文章编号: 1000-5862(2014)05-0454-05

离群点检测方法及其在大数据时代下的改进方法研究

苗永春,程　艳*
(江西师范大学计算机信息工程学院,江西 南昌　330022)

摘要: 通过对当前有代表性的离群数据检测方法的分析和比较,总结了各方法的特性及优缺点. 针对大数据的数据量大、维数高的特性,分析了离群点检测方法的改进策略,并以 T-ODCD 算法和 AROD 算法为例,进一步说明离群点检测改进策略.

关键词: 大数据; 离群点检测方法; 改进策略
中图分类号: TP 391; TP 311　　**文献标志码**: A

0　引言

随着云计算、物联网及社交网络等技术的兴起,数据的种类和规模正在不断增长和积累,大数据时代已到来. 大数据呈现出 4 种特性[1]: 规模性(volume)、多样性(variety)、高速性(velocity)和价值性(value),数据像从"池塘"变成"海洋",不仅数据的量大,数据的维数也剧增. 对合并后的小型数据集合进行离群点挖掘,可以获得许多额外的信息,可用来避免疾病扩散、网络入侵检测、信用卡恶意透支、贷款证明的审核等,这些用途正是大数据时代下离群点挖掘盛行的原因.

离群点检测是数据挖掘技术中一个重要的研究领域,也被称为离群点挖掘,其目的是试图捕获那些显著偏离多数模式的异常情况. 离群点检测在许多应用中都是重要的,如医疗处理、公共安全、工业损坏检测、图像处理、传感器/视频网络监视和入侵检测等. 早期的离群点检测算法是针对整个数据集,检测的是全局离群点[2-4]. 后来,研究发现: 在现实世界中,数据集本身具有复杂性、多变性及不完整性,而且在较多场合,更多考虑领域的局部情况,为此,提出局部离群点检测算法[4]. 随着大数据时代的到来,数据的来源、数据量及维数急剧增加,离群点检测面临着一系列挑战.

本文将离群点检测方法[5-7]分为基于统计的、基于距离的、基于密度的、基于聚类的和基于分类的离群点检测方法,并分析了这些方法各自的优缺点. 针对大数据,为改进的离群点检测方法,当今研究者多把研究焦点聚集到采样点的预处理上,笔者根据近几年相关研究总结得出改进策略: 数据集的预处理分为剪枝和属性约简,把对复杂的高维、大数据量的离群点检测问题转化为传统的离群点检测问题,将复杂问题简单化,并以 T-ODCD 算法和 AROD 算法为例说明对应的改进策略.

1　离群点检测方法

1.1　基于统计的离群点检测方法

离群点检测的研究最早始于统计领域. 基于统计的方法[8]的主要思想为对于数据的正常性做出假设. 假定数据集中的正常对象服从某种分布或概率模型,通过不一致检验把那些严重偏离分布曲线的对象视为离群点,或低概率区域中的对象是离群点.

针对给定的数据集,该方法需要学习一个拟合的生成模型. 根据如何学习生成模型,该类方法又进一步划分成 2 个主要类型: 参数方法和非参数方法.

(i) 参数方法[9]: 假定正常的数据对象服从一个以 o 为参数的参数分布. 该参数分布的概率密度函数 $f(x,o)$ 给出对象 x 被该分布产生的概率. 该值越小,x 越可能是离群点. 该方法主要包括基于高斯模型的和基于回归模型的检测方法[11].

(ii) 非参数方法[10]: 并不假定先验统计模型,

收稿日期: 2014-05-17
基金项目: 国家社科基金教育学青年课题"教育虚拟社区的群集智能化构建方法研究"(CCA110109)资助项目.
通信作者: 程　艳(1976-),女,江西婺源人,教授,博士,主要从事智能计算机辅助教育、教育数据挖掘和虚拟学习社区研究.

而是试图从输入数据中学习“正常数据”的模型. 该方法主要包括基于直方图的和基于核函数的检测方法[11].

