基于人工神经网络的多维离群点检测算法

梁 兵,卢建军,卫 晨

(西安邮电大学 通信与信息工程学院,陕西 西安 710121)

摘 要:为了更加智能地检测离群点,克服传统离群点检测算法的机械性,提升多维数据集合离群点挖掘效率,在传统的离群数据挖掘算法的基础上,提出了一种基于人工神经网络的多维离群点检测算法。仿真实验结果表明,该算法具有对用户依赖性小、检测精度高的优点,为检测离群点提供了一种新的路径。

关键词:人工神经网络;多维数据;智能化;熵权

中图分类号:TN391.5 **文献标识码**:A **文章编号**:1674-7720(2014)05-0076-03

Outlier detecting algorithm for multidimensional datasets based on ANN

Liang Bing, Lu Jianjun, Wei Chen

(School of Communication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China)

Abstract: To detect outlier more intelligently, avoid the mechanical character of traditional algorithm of outlier detecting and improve the efficiency of outlier mining for multidimensional data sets, this paper proposes an algorithm of outlier detecting for multidimensional data sets based on ANN founded on the pros and cons of traditional algorithm of outlier mining. The simulation results indicate that the algorithm has the advantages of less dependent to users and higher accuracy. It opens up a new approach for the outlier detecting.

Key words: ANN; multidimensional data; intelligentialize; entropy

离群点(Outlier)就是明显偏离其他数据、不满足数据的一般模式或行为、与存在的其他数据不一致的数据[1]。离群点检测的目的在于从海量数据中找出具有明显异常行为的数据。离群点的检测应用于多个行业,如通信盗用、网络病毒检测、疾病诊断等方面。目前有一些高效的离群点检测挖掘算法,比如基于统计的、距离的、深度的、密度的方法,参考文献[2]-[6]中较为详细地介绍了这些方法和各自的局限性。

这些传统方法虽然有时针对各自的检测对象具有良好性能,但是前提是必须对数据集有很深入的了解,比如用基于统计的方法,需要预先知道数据集属于什么分布。这些传统方法没有智能挑选的能力,不会从复杂数据集中找出潜藏的规则。如有一组数据 A=[1 2 4 8 15 32],如果按照基于距离的离群点检测方法检测,最有异常行为的数据是 32,但是如果经过训练与预测,可发现 15 这个点在这里才具有最异常的行为。因此,找出数据集中潜在的规则是很有现实意义的。人工神经网络使得解决这一问题变成了一种可能。

高维空间点的数据特性决定了其检测与低维数据集有很大的区别。首先,与低维空间不同的是高维空间中的数据分布比较稀疏,造成高维空间中数据之间的距离尺度及区域密度不再具有直观的意义[7]。从一个数据点来看,其他点到它的距离落在一个范围很小的区间内,很难给出一个合适的近似度阈值来确定哪些点与它相似,哪些点不是。另外,对高维数据的估计需要的样本个数与维数构成指数增加的关系,这在机器学习中称作著名的维数灾难(Curse of Dimensionality)。大量的数据分析问题本质上是非线性的,甚至是高度非线性,对此不能利用已有的快速成熟的线性模型进行研究[8]。

因此引入熵权的概念,通过它能知道每个属性对于离群点的贡献程度,较好地解决了非线性问题,而且分开对于每个属性值进行预测,然后做一个统计求和,对于位于维数灾难有了较好的解决。

1 相关工作

1.1 人工神经网络

人工神经网络(ANN)是一种应用类似于大脑神经突触连接的结构进行信息处理的数学模型[9]。ANN是一个由大量简单的处理单元组成的高度复杂的大规模非线性自适应系统[10]。它是对巨量信息并行处理和大规模平行计算的基础,既是高度非线性的动力学系统,又是自适应组织系统,可用来描述认知、决策及控制的智能行为。对于处理大量原始数据而不能用规则或公式描述的问题,ANN则表现出极大的灵活性和自适应性。

