \cdots & d_{1P} \\ \cdots & & \cdots \\ d_{L1} & \cdots & d_{LP} \end{bmatrix}, \quad i = 1, 2, \cdots, K \tag{4}$$

式中:P 为每个聚类中的训练样本数;d_{lp} 表示第 l 个分类器对第 i 个聚类中的 p 样本分类结果:$d_{lp} = \begin{cases} 1 & \text{分类杖} \\ 0 & \text{分类次} \end{cases}$。下面针对每个聚类操作:

1)根据 D_i,计算 L 个分类器的精度和差异度,并将分类器分别依精度和差异度按降序排列在此处,每个分类器的精度用双误度量公式[8-9]来表示:

$$P_i = -\sum_{j=1, j\neq i}^{L} E_{i,j} \tag{5}$$

$$E_{i,j} = \frac{N^{00}}{N^{00} + N^{01} + N^{10} + N^{11}} \tag{6}$$

式中:N^{00} 表示 2 个分类器同时分类错误的样本个数;N^{11} 表示 2 个分类器同时分类正确的样本个数;N^{10} 表示第 1 个分类器分类正确,而第 2 个分类器分类错误的样本数;N^{01} 表示第 1 个分类器分类错误,而第 2 个分类器分类正确的样本数。分类器间的差异度由不一致度量公式[9]衡量:

$$M_{i,j} = \frac{N^{01} + N^{10}}{N^{00} + N^{01} + N^{10} + N^{11}} \tag{7}$$

2)从生成的精度和差异度序列中选择出前$(L/2)$个分类器,分别为 $H_i^{1.\mathrm{acc}}$ 和 $H_i^{1.\mathrm{div}}$,取 $H_i^1 = H_i^{1.\mathrm{acc}} \cap H_i^{1.\mathrm{div}}$ 为最终局部分类器组。

注:如果 H_i^1 中的分类器个数为偶数,则在 H_i^1 中加入精度序列中的第$(L/2)+1$ 个分类器。

3)根据 $\boldsymbol{D}_i$,寻找利用 H_1 根据多数投票融合策略组成的集成对聚类中的验证样本错分的样本,形成错分样本集合 C_i^e。

4)根据 $\boldsymbol{D}_i$,计算 L 个分类器对错分样本集合 C_i^e的分类精度和差异度,并对分类器分别依精度和差异度进行降序排列。

5)从生成的精度和差异度序列中选择出前$(L/2)$个分类器,分别为 $H_i^{2.P}$ 和 $H_i^{2.M}$,取 $H_i^2 = Q_i^{2.P} \cap Q_i^{2.M}$ 为最终错分局部分类器组。

注:如果 H_i^2 中的分类器个数为偶数,则在 H_i^2 中加入精度序列中的第$(L/2)+1$ 个分类器。

步骤 4 通过**步骤 3**,每个聚类都形成 2 个分类器组:H_i^1、H_i^2,根据多数投票融合策略组成分类器集成 f_i^1 f_i^2。

测试阶段:

步骤 5 对于新测试样本 $x \in T$,找出与其距离最近的聚类中心 $v_{i^*} = \arg\min_i (d(x, v_i))$,式中 $d(\cdot, \cdot)$ 使用欧氏距离。

步骤 6 在所属聚类中,找到新样本的近邻样本,为简单起见,这里只寻找到距其最近的样本 $x^* \in C_{i^*}$,并检测 x^* 是否属于所属聚类中的错分样本集 $C_{i^*}^e$。

如是,用 $f_{i^*}^1$ 集成对待测样本进行分类;否则用 $f_{i^*}^2$ 集成对其分类。

4 实验分析

本文所提 DC – SVM 算法中,设聚类数 $k = 4$,为了验证上述所提算法的有效性,使用 Bagging – SVM, Single – SVM,以及文献[10]中所提出的 CSv – SVM 算法作参考。为便于比较,对不同数据、不同算法采用同样的参数组合,即所有算法所选择的 SVM 基分类器的核函数选择为高斯核函数 $K(x, x_j) = \exp\{-\| x - x_i \|^2 / 2\sigma^2\}$,核参数 $\sigma = 4$, $C = 100$。原始基分类器数定为 31。对于文献[10]中的算法,选择高精度分类器数

