文章编号:1001 - 9081(2008)08 - 2091 - 03

集成学习中基于离散化方法的基分类器构造研究

蔡 铁,伍 星,李 烨
(深圳信息职业技术学院 信息技术研究所,广东 深圳 518029)
(cai.tie@163.com)

摘 要:为构造集成学习中具有差异性的基分类器,提出基于数据离散化的基分类器构造方法,并用于支持向量机集成。该方法采用粗糙集和布尔推理离散化算法处理训练样本集,能有效删除不相关和冗余的属性,提高基分类器的准确性和差异性。实验结果表明,所提方法能取得比传统集成学习算法 Bagging 和 Adaboost 更好的性能。

关键词:集成学习; 基分类器; 离散化; 支持向量机集成
中图分类号: TP181 **文献标志码**:A

Research on construction of base classifiers based on discretization method for ensemble learning

CAI Tie, WU Xing, LI Ye
(*Institute of Information Technology, Shenzhen Institute of Information Technology, Shenzhen Guangdong* 518029, *China*)

Abstract: Construction method of base classifiers based on data discretizaion was proposed to produce individual classifiers with good diversity in ensemble learning. And then it was used in support vector machines ensemble. Using the rough sets and Boolean reasoning algorithm to process the training samples, this method can eliminate the irrelative and redundant attributes to improve the accuracy and diversity of base classifiers. Experimental results show that the presented method can achieve better performance than the traditional ensemble learning methods such as Bagging and Adaboost.

Key words: ensemble learning; base classifiers; discretization; Support Vector Machine (SVM) Ensemble

0 引言

集成学习是近年来机器学习领域的研究热点之一,它通过训练和组合多个准确而有差异的分类器,可以显著提高学习系统的泛化能力。集成学习包括基分类器的构造和基分类器的组合两个部分,其中基分类器的构造极其重要。文献[1]提出分类器有效集成的关键是各基分类器应当是有差异的,文献[2]也提出分类器集成有效的充分必要条件是个体分类器应具有较高的正确率及一定的差异性。这种差异性要求各基分类器是相互独立的,如果基分类器之间是负相关的,则集成可以获得更好的泛化性能[3]。基于这种思想,国内外研究人员提出了多种基分类器的构造方法。从不同训练样本子集的角度,文献[4]提出了 Bagging 方法,文献[5]提出了基于 Boosting 技术的 Adaboost 方法;从不同特征子集的角度,文献[6]提出了随机特征子空间方法构造基分类器;从学习机内部构造的角度,文献[7]提出将随机性引入基分类器的参数设置;从输出编码的角度,文献[8]提出了纠错输出编码方法。此外,还可以将上述多类方法进行结合,提高基分类器的差异性[9]。

本文从构造有差异的基分类器出发,采用基于粗糙集与布尔推理的离散化算法(Rough Set and Boolean Reasoning Approach, RSBRA)[10]处理训练样本集,提出了基于 RSBRA 数据离散化的基分类器构造方法,并用于支持向量机集成(SVM Ensemble)算法中,可有效提高集成学习的性能。

1 基于离散化方法构造基分类器

数据离散化是一个将连续属性转化为离散属性的过程。虽然支持向量机可以直接处理连续数据属性,但是在数据离散化过程中,如果选取的断点集合不同,会产生不同的离散化结果,因此可以采用不同的离散化数据训练支持向量机,构造有差异的基分类器。

离散化的实质是将数据空间划分为有限个区域,每个区域中的对象具有相同的类别。一个典型的离散化过程包括确定候选断点、从候选断点中选择实际断点、以及利用实际断点对初始数据离散化三个部分。设决策表 $T = (U, C \cup D, V, f)$,其中 U 为非空有限对象集,$C \cup D$ 为非空有限属性集,C 为条件属性,$D = \{d\}$ 为决策属性,V 是各属性值域的并集,f 为信息函数。对于 $\forall a \in C$,值域 $V_a = [l_a, r_a)$,P_a 为 V_a 上的一个划分,即:

$$P_a = \{[c_0^a, c_1^a), [c_1^a, c_2^a), \cdots, [c_k^a, c_{k+1}^a)\} \quad (1)$$

