Dynamical Systems and Control 动力系统与控制, 2022, 11(1), 1-10
Published Online January 2022 in Hans. http://www.hanspub.org/journal/dsc
https://doi.org/10.12677/dsc.2022.111001

Hans汉斯

基于支持向量机的混杂数据过程控制

孙　静，郭叙成

清华大学，经管学院，现代管理研究中心，北京

收稿日期：2021年12月17日；录用日期：2021年12月31日；发布日期：2022年1月18日

摘　要

近年来，随着制造系统智能化和复杂化程度提升，获取的数据呈现出数据来源多元化、数据量激增、数据类型混杂等特点。本文研究了计量数据(measurable data)与属性数据(attribute data)的混杂数据过程控制与诊断问题。借助支持向量机(Support Vector Machine, SVM)，探讨了多元混杂数据出现均值偏移时的过程控制与诊断。沿着从多元正态数据到混杂数据的研究思路，探讨了SVM准确率和支持向量个数，在偏移量、子组大小和相关系数变化的情况下，对比四类核函数的SVM的过程控制能力；采用二分类SVM和多分类M-SVM，分析了过程控制和过程诊断能力；并与Hotelling多元控制图进行性能比较。研究发现，利用SVM可以有效实现混杂数据的过程控制和诊断。利用SVM对混杂数据进行过程控制，其性能远优于经典的Hotelling多元控制图。

关键词

混杂数据，多元过程控制，支持向量机

Process Control of Mixed Data Based on Support Vector Machine

Jing Sun, Xucheng Guo

Research Center of Contemporary Management, School of Economics and Management, Tsinghua University, Beijing

Received: Dec. 17th, 2021; accepted: Dec. 31st, 2021; published: Jan. 18th, 2022

Abstract

Recently, as manufacturing system is becoming increasingly automated and intelligent, the data obtained from manufacturing system also presents the characteristics of diversified data sources, large data volumes, and mixed data types. In this paper, it is discussed on process control and di-

文章引用: 孙静, 郭叙成. 基于支持向量机的混杂数据过程控制[J]. 动力系统与控制, 2022, 11(1): 1-10.
DOI: 10.12677/dsc.2022.111001

agnoses of mixed data in order to handle both measurable data and attribute data simultaneously. Support Vector Machine (SVM) is applied to realize the process control and diagnosis for multivariate mixed data when the process mean out of control. From multivariate normal data to mixed data, the accuracy of SVM is studied from the viewpoints of the shift, subgroup size, and the change of correlations to make comparison among four different kernels, and then C-SVM and M-SVM are provided to monitor the process in control or not and identify which parameter(s) to be out of control. Furthermore, compared SVM with Hotelling's multivariate control chart, it is concluded that SVM has excellent performance on monitoring the process in control or not when it is applied to mixed data.

Keywords

Mixed Data, Multivariate Process Control, Support Vector Machine

1. 引言

现代制造过程由于多个需要控制的特性可能存在关联，对多变量进行综合分析成为客观必然。Hotelling [1]率先明确了多元统计过程控制(Multivariate Statistical Process Control, MSPC)的必要性，后人在此基础上提出了多元过程控制的有效方法，目前 Hotelling 控制图仍是使用最为广泛的一种多元控制图。统计过程控制理论中，通常使用平均链长(Average Run Length, ARL)作为评价控制图性能的重要指标，以表示控制图从开始进行控制到发出警报信号为止所抽取的平均样本数，这是链长 RL (Run Length)的算术平均。

如今，数据类型混杂是制造业和服务业都需要面对的科学问题。然而，传统多元过程控制理论往往基于多元正态分布的假设前提，用于计量数据的过程控制。近年来，逐渐有关于多属性(Multi-Attribute)数据的研究成果[2] [3]，而计量数据(Variable/Measurable Data)与属性数据(Attribute Data)混杂领域的研究成果则较为匮乏[4]。

