文章编号：2095-2163(2020)11-0042-05 中图分类号：TP391.7 文献标志码：A

基于 GWO-SVR 的冠心病住院费用预测

张 慧[1]，贺 松[2]，张 硕[1]，黄 旭[1]，席欢欢[1]
（1 贵州大学 大数据与信息工程学院，贵阳 550025；2 贵州大学 医学院，贵阳 550025）

摘 要：冠心病，作为世界上威胁中老年人健康最常见的疾病之一，近年来诊疗费用不断攀升。因此对冠心病住院费用进行准确的预测，对于着力控制其医疗费用增长具有重要意义。本文运用灰狼优化算法(Grey Wolf Optimizer ，GWO)对支持向量机回归(Support Vector Regression，SVR)模型的惩罚系数 C 和核函数方差 g 进行优化，实现了基于 GWO-SVR 的冠心病住院费用预测模型。研究结果表明，相较于原始 SVR 模型，差分进化算法(Differential Evolution，DE)、布谷鸟搜索算法(Cuckoo Search，CS)、粒子群算法(Particle swarm optimization，PSO)优化的 SVR 模型，灰狼优化算法可以在最短时间内实现参数优化，并且能更加精准有效的预测出冠心病住院费用变化的趋势。

关键词：住院费用预测；冠心病；灰狼优化算法；支持向量机回归

Prediction of hospitalization costs for coronary heart disease based on GWO-SVR

ZHANG Hui[1], HE Song[2], ZHANG Shuo[1], HUANG Xu[1], XI Huanhuan[1]
(1 College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China;
2 College of Medical, Guizhou University, Guiyang 550025, China)

【Abstract】 Coronary heart disease (CHD), as one of the most common diseases threatening the health of the middle-aged and elderly people in the world, has been increasing the cost of diagnosis and treatment in recent years. Therefore, accurate prediction of the hospitalization cost of coronary heart disease is of great significance for controlling the growth of medical expenses. In this paper, the grey Wolf optimization algorithm is used to optimize the punishment coefficient C and kernel variance G of the SVM regression model, so as to realize a gwO-SVR based hospital cost prediction model for coronary heart disease. The results show that, compared with the SVR parameter models optimized by differential evolution algorithm (DE), Cuckoo search algorithm (CS) and particle swarm optimization (PSO), the grey Wolf optimization algorithm can realize parameter optimization in the shortest time, and can more accurately and effectively predict the trend of changes in the hospitalization costs of coronary heart disease.

【Key words】 Hospital cost projections; Coronary heart disease; GWO; SVR

0 引 言

随着健康中国战略的不断推进，人们更加注重医疗卫生建设，全面提高全民健康水平已成为各方关注的焦点。但就目前来看，中国医疗费用持续攀升，“看病难，看病贵”依然是医疗卫生事业改革的痛点和难点[1]。冠心病是全球死亡率最高的疾病之一，据世界卫生组织 2011 年的报告指出，中国的冠心病死亡人数已列世界第二位，且发病呈年轻化趋势[2]。因此，研究冠心病住院费用的增长趋势，分析影响病人住院费用的显著影响因素，运用机器学习算法对住院费用进行精准预测，对协调提升全民健康水平和合理控制医疗费用增长具有重大意义。

支持向量机(Support Vector Machine，SV 是 M)对于非线性问题具有较强的拟合能力，简单且泛化能力好，有较强的鲁棒性[3]。但通常在选取最优核函数时，存在许多随机性，经常无法达到良好的预测效果[4]。目前常用的参数优化算法主要有：差分进化算法(Genetic Algorithms，DE)[5]、粒子群优化算法(Particle Swarm Optimization，PSO)[6]和布谷鸟搜索算法(Cuckoo Search，CS)[7]等。本文采用 2014 年 Mirjalili 等提出的灰狼算法 GWO(Grey Wolf Optimizer)[8]，对 SVM 回归模型进行参数优化。相较于上述几种优化算法，灰狼优化算法无论在参数寻优速度方面，还是在预测住院费用拟合效果上都具有突出表现。

