DOI: 10.11992/tis.201904047
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.tp.20190722.1417.006.html

采用划分融合双向控制的粒度支持向量机

赵帅群[1], 郭虎升[1,2], 王文剑[2]
（1. 山西大学 计算机与信息技术学院, 山西 太原 030006; 2. 山西大学 计算智能与中文信息处理重点实验室, 山西 太原 030006）

摘　要: 粒度支持向量机 (granular support vector machine, GSVM) 引入粒计算的方式对原始数据集进行粒度划分以提高支持向量机 (support vector machine, SVM) 的学习效率。传统 GSVM 采用静态粒划分机制, 即通过提取划分后数据簇中的代表信息进行模型训练, 有效地提升了 SVM 的学习效率, 但由于 GSVM 对信息无差别的粒度划分导致对距离超平面较近的强信息粒提取不足, 距离超平面较远的弱信息粒被过多保留, 影响了 SVM 的学习性能。针对这一问题, 本文提出了采用划分融合双向控制的粒度支持向量机方法 (division-fusion support vector machine, DFSVM)。该方法通过动态数据划分融合的方式, 选取超平面附近的强信息粒进行深层次的划分, 同时将距离超平面较远的弱信息粒进行选择性融合, 以动态地保持训练样本规模的稳定性。通过实验表明, 采用划分融合的方法能够在保证模型训练精度的条件下显著提升 SVM 的学习效率。

关键词: 支持向量机; 粒度支持向量机; 划分; 融合; 强信息粒; 弱信息粒; 动态机制; 双向控制
中图分类号: TP18　　**文献标志码:** A　　**文章编号:** 1673−4785(2019)06−1243−12

中文引用格式: 赵帅群, 郭虎升, 王文剑. 采用划分融合双向控制的粒度支持向量机 [J]. 智能系统学报, 2019, 14(6): 1243–1254.
英文引用格式: ZHAO Shuaiqun, GUO Husheng, WANG Wenjian. Granular support vector machine with bidirectional control of division-fusion[J]. CAAI transactions on intelligent systems, 2019, 14(6): 1243–1254.

Granular support vector machine with bidirectional control of division-fusion

ZHAO Shuaiqun[1], GUO Husheng[1,2], WANG Wenjian[2]
(1. School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China; 2. Key Laboratory of Computational Intelligence and Chinese Information Processing, Shanxi University, Taiyuan 030006, China)

Abstract: Granular support vector machine (GSVM) introduces the method of granular computing to divide the original dataset; therefore, GSVM improves the efficiency of the support vector machine (SVM). The traditional GSVM adopts the static granules partitioning mechanism to extract representative information from the divided data clusters for model training, which can effectively increase the learning efficiency of the SVM. However, the GSVM uses the same processing way for different information granules, which may lead to a decline in the generalization ability because of two reasons: (i) No sufficient valid information is extracted from the strong information granules that are close to the hyper-plane, and (ii) excess of the weak information of granules far from the hyper-plane is reserved. These all reduce the learning performance of the SVM. To address this problem, this study proposes a division and fusion SVM model based on dynamical granulation, namely DFSVM. With the DFSVM, the information from the strong information granules near the hyper-plane is divided in depth, and weak information from weak information granules far from the hyper-plane is selectively merged to dynamically maintain the stability of the size of the training samples. The experiments demonstrate that this model can significantly improve the SVM learning efficiency, ensuring the training precision of the model.

Keywords: support vector machine (SVM); granular support vector machine (GSVM); division; fusion; strong information granule; weak information granule; dynamic mechanism; bidirectional control

收稿日期: 2019−04−19.　**网络出版日期:** 2019−07−23.
基金项目: 国家自然科学基金项目 (61673249, 61503229, U1805263); 山西省回国留学人员科研基金项目 (2016-004).
通信作者: 王文剑. E-mail: wjwang@sxu.edu.cn.

支持向量机 (support vector machine, SVM) 是由 Vapnik 等[1] 提出的基于统计学习理论和结构风险最小化准则的一种学习策略, 在小样本多维

度的数据分类和回归问题方面表现出了优良的泛化性能，广泛应用于机器学习、模式识别、模式分类、图像处理等领域[2-8]。目前在大规模数据处理方面，SVM 仍存在一些不足。主要问题是当样本数 n 较大时，会消耗大量的内存空间和运算时间，严重降低了 SVM 的学习效率，限制了 SVM 在大规模数据集上的应用。

粒度支持向量机的含义最早由 Tang 等[9] 提出，其主要思想是首先构建粒度空间获得一系列信息粒，然后在每个信息粒上进行 SVM 学习，最后聚合信息粒上的信息获得最终的决策函数。依据粒划分方式的不同，衍生出了基于聚类的 GSVM、基于分类的 GSVM 以及基于关联规则的 GSVM 等方法[10-19]。GSVM 采用粒化的方式压缩数据集的规模，以提高 SVM 的学习效率，而目前的 GSVM 大都在静态层级进行划分，即只对信息粒进行有限次的浅层次划分，丢失了大量对分类起关键作用的样本信息，且冗余信息较多，降低了模型的性能。尽管已经提出的动态粒度支持向量机 (dynamic granular support vector machine, DGSVM)[20]，以及动态支持向量回归机 (dynamic granular support vector regression, DGSVR)[21]，采用动态的方式对重要信息粒深层次划分，对无关信息粒则进行浅层次划分，但 DGSVM 随着粒划分过程会使数据规模不断增加，使得 SVM 的效率有所降低。

为了进一步提升 SVM 在大规模数据集上的应用能力，本文提出了采用划分融合双向控制的粒度支持向量机方法。在 SVM 分类过程中，对分类起关键重要的信息分布于超平面附近，称为强信息区，超平面远端的信息对分类影响较小，称为弱信息区，本文提出的方法通过对强信息区的强信息粒进行深度划分，同时融合弱信息区的弱信息粒，使训练数据始终动态保持在较小规模。该方法分为两个阶段，首先通过聚类算法对原始数据集进行初始粒划分，挑选粒中代表信息组成新的训练集训练得到初始分类超平面，然后通过迭代划分融合的方式深度划分强信息粒，同时融合远端弱信息粒。实验表明，该方法能够在保证模型精度的条件下显著提升 SVM 的学习效率。

1　粒度支持向量机

粒度支持向量机引入粒计算的概念，对复杂问题进行抽象和简化，以较低的代价来得到问题的满意近似解。在多种的粒划分方式中，基于聚类的粒度支持向量机 (clustering-based granular support vector machine, CGSVM) 是当前研究的热点之一[22-25]。CGSVM 通过聚类算法将大规模数据集分解成多个小规模数据簇，簇内信息具有高度相似性，而簇间信息相似度较低，挑选出每个簇中具有代表性的样本信息作为新的训练样本，整合所有挑选出的样本训练得到新的模型。

CGSVM 只采用了少量代表样本作为训练集，有效地加速了 SVM 学习过程。但 CGSVM 本身也存在一些不足，在 SVM 学习过程中，距离分类超平面较近的信息对分类起关键作用，而距离超平面较远的信息几乎不影响模型训练过程。CGSVM 在数据处理过程中没有区分不同信息粒对分类的影响程度，对所有信息粒都进行同等层次的划分，导致对重要信息提取不足且仍存在过多的冗余信息。如图 1 中，距离超平面较近的 G_1^+、G_2^+、G_3^+、G_1^-、G_2^-、G_3^- 中包含较多支持向量信息，对分类起到了关键作用，距离超平面较远的 G_5^+、G_6^+、G_5^-、G_6^- 对分类影响较小。尽管 DGSVM 通过对超平面附近重要信息粒深度划分，但远端的冗余信息仍然被保留，在动态划分过程中数据规模会不断增加，导致训练时间也不断地提高。

