第 14 卷 第 3 期 2014 年 1 月
1671—1815(2014)03-0081-06

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 14 No. 3 Jan. 2014
ⓒ 2014 Sci. Tech. Engrg.

基于支持向量机的不平衡样本分类研究

丁福利 孙立民*
(烟台大学计算机学院,烟台 264005)

摘 要 分类问题是机器学习领域的重要研究方向之一。支持向量机是一种基于结构风险最小化的学习机器,在解决分类问题上有着出色的效果。但基于支持向量机的分类器在处理不平衡样本时,对少类样本分类准确率偏低。诸多研究在对此问题做分析时往往把主要原因归结为各类样本间数量上的不平衡,而没有充分考虑样本点在特征空间上的分布情况。针对此问题做出原因分析,并给出结论:样本的不平衡性主要是由特征空间下各类样本的分布所决定的,而和数量上的不平衡关系较小。通过实验验证结论的科学有效性。

关键词 支持向量机 不平衡样本集 特征空间 样本分布
中图法分类号 TP391.9; **文献标志码** A

分类准确率是反映分类器性能的重要指标。传统的分类算法以提高整体分类准确率作为目标,且假定数据集中各类样本数是平衡的。然而在实际问题中存在大量不平衡样本集:某一类的样本数量远远少于其他类样本数量。例如,信用卡欺诈行为检测,网络入侵检测,医学疾病诊断[1]等。不平衡样本集分类的普遍问题是,总体分类准确率可以很高,而少类样本的分类准确率较低。极端情况下,把全部样本都分类为多数类,依然能获得较高的整体分类精度。然而很多实际问题中,少数类的分类准确率往往比多数类的分类准确率更为重要。比如,在癌症检测中,健康细胞相对于癌细胞是多数类,对癌细胞的正确分类更重要。因此,提高少数类的分类准确率成为分类问题中的一个研究热点。

支持向量机是以统计学习理论和结构风险最小化原则为基础的一种学习机器[2],推广能力好,在分类,回归等领域有着广泛的应用。但是支持向量机在处理不平衡样本集时,对少类样本的分类效果也不理想。为此,研究者们提出了很多处理不平衡数据分类问题的方法。Kubat 等[3]提出了一种启发式欠采样方法,用于去除多类样本中的噪声与冗余。Chawla 等提出了 SMOTE[4]方法,在相距较近的少类样本之间插入人造的少类样本,用来使得少类样本的数量同多类样本的数量达到平衡。吴洪兴等[5]利用遗传交叉运算,生成新的少类样本。Vempulos 等[6]对两类样本施加不同的惩罚因子的值,给少类样本以较大的惩罚因子,并给多类样本较小的惩罚因子,用来降低样本不平衡对分类器的影响。这些方法虽然从一定程度上能提高少类样本的分类准确率,但都没有从根本上分析造成分类准确率不平衡的原因,仅从现象的层次上去解决分类正确率不平衡问题。本文将分析论证,并通过实验验证不平衡样本集分类准确率不平衡的原因,并给出此类问题的解决方案。

1 支持向量机

设已知训练样本集为 T,$T=\{(x_1,y_1),(x_2,y_2),\cdots\cdots,(x_l,y_l)\}\in(X\times Y)^l$。其中 $x_i\in X=R^n$,$y_i=\{1,-1\}$,$i=1,2,\cdots,l$。支持向量机首先通过引入核函数[2]将输入空间中的 x_i 映射到特征空间中的 $\Phi(x_i)$,并构造最优化问题

$$\min\frac{1}{2}\|W\|^2+C\sum_{i=1}^{l}\varepsilon_i \tag{1}$$

$$\text{s.t}\begin{cases}y_i[W\cdot\Phi(x_i)+b]\geqslant 1-\varepsilon_i, & i=1,2,\cdots,l\\ \varepsilon_i\geqslant 0, & i=1,2,\cdots,l\end{cases} \tag{2}$$

式(2)中,W 是特征空间中的权向量,C 是惩罚因子,用来调节置信区间和经验风险的权重[7],ε_i 为松弛变量,b 为函数的阈值。

通过引入朗格朗日乘子,将上述式子转化为原问题的对偶问题[2]

$$\min\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_j\alpha_i\alpha_jK(x_i,x_j)-\sum_{i=1}^{l}\alpha_i \tag{3}$$

