计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(3):180−186 [doi: 10.15888/j.cnki.csa.007322]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于层次化深度学习的医疗数据库离群数据检测算法①

李晓峰[1], 王妍玮[2], 李　东[3]

[1](黑龙江外国语学院 信息工程系, 哈尔滨 150025)
[2](普渡大学 机械工程系, 西拉法叶市 IN47906)
[3](哈尔滨工业大学 计算机科学与技术学院, 哈尔滨 150001)

摘　要: 对医疗数据库中存在的离散数据进行检测时, 由于缺少数据过滤等过程而导致检测执行时间较长、检测效率低、离散点检测率低等问题, 为此提出基于层次化深度学习的医疗数据库离散数据检测算法. 首先, 采用动态网格划分法划分空间中的稀疏区域和稠密区域, 降低数据检测的规模, 缩短检测执行时间; 然后, 通过层次化深度学习过程融合专家知识和数据的属性取值分布信息, 实现医疗数据库中离散数据的检测. 实验结果表明, 该算法可以在较短的时间内准确完成医疗数据库中离散数据的检测, 且相较于传统算法来说更具有应用优势.

关键词: 层次化深度学习; 医疗数据; 离群点; 离群数据检测; 动态网格划分

引用格式: 李晓峰,王妍玮,李东.基于层次化深度学习的医疗数据库离群数据检测算法.计算机系统应用,2020,29(3):180–186. http://www.c-s-a.org.cn/1003-3254/7322.html

Medical Database Outlier Data Detection Algorithm Based on Hierarchical Deep Learning

LI Xiao-Feng[1], WANG Yan-Wei[2], LI Dong[3]

[1](Department of Information Engineering, Heilongjiang International University, Harbin 150025, China)
[2](Department of Mechanical Engineering, Purdue University, West Lafayette IN47906, USA)
[3](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: When using the current algorithm to detect the discrete data in the medical database, problems such as long execution time, low detection efficiency and low detection rate of discrete points are caused by the lack of data filtering and other processes. Therefore, an algorithm for detecting discrete data in the medical database based on hierarchical deep learning is proposed. Firstly, the dynamic grid method is used to divide the sparse and dense areas in the space, so as to reduce the size of data detection and shorten the detection execution time. Then, the expert knowledge and data attribute value distribution information are integrated through the hierarchical deep learning process, and realize the detection of discrete data in medical database. Experimental results show that this algorithm can accurately complete the detection of discrete data in the medical database in a relatively short time, and has more advantages in application compared with the traditional algorithm.

Key words: hierarchical deep learning; medical data; outlier; outlier data detection; dynamic mesh generation

1　引言

数据库管理系统和信息技术在近年来得以快速发展, 人们收集和产生数据的能力不断提高, 医疗数据库中存在的数据量呈直线增长. 过去对数据的检测分析

① 基金项目: 国家自然科学基金 (61803117); 教育部科技发展中心产学研创新基金 (2018A01002); 国家科技部创新方法专项 (2017IM010500)
Foundation item: National Natural Science Foundation of China (61803117); Industry-University-Research Innovation Fund of Science and Technology Development Center, MOE (2018A01002); Special Fund for Innovation Methodology of Ministry of Science and Technology (2017IM010500)
收稿时间: 2019-08-03; 修改时间: 2019-09-02, 2019-09-11; 采用时间: 2019-09-18; csa 在线出版时间: 2020-02-28

主要通过分析员完成, 在专家意见的基础上通过数据分析在医疗数据库中获取和查询数据, 由分析员决定数据分析的结果. 但由于数据库中的数据急剧膨胀, 数据的复杂性和时效性也不断增强, 传统方法已经不能满足人们的要求. 为了从医疗数据库中获取有用的信息, 需要改进现有的数据检测技术.

在医疗数据库中存在一些与其他数据行为不同, 或是与其他数据差异较大的数据, 被称为离群数据. 离群数据中通常存在有用的信息, 因此需要对医疗数据库中存在的离群数据进行检测, 众多学者进行了相关研究, 并取得了一定的成果.

