基于统计聚类 RBF 神经网络的孤立点检测研究

周 凯
(中国科学院研究生院 北京 100738)

摘 要 孤立点挖掘是数据挖掘的一个重要领域,而统计分析方法在孤立点检测中具有天然的优势。本文将统计聚类方法融入 RBF 神经网络,提出了一种基于统计聚类 RBF 神经网络的新的孤立点检测算法—SCRBF。该算法包括两部分,先用统计聚类方法对神经网络进行初始化,然后根据网络的训练情况进行隐单元的简化,提高了神经网络的泛化能力,同时也降低了过拟合现象的出现概率。与 LSC 算法的对比实验表明,该算法是有效的。

关键词 统计方法,聚类,RBF 神经网络,孤立点检测

A New Isolated Point Detecting Algorithm Based on Statistical Clustering RBF Neural Network

ZHOU-Kai
(Graduate School of China Academy of Sciences, Beijing100738)

Abstract Mining isolated point is an important field in Data Mining. Methods of statistical analysis have natural advantage in detecting isolated points. In this paper, statistical clustering is first integrated into RBF Neural Network and a new isolated point detecting algorithm based on statistical clustering RBF Neural Network, SCRRBF is proposed, which has two steps. The first step is initializing the neural network using statistical clustering, and the second is to reduce the concealing units of neural network according to the training situation. Using this, the generalization of neural network can be improved and the Over-fitting phenomenon can be reduced. With experimental contract to LSC algorithm, SCRBF is effective.

Keywords Statistical method, Clustering, RBF neural network, Isolated point detecting

1 引言

随着交叉学科研究的不断深入,目前统计分析已经成为数据挖掘领域的一种重要方法,因为数据挖掘本身就是从一堆看似杂乱无章的数据中发现规律,这与统计学中的统计分析方法有着非常大的相似之处。目前形成的很多数据挖掘方法、算法也都引入了统计分析的方法。因此,我们可以把统计方法看作是数据挖掘的参数方法,这也就暗含了可以获得或收集到正确的函数类型,参数数目以及参数可能的值。统计学为数据分析提供了大量的技术方法和结论,因为在实际应用合理的条件下可以证明数据挖掘中所使用的估计和搜索过程是一致的。在统计学里,需要对数据彻底地了解来获得正确的参数化模型。而在实际运用中,基于统计的数据挖掘方法可以和另一种重要的数据挖掘方法——机器学习方法相互补充,也可以融合在一起使用。因此,我们可以得出这样的初步结论:统计方法与计算机技术的融合是现代经济条件和科学技术发展的必然趋势。

2 孤立点检测及研究现状

在实施数据挖掘应用中可能会碰到这样的情况:经常存在一些数据对象,它们不符合数据的一般模型,与数据的其它部分不同或不一致,这样的数据对象被称为孤立点。孤立点的产生可能是度量或执行错误所导致的,也可能是固有的数据变异性的结果。孤立点检测的研究对象是数据集中偏离绝大多数对象的很小一部分数据。孤立点检测在数据挖掘领域是一项重要的挖掘技术,也是一项重要任务,有着广泛的应用。一方面,孤立点检测能用于欺诈检测,探测不寻常的信用卡使用或电信服务;或者在市场分析中用于确定极低或极高收入客户的消费行为;或者在医疗分析中用于发现对多种治疗方式的不寻常的反映。因此,在某些许多数据挖掘应用中,研究孤立点可能比研究聚类更有用、更重要。另一方面,在进行原始数据预处理时,由于错误数据往往表现为孤立点,因此通过检测并去除数据源中的孤立点可以达到数据清理的目的,从而提高原始数据的数据质量。

一个典型的孤立点检测问题可以描述如下:给定一个 n 个数据点或对象的集合,及预期的孤立点的数目 k,发现与剩余的数据相比是显著相异的、异常的或不一致的头 k 个对象。一般的方法是将孤立点检测看成两个子问题:①在给定的数据集合中定义什么样的数据可以被认为是不一致的;②找到一个有效的方法来检测这样的孤立点。

在数据挖掘中,孤立点检测算法大体上可分为以下几类:统计学方法、基于距离的方法、基于偏离的方法和基于密度的方法。基于密度的方法能够挖掘出比基于距离异常算法更多的异常数据,而且能够找到局部异常。局部稀疏系数(LSC)就是其中一种[1],但 LSC 算法具有先天的缺陷。一是该算法需要计算数据集中每个对象的局部稀疏率和局部稀疏系数,当数据集很大时,计算每个对象的局部稀疏率和局部稀疏系数将耗费很大的计算量;二是受噪声影响较大。现有的很多孤立点检测算法大都是采用每样本数据调整的训练方式,即在每一个样本进入网络之后,就进行参数调整。而样本