基于统计的离群点检测方法适用于单变量的服从特定概率模型的数据集. 其优点为该方法建立在标准的统计学技术之上,具有稳定的基础; 对于单个属性的离群点检测,当具有充分的数据和所需的先验知识时,该方法检测效果较好. 其缺点是对于多维数据集,该方法检测效果会变差,对于很难估计真实的分布的高维数据,该方法不适用[12]; 在许多情况下,数据集服从的分布或概率模型是未知的,用不同的模型检测出来的离群点可能不一致; 基于统计的方法的有效性较大程度上依赖于对待挖掘的数据集所做的统计模型假定是否成立[13]. 为了改进这些不足之处,发展出了基于计算统计学的方法,被称为基于深度的方法.

1.2　基于距离的离群点检测方法

基于距离的方法最早由 E. M. Knorr 等[14-15]提出,其主要思想为对于待要分析的数据集 $DB(pct, dis_{\min})$,用户可以指定一个距离阈值 $dis_{\min}$ 来定义对象的合理邻域,对于每个对象 O,可以考察 O 的 $dis_{\min}$-邻域中的其他数据对象. 如果数据集 DB 中大多数对象都远离 O,即至少有 pct 部分的数据对象与 O 的距离大于 $dis_{\min}$,则该对象 O 被视为离群点.

Rastogi & Ramaswamy[16] 在基于上面对距离的离群点定义的基础上,提出基于距离的 k-最近邻(k-NN) 离群检测算法. 该算法的一个主要缺陷是每计算对象 O 的第 k 个最近邻点的距离值,就要扫描一次数据集,计算效率低. 针对该缺陷,提出基于索引的(index-based) 算法引进索引的思想来提高算法的效率[17-18]、嵌套循环(Nested Loop,简称 NL) 算法主要从减少操作的 I/O 次数方面来改善算法的效率[14,17] 和基于网格(cell-based) 的算法通过结合点的局部密度方法来提高离群检测的效率[16]. 该方法比较适用于数据对象的属性维数比较少且参数 pct 和 $dis_{\min}$ 的值比较容易确定的数据集.

其优点是该方法比较简单,只要能定义反应数据之间彼此差别的距离函数,就可以采用该方法. 其缺点是该方法中指定的距离阈值是全局阈值,对于不同密度的数据集,它检测出离群点的准确度低[19]; 如果需要确定的距离阈值 $dis_{\min}$ 和参数 pct 的先验知识不足,则对其运用造成一定的困难,尤其对不同聚类密度数据集而言,距离阈值 $dis_{\min}$ 差别会较大,指定不同的距离阈值 $dis_{\min}$,离群点检测结果也常常会出现不一致的现象[20]; 由于遍历邻域内的数据对象需要一定的时间复杂度,因此,难以用于大规模数据集.

1.3　基于密度的离群点检测方法

基于密度的方法[21] 主要思想为假定正常数据对象周围的密度与其邻域周围的密度类似,而离群点对象周围的密度显著不同于其邻域周围的密度. 需要把对象周围的密度与对象邻域周围的密度进行比较,把低密度的对象视为离群点. 一般使用每个对象到第 k 个最近邻的距离大小来度量密度,定义密度为到 k 个最近邻的平均距离的倒数. 如果数据对象的该值大,则密度得分就高,离群程度较大.

该检测方法的一个典型的例子是 M. M. Breuning[22] 等提出基于局部离群因子的离群点检测算法,除此之外,还有基于平均密度的离群点检测方法[23] 和 C. C. Aggarwal[24] 提出的一个结合子空间投影变换的基于密度的高维离群检测算法. 根据算法特性,它更适用于聚类特性比较明显,求局部密度时的 I/O 代价比较低的数据集.

