

基于动态数据的支持向量分类机模型^{*})

原峰山^{1,2} 朱思铭¹

(中山大学数学与计算科学学院 广州 510275)¹ (广州航海高等专科学校网络中心 广州 510725)²

摘要 在支持向量分类机模型中,要从样本中选出具有代表性的数据作为输入值。一般地,选定的数据是静态的,但实际情况中数据是动态的。本文提出了基于动态数据的支持向量分类机模型,并给出了相应的证明和分析。

关键词 动态数据,支持向量机,分类机模型

Classified Support Vector Machines Model Based on Dynamic Data

YUAN Feng-Shan^{1,2} ZHU Si-Ming¹

(School of Mathematics and Computing Science, Zhongshan University, Guangzhou 510275)¹

(Guangzhou Maritime College, Guangzhou 510725)²

Abstract In model of classified Support Vector Machines, the typical data selected from samples can be used as import data which is generally static. But in fact, data is dynamic. In this paper, classified Support Vector Machines model based on dynamic data are raised.

Keywords Dynamic data, Support Vector machines, Classified model

1 引言

常规的支持向量机建模时,要从样本中选出具有代表性的数据作为输入值,依靠构造支持向量机建立系统输入与输出之间的映射关系,建立的模型一般不再变化,尤其是选定的数据更不会改变。这种模型简单、易用,并且数据处理也较为简单,我们不妨称这种模型为“静态支持向量机”。但是,在实际应用和工程中,很多情况下数据都是随着时间的变化而变化的,将这种模型应用于时变的系统时,往往就会与实际情况不相符,如图 1 所示。本例是以分类问题作为说明对象的。图中表示了两类将要分类的两类样本,其中有样本在一定范围内,即 $V_0 \leq V \leq V_1$,对于“静态支持向量机”,假设 $0 \leq t \leq T$ 区域为研究问题的区域,那么一旦模型建立起来,取值的范围也不会随着时间的变化而变化,它是一个固定的模式,当然所取的样本值也在 $0 \leq t \leq T$ 和 $V_0 \leq V \leq V_1$ 所框定的矩形内。在我们的实际问题中,这种模型显然是有问题的。

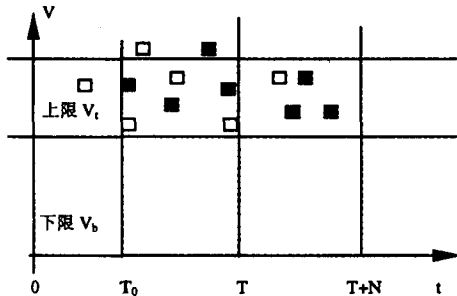


图 1

在一个被抽样的 SVM 系统中,有两种取值的可能性。第一种是:被测样本是连续信号,来自于不同的物理信号,如

温度、压力、流量、电流、电压…等。对于这种信号,可以依据抽样定理进行抽样,即:若连续信号 $x(t)$ 是有限带宽的,并且其频谱的最高频率为 f_c ,当抽样频率 f_s 满足: $f_s \geq 2f_c$ 时,抽出的样本 $x(nT_s)$ (注: T_s 为抽样周期) 可以恢复出 $x(t)$,即 $x(nT_s)$ 保留了 $x(t)$ 全部信息。因此,对连续信号,只要它是有限带宽的,就能将它变为离散的数据量。第二种情况,被测样本是离散值,数据在时间轴的离散点上取值,在此估且称其为离散时间信号,记为 $x(nT_s)$, T_s 含义同前, n 为正数,即:

$$x(nT_s), n = -N_1, \dots, -1, 0, 1, \dots, N_2$$

N_1, N_2 是 n 的取值范围,若做归一化处理,令 $T_s = 1$,则 $x(nT_s)$ 简记为 $x(n)$,它只是正数 n 的函数,又称为离散时间序列。

