一种有效的可视化孤立点发现与预测新途径*)

汪加才　张金城　江效尧
(南京审计学院计算机科学与技术系　南京 210029)

摘　要　孤立点发现是数据挖掘活动的重要组成部分,被广泛应用于电子贸易、信用卡等领域的欺诈检测。由于优良的拓扑结构保持和概率分布保持特性,SOM (Self-Organizing Maps)可作为一种有效的降维工具供分析人员获取隐藏于数据中的分布结构信息。在分析了当前基于距离的孤立点发现的基础上,提出了一种基于 SOM 的孤立点发现与预测新途径,具有可扩展性、可预测性、交互性、简明性等特征。实验结果表明,基于 SOM 的孤立点发现与预测是有效的。

关键词　孤立点发现,孤立点预测,SOM,交互式数据挖掘

An Effective and Efficient Approach to Detect and Predict Outliers Visually

WANG Jia-Cai　ZHANG Jin-Cheng　JIANG Xiao-Yao
(Department of Computer Science and Technology, Nanjing Audit University, Nanjing 210029)

Abstract　Outlier detection is an integral part of data mining and is critical important to some areas such as monitoring of criminal activities in electronic commerce, credit card fraud, etc. Due to the topological structure and probabilistic distribution preserving nature, SOM (Self-Organizing Maps) has been used as a tool for mapping high-dimensional data into a two dimensional feature map and gaining some idea of the structure of the data by observing the map. Based on the analysis of the existing distance-based outlier detection algorithms, a SOM based approach to detect and predict outliers is proposed, which has an obvious superiority in scalability, predictability, interactiveness, conciseness. Experimental results on real database show that the SOM based outlier detection and prediction is effective.

Keywords　Outlier detection, Outlier prediction, SOM, Interactive data mining

1　引言

孤立点发现 (Outlier Detection,或称之为离群点发现)用来发现数据集 D 中小部分对象,这些对象与数据中的一般行为或数据模型有着明显的不同[1]。其研究成果可广泛地应用到诸多领域中,包括信用卡诈骗检测、网络入侵检测[2]、电子贸易、医药研究、数据清洗、计算机辅助审计等。

早期的孤立点发现研究多见于统计领域。基于统计的方法一般只适用于单变量的数据集,虽然某些算法也可以检测多变量数据,但需要事先指定(假定)数据服从的分布模型,这两个缺点极大地限制了它的应用。

近年来,研究人员又提出了各种各样的方法[3]。其中,数据库界所提出的用于孤立点发现的方法大致有三种:最直接的方法是利用数据聚类分析,把聚类后不属于任何聚簇(Cluster)的数据对象作为孤立点[4];另一种方法是基于距离来定义孤立点对象[5~7],一般根据数据对象的最近邻居来判断其是否为孤立点。本方法的优点在于不需事先知道数据的分布模型,因而可以应用于任何可以用某种距离机制量度的特征空间。其缺点是定义孤立点对象的参数往往是全局性的;针对基于距离的方法存在的问题,后来提出了基于密度的局部孤立点发现方法[8,9]。

孤立点发现的关键是从数据集 D 中按照既定方法找出孤立程度最大的前 n 个数据点(设 w^* 为第 n 大孤立程度值)。孤立点预测(Outlier Prediction)则是针对新到来的数据,如果相对于 D 它的孤立程度大于 w^*,则判定其为孤立点[10]。尽管可以用孤立点发现的方法进行孤立点预测,但由于需要全局搜索 D 而导致效率低下,导致这方面的研究成果较少。文[10]提出了一种 D 的压缩表示方案——孤立点发现解集(outlier detection solving set),并以此为学习模型来进行孤立点预测。实质上,此解集仅是 D 的子集而已,它是以牺牲发现或预测精确性为代价的,而且也不适合于 D 为高维度、分布稀疏的场合。

Kohonen SOM(Self-organizing Maps)[11]是一种广泛应用的聚类算法,它具有拓扑结构保持、概率分布保持、可视化等优良特性[12]。基于这些特性,本文提出了一种有效的基于 SOM 的孤立点发现和预测方法,即以 SOM 作为 D 的分布及邻接索引,使孤立点的搜索空间由全域转为局域。同时,分析人员可根据 SOM 的标记图和距离矩阵图[12]选取合适的孤立点发现方法,或在 SOM 上动态选取稀疏分布区域或远离聚类中心的数据对象进行深入分析,具有较好的交互特征。