其优点是对于密度分布不均匀的数据集,能够更好地检测出那些位于稠密簇周边的离群点(局部离群点); 不需要知道数据集的先验知识,并且可以同时检测出全局离群点和局部离群点[25]. 其缺点是由于算法中用到的计算复杂度较大,因此,该检测方法的时间和空间效率不高; 数据的稀疏性和离群意义难以解释,则对参数 k 的选择很困难. 对于规模较大的数据,I/O 的也较高.

1.4　基于聚类的离群点检测方法

基于聚类的方法[26] 主要思想为如果对象不属于任何簇或与最近簇之间的距离都很远,则视该对象为离群点; 如果某簇包含的数据对象较小且又稀疏,则该簇中的所有数据对象均为离群点.

由定义可知,该算法既可以发现簇,也可以发现离群点,但是其主要的目标是发现簇,而离群点就是没有被包含在簇内的对象. 该方法一个显著的特点是首先采用特定的聚类算法处理所有输入的数据对象得到聚类,然后在聚类的基础上来评估各对象属于簇的程度,从而检测出离群点. 依据其特点,该方法比较适用于聚类特性明显,容易用聚类算法发现簇的数据集. 其主要的代表方法[27] 有基于对象离群因子的方法和基于簇的离群因子的方法.

其优点为该方法对许多类型的数据均有效,并且是以无监督方法检测离群点; 由于与整个数据集

包含的对象总数相比,簇中包含的对象数目小了很多,因此,在离群点检测阶段,比较对象与簇之间的关系,可以更快地确定该对象是否是离群点. 其缺点是它的有效性高度依赖于聚类算法,且所使用的聚类算法产生的簇的质量对检测出离群点的质量影响很大[27];离群点也非常依赖于所用的簇的个数和数据中离群点的存在性;有些聚类方法强制规定每个数据点都依附某个簇,当离群点恰好依附于一个稠密的簇时,容易漏检;大多聚类算法需要的时间复杂度为$O(dN^2)$[28],对于大型数据集,该方法开销较大,有可能成为制约算法应用的瓶颈.

1.5 基于分类的离群点检测方法

针对分类标签已知的数据集,其包含一些标记为"正常",而其他标记为"离群点"的样本. 基于分类的方法[29]主要思想为对分类标示已知的数据集,经过训练和学习,找出区分数据类的模型,即构建一个可以区分正常类和离群点类的分类器. 对于被检测的对象,考察其被分成正常类,还是离群点类.

由于样本数据的不平衡性,即正常样本的数量可能远远高于离群点样本的数量,离群点样本数量的不足,使得很难构造一个准确的分类器. 另外离群点样本的表示不充分,如实际中,新的离群点不时地出现,导致无法枚举所有离群点. 为了解决上述问题,基于分类的方法通常构建一类模型,即构建一个仅描述正常类的分类器,不属于正常类的任何样本都被视为离群点. 根据训练集中正常类标签的多少,该方法可以进一步划分为"多类别离群分类检测法"和"单类别离群分类检测法". 基于分类的离群点检测方法主要包括:基于神经网络的方法[30]、基于贝叶斯网络的方法[31]、基于支持向量机的方法[32]和基于规则的方法[33].

其优点是该方法使用正常类的模型(一类模型)检测离群点,可以检测可能不靠近训练集中的任何离群点的新离群点;该方法一旦构建好分类模型,离群点检测过程就较快. 其缺点是该方法的有效性不仅高度依赖分类算法,还依赖于有代表性的正常类标签的数量;在实际应用中,难以获得高质量的训练数据,这使得此方法在应用中受到制约.

2 离群点检测方法的改进策略

当数据量增长到一定规模以后,可以从小量数据中挖掘出有效信息的算法并一定适用于大数据,针对大数据规模大、维数高的特性,在传统的离群点检测方法的基础上,提出了 2 种改进策略,以便进一步地深入研究奠定了基础.

2.1 剪枝策略

离群点检测方法的时间复杂度和数据集规模有着密切的关系,大数据的数据量越大,计算量越大,算法的时空效率越低[34]. 剪枝策略[4,34-36]是指离群点占整个数据集的小部分,在离群点检测前,剪掉那些不包含离群点的数据对象类,对余下的数据进行离群点检测. 研究发现:这种通过减小数据量,进而降低计算量,对分布密度显著不同的数据集,挖掘的效果佳.

