第 36 卷 第 2 期 计 算 机 工 程 2010 年 1 月
Vol.36 No.2 Computer Engineering January 2010
·人工智能及识别技术· 文章编号：1000—3428(2010)02—0181—03 文献标识码：A 中图分类号：TP301.6

处理非平衡数据的粒度 SVM 学习算法

郭虎升[1]，亓 慧[1,2]，王文剑[1]
(1. 山西大学计算机与信息技术学院计算智能与中文信息处理教育部重点实验室，太原 030006；2. 太原师范学院计算机系，太原 030012)

摘 要：针对支持向量机对于非平衡数据不能进行有效分类的问题，提出一种粒度支持向量机学习算法。根据粒度计算思想对多数类样本进行粒划分并从中获取信息粒，以使数据趋于平衡。通过这些信息粒来寻找局部支持向量，并在这些局部支持向量和少数类样本上进行有效学习，使 SVM 在非平衡数据集上获得令人满意的泛化能力。

关键词：粒度支持向量机；非平衡数据；信息粒；局部支持向量

Granular SVM Learning Algorithm for Processing Imbalanced Data

GUO Hu-sheng[1], QI Hui[1,2], WANG Wen-jian[1]
(1. Key Laboratory of Computational Intelligence and Chinese Information Processing, Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006; 2. Department of Computer, Taiyuan Normal College, Taiyuan 030012)

【Abstract】This paper presents a Granular Support Vector Machine(GSVM) learning algorithm in order to improve the performance of SVM on imbalanced datasets. The GSVM divides some granules for majority data based on granular computing theory and extracts information granules. So the data becomes balanced, then GSVM finds local support vectors from those granules. SVM learns on these LSVs together with minority data. The satisfactory generalization performance can be obtained on imbalanced data.

【Key words】Guanular Support Vector Machine(GSVM); imbalanced data; information granule; local support vectors

1 概述

支持向量机是一类通用有效的机器学习方法[1]，能够非常成功地解决分类和回归问题，目前已成为机器学习的研究热点，并在很多领域如手写数字识别、人脸图像识别、时间序列预测等得到成功的应用。在实际应用中，经常会遇到许多数据分布不均衡的问题，如蛋白质同源性检测、疾病诊断、信用卡欺骗检测以及网页信息提取等。对于这类问题，人们往往更关心的是寻找少数类数据中所蕴含的重要信息。目前，对于非平衡数据学习问题的研究已取得了一些成果，如调整不同类的参数或权值[2]，以及直接采用传统 SVM 提取局部支持向量来获取重要数据等[3]。这些方法多是基于权值调整或边缘校正策略，因而往往精度不高，或精度较高但时间复杂度也很高。通过粒度计算[4]的思想分割原始数据集并进行有效的压缩，取其中重要的数据进行训练，以提高 SVM 的训练效率是一种简单、直观而有效的方法。

2 GSVM 学习算法

2.1 算法设计

对于高度非平衡的数据集，SVM 在训练中会使很多少数类样本归于多数类，有时甚至错误地将所有数据都归入到多数类中，从而使本来属于 2 类的数据划分为一个类，无法提取出那些虽然属于少数类但却非常重要的样本，如图 1 所示。

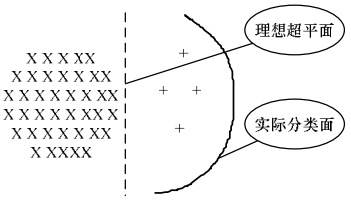


图 1 一般的非平衡数据分类图

考虑到非平衡数据的这种不易分性，通过划分粒的方法使数据趋于平衡；由于有些粒离“理想”超平面较远，对于分类没有影响，可以被安全地删除，而另一些粒中含有支持向量，离超平面相对较近，称之为信息粒，如图 2 所示。