本文第 2 节分析了基于距离的孤立点发现的相关工作,第 3 节介绍了 SOM 模型及其在数据可视化中的应用,第 4 节给出了基于 SOM 的孤立点发现实验结果,最后是结论。

2　基于距离的孤立点发现

基于距离的孤立点(distance-based outlier)取决于数据对象邻域的定义。即便是对给定的距离量度函数,对孤立点也有不同的定义。

*)江苏省高校自然科学基金资助项目(06KJD520093,04KJB520059)。汪加才　教授,博士,主要研究方向:数据挖掘,商业智能等。

定义 1[5] 在包含有 N 个数据对象的数据集 D 中，o 是孤立点，仅当 D 中至少有 pct 部分对象与 o 的距离大于 d。换句话说，如果 o 在 d 范围内有不多于 $k=N(1-\text{pct})$ 个邻居，则 o 是一个带参数 pct 和 d 的 DB(pct,d)孤立点。

DB(pct,d)孤立点发现方法不要求用户预先知道数据集服从哪种统计分布模型。实际上，对于恰当定义的 pct 和 d，一个可以被给定的不一致检验测出的孤立点同样可以利用 DB(pct,d)检测出；同时，它克服了基于统计的检测仅能检测单个属性的缺点。

文[5]给出的基于块嵌套循环(block nested-loop)的 DB(pct,d)算法比较容易理解，其计算复杂度是 $O(\delta N^2)$，当维度 δ 增加时，复杂度的增加是线性的。基于单元格(cell-based)的 DB(pct,d)算法将 δ 维空间划分为边长为 $d/(2\sqrt{\delta})$ 的单元格，并以单元格为单位进行检测。其计算复杂度是 $O(m(2\sqrt{\delta}+1)^{\delta}+N)$，其中 m 是单元个数，因此该算法仅适合于大数据集、低维度的场合。

定义 2[6] D_k 孤立点是数据集中那些到其第 k 个最近邻居的距离最大的 n 个对象。

若 $D^k(p)$ 表示数据点 p 与其第 k 个最近邻居的距离，则处于分布稀疏区域的数据点将具有较大的 D^k 值，而属于聚类中的类内数据点(intra-cluster point)将具有较低的 D^k 值。D_k 孤立点发现方法基于各数据点的 D^k 排列，克服了 DB(pct,d)法缺少孤立程度信息的不足。同时，D_k 法无须用户指定距离参数 d。

文[6]给出了一种基于划分的发现算法。首先利用聚类算法划分数据集；然后计算各划分(Patition)P 的 D^k 边界(P.lower,P.upper)，使 P 中的每个点 p，满足 P.lower$\leqslant D^k(p)\leqslant P$.upper，并利用此信息确定 P 中是否可能包含孤立点；最后仅在可能包含孤立点的划分中计算和寻找孤立点。由于所要寻找的孤立点数目 n 相对较少，该方法可通过排除包含大量数据点的划分而降低计算量。实验显示，该方法关于 N 和 $\delta(\leqslant 10)$ 的可扩展性均较好。但是，由于 $D^k(p)$ 并没有包含 p 点所有 k 个最近邻的全部信息，因而它并不能很好地反映其邻域的紧密或稀疏状况。

定义 3[7] w_k 孤立点是数据集中那些与其 k 个最近邻距离之和最大的 n 个对象。

对于数据点 p，p 与其 k 个最近邻距离的和称为 p 的权，记为 $w_k(p)$。显然，$w_k(p)$ 比 $D^k(p)$ 更精确地度量了 p 的邻域稀疏程度。

为使计算各点 w_k 值的过程具有可扩展性，文[7]提出了 HilOut 孤立点发现算法。该算法采用了先求近似解，然后再从中获取精确解的策略。为避免直接求解每对点之间距离，HilOut 算法利用 Hilbert 空间填充曲线(Hilbert space filling curve)将数据集线性化，并基于此线性化的数据集上的前驱关系和后继关系，可快速地找出各点的 k 个近似最近邻。