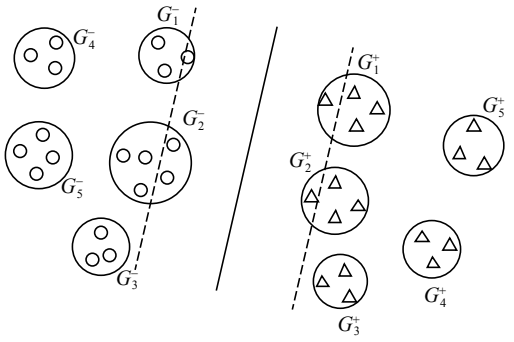


图 1　CGSVM 粒度划分
Fig. 1　CGSVM granular division

2　DFSVM 模型

现阶段 CGSVM 通过静态的、浅层次的方式，对粒划后的信息粒进行无差别的信息提取，导致对分类起关键作用的信息提取不足且还保留了大量对分类影响较小的冗余信息。本文提出的方法采用多层次的划分策略，由于超平面附近的样本信息有较大概率成为支持向量，距离超平面较远的样本信息对分类几乎没有影响，因此，DFS-

VM 采取动态迭代划分的方式, 对超平面附近可能成为支持向量的信息粒深度划分, 同时融合距离超平面较远的冗余信息, 不断更新超平面以获得更多潜在有效的分类信息, 该方法能够将训练集始终固定在一个较小的规模, 加速了 SVM 的训练过程。

2.1　初始粒划分

给定原始数据集 $D=\{X,y\}=\{(x_1,y_1),(x_2,y_2),\cdots,(x_e,y_e)\}$, $y_e\in\{1,-1\}$, $x_e\in R^l$, DFSVM 首先通过聚类算法将数据集中的正类与负类样本分别划分为 k 个粒, 通过初始粒划分方式得到新的信息粒集:

$$D'=\{(G_1^+,G_2^+,\cdots,G_k^+)\cap(G_1^-,G_2^-,\cdots,G_k^-)\}$$

式中: G_k 表示通过划分得到的信息粒。SVM 通过核函数 $K(x,y)=\varphi(x)\varphi(y)$ 将数据映射到 N 维核空间, 将数据集经过初次划分在 N 维空间形成的粒称为超粒, 第 i 个超粒的中心 u_i 和半径 γ_i 为

$$\mu_i=\frac{\sum_{p=1}^{n_i}\varphi(x_p)}{n_i}=\frac{1}{n_i}\sqrt{\sum_{p=1}^{n_i}\sum_{q=1}^{n_i'}K(x_p,x_q)} \tag{1}$$

$$\gamma_i=\frac{\max(x_l)-\min(x_s)}{2}=\frac{1}{2}\sqrt{\varphi(x_l)^2-2\varphi(x_l)\varphi(x_m)+\varphi(x_m)^2}=\sqrt{K(x_l,x_l)-2K(x_l,x_m)+K(x_m,x_m)} \tag{2}$$

式中: $\varphi(x_l)$ 和 $\varphi(x_m)$ 代表核空间上下边界, 超粒半径通过其平均值衡量, 样本 $\varphi(x_s)$ 到任意超粒 G_i 中心 μ_i 的距离可表示为

$$\mathrm{dis}(\varphi(x_s),\mu_i)=\sqrt{K(x_s,x_s)-\frac{2}{n_i}\sum_{j=1}^{n_i}K(x_s,x_p)+\frac{1}{n_i^2}\sum_{p=1}^{n_i}\sum_{q=1}^{n_i'}K(x_p,x_q)} \tag{3}$$

通过初始粒划分将原始数据集划分为 G_1、G_2、$\cdots G_k$ 个粒, 提取每个粒中的代表信息以训练获得初始分类超平面。

2.2　动态划分融合方法

通过初始粒划过程获得超平面 $y=\boldsymbol{W}^{\mathrm{T}}\cdot\varphi(x)+b$, 在 SVM 模型分类过程中, 对分类起关键作用的样本信息主要分布在最大间隔内部以及间隔线附近, 该区域的样本在模型训练过程中会被多次遍历, 而位于超平面相对较远的样本无需过多的遍历即可将其分类正确。因此, 基于以上条件将样本划分为强信息区与弱信息区。给出两个参数 β^+ 和 β^-, 其中 $\beta^+>\beta^-$。当样本与超平面之间的距离满足 $D'\leqslant\gamma/2+\beta^-$ 时, 样本点对分类超平面具有重要影响, 划分为强信息区。同理, 样本与超平面之间的距离 $D'\geqslant\gamma/2+\beta^+$ 时, 认为样本点对分类超平面影响较小, 划分为弱信息区。其中 β^- 可在 0 至 $\gamma/2$ 之间选取, β^+ 可在 $\gamma/2$ 至 γ 间选取。强信息区的信息有较大可能在迭代融合划分过程中成为支持向量, 弱信息区的数据则对分类影响较小, 对强信息粒区域采用划分方式提取分类信息, 对弱信息区采用融合方式减少冗余信息。其中, 超平面最大间隔 γ 为

$$\gamma=\frac{2}{\|\boldsymbol{W}\|} \tag{4}$$

针对每个划分好的信息粒, 选择中心点 μ_i 作为代表点计算该粒到超平面之间的距离, 公式如下:

$$D'=\frac{\boldsymbol{W}^{\mathrm{T}}\cdot\varphi+b}{\|\boldsymbol{W}\|}=\frac{\sum_{i=1}^{n'}\alpha_i y_i k(x_i,\mu_i)+b}{\sqrt{\left(\sum_{i=1}^{n'}\alpha_i y_i x_i\right)^2}} \tag{5}$$

动态划分过程通过衡量粒与超平面之间距离来选取候选粒进行深度划分。但由于不同粒的大小、粒内部数据分布等差异, 密度较大的粒中信息分布集中、重叠度大, 含有更多潜在成为支持向量的信息; 密度较小的粒中信息分布稀疏, 包含的支持向量信息少。因此, 对超平面附近密度较大的信息粒优先选择在当前迭代过程中划分, 密度相对较小的信息粒可能成为后续划分过程中的候选粒。为了衡量每个粒的差异程度, 给出粒密度的定义:

$$\rho_i=\frac{n_i}{\gamma_i}=\frac{n_i}{\sqrt{K(x_l,x_l)-2K(x_l,x_m)+K(x_m,x_m)}} \tag{6}$$

式中: n_i 为第 i 个粒中的样本数; γ_i 为第 i 个粒的半径。

图 2 表示 DFSVM 动态粒划过程, 其中 G_1^+、G_2^- 被选为当前最优分类信息粒, G_1^+ 被划分为 G_{d1}^+、G_{d2}^+, G_2^- 被划分为 G_{d1}^-、G_{d2}^-。同时将 G_4^- 和 G_5^- 融合为 G_m^-, G_4^+ 和 G_5^+ 融合为 G_m^+

