基于信息几何的卷烟焦油 SVM(支持相量机)预测

王德吉[1,2]，李广才[1]，栗卫军[1]

1 中国烟草总公司职工技术培训中心培训处，郑州市北林路 7 号 450008；

2 中科院合肥智能机械研究所，合肥市科学岛 1130 信箱 230031

摘 要：由于样本少且是非线性关系，卷烟配方中焦油量预测非常困难，为此引入能较好的解决小样本非线性预测问题的 SVM(支持向量机)进行卷烟焦油预测，其中的核函数选择是关系到预测精度的关键，而现有方法仅能试验试凑。本文从信息几何的角度，通过保角映射，给出核函数构造的一般方法，提高了预测精度和效率并通过试验证明该模型能更准确的预测卷烟焦油量。

关键词：SVM(支持向量机)；卷烟焦油量；信息几何

doi：10.3969/j.issn.1004-5708.2009.04.005

中图分类号：TS411 **文献标识码**：A **文章编号**：1004-5708(2009)04-0022-03

Cigarette tar delivery prediction by SVM based on geometric information

WANG De-ji[1,2], LI Guang-cai[1], LI Wei-jun[1]

1 Training Centre of CNTC, Zhengzhou 450008, China;

2 Institute of Intelligent Machines, Chinese Academy of Science, Hefei 230031, China

Abstract: Cigarette tar delivery is difficult to predict because of inadequate samples and nonlinear relationship among variables. Hence a SVM (Support Vector Machine) forecast method based on information geometry was proposed. In SVM based method, kernel function is very important to the prediction accuracy. A new method was given by conformal mapping from the perspective of geometric information, which can enhance accuracy and efficiency. Experiment was carried out to test the performance of the presented method. Result showed that the SVM based on information geometry can predict tar delivery better than currently used method.

Key words: support vector machine; cigarettes car content; information geometry

为控制成品卷烟的焦油量，在投产前，往往需要预测成品卷烟的焦油量，国内外少数厂家通过配方试验采集数据，然后利用多元回归的方法建立预测成品卷烟焦油量的数学模型[1-2]，但回归分析一般需要适当的大样本才能获得较为可靠的统计结果。由于成本等原因，样本数不足往往使普通回归分析方法难以奏效。本文提出从信息几何与 SVM(支持向量机)的角度分析建立预测模型，并以某烟厂 22 种常用的不同地区和等级的烟叶进行的 20 项配方试验所得的数据验证模型的正确性[3]。

1 支持向量机

支持向量机是 Vapnik 及其同事在 1992 年的 COLT 会议上首次提出的，针对有限样本预测，通过非线性映射将输入空间映射到一个高维特征空间，在这个空间中构造最优分类超平面的实现过程[4-6]。

设样本数据为 (x_i, y_i)，$i = 1, \ldots, n$，x_i 为输入向量，y_i 是期望值，根据 SVM 相关理论书籍可以得到预测公式如下：

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i \cdot x_j) + b \qquad (1)$$

α，α^* 是 Lagrange 算子，b 为偏置量，$K(x, x^*)$ 是内积核函数，它必须服从 Mercer 条件。它可隐含地将输入空间的数据映射到高维特征空间，通过非线性变换转化为某个高维空间的线性问题。但至今为止，核函数的选择都没有很好的指导方法，只能通过试凑，下面论文将从信息几何的角度给出一般的指导方法。

2 基于信息几何的核函数设计

信息几何是采用微分几何方法来研究统计学的理

作者简介：王德吉，男，博士，高级培训师，主要从事烟草信息化及自动化研究，Tel：0371-65857981，E-mail：wangdeji@yahoo.cn

基金项目：国家自然科学基金资助项目(No. 60075022)

收稿日期：2008-10-07

论。自 1975 年 Efron 首先在统计学中采用微分几何方法以来,许多统计学家在这方面进行了大量的工作。特别是由于甘利俊一(Amari S)和 Zhu H 等人的杰出工作,使得信息几何理论得到学术界的广泛关注。为了说明核函数的几何结构,首先给出微分流形、子流形和嵌入的概念[7]。

定义 1:微分流形是一类拓扑空间,除具有通常的拓扑结构外,还添上了微分结构。微分几何学的研究是建立在微分流形上的。三维欧氏空间 R^3 中的曲面是二维的微分流形,但微分流形的概念远比这广泛得多,非但维数不限于二维,而且流形也不必作为 n 维欧氏空间 R^n 中的曲面来定义。此外,一般微分流形也不一定有距离的概念。

