# 带置信度的混合压缩相符预测器模型研究

王华珍　林成德　杨　帆　庄进发

(厦门大学 自动化系，福建 厦门 361005)

摘要：提出一种改进的混合压缩相符预测器. 混合压缩相符预测器采用两阶段压缩过程：先将部分序列样本压缩成模型以知识的形式保存；再将知识传递给后续样本用于置信预测. 混合压缩相符预测器不仅能提高计算效率，还提供巧妙的邻近性度量，从而极大地提高了第二阶段的预测效率. 以田纳西-伊斯曼化工过程为例，验证了该方法的有效性和高效性.

关　键　词：故障检测；置信度；在线推理；转导机器；相符预测器；工业大系统

中图分类号：TP19　　文献标识码：A　　文章编号：1671-4512(2009)01-0088-04

# The model research of hybrid-compression conformal predictor with confidence

*Wang Huazhen　Lin Chengde　Yang Fan　Zhuang Jinfa*

(Department of Automation, Xiamen University, Xiamen 361005, China)

**Abstract**: Conformal predictor is extended to hybrid-compression conformal predictor (HCCP) in order to improve the computational efficiency. HCCP executes compression in two stages: **a**. a compression expert is assigned to compress part of the sequence of data; **b**. it transmits the extracted information to the successive transductive prediction. As a result, HCCP yields competitive computational efficiency, as well as maintaining the predictive efficiency, due to the ingenious proximity between the examples produced in the second stage. A case study of Tennessee Eastman Process was provided to illustrate the advantage of the proposed method.

**Key words**: fault test; confidence; online conferencing; transducers; conformal predictor; large-scale industrial system

故障检测是工业大系统异常事件管理的核心问题. 故障检测具有模式识别的特点，许多学者就基于历史数据的故障检测算法展开研究[1]. 关于相符预测器(CP)的研究也有很多成果[2~7]. 本文提出一种改进算法，即混合压缩相符预测器(HCCP)，它采用两阶段压缩过程，先将部分序列样本压缩成模型以知识的形式保存，再将知识传递给后续样本用于置信预测，其创新性和优势在于：**a**. 降低存储规模. HCCP 将部分序列样本以模型知识形式存储，原始样本不再需要，因此提高了存储效率，并提高计算效率. **b**. 提供巧妙的邻近性度量. 随机森林模型能产生邻近度来衡量样本之间的关系. HCCP 第二阶段样本被映射到邻近度空间，相当于把随机森林信息融入样本奇异度测量，没有因为样本信息浪费而导致预测效率下降. 而且样本在邻近度空间里界线分明，极大地提高了第二阶段的预测效率.

## 1　HCCP 算法原理

### 1.1　经典相符预测器理论

学习问题描述如下：研究对象产生样本序列

收稿日期：2008-03-14.

作者简介：王华珍(1975-)，女，博士研究生，E-mail: hzwang@xmu.edu.cn

基金项目：国家高技术研究发展计划资助项目(2006AA01Z129).

$(x_1, y_1)$, …, $(x_{n-1}, y_{n-1}) = z_1, z_2, \cdots, z_{n-1} = Z^{(n-1)}$ 和一个没有类别的测试数据 $x_n$. $X$ 为属性空间, $x_i \in X\ (i = 1, 2, \cdots, n)$; $Y$ 为类别空间, $y_i \in Y\ (i = 1, 2, \cdots, n-1)$, $Z = X \times Y$ 为样本空间. $y \in Y$ 是 $x_n$ 的假设类别, $z_n = (x_n, y)$. 上述学习问题记为 $E$.

样本奇异函数的设计决定了相符预测器(CP)的预测有效性[8~10]. 样本奇异函数 $A_n$ 是样本空间到实数空间的一种与样本出现顺序无关的映射, 它将样本序列 $z_1, z_2, \cdots, z_n$ 映射成样本奇异值序列 $\alpha_1, \alpha_2, \cdots, \alpha_n$. 样本奇异值反映样本对样本序列形成的分布的隶属程度, 奇异值越大样本的隶属度越小.

