计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191200132

基于粒子群优化的SVM多分类的电动车价格预测研究

李宝胜[1]　秦传东[1,2]
1 北方民族大学数学与信息科学学院　银川 750021
2 宁夏智能信息与大数据处理重点实验室　银川 750021

摘　要　随着新能源汽车的推广，电动汽车逐渐进入千家万户，而影响电动汽车价格的因素较多。文中对影响电动汽车价格的20个属性进行主成分分析研究，先用Pearson相关系数法和PCA算法对数据进行预处理，获得比较重要的样本属性，然后对研究后的新数据进行多分类有监督学习。在支持向量机模型的基础上，用粒子群算法对支持向量机(Support Vector Machine，SVM)模型的参数进行优化选择，实现了对电动汽车的多分类研究，实验表明所建立的模型对电动汽车的多分类效果明显。

关键词：电动汽车；多分类问题；支持向量机；粒子群算法
中图法分类号　TP305

Study on Electric Vehicle Price Prediction Based on PSO-SVM Multi-classification Method

LI Bao-sheng[1] and QIN Chuan-dong[1,2]
1 School of Mathematics and Information Science，North Minzu University，Yinchuan 750021，China
2 Ningxia Key Laboratory of Intelligent Information and Big Data Processing，Yinchuan 750021，China

Abstract　With the promotion of new energy vehicles，electric vehicles have gradually entered thousands of households. There are many factors that affect the price of electric vehicles. Twenty attributes that affect the price of electric vehicles are studied by principle component analysis. First of all，the data are preprocessed by Pearson correlation coefficient method and PCA algorithm to obtain more essential sample attributes. Then，the new data are studied by multi-classification supervised learning. Based on the SVM model，the particle swarm optimization algorithm is used to optimize the parameters of the support vector machine model，and the multi-classification research of electric vehicle is realized successfully. The experimental results show that the multi-classification SVM model has significant effect.

Keywords　Electric vehicle，Multi-classification problem，Support vector machine，Particle swarm optimization algorithm

1　引言

近年来，电动汽车领域快速发展，发展电动汽车对促进节能减排、带动产业转型升级有重要意义[1]。然而，根据中国汽车工业协会调查，我国电动汽车普遍售价高，2019年上半年电动汽车平均售价比燃油车高出81%，导致消费者对电动汽车的接受程度较低。因此，对电动汽车价格进行分类研究具有一定的现实意义。

支持向量机[2]最早是基于二分类问题而设计的，是一种经典的二分类模型，相较于神经网络而言，其在解决线性不可分问题时优势突出。但是实际工作中，总会遇到多分类问题，例如金融信贷评级、生物医学诊断、自然语言处理等。由文献[3]和文献[4]可知，解决多分类问题的方法主要有两种：一种是将多个分类求解问题综合成一个最优化求解问题，同时考虑所有的分类问题[5]；另一种是组合二分类器共同解决多分类问题。第一种方法由于参数众多，目标函数十分复杂，难以实现。因此，将第二种作为常用方法，其构造方法有一对一(OVO)分类器、一对余类(OVR)和纠错输出编码分类器等[6]。

SVM在选取参数时涉及人为主观因素，比如，多项式核中的惩罚因子C与d等参数的选择对模型的性能影响很大。如何对SVM参数进行优化成为学者研究的重点。目前，主流的参数优化方法有Holland创立的遗传算法(Genetic Algorithm，GA)[7]，Dorigo等提出的蚁群算法(Ant Colony Optimization，ACO)[8]，Kennedy等提出的粒子群算法(Partice Swam Optimization，PSO)[9]。还有一些其他优化算法，例如，利用交叉验证法优化参数，通过增量学习和选择性学习来对SVM的参数进行优化等。

由于粒子群具有全局优化的优点，并且粒子具有记忆性等显著特点，因此，将粒子群优化算法与多分类支持向量机模型结合，通过让机器自己去寻找最优的参数，既避免了人为因素对模型的干扰，也能提高模型的性能。

2　相关方法

2.1　SVM的基本模型

支持向量机的基本思想是构建一个超平面来划化二分类的数据[10]，为了使间隔最大化(超平面与最近的样本点的距

本文已加入开放科学计划(OSID)，请扫描上方二维码获取补充信息。
基金项目：宁夏先进智能感知控制技术创新团队(NSFC61362033，NXJG2017003，NXYLXK2017B09)
This work was supported by the Ningxia Advanced Intelligent Perception Control Technology Innovation Team (NSFC61362033，NXJG2017003，NXYLXK2017B09).
通信作者：秦传东(qcd369@163.com)