因此,实际问题中要分析的样本系统,无论是连续系统还是离散系统,都可以视为离散系统来处理。由于这类问题是面向支持向量机的,建立起来的系统属于与时间有关离散系统,我们将其称为基于动态支持向量机的离散时间系统。

2 动态数据支持向量机模型的建立

2.1 静态条件下基于线性决策规则的最优超平面

根据 Vapnik 的定义^[2],假定训练数据

$(x_1, y_1), \dots, (x_l, y_l), x \in R^l, y \in \{+1, -1\}$ 可以被一个超平面 $(w \cdot x) - b = 0$ 分开,如果这个向量集合被超平面没有错误地分开,并且离超平面最近的向量与超平面之间的距离是最大的,则这个向量的集合称为最优超平面,支撑构造这个最优超平面的向量集合为支持向量,形成的模型成为支持向量机。

2.2 动态支持向量分类机模型的建立

2.2.1 支持向量机一个处理单元模型的分析 根据支持向量机的模型,在这里抽出一个处理单元分析基于时间变

^{*})基金项目:国家自然科学基金资助项目(编号:10071097)。原峰山 高级工程师,博士生导师,研究方向:应用数学,人工智能与计算机网络。

朱思铭 教授,博士生导师,研究方向:人工智能与计算机网络。

化的 SVM。

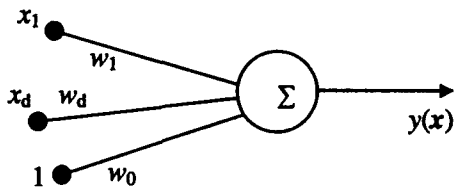


图 2

对于分类问题,这里选择线性决策函数来分类。用 w_0 表示整个决策的偏差,用向量 x 表示特征向量的集合,代表输入变量的特征,用符号 w 表示整个权重向量,那么输出变量可表示为一个线性函数,系统的整体就是线性网络:

$$y(x) = w \cdot x + w_0 \quad (1)$$

输出是对所有的输入乘以权重后进行累加求和。

对于分类的情况,有函数单元 $y_k(x)$ 和权重向量 w_k ,它们之间存在着——对应的关系,对每一个样本,满足:

$$y_k(x_i) = w_k \cdot x_i = \sum_{j=1}^n w_{k,j} x_{i,j} + w_0 \quad (2)$$

显然,对一个单元或者是一个系统而言,输入输出之间满足线性关系。当输入向量有可能随时间变化时,系统就是线性离散时间系统。

2.2.2 线性离散时间系统的性质^[1] 离散时间系统的输入输出关系可以简单地用图 3 表示。

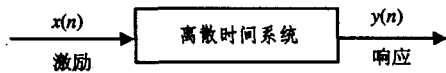


图 3

响应与激励之间的关系可表示为:

$$x(n) \rightarrow y(n) \quad (3)$$

本模型的基础是线性离散时间系统,当离散系统满足齐次性和叠加性时,即:对任意一组常数 c_k ($1 \leq k \leq K$),必定成立:

$$\{x_k(n) \rightarrow y_k(n), (1 \leq k \leq K)\} \Rightarrow \left\{ \sum_{k=1}^K c_k x_k(n) \rightarrow \sum_{k=1}^K c_k y_k(n) \right\} \quad (4)$$

式中“ $\Rightarrow$ ”符号左边表示前提,右边表示推论。

在同样起始状态下,系统响应特性与激励施加于系统的时刻无关时称为时不变离散时间系统。也就是说,若激励时延 T 时间,响应也时延相同的时间 T 。即对任意 T ,满足:

$$\{x(n) \rightarrow y(n)\} \Rightarrow \{x(n-T) \rightarrow y(n-T)\} \quad (5)$$

系统的稳定性:如果一个系统对任意有界激励 ($|x(n)| < \infty (\forall n)$) 零状态响应 ($|y(n)| < \infty (\forall n)$) 都是有界的,那么该系统就是“稳定系统”。

