

高流量负荷下基于支持向量机的空间数据聚类方法

王 威

(商洛学院 数学与计算机应用学院, 陕西 商洛 726000)

摘 要: 目前通常采用聚合法和分裂法对空间数据进行分类处理及分层次解析, 其分析速度不高, 局限性较大. 为了加快空间数据聚类速度, 提高数据分类的准确率, 提出一种基于支持向量机的空间数据聚类方法, 首先对空间数据进行预处理, 根据小波多窗普特特征过滤筛选冗余数据, 其次划分过滤后数据的子区间, 作为查询数据时的限定条件提取数据, 根据空间数据的具体特征, 将描述空间数据个体的集合划分为一系列相互独立的组, 确定数据集的空间分布规律, 实现聚合空间数据. 实验结果证明, 所提方法对空间数据具有很高的分类效率, 且准确率较高.

关键词: 空间数据聚类; 支持向量机; 空间分布规律

中图分类号: TP391

文献标识码: A

文章编号: 1000—7180(2017)08—0137—04

Spatial Data Clustering Method Based on Support Vector Machine Under High Traffic Load

WANG Wei

(School of Mathematics and Computer Application, Shangluo College, Shangluo 726000, China)

Abstract: At present, the method of aggregation and splitting is used for spatial data classification and hierarchical analysis. The analysis speed is not high and the limitation is large. In order to speed up the spatial data clustering speed, improve the accuracy of data classification, this paper presents a spatial data clustering method based on support vector machine, firstly, spatial data preprocessing based on wavelet multi window feature pute filtering sieving redundant data, then filtered according to the number of divided subintervals, as defining data extraction condition query data when, according to the specific characteristics of spatial data, the spatial data describing the individual collection into a series of independent groups, determine the spatial distribution of the data set, to realize spatial data aggregation. The experimental results show that the proposed method has high classification efficiency and high accuracy.

Key words: spatial data clustering; support vector machine; spatial distribution

1 引言

由于互联网行业的迅猛发展, 空间数据聚类^[1]在工程学领域有着至关重要的地位. 随着大数据时代的到来, 各种电子类文本数据数量激增, 不仅包含有效信息资源, 还充斥着大量的虚假广告、病毒软件等. 因此人们对空间数据的分析和处理要求不断提高. 支持向量机^[2]作为一种基于统计学理论的学习方法, 是数据挖掘的新方法, 它通过数据库管理系

统^[3-6]对数据进行存储、分析分类.

文献[7]提出基于距离的空间数据分割聚类法, 首先减小空间数据的维数, 缩小低维空间的噪点从而接近有效空间数据, 实现空间数据的分类, 但是这种方法实用性较差, 很多领域都不适用. 文献[8]提出基于数理统计的空间数据聚类方法, 通过对空间数据对象的不确定信息进行统计、分析和分类处理^[4-6], 利用方差结构^[9]、变异函数的相似程度^[10]等实现空间数据的分类, 但是该方法易受到空间数据

收稿日期: 2016—11—08; **修回日期:** 2016—12—18

基金项目: 陕西省教育厅科技计划项目(16JK1240); 商洛学院科学与技术研究项目-服务地方专项(15SKY-FWDF008)

提取的不确定性因素影响。

针对上述问题,提出一种基于支持向量机的空间数据聚类方法,首先对空间数据进行预处理,其次将过滤后得到的数据通过划分子区间,根据空间数据的具体特征,确定数据集的空间分布规律,实现空间数据的聚类。

2 基于支持向量机的空间数据聚类方法

2.1 空间数据的过滤

对空间数据进行预处理,首先对描述的数据信息进行如下说明: $f_x(X,t)$ 、 $f_\theta(X,t)$ 是空间数据中不确定的有界函数,其估计函数分别表示为 $\hat{f}_x(X,t)$ 和 $\hat{f}_\theta(X,t)$,它们之间的关系如下:

$$f(X,t) = \hat{f}_\theta(X,t) + \Delta f_x(X,t) \quad (1)$$

$$f_x(X,t) = \hat{f}_x(X,t) + \Delta f_x(X,t) \quad (2)$$

同时满足:

$$|\Delta f_\theta(X,t)| \leq F_x(X,t) \quad |\Delta f_\theta(X,t)| \leq F_\theta(X,t) \quad (3)$$

假设支持向量机空间数据的干扰信息为数据干扰误差,计算空间数据的指向性增益为

$$c(v) = \int_0^v \cos\left(\frac{\pi}{2}x^2\right)dx \quad (4)$$

根据上述公式得到空间数据的干扰信息为

$$|s(f)| = A\sqrt{\frac{1}{2k}\{[c(v_1) + c(v_2)]^2 + [s(v_1) + s(v_2)]^2\}} \quad (5)$$

