doi:10.13582/j.cnki.1672-7835.2020.01.014

基于支持向量机的养老保障满意度非线性模型

李熠煜[1],禹宁瑶[1,2]

(1.湘潭大学 公共管理学院,湖南 湘潭 411105;2.湖南科技大学 人文学院,湖南 湘潭 411201)

摘　要:目前对养老保障满意度的研究所采用的统计方法,都是基于养老保障满意度与影响因素之间的线性关系,且未对所建模型进行检验及理论预测。由于事物之间关系复杂,变量之间往往呈现非线性关系。采用支持向量机算法结合粒子群优化算法,建立养老保障满意度非线性模型。用于研究的养老保障满意度样本数为 8 339 份。结果显示,基于支持向量机的分类模型对养老保障满意度预测精度高于 76%,预测性能优于二元逻辑回归预测结果。表明养老保障满意度与受教育程度、受教育满意度、家庭经济状况满意度、总体生活满意度、对社会总体评价等 5 个影响因素之间存在非线性关系。因此,应用支持向量机算法建立养老保障满意度非线性模型是可行的。

关键词:养老保障;满意度;支持向量机;粒子群优化算法;二元逻辑回归

中图分类号:C913.6　　**文献标志码:**A　　**文章编号:**1672-7835(2020)01-0104-05

本文采用 SVM 算法建立居民养老保障满意度与自变量关系模型,并对模型进行理论预测,检验模型预测能力。

一　建模方法与数据来源

(一)建模方法

支持向量机作为一种新颖而独特的机器学习算法,在数据挖掘、模式识别中应用很广,而分类是数据挖掘中的一项非常重要的任务。用图 1 来说明 SVM 的分类思想。在图 1 的(a)图中,符号“·”“×”分别表示两类样本。该图中可画许多条线将它们分开,如 A、B、C 3 条线。其中 A 线离两组样本更远,将样本错误分类的风险小,因此 A 线为最优分类线。在图 1 的(b)图中,同样有两组样本,H_1、H_2 之间距离则为分类间隔(margin)。H 线与 H_1、H_2 平行,H 线在分开两组样本时,使得 margin 最大。因此,H 线最优为分类线。当样本点非线性可分时,可用核函数将低维空间中的点映射到高维空间中,通过线性可分情形寻找可区分数据的超平面。在高维特征空间,将两组样本分开的最优“平”面称为最优超平面。离分隔超平面最近的点(训练样本)就叫做支持向量。

对于 n 维训练数据(x_i; y_i),$i = 1, \cdots, l$; $x_i \in R^n$; $y \in \{1, 2\}$,1、2 表示因变量 y(如满意度)类标值,SVM 需要解决的最优超平面问题表示为:

$$\min_{w,b,\xi} J(w,\xi,b) = \frac{1}{2}w^{\mathrm{T}}w + C\sum_{i}^{l}\xi_i \tag{1}$$

$$y_i(w^{\mathrm{T}}\varphi(x_i) + b) \geqslant 1 - \xi_i \tag{2}$$

$$\xi_i \geqslant 0 \tag{3}$$

上式中 w 为权失量,b 为偏置量。$\xi(\geqslant 0)$ 为松驰变量,当 $0<\zeta_i<1$ 时数据点 x_i 可以被正确分类;而当 $\zeta_i \geqslant 1$ 时数据点 x_i 被错分。C 为惩罚参数,用于调节对错分样本惩罚程度。当 C 值过大或过小时分别会产生过小或过大的训练错误。

当训练数据集矢量 x_i 被核函数映射到高维特征空间、求取最优线性分类面时,优化目标成为:

$$\min W(\alpha) = -\sum_{i}\alpha_i + \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) \tag{4}$$

收稿日期:2019-10-12

基金项目:国家社会科学基金项目(16BZZ055)

作者简介:李熠煜(1972—),女,湖南资兴人,博士,教授,主要从事行政文化与非政府组织研究。

$$\sum_i \alpha_i y_i = 0 \quad (5)$$

$$0 \leqslant \alpha_i \leqslant C \quad (6)$$

式中 $\alpha_i(\geqslant 0)$ 为 Lagrange 乘子。根据最优化理论中的 KKT (Karush-Kuhn-Tucker)条件,只有少量样本的 α_i 值不为零。

