文章编号: 1001-0920 (2013) 11-1667-07

# 一种面向演进数据流的结合相似准则和反例信息的分类方法

倪彤光[1,2], 王士同[1], 邓赵红[1], 王 骏[1]

(1. 江南大学 数字媒体学院，江苏 无锡 214122；2. 常州大学 信息科学与工程学院，江苏 常州 213164)

**摘 要:** 提出一种面向演进数据流数据的分类方法, 在有效利用相邻演进窗内数据间相似性信息的基础上, 通过引入反例信息, 构建一种面向演进数据流的增强型演进分类器优化目标函数, 从而推导出面向演进数据流的分类方法. 该方法在保有最大间隔原则和全局优化特性的同时, 充分考虑了反例信息对待解分类平面的影响. 在模拟和真实数据集上进行实验, 结果表明了所提出方法的有效性.

**关键词:** 演进数据流；分类；支持向量机；反例

**中图分类号:** TP391.4 **文献标志码:** A

# Classification method based on similarity criterion and counterexample information for evolutionary data streams

*NI Tong-guang*[1,2], *WANG Shi-tong*[1], *DENG Zhao-hong*[1], *WANG Jun*[1]

(1. School of Digital Media，Jiangnan University，Wuxi 214122，China；2. School of Information Science and Technology，Changzhou University，Changzhou 213164，China. Correspondent：NI Tong-guang，E-mail：hbxtntg-12@163.com)

**Abstract:** A classification method for evolutionary data streams is proposed. By utilizing the similarity criterion between the data distribution within the adjacent evolutionary windows and the related knowledge of counterexamples effectively, a enhanced objective function for the optimization problem is proposed. Meanwhile, the solution for the optimization problem is also derived. Both the maximal margin criterion in each evolution window and global optimization of the whole evolutionary data stream are considered, the counterexamples are also fully utilized. The new method learns decision hyperplanes successfully. The experiments on the artificial and real datasets demonstrate the effectiveness of the proposed method.

**Key words:** evolutionary data streams；classification；support vector machine；counterexample

## 0 引 言

演进数据流是指数据本身是不断发生演进变化的数据流[1-3], 近年来成为数据挖掘领域的研究热点. 根据变化发生的剧烈程度, 演进数据流中涉及的变化可分为突变和渐变两种形式[4]. 许多学者已提出了大量的算法与模型用以解决演进数据流的数据突变问题, 这些方法大致可以分成 3 类: 时间窗法、逐渐遗忘策略和组合法. 文献 [5] 提出用滑动窗 (SW) 方法为不同宽度的窗口建立不同的分类器, 并使用多个支持向量机来寻找数据流分类最佳时间间隔; 文献 [6] 提出渐进遗忘 (GF) 算法, 通过给样本点赋予不同权重来降低旧知识对当前分类器的影响; 文献 [7] 使用统计测试来预测演进数据流在下一时间片是否出现数据突变; 文献 [8] 提出基于最近邻的自适应调节分类方法来解决演进数据流中的数据突变漂移问题; 文献 [9] 使用集成分类器的方法分类演进数据流; 文献 [10] 利用自适应一对多分类决策树来解决同样的问题. 最近, 文献 [11] 提出了一种时间自适应的支持向量机 (TA-SVM) 模型来解决演进数据流分类时的数据渐变问题. 基于相邻时间窗的耦合特性, 该方法基于多分类器组合序列来构造新的目标学习准则, 以此达到寻找兼顾全局和局部最优的模式分割平面的目的. TA-SVM 可以看成 SVM 的变体[12-13], 其目标学习准则一方面考虑了每个时间窗口中的基分类器达到最优, 另一方面还从全局的角度考虑了相邻基分类器之间的耦合关系. 由于 TA-SVM 在局部优化与全局

**收稿日期:** 2012-07-17；**修回日期:** 2012-10-10.

**基金项目:** 国家杰出青年科学基金项目(60903100, 60975027)；江苏省自然科学基金项目(BK2009067).

**作者简介:** 倪彤光(1978−), 男, 讲师, 博士生, 从事模式识别、人工智能的研究；王士同(1964−), 男, 教授, 博士生导师, 从事模式识别、人工智能等研究.

优化之间进行了较好的平衡, 该方法对于演进数据流可取得比已有相关方法更好的效果.

遗憾的是, 已有的这些方法均未同时考虑演进数据流中存在的渐变和突变两种情况, 而在真实世界中事物的演进变化往往同时包含渐变和突变, 例如渐变可理解为事物的量变, 突变可理解为量变到达一定程度发生的质变. 以上述分析作为出发点, 本研究通过在目标学习准则中同时结合相似准则和反例信息, 提出一种面向演进数据流分类问题的新方法SCC-SVM(similarity criterion and counterexamples based SVM). 在SCC-SVM中, 相似准则对应于渐变情况, 而反例信息对应于突变情况. 针对演进数据流的演进变化特性, 在构造目标学习准则时, 加入相似性度量准则项的同时引入反例的相关知识, 从而得到一个全新的目标函数; 而后将其转化为二次规划(QP)对偶问题进行求解. SCC-SVM方法首次将数据流演进变化产生的反例信息引入到分类学习过程中. 理论分析和实验结果均证实了该方法的先进性和有效性.

## 1 TA-SVM

时间自适应支持向量机(TA-SVM)是近期广为关注的面向演进数据流的分类方法, 为了在后文中引出SCC-SVM方法, 下面将对TA-SVM作简要描述.

