海量散乱点云数据的模糊聚类挖掘方法研究

陆兴华,刘文林,吴宏裕,冯飞龙
(广东工业大学华立学院,广东 广州 511325)

摘 要:物联网和云计算环境下海量散乱点云数据挖掘容易受到关联规则项的干扰,数据挖掘的模糊聚类不好。为了提高海量散乱点云数据挖掘能力,提出一种基于支持向量机的大数据分类挖掘技术。采用分段向量量化编码技术进行海量散乱点云数据空间存储结构分析,结合闭频繁项集检测方法进行海量散乱点云数据的信息融合处理,对高维融合数据进行语义特征分析和关联规则特征提取,对提取的海量散乱点云数据的关联规则采用支持向量机分类器进行模式识别,结合尺度分解方法对分类输出的海量散乱点云数据进行降维处理,采用模糊聚类方法实现对海量散乱点云数据的分类挖掘。仿真结果表明,采用该方法进行海量散乱点云数据挖掘的聚类性能较好,数据挖掘的精度较高。

关键词:海量散乱点云数据;挖掘;模糊聚类;特征提取
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2019)11-0012-05
doi:10.3969/j.issn.1673-629X.2019.11.003

Research on Fuzzy Clustering Mining Method for Massive Scattered Point Cloud Data

LU Xing-hua,LIU Wen-lin,WU Hong-yu,FENG Fei-long
(Huali College Guangdong University of Technology,Guangzhou 511325,China)

Abstract:In the environment of Internet of things and cloud computing,the mining of massive scattered point cloud data is easily disturbed by association rule items,and the fuzzy clustering of data mining is poor. In order to improve the ability of massive scattered point cloud data mining,we present a classification and mining technique for big data based on support vector machine (SVM). The segmented vector quantization coding technique is used to analyze the spatial storage structure of mass scattered point cloud data,and the information fusion processing of mass scattered point cloud data is carried out by combining the detection method of closed frequent itemsets. Semantic feature analysis and association rule feature extraction are carried out for high-dimensional fusion data. Support vector machine classifier is used to recognize the association rules of massive scattered point cloud data. Based on the scale decomposition method,the dimensionality of the massive scattered point cloud data is reduced,and the classification mining of the massive scattered point cloud data is realized by using the fuzzy clustering method. The simulation shows that the clustering performance of this method is better and the precision of data mining is higher.

Key words:massive scattered point cloud data;mining;fuzzy clustering;feature extraction

0 引 言

随着云计算技术和物联网技术的快速发展,在物联网环境中通过云存储方式进行海量散乱点云数据的集成处理,通过模糊聚类方法实现散乱点云数据的信息融合和自适应调度,提高云计算和云组合服务的质量。海量散乱点云数据的准确挖掘和分类管理是保障云服务质量的关键,采用智能挖掘和信息处理算法进行海量散乱点云数据的优化挖掘和调度,提高用户进行数据检索和管理的能力,并根据海量散乱点云数据的挖掘结果,构成最优的服务组合,提高数据检索和调度的准确性[1]。

对海量散乱点云数据的挖掘是建立在对大规模数据集的特征提取和关联规则特征分析基础上的。根据网络传输的流量特征进行海量散乱点云数据挖掘,采用相关的信息处理和数据检测方法,提高海量散乱点云数据挖掘的准确性和抗干扰能力[2]。传统方法中,

收稿日期:2018-12-18 **修回日期**:2019-04-22 **网络出版时间**:2019-06-27
基金项目:2019 年“攀登计划”广东大学生科技创新培育专项资金立项项目(pdjh2019b0616)
作者简介:陆兴华(1981-),男,硕士,副教授,通讯作者,研究方向为嵌入式技术、无人机飞行稳定性控制方法、机器人运动控制方法;刘文林(1998-),男,研究方向为计算机图形学、计算机算法应用。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190627.1105.044.html

