基于信息粒化的 SVM 时序回归预测①

彭 勇[1], 陈俞强[1,2]

[1](东莞职业技术学院 计算机工程系, 东莞 523808)
[2](广东工业大学 自动化学院, 广州 510006)

摘 要: 为了提高 SVM 的学习效率和泛化能力, 首先利用一种信息粒化算法对原始数据进行预处理, 该算法能将样本空间划分为多个粒(子空间), 降低样本规模, 节省时间复杂度. 然后将模糊粒化后的信息利用 SVM 进行回归分析, 同时利用交叉验证选出最优的分类器调节参数, 可降低分类器的复杂性和提高分类器的泛化能力, 避免出现过学习和欠学习. 最后通过预测上证指数的实验验证了该算法具有优越的特性, 能够较为准确的进行时序回归预测.

关键词: 信息粒化; 支持向量机; 泛化能力; 回归预测

Time Series Regression and Prediction Based on Information Granulation and SVM

PENG Yong[1], CHEN Yu-Qiang[1,2]

[1](Department of Computer Engineering, Dongguan Polytechnic, Dongguan 523808, China)
[2](Faculty of Automation, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: In order to improve learning efficiency and generalization ability of SVM, firstly, the raw data is preprocessed by using an information granulation algorithm. This algorithm can divide sample space into multiple particles (subspace), reduce the sample size and save the time complexity. And then, the granulated information take SVM to carry on the regression analysis, while take cross validation to select the optimal classifier adjustable parameters, which can reduce the complexity of the classifier and improve the generalization capability of the classifier and avoid Over learning and less learning. Finally, the test results on forecasting the Shanghai Composite Index have proved that the system has a good-performance and make time series regression prediction precisely.

Key words: information granulation; SVM; generalization ability; regression prediction

支持向量机 (SVM)是由 Vapnik 于 20 世纪 90 年代初提出的一种新的机器学习方 法. SVM 算法利用凸二次规划(quadratic programming, QP)、Mercer 核、稀疏解和松弛向量等多项技术, 具有结构简单、训练误差低和泛化能力好等优点, 因而被广泛地应用于分类学习问题中. 但是当样本规模较大时, SVM 算法在二次寻优过程中需要存储核矩阵和进行大量矩阵运算, 使得运算时间过长.

信息粒化(IG, Information Granulation)最早有 LotfiA. Zadeh 教授提出的, 就是将一个整体分解成一个个的部分进行研究, 每一部分为一个信息粒[1]. Zadeh 教授指出: 信息粒就是一些元素的集合, 这些元素由于难以识别、或相似、或接近或某种功能结合在一起. 在这样的理论背景下, 将信息粒化思想引入 SVM 中, 不仅可以丰富 SVM 理论, 而且可以推广 SVM 的应用领域, 本文主要是通过常用的粒划分方法构建粒空间获得一系列信息粒, 然后在每个信息粒上进行学习, 最后通过聚合信息粒上的信息获得最终的 SVM 决策函数.

1 模糊信息粒化方法模型

模糊信息粒就是以模糊集形式表示的信息粒. 用模糊集方法对时间序列进行模糊粒化, 主要分为两个

① 基金项目:广东省自然科学基金(0002014014)
收稿时间:2012-10-10;收到修改稿时间:2012-11-11

© 中国科学院软件研究所 http://www.c-s-a.org.cn

步骤: 划分窗口和模糊化. 划分窗口就是将时间序列分割成若干小子序列, 作为操作窗口; 模糊化则是将产生的每一个窗口进行模糊化, 生成一个个模糊集, 也就是模糊信息粒. 这两种广义模式结合在一起就是模糊信息粒化, 称为 f—粒化. 在 f—粒化中, 最为关键的是模糊化的过程, 也就是在所给的窗口上建立一个合理地模糊集, 使其能够取代原来窗口中的数据, 表示相关的人们所关心的信息[4].