定义 2:如果满足下列条件,称 M 是 W 的子流形。其中 W 是 M 的流形,M 是 W 的子集合。$[\zeta^i]=[\zeta^1,\ldots,\zeta^n]$,$[\zeta^a]=[\zeta^1,\ldots,\zeta^m]$,,分别是对应于 W 和 M 的坐标系。

$$n=\dim W \tag{2}$$

$$m=\dim M \tag{3}$$

(i) 受限于 M 的每个 ζ^i 记为 $\zeta^i|_M$ 是 M 上的 C^∞ 级(无穷可微)函数;

(ii) 设 $B_a^i=(\frac{\partial \zeta^i}{\partial \zeta^a})_p$(更确切地记为 $(\frac{\partial \zeta^i|_M}{\partial \zeta^a})_p$)及 $B_a=[B_a^1,\ldots,B_a^n]\in R^n$。对 M 中的每个点 p,$\{B_1,\ldots,B_m\}$ 是线性无关的($m\leqslant n$)。

(iii) 对 M 中的任意开子集 V,存在 W 开子集 U,使得 $V=M\cap U$。

值得注意的是上述条件对于坐标系 $[\zeta^i]$,$[\zeta^a]$ 的选择是无关的。事实上,条件(ii)和(iii)给出了 M 是 W 的嵌入 $t:M\to W$,$t(p)=p$,$\forall p\in M$,表明嵌入 t 是 C^∞ 映射。

根据上述观点,非线性映射 $\phi(x)$ 是一个子流形,它定义了从输入空间 S 到特征空间 F 的一个嵌入。一般 F 为再生核 Hilbert 空间(RKHS),RKHS 是 Hilbert 空间的子空间,因此,可以在 S 空间中引入一个黎曼度量 G_{ij}[7]。

$$G_{ij}=<\frac{\partial}{\partial x_i}\phi(x),\frac{\partial}{\partial x'_j}\phi(x')> \tag{4}$$

考虑到

$$K(x,x')=<\phi(x),\phi(x')> \tag{5}$$

$$G_{ij}=\frac{\partial}{\partial x_i}\frac{\partial}{\partial x'_j}K(x,x')|_{x'=x} \tag{6}$$

为方便,省略了上式关于 i 和 j 的求和,这种表达法便是通常的爱因斯坦归约。令 $G(x)=\det|G_{ij}(x)|$,可以看出 $G(x)$ 表示了 S 的局部区域在映射 ϕ 的伸缩情况,故称为伸缩因子。

基于以上分析,为了提高 SVM 的函数逼近和回估的精度,可以构建核函数,使得像空间数据边界超平面附近的体积放大,即放大其黎曼度量 G_{ij},这可以通过引入核函数共形变换来实现。

$$\tilde{K}(x,x')=D(x)D(x')K(x,x') \tag{7}$$

由上式,可得到变换后的黎曼度量:

$$\tilde{G}_{i,j}(x)=\frac{\partial D(x)}{\partial x_i}\frac{\partial D(x)}{\partial x'_j}+[D(x)]^2G_{i,j}(x) \tag{8}$$

基于上述讨论,如果适当选取一个保角映射,就可以在保持原来空间拓扑结构不变的情况下,对非线性数据中的重要样本点附近的区域实现有效放大,从而提高预测效果。

定理:新的核函数 $\tilde{K}(x,x')$ 满足 Mercer 定理,可以作为 SVM 中的核函数。

证明:设 A 为 R^n 的紧子集,$h(x)\in L_2(A)$。

由 $\tilde{K}(x,x')=D(x)D(x')K(x,x')$,

易见是连续并且对称。

设 A 为 R^n 的紧子集,$h(x)\in L_2(A)$。

因为 $D(x)>0$,因此必存在一个正数 β,使得 $D(x)\geqslant\beta>0$,从而

$$\begin{aligned}&\iint_{A\times A}\tilde{K}(x,x')h(x)h(x')\,dxdx'\\&=\iint_{A\times A}D(x)D(x')K(x,x')h(x)h(x')\,dxdx'\\&\geqslant\beta^2\iint_{A\times A}K(x,x')h(x)h(x')\,dxdx'\geqslant 0\end{aligned} \tag{9}$$

由此可见 $\tilde{K}(x,x')$ 半正定。新的核函数 $\tilde{K}(x,x')$ 满足 Mercer 定理,因此可作为 SVM 的核函数。

3　试验

选择 22 种不同地区和等级的烟叶,利用相同卷烟纸和滤嘴将各种烟叶卷制成单料卷烟,检测得各种烟叶单克烟丝烟气中焦油量($mg\cdot g^{-1}$):彭水 23.01,楚雄 19.13,大理 27.01,昆明 26.73,陆良 25.94,罗平 25.15,沾益 21.75,思茅 28.33,宣威 24.52,会东 23.38,叙永 27.06,平陆 26.37,辽宁 29.85,津巴布韦 26.26,桂阳 28.82,临颖 28.40,蓝山 31.49,三门峡 28.93,兴义 29.32,大龙 29.61,福建 26.49,巴西 26.55。利用这 22 种烟叶按表 1 组成 20 个配方,用相同卷烟纸和滤嘴卷制成相同规格的卷烟,然后进行检测。