2.3　DFSVM 算法

DFSVM 模型的数据处理过程分为两个阶段: 1) 对原始数据进行初始粒划分, 然后通过式 (1) 计算得到每个粒的粒心, 将所有粒心作为训练集训练得到初始分类超平面; 2) 利用动态划分融合的思想, 对信息粒不断迭代处理以获得最优分类超平面。首先通过式 (4) 与参数 β^+、β^- 划分强信息区与弱信息区, 利用式 (5)、(6) 计算这两个区域内每个粒与超平面的距离和自身的粒密度, 挑选强信息区距超平面较近且粒密度大的粒在当前迭

代过程进行划分，挑选弱信息区距超平面较远且粒密度小的粒在当前迭代过程进行融合，用划分后的超粒代替原始超粒。在该方式下，数据规模能够保持在较低水平，SVM 的学习效率也得到有效的提升。

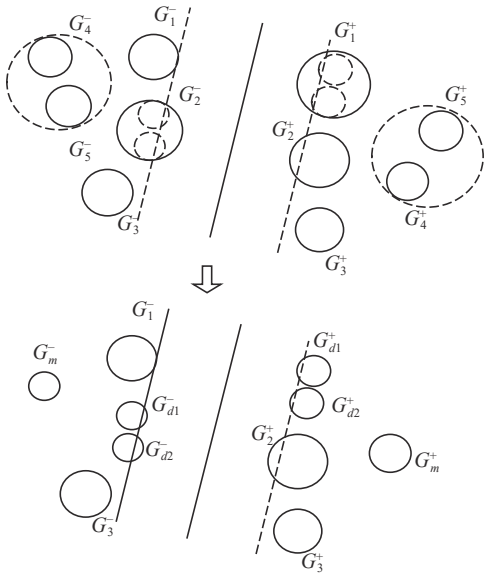


图 2　动态划分融合过程

Fig. 2　Dynamic division and fusion process

本文提出的 DFSVM 针对传统 SVM 无法高效的处理大规模数据以及 CGSVM 静态划分的不足进行了改进，探讨的目标是 DFSVM 是否能够在保证精度损失较少的情况下有效提升 SVM 的学习效率。本文在不同的参数下做了大量实验，基本算法描述如下：

算法　采用划分融合双向控制的粒度支持向量机

输入　原始数据集 D，初始粒化参数 k，动态粒化参数 m，迭代粒化参数 d，停止条件 t(预先设定的模型迭代次数)；

输出　划分融合过程得到的模型测试结果集。

1) 用聚类算法将数据集 D 中每一类划分为 k 个粒 $G_1,G_2,\cdots,G_k$；

2) 将划分后的每个粒中心加入到训练集中训练得到初始分类超平面 f'；

3) 通过式 (4) 和式 (6) 计算强信息区的信息粒与超平面的距离 D_i 以及粒密度 ρ_i，挑选当前需要划分的 d 个信息粒，并将这些信息粒分别深度划分为 m 个子粒；

4) 通过式 (4) 和式 (6) 计算弱信息区信息粒、超平面的距离 D_i 与粒密度 ρ_i，挑选出当前需要融合的 $d\times m$ 个弱信息粒；

5) 将更新后的信息粒代替原信息加入到训练集并更新分类超平面，同时记录模型测试结果；

6) 重复 4)~6)，直到满足停止条件 t；

7) 记录模型结果集，算法结束。

传统 SVM 模型训练的时间复杂度和空间复杂度分别为 $o(n^3)$ 和 $o(n^2)$，其中 n 为数据的规模。SVM 在模型训练过程中，需要存储和计算大规模的核矩阵，随着数据规模的增长，效率会大大降低。DFSVM 算法采用动态划分融合双向控制的方式对数据集进行迭代划分，始终将训练集维持在较小的规模，提高了模型的学习效率。尽管 DFSVM 在划分过程中会多次训练超平面，但训练总耗时仍然较少，并进一步改进了 CGSVM 静态单层划分对重要信息提取不足的缺点，针对于强信息粒进行信息提取，同时融合冗余的弱信息粒，降低训练规模的同时提升 CGSVM 的训练精度。DFSVM 模型在保证较高分类精度的条件下，有效地提升了模型的学习效率。

3　实验和分析

3.1　实验数据集

本文实验在多个 UCI 数据集和标准数据集上进行实验，见表 1，SVM 选用高斯核函数，在多种参数下进行实验。实验在一台 CPU 为 2.50 GHz，内存 8 GB 计算机上运行，实验平台为 Matlab2016a。

表 1　实验数据集

Table 1　Experimental data sets

数据集	样本总数	训练集	测试集	特征维度	数据比例
banana	8 726	6 982	1 744	2	1:1
thyroid	3 220	2 576	644	5	1:1
image	9 900	7 920	1 980	18	1:1
german	3 000	2 400	600	20	1:1
diabetis	5 360	4 288	1 072	8	1:1
spambase	3 200	2 560	640	57	1:1
splice	15 270	12 216	3 054	60	1:1
kdd-1999	100 000	80 000	20 000	41	1:1