将模糊信息粒化方法运用到原始数据预处理中, 在原空间上进行粒度划分, 进而在各个粒上进行 SV 训练, 可以将一个线性不可分问题转化为一系列线性可分问题, 从而获得多个决策函数; 同时, 这一学习机制也使得数据的泛化性能增强, 即可在 SVM 的训练中得到间隔更宽的超平面. 也就是说将一个大规模 QP 问题, 通过粒度划分, 分解为一系列小的 QP 问题; 同时, 最后实现对原问题的求解.

对于给定的时间序列, 考虑单窗口问题, 即把整个时序 X 看成是一个窗口进行模糊化, 模糊化的任务是在 X 上建立一个模糊粒子 P, 既一个能够合理描述 X 的模糊概念 G(以 X 为论域的模糊集合), 确定了 G 也就确定了模糊粒子 P:

$$g \overset{\Delta}{=} x \text{ is } G \tag{1}$$

所以模糊化过程的本质就是确定一个函数 A 的过程, A 是模糊概念 G 的隶属函数, 既 $A=\mu_G$. 通常粒化时首先确定模糊概念的基本形式, 然后确定具体的隶属函数 A.

一般情况下, 模糊粒子 P 可以代替模糊概念 G, 既 P 可简单描述为:

$$P = A(x) \tag{2}$$

常用的模糊粒子有以下几种基本形式[5]: 三角形, 高斯型, 抛物型. 本文选用三角型模糊粒子, 其隶属函数如下:

$$A(x,a,m,b)=\begin{cases}0, x<a\\ \dfrac{x-a}{m-a}, a\le x\le m\\ \dfrac{b-x}{b-m}, m<x\le b\\ 0, x>b\end{cases} \tag{3}$$

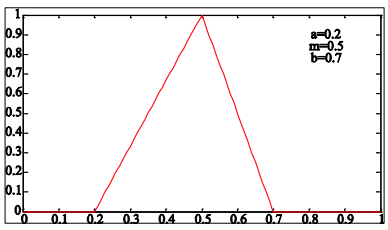

图 1　三角型隶属函数

建立模糊粒子的基本思想主要有: ①模糊粒子能够合理地代表原始数据; ②模糊粒子要有一定的特殊性. 既无论使用那种形式的模糊集来建立模糊粒子, 都要满足上面建立模糊粒子的基本思想. 为了满足上述的两个要求, 找到两者之间的最佳平衡, 可以考虑建立如下关于 A 的一个函数:

$$Q_A=\frac{M_A}{N_A} \tag{4}$$

其中, M_A 满足建立模糊粒子的基本思想①, N_A 满足建立模糊粒子的基本思想②.

当取 $M_A=\sum_{x\in X}A(x)$、$N_A=measure(supp(A))$ 时

$$Q_A=\frac{\sum_{x\in X}A(x)}{measure(supp(A))} \tag{5}$$

则为满足模糊粒子的基本思想, 只需 Q_A 越大越好. M_A 是模糊集合 A 的能量值, 表示隶属函数的泛化性; N_A 表示模糊集合 A 的语义性度量, $supp(A)$是模糊集合的支持集. 通过最大化指标 Q 确定细粒度的初始隶属函数参数 a,m,b 值.

由于不同的特征对分类的重要度不同, 所以下一步需要结合分类性能对每个特征的粒化结果分别进行优化处理. 设特征 f_i 模糊粒化后的结果为 $G_i, G_i=\{g_{i1},g_{i2},\cdots,g_{im_i}\}$ $(i=1,\cdots,n)$, m_i 表示第 i 个特征包含的隶属函数个数, 并将该结果作为 SVM 的训练集样本.

2　SVM回归拟合模型

支持向量机的回归拟合基本思想是寻找一个最优分类面使得所有训练样本离该最优分类面的误差最小[6].

不是一般性的, 设含有 l 个训练样本的训练集样本对为 $\{(x_i,y_i),i=1,2,\cdots,l\}$, 其中, $x_i(x_i\in R^d)$ 是第 i 个训练样本的输入列向量, $x_i=[x_i^1,x_i^2,\cdots,x_i^d]^T$, $y_i\in R$ 为对应的输出值.

设在高维特征空间中建立的线性回归函数为:

$$f(x)=w\phi(x)+b \tag{6}$$

其中 $\phi(x)$ 为非线性映射函数.