其中,$l_a = c_0^a < c_1^a < \cdots < c_{k+1}^a = r_a$,$V_a = [c_0^a, c_1^a) \cup [c_1^a, c_2^a) \cup \cdots \cup [c_k^a, c_{k+1}^a)$。$c_i^a (i = 1, 2, \cdots, k)$ 称为一个断点,$C_a = \{c_1^a, c_2^a, \cdots, c_k^a\}$ 为 V_a 的断点集,每一个断点集 C_a 唯一确定了一个划分 P_a。任意的 $P = \{P_a : a \in C \cup D\}$ 定义了一个新决策表 $T^P = (U, C^P \cup D, V^P, f^P)$,这称为 T 的 P 离散化。

基于粗糙集和布尔推理的离散化方法是一种有效的启发式监督离散化算法,它采用最大分辨能力启发式方法直接实现[10]。RSBRA 的直接实现复杂度很高,在实际应用中,必须采用 RSBRA 的高效实现算法,它可降低时间和空间复杂度,适用于大数据集的离散化。高效的 RSBRA 实现算法以分辨不同类别样本的能力(能分辨的样本数)为标准选择实际断点,当决策表中的全部样本能够由已选择的断点分辨时结束断点的选择。在 r 类分类问题中,给定样本子集 $X \subseteq U$,对于

收稿日期:2008 - 03 - 21;**修回日期**:2008 - 06 - 30。
基金项目:国家自然科学基金资助项目(60772163);深圳市科技计划项目(SZKJ0708)。
作者简介:蔡铁(1977 -),男,湖南长沙人,讲师,博士,主要研究方向:信号处理、模式识别、语音处理与识别; 伍星(1980 -),女,湖南娄底人,讲师,硕士,主要研究方向:网络信息系统、信息安全; 李烨(1974 -),男,湖南新化人,讲师,博士,主要研究方向:模式识别、机器学习。

某一断点 c_m^a，$(a \in C, 1 \leqslant m \leqslant n_a)$，$n_a$ 为样本子集的样本数目，令：

$$l_j^X(c_m^a) = |\{x \in X \mid [a(x) < c_m^a] \wedge [d(x) = j]\}| \quad (2)$$

$$r_j^X(c_m^a) = |\{x \in X \mid [a(x) > c_m^a] \wedge [d(x) = j]\}| \quad (3)$$

其中 $j = 1,2,\cdots,r$，则 c_m^a 能分辨的 X 中的样本数为：

$$W^X(c_m^a) = \sum_{j=1}^{r} l_j^X(c_m^a) \cdot \sum_{j=1}^{r} r_j^X(c_m^a) - \sum_{i=1}^{r} [l_i^X(c_m^a) \cdot r_i^X(c_m^a)] \quad (4)$$

对于集合内的所有子集，c_m^a 能分辨的总的样本数为：

$$W_P(c_m^a) = W^{X_1}(c_m^a) + W^{X_2}(c_m^a) + \cdots + W^{X_n}(c_m^a) \quad (5)$$

RSBRA 的高效实现算法如下。

输入：一致决策表 T。

输出：实际断点集 P。

步骤：

1）令决策表集 $L = \{U\}$，$X = L$，实际断点集 $P = \varnothing$。将 L 中属性按属性值 $a(x)$ 大小排序，得序列 $v_1^a < v_2^a < \cdots < v_{n_a}^a$，其中 $\{v_1^a, v_2^a, \cdots, v_{n_a}^a\} = \{a(x) \mid x \in U\}$，$n_a$ 为样本数目，则 a 上所有得候选断点可取为属性值的平均值：

$$C_1 = \cup_{a \in C} C_a = \cup_{a \in C} \left\{ c_i^a = \frac{v_i^a + v_{i+1}^a}{2}, 1 \leqslant i \leqslant n_a - 1 \right\} \quad (6)$$