周 凯 硕士研究生,主要研究方向为计算机在统计学方面的应用。

中数据一般都是存在噪声的。每样本调整的训练方式，使网络更容易受到少数受噪声影响较严重的样本不良影响，降低算法效率和网络泛化能力。所以，如果网络建立之初学习的前几个样本刚好受噪声影响较大，那么对于网络的健康生长将是十分不利的。

3 基于统计聚类 RBF 神经网络的孤立点检测

近年来，神经网络技术在数据挖掘中的应用日益广泛。神经网络最有用的特性之一就是天生具有的学习能力使它能够实现输入空间到输出空间的映射。在各种神经网络方法中，径向基函数(RBF)神经网络最为常用，它是以函数逼近理论为基础的一类前向网络。有关 RBF 神经网络的具体概念请参见有关文献[2]，此处不做赘述。本文将统计学方法引入基于 RBF 神经网络的孤立点检测，提出了一种新的孤立点检测方法——基于统计聚类 RBF 神经网络的孤立点检测算法(SCRBF)。该算法一方面利用统计聚类方法对网络进行初始化，发现不活跃的隐层神经元，从而降低甚至排除孤立点(噪声数据)作为有用的隐层神经元进入神经网络的可能性；另一方面通过一定的简化策略降低 RBF 神经网络的复杂性，从而提高 RBF 神经网络用于孤立点检测的效果。

3.1 算法描述

(1)基于统计聚类方法的网络初始化

网络初始化的目的就是在现有样本集或子集上，运用统计聚类方法初步确定隐层中心，建立初始结构，然后再运行顺序学习算法，这样将获得更好的效率和效果。很多聚类技术可以用于训练样本的分类和隐层中心的选取。为了更好地回避噪声的影响，本文采用的是 K-中心聚类算法。采用 K-中心聚类算法从初始训练样本集中选出 k 个聚类中心之后，就能够以它们为隐层基函数中心，初步确立 RBF 网络结构。各隐层基函数的宽度 σ 可以取为与最近一个其他隐节点中心距离的 1.5～2 倍。之后，再循环启用此算法进行下一步的网络训练。在初始训练样本集较大的情况下，还可以随机抽取一个较小子集进行初始化。

假设需要将 N 个数据点 x^p，$p=1,2,\cdots,N$ 分到 K 个聚类当中并找出聚类中心。K-均值聚类算法的目标就是最小化样本-聚类中心距离平方和：

$$J=\sum_{j=1}^{n}\sum_{p\in S_j}\| x^p-\mu_j \|^2$$

其中，μ_j 是所在类 S_j 的聚类中心，K-均值聚类算法中取为所有类内数据点的几何平均值：

$$\mu_j=\frac{1}{N_j}\sum_{p\in S_j}x^p$$

可以用以下整批次的方式实现 K-均值聚类算法。首先，随机地把样本数据点分配到 K 类中，并计算各类的聚类中心。然后，再将各数据点重新分配到距离最近的聚类当中去。重复重新分配的过程，直到在分配时不再有样本点从一类归入另一类中为止。

K-聚类算法的一个问题是，必须预先指定聚类数 k。这里不妨把它设为一个与训练集规模相称的较小数目，因为在随后的学习过程中，RBF 神经网络可以自动增加所需要的隐节点。

(2)隐单元简化机制

尽管我们已通过统计聚类方法对网络进行初始化，但由于聚类方法不可能将所有噪声数据剔除出去，因此如果任由 RBF 神经网络增加隐层单元的数量，就有可能产生过拟合的问题。对每一个新进训练样本，传统 RBF 神经网络只依据网络输出误差和最近隐单元距离来判断是否为新样本。对于数据量较大的样本而言，总不可避免地会有“漏网之鱼”——噪声数据进入网络，当一个被噪声污染较严重的样本进入网络，就很可能被选为新隐节点的中心。因为这个新节点与最近现有隐单元的距离也许大于规定的阈值，所以网络为了拟合噪声，而错误地引入了新的隐节点。

为此我们在算法第二步中采用隐单元简化机制。它的中心思想是通过研究某个隐单元对于输出层的贡献，判断其是否应该被删除。假定某个 RBF 神经网络的隐层神经元个数为 N_h、每个隐层神经元基函数的中心为 c_i、宽度为 σ_i，隐层到输出层的权值为 w(向量)，如果在连续 n_p 个训练样本期间，标准化输出向量 r_k^i 的所有分量都小于一个预先选定的阈值，就说明隐单元 k 对于整个网络输出的贡献已经十分不显著，应该将其从网络中删除。当然，这其中有一个阈值确定的问题。