Hauskrecht M 等[1]通过对数据离群点检测实现异常患者管理, 该方法通过使用 EMR 存储库来学习将患者状态与病人管理操作相关联的统计模型, 使用电子病历保存患者信息, 通过与以往病历的异常分析, 获取异常患者行为, 但该方法的计算代价较大; Yu YW 等[2]提出了一种新的基于邻域轨迹离群点的分类方法, 对研究对象真实数据集进行理论分析和实证研究, 验证了本文方法在捕获不同类型数据的有效性, 但该方法的离群点检测率不高, 且误差率较高; Jobe JM 等[3]提出一种基于计算机的数据集群方法, 将 Rousseuw 的最小协方差行列式方法的重加权版本与最初基于多步聚类的算法结合起来, 找出离群点, 实验结果表明, 该方法稳健性较好, 但是离群点检测率较低, 计算代价大; 邹云峰等[4]提出基于局部密度的数据库离散数据检测算法, 该算法将弱 k 近邻点和强 k 近邻点概念引入离散数据检测中, 对邻近数据点在数据库中的离群相关性进行分析, 根据分析结果区别对待数据点, 通过数据点离群性预判方法完成医疗数据库离群数据的检测, 该算法检测离散数据的执行时间较长, 存在检测效率低的问题. 李少波等[5]提出基于密度的数据库离群数据检测算法, 该算法在离群数据检测过程中引入滑动时间窗口, 通过滑动时间窗口划分数据, 计算数据的信息熵, 根据计算结果对数据进行筛选和剪枝, 通过离群因子对筛选后的数据进行判断, 完成数据库离散数据的检测, 该算法计算得到的离群因子存在误差, 不能准确的对医疗数据库中的数据进行判断, 存在离散点误差率高的问题. 魏畅等[6]提出基于约简策略的数据库离散数据检测算法, 该算法在马氏距离标准的基础上对数据集进行简约处理, 通过数据流时间相关性和数据分布密度准则构建决策模型, 通过决策模型对数据库中存在的离散数据进行检测, 该算法构建的决策模型精准度较低, 导致离散点检测率低. 尹娜等[7]提出了一种基于混合式聚类算法的离群点挖掘在异常检测中的应用方法, 该方法通过 k-中心点算法找出簇中心, 在此基础上去除其中较隐秘的数据样本, 再结合基于密度的聚类算法计算出离群数据的异常度, 从而判断出离群点. 但是该算法在挖掘隐秘样本时出错率较高, 致使最终的检测结果存在较大误差.

针对目前现有方法中存在的离群数据检测过程执行时间较长、检测效率低、离群点检测率低的问题, 提出基于层次化深度学习的医疗数据库离群数据检测算法. 在对空间中的稀疏区域和稠密区域进行划分再合并, 实现数据过滤, 通过层次化深度学习过程融合专家知识增强对离群数据的多层感知, 实现对离群数据的检测, 达到降低算法计算代价、降低耗时、提高检测率和准确率的目的.

2 动态网格划分与合并

医疗数据库中存在海量的数据, 在对其中的离群点检测之前, 本文基于层次化深度学习的医疗数据库离群数据检测算法首先使用动态网格划分方法对医疗数据库中的数据进行筛选, 构建候选离群数据集, 以此来达到缩小检测规模、减少检测执行时间的目的.

动态网格划分方法是根据医疗数据库中数据流的密度特点对数据做网格分裂及合并处理, 按照密度大小对数据库空间中的数据进行分类, 划分为稀疏区域和稠密区域, 对稠密区域中存在的大量主体数据进行分析, 存储有较大概率成为离群点的数据并构建候选离群点集合[8,9].

将较小的权重赋予给历史数据, 降低历史数据对网格划分的影响, 使当前数据在数据库中的分布情况能够更好的通过网格进行反应[10].