假设存在一个包含$N$个样本数据的演进数据流$T=[(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_N,y_N)]$. 其中: $\boldsymbol{x}_l,y_l$分别表示在第$l$时刻样本空间中的向量和分类标签, $y_l\in\{-1,1\}$, $\boldsymbol{x}$和$y$随时间发生演进变化. 数据集$T$分为$n$个连续的互不重叠的个子集$T_\mu$, $\mu=1,2,\cdots,n$, $n\leqslant N$, 对应$n$个演进窗. TA-SVM的目标是求解一个$(\boldsymbol{w},b)$的序列, 其中每个$(\boldsymbol{w}_\mu,b_\mu)$对应演进窗$T_\mu$的最优分类超平面, 则第$i$个样本对应的分类超平面由$(\boldsymbol{w}_{\mu_i},b_{\mu_i})$确定. 由文献[11]可知, 在综合考虑基分类器个体优化及基分类器间相互影响的基础上, 若采用线性SVM分类器作为基分类器, TA-SVM方法的目标函数如下:

$$
\begin{aligned}
&L(\boldsymbol{w}_\mu,b_\mu,\xi)=\\
&\min_{\boldsymbol{w}_\mu,b_\mu,\xi}\frac{1}{n}\sum_{\mu=1}^{n}\|\boldsymbol{w}_\mu\|^2+C\sum_{i=1}^{N}\xi_i+\\
&\frac{\gamma}{n-1}\sum_{\mu=1}^{n-1}\left(\|\boldsymbol{w}_\mu-\boldsymbol{w}_{\mu+1}\|^2+(b_\mu-b_{\mu+1})^2\right);\\
&\text{s.t.}\quad y_i(\boldsymbol{w}_{\mu_i}\boldsymbol{x}_i+b_{\mu_i})-1+\xi_i\geqslant 0,\\
&\qquad \xi_i\geqslant 0,\ \mu=1,2,\cdots,n,\ i=1,2,\cdots,N. \qquad (1)
\end{aligned}
$$

由式(1)可以看出, TA-SVM是在全部基分类器误差均值的基础上加上基分类器相似性度量准则项

$$\frac{\gamma}{n-1}\sum_{\mu=1}^{n-1}(\|\boldsymbol{w}_\mu-\boldsymbol{w}_{\mu+1}\|^2+(b_\mu-b_{\mu+1})^2),$$

引入该项的目的是兼顾每个基分类器可达到最优和使所有分类器的最优分割超平面的平均相似度达到最大化, 较好地解决了演进数据流中的渐变问题. 但是, 式(1)也在一定程度上说明了TA-SVM方法存在的缺陷, 即认为演进数据流概念的演进变化完全是渐进式发生的, 也就是说相邻演进时间窗的数据分布均具有相似性, 没有考虑数据发生突变的情况, 而演进数据流在演进变化过程中数据分布也可能呈现相反的变化趋势. 相邻演进时间窗数据分布的突变在一定程度上也能指导分类学习, 比如演进时间窗$T_\nu$中的数据分布存在很大概率不会与前一相邻时间窗$T_\mu$内数据的分布相似, 本文称这种可利用的知识为反例信息. 为了能有效地利用这类反例信息提升算法的有效性, 本文将在后文提出相关的新方法.

## 2 结合相似准则与反例信息的演进数据流支持向量机(SCC-SVM)

针对已有方法中反例信息往往被忽视的情况, 本文在保有TA-SVM方法中的相似性度量准则项的基础上, 加入了反例的相关归纳知识. 通过在准则中融入反例信息差异准则项, 提出了一种结合相似准则和反例的演进数据流分类新方法(SCC-SVM). 该方法能在确定决策超平面时不但可在每个演进窗内基于最大间隔原理, 而且兼顾到整个演进数据流上相似性准则信息和反例信息对样本分类的指导作用.

### 2.1 结合相似准则和反例信息的目标函数构造

假定存在演进数据流$T$, 由$n$个按时间顺序采集的子数据集$U_\mu$组成, $\mu\in 1,2,\cdots,n$, 对应$n$个演进窗. 第$\mu$个子数据集$U_\mu$中的数据点所对应的下标记为$i,\mu_i=\mu$. $U$为$n\times N$矩阵, 用于标示样本属于哪个演进窗, 如若第$j$个点属于第$\mu$个演进窗, 则$j\in U_\mu$, $U_{\mu_j}=1$, 其余取值为0. 若相邻两个演进时间窗内数据分布$h_i$和$h_j$之间的变化是突然地, 则称$h_j$为$h_i$的反例. 从反相似角度方面得到了描述数据集演进变化的可用知识信息, 利用这些反例信息结合相似准则也就达到提高分类器精度的目的. 一般情况下, 演进数据流整体上反例数目$m$小于$n$.