2）对 $\forall c \in C_1$，计算 $W_P(c)$。

3）选择 W_P 值最大的候选断点 $c_{\max}$ 加入到 P 中，并从 C_1 中删除该断点。

4）对于 $X \in L$，若 $c_{\max}$ 将 X 分割为 X_1 和 X_2，那么从 L 中删除 X，并将 X_1 和 X_2 加入到 L 中。

5）若对于 $\forall X_i \in L$，X_i 中所有样本属于同一类别，则算法结束，否则转到2）。

2 在支持向量机集成中的应用

为了构造支持向量机集成，需要多个基分类器的训练集。本文通过选取不同的断点集，产生不同的离散化数据来形成不同的训练集合。由于 RSBRA 算法在每选择一个实际断点都重新计算各剩余断点的分辨能力，但算法结束时，剩余断点的分辨能力都为0。因此，可根据一个断点的分辨能力和一个预设的基本概率，来确定该断点增加到实际断点集的概率。在产生断点集时，以剩余断点对原始决策表的分辨能力为依据结合基本概率进行断点集选择，从而产生不同的断点集。

基于 RSBRA 离散化方法构造基分类器的 SVM 集成算法（RSBRA SVM Ensemble，RSVMEN）描述如下：

输入：训练集 $S = \{(x_i, y_i)\}$，$i = 1,2,\cdots,m$；学习机 L，基分类器数目 T；剩余断点的基本选择概率 p。

输出：集成结果 N^*。

步骤：

1）采用 RSBRA 算法对原始数据集 S 进行离散化，获得实际断点集 P_{prac}^0，剩余断点集 P_r^0，剩余断点的分辨能力 $W_{P_r^0}$。

2）从原始数据 S 中删除在 P_{prac}^0 中无实际断点的属性，记为 S^*。

3）计算剩余断点的选择概率：$p_s = p \times W_{P_r^0} / \max(W_{P_r^0})$。

4）for $t = 1,2,\cdots,T$

①以概率 p_s 从剩余断点集 P_r^0 中选择断点组成集合 P_{add} 加入到断点集 P_{prac}^{t-1} 中，即：

$$P_{\text{prac}}^t = P_{\text{prac}}^{t-1} \cup P_{\text{add}} \quad (7)$$

②采用 P_{prac}^t 离散化数据集 S^*，得新的数据集 S^{**}；

③删除 S^{**} 中的重复样本，采用 S^{**} 训练支持向量机：$N_t = L(S^{**})$；

5）采用多数投票法组合基分类器：$N^*(x) = \arg\max_{y \in Y} \sum_{t; N_t(x) = y} 1$。

3 实验结果与分析

实验采用三个基准数据集对本文算法进行性能研究，并与支持向量机方法、Bagging 和 Adaboost 集成学习算法进行了对比实验。使用的三个基准数据集包括 UCI 数据仓库[11]中的 Wisconsin Breast Cancer Database（简称数据集 Breast）、Glass Identificaition Database（简称数据集 Glass）和 Statlag[12]中的 Heart Disease Database（简称数据集 Heart），其中 Breast 数据集删除了16个属性值不全的样本。

表1 3种数据集平均测试准确率比较 %

数据集	SVM	Bagging	Adaboost	RSVMEN
Breast	96.634	96.293	95.854	97.415
Glass	67.656	68.281	66.875	70.625
Heart	81.852	78.642	76.914	84.444

表2 Breast 数据集的 win-tie-loss 比较

方法	SVM	Bagging	Adaboost	RSVMEN
SVM		2-2-6	1-2-7	7-2-1
Bagging			1-3-6	9-1-0
Adaboost				9-0-1

表3 Glass 数据集的 win-tie-loss 比较

方法	SVM	Bagging	Adaboost	RSVMEN
SVM		5-2-3	5-1-4	6-1-3
Bagging			2-3-5	6-0-4
Adaboost				5-1-4

表4 Heart 数据集的 win-tie-loss 比较

方法	SVM	Bagging	Adaboost	RSVMEN
SVM		2-0-8	1-1-8	9-0-1
Bagging			2-2-6	9-0-1
Adaboost				10-0-0