设$Cell(C,\vec{S}^1,\vec{S}^2,n_C,O_C,t_{la})$为六元组, 用来表示网格, 其中$C$代表的是$k$维数据空间中存在的超方体; n_C代表的是落在超方体C中的数据点总数; O_C代表的是网格中存在的候选离群点集合; t_{la}代表的是时间; 网格统计信息$\vec{S}^1=[s_1^1,\cdots,s_k^1]$, 其中元素$s_k^1$的计算公式如下:

$$s_k^1=\sum_C \theta^{t_a-t_c} r_i \tag{1}$$

式中, r_i代表的是数据点. 网格统计信息$\vec{S}^2=[s_1^2,\cdots,s_k^2]$,

元素s_k^2的计算公式如下:

$$s_k^2=\sum_C \theta^{t_a-t_c} r_i^2 \tag{2}$$

网格统计信息$\vec{S}^1$、$\vec{S}^2$在t时刻满足下式:

$$\vec{S}_t^1=\theta^{t-t_{la}}\times\vec{S}_{t_{la}}^1 \tag{3}$$

$$\vec{S}_t^2=\theta^{t-t_{la}}\times\vec{S}_{t_{la}}^2 \tag{4}$$

设t_c代表的是当前时间. 根据上述性质, 增量更新数据$\vec{r}$在网格C中对应的统计信息如下:

$$n_C=\theta^{t_c-t_{la}}\times n_C+1 \tag{5}$$

$$s^1=\theta^{t_c-t_{la}}\times s_i^1+r_i \tag{6}$$

$$s^2=\theta^{t_c-t_{la}}\times s_i^2+r_i^2 \tag{7}$$

$$t_{la}=t_c \tag{8}$$

在初始化处理时, 对数据的网格进行分割, 获得初始网格, 根据网格统计信息$\vec{S}^1$、$\vec{S}^2$, 可以计算得到数据在网格中对应的平均值μ_i和标准偏差σ_i:

$$\mu_i=\frac{S_i^1}{n_C} \tag{9}$$

$$\sigma_i=\sqrt{\frac{S_i^2-2\mu_i S_i^1}{n_C}+\mu_i^2} \tag{10}$$

如果网格的密度达到设定的阈值时, 分割网格. 将数据聚集并划分到对应的网格中是网格分裂合并的原则[11]. 所以保存每个维度上网格对应的方差和均值, 选择最大方差相应的维度, 在均值处做划分处理, 可以在两个新生成的网格中划入数据.

将$Cell(C,\vec{S}^1,\vec{S}^2,n_C,O_C,t_{la})$视为分割后的网格, $\sigma_j=\max(\sigma_1,\sigma_2,\cdots,\sigma_k)$, 在该维度上分割网格, 针对其余维度, 保持对应的$\vec{S}^1$、$\vec{S}^2$不发生变化[12]. 假设在正态分布下, 对网格$Cell$在平均值μ_j处进行分割, 获得$Cell_1$、$Cell_2$, 此时:

$$n_{C_1}=n_{C_2}=\frac{n_C}{2} \tag{11}$$

$$t_{la_1}=t_{la_2}=t_{la} \tag{12}$$

$$\sigma_j^{C_1}=\sqrt{\int_{\min_j}^{\mu_j} x^2\frac{1}{\sqrt{2\pi}\sigma_j}e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}}dx-\left(\mu_j^{C_i}\right)^2} \tag{13}$$

$$\sigma_j^{C_2}=\sqrt{\int_{\mu_j}^{\max_j} x^2\frac{1}{\sqrt{2\pi}\sigma_j}e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}}dx-\left(\mu_j^{C_2}\right)^2} \tag{14}$$

式中, $\min_j$代表的是第j维度上在网格中存在的最小值; $\max_j$代表的是第j维度上在网格中存在的最大值.

对分割得到的网格$Cell_1(C_1,\vec{S}_{C_1}^1,\vec{S}_{C_1}^2,n_{C_1},O_{C_1},t_{la1})$、$Cell_2(C_2,\vec{S}_{C_2}^1,\vec{S}_{C_2}^2,n_{C_2},O_{C_2},t_{la2})$进行合并, 得到网格$Cell'(C,\vec{S}^1,\vec{S}^2,n_C,O_C,t_{la})$, 在网格$C'$中对应的统计信息如下:

$$n_C'=\theta^{t_c-t_{la1}}\times n_{C_1}+\theta^{t_c-t_{la2}}\times n_{C_2} \tag{15}$$

$$\vec{S}^{1\prime}=\theta^{t_c-t_{la1}}\times\vec{S}_{C_1}^1+\theta^{t_c-t_{la2}}\times\vec{S}_{C_2}^1 \tag{16}$$

$$\vec{S}^{2\prime}=\theta^{t_c-t_{la1}}\times\vec{S}_{C_1}^2+\theta^{t_c-t_{la2}}\times\vec{S}_{C_2}^2 \tag{17}$$

$$t_{la}'=t_c \tag{18}$$

通过对网格进行划分再合并, 能够去除数据集中的非离群数据, 保证剩余的数据均为离群数据, 从而实现数据过滤, 有效降低算法计算代价和复杂度, 节约耗时提高医疗数据库离群数据检测的效率.