2.2 属性约简策略

由于传统的离群点检测方法仅仅为了寻找到离群点,不会关注离群点里面包含的内在信息,并且高维数据空间对象间的距离往往并不明确[37]. 因此,针对大数据的高维特性,传统的离群点检测方法的准确性、有效性及适用性均很低.

解决该问题的关键是对数据空间的维度进行划分和归约来进行优化[38],即把对高维数据的离群点检测转为传统的离群点检测或者对子空间检测离群点,但需要确保经过约简之后的属性集合和全属性集上发现的大部分离群点基本一致[39].

3 改进策略的例证

研究人员一般对大数据剪枝,对高维大数据约维,再扩充传统的离群点检测方法,使其适用到当前的离群点检测应用中. 下面以具体的方法为例,来说明离群点检测方法的改进策略.

3.1 剪枝策略

剪枝策略以基于聚类划分的两阶段离群点检测方法(T-ODCD)为例来说明,T-ODCD 算法[20]对传统的基于距离的离群点检测方法的扩充,采用基于聚类和距离相结合的办法进行局部离群点的检测,聚类阶段是剪枝策略的关键阶段. 笔者总结出 T-ODCD算法的流程图如图 1 所示.

从图 1 可以清晰地观察出 T-ODCD 算法,首先将数据集划分成若干个微聚类. 再利用信息熵去判断得到的微聚类中是否包含离群点,如果不包含离群点,则剔除该微聚类[20,40]. 最后,利用基于距离的方法在剩余微聚类中挖掘离群点. 前 2 步的主要目的是避免从不包含离群点的微聚类中强行挖掘离群点,对整体数据集进行剪枝处理,降低数据量,从而降低了基于距离的离群点检测方法中的计算量.

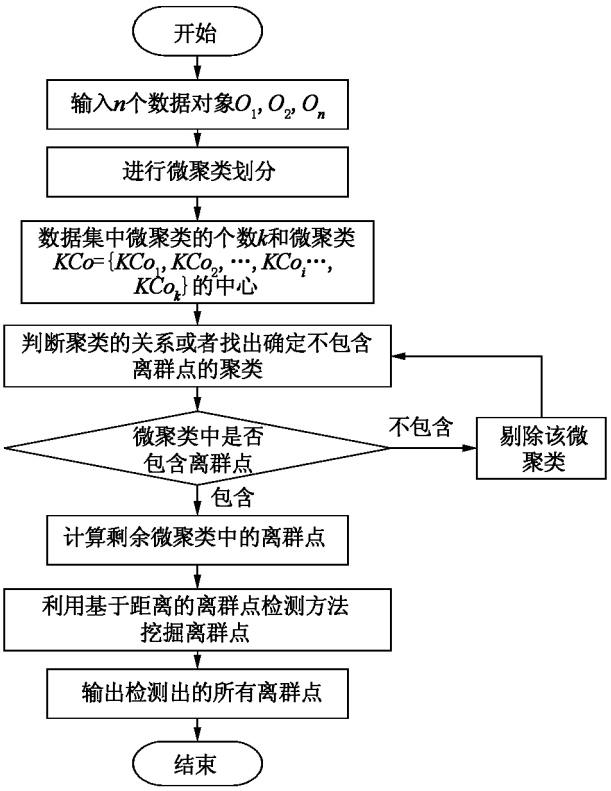

图 1　T-ODCD 算法的流程图

3.2　属性约简策略

基于属性约简的离群点检测方法，引入基于信息熵的属性划分，对非重要属性进行约简. 其基本思想[39,41]：首先计算每个属性信息熵，将其作为加权距离的权值; 其次依据属性划分熵值和数据集的信息熵对属性重要程度进行划分，对非重要属性进行约简; 最后结合数据的离群度计算方法，对离群度进行降序排序，选取前 k 个离群度最高的对象作为离群点. 总结出 AROD 算法的流程图如图 2 所示.