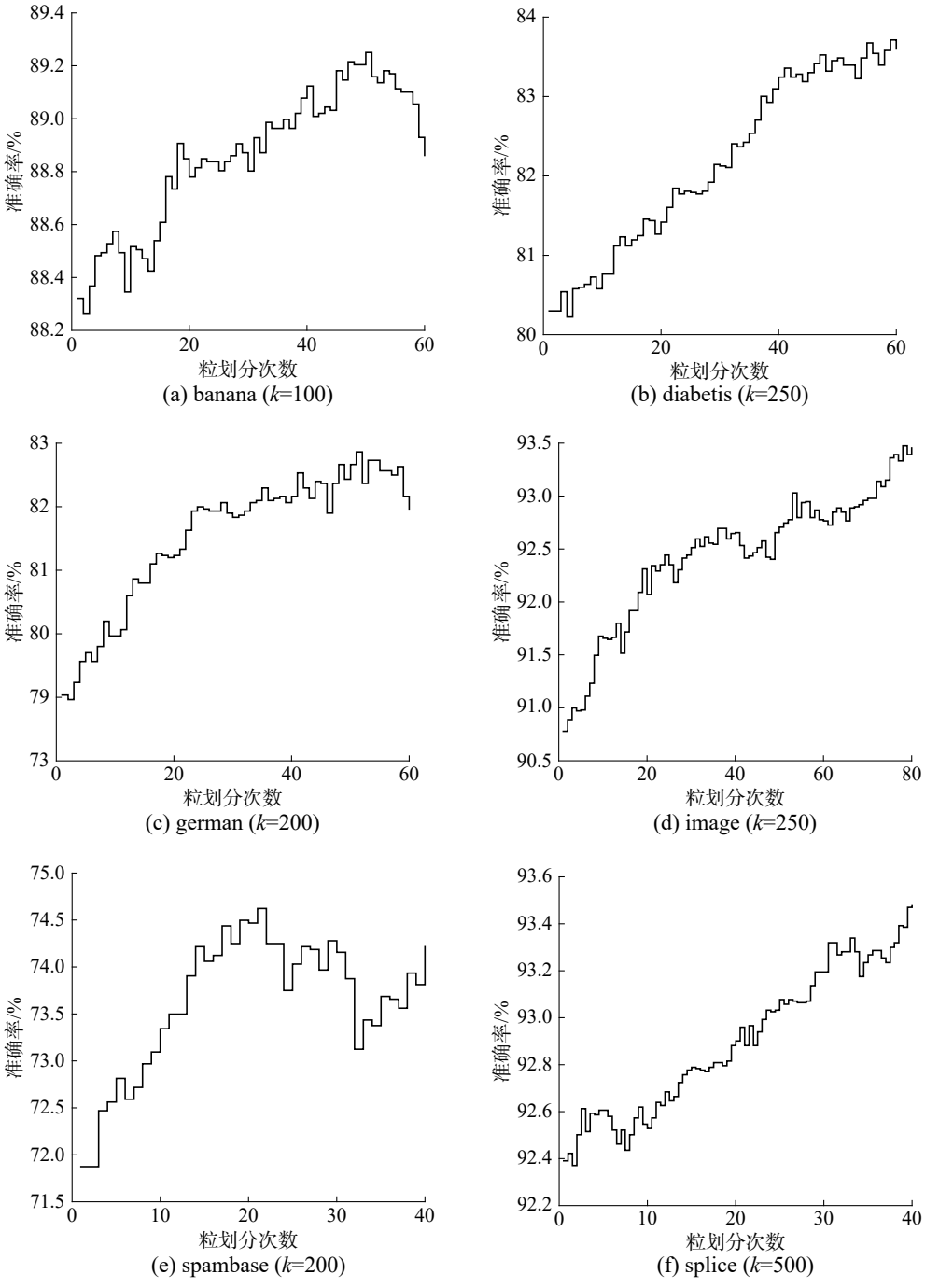
3.2　动态粒划分结果分析

本文提出的采用划分融合双向控制的粒度支持向量机模型，在粒划分过程中逐步提取潜在的支持向量信息，通过信息融合清除掉过多的冗余信息，提升 SVM 的学习效率。本小节实验验证 DFSVM 粒划分融合过程中对 SVM 泛化能力的影响。

由于初始参数 k 决定了动态划分融合阶段的数据规模, k 值过小会导致学习性能的下降, 过大会增加时间消耗, 因此对于不同数据集需要选择合适的参数值, 在 3.4.3 节中有相关参数讨论。为了尽可能观测粒划过程中预测准确率的变化, 本节实验设定迭代粒划参数 $d = 1$, 动态粒化参数 $m = 2$, 既每次将一个强信息粒划分为 2 个子粒, 同时将远端的两个弱信息粒进行融合, 图中初始结果即为 CGSVM 结果, SVM 惩罚因子 $c = 1$, 高斯核参数 $g = 1/k'$(k' 为特征数)。

从图 3 中可以看出, 在对数据集迭代划分融合过程中, SVM 的分类准确率逐步提高, 但不同数据集的变化情况也存在差异。

实验结果表明本文提出的方法能够充分提取数据集中的关键信息, 有效地提升了模型的学习效率。在有限次的数据处理过程中, 数据分布增强了对 SVM 的适应性, 但随着划分次数增加, 数据分布的改变可能导致 SVM 过拟合化, 降低模型性能, 如 spambase 数据显示出迭代次数大于 20 时, 准确率有明显下降趋势。实验表明采用划分融合双向控制的粒度划分方法在一定程度上具有普适性。



(a) banana (k=100)　(b) diabetis (k=250)

(c) german (k=200)　(d) image (k=250)

(e) spambase (k=200)　(f) splice (k=500)

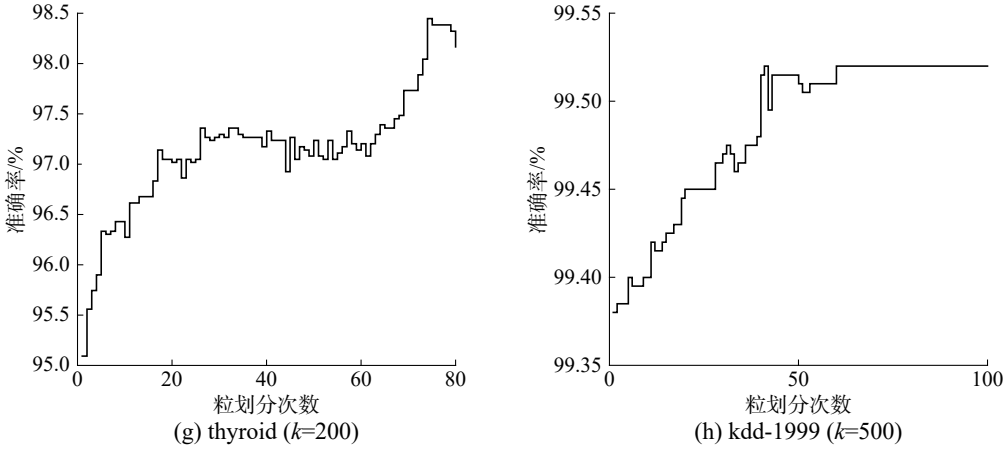


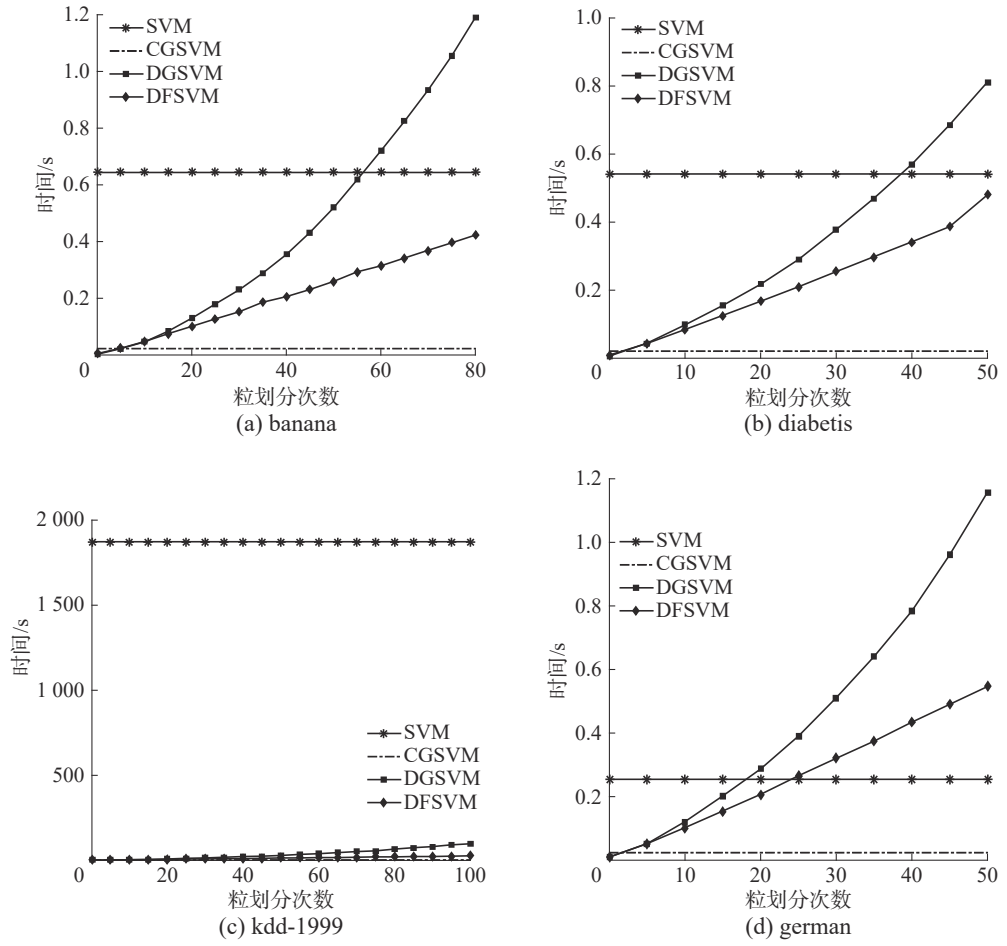
图 3　粒划分过程中精度变化

Fig. 3　Accuracy change during granules division process

3.3　模型精度与时间结果分析

针对在迭代过程中模型预测准确率和时间变化与传统 SVM、CGSVM、DGSVM 进行对比，参数选取与 4.2 节中实验相同，DGSVM 平均每次划分数据增量为 4，图 4 为时间对比图，图 5 为准确率对比图。

