文章编号: 1004 -4574(2011)06 -0167 -05

基于支持向量机的水资源安全评价

畅明琦[1,2],刘俊萍[3],马 惟[2]

(1. 长安大学 水与发展研究院,陕西 西安 710064; 2. 中国灌溉排水发展中心,北京 100054;
3. 浙江工业大学 建筑工程学院,浙江 杭州 310032)

摘 要: 支持向量机以统计学习理论为基础,采用结构风险最小化准则,将学习问题转化为一个凸二次规划问题,能够得到全局最优解,适合解决小样本、非线性分类及回归问题。根据水资源安全的内涵,筛选出具有代表性的指标,组成水资源安全评价指标体系。建立了基于支持向量机的水资源安全评价模型,将安全标准划分为良好、安全、临界、不安全、危险5个等级。根据水资源安全评价标准及所属评价等级值,随机生成样本集,180个样本作为训练样本,构造了5个两类支持向量分类器,20个样本作为检验样本,检验样本分类全部正确。将模型应用于山西省11个城市的水资源安全评价,结果表明,该方法是有效、可行的。

关键词: 统计学习理论; 支持向量机; 模式分类; 水资源安全
中图分类号: TV213.4 **文献标志码:** A

Water resources security assessment based on support vector machine

CHANG Ming-qi[1,2], LIU Jun-ping[3], MA Wei[2]

(1. Research Institute of Water Development, Chang'an University, Xi'an 710064, China; 2. China Irrigation and Drainage Development Center, Beijing 100054, China; 3. College of Civil Engineering and Architecture, Zhejiang University of Technology, Hangzhou 310032, China)

Abstract: Based on statistical learning theory, support vector machine(SVM) can transform the learning process into a convex quadratic planning problem to get a global optimization by using the rule of structure risk minimization, which is appropriate to solving small sample, nonlinear classification and regression. Based on the concept of water resources security, representative indicators were selected for the water resources security assessment indicator system. Water resources assessment model based on support vector machine was established. Water resources security standards were divided into five grades, named good, safe, critical, not safe and dangerous. Sample sets were formed by stochastic method according to water resources security standards and their grade values. 180 samples were used for training to construct 5 two-classification support vector classifiers. Twenty samples were used for testing and all of which can be classified correctly. Applying the model to 11 cities in Shanxi Province, the results show that the algorithm is reasonable and feasible.

Key words: statistical learning theory; support vector machine; pattern classification; water resources security

传统统计模式识别的方法都是建立在样本数目足够多的前提下进行研究的,所提出的各种方法只有在样本数趋向无穷大时其性能才有理论上的保证。而在多数实际应用中,样本数目通常是有限的,这时很多方法都难以取得理想的效果。统计学习理论是一种专门的小样本统计理论,为研究有限样本情况下的统计模式识别和更广泛的机器学习问题建立了一个较好的理论框架,同时也发展了一种新的模式识别方法——支

收稿日期: 2010 -06 -31; 修回日期: 2010 -12 -25
基金项目: 教育部国家外国专家局111创新引智计划(B08039); 全球环境基金(GEF)(MWR -9 -2 -1)
作者简介: 畅明琦(1962 -),男,教授级高级工程师,博士,主要从事水资源安全系统研究. E-mail: cmq93@163.com

持向量机(support vector machine,SVM),能够较好地解决小样本学习问题。目前,统计学习理论和支持向量机已经成为国际上机器学习领域新的研究热点[1-3]。

1 支持向量机分类

1.1 线性分类

SVM是从线性可分情况下的最优分类面发展而来的。基本原理如下[4-6]:

设给定训练样本集$\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,其中$x_i \in R^d, i \in 1,2,\cdots,n$是$n$个$d$维向量,$y_i \in \{1,-1\}$或$y_i \in \{1,2,\cdots,k\}$或$y_i \in R, i \in 1,2,\cdots,n$,。通过训练学习寻求模式$f(x)$,使得不但对于训练样本集满足$y_i = f(x_i)$,而且对于预测数据集$\{x_{n+1},x_{n+2},\cdots,x_m\}$,同样能得到满意的对应预测值$y_i$,模式$f(x)$称为支持向量机。当$y_i \in \{1,-1\}$时为最简单的两类分类,$y_i \in \{1,2,\cdots,k\}$时为类分类,$y_i \in R$时为函数估计,即回归分析。