(3)SCRBF 算法步骤

基于以上讨论，整个 SCRBF 孤立点检测算法可以用下列步骤来描述：

输入：原始数据集的每一个输入样本(x_i,y_i)，初始聚类数目 k，距离阈值 δ，训练误差 θ；

输出：孤立点集。

①初始选定聚类中心 μ_j；

②计算各子类与 μ_j 的距离，并与聚类距离阈值 σ 相比较；

③if$\mu_j>\delta$，则 $\mu_j=\mu_j+1$；

④重复②、③，直至聚类结束，得初始网络隐单元集 $S_i=S_i+\{\mu_j\}$；

⑤计算所有隐单元的输出向量 O_k^i，$(k=1,\cdots,N_h)$；

⑥为每一个隐单元计算归一化输出向量 r_k^i；

⑦计算网络输出 $\hat{y}(x_i)=[\hat{y}_1(x^i),\cdots,\hat{y}_l(x^i),\cdots,\hat{y}_n(x^i)]^T$；

⑧选择能获得最大误差下降比的 x_k 当作 RBF 新隐单元的中心点，并计算网络输出加权值及训练误差 θ_k；

⑨if$\theta_k>\theta$，则重复⑤、⑥、⑦步，直至满足要求；

⑩删除所有归一化输出向量 r_k^i 的每个分量都不符合误差下降比的隐单元。

至此，网络训练完成。将其作用于实际数据集，即可得出孤立点集。注：RBF 函数的宽度值可选用常用的高斯函数求得：$\delta=\sqrt{\frac{d_{\max}^2}{j+1}}$。其中 $d_{\max}$ 表示训练样本中输入向量间的最大距离。

3.2 实例应用与结果分析

为了与 LSC 算法进行对比，本文仍然采用文[1]中的入侵检测实验数据集作为测试数据。此数据集是一个常用的孤立点检测测试数据集。它源于美国空军局域网的仿真环境，每个实例包含 41 个属性，均已进行分类标识。实验环境：CPU P4 2.0，内存 512 M，系统为 Win2k SP4，用 MATLAB 语言编程实现 SCRBF 算法。考虑到实验用机的数据处理能力，本次实验从数据集中选取数据 7844 条。为了具有可比性，本文将进行孤立点检测率的比较，即检测出的孤立点占样本数据量的比例。样本数据的选取尽可能多地包含所有类

(下转第 271 页)

应的参考值具有良好的时效性。

3.3 Web应用的优化和演化

在上述实践经验的有效指导下，我们能够对Web应用进行系统优化[12]，并有利于维护和演化等方面工作的开展，具体体现在以下几个方面：

首先是通过常见错误分类来方便故障的查找和排除，为测试人员提供非常有价值的经验和参考，有效提高测试人员的业务水平，进而改善测试工作的效率；

其次是在设计模式的指导下，Web应用的逻辑结构更趋于合理、内容组织更加有层次且专业规范，能够更好地适应用户的需求，并且对整个系统性能的改善也起到积极的作用；

再次，通过计算Web应用度量指标的实际取值以及对参考值的及时更新，不仅能够明确该Web应用的质量水平和测试效率，还可以为其它Web应用的度量发挥参照作用。

小结 如何有效地处理Web应用测试结果，目前所受到的关注度还不是太高。但作为测试流程中的最后一步，Web应用测试结果的分析比较还是必不可少和非常重要的。因为只有通过测试结果评判后，才能够决定测试是否可以终止；另外，通过对测试结果的统计分析，还可以得到一些故障排查的线索、设计测试中的经验等，从而有利用于系统的不断进化。

本文主要研究如何对测试结果进行分析比较、如何利用测试结果反馈作用于Web应用以取得良好的发展演化效果。主要利用语义标注和XML描述技术实现Web页面中数据与显示信息的分离，从而方便测试结果与预期结果的比较；将测试结果反馈给Web应用本身，以便于故障修复并提取一些操作经验，然后应用到整个系统的优化工作中去，以指导其发展和演化。

参考文献

1 Cai K Y. Optimal software testing and adaptive software testing in the context of software cybernetics. Information and Software Technology, 2002, 44: 841～855

2 Handschuh S, Volz R, Staab S. Annotation for the deep Web. IEEE Intelligent Systems, 2003, 18(5): 42～48

3 Kallepalli C, Tian J. Measuring and Modeling Usage and Reliability for Statistical Web Testing. IEEE Trans Software Engineering, 2001, 27(11): 1023～1036