当给定置信度 $1-\varepsilon$($\varepsilon$ 是重要性水平), CP 对 $x_n$ 的预测结果为

$$\Gamma^{\varepsilon,\tau_n}(z_1, z_2, \cdots, z_{n-1}, x_n, \tau_n) = \{y \in Y: p_y = |\{i = 1, 2, \cdots, n; \alpha_i > \alpha_n\}| + \tau_n |\{i = 1, 2, \cdots, n; \alpha_i = \alpha_n\}|/n > \varepsilon\}, \quad (1)$$

式中: $y$ 是 $x_n$ 可能取得的类别; $\tau_n (n \in N)$ 是在 $[0, 1]$ 上服从均匀分布的随机变量; $p_y$ 即 $P$ 值, 是 $y$ 成为被测数据真实类别的置信度. 凡是满足 $p_y > \varepsilon$ 的 $y$ 都是 $\Gamma^{\varepsilon,\tau_n}$ 元素, 因此 CP 输出的是含有多个类别的预测集. 若 $x_n$ 的真实类别没有出现在预测集 $\Gamma^{\varepsilon,\tau_n}$ 中, 则预测错误. CP 理论最突出的特点是具有可校准性, 即预测的准确率对等于用户预先设定的置信度.

采用归纳式相符预测器 (ICP) 对样本进行预测时, 用部分序列样本学习一个通用规则, 这个规则被用来测量剩余样本的奇异度. 即 ICP 选取前 $m_k$ 个样本 $z_1, z_2, \cdots, z_{m_k}$ 构建规则 $D$, 剩余的 $n-m_k$ 个样本的奇异值 $\alpha_{m_k+1}, \alpha_{m_k+2}, \cdots, \alpha_n$ 计算方法为

$$\alpha_i = A_{m_k+1}(z_i, [z_1, z_2, \cdots, z_{m_k}]);\quad i = m_{k+1}, m_{k+2}, \cdots, n-1;$$
$$\alpha_n = A_{m_k+1}(z_n, [z_1, z_2, \cdots, z_{m_k}]),$$

式中: $[\cdot]$ 为数据包, 其包含的元素可以任意交换位置; $m_k\ (k = 1, 2, \cdots, \infty)$ 为训练规模且满足 $m_k < n < m_{k+1}$. 随着学习样本量不断增大, ICP 相应地扩大训练规模的值, 即 $m_1, m_2, \cdots$ 是严格递增的有限或无限正整数序列. 由于 ICP 只计算了 $n-m_k$ 个奇异值, 因此在利用式(1)计算 $P$ 值时, 分母 $n$ 相应改成 $n-m_k$.

### 1.2　HCCP 算法原理

归纳式相符预测器的算法思想为本文的 HCCP 提供了参考, 但是 HCCP 弥补了 ICP 的不足. HCCP 由压缩模型 $M$(类似 ICP 的通用规则 $D$)和剩余的 $n-m_k$ 个样本共同参与计算 $z_{m_k+1}, z_{m_k+2}, \cdots, z_n$ 的奇异值, 即

$$\alpha_j = A_n(z_j, [z_{m_k+1}, z_{m_k+2}, \cdots, z_j, z_{j+1}, \cdots, z_n], M),$$
$$j = m_k + 1, m_k + 2, \cdots, n-1,$$
$$\alpha_n = A_n(z_n, [z_{m_k+1}, z_{m_k+2}, \cdots, z_{n-1}], M).$$

这时 HCCP 对 $x_n$ 的预测结果为

$$\Gamma^{\varepsilon}_{M}(z_1, z_2, \cdots, z_{n-1}, x_n, \tau_n, m_k) = \{y \in Y: p_y = |\{j = m_k + 1, m_k + 2, \cdots, n; \alpha_j > \alpha_n\}| + \tau_n |\{j = m_k + 1, m_k + 2, \cdots, n; \alpha_j = \alpha_n\}|/(n - m_k) > \varepsilon\}. \quad (2)$$

为了说明 CP, ICP 和 HCCP 的异同点, 图 1 给出 3 种算法的示意图. 由图 1 可以看出, 3 种算法最根本的区别在于计算样本奇异值的方法不同. 当对第 $n$ 个样本进行预测时, CP 利用所有 $n$ 个样本计算这个样本的奇异值; ICP 没有用到 $z_{m_k+1}, z_{m_k+2}, \cdots, z_{n-1}$ 的信息, 因此 ICP 的预测有效性略差; HCCP 将 $n$ 个样本的信息以两种方式提供给这个样本进行奇异值计算.