对线性可分样本,d维空间中线性判别函数的一般形式为$g(x) + w + x + b$,分类面方程为:

$$w + x + b = 0. \tag{1}$$

将判别函数进行归一化,使两类所有样本都满足$|g(x)| \geqslant 1$,使离分类面最近的样本的$|g(x)| = 1$,分类间隔等于$2/\|w\|$,因此使间隔最大等价于使$\|w\|$或$\|w\|^2$最小,而要求分类线对所有样本正确分类,就是要求它满足

$$y_i(w \cdot_{xi} + b) = 1, i = 1,2,\cdots,n, \tag{2}$$

满足上述条件且使最小的分类面就是最优分类面。最优分类面不但能将两类样本正确分开,而且使分类间隔最大。如图1所示,H为分类面,H_1和H_2分别为过各类中距分类面最近的样本且平行于分类面的面,它们之间的距离称为分类间隔。所谓最优分类面就是要求分类面不但能将两类正确分开(训练错误率为0),而且使分类间隔最大。H_1,H_2上的训练样本就是使式中使等号成立的样本,这些样本称为支持向量(support vectors),即为图1中用圆圈标出的点。

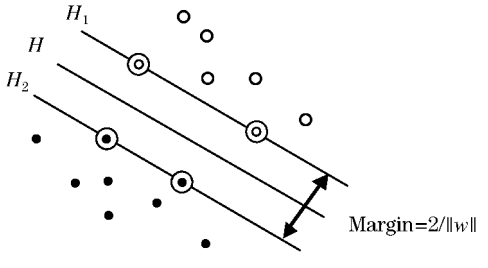

图1 最优秀分类面示意图

Fig. 1 Sketch of optimum classification surface

经理论推导,最优分类函数是

$$f(x) = \mathrm{sgn}\{(w^* \cdot x + b^*\} = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i^* y_i(x_i \cdot x) + b^*\right). \tag{3}$$

式中:sgn(·)为符号函数;α_i^*为最优解;$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$为最优分类面的权系数向量,是训练样本向量的线性组和;b^*为分类的域值。由于非支持向量对应的α_i均为0,因此式中的求和实际上只对支持向量进行。

1.2 非线性分类

对非线性问题,可以通过非线性变换将其转化为某个高维空间中的线性问题,在变换后空间中求最优分类面。在最优分类面中采用适当的内积函数$K(x_i \cdot x_j)$就可以实现某一非线性变换后的线性分类,相应的分类函数也变为

$$f(x) = \mathrm{sgn}\{(w^* \cdot x + b^*\} = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i^* y_i K(x_i \cdot x) + b^*\right), \tag{4}$$

式中符号意义同前。

在特征空间H中构造最优超平面时,在分类算法中用$K(x_i,x_j)$代替(x_i,x_j)。从理论上讲,满足Mercer条件的对称函数都可以作为核函数。

2 水资源安全支持向量机评价模型

2.1 水资源安全评价指标及分级标准

水资源安全系统是一个复杂的动态的巨系统,涉及自然和社会经济复合系统中的诸多子系统,根据水资

源安全的概念与内涵,筛选出具有代表性的 21 个指标,即人均水资源量、亩均水资源量,...,道德意识,建立了水资源安全评价指标体系,见表 1。将水资源安全标准划分为 5 级,即良好、安全、临界安全、不安全、危险,对应每一级别,每个指标都有各自的阈值范围,根据山西省实际情况,结合国际、国家、标准,综合分析相关研究成果,通过分析历史资料,确定评价标准。用这 5 个等级,评判山西省各地市水资源安全状态所处的级别。

