

高维空间中离群点检测算法研究

刘培奇, 孙 靖, 段中兴

(西安建筑科技大学 信息与控制工程学院, 陕西 西安 710055)

摘 要: 提出一种基于改进粒子群优化算法的离群点检测算法, 解决高维环境下离群点挖掘效率偏低的问题. 新算法能够充分发挥粒子群优化算法全局搜索的优势, 并具有 k 均值算法快速收敛的特点, 可避免粒子群优化算法的早熟, 减小确定 k 均值算法聚类中心的计算量等问题. 实验表明, 该算法在高维环境下可快速有效的挖掘出离群数据的离群支持度, 有较好的挖掘效率、准确率和实用性.

关键词: 数据挖掘; 离群点检测; 高维数据; PSO 算法; k 均值算法

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2013)07-0068-04

Research on Outliers Detection Algorithms in High-Dimensional Space

LIU Pei-qi, SUN Jing, DUAN Zhong-xing

(School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China)

Abstract: A kind of outlier detection algorithm based on the improved particle swarm optimization algorithm is proposed in the paper here. And the paper focuses on solving the problem that mining algorithms had low efficiency in high dimensional environment. New algorithm using the weight of evolutionary algorithm and moving step function improve particle swarm optimization algorithm, and new algorithm unite k-means algorithm in it. At the same time, the algorithm uses the global search of particle swarm optimization algorithm and the rapid convergence of k-means algorithm advantages. So the algorithms can avoid the premature convergence of the particle swarm optimization algorithm and reduce the amount of calculation clustering center for the k-means algorithm. Experiments show that this algorithm, in high dimensional environment, can mining outliers the support degree of outlier data, which making outlier detection efficiently, accurately and practicably.

Key words: data mining; outlier detection; high dimensional data; particle swarm optimization; k-mean

1 引言

随着数据挖掘技术的不断改进, 对数据库中离群点的挖掘具有重要意义. 离群点挖掘包括离群点检测和离群点分析两部分. 本文主要研究离群点检测问题. 离群点检测主要分为三大类^[1-2], 其中基于统计学的鲁棒回归法、状态估计法, 主要使用概率模型完成检测, 对于已知概率模型的系统检测高效效果好, 但对不确定概率模型的系统 and 异常数据的处理效果不好; 基于偏差的方法是通过数据的特性

检验发现离群数据, 此方法可对多种数据检测但必须预先确定相异度函数, 所以相异度函数的选取直接影响挖掘结果, 在实际中使用困难; 基于距离的方法主要通过计算数据点和聚类中心的距离寻找远离类的的数据点, 避免了统计学中的大量概率计算和偏差检验中的相异度函数的确定. 但是此方法的缺点在于需用户确定初始的中心 C 和离群距离 D. 这两个参数的确定需要多次尝试计算工作量较大, 且单一度量距离很难准确说明数据点离群的原因造成离群分析的困难.

收稿日期: 2012-10-20; **修回日期:** 2012-11-22

基金项目: 国家自然科学基金项目(51178373); 陕西省教育厅自然科学计划项目(12JK0743)

经过大量实验分析,在高维离群点检测中应该选用一种高效、实用、对异常数据不敏感的离群点算法^[3].根据这样的思想,本文提出一种改进粒子群 k 均值离群点检测算法(Particle swarm optimizer k-mean outlier detection, PSOKOD)算法确定离群点的适应度和验证离群点的离群程度.并用仿真数据和私有种子(Private tracker, PT)共享网站的种子数据库对算法进行试验测试.最后,对离群点挖掘结果以及离群原因进行了分析,验证了方法的有效性.

2 PSOKOD 算法的设计与分析

PSOKOD 算法是标准 PSO 算法与 k 均值算法的有效结合.为了进一步提高标准 PSO 算法的收敛速度,通过在惯性权重中引入最大未改变代数和未改变代数改进惯性权重.因此 PSOKOD 算法包括对标准 PSO 算法的改进和改进后的 PSO 算法与 k 均值算法的结合两部分.

2.1 PSO 算法改进思想

PSO (Particle Swarm Optimizer, PSO) 算法是 Eberhart 博士和 Kennedy 博士首先提出的一种进化计算技术^[4].为了提高 PSO 算法的收敛性能,Shi Y 和 Eberhart R C 于 1998 年提出标准 PSO 算法^[5-6].