2.2.3 动态支持向量机系统分类模型的建立 根据构造广义最优超平面的概念,借助于支持向量机中分类的概念,惩罚因子 ξ_i 是超平面纠偏引起的。设已知训练集 $T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (X \times Y)^l$, 其中 $x_i \in X = R^d$, $y_i \in Y = \{1, -1\}$, $i = 1, \dots, l$, 当系统的优化问题被转化为与时间有关时,可在其最小化关系式中加入时间变量因子 h_i , 那么原始的最优化问题为:

最小化:

$$\begin{cases} \Phi(w, \xi) = \frac{1}{2} (w \cdot w) + C \sum_{i=1}^n h_i \xi_i \\ \text{约束条件为: } y_i [w \cdot x_i - b] \geq 1 - h_i \xi_i, \xi_i \geq 0, i = 1, \dots, n \end{cases} \quad (6)$$

其中, x_i 为特征向量, w 为权向量, C 为正规化参数,也是惩罚参数,满足 $C > 0$, ξ 为误差的松弛变量。 h_i 为与时间有关的误差权值,并且分离间隔为 1。

将上述问题的优化解由拉格朗日泛函的鞍点给出,则:

$$\begin{aligned} L(w, b, \xi, a, h) = & \frac{1}{2} w \cdot w + C \sum_{i=1}^n h_i \xi_i - \sum_{i=1}^n a_i [y_i (w \cdot x_i - b) - 1 + h_i \xi_i] - \sum_{i=1}^n h_i \xi_i \\ \text{其中, } & a_i \geq 0, h_i \geq 0. \end{aligned} \quad (7)$$

函数 $L(w, b, \xi, a, h)$ 应对 a_i 和 h_i 最大化,且对 w, b 和 ξ_i 最小化。因此函数 $L(w, b, \xi, a, h)$ 的极值应满足条件:

$$\frac{\partial}{\partial w} L(w, b, \xi, a, h) = 0, \text{ 得:}$$

$$w = \sum_{i=1}^n a_i y_i x_i \quad (8)$$

$$\frac{\partial}{\partial b} L(w, b, \xi, a, h) = 0, \text{ 得:}$$

$$\sum_{i=1}^n a_i y_i = 0 \quad (9)$$

$$\frac{\partial}{\partial \xi_i} L(w, b, \xi, a, h) = 0, \text{ 得:}$$

$$C - \sum_{i=1}^n a_i - 1 = 0 \quad (10)$$

$\frac{\partial}{\partial h_i} L(w, b, \xi, a, h) = 0$, 同样可得: $C - \sum_{i=1}^n a_i - 1 = 0$, 成为约束条件。

将式(8)代入式(7),可以得到优化问题的对偶形式,最小化函数为:

$$\begin{cases} W(a) = \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i,j=1}^n a_i a_j y_i y_j (x_i \cdot x_j) \\ \text{其约束条件为: } \sum_{i=1}^n a_i y_i = 0 \quad 0 \leq a_i \leq C, \\ i = 1, 2, \dots, l \end{cases} \quad (11)$$

这是一个典型的二次优化问题, C 成为重要的约束条件。

判别函数为:

$$f(x) = \text{sgn} \left[\sum_{\text{支持向量}} a_i y_i K(x_i, x) - b \right] \quad (12)$$

3 动态支持向量机系统模型的分析

对以上的模型可作如下分析:

(1) 从式(6)中可以看出,由于 $\Phi(w, \xi)$ 代表支持向量机中最大化的分离裕度,因此 h_i 对分离裕度有直接的影响。而由拉格朗日方法处理后,得出的优化问题的最小化函数(11)成为极值问题对应于原问题的的对偶,这种转化是支持向量机中常用的方法。可以看出,引入时间变量因子 h_i 对提出优化问题有影响,但没有对优化结果的公式有直接影响,将极大地方便利用原有的支持向量机理论解决基于离散时间系统动态支持向量机的问题。