对海量散乱点云数据的挖掘主要采用分集检测和谱分析方法[3],采用自相关特征谱分解方法进行海量散乱点云数据的信息融合和相关性检测,结合模糊数值分析和簇聚类方法实现海量散乱点云数据挖掘。根据上述原理,相关人员进行了数据挖掘算法研究。文献[4]中提出一种基于简化梯度算法的海量散乱点云数据挖掘模型,采用相关检测器进行 3D 云数据的干扰滤波,结合简化梯度算法进行云数据的输出信道均衡设计,提高数据挖掘的抗干扰能力,但该方法存在带宽受限和维数较大等问题;文献[5]中提出一种基于模糊指向性聚类的海量散乱点云数据挖掘方法,采用模糊 K 质心方法进行海量散乱点云数据的模糊加权,在保留海量散乱点云数据集内在的不确定性的条件下实现数据优化聚类,提高数据挖掘的模糊决策性,但该方法存在计算开销较大和复杂度较高的问题。

针对上述问题,文中提出一种基于支持向量机的大数据分类挖掘技术。首先采用分段向量量化编码技术进行海量散乱点云数据空间存储结构分析,结合闭频繁项集检测方法进行海量散乱点云数据的信息融合处理,然后对高维融合数据进行语义特征分析和关联规则特征提取,结合尺度分解方法对分类输出的海量散乱点云数据进行降维处理,采用模糊聚类方法实现对海量散乱点云数据的分类挖掘。最后通过仿真证明了该方法的有效性。

1　海量散乱点云数据的数据结构分析和特征提取

1.1　海量散乱点云数据的数据结构分析

为了实现对海量散乱点云数据的优化挖掘,首先分析海量散乱点云数据的数据结构和相似度特征信息。采用 C4.5 决策树模型,构建海量散乱点云数据的分类决策模型[6],进行海量散乱点云数据的相似度分解,如图 1 所示。

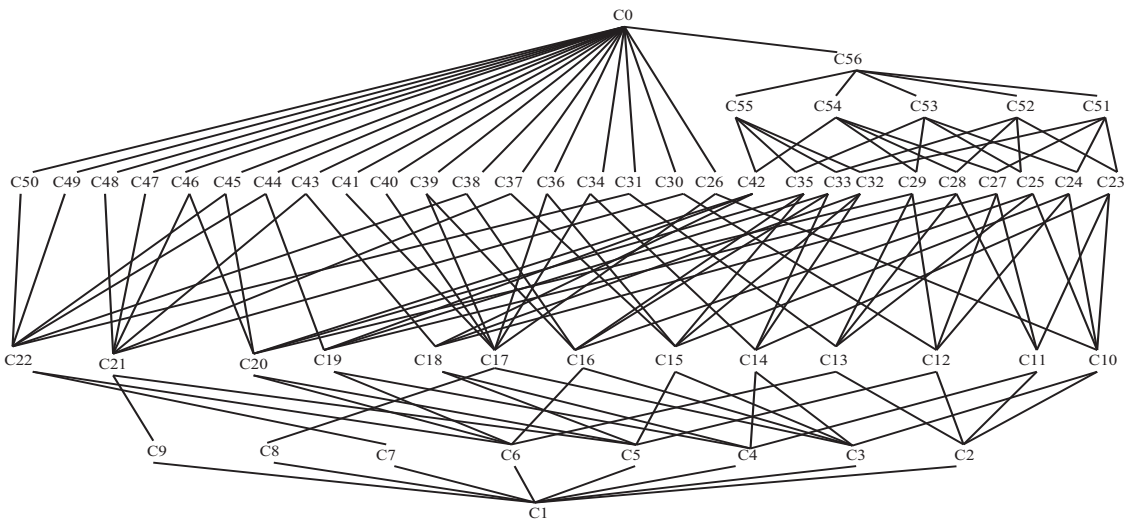

图 1　海量散乱点云数据的数据结构分解决策树模型

根据图 1 的决策树模型,对海量散乱点云数据进行模糊特征识别和数据分类,构造海量散乱点云数据的混合属性模糊分类模型[7],根据数据的混合分类属性进行相似度分析,对模糊信息的分段属性集 X 进行奇异值(SVD)分解:

$$X = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}} \tag{1}$$

其中,$\boldsymbol{U} \in R^{m\times m}$ 为语义映射的量化分解矩阵,$\boldsymbol{V} \in R^{M\times M}$ 为海量散乱点云数据的类间闭频繁项矩阵,且 $\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{U}^{-1}$,$\boldsymbol{V}^{\mathrm{T}} = V^{-1}$;$\boldsymbol{D} \in R^{m\times M}$,且满足 $\boldsymbol{D} = [\sum 0]$。在本体映射下,海量散乱点云数据分布式特征量的加权值为 $\sum = \mathrm{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_m})$,$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m$。在对海量散乱点云数据的相似度分析的基础上,提取海量散乱点云数据的数值属性特征和分类属性特征。假设 X 为具有 m 个属性的海量散乱点云数据集,第 i 个属性值的海量散乱点云数据 $y(k)$ 和分类训练数据集 $\varphi(k)$ 可以表示为:

$$\begin{cases} y(k) = s_1(k) + n_1(k) \\ \varphi(k) = s_2(k) + n_2(k) \end{cases} \tag{2}$$

$$\begin{cases} s_1(k) = AA_H \mathrm{e}^{\mathrm{j}(\Omega k + \theta_H)} \\ s_2(k) = AA_{H_B} \mathrm{e}^{\mathrm{j}(\Omega k + \theta_{H_B})} \end{cases} \tag{3}$$

其中,A_H、A_{H_B} 和 θ_H、θ_{H_B} 分别是前 p 个元素是数值属性值以及系统函数 $H(z)$ 和 $H_B(z)$ 的离散化数值属性和向量量化特征量。

求得海量散乱点云数据的语义概念集的分布矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$,取非零特征值作为训练子集,进行数据信息流模型重构。采用混合相似度特征分析方法,对海量散乱点云数据进行特征重组和向量量化分析,得到云数据特征重组后输出的平均互信息特征表达式为:

$$I(Q,S)=\sum_i\sum_j p_{sq}(s_i,q_j)\log_2[p_{sq}(s_i,q_j)/p_s(s_i)] \tag{4}$$

其中，$p_{sq}(s_i,q_j)$ 表示海量散乱点云数据的语义本体概念集 s_i 和数据概念集 q_j 的联合分布概率。

定义海量散乱点云数据的簇中的信息分布模型为 $[s,q]=[x(t),x(t+\tau)]$，得到模糊信息的闭频繁项，结合闭频繁项集检测方法进行海量散乱点云数据的信息融合处理[8]。

1.2 闭频繁项特征提取

假定当前海量散乱点云数据分布节点的数目为 n，$N_1,\cdots,N_n$，待挖掘的大数据在链路层中的负载为 $L_1,\cdots,L_n$，海量散乱点云数据挖掘输出的特征估计值为 $P_1^{\min},\cdots,P_n^{\min}$，在线性规划模型中对海量散乱点云数据进行特征分解和闭频繁项挖掘，用一个四元组 G 表示海量散乱点云数据的模糊分布式结存储中心，即 $G=(V,E,W,C)$。假设 d 为海量散乱点云数据交互的相空间嵌入维数，采用多个非线性成分联合统计方法进行海量散乱点云数据的高维特征空间重构，结合模糊聚类方法进行大数据的自适应分类处理[9]，构建一个关联规则项约束方程表达海量散乱点云数据的信息流模型：

$$x_n=x(t_0+n\Delta t)=h[z(t_0+n\Delta t)]+\omega_n \tag{5}$$

其中，$h(\cdot)$ 为海量散乱点云数据分布式时间序列，表示为一个具有多维数据结构模型的函数；ω_n 为大数据的测量误差。

构建海量散乱点云数据分布的时态结构模型，将挖掘的海量散乱点云数据按照五元组进行关联规则项特征重建，海量散乱点云数据的分布结构模型的分布函数描述式为：

$$X_p(u)=\begin{cases}p\sqrt{\dfrac{1-j\cot\alpha}{2\pi}}e^{\frac{u^2}{2}\cot\alpha}\displaystyle\int_{-\infty}^{+\infty}x(t)e^{\frac{t^2}{2}\cot\alpha-jtu\csc\alpha}\mathrm{d}t,\alpha\neq n\pi\\ x(u),\alpha=2n\pi\\ x(-u),\alpha=(2n\pm1)\pi\end{cases} \tag{6}$$

其中，p 为分布式海量散乱点云数据存储结构的阶数；α 为统计信息采样的频繁项集。

采用统计回归分析方法进行海量散乱点云数据的闭频繁项检测[10]，检测模型表达如下：

$$\min(f)=\sum_{i=1}^m\sum_{j=1}^n C_{ij}X_{ij}$$

$$\text{s.t}\begin{cases}\sum_{j=1}^m X_{ij}=a_i,i=1,2,\cdots,m\\ \sum_{i=1}^m X_{ij}=b_i,j=1,2,\cdots,n\\ X_{ij}\geqslant 0,i=1,2,\cdots,m,j=1,2,\cdots,n\end{cases} \tag{7}$$