由于TA-SVM的目标函数表达式(1)仅考虑了基分类器最优分类超平面间的相似情况, 本文为了综合考虑演进数据流中渐变和突变情况, 给出如下SCC-SVM原始优化问题:

$$
\begin{aligned}
&L(\boldsymbol{w}_\mu,\boldsymbol{w}_\nu,b_\mu,b_\nu,\xi)=\\
&\min_{\boldsymbol{w}_\mu,\boldsymbol{w}_\nu,b_\mu,b_\nu,\xi}\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{w}_\mu\|^2+C\sum_{i=1}^{N}\xi_i+
\end{aligned}
$$

$$\frac{\eta}{n-m-1}\sum_{\mu\in N_+}(\|\boldsymbol{w}_\mu-\boldsymbol{w}_{\mu+1}\|^2+(b_\mu-b_{\mu+1})^2)-$$
$$\frac{\delta}{m-1}\sum_{v\in N_-}(\|\boldsymbol{w}_\nu-\boldsymbol{w}_{\nu+1}\|^2+(b_\nu-b_{\nu+1})^2);$$
$$\text{s.t.}\quad y_i(\boldsymbol{w}_{\mu_i}\boldsymbol{x}_i+b_{\mu_i})-1+\xi_i\geqslant 0,$$
$$\xi_i\geqslant 0,\ i=1,2,\cdots N. \tag{2}$$

其中: $N_+$ 为相邻演进窗数据分布为相似关系的演进窗集合; 而 $N_-$ 为相邻演进窗数据分布为反例关系的演进窗集合; $\eta$ 和 $\delta$ 为调节权重, $\eta>0,\ \delta>0$.

下面对式 (2) 作如下分析和说明.

1) 目标函数 (2) 后两项分别代表数据的相似性准则项和反例差异准则项. 它们将整个演进分类器序列耦合串联起来, 使每个演进窗内支持向量机的求解都依赖于其他的支持向量机.

2) $\eta$ 控制渐变演进窗口中基分类器间的相似程度对整体演进数据流的影响. 如果 $\eta$ 很小, 则会削弱相似集分类器序列间的相似性; 如果 $\eta$ 值很大, 则效果相反, 将产生一个几乎相同的支持向量机序列. $\delta$ 控制反例差异信息对分类效果的影响程度. 通过调节 $\eta$ 和 $\delta$ 的值, 可很好地反映普遍存在于演进数据流中的渐变和突变的特性.

3) 合适的 $\eta$ 和 $\delta$ 的值可由交叉验证策略来估计.

4) 在支持向量机目标函数中引用耦合项, 若对于数据流是单样本到达的情况, 也就是 $n=N$ 时, 此公式仍然有效. $N/n$ 相应代表分类器的个数, 通过调整这个比值可调整算法的复杂度.

借鉴文献 [11] 的求解策略, 对于给定演进数据流 $T$, 结合本节相似性准则和反例信息的描述, 定义基分类器分布矩阵 $R=[R_{\mu\nu}]_{n\times n}$, 记

$$R_{\mu\nu}=\begin{cases}\eta,\ \nu \text{ is in the similar to } \mu;\\ -\delta,\ \nu \text{ is the counterexample of } \mu;\\ 0,\ \text{otherwise.}\end{cases}$$

则 SCC-SVM 的原始优化问题 (2) 可表示为

$$L(\boldsymbol{w}_\mu,b_\mu,\xi)=$$
$$\min_{\boldsymbol{w}_\mu,b_\mu,\xi}\frac{1}{2n}\sum_{\mu=1}^{n}\Big(\frac{1}{2}\sum_{\nu=1}^{n}R_{\mu\nu}(\|\boldsymbol{w}_\mu-\boldsymbol{w}_\nu\|^2+$$
$$(b_\mu-b_\nu)^2)+\|\boldsymbol{w}_\mu\|^2\Big)+C\sum_{i=1}^{N}\xi_i;$$
$$\text{s.t.}\quad y_i(\boldsymbol{w}_{\mu_i}\boldsymbol{x}_i+b_{\mu_i})-1+\xi_i\geqslant 0,$$
$$\xi_i\geqslant 0,\ i=1,2,\cdots,N. \tag{3}$$

SCC-SVM 的原始优化问题 (3) 与文献 [11] 中的 TA-SVM 方法具有相似的描述形式, 根据对偶优化理论, 可给出如下对偶优化问题.

**定理 1**　SCC-SVM 的原始优化问题 (3) 的对偶问题为

$$L(\boldsymbol{\beta})=\max_{\boldsymbol{\beta}}-\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\beta}-\mathbf{1}^{\mathrm{T}}\boldsymbol{\beta};$$
$$\text{s.t.}\ \sum_{i=1}^{N}y_i\beta_i=0,\ 0\leqslant\beta_i\leqslant C. \tag{4}$$

其中

$$\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_N)^{\mathrm{T}},\ \mathbf{1}=(1_1,1_2,\cdots,1_N)^{\mathrm{T}}.$$
$$\boldsymbol{Q}=(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{U})\odot\boldsymbol{K}+(\boldsymbol{U}^{\mathrm{T}}(\boldsymbol{S}-\boldsymbol{I}/n)^{+}\boldsymbol{U})\odot\boldsymbol{Y}.$$
$$\boldsymbol{U}=[U_{\mu_j}]_{n\times N},U_{\mu_j}=\begin{cases}1,\ j\in U_{\mu_j};\\ 0,\ \text{otherwise.}\end{cases}$$
$$\boldsymbol{S}=[S_{\mu\nu}]_{n\times n},S_{\mu\nu}=\begin{cases}(1+\sum_{\kappa}R_{\mu\kappa})/n,\ \mu=\nu;\\ -R_{\mu\nu}/n,\ \text{otherwise.}\end{cases}$$
$$\boldsymbol{K}=[K_{ij}],\ K_{ij}=y_iy_j\boldsymbol{x}_i\boldsymbol{x}_j,\ \boldsymbol{Y}=\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}.$$

$\odot$ 表示 Hadamard 矩阵乘积.

