一种基于密度偏差抽样的孤立点检测算法*)

余建桥　葛继科　李　娅
(西南农业大学信息学院　重庆400716)

摘　要　孤立点检测是一项有价值的、重要的知识发现任务。在对大规模数据集中的孤立点数据进行检测时，样本数据集的选择技术至关重要。本文提出了一种新的基于密度的偏差抽样技术作为数据约简的手段，并给出了基于密度偏差抽样的孤立点检测算法，该算法可以用来识别样本数据集低密度区域中的孤立点数据，并从理论和实验两个方面对其进行分析评估，分析与实践证明该算法是有效的。

关键词　大规模数据集，偏差抽样，孤立点检测

Outlier Detection Algorithms Based on Density Biased Sampling

YU Jian-Qiao　GE Ji-Ke　LI Ya
(Information College, Southwest Agricultural University, Chongqing 400716)

Abstract　Outlier detection is a meaningful and important knowledge discovery task. The choice of sampling data set is very important during the process of outlier detection in large data sets. We propose a new density biased sampling as a data reduction technique to speed up the operation of outlier detection in large data sets, and introduce an algorithm based on density biased sampling. The algorithm can identify outliers of the sparse region. Finally, by evaluating the proposed method and presenting a experimental evaluation, we verify the utility of our approach.

Keywords　Large data set, Biased sampling, Outlier detection

1　引言

在数据集中，常常存在一些数据对象，它们不符合数据的一般模型，即与数据的其他部分不同或不一致，这样的数据对象被称为孤立点(outlier)[1]。

在大规模数据集中进行孤立点检测是数据挖掘中重要的数据分析任务。由于大规模数据集具有海量的记录和很多属性，所以在进行数据挖掘时一般需要多遍扫描、检测所要分析的数据集。如何减少系统开销，提高数据挖掘算法的效率，成为当前研究的热点。

在进行孤立点检测时，现有算法大多数基于简单随机抽样，每个点都以相同的概率被包含在样本中，当样本较小时，存在样本准确性不高的问题，为保证样本的准确性，必需使用大样本，这不但降低了抽样效率，而且增加了后续工作的复杂度。因此，应用简单随机抽样对大规模数据集进行抽样时，需要消耗大量的系统开销。偏差抽样是数据挖掘中产生的一种新的抽样策略[2]，其核心思想是依据数据分布的密度情况来生成样本，与简单随机抽样相比，数据点被添加到样本中的概率逐点不同。偏差抽样中的抽样概率取决于预先指定的数据特性和具体的分析目标。一旦得到了涉及数据特性和分析目标的抽样概率分布，就能快速准确地从数据集中抽取出偏差样本。因此，在大规模数据集中应用偏差抽样技术进行抽样，可以有效克服简单随机抽样技术中存在的准确性不高和效率低下的问题。

2　样本数据集的获取

在文[2]中，Palmer C R 等人提出的基于密度的偏差抽样技术(density biased sampling)，可以有效地实现数据的约简。改进该抽样技术，可以使得到的新的密度偏差抽样技术具有更大的灵活性。

2.1　改进的偏差抽样技术

设空间域为$[0,1]^d$，D是一个有n个数据点的d维数据集，设$f(x_1,\cdots,x_d)$是数据集D的一个密度估算函数。即，对于一个给定的区域$R\subset[0,1]^d$，积分$n\int_R f(x_1,\cdots,x_d)\mathrm{d}x_1\cdots\mathrm{d}x_d$近似等于$R$中的点的集合。围绕点$x$的本地密度值为$LD(x)=1/V$，其中$I=\int_B f(x)\mathrm{d}x$，$B$是环绕在点$x$周围的半径为$\varepsilon(\varepsilon>0)$的球，$V$是球的体积，$V=\int_B \mathrm{d}x$。在此，假设描述数据集密度的函数曲线是连续的和光滑的(它是一个多重复合函数)。在一个点周围的一个球的本地密度近似代表了这个球的密度函数的平均值，因此，对于一个很小的ε，对应的期望偏差是很小的。

密度偏差抽样技术的目标是找到数据集D的一个具有如下性质的偏差样本[2]。

性质1　D中的任一数据点x被包含在样本中的概率是数据点x周围的数据空间的一个本地密度函数。

性质2　期望的样本长度应当是b，b是由用户自定义的一个参数。

在密度偏差抽样中，要想得到一个给定点$(x_1,\cdots,x_d)$被包含在样本中的概率，其实就是获得环绕这个点的一个本地密度函数，这由$f(x_1,\cdots,x_d)$的给定值决定。为了使偏差抽样技术具有灵活性，引入一个改进的密度估算函数$f'(x_1,\cdots,x_d)$，使得