(2) 式(7)中的 ξ_i 是一个非负的松弛变量,作为一个数据点与理想可分情况下偏差的一种惩罚。它的几何意义为:

对于落在决策区域线一侧但是又处于分离区域内的样本点, $\xi_i < 1$ 。对于落在错误决策内的样本点会加上更大的惩罚,此时 $\xi_i > 1$ 。

(3) 式(6)中 $C \sum_{i=1}^n h_i \xi_i$ 项,设 $h_i \xi_i = A$ 。 ξ_i 的作用除了上述的以外,它与 h_i 的乘积一起形成了实际意义上的惩罚因子总和 $\sum_{i=1}^n h_i \xi_i$, 最小化函数 $\Phi(w, \xi)$ 受 C 和 A 的共同影响,两者的影响是相反的。 $C \sum_{i=1}^n h_i \xi_i$ 式中,对于较小的 C 值, A 的影响较

大。由于 $h_i \xi_i = A$, 在惩罚性松弛因子 ξ_i 一定时, 加大 h_i 可提高 A 的影响。这就是说, 得到的解在最小化错误率时倾向于得到一个较小的超平面间隔。对于较大的 C 值, A 在最小化中的影响较小, 可以容忍较大的分类错误率, 从而使分类间隔较大。

(4) C 的影响: 有一部分影响 (2) 已经列出。从推导出的优化二次泛函表达式 (11) 中可以看出: 它与不加惩罚系数的结果是一样的, 唯一不同的是它应满足更加严格的条件限制: $\sum_{i=1}^l a_i y_i = 0, 0 \leq a_i \leq C$ 。因此, 被优化的泛函的取值受到了限制, 并且 C 成为优化问题的决定性因素之一。一般情况下, C 值的选择必须通过实验的方法来试, 有时可能具有不止一个最优值。

(5) 基于离散时间系统的动态支持向量机分析

支持向量机二次优化问题确定广义最优超平面系数的公式为:

$$w = \sum_{i=1}^l a_i y_i x_i \quad (13)$$

现在, 考虑离散时间系统输入和输出受时间影响, 故可以将上述表达式转化为:

$$\sum_{i=1}^l a_i w(n-t) = \sum_{i=1}^l a_i y_i x(n-t) \quad (14)$$

根据上式, 先作如下定义:

定义 1 系统的传递函数^[1]:

$\because y_i \in \{-1, 1\}$, 故可将上式变为 $\sum_{i=1}^l a_i w(n-t) = \sum_{i=0}^T b_i x(n-t)$, 对两边求 Z 变换, 并设 $x(n)$ 是因果序列, 且系统处于零状态响应, 则有:

$$W(z) \cdot \sum_{i=0}^T a_i z^{-i} = X(z) \cdot \sum_{i=0}^T b_i z^{-i} \quad (15)$$

$\therefore$ 定义 $H(z) = \frac{W(z)}{X(z)} = \frac{\sum_{i=0}^T b_i z^{-i}}{\sum_{i=0}^T a_i z^{-i}}$ 为离散系统的传递函数, 若对上式分子、分母进行因式分解, 有:

$$H(z) = G \cdot \frac{\prod_{i=1}^T (1 - z_i z^{-1})}{\prod_{i=1}^T (1 - p_i z^{-1})} \quad (16)$$

定义 2 系统的单位冲击响应^[1]: 指系统对单位冲激序列 $\delta(n)$ 的零状态响应, 记为 $h(n)$, 即

$$\delta(n) \rightarrow h(n) \quad (17)$$

对于离散时间系统, 可以知道, 有如下特性:

特性 1 系统的零状态响应等于激励与单位冲击响应之间的卷积。

特性 2 传递函数 $H(z)$ 与单位冲激响应 $h(n)$ 是一对 Z 变换对。

4 离散时间系统的动态支持向量机的几点结论

a) 离散时间系统的最优超平面及其分离裕度

若给定一个超平面, 最近的样本到它的距离称为分离裕度。在决策函数为线性的条件下, 当分离裕度 $\frac{2}{\|w\|}$ 为最大值时, 由支持向量确定的超平面为最优超平面。显然, 决定最优超平面的主要因素为 w , 而 w 与样本向量 x 的关系是离散