**证明**　式 (3) 对应的拉格朗日函数为

$$L=$$
$$\frac{1}{2n}\sum_{\mu=1}^{n}\Big(\|\boldsymbol{w}_\mu\|^2+\frac{1}{2}\sum_{\nu=1}^{n}R_{\mu\nu}(\|\boldsymbol{w}_\mu-\boldsymbol{w}_\nu\|^2+(b_\mu-b_\nu)^2)\Big)-$$
$$\sum_{i=1}^{N}\beta_i(y_i(\boldsymbol{w}_{\mu_i}\boldsymbol{x}_i+b_{\mu_i})-1+\xi_i)+C\sum_{i=1}^{N}\xi_i-\sum_{i=1}^{N}\gamma_i\xi_i. \tag{5}$$

其中: $\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_N),\boldsymbol{\gamma}=(\gamma_1,\gamma_2,\cdots,\gamma_N)$ 为拉格朗日系数.

根据 Karush-Kuhn-Tucker(KKT)[14] 条件:

$$\partial L/\partial\xi_i=0, \tag{6}$$
$$\partial L/\partial\boldsymbol{w}_\mu=0, \tag{7}$$
$$\partial L/\partial b_\mu=0, \tag{8}$$

定义基分类器加权矩阵 $\boldsymbol{S}=[S_{\mu\nu}]_{n\times n}$, 由式 (7) 可得

$$\boldsymbol{w}_\mu=\sum_{j\in U_\mu}S^{-1}_{\mu\mu_j}\beta_jy_j\boldsymbol{x}_j. \tag{9}$$

定义核矩阵 $\boldsymbol{K}$, 可进一步得到

$$\sum_{\mu=1}^{n}\|\boldsymbol{w}_\mu\|^2=\boldsymbol{\beta}^{\mathrm{T}}((\boldsymbol{U}^{\mathrm{T}}\boldsymbol{S}^{-2}\boldsymbol{U})\odot\boldsymbol{K})\boldsymbol{\beta}. \tag{10}$$

令 $\boldsymbol{D}$ 为 $n\times n$ 的对角矩阵, 记

$$D_{\mu\nu}=\begin{cases}\sum_{\kappa}R_{\mu\kappa},\ \mu=\nu;\\ 0,\ \text{otherwise.}\end{cases}$$

则有

$$\sum_{\mu=1}^{n}\sum_{\nu=1}^{n}R_{\mu\nu}\|\boldsymbol{w}_\mu-\boldsymbol{w}_\nu\|^2=$$
$$2\boldsymbol{\beta}^{\mathrm{T}}((\boldsymbol{U}^{\mathrm{T}}\boldsymbol{S}^{-1}(\boldsymbol{D}-\boldsymbol{R})\boldsymbol{S}^{-1}\boldsymbol{U})\odot\boldsymbol{K})\boldsymbol{\beta}. \tag{11}$$

由式 (8) 可得

$$\Big(b_\mu \sum_{v=1}^{n} R_{\mu\nu} - \sum_{v=1}^{n} R_{\mu\nu} b_\nu\Big)\Big/ n = \sum_{i\in U_\mu} \beta_i y_i. \quad (12)$$

由于 $\boldsymbol{D}-\boldsymbol{R}$ 的行列式为 0, 矩阵 $\boldsymbol{D}-\boldsymbol{R}$ 奇异, 令 $h_i = \beta_i y_i$, 有如下关系成立:

$$\boldsymbol{b} = n(\boldsymbol{D}-\boldsymbol{R})^{+}\boldsymbol{U}\boldsymbol{h}. \quad (13)$$

利用此关系推导可得

$$\frac{1}{4n}\sum_{\mu=1}^{n}\sum_{\nu=1}^{n} R_{\mu\nu}(b_\mu - b_\nu)^2 - \sum_{i\in U_\mu}\beta_i y_i b_{\mu_i} = -\frac{n}{2}\boldsymbol{\beta}^{\mathrm{T}}((\boldsymbol{U}^{\mathrm{T}}(\boldsymbol{D}-\boldsymbol{R})^{+}\boldsymbol{U})\odot\boldsymbol{Y})\boldsymbol{\beta}, \quad (14)$$

其中 $(\cdot)^{+}$ 表示广义逆矩阵.

$\boldsymbol{D}, \boldsymbol{S}$ 和 $\boldsymbol{R}$ 之间满足如下关系:

$$\boldsymbol{S} = (\boldsymbol{I} + (\boldsymbol{D}-\boldsymbol{R}))/n, \quad (15)$$

其中 $\boldsymbol{I}$ 为单位阵.

将式 (6), (10), (11), (14) 和 (15) 代入 (5), 经化简后可得

$$L = -\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}((\boldsymbol{U}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{U})\odot\boldsymbol{K} + (\boldsymbol{U}^{\mathrm{T}}(\boldsymbol{S}-\boldsymbol{I}/n)^{+}\boldsymbol{U})\odot\boldsymbol{Y})\boldsymbol{\beta} + \sum_{i=1}^{N}\beta_i. \quad (16)$$

令新核矩阵 $\boldsymbol{Q} = (\boldsymbol{U}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{U})\odot\boldsymbol{K} + (\boldsymbol{U}^{\mathrm{T}}(\boldsymbol{S}-\boldsymbol{I}/n)^{+}\boldsymbol{U})\odot\boldsymbol{Y}$, 将其代入式 (15), 由此定理得证. □

### 2.2 算法描述

根据前节推导所得到的定理, 这里给出如下 SCC-SVM 分类器算法:

输入: 包含 $n+1$ 个子集演进数据集 $U$, 对应 $n+1$ 个演进时间窗, 前 $n$ 个子集数据构成训练集 $\mathrm{Train}_n$, 第 $n+1$ 个子集数据为待分类数据 $\mathrm{Test}_{n+1}$;

输出: 分类决策函数 $g_{n+1}(\boldsymbol{x})$;

预处理: 考察演进数据集, 确定相似集和反例集.