4　结论与展望

本文通过对离群点检测方法的分析可知，传统方法本身存在不足，并且针对大数据的数据量大、维数高的特性，传统方法效率低，准确性低的问题更加突出. 为此总结出当前 2 种离群点改进策略: 剪枝策略和属性约简策略，并通过 T-ODCD 算法和 AROD 算法为例进行分析，以便研究者更进一步深入研究.

大数据时代的到来，数据呈现爆炸式的增长，人们正被数据洪流所包围，从大规模数据集中检测出离群点信息犹如从大海捞针，如何通过剪枝规则来加速大规模数据集中离群点的检测面临很大的挑战. 数据的多样性是大数据时代的显著特征之一，这也就是意味着除了结构化数据，半结构化和非结构化数据也将是大数据时代的重要数据类型组成部分[42]，因此流式数据的离群点检测也是一个热点. 最新文献表明，地学数据的离群检测算法、动态环境下异常的增量式挖掘算法、长时间序列离群检测算法以及基于人工智能的离群检测算法将是未来一段时间内离群数据挖掘领域的一个主要研究方向.

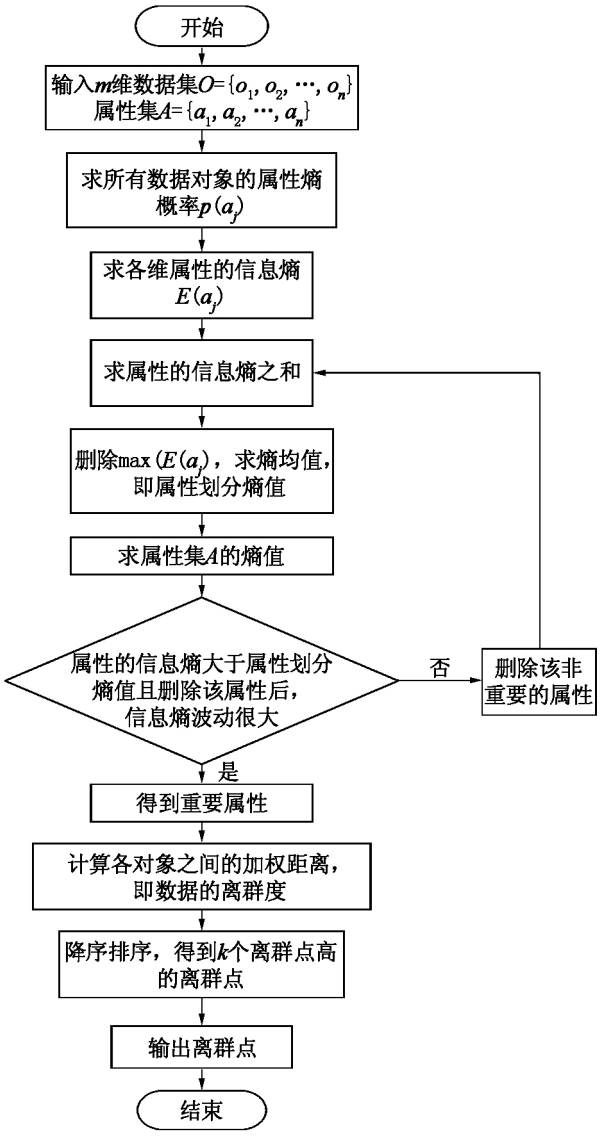

图 2　AROD 算法的流程图

5　参考文献

[1] Barwick H. The “four Vs ” of big data. Implementing information infrastructure symposium [EB/OL]. [2012-10-02]. http://www.computerworld.com.au/article/396198/iiis_four_vs_big_data/.

[2] Han Jiawei, Kamber. Data mining: concepts and techniques [M]. 2ed. San Francisco: Morgan Kaufmann, 2006.