在标准 PSO 中每个优化问题的解称为一个“粒子”,所有粒子都有一个由优化函数决定的适应值(fitness value).粒子的速度、位置和惯性权重的更新方法见公式(1)到(3).

$$v_{id}^{k+1} = wv_{id}^k + c_1 rand_1^k (pbest_{id}^k - x_{id}^k) + c_2 rand_2^k (gbest_{id}^k - x_{id}^k) \quad (1)$$

$$x_{id}^{k+1} = x_{id}^k + v_{id}^{k+1} \quad (2)$$

$$w = w_{\max} - \frac{w_{\max} - w_{\min}}{k_{\max}} \times k \quad (3)$$

公式中, V_{id} 是粒子的速度, X_{id} 是当前粒子的位置. $pbest_{id}$ 是个体极值; $gbest_{id}$ 是全局极值; $rand_1$ 和 $rand_2$ 是介于(0,1)之间的随机数. c_1, c_2 是学习因子. $w_{\max}$ 为初始权重, $w_{\min}$ 最终权重, $k_{\max}$ 为最大迭代次数, k 为当前迭代次数.改进后的粒子群优化算法,初期能较快的接近全局最优点区域,而在后期具有良好的局部搜索性能.

针对标准 PSO 算法中原惯性权重适用于固定进化代数为结束条件的 PSO 算法,但对于要求精度要求较高的 PSO 算法的问题,本文对惯性权重进行如下改进,改进后的惯性权重计算公式见公式(4):

$$w = w_{\max} - \frac{w_{\max} - w_{\min}}{T_{\max}} \times T_{\text{best}} \quad (4)$$

式中, $T_{\max}$ 为 gbest 被规定的最大未改变代数, T_{best} 为当前 gbest 未改变代数.

借鉴在高速运动中的粒子会出现迁跃现象,将此现象引入算法中增强算法的全局搜索能力.改进后的公式见(5).

$$x_{id}^{k+1} = x_{id}^k + v_{id}^{k+1} + T(\delta) \quad (5)$$

其中, $T(\cdot)$ 是迁跃函数.对于随机数 $R \in [0, 1]$, 当 $R < \delta$ 时, $T(\delta) = (1-R)V_{\max}$; 当 $R \geq \delta$ 时, $T(\delta) = 0$. 式中的迁跃因子 $\delta \in [0, 1]$ 是用户根据实际需要而设定的,一般取值的范围在 $\delta < 0.001$.

2.2 PSOKOD 算法设计

在数据检测前,应对数据进行筛选和规约.对于非数值格式的属性,如果是类别属性将转化为数值编码,将不能运算的属性剔除,然后,根据数据属性值的上下界确定 PSO 算法边界.

融合 k 均值算法后的 PSOKOD 检测算法的主要思想:

S1: 根据数据库中的属性值定义空间边界.

S2: 在解空间 S 中产生 n 个随机粒子 $x_1, x_2, \dots, x_n$ 形成初始粒子群 $X(n)$.

S3: 随机生成 $v_1, v_2, \dots, v_n$ 初始化速度变化矩阵 $V(n)$.

S4: 计算适应值 $F(n) = 1/(D+1)$ 对每个离子评价.计算适应值采用基于欧氏距离的计算公式(6),其中 $dis(p_i, f_i)$ 是点 p_i 在其领域 f_i 中欧氏距离之和.

$$D = \sum_{i=1}^n dis(p_i, f_i) \quad (6)$$

S5: 每个离子用当前适应值和自身历史最好 pbest 比较.

S6: 如果离子优于历史最好值则记录并更新 pbest.

S7: 每个离子用当前适应值和全体离子的历史最好值 gbest 比较.

S8: 如果离子优于历史最好值则记录并更新 gbest.

S9: 使用速度公式(1)更新离子的速度 V_{i+1}

S10: 使用位置公式(5)位置 X_{i+1} 产生新粒子群 $X(n+1)$.

S11: 用 PSO 算法中得出的解集更新离群点集合并指导 k 均值聚类的中心,

S12: 达到设定条件执行一次 k 均值聚类算法.