表 1 22 种不同地区等级的烟叶组成 20 种配方 (mg)

配方	彭水	楚雄	大理	昆明	陆良	罗平	沾益	思茅	宣威	会东	叙永	平陆	辽宁	津巴布韦	桂阳	临颍	蓝山	三门峡	兴义	大龙	福建	巴西	合计
1	0	0	4	4	6	8	10	0	0	4	4	8	8	10	10	0	4	6	8	8	10	10	122
2	0	4	6	8	10	4	10	0	4	6	8	6	6	8	10	0	0	4	4	6	8	10	122
3	4	6	10	0	4	10	8	4	6	10	0	10	0	4	10	0	4	6	8	10	4	10	128
4	4	8	0	6	10	6	8	4	8	0	6	4	8	10	8	4	6	10	0	4	10	8	132
5	6	10	4	10	4	0	6	6	10	4	10	10	0	6	8	4	8	0	6	10	6	8	136
6	6	0	8	0	8	10	6	6	0	8	0	4	8	0	6	6	10	4	10	4	0	6	110
7	8	4	10	6	0	6	8	8	4	10	6	8	4	10	6	6	0	8	0	8	10	6	136
8	8	6	0	10	8	0	4	8	6	0	10	0	10	6	4	8	4	10	6	0	6	4	118
9	10	8	6	4	0	8	0	10	8	6	4	8	4	0	4	8	6	0	10	8	0	4	116
10	10	10	8	8	6	4	0	10	10	8	8	0	10	8	0	10	8	6	4	0	8	0	136
11	10	8	6	4	4	0	0	10	8	6	6	6	6	4	0	10	10	8	8	6	4	0	124
12	10	4	10	8	6	4	0	10	4	0	10	4	4	0	0	10	8	6	4	4	0	0	106
13	8	10	4	0	10	6	4	8	10	8	4	8	6	4	0	10	4	10	8	6	4	0	132
14	8	6	10	6	0	8	4	8	6	0	10	0	10	6	4	8	10	4	0	10	6	4	128
15	6	0	4	10	8	10	6	6	0	8	4	6	0	8	4	8	6	10	6	0	8	4	122
16	6	10	8	0	8	0	6	6	10	4	8	10	4	10	6	6	0	4	10	4	10	6	136
17	4	6	0	6	10	4	8	4	6	10	0	0	8	0	6	6	10	8	0	8	0	6	110
18	4	0	8	10	0	6	8	4	0	4	8	6	10	4	8	4	6	0	6	10	4	8	118
19	0	8	0	4	6	8	10	0	8	10	0	10	0	6	8	4	0	8	10	0	6	8	114
20	0	4	6	8	8	10	10	0	4	6	6	4	6	8	10	0	8	0	4	6	8	10	126

通常,烟厂根据单料烟叶中所测焦油量,使用如下简单求和公式作为单支卷烟焦油量的估计公式。其中,y 为单支香烟焦油估计含量/mg;y_i 为第 i 种烟叶单克焦油量/mg;x_i 为第 i 种烟叶在配方中的质量/mg;m 为配方总质量/mg;$w = 0.7$ g,为单支香烟质量。

$$y = \frac{w}{m}\sum_{i=1}^{22} y_i x_i = \sum_{i=1}^{22} \frac{w}{m} y_i x_i = \sum_{i=1}^{22} X_i \qquad (10)$$

上式中 $X_i = \frac{w}{m} y_i x_i$ $(i = 1,2,\ldots,22)$ 表示单支香烟中由第 i 种烟叶产生的焦油量。经过公式(10)的计算,可得出 0.7 g 单支香烟各种烟叶产生的焦油量如表 2 所示,对其进行简单求和,就可得简单求和焦油量值预测值。对于 SVM 预测,就变成了能否利用表 2 的 20 个观测值建立 y 和 $X_1, X_2, \ldots, X_{22}$ 的预测关系。