近来支持向量机(Support Vector Machine, SVM)对控制图的模式识别有不俗的表现[5] [6]。研究发现 SVM 具备同时处理属性数据和计量数据的能力，故 SVM 可用于数据类型混杂情况下的过程控制研究。然而，文献[4]指出在统计过程控制 SPC 领域，现有 SVM 的应用与研究还都是针对计量数据展开的。故而，本文利用 SVM 对混杂数据的过程控制与诊断进行研究。为了深入讨论基于 SVM 的过程控制与诊断，本文沿着从多元计量数据到混杂数据的研究思路，最后，将基于 SVM 的控制方法与 Hotelling 多元控制图就控制性能进行对比研究。

2. 基于 SVM 的计量数据的过程控制与诊断

模拟生成 k 维数据 $X=\left(x_1,x_2,\cdots,x_k\right)'$，数据彼此独立。过程异常是数据均值出现偏移且偏移发生在某个变更点(change point)的情况。由于 k 维数据的每个维度都有出现异常的可能，即存在受控和失控两种状态，故共有 2^k 种情况，包括：k 维数据均处于受控状态的情况，以及 $\left(2^k-1\right)$ 种存在失控数据的情况。需要随机生成等量的受控和失控子组，每个子组由 n 个观测构成。换言之，共生成 $l=\left(2^{k+1}-2\right)\times m$ 个 k 维子组，其中：$\left(2^k-1\right)\times m$ 个处于受控状态的子组，$\left(2^k-1\right)\times m$ 个处于失控状态的子组。训练集与测试集各占 50%。对于服从正态分布的计量数据，利用每个子组内的各变量的平均值、标准差、变量间的相

关系数，构成输入 SVM 的特征向量。

随后混杂数据的子组，由于随机生成 0-1 分布的数据，其子组的样本标准差有可能为 0，故不同于正态数据的处理，使用变量间的协方差代替相关系数，即利用每个子组内的各变量的平均值、标准差、以及变量间的协方差，构成输入 SVM 的特征向量。

选用支持向量机的四种核函数，通过网格搜索(Grid Search)对惩戒因子和核函数系数进行优化，搜索范围从 2^{-4} 到 2^{4}，步长为 0.5。取 5 折交叉验证准确率最高的结果作为参数进行 SVM 训练，得到分类结果。

2.1. 基于二分类 SVM 的过程控制

二维正态数据 $\left(x_1, x_2\right)'$ 的设置如表 1 所示。失控样本的均值偏移量 $\Delta\mu$ 取±1 和±2，对应着 1 个标准差和 2 个标准差的偏移。变量间相关系数 ρ 取±0.3 和±0.7，以考虑强相关与弱相关、正相关与负相关的情况。每个子组经过特征提取，向 SVM 输入 5 个特征变量 $\left(\bar{x}_{1i}, \bar{x}_{2i}, s_{1i}, s_{2i}, \rho_{12i}\right)$，即子组内各变量的平均值、标准差、以及变量间的相关系数。m 取 1000，共生成 6000 个二维子组，其中：受控状态子组与失控状态子组分别为 3000 组。子组大小 n 取 5、10、20。

Table 1. Determination of 2-dimensional normal data
表 1. 二维正态数据的设置

种类	状态	分布设定	数据量
1	统计受控状态	$x_1 \sim N(0,1), x_2 \sim N(0,1)$	$3m$ 个子组，每个子组 n 个观测
2	失控状态	$x_1 \sim N(\Delta\mu,1), x_2 \sim N(0,1)$	m 个子组，每个子组 n 个观测
3	失控状态	$x_1 \sim N(0,1), x_2 \sim N(\Delta\mu,1)$	m 个子组，每个子组 n 个观测
4	失控状态	$x_1 \sim N(\Delta\mu,1), x_2 \sim N(\Delta\mu,1)$	m 个子组，每个子组 n 个观测

$n=10$，$\rho=0.3$ 时均值偏移量变化的分析结果如表 2 所示。对比四种核函数的结果可见，Sigmoid 核函数与线性核函数表现不佳，而且均使用了几乎所有的数据作为支持向量；多项式核函数和 RBF 核函数的准确率则较高，其中多项式使用的支持向量个数更少。使用同一种核函数时，对比不同均值偏移量的分析结果可见，异常样本的均值偏移量越大，SVM 准确率越高。