基金项目：贵州省数字健康管理工程技术研究中心项目(黔科合 G 字[2014]4002 号)。

作者简介：张 慧(1994-)，女，硕士研究生，主要研究方向：医疗大数据、数据分析、机器学习；贺 松(1974-)，男，硕士，副教授，主要研究方向：医疗大数据；张 硕(1993-)，男，硕士研究生，主要研究方向：计算机应用与网络安全；黄 旭(1995-)，男，硕士研究生，主要研究方向：数据挖掘；席欢欢(1994-)，男，硕士研究生，主要研究方向：医学图像处理。

通讯作者：贺 松 Email：1261908483@qq.com

收稿日期：2020-09-08

1　理论方法

1.1　GWO 优化算法

1.1.1　基本原理

GWO 优化算法的本质就是模拟大自然灰狼群体中严格的等级制度和狩猎行为。通过把灰狼群体划分成 α、β、δ、ω4 个等级，向猎物进行前进搜索，并依次作为最优解、次优解、次次优解和底层进行数学建模的过程。如图 1 所示。

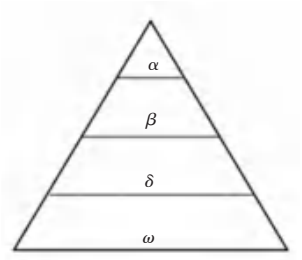

图 1　GWO 灰狼群体等级分类

Fig. 1　GWO grey Wolf group classification

1.1.2　GWO 优化算法狩猎过程

GWO 优化算法是一种通过模仿狼群包围、追捕、攻击 3 大步骤而形成狩猎行为过程。

(1)包围猎物。当灰狼一旦发现猎物，便会迅速向猎物靠近。灰狼与猎物间的距离以及灰狼位置的更新可由式(1)、式(2)得到。

$$\boldsymbol{D} = \{(\boldsymbol{D}_i = | C\boldsymbol{X}_i^p(t) - \boldsymbol{X}_i(t) |) \mid i = 1,2,\cdots,d\}. \tag{1}$$

$$X(t+1) = \boldsymbol{X}_i^p(t) - \boldsymbol{AD}. \tag{2}$$

式中，d 代表搜索空间维度；t 代表灰狼当前的迭代次数；$X(t)$ 代表第 t 代灰狼位置向量；$\boldsymbol{X}_i^p(t)$ 代表第 t 代猎物所在位置向量，即全局最优解向量；$\boldsymbol{D}$ 代表灰狼与猎物之间的距离向量，$\boldsymbol{A}$ 和 $\boldsymbol{C}$ 代表的是系数向量，由式(3) ~ 式(5) 得到。

$$a = 2 - 2 \times \frac{t}{t_{\max}}, \tag{3}$$

$$A = 2ar_1 - a. \tag{4}$$

$$C = 2r_2. \tag{5}$$

式中，$t_{\max}$ 代表最大迭代次数；a 在迭代过程中线性从 2 下降至 0；r_1 与 r_2 均为[0,1] 上的随机变量。

(2)追捕猎物。狼群进行狩猎行为，通常是按照其适应度大小进行排序。可以依次获取到 α 狼的位置，即最优解 X_α；β 狼的位置，即次优解 X_β；δ 狼的位置，即次次优解 X_δ。

α、β、δ、ω 灰狼的实时位置更新公式由式(6) ~ 式(8)得到。

$$\begin{aligned} D^\alpha &= \{(D_i^\alpha = | CX_i^\alpha(t) - X_i(t) |) \mid i = 1,2,\cdots,d\}, \\ D^\beta &= \{(D_i^\beta = | CX_i^\beta(t) - X_i(t) |) \mid i = 1,2,\cdots,d\}, \\ D^\delta &= \{(D_i^\delta = | CX_i^\delta(t) - X_i(t) |) \mid i = 1,2,\cdots,d\}. \end{aligned} \tag{6}$$

$$\begin{aligned} X_1 &= X^\alpha - AD^\alpha \\ X_2 &= X^\beta - AD^\beta \\ X_3 &= X^\delta - AD^\delta \end{aligned} \tag{7}$$