2013 年 8 月 12 日收到,9 月 20 日修改 山东省自然科学基金(2009ZRB019CE)资助

第一作者简介:丁福利(1990—),男,硕士研究生,研究方向:机器学习。E-mail:dingfuli@126.com。
* 通信作者简介:孙立民(1960—),男,教授,研究生导师,研究方向:机器学习,模式识别。E-mail:cslmsun@126.com。

$$\text{s.t}\begin{cases}\sum_{i=1}^{l} y_i\alpha_i = 0\\ 0 \leqslant \alpha_i \leqslant C, i = 1,2,\cdots,l\end{cases} \tag{4}$$

求解这个最优化问题,得最终函数表达式

$$f(x) = \text{sgn}\left[\sum_{i=1}^{l} y_i\alpha_i K(x_i,x) + b\right] \tag{5}$$

2 分类正确率不平衡原因分析

在各类样本数量上相差较大的数据集中,样本较多的类称为多数类,样本较少的类称为少数类。在针对不平衡数据集的研究上,往往把各类样本分类准确率的不平衡的原因归结于各类样本数量上的不平衡。文献[8—10]指出,由于各类样本数量上相差较大,而支持向量机是以增加整体分类准确率为目标,为追求更高的分类正确率,在分类的过程中偏好数量上居多的多数类,结果多数类的分类准确率较高,而少数类分类准确率较低。文献[8,9]中指出,这种分类准确率的不平衡性主要是由各类样本数量上的差异所导致的。事实上,这个结论并不恰当,这可以从图 1 所示的例子得到简单的反证。

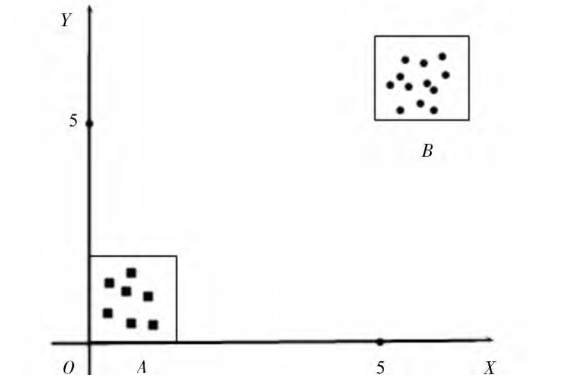

图 1 不平衡样本集

图 1 是构造的不平衡样本集。在区域 A 内($x,y \in [0,1]$)随机生成 20 个样本点,即少类样本点,在区域 B 内($x,y \in [5,6]$)随机生成 2 000 个样本点,即多类样本点。很显然,此样本集尽管不平衡度较大,但少类样本的分类正确率依然能达到 100%。这可从表 1 的实验数据得到进一步验证。

表 1 分类准确率

Data set	acc^+	acc^-	acc^-/acc^+	N^-/N^+
Data set constructed in this paper	100%	100%	1	0.01
Letter	99.87%	84.77%	0.85	0.04
German Credit	91.78%	42.59%	0.46	0.43
Cmc	100%	0%	0	0.53

表 1 中,Cmc,German Credit,Letter 是从 UCI 数据库[11]中选取的三个数据集。将 Cmc 和 Letter 数据集的其中一类样本记为少类样本,剩余样本记为多类样本,以此来构造出不平衡样本集。并将上述四个样本集随机抽取 80% 做训练样本,20% 做测试样本。基于支持向量机建模进行分类预测,分类准确率如表所示。

表 1 中,acc^- 和 acc^+ 分别表示少类样本和多类样本的分类准确率,acc^-/acc^+ 表示二者的比值。N^- 和 N^+ 分别表示少类样本和多类样本的个数,N^-/N^+ 是二者的比值。由表 1 可知,两类样本分类正确率的平衡度与两类样本数量的差异没有必然的联系,此处的实例表明:两类样本数量上的差异变大时,分类准确率的差异不升却反降,这与文献[8,9]所给出的结论相矛盾。因此,不平衡样本集分类准确率的差异并不仅仅与两类样本的数量差异有关。