实现非线性变换求解需要定义适当的核函数,本文用高斯径向基函数(RBF)作为 SVM 的核函数:

$$K(x_i, x_j) = \exp(-\gamma x_i - x_j^2) \quad (7)$$

式中 γ 为内核参数,当 γ 值过大或过小时分别会产生过小或过大的训练错误。

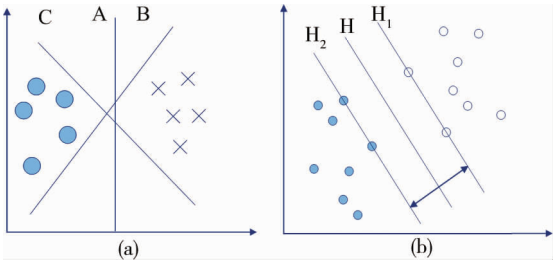

图 1　SVM 线性可分情形最优分类线

SVM 参数值 C 与 γ 大小影响 SVM 预测性能,本文采用粒子群优化算法(particle swarm optimization, PSO)搜寻最佳 C 与 γ 参数值。PSO 算法模拟鸟类捕食过程的社会行为,是一种全局随机优化技术。在 D 维空间,N 个粒子中,每个粒子 i 的位置 $x_i = (x_{i1}, x_{i2}, \cdots, x_{iD})$;速度 $v_i = (v_{i1}, v_{i2}, \cdots, v_{iD})$;粒子 i 经历的最佳位置 $pBest_i = (p_{i1}, p_{i2}, \cdots, p_{iD})$;种群经历的最佳位置 $p_g = (p_{g1}, p_{g2}, \cdots, p_{gD})$。每个粒子通过追踪 $pBest_i$ 与 p_g,按照公式(8)与(9)更新自己速度 v_i 和位置 x_i:

$$v_{id}(\text{new}) = w \times v_{id}(\text{old}) + c_1 \times r_1 \times (p_{id} - x_{id}) + c_2 \times r_2 \times (p_{gd} - x_{id}) \quad (8)$$

$$x_{id}(\text{new}) = x_{id}(\text{old}) + \mu \times v_{id}(\text{new}) \quad (9)$$

其中式(9)中的 w 为惯性权重,调节解空间的搜索范围;r_1 和 r_2 为取值[0, 1]之间的随机函数,增加搜索的随机性;非负常数 c_1、c_2 为学习因子(或加速因子),按经验均取值,可以取值 2 或[0, 4],调节学习最大步长。从随机解出发,通过反复迭代运算,最优解的误差满足设定的阈值或达到迭代次数,优化结束。

(二)数据来源

本研究所使用的数据来源于中国社会科学院 2015 年的“中国社会状况综合调查”(CSS2015)①

因变量为养老保障满意度,分为 10 个等级,1 分表示非常不满意,10 分表示非常满意。由于满意度是一种主观性极强的评价,测量难度大。满意度等级划分越多,相邻或相近等级越难区分。比如划分 10 等级时,2 与 3,3 与 4,2 与 4,……,它们难以截然划分。由于 CSS2015 养老保障的满意度划分为 10 等级,2 分与 9 分的满意度水平分别接近最低与最高水平,因此这 2 个等级可以明显区分。于是本文将“2”分的等级满意度水平归属于为“不满意”,类标值设置为“1”;“9”分的等级满意度水平设置为“满意”,类标值设置为“2”。那么,在 CSS2015 养老保障表中,“1”分满意度水平的样本属于“1”类;“10”分满意度水平的样本属于“2”类。满意度水平分数为 3~8 的样本,其类标值则可能为“1”(“不满意”),也有可能为“2”(“不满意”)。

选取的变量有:家庭成员数;出生年份;受教育程度;自有住房套数;医疗保健支出;个人总收入;家庭总收入;幸福指数;受教育满意度;社交生活满意度;休闲娱乐文化活动满意度;家庭关系满意度;家庭经济状况满意度;居住地的环境状况满意度;总体生活满意度;对社会总体评价,共 16 个自变量。对“不适用”,“拒绝回答”,“不知道”,样本进行删除之后,得到有效样本 8 339 份。