$$X_{(t+1)} = \frac{X_1 + X_2 + X_3}{3}. \tag{8}$$

式中，X_1、X_2、X_3 依次代表灰狼 α、β、δ 的实时更新位置，$X_{(t+1)}$ 代表更新后的最优解向量。灰狼算法最优解的更新过程如图 2 所示。

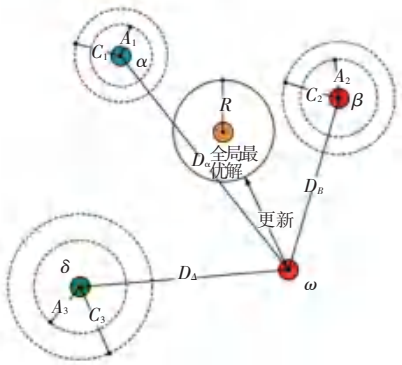

图 2　GWO 算法最优解向量更新过程

Fig. 2　GWO algorithm optimal solution vector update process

(3)攻击猎物。当猎物停止移动时，灰狼便开始发起进攻。进攻行为的发起主要是通过式(3)中 a 迭代次数，从而间接控制式(4) 中 A 的取值来完成的。当 $|A| \leqslant 1$ 时，灰狼群对猎物进行攻击，对应局部搜索；当 $|A| > 1$ 时，灰狼群将远离猎物散去，再次进行全局搜索。

1.2　支持向量机回归

支持向量机回归(SVR)，是支持向量机(SVM)的重要应用分支之一，它是通过 SVM 方法进行拟合曲线，做出相应的回归分析的模型[9]。其核心思想是寻找到一个回归平面，使得一个集合内所有的实验数据距离该平面的距离最近。

SVR 主要是通过给定不敏感损失函数 ε，采用适合的核函数，进行样本训练。通过对惩罚因子 c 和核函数中的方差 g 的计算，获取到不为 0 零的参数所对应的支持向量，继而通过训练样本进行建模，并利用该模型对测试样本进行预测的过程[10]。

$$f(x) = \boldsymbol{\omega}T(x) + \boldsymbol{b}. \tag{9}$$

式(9)中，$\boldsymbol{\omega}$ 代表权向量，$\boldsymbol{b}$ 代表偏置向量。

$$\min \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j - \alpha_j)K(X_i, X_j) +$$

$$\varepsilon\sum_{i=1}^{n}(\alpha_i^* - \alpha_i) - \sum_{i=1}^{n} y_i(\alpha_i^* - \alpha_i)$$

$$\sum_{i}^{n}(\alpha_i^* + \alpha_i) = 0, 0 \leqslant \alpha_i^*, \alpha_i \leqslant \frac{C}{n}. \qquad (10)$$

式(10)中引入的是拉格朗日函数及松弛变量。C 代表 SVR 模型的惩罚因子,用其来表达超出误差 ε 的惩罚程度;K(X,Y)代表 SVR 模型的核函数。SVR 的回归模型由式(11)得到:

$$f(x) = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)K(X_i, X) + b. \qquad (11)$$

2 基于 GWO-SVR 的预测模型构建

在 SVR 回归预测中,惩罚因子 c 和核函数 g 的选取,将直接影响支持向量机的预测拟合效果。为进一步提高模型的准确率,需要借助灰狼优化算法,获取最优参数 Bset_c 和 Best_g,再进行 SVR 回归预测。

GWO-SVR 预测模型的构建过程如下:

步骤 1 读取某二甲医院冠心病 727 份病案首页数据进行预处理。按照比例随机划分数据集,最终生成 650 份训练集数据和 77 份测试集数据,并对所有数据进行归一化处理。