根据支持向量机空间数据的丢失数据流特征,进行数据特征分解,得到支持向量机空间干扰数据的去均值量为

$$\tilde{y} = \frac{1}{p_s} \sum_{i=1}^{\infty} \lambda_s^i y_{(x-i)} \quad (6)$$

$$\bar{y} = \frac{1}{p_s} \sum_{i=1}^{\infty} \lambda_L^i y_{(x-i)} \quad (7)$$

式中, λ_s 和 λ_L 表示支持向量机空间数据卷盲因子($0 \leq \lambda_s \leq 1, 0 \leq \lambda_L \leq 1$),且 $\lambda = 2^{-1/h}$, $h_L \gg h_s$ (通常为 100 倍以上)。

$$p_s = \sum_{i=1}^{\infty} \lambda_s^i = \frac{\lambda_s}{1 - \lambda_s} \quad (8)$$

$$p_L = \sum_{i=1}^{\infty} \lambda_L^i = \frac{\lambda_L}{1 - \lambda_L} \quad (9)$$

以式(8)、(9)滤波处理,得到支持向量机空间数据的干扰数据流,以此作为输入向量,进行空间数

据过滤。

2.2 空间数据的提取

支持向量机的空间数据提取作用,将需要提取的空间数据每个子区间上限值标定为 $c_i, c_0 = 0$,则空间数据的每个子区间表示为 $[c_{i-1}, c_i]$,同理,将样本训练集中的每个数据子区间上限表示为 a_j ,则样本集中的每个数据子区间表示为 $[a_{j-1}, a_j]$,令 $a_0 = 0$,空间数据样本区间 $[a_{j-1}, a_j]$ 的个数为 b_j ,则 c_1 的计算方法步骤如下:

$$(1) \text{ 假设 } b_1 \geq k, \text{ 则 } c_1 = c_0 + \frac{k}{b_1} \times (a_1 - a_0).$$

$$(2) \text{ 在空间数据样本区间假设 } b_1 < k \text{ 并且 } b_1 + b_2 \geq k, \text{ 则 } c_1 = (a_1 - c_0) + \frac{k - b_1}{b_2} \times (a_2 - a_1).$$

$$(3) c_1 = (a_2 - c_0) + \frac{k - b_1 - b_2}{b_3} \times (a_3 - a_2)$$

(4) 自上而下推导,设满足公式 $\sum_{i=1}^q b_i \geq k$ 的最小整数时, $q > 2$ 则,求得:

$$c_1 = (a_{q-1} - c_0) + \frac{k - \sum_{i=1}^{q-1} b_i}{b_q} \times (a_q - a_{q-1}).$$

计算得到数据 c_1 ,则按照以上步骤可求得 c_2 :

(1) 设 a_t 为空间数据集中序列 $a_0, a_1, a_2, \dots$ 中的第一个大于 c_1 的数据元素。

(2) 如果子区间 $[c_1, a_t]$ 中的空间数据记录个数

$$\text{大于等于 } k, \text{ 则 } c_1 = a_t + \frac{k - \frac{a_t - c_1}{a_t} - a_{t-1} \times b_t}{b_{t-1}} \times (a_{t+1} - a_t).$$

(3) 如果子区间 $[c_1, a_t]$ 和 $[c_1, a_{t+1}]$ 中的空间数据记录个数都小于 k , 设 $q (q > t + 1)$, 为满足数据

$$\frac{a_t - c_1}{a_t - a_{t-1}} \times b_t + \sum_{l=t+1}^q b_l > k \text{ 的最小整数集合, 则 } c_2 =$$

$$a_t + \frac{k - \frac{a_t - c_1}{a_t - a_{t-1}} \times b_t - \sum_{l=t+1}^{q-1} b_l}{b_q} \times (a_{t+1} - a_t).$$

设 a_t 为空间数据集中序列 $a_0, a_1, a_2, \dots$ 中的第一个大于 c_1 的数据元素。

$$q = \begin{cases} t, [c_{i-1}, a_i] \text{ 区间中空间数据记录个数} \geq k & \text{①} \\ t+1, \text{ 不满足条件(1) 且区间中空间数据记录个数} \geq k & \text{②} \\ \min(h), \text{ 不满足条件(1)(2) 且 } h \text{ 满足} & \\ \frac{a_i - c_i}{a_i - a_{i-1}} \times b_i + \sum_{i=t+1}^h b_i \geq k & \text{③} \end{cases}$$