4 Liu C H. A Formal Object-Oriented Test Model for Testing Web Applications. 2002, Doctor Dissertation

5 Ricca F, Tonella P. Web Testing: a Roadmap for the Empirical Research. In: Proc. of the 2005 7th IEEE Int. Symposium on Web Site Evolution (WSE'05), 2005. 63～70

6 Soo Von-Wun, Lee Chen-Yu, Li Chung-Cheng, et al. Automated Semantic Annotation and Retrieval Based on Sharable Ontology and Case-based Learning Techniques. In: Proc. of the 2003 Joint Conference on Digital Libraries (JCDL'03), 2003. 61～72

7 Warren P, Gaskell C, Boldyreff C. Preparing the ground for website metrics research. In: Proc. of the 3rd International Workshop on Web Site Evolution, 2001. 78～85

8 Xu Lei, Xu Baowen. Research on the Analysis and Measurement for Testing Results of Web Applications. Second International Workshop on Web Computing in Cyberworlds (WCCW2005), 2005. 559～565

9 Xu Lei, Xu Baowen, Nie Changhai, Huowang Chen, Hongji Yang. A Browser Compatibility Testing Method Based on Combinatorial Testing. Proc. of the Int. Conference on Web Engineering (ICWE), 2003. 310～313

10 许蕾，徐宝文. 用户行为获取方法在Web性能测试中的应用研究. 软件学报，2003，14（增刊）：115～120

11 许蕾，徐宝文，陈振强. Web测试综述. 计算机科学，2003，30(3)：100～104

12 阳小华，周龙镶. 基于用户访问模式的WWW浏览路径优化. 软件学报，2001，12(6)：846～850

13 张卫丰，徐宝文. 带反馈自适应Web搜索引擎研究. 计算机科学，2004，31(9)：3～8

（上接第197页）

别，并且保证每种类别有一定数量的实例个数，以满足RBF神经网络的训练需求。为了具有普适性，在对数据集进行分析的基础上，分别选定两组阈值参数进行试验，由于篇幅所限，此处只给出试验结果比较情况，如表1所示。

表1 SCRBF与LSC算法试验结果比较

算法	阈值参数1	孤立点检测率	阈值参数2	孤立点检测率
SCRBF	k=100,δ=0.10,θ=0.03	72.3%	k=200,δ=0.15,θ=0.05	45.5%
LSC	误差=0.03	68.4%	误差=0.05	43.9%

两组不同阈值参数的实验结果都表明，在误差率保持相等的情况下，SCEBF算法的孤立点检测率都要高于LSC算法。而且，内在的隐单元简化机制使同样的问题SCRBF所需的网络结构相对较小，具有较低的网络复杂性。但本实验比较也存在不足之处，一是受限于实验用机的性能，所选取的训练集较小，SCRBF在大训练集中的应用效果还不能完全确定；二是在本实验中，由于数据集限制，未出现噪声数据进入网络的情况，使SCRBF算法的一个重要的优越性未能体现出来。但这也从另一个侧面验证了采用统计聚类方法对神经网络进行初始化，可以有效地降低引进噪声样本点为隐单元的可能性。

结束语 本文提出的SCRBF孤立点检测算法创造性地将统计聚类方法引入孤立点检测中，既克服了LSC算法的泛化能力弱的缺陷，又通过合理的隐单元简化机制防止了传统RBF神经网络容易过拟合的缺点。对比实验表明，SCRBF算法与LSC算法相比，在检测孤立点问题上具有很大的优越性。SCRBF用于孤立点检测还有一个优点就是，可以在完成其他数据检测任务（如分类和聚类）的同时，完成孤立点检测的任务。由于结合了统计方法，使人工神经网络易产生的随机性和过拟合作用造成的不利影响减小，从而提高了孤立点检测的可靠性。

参考文献

1 Agyemang M. Local Sparsity Coefficient- Based Mining of Outliers. Windsor, Ontario, Canada: University of Windsor, 2002

2 徐秉铮，张百灵，韦岗. 神经网络理论与应用. 广州：华南理工大学出版社，1994

3 蒋建春，马恒太. 网络安全入侵检测：研究综述. 软件学报，2000，11(11)：1460～1466

4 李俭川，秦国军，温熙森. 神经网络学习算法的过拟合问题及解决方法. 振动、测试及诊断，2002(4)：260～265

5 A xelsson S. The base-rate fallacy and its implications for the difficulty of intrusion detection. In: Tsudik, G, ed. Proceedings of the 6th Conference on Computer and Communication Security. New York: ACM Press, 1999. 1～7