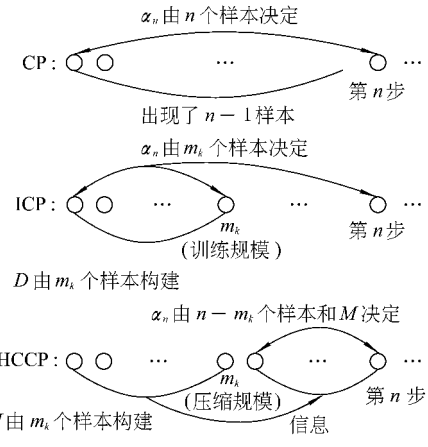


图 1　3 种不同的信息利用方式

### 1.3　HCCP 算法过程

经典 CP 算法将传统的机器学习方法引入 CP 的模型框架里, 来计算样本的奇异值. $K$-近邻算法($K$-Nearest Neighbors, $K$NN)是其中应用比较成功的例子[10]. 本文也将 $K$NN 引入 HCCP 模型框架里, 用于第二阶段的压缩过程, 这种算法命名为 HCCP-$K$NN.

HCCP-$K$NN 第一阶段的压缩过程为: 构建随机森林(RF)模型. RF 是一种性能优良的组合分类树算法, 它巧妙地为数据提供邻近度: 数据代入 RF 进行计算, 当每一棵树对数据进行判断时, 若有两个数据同时被分到这棵树的同一个终节点上, 则它们的邻近度为 1, 否则为零; 累计计算森林中所有树对这两个数据的邻近度, 结果除以树的数目, 即得这两个数据的邻近度. $N$ 个数据之

间的邻近度构成一个 $N \times N$ 矩阵, 记为 $(\boldsymbol{x}(i, j))_{N\times N}$. 当两个数据的邻近度被一棵树判断为 1 时, 意味着它们所有的属性中恰为这棵树的分裂属性的属性值全部相同或相近, 否则为零. 这种离散化取值方法使得样本在邻近度空间的差异间隔较大, 样本被有效区分开来, 类分开性大大提高. 基于 RF 的邻近度已应用在很多领域中.

HCCP-$K$ NN 第二阶段的压缩过程为: 利用 RF 将剩余样本 $z_{m_k+1}, z_{m_k+2}, \cdots, z_n$ 映射到邻近度空间, 在邻近度空间里计算样本奇异值, 即

$$\alpha_i = \sum_{j=1}^{k}(1 - x_{ij}^{y_i}) \Big\backslash \sum_{j=1}^{k}(1 - x_{ij}^{-y_i}), \qquad (3)$$

式中: $x_{ij}^{y_i}$ 表示样本 $i$ ($i = m_{k+1}, m_k+2, \cdots, n$) 与序号为 $m_k+1, m_k+2, \cdots, n$ 的样本中, 类别为 $y_i$ 的所有样本的邻近度里面第 $j$ 个最大的邻近度; $x_{ij}^{-y_i}$ 表示样本 $i$ 与序号为 $m_k+1, m_k+2, \cdots, n$ 的样本中, 类别不为 $y_i$ 的所有样本的邻近度里面第 $j$ 个最大的邻近度.

HCCP-$K$ NN 的算法过程描述如下.

输入: 学习样本 $((x_1, y_1), (x_2, y_2), \cdots, (x_{n-1}, y_{n-1}))$ 和待测数据 $x_n$, 其中 $((x_1, y_1), (x_2, y_2), \cdots, (x_{m_k}, y_{m_k}))$ 为训练样本集 $T$, $((x_{m_k+1}, y_{m_k+1}), (x_{m_k+2}, y_{m_k+2}), \cdots, (x_{n-1}, y_{n-1}))$ 为校验样本集 $V$.

输出: $x_n$ 的类别集合 $\Gamma_M^{\varepsilon}$.

**a** 利用 $((x_1, y_1), (x_2, y_2), \cdots, (x_{m_k}, y_{m_k}))$ 构建 RF 模型 $M$;

**b** 将数据 $(x_{m_k+1}, x_{m_k+2}, \cdots, x_n)$ 代入 $M$, 计算邻近度矩阵 $(\boldsymbol{x}(i, j))_{(n-m_k)\times(n-m_k)}$;

**c** 初始化预测集 $\Gamma_M^{\varepsilon} = \varnothing$ 和校验样本集

$V = ((x_{m_k+1}, y_{m_k+1}), (x_{m_k+2}, y_{m_k+2}), \cdots, (x_{n-1}, y_{n-1}))$;

**d** for $j = 1$ to $c$ do ($c$ 是样本的类别数), 指定 $j$ 是 $x_n$ 的假设类别; 利用式 (3) 计算 $((x_{m_k+1}, y_{m_k+1}), (x_{m_k+2}, y_{m_k+2}), \cdots, (x_{n-1}, y_{n-1}), (x_n, j))$ 各个样本的奇异值; 利用式 (2) 计算 $(x_n, j)$ 的 $p$ 值, 得到 $p_n^j$.

