阴影集的模糊支持向量机样本选择方法

苏小红，赵玲玲，谢 琳，马培军

（哈尔滨工业大学 计算机科学与技术学院，150001 哈尔滨）

摘 要：样本选择可以提高模糊支持向量机训练速度并在一定程度上提高其抗噪能力，但存在有效样本选择困难和选样率高的问题，利用阴影集对模糊集的分析能力，提出一种新的基于阴影集的模糊支持向量机样本选择方法，将模糊集合划分为可信任、不可信任及不确定3个子集，仅在可信任和不确定子集中选样，并分别采用子空间样本选择和边界向量提取的方法选样. 实验结果表明，该方法在保持分类器泛化能力的前提下可以有效降低选样率和训练时间. 因该方法去除了样本中的不可信任数据，所以当训练样本中含有噪声时，还可以有效提高分类器的分类性能.

关键词：模糊支持向量机；样本选择；阴影集

中图分类号：TP183 **文献标志码**：A **文章编号**：0367－6234(2012)09－0078－07

Shadowed sets-based sample selection method for fuzzy support vector machine

SU Xiao-hong, ZHAO Ling-ling, XIE Lin, MA Pei-jun

(School of Computer Science and Technology, Harbin Institute of Technology, 150001 Harbin, China)

Abstract: Sample selection can speed up the training of Fuzzy Support Vector Machine(SVM). However, it is difficult to select effective sample and the selection ratio is very high. This paper proposes a new sample selection method for Fuzzy SVM based on shadowed sets. We divide the fuzzy sets into three subsets, i. e. trustable data sets, trustless data sets and uncertain data sets. The samples are only selected in trustable data sets and uncertain data sets by using the subspace selection algorithm and the border vector extraction method respectively. Experimental results show that the training time and selection ratio is significantly reduced without any decrease in generalization ability by using the samples chosen by the proposed method. Furthermore, it improves the prediction performance of the classifiers when the data sets contain noises.

Key words: fuzzy support vector machine; sample selection; shadowed sets

系统的性能退化状态识别是智能维护(Intelligent Maintenance System, IMS)[1]系统的核心之一. 目前，系统的退化状态识别方法主要有3类[2]. 基于模型的方法因系统的性能退化机理复杂、很难建立退化模型，而实际应用价值不大[3]. 神经网络虽应用广泛，但因系统结构复杂，识别过程受很多因素的影响[4]，因此，识别退化状态时，易发生“维度灾难”问题[5]. 支持向量机(Support Vector Machine, SVM)由 V. Vapnik[6]于1995年提出，该方法建立在结构风险最小化的基础上，有良好的泛化能力，能较好的解决小样本、非线性、高维度问题，因此广泛应用于故障识别中.

将SVM用于退化状态识别主要有两个难点：首先，训练集中人工标定的退化状态中可能会有错误标定，即训练集中存在噪声，因此，应提高SVM的抗噪能力. 其次，系统的退化数据较多，因此，应提高SVM在大样本上的训练速度.

实际上，系统的退化是一个连续的过程，明确标定每个样本点的状态是不合理的，因此，退化状态识别应使用模糊支持向量机(Fuzzy Support Vector Machine, FSVM)[7]. 样本选择是减少FSVM训练时间，一定程度提高其抗噪能力的一个

收稿日期：2011－05－24.
基金项目：国家自然科学基金资助项目(61175027).
作者简介：苏小红(1966—)，女，教授，博士生导师；
马培军(1963—)，男，教授，博士生导师.
通信作者：苏小红，sxh@hit.edu.cn.

有效且直接的方法.

本文利用阴影集[8-9]对模糊集的分析能力，提出一种基于阴影集划分的 FSVM 样本选择方法，在不同置信区间采用不同的样本选择方法. 实验表明，该方法在保证 FSVM 泛化能力的前提下，大大减少了训练时间，还提高了含噪声数据的 FSVM 的预测精度.

1　传统 SVM 与 FSVM 样本选择方法

传统 SVM 的样本选择方法大致包含以下 3 种：随机选样的方法[10]、保留典型样本的方法[11-12]、保留支持向量的方法[13-15]. 其中，保留典型样本方法主要采用的是对样本进行聚类，保留聚类中心为典型样本的策略. 保留 SVM 的方法中，文献[15]主要研究的是可分支持向量机的样本选择方法，该方法是所有的样本选择方法中，选样率最低的一种方法，且能保证选样后 SVM 的泛化能力.