[3] 薛安荣，姚林，鞠时光，等. 离群点挖掘方法综述 [J]. 计算机科学，2008，35(11)：13-27.

[4] 薛安荣,鞠时光,何伟华,等. 局部离群点挖掘算法研究[J]. 计算机学报,2007,30(8):1456-1463.
[5] 黄洪宇,林甲祥,陈崇成,等. 离群数据挖掘综述[J]. 计算机应用研究,2006,8:8-11.
[6] Hawkins D. Identification of outliers [M]. London: Chapman and Hall,1980.
[7] 徐翔,刘建伟,罗雄麟. 离群点挖掘研究[J]. 计算机应用研究,2009,26(1):34-39.
[8] Barnett V,Lewis T. Outliers in statistical data [M]. New York: John Wiley & Sons,1994.
[9] 金义富,邓明. 基于统计的离群数据挖掘与分析[J]. 湛江师范学院学报,2007,28(6):71-73.
[10] 李志云. 数据挖掘中离群点检测的非参数方法研究[J]. 微型电脑应用,2013,29(8):46-47.
[11] Paul S T,Fung K Y. A Generalized extreme studentized residual multiple-outlier-detection procedure in linear regression [J]. Techno-metrics,1991,33:339-348.
[12] 史东辉,张春阳,蔡庆生. 离群数据的挖掘方法研究[J]. 小型微型计算机系统,2001,22(10):234-236.
[13] 杨茂林. 离群检测算法研究[D]. 武汉:华中科技大学,2012.
[14] Knorr E M,Ng R T. Algorithms for mining distance-based outliers in large datasets [C]//New York: Proc of Int Conf Very Large Data-bases(VLDB'98),1998:392-403.
[15] Knorr E,Ng R. Findingintensional knowledge of distance-based outliers [C]//Scotland: Proc of the 25 th VLDB Conference Edin-burgh,1999:211-222.
[16] Angiulli F,Pizzuti C. Fast outlier detection in high dimensional spaces [EB/OL]. [2012-10-16]. http://www.researchgate.net/publication/220699183_Fast_Outlier_Detection_in_High_Dimensional_Spaces.
[17] Bay S D,Schwabacher M. Mining distance-based outliers in near linear time with randomization and a simple pruning rule [C]. Washington,DC: Sigkdd,2003.
[18] An Jiawei,Kamber M. Datamining: concepts and techniques [M]. New York: Academic Press,2001.
[19] 胡彩平,秦小麟. 一种基于密度的局部离群点检测算法 DLOF[J]. 计算机研究与发展,2010,47(12):2110-2116.
[20] 杨福萍,王洪国,等. 基于聚类划分的两阶段离群点检测算法[J]. 计算机应用研究,2013,30(7):1943-1945.
[21] Spiros Papadimitriou,Hiroyuki Kitagawa,et al. LOCI: fast outlier detection using the local correlation integral [EB/OL]. [2013-10-12] 10.1109/ICDE.2003.1260802.
[22] Breuning M M,Kriegel H P,Ng R T,et al. LOF: identifying density-based local outliers [C]. Dallas: ACM Press,2000:93-104.
[23] 施化吉,周书勇,李星毅,等. 基于平均密度的孤立点检测研究[J]. 电子科技大学学报,2007,36(6):1286-1288.
[24] Aggarwal C C,Yu P. Findinggeneralized projected clusters in high dimensional spaces [C]. Dallas: ACM Press,2000:70-81.
[25] 张卫旭,尉宇. 基于密度的局部离群点检测算法[J]. 计算机与数字工程,2010,38(10):11-14.
[26] Ng R,Han J. Efficient and effective clustering methods for spatial data mining [C]. California: Morgan Kaufmann Publishers Inc,1994,144-155.
[27] 蒋盛益,李霞,郑琪. 数据挖掘原理与实践[M]. 北京:电子工业出版社,2011.
[28] Xu R,Wunsch II D. Survey of clustering algorithms [J]. IEEE Transactions on Neural Networks,2005,16(3):645-678.
[29] Das K,Schneider J. Detecting anomalous records in categorical dataset [C]. New York: ACM,2007,220-229.
[30] Markou M,Singh S. Novelty detection: a review-part2: neural network based approaches [J]. Signal Processing,2003,83(12):2499-2521.
[31] Wong W K,Moore A,Cooper G,et al. Bayesian network anomaly pattern detection for disease outbreaks [C]. Washington DC: AAAI Press,2003:808-815.
[32] Ratsch Q Mika S,Scholkopf B. Constructing boosting algorithms from svms: An application to one-class classification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2002,24(9):1184-1199.
[33] Mahoney M V,Chan P K. Learning rules for anomaly detection of hostile network traffic [C]. Washington DC: IEEE,2003:601-604.
[34] 崔贯勋,朱庆生. 一种改进的基于密度的离群数据挖掘算法[J]. 计算机应用,2007,27(3):560-573.
[35] 古平,刘海波,罗志恒. 一种基于多重聚类的离群点检测算法[J]. 计算机应用研究,2013,30(3):751-754.
[36] 赵战营,成长生. 基于聚类分析局部离群点挖掘改进算法的研究与实现[J]. 计算机应用与软件,2010,27(11):255-258.
[37] Agrawal R,Gehrke J,Gunopulos D,et al. Automatic subspace clustering of high dimensional data for data mining applications [EB/OL]. [2013-10-17]. http://wenku.baidu.com/link?url=GuhDQJR7Xnz0D_PifjZVa1jMJtC-iFqlbh_qphD8egqzM_2fkYZJLCaj8sfpFuJ5gocOgVM3vv-U2c_NX_AlhbEd0BhLCW4bagPjP3CYF1Qmq.
[38] 吴晓燕. 高维数据空间中离群点检测算法的研究[D]. 南京:南京财经大学,2010.
[39] 王芳. 基于属性重要度的属性约简算法研究[D]. 成都:电子科技大学,2011.
[40] Ye Zhengwang. The research of intrusion detection algorithms based on the clustering of information entropy [C]. Wuhan: Hubei University of Technology,2010:552-555.
[41] 陈源,曾德胜,谢冲. 基于聚类的属性约简方法[J]. 计算机系统应用,2009,(5):173-176.
[42] 孟小峰,慈祥. 大数据管理、概念技术与挑战[J]. 计算机研究与发展,2013,50(1):146-169.