按照上述思路可以求取 q 值,进而可以总结出

数据 c_i 在区间的取值公式为

$$c_i = \begin{cases} c_{i-1} + \frac{k}{b_i} \times (a_i - a_{i-1}), q = \text{tand} i > 0 \\ a_i + \frac{k - \frac{a_i - c_{i-1}}{a_i - a_{i-1}}}{b_{i+1}} \times (a_{i+1} - a_i), q = t + 1 \text{ and } i > 0 \end{cases}$$

根据上述公式推理算法可以把有用的空间数据从区间训练集提取出来。

2.3 支持向量机的空间数据聚类

空间数据聚类主要根据数据的具体特征进行分类,并使得属于同一类别的数据之间差异尽量小,不同类别数据之间的差异尽可能大,使得不可分割的空间数据获得新的特点,这样更有利于对空间数据的聚类。

假设 $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ 为参与的训练样本,其中 $x_i \in R^d$ 表示输入; $y_i \in \{\pm 1\}$ 表示输出。设最优决策面方程为

$$w^T x_1 + b = 0 \quad (10)$$

$y_1(w^T x_1 + b) \geq 1 - \xi_i$ 为限定性条件,其中 ξ_i 代表线性条件不可分空间数据松弛变量。因此利用乘子法将求解问题转换为限定性优化问题:

$$Q(a) = \sum_{i=1}^N a_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_i a_j y_i y_j K(x_i x_j) \quad (11)$$

其中满足限定性条件的乘子是 $\{a_i\}_{i=1}^N$, 限定性条件表示为 $\sum_{i=1}^N a_i y_i = 0, 0 \leq a_i \leq C, i = 1, 2, 3, \dots, N; K(x, x_i)$ 代表空间数据核函数满足 Mercer 定理。在数据特征空间中,惩罚系数 C 的数值大小变化影响决策的偏离程度。若任意空间数据对象为 s_k , 对空间数据聚类 $C_i (1 \leq i \leq c)$ 的支持向量机隶属度表示为 $PM(s_k, C)$, 通过式(12) 计算:

$$PM(s_k, C) = \frac{(d_{ik})^{-1/(m-1)}}{\sum_{j=1}^c (d_{jk})^{-1/(m-1)}} \quad (12)$$

式中, d_{ik} 是空间数据 s_k 的属性变量 $(x_k, y_k)^T$, 与空间数据聚类 C_j 中心数据的属性变量 $v_j = (v_{j1}, \dots, v_{jd}, v_{jd+1})^T$ 之间的距离。

如果在对应的 $a_j^* > 0, j = 1, \dots, l$ 训练集合 T 中输入 x , 设 $a^* = (a_1^*, \dots, a_l^*)^T$ 代表空间数据任意解集, 则可以按照式(13) 计算原始问题:

$$w^* = \sum_{i=1}^l y_i a_i^* x_i = 0$$

$$b^* = y_j - \sum_{i=1}^l y_i a_i^* (x_i, x_j) \quad (13)$$

式中, $j \in \{j | a_j^* > 0\}$ 或者 $w = \sum_{i=1}^l y_i a_i^* x_i$ 。

支持向量机对空间数据的分类如下:

第 1 步: $x_1 \in X = R^n, y_1 \in Y = \{1, -1\}, i = 1, \dots, l$ 假设已知空间数据训练集合为 $T = \{(x_1, y_1), \dots, (x_l, y_l) \in (X \times Y)^l\}$;

第 2 步: 创建求解最优化数据方程问题(2.3.7), 求得最优解为 $a^* = (a_1^*, \dots, a_l^*)^T$;

第 3 步: 计算 $w^* = \sum_{i=1}^l a_i^* y_i x_i$, 选择 a^* 的一个支持向量机空间数据正分量 a_i^* , 据此计算 $b^* = y_j - \sum_{i=1}^l y_i a_i^* (x_i, x_j)$;

第 4 步: 把 $w^* = \sum_{i=1}^l a_i^* y_i x_i$ 带入到超平面方程, 求得空间数据决策函数:

$$f(x) = \text{sgn}(\sum_{i=1}^l a_i^* y_i (x_i \cdot x) + b^*) \quad (14)$$

根据式(14)完成对空间数据的聚类。

3 实验结果分析

实验采用直方图相交核, 支持向量机采用软件包 Libsvm-9-0-2。空间数据聚类标记采用两种方法, 所有试验运行 6 次, 最后结果取 6 次的平均值。不同空间数据聚类方法在 scene 数据库的性能对比见表 1。