FSVM 与传统 SVM 的最大区别在于其训练样本含有模糊隶属度. 随机选样方法可直接应用于 FSVM，但不能保证选样后 FSVM 的泛化能力. 模糊隶属度是样本点的重要信息，而保留典型样本的方法是根据样本点的位置信息决定典型样本，因此选出的典型样本在模糊集中并不一定是典型的. FSVM 中，模糊隶属度不同的样本点，分类面对它们的容忍程度是不一样. 而支持向量是由样本点到分界面的距离以及样本点的模糊隶属度共同决定的. FSVM 中，可以确定的非支持向量只有类别中心所在的样本点，以及相较与类别中心，距离另一类更远的点. 文献[16]就是基于这个思想，提出了基于边界向量提取的模糊支持向量机. 但该方法的选样率较高，理论上需选择 1/2 的样本.

对模糊集直接进行样本选择是很困难的，因此，挖掘出模糊集中的确定信息，从而将传统 SVM 的样本选择方法应用于 FSVM 中，是非常有意义的.

2　模糊集

定义 1（模糊集）[17]　设 A 是论域 X 到 $[0,1]$ 上的一个映射关系，即

$$A:\ X\to[0,1],x\mapsto A(x). \tag{1}$$

则称 A 是 X 上的模糊集.

定义 2（模糊隶属度）　式(1)中，$A(x)$ 称为 x 对模糊集 A 的模糊隶属度，或直接称其为模糊集的隶属函数.

以下，记 X 上全体模糊集所构成的集合为 $F(X)$. 如果，$A\in F(X)$，且 $A:X\to\{0,1\}$，则 A 为经典集，记为 $A\in P(X)$.

定义 3（模糊度）　若映射

$$d:F(X)\to[0,1].$$

满足条件：

1）当且仅当 $A\in P(X)$ 时，$d(A)=0$；

2）当且仅当 $A(x)\equiv 0.5$ 时，$d(A)=1$；

3）对于 $\forall x\in X$，当 $B(x)\leqslant A(x)\leqslant 0.5$ 时，$d(B)\leqslant d(A)$；

4）对于 $A\in F(X)$，有 $d(A)=d(A^c)$.

则称映射 d 为 $F(X)$ 上的一个模糊度，$d(A)$ 为模糊集 A 的模糊度.

常用的几种模糊度计算公式如下：

令 $A=\{(x_1,A(x_1)),(x_2,A(x_2)),\cdots,(x_n,A(x_n))\}$ 表示一个模糊集.

$$A_{0.5}(x_i)=\begin{cases}1,A(x_i)\geqslant 0.5;\\0,A(x_i)<0.5.\end{cases}$$

模糊集 A 的 Hamming 模糊度为

$$d(A)=\frac{2}{n}\sum_{i=1}^{n}|A(x_i)-A_{0.5}(x_i)|.$$

模糊集 A 的 Euclid 模糊度为

$$d(A)=\frac{2}{\sqrt{n}}\Big(\sum_{i=1}^{n}|A(x_i)-A_{0.5}(x_i)|^2\Big)^{\frac{1}{2}}.$$

虽然 Hamming 模糊度的计算简单，但误差较大. Euclid 模糊度相比于 Hamming 模糊度更准确. 因此本文采用 Euclid 作为度量模糊集合的模糊度.

3　阴影集

阴影集由 W. Pedrycz 于 1998 年提出，它主要用于解决模糊逻辑中的一个矛盾问题：用确定的模糊隶属度来描述不确定的集合. 阴影集的一个主要依据是，模糊集中隶属度为 0.5 附近的样本，其类别信息很难确定，而隶属度接近于 1 或隶属度接近于 0 的样本，其类别信息是可确定的. 因此，阴影集映射将隶属度足够高的样本的隶属度提高为 1，将隶属度足够低的样本的隶属度降低为 0，而将其他样本的隶属度放宽为 $[0,1]$ 区间. 从而实现提取出了模糊集中的确定信息.