1.2 BP神经网络的基本结构以及工作范式

BP网络是误差反向传播神经网络的简称,由输入层、隐含层、输出层组成。每一层由一个或多个神经元组成。隐含层可以包括BP网络的结构,如图1所示。

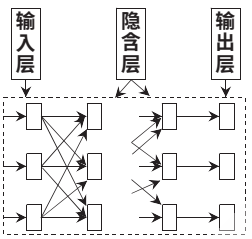

图1 BP结构模型

BP神经网络的输入层接收输入样本信息,隐含层对输入信息进行处理,输出层负责处理后的结果。如果输出层结果与预测值有误差或者误差大于给定阈值,则网络将误差反向通过输出层传递给隐含层,经过隐含层处理后,传递给输入层,期间相邻网络层之间的连接权值经过多次的权值修正。由此通过多次传输与反向传输,相邻层之间的连接权值通过不断修正,从而将误差控制到给定阈值范围之内,至此,学习结束。权值不断调整的过程就是网络学习的过程。BP神经网络最直接的优点就是与大脑认知具有一定的相似性,如容错性、学习能力、非线性等。

1.3 相关定义与公式

定义1 r_{ji}称为第j个对象在i个属性上的值,且$r_{ji}\in[0,1]$,则在n个对象d维属性中,第i维属性的熵定义为:

$$H_i=-k\sum_{j=1}^{n}p_{ji}\ \ln\ p_{ji} \tag{1}$$

式中,$p_{ji}=\dfrac{r_{ji}}{\sum_{j=1}^{n}r_{ji}}$,$k=\dfrac{1}{\ln\ d}$。

定义2 第i维属性的熵权$\tilde{\omega}_i$定义为:

$$\tilde{\omega}_i=\frac{1-H_i}{d-\sum_{i=1}^{d}H_i} \tag{2}$$

式中,$0\leqslant\tilde{\omega}_i\leqslant1$,$\sum_{i=1}^{d}\tilde{\omega}_i=1$。

定义3 每个对象的离群程度O_j定义为:

$$O_j=\sum_{i=1}^{m}\frac{|r_{ji}-r'_{ji}|}{r'_{ji}}\tilde{\omega}_i \tag{3}$$

式中,r_{ji}和r'_{ji}分别表示对象j在i属性上的原始值和预测值。

为了防止熵值计算中对数计算无穷大的情况,必须进行极差变换,将极差映射到0.1~0.9之间,数据预处理所用到的极差公式为:

$$x_{ik}=0.8\frac{(x''_{ik}-x''_{k\min})}{(x''_{k\max}-x''_{k\min})}+0.1 \tag{4}$$

式中,$x''_{k\max}$和$x''_{k\min}$分别表示最大值和最小值。

2 算法描述及伪代码

本文算法(BAOA)将所选数据分为训练数据和检测数据(预测数据)。算法将训练数据当做全部非离群点进行训练而找出隐藏规则,然后将这规则应用于检测数据的预测。所选训练数据通常占全部数据比率为8.5~11.5%左右(此时数据量也比较大),这样既可以保证训练的有效性(找出隐藏规则),同时又能保证丢失掉的训练数据中的离群点(如果存在)对于全部离群点来说影响又不大。该算法除在训练点数据个数的选取上较为新颖且有实际意义外,而且中间加入判定有无预测值的算法,对于没有预测值的数据点赋予一个经验值,这样更能维持数据监测的稳定性。

该算法首先对原始数据集中每一个属性对应的值进行极差变换,然后计算每一个属性的熵权,而后对数据集中的训练数据的每一个非空间属性按照顺序排列后经过所选人工神经网络模型进行训练,然后对于剩下的所有数据(检测数据)的每一个属性按照顺序排序后经过所选神经网络模型进行预测,然后经过算法的判断函数,将没有预测值的属性值人工赋予一个预测值(在经验波动范围内),保证每个待检测的数据点都有预测值。而后将预测值作为标准值,通过计算每一个属性值自身的的偏差,再结合每一个属性熵权对它进行处理,得出每一个数据点的离群程度大小,最后按照离群程度从大到小的顺序进行排序。

3 仿真

仿真操作系统和软件:win7-32、Matlab

仿真对象:葡萄酒识别数据

所选数据描述:所选数据来源于由C.Blake于1998年9月21日更新的数据集,它分为低中高三种,个数分别为63,1319,27。有12属性,分别为:酒精、苹果酸、灰、镁、总酚类、黄酮、Nonflavanoid酚类、原花色素、颜色强度、色相、0D280/0D315稀释葡萄酒、脯氨酸。