Table 2. Results of 2-dimensional normal data when the shift of process mean is changed ($n=10$, $\rho=0.3$)
表 2. 均值偏移量不同时二元正态数据的分析结果($n=10$，$\rho=0.3$)

核函数	均值偏移量	受控误判数	失控误判数	准确率	SVs
Linear	1	292	1015	56.438%	2980
	2	239	866	63.186%	2972
Polynomial	1	93	144	92.110%	591
	2	2	2	99.881%	23
RBF	1	96	143	92.019%	655
	2	2	1	99.900%	110
Sigmoid	1	720	984	43.214%	3000
	2	1088	1214	23.257%	3000

$n=10$、均值偏移量$\Delta\mu$为1个标准差时相关系数变化的分析结果如表3所示。对比四种核函数的结果可见，不论变量间的相关关系是正相关还是负相关、是强相关还是弱相关，Sigmoid核函数与线性核函数都表现不佳，而且均使用了几乎所有的数据作为支持向量。对于表现较好的多项式核函数和RBF核函数，则变量间相关性越强，SVM准确率越高；正相关还是负相关的影响不大。

Table 3. Results of 2-dimensional normal data when the correlation is changed ($n=10$, $\Delta\mu:1\sigma$)
表3. 相关系数不同时二元正态数据的分析结果($n=10$，$\Delta\mu$为1个标准差)

核函数	相关系数	受控误判数	失控误判数	准确率	SVs
Linear	−0.7	286	948	58.867%	2984
	−0.3	312	959	57.638%	2981
	0.3	292	1015	56.438%	2980
	0.7	269	972	58.638%	2984
Polynomial	−0.7	41	77	96.071%	363
	−0.3	88	147	92.157%	582
	0.3	93	144	92.110%	591
	0.7	40	71	96.276%	350
RBF	−0.7	46	77	95.886%	433
	−0.3	104	128	92.262%	643
	0.3	96	143	92.019%	655
	0.7	44	77	95.952%	426
Sigmoid	−0.7	928	802	42.329%	3000
	−0.3	632	1065	43.443%	3000
	0.3	720	984	43.214%	3000
	0.7	1033	752	40.500%	3000

$\rho=0.3$，均值偏移量$\Delta\mu$为1个标准差时子组大小变化的分析结果如表4所示。对比四种核函数的结果可见，不论子组大小n为5、10、20，Sigmoid核函数与线性核函数都表现不佳，而且均使用了几乎所有的数据作为支持向量。对于表现较好的多项式核函数和RBF核函数，则随着子组大小的减少，SVM准确率会降低，所需的支持向量数增加。

Table 4. Results of 2-dimensional normal data when the subgroup size is changed ($\rho=0.3$, $\Delta\mu:1\sigma$)
表4. 子组大小不同时二元正态数据的分析结果($\rho=0.3$，$\Delta\mu$为1个标准差)

核函数	子组大小	受控误判数	失控误判数	准确率	SVs
Linear	5	431	993	52.538%	2971
	10	292	1015	56.438%	2980
	20	168	924	63.591%	2992
Polynomial	5	186	357	81.891%	1265
	10	93	144	92.110%	591
	20	18	31	98.367%	192

Continued

RBF	5	221	321	81.924%	1319
	10	96	143	92.019%	655
	20	21	28	98.352%	232
Sigmoid	5	764	956	42.662%	3000
	10	720	984	43.214%	3000
	20	602	892	50.229%	3000

综合表 2、表 3 和表 4 的分析结果可见，利用基于 Sigmoid 核函数与线性核函数的 SVM 进行过程控制皆表现不佳，故而，随后的分析都是利用基于多项式核函数 Polynomial 和 RBF 核函数的 SVM 进行过程控制与诊断。

2.2. 基于多分类 M-SVM 的过程诊断

前面利用二分类 SVM 对受控样本和失控样本进行区分，以实现基于 SVM 的过程控制。本节将利用多分类 M-SVM，不仅要对受控状态与失控状态进行区分，而且要对出现了哪类失控状态进行诊断。