从模糊集到阴影集实质上是一个三值映射.

$$\Psi:A\xrightarrow{\alpha}\{0,(0,1),1\}.$$

式中：A 为模糊样本集；Ψ 为由模糊集到阴影集的映射，称为模糊映射；α 为确定阴影集划分的阈值；映射 Ψ 将模糊隶属度 $<\alpha$ 的样本映射为 0；模糊隶属度 $>1-\alpha$ 的样本映射为 1；其余样本映射

为(0,1).

在阴影集的基础上,本文提出可信任数据、不确定数据以及不可信任数据的概念. 利用阴影集映射将原有的模糊集进行区域划分.

定义4(可信任数据) 对于 $\forall(x,A(x))\in A$,A 为模糊集,$A(x)$ 称为 x 对模糊集 A 的模糊隶属度.

如果

$$\Psi(x,A(x))=1.$$

即

$$A(x)>1-\alpha.$$

则称 x 为可信任数据.

定义5(不可信任数据) 对于 $\forall(x,A(x))\in A$,A 为模糊集,$A(x)$ 称为 x 对模糊集 A 的模糊隶属度.

如果

$$\Psi(x,A(x))=0.$$

即

$$A(x)<\alpha.$$

则称 x 为不可信任数据.

定义6(不确定数据) 对于 $\forall(x,A(x))\in A$,A 为模糊集,$A(x)$ 称为 x 对模糊集 A 的模糊隶属度.

如果

$$\Psi(x,A(x))=(0,1).$$

即

$$\alpha\leqslant A(x)\leqslant 1-\alpha.$$

则称 x 为不确定数据.

4 基于阴影集划分的样本选择方法

4.1 划分阈值的确定

阈值的确定是阴影集划分中最关键的一步,合理的阈值选择是后续分区域样本选择方法正确性的保证.

文献[8-9]在给出阴影集定义的同时,也给出了划分阈值的确定原则:由模糊集到阴影集的映射应该保持整个集合不确定性的平衡. 映射后,模糊集中隶属度 $<\alpha$ 的样本点和隶属度 $>1-\alpha$ 的样本点的不确定性都消除了,这部分的变化应该由阴影集(即本文定义的不确定数据集合)来补偿. 这里,不确定性是由样本点的模糊隶属度来衡量的.

理论上,如果已知模糊集的隶属度函数且知道样本点的分布,在样本点无穷时,不确定性的平衡是可以达到的. 但实际样本点的分布是很难预知的,仅知有限点的模糊隶属度,这时,可能找不到一个阈值,可保持不确定性的平衡. 因此,划分的阈值一般取使不确定性改变最小的值. 由此,得到阈值 α 的确定公式为

$$V(\alpha)=\left|\sum_{i:A(x_i)\leqslant\alpha}A(x_i)+\sum_{i:A(x_i)\geqslant 1-\alpha}1-A(x_i)-\operatorname{card}\{x_i\in X\mid\alpha<A(x_i)<1-\alpha\}\right|,$$

$$\alpha^*=\arg\min_{\alpha}V(\alpha).$$

上述阈值确定的方法存在一定的问题. 首先,模糊理论中,集合的模糊性是由模糊度来衡量的. 由文献[8-9]提出的对模糊集不确定性的度量实际上是没有归一化的 Hamming 模糊度的计算,但是文献[17]中指出,Hamming 模糊度的计算有很大的误差,这会给实际计算的阈值带来很大的误差. 它可能使计算的阈值小于实际的阈值,也可能使计算的阈值大于实际的阈值. 实际上,这是和样本本身的分布有关的. 其次,文献[8-9]以不确定数据集合来衡量映射后集合不确定性的减少是不合理的. 实际上,不确定数据对应的模糊集中的样本,其不确定性本身就很大,虽然映射后,会增加这部分数据的不确定性,但是这个增量是<1的. 这个部分的计算也会产生较大的误差,并且,这个误差会使计算的阈值比实际的阈值大.

本文指出,由模糊集到阴影集的映射应该保持集合的模糊度不变. 在此基础上,给出新的阈值计算为

$$V(\alpha)=|d(A)-d(B)|,$$

$$\alpha^*=\arg\min_{\alpha}V(\alpha).$$

式中:$d(A)$、$d(B)$ 分别为模糊集 A 及其对应的阴影集 B 的模糊度.

