一种新的模糊多球分类算法及其集成方法*)

顾 磊 吴慧中 肖 亮

(南京理工大学计算机科学与技术学院 南京 210094)

摘 要 本文提出了一种新的模糊多球分类算法。该算法在训练阶段为每一个模式类构造多个球,覆盖其所有的训练样本,并且在识别阶段利用一个模糊隶属函数来归类测试样本。此外,在提出的分类算法的基础上,还给出了它的集成方法。最后,我们采用了四个真实数据集进行实验,实验结果表明本文提出的算法具有较好的分类性能,是一种行之有效的分类算法。

关键词 模式分类,山峰函数,模糊隶属函数,分类集成

A Novel Fuzzy Multiple Spheres Classification Algorithm and its Ensemble Method

GU Lei WU Hui-zhong XIAO Liang

(School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract In this paper, a novel fuzzy multiple spheres classification algorithm is proposed. In the training process all training samples of each class are covered by the constructed multiple spheres. Each of sphere encompasses as many samples with the same class. In the classification process a fuzzy membership function is defined to label the testing samples. Moreover, an ensemble method based on the proposed classification algorithm is presented. Finally, experiments on four real datasets show that our approach is valid and has encouraging pattern classification performance.

Keywords Pattern classification, Mountain function, Fuzzy membership function, Classification ensemble

1 引言

模式分类是一种模式识别技术,它主要是根据决策理论来建立各模式类别间的决策边界。整个模式分类算法一般分为两个阶段,即训练阶段和识别阶段,也称为学习阶段与测试阶段。训练阶段的主要任务是使用分类器划分训练样本的特征空间;而识别阶段的任务则是根据决策规则及相似性的度量指定测试样本所属的类别[1,2]。不同的模式分类算法其理论基础与设计思想是不同的。例如:K 近邻(K-Nearest Neighbor, KNN)[3]技术力图在训练样本集中找到测试样本的 K 个近邻,并指定测试样本属于 K 近邻内出现频率最高的模式类;多层感知器(Multi-Layer Perceptrons, MLP)[4]算法使用梯度下降法来最小化网络实际输出与期望输出间的差异;支持向量机(Support Vector Machines, SVM)[5]分类算法的设计中则引入了"核技巧"[6]的思想。"核技巧"可以利用一个非线性的核函数来代替欧式点积,使得原本分界模糊的两类样本在高维特征空间中变得清晰可分。

尽管 KNN, MLP, SVM 算法可以获得较好的分类效果,但是在实际应用中它们对训练样本的识别往往无法令人满意。因此,本文提出了一种新的模糊多球分类算法,从四个真实数据集上进行的实验可以看出,提出的新分类算法不仅可以识别所有的训练样本,而且对整个数据集中样本的识别也明显优于 KNN, MLP 和 SVM。此外,对于新算法我们还给出了一个有效的集成方法。

2 新的模糊多球分类算法

2.1 算法的训练阶段

我们给出新算法在训练阶段的实现步骤:

Step1 令 $C=1, K=1$ 且 N 为模式类别总数;

Step2 令 P 一个集合,集合中元素为模式类 C 中所有训练样本点;

Step3 选择操作,即从集合 P 中选择一个元素 x_C 作为球 S_K(S_K 为超球,本文中简称超球为球)的球心;

Step4 发现一个离 x_C 最近的异类样本点 y_T,这里样本点 y_T 属于模式类 T 且 $C\neq T$。令 d_1 为点 x_C 和点 y_T 间的欧式距离;

Step5 从集合 P 中寻找一个属于模式类 C 且距离 x_C 最远的样本点,且同时此点还需满足其到 x_C 的欧式距离 d_2 必须小于 d_1,若找到了这样一个最远点,则令球 S_K 的半径 R_K 为 d_2,否则令 S_K 的半径 R_K 为 $d_1/2$;

Step6 得到一个球心为 x_C、半径为 R_K 且属于模式类 C 的球 R_K,再从集合 P 中去除球 S_K 中覆盖的所有训练样本点;