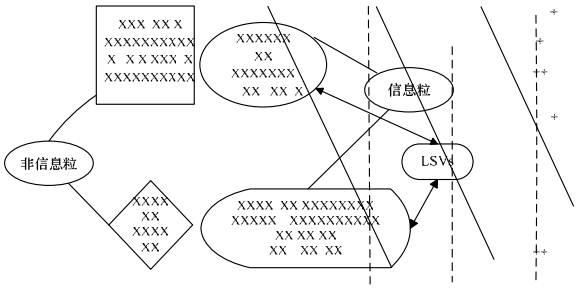


图 2 GSVM 对非平衡数据的分类

如果只采用这些粒中的某些点去代替整个粒中的样本训练必然会降低训练的精度，因此可以采用其中较多的甚至是全部样本进行下一次训练，提取每个粒中的支持向量作为局部支持向量。由于有多个粒，并且分布相对简单集中，因此可以采用线性 SVM 并行训练它们训练的时间开销不会增加很多。当然，如果样本量比较大且这些 LSVs 分布又相对集中时，得到的 LSVs 可能会比较多，所以根据实际需要再次

基金项目：国家自然科学基金资助项目(60673095)；国家“863”计划基金资助项目(2007AA01Z165)；教育部新世纪优秀人才支持计划基金资助项目(NCET-07-0525)；教育部科学技术研究基金资助重点项目(208021)；山西省青年学术带头人支持计划基金资助项目
作者简介：郭虎升(1986－)，男，硕士研究生，主研方向：机器学习；亓 慧，助教、硕士研究生；王文剑，教授、博士、博士生导师
收稿日期：2009-08-04 **E-mail：**chaofei142@163.com

进行分粒训练。此时的样本更靠近理想超平面，对于分类的平均作用比原来样本要大得多，训练的精度并不会因为这种粒划分代替而降低。

为方便描述，设 G 为训练集，其中，G^- 为负类训练样本集；G^+ 为正类训练样本集；G'为测试集；n 为 G^- 中样本的个数；p 为 G^+ 中样本的个数。不失一般性，本文中都假设负类样本远多于正类样本，即 $n>>p$。

GSVM 学习算法的主要步骤如下：

Step1 给定训练集 G 和测试集 G'。

Step2 对负类样本 G^- 进行粒划分，粒度大小为 K，其中，K 可以取为正类样本数 p 附近的值。

Step3 在粒划分之后的数据集上利用近邻点法产生压缩的负类数据集，并将其与所有正类数据 G^+ 结合来训练 SVM。

Step4 选择存在支持向量的粒作为信息粒。

Step5 将这些信息粒中的数据分别与正类数据集合并作为新的训练集，采用线性 SVM 训练，并筛选每个粒中的 LSVs。

Step6 将这些 LSVs 合并，进行聚类，取类中心与所有的正类数据结合，采用 RBF 核 SVM 进行训练，得到最终的分类超平面。

Step7 算法结束。

在进行粒划分时可以采用不同的方法，如各种聚类算法、关联规则、粗糙集等，为简单起见，本文采用 k-均值聚类进行划分。同时，为检验 GSVM 算法的有效性，与传统 SVM 算法以及基于一次 k-均值聚类的 SVM(Clustering based SVM, CSVM)算法进行比较。

算法 CSVM 的主要步骤如下：

Step1 给定训练集 G 和测试集 G'。

Step2 对负类样本 G^- 进行 k-均值聚类，其中，k 可以取为正类样本数 p 附近的值。

Step3 在聚类之后的数据集上利用近邻点法产生压缩的负类数据集，并将其与所有正类数据结合来训练 SVM。

Step4 算法结束。

2.2 算法评价指标

考虑到非平衡分类问题的特殊性及复杂性，以及大部分非平衡分类问题的实际应用，采用 4 个指标来衡量学习器的学习性能。

(1)*g_means*

对于非平衡数据集，常采用文献[5]提出的 *g_means* 值作为衡量指标，它可以有效地衡量非平衡数据的分类精度，一般 *g_means* 值越大分类效果越好。*g_means* 的定义为

