文章编号:1671-4598(2018)05-0046-04 DOI:10.16526/j.cnki.11—4762/tp.2018.05.012 中图分类号:TP319 文献标识码:A

基于互相关分析和SOM神经网络的异常值检测平台

路 辉,刘雅娴
(北京航空航天大学 电子信息工程学院,北京 100191)

摘要:异常值的检测问题是时下数据挖掘领域的研究热点;目前已经有许多种成熟的异常值检测方法,但当数据是高维混合型属性,或者存在成片孤立点时,这些方法就变得很不理想甚至不再适用;因此,针对这些现有方法的不足之处,提出了新的孤立点检测方法,并设计了时域和空域的异常值检测平台;对于时间和空间序列数据集,该平台分别采用基于互相关分析和自组织竞争(self—organizing maps,SOM)神经网络的异常值检测方法;经实验验证,检测平台具有较高的检测率和可靠性;同时,在搭建该平台时充分考虑了模块化和层次化的方式,使得平台具有良好的可扩展性和开放性。

关键词:异常值检测;系统软件平台;互相关分析;SOM神经网络

An Outlier Detection Platform Based on Cross—Correlation Analysis and SOM Neural Network

Lu Hui, Liu Yaxian
(College of Electronic Information Engineering, Beihang University, Beijing 100191, China)

Abstract: The outlier detection problem has become the focus of research in the field of data mining. At present, there are many kinds of mature outlier detection methods. But when the data has high dimensional mixed property, or there are assembled outliers, the result of these methods become unsatisfactory or imapplicable. Therefore, in view of the shortcomings of these existing methods, a new outlier detection method is proposed in this paper. Meanwhile, we designed the outlier detection platform in time and space domain. For time and spatial series datasets, the platform is based on cross—correlation analysis and self—organizing maps (SOM) neural network clustering. It can be proved by experiments that the platform has higher detection rate and reliability. By the way, the platform is built in a modular and hierarchical way with good openness and extensibility.

Keywords: outlier detection; software platform; cross—correlation analysis; SOM neural network

0 引言

随着互联网的爆炸式发展和云时代的来临,人们利用网络获取收集信息的同时,数据量呈现出指数爆炸式的增长,大数据吸引了越来越多的关注。大数据的5 V特点:Volume(大量)、Velocity(高速)、Variety(多样)、Value(低价值密度)、Veracity(真实性),使得传统常规的小规模数据处理方法变得不再适用,数据挖掘技术应运而生[1]。

数据挖掘指的是通过特定算法发掘大量数据中的重要隐含信息的过程,其通常包含四种类型,分别是关联规则发现、类别的描述、类别的判断和孤立点发现[2]。孤立点发现是数据挖掘中的一个重要的分支,已经成为数据挖掘领域的研究热点,并且在天气预报、股市分析、市场调研、药品研发[3]、医疗保险[4]、网络入侵检测[5]、商业欺诈检测[6]、飞参数据、无线传感器网络[7]等领域得到了广泛的应用。

孤立点通常是指数据集中不符合一般模型的异常对象。由于孤立点既有可能是噪声也有可能隐藏着比一般数据更为重要的信息,随意删除孤立点数据可能导致这些有价值信息的丢失,所以通过孤立点检测技术发现并利用在孤立点中的有用信息具有非常重要的意义[8]。孤立点检测也称孤立点挖掘,是从海量数据中找到异常行为,并发现异常行为中所蕴含的信息

的过程,这无疑是一项极富挑战性的工作。目前对异常值检测技术的研究主要分为两个方面:理论方法研究和软件平台搭建。本文对异常值检测课题的研究也将从这两个方面展开。

1 系统结构及原理

本文搭建的异常值检测平台可以实现时间序列和空间序列的异常值检测。当输入数据集为时间序列时,采用基于互相关分析的异常值检测方法;当输入数据集为空间序列时,采用基于SOM神经网络聚类的异常值检测方法。具体系统平台结构如图1所示。

2 异常值检测算法

2.1 时间序列异常值检测

时间序列是指将同一统计指标的数值按其发生的时间先后顺序排列而成的数列,通常时间序列的特点是具有一定的趋势性和相关性,时间序列分析的主要目的是就是根据这两个特点,利用已有的历史数据对未来进行预测[9]。如果时间序列中