下面将进一步分析论证,样本的不平衡并不是分类正确率不平衡的主因,后者与特征空间中样本的重叠度呈正相关,而与样本的类间距呈负相关。

图 2 不平衡样本集

文献[12]指出,即使在一些数量严重不平衡的样本集中,只要多类样本的区域和少类样本的区域分隔较大,分类器仍然具有良好效果,因此把区域间隔也作为产生分类准确率不平衡问题的重要因素。事实上,衡量两类样本可分性的标准有:重叠度以及类间距。当两类重叠度较大,并且两类类间距较小时,说明两类区域重合得比较严重。在这种情况下,一旦出现两类之间数量差异较大的情况(图 2),那么分类器为追求整体分类准确率,将会把大部分甚至全部少类样本预测成多类样本。从而导致多类和少类样本分类正确率的悬殊。反之,当两类重叠度较小,并且两类类间距较大时(图 1),很容易就能在两类样本点之间做出分隔超平面,将两类样本点很好地区分开来。此时,即使多类个数远多于少类样本个数,但多类样本中的支持向量并不会因此增加。综上所述,可以得到如下结论:在样本集为不平衡样

本集的前提下,多类样本和少类样本的分类准确率的差异主要取决于两类样本重叠度和类间距,并且分类准确率的悬殊程度与重叠度成正比,与类间距成反比。下面将进一步通过实验进行分析论证。

3　正确率平衡度与类间可分性关系

3.1　类重叠度计算

重叠度的数学表示采用 CCSL 算法[13],CCSL 算法具体步骤如下:

首先,两类样本 C_p 和 C_q 在样本点 X_k 处的模糊度[13]为

$$\mu(X_k: C_p, C_q) = \min\left\{\frac{\|X_k - V_p\|}{\|X_k - V_p\| + \|X_k + V_q\|}, \frac{\|X_k - V_q\|}{\|X_k - V_p\| + \|X_k + V_q\|}\right\} \tag{6}$$

式(6)中,V_p 和 V_q 分别是类 C_p 和 C_q 的中心。$\|X_k - V_p\|$ 和 $\|X_k - V_q\|$ 分别表示 X_k 到两类中心点的距离。$\mu(X_k: C_p, C_q)$ 是样本点 X_k 属于类 C_p 和类 C_q 的模糊度。类 C_p 和类 C_q 划分越清楚,模糊度越低,其值分布在 0 ~ 0.5。

由于支持向量机通过核函数将 x_i 映射到特征空间中的 $\Phi(x_i)$,因此在特征空间上,样本点到类中心点的距离计算过程如下。

特征空间上 X_k 到类 C_p 中心点距离

$$D = \left\|\Phi(x_k) - \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\right\| = \sqrt{\left[\Phi(x_k) - \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\right]^2} = \sqrt{\Phi(x_k)\Phi(x_k) - \frac{2}{n}\sum_{i=1}^{n}\Phi(x_i)\Phi(x_k) + \left[\frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\right]^2} = \sqrt{K(x_k, x_k) - \frac{2}{n}\sum_{i=1}^{n}K(x_i, x_k) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(x_i, x_j)} \tag{7}$$

式(7)中,n 为类 C_p 的样本个数,类 C_p 在特征空间上的中心点为 $\frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)$。同理可以得到样本点 X_k 到类 C_q 中心点的距离。并根据点到类中心点的距离计算得到类 C_p 和类 C_q 在样本点 X_k 处的模糊度 $\mu(x_k: C_p, C_q)$。

由于,归属模糊的样本点比归属清楚的样本点有更大的权重,所以两个类的重叠度应该在归属模糊的样本点处定义更大的值[13]。

因此,特征空间上类 C_p 和类 C_q 在样本点 X_k 处的重叠度

$$CSL(X_k: C_p, C_q) = \begin{cases} (2\times\mu(X_k: C_p, C_q))^2, & 0 < \mu(X_k: C_p, C_q) \leqslant 0.1 \\ (2\times\mu(X_k: C_p, C_q))^1, & 0.1 \leqslant \mu(X_k: C_p, C_q) \leqslant 0.3 \\ (2\times\mu(X_k: C_p, C_q))^{0.25}, & 0.3 \leqslant \mu(X_k: C_p, C_q) \leqslant 0.5 \end{cases} \tag{8}$$