S13: 检查算法终止条件是否达成,若达成终止条件记录离群点,否则跳转到 S4 步骤.

S14:算法结束

PSOKOD 算法用类 C 代码描述为:

算法输入: DS, K, N, δ // DS 为数据集, K 为初始聚类个数, N 为粒子群规模, δ 为迁跃概率

算法输出: OL // OL 为离群点集合

Begin

Initialization $X(n)$;

Initialization $V(n)$;

For ($t \geq T_{\max}$ || $e \leq E_{\min}$; $i++$) {

For ($i \leq n$; $i++$) {

fitness value = $F(n)$; // 计算适应值

If ($F(n) > pbest(X_i)$) {

$pbest(X_i) = F(n)$;

sort ($gbest[]$);

For ($i \leq j$; $i++$) {

If ($F(n) > gbest[j]$) {

$gbest[j] = F(n)$;

break; } }

For ($i \leq n$; $i++$) {

Using the equation (1) and equation (4)

updates the v_{id} ;

If ($V > V_{\max}$)

{ $V = V_{\max}$; }

Using the equation (5) updates the x_{id} ; }

Updated outliers point set with $gbest$;

K-mean algorithm; }

END

3 PSOKOD 算法分析

PSOKOD 算法和 PSO 算法相比具有很强的优势. 下面从 4 个方面对新算法进行分析.

3.1 PSOKOD 算法的可操作性

从和其他离群点检测算法相比本算法的优势在于初始设定比较简单, 其主要参数设置有初始聚类个数、粒子群规模和迁跃概率. 本文根据问题规模采用不同粒子群大小, 用属性个数决定编码长度, 并采用实数编码形式初始设定简单. 在算法中, 第 S1 步骤使用各维度的边界值定义多维空间的边界; S2 步骤将属性转化为整形编码格式如: $A_1 A_2 A_3 \dots A_n$ 的离子编码格式其中 A_i 是随机选中一条数据的属性转化后直接将编码拼接的实数格式. 最后根据 S1 步的边界随机生成个离子的初始速度 $v_1, v_2, \dots, v_n$, 并由其初始化速度变化矩阵 $V(n)$ 完成算法的初始化操作.

3.2 引入迁跃函数

改进后的算法引入迁跃函数进一步提高了算法的全局搜索能力.

迁跃函数是根据物理实验中离子会在空间中消失又从另一个地方出现的物理现象启发得到的. 迁跃函数的作用是在算法收敛的后期保证算法有一定的跳出局部最优解的能力. 迁跃函数在计算离子位置的时候以 δ 概率出现, 使离子突然发生较大位移, 是一个随机位置移动, 保证算法不会早熟. 在改进算法中通过在 x_{id} 中引进 $T(\delta)$ 函数, 改进算法性能.

算法中的 x_{id} 反映算法的全局搜索区域. x_{id} 越大, 说明算法的全局搜索能力越强. 反之, 其全局搜索能力就弱. 在 PSOKOD 算法中, 给迭代公式中增加迁跃函数 $T(\delta)$. 一般情况下, 因 δ 值很小, 因此, $T(\delta) = 0$, 对 x_{id} 没有影响, 当随机数大于 δ 时, PSOKOD 算法的 x_{id} 必然大于 PSO 算法的 x_{id} , 离子将有较大机会跳出局部最小解, 从而改善了算法的全局搜索能力.

3.3 优化惯性权重计算公式

对惯性权重计算公式的优化, 惯性权重的作用主要是让离子有速度的记忆性, W 越大离子对速度的继承越多, 相反 W 越小离子对速度的继承越小. 在 PSOKOD 算法中主要将以固定的速率减小的惯性参数修改为自适应的不同速率的减小过程. 原有的惯性公式在确定 $K_{\max}$ 后, w 与 $K_{\max}$ 是一个线性关系, 其最大值为 $W_{\max}$, 最小值为 $W_{\min}$, 减小速率是由 $K_{\max}$ 决定的. 线性减小的过程使算法的适应性弱, 制约了算法的收敛速度. 在 PSOKOD 算法中, 通过引入参数 K_{best} , 将惯性参数公式变为自适应的变化速率减小过程, 使惯性权重可以在初始时加快寻找最优结果的区域, 在后期惯性参数会渐渐变小, 使算法能更好地搜索全局最优解. 图 1、图 2 为改进算法和原算法求解 100 次 Six-Hump Camel-Back Function 平均结果的对比.