模糊集中隶属度为0.5的样本,样本信息是完全未知,与阴影集中不确定数据的含义相同,因此,本文将不确定数据模糊度的计算等同于模糊集中隶属度为0.5的样本的模糊度的计算. 由此,给出划分阈值为

$$V(\alpha)=\left|\frac{2}{\sqrt{n}}\sqrt{\sum_{i=1}^{n}|A(x_i)-A_{0.5}(x_i)|^2}-\sqrt{\operatorname{card}\{x_i\in A\mid\alpha<A(x_i)<1-\alpha\}/n}\right|,$$

$$\alpha^*=\arg\min_{\alpha}V(\alpha). \tag{2}$$

由式(2)知,划分阈值只是用于对模糊样本的模糊隶属度进行划分,因此,这里划分阈值可仅考虑有限种情况,它是由样本的隶属度决定的.

给定模糊集 A,划分阈值的取值集合为

$$\{\alpha_i\mid\alpha_i=\min\{A(x_i),1-A(x_i)\}\}.$$

分析可知,随着所选阈值 α 的增大,不确定数据会逐渐减少,对应的阴影集的模糊度会由1逐渐减少

为 0. 即随着阈值 α 的增大，$V(\alpha)$ 会先增大后减小. 因此，当本文从小到大计算每一个 α 对应的 $V(\alpha)$ 时，只要在某一次计算中，$V(\alpha)$ 的值变大，就可以认定之后 $V(\alpha)$ 会逐步增加，最佳阈值即是上次计算中的 α.

本文给出的计算最佳阈值的算法如图 1 所示.

```
算法：best_Threshold(blongs[], n)
输入：blongs[]表示模糊样本集每个样本对应的模糊隶属度，n
表示模糊集中样本的个数.
输出：阴影集划分的阈值 a_best
1. 令 a_last 保存上次处理的阈值.
2. V_last 保存上一次计算的 V(α).
3. fuzzy_f 记录原模糊集的模糊度.
4. fuzzy_s 记录阴影集的模糊度.
5. a[] 保存阈值的可能取值.
6. low_last、high_last 分别保存上次处理阈值时，排序后的
   blongs[]中大于 a_last 的最小下标，以及 blongs[]中小于 1 -
   a_last 的最大下标.
7. num_shadowed 记录阴影中不确定数据的个数.
8. 初始化a_last = -1, V_last = inf, num_shadowed = n.
        low_last = 0, high_last = n - 1.
9. 确定划分阈值的集合为
   a[] = {a[i] | a[i] = min{blongs[i], 1 - blongs[i]}}.
10. 对 a[] 从小到大排序，并剔除其中重复的阈值.
11. 对 blongs[]从小到大排序.
12. 计算模糊集的模糊度为
    fuzzy_f = 2/n * sqrt(
              sum{pow(blongs[i], 2) | blongs[i] < 0.5} +
              sum{pow((1 - blongs[i]), 2) | blongs[i] >=
              0.5}).
13. for 每一个划分阈值 a[i].
14.     while blongs[low_last] <= a[i].
15.         low_last++.
16.     while blongs[high_last] >= 1 - a[i].
17.         high_last--.
18.     num_shadowed = high_last - low_last + 1.
19.     计算对应阴影集的模糊度为
            fuzzy_s = sqrt(num_shadowed/n).
20.     计算 V(a).
21.         V(a) = fabs(fuzzy_s - fuzzy_f).
22.     if V(a) < V_last.
23.         更新 a_last 和 V_last.
24.     else    return a_best = a_last.
```

图 1　最佳阈值的计算算法

可以看到，在对 blongs[]进行排序后，最多扫描 blongs[]一遍，因此，搜索最佳阈值的时间复杂度为 $O(n)$，算法的时间复杂度为 $O(n\lg n)$.

4.2　可信任数据集合的样本选择方法

可信任数据集合采用子样本空间样本选择的方法. 这里的子空间指的是已选数据集张成的空间. 该方法是一种类内迭代的选样方法，每次迭代中选择距离子空间最远的样本点添加到已选数据集中，直到满足误差界或选样个数. 本质上说，该方法是对原样本集空间维数的一个快速逼近，选样个数最多是特征空间的维数. 该方法不能完全逼近原数据集的凸包，即不能找到可信任数据中的所有支持向量，但实际中并不需要精确逼近原数据集的凸包就能保证支持向量机的泛化能力.