$$g\_means=\sqrt{\frac{TN}{TN+FP}\times\frac{TP}{TP+FN}} \tag{1}$$

各符号的含义如表 1 所示。

表 1 相关符号的含义

预测样本类别	真实负类样本	真实正类样本
预测的负类样本	*TN*	*FN*
预测的正类样本	*FP*	*TP*

(2)*impor_rate*

考虑到非平衡分类问题的特殊性，在实际中往往更重视是否能将少数类的数据提取出来，而只依靠 *g_means* 值不能够反应这一特征。因此，采用重要样本的提取率 *impor_rate* 来度量少数重要样本的提取度，一般 *impor_rate* 值越大越好。其定义为

$$impor\_rate=\frac{TP}{TP+FN} \tag{2}$$

(3)*sacri_rate*

在非平衡数据分类问题中，提取重要的样本是关键的，但是提取这些重要的样本必然会导致部分非重要样本(即多数类样本)分类错误，因此，必须要考虑这些重要样本的提取究竟是以牺牲多少非重要样本为代价的。本文采用 *sacri_rate* 来度量每提取一个重要样本所要错分的非重要样本的相对牺牲率，一般 *impor_rate* 值越小越好。其定义为

$$sacri\_rate=\frac{FP}{TP\times\eta} \tag{3}$$

其中，η 为一常数，可以通过调整 η 值使 *sacri_rate* 处在[0,1]内。

(4)*accuracy*

accuracy 为所有样本的传统的正确分类率，其定义为

$$accuracy=\frac{TP+TN}{TP+FP+TN+FN} \tag{4}$$

对于分布较为平衡的数据来讲，*accuracy* 能准确反映学习器的泛化性能，而对于非平衡数据，这一指标实际意义并不大，因为它反映的是多数类样本的分类测试结果，所以本文采用此指标只是作为一般参考。

3 实验结果及分析

3.1 数据集说明

为验证 GSVM 学习算法的效率，在 3 个典型的 UCI 数据集(见表 2)上进行测试。实验中采用高斯核函数，其中正则参数 C 取 1 000，核参数 σ 取 1.0。

表 2 实验采用的数据集

数据集	训练集规模(负类/正类)	测试集规模(负类/正类)	维数
Thyroid	1 820(1 800/20)	1 011(1 000/11)	5
Spambase	1 530(1 500/30)	765(750/15)	57
Australian Sign Language(ASL)	130(100/30)	120(93/27)	22

3.2 实验结果分析

对 GSVM 算法的有效性进行验证。在实验中，影响算法表现的主要因素是粒度参数(聚类参数)k 的设定。为此，在数据集 Thyroid 上对粒度 k 对算法的影响进行了分析，表 3 为 GSVM 算法与 CSVM 算法的实验结果比较。

表 3 数据集 Thyroid 上的训练和测试结果

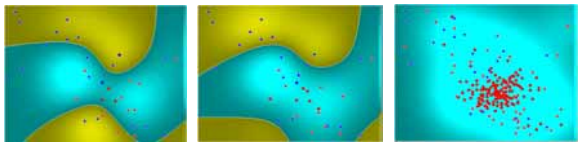
k	*g_means*		*impor_rate*		*sacri_rate*		*accuracy*	
	GSVM	CSVM	GSVM	CSVM	GSVM	CSVM	GSVM	CSVM
10	0.714	0.629	0.546	0.455	0.110	0.262	0.930	0.865
15	0.661	0.582	0.455	0.364	0.078	0.173	0.956	0.925
20	0.769	0.595	0.636	0.364	0.103	0.070	0.925	0.965
25	0.592	0.509	0.364	0.273	0.088	0.173	0.959	0.941
30	0.409	0.409	0.182	0.182	0.410	0.390	0.910	0.914
35	0.299	0.409	0.091	0.182	0.140	0.395	0.976	0.930
40	0.594	0.418	0.364	0.182	0.078	0.205	0.962	0.951
45	0.515	0.511	0.273	0.273	0.097	0.137	0.963	0.952
50	0.291	0.296	0.091	0.091	0.710	0.390	0.920	0.962