收稿日期:2018-03-09; **修回日期**:2018-03-30。

作者简介:路 辉(1977-),女,黑龙江省肇东市人,教授,博士研究生导师,主要从事无线电导航、信息系统测试与性能评估技术方向的研究。

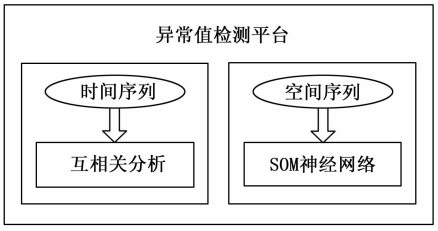

图 1　异常值检测平台结构

存在异常值，将会严重影响到数据的预测效果。因此，时间序列异常值的检测很有意义。

对于混合型属性数据集，我们先将其分为数值型属性数据和非数值型属性数据两部分分别处理。对于非数值型属性数据进行出现频率的统计；对于数值型属性先进行利用线性插值将成片孤立点转换成单独孤立点，再对相邻采样时刻序列求互相关函数，最后用多级最大类间方差算法自适应地选取阈值，并将孤立点分级输出。当检测出的孤立点数目达到预设的比例时，检测算法停止并输出检测结果。具体的算法原理框图见图 2。

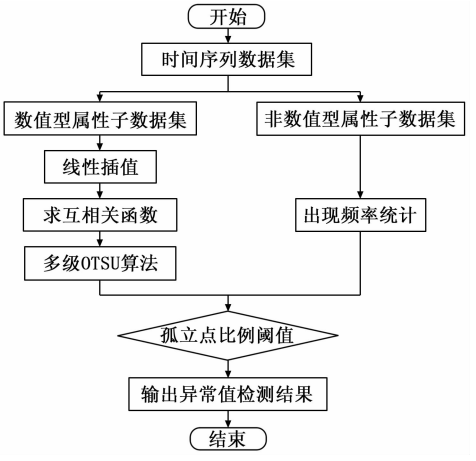

图 2　基于互相关分析的异常值检测算法原理框图

实际系统中的异常值按其分布范围可以分为两类：一类是单独出现的异常值点，另一类是小范围内成片出现的异常值群。对于成片出现的异常值群采用了线性插值的方法对数据做预处理。利用剪枝策略确定数值型属性数据集中一定不是异常值的正常点，并利用这些正常点对原始数据进行线性插值，使得小范围聚集的异常值数据之间插入正常数据点，转化为单独孤立点的情况进行检测。

对于只含有单独孤立点的时间序列数据集，采用互相关分析的方法进行孤立点检测。每个采样时刻处的各个维度的数据值都可以组成一个采样时刻序列。选取前两个采样时刻序列作为一个滑动计算窗，计算这两个采样时刻序列的互相关系数，得到互相关函数的第一个点。将活动窗口按照步长为 1 的间隔移动，长度为 L 的多维数值型属性序列就可以生成一个点数为 L－1 的互相关函数。由于互相关系数反映的是相邻采样时刻序列的相似程度，所以互相关函数的谷值处可能存在孤立点。

由于正常点集与异常点集之间可能存在重合的位置关系，异常点和正常点之间的分界可能是模糊的，所以有时把一个数据点绝对的定义为异常点或正常点是不科学的，这时用孤立点分级的概念就可以定义孤立点的孤立程度，避免一概而论。因此孤立点检测系统阈值的自适应选择和孤立点的分级输出是必须要解决的两个问题。最大类间方差算法，简称 OTSU 算法，最先在数字图像处理邻域中用于划分前景和背景。OTSU 算法可以将样本点以类间方差最大，类内方差最小的原则分为两类，并自适应地给出分类阈值。因此，算法对互相关函数幅值采用多级最大类间方差算法分类，将第一次分类产生的孤立点记为 1 级孤立点，并将此次分类产生的正常点集进行二次分类，如此重复下去，逐步筛选出各级孤立点分级输出，直到产生的孤立点个数到达预先设定的比例，并将最后一次最大类间方差算法的阈值记为最终的互相关系数阈值。

在非数值型空间内，异常点的非数值型属性值相比正常点出现的更不频繁。基于这种考虑，我们先找出数据集中有几种非数值属性值，再对各个非数值型属性值出现的频数做出统计，得到非数值型属性频数表。依次找出出现频数最少的非数值型属性数据，直到累计频数超过预先设定的孤立点比例阈值，停止算法并输出对应的非数值型属性异常值。

2.2　空间序列异常值检测

典型 SOM 网共有两层，输入层模拟的是人眼的视网膜，用于接收外界数据，一般为单层神经元排列；输出层模拟的是人脑的大脑皮层，用于处理输入层获得的数据信息，输出层的每个神经元都与它周围的其他神经元相互连接，排列成二维的棋盘状平面[10]。SOM 神经网络可以将任何高维数据集映射到二维的输出神经元平面上，映射的获胜神经元相距越近则它们所对应的原始数据对象就越相似。