4.3　不确定数据集合的样本选择方法

不确定数据集合采用基于边界向量提取方法进行样本选择. 首先，该方法利用支持向量描述的方法确定类别的中心，认为这两个中心肯定不是支持向量；然后以这两个中心的连线为直径确定一个圆，并定义圆内的样本为边界向量. 模糊支持向量机的支持向量肯定位于圆内，选择边界向量进行训练，因此减少了训练时间.

4.4　基于阴影集划分的样本选择方法

本文提出的样本选择基本流程如图 2 所示. 其基本思路为：先利用阴影集思想判定模糊集中的样本是可信任、不可信任还是不确定数据；其次，由于可信任数据是样本集中可确定类别的样本，两类的可信任数据集是可分的. 因此，对可信任数据集采用子空间样本选择方法，对不确定数据采用边界向量提取方法；最后将两部分样本合并得到最终的精简模糊样本集.

5　模糊隶属度的确定

本文在文献[18]的基础上，修改隶属度映射公式为

$$\xi_i = \begin{cases} (1-\lambda)\left(\dfrac{1-\dfrac{d_i}{r}}{1+\dfrac{d_i}{r}}\right)^{\frac{1}{\sigma_1}} + \lambda, d_i \leqslant r; \\ \lambda(1+d_i-r)^{\sigma_2}, d_i > r. \end{cases}$$

式中：ξ_i 为某样本点的模糊隶属度；d_i 为该样本点距离该类中心的距离；r 为包含该类大多数样本点的最小超球的半径；λ 为最小包围球内外样本点的临界隶属度，一般取为 0.4；σ_1、σ_2 分别为控制隶属度变化范围的一个尺度因子，这里分别取为 2.0 和 4.0.