若 $p_n^j > \varepsilon$, 则 $\Gamma^{\varepsilon} \cup j$.

**e** 获得 $x_n$ 的真实类别 $y_n$, 扩充校验样本集 $V = V \cup (x_n, y_n)$, 扩充后的 $V$ 作为下一个测试数据的校验样本集.

在给定 $\varepsilon$ 后, 下列的指标用于计算 HCCP-$K$ NN 的预测效率: 确定率, 指含单个元素的预测集的比率; 不确定率, 指含有多于一个元素的预测集的比率; 空集率, 指空集的比率.

## 2 实验与讨论

### 2.1 实验设置

田纳西-伊斯曼化工过程 (TEP) 设有 52 个监测器, 用来检测生产过程的压强、温度等, 数据有 52 个属性. 文献 [2] 指出, 为了表现故障的分布特性, 当采样间隔设置为 3 min 时, 每类故障累积的样本量不能少于 480 个. 在本文实验中, 每类的采样样本量为 800 个 (采样间隔为 3 min); 数据共有 22 个类别, 包括 21 个故障类和一个正常类. 因此, 总共有 17 600 个样本用于下面的仿真实验.

训练样本集 $T$、校验样本集 $V$ ($V$ 可以初始化为空集)、测试样本集 $S$ 中的样本随机交换 10 次, 实验结果取平均值. 构建 RF 模型时, 参数设置尽量简单, 森林中树的数目为 1 000, 节点候选特征的个数为 $\lfloor\sqrt{52}\rfloor$, 这些设置在各种实验中都保持不变. 为了简化计算, 进一步提高算法的计算效率, $K$NN 的参数 $K$ 被设置为 1 (大部分 CP-$K$ NN 和 ICP-$K$ NN 算法[5, 6, 10] 都取 $K = 1$).

### 2.2 HCCP-$K$ NN 的校准能力分析

初始化 $T$ 的规模为 10 560, $V$ 为空集, $S$ 的规模为 7 040. 运用 HCCP-$K$ NN 算法进行在线学习, 在各种 $\varepsilon$ 给定的条件下, HCCP-$K$ NN 累积错误次数与测试数据数目的关系如图 2 所示.

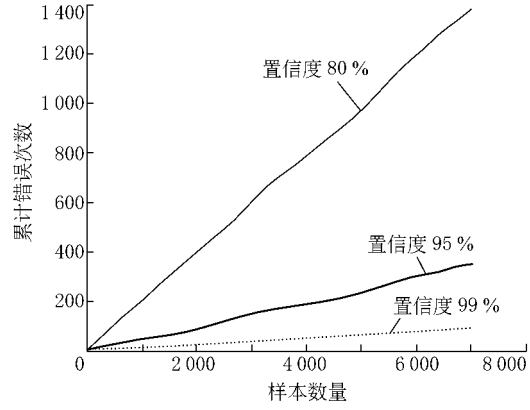


图 2 HCCP-$K$ NN 的校准性

从图 2 可以看出, 随着测试数据的增加, 错误次数线性增加. 直线的斜率等于对应的 $\varepsilon$, 即99 % 置信度对应的直线斜率为 1 %, 95 % 置信度对应的直线斜率为 5 %, 80 % 置信度对应的直线斜率为 20 %. 这说明 HCCP-$K$ NN 的预测置信度是有效的, 预测准确率对等于预先设置的置信度, 实现了预测风险的可控性.

### 2.3 $K$NN 比较

CP-$K$ NN 的学习样本规模为 10 560, ICP-

KNN 和 HCCP-*K*NN 具有相同的规模, 训练集为 6 000, 校验集为 4 560 的规模. 在各种 ε给定的条件下, 图 3 表示 3 种算法的确定率与置信度 1- ε的关系. 从图 3 可以看出, HCCP-*K*NN 的预测效率不仅比 ICP-*K*NN 好, 也比 CP-*K*NN 好, 它的确定率曲线非常接近对角线. 这说明大部分预测集包含的类别个数只有一个, 并且这个类别几乎都是真实类别. 预测集的规模越小, 算法的预测效率越高. HCCP-*K*NN 不仅充分利用了样本的信息, 把样本变换到邻近度空间, 而且样本分布更有利于分类, 使得分类准确率更高.