结合闭频繁项集检测方法进行海量散乱点云数据的信息融合处理，构造海量散乱点云数据挖掘的线性规划模型[11]。

2 数据模糊聚类挖掘实现

2.1 数据模糊聚类处理

在采用分段向量量化编码技术进行海量散乱点云数据空间存储结构分析的基础上，对高维融合数据进行语义特征分析和关联规则特征提取和模糊聚类处理。采用分段向量量化编码技术进行海量散乱点云数据空间存储结构分析和关联规则特征提取[12]，构建需要挖掘的海量点云数据的量化编码分析模型：

$$\begin{cases}\min\displaystyle\sum_{1\leqslant i\leqslant Ke\subseteq k(e)}\sum\frac{f(e(i))}{C(e,i)}\\ 0\leqslant f(e,i)\leqslant C(e,i)\\ F=\text{const}\\ \displaystyle\sum_{1\leqslant i\leqslant K,e\subseteq k(e)}\frac{f(e(i))}{C(e,i)}+\sum_{e\subseteq k(e)}\frac{f(e'(i))}{C(e',i)}\leqslant k(v)\end{cases} \tag{8}$$

给出海量散乱点云数据数据结构的特征标识函数 $P_C=\sum_{i=0}^n\sum_{j=0}^n\alpha(i,j)P(i,j)$。假设每个分类属性值的初始码元为 $C_0=C_{N/2}=0$、$C_{N-n}=C_n^*$，$n=0,1,\cdots,N/2-1$，海量散乱点云数据的数据对象和簇中心分布关系模型为：

$$P_r=\frac{P_t}{(4\pi)^2(d/\lambda)^\gamma}[1+\alpha^2+2\alpha\cos(4\pi h^2/d\lambda)] \tag{9}$$

根据数据的不同属性在聚类的差异性，进行海量散乱点云数据特征识别[13]。数值属性特征和分类属性特征分别为：

$$R_\beta X=U\{E\in U/R\mid c(E,X)\leqslant\beta\} \tag{10}$$

$$R_\beta X=U\{E\in U/R\mid c(E,X)\leqslant 1-\beta\} \tag{11}$$

对于第 ii 个分类属性的两个数据块 m_i 和 m_j，得到数据分类的模糊质心 $u\in L_{t,x}^{d+1}(K\times \mathrm{IR}^d)$，采用模糊 C 均值聚类分析方法，得到数据模糊聚类的迭代过程为：

$$S_b=\sum_{i=1}^e p(\omega_i)(u_i-u)(u_i-u)^{\mathrm{T}} \tag{12}$$

$$S_\omega=\sum_{i=1}^e p(\omega_i)E[\frac{(u_i-u)(u_i-u)^{\mathrm{T}}}{\omega_i}] \tag{13}$$

$$S_i=S_b+S_\omega \tag{14}$$

其中，$p(\omega_i)$ 为数据挖掘的分配规则向量集；$\mu=E(x)$ 为散乱点云数据的分布稀疏度。

2.2 基于支持向量机的数据挖掘

文中提出一种基于支持向量机的大数据分类挖掘技术，采用自适应加权算法，得到支持向量机进行大数据特征分类器的加权系数为：

$$w_{sij}(n_0+1)=w_{sij}(n_0)-\eta_{sij}\frac{\partial J}{\partial w_{sij}} \tag{15}$$

采用支持向量机的学习算法[14],得到海量散乱点云数据分类的自适应学习过程为:

$$\alpha_{\text{desira}}^{i}=\alpha_1\cdot\frac{\text{Density}_i}{\sum_i \text{Density}_i}+\alpha_2\cdot\frac{\text{AP}_i}{\text{AP}_{\text{init}}} \tag{16}$$

在 $B\Rightarrow D$, $A\cap B\Rightarrow D$ 等规则约束项下,得到海量散乱点云数据模糊挖掘的量化参数满足:

$$\begin{cases}\alpha_1+\alpha_2=1,\alpha_1,\alpha_2\in[0,1]\\ \alpha_2=\dfrac{\max\limits_i(\text{AP}_i)-\min\limits_i(\text{AP}_i)}{\text{AP}_{\text{init}}}\end{cases} \tag{17}$$