二　结果与讨论

在 8 339 份有效样本中,属于 1、2、3、4、5、6、7、8、9、10 分满意度水平的样本数量分别是 420、300、507、545、1 626、1 286、1 161、1 266、382、846 份。将 300 份 2 分满意度水平的样本(类标值“1”)与 382 份 9 分满意度水平的样本(类标值“2”)作为因变量,上述 16 个自变量组成样本数据集,采用 IBM SPSS Statistics 19/二元逻辑回归/Wald 方法寻找养老保障满意度影响因素,结果得到分类表(见表 1)。表 1 显示,步骤 4 得到的整体回归分类正确率最高,为 77.0 %,但对类标值 1 的预测偏低。步骤 5 得到的整体回归分类正确率次之,为 76.8%,对类标值 1 的预测准确率高于步骤 4 与步骤 6 的预测

①资料出处说明:本论文使用数据全部来自中国社会科学院、中国社会科学院—上海市人民政府上海研究院资助的《2015 年中国社会状况综合调查》。该调查由中国社会科学院社会学研究所执行,项目主持人为李培林。作者感谢上述机构及其人员提供数据协助,本论文内容由作者自行负责。

水平。步骤 6 得到的整体回归分类正确率偏低,仅 75.8%;且对类标值 1 的预测也偏低。因此,我们将步骤 5 得到的自变量(受教育程度;受教育满意度;家庭经济状况满意度;总体生活满意度;对社会总体评价)用于本文的建模。各变量的赋值与描述性统计如表 2 所示。

本文采用 Kennard-Stone 算法①②对将由养老保障 2 分满意度水平的 300 份样本(类标值"1")与养老保障 9 分满意度水平的 382 份样本(类标值"2")划分为训练集与测试集。训练集用来建立模型,测试集用来对模型进行检验。训练集(482 份样本)含类标值"1"的样本为 234 份,含类标值"2"样本为 248 份;测试集(200 份样本)含类标值"1"的样本为 66 份,含类标值"2"样本为 134 份。

采用 IBM SPSS Statistics 19/二元逻辑回归/进入方法对训练集进行建模,自变量特征见表 3。表 3 中 Sig. 值显示,所有的自变量均有 Sig. < 0.05,即方程中的变量均对养老保障满意度有着显著影响。具体解释分析如下:

表 1　二元逻辑回归分类表

已观测			已观测		
			养老保障满意度		百分比校正
			1	2	
步骤 1	养老保障满意度	1	211	89	70.3
		2	105	277	72.5
	总计百分比				71.6
步骤 2	养老保障满意度	1	196	104	65.3
		2	75	307	80.4
	总计百分比				73.8
步骤 3	养老保障满意度	1	210	90	70.0
		2	72	310	81.2
	总计百分比				76.2
步骤 4	养老保障满意度	1	203	97	67.7
		2	60	322	84.3
	总计百分比				77.0
步骤 5	养老保障满意度	1	210	90	70.0
		2	68	314	82.2
	总计百分比				76.8
步骤 6	养老保障满意度	1	207	93	69.0
		2	72	310	81.2
	总计百分比				75.8

表 2　各变量的赋值与描述性统计

变量名称	测量尺度	变量赋值	均值	标准差
受教育程度	定序变量	未上学=1;小学=2;初中=3;高中=4;中专=5;职高技校=6;大学专科=7;大学本科=8;研究生=9;其他=10	3.392 3	1.961 24
受教育满意度	定序变量	从非常不满意到非常满意赋值 1~10	5.131 0	2.378 32
家庭经济状况满意度	定序变量	从非常不满意到非常满意赋值 1~10	5.246 0	2.120 52
总体生活满意度	定序变量	从非常不满意到非常满意赋值 1~10	6.431 3	1.935 44
对社会总体评价	定序变量	从非常不满意到非常满意赋值 1~10	6.456 4	1.625 35
养老保障满意度	定序变量	从非常不满意到非常满意赋值 1~10	6.082 3	2.369 44

(1)养老保障满意度受教育程度的影响。教育程度未上学、小学、初中、高中、中专、职高技校、大学专科、大学本科、研究生、其他,赋值分别是 1、2、3、4、5、6、7、8、9、10。各值所占百分比分别为:11.3、24.9、32.2、12.7、4.2、0.5、7.7、5.7、0.7、0.1,平均分值 3.39,标准差 1.96 分。小学、初中人员占多数。受教育程度越高,对养老保障满意度越低。这可能与教育程度越高的人员对养老保障的希望值越高,而实际情况与心理感受相差较远,因而呈现负相关。