基于 SOM 神经网络的异常值检测方法如下：第一，对迭代次数、输出结点权值向量、学习率和邻域半径进行初始化。对输入向量进行归一化。第二，对于每一个输入向量，利用欧式距离相似性度量准则寻找与其对应的竞争获胜神经元。第三，利用梯度下降法，对获胜节点和其邻域范围内神经元集合中的每一个节点都进行权值更新。第四，在所有输入向量均遍历结束，并且所有获胜神经元权值更新之后，更新学习率和邻域函数。第五，当神经网络的训练次数达到预设的最大次数时，退出训练学习过程，即可得到训练好的 SOM 神经网络。否则转入第二步。第六，将 4－邻域内没有其他获胜神经元，并且类规模很小的孤立聚类视为孤立点输出。算法原理的整体框图见图 3。

3　异常值检测平台

3.1　检测平台概述

本系统在 MATLAB 开发环境下，实现了基于互相关分析的时间序列孤立点检测算法和基于 SOM 神经网络聚类的非时间序列孤立点检测算法，提供了相应的数据集输入接口以及检测结果输出界面。

系统设计功能主要分为两个方面：（1）时间序列异常值检测；（2）非时间序列异常值检测。总体框图见图 4。

软件可以用互相关分析法检测时间序列数据集中的异常值，也可以用 SOM 神经网络聚类法检测非时间序列数据集中的异常值。针对检测结果提供了算法评价功能，显示出每次检

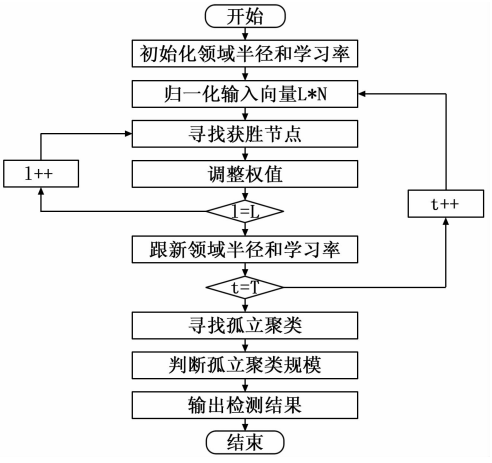

图 3 SOM 神经网络聚类算法原理框图

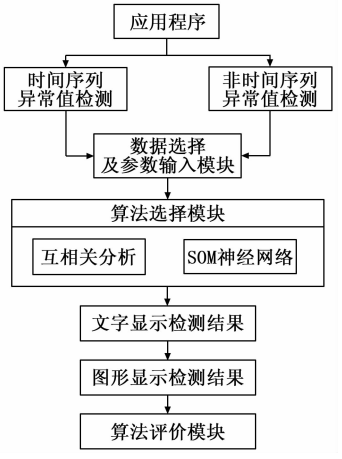

图 4 系统软件功能框图

测得到的真实值个数、真正异常值个数、伪异常值个数以及误判为正常值数据个数，并利用这四个数据计算出评价系数，反映孤立点检测结果的优劣。结果分析时，系统不仅可以用文字描述出异常值在数据集中的位置，而且提供图形化显示结果功能，并将异常值标记出来。

为了增加系统软件的可扩展性和开放性，设计时所遵从的总体原则和思路如下：提供数据输入接口，使用户可以灵活选择待检测的数据集；将每个算法封装成单独的模块和子界面，使用户可以在系统软件上增加新的模块和算法，满足程序设计的模块化和层次化要求；检测结果不仅有文字描述，还有图形显示功能，更加具体生动，有说服力，便于用户灵活分析检测结果，采用正确的方法处理检测到的异常值；系统软件提供检测结果评价和分析功能，帮助用户合理的调整孤立点比例阈值，使检测效果达到最佳；友好的人机交互界面设计，便于用户操作。