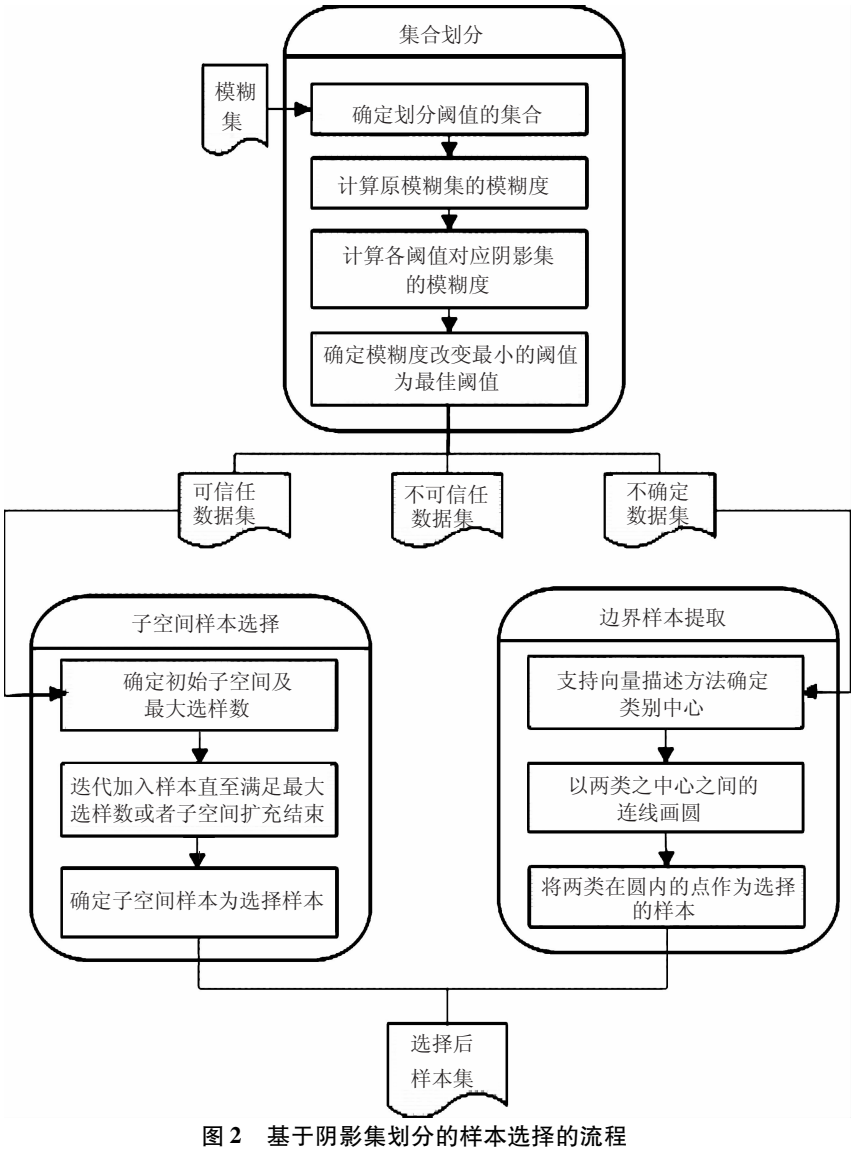

图 2　基于阴影集划分的样本选择的流程

6　实验结果与分析

实验在 Pentium IV 1.6 GHz CPU,1.00 GB 内存的 PC 上执行. 为方便比较,FSVM 中,核函数取为径向基函数,FSVM 都为 C-FSVM 模型. 参数利用 libsvm[19] 工具包选取对传统 SVM 分类效果最好的参数形式.

6.1　实验 1:仿真数据实验

为直观给出本文方法选样点的位置信息,该实验主要分析 2 维数据的线性不可分问题. 由 MATLAB 生成的两类分布不同的随机数. 正类和反类分别为以[0.25,0.50]和[0.75,0.50]为期望的 50 个二维正态随机数集合.

训练样本点及数据划分如图 3 所示. 集合中样本点按照上述模糊隶属度公式确定样本的隶属度. 求解出最优的划分阈值后,就得到了对应的阴影集. 可见,两个类的可信任数据是可分的,不确定数据在可信任数据外围,有部分的交叉. 不可信任数据则是离中心较远的数据,可视为噪声,直接舍去. 在两类中,一部分样本点远离另一类,也被认为是不可信任数据而舍弃了. 这是因为模糊隶属度是按样本点到类中心的距离计算的,这部分数据离中心较远,被赋予了较小的模糊隶属度. 但由于这部分数据肯定不是支持向量,不管其舍弃与否,都对 FSVM 的训练结果没有影响.

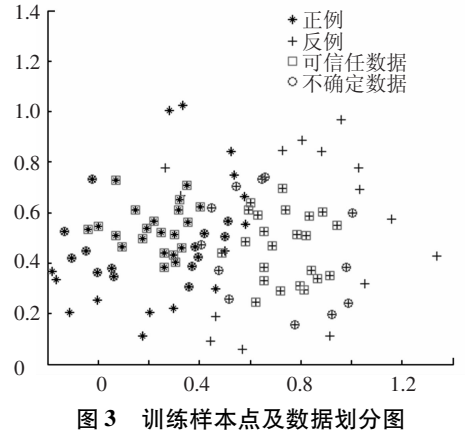

图 3　训练样本点及数据划分图

选样前后分类效果对比如图 4 所示. 图中，较细的线表示选样前样本训练得到的分类面，较粗的线表示选样后样本训练得到的分类面. 从分类效果上来看，用原模糊集进行训练得到的分类面较好，错分了 9 个样本点，选样后的样本训练得到的分类面错分了 10 个样本点. 从泛化能力上来看，实际的分类面应该是过(0.5,0)的垂直线. 选样后的样本点训练得到的分类面要稍优于不选样训练得到的分类面.

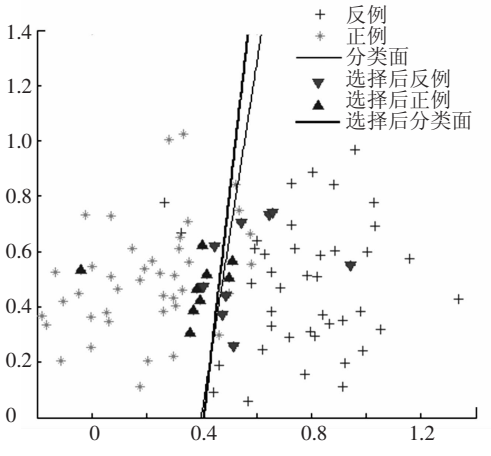

图 4　选样前后分类效果对比

6.2　实验 2：真实数据集实验

本文选择 Statlog[20] 中的 german. numer 数据集，对数据归一化后，计算模糊隶属度，得到需要的模糊集.