图 4 中的实验结果表明，随着迭代次数的增加，DGSVM 的训练时间增加率快于 DFSVM。实验在 german、thyroid、spambase 数据集上的预测准确率没有在有效粒划次数内达到最优，在其他数据集上都达到了最优值。图 5 中结果表明 DGSVM 的精度达到的峰值要高于 DFSVM，但时间消耗上要接近 DFSVM 的两倍，且高于传统 SVM 的训练时间。DGSVM 与 DFSVM 在传统 SVM 基础上通过数据压缩的方式降低了数据规模，提升了模型效率，而迭代次数会影响 DGSVM 与 DFSVM 的



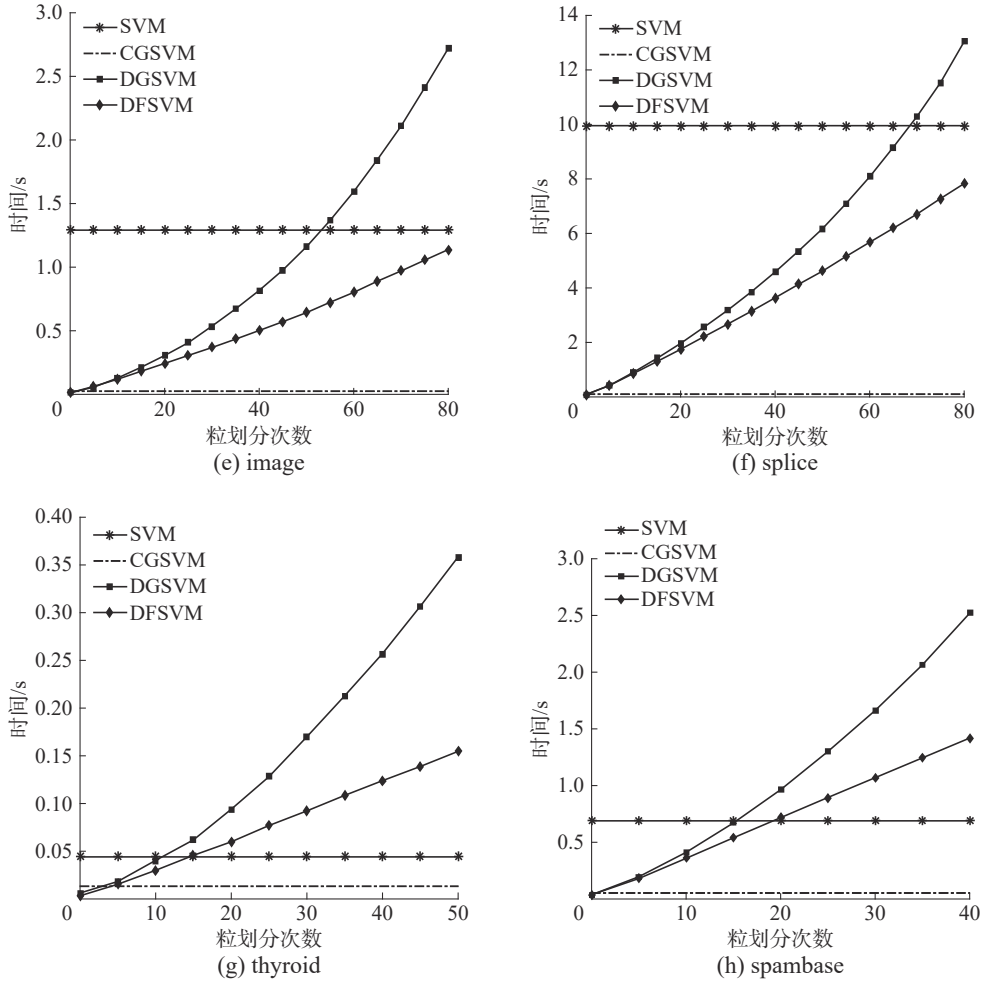


图 4　不同方法模型训练时间对比

Fig. 4　Comparison of model training time on different methods

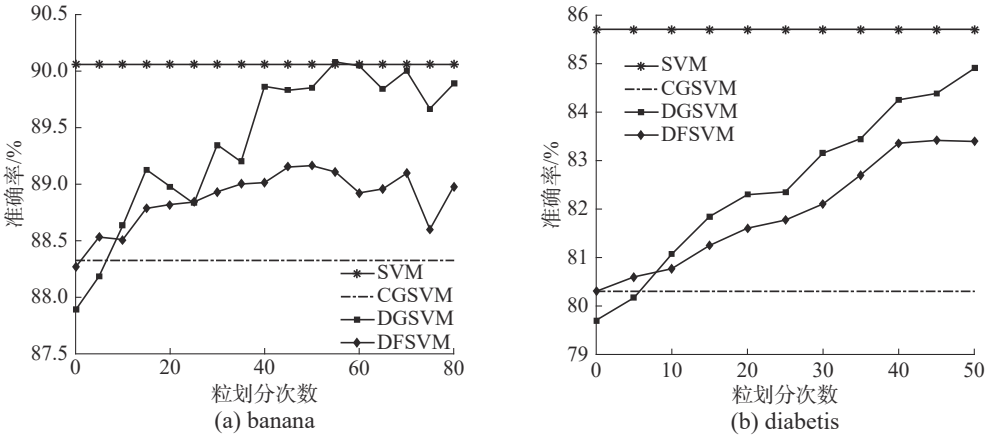
学习效率。DFSVM 通过划分融合的方式动态保持了数据规模的稳定, 而 DGSVM 的数据规模在划分的过程中不断增大, 导致训练时间增加。DFSVM 在时间上有明显的提升, 与 DGSVM 相比仍然损失了一些精度。

3.4　参数对 DFSVM 的影响

3.4.1　迭代参数与粒划分参数分析

DFSVM 迭代过程中参数 d 控制每次划分的粒数目, 参数 m 控制每个粒进行深度划分的数目, 其他参数与 3.2 节中设置相同。实验中准确率、时间和迭代次数分别采用模型训练结果达到稳定时的平均水平进行对比, 见表 2, 其中 acc 表示模型准确率, t 表示所用时间, h 表示动态迭代划分次数。