表 2 单支香烟焦油量实测值与简单求和预测值

配方	彭水 X_1	楚雄 X_2	大理 X_3	昆明 X_4	陆良 X_5	罗平 X_6	沾益 X_7	思茅 X_8	宣威 X_9	会东 X_{10}	叙永 X_{11}	平陆 X_{12}	辽宁 X_{13}	津巴布韦 X_{14}	桂阳 X_{15}	临颍 X_{16}	蓝山 X_{17}	三门峡 X_{18}	兴义 X_{19}	大龙 X_{20}	福建 X_{21}	巴西 X_{22}	实测值	简单求和预测值
1	0.00	0.00	0.62	0.61	0.89	1.15	1.25	0.00	0.00	0.54	0.62	1.21	1.37	1.51	1.65	0.00	0.72	1.00	1.35	1.36	1.52	1.52	19.20	18.89
2	0.00	0.44	0.93	1.23	1.49	0.58	1.25	0.00	0.56	0.80	1.24	0.91	1.03	1.21	1.65	0.00	0.00	0.66	0.67	1.02	1.22	1.52	18.00	18.41
3	0.50	0.63	1.48	0.00	0.57	1.38	0.95	0.62	0.80	1.28	0.00	1.44	0.00	0.57	1.58	0.00	0.69	0.95	1.28	1.62	0.58	1.45	19.30	18.37
4	0.49	0.81	0.00	0.85	1.38	0.80	0.92	0.60	1.04	0.00	0.86	0.56	1.27	1.39	1.22	0.60	1.00	1.53	0.00	0.63	1.40	1.13	18.60	18.49
5	0.71	0.98	0.56	1.38	0.53	0.00	0.67	0.87	1.26	0.48	1.39	1.36	0.00	0.81	1.19	0.58	1.30	0.00	0.91	1.52	0.82	1.09	19.90	18.42
6	0.88	0.00	1.38	0.00	1.32	1.60	0.83	1.08	0.00	1.19	0.00	0.67	1.52	0.00	1.10	1.08	2.00	0.74	1.87	0.75	0.00	1.01	19.80	19.03
7	0.95	0.39	1.39	0.83	0.00	0.78	0.90	1.17	0.50	1.20	0.84	1.09	0.61	1.35	0.89	0.88	0.00	1.19	0.00	1.22	1.36	0.82	19.40	18.35
8	1.09	0.68	0.00	1.59	1.23	0.00	0.52	1.34	0.87	0.00	1.61	0.00	1.77	0.93	0.68	1.35	0.75	1.72	1.04	0.00	0.94	0.63	19.60	18.75
9	1.39	0.92	0.98	0.65	0.00	1.21	0.00	1.71	1.18	0.85	0.65	1.27	0.72	0.00	0.70	1.37	1.14	0.00	1.77	1.43	0.00	0.64	19.70	18.58
10	1.18	0.98	1.11	1.10	0.80	0.52	0.00	1.46	1.26	0.96	1.11	0.00	1.54	1.08	0.00	1.46	1.30	0.89	0.60	0.00	1.09	0.00	19.00	18.46
11	1.30	0.86	0.91	0.60	0.59	0.00	0.00	1.60	1.11	0.79	0.92	0.89	1.01	0.59	0.00	1.60	1.78	1.31	1.32	1.00	0.60	0.00	18.90	18.79

(续表 2)

12	1.52	0.51	1.78	1.41	1.03	0.66	0.00	1.87	0.65	0.00	1.79	0.70	0.79	0.00	0.00	1.88	1.66	1.15	0.77	0.78	0.00	0.00	18.40	18.95
13	0.98	1.01	0.57	0.00	1.38	0.80	0.46	1.20	1.30	0.99	0.57	1.12	0.95	0.56	0.00	1.51	0.67	1.53	1.24	0.94	0.56	0.00	19.90	18.35
14	1.01	0.63	1.48	0.88	0.00	1.10	0.48	1.24	0.80	0.00	1.48	0.00	1.63	0.86	0.63	1.24	1.72	0.63	0.00	1.62	0.87	0.58	19.50	18.88
15	0.79	0.00	0.62	1.53	1.19	1.44	0.75	0.98	0.00	1.07	0.62	0.91	0.00	1.21	0.66	1.30	1.08	1.66	1.01	0.00	1.22	0.61	19.70	18.65
16	0.71	0.98	1.11	0.00	1.07	0.00	0.67	0.87	1.26	0.48	1.11	1.36	0.61	1.35	0.89	0.88	0.00	0.60	1.51	0.61	1.36	0.82	20.10	18.27
17	0.59	0.73	0.00	1.02	1.65	0.64	1.11	0.72	0.94	1.49	0.00	0.00	1.52	0.00	1.10	1.08	2.00	1.47	0.00	1.51	0.00	1.01	20.50	18.58
18	0.55	0.00	1.28	1.59	0.00	0.90	1.03	0.67	0.00	0.55	1.28	0.94	1.77	0.62	1.37	0.67	1.12	0.00	1.04	1.76	0.63	1.26	19.30	19.04
19	0.00	0.94	0.00	0.66	0.96	1.24	1.34	0.00	1.20	1.44	0.00	1.62	0.00	0.97	1.42	0.70	0.00	1.42	1.80	0.00	0.98	1.30	19.00	17.96
20	0.00	0.43	0.90	1.19	1.15	1.40	1.21	0.00	0.54	0.78	0.90	0.59	1.00	1.17	1.60	0.00	1.40	0.00	0.65	0.99	1.18	1.48	19.50	18.54