上述发现方法均基于各数据点本身的邻域来判别其是否是孤立点，其检测标准是全局的、绝对的。而基于 LOF(Local Outlier Factor)[8]的孤立点发现方法则通过考察数据点 p 与其 k-邻域中其它诸点 o(p 与 o 的距离小于等于 $D^k(p)$)的差异来反映其孤立程度，其检测标准是局部的、相对的。

定义 4[8] 数据点 p 相对于数据点 o 的 k-可达距离(k-reach- ability distance，记为 k-rd)为 p 与 o 的直接距离以及 $D^k(o)$ 中的较大者。p 的局部可达密度(local reachability density，记为 k-lrd)为 p 相对于其 k-邻域中各最近邻居 o 的 k-rd 平均值之倒数。p 的局部孤立因子 $\text{LOF}_k(p)$ 被定义为 p 各最近邻居 o 的 k-lrd 平均值与 p 本身的 k-lrd 之比。

显然，数据点的局部可达密度越低，或其邻域的平均局部可达密度越高，则该点的局部孤立因子就越大，其孤立程度就越强。

3 SOM 的可视化

3.1 SOM 及其特征

SOM 由输入层和竞争输出层组成[11]。将 M 个输出神经元排列成规则的二维阵列 A。设输入向量为 δ 维，对应着 δ 个输入神经元。每个输出神经元均有一权向量与 δ 个输入神经元相连接。对于给定输入向量，训练过程不仅要调节获胜单元(与输入向量距离最近的输出神经元，也称为最佳匹配单元)的权向量，而且还要调节获胜单元相邻接的输出单元的权向量。

因此，训练收敛的 SOM 可以将高维的输入空间 D 按向量间的相似程度映射到低维的输出空间 A 中，并具有以下两个特性：(1)拓扑结构保持：将近似的输入向量映射到 A 的相近神经元，或反之，A 中相近的神经元对应近似的输入向量；(2)概率分布保持：D 中输入向量分布密度高的区域在 A 中也存在较多的相应神经元。

3.2 基于 SOM 的数据可视化

数据可视化的基本目的是通过有效的图形特征(位置、形状、颜色等)使人们容易从图形所呈现的大量细节信息中获取数据特征的定性概念。根据不同的分析目标需求，SOM 可以用不同的可视化形式表示，大致可分为 SOM 标记图和距离矩阵图两类[12]。

3.2.1 SOM 标记图

设 SOM 的输出层为规则的二维阵列，则可为每个输出单元做上有意义的标记(如输出标记、属性标记、命中标记等)，产生 SOM 标记图。

(1)输出标记图。SOM 理论认为，每个输出单元可看作为一组相似事例的聚类中心。采用最近邻分类思想，输出单元的输出标记可以用与其最近邻的 k 个事例的输出表决结果来表示。

(2)命中标记图。对于每个事例，可在 A 中确定对应的获胜单元。因此，SOM 是对事例集 D 按事例间的相似程度所进行的一个完全划分，输出单元所对应的事例子集称为 Voronoi 集(简称为 V 集)。将每个输出单元获胜次数标注出来，可直观地观察输入空间事例的分布情况。

(3)属性标记图。类似地，可用各 V 集中的事例来评价 δ 个属性中与聚类中心对应权分量的接近程度。最接近的一个或前几个属性可以作为该输出单元的属性特征。

3.2.2 距离矩阵图

距离矩阵记录了 SOM 中相邻输出单元间的距离信息，是一种广为采用的从 SOM 中探测数据类聚的可视化技术。由于 SOM 的拓扑结构保持特征，相邻输出单元间的距离能反映输入空间中两组相邻的相似事例间的位置关系。在距离矩阵中，每个输出单元的距离值为其至直接相邻单元距离的平均值；而文[13]所定义的 U-距离矩阵(Unified Distance Matrix)则既有输出单元至直接相邻单元间的平均距离，又含有它与各个直接相邻单元间的个体距离。将不同距离值以不同的颜色、灰度、形状表示出来，可直观地观察数据在 SOM 上的分布和聚集情况。