2.2　训练样本生成

根据水资源安全评价等级标准范围,采用随机技术模拟生成足够数量的评价样本[8-10]。设第 k 个评价等级第 j 个评价指标的上限和下限分别为 U_j^k 和 L_j^k,y_j^k 为第 i 个样本对应的评价等级,第 k 级第 j 个指标生成的第 i 个样本为

$$x_{ij}^k = \text{rand} \cdot (L_j^k - U_j^k) + U_j^k. \tag{5}$$

式中: i 为样本个数,$i = 1, 2, \cdots, n_k$,n_k 为属于第 k 个评价等级的样本总数; j 为指标个数; k 为等级数。由 x_{ij} 和 y_i 组成训练样本,x_{ij}为训练样本的输入样本,y_i 为训练样本的输出样本。

2.3　模型训练及检验

每个等级生成 40 个样本,共生成 200 个样本,将 200 个样本随机重新排列,180 个样本作为训练样本,20 个样本作为检验样本。

水资源安全评价属于多分类问题,经典的支持向量机算法只给出了两类分类的算法,本文采用一对多算法,构造 5 个两类分类器,SVM1 为良好与其他类别的分类器,SVM2 为安全与其他类别的分类器,SVM3 为临界安全与其他类别的分类器,SVM4 为不安全与其他类别的分类器,SVM5 为危险与其他类别的分类器。每个分类器是将某一类样本当作一个类别,其他类别的样本当作另一个类别,每一类别与其他的类别之间构造两分类函数,如 SVM1 为良好与其他类别的分类器,即将良好样本的输出设为 1,其他样本的输出设为 -1,经过 180 个训练样本的学习,选取 $C = 1000$,$e = 0.01$,$r = 0.1$,20 个检验样本的检验结果见表 1。

表 1　SVM1 检验结果

Table 1　Check-up results of SVM1

样本	人均水资源量/(m^3/人)	亩均水资源量/(m^3/亩)	全社会 GDP/万元	人均 GDP/(万元/人)	地表水开发利用程度/%	地下水开发利用程度/%	Ⅲ级河段长/总河段长/%	Ⅲ级面积/总区域面积/%	降水深/mm	蒸发量/mm
1	385.3	187.63	3 352.1	1.872 2	1.027 2	1.039	0.590 52	0.794 04	466.24	1 343.2
2	495.23	276.25	8 540.9	3.633 6	0.751 98	0.825 46	0.778 22	0.962 1	670.56	960.69
3	381.86	185.01	4 544.1	1.584	1.058 5	1.034 7	0.557 92	0.702 44	474.83	1 311.3
4	413.77	178.39	3 954.1	1.497	1.357 4	1.003 9	0.507 01	0.79501	411.35	1 345.3
5	513.43	266.97	8 515.5	4.568 4	0.712 72	0.887 27	0.799 35	0.972 19	608.64	940.04
6	384.34	198.74	4 291.5	1.641 5	1.31	1.096 5	0.528	0.709 89	466.06	1 392.1
7	575	546.8	9 149.4	7.953 4	0.541 09	0.374 82	0.819 61	1	943.9	234.3
8	806.41	482.93	9 731.4	5.512 1	0.172 46	0.468 05	0.927 89	1	709.94	891.62
9	35.056	52.255	2 799.6	0.1527	9.800 4	7.002 2	0.220 06	0.276 35	50.064	1 503.4
10	518.54	289.5	8 964.7	3.624 1	0.773 31	0.895 14	0.744 12	0.952 84	677.45	912.96
11	170.31	74.25	809.9	0.308 49	9.045 5	7.497 2	0.399 48	0.614 49	227.04	2 394.3
12	396.21	187.57	3 109.6	2.380 4	1.022 3	1.032 1	0.586 16	0.764 68	471.94	1 328.9
13	504.95	285.44	8 134.8	4.481 2	0.677 1	0.847 45	0.729 23	0.943 49	664.46	919.55
14	409.91	193.32	4393.8	1.850 3	1.377 7	1.053	0.568 59	0.743 15	388.16	1 207.3
15	480.07	217.93	5 564.6	3.271 4	0.950 62	0.908 85	0.659 39	0.852 08	505.85	1 123
16	540	290	9 105	4.63	0.6	0.8	0.8	1	680	900
17	486.25	242.67	6 345.3	3.324 3	0.860 24	0.972 51	0.697 79	0.811 79	586.61	1 130.7
18	202.25	129.9	1 084.5	0.253 27	2.970 3	9.647 2	0.096 854	0.606 08	70.784	2 841.4
19	468.67	253.64	5 360.2	3.031 9	0.859 31	0.911 34	0.670 07	0.861 87	514.59	1 082.7
20	389.12	193.09	3 562.2	1.466 8	1.415 5	1.007 6	0.570 26	0.718 42	487.85	1 228.7

注: 1 亩 = 1hm^2/15.