在 SVM 模型中,取参数 $c = 2000$, $\varepsilon = 0.00001$ 核函取为

$$K(x, x') = \exp\left(-\frac{||x - x'||^2}{2\rho^2}\right) = exp\left(-\frac{||x_i - x_j||^2}{2 \times 0.75}\right) \tag{11}$$

将保角映射取为：

$$D(x) = \frac{1}{n}\sum_{i=1}^{n}\exp\left(-\frac{||x - o_i||^2}{\beta_i^2}\right) \tag{12}$$

其中 n, o_i, β_i 分别为分类点数目,第 i 类的中心与对应的宽度。

由表 3 可知,通过简单求和方法和基于信息几何的 SVM 预测计算出的预测误差平方和分别为 20.4571 和 0.8571,比值为 23.8678；预测误差绝对平均值分别为 0.8595 和 0.1571。比值为 5.4709,可见,基于信息几何 SVM 预测比原预测公式能够更好地预测卷烟焦油量。

表 3　SVM 焦油量预测与简单求和焦油量预测对比表

实测值	SVM		简单求和	
	预测值	误差值	预测值	误差值
19.2	19.5337	-0.0337	18.89	0.31
18.0	18.3609	-0.0609	18.41	-0.14
19.3	19.7540	-0.4540	18.37	0.93
18.6	18.6070	-0.0070	18.49	0.11
19.9	19.4769	0.4231	18.42	1.48
19.8	19.6651	0.1349	19.03	0.77
19.4	19.3878	0.0122	18.35	1.05
19.6	19.4850	0.1150	18.75	0.85
19.7	19.5579	0.1421	18.58	1.12
19.0	19.2377	-0.2377	18.46	0.54
18.9	19.2100	-0.0100	18.79	0.11
18.4	18.1973	0.2027	18.95	-0.55
19.9	19.7256	0.1744	18.35	1.55
19.5	19.5810	-0.0810	18.88	0.62
19.7	19.8146	-0.1146	18.65	1.05
20.1	20.1798	-0.0798	18.27	1.83
20.5	20.1621	0.3379	18.58	1.92
19.3	19.3421	-0.0421	19.04	0.26
19.0	18.8839	0.1161	17.96	1.04
19.5	19.8629	-0.3629	18.54	0.96

4　结论

卷烟焦油量的预测是卷烟焦油量控制的重要前提和保证,本文从信息几何和支持向量机的角度,建立了基于信息几何的卷烟焦油量 SVM 预测模型。该模型能在小样本的情况下,准确预测。特别是核函数的构造,通过共性变换,给出了 SVM 模型预测的核函数选择的一般方法。与其他预测模型相比更加科学、客观,在卷烟的配方设计中有较好的应用价值,该方法对解决类似的实际问题具有一定的意义。

参考文献

[1] 王强, 陈英武, 李孟军. 基于支持向量机的卷烟焦油预测[J]. 计算机工程与应用, 2007, 43(9)：234-235.

[2] 徐雅静, 汪远征, 王建民. 卷烟焦油含量预测的数学模型[J]. 郑州轻工业学院学报：自然科学版, 2005, 20(5)：35-38.

[3] 钟科军, 蒋腊梅, 黄建国. 卷烟降焦综合技术方法与实践[J]. 烟草科技, 2005(3)：35-37.

[4] VapniK V N. The Nature of Statistical Learning Theory[M]. Springer, New York, 1998.

[5] Mika S, Rätsch G, Weston J, et al. Fisher discriminant Analysis with Kernels[J]. Neural networks for signal processing, 1999, 8：41-48.

[6] Hammer B. A note on the universal approximation capability of support vector machine[J]. Neural processing letters, 2003, 17：45-53.

[7] WANG De-ji, XIONG Fan-lun, WANG Ru-jing, et al. The metamer number prediction based on improved SVM[J]. Pattern Recognition and Artificial Intelligence, 2006, 19(4)：557-560.