3.3 SOM 图例

图 1 是 Iris 数据集的 3 个 SOM 图形(SOM 输出层为 6 * 6 的正方形结构)。Iris 数据集包括 3 类事例(1-Setosa, 2-Versicolor, 3-Virginica,每类各含 50 个事例),每个事例均由 4 个属性(A1-SepalL, A2-SepalW, A3-PetalL, A4-PetalW)所描述。图 1(a)是 150 个事例在 SOM 中的分布情况;图 1(b)是 SOM 各 V 集的距离偏离,其值为 V 集各事例与对应权向量距离的平均(为具有可比性,将最大平均距离值设为 1),反映了 V 集内部分布的稀疏程度;图 1(c)是 SOM 的 U-距离矩阵图,图块颜色的深浅体现了各输出单元与相邻单元距离的大小,反映了神经元间的接近程度。由图 1(c)可看出,数据集分布大致分为三部分,即上部(前两行)、左下部、右下部;同时,由图 1(a)和图 1(b)可明显看出,上部 10 个输出单元所覆盖的 50 个事例分布比较均匀,下部 4 行的 11 个边缘单元内的事例分布较为稀疏。

(a)数据点分布

1	13	8	11	6	6
0	0	1	2	1	1
7	1	1	0	0	0
8	8	8	1	1	3
7	4	1	6	3	2
5	10	3	9	3	9

(b)距离偏离

0	0.2	0.1	0.1	0.1	0.3
0	0	0	0.1	0	0
0.4	0	0	0	0	0
0.2	0.2	0.4	0	0	0.3
0.8	0.3	0	0.2	0.09	0.7
0.6	0.3	0.6	0.3	0.3	1

(c) U-matrix

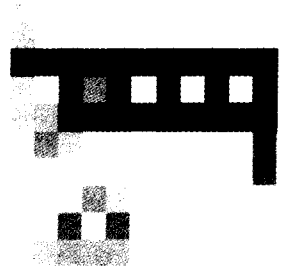

图 1 Iris 数据集的 SOM 图(A=6 * 6)

4 基于 SOM 的孤立点发现

4.1 基于 SOM 的孤立点发现与预测流程

前文基于距离的孤立点发现方法,均具有定义简明的优点,而在可扩展性、交互性和可预测性等方面存在一定的不足。其根源在于邻域定义的主观性(如所指定的参数 k)和邻域对象发现的低效率(全域搜索)。为使算法具有可扩展性,人们提出了基于单元格计数[5]、网格单元计数[9]、划分删减[6]、Hilbert 空间填充曲线映射[7]等近似求解方案,但大大增加了算法设计与实现的复杂性,损害了结果的准确性和算法实现的简明性。同时,无论是邻域半径参数,还是最近邻个数参数,应随求解领域的不同而不同。遗憾的是,利用前文所列方法,并不能在运行算法前得到这些关键参数选取的线索。

实际上,与其他数据挖掘活动一样,孤立点发现仅是数据分析活动中的一个方面。其方法的准确利用及结果的有效获取建立于用户对数据分布特征的正确认识。我们提出的基于 SOM 的孤立点发现或预测新途径可按如下步骤进行:

步骤 1 规范化数据集 D,并随机生成指定大小或指定比率的抽样集 X;

步骤 2 以 X 为训练集生成 SOM 模型;

步骤 3 以 D 为数据源,标注 SOM 模型,获取 SOM 标记图和距离矩阵图;

步骤 4 以输出单元的 V 集和其直接邻接单元的 V 集作为候选最近邻集,按前文的孤立点定义进行分析与判断;

步骤 5 对于任一待预测数据,首先确定其最佳匹配单元,然后,以该单元的 V 集和其直接邻接单元的 V 集作为候选最近邻集,按上述定义进行分析与判断;

步骤 6 在 SOM 上标出所发现的孤立点,用户可结合 SOM 图逐个分析各数据点。

显然,上述基于 SOM 的孤立点发现途径可带来如下好处:

(1)可扩展性。以 SOM 模型快速确定孤立点候选集和候选最近邻集,大大缩小了搜索空间。

(2)可预测性。通过在 SOM 中寻找最佳匹配单元,以及利用各种 SOM 标记图和距离矩阵图,可快速确定一新的数据点是否为孤立点及其孤立程度,实现预测功能。

(3)交互性。SOM 为我们提供了展示隐藏于数据集中总体分布信息的途径,而各数据点的孤立指标则从微观层面上表示了其偏离正常分布或局部分布的程度。基于 SOM 的数据分析,用户可动态地选取感兴趣数据区域或数据点进行深入分析。