Step 1: 构造 $\mathrm{Train}_n$ 的数据分布情况矩阵 $\boldsymbol{U}$;

Step 2: 构造 $\mathrm{Train}_n$ 中 $n$ 个基分类器加权矩阵 $\boldsymbol{S}$;

Step 3: 选择参数 $\eta$ 和 $\delta$, 根据定理 1 求解拉格朗日系数 $\beta$;

Step 4: 根据式 (9) 和 (12), 计算决策超平面法向量序列 $(\boldsymbol{w}_i, b_i), i = 1, 2, \cdots, n$;

Step 5: 取序列中最后一个分类器参数 $(\boldsymbol{w}_n, b_n)$, 输出 $\mathrm{Test}_{n+1}$ 上的分类函数 $g_{n+1}(\boldsymbol{x}) = \boldsymbol{w}_n^{\mathrm{T}}\boldsymbol{x} + b_n$.

在时间复杂度方面, 该方法主要计算量集中在求解新的核矩阵 $\boldsymbol{Q}$ 和二次规划求极值问题上. 二次优化问题的时间复杂度是 $O(N^2)$, $N$ 为样本点个数. 因为本文仅考虑最近两个相邻的演进分类器之间的相互作用, 且核矩阵 $\boldsymbol{Q}$ 中

$$(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{U})_{ij} = \boldsymbol{S}^{-1}_{\mu_i\mu_j},$$
$$(\boldsymbol{U}^{\mathrm{T}}(\boldsymbol{S}-\boldsymbol{I}/n)^{+}\boldsymbol{U})_{ij} = (\boldsymbol{S}-\boldsymbol{I}/n)^{+}_{\mu_i\mu_j},$$

参照文献 [11] 中的相关结论可知, $\boldsymbol{Q}$ 矩阵求解的时间复杂度也为 $O(N^2)$.

## 3 讨 论

### 3.1 SCC-SVM 与 SVM 的关系

若演进数据流中相邻时间窗样本分布情况均是相同的, 即当 $\|\boldsymbol{w}_\mu - \boldsymbol{w}_{\mu+1}\|^2 + (b_\mu - b_{\mu+1}^2) = 0$ 时, 式 (2) 变为

$$L(\boldsymbol{w}_\mu, b_\mu, \xi) = \min_{\mathbf{w}_\mu, b_\mu, \xi} \frac{1}{n}\sum_{\mu=1}^{n}\|\boldsymbol{w}_\mu\|^2 + C\sum_{i=1}^{N}\xi_i;$$
$$\text{s.t.}\quad y_i(\boldsymbol{w}_{\mu_i}\boldsymbol{x}_i + b_{\mu_i}) - 1 + \xi_i \geqslant 0,$$
$$\xi_i \geqslant 0,\ \mu = 1, 2, \cdots, n,\ i = 1, 2, \cdots, N. \quad (17)$$

式 (27) 对应的拉格朗日问题为

$$L(\boldsymbol{w}_\mu, b_\mu, \xi) = \frac{1}{2n}\sum_{\mu=1}^{n}\|\boldsymbol{w}_\mu\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i(y_i(\boldsymbol{w}_{\mu_i}\boldsymbol{x}_i + b_{\mu_i}) - 1 + \xi_i) - \sum_{i=1}^{N}\beta\xi_i. \quad (18)$$

通过求解对应的优化问题可得

$$\boldsymbol{w}_\mu = \sum_{j\in U_\mu}\alpha_j^* y_j \boldsymbol{x}_j, \quad (19)$$

$$b_\mu{}^* = y_k - \sum_{j\in U_\mu}\alpha_j^* y_j(\boldsymbol{x}_j\boldsymbol{x}_k), \forall k \in \{k|\alpha_k^* > 0\}. \quad (20)$$

由此可见, 原目标学习准则可分解为 $n$ 个独立支持向量机之和的形式.

### 3.2 SCC-SVM 与 TA-SVM 的关系

若演进数据流在所考察区间内仅出现缓慢变化情况, 没有突变情况出现, 即相邻演进时间窗内分类器间的分布都是相似的, 反例情况还未发生的情况下, 则相似度矩阵 $\boldsymbol{R}$ 变为

$$R_{\mu\nu} = \begin{cases} \eta, & \nu \text{ is in the similar to } \mu; \\ 0, & \text{otherwise.} \end{cases}$$

将其代入式 (2) 就等同于 TA-SVM 所适应的场景情况. 对于一般演进数据流, 如果出现反例情况, 则 TA-SVM 算法不能很好地解决这个问题, 详情请参见第 4 节的实验部分. TA-SVM 算法可以看成本文所提出方法的一个特例.

## 4 实验与分析

为了验证本文方法在解决演进数据流时更为有效, 本文将利用 SCC-SVM 方法分别在人造数据集 (STAGGER 数据集, 旋转超平面数据集 Rotating HyperPlane) 与真实数据集 (Electricity Price Dataset) 上进行测试, 并与相应的算法进行比较和分析. 通过测试人造数据集来说明本文方法在抉择分类函数过程中所依据的基本原理和方法. 另外, 利用测试真

实数据集表明本文方法作为一种解决演进数据流的新方法的有效性.