根据以上定义,把类 C_p 和类 C_q 之间重叠度定义为两类上所有样本点重叠度的平均值。因此两个类 C_p 和 C_q 之间的重叠度

$$CSL(C_p, C_q) = \frac{1}{N}\sum_{i=1}^{N}CSL(X_k: C_p, C_q) \tag{9}$$

这样,类 C_p 和类 C_q 之间归属模糊的数据越少,重叠度的值越小。反之亦然。

3.2　类间距计算

类 C_p 和类 C_q 在特征空间下的类间距离采用类中心间距来表示[14]。特征空间下类 C_p 的中心点为 $\frac{1}{n_p}\sum_{i=1}^{n_p}\Phi(x_i)$,$n_p$ 为类 C_p 的样本点个数。类 C_q 的中心为 $\frac{1}{n_q}\sum_{i=1}^{n_q}\Phi(x_i)$,$n_q$ 为类 C_q 的样本点个数。类 C_p 和类 C_q 在特征空间下的类间距离

$$D(C_p, C_q) = \left\|\frac{1}{n^p}\sum_{i=1}^{n_p}\Phi(x_i) - \frac{1}{n^q}\sum_{i=1}^{n_q}\Phi(x_i)\right\| = \sqrt{\left[\frac{1}{n^p}\sum_{i=1}^{n_p}\Phi(x_i) - \frac{1}{n^p}\sum_{i=1}^{n_p}\Phi(x_i)\right]^2} = \sqrt{\frac{1}{n_p^2}\sum_{i=1}^{n_p}\sum_{j=1}^{n_p}K(x_i, x_j) + \frac{1}{n_q^2}\sum_{i=1}^{n_q}\sum_{j=1}^{n_q}K(x_i, x_j) - \frac{2}{n_p\times n_q}\sum_{i=1}^{n_p}\sum_{j=1}^{n_q}K(x_i, x_j)} \tag{10}$$

将类间距和重叠度的比值 $D(C_p, C_q)/CSL(C_p, C_q)$ 简记为 D/CSL。

3.3　实验验证

为验证分类准确率不平衡度与类间距和重叠度比值的关系,从 UCI 数据库[11]下载 Iris,Haberman(简记为 Haber),Wine,Blood Transfusion(简记为 Blood),Pendigits,German Credit(简记为 German)数据集,并从 Libsvm[15]网站上下载 Cmc,Letter 数据集。这些数据集都是研究不平衡样本集问题常用的数据集(Iris and Wine in 文献[16,17],German Credit in 文献[18],Cmc,Haberman,Letter in 文献[19],Haberman and Blood Transfusion in 文献[20])。在这些数据集中,类别数量多于两个类别的,将其中一类归为少类样本,剩余类归为多类样本,以此来构造出不平衡样本集。将这些不平衡样本集随机抽取 80% 用于做训练样本,其余的 20% 用于做测试样本,并随机抽取 10 次做实验,得到 10 次实验结果的平均值。表 2 ~ 表 4 分别是在三种核函

数下所得的实验结果。表中第 1 列是数据集的名称,第 2,第 3 列分别代表多类样本和少类样本的分类准确率,第 4 列表示少类样本的分类准确率和多类样本分类准确率的比值,比值接近于 1,说明两者的准确率较为平衡,比值越接近于 0,说明两者准确率很不平衡。表中第 5 列是少类样本个数和多类样本个数的比值。第 6 列是两类样本在特征空间下类间距和重叠度的比值。

表 2 实验结果(线性核函数)

数据集	acc^+	acc^-	acc^-/acc^+	N^-/N^+	D/CSL
Iris	100%	90%	0.90	0.50	0.52
Haber	100%	0%	0	0.36	0.01
Wine	100%	0%	0	0.50	0.12
Blood	100%	0%	0	0.31	0.01
Pendigits	99.26%	57.35%	0.58	0.12	0.15
German	100%	0%	0	0.43	0.04
Cmc	100%	0%	0	0.53	0.03
Letter	100%	0%	0	0.04	0.04

表 3 实验结果(多项式核函数)

数据集	acc^+	acc^-	acc^-/acc^+	N^-/N^+	D/CSL
Iris	100%	90%	0.90	0.50	0.52
Haber	100%	0%	0	0.36	0.01
Wine	100%	0%	0	0.50	0.12
Blood	100%	0%	0	0.31	0.01
Pendigits	99.26%	57.35%	0.58	0.12	0.15
German	100%	0%	0	0.43	0.04
Cmc	100%	0%	0	0.53	0.03
Letter	100%	0%	0	0.04	0.04