表 1 不同空间数据聚类方法在 scene 数据库的性能对比

空间数据聚类方法	RI	时间	参数
基于层次空间(完全图)	0.706 8	37 321.617 9 s	$C=0.1$
基于距离空间分割(k 相邻法)	0.723 5	153.864 1 s	$C=0.1$
基于数量统计	0.752 3	121.548 3 s	$N=8$
基于支持向量机	0.832 5	18.924 5	$\text{eps}=0.1$

在数据库中, 本文方法在时间上只用了 18.924 5 s, 因此该方法对空间数据的分类在速度上有所提升。

本次实验将本文方法与三种传统方法在数据分类的效率和准确率方面进行了分析对比。由图 1 得, 基于支持向量机的空间数据聚类方法对数据的分类速度远高于其他三种方法; 而图 2 中的准确率, 本文方法也高于其他对比方法。

通过图 1、图 2 可以看出, 本文方法不但在分类准确率上呈现上升状态, 而且在速度上也保持平稳

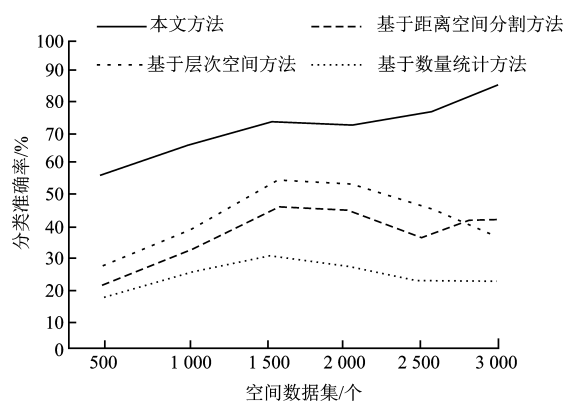


图 1 不同方法对空间数据聚类的准确率影响

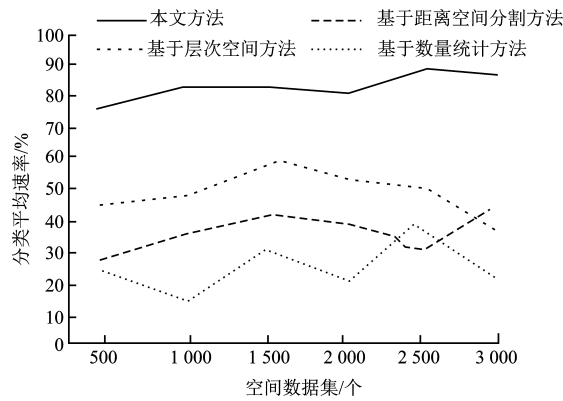


图 2 不同方法对空间数据分类速度的影响

上升的状态.

4 结束语

本文提出的基于支持向量机的空间数据聚合方法有很高的效率并且大大提升了分类的准确率. 实验结果证明该方法实用性更好.

参考文献:

[1] 白海钊,鲍长春,刘鑫. 基于局部最小二乘支持向量机的音频频带扩展方法[J]. 电子学报, 2016, 44(9): 2203-2210.

[2] 杨勃. 优化样本分布的最接近支持向量机[J]. 电子学报, 2014, 42(12): 2429-2434.

[3] 陈素根,吴小俊. 基于特征值分解的中心支持向量机算法[J]. 电子与信息学报, 2016, 38(3): 557-564.

[4] 温浩杰,于京杰. 医院科学评审监测系统的设计与实现[J]. 电子设计工程, 2016, 33(1): 36-38.

[5] 李华中,杨景花. 基于改进支持向量机的推荐系统托攻击检测方法[J]. 现代电子技术, 2016, 39(18): 96-98.

[6] 孟丽丽,宋锋. Web 网络大数据分类系统的设计与改进[J]. 现代电子技术, 2016, 39(22): 36-40.

[7] 叶春明,王珍,陈思,等. 基于节点行为特征分析的网络流量分类方法[J]. 电子与信息学报, 2014, 36(9): 2158-2165.

[8] 赵永威,李弼程,柯圣财. 基于弱监督 E2LSH 和显著图加权的目标分类方法[J]. 电子与信息学报, 2016, 38(1): 38-46.

[9] 陈仁太. 基于数据分类思维的网络危险信号检测系统设计与实现[J]. 现代电子技术, 2016, 39(10): 142-146.

[10] 张素莉. 基于部分连接神经网络的序列数据分类算法研究[J]. 现代电子技术, 2016, 39(9): 111-113.

作者简介:

王 威 男, (1980-), 硕士研究生, 讲师. 研究方向为空间数据处理、机器学习、智慧旅游.
E-mail: wei_wang888@163. com.