3 医疗数据库离群数据层次深度学习检测

医疗数据库中, 针对数据类别的确定有多种方式, 可依据不同设备采集到的数据进行分类, 可依据不同种类疾病进行数据分类, 还可依据不同身体部位进行数据分类等, 只有依据同一分类方式获取得到的医疗数据才具有实际意义. 因此, 本文提出了基于深度学习的医疗数据分类和检测框架, 在每一分类层次上都能够实现数据检测, 即采用层次化深度学习方法对医疗数据库中存在的离群数据进行检测.

现有的离群数据检测算法一般都是根据专家经验设定对象邻域半径, 结果随机性和主观性较大[13]. 本文所提的基于层次化深度学习的医疗数据库离群数据检测算法中, 深度学习是基于模拟人脑进行学习的一种神经网络, 本文采用一种基于卷积神经网络的深度网络结构进行离群数据检测; 层次化是指包含了专家知识层次和数据属性取值分布信息层次两部分, 依据这两者构建深度网络分类器, 有效感知离群数据, 提高离群数据检测结果的准确率. 基于层次化深度学习的离群数据检测结构框架如图 1 所示.

根据图 1 可知, 层次化深度学习检测框架中, 基于专家知识和数据属性取值分布信息这两个层次分类,

构建了深度网络分类器. 接下来主要通过对数据差异度量来训练分类器, 从而实现离群数据检测, 具体过程如下:

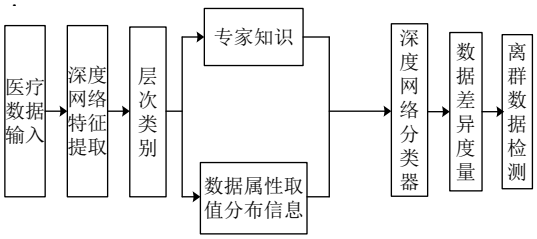

图 1 层次化深度学习检测框架

医疗数据库离群数据存在混合型属性值和数据型属性值, 为了有效的对两者之间存在的差异进行度量, 主要通过度量邻域距离实现[13,14]. 设$HEOM_B(x,y)$代表的是重叠度量值, 其计算公式如下:

$$HEOM_B(x,y)=\sqrt{\sum_{i=1}^{k}d_{c_{ji}}(x,y)^2} \tag{19}$$

式中, 参数$d_{c_{ji}}(x,y)$的计算公式如下:

$$d_{c_{ji}}(x,y)=\begin{cases}0, & \text{if } c_{ji}\text{为符号型属性,且}f(x,c_{ji})=f(y,c_{ji})\\ 1, & \text{if } c_{ji}\text{为符号型属性,且}f(x,c_{ji})\neq f(y,c_{ji})\\ |f(x,c_{ji})-f(y,c_{ji})|, & \text{if } c_{ji}\text{为数值型属性}\end{cases} \tag{20}$$

通过式 (22) 确定邻域半径ε_{cj}:

$$\varepsilon_{cj}=\begin{cases}0, \text{ if } c_j\text{为符号型属性}\\ \dfrac{std(c_j)}{\lambda}, \text{ if } c_j\text{为数值型属性}\end{cases} \tag{21}$$

式中, $std(c_j)$代表的是属性c_j取值时对应的标准差, 可以通过该标准差对属性均值的分散程度进行衡量[15]. 如果标准差$std(c_j)$的值较大时, 表明在属性c_j上大部分数据的均值和取值之间存在的差异较大; 如果$std(c_j)$的值较小时, 表明在属性c_j上大部分数据的均值和取值之间存在的差异较小[16,17].

λ代表的是专家设定的参数, 邻域半径的大小可以通过参数λ进行调整[18].