表 4 实验结果(径向基核函数)

数据集	acc^+	acc^-	acc^-/acc^+	N^-/N^+	D/CSL
Iris	100%	100%	1	0.50	0.81
Haber	100%	0%	0	0.36	0.06
Wine	100%	84.62%	0.85	0.50	0.37
Blood	100%	0%	0	0.31	0.08
Pendigits	99.03%	77.64%	0.78	0.12	0.33
German	92.36%	32.56%	0.35	0.43	0.13
Cmc	100%	0%	0	0.53	0.06
Letter	99.87%	77.85%	0.78	0.04	0.30

图 3 ~ 图 5 是根据表 2 ~ 表 4 所绘制的图。其中,横坐标分别表示实验所用到的 8 个数据集,Line 1 表示 acc^-/acc^+,Line 2 表示 N^-/N^+,Line 3 表示 D/CSL,即特征空间下两类类间距和重叠度的比值。分析表 2 ~ 表 4 及图 3 ~ 图 5 可以看出,多类和少类样本分类准确率的比值主要与两类类间距和重叠度的比值有关,而与两类样本数量的比值关系不大。甚至有些样本,当两类样本的个数相差悬殊时,两者的分类准确率却十分接近(如 Letter 数据集,Pendigits 数据集,以及本文根据图 1 所构造的数据集)。以上实验进一步验证了本文的结论:分类准确率的不平衡度主要与特征空间中类间距和重叠度的比值有关,并且它与类间距成反比,与重叠度成正比。

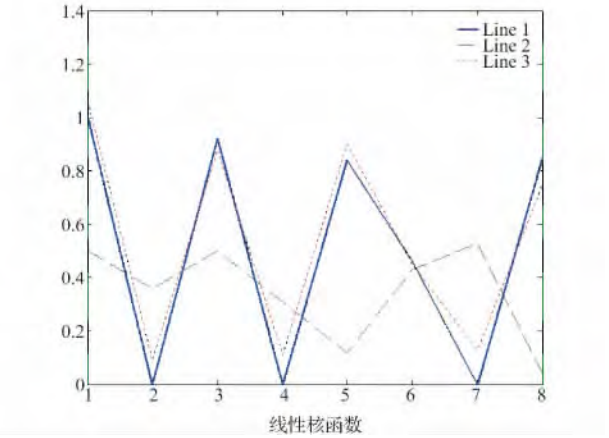

图 3 实验结果(线性核函数)

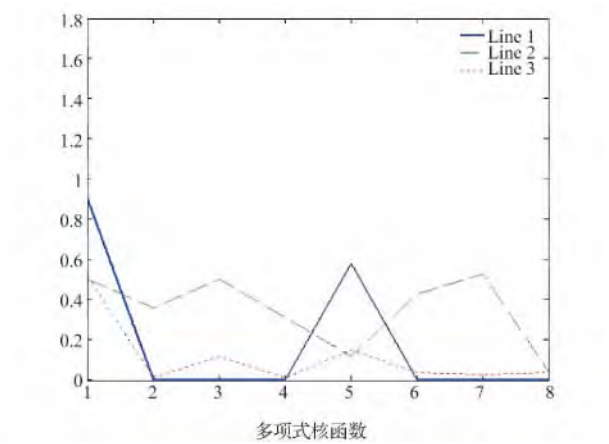

图 4 实验结果(多项式核函数)

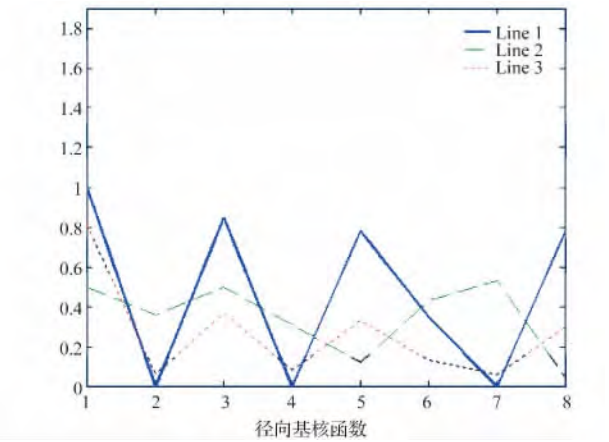

图 5 实验结果(径向基核函数)