实验共采用了单个 SVM、Bagging、Adaboost 以及 RSVMEN 四种方法进行对比研究，其中 Bagging、Adaboost 以及 RSVMEN 各包含了21个 SVM 作为基分类器，并都采用多数投票法对基分类器进行组合。每个 SVM 的参数 C 和 γ 通过对训练集进行网络搜索和5倍交叉验证确定，C 和 γ 的搜索范围分别取为 $[-4,14]$ 和 $[-14,4]$，搜索步长为2。剩余断点的基本选择概率设为0.3。实验在三个数据集上分别进行10次，每次随机选择70%的样本作为训练集，其余的作为测试集。各方法分别将十次实验的平均测试值作为最后的结果，如表2～4所示。“win”表示10次实验中当前方法的结果显著优于所比较方法的次数；“tie”表示当前方法的结果与所比较方法没有显著差异的实验次数；“loss”表示当前方法的结果显著差于所比较方法的次数。

从表1～4的实验结果可以看出，Bagging 和 Adaboost 两种集成算法并没有获得优于单个 SVM 的分类准确度，其中 Bagging 方法仅在 Glass 数据集上略优于单个 SVM，而在其余两个数据集上都比单个 SVM 差，Adaboost 方法则在三个数据集上都比单个 SVM 差。RSVMEN 方法充分挖掘了原始数据集的内在信息，通过特征选择和删除不相关的属性，可构造准

确而有差异的基分类器,有效地提高了集成学习的性能,在三个数据集上的分类准确度都远优于其他三种方法。

基分类器构造和基分类器组合的方法都影响集成学习的性能。在不同集成规模下(不同的基分类器数目),RSVMEN 算法在 Glass 数据集上的性能如图 1 所示。随着基分类器数目的增加,RSVMEN 的学习性能一直保持优于单个 SVM,表明了本文集成学习算法的优越性。但基分类器数目的增加并没有保证 RSVMEN 性能的递增,这是由于 RSVMEN 采用的是简单投票的组合方法,没有对基分类器进行选择,这也验证了文献[13]的结论,即在一组个体学习器中,进行选择性集成比用所有个体集成性能更优。在后续的研究中,将考虑使用选择性组合方法进行集成,使本文集成算法随着基分类器数目的增加而递增。

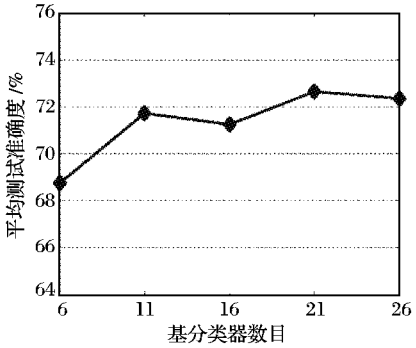

图 1　不同基分类器数目下 RSVMEN 算法的性能(Glass 数据集)

4　结语

构造准确而有差异的基分类器是提高集成学习泛化性能的关键。针对这一问题,本文基于数据离散化方法处理训练样本集,通过构造不同的断点集以生成有差异的基分类器,有效地提高了集成学习的性能。支持向量机集成的实验结果表明,本文方法明显优于单个支持向量机以及 Bagging、Adaboost 等集成学习算法。

参考文献:

[1] JI CHUAN-YI, MA SHENG. Combination of weak classifiers[J]. IEEE Transaction on Neural Networks, 1997, 8(1): 32 - 42.

[2] DIETTERICH T G. An experimental comparison of three methods for constructing ensembles of decision tress: Bagging, boosting and randomization[J]. Machine Learning, 2000, 40(2): 139 - 158.

[3] KUNCHEVA L I, WHITAKER C J, SHIPP C A, *et al*. Is independence good for combining classifiers[EB/OL]. [2007 - 12 - 23]. http://ict. ewi. tudelft. nl/ ~ duin/papers/icpr_00_cclass. pdf.

[4] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123 - 140.

[5] SCHAPIRE R E. The boosting approach to machine learning: An overview[EB/OL]. [2007 - 12 - 26]. http://www. ccls. columbia. edu/compbio/geneclass/non_html_files/Schapire_boosting_review. pdf.