在对SCC-SVM的实验研究中主要引入SW-SVM[5]和TA-SVM[11]两类算法进行比较: 利用与SW-SVM的比较来说明所提出方法相对固定窗宽的滑动窗支持向量机在应对演进数据流演进过程中数据发生变化后的应对能力较强; 通过与TA-SVM的比较来说明本文方法在结合反例信息后对分类器精度提高方面的有益作用.

如无特别说明, 所有实验均通过网格搜索的方式来确定优化的实验参数. 实验中采用线性核和高斯核两种核函数, 如果采用高斯核 $k(\boldsymbol{x},\boldsymbol{x}_i)=\exp(-\|x-x_i\|^2/2\sigma^2)$, 则核函数中的 $2\sigma^2$ 选择以源领域样本的平均2范数的平方 $s$ 为基准, 并在网格 $\{s/64, s/32, s/16, s/8, s/4, s/2, s, 2s, 4s, 8s, 16s, 32s, 64s\}$ 中搜索直至最优; SW-SVM, TA-SVM 和 SCC-SVM 的正则化参数在网格$\{2^{-2},2^{-1},2^0,2^1,2^2,2^3,2^4,2^5,2^6,2^7,2^8,2^9,2^{10}\}$中搜索直至最优; TA-SVM 中分类超平面相似性权重调节参数 $\gamma$ 在集合 $\{10^2,10^3,10^4,10^5,10^6,10^7\}$ 中选取[6]; 对于SCC-SVM, 相似性准则调节参数 $\eta$ 和反例调节参数 $\delta$ 均在集合 $\{10^1,10^2,10^3,10^4,10^5,10^6,10^7\}$ 中. 实验重复20次, 取其最佳精度作为算法实验结果. 所有实验均在Intel Core2, 1.6 GHz主频, 3G RAM, Windows XP系统, Matlab 2009a的平台上实现.

### 4.1　测试人造数据集

通过利用两种不同的验证策略(突变型、缓慢变化和突变型共有)来说明本文方法的基本原理及有效性分析.

#### 4.1.1　测试Stagger数据集

本文采用概念漂移问题经典数据集Stagger对SCC-SVM进行了性能分析. Stagger数据集的实例空间由如下3个属性描述:

$$\text{size}=\{\text{small},\text{medium},\text{large}\},$$

$$\text{color}=\{\text{red},\text{green},\text{blue}\},$$

$$\text{shape}=\{\text{square},\text{circular},\text{triangular}\}.$$

数据是3个属性27个值的随机组合, 类别标签class $\in\{-1,+1\}$. 另外3个目标概念采用如下形式:

1) size = small 并且 color = red;

2) color = green 或 shape = circular;

3) size = (medium 或 large).

从上面关于Stagger数据集的分析可以看出, Stagger数据集是一种突变型数据集. 本文实验中随机产生120个训练实例, 根据当前的概念给每个实例分配一个类别, 每40个训练实例同属于一个概念. 模拟演进数据经过40步的渐进变化然后发生了突变. 在每个时间步, 分类器从一个实例学习知识, 并且对包含100个实例的测试集进行预测准确性测试. 测试实例也是根据当前概念随机产生. 在同一概念内数据间是符合相似性准则的, 概念变化后时间步的数据可以看作之前时间步中数据的反例. 实验中TA-SVM和SCC-SVM算法的基分类器个数和训练样本数相等, 即$n=N$. SW-SVM算法采用固定长度的滑动窗, $t=5$.

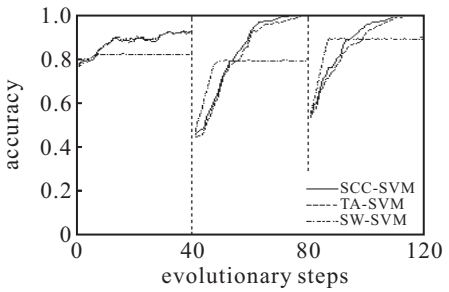



**图 1**　Stagger数据集上的平均精度比较

由图1可知, 40步以前由于演进数据流未发生突变, 也就是演进窗间的样本数据分布未出现反例情况, 本文方法和TA-SVM算法的精度相比十分相近. 在40时刻后由于出现了反例情况, 分类精度骤然下降; 但随着时间的推移, 本文方法SCC-SVM和TA-SVM的精确度均迅速回升. 然而, 由于本文方法充分考虑到了演进数据流的特性, 特别是学习了反例的知识, 相比TA-SVM在精度提升方面更为迅速, 并更快趋于稳定. SW-SVM算法里滑动窗的宽度较小, 因此对演进数据流中出现反例情况的应对能力较强, 反映较快; 但由于在小窗口下预测所需的知识不充分, 导致其精度率并不高.