定义 ε 线性不敏感损失函数

$$L(f(x),y,\varepsilon)=\begin{cases}0, & |y-f(x)|\le\varepsilon\\ |y-f(x)|-\varepsilon, & |y-f(x)|>\varepsilon\end{cases} \tag{7}$$

© 中国科学院软件研究所 http://www.c-s-a.org.cn

其中 $f(x)$为回归函数返回的预测值; y 为对应的真实值.

回归拟合的过程就是解出 w, b 的过程, 既

$$\begin{cases} \min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\varepsilon_i + \varepsilon_i^*) \\ s.t. \begin{cases} y_i - w\Phi(x_i) - b \le \varepsilon_i + \varepsilon_i^* \quad ,i = 1,2,\cdots,l \\ -y_i + w\Phi(x_i) + b \le \varepsilon_i + \varepsilon_i^* \\ \varepsilon_i \ge 0, \varepsilon_i^* \ge 0 \end{cases} \end{cases} \tag{8}$$

其中 $\varepsilon_i, \varepsilon_i^*$ 为松弛变量, C 为惩罚因子, C 越大表示对训练误差大于 ε 的样本惩罚越大, ε 规定了回归函数的误差要求, ε 越小表示回归函数的误差越小.

求解式(8), 并转换为对偶形式:

$$\begin{cases} \max\limits_{a,a^*}[-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(a_i - a_i^*)(a_j - a_j^*)K(x_i, x_j) - \\ \sum_{i=1}^{l}(a_i + a_i^*)\varepsilon + \sum_{i=1}^{l}(a_i - a_i^*)y_i] \\ s.t. \begin{cases} \sum_{i=1}^{l}(a_i - a_i^*) = 0 \\ 0 \le a_i \le C \\ 0 \le a_i^* \le C \end{cases} \end{cases} \tag{9}$$

其中, $K(x_i, x_j) = \Phi(x_i)\Phi(x_j)$ 为核函数.

设求式(9)得到的最优解为 $a = [a_1, a_2, \cdots, a_l], a^* = [a_1^*, a_2^*, \cdots, a_l^*]$, 则有

$$w^* = \sum_{i=1}^{l}(a_i - a_i^*)\Phi(x_i) \tag{10}$$

$$b^* = \frac{1}{N_{nsv}}\{\sum_{0<a_i<C}[y_i - \sum_{x_i \in SV}(a_i - a_i^*)K(x_i, x_j) - \varepsilon] + \sum_{0<a_i<C}[y_i - \sum_{x_j \in SV}(a_j - a_j^*)K(x_i, x_j) + \varepsilon]\} \tag{11}$$

其中, N_{nsv} 为支持向量个数.

于是, 回归函数为:

$$f(x) = w^*\Phi(x) + b^* = \sum_{i=1}^{l}(a_i - a_i^*)\Phi(x_i)\Phi(x) + b^* = \sum_{i=1}^{l}(a_i - a_i^*)K(x_i, x) + b^* \tag{12}$$

其中, 只有部分参数 $(a_i - a_i^*)$ 不为 0, 其对应的样本 x_i 即为问题中的支持向量.

其基本结构如图 2 所示, 输出是中间节点的线性组合, 每个中间节点对应一个支持向量[7].

从上述模型可以看出 SVM 分类器通过求解式(9)来获得最优分类超平面, 如随着数据规模增大, SVM 学习效率迅速下降, 在特定空间泛化能力受到限制. 通过模糊信息粒化方法模型提前对特征空间进行划分, 使其成为为一系列子空间, 求解目标转向为子空间构造 SVM 分类器. 由此直观上就可降低样本规模, 节省时间复杂度, 更重要的是可降低分类器的复杂性.