(下转第 495 页)

Abstract: This was the first time to use HPLC to detect the contents of huperzine A(HupA) in the whole plant and different organs of *Huperzia serrata* from Mount Lushan and Mount Jinggang in Jiangxi Province at different seasons. The results showed that the whole plant content of HupA collected from Mount Lushan and Mount Jinggang increased with season changed and reached the maximum in December which reached to 211. 9 μg · g^{-1} and 325. 9 μg · g^{-1}. The content of HupA in different organs collected from two sample sites were various. Stem and leaf contained more HupA,while only a little HupA was detected in root. The results showed the temporal and spatial dynamic change rule of HupA in *H. serrata*,and may provided scientific reference for the reasonable use of rare plant resources.

Key words: *Huperzia serrata*; huperzine A; content detection; temporal and spatial variation

(责任编辑: 刘显亮)

(上接第 458 页)

The Outlier Detection Method and Its Improvement in the Eea of Big Data

MIAO Yong-chun,CHENG Yan*

(College of Computer Information and Engineering,Jiangxi Normal University,Nanchang Jiangxi 330022,China)

Abstract: The paper compared and analyzed major outlier detection method and their features and merit and demerit were summarized. In addition,in view of the large amount of data and high dimension of the big data,improvement strategies of outlier detection method were analyzed. Improvement strategies of outlier detection were further illustrated by T-ODCD and AROD algorithms.

Key words: big data; outlier detection method; improvement strategies

(责任编辑: 冉小晓)