(2)养老保障满意度与受教育满意度正相关。对教育程度满意度同样分为 10 个等级,1 分为非常不满意,10 分为非常满意。从 1-10 分所占百分比分别为:9.2、6.4、10.5、9.6、22.7、14.3、9.6、10.2、2.7、5.0,平均分值 5.13,标准差 2.38 分。对教育程度满意度越高,对养老保障满意度也越高。

(3)养老保障满意度与家庭经济状况满意度也正相关。对家庭经济状况满意度从非常不满意到非常满意赋值 1~10。从 1~10 分各分段所占百分比分别是:6.0、5.8、9.1、9.9、26.0、16.5、11.1、10.6、2.6、2.6,平均分值 5.25,标准差 2.12 分。满

①Daszykowski M, Serneels S, Kaczmarek K, Espen P V, Croux C, Walczak B."TOMCAT: A MATLAB toolbox for multivariate calibration techniques",*Chemometrics and Intelligent Laboratory Systems*, 2007, 85:269-277.

②Yu X, Wang Y, Yang H, Huang X. "Prediction of the binding affinity of aptamers against the influenza virus",*SAR and QSAR in Environmental Research*, 2019, 30(1): 51-62.

意度为中等(5 分、6 分)的人员占多数。对家庭经济状况满意度越高,对养老保障满意度也越高。

(4)养老保障满意度与总体生活满意度正相关。对生活满意度划分为 10 个等级,从 1 分的非常不满意到 10 分的非常满意,各分段所占百分比分别是:1.5、1.8、3.8、5.0、20.9、17.9、16.7、19.7、7.0、5.8,平均分值 6.43,标准差 1.94 分。对生活的满意度越高,养老保障满意度也越高,相关系数 R 为 0.339.

(5)养老保障满意度与对社会总体评价成正比,相关系数 R 达 0.398。对社会的评价按照从非常不满意到非常满意赋值 1~10。各分段所占百分比分别是:0.8、0.7、2.0、4.9、19.8、23.2、22.0、18.4、4.4、3.9、平均分值 6.45,标准差 1.62 分。对社会的总体评价越好,对养老保障满意度也越好。

基于上述 5 个变量的二元逻辑回归分类模型对测试集、以及养老保障满意度水平为 1 分、3~8 分、10 分得样本进行理论预测。训练集、测试集预测结果见表 4。模型对养老保障满意度水平 1~10 分样本预测结果示于图 2,含类标值“2”样本数目依次为 91、15、128、199、609、663、773、966、102、667;其含量(%)依次为 21.7、22.7、25.2、36.5、37.4、51.6、66.6、76.3、76.1、78.8。按照实际情况,养老保障满意度水平分数值越高,满意度也越高,样本集中含类标值“2”样本分数也应该依次升高。但养老保障满意度水平为 8 分、9 分的样本集中,含类标值“2”样本分数分别是 76.3、76.1,前者高于后者,与实际情形不符。因此采用 SVM 进一步建模。

SVM 建模在 MATLAB R2014a 平台上运行。在建模过程中,参数值设置如下:认知学习因子 c_1 初始值和社会学习因子 c_2 初始值设置为 1.5;种群最大数量为 20;最大迭代次数为 200;SVM 参数搜索范围 C 是 10~100;SVM 参数 γ 搜索范围是 0~0.1。采用与二元逻辑回归相同的训练集优化 SVM 参数 C 与 γ,以 5-折进行交叉验证。所到最佳参数:$C = 8.002\ 7$;$\gamma = 8.104\ 5 \times 10^{-4}$。

表 3　二元逻辑回归模型中的变量特征

变量	B	S.E,	Wals	df	Sig.	Exp (B)
受教育程度	−0.231	0.062	13.622	1	0.000	0.794
受教育满意度	0.244	0.049	24.810	1	0.000	1.276
家庭经济状况满意度	0.127	0.057	5.070	1	0.024	1.136
总体生活满意度	0.193	0.062	9.654	1	0.002	1.213
对社会总体评价	0.522	0.074	50.371	1	0.000	1.685
常量	−5.614	0.606	85.886	1	0.000	0.004

基于最佳模型参数,对测试集,以及养老保障满意度水平 1 分、3~8 分、10 分的样本进行预测。训练集、测试集的拟合结果见表 4。表 4 显示,SVM 模型所得训练集、测试集整体预测精度高于二元逻辑回归模型预测结果。SVM 模型对养老保障满意度水平 1-10 分样本预测结果示于图 3,含类标值“2”样本其含量(%)依次为 21.0、21.2、25.0、34.7、37.5、51.7、66.8、75.6、76.9、77.1。这些类标值“2”含量数据排列符合实际情况:养老保障满意度水平分数值越高,满意度也越高。另外,通过对比图 2、图 3 可以看出,图 3 满意度含量随着满意度分值提高呈现增长趋势。