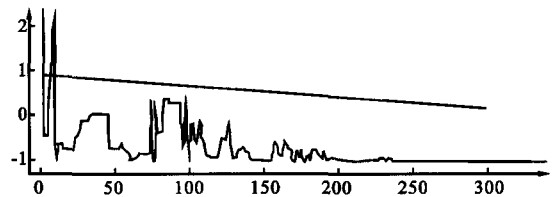


图 1 改进前的算法收敛情况和惯性权重减小的趋势

图中横轴为算法迭代次数, 纵轴为函数最优解. 在图 1 中, 直线为从 0.9 减小到 0.2 的惯性权重变化过程, 曲线为函数求解的进化过程, 其进化的速度直接受直线的影响. 图 2 中, 反映了使用改进公式

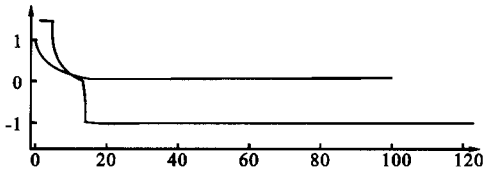


图2 改进后的算法收敛情况和惯性权重减小的趋势

(5)后的函数最优解变化过程,其中从0.9减小到0.2的惯性权重变化过程改进为多次变化率的曲线,改进了函数求解过程,从图形对比中能轻松发现在改进前算法在前期最优解有较大的震荡后期算法收敛的过程也很缓慢;改进后的算法虽然在初期有短暂的停顿但是收敛过程相对迅速收敛过程没有反复震荡的过程。

综上所述,改进后的惯性权重算法的收敛速度得到明显改善。

3.4 融合K均值聚类算法

PSOKOD算法使PSO算法和K均值聚类算法相互促进,使离群数据的检测更高效快速。主要解决确定初始中心C和离群距离D两个参数的时间复杂度。

PSOKOD算法使改进PSO算法和K均值聚类算法相融合,采用将K均值聚类算法植入PSO算法中的方式结合两种算法。K均值聚类算法的缺点在于需要用户确定两个重要参数类中心个数C和离群距离D,而这两个参数的设定影响算法的效果和算法的性能。PSOKOD算法将解集以自然数编码的方式注入K均值聚类算法中,即K均值聚类算法不在依赖初始设定值。算法会经过公式(6)计算适应值F来确定K均值聚类算法需要的类中心C和离群距离D。

3.5 算法在高维空间中的优势

在高维空间中K均值聚类使用的时间会随着位数增长迅速增长,本文算法在对两算法融合过程中加入限制条件。在算法S12步骤时最优解未改变代数达到设定的收敛条件的十分之一时加入K均值聚类算法进行详细的局部搜索。不盲目的增加算法的时间,也减小了算法对陷入局部最优解的可能。优化了K均值聚类算法在高维数据下的初始参数,从而达到提高算法在高维空间中快速检测离群点的目的。

通过以上分析可以看出,改进后的PSOKOD算法在算法的收敛速度和算法的全局搜索能力上都有改进,改进后的算法在高维离群点的检测上表现出优异的性能。

4 实验设计及分析

本文实验工作是在PC机上实现的。硬件环境是I5-2450 2.5GHz的CPU,DDR3 4GB的内存,Window 7的操作系统。实验使用仿真数据生成的200个二维数据,可以看出,离群支持度对离群点的描述准确,PSOKOD算法能在挖掘出离群点同时提供准确的离群支持度,说明算法有较好的实用性。

另外,本文在研究过程中,采用PSOKOD算法、GAK(Genetic Algorithm K-mean)算法和改进前的PSOKOD算法对教育网西安建筑科技大学的私用共享网站种子数据库中2011年的数据进行了离群点检测,筛查网站中的作弊用户,网站主要为教育网的接入用户提供资源共享的一个平台。实验通过将种子表(tracker)和用户表(user)关联后的数据表进行离群点检测,数据库维度为46。其中算法参数设置为初始聚类个数10,粒子群规模为45个离子,迁跃概率为0.001固有的学习因子 C_1, C_2 设为1,由惯性权重计算公式(4),从 $W_{\max}=0.9$ 递减到 $W_{\min}=0.2$,算法停止条件为最优解停止更新2000代。实验的挖掘对比结果可以看出,PSOKOD算法在改进后,算法所用时间明显降低。改进后的算法比GAK算法所用时间平均降低18%,且发现45条可疑数据,经过人工确认其中43个有不同程度的作弊行为,相符情况达到95.5%,证明本文算法在高维数据库情况下表现出了相对较好的挖掘效率和实用性。