Step7 如果 P 非空,则令 $K=K+1$,再转至 Step3,否则转至 Step8;

Step8 如果 $C\leqslant N$,则令 $K=K+1$ 且 $C=C+1$,再转至 Step2;

Step9 扩球操作,即对每一个球根据其密度扩大球的半径;

Step10 结束训练。

新算法在训练阶段构造多个球来覆盖训练样本点。尽管同一个模式类的训练样本点可以被几个球所覆盖,但是一个球仅仅覆盖一个模式类的样本点。因此本文提出的算法很容易实现对训练样本点的全部识别。

新算法训练阶段的第 Step3 步为选择操作,选择操作的目的是选择一个样本点作为球 S_K 的球心。这里我们利用公式(1)的山峰函数[7],选取具有最大峰值的样本点作为球心:

$$M(x_C)=\sum_{i=1}^{u} e^{-\|x_C-x_i\|^2} \tag{1}$$

其中 $x_i\in P, u$ 为集合 P 中元素的总数,$\|x_C-x_i\|$ 为点 x_C 与 x_i 间的欧式距离。

*)国家自然科学基金(No. 60672074);江苏省自然科学基金(No. BK2006569)。顾 磊 博士研究生,主要研究方向为模式识别、机器学习;吴慧中 教授,博士生导师;肖 亮 副教授,博士。

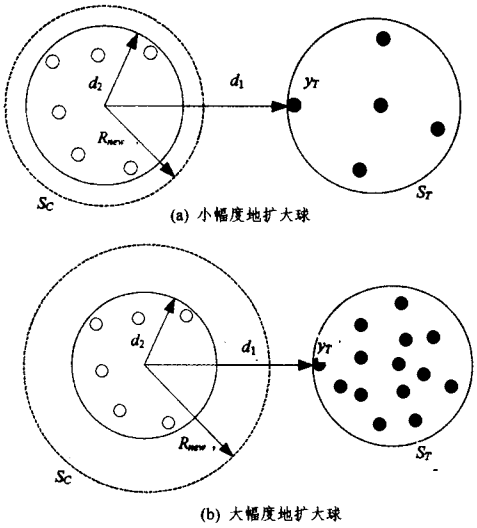


图 1 扩球操作

新算法训练阶段的第 Step9 步为扩球操作，扩球操作即对每个球，根据球的密度扩大球的半径，其目的是在两个属于不同模式类的球之间构造较好的分界面。我们给出如下公式来估计球 S_K 的密度：

$$g(S_K)=\frac{B_K}{V_K} \tag{2}$$

其中 B_K 为球 S_K 中覆盖的样本点的总数，V_K 为球 S_K 的体积。密度越小的球，其球内样本点分布得越离散；密度越大的球，样本点分布得越聚集。本文为了简化计算，直接利用 $g(S_K)=B_K/R_K$ 来计算球的密度，其中 R_K 为球 S_K 的半径。此外，球的扩大并不是无限制的，这里我们利用球本身的密度以及在构造球时第 Step4 步中找到的那个异类样本点所在球的密度来制约球的扩大。假定在构造球 S_C 时，在第 Step4 找一个异类的样本点 y_T 且 y_T 被球 S_T 所覆盖，如图 1(a)所示，若 S_T 密度较小，则 S_C 的扩大幅度应较小；否则如图 1(b)所示，若 S_T 密度较大，则 S_C 可得到较大幅度的扩大。下面给出一个扩大球的公式：

$$R_{new}=d_1\cdot\frac{g_C}{g_C+g_T}+d_2\cdot\frac{g_T}{g_C+g_T} \tag{3}$$

公式(3)中要扩大的球 S_C 其密度为 g_C，在构造 S_C 时找到一个异类点，其所在球的密度为 g_T，而 d_1 和 d_2 则可在 Step4 和 Step5 中求得。