时间系统的输入输出关系。 x 的状态决定了 w , 也就决定了能否得到系统的稳定, 从而决定最优超平面的分离裕度能否稳定下来。

b) 获得稳定的分离裕度的充分必要条件

获得稳定的 w , 即能获得稳定的分离裕度。满足这一点的充分必要条件是: 传递函数 $H(z)$ 的逆 Z 变换“绝对可和”^[1], 即满足

$$\sum_{n=-\infty}^{\infty} |z^{-1}(H(z))| < \infty, \quad \text{或} \quad \sum_{n=-\infty}^{\infty} |h(n)| < \infty \quad (18)$$

证明, 先证充分性。

由特性 1 可知:

$w(n) = x(n) * h(n) = \sum_{k=-\infty}^{\infty} h(k)x(n-k)$, 由于系统的单位冲激响应绝对可和, 可设:

$$\sum_{k=-\infty}^{\infty} |h(k)| = A < \infty \quad (19)$$

任意给定一个有界激励 $x(n)$, 即 $|x(n)| < B < \infty (\forall n)$, 则

$$|w(n)| = |x(n) * h(n)| = \left| \sum_{k=-\infty}^{\infty} h(k)x(n-k) \right| \leq \sum_{k=-\infty}^{\infty} |h(k)| \cdot |x(n-k)| \leq B \sum_{k=-\infty}^{\infty} |h(k)| = AB < \infty \quad (20)$$

因此系统是稳定的。

再证必要性。已知系统是稳定的, 那么只要输入是有界的, 输出也必定是有界的。现在用反证法证明其单位冲激函数应绝对可和。假定系统的单位冲激响应不是绝对可和的, 则:

$$\sum_{k=-\infty}^{\infty} |h(k)| = \infty \quad (21)$$

取一个有界激励

$$x(n) = \begin{cases} 0, & h(-n) = 0 \\ \text{sgn}(h(-n)), & h(-n) \neq 0 \end{cases} \quad (22)$$

显然, 有 $|x(n)| \leq 1, \forall n$, 输入是有界的。现在求:

$$w(0) = (x * h)(0) = \sum_{k=-\infty}^{\infty} h(k)x(0-k) = \sum_{\substack{k=-\infty \\ h(k) \neq 0}}^{\infty} h(k)x(-k) \\ k) = \sum_{\substack{k=-\infty \\ h(k) \neq 0}}^{\infty} h(k)\text{sgn}(h(k)) = \sum_{\substack{k=-\infty \\ h(k) \neq 0}}^{\infty} |h(k)| = \sum_{k=-\infty}^{\infty} |h(k)| = \infty \quad (23)$$

即 $w(0)$ 无界。与假设矛盾, 故必要性成立。

因此, 在本文构造的基于离散时间系统的动态向量机中, 分离裕度最大时可构造最有超平面, 但是必须有一个“度”, 即有一个稳定的分离裕度才能保证系统的稳定性, 前提是传递函数 $H(z)$ 的逆 Z 变换是有界的, 或者说系统对单位冲激序列 $\delta(n)$ 的零状态响应 $h(n)$ 应该是有界的。

结束语 本文利用离散时间系统的理论用于构造数据样本受时间影响的支持向量机, 并作了一些基本分析, 得到了一些重要的结论, 无疑是对支持向量机理论在一定程度上的改进, 也是在此领域中的初次尝试。

参考文献

- 1 郑方, 徐明星. 信号处理原理. 北京: 清华大学出版社, 2000
- 2 张学工译. 统计学习理论的本质. 北京: 清华大学出版社, 2000
- 3 <http://www.csie.ntu.edu.tw/~cjlin/libsvm>