设$VDM(x,y)$代表的是差异度量值, 其计算公式为:

$$VDM(x,y)=\sum_{P}d_f(x_f,y_f) \tag{22}$$

式中, x, y为对象集中存在的对象; P 代表的是对象集对应的特征集; $d_f(x_f,y_f)$代表的是x_f、y_f之间存在的距离.

为了确定数据在数据库中的离群程度, 离群度量数据型属性的取值[19,20]. 用$NVDM(x_i,x_j)$代表某存在对象x_i和y_i之间的邻域值差异度量值, 设NOF代表的是邻域离群因子, 其计算公式如下:

$$NOF=\sum_{i,j=1}^{n}\sqrt{NVDM(x_i,x_j)-VDM(x,y)} \tag{23}$$

设μ代表的是预设的离群点判定阈值, 对比邻域离群因子NOF与阈值μ的大小. 如果满足如下条件, 则该数据为离群数据, 否则为离群数据. 对所有的数据判断完, 即完成了对医疗数据库中离群数据的检测.

$$NOF>\mu \tag{24}$$

4 实验分析与结果

为了验证基于层次化深度学习的医疗数据库离群数据检测算法的整体有效性, 需要对其进行测试.

实验条件设置如表 1 所示.

表 1 实验条件设置情况

参数	配置
实验平台	Matlab
CPU	Intel Core i5-2400 CPU、4 GB 内存、8.5 GB 硬盘
开发环境	MyEclipse2014+JDK1.7
编程环境	Matlab R2015b
操作系统	Windows8

实验数据: 本文使用 UCI 机器学习库中的 Annealing 和 Wisconsin Breast Cancer 数据集 (网址: http://archive.ics.uci.edu/ml/). 为增强实验说服力, 将本文所提的基于层次化深度学习的医疗数据库离群数据检测算法 (算法 1) 与文献[2](算法 2)、文献[3](算法 3)、文献[4]中的基于局部密度的数据库离散数据检测算法 (算法 4)、文献[5]中的基于密度的数据库离群数据检测算法 (算法 5)、文献[6]中的基于约简策略的数据库离散数据检测算法 (算法 6)、文献[7]中的基于混合式聚类算法的离群点挖掘在异常检测中的应用方法 (算法 7) 进行对比测试.

实验选取的评价指标及计算方式如下:

(1) 计算代价: 数据在实际应用中, 由于过滤不佳或其他问题, 易导致错误率增加, 加大计算代价, 本实验以计算代价为指标进行分析, 选取代价权值体现不同算法的计算代价情况, 代价权值越高, 计算代价越大.

(2) 检测时间: 在迭代次数相同的条件下, 测试本文算法和算法 4、算法 5、算法 6、算法 7 等 5 种不同算法检测离群数据的执行时间, 执行时间越短证明检

测效率越高.

(3) 离群点检测率: 为了进一步验证本文所提的基于层次化深度学习的医疗数据库离群数据检测算法的整体有效性, 将离群点检测率作为对比指标进行实验, 计算方法如下:

设 L 代表的是离群点检测率, 其计算公式如下:

$$L = \frac{N_l}{N_z} \tag{25}$$

式中, N_l代表的是检测出正确的离群点总数; N_z代表的是数据集中存在的离群点总数.

(4) 离群点误差率: 将离群点误差率作为对比指标, 对基于层次化深度学习的医疗数据库离群数据检测算法、算法 2、算法 5、算法 6、算法 7 进行测试.

设 W 代表的是离群点误差率, 其计算公式如下:

$$W = \frac{M_1 - M_2}{S - M_2} \tag{26}$$

式中, M_1代表的是输出的离群点总数; M_2代表的是正确离群点总数; S代表的是数据集总数.

4.1 计算代价对比

对本文基于层次化深度学习的医疗数据库离群数据检测算法与算法 2、算法 3、算法 4 进行对比, 结果如图 2 所示.