(4)适应性。可通过各种 SOM 标记图和距离矩阵图探测数据实际分布情况,并选取合适的邻域半径和最近邻数目。同时,由于搜索效率的提高,可以使运行参数的最优化成为可能。

(5)简明性。由于搜索空间的缩小,各种孤立点发现算法可以基于其定义直接实现,而无须再采用复杂的近似替代方法。

4.2 有效性实验

取 $k=5$,找出 Iris 数据集的前 15 个(占数据集的 10%) D^k 孤立点、w_k 孤立点、LOF 孤立点,并在各 V 集中确定这些孤立点所在单元。其分布情况如图 2(a)-(c)所示。分析图 1 和图 2 可见,D^k 孤立点和 w_k 孤立点基本来自距离偏离较大的单元(当 V 集点数为 1 时距离偏离定义为 0,当然也可以定义为最大值 1);而 LOF 孤立点则来自本身距离偏离较大或与邻接单元距离较大(通过 U-距离矩阵图)的单元。

设 o_i 为采用某孤立点发现方法 $F(F\in\{D^k, w_k, LOF_k\})$ 在 D 中得到的第 i 个孤立点,其孤立值为 $F_k^*(o_i)$。以 o_i 所在 SOM 单元及其直接邻接单元的 V 集作为寻找其 k-最近邻居的局部数据子集,其孤立值为 $F_k(\mathrm{SOM}, o_i)$。$F_k^*(o_i)$ 与 $F_k(\mathrm{SOM}, o_i)$ 的差值反映了利用 SOM 发现孤立点的精确程度。图 3 是在 $k=1-10$,SOM 网络结构分别为 5 * 5, 6 * 6, 7 * 7, 8 * 8 下前 15 个孤立点的平均误差(为具有可比性,将 max $(F_k^*(o_i), F_k(\mathrm{SOM}, o_i)$取为 100)。

由图3可见，利用SOM输出单元的邻接关系，D^k、w_k 均可以在较小的搜索空间中得到与全域搜索相近的结果(平均误差均小于2%)。图3(d)每个SOM V 集的平均容量仅为2.34个实例，当 k 超过6时，基于SOM的 LOF_k 方法的平均误差将快速增加。事实上，此时 o_i 的局部数据子集可能只包括 o_i 的 k-最近邻居的部分最近邻居。因此，当SOM V 集的平均容量大大小于 k 时，需要考虑间接邻接单元的 V 集以扩充局部数据子集，进而提高 LOF_k 方法的孤立点发现正确率。

(a)Dk 孤立点分布

1	0	0	0	0	0
0	0	0	0	0	1
1	0	0	0	0	0
0	0	0	0	0	0
4	0	0	0	0	1
1	0	1	0	1	4

(b)Wk 孤立点分布

1	0	0	0	0	0
0	0	0	0	0	1
2	0	0	0	0	0
0	0	0	0	0	1
3	0	0	0	0	1
0	0	1	0	1	4

(c)LOF 孤立点分布

1	1	0	1	0	0
0	0	1	2	0	1
2	0	0	0	0	0
1	0	0	0	0	0
1	0	0	0	0	1
0	0	0	0	1	2

图2 Iris数据集孤立点在SOM图中的分布图($k=5$，$n=10$)