由表 2 中数据可以看出, 随着参数 d、m 的增大, 每次参与划分和融合的数据增多, 模型能够



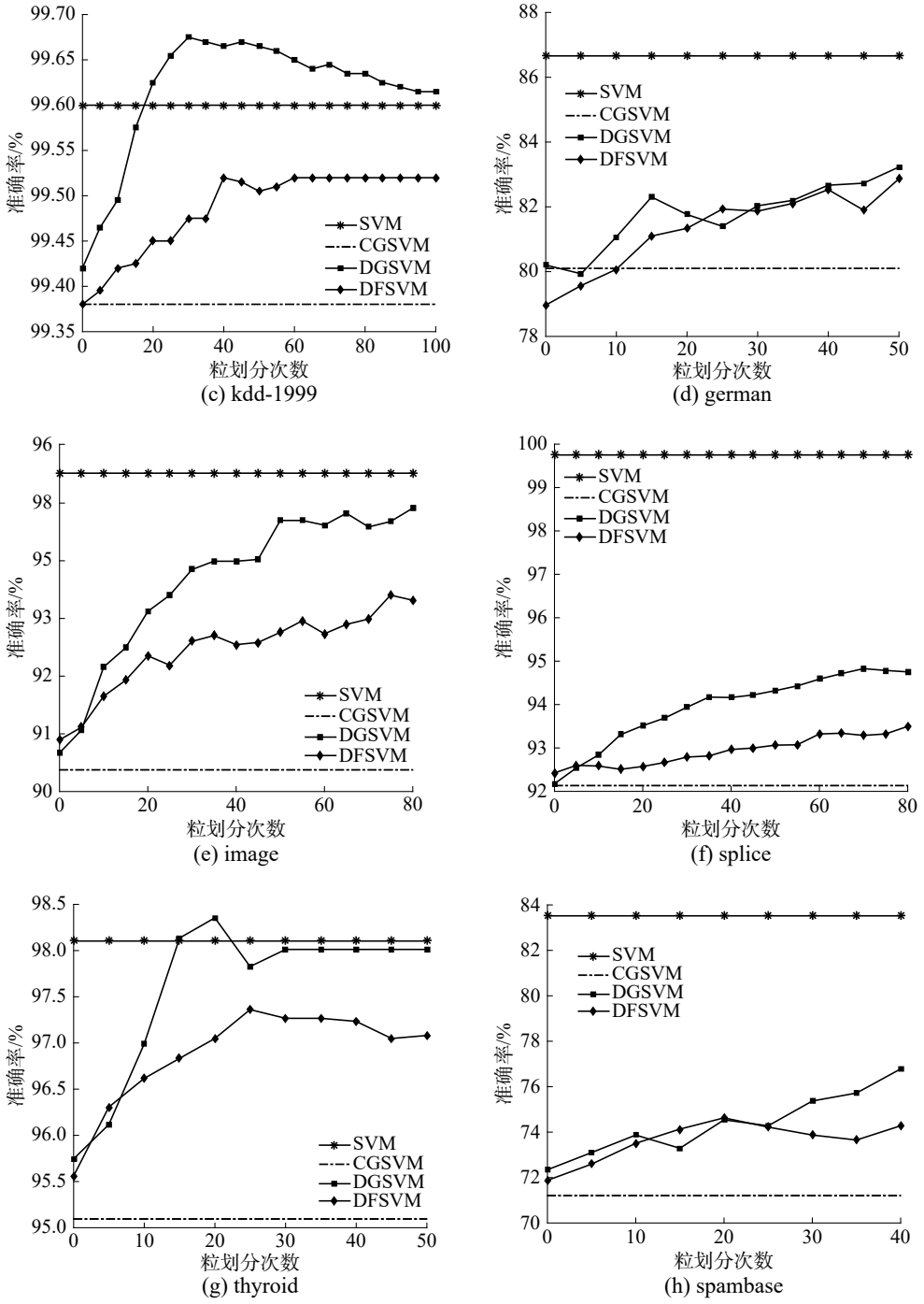


图 5　不同方法测试精度对比

Fig. 5　Accuracy comparison on different methods

在较少的迭代次数内收敛到最优值。由于数据集规模与分布的不同，结果存在一定的差异，预测结果波动范围较小，表明参数 d、m 在取值较大时能够降低算法迭代次数，有效缩短模型训练时间。

3.4.2　SVM 模型参数分析

本实验中主要调节 SVM 中参数惩罚因子 c 以及高斯核参数 g。实验选取不同 c、g 参数值进行实验，讨论惩罚因子及核参数对实验结果的影响，其余参数与 3.2 节中设置相同，模型预测结果见图 6，c、g 参数取值见表 3。如图 6，参数 c、g 的变化影响数据的最优性能，所有数据集都能够通过惩罚参数和核参数的调节来提高 DFSVM 的性能，而且大部分数据集在迭代过程中都表现出较好的稳定性，thyroid、spambase 数据集出现了一些离群点，但不影响总体结果。