#### 4.1.2　旋转超平面

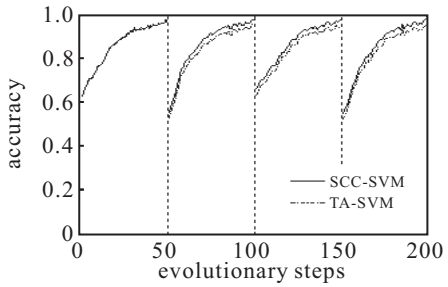
为了验证本文方法SCC-SVM的有效性, 下面进一步利用一个兼有渐变与突变特性的旋转超平面数据集来进行定性分析. 旋转超平面数据集是均匀分布在 $d$ 维的超立方体 $[-1,1]^d$ 区间上, 超平面满足 $\sum_{i=1}^{d}\boldsymbol{w}(j)\boldsymbol{x}_i=0$, $\boldsymbol{x}_i$ 为向量 $\boldsymbol{x}$ 的第 $i$ 维坐标. 在本次实验中部分数据集的生成皆遵循如下规则生成:

$$\begin{cases} w_1(j)=\cos(2\pi j/500)/10, \\ w_2(j)=\sin(2\pi j/500), \\ w_{3,4,\cdots,d}(j)=0. \end{cases}$$

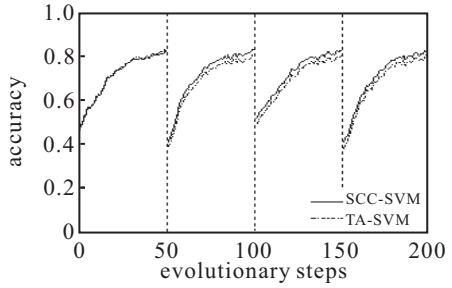
如果 $\sum_{i=1}^{d}\boldsymbol{w}(j)\boldsymbol{x}_i\geqslant 0$, 则标记为正样例, 反之标记为负样例. 实验中取 $d=3$, 并随机生成4种不同的权重集合, 根据上述的描述此数据集的反例个数是3, 其余均符合相似性准则. 这里以两个角度设计实验并进行了实验分析, 分别叙述如下.

**实验 1**　为了考察所提方法SCC-SVM的分类准

确度和抗噪性, 本实验使用相同的规则生成两个Rotating Hyperplane数据流, 每个数据流1 000个数据, 其中包含4个概念, 3次数据概念突变, 每个概念内250个数据. 将数据集分为50个小块, 每小块5个样本, 也就是$N/n = 5$演进窗宽度, 进行的是5次的渐变学习. 其中一个数据集加10%的噪声, 即随机10%的样本用错误标签替换正确的标签. 在每个时间步都使用一个独立的验证集和100个数据的测试集.




(a) 不带噪声




(b) 带10%噪声

**图 2** Hyperplane数据集平均精度比较

由图2可知, 在移动超平面数据集上, 观察前50个演进窗, 由于数据流并无反例情况出现, 所提方法和TA-SVM方法的精度大致相同. 当反例的情况出现后, SCC-SVM比TA-SVM方法能够更快地捕捉到演进数据流中产生质变的信息, 并使精度得到快速提升. 在10%噪音的情况下, SCC-SVM算法整体比TA-SVM的精确度也要高一些.

**实验 2** 本实验主要考察$N/n$的值对实验结果的影响. 按照实验1的规则生成旋转超平面数据集和对应的验证集和测试集. 分别取$N/n = [1, 2, 5, 10, 15, 20]$, 测试SCC-SVM和TA-SVM在整个数据流上各个时刻的平均精度随$N/n$的值变化的趋势.

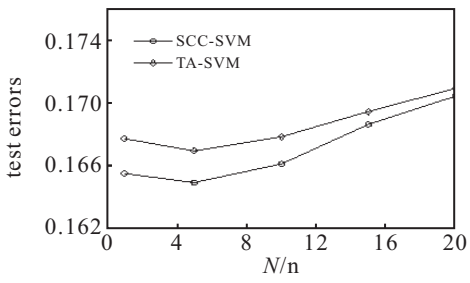



**图 3** 不同分类器个数下算法错误率比较

由图3可知, 当$N/n$=5时, SCC-SVM和TA-SVM算法都取得最高精度, 随$N/n$值增大, 两种算法都会下降. 这是因为随着演进窗包含的点增多, 演进分类器数目减少, SCC-SVM方法对相似和反例情况的矫正能力下降, 导致精度下降. TA-SVM算法也存在类似的问题.

### 4.2 真实数据集

为了在真实数据集上对本文算法的有效性进行分析与验证, 本文采用著名数据集Elec2[15]来测试算法在演进数据集上的分类效果. Elec2来源于真实反映的电力价格波动的数据, 因此并不知道漂移发生的确切时间和数量. 该数据集一共有45 312个条目, 是澳大利亚New South Wales电力供应商的1996年5月7日至1998年12月5日每半小时所采集的记录. 记录有9个属性: 前3个表示日期, 分别是日期(年月日)、星期几(1～7)和一天中的时间段(1～48); 后面5个个属性分别是New South Wales地区的电力需求、New South Wales地区的电价、Victoria地区的电力需求、Victoria地区的电价和两个地区间电力传输的数目; 最后1个是数据的类别, 分别表示电价的升高和下降.

本文采用与文献[15]一致的实验策略, 在实验中抽取Elec2中的5个属性, 分别是星期几(1～7)、一天中的时间段(1～48)、New South Wales地区的电力需求、Victoria地区的电力需求和两个地区间电力传输的数目. 随机截取15周的数据片段, 共5 040个样本. 每个演进分类器的数据长度是1周的电力采集数据, 也就是336个数据. 实验中使用测试集前面所有的数据作为训练集, 验证集是测试周的前一周数据.

由于真实数据的内在联系是未知的, 根据常识, 本文假设相邻的演进时间窗内要考察的数据部分大多数是相似的, 但由于气候、人为等原因, 必然会出现反例情况. 为了应用本文的SCC-SVM方法, 首先进行初始化, 对反例位置的预测. 根据此数据的特点, 这里采用了一个最简单策略, 就是用当前测试周的数据平均值和整个数据集的平均值相除, 如果大于某个阈值便可认为是反例.