对于 GSVM 算法，在 k=20 时，*g_means* 和 *impor_rate* 均达到最大值；而对于 CSVM 算法，在 k=10 时，*g_means* 和 *impor_rate* 达到最大值。然而这 2 个值都远小于 GSVM 所取得的值。从整体看，在相同的粒度下，GSVM 大多数情况下训练和测试的 4 个衡量指标都比 CSVM 好。

直接采用 SVM 在数据集 thyroid 上进行训练和测试，*g_means* 与 *impor_rate* 这 2 个指标均为 0，此时 *sacri_rate* 值无意义，而 *accuracy* 为 0.989 12。这种方法把少数类样本全部错误地划分到多数类中，而由于少数类样本所占总样本的

比例太小，因此 *accuracy* 值虽然较大，但它只是一个看起来较好的“虚假”精度值，其效果实际上是最差的。

3 种方法得到的分类超平面如图 3 所示。



(a)GSVM (b)CSVM (c)SVM

图 3 在数据集 Thyroid 上训练得到的超平面

从上面的实验可以得知，对于数据集 Thyroid，本文提出的 GSVM 算法要明显优于 CSVM 算法与传统 SVM 算法，并且在该数据集中 k 取 20 可达到一个较好的分类效果。

表 4 与表 5 分别给出了 GSVM 算法和 CSVM 算法在数据集 Spambase 和 ASL 上的实验结果比较。

表 4 数据集 Spambase 上的训练和测试结果

k	*g_means*		*impor_rate*		*sacri_rate*		*accuracy*	
	GSVM	CSVM	GSVM	CSVM	GSVM	CSVM	GSVM	CSVM
10	0.299	0.441	0.875	0.636	0.869	0.681	0.099	0.284
15	0.582	0.661	0.545	0.625	0.433	0.412	0.566	0.635
20	0.582	0.670	0.500	0.625	0.555	0.396	0.613	0.667
25	0.700	0.605	0.625	0.455	0.292	0.270	0.699	0.735
30	0.524	0.604	0.364	0.455	0.425	0.282	0.693	0.755
35	0.556	0.550	0.375	0.364	0.403	0.302	0.746	0.780
40	0.652	0.562	0.500	0.364	0.260	0.282	0.782	0.817
50	0.491	0.574	0.273	0.375	0.270	0.287	0.820	0.821

表 5 数据集 Spambase 上的训练和测试结果

k	*g_means*		*impor_rate*		*sacri_rate*		*accuracy*	
	GSVM	CSVM	GSVM	CSVM	GSVM	CSVM	GSVM	CSVM
15	0.452	0.452	1.000	1.000	0.274	0.274	0.383	0.383
20	0.577	0.452	1.000	1.000	0.230	0.274	0.483	0.383
25	0.577	0.568	1.000	1.000	0.230	0.233	0.483	0.475
30	0.568	0.558	1.000	1.000	0.233	0.233	0.475	0.475

从表 4 与表 5 中可以看出，算法 GSVM 的 4 个指标均优于算法 CSVM。对于数据集 Spambase，当 k=25 时算法 GSVM 可达到一个较好的效果；而对于数据集 ASL，当 k=20 时算法 GSVM 可达到一个较好的效果并且重要样本的提取率为 100%。

4 结束语

支持向量机由于其出色的学习性能，目前已成为机器学习领域的一个研究热点，并在诸多实际领域得到了成功的应用。本文提出的 GSVM 算法可以对非平衡数据集进行有效的分类，从而使更多的少数类样本以更小的代价被正确分类。在未来的工作中，可以考虑将 GSVM 学习算法与核函数及参数选择相结合，从而使非平衡分类问题的效率得到进一步的提高。

参考文献

[1] Vapnik V. Statictical Learning Theory[M]. New York, USA: Wiley, 1998.