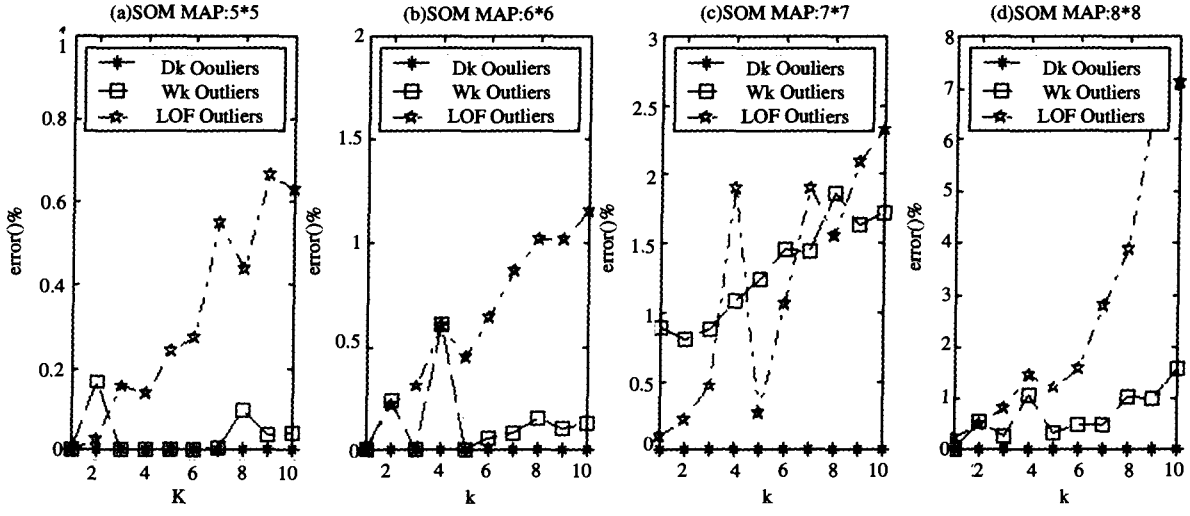

图3 基于SOM的Iris数据集孤立点发现

结论 已有的基于距离的孤立点发现方法都需要进行全空间的邻域查找操作。为使算法具有可扩展性，往往需采用近似求解方案，但大大增加了算法设计与实现的复杂性，损害了结果的准确性。基于SOM的拓扑结构保持、概率分布保持、可视化等优良特性，可使用户在数据分析的基础上，有针对性地选取感兴趣区域进行深入分析，具有交互性的特点。同时，由于可在SOM的局部邻域内寻找 k-最近邻居，可以根据孤立点定义进行算法的设计与实现，使其具有可扩展性、简明性等特征。

参考文献

1 Han J, Kamber M. Data Mining, Concepts and Technique. San Francisco: Morgan Kaufmann, 2001

2 Eskin E. A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data. Applications of Data Mining in Computer Security, Kluwer, 2002

3 Jin W, Tung A K H, Han J. Mining Top-n Local Outliers in Large Databases. In: Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '01), 2001

4 Yu D, Sheikholeslami G, Zhang A. Findout: finding out outliers in large datasets. Knowledge and Information Systems, 2002, 4(4): 387～412

5 Knorr E, Ng R. Algorithms for Mining Distance-Based Outliers in Large Datasets. In: Proc. Int'l Conf. Very Large Databases (VLDB '98), 1998. 392～403

6 Ramaswamy S, Rastogi R, Shim K. Efficient Algorithms for Mining Outliers from Large Data Sets. In: Proc. Int'l Conf. Management of Data (SIGMOD '00), 2000. 427～438

7 Angiulli F, Pizzuti C. Outlier Mining in Large High-Dimensional Data Sets. IEEE Trans. Knowledge and Data Eng., 2005, 2(17): 203～215

8 Breunig M M, Kriegel H, et al. LOF: Identifying Density-Based Local Outliers. In: Proc. Int'l Conf. Managment of Data (SIGMOD '00), 2000

9 Papadimitriou S, Kitagawa, et al. LOCI: fast outlier detection using the local correlation integral. In: Proc. 19th International Conference on Data Engineering, 2003. 315～326

10 Angiulli F, Basta S, Pizzuti C. Distance-based detection and prediction of Outlier. IEEE Trans. Knowledge and Data Eng., 2006, 2(18): 145～160

11 Kohonen T. Self-organizing maps. Springer-Verlag, Berlin, 1997

12 汪加才，等. VISMiner：一个交互式可视化数据挖掘原型系统. 计算机工程，2003(1)：17～19

13 Ultsch A, Siemon H P. Kohonen's Self-organizing Feature Maps for Exploratory Data Analysis. In: Proc. INNC'90, International Neural Network Conference, Dordrecht, Netherlands, 1990. 305～308

14 Aggarwal C, Yu P. Outlier detection for high dimensional data. In: Proceedings of the SIGMOD, 2001. 37～46