2.2 算法的识别阶段

在识别阶段，我们定义了一个模糊隶属函数来判断测试样本对每个球的隶属度。令球 S_C 的半径为 R_C，L_C 为球 S_C 中所有样本点到球 S_C 球心的平均距离，球 S_C 中的所有样本点属于模式类 C，即 S_C 属于模式类 C，则模糊隶属函数球 $u_C(x)$ 为：

$$u_C(x)=\begin{cases}\frac{1}{3}\times\left(\frac{1}{1+\alpha((d(x,S_C)-R_C)/R_C)}\right) & \text{if } d(x,S_C)\geqslant R_C\\ \frac{1}{3}+\frac{1}{3}\times\left(\frac{1}{1+\alpha((d(x,S_C)-L_C)/L_C)}\right) & \text{if } L_C\leqslant d(x,S_C)<R_C\\ \frac{2}{3}+\frac{1}{3}\times\left(\frac{1-d(x,S_C)/L_C}{1+\beta(d(x,S_C)/L_C)}\right) & \text{if } d(x,S_C)<L_C\end{cases} \tag{4}$$

其中 $\alpha,\beta>0$，x 为一个测试样本点，$d(x,S_C)$ 为 x 到球 S_C 球心的欧式距离。公式(4)中隶属度值不仅与测试样本点到球心的距离有关，而且与球的半径以及球中所有点到球心的平均距离有关。当测试样本点落在球外时，隶属度值 u_C 的值域为(0,1/3]；当测试样本点落在平均距离与球边缘之间的区域时，隶属度值 u_C 的值域为(1/3,2/3]；当测试样本点落入平均距离以内的区域时，隶属度值 u_C 的值域为(2/3,+∞)。此外，公式(4)中本文取 $\alpha=\beta=1$，且若 $L_C=0$，则令 $L_C=R_C/2$。

3 新算法的集成方法

假定第 1 次执行上一节中介绍的新算法可以得到一个集合 $A_1=\{S_{11},S_{12},\cdots,S_{1q}\}$，其中 A_1 的每个元素表示一个新算法构造的球，q 为 A_1 中元素总数。由于新算法第 Step3 步利用山峰函数来选择球心，则第 1 次，第 2 次，…，甚至第 n 次执行新算法，所得到的集合 $A_1,A_2,\cdots,A_n$ 都是相同的，即 $A_1=A_2,\cdots,=A_n$。而若我们在第 Step3 步随机选择一个训练样本点作为球心，则每次执行新算法得到的集合 $A_i(i=1,2,\cdots,n)$互不相等，即每次执行新算法得到的球，异构性大，利于集成。我们给出新算法的集成方法如下：

① 令 $w=1$，O 为新算法执行的总次数，再令 Trn 与 Tst 分别为训练样本集和测试样本集；

② 用新算法来训练集 Trn 中的样本，其中新算法第 Step3 步采用随机选择样本点作为球心的策略；

③ 对 Tst 集中的样本进行测试，得到结果 E_w；

④ 令 $w=w+1$；如果 $w\leqslant O$，则转至②，否则转至⑤；

⑤ 对 Tst 集中的每个测试样本，根据新算法 O 次执行的结果 $E_w(w=1,2,\cdots,O)$进行投票，从而产生最终的结果。

4 实验

4.1 实验数据

我们首先从 UCI 数据库[8]中选择一个数据集，即 Balance 数据集。Balance 数据集由 625 样本组成，同样也来自三个不同的类；其次我们从 Statlog 数据库[9]选取一个数据集，即 Australian 数据集。Australian 中有 690 个样本，且数据集中的样本被分为两个不同的类别；最后我们从 LibSVM 数据库[10]中选取两个数据集，即 Diabetes 和 SVMguide2 数据集，Diabetes 中有 768 个分别来自于两个不同类别的样本，而 SVMguide2 则由 391 个样本组成，这些样本被分为三个不同的类别。

实验前，除了已被标准化的 Diabetes 和 Svmguide2 数据集之外，其它的如 Balance 和 Australian 数据集中每个数据的每个属性值都标准化到[−1,1]的范围内，这里我们采用最大-最小标准化方法。