3.4.3　初始聚类参数 k

动态划分首先要通过初始聚类参数 k 对数据进行压缩，压缩过小会因欠拟合而降低模型精

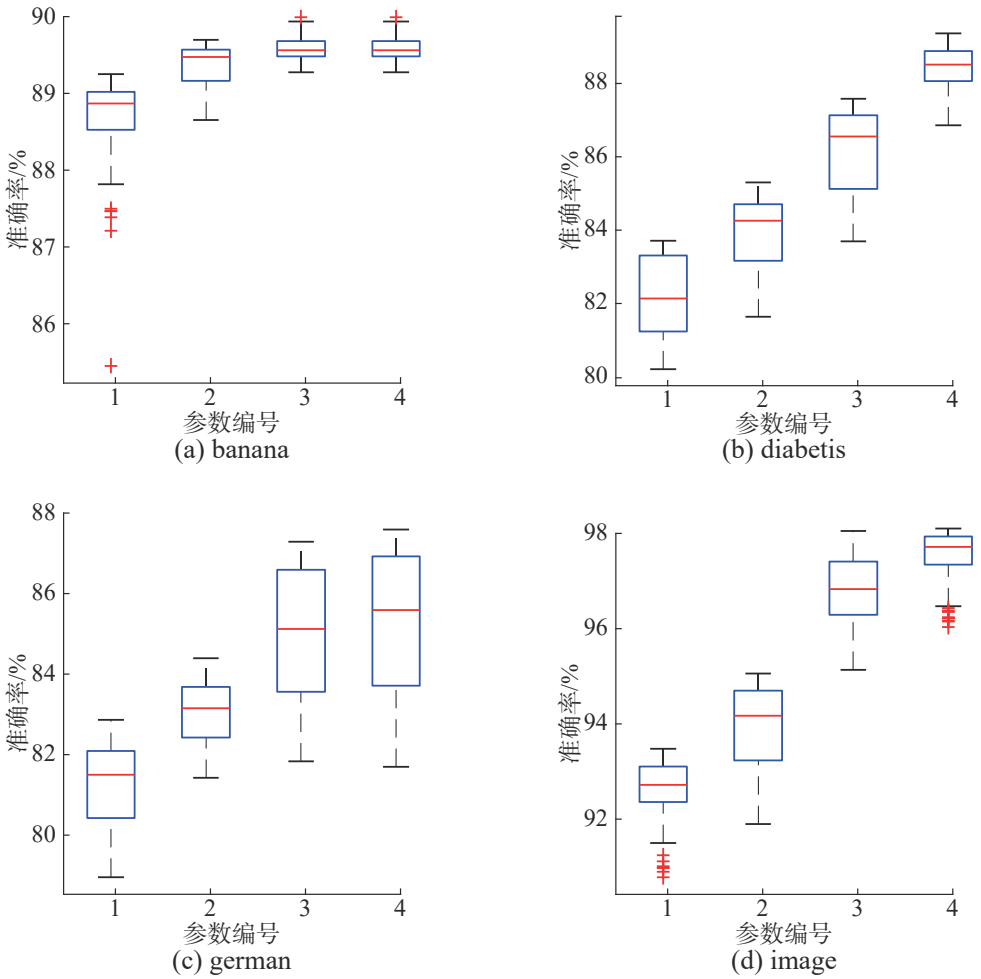
表 2　迭代参数 d 与粒划参数 m 实验结果

Table 2　The result on iteration parameter d and dividing parameter m

数据集	d=1, m=2			d=2, m=2			d=2, m=3			d=3, m=3		
	acc	t/s	h	acc	t/s	h	acc	t/s	h	acc	t/s	h
banana	89.250	0.253 4	50	89.318	0.112 6	25	88.513	0.079 2	16	89.341	0.052 0	10
diabetis	83.675	0.425 0	50	83.731	0.449 6	45	82.603	0.223 6	25	82.425	0.184 0	20
german	82.356	0.536 4	50	82.210	0.309 4	30	81.000	0.151 4	15	81.600	0.098 2	10
image	93.269	1.202 6	85	93.275	0.760 0	55	92.865	0.324 0	25	93.061	0.124 0	10
spambase	74.219	0.681 0	20	73.344	0.601 0	15	73.031	0.304 0	10	74.031	0.300 2	10
splice	93.863	12.384	120	94.353	7.608 4	80	94.357	7.085 8	70	94.178	6.322 4	60
thyroid	98.294	0.255 0	80	98.217	0.116 8	35	98.012	0.120 0	35	98.509	0.065 8	20
kdd-1999	99.520	15.350 0	60	99.660	6.230 0	30	99.645	7.180 0	30	99.630	4.651 0	20

度, 压缩过大则可能造成数据冗余而降低模型效率, 因此, 本节实验选取不同的参数 k 进行了实验分析, 其他参数与 3.2 节中设置相同, 测试结果见图 7。

由于不同数据集规模和分布差异, 参数 k 的选取也不同。从图 7 中可以看出, k 值在一定范围内增加会使模型准确率有所提升, 在 splice 和 german 数据结果中, 不同的参数 k 对应的曲线具有明显差异性, 但对于 diabetis 和 image 数据集, 参数 k 存在相对最优值, 即 k 高于某一值后对模型结果提升不明显。当 k 值较小时, 甚至会显著降低模型性能, 如 german 数据集在 k 取 100 时, 结果变差。实验表明, k 值的选取对模型结果有一定的影响。



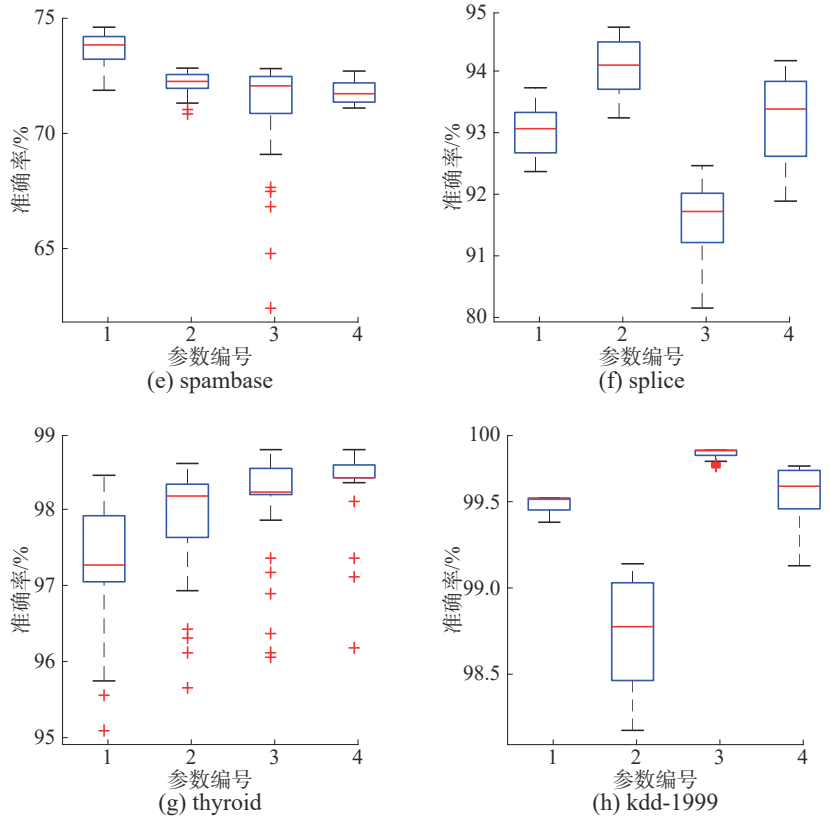


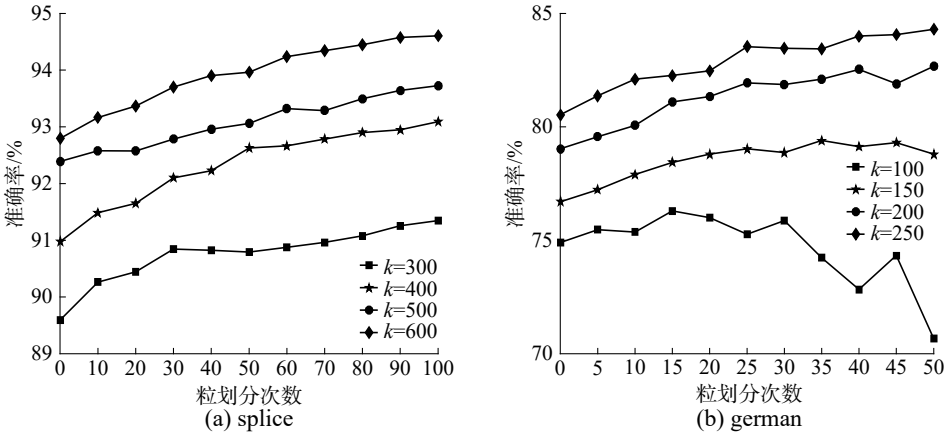
图 6　惩罚因子 c 与高斯核参数 g 的影响

Fig. 6　The effect of cost parameter c and RBF kernel parameter g

表 3　SVM 模型参数取值

Table 3　The value of the SVM model parameters

编号	banana		diabetis		german		image		spambase		splice		thyroid		kdd-1999	
	c	g	c	g	c	g	c	g	c	g	c	g	c	g	c	g
1	1	1/2	1	1/8	1	1/20	1	1/18	1	1/57	1	1/60	1	1/5	1	1/41
2	2	1/2	2	1/8	2	1/20	2	1/18	2	1/57	2	1/60	2	1/5	1/2	1/41
3	1	1	1	1/4	1	1/10	1	1/9	1	1/25	1	1/100	1	1/2.5	1	1/20
4	2	1	2	1/4	2	1/10	2	1/9	2	1/25	2	1/100	2	1/2.5	1/2	1/20



