DOI: 10.11992/tis.201904048
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20190904.1734.002.html

基于模糊核聚类粒化的粒度支持向量机

黄华娟[1], 韦修喜[1], 周永权[1,2]
（1. 广西民族大学 信息科学与工程学院, 广西 南宁 530006; 2. 广西民族大学 广西高校复杂系统与智能计算重点实验室, 广西 南宁 530006）

摘　要: 针对传统的粒度支持向量机 (granular support vector machine, GSVM) 将训练样本在原空间粒化后再映射到核空间, 导致数据与原空间的分布不一致, 从而降低 GSVM 的泛化能力的问题, 本文提出了一种基于模糊核聚类粒化的粒度支持向量机学习算法 (fuzzy kernel cluster granular support vector machine, FKC-GSVM)。FKC-GSVM 通过利用模糊核聚类直接在核空间对数据进行粒的划分和支持向量粒的选取, 在相同的核空间中进行支持向量粒的 GSVM 训练。在 UCI 数据集和 NDC 大数据上的实验表明: 与其他几个算法相比, FKC-GSVM 在更短的时间内获得了精度更高的解。

关键词: 模糊核聚类; 粒化; 支持向量机; 粒度支持向量机; 原空间; 核空间; 支持向量; 聚类
中图分类号: TP18　　**文献标志码:** A　　**文章编号:** 1673−4785(2019)06−1271−07

中文引用格式: 黄华娟, 韦修喜, 周永权. 基于模糊核聚类粒化的粒度支持向量机 [J]. 智能系统学报, 2019, 14(6): 1271–1277.
英文引用格式: HUANG Huajuan, WEI Xiuxi, ZHOU Yongquan. Granular support vector machine based on fuzzy kernel clustering granulation[J]. CAAI transactions on intelligent systems, 2019, 14(6): 1271–1277.

Granular support vector machine based on fuzzy kernel clustering granulation

HUANG Huajuan[1], WEI Xiuxi[1], ZHOU Yongquan[1,2]
(1. College of Information Science and Engineering, Guangxi University for Nationalities, Nanning 530006, China; 2. Guangxi Higher School Key Laboratory of Complex Systems and Intelligent Computing, Guangxi University for Nationalities, Nanning 530006, China)

Abstract: For the traditional granular support vector machine (GSVM), the training samples are granulated in the original space and then mapped to the kernel space. However, this method will lead to the inconsistent distribution of the data between the original space and the kernel space, thereby reducing the generalization of GSVM. To solve this problem, a granular support vector machine based on fuzzy kernel cluster is proposed. Here, the training data are directly granulated, and support vector particles are selected in kernel space. The support vector particles are then trained in the same kernel space by the GSVM. Finally, experiments on UCI data sets and NDC big data sets show that FKC-GSVM achieves more accurate solutions in a shorter time than other algorithms.

Keywords: fuzzy kernel cluster; granulation; support vector machine; granular support vector machine; original space; kernel space; support vector; clustering

收稿日期: 2019−04−18.　**网络出版日期:** 2019−09−05.
基金项目: 国家自然科学基金资助项目 (61662005); 广西自然科学基金项目 (2018JJA170121); 广西高校中青年教师科研基础能力提升项目 (2019KY0195).
通信作者: 韦修喜. E-mail: weixiuxi@163.com.

支持向量机 (support vector machine, SVM) 自 1995 年由 Vapnik 提出以来就受到理论研究和工程应用 2 方面的重视, 是机器学习的一个研究

方向和热点，已经成功应用到很多领域中[1-3]。SVM 的基本算法是一个含有不等式约束条件的二次规划 (quadratic programming problem, QPP) 问题，然而，如果直接求解 QPP 问题，当数据集较大时，算法的效率将会下降，所需内存量也会增大[4-8]。因此，如何克服 SVM 在处理大规模数据集时的效率低下问题，一直是学者们研究的热点。

为了更好地解决大规模样本的分类问题，基于粒度计算理论[9-10] 和统计学习理论的思想，Tang 等于 2004 年首次提出粒度支持向量机 (granular support vector machine, GSVM) 这个术语。GSVM 的总体思想是在原始空间将数据集进行划分，得到数据粒。然后提取出有用的数据粒，并对其进行 SVM 训练[11-12]。与传统支持向量机相比，GSVM 学习机制具有以下优点：针对大样本数据，通过数据粒化和对有用粒子 (支持向量粒) 的提取，剔除了无用冗余的样本，减少了样本数量，提高了训练效率。然而，Tang 只是给出了 GSVM 学习模型的一些设想，没有给出具体的学习算法。2009 年，张鑫[13] 在 Tang 提出的 GSVM 思想的基础上，构建了一个粒度支持向量机的模型，并对其学习机制进行了探讨。此后，许多学者对支持向量机和粒度计算相结合的具体模型进行了研究，比如模糊支持向量机[13]、粗糙集支持向量机[14]、决策树支持向量机[15]和商空间支持向量机[16] 等。但这些模型的共同点都是在原始空间直接划分数据集，然后再映射到高维空间进行 SVM 学习。然而，这种做法很有可能丢失了大量包含有用信息的数据粒，其学习算法的性能会受到影响。为此，本文采用模糊核聚类的方法将样本直接在核空间进行粒的划分和提取，然后在相同的核空间进行 GSVM 训练，这样保证了数据分布的一致性，提高了算法的泛化能力。最后，在标准 UCI 数据集和 NDC大数据上的实验结果表明，本文算法是可行的且效果更好。