利用多项式核函数和 RBF 核函数的 M-SVM 进行异常诊断，此时 $\rho=0.3$，$n=10$。多分类 M-SVM 共需训练 $4(4-1)/2=6$ 次二分类 SVM，投票以确定最终分类。均值偏移量 $\Delta\mu$ 不同时基于 M-SVM 的诊断结果如表 5 所示。

Table 5. Results of diagnosis based on M-SVM when the shift of process mean is changed ($\rho=0.3$，$n=10$)
表 5. 均值偏移量 $\Delta\mu$ 不同时基于 M-SVM 的诊断结果($\rho=0.3$，$n=10$)

核函数	均值偏移量 $\Delta\mu$	“1”受控准确率	“2”失控准确率	“3”失控准确率	“4”失控准确率	总准确率
Polynomial	1	95.276%	79.229%	79.514%	79.571%	87.357%
	2	99.952%	99.600%	99.714%	99.114%	99.714%
RBF	1	93.914%	80.829%	80.114%	82.600%	87.548%
	2	99.924%	99.657%	99.771%	99.943%	99.857%

由表 5 可见，均值偏移量 $\Delta\mu$ 增大时，不论是判断受控状态的准确率、还是判断 2、3、4 这三种失控状态的准确率，以及总准确率都会增加，这与逻辑判断相符。有趣的是，基于多项式核函数的 M-SVM 在判断受控状态时表现略优于基于径向基 RBF 核函数的 M-SVM；而在诊断 2、3、4 这三种失控状态时，基于径向基 RBF 核函数的 M-SVM 表现略佳；从总准确率上看，基于径向基 RBF 核函数的 M-SVM 略优于基于多项式核函数的 M-SVM。

对比二分类 SVM 的准确率(见表 2)与多分类 M-SVM 的总准确率(见表 5)，不论是利用多项式核函数还是径向基 RBF 核函数，二分类 SVM 的准确率都优于 M-SVM，尤其当均值偏移量 $\Delta\mu$ 为 1 个标准差的小偏移时，二分类 SVM 的优势更明显。故而，当需要对受控状态与失控状态进行控制时，二分类 SVM 的控制方法更为有效；当需要进一步对不同的失控状态进行诊断时，建议使用 M-SVM 的诊断方法。

3. 基于 SVM 的混杂数据的过程控制与诊断

3.1. 基于二分类 SVM 的混杂数据的过程控制

本文对三维混杂数据进行研究。三维混杂数据依次服从正态分布、0-1 分布和泊松分布，数据之间存

在相关性。即三维数据 $X=(x_1,x_2,x_3)'$，其中：$x_1\sim N(\mu,1),x_2\sim B(1,p),x_3\sim P(\lambda)$。

由于随机生成 0-1 分布的子组，其子组的样本标准差有可能为 0，导致相关系数无意义，故不同于前面对正态数据的处理，使用变量间的协方差代替相关系数，即利用每个子组内的各变量的平均值、标准差、以及变量间的协方差，构成输入 SVM 的特征向量 $(\bar{x}_{1i},\bar{x}_{2i},\bar{x}_{3i},s_{1i},s_{2i},s_{3i},S_{12i},S_{13i},S_{23i})$。考虑到受控状态以及所有可能存在的失控状态，共需生成 8 类三维混杂数据，具体设置如表 6 所示。

Table 6. Determination of 3-dimensional mixed data
表 6. 三维混杂数据的设置

种类	状态	分布设定	数据量
1	统计受控状态	$x_1\sim N(0,1),x_2\sim B(1,p),x_3\sim P(\lambda)$	$7m$ 个子组，每个子组 n 个观测
2	失控状态	$x_1\sim N(\Delta\mu,1),x_2\sim B(1,p),x_3\sim P(\lambda)$	m 个子组，每个子组 n 个观测
3	失控状态	$x_1\sim N(0,1),x_2\sim B(1,p+\Delta p),x_3\sim P(\lambda)$	m 个子组，每个子组 n 个观测
4	失控状态	$x_1\sim N(0,1),x_2\sim B(1,p),x_3\sim P(\lambda+\Delta\lambda)$	m 个子组，每个子组 n 个观测
5	失控状态	$x_1\sim N(\Delta\mu,1),x_2\sim B(1,p+\Delta p),x_3\sim P(\lambda)$	m 个子组，每个子组 n 个观测
6	失控状态	$x_1\sim N(\Delta\mu,1),x_2\sim B(1,p),x_3\sim P(\lambda+\Delta\lambda)$	m 个子组，每个子组 n 个观测
7	失控状态	$x_1\sim N(0,1),x_2\sim B(1,p+\Delta p),x_3\sim P(\lambda+\Delta\lambda)$	m 个子组，每个子组 n 个观测
8	失控状态	$x_1\sim N(\Delta\mu,1),x_2\sim B(1,p+\Delta p),x_3\sim P(\lambda+\Delta\lambda)$	m 个子组，每个子组 n 个观测