比较表 2 ~ 表 4 的数据还可以发现,不同的核函数,对 acc^-/acc^+ 的值有影响。比如 wine 数据集,采用线性核,二项式核和径向基核时,这个比值分别是 0.92,0,0.85。这说明分类正确率不平衡度与核

函数的选取也有着较大关系。

为进一步考查 acc^-/acc^+ 与 N^-/N^+ 及 D/CSL 之间的关系，还可以通过计算两者的线性相关性[21]来验证。

设两个序列 x_i 和 y_i，$i=1,2,3,\cdots,n$。他们的相关系数计算公式[22]为

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{11}$$

式(11)中，r_{xy} 为两序列的相关系数。n 为两序列的个数。相关系数的取值在 $[-1,1]$。相关系数越接近于 1，表示正比关系越明显。当相关系数为 0 时，表示两者不相关[22]。

根据式(11)，分别计算 acc^-/acc^+ 与 N^-/N^+ 和 D/CSL 的相关系数，如表 5 所示。从表 5 数据可以看出，acc^-/acc^+ 与 D/CSL 有很强的相关性，而与 N^-/N^+ 的相关性比较弱。这进一步支持了上述的结论。

表 5　相关系数表

所选核函数	相关系数	
	acc^-/acc^+ 和 N^-/N^+	acc^-/acc^+ 和 D/CSL
线性核函数	0.22	0.99
多项式核函数	0.01	0.91
径向基核函数	0.20	0.86

4　结论

不平衡样本集分类时普遍存在的问题是少类样本分类正确率比较低，分类器为追求整体分类正确率的提高，往往把少类样本分类为多类。支持向量机在处理不平衡样本集时，分类效果也不理想。本文对产生这一问题的根本原因进行分析论证，并通过实验验证，最终得到如下结论：在样本集为不平衡样本集的前提下，分类准确率的平衡度主要与特征空间下类间距和重叠度的比值有关，而与两类间的数量差异关系较小。通过实验还发现，核函数的选择也与分类正确率平衡度有较大关系。因此，对于不平衡样本集，解决分类正确率不平衡问题的思路应该从降低特征空间中两类的重叠度及增大类间距入手，才能使问题得到根本解决。首先，选择合适的核函数对解决此类问题很重要。因为核函数决定了输入空间的样本映射到特征空间后的分布，合适的核函数会使样本的重叠度降低，类间距增大。其次，欠采样算法也不失为解决此类问题的一个比较有效的方法。尽管欠采样方法由于删减大量多类样本点会造成信息丢失，但这种方法也的确可以降低重叠度，增大类间距。今后将进一步研究降低类重叠度，增大类间距的系列方法，以进一步解决不平衡样本集分类问题所遇到的困难。

参　考　文　献

1　Li Derchiang，Liu Chiaowen，Hu S C. A learning method for the class imbalance problem with medical data sets. Computers in Biology and Medicine，2010；5(40)：509—518

2　邓乃扬，田英杰. 数据挖掘的新方法—支持向量机. 北京：科学出版社，2004

3　Kubat M，Matwin S. Addressing the course of imbalanced training sets：one-sided select. Proc of the 14th International Conference on Machine Learning(ICML'97). [s. 1.]：Morgan Kaufmann，1997：179—186

4　Chawla N V，Bowyer K W，Lawrence O H，*et al*. SMOTE：synthetic minority over-sampling technique. Journal of Artificial Intelligence Research，2002；16：321—357

5　吴洪兴，彭　宇，彭喜元. 适用于不平衡样本数据处理的支持向量机方法. 电子学报，2006；34(12A)：2395—2398

6　Zhou Zhihua，Liu Xuying. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans on Knowledge and Data Engineering，2006；18(1)：63—77

7　肖　建，于　龙，白裔峰. 支持向量回归中核函数和超参数选择方法综述. 西南交通大学学报，2008；43(3)：297—303

8　文传军，詹永照. 基于样本投影分布的平衡不平衡数据集分类. 计算机应用研究，2009；26(8)：3131—3133

9　吴洪兴，彭　宇，彭喜元，等. 适用于不平衡样本数据处理的支持向量机方法. 电子学报，2006；34(z1)：2395—2398

10　姚程宽. 不平衡样本集的支持向量机模型选择. 南京：南京师范大学，2007

11　机器学习领域所用数据集. http://archive.ics.uci.edu/ml/

12　Prati R C，Batista G E A P A，Monard M C，*et al*. Class imbalances versus class overlapping：an analysis of a learning system behavior. Third Mexican International Conference on Artificial Intelligence (MICAI 2004)，2004：312—321