1　粒度支持向量机

张鑫[17] 在 Tang 提出的 GSVM 思想的基础上，构建了一个粒度支持向量机的模型。

设给定数据集为 $\boldsymbol{X}=\{(\boldsymbol{x}_i,y_i),i=1,2,\cdots,n\}$，$n$ 为样本的个数；y_i 为 $\boldsymbol{x}_i$ 所属类的标签。采用粒度划分的方法 (聚类、粗糙集、关联规则等) 划分 $\boldsymbol{X}$，若数据集有 l 个类，则将 $\boldsymbol{X}$ 分成 l 个粒，表示为：

$$(\boldsymbol{X}_1,\boldsymbol{Y}_1),(\boldsymbol{X}_2,\boldsymbol{Y}_2),\cdots,(\boldsymbol{X}_i,\boldsymbol{Y}_i),\cdots,(\boldsymbol{X}_l,\boldsymbol{Y}_l)$$

若每个粒包含 l_i 个点，Y_i 表示第 i 个粒的类别，则有：

$$\boldsymbol{X}=\{(\boldsymbol{X}_i,\boldsymbol{Y}_i),i=1,2,\cdots,l\}$$

其中：

$$\boldsymbol{X}_1=\begin{Bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\\ \vdots\\ \boldsymbol{x}_{l_1}\end{Bmatrix},\boldsymbol{X}_2=\begin{Bmatrix}\boldsymbol{x}_{l_1+1}\\ \boldsymbol{x}_{l_1+2}\\ \vdots\\ \boldsymbol{x}_{l_1+l_2}\end{Bmatrix},\cdots,\boldsymbol{X}_l=\begin{Bmatrix}\boldsymbol{x}_{l_1+l_2+\cdots+l_{l-1}+1}\\ \boldsymbol{x}_{l_1+l_2+\cdots+l_{l-1}+2}\\ \vdots\\ \boldsymbol{x}_{l_1+l_2+\cdots+l_{l-1}+l_l}\end{Bmatrix}$$

则在 GSVM 中，最优化问题变为：

$$\begin{aligned}&\min_{w,b}\ \frac{1}{2}\boldsymbol{w}^2\\ &\text{s.t.}\quad \boldsymbol{Y}_j((\boldsymbol{w}\cdot\boldsymbol{X}_j)+b)\geqslant 1,\quad j=1,2,\cdots,l\end{aligned}\tag{1}$$

将上述问题根据最优化理论转化为其对偶问题：

$$\begin{aligned}&\max_a \boldsymbol{W}(a)=\max_a -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}a_ia_j\boldsymbol{Y}_i\boldsymbol{Y}_j(\boldsymbol{X}_i\boldsymbol{X}_j)+\sum_{i=1}^{l}a_i\\ &\text{s.t.}\quad \sum_{i=1}^{l}a_i\boldsymbol{Y}_i=0,\quad 0\leqslant a_i\leqslant C\end{aligned}\tag{2}$$

解得最优解 $\boldsymbol{a}^*=[\boldsymbol{a}_1{}^*\ \boldsymbol{a}_2{}^*\cdots\boldsymbol{a}_l{}^*]^{\mathrm{T}}$，计算最优权值向量 $\boldsymbol{w}^*=\sum_{j=1}^{l}a_j{}^*y_j\boldsymbol{X}_j$ 和最优偏置 $b^*=y_i-\sum_{j=1}^{l}y_ja_j^*(\boldsymbol{X}_j\boldsymbol{X}_i)$，$i\in\{i|a_i^*>0\}$。因此得到最优分类超平面 $(\boldsymbol{w}^*\boldsymbol{X})+b^*=0$，而最优分类函数为：

$$\begin{aligned}f(\boldsymbol{x})=&\operatorname{sgn}\{(\boldsymbol{w}^*\boldsymbol{X})+b^*\}=\\ &\operatorname{sgn}\{(\sum_{j=1}^{l}a_j^*y_j(\boldsymbol{X}_j\boldsymbol{X}_i))+b^*\},\boldsymbol{X}\in\mathbf{R}^n\end{aligned}\tag{3}$$