数据的统计量化集为 $(u,v)\in E$,设 $A\subset V$, $B\subset V$ 且 $A\cap B=\varnothing$,采用支持向量机分类器进行模式识别,实现对海量散乱点云数据重组和数据结构重排。对高维融合数据进行语义特征分析和关联规则特征提取,对提取的海量散乱点云数据的关联规则采用支持向量机分类器进行模式识别[15],数据准确挖掘的概率密度函数为:

$$P_{\text{S}}=p_{2\text{D}}^{k}(1-p_{2\text{D}})^{N-1-k}\sum_{i=1}^{\infty}\lambda_{\text{S}}^{i}=\frac{\lambda_{\text{S}}}{1-\lambda_{\text{S}}} \tag{18}$$

其中, λ_{S} 为在采样时刻进行数据采集的相似度系数; $p_{2\text{D}}$ 为簇中的信息分布概率密度。

海量散乱点云数据簇中心之间的相异度为:

$$\text{DisSim}(A,B)=1-\left|\frac{\text{SameDis}(A)-\text{SameDis}(B)}{\text{Dis}(A)+\text{Dis}(B)}\right| \tag{19}$$

其中, $\text{Dis}(A)$ 表示聚类中心的欧氏距离; $\text{Dis}(B)$ 表示语义本体集。

采用基于模糊质心相异性度量方法构建海量散乱点云数据的分类模糊集。根据上述分析,实现了海量散乱点云数据的模糊聚类挖掘。

3　仿真实验分析

通过仿真实验测试文中方法在实现海量散乱点云数据优化挖掘中的应用性能。实验采用 Matlab 设计,测试数据集选用 KTT 数据集,实验中的大数据样本库采用 Olivetti-Oracle Research Lab (ORL)海量散乱点云数据库,每个高维融合数据子块阈值 $Y_{\text{HW}}=0.15$,对海量散乱点云数据采样的占空比为 0.34,样本训练集规模为 26 kbps,海量散乱点云测试集为 100 kbps,稀疏度为 0.56。根据上述仿真环境和参数设定,进行海量散乱点云数据模糊聚类和挖掘仿真,得到数据采样的时域分布如图 2 所示。