步骤 2 设置 GWO 算法参数。种群规模设置为 N,最大迭代次数设置为 T;设定惩罚系数 C 和核函数方差 g 取值范围。

步骤 3 初始化狼群位置,即每个灰狼的个体位置均由惩罚系数 c 和核函数参数 g 的值决定。

步骤 4 计算每头狼相应的适应度,按照适应度函数值大小排序。

步骤 5 将灰狼群划分成 α、β、δ、ω4 个等级。

步骤 6 根据公式(6)~公式(8)更新灰狼群中每个个体的位置;将灰狼新位置适应度与上一次迭代的适应度进行比较,判断是否替换适应度。

步骤 7 判断当前迭代次数是否为 $t \geqslant T$。若是,输出全局最优值 Best_c 和 Best_g;否则,跳转至步骤四,继续进行参数优化。

步骤 8 输出 α 灰狼位置,即得到最优的 Best_c 和 Best_g 参数。

步骤 9 通过最优的参数 Best_c 和 Best_g 建立 SVR 回归模型,通过测试集进行预测,分析实验结果。

GWO-SVR 实现的流程图如图 3 所示。

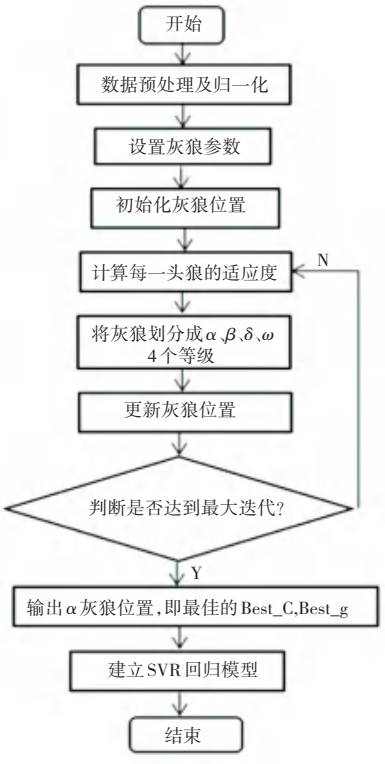

图 3 灰狼算法优化 SVR 参数流程图

Fig. 3 Gray Wolf algorithm to optimize the flow chart of SVR parameters

3 实验结果分析

3.1 数据来源

本研究选用了某二甲医院冠心病病人的病案首页数据,共计 743 例。该数据能够最直观的反映冠心病病人在诊疗过程中最真实有效的费用。经过对目标样本进行数据清洗,缺失项填补、删除信息不全及不符合逻辑项共计 16 例,共纳入有效项 727 例,纳入率为 97.84%。

3.2 住院费用影响因素分析

通过对 727 份冠心病住院费的住院年份、患者年龄、入院途径、并发症合并症级别、是否手术等多个因素进行多重回归分析,结果证明:回归方程模型具有统计学意义($F = 545.336$,$P = 0.000$),方程决定系数 $R^2 = 0.892$。说明住院费用 89.2%的变异可由表 1 中 9 个变量解释。其中对冠心病住院费用影响最大的前 3 位因素分别为:是否造影检查、是否手术、特级护理时间。

3.3 数据预处理及参比模型

将 727 份病案首页数据按照 9∶1 的比例,随机划分成 650 份训练集和 77 份测试集数据。实验中将上述 9 大显著影响住院费用的因素作为输入数据,将住院费用作为输出数据。由于医院病案数据因病人而异,变化幅度相对较大,有时具有突变性,而 SVR 模型对[0,1]的数据又非常敏感,所以在建

模之前，需要借助对病案数据进行归一化处理，以此来提高回归模型的效率。归一化方法如式(12)：

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \quad (12)$$

表 1　住院影响因素的多重回归分析

Tab. 1　Multiple regression analysis of influencing factors of hospitalization

变量	B	SE	Beta	t	p
住院年份	495	272	0.026	0.2.04	0.041
患者年龄	435	112	0.024	1.71	0.038
入院途径	496	295	0.026	2.04	0.041
并发症级别	1 196	168	0.097	7.10	0.000
特级护理	1 792	114	0.250	15.69	0.000
一级护理	744	70	0.285	10.58	0.000
是否手术	6 195	413	0.227	14.99	0.000
造影检查	28 223	703	0.615	40.13	0.000
离院方式	474	143	0.026	2.12	0.042

为了验证 GWO-SVR 模型对冠心病住院费用预测的准确性和有效性，本文除了与原始 SVR 模型形成对照外，还通过与差分进化算法优化的 DE-SVR 模型、布谷鸟算法优化的 CS-SVR 模型和粒子群优化算法的 PSO-SVR 模型形成对照模型，进行了比较分析。

为公平起见，将 SVR 参数取值范围设置为：$c \in [1,100]$，$g \in [1,100]$；所有优化算法的种群个数设为 20，最大迭代次数设为 20。其中，CS 的缩放因子上界为 0.8，下界为 0.2，交叉概率为 0.2；PSO 的 $c_1 = c_2 = 1.5$。