当数据集是线性不可分时，GSVM 不是在原始空间构造最优分类面，而是映射到高维特征空间，然后再进行构造，具体为：

将 $\boldsymbol{X}$ 从 $\mathbf{R}^n$ 变换到 $\boldsymbol{\Phi}$：

$$\boldsymbol{X}\rightarrow\boldsymbol{\Phi}(\boldsymbol{X})=[\boldsymbol{\Phi}_1(\boldsymbol{X})\,\boldsymbol{\Phi}_2(\boldsymbol{X})\cdots\boldsymbol{\Phi}_l(\boldsymbol{X})]^{\mathrm{T}}$$

以特征向量 $\boldsymbol{\Phi}(\boldsymbol{X})$ 代替输入向量 $\boldsymbol{X}$，则可以得到最优分类函数为：

$$\begin{aligned}f(\boldsymbol{X})=&\operatorname{sgn}(\boldsymbol{w}\boldsymbol{\Phi}(\boldsymbol{X})+b)=\\ &\operatorname{sgn}(\sum_{i=1}^{l}a_iy_i\boldsymbol{\Phi}(\boldsymbol{X}_i)\boldsymbol{\Phi}(\boldsymbol{X})+b)\end{aligned}\tag{4}$$

利用核函数来求解向量的内积，则最优分类函数变为：

$$\begin{aligned}f(\boldsymbol{X})=&\operatorname{sgn}(\boldsymbol{w}\boldsymbol{\Phi}(\boldsymbol{X})+b)=\\ &\operatorname{sgn}(\sum_{i=1}^{l}a_iy_ik(\boldsymbol{X}_i,\boldsymbol{X})+b)\end{aligned}\tag{5}$$

其中，$k(\boldsymbol{X}_i,\boldsymbol{X})$ 为粒度核函数。当 $a_i>0$，根据以上分析，可知 $\boldsymbol{X}_i$ 是支持向量。显然地，式 (5) 的形式和 SVM 的最优分类函数很一致，确保了最优解的唯一性。

2　基于模糊核聚类粒化的粒度支持向量机

2.1　问题的提出

在研究中发现，只有支持向量才对 SVM 的训

练起积极作用, 它们是非常重要的, 对于 SVM 是不可或缺的, 而其余的非支持向量对于分类超平面是不起作用的, 甚至可能产生负面影响, 比如增加了核矩阵的容量, 降低了 SVM 的效率。GSVM 也存在同样的问题, 只有支持向量粒才对 GSVM 的训练起决定性作用。可以通过理论证明来说明这个观点的正确性。

定理 1　粒度支持向量机的训练过程和训练结果与非支持向量粒无关。

证明　定义 $I_{\mathrm{sv}}=\{i|a_i>0\}$ 和 $I_{\mathrm{nsv}}=\{i|a_i=0\}$ 分别为支持向量粒和非支持向量粒对应样本序号的索引集, 支持向量粒的个数记为 l'。引入只优化支持向量粒对应样本的问题

$$\begin{aligned}&\max_{\boldsymbol{a}} g(\boldsymbol{a})=\max_{\boldsymbol{a}}-\frac{1}{2}\sum_{i=1}^{l'}\sum_{j=1}^{l'}a_ia_j\boldsymbol{Y}_i\boldsymbol{Y}_j(\boldsymbol{X}_i\boldsymbol{X}_j)+\sum_{i=1}^{l'}a_i\\&\text{s.t.}\quad\sum_{i=1}^{l}a_i\boldsymbol{Y}_i=0,\ 0\leqslant a_i\leqslant C,\ i=1,2,\cdots,l'\in \mathrm{I}_{\mathrm{sv}}\end{aligned}\tag{6}$$

要证明定理 1, 只需要证明式 (2) 和式 (6) 同解。用反正法, 假设式 (6) 存在一个最优解 $\boldsymbol{a}'$ 使得 $g(\boldsymbol{a}')>g(\boldsymbol{a}^*)$。由于 $\boldsymbol{a}^*$ 是式 (2) 的最优解, 也即 $\boldsymbol{a}^*$ 是式 (6) 的可行解, 同样, $\boldsymbol{a}'$ 也是式 (2) 的可行解。由于 $\boldsymbol{a}^*$ 是式 (2) 的最优解, 可得 $w(\boldsymbol{a}^*)>w(\boldsymbol{a}')$。又因为

$$w(\boldsymbol{a}')=\max_{\boldsymbol{a}}-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}a_i'a_j'y_iy_jk(\boldsymbol{X}_i,\boldsymbol{X}_j)+\sum_{i=1}^{l}\boldsymbol{a}'=$$

$$\max_{\boldsymbol{a}}-\frac{1}{2}\sum_{i=1}^{l'}\sum_{j=1}^{l'}a_i'a_j'y_iy_jk(\boldsymbol{X}_i,\boldsymbol{X}_j)+\sum_{i=1}^{l'}\boldsymbol{a}'=g(\boldsymbol{a}')$$

$$w(\boldsymbol{a}^*)=\max_{\boldsymbol{a}}-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}a_i^*a_j^*y_iy_jk(\boldsymbol{X}_i,\boldsymbol{X}_j)+\sum_{i=1}^{l}\boldsymbol{a}^*=$$

$$\max_{\boldsymbol{a}}-\frac{1}{2}\sum_{i=1}^{l'}\sum_{j=1}^{l'}a_i^*a_j^*y_iy_jk(X_i,X_j)+\sum_{i=1}^{l'}\boldsymbol{a}^*=g(\boldsymbol{a}^*)$$