3.2 人机交互界面

系统主界面将数据集分为时间序列和非时间序列两大类供用户选择。当用户选择时间序列并点击开始按键时，会进入互相关分析法子界面。互相关分析法子界面见图 5。

该界面由 7 个部分组成。输入数据选择区可以让用户在计算机上选择一个 Excel 数据文件作为孤立点检测的原始数据；

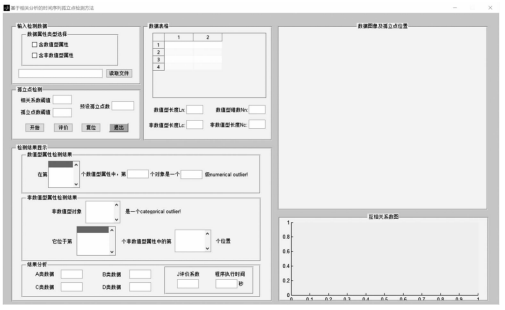

图 5 互相关分析子界面

原始数据表格区可以以表格的形式在界面上展示原始数据，方便用户查看；用户系统控制区可以让用户输入判决阈值和预设孤立点个数，控制检测系统的启动、停止和复位；检测结果显示区可以告诉用户检测出的孤立点的数据值和其在数据集中的具体位置；检测算法评价区可以通过构造混淆矩阵计算四类数据个数来评价孤立点检测算法的检测率和误报率；互相关函数图象区可以画出数值型属性数据相邻采样时刻的互相关函数图像；孤立点数据图像区可以画出各个属性上数据变化的折线图，并在其上标出孤立点的位置。

当用户选择非时间序列并点击开始按键时，会进入 SOM 神经网络聚类子界面。SOM 神经网络聚类子界面见图 6。

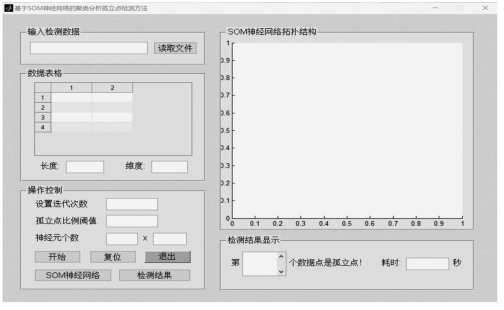

图 6 SOM 神经网络子界面

该界面由 5 个部分组成。与互相关分析子界面类似，该界面同样可以选择输入数据、展示原始数据、控制相关参数、显示检测结果，同时还可以画出神经网络的拓扑结构，标出获胜神经元和孤立点。

4 实验结果与分析

4.1 时间序列异常值检测实验

以 10 个国家（澳大利亚、巴西、加拿大、中国、埃及、法国、德国、日本、英国、美国）从 1950 年到 2008 年的人口变化数据为例验证基于互相关分析的时间序列异常值检测方法。同时，在数据集后补充了 2 个非数值型属性，每个非数值型属性同样也有 59 个数据对象。

为了验证基于互相关分析的算法，在各个属性中人为伪造了 8 个异常值，利用基于互相关分析的算法检测该数据集的异常值情况。8 个异常数据点的位置分别为：第 4 个属性的第 9 个对象、第 7 个属性的第 19 个对象、第 9 个属性的第 37 个对象、第 10 个属性的第 46～48 个对象、第 11 个属性的第 8 个对象和第 12 个属性的第 31 个对象。其中第 10 个属性的第 46－48 个对象为小范围内聚集的成片孤立点，第 11 个属性的第