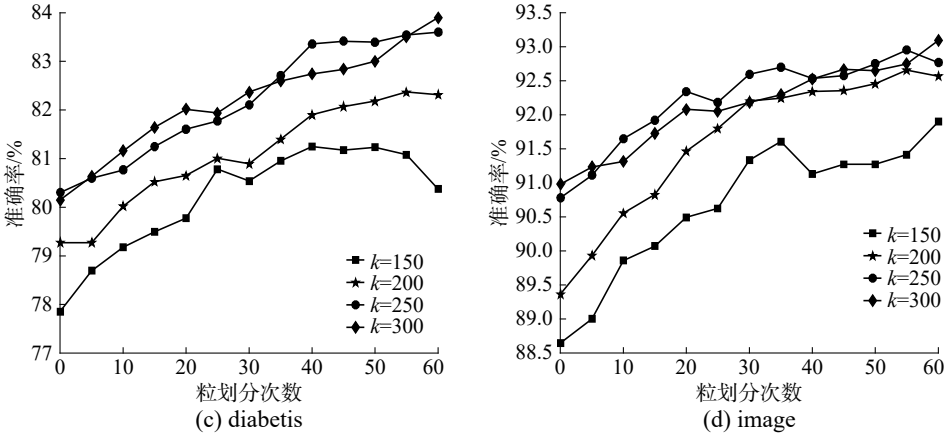


图 7　初始聚类参数 k 对测试结果的影响

Fig. 7　The effect of initial clustering parameter k on the experiment

4　结束语

本文在动态粒度支持向量机的基础上结合划分与融合的思想, 扩展了 SVM 在大规模数据集上应用的能力, 通过多种参数共同调节, 能够保证在精度损失较小的情况下, 提升 SVM 的学习效率。但在采用划分与融合的思想在数据处理过程中可能会改变数据集的分布, 限制了数据迭代划分次数, 参数调节也增加了模型的复杂度。在未来的工作中, 会继续针对该模型在实际应用问题中进行探讨, 在简化模型的同时保证模型的泛化性能。

参考文献:

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

[2] YUAN Ruixi, LI Zhu, GUAN Xiaohong, et al. An SVM-based machine learning method for accurate internet traffic classification[J]. Information systems frontiers, 2010, 12(2): 149–156.

[3] CHEN G Y, XIE W F. Pattern recognition with SVM and dual-tree complex wavelets[J]. Image and vision computing, 2007, 25(6): 960–966.

[4] REYNA R A, ESTEVE D, HOUZET D, et al. Implementation of the SVM neural network generalization function for image processing[C]//Proceedings of the 5th IEEE International Workshop on Computer Architectures for Machine Perception. Padova, Italy, 2000: 147–151.

[5] LIU Yang, WEN Kaiwen, GAO Quanxue, et al. SVM based multi-label learning with missing labels for image annotation[J]. Pattern recognition, 2018, 78: 307–317.

[6] XIONG Xiaoxia, CHEN Long, LIANG Jun. A new framework of vehicle collision prediction by combining SVM and HMM[J]. IEEE transactions on intelligent transportation systems, 2018, 19(3): 699–710.

[7] BISHWAS A K, MANI A, PALADE V. An all-pair quantum SVM approach for big data multiclass classification[J]. Quantum information processing, 2018, 17(10): 282.

[8] ZHOU Xueliang, JIANG Pingyu, WANG Xianxiang. Recognition of control chart patterns using fuzzy SVM with a hybrid kernel function[J]. Journal of intelligent manufacturing, 2018, 29(1): 51–67.

[9] TANG Yuchun, JIN Bo, SUN Yi, et al. Granular support vector machines for medical binary classification problems[C]//Proceedings of 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. La Jolla, USA, 2004: 73–78.

[10] YU H, YANG J, HAN Jiawei. Classifying large data sets using SVMs with hierarchical clusters[C]//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2003: 306–315.

[11] WANG Wenjian, XU Zongben. A heuristic training for support vector regression[J]. Neurocomputing, 2004, 61: 259–275.

[12] MAO Xueyu, SARKAR P, CHAKRABARTI D. Overlapping Clustering Models, and One (class) SVM to Bind Them All[J]. arXiv: 1806.06945, 2018.

[13] DING Shifei, QI Bingjuan. Research of granular support vector machine[J]. Artificial intelligence review, 2012, 38(1): 1–7.

[14] GUO Husheng, WANG Wenjian, MEN Changqian. A novel learning model-Kernel granular support vector machine[C]//Proceedings of 2009 International Conference on Machine Learning and Cybernetics. Hebei, China, 2009: 930–935.

[15] 程凤伟, 王文剑, 郭虎升. 动态粒度 SVM 学习算法 [J]. 模式识别与人工智能, 2014, 27(4): 372–377.
CHENG Fengwei, WANG Wenjian, GUO Husheng. Dynamic granular support vector machine learning algorithm[J]. Pattern recognition and artificial intelligence, 2014, 27(4): 372–377.

[16] HASSAN R, OTHMAN R M, SHAH Z A. Granular support vector machine to identify unknown structural classes of protein[J]. International journal of data mining and bioinformatics, 2015, 12(4): 451–467.

[17] GUO Husheng, WANG Wenjian. Granular support vector machine: a review[J]. Artificial intelligence review,

2019, 51(1): 19–32.
[18] MA Zhixian, LI Weitian, WANG Lei, et al. X-ray astronomical point sources recognition using granular binary-tree SVM[C]//Proceedings of the 13th International Conference on Signal Processing. Chengdu, China, 2017: 1021–1026.
[19] GUO Husheng, WANG Wenjian. Support vector machine based on hierarchical and dynamical granulation[J]. Neurocomputing, 2016, 211: 22–33.
[20] 郭虎升, 王文剑. 动态粒度支持向量回归机 [J]. 软件学报, 2013, 24(11): 2535–2547.
GUO Husheng, WANG Wenjian. Dynamical granular support vector regression machine[J]. Journal of software, 2013, 24(11): 2535–2547.
[21] YAO Y. Perspectives of granular computing[C]//Proceedings of 2005 IEEE International Conference on Granular Computing. Beijing, China, 2005: 85–90.
[22] TANG Yuchun, JIN Bo, ZHANG Yanqing. Granular support vector machines with association rules mining for protein homology prediction[J]. Artificial intelligence in medicine, 2005, 35(1/2): 121–134.
[23] LI Boyang, WANG Qiangwei, HU Jinglu. A fast SVM training method for very large datasets[C]//Proceedings of 2009 International Joint Conference on Neural Networks. Atlanta, USA, 2009: 1784–1789.
[24] LI Xiaoou, YU Wen. Fast support vector machine classification for large data sets[J]. International journal of computational intelligence systems, 2014, 7(2): 197–212.
[25] LI Xiaoou, CERVANTES J, YU Wen. A novel SVM classification method for large data sets[C]//Proceedings of 2010 IEEE International Conference on Granular Computing. San Jose, USA, 2010: 297–302.

作者简介：

赵帅群，男，1993 年，硕士研究生，主要研究方向为机器学习。

郭虎升，男，1986 年，副教授，博士，主要研究方向为机器学习与数据发掘。主持国家自然科学基金项目 1 项、省部级项目多项。发表学术论文 30 余篇。

王文剑，女，1968 年，教授，博士，主要研究方向为计算智能、机器学习与数据挖掘。主持国家自然科学基金项目 4 项、省部级项目及企事业委托项目 20 余项。发表学术论文 150 余篇。