可得 $w(\boldsymbol{a}')=g(\boldsymbol{a}')>g(\boldsymbol{a}^*)=w(\boldsymbol{a}^*)$, 即 $w(\boldsymbol{a}')>w(\boldsymbol{a}^*)$, 这与 $\boldsymbol{a}^*$ 是式 (2) 的最优解得出的 $w(\boldsymbol{a}^*)>w(\boldsymbol{a}')$ 相矛盾。定理 1 得证。注: $\boldsymbol{a}'$ 是 l' 维向量, 代入 w 的时候拓展为 l 维向量。

要想迅速地得到支持向量粒, 节省粒化的时间, 首先了解支持向量的特征, 文献 [13] 对其特征做了归纳总结。

1) 现实中, 支持向量一般都是稀疏地聚集在训练数据集的边缘。

2) 根据第一个特征, 则每个类中心附近的数据不会是支持向量, 即, 离支持向量机超平面较近的数据比较可能是支持向量, 这就为支持向量的选取提供了快速的获取方法。

2.2　问题分析

图 1 中, 红色部分的数据是 GSVM 的支持向量粒, 它们决定了分类超平面。并且从中可以看出, 对于每一类, 离类中心越远的点, 就越有可能是支持向量粒。并且, 从图 1 中还可以看出, 落在每一个环上的样本, 它们离类中心的距离差不多相等。离类中心越远的环就越有可能含有多的支持向量粒。基于这个思想, 本文先把样本映射到核空间, 按类标签的个数进行粗粒划分, 确保相同标签的样本都在同一个粗粒中。然后, 对于每一个粗粒, 采用模糊聚类的方法进行粒化, 具有相同隶属度的样本归为一个粒, 进行细粒划分。每一个细粒就对应图 1 中的一个环, 从图中可以看出, 离粗粒中心越远的环, 越靠近分类超平面, 其是支持向量粒的可能性越大。而离粗粒中心越远的环, 其隶属度越小。因此, 给定一个阈值, 当细粒的隶属度小于给定的阈值, 就说明其处于粗粒的边缘, 是支持向量粒, 进而提取出支持向量粒.

图 1　支持向量分布图

Fig. 1　The distribution of support vectors

2.3　模糊核聚类

模糊核聚类 (fuzzy kernel cluster, FKC) 的主要思想是先将数据集映射到高维空间, 然后直接在高维空间进行模糊聚类。而一般的聚类算法是直接在原始空间进行聚类划分。与其他的聚类算法相比, 模糊核聚类引入了非线性映射, 能够在更大程度上提取到有用的特征, 聚类的效果会更好。

设原空间样本集为 $\boldsymbol{X}=(x_1,x_2,\cdots,x_N)$, $x_j\in\mathbf{R}^d$, $j=1,2,\cdots,N$。核非线性映射为 $\boldsymbol{\Phi}: x\to\boldsymbol{\Phi}(x)$, 在本文中, 采用 Euclid 距离作为距离测量方法, 由此得到模糊核 C-均值聚类:

$$\begin{aligned}J_m=&(\boldsymbol{X},\boldsymbol{U},\boldsymbol{V})=\\&\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^m\left\|\boldsymbol{\Phi}(x_j)-\boldsymbol{\Phi}(v_i)\right\|^2=\\&\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^m[k(x_j,x_j)-2k(x_j,v_i)+k(v_i,v_i)]=\\&\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^md_{Kij}^2(x_j,v_i),\ 2\leqslant C<N\end{aligned}\tag{7}$$

式中: C 是事先确定的簇数; $m\in(1,\infty)$ 是模糊加权指数, 对聚类的模糊程度有重要的调节作用; v_i 为第 i 类的类中心; $\phi(v_i)$ 为该中心在相应核空间中的像。

按模糊 C-均值优化方法, 隶属度设计为

$$u_{ij}=\frac{[1/d_{Kij}^2(x_j,v_i)]^{1/(m-1)}}{\sum_{j=1}^{C}[1/d_{Kij}^2(x_j,v_i)]^{1/(m-1)}} \tag{8}$$

且有

$$\boldsymbol{\Phi}(v_i)=\frac{\sum_{k=1}^{N}u_{ik}^m\boldsymbol{\Phi}(x_k)}{\sum_{k=1}^{N}u_{ik}^m},\quad i=1,2,\cdots,C \tag{9}$$