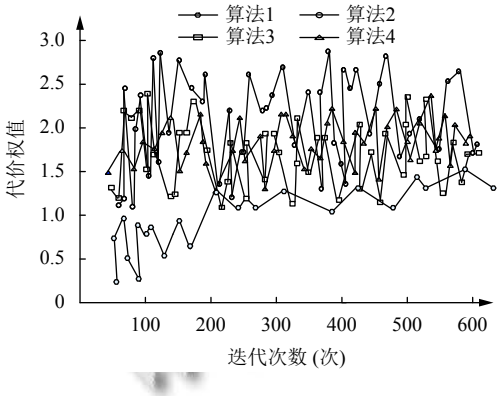

图 2 计算代价对比

分析图 2 可以看出, 本文基于层次化深度学习的医疗数据库离群数据检测算法的计算代价明显较低, 代价权值不超过 1.5, 而算法 2、算法 3、算法 4 的代价权值集中在 1.0~3.0 之间, 算法 2 最高, 代价权值多在 2.5 以上, 由此可以看出, 本文算法的计算代价小, 具有一定的优势. 因为本文算法通过对网格进行划分再合并, 去除了数据集中的非离群数据, 即进行了数据过滤, 有效提高了数据质量, 降低了计算代价.

4.2 检测时间对比

在迭代次数相同的条件下, 5 种不同算法检测离群数据的执行时间测试结果如图 3 所示.

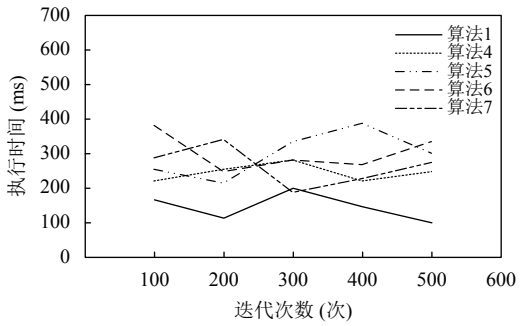

图 3 5 种不同算法的执行时间对比

分析图 3 可知, 随着迭代次数的不断增加, 不同算法的在检测离群数据时的执行时间也在不断发生变化. 其中, 本文所提的基于层次化深度学习的医疗数据库离散数据检测算法在多次迭代中的最多执行时间为 200 s, 其执行时间折线仅在迭代次数为 300 次时与基于混合式聚类算法的离群点挖掘在异常检测中的应用方法的执行时间折线相交, 证明该算法的执行时间明显少于基于局部密度的数据库离散数据检测算法、基于密度的数据库离群数据检测算法、基于约简策略的数据库离散数据检测算法、基于混合式聚类算法的离群点挖掘在异常检测中的应用方法的执行时间. 这是主要因为基于层次化深度学习的医疗数据库离群数据检测算法采用动态网格划分方法对数据进行筛选, 有效缩小了数据检测的范围和规模, 因此节省了检测数据所用的时间, 大大提高了检测效率.

4.3 离群点检测率对比

对基于层次化深度学习的医疗数据库离群数据检测算法、算法 2、算法 3、算法 6、算法 7 进行测试.

基于层次化深度学习的医疗数据库离群数据检测算法、算法 2、算法 3、算法 6、算法 7 的离群点检测率计算结果如表 2 所示.

表 2 5 种不同算法的离群点检测率测试结果 (%)

算法	迭代次数 (次)				
	100	200	300	400	500
算法 1	97	96	98	99	98
算法 2	82	86	83	81	83
算法 3	79	78	74	75	73
算法 6	70	71	68	66	72
算法 7	85	89	76	73	88

为了更直观、清晰地对比不同算法的离群点检测率, 将表 2 中的数据用折线图的形式表现, 如图 4 所示.

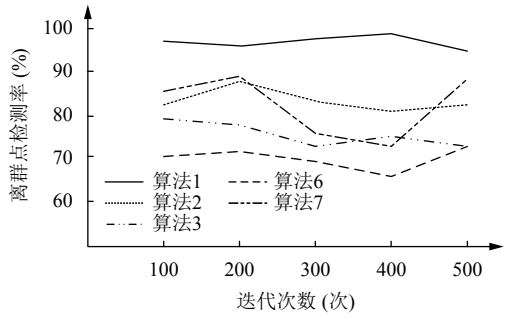

图 4 5 种不同算法的离群点检测率对比

分析表 2 和图 4 中的数据可知, 在 5 次不同迭代中, 本文所提的基于层次化深度学习的医疗数据库离群数据检测算法的平均离群点检测率为 97.6%, 算法 4 的平均离群点检测率为 83.0%, 算法 5 的平均离群点检测率为 75.8%, 算法 6 的平均离群点检测率为 69.4%, 算法 7 的平均离群点检测率为 82.2%. 对比 5 种不同算法的离群点检测率可知, 基于层次化深度学习的医疗数据库离群数据检测算法的离群点检测率始终高于另外 4 种算法, 进一步证明了本文所提算法的有效性. 究其原因, 是因为本文算法基于多层次深度学习进行离群数据检测, 融合了卷积神经网络和层次分类两者的优势, 有效提高了算法的离群点检测率.