[6] SKURICHINA M, DUIN R P W. Bagging , boosting and random subspace method for linear classifiers[J]. Pattern Analysis and Applications, 2002, 5(2): 121 - 135.

[7] PARMANTO B, MUNRO P W, DOYLE H R. Improving committee diagnosis with resampling techniques[C]// Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 1996, 8: 882 - 888.

[8] DIETTERICH T G, BAKIRI G. Solving multiclass learning problems via error-correcting output codes[J]. Journal of Artificial Intelligence Research, 1995, 3(2): 263 - 286.

[9] SHARKEY A J C, SHARKEY N, GERECKE U, *et al*. The "test and select" approach to ensemble combination[C]// Proceedings of 1st International Workshop on Multiple Classifier Systems. London: Springer-Verlag, 2000, 1857: 30 - 44.

[10] NGUYEN H S, SKOWRON A. Quantization of real value attributes: Rough set and boolean reasoning approach[C]// Proceedings of the 2nd Joint Annual Conference on Information Sciences, Society for Information Processing. Canada: Springer-Verlag, 1995: 34 - 37.

[11] NEWMAN D J, HETTICH S, BLAKE C L, *et al*. UCI Repository of machine learning databases[M]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.

[12] BRAZDIL P. Statlog repository[EB/OL]. [2007 - 10 - 22]. http://www. liacc. up. pt/ML/statlog/ datasets . html.

[13] ZHOU ZHI-HUA, WU JIAN-XIN, TANG WEI. Ensembling neural networks: many could be better than all[J]. Artificial Intelligence, 2002, 137(1/2): 239 - 263.

推荐 CCF 优秀博士学位论文的通知

为推动中国计算机领域的科技进步,鼓励创新性研究,促进青年人才成长,中国计算机学会(CCF)自 2006 年起设优秀博士学位论文奖。2008 年度优秀博士学位论文推荐工作即日启动,现将有关内容通知如下。

1. 参评条件:

1)本次优秀博士学位论文的评选范围为 2006 年 7 月 1 日至 2008 年 6 月 30 日期间在中国获得计算机科学与技术学科相关专业博士学位的学位论文;

2)参加评选的博士学位论文须经具有计算机科学与技术学科博士点的高校计算机学院(系)或研究机构推荐,每个具有一级学科博士点单位推荐参评学位论文不超过 2 篇,其他不具有一级学科博士点的单位限推荐 1 篇,已经参评过的论文不得再被推荐。

3)具体参评条件和约束条件见学会网站上"中国计算机学会优秀博士学位论文奖条例"。

2. 参评申报材料:

1)印刷版论文 2 份;

2)电子版论文 1 份;

3)CCF 优秀博士学位论文推荐表(必须有作者答辩时所在单位(如系、院、所等)负责人签字、单位盖章),见学会网站;

4)其他有关证明材料;

5)评审费:1000 元/篇(CCF 会员 800 元/篇)。

3. 申报材料和评审费须于 2008 年 9 月 4 日 17:00 前报送到 CCF,过期无效。

4. 评选时间安排:

1)受理:2008 年 7 月 22 日至 2008 年 9 月 4 日。

2)格式和资质审查:2008 年 9 月 5 日 - 9 月 12 日。

3)初评:2008 年 9 月 13 日 - 10 月 12 日,CCF 组织小同行专家对申报材料进行初评,从中评选出不超过 30 篇入围候选优秀博士学位论文。

4)初评公示:2008 年 10 月 13 日 - 11 月 12 日。

5)终评:2008 年 11 月 13 日 - 12 月 12 日,CCF 终评委员会进行终评,评出获奖者。获奖总数不超过 10 篇,另有不超过 5 篇论文获提名奖。

6)终评公示:2008 年 12 月 13 日 - 2009 年 1 月 12 日。

通信地址:

北京 2704 信箱中国计算机学会,邮编:100190。

联系人:

孙文韬　电话:010 - 62562503 - 20　Email:ccf-ed5@ict. ac. cn

李乐强　电话:010 - 62562503 - 14　Email:ccf-aw@ict. ac. cn