4.2 实验结果

首先我们将提出的用山峰函数来选择球心的模糊多球分类算法（Fuzzy Multiple Spheres Classification Algorithm based the Mountain Function，FMSCA_MF）同 K-NN，MLP 和 SVM 算法在四个真实数据集（Balance，Australian，Diabetes 和 SVMguide2）上进行比较。在每个数据集上，随机抽取 60%的样本作为训练样本，整个数据集的样本作为测试样本，这样的随机抽取进行 10 次，将 10 次的识别结果进行平均，作为每个算法最终的识别结果。此外在每个数据集上，SVM 采用高斯函数且为 SVM 选择合适的核参数，K-NN 从 K=5，10，15 中选择较好的 K，MLP 由输入层、输出层和具有 10 个结点的隐藏层组成。表 1 中给出了四种算法对训练样本识别的结果，表 2 给出了四种算法对整个真实数据集的样本识别的结果，其中较高的分类正确率已用粗体表示出来。从表 1 可以看出本文提出的 FMSCA_MF 算法对训练样本可以做到 100%的识别，而这是另外三种算法无法实现的。此外，从表 2 可以看出 FMSCA_MF 在整个数据集上较 K-NN，MLP 和 SVM 有更好的识别效果。

表 1 对训练样本进行识别时四种算法获得的正确率的比较

Data sets	BPNN	K-NN	SVM	FMSCA_MF
Australian	0.8184	0.7633	0.9687	1.0000
Balance	0.9260	0.9040	0.9753	1.0000
Diabetes	0.8923	0.8282	0.9381	1.0000
SVMguide2	0.9977	0.8553	0.8529	1.0000

表 2 对整个数据集中样本进行识别时四种算法获得的正确率的比较

Data sets	BPNN	K-NN	SVM	FMSCA_MF
Australian	0.7313	0.7173	0.8182	0.8425
Balance	0.9006	0.8976	0.9516	0.9530
Diabetes	0.8245	0.7962	0.8544	0.8906
SVMguide2	0.8269	0.8188	0.8309	0.9023

表 3 三种方法分类正确率的比较

Data sets	FMSCA_MF 的分类正确率	FMSCA _RS 的平均分类正确率	集成方法的分类正确率
Australian	0.5942	0.5913	0.5942
Balance	0.8800	0.8720	0.8800
Diabetes	0.7296	0.7264	0.7427
SVMguide2	0.7179	0.7391	0.7628

其次，在每个真实数据集上，我们抽取 60%的样本作为训练样本，余下的 40%作为测试样本。在每个真实数据集上我们利用随机选择球心的方法来执行新的模糊多球分类算法(Fuzzy Multi-spheres Classification Algorithm based the Random Selection，即 FMSCA_RS)，总共执行 10 次。我们将 10 执行后集成的投票结果、FMSCA_RS10 次执行的平均结果以及在同样的数据集上使用 FMSCA_MF 算法的结果在表 3 中进行了比较，其中较高的分类正确率已用粗体表示出来。从表 3 可以看出本文提出的集成方法其分类性能好于 FMSCA_MF 和 FMSCA_RS 的平均表现。

结束语 本文提出了一种新的模糊多球分类算法，该算法在训练阶段为每一个模式类构造多个球覆盖其所有的训练样本，并且在识别阶段利用一个模糊隶属函数来归类测试样本。此外，在提出的分类算法的基础上，还给出了它的集成方法。最后，在四个真实数据集上进行实验，实验的结果表明本文提出的算法及其集成方法不仅可以识别全部训练样本，而且具有较好的分类性能。

参 考 文 献

[1] Kulkarni S R, Lugosi G, Venkatesh S S. Learning pattern classification-a survey[J]. IEEE Trans. Inf. Theory, 1998, 6(44): 2178-2201

[2] Jain A K, Duin R P W, Mao J. Statistical pattern recognition: a review[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2000, 1(22): 4-33

[3] Dasatrthy B V. Nearest neighbor(NN) norms: NN pattern classification techniques[M]. Los Alamitos, CA: IEEE Computer Society Press, 1990

[4] Qiang G Q, Zhang P. Neural networks for classification: a survey[J]. IEEE Trans. Syst. Man Cybern. Part C, 2000, 4(30): 451-458