续表 1

Table 1 Continued

样本	径流深/mm	地下水径流模数/($m^3 \cdot (s \cdot km^2)^{-1}$)	工业用水水平/(m^3/万元)	农业用水水平/(m^3/亩)	城市生活用水水平/(L/人·d)	农村生活用水水平/(L/人·d)	污水利用率/%	法规制度	水价	公共权力	道德意识	实际模型分类	分类
1	45.491	3.520 9	95.298	300.21	204.63	101.87	0.333 6	0.615 06	0.601 76	0.628	0.664 81	-1	-1
2	72.44	4.086 5	77.961	267.28	175.1	58.78	0.554 39	0.860 91	0.853 73	0.829 44	0.827 79	-1	-1
3	40.726	3.502 5	99.885	316.06	211.61	107.4	0.255 95	0.676 8	0.690 42	0.642 33	0.614 44	-1	-1
4	48.738	3.571 1	99.833	313.09	204.74	101.56	0.344 41	0.694 29	0.682 89	0.675 83	0.668 51	-1	-1
5	67.794	4.723 8	68.512	265.53	177.69	59.907	0.511 4	0.898 94	0.874 61	0.898 54	0.845 88	-1	-1
6	41.226	3.620 4	94.395	318.82	213.12	105.86	0.307 69	0.624 69	0.670 39	0.633 41	0.653 76	-1	-1
7	77.412	6.713 8	12.841	210.05	149.84	46.504	0.826 3	0.934 15	0.929 3	0.908 97	0.907 1	1	1
8	90.366	7.559 5	0.479 44	80.571	127.48	32.726	0.996 81	0.931 42	0.979 3	0.947 75	0.951 76	1	1
9	12.276	0.169 57	123.17	391.08	304.37	176.91	0.078 08	0.144 6	0.589 69	0.274 26	0.280 78	-1	-1
10	73.531	4.601 5	62.255	277.22	173.94	51.155	0.500 72	0.856 65	0.891 32	0.836 98	0.874 73	-1	-1
11	33.903	1.509 5	105.78	538.83	257.14	265.95	0.168 33	0.120 72	0.019 273	0.594 28	0.195 79	-1	-1
12	47.456	3.623 7	92.54	303.54	202.24	119.78	0.310 24	0.613 64	0.641 65	0.622 49	0.658 47	-1	-1
13	67.039	4.389 1	65.026	263.16	174.35	50.558	0.596 29	0.846 42	0.885 29	0.821 39	0.899 11	-1	-1
14	42.359	3.513 3	92.108	318.31	219.66	119.06	0.332 51	0.685 53	0.676 13	0.651 38	0.639 8	-1	-1
15	64.28 4	3.786 4	85.13	289.25	183.92	79.736	0.497 41	0.731 19	0.720 56	0.769 08	0.791 66	-1	-1
16	76	4.76	58	260	160	50	0.65	0.9	0.9	0.9	0.9	1	-1
17	55.641	3.982 8	81.496	297.57	193.5	92.835	0.355 97	0.743 55	0.769 4	0.746 13	0.793 73	-1	-1
18	34.632	3.248 5	141.3	333.67	320.39	299.92	0.2149 5	0.525 77	0.4900 7	0.503 14	0.1659 2	-1	-1
19	62.575	3.897 3	83.642	286.65	196.44	77.249	0.389 85	0.740 39	0.787 4	0.700 03	0.764 76	-1	-1
20	48.562	3.635 8	99.201	306.64	218.77	109.98	0.252 33	0.694 26	0.699 86	0.612 38	0.670 09	-1	-1