取 $\mu=0,p=0.2,\lambda=3$。本文期望对正态近似效果较差的混杂数据进行研究，故数据的分布参数以及子组大小的选定尽量远离正态近似的条件，n 取 10。

对于失控数据的参数确定，选择均值偏移量为 1 个标准差。正态数据 x_1 出现 $\Delta\mu=1$ 的偏离；0-1 数据 x_2 的均值偏移量为 $\Delta p=\sqrt{p(1-p)}$，由 $p=0.2$ 得到 $\Delta p=\sqrt{0.2(1-0.2)}=0.4$；泊松数据 x_3 的均值偏移量为 $\Delta\lambda=\sqrt{\lambda}$，由 $\lambda=3$ 得到 $\Delta\lambda=\sqrt{3}$。

考虑到相关性强弱的情况，分别设置为 R_1 和 R_2：

$$R_1=\begin{bmatrix}1 & 0.3 & 0.3\\ 0.3 & 1 & 0.3\\ 0.3 & 0.3 & 1\end{bmatrix},R_2=\begin{bmatrix}1 & 0.7 & 0.7\\ 0.7 & 1 & 0.7\\ 0.7 & 0.7 & 1\end{bmatrix}$$

m 取 1000。共生成 14000 个子组，其中：受控子组和失控子组各 7000 组。基于二分类 SVM 的混杂数据分析结果如表 7 所示。

由表 7 可见，利用二分类 SVM 对混杂数据处于受控状态还是失控状态进行控制，不论相关性强弱，其判断的准确率始终高于 93%。相比于，利用二分类对正态数据的控制，即表 2、表 3 和表 4 中 n 取 10、均值偏移量为 1 个标准差的分析结果，基于二分类 SVM 的混杂数据的控制准确度与正态数据的准确度基本持平，略有提升。

由表 7 可见，相关性的增强会带来准确率的提升与支持向量数的下降，与逻辑判断相符。对比多项式 polynomial 核函数与径向基 RBF 核函数的分析结果可见，在支持向量数上，多项式 polynomial 核函数明显少于径向基 RBF 核函数；在准确率上，弱相关时径向基 RBF 核函数略高，强相关时多项式 polynomial 核函数略高，相差幅度有限。

Table 7. Results of C-SVM for mixed data
表 7. 基于二分类 SVM 的混杂数据分析结果

相关性	核函数	准确率	SVs
弱相关性	Polynomial	93.019%	1222
	RBF	93.149%	1537
强相关性	Polynomial	96.357%	760
	RBF	96.202%	1036

3.2. 基于多分类 M-SVM 的混杂数据的过程诊断

多分类 M-SVM 共需要训练$8(8-1)/2=28$次二分类 SVM，投票以确定最终分类。基于 M-SVM 的诊断结果如表 8 所示。

Table 8. The accuracy of diagnosis based on M-SVM for mixed data
表 8. 基于 M-SVM 的混杂数据诊断准确率

种类	弱相关性		强相关性	
	Polynomial	RBF	Polynomial	RBF
1	95.757%	96.000%	98.440%	98.706%
2	74.640%	72.920%	88.640%	85.980%
3	65.320%	61.220%	83.000%	77.840%
4	67.580%	65.860%	85.140%	82.420%
5	80.660%	80.760%	88.320%	87.820%
6	75.400%	74.740%	89.520%	88.900%
7	80.600%	78.200%	87.020%	87.060%
8	80.320%	78.200%	93.680%	91.680%
总准确率	85.344%	84.564%	93.173%	92.331%