[5] Burges J C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2: 121-167

[6] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods[M]. Cambridge, U. K: Cambridge Univ. Press, 2000

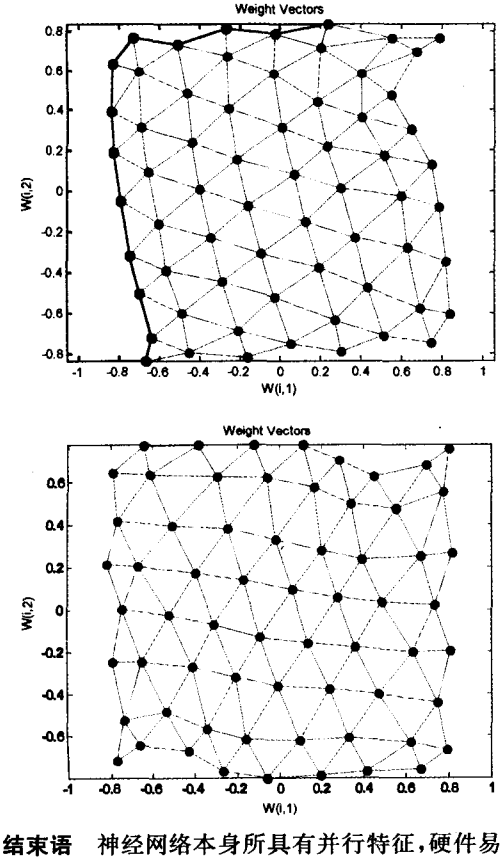
[7] Yager P R, Filev D P. Approximate clustering via the mountain method[J]. IEEE Trans. Syst. Man Cybern., 1994, 8(24): 1279-1284

[8] UCI Machine Learning Repository[DB]. http://www.ics.uci.edu/-mlearn/MLRepository.html

[9] Statlog coolection[DB]. http://www.niaad.liacc.up.pt/old/statlog/datasets.html

[10] LibSVM Website[DB]. http://www.csie.ntu.edu.tw/-cjlin/libsvmtools/dataset

(上接第 182 页)



结束语 神经网络本身所具有并行特征，硬件易实现性等，SOFMF 算法很好地克服了许多聚类算法存在的问题，在时间复杂度上具有良好的性能。针对不同的用户，可以得到不同层次的聚类结果，以便详细地进行分析。

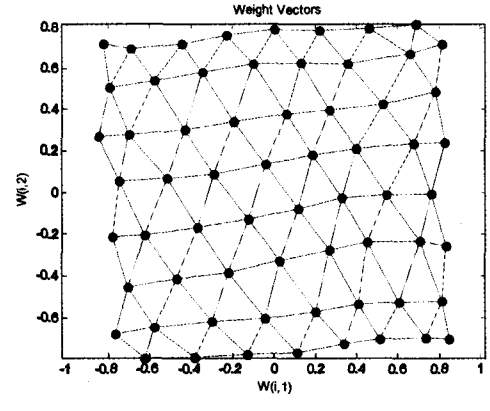


图 5 针对不同的步长，训练网络，相应的权值分布图

参 考 文 献

[1] Ezequied L R, et al. Invariant pattern identification by self-organising networks. Pattern Recognition Letters, 2001, 22: 983-990

[2] Dunkel B, Soparkar N. Data Organization and Access for Efficient Data Mining. ICDE, 1999

[3] Han J, Fu Yongjian. Mining Multiple-Level Association Rules in Large Database. IEEE Trans. on Knowledge and Data Engineering, 1999, 11(5): 798-805

[4] 刘君强，孙晓莹，潘云鹤. 关联规则挖掘技术研究的新进展. 计算机科学，2004，31(1)：110-113

[5] Bloch, Isabelle. Fuzzy relative position between objects in image processing: A morphological approach. IEEE Transactions on Patten Analysis and Machine Intelligence, 1999, 21(7): 657-664

[6] Brin S, Motwai R J D, Ullman, et al. Dynamic Itemset Counting and Implication Rules for Market Basket Data // ACM SIGMOD Conference on Management of Data. 1997: 265-276