8 个对象和第 12 个属性的第 31 个对象为非数值型属性孤立点。检测结果如图 7 所示。

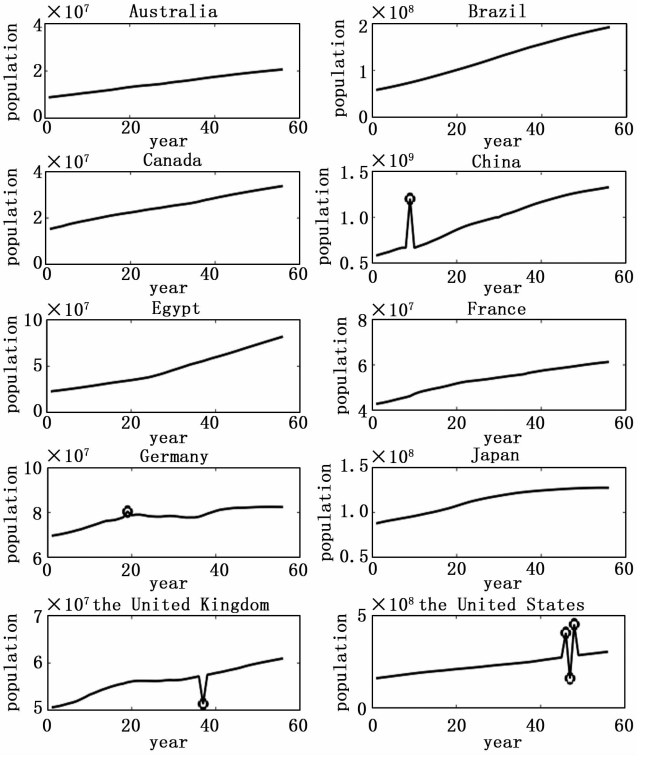

图 7　时间序列数据集异常值检测结果

经验证，算法可以检测出所有的数值型属性和非数值型属性的异常值，没有出现虚警漏警的现象，评价系数为 1。

4.2　空间序列异常值检测实验

以每 30 个数据为一类，共分 5 类的聚类分析数据为例验证基于 SOM 神经网络的非时间序列异常值检测方法。聚类分析数据共有 5 个属性，150 个对象。为了便于验证孤立点检测的效果，在数据集中人为加入 4 个孤立点，分别位于第 49 个数据对象、第 50 个数据对象、第 120 个数据对象和第 140 个数据对象处。其中第 49 个数据对象和第 50 个数据对象处的孤立点分布趋于一致，属于小范围内聚集的成片孤立点，第 120 个和第 140 个数据对象处的孤立点为单独出现的孤立点。得到的孤立点检测结果如图 8。

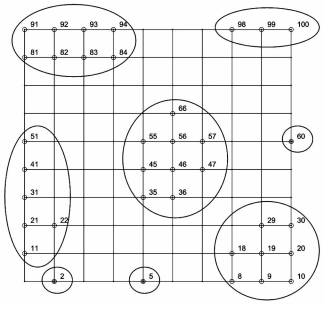

图 8　空间序列数据集异常值检测结果

其中，网线代表 SOM 神经元之间的连接关系，网线的交点代表 SOM 神经网络输出层的所有神经元，数字标记处的神经元代表在训练过程中获胜的神经元（右上角的数字仅代表获胜神经元编号，只为标记使用，无其他含义）。此拓扑结构图的获得过程详见异常值检测算法章节中的图 3。不难看出，获胜神经元按其分布位置可以自然地分为 8 类，2 号、5 号和 60 号神经元各自被划分为一类。这三个规模较小的孤立聚类即为算法检测出的孤立点。其中第 2 号神经元代表第 120 数据对象处的单独孤立点，第 5 号神经元代表第 140 数据对象处的单独孤立点，第 60 号神经元代表第 49 个数据对象和第 50 个数据对象处的小范围聚集成片孤立点。检测结果直观，检测效果良好。

5　结束语

本文针对异常值检测这一热点研究方向，结合现有方法的不足之处，设计了一种时域空域相结合的异常值检测系统。异常值检测系统是在模块化、层次化的基础上搭建了统一的孤立点检测软件平台，具有一定得开放性和可扩展性，可以对各类属性进行异常值分析和数据处理，也可以在软件平台上加入其他的孤立点检测算法，还可以扩展到其他领域的数据。该平台对于时间序列数据集和空间序列数据集分别采用了基于互相关分析和 SOM 神经网络的方法来进行异常值检测。经验证，异常值检测系统能够较好地完成时域和空域的异常值检测任务，具有较高的检测效率和可靠性。

参考文献：

[1] 顾　荣. 大数据处理技术与系统研究 [D]. 南京：南京大学，2016.

[2] 谢方方. 基于距离的孤立点挖掘在计算机取证中的应用研究 [D]. 山东师范大学，2014.

[3] Xie Z，Li X，Wu W，et al. An Improved Outlier Detection Algorithm to Medical Insurance [M]. Intelligent Data Engineering and Automated Learning－IDEAL 2016. Springer International Publishing，2016.

[4] Christy A，Gandhi G M，Vaithyasubramanian S. Cluster Based Outlier Detection Algorithm for Healthcare Data [J]. Procedia Computer Science，2015，50：209－215.

[5] Tao Li. Novel heuristic dual－ant clustering algorithm for network intrusion outliers detection [J]. Optik－International Journal for Light and Electron Optics，2015，126 (4)：494－497.

[6] Carneiro N，Figueira G，Costa M. A data mining based system for credit－card fraud detection in e－tail [J]. Decision Support Systems，2017，95.

[7] Abid A，Masmoudi A，Kachouri A，et al. Outlier Detection in Wireless Sensor Networks Based on OPTICS Method for Events and Errors Identification [J]. Wireless Personal Communications，2017 (1)：1－13.

[8] 鄢团军，刘　勇. 孤立点检测算法与应用 [J]. 三峡大学学报（自然科学版），2009，31 (1)：98－103.

[9] 杨　樱. 基于强化学习的绩优股票预测系统研究 [D]. 秦皇岛：燕山大学，2006.

[10] Cao C，Zhang W，Wang Z，et al. The Diagnosis Method of Stator Winding Faults in PMSMs Based on SOM Neural Networks [J]. Energy Procedia，2017，105：2295－2301.