注: 1 亩 = 1hm^2/15。

根据 SVM1 分类模型得到检验样本的分类完全正确,见表 1 所示。

同样利用 180 个训练样本,得到其余 4 个支持向量机分类器,SVM2 为安全与其他类别的分类器,SVM3 为临界安全与其他类别的分类器,SVM4 为不安全与其他类别的分类器,SVM5 为危险与其他类别的分类器。利用训练好的分类器对 20 检验样本进行分类,4 个分类器的误判率均为 0。模型计算与预先设定的评价等级完全一致,说明所选参数较合理。

2.4 模型预测

收集山西省 11 个城市,即太原,大同,阳泉,长治,晋城,朔州,忻州,吕梁,晋中,临汾,运城的 21 个评价指标,组成 11 ×21 的评价样本集。由于各指标具体量纲不同,无法进行直接比较,为了消除各评价指标的量纲影响,需对各评价指标的原始数据进行标准化处理,见表 2。将这个评价矩阵作为预测样本,利用训练好的 5 个支持向量机分类器,评价 21 个地市的水资源安全状况及所属的等级。

SVM1 的输出结果为 [-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1],根据 SVM1 的输出,可以判别出 11 个城市的水资源安全级别均不属于 1 级(良好)。

SVM2 的输出结果为 [-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1],根据 SVM2 的输出,可以判别出 11 个城市的水资源安全级别均不属于 2 级(安全)。

SVM3 的输出结果为 [1 -1 1 1 1 -1 -1 -1 -1 -1 -1],据此可以判别出太原市、阳泉市、长治市和晋城市的水资源安全级别均属于 3 级(临界安全)。

SVM4 的输出结果为 [-1 1 -1 -1 -1 1 1 1 1 1 1],据此可以判别出除属于临界安全的以上 4 个城市外,其余城市的水资源安全级别均属于 4 级(不安全)。

SVM5 的输出结果为 [-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1],据此可以判别出 11 个城市的水资源安全级别均不属于 5 级(危险)。

表 2　标准化处理后的指标集

Table 2　Standardized index set

评价指标	太原市	大同市	阳泉市	长治市	晋城市	朔州市	忻州市	吕梁	晋中市	临汾市	运城市
人均水资源量	0.100 0	0.294 5	0.389 3	0.465 4	0.752 6	0.624 2	0.900 0	0.454 6	0.481 9	0.464 6	0.298 0
亩均水资源量	0.367 1	0.100 0	0.356 5	0.354 6	0.900 0	0.137 9	0.248 6	0.152 9	0.315 2	0.770 7	0.173 4
全社会 GDP	0.900 0	0.363 0	0.133 9	0.362 1	0.285 1	0.100 0	0.100 6	0.102 6	0.266 9	0.361 4	0.368 4
人均 GDP	0.900 0	0.422 1	0.603 9	0.396 1	0.526 0	0.432 5	0.136 4	0.100 0	0.313 0	0.281 8	0.209 1
地表水开发利用程度	0.384 9	0.648 0	0.467 1	0.784 9	0.900 0	0.297 3	0.784 9	0.680 8	0.779 5	0.620 6	0.100 0
地下水开发利用程度	0.144 0	0.496 3	0.540 4	0.745 9	0.753 2	0.797 3	0.900 0	0.525 7	0.261 5	0.789 9	0.100 0
Ⅲ级河段长/总河段长	0.616 7	0.100 0	0.191 7	0.560 0	0.280 8	0.525 0	0.475 0	0.472 3	0.309 1	0.164 8	0.900 0
Ⅲ级面积/总区域面积	0.533 3	0.816 7	0.516 7	0.816 7	0.800 0	0.400 0	0.900 0	0.783 3	0.683 3	0.566 7	0.100 0
降水深	0.316 4	0.160 4	0.528 8	0.716 9	0.900 0	0.100 0	0.349 6	0.427 6	0.459 4	0.571 7	0.699 0
蒸发模数	0.479 0	0.386 3	0.184 2	0.504 2	0.605 3	0.428 4	0.495 8	0.163 2	0.900 0	0.521 1	0.100 0
径流深	0.100 0	0.210 8	0.672 1	0.489 6	0.900 0	0.162 7	0.308 0	0.253 5	0.309 8	0.472 3	0.303 4
地下水径流模数	0.538 9	0.176 0	0.464 6	0.343 6	0.900 0	0.430 0	0.355 7	0.100 0	0.144 9	0.264 2	0.580 4
工业用水水平	0.806 3	0.760 9	0.578 4	0.739 6	0.900 0	0.100 0	0.668 0	0.880 5	0.898 5	0.696 1	0.806 3
农业用水水平	0.543 6	0.664 4	0.100 0	0.414 0	0.900 0	0.655 0	0.555 9	0.498 5	0.533 2	0.417 5	0.668 0
城市生活用水水平	0.100 0	0.791 9	0.834 3	0.695 4	0.486 0	0.896 3	0.877 7	0.900 0	0.791 9	0.742 5	0.748 4
农村生活用水水平	0.661 9	0.824 9	0.100 0	0.824 9	0.561 1	0.900 0	0.692 0	0.835 7	0.694 1	0.794 9	0.694 1
污水利用率	0.900 0	0.732 3	0.100 0	0.332 3	0.332 3	0.358 1	0.603 2	0.241 9	0.216 1	0.177 4	0.345 2
法规制度	0.900 0	0.671 4	0.442 9	0.385 7	0.328 6	0.100 0	0.100 0	0.100 0	0.385 7	0.328 6	0.385 7
水价	0.900 0	0.544 4	0.455 6	0.277 8	0.277 8	0.100 0	0.277 8	0.100 0	0.366 7	0.544 4	0.633 3
公共权力	0.900 0	0.442 9	0.214 3	0.157 1	0.157 1	0.100 0	0.100 0	0.100 0	0.328 6	0.214 3	0.442 9
道德意识	0.900 0	0.536 4	0.536 4	0.354 6	0.354 6	0.281 8	0.172 7	0.100 0	0.245 5	0.172 7	0.354 6