由表 8 可见，相关性增强时，不论是判断受控状态的准确率、还是判断 2 至 8 这七种失控状态的准确率、以及总准确率都会明显增大，这与逻辑判断相符。

值得注意的是，基于多项式 polynomial 核函数的 M-SVM 在判断受控状态时表现略差于基于径向基 RBF 核函数的 M-SVM，该结论与正态数据得到的结论相反；在诊断 2 至 8 这七种失控状态且相关性弱时，基于多项式 polynomial 核函数的 M-SVM 表现更佳，这也与正态数据得到的结论相反；从总准确率上看，基于径向基 RBF 核函数的 M-SVM 略差于基于多项式核函数的 M-SVM，亦与正态数据得到的结论相反。由此可见，在利用 M-SVM 进行诊断时，对核函数的考虑需要特别审慎。

对比二分类 SVM 的准确率(见表 7)与多分类 M-SVM 的总准确率(见表 8)，不论是利用多项式核函数还是径向基 RBF 核函数，二分类 SVM 的准确率都优于 M-SVM，尤其是弱相关时，二分类 SVM 的优势更明显。故而，当需要在受控状态与失控状态之间进行控制时，二分类 SVM 的控制方法更为有效；当需要进一步对不同的失控状态进行诊断时，建议使用 M-SVM 的诊断方法，这与正态数据给出的建议一致。

4. 与 Hotelling 多元控制图的性能对比

本文旨在研究计量数据与属性数据的混杂数据过程控制与诊断问题。借助支持向量机，探讨了多元

混杂数据出现均值偏移时的过程控制与诊断，至此，需要进行混杂数据过程控制与诊断方法的性能分析。若仅是进行多元正态数据或是多元属性数据的控制方法的性能对比，那么，都有多种统计控制方法可进行对比研究。然而，进行混杂数据的控制方法的性能对比，那么，只有选择基于 Hotelling 统计量的多元控制图。

基于 Hotelling 统计量设计的多元控制图是应用最广泛、对存在关联关系的多变量进行联合控制的有效方法。Hotelling 统计量的统计特性源于存在关联关系的多个变量服从多元正态分布的基本假设。面对混杂数据时，则借助正态近似，利用多元控制图进行控制。下面就本文提出的基于 SVM 的控制方法与 Hotelling 多元控制图[7]就控制性能进行对比。

4.1. 性能对比的设计

对比基于 SVM 的控制方法与 Hotelling 多元控制图的控制性能，需要对相同的研究对象进行分析，即数据集相同；同时，待考核的性能指标必须是可比的。

统计过程控制通常使用平均链长 ARL (average run length)作为评价控制性能的重要指标。以 Hotelling 多元控制图为例，ARL(0)和 ARL(1)分别表示过程处于受控状态和失控状态时，连续两次出现落在上控制限 UCL(upper control limit)之外的出界点之间的平均链长。显然，基于 SVM 的控制方法本身并没有考虑数据的序列问题，不存在平均链长 ARL 的概念。本文将借助控制方法中普遍存在的两类错误概率 α 和 β，来建立多元控制图的 ARL(0)与 ARL(1)相对于基于 SVM 控制方法对受控状态与失控状态的分类准确率 *Accuracy* 之间的对应关系，具体表述如下：

$$Accuracy(0)=1-\alpha=1-1/\mathrm{ARL}(0) \tag{1}$$

$$Accuracy(1)=1-\beta=1/\mathrm{ARL}(1) \tag{2}$$

为了与 ARL(0)、ARL(1)相统一，这里使用的 *Accuracy*(0)、*Accuracy*(1)分别代表受控状态、失控状态的分类准确率。

利用平均链长 ARL 分析控制图的控制性能的常规思路是：在给定受控状态 ARL(0)的情况下，比较失控状态的 ARL(1)，进而，对控制图发现过程出现异常的灵敏程度进行对比研究。本文首先对训练集进行研究，利用训练集中的受控状态子组和失控状态子组，学习得到二分类 SVM 模型；接着，利用得到的二分类 SVM 模型，对训练集中的受控子组进行判断，得到受控状态子组的准确率 *Accuracy*(0)；随后，利用公式(1)，计算得到第 I 类错误的概率 α。至此，为下一步控制性能的对比研究做好准备。