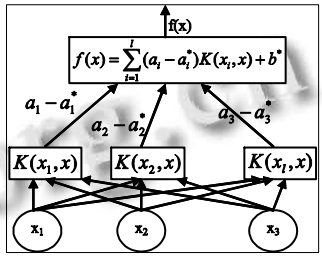

图 2 SVR 结构图

3 基于信息粒化的SVM

传统的 SVM 分类器将整个特征空间用一个连续的分类超平面划分, 而粒度SVM机制将整个特征空间划分为一系列子空间, 目标转向为子空间构造SVM分类器. 如此就可降低样本规模, 节省时间复杂度, 更重要的是可降低分类器的复杂性, 却保证了分类精度, 而且不会发生过学习现象并且提升了分类器的泛化能力. 同时这种学习方法本质上可以并行实现, 从而获得更高的学习效率[8,9]. 其算法流程如下:

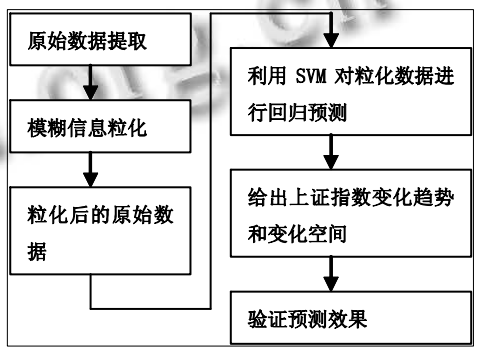

图 3 信息粒化的 SVM 流程图

从算法流程可以看出, 模糊信息粒化主要是对 SVM 的原始数据进行预处理, 关键在于粒化后的特征空间有效地简化了模糊特征和类别之间的关系, 粒化结果能达到最好的分类性, 从而提高 SVM 的学习效率.

4 实验与仿真

为了验证上述算法的效果本文采取该算法预测上

© 中国科学院软件研究所 http://www.c-s-a.org.cn

证开盘指数，首先收集从 1991 年 1 月 11 日到 2009 年 9 月 12 日这段期间的上证指数[10](如图 4 所示)，预测下 5 个交易日的变化趋势和变化空间.

首先利用模糊粒化模型对原始数据进行信息粒化，结果如图 5 所示 . 这里 min, M, max 分别为模糊粒子的三个参数，对应的是原始数据变化的最小值、大体平均水平和最大值.

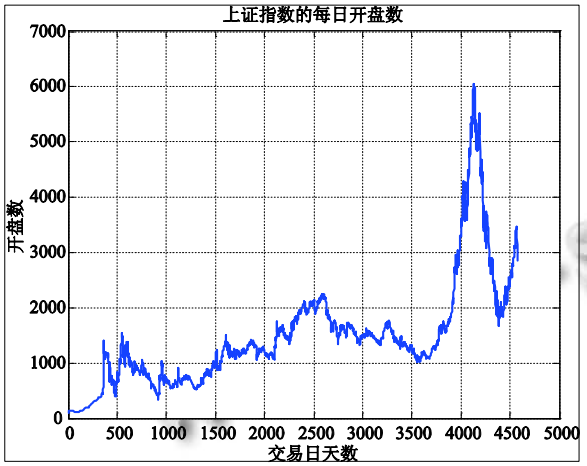

图 4　上证指数每日开盘数结果图

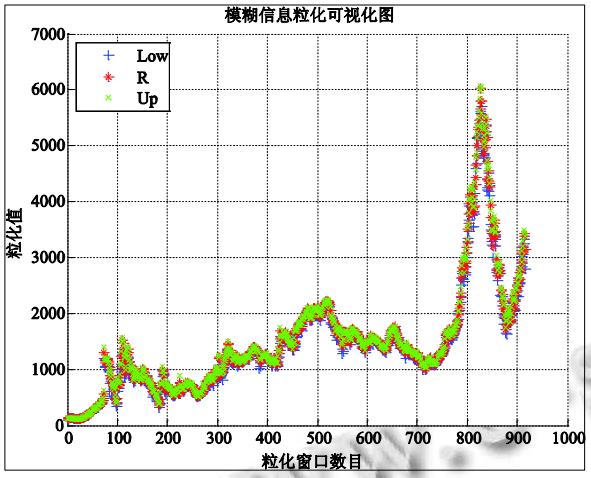

图 5　粒化结果图

由于对 min, M, max 三个模糊粒子进行 回归预测的过程类似，这里只给出模糊粒子 M 的运行结果 ，首先将 M 进行归一化处理，结果如图 6.