3　结语

支持向量机是在统计学理论的基础上发展起来的一种有限样本下的学习算法,具有严格的理论基础,能较好地解决了小样本、非线性、高维数和局部极小点等实际问题。本文建立了水资源安全评价指标体系和 5 级评价标准,根据评价标准,随机生成评价指标作为训练样本,采用一对多算法构造了 5 个两分类分类器,结果表明,支持向量机具有出色的较好的分类效果,在水资源安全评价中有较好的应用前景。

参考文献:

[1] 边肇祺,张学工. 模式识别 [M]. 2 版. 北京: 清华大学出版社,2000: 284 - 300.

[2] Vapnik V N. The nature of statistical learning theory [M]. New York: Springer, 1995: 70 - 256.

[3] Nello C, John S T. An introduction to support vector machines and other kernel - based learning methods [M]. London: Cambridge university press, 2000.

[4] Corinna C, Vapnik V N. Support - vector network [J]. Machine learning, 1995, 20: 273 - 297.

[5] Lee Y K, Lin Y, Wahba G. Multicategory support vector machines: theory and application to the classification of microarray data and satellite radiance data [J]. Journal of the American Statistical Association, 2004, 99(465): 67 - 81.

[6] Tibshirani R, Hastie T. Margin trees for high - dimensional classification [J]. Journal of Machine Learning Research, 2007, 8: 637 - 652.

[7] 吕干云,程浩忠,董立新,等. 基于多级支持向量机分类器的电力变压器故障识别 [J]. 电力系统及其自动化学报, 2005, 17(1): 19 - 22.

[8] 金菊良,丁晶,魏一鸣,等. 区域水资源可持续利用系统评价的插值模型 [J]. 自然资源学报, 2002, 17(5): 610 - 616.

[9] 宋松柏,蔡焕杰. 区域水资源可持续利用评价的人工神经网络模型 [J]. 农业工程学报, 2004, 20(6): 89 - 93.

[10] 卢敏,张展羽,冯宝平,等. 基于支持向量机的区域水安全预警模型及应用 [J]. 计算机工程, 2006, 32(15): 44 - 45.