离)，得到以下优化问题：

$$\begin{aligned}&\min\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i\\&\text{s.t.}\ \ y_i((w\cdot x_i)+b)\geqslant 1-\xi_i,i=1,\cdots,l\\&\qquad\xi_i\geqslant 0,i=1,\cdots,l\end{aligned}\tag{1}$$

其中，$C>0$，ξ_i 为松弛变量。

式(1)相应的拉格朗日函数为：

$$L(w,b,\xi,\alpha,\gamma)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{l}\alpha_i(y_i((w\cdot x_i)+b)-1+\xi_i)+C\sum_{i=1}^{l}\xi_i-\sum_{i=1}^{l}\gamma_i\xi_i$$

其中，α_i，γ_i 为拉格朗日乘子。

对 L 关于 w，b，ξ 求极小值，得到：

$$\begin{aligned}&\sum_{i=1}^{l}\alpha_i y_i=0\\&w=\sum_{i=1}^{l}\alpha_i y_i x_i\\&C-\alpha_i-\gamma_i=0\end{aligned}\tag{2}$$

将式(2)代入拉格朗日函数中，得到对偶问题：

$$\begin{aligned}&\max_{\alpha}\ \sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}y_i y_j\alpha_i\alpha_j K(x_i,x_j)\\&\text{s.t.}\ \ \sum_{i=1}^{l}\alpha_i y_i=0\\&\qquad 0\leqslant\alpha_i\leqslant C,i=1,\cdots,l\end{aligned}\tag{3}$$

决策函数为：

$$f(x)=\text{sign}[(\sum_{i,j=1}^{l}\alpha_i y_i K(x_i,x_j))+b]\tag{4}$$

将二分类模型推广到多类分类问题，有第 1 节中介绍的几种方法，鉴于选取的数据仅有 4 类，初步选取一对一(OVO)分类器方法。这种方法是将一个 m 类问题构造成 $m(m-1)/2$ 个二分类器[11]来解决多分类问题，虽然分类器个数有所增加，但是契合了支持向量机二元分类的特性，其准确率应该会达到理想的目标。

对于来自第 i 类和第 j 类的数据，通过解决下面的二分类问题：

$$\begin{aligned}&\min\ \frac{1}{2}\|w^{ij}\|^2+C\sum_{m=1}^{M}\xi_m^{ij}\\&\text{s.t.}\ \begin{cases}w^{ij}\cdot x_m+b^{ij}\geqslant 1-\xi_m^{ij}, & y_m=i\\ w^{ij}\cdot x_m+b^{ij}\leqslant -1+\xi_m^{ij}, & y_m=j\end{cases}\\&\qquad\xi_m^{ij}\geqslant 0\end{aligned}\tag{5}$$

其中，$M=m(m-1)/2$ 为二分类器的个数。然后基于投票策略进行判定。

如果 $\text{sign}((w^{ij}\cdot x)+b^{ij})$ 表明 x 在第 i 类数据中，则第 i 类获得一票，否则，第 j 类增加一票，得票最多的类别就是要判定的类别。

2.2 SVM 的参数选择

支持向量机由于核函数[12]设定的不同，所需要的参数也有所不同，常用的核函数及其对应的参数如表 1 所列。

表 1 核函数公式及参数

Table 1 Formula and parameters of Kernel function

核函数	公式	参数
线性核	$(x\cdot x')$	C
径向基核	$\exp(-\|x-x'\|^2/\sigma^2)$	C，σ
多项式核	$((x,x')+1)^d$	C，d

由表 1 可知，不同的核函数对应于不同的学习模型，可以根据不同的要求选择最适合的核函数。

为了提高模型的性能，应当设定最优的参数。可以利用启发式优化方法让机器进行参数选择；也可以利用穷举方法，通过网格搜索方法来寻找最优参数。一般认为后者方法过于粗鲁，虽然能够得到最优参数，但是计算代价过高，实际工作中对时间和空间的要求比较高。因此，一般选择启发式优化方法进行参数的选择。本文选用粒子群算法进行参数优化，并加入交叉验证方法，最后利用混淆矩阵进行评估，既避免了庞大的计算量，也能够在一定程度上找到设定最优的参数，提高 SVM 模型的性能。

2.3 基于粒子群的 Multi-class SVM 实现

粒子群算法(PSO)[13]是基于群体的智能优化搜索方法。首先生成初始化种群，每个粒子作为可行空间中的解，都在可行空间中运动，并由目标函数确定对应的适应值，粒子的运动速度决定它们的飞行方向和距离，一般情况下，由于粒子特殊的记忆能力帮助其跟随最优粒子在空间中搜索，每一次迭代中，粒子会通过两个极值来更新自己，一个是局部的最优解 P_{id}，另一个是全局的最优解 P_{gd}。