为了得到比较理想的回归预测效果，本文采取交叉验证的方法来寻找回归的最佳参数，交叉验证是用来验证分类器性能的一种统计分析方法，基本思想是在某种意义下将原始数据进行分组，一部分作为训练集，另一部分作为验证集. 其方法是首先用训练集对分类器进行训练，再利用验证集测试得到的模型，以得到的分类准确率作为评价分类器的性能指标. 交叉验证首先进行粗略的寻找，实验结果见图 7，观察粗略寻找结构后再进行精细选择，实验结果见图 8.

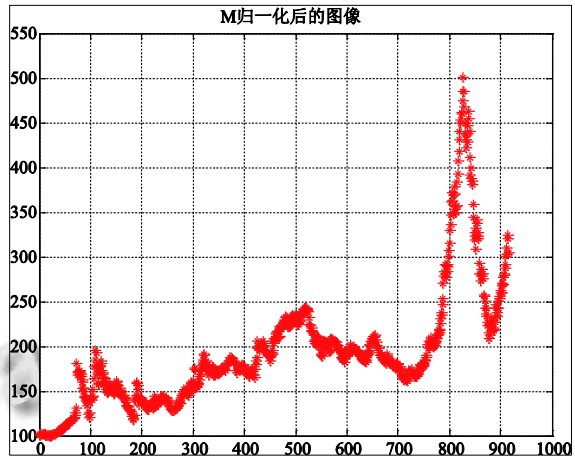

图 6　M 归一化后图像

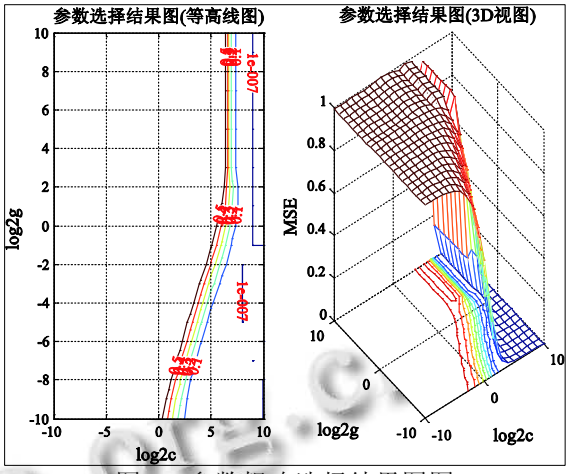

图 7　参数粗略选择结果图图

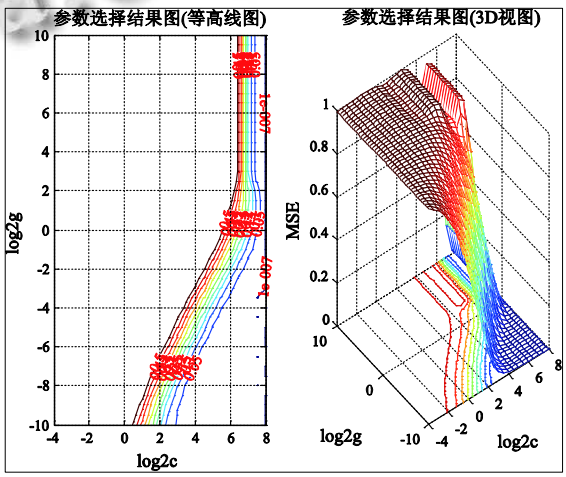

图 8　参数精细选择结果图

得到最佳参数 C=256, g=0.02209.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

利用上述参数进行进行训练和预测，可以得拟合效果图 9 和误差可视化图 10.