13　瞿　俊，姜青山，翁芳菲，等. 基于重叠度的层次聚类算法. 计算机研究与发展，2007；44(z2)：181—186

14　毛国君，段立娟，王　实，等. 数据挖掘原理与算法. 北京：清华大学出版社，2005

15　Chang C C，Lin C J. LIBSVM：a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST)，2011；2(3)：27

16　陶晓燕，姬红兵，马志强. 基于样本分布不平衡的近似支持向量机. 计算机科学，2007；34(5)：174—176

17　陶晓燕，姬红兵，董淑福. 用于非平衡样本分布的近似支持向量机. 模式识别与人工智能，2007；20(4)：552—557

18　郑恩辉，许　宏，李　平，等. 基于 v-SVM 的不平衡数据挖掘研究. 浙江大学学报(工学版)，2006；40(10)：1682—1687

19　林舒杨，李翠华，江　戈，等. 不平衡数据的降采样方法研究. 计算机研究与发展，2011；48(z2)：47—53

20　赵自翔，王广亮，李晓东. 基于支持向量机的不平衡数据分类的改进欠采样方法. 中山大学学报(自然科学版)，2012；51(6)：10—16

(下转第 92 页)

The Simulation Study on Image Algorithm Based on the Reversible Linear Memory Cellular Automata Integration with Time Delay

HUANG Chan[1], ZENG Xiao-long[2]
(Department of Computer[1], Department of Modern Education Technology[2], Ganzhou Teachers College, Ganzhou 341000, P. R. China)

[Abstract] To the cellular automata encryption system does not having the memory function, resulting in the low security and encryption speed, and all of the current image encryption algorithms were ignored the time delay phenomenon to cause it not reflecting the actual encryption process. The reversible linear memory cellular automata (CA) and time delay function were proposed to overcome above problems, and also the non-linear coupled permutation method was introduced into this algorithm. Firstly, iterating the one dimension piecewise linear chaotic map to get the array, and then the non-linear coupled permutation method was designed according to the array, as a result, the permutation array was got, used the permutation array to permute the plain-image to change the pixel position; secondly, the time delay was introduced into the logistic map, the pseudo-random number on time delay obtained by logistic map integration with time delay function was iterated by evolution rule of reversible linear memory cellular automata to get the iteration array, then this array was used to diffuse the permutation image to change the pixel value by pixel diffusion mechanism. The simulation software MATLAB was used to test the algorithm, the results showe that this image encryption algorithm based on the reversible linear memory cellular automata integration with time delay has an excellent encryption performance, the encryption system is highly; the key space was huge, and the anti-attack property significantly increased.

[Key words] reversible linear memory cellular automata time delay non-linear coupled permutation evolution rule pixel diffusion mechanism

(上接第 85 页)

21 李占文,骆有庆,苓建强,等. 宁夏灵武枣区枣桃小食心虫发生规律与气候相关性. 宁夏大学学报(自然科学版),2011;32(4):395—399

22 徐永群,邓三尧,李兴旺,等. 阵列相关系数比对法在中药鉴别中的应用研究. 光谱学与光谱分析,2007;27(11):2239—2242

Unbalanced Sample Set Classification Based on Support Vector Machine

DING Fu-li, SUN Li-min
(School of Computer Science, Yantai University, Yantai 264005, P. R. China)

[Abstract] Classification is an important field of machine learning. SVM is a learning machine based on structural risk minimization, it is very good at solving classification. However its classification accuracy for the minority class of the unbalance sample set is very low. Many researchers give their analysis on it, they often consider the problem is caused by the sample unbalance in quantity. They did not consider the distribution of sample points in the feature space. The reasons are analyzed for this problem, and given the conclusion. The unbalance of classification accuracy is mainly determined by the sample distribution in the feature space, it has a smaller relationship with the imbalance in quantity. The experiment results validate conclusion.

[Key words] support vector machine unbalanced sample set feature space sample distribution