4.4 离群点误差率对比

基于层次化深度学习的医疗数据库离群数据检测算法、算法 2、算法 5、算法 6、算法 7 的离群点误差率计算结果如表 3 所示.

表 3 5 种不同算法的离群点误差率计算结果

算法	迭代次数 (次)				
	100	200	300	400	500
算法 1	0.11	0.14	0.13	0.10	0.12
算法 2	0.31	0.25	0.27	0.33	0.28
算法 5	0.30	0.32	0.27	0.28	0.29
算法 6	0.36	0.33	0.37	0.39	0.30
算法 7	0.29	0.44	0.39	0.27	0.19

为了更直观地对比不同算法的离群点误差率, 将表 3 中的数据用折线图的形式表现, 如图 5 所示.

分析表 3 和图 5 可知, 在五次不同迭代中, 本文所提的基于层次化深度学习的医疗数据库离群数据检测算法的平均离群点误差率为 0.12%; 算法 2 的平均离群点误差率为 0.288%; 算法 5 的平均离群点误差率为 0.292%; 算法 6 的平均离群点误差率为 0.35%, 算法 7 平均离群点误差率为 0.316%. 对比 5 种不同算法的平均离群点误差率可知, 基于层次化深度学习的医疗数据库离群数据检测算法的离群点误差率始终低于另外 4 种算法, 证明了本文所提算法的有效性. 本文算法融合专家知识和数据的属性取值分布信息, 从多个层次感知离群数据信息, 从而降低了离群数据检测误差.

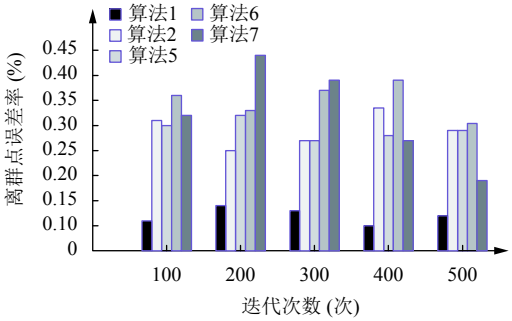

图 5 5 种不同算法的离群点误差率对比

综上所述, 本文所提的基于层次化深度学习的医疗数据库离散数据检测算法的离群点检测率较高、离群点误差率较低. 这主要是因为基于层次化深度学习的医疗数据库离群数据检测算法在过滤离群数据时, 采用动态网格划分法降低数据检测的计算代价, 缩短了检测执行时间, 而在计算邻域半径时, 融合专家知识和数据的属性取值分布信息, 降低了检测误差, 大大提高了基于层次化深度学习的医疗数据库离群数据检测算法的有效性.

5 结语

医疗信息量的不断增长以及信息技术的飞速进步, 使医疗数据库中积累了大量数据. 如何在医疗数据库中及时、高效、准确的获取信息, 是目前亟需解决的问题之一. 针对当前医疗数据库离群数据检测算法存在检测效率低、离群点检测率低和离群点误差率高的问题, 本文提出基于层次化深度学习的医疗数据库离群数据检测算法, 可以精准的在短时间内完成医疗数据库中离群数据的检测, 解决了当前医疗数据库离群数据检测算法中存在的问题, 具有计算代价小、检测耗时短、离群点检测率高、离群点误差率低的优点, 为数据检测、挖掘技术的发展奠定了基础. 在未来的研究阶段, 将深入对不同属性的离群数据进行精细检测, 进一步提高检测效果.

参考文献

1 Hauskrecht M, Batal I, Hong C, *et al*. Outlier-based detection of unusual patient-management actions: An ICU study. Journal of Biomedical Informatics, 2016, 64: 211–221. [doi: 10.1016/j.jbi.2016.10.002]

2 Yu YW, Cao L, Rundensteiner EA, *et al*. Outlier detection over massive-scale trajectory streams. ACM Transactions on Database Systems, 2017, 42(2): 10.