假设粒子群在一个 n 维的解空间中搜索，种群记作：

$$X=\{X_1,X_2,\cdots,X_m\}\tag{6}$$

粒子数为 m，每个粒子的位置记作：

$$X_i=\{x_{i1},x_{i2},\cdots,x_{in}\}\tag{7}$$

每个粒子的速度都是随机生成的，记作：

$$V_i=\{v_{i1},v_{i2},\cdots,v_{in}\}\tag{8}$$

每个粒子的位置记作：

$$P_i=\{p_{i1},p_{i2},\cdots,p_{in}\}\tag{9}$$

当两个最优解被找到之后，每个粒子根据下面的公式来更新速度。

$$v_{id}^{t+1}=wv_{id}^{t}+c_1r_1(p_{id}^{t}-x_{id}^{t})+c_2r_2(p_{gd}^{t}-x_{gd}^{t})\tag{10}$$

$$x_{id}^{t+1}=x_{id}^{t}+v_{id}^{t+1},i=1,\cdots,m\tag{11}$$

其中，w 是惯性权重，r_1 和 r_2 是 0～1 之间的随机数，c_1 是学习因子，c_2 是社因子。当学习因子大于社会因子时，粒子很容易陷入局部搜索；当社会因子大于学习因子时，有助于粒子向全局最优搜索，因此，取 $c_1=1.6$，$c_2=2$。

w 是由 Shi 等[14-15]在 1998 年的论文中引入的概念，其大小决定了对粒子当前速度的继承，适当的选择可以使粒子具有均衡搜索和开发能力，权重有几种取法，例如：固定权重、时变权重、模糊权重和随机权重等。时变权重可以随着迭代次数动态调节，而且，为使得粒子在前期搜索能力较强，飞行后期具有较好的开发能力，选取时变权重，权重取值范围为 $[w_{\max},w_{\min}]$，最大迭代次数为 $iter\_max$，公式如下：

$$w_i=w_{\max}-\frac{w_{\max}-w_{\min}}{iter\_max}\times i\tag{12}$$

为了避免粒子速度过大，设置速度的上限 $v_{\max}$ 和下限 $v_{\min}$。

$$\begin{cases}v_{\max}=(X_{\max}-X_{\min})/2\\ v_{\min}=-(X_{\max}-X_{\min})/2\end{cases}\tag{13}$$

由于本文选择的核函数是多项式核函数，在将粒子群算法应用于多分类支持向量机的参数寻优时，粒子就是惩罚参数 C 和 d。采用 PSO 调参的目的就是使 Multi-class SVM 分类准确率最大，因此，将经过交叉验证之后的最大分类准确率作为粒子群的适应度函数。基于粒子群优化 Multi-class SVM 的参数流程如图 1 所示。

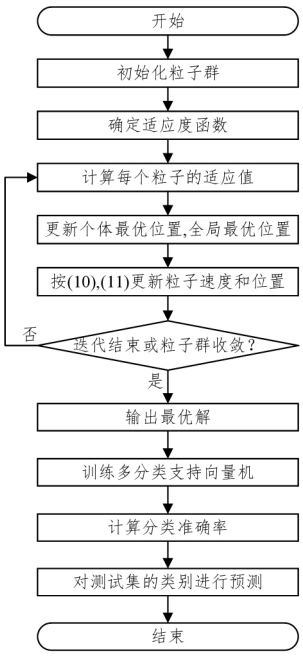

图 1　PSO-MSVM 流程图

Fig. 1　Flowchart of PSO-MSVM

3　数据预处理

本次研究所使用的数据来源于全国首届研究生工业与金融大数据建模邀请赛中的电动汽车价格分类预测数据，数据共有 20 个特征，4 个类别。通过对数据特征的观察发现，部分特征之间可能存在强相关，因此，进行 Pearson 相关系数检验，在计算相关系数之前，首先对数据归一化处理，以消除不同量纲造成的影响。

表 2 是经过计算得到的特征之间的 Pearson 相关系数大于 0.5 的 4 对特征的现实含义。可以看出，特征 5 与特征 11、特征 12 与特征 13、特征 15 与特征 16 以及特征 6 和特征 18 之间相关性较强，其现实意义也比较相近。