$N=15$,在高精度分类器中选择具有较大差异的分类器数 $J=9$。

选用 UCI 机器学习测试数据集中的 5 个数据集进行实验,有关数据集的描述见表 1。为提高实验结果的可靠性,每次实验中,对各组数据采用随机采样的形式划分训练集和测试集,训练集取全集的 2/3,余下的作为测试集使用,取 20 次实验的平均精度和方差作为衡量指标。

表 1　UCI 数据集

Tab. 1　UCI data set

数据集	个数	类别	数据集特征	
			数值型	名义型
Ionosphere	351	2	34	–
Image Segmentation	2 310	6	19	–
Autos	205	7	15	10
Pima – indians	768	2	8	–
Hepatitis	155	2	6	13

表 2 中给出了采用不同数据集,其中的数据分别对应着各个算法对每个数据集的平均分类精度(CM)和方差值(SD)。为说明各种集成算法的性能,将分类性能优于 Single – SVM 的集成称作有效集成,称实验中对某个数据集得到最优性能的集成为最优集成。表 2 中加框数据对应有效集成、黑粗体数据对应最优集成。通过表 2 可以看到,当基分类器的数目足够大时,简单 Bagging – SVM 精度要高于 Single – SVM;本文提出的 DC – SVM 算法精度要好于文献[10]所提出的 CSv – SVM 算法及 Single – SVM 算法,与具有高数目基分类器的 Bagging – SVM 算法相当,证明了此算法的有效性。

表 2　各种分类方法精度比较

Tab. 2　Comparison of classification methods

数据集	Single – SVM		Bagging – SVM		CSv – SVM		DC – SVM	
	精度	方差	精度	方差	精度	方差	精度	方差
Ionosphere	95. 726	–	96. 496	1. 023 3	**96. 581**	0. 697 86	97. 009	0. 726 36
Image Segmentation	91. 691	–	**91. 976**	0. 213 6	**91. 898**	0. 174 4	92. 184	0. 169 5
Autos	84. 058	–	85. 362	2. 100 2	83. 768	2. 800 3	84. 783	3. 222 6
Pima – Indians	78. 125	–	**78. 594**	0. 684 06	**78. 477**	0. 723 8	78. 828	0. 731 95
Hepatitis	76. 923	–	**79. 615**	2. 432 5	**80. 192**	2. 407	81. 346	3. 274 9

表 3 为 4 种算法分别在 5 个数据集上的运行时间。

表 3　各种分类方法运行时间比较

Tab. 3　Comparison of run time of classification methods　　s

数据集	Bagging – SVM		CSv – SVM		DC – SVM	
	训练	测试	训练	测试	训练	测试
Ionosphere	30. 109	4. 562 5	42. 753	1. 698 4	42. 753	1. 054 7
Image Segmentation	5 890. 9	159. 98	6 336. 2	57. 641	6 336. 2	58. 914
Autos	9. 273 4	2. 310 9	15. 469	0. 857 81	15. 469	0. 532 81
Pima – Indians	200. 26	33. 381	290. 27	11. 903	290. 27	13. 113
Hepatitis	5. 053 1	0. 878 13	7. 478 1	0. 334 37	7. 478 1	0. 248 44

从表中可以看出,本文提出的 DC – SVM 和文献[10]所提出的 CSv – SVM 算法虽然训练时间比较长,但在测试过程中,其运行时间相对较短,且与测试样本的数量呈线性关系。另外,从表 2 和表 3 可以看出,本文提出的 DC – SVM 算法较 Bagging – SVM 算法,在保证精度的同时,测试时间大幅度减少,和 CSv – SVM 算法相比,测试时间相当,但精度有所提升。