6.2.1　子实验 1：阈值比较

这里，模糊度的计算分别采用 Hamming 和 Euclid 两种方法. 文献[9]和本文方法得到的阈值如表 1 所示.

表 1　阈值比较

文献[9]	本文方法	
	Hamming 模糊度	Euclid 模糊度
0.439 106	0.404 333	0.413 917

可以看到，文献[9]确定的阈值比本文确定的阈值要大，映射增大了集合的模糊度.

6.2.2　子实验 2：可信任数据选样数对结果的影响

对可信任数据，选样的最大数量是人为定义的. 不难发现，随着选择样本的最大数量的增大，对测试集的识别率会提高，但选样时间也会增大. 每类样本的最大选样数量和预测精度的关系如图 5 所示.

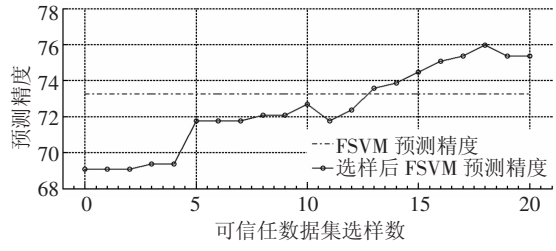

图 5　可信任数据集选样数与预测精度关系

这里，原样本训练得到的 FSVM 的预测精度为 73.273 3%. 可以看出，对可信任数据，当每类最大选样点逐渐增加时，预测的正确率呈上升趋势，并且对于 german. numer 数据集，当每类选样数为 18 时，对测试集的预测精度最高. 其后，随着选样数的增加，预测精度基本保持. 可以认为，对可信任数据，只要一个很小的选样数，就能保证预测的精度.

当每类选样数为 18 时，利用本文方法进行样本选择的时间为 0.031 s，挑选后的样本利用 FSVM 训练的时间为 0.032 s，该方法的总时间为 0.063 s. 而不进行样本选择的训练 FSVM 的时间为 0.3 s. 可见选样后大大减少了样本的训练时间.

6.3　实验 3：结果分析

实验采用的数据集包括 TKH96a[21] 中 fourclass，Statlog 中的 german. numer，NIPS 2003 Feature Selection Challenge[22] 中的 gisetts. 数据集说明如表 2 所示.

表 2　数据集说明

数据集	属性维度	训练集		测试集	
		正例	反例	正例	反例
gisetts	13	80	100	40	50
fourclass	2	205	370	102	185
german. numer	24	200	467	100	233

这里主要从预测精度以及运行时间上验证本文算法的有效性. 实验比较了选样前后，训练 FSVM 的时间，以及对测试集的预测精度. 这里，选样率指选择的样本数占原模糊集样本数的比例. 实验结果如表 3 所示.

表 3　实验对比结果

数据集	选样率	选样后 FSVM			不选样 FSVM	
		准确率/%	选样时间/s	训练时间/s	准确率/%	训练时间/s
gisetts	0.19	82.222	0.004	0.004	78.888 9	0.027
fourclass	0.20	100	0.019	0.017	100	0.079
german. numer	0.18	75.976	0.018	0.037	73.273 3	0.306

对 fourclass 数据集的可信任数据,当每类最大选样数为 23 时,预测准确率为 100%;当每类最大选样数为 16 时,预测准确率为 99.651 6%(错误预测一个样本).对可信任数据,若每类选样 16 个,则在基本保持预测精度的前提下,还可进一步减少运行时间.

一般来说,边界向量提取的方法的选样率在 0.5 左右,本文提出的方法将选样率降低至 0.2 左右,大大降低了选样率.可以看出,采用本文方法,可大大减少 FSVM 的训练时间,且能保证预测的准确率.