用来测试和比较的方法及其平均精度和运行时间如表1所示.

**表 1** 不同方法的预测精度和运行时间比较

| 方法 | 采用的核函数 | 精度/% | 训练时间/s | 决策时间/s |
|---|---|---|---|---|
| SVM | | 63.5 | 32.15 | 1.25 |
| TA-SVM | Linear | 65.3 | 90.67 | 1.31 |
| SCC-SVM | | **66.2** | 98.09 | 1.33 |
| SVM | | 66.2 | 35.56 | 1.28 |
| TA-SVM | Gaussian | 68.9 | 210.31 | 1.32 |
| SCC-SVM | | **70.7** | 260.79 | 1.35 |

由表 1 可见, 采用线性核函数时, 本文算法 SCC-SVM 和 TA-SVM 都领先于传统的 SVM 算法. 在使用高斯核作为核函数时, 3 种方法都提高了分类精度, 本文方法由于引入了反例知识, 较之其余两种方法取得了更好的效果, 达到了 70% 以上, 这也再次验证了反例信息的重要性和可利用性.

## 5 结　论

本文针对演进数据流的分类问题, 通过在传统支持向量机的代价函数中加入相似性度量准则项和反例信息项, 形成新的差异惩罚函数, 构造一个新的演进数据流分类器, 求解时不仅在本身演进窗内达到最优, 而且考虑整个演进数据流的具体情况. 在本文方法中, $\eta$ 和 $\delta$ 这两个参数可分别调整分类器相似和反例对整体分类器序列求解时的影响. 若数据流分布改变非常频繁, 此时数据流中数据间的相似性准则几乎不起作用, 设置 $\eta$ 很小而 $\delta$ 很大, 则由数据之间存在的反例集来矫正整个支持向量机分类面序列的求解. 在极端情况下, 如果仅知道演进数据流中突变情况, 应用本文方法也能对分类面起到一定的矫正作用, 这将成为今后研究的重点. 另外, 如何使用更为有效的方法来获取演进数据流中相似信息与反例信息和在支持向量机中引入多核技术, 也是值得关注及研究的方向.

### 参考文献(References)


[1] Leite D, Costa P, Gomide F. Evolving granular neural network for semi-supervised data stream classification[C]. Int Joint Conf on Neural Networks. Barcelona, 2010: 1-8.

[2] Elwell R, Polikar R. Incremental learning of concept drift in nonstationary environments[J]. IEEE Trans on Neural Networks, 2011, 22(10): 1517-1531.

[3] Masud M M, Gao J, Khan L, et al. Classification and novel class detection in concept-drifting data streams under time constraints[J]. IEEE Trans on Knowledge and Data Engineering, 2011, 23(6): 859-874.

[4] 欧阳震诤. 不平稳数据流的分类技术研究[D]. 长沙: 国防科学技术大学 计算机学院, 2009.
(Ouyang Z Z. Research of classical technique for imbalance data stream[D]. Changcha: College of Computer，Defence Science University, 2009.)

[5] Klinkenberg R, Joachims T. Detecting concept drift with support vector machines[C]. The 17th Int Conf on Machine Learning. Sannateo, 2000: 487-494.

[6] Koychev I. Gradual forgetting for adaptation to concept drift[C]. Proc of ECAI Workshop Current Issues Spatio-Temporal Reason. Berlin, 2000: 101-106.

[7] Koychev I, Lothian R. Tracking drifting concepts by time window optimization[C]. The 25th SGAI Int Conf on Innov Technology Application Artif Intell. New York, 2005: 46-59.

[8] Alippi C, Roveri M. Just-in-time adaptive classiers in non-stationary conditions[C]. Proc of Int Joint Conf on Neural Netw. Orlando, 2007: 1014-1019.

[9] Gao J, Ding B, Han J, et al. Classifying data streams with skewed class distributions and concept drifts[J]. IEEE Internet Computing, 2008, 12(6): 37-49.

[10] Hashemi S, Yang Y, Mirzamomen Z, et al. Adapted one-versus-all decision trees for data stream classification[J]. IEEE Trans on Knowledge and Data Engineering, 2009, 21(5): 624-637.

[11] Guillermo L G, Lucas C U H. Alejandro Ceccatto, et al. Solving nonstationary classification problems with coupled support vector machines[J]. IEEE Trans on Knowledge on Neural Network, 2011, 22(1): 37-51.

[12] 刘忠宝, 王士同. 基于熵理论和核密度估计的最大间隔学习机[J]. 电子与信息学报, 2011, 33(9): 2187-2191.
(Liu Z B, Wang S T. A maximum argin learning machine based on entropy concept and kernel density estimation[J]. J of Electronics & Information Technology, 2011, 33(9): 2187-2191.)

[13] 陶剑文, 王士同. 具有磁场效应的大间隔支持向量机[J]. 电子与信息学报, 2011, 33(5): 1055-1061.
(Tao J W, Wang S T. Maximal margin support vector machine with magnetic field effect[J]. J of Electronics & Information Technology, 2011, 33(5): 1055-1061.)

[14] Scholkopf B, Herbrich R, Smola A J. A generalized representer theorem[C]. Proc of Computational Learning Theory’2001. Amsterdam: Springer Press, 2001: 416-426.

[15] Harries M. Splice-2 comparative evaluation: Electricity pricing[R]. Sydney: University of New South Wales, 1999.