3 Jobe JM, Pokojovy M. A cluster-based outlier detection scheme for multivariate data. Journal of the American Statistical Association, 2015, 110(512): 1543–1551. [doi: 10.1080/01621459.2014.983231]

4 邹云峰, 张昕, 宋世渊, 等. 基于局部密度的快速离群点检测算法. 计算机应用, 2017, 37(10): 2932–2937. [doi: 10.11772/j.issn.1001-9081.2017.10.2932]

5 李少波, 孟伟, 璩晶磊. 基于密度的异常数据检测算法 GSWCLOF. 计算机工程与应用, 2016, 52(19): 7–11. [doi: 10.3778/j.issn.1002-8331.1603-0323]

6 魏畅, 李光辉. 基于约简策略与自适应 SVDD 的无线传感网络离群检测方法. 传感技术学报, 2017, 30(9): 1388–1395. [doi: 10.3969/j.issn.1004-1699.2017.09.015]

7 尹娜, 张琳. 基于混合式聚类算法的离群点挖掘在异常检测中的应用研究. 计算机科学, 2017, 44(5): 116–119, 140. [doi: 10.11896/j.issn.1002-137X.2017.05.021]

8 Rahmani M, Atia GK. Randomized robust subspace recovery and outlier detection for high dimensional data matrices. IEEE Transactions on Signal Processing, 2017, 65(6): 1580–1594. [doi: 10.1109/TSP.2016.2645515]

9 段培永, 崔冲, 张洁珏. 一种改进的局部离群数据检测算法. 黑龙江大学自然科学学报, 2017, 34(4): 474–480.

10 Jiang Z, Shen XH, Ge Y, *et al*. Approximate Gibbs algorithm for blind data detection in two-way relay networks. IET Communications, 2017, 11(8): 1230–1240. [doi: 10.1049/iet-com.2016.0597]

11 丁天一, 张旻, 方胜良. 一种相似度剪枝的离群点检测算法. 小型微型计算机系统, 2018, 39(8): 1680–1684. [doi: 10.3969/j.issn.1000-1220.2018.08.009]

12 韩崇, 袁颖珊, 梅焘, 等. 基于 K-means 的数据流离群点检测算法. 计算机工程与应用, 2017, 53(3): 58–63. [doi: 10.3778/j.issn.1002-8331.1607-0236]

13 韩东明, 郭方舟, 潘嘉铖, 等. 面向时序数据异常检测的可视分析综述. 计算机研究与发展, 2018, 55(9): 1843–1852. [doi: 10.7544/issn1000-1239.2018.20180126]

14 袁钟, 冯山. 基于邻域值差异度量的离群点检测算法. 计算机应用, 2018, 38(7): 1905–1909.

15 Rizk H, Elgokhy S, Sarhan A. A hybrid outlier detection algorithm based on partitioning clustering and density measures. Proceedings of 2015 Tenth International Conference on Computer Engineering & Systems. Cairo, Egypt. 2015. 175–181.

16 迟荣华, 黄少滨, 吕天阳. 基于频繁密度分布模式的不确定数据流查询方法. 哈尔滨工程大学学报, 2018, 39(6): 1052–1058.

17 Huang JL, Zhu QS, Yang LJ, *et al*. A novel outlier cluster detection algorithm without top-n parameter. Knowledge-Based Systems, 2017, 121: 32–40. [doi: 10.1016/j.knosys.2017.01.013]

18 钱晓军, 范冬萍, 吉根林. 物联网差异数据库中的故障数据快速挖掘仿真. 计算机仿真, 2016, 33(1): 301–304. [doi: 10.3969/j.issn.1006-9348.2016.01.064]

19 刘莘, 张绍良, 王飞, 等. 基于地统计学的空间离群点检测算法的研究. 计算机应用研究, 2016, 33(12): 3700–3704. [doi: 10.3969/j.issn.1001-3695.2016.12.040]

20 王习特, 申德荣, 白梅, 等. BOD: 一种高效的分布式离群点检测算法. 计算机学报, 2016, 39(1): 36–51. [doi: 10.11897/SP.J.1016.2016.00036]