表 2　强相关的的特征及含义

Table 2　Characteristics and meaning of strong correlation

特征	含义
feat11，feat5	前置摄像头百万像素，主要相机百万像素
feat12，feat13	像素分辨率高度，像素分辨率宽度
feat15，feat16	屏幕高度，屏幕宽度
feat6，feat18	是否支持 4G，是否有 3G

图 2 中，横坐标代表相关性较大的两对特征，纵坐标代表 Pearson 相关系数值。可以看出，共有 4 对特征的相关系数大于 0.5，而且由表 2 可以看出，每对特征的实际意义相似，这也与我们主观判断的结果相符合。然后，利用主成分分析 PCA 进行特征降维处理，在保留 92.5% 的方差贡献率下，得到新的 16 个主要特征。

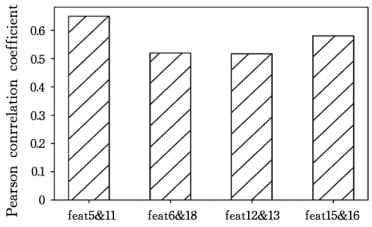

图 2　具有强相关性的特征

Fig. 2　Strongly related features

4　数值实验

由于本文所选取的电动汽车数据共有 4 个类别，因此，在经过一对一分类之后，即 C_4^2，共得到 6 个二分类器。然后，分别计算它们的混淆矩阵，并在这些混淆矩阵的基础上综合考察分类准确率、分类精度和召回率，即直接在每个混淆矩阵上计算出的结果之上再求平均。

$$accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{13}$$

$$precision=\frac{TP}{TP+FP} \tag{14}$$

$$recall=\frac{TP}{TP+FN} \tag{15}$$

其中，TP，TN，FP，FN 分别为二分类中的真正例、真反例、假正例、假反例。

由于 PSO 属于启发式随机搜索，每次计算出的混淆矩阵可能会略有不同，从中选择相对较好的混淆矩阵进行计算。将迭代次数设置为 200，种群设定为 50。利用 k 折交叉验证法，通常 k 取值为 10，即 10 折交叉验证，结果如表 3 所列。

表 3　各二分类准确率、精度及召回率

Table 3　Binary classification accuracy，precision and recall rate

（单位：%）

类别	*accuracy*	*precision*	*recall*
0-1	96.20	95.78	95.78
0-2	100	100	100
0-3	100	100	100
1-2	96.89	94.91	99.49
1-3	100	100	100
2-3	96.98	95.49	98.11
Mean	98.34	97.70	98.80

由表 3 可知，Multi-class SVM 在的综合平均的分类准确率为 98.34%，在预测的是正类的所有结果中，预测正确的比重的平均值为 97.70%；在预测是负类的所有结果中，预测正确的比重的平均值是 98.80%。准确率、精度及召回率越大，则表明模型性能越好，因此，可以认为 Multi-class SVM 模型较为优越。

由图 3 可知，整体而言模型的准确率随着核参数 d 的增加而降低，当 d 在 2 附近时，准确率最高，在 95% 以上。

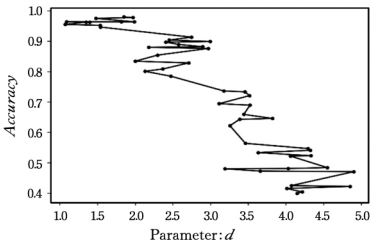

图 3　准确率的变化趋势图

Fig. 3　Trend chart of accuracy

分类支持向量机的分类方法有许多种，不同的分类方法适用于不同的场景，为验证一对一（OVO）分类器方法的有效性，将选择一对余（OVR）[16] 方法和 Crammer 等[17] 提出的直接法（Crammer-Singer，CS）作为对照。

同时，为验证 Multi-class SVM 的有效性，将其与随机森林、决策树（CART）、Adaboost 集成算法及朴素贝叶斯分类器进行对比。由图 4 可以看出，相较于一对余和 Crammer-Sin-

ger[17]方法，一对一方法的分类准确率更高，在 97% 左右，这说明一对一方法更适合于电动汽车多分类数据研究。

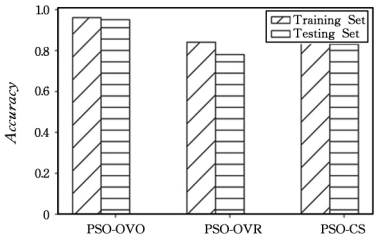

图 4　不同的多分类方法比较

Fig. 4　Comparison of different multi-classification methods

由表 4 可以看出，经过粒子群算法对多分类支持向量机参数 C 和 d 优化之后，其分类效果相较于其他 4 类算法更加优越，在准确率、精度和召回率 3 种评价度量方式下，都略高于其他几种分类算法，这说明 PSO-MSVM 模型在进行多分类时的精确性较高。