为了最小化目标函数，需要计算 $k(x_j,v_i)$ 和 $k(v_i,v_i)$，由 $k(x_i,x_j)=<\phi(x_i),\phi(x_j)>$ 可得：

$$k(x_j,v_i)=<\boldsymbol{\Phi}(x_j),\boldsymbol{\Phi}(v_i)>=\frac{\sum_{k=1}^{N}u_{ik}^m k(x_k,x_j)}{\sum_{k=1}^{N}u_{ik}^m} \tag{10}$$

$$k(v_i,v_i)=<\boldsymbol{\Phi}(v_i),\boldsymbol{\Phi}(v_i)>=\frac{\sum_{k=1}^{N}\sum_{s=1}^{N}u_{ik}^m u_{is}^m k(x_k,x_s)}{\left(\sum_{k=1}^{N}u_{ik}^m\right)^2} \tag{11}$$

把式 (9)~(11) 代入式 (7)，可以求出模糊核 C-均值聚类的目标函数值。

模糊核 C-均值聚类的算法步骤如下：

1) 初始化参数：ε、m、T 和 C；

2) 对训练数据集进行预处理；

3) 设置 $v_i(i=1,2,\cdots,C)$ 的初始值；

4) 计算隶属度 $u_{ij}(i=1,2,\cdots,C;j=1,2,\cdots,N)$；

5) 计算新的 $k(x_j,v_i)$ 和 $k(v_i,v_i)$，更新隶属度 u_{ij} 为 $\hat{u}_{ij}$；

6) 若 $\max_{j,i}\left|u_{ij}-\hat{u}_{ij}\right|<\varepsilon$ 或迭代次数等于预定迭代次数 T 则算法停止，否则转到 4)。

2.4　FKC-GSVM 的算法步骤

目前，已有的粒度支持向量机算法模型大都是直接在原始空间对数据集进行粒化和提取，然后映射到核空间进行 SVM 的训练。然而，不同空间的转换，很有可能丢失了数据集的有用信息，降低学习器的性能。为了避免因数据在不同空间分布不一致而导致泛化能力不高的问题，本文采用模糊核聚类的方法直接在核空间对数据集进行粒化、提取和 SVM 的训练。基于以上思想，本文提出了基于模糊核聚类粒化的粒度支持向量机 (fuzzy kernel cluster granular support vector machine, FKC-GSVM)。FKC-GSVM 算法包括 3 部分：采用模糊核聚类进行粒度的划分；设定阈值，当每个粒子的隶属度大于规定的阈值时，认为这个粒子为非支持向量粒，丢弃，而剩余的粒子为支持向量粒；在核空间对支持向量粒进行 SVM 训练。具体的算法步骤如下：

1) 粗粒划分：以类标签个数 l 为粒子个数，对训练样本进行粗粒的划分，得到 l 个粒子；

2) 细粒划分：采用模糊核聚类分别对 l 个粒子进行细粒划分，计算每个粒子的隶属度；

3) 支持向量粒的提取：给定一个阈值，当一个粒子的隶属度小于给定的阈值，提取这个粒子 (支持向量粒)，提取出来的支持向量粒组成了一个新的训练集；

4) 支持向量集的训练：在新的训练样本集上进行 GSVM 训练；

5) 泛化能力的测试：利用测试集测试泛化能力。

2.5　FKC-GSVM 算法性能分析

下面，从 2 个方面对 FKC-GSVM 的算法性能进行分析：

1) FKC-GSVM 的收敛性分析

与 SVM 相比，FKC-GSVM 采用核空间代替原始空间进行粒化，提取出支持向量粒后在相同的核空间进行 GSVM 训练，其训练的核心思想依然是采用支持向量来构造分类超平面，这与标准支持向量机相同。既然标准支持向量机是收敛的，则 FKC-GSVM 也是收敛的。但是由于 FKC-GSVM 剔除了大量对训练不起积极作用的非支持向量，直接采用支持向量来训练，所以它的收敛速度要快于标准支持向量机。

2) FKC-GSVM 的泛化能力分析

评价一个学习器性能好坏的重要指标是其是否具有较强的泛化能力。众所周知，由于 SVM 采用结构风险最小 (SRM) 归纳原则，因此，与其他学习机器相比，SVM 的泛化能力是很突出的。同样，FKC-GSVM 也执行了 SRM 归纳原则，并且直接在核空间选取支持向量，确保了数据的一致性，具有更好的泛化性能。

3　实验结果及分析

3.1　UCI 数据集上的实验

为了验证 FKC-GSVM 的学习性能，本文在 Matlab7.11 的环境下对 5 个常用的 UCI 数据集进行实验，这 5 个数据集的描述如表 1 所示。在实验中，采用的核函数为高斯核函数，并且采用交叉验证方法选取惩罚参数 C 和核参数 σ，聚类数 c 设为 20。影响算法表现的主要因素是阈值 k 的设定，为此，对不同的阈值对算法的影响进行了分析。

表 1　实验采用的数据集

Table 1　Datasets used in experiments

数据集	Abalone	Contraceptive Method Choice	Pen-Based Recognition of Hand-written Digits	NDC-10k	NDC-1l
#训练集	3 177	1 000	6 280	10 000	100 000
#测试集	1 000	473	3 498	1 000	10 000
维度	8	9	16	32	32