本文基于 SVM 的预测结果优于二元逻辑回归预测结果;且 SVM 的预测结果符合实际情况(养老保障满意度水平分数值越高,满意度也越高)。因此,养老保障满意度与 5 个影响因素(受教育程度;受教育满意度;家庭经济状况满意度;总体生活满意度;对社会总体评价)之间存在非线性关系,采用 SVM 预测养老保障满意度水平是成功的。

表 4　模型预测结果比较

算法		已观测		已预测		
				养老保障满意度		百分比
				1	2	校正
BLR	训练集	养老保障满意度	1	177	57	75.6
			2	50	198	79.8
		总计百分比				77.8
	测试集	养老保障满意度	1	51	15	77.3
			2	32	102	76.1
		总计百分比				76.5
SVM	训练集	养老保障满意度	1	179	55	76.5
			2	51	197	79.4
		总计百分比				78.0
	测试集	养老保障满意度	1	52	14	78.8
			2	31	103	76.9
		总计百分比				77.5

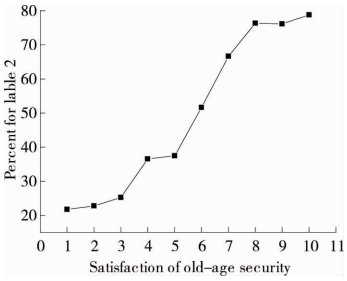

图 2　二元逻辑回归预测结果与满意度关系

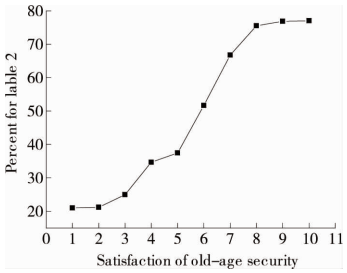

图 3　支持向量机预测结果与满意度关系

结论

基于 CSS2015 的 8 339 份养老保障满意度样本,采用 PSO-SVM 算法成功建立了养老保障满意度与 5 个影响因素(受教育程度;受教育满意度;家庭经济状况满意度;总体生活满意度;对社会总体评价)之间的关系模型。SVM 模型对训练集、测试集满意度预测精度分别是 78.0%、77.5%,分别高于二元逻辑回归预测结果 77.8%、76.5%。并且,SVM 模型对养老保障满意度水平 1-10 分样本预测结果中,含类标值"2"样本其含量(%)依次为 21.0、21.2、25.0、34.7、37.5、51.7、66.8、75.6、76.9、77.1,与养老保障满意度水平分数值越高、满意度也越高的实际情况相符。研究证实,养老保障满意度与 5 个影响因素之间是非线性关系,采用 SVM 预测养老保障满意度水平是成功的。

Nonlinear Model for Satisfaction of Old-Age Security Based onSupport Vector Machine

LI Yi-yu[1] & YU Ning-yao[1,2]

(1. School of Public Administration, Xiangtan University, Xiangtan 411105, China;
2. School of Humanities, Hunan University of Science and Technology, Xiangtan 411201, China)

Abstract: Currently, the statistical methods used for the satisfaction of the old-age security are based on the assumption that there are linear relationships between the satisfaction of the old-age security and the influencing factors. Moreover, these models have not been tested and predicted theoretically. In fact, these relationships should be nonlinear due to the complexity of the relationships. This paper is the first report on the development of the nonlinear model for satisfaction ofthe old-age security, by applying support vector machine (SVM), together with particle swarm optimization (PSO). The number of samplesfor the satisfaction of old-age security is 8339. Results show that the accuracy of SVM classification in predicting satisfaction of old-age security is above 76%, and the prediction performance based on SVM classification is better than that of binary logistic regression (BLR). Results indicate that there are non-linear relationships between the satisfaction of the old-age security and the five influencing factors (educational level, educational satisfaction, family economic status satisfaction, overall life satisfaction, and overall social evaluation). The feasibility of applying PSO-SVM to build nonlinear model for satisfaction of old-age security has been demonstrated.

Key words: old-age security; satisfaction; support vector machine; particle swarm optimization; binary logistic regression

(责任校对　朱正余)