3.4　实验结果分析

图 4、图 5 分别描述了在原始 SVR 模型和 GWO-SVR 模型下测试集的预测结果。

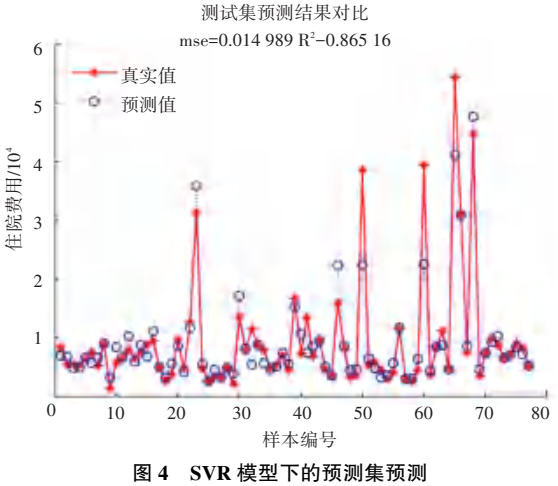

图 4　SVR 模型下的预测集预测

Fig. 4　Prediction set prediction under the SVR model

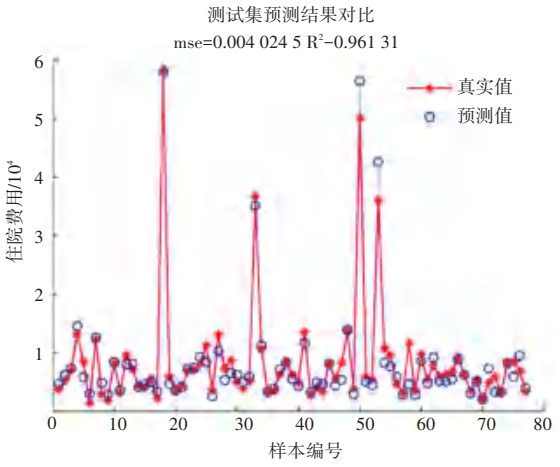

图 5　GWO-SVR 模型下的预测集预测

Fig. 5　Prediction set prediction under the GWO-SVR model

由图可以清晰地看到，在灰狼算法的优化下，SVR 回归模型能够更好的对预测集数据进行预测。为了验证本文所提出的 GWO-SVR 模型的有效性和准确性。表 2 给出了不同优化算法模型对惩罚因子 Best_c 和核函数方差 Best_g 的寻优结果。

通过均方误差 *MSE*、决定系数 R^2、*SVR* 参数迭代寻优时间 *T* 3 个指标，评价不同模型的回归预测性能。详尽数据见表 3。

表 2　不同优化算法下的 Best_c 和 Best_g

Tab. 2　Best_C and Best _g under different optimization algorithms performance

模型	Best_C	Best_g
SVR	4	0.0156
DE-SVR	10.5629	0.01
CS-SVR	12.8877	0.1242
PSO-SVR	0.01	26.6843
GWO-SVR	100	0.0140

表 3　不同模型下的回归预测指标

Tab. 3　Regression prediction indexes of different models

模型	*MSE*	R^2	*T*
SVR	0.015 0	0.865 2	159.86
DE-SVR	0.016 3	0.855 6	8.34
CS-SVR	0.013 8	0.899 6	21.69
PSO-SVR	0.011 4	0.925 2	26.68
GWO-SVR	0.004 0	0.961 3	1.16

通过表 3 的结果可以发现，原始 SVR 模型参数寻优速度较慢，而上述 4 种优化算法均可以显著提升参数寻优速度；另外 CS-SVR、PSO-SVR、GWO-SVR3 种预测模型相较于原始 SVR 模型，对均方误差 *MSE*、决定系数 R^2 都有一定的提升。而 DE-SVR 预测模型的均方误差和决定系数 R^2 不降反升，有可

能是参数设置原因,但参数寻优时间相较于原始 SVR 模型被明显缩短。总体来说,本文所引入的的 GWO-SVR 模型表现最佳,不仅使预测结果的均方误差最小,还具有较强的全局空间搜索能力,在 SVR 参数寻优时仅需要 1.16 s 就可以得到最理想的 (c,g) 解向量,决定系数 R^2 相较于原始 SVR 模型也具有最明显的提升,有效提高了冠心病住院费用的预测精度。

4 结束语

本文通过多重回归分析和归一化处理,分析出影响住院费用的显著影响因素的同时,又提高了 SVR 模型对数据的敏感性,为进一步提高冠心病住院费用预测的精度。本研究引入灰狼算法,对支持向量机回归的惩罚因子 C 和核函数方差 g 进行进一步寻优,算法的全局搜索能力显著提升。GWO 可以迅速锁定 SVR 中最优(C,g),因此在进行预测模型的搭建时,住院费用的预测结果得到了很大的提升。通过与原始 SVR 模型、DE-SVR、CS-SVR、PSO-SVR 的形成对照模型发现,基于灰狼算法优化的 SVR 模型表现最佳,均方误差仅为 0.004 0,参数寻优迭代速度最快仅需 1.16 s、在决定系数 R^2 上也由原来的 0.865 2 显著提升至0.961 3。

参考文献

[1] 黄云霞,杨练,李胜,等. 四川省卫生总费用趋势预测及方法探讨[J]. 中国卫生统计,2015,32(5):836-838.