为了比较数据集在原空间粒化和在核空间粒化的不同效果, 本文采用基于模糊聚类的粒度支持向量机 (FCM-GSVM)、基于模糊核聚类的粒度支持向量机 (FKC-GSVM) 和粒度支持向量机 (GSVM) 等 3 种算法对 5 个典型的 UCI 数据集进行了测试, 测试结果如表 2 所示。为了更直观地看出 FKC-GSVM 在不同阈值条件下的分类效果, 给出了 Contraceptive Method Choice 数据集在不同阈值条件下采用 FKC-GSVM 分类的效果图, 如图 2~图 5 所示。

表 2　FCM-GSVM 与 FKC-GSVM 测试结果比较

Table 2　Comparison of test results between FCM-GSVM and FKC-GSVM　%

数据集	算法	阈值 k			
		0.9	0.85	0.8	0.75
Abalone	GSVM	80.1	75.6	76.7	65.6
	FCM-GSVM	82.9	79.1	75.9	68.8
	FKC-GSVM	**94.8**	**85.4**	**79.2**	**75.9**
Contraceptive Method Choice	GSVM	82.1	76.4	78.4	64.9
	FCM-GSVM	85.4	82.7	79.3	69.8
	FKC-GSVM	**91.6**	**87.2**	**85.6**	**81.5**
Pen-Based Recognition of Handwritten Digits	GSVM	81.2	77.1	70.2	63.8
	FCM-GSVM	84.6	80.6	75.6	70.3
	FKC-GSVM	**90.7**	**83.3**	**80.4**	**72.9**
NDC-10k	GSVM	85.2	82.1	79.7	69.3
	FCM-GSVM	86.5	83.8	81.4	79.6
	FKC-GSVM	**89.6**	**86.1**	**84.6**	**82.3**
NDC-1l	GSVM	80.2	72.4	73.5	66.6
	FCM-GSVM	83.7	79.3	76.9	72.9
	FKC-GSVM	**86.5**	**83.8**	**82.3**	**79.2**

FCM-GSVM 和 GSVM 是在原空间进行粒度划分和支持向量粒的提取, 然后把支持向量粒映射到高维空间进行分类, 而 FKC-GSVM 是直接在核空间进行粒度划分和支持向量粒的提取, 然后在相同的核空间进行分类。从表 2 的测试结果可以看出, 由于 FCM-GSVM 和 GSVM 可能导致数据在原空间和核空间分布不一致, 在相同的阈值条件下, 其分类效果要比 FKC-GSVM 的分类效果差, 这说明 FKC-GSVM 的泛化能力比 FCM-GSVM 的泛化能力强。

为了分析在不同阈值条件下 FKC-GSVM 的泛化性能, 本文给出了 0.9、0.85、0.8、0.75 四个不同阈值条件下的实验。从表 2 可以看出, 阈值越小, FKC-GSVM 的分类效果越差, 这是因为阈值越小, 选取的支持向量粒就越少, 这一过程可能丢失了一些支持向量, 影响了分类效果。但是阈值越小, 大大压缩了训练样本集, 算法训练的速度得到了很大的提高。因此, 对于大规模样本来说, 只要在能接受的分类效果的范围内, 选取合适的阈值, 采用 FKC-GSVM 就能快速地得到需要的分类效果。图 2~5 是 Contraceptive Method Choice 数据集在不同阈值条件下采用 FKC-GSVM 分类的效果图, 从这几个图中可以很直观地看出, FKC-GSVM 的分类效果还是比较令人满意的。