5 结束语

本文针对数据挖掘中的高维离群点检测效率较低的问题,对离群点检测算法进行了改进。在改进中重点发挥PSO智能检测方法的寻优能力和算法融合优势。根据改进思想,本文利用惯性参数和迁跃函数改进PSO算法,并应用于离群点检测算法中。实验结果证明PSOKOD算法在高维环境下能高效的检测离群点。改进算法可应用于数据库中特殊数据的检测问题,并在教育网中的润西草堂的数据库挖掘中进行作弊数据检测中帮助检测出了违规账号,取得良好效果。

参考文献:

- [1] Kriegel HP, Schubert M, Zimek A. Angle-based outlier detection in high-dimensional data[C]//Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. China: Beijing, 2008, 444-452.

(下转第77页)

- 5, 1969, 4(7):627—631.
- [3] Mengaliu, D'Andrea A N. Synchronization techniques for digital receivers[M]. New York: Plenum, 1997: 205—228.
- [4] 曾凤池. 数字化考斯托斯锁相环的分析与研制[J]. 自动化学报, 1981, 7(4):312—318.
- [5] 张安安, 杜勇, 韩方景. 全数字 Costas 环在 FPGA 上的设计与实现[J]. 电子工程师, 2006, 32(1):18—20.
- [6] 樊昌信, 曹丽娜. 通信原理[M]. 2 版. 北京:国防工业出版社, 2007.
- [7] Lindsey W. Phase noise effects on space shuttle communications link performance[J]. IEEE Trans Commun, 1978, 26(11):1117—1124.
- [8] 张才霞, 王少云, 刘海颖. 一种基于 Costas 环的 BPSK 解调设计[J]. 信息化研究, 2010, 6(11):34—40.
- [9] Chung B Y. Performance analysis of an all—digital BPSK direct—sequence spread—spectrum IF receiver architecture[J]. IEEE Journal on Selected Areas in Communications, 1993, 11(7):1096—1107.
- [10] 张欣. 扩频通信数字基带信号处理算法及其 VLSI 实现[M]. 北京:科学出版社, 2004.
- [11] 张厥盛, 郑继禹, 万心平. 锁相技术[M]. 西安:西安电子科技大学出版社, 1998.
- [12] Nguyen T M. The behavior of costas loop in the presence of space telemetry signals[J]. IEEE Tran Communi, 1992, 40(1):190—198.

作者简介:

苏 洲 男,(1987—),硕士研究生. 研究方向为基于 FPGA 的数字通信和信号处理.

(上接第 71 页)

- [2] Maimon O, Rokach L. Data mining and knowledge discovery handbook[M]. New York:Kluwer Academic Publishers, 2005: 374—380.
- [3] Jewie Han, Michelin Kamber. Data mining: concepts and techniques[M]. 2nd ed. 北京:机械工业出版社, 2008:237—245.
- [4] Kennedy J, Eberhart R.. Particle swarm optimization [C]//Proc. IEEE International Conference on Neural Networks, IV. Piscataway, NJ:IEEE Service Center, 1995:1942—1948.
- [5] Shi Y, Eberhart R C. Empirical studying on particle swarm optimization[C]//Proc. on evolutionary computation. Piscataway, NJ:IEEE Service Center, 1999: 1945—1950.
- [6] Shi Y, Eberhart R C. Fuzzy adaptive epartiele swarm optimization[C]//Proe. on evolutionary computation. Seoul, Korea, IEEE Service Center, 2001. 741—750.

作者简介:

刘培奇 男,(1959—),博士,副教授. 研究方向为机器学习、数据挖掘、自然语言处理.

孙 靖 男,(1986—),硕士研究生. 研究方向为自然语言理解、数据挖掘.