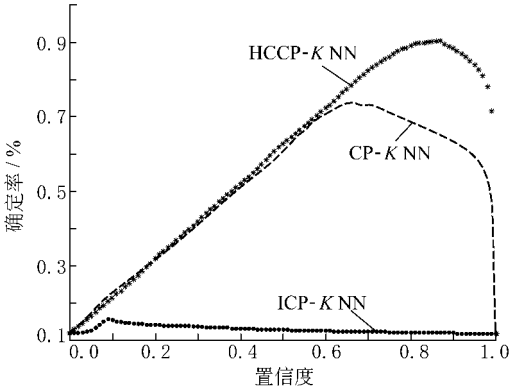


图 3　3 种算法的预测效率比较

HCCP-*K*NN 引入随机森林模型(RF) 从而提高了存储效率. 这种优势表现在 HCCP-*K*NN 的计算效率比 CP-*K*NN 高. 经典 CP-*K*NN 的计算复杂度为 $O(|Y|n(n-1)T_A(n,m))$, 其中: $Y$ 是类别的集合; $n$ 是学习样本的数目; $m$ 是数据属性的个数; $T_A(\cdot)$ 是计算 $p$ 值的复杂度. 对于 HCCP-*K*NN, RF 产生邻近度矩阵的复杂度 $O(T_M)$; 第二阶段的置信预测过程类似于经典 CP 算法, 其复杂度为 $O(|Y|S(S-1)T_A(V,m)S$ 为测试集规模. 因此 HCCP-*K*NN 复杂度为 $O(T_M)+O(|Y|S(S-1)T_A(S,m))$. 由于 $T_M$ 通过归纳推理进行计算, 计算速度快, 因此复杂度的第一项不占很大的比值, 而第二项由于测试集 $S$ 规模较小, 其值可以控制, 即通过适当增大训练样本 $T$ 的规模来提高 HCCP-*K*NN 的计算效率.

HCCP-*K*NN 的检测结果是由置信度控制的故障类别集合, 集合里的所有故障类型都可能出现. 这种多类别输出方式使得预测结果的信息量更大, 使用者具有更多的选择余地. 虽然有很多机器学习方法应用在TEP过程中, 但只有HCCP-*K*NN 算法允许事先控制算法出错的风险, 从而给用户带来极大的便利, 这也是被众多模式识别算法忽略的特性.

## 参　考　文　献

[1] Venkat V, Rengaswamy R, Yin, K, et al. A review of process fault detection and diagnosis, part ㈣: process history based methods[J]. Computers and Chemical Engineering, 2003, 27(1): 327-346.

[2] Lyman P R, Georgakis C. Plant-wide control of the Tennessee Eastman problem[J]. Computers and Chemical Engineering, 1995, 19(2): 321-331.

[3] Abhijit K, Jayaraman V K, Kulkarni B D. Knowledge incorporated support vector machines to detect faults in tennessee eastman process[J]. Computers & Chemical Engineering, 2005, 29(10): 2 128-2 133.

[4] Gammerman A, Vovk V. Prediction algorithms and confidence measures based on algorithmic randomness theory[J]. Theoretical Computer Science, 2002, 287(3): 209-217.

[5] Vovk V, Gammerman A, Shafer G. Algorithmic learning in a random world[M]. Berlin: Springer, 2005.

[6] Papadopoulos H. Qualified predictions for large data sets[D]. London: Computer Learning Research Centre, Royal Holloway, University of London, 2004.

[7] Papadopoulos H, Vovk V, Gammerman A. Conformal prediction with neural networks[C]// Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence. Washington, DC: IEEE Computer Society, 2007: 29-31.

[8] Vanderlooy S, Laurens V, Maaten D, et, al. An off-line learning with transductive confidence machines: an empirical evaluation, MICC-IKAT 07-03[R]. The Netherlands: Universiteit Maastricht, 2007.

[9] Bellotti T, Luo Z Y, Gammerman A, et al. Qualified predictions for microarray and proteomics pattern diagnostics with confidence machines[J]. International Journal of Neural Systems, 2005, 15(4): 247-258.

[10] Proedrou K, Nouretdinov I, Vovk V, et al. Transductive confidence machines for pattern recognition [C]// Proceedings of the 13th European Conference on Machine Learning. Berlin: Springer-Verlag, 2002: 381-390.

[11] Breiman L. Random forests[J]. Machine Learning, 2001, 45(3): 5-32.