5　结束语

本文提出了一种基于差异聚类的动态 SVM 选择集成算法,利用它可以改善当局部分类器无法对待测样本正确分类时无法避免错分的情况,从而提高了集成的分类性能。通过对标准 UCI 数据集的测试结果验证了此算法的性能。

参考文献:

[1]　Ko A, Sabourin R, Britto Jr A. From Dynamic Classifier Selection to Dynamic Ensemble Selection[J]. Pattern Recognition,

2008,41 (5): 1718 - 1731.
[2] Giacinto G, Roli F. Dynamic Classifier Selection Based on Multiple Classifier Behaviour[J]. Pattern Recognition, 2001, 34 (9): 1879 - 1881.
[3] Kuncheva L I. Cluster - and - selection Model for Classifier Combination[C]//Proceedings of International Conference on Knowledge Based Intelligent Engineering Systems and Allied Technologies. Sussex UK: University of Brighton, 2000: 185 - 188.
[4] Liu R, Yuan B. Multiple Classifier Combination by Clustering and Selection[J]. Information Fusion, 2001, 2 (3): 163 - 168.
[5] Shin H W, Sohn S Y. Seleeted tree Classifier Combination Based on Both Accuracy and Error Diversity[J]. Pattern Recognition, 2005, 38(2): 191 - 197.
[6] Woods K, Kegelmeyer W P, Bowyer K. Combination of Multiple Classifiers Using Local Accuracy Estimates[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19 (4): 405 - 410.
[7] 张健沛,程丽丽. 基于全信息相关度的动态多分类器融合[J]. 计算机科学, 2008,35(3): 188 - 190.
ZHANG Jianpei, CHENG Lili. Dynamic Multiple Classifiers Combination Based on Full Information Correlation[J]. Computer Science, 2008, 35(3): 188 - 190. (in chinese)
[8] Kuncheva L I. Switching between Selection and Fusion in Combining Classifiers: An Experiment[J]. IEEE Transactions on Systems Man and Cybernetics - part B: Cybernetics, 2002, 32(2): 146 - 156.
[9] Kuncheva L, Whitaker C. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy [J]. Machine Learning, 2003, 51(2): 181 - 207.
[10] Canuto A M P, Soares R G F, Santana A, et al. Using Accuracy and Diversity to Select Classifiers to Build Ensembles[C]// Proceedings of International Joint Conference on Neural Networks. Canada: Sheraton Vancouver Centre Hotel, 2006: 2289 - 2295.

(编辑:田新华)

A New Dynamic SVM Selected Ensemble Algorithm

LIAO Yong[1,2], WANG Xiao - dan[1], QI Jun - jie[2]
(1. Missile Institute, Air Force Engineering University, Sanyuan 713800, Shaanxi, China; 2. Unit 95824, Beijing 100195, China)

Abstract: Dynamic Selection of integration algorithm is usually accompanied with the situation that there is no way to avoid the misclassification when the local classifier can not classify the test pattern correctly, accordingly a novel dynamic SVM selection ensemble algorithm based on diversity - clustering is proposed. Clustering is applied to training samples firstly in this method. To every clustering, appropriate classifier ensemble is selected based on accuracy and diversity, and the sample areas which are misclassified by the classifier ensemble for every clustering is demarcated, and a set of classifier ensemble for it is designed. During testing, the test sample is classified by the appropriate classifier ensemble based on the clustering to which it belongs and the distance between it and the misclassified sample areas. Using this method can remarkably reduce the blind regions while the test sample is very close to the misclassified areas mentioned above. Experimental results show the effectiveness of this method. Compared with Bagging - SVM and literature [10] on UCI data set, the testing speed can be guaranteed and simultaneously the classification accuracy can be effectively improved by using this algorithm.

Key words: diversity - clustering; support vector machine; dynamic ensemble