表 4　与其他算法的对比结果

Table 4　Comparison of other classification algorithm

（单位：%）

算法	*accuracy*	*precision*	*recall*
决策树(CART)	92.19	92.92	91.13
随机森林	91.67	90.64	92.40
Adaboost	88.69	89.97	87.52
朴素贝叶斯	92.80	92.43	92.82
PSO-MSVM	98.34	97.70	98.80

结束语　本文针对电动汽车多分类数据建立 Multi-class SVM 模型，采用多项式核函数和一对一分类器算法，并将交叉验证分类准确率作为粒子群的适应度函数值，进行最优参数的搜索。数值实验表明，将优化后的参数代入模型，其分类性能显著提高。与其他分类算法的分类效果进行比较，结果表明，利用 PSO 优化核参数的多分类支持向量机的分类准确率更高。但是，利用 PSO 算法进行参数搜索时，容易陷入局部最优，同时，将交叉验证的分类准确率作为适应度值，会造成计算量的增加及运行速度的下降。

参 考 文 献

[1] LIN Q Y, QIU G Y, ZENG H, et al. Research on Price subsidy and Sustainability of Pure Electric vehicles in China based on Learning Curve[J]. Management Modernization, 2019, 39(3): 39-43.

[2] CORTES C, VAPNIK V. Support-Vector Networks [J]. Machine Learning, 1995, 20: 273-297.

[3] 邓乃扬，田英杰. 数据挖掘中的新方法－支持向量机[M]. 北京：科学出版社，2004.

[4] 杨晓峰，郝志峰. 支持向量机的算法设计与分析[M]. 北京：科学出版社，2013.

[5] WETSON J, WATKINS C. Support vector machines for multi-class pattern recognition[R]. Proceedings of the 7th European Symposium on Artificial Neural Networks. 1999.

[6] TOMAR D, AGARWAL S. A comparison on multi-class classification methods based on least squares twin support vector machine[J]. Knowledge-Based Systems. 2015, 81(Jun.): 131-147.

[7] ZHENG C H, JIAO L H. Automatic parameters selection for SVM based on GA[C]// Proceedings of 5th World Congress on Intelligent Control and Automation. Piscataway: IEEE Press, 2004: 1869-1872.

[8] ZHANG X L, CHEN X F, HE Z J. An ACO-based algorithm for parameter optimization of support vector machines[J]. Expert Systems with Applications, 2010, 37(9): 6618-6628.

[9] RANAEE V, EBRAHIMZADEH A, GHADERI R. Application of the PSO-SVM model for recognition of control chart patterns [J]. ISA Transactions. 2010, 49(4): 577-586.

[10] TREVORHASTIE, TIBSHIRANI R, FRIEDMAN J M. The elements of statistical learning[M]. 12th Springer series in statistic, 2017.

[11] ARDJANI F, SADOUNI K. Optimization of SVM Multiclass by Particle Swarm (PSO-SVM)[J]. IJMECS, 2010, 2(2): 32-38.

[12] JU X C, TIAN Y J, LIU D L, et al. Nonparallel Hyperplanes Support Vector Machine for Multi-class Classification[J]. Procedia Computer Science, 2015, 51: 1574-1582.

[13] KAYA D. Optimization of SVM Parameters with Hybrid CS-PSO Algorithms for Parkinson's Disease in LabVIEW Environment[J]. Parkinson's Disease, 2019, 5: 1-9.

[14] SHI Y, EBERHART R C. A modified particle swarm optimizer, [C] // 1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congresson Computational Intelligence (Cat. No. 98TH8360). Anch-orage, AK, USA, 1998: 69-73.

[15] SHI Y, EBERHART R C. Empirical study of particle swarm optimization[C]// Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406). Washington, DC, USA, 1999: 1945-1950.

[16] LIU Y, ZHENG Y F. One-against-all multi-class SVM classification using reliability measures[C]// Proceedings. 2005 IEEE International Joint Conference on Neural Networks. Montreal, 2005: 849-854.

[17] CRAMMER K, SINGER Y. On the Algo-rithmic Implementation of Multiclass Kernel-based Vector Machines[J]. Journal of Machine Learning Research, 2001(2): 265-292.

LI Bao-sheng, born in 1996, postgraduate. His main research interests include big data analysis and machine learning.

QIN Chuan-dong, born in 1976, Ph.D, associate professor. His main research interests include machine learning and intelligent information processing.