基于 SVM 的控制方法在前文中介绍了各变量分布参数的设置。进行对比研究的 Hotelling 多元控制图，在同样的分布参数设置下设计控制图，即 χ^2 控制图。利用第 I 类错误的概率 α，得到 χ^2 控制图的上控制限 UCL。

针对相同的测试集，使用 χ^2 控制图的上控制限 UCL，得到失控状态的 ARL(1)；使用二分类 SVM 的控制方法，得到失控状态的准确率 *Accuracy*(1)。借助公式(2)，对 ARL(1)和 *Accuracy*(1)进行适当转换，对控制性能进行比较。

4.2. 正态数据的控制性能对比

对正态数据的控制性能进行对比研究。二维数据 $\left(x_1, x_2\right)'$ 的设置如表 1 所示。失控样本的均值偏移量 $\Delta\mu$ 取 1 个标准差的偏移，变量间相关系数 ρ 取 0.7，子组大小 n 取 10，m 取 1000，共生成 6000 个子组，其中：受控状态子组与失控状态子组各 3000 组。并等分为训练集与测试集。

首先，利用训练集的 3000 个子组，学习得到二分类 SVM 模型；接着，利用该二分类 SVM 模型，

对训练集中的 1500 个受控子组进行判断，得到受控状态的准确率 *Accuracy*(0) = 97.4%，即第 I 类错误的概率 $\alpha = 0.026$。随后，由第 I 类错误的概率 α，可以得到 χ^2 控制图的上控制限 UCL = 7.30。最后，针对测试集中的 1500 个失控子组，利用 χ^2 控制图的上控制限 UCL，得到失控状态的 ARL(1)；利用二分类 SVM 的控制方法，得到失控状态的准确率 *Accuracy*(1)。分析结果如表 9 所示。

Table 9. Comparison of process control performance for normal data ($\rho = 0.7$， $n = 10$， $\Delta\mu : 1\sigma$)
表 9. 正态数据的控制性能比较($\rho = 0.7$， $n = 10$， $\Delta\mu$ 为 1 个标准差)

过程控制方法	α	准确率	ARL
SVM	0.026	94.705%	1.0559
χ^2 控制图	0.026	94.810%	1.0547

由表 9 可见，基于 SVM 的控制和 χ^2 控制图的准确率以及 ARL 都非常相近。从数值来看，χ^2 控制图的准确率和ARL都略优于基于SVM的控制方法，说明Hotelling多元控制图至今广为使用的客观必然。

4.3. 混杂数据的控制性能对比

对混杂数据的控制性能进行对比研究。三维混杂数据 $(x_1, x_2, x_3)'$ 的设置如表 6 所示。受控状态子组取 $\mu = 0$， $p = 0.2$， $\lambda = 3$；失控状态子组的参数偏移量为 1 个标准差的偏移，且仅产生正向的偏移，得到 $\Delta\mu = 1$， $\Delta p = 0.4$， $\Delta\lambda = \sqrt{3}$。变量间相关性强，为 R_2。子组大小 n 取 10。m 取 1000，共生成 14000 个子组，其中：受控状态子组与失控状态子组各 7000 组。并等分为训练集与测试集。

首先，利用训练集的 7000 个子组，学习得到二分类 SVM 模型；接着，利用该二分类 SVM 模型，对训练集中的 3500 个受控子组进行判断，得到受控状态的准确率 *Accuracy*(0) = 97.5%，即第 I 类错误的概率 $\alpha = 0.025$。随后，由第 I 类错误的概率 α，可以得到 χ^2 控制图的上控制限 UCL = 9.35。最后，针对测试集中的 3500 个失控子组，利用 χ^2 控制图的上控制限 UCL，得到失控状态的 ARL(1)；利用基于二分类 SVM 的控制方法，得到失控状态的准确率 *Accuracy*(1)。分析结果如表 10 所示。