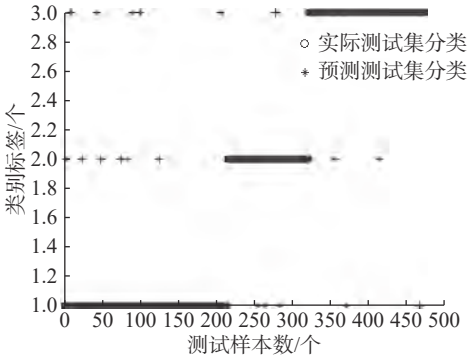

图 2　FKC-GSVM 在阈值为 0.9 条件下的分类效果

Fig. 2　The classification results of FKC-GSVM under the threshold 0.9

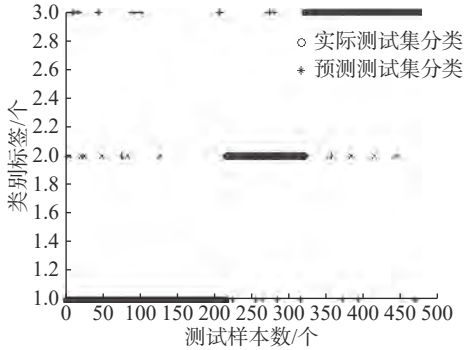

图 3　FKC-GSVM 在阈值为 0.85 条件下的分类效果

Fig. 3　The classification results of FKC-GSVM under the threshold 0.85

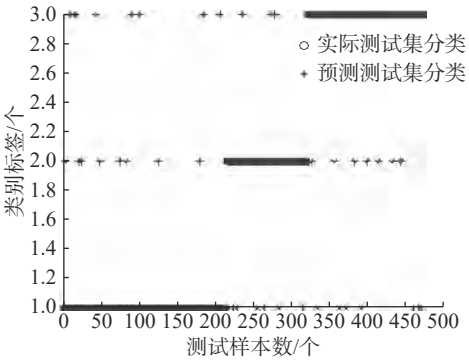

图 4　FKC-GSVM 在阈值为 0.8 条件下的分类效果

Fig. 4　The classification results of FKC-GSVM under the threshold 0.8

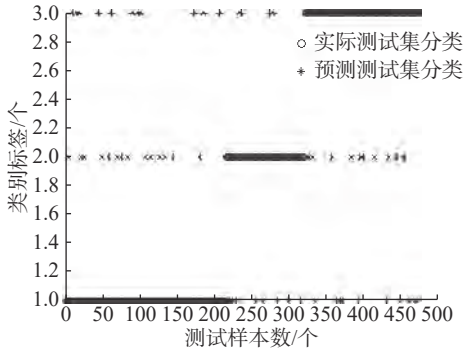

图 5　FKC-GSVM 在阈值为 0.75 条件下的分类效果

Fig. 5　The classification results of FKC-GSVM under the threshold 0.75

3.2　NDC 大数据集上的实验

为了测试 FKC-GSVM 处理大数据集的性能，在实验中，采用的数据集是 NDC 大数据集 [20]，是由 David Musicant's NDC 数据产生器产生的，NDC 数据集的描述如表 3 所示。在实验中，FKC-GSVM 的测试结果将与现在比较流行的孪生支持向量机 (twin support vector machines, TWSVM) 的测试结果[20] 从测试精度和运行时间 2 方面进行对比。其中，FKC-GSVM 的运行环境、参数设置方法和实验 3.1 一样，阈值 $k=0.9$；TWSVM 的惩罚参数和核参数的选取都是从 $\{2^{-8},2^{-7},\cdots,2^{7}\}$ 这个范围内采用网格搜索算法进行选择。表 4 显示的是 FKC-GSVM 和 TWSVM 两种算法的运行结果。

表 3　实验采用的 NDC 数据集

Table 3　NDC datasets used in experiments

数据集	训练集	测试集	维度
NDC-3L	300 000	30 000	32
NDC-5L	500 000	50 000	32
NDC-1M	1 000 000	100 000	32
NDC1	5 000	5 000	100
NDC2	5 000	5 000	1 000

表 4　2 种算法对 NDC 数据集的测试结果

Table 4　Comparison of two algorithms on NDC datasets

数据集	FKC-GSVM			TWSVM		
	训练正确率/%	测试正确率/%	CPU时间/s	训练正确率/%	测试正确率/%	CPU时间/s
NDC-3l	82.12	78.56	2.89	79.54	78.76	23.10
NDC-5l	79.65	78.08	5.131	78.84	77.12	90.35
NDC-1m	82.12	80.27	50.734	–	–	–
NDC1	89.68	86.31	1.152	86.72	84,52	19.574
NDC2	90.67	86.95	60.364	–	–	–

"–"表示训练时间过高，实验无法进行

从表 3 中可以看出，本实验测试的对象为 5 种数据集，NDC-3L 的训练样本数为 300 000 个，而 NDC-1m 的样本增加到了 1 000 000 个，同样，测试样本也从 30 000 增加到了 100 000，特征数都是 32 维。这 3 个数据集主要是为了测试算法在处理维度一样而数据量不断增加时候的学习性能。为了进一步测试学习算法处理高维样本的性能，NDC1 和 NDC2 这 2 个数据集的维数分别是 100 和 1 000，设置他们的训练样本量和测试样本量都一样，都是 5 000。

实验结果如表 4 所示，从中可以看出，当数据集为 NDC-1m 时，由于训练时间过高，采用 TWSVM 算法无法将实验进行下去。然而，FKC-GSVM 在处理 NDC-1m 数据集时能够在合理的运行时间内得到较满意的精度解，这表明了 FKC-GSVM 在处理大数据时是具有优势的。同样，在处理 NDC1 和 NDC2 这 2 个高维数据集时，从表 4 可以明显看出，FKC-GSVM 处理高维数据的效果也是不错的。实验结果充分说明了 FKC-GSVM 的学习能力比 TWSVM 的强，更适合于处理大数据集。

4　结束语

GSVM 是将训练样本在原空间粒化后再映射到核空间，这将导致数据与原空间的分布不一致，从而降低了 GSVM 的泛化能力。为了解决这个问题，本文提出了一种基于模糊核聚类粒化的粒度支持向量机方法 (FKC-GSVM)。FKC-GSVM 通过利用模糊核聚类直接在核空间对数据进