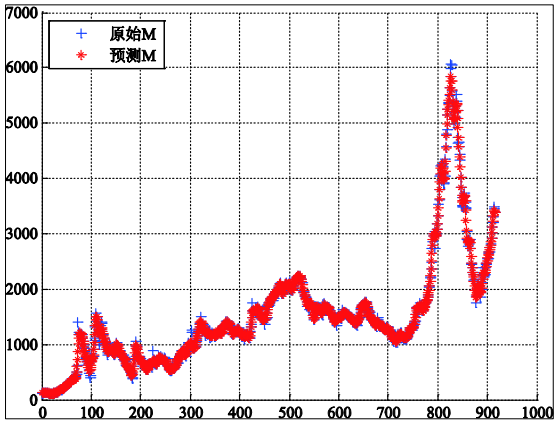

图 9 拟合结果图

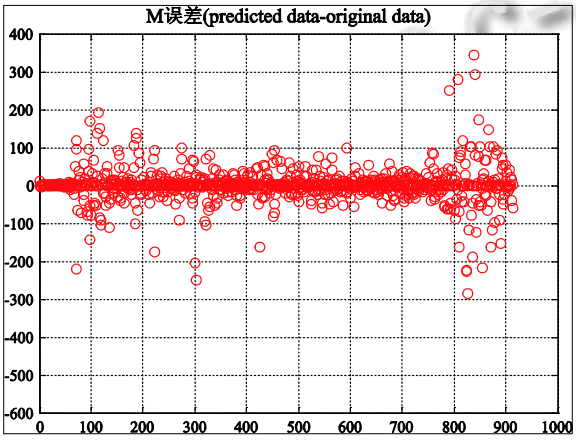

图 10 误差可视化图

为了进一步验证基于信息粒化的 SVM 模型的优势，我们同时采取传统 SVM 模型同样对上证开盘指数进行预测，但是由于训练样本较大,从而导致训练时间过长和效果不佳，具体实验结果见图 11 和图 12，从图 12 可以看出，传统 SVM 在大样本情况下，训练效果不佳.

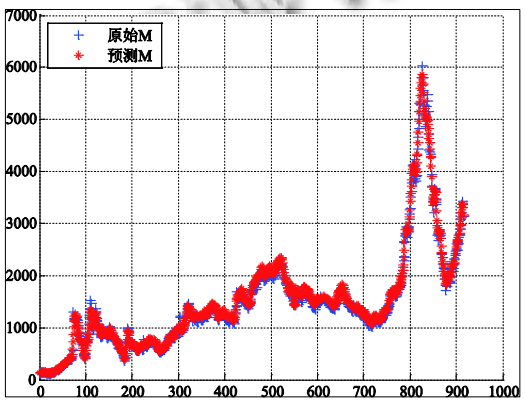

图 11 M 拟合结果图

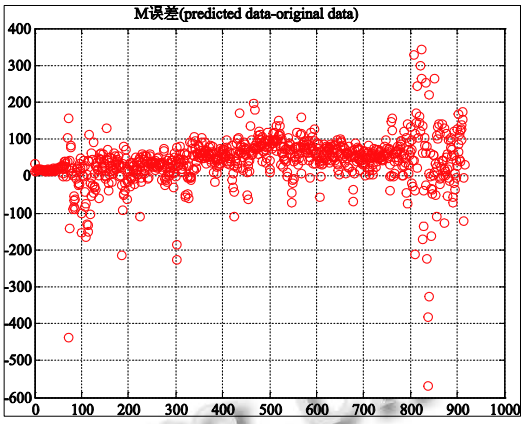

图 12 误差可视化图

对 max 和 min 也进行回归预测，得到的结果如下表 1:

表 1 上证指数变化趋势和变化空间预测表

日期	13	14	17	18	19	实际范围：[2796.3，3138.2，3380.2]
实际	311 2.6	313 8.2	299 4.9	284 5.3	291 6.1	预测范围：[2796.8，2950.0，3267.3]

5 结语

本文将粒度计算思想引入 SVM 中，用以改进传统 SVM 分类器的训练速度极大地受到训练集规模的影响、在特定的空间中泛化能力受到限制、应用领域有待于进一步拓展等缺陷，对上证指数进行了预测. 通过表 1 可以看出, 5 天内上证指数都在我们的预测范围内，这说明基于信息粒化的支持向量机回归预测方法有较好的预测效果，该研究成果不仅可以丰富 SVM 的理论和方法研究，同时 SVM 在非平衡数据处理的成功应用也有望进一步拓展 SVM 的应用领域.