Table 10. Comparison of process control performance for mixed data (correlation matrix: R_2)
表 10. 混杂数据的控制性能比较(相关系数矩阵为 R_2)

过程控制方法	α	准确率	ARL
SVM	0.025	95.971%	1.0420
χ^2 控制图	0.025	37.351%	2.6773

由表 10 可见，相比于 χ^2 控制图，基于 SVM 的控制方法其控制性能得到了极其显著地提升，其准确率与 ARL 提高了 1.5 倍多。显然，当面对混杂数据，尤其在混杂数据不能满足正态近似条件时，基于 SVM 的控制方法将大幅度改进控制性能，是远优于广为使用的 Hotelling 多元控制图的一种有效的控制途径。

5. 结论

本文利用支持向量机 SVM 实现了对混杂数据的过程控制与诊断，研究发现：

1) 相比于线性核函数和 Sigmoid 核函数，利用基于多项式 polynomial 核函数和径向基 RBF 核函数的支持向量机进行控制，均有较高的准确率。子组大小越大、变量间相关性越强以及均值偏移量越大，则

准确率越高。

2) 对比二分类 SVM 的准确率与多分类 M-SVM 的总准确率可见，二分类 SVM 优于 M-SVM。故而，当需要在受控状态与失控状态之间进行控制时，二分类 SVM 的控制方法更为有效；当需要进一步对不同的失控状态进行诊断时，建议使用基于 M-SVM 的诊断。在利用 M-SVM 进行诊断时，对核函数的考虑需要特别审慎，建议同时尝试多项式 polynomial 核函数和径向基 RBF 核函数。

相比于 Hotelling 多元 χ^2 控制图，在控制多元正态数据时，基于 SVM 的控制方法与 χ^2 控制图的控制性能非常相近。在控制混杂数据时，基于 SVM 的控制方法则大幅提升控制性能，故而，基于 SVM 的控制是一种更为稳健地控制混杂数据的方法。利用多分类 M-SVM，可对多元数据中出现哪种失控状态进行有效诊断。然而，Hotelling 多元控制图只能对是否出现失控状态进行判断，而无法进行诊断。

基金项目

国家自然科学基金资助项目(NSFC-71672100)。

参考文献

[1] Hotelling, H. (1947) Multivariate Quality Control Illustrated by the Air Testing of Sample Bombsights, in Techniques of Statistical Analysis. McGraw Hill, New York, 111-184.

[2] Niaki, S.T.A. and Khedmati, M. (2013) Estimating the Change Point of the Parameter Vector of Multivariate Poisson Processes Monitored by a Multiattribute T^2 Control Chart. *International Journal of Advanced Manufacturing Technology*, **64**, 1625-1642. https://doi.org/10.1007/s00170-012-4128-x

[3] Niaki, S.T.A. and Khedmati, M. (2014) Monotonic Change-Point Estimation of Multivariate Poisson Processes Using a Multi-Attribute Control Chart and MLE. *International Journal of Production Research*, **52**, 2954-2982. https://doi.org/10.1080/00207543.2013.857797

[4] Weese, M., Martinez, W., Megahed, F.M., *et al*. (2016) Statistical Learning Methods Applied to Process Monitoring: An Overview and Perspective. *Journal of Quality Technology*, **48**, 4-24. https://doi.org/10.1080/00224065.2016.11918148

[5] Zhou, X., Jiang, P. and Wang, X. (2018) Recognition of Control Chart Patterns Using Fuzzy SVM with a Hybrid Kernel Function. *Journal of Intelligent Manufacturing*, **29**, 51-67. https://doi.org/10.1007/s10845-015-1089-6

[6] Zhang, M., Yuan, Y., Wang, R., *et al*. (2020) Recognition of Mixture Control Chart Patterns Based on Fusion Feature Reduction and Fireworks Algorithm-Optimized MSVM. *Pattern Analysis & Applications*, **23**, 15-26. https://doi.org/10.1007/s10044-018-0748-6

[7] Montgomery, D.C. (2013) Introduction to Statistical Quality Control. 7th Edition, Wiley, New York, 514-523.