7 结 论

1)利用阴影集的思想挖掘中的确定信息,将 SVM 的样本选择方法应用于 FSVM 中,大大减少了选样率.

2)给出了新的阴影集阈值确定方法,并给出了基于阴影集的模糊支持向量机样本选择算法.

3)实验表明,本文提出的选样方法在保证 FSVM 的泛化能力的前提下,大大降低了选样率,减少了 FSVM 的训练时间.

参考文献:

[1] LEE J. Measurement of machine performance degradation using a neural network model [J]. Computers in Industry, 1996, 30(3): 193 – 209.

[2] 郭磊,陈进. 设备性能退化评估与预测研究综述[J]. 振动与冲击,2008,27(S):139 – 142.

[3] VENKATASUBRAMANIAN V, RENGASWAMY R, YIN K, *et al.* A review of process fault detection and diagnosis part I: quantitative model-based methods [J]. Computers and Chemical Engineering, 2003, 27(3): 293 – 311.

[4] 柴天佑,丁进良,王宏,等. 复杂工业过程运行的混合智能优化控制方法[J]. 自动化学报,2008, 34(5): 505 – 515.

[5] 吕琛,王桂增,张泽宇. PWM 型 VLSI 神经网络在故障诊断中的应用[J]. 自动化学报,2005,31(2): 195 – 201.

[6] VAPNIK V. The nature of statistical learning theory [M]. New York: Springer, 1995.

[7] LIN Chunfu, WAN Shengde. Fuzzy support vector machine [J]. IEEE Trans. on Neural Networks, 2002, 13(2):464 – 471.

[8] PEDRYCZ W. Shadowed sets: representing and processing fuzzy sets [J]. IEEE Trans on Systems, Man and Cybernetics— Part B: Cybernetics, 1998, 28(1): 103 – 109.

[9] PEDRYCZ W. From fuzzy sets to shadowed sets: interpretation and computing[J]. International Journal of Intelligence System, 2009, 24(1):48 – 61.

[10] LEE Y J, MANGASARIAN O L. RSVM: reduced support vector machines[R]. Wiscosin: University of Wisconsin, 2000. Chicago, 2001.

[11] ZHENG Songfeng, LU Xiaofeng, ZHENG Nanning, *et al.* Unsupervised clustering based reduced support vector machine [C]//Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP). Piscataway, NJ: IEEE, 2003: 821 – 824.

[12] ALMEIDA M B, BRAGA A P, BRAGA J P. Svm-km: speeding svms learning with a priori cluster selection and *k*-means [C]//Proceeding of the 6th Brazilian Symposium on Neural Networks. Washington, DC: IEEE Computer Society, 2000: 162 – 167.

[13] 李红莲,王春花,袁保宗,等. 针对大规模训练集的支持向量机学习策略[J]. 计算机学报,2004, 27(5): 140 – 144.

[14] LI Yuangui, HU Zhonghui, CAI Yunza, *et al.* Support vector based prototype selection method for nearest neighbor rules [M]. Berlin, Germany: Springer, 2005:528 – 535.

[15] 姜文瀚. 模式识别中的样本选择研究及应用[D]. 南京:南京理工大学, 2007.

[16] 吴青,刘三阳,杜喆. 基于边界向量提取的模糊支持向量机方法[J]. 模式识别与人工智能, 2008, 21(3): 332 – 337.

[17] 胡宝清. 模糊理论基础[M]. 武汉:武汉大学出版社, 2004.

[18] 张翔,肖小玲,徐光祐. 基于样本之间紧密度的模糊支持向量机方法[J]. 软件学报,2006,17(5):951 – 958.

[19] CHANG Chih-chung, LIN Chih-jen. LIBSVM: a library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 27.

[20] BRAZDIL P. Statlog dataset [EB/OL]. http://www.liacc.up.pt/ML/old/statlog/datasets.html.

[21] HO T K, KLEINBERG E M. Building projectable classifiers of arbitrary complexity [C]//Proceedings of the 13th International Conference on Pattern Recognition. Washington, DC: IEEE Computer Society, 1996:880 – 885.

[22] GUYON I, GUNN S, HUR A B, *et al.* Result analysis of the NIPS 2003 feature selection challenge [J]. Neural Information Processing Systems, 2005, (17):545 – 552.

(编辑 张 红)