[2] 刘鸿闽. 胺碘酮治疗冠心病室性心律失常的临床疗效[J]. 齐齐哈尔医学院学报,2012,33(22):3081.

[3] MA Y. Supportvector machines applications[M]. New York: Springer,2014.

[4] 马小平,李博华,张旭,等. 基于 GWO 优化的 CS-SVM 轴承故障诊断[J]. 煤矿机械,2019,40(5):171-173.

[5] 刘娇, 史国友, 杨学钱,等. 基于 DE-SVM 的船舶航迹预测模型[J]. 上海海事大学学报, 2020, 41(1):34-39,115.

[6] SELAKOV A, CVIJETINOVI C D, MILOVIC L, et al. Hybrid PSOSVM method for short-term load forecasting during periods with significant temperature variations in city of Burbank[J]. Applied Soft Computing,2014,16:80-88.

[7] TANG M Z, YANG C H, GUI W H. Fault detection based on modified QBC and CS-SVM[J]. Control and Decision,2012,27(10):1489-1493.

[8] MIRJALILI S, MIRJIALILI S M, LEWIS A. Grey wolf optimizer[J]. Adances in Engineering Software,2014,69(3):46-61.

[9] VLADIMIR C, YUNQIAN M. Practical Selection of SVM Parameters and Noise Estimation for SVM Regression[J]. Neural Networks,2004,17(1):113-126.

[10]陈建飞,郑汉钦. 基于 SVR 的大坝坝体渗压监测预测[J]. 建材与装饰,2019,(16):283-284.

(上接第 41 页)

通过在公开数据集 SIFT1M 上的实验结果表明,在近似最近邻搜索应用中,本文提出的最小化均方误差多阶段码本训练方法优于多阶段向量量化方法,可以进一步地减小向量编码的量化误差,提高查询精度,证明了该方法的可行性和有效性。

参考文献

[1] INDYK P, MOTWANI R. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality[J]. Theory of Computing, 2000, 1(1): 604-613.

[2] GERSHO A, GRAY R, GERSHO A, et. al. Vector quantization and signal compression[J]. springer international, 1992, 159(1):407-485.

[3] HERVÉ JÉGOU, DOUZE M, SCHMID C. Product Quantization for Nearest Neighbor Search[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010, 33(1):117-128.

[4] JUANG B H, GRAY A. Multiple stage vector quantization for speech coding [C]// IEEE International Conference on Acoustics, Speech, & Signal Processing. IEEE, 1982.

[5] CHEN Y, GUAN T, WANG C. Approximate Nearest Neighbor Search by Residual Vector Quantization[J]. Sensors, 2010, 10(12):11259-11273.

[6] BABENKO A, LEMPITSKY V. Additive Quantization for Extreme Vector Compression[C]// Computer Vision & Pattern Recognition. IEEE, 2014.

[7] YUAN J , LIU X. A novel index structure for large scale image descriptor search[C]// IEEE International Conferenceon Image Processing. IEEE, 2013: 1937-1940.

[8] NOROUZI M, FLEET D J. Cartesian k-means[C]//Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. 2013: 3017-3024.

[9] GE T, HE K, KE Q, et al. Optimized product quantization for approximate nearest neighbor search[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 2946-2953.

[10]BABENKO A, LEMPITSKY V. The inverted multi-index[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 37(6): 1247-1260.

[11] ZHANG T, DU C, WANG J. Composite Quantization for Approximate Nearest Neighbor Search[C]//ICML. 2014, 2: 3.

[12]TORRALBA A, FERGUS F, FREEMAN W T. 80 million tiny images: A large database for non-parametric object and scene recognition. IEEE Trans. Patt. Anal. Mach. Int. 2008, 30,1958-1970.