所选ANN网络:BP网络

输入个数J:4

输出个数K:1

隐含层个数Y:6

处理说明：在训练和预测时，每次都是对属性值排序后进行训练和预测，这样更容易找出隐藏规则，计算效率更高，预测效果更好。后 $s+1$ 到 n 个数据点每个属性预测时，前 J 个作为输入值时，它没有对应的预测值。对此进行的处理是此时赋予它一个合适的值（波动大小在经验范围内），此次仿真过程中是赋予一个和原始值一样的值作为预测值。虽然后 $s+1$ 到 n 个数据点每个对象按照每个属性每次排序后对应的前 J 个值 id 不一样，但是因为数据海量，且维数较多，这样处理后对于离群点的预测并无大的影响。图 2 为后 800 个葡萄酒样本中脯氨酸的属性值的真实值和预测值。

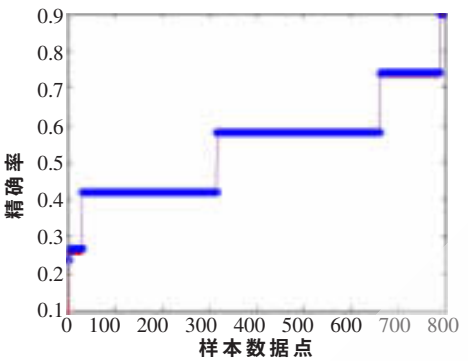

图 2 脯氨酸的真实值和预测值

对后边 800 个数据前 45 个输出后，对照原始数据集得知，在训练点个数为 160（11%）时，有 42 个点为低等或者高等，离群点正确率达到了 93.33%。对比几种高效的多维离群点检测算法，可以发现这一算法的离群点检测准确率更高。将 LOF、SPOD 和本文算法 BAOA 的算法精确度在不同训练数据下进行比较，可以发现本文这种算法精确率更高，如图 3 所示。

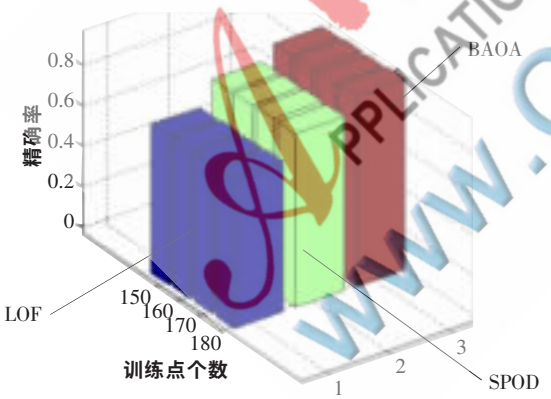

图 3 三种算法的比较

本文针对高维空间中数据的特点，提出了一种智能找出隐藏规则并且自动检测离群点的算法。对于多维复杂且对离群点特征没有明显约束的数据集，ANN 表现出了它的优越性。仿真结果表明，通过 ANN 建立的多维离群点检测，具有传统方法无可比拟的智能性，而且检测精度较高。为各位离群点检测相关专业人员和业务爱好者提供了一种思路。

参考文献

[1] HAWKINS D M. Identification of outliers [M]. London: Chapman and Hall, 1980.

[2] HAN J, KAMBER M, PEI J. Data mining: concepts and techniques[M]. Morgan kaufmann, 2006.

[3] WANG L, ZOU L. Research on algorithms for mining distance based outliers [J]. Chinese Journal of Electronics, 2005, 14(3): 384-387.

[4] SHEKHAR S, LU C T, ZHANG P. A unified approach to detecting spatial outliers [J]. GeoInformatica, 2003, 7(2): 139-166.

[5] AGGARWAL C C, YU P S. Finding generalized projected clusters in high dimensional spaces[M]. ACM, 2000.

[6] 魏黎，宫学庆，钱卫宁，等.高维空间中的离群点发现[J].软件学报，2002，13(2)：280-290.

[7] SHEKHAR S, LU C T, ZHANG P. A unified approach to detecting spatial outliers[J]. GeoInformatica, 2003, 7(2): 139-166.

[8] 傅荟璇，赵红.MATLAB 神经网络应用设计[M].北京：机械工业出版社，2010.

[9] 钟义信.知识理论与神经网络[M].北京：清华大学出版社，2009.

[10] 闵剑. 人工神经网络在石化项目绩效评价中的应用研究[D].北京：清华大学，2009.

（收稿日期：2013-12-19）

作者简介：

梁兵，男，1987 年生，硕士研究生，主要研究方向：煤炭企业信息化、数据挖掘。

卢建军，男，1962 年生，教授，主要研究方向：企业信息化发展战略与建设模式的研究，企业专用综合信息网组网技术及业务应用。

卫晨，男，1983 年生，助教，硕士，主要研究方向：电子商务。