参考文献

1 郎丛妍,须德,李兵.一种基于模糊信息粒化的视频时空显著单元提取方法.电子学报,2007,35(10):2023–2028.

2 Zadeh LA. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets and Systems,1997,90(2):111–127.

3 Skowron A. Toward intelligent systems: calculi of information granules. Proc. of International Workshop on Rough Set Theory and Granular Computing (RSTGC-2001), Bulletin of International Rough Set Society.Japan:[s.n.], 2001.

(下转第 206 页)

© 中国科学院软件研究所 http://www.c-s-a.org.cn

是可以接受的.

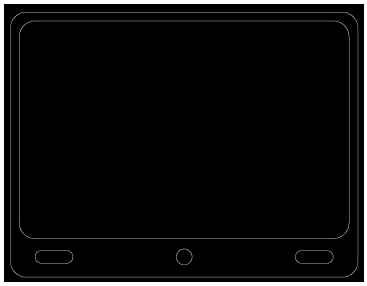

图 5 图元数量为 3000 的算例

表 2 优化效率表

图元数量	优化前耗时	优化后耗时	优化效率
500	0.047	0.0001	99%
3000	0.422	0.016	96%
10000	6.17	0.109	97%

5 结语

通过上述分析, 本文所提出的基于 HPGL 优化算法. 在耗时与路径优化都是可行且有效的. 该算法针对 HPGL 的大量图元进行处理, 在最坏情况下仍可以用空间换时间思想, 建立大张哈希表, 是整体平均复杂度为 O(n). 这在实际的生产过程中是能被操作人员所接受的. 通过对 HPGL 的处理, 将问题转化为基于 DXF 文件的路径优化问题, 采用文献[6]中的方法即可减少大量无效空行程, 提高了绘图效率, 降低了经济成本.

参考文献

1 穆海华,彭芳瑜,陈吉红.基于 DXF 的大型数控平面绘图机 CAM 系统.机床与液压,2003,107−109.

2 莫蓉.基于 DXF 的绘图轨迹优化及仿真系统研究实现[学位论文].西安:西北工业大学,2006.

3 胡胜红.基于图形计算几何技术的激光加工优化算法研究[学位论文].武汉:华中科技大学,2007.

4 吕欣泽.基于 DXF 文件的激光雕刻系统设计与实现[学位论文].天津:天津师范大学,2010.

5 陈树敏,刘强,方少亮等.DXF 排样切割中的应用.计算机应用与软件,2012,143−146.

6 龚清洪,常智勇,莫容,等.基于 DXF 文件的图元优化排序.计算机应用,2006,26(1):169−171.

7 胡胜红.对面域作图的 DXF 文件优化激光加工路径.工程图学学报,2010,(6):106−110.

8 马凯,杨泽林,吕静.基于 DXF 文件的 CAD /CAM 刀具路径优化与生成.机床与液压,2011,39(10):39−42.

9 甘明,陈小亮,张科威.基于 DXF 的数控切割加工优化算法的研究与实现.煤矿机械,2010,31(11):130−132.

(上接第 167 页)

4 修保新,刘忠,张维明,阳东升.基于信息粒化理论的主体间任务分配方法.国防科技大学学报,2007,29(3):71−75.

5 侯方国,胡圣武.模糊信息粒化理论在空间信息系统地位的探讨.测绘与空间地理信息,2004,27(3):14−16.

6 Chapelle O, Vapnik V, Bousquet O, Mukherjee S. Choosing multiple parameters for support vector machines.Machine Learning,2002,46(1-3).

7 Vladimir C, Ma YQ. Practical selection of SVM parameters and noise estimation for SVM regression.Neural Networks, 2004,17(1):113−126.

8 Keerthi SS, Lin CJ. Asymptotic behaviors of support vector machines with gaussian kernel. Neural Computation,2003,15 (7):1667−1689.

9 Lin HT, Lin CHJ. A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. Taipei: Department of Computer Science and Information Engineering National Tai wan University, 2005.

10 严晓明.基于优化 GA 属性约简的上证指数预测.福建师范大学学报(自然科学版),2011,27(5):29−33.

© 中国科学院软件研究所 http://www.c-s-a.org.cn