行粒的划分和支持向量粒的选取, 然后在相同的核空间中进行支持向量粒的 GSVM 训练, 在标准数据集的实验说明了 FKC-GSVM 算法的有效性。但是阈值参数的选取仍具有一定的随意性, 影响了 FKC-GSVM 的性能。如何自适应地调整合适的阈值, 将是下一步要研究的工作内容。

参考文献:

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

[2] 丁世飞, 张健, 张谢锴, 等. 多分类孪生支持向量机研究进展 [J]. 软件学报, 2018, 29(1): 89–108.
DING Shifei, ZHANG Jian, ZHANG Xiekai, et al. Survey on multi class twin support vector machines[J]. Journal of software, 2018, 29(1): 89–108.

[3] AN Yuexuan, DING Shifei, SHI Songhui, et al. Discrete space reinforcement learning algorithm based on support vector machine classification[J]. Pattern recognition letters, 2018, 111: 30–35.

[4] 谢娟英, 谢维信. 基于特征子集区分度与支持向量机的特征选择算法 [J]. 计算机学报, 2014, 37(8): 1704–1718.
XIE Juanying, XIE Weixin. Several feature selection algorithms based on the discernibility of a feature subset and support vector machines[J]. Chinese journal of computers, 2014, 37(8): 1704–1718.

[5] YAO Y Y. Granular computing: basic issues and possible solution[C]//Proceedings of the 5th Joint Conference on Information Sciences. Atlantic City, USA, 2000: 186–189.

[6] DING Shifei, XU Li, ZHU Hong, et al. Research and progress of cluster algorithms based on granular computing[J]. International journal of digital content technology and its applications, 2010, 4(5): 96–104.

[7] 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, CA, USA, 2004: 73–78.

[8] 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.

[9] 冯昌, 廖士中. 随机傅里叶特征空间中高斯核支持向量机模型选择 [J]. 计算机研究与发展, 2016, 53(9): 1971–1978.
FENG Chang, LIAO Shizhong. Model selection for Gaussian kernel support vector machines in random Fourier feature space[J]. Journal of computer research and development, 2016, 53(9): 1971–1978.

[10] 段丹青, 陈松乔, 杨卫军, 等. 使用粗糙集和支持向量机检测入侵 [J]. 小型微型计算机系统, 2008, 29(4): 627–630.
DUAN Danqing, CHEN Songqiao, YANG Weiping, et al. Detect intrusion using rough set and support vector machine[J]. Journal of Chinese computer systems, 2008, 29(4): 627–630.

[11] 李涛, 刘学臣, 张帅, 等. 基于混合编程模型的支持向量机训练并行化 [J]. 计算机研究与发展, 2015, 52(5): 1098–1108.
LI Tao, LIU Xuechen, ZHANG Shuai, et al. Parallel support vector machine training with hybrid programming model[J]. Journal of computer research and development, 2015, 52(5): 1098–1108.

[12] 丁世飞, 黄华娟. 最小二乘孪生参数化不敏感支持向量回归机 [J]. 软件学报, 2017, 28(12): 3146–3155.
DING Shifei, HUANG Huajuan. Least squares twin parametric insensitive support vector regression[J]. Journal of software, 2017, 28(12): 3146–3155.

[13] 张鑫. 粒度支持向量机学习方法研究 [D]. 太原: 山西大学, 2009.
ZHANG Xin. Research on granular support vector machine learning method[D]. Taiyuan: Shangxi University, 2009.

[14] DING Shifei, AN Yuexuan, ZHANG Xiekai, et al. Wavelet twin support vector machines based on glowworm swarm optimization[J]. Neurocomputing, 2017, 225: 157–163.

[15] KUMAR M A, GOPAL M. Application of smoothing technique on twin support vector machines[J]. Pattern recognition letters, 2008, 29(13): 1842–1848.

[16] 黄华娟. 孪生支持向量机关键问题的研究 [D]. 徐州: 中国矿业大学, 2014.
HUANG Huajuan. Research on the key problems of twin support vector machines[D]. Xuzhou: China University of Mining and Technology, 2014.

[17] 郭虎升, 王文剑, 张鑫. 基于粒度核的支持向量机学习算法 [C]// 第三届中国粒计算联合会议. 河北, 石家庄, 2009.95–97: 155.

作者简介:

黄华娟, 女, 1984 生, 副教授, 博士, 主要研究方向为机器学习与数据挖掘。主持国家自然科学基金项目、广西自然科学基金项目各 1 项。发表学术论文 20 余篇。

韦修喜, 男, 1980 生, 讲师, 主要研究方向为人工智能。主持广西高校中青年教师科研基础能力提升项目 1 项。发表学术论文 10 余篇。

周永权, 男, 1962 年生, 教授, 博士, 主要研究方向为计算智能。主持国家自然科学基金项目 3 项。发表学术论文 100 余篇。