采用分段向量量化编码技术进行海量散乱点云数据的信息融合,实现数据模糊聚类和挖掘,得到的挖掘结果输出如图 3 所示。

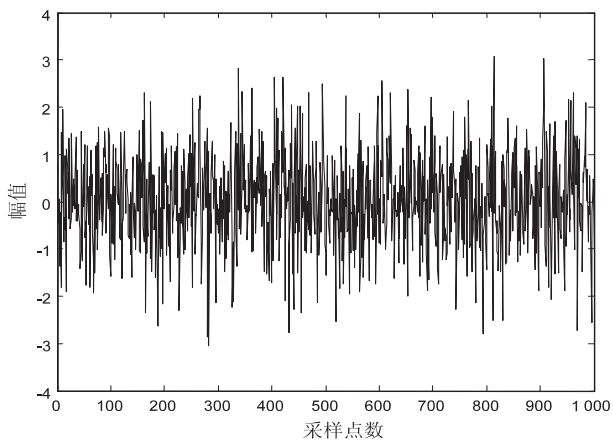

图 2　数据采样的时域分布

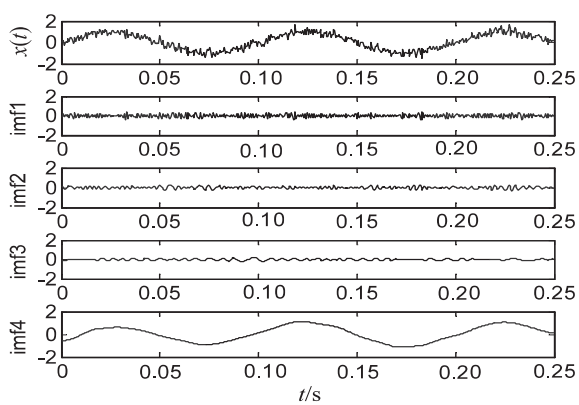

图 3　数据模糊聚类挖掘输出

分析图 3 得知,采用文中方法能有效实现对海量散乱点云数据的分类挖掘,特征的聚类性较好。测试不同方法进行数据挖掘的召回率,得到的对比结果如图 4 所示。

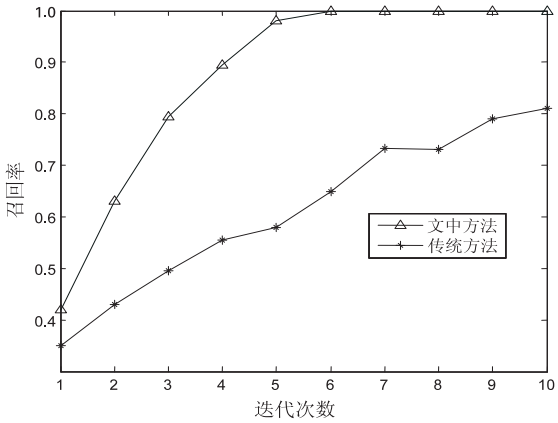

图 4 数据挖掘的召回性对比

分析图 4 得知,文中方法进行数据挖掘的召回率较高,说明数据挖掘精度较高,挖掘的收敛性较好,具有很好的模糊聚类挖掘性能。

4 结束语

文中提出一种基于支持向量机的大数据分类挖掘技术。采用分段向量量化编码技术进行海量散乱点云数据空间存储结构分析,结合闭频繁项集检测方法进行海量散乱点云数据的信息融合处理,对高维融合数据进行语义特征分析和关联规则特征提取。对提取的海量散乱点云数据的关联规则采用支持向量机分类器进行模式识别,结合尺度分解方法对分类输出的海量散乱点云数据进行降维处理,采用模糊聚类方法实现对海量散乱点云数据的分类挖掘。仿真结果表明,该方法进行数据挖掘的召回性能较好,挖掘精度较高。

参考文献:

[1] 何 华,李宗春,李国俊,等. 散乱点云的自适应 α-shape 曲面重建[J]. 计算机应用,2016,36(12):3394-3397.

[2] 袁 泉,郭江帆. 新型含噪数据流集成分类的算法[J]. 计算机应用,2018,38(6):1591-1595.

[3] YIN Chuanlong, ZHU Yuefei, FEI Jinlong, et al. A deep learning approach for intrusion detection using recurrent neural networks[J]. IEEE Access,2017,5:21954-21961.

[4] 谷 琼,袁 磊,熊启军,等. 基于非均衡数据集的代价敏感学习算法比较研究[J]. 微电子学与计算机,2011,28(8):146-149.

[5] 杨雅辉,黄海珍,沈晴霓,等. 基于增量式 GHSOM 神经网络模型的入侵检测研究[J]. 计算机学报,2014,37(5):1216-1224.

[6] 高 妮,贺毅岳,高 岭. 海量数据环境下用于入侵检测的深度学习方法[J]. 计算机应用研究,2018,35(4):1197-1200.

[7] 毛文涛,田杨阳,王金婉,等. 面向贯序不均衡分类的粒度极限学习机[J]. 控制与决策,2016,31(12):2147-2154.

[8] 李 涛,王次臣,李华康. 知识图谱的发展与构建[J]. 南京理工大学学报:自然科学版,2017,41(1):22-34.

[9] FERCOQ O,RICHTÁRIK P. Accelerated,parallel and proximal coordinate descent[J]. SIAM Journal on Optimization,2014,25(4):1997-2023.

[10] LOW Y,BICKSON D,GONZALEZ J,et al. Distributed GraphLab:a framework for machine learning and data mining in the cloud[J]. Proceedings of the VLDB Endowment,2012,5(8):716-727.

[11] 陈 光. 基于大数据的数据服务应用研究[J]. 计算机技术与发展,2018,28(8):129-134.

[12] 郭华平,董亚东,毛海涛,等. 一种基于逻辑判别式的稀有类分类方法[J]. 小型微型计算机系统,2016,37(1):140-145.

[13] 史皓良,吴禄慎,余喆琦,等. 散乱点云数据特征信息提取算法[J]. 计算机工程,2017,43(8):279-283.

[14] KHALILI A,SAMI A. SysDetect:a systematic approach to critical state determination for industrial intrusion detection systems using Apriori algorithm[J]. Journal of Process Control,2015,32:154-160.

[15] LONG Mingsheng,WANG Jianmin,DING Guiguang,et al. Adaptation regularization:a general framework for transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering,2014,26(5):1076-1089.