[2] 蒋 莎, 张晓龙. 一种用于非平衡数据的 SVM 学习算法[J]. 计算机工程, 2008, 34(20): 198-199.

[3] Tang Yuchun. Granular Support Vector Machines Based on Granular Computing, Soft Computing and Statistical Learning[D]. Atlanta, USA: Georgia Stage University, 2006.

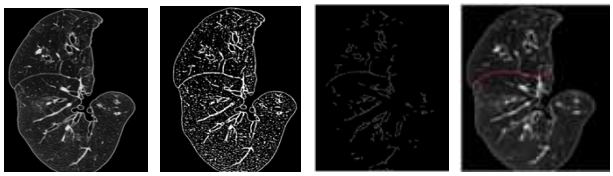
[4] Yao Y Y. On Modeling Data Mining with Granular Computing[C]// Proc. of the 25th Annual International Conference on Computer Software and Applications. Chicago, USA: [s. n.], 2001.

[5] Kubat M, Matwin S. Addressing the Curse of Imbalanced Training Sets: One-sided Selection[C]//Proc. of the 14th International Conference on Machine Learning. Nashville, Tennessee, USA: [s. n.], 1997.

编辑 顾逸斐

(上接第 180 页)

本文提出的自动方法基于这样的假设，裂纹上的像素值必须要比周围的结构高，而且裂纹上的像素灰度值变化不大。采用对肺部区域进行预处理的局部自适应分割方法，对于裂纹的强度必须要比周围结构高的要求很强。对于裂纹灰度值跟周围结构对比度不强的问题，可以通过非监督[6]的方法来增强。对于因为病变或者是 CT 图像分辨率低的原因而使自动检测失败的情况，可以手动给出裂纹的端点，动态规划也可以给出很好的结果。



(a)肺部图像 (b)自适应分割结果 (c)中轴变换结果 (d)检测结果

图 7 实验图像

6 结束语

本文提出了一个基于动态规划的在二维 CT 图像上检测肺部裂纹的自动方法，实验表明它可以很好地检测出裂纹，为进一步定量分析肺部结构提供了重要的基础。

参考文献

[1] Kubo M, Niki N, Nakagawa S, et al. Extraction Algorithm of Pulmonary Fissures from Thin-section CT Images Based on Linear Feature Detector Method[J]. IEEE Trans. on Nuclear Science, 1999, 46(6): 2128-2133.

[2] Kuhnigk J, Hahn H K, Hindennach M, et al. Lung Lobe Segmentation by Anatomy-guided 3D Watershed Transform[C]// Proc. of the International Society for Optical Engineering Medical Imaging Conference. San Diego, USA: [s. n.], 2003: 1482-1490.

[3] Zhang Li, Hoffman E A, Reinhardt J M. Altas-driven Lung Lobe Segmentation on Volumetric X-ray CT Images[J]. IEEE Trans. on Medical Imaging, 2006, 25(1): 1-16.

[4] Dehmeshki J, Amin H, Valdivieso M, et al. Segmentation of Pulmonary Nodules in Thoracic CT Scans: A Region Growing Approach[J]. IEEE Trans. on Medical Imaging, 2008, 27(4): 467-480.

[5] Hu Shiying, Hoffman E A, Reinhardt J M. Automatic Lung Segmentation for Accurate Quantitation of Volumetric X-ray CT Images[J]. IEEE Trans. on Medical Imaging, 2001, 20(6): 490-498.

[6] Rikxoort E M, Ginneken B, Klik M, et al. Supervised Enhancement Filters: Application to Fissure Detection in Chest CT Scans[J]. IEEE Trans. on Medical Imaging, 2008, 27(1): 1-10.

编辑 顾逸斐