$$f'(x_1,\cdots,x_d)=(f(x_1,\cdots,x_d))^a \tag{1}$$

其中a为实数。因此，可以通过a来控制抽样的进程。

设

*)基金项目：重庆市教委资助项目(030201)。余建桥　教授，硕士生导师，研究方向为数据库技术、人工智能；葛继科、李　娅　硕士研究生，研究方向为数据挖掘。

$$k=\sum_{(x_1,\cdots,x_d)\in D} f^t(x_1,\cdots,x_d) \tag{2}$$

并使

$$f^*(x_1,\cdots,x_d)=\frac{n}{k}f^t(x_1,\cdots,x_d) \tag{3}$$

f^*的附加特性是$\sum_{(x_1,\cdots,x_d)\in D} f^*(x_1,\cdots,x_d)=n$。

由密度偏差抽样的特性可知，在抽样过程中，每个点$(x_1,\cdots,x_d)\in D$被包含在样本中的概率是：$\frac{b}{n}f^t(x_1,\cdots,x_d)$，显然满足性质1；因为样本期望长度$\sum_{(x_1,\cdots,x_d)\in D} f^*\frac{b}{n}(x_1,\cdots,x_d)=\frac{b}{n}\sum_{(x_1,\cdots,x_d)\in D} f^*(x_1,\cdots,x_d)=b$，所以也满足性质2。

从密度估算函数$f^t(x_1,\cdots,x_d)$可知，改变a的值，不但可以得到高密度区域($a>0$)的样本，也可以得到低密度区域($a<0$)的样本。对a取不同的值进行分析：

① $a=0$每一点都以$\frac{b}{n}$的概率被包含在样本中，这样就得到了简单随机抽样[3]。

② $a>0$设x和y分别是位于数据集中不同密度区域的点，如果$f^a(x)>f^a(y)$，那么$f(x)>f(y)$。表示在高密度区域的抽样率要比低密度区域的抽样率高。准确来说，在数据空间中，如果某区域的密度高于平均密度，那么，偏差抽样在该区域的抽样率比简单随机抽样的抽样率要高；同理，如果该区域的密度低于平均密度，那么，偏差抽样在该区域的抽样率比简单随机抽样的抽样率要低。

③ $a<0$同②，如果$f^a(x)>f^a(y)$，那么$f(x)<f(y)$。这种情况与$a>0$恰好相反，低密度区域的抽样比简单随机抽样高，而高密度区域的抽样比简单随机抽样低。

假设密度估算函数f有效，对数据集进行扫描，根据公式(1)、(2)，可以计算出k的值，根据得到的k值和公式(3)，可以计算出数据集中的每一点x的值$f^*(x)$。由此，通过一次扫描，就可以得到期望的样本数据集。设样本数据集的平均密度为ρ，$\rho=\frac{1}{n}\sum_{i=1}^{n} f^*(x_{i1},\cdots,x_{id})$，当$f^*(x)>m\rho$，认为点$x$属于高密度区域；当$f^*(x)<m\rho$时，认为点$x$属于低密度区域，其中临界值$m$可由用户设置，$m\in(0,1]$。

2.2 样本密度值的计算

为了高效执行一次扫描技术，需估计出每个区域R的积分，在通常情况下，不能直接分析估算出这些积分，但是可以用蒙特卡罗法在数量上估算出积分的近似值。

据前所述，$f(x)$是D上的一个密度估算函数，为了阐述简单，假设$f(x)$是定义在区间[0,1]上的一维空间函数，希望得到积分$I=\int f(x)\mathrm{d}x$，应用蒙特卡罗法计算密度估算函数的积分。借用简单随机抽样技术，随机得到在区间[0,1]上的一些点：$\xi_1,\cdots,\xi_N$，$f(\xi_1),\cdots,f(\xi_N)$，是对应于这些点的函数值。通过下面这个公式可以得到I的近似值S，$S=\frac{\sum_{i=1}^{N} f(\xi_i)}{N}$，$S$是$f(\xi_i)$的均值，其中$1\leqslant i\leqslant N$，设$\mu$等于$E(f(\xi_i))$，即$\mu$为$f(\xi_i)$的期望值，$\sigma^2$为$f(\xi_i)S$的方差，置信区间为$|D-\mu|$，置信度为$1-\varepsilon$，可以得到绝对误差$\delta=|S-\mu|=\sigma*\sqrt{(\frac{1}{\varepsilon*N})}$。在实际应用中，并不知道$\sigma$的值，可以用$D$的标准差来代替。

3 孤立点的检测

3.1 算法描述

利用密度偏差抽样技术对大规模数据集进行扫描后，可以得到两种类型的数据区域。一种是数据密度比较高的区域；另一种是数据密度比较低的区域，只对该区域进行考察来发现孤立点，为此，给孤立点做了如下定义：在样本数据集D中，对于某一对象o，如果最多有p个数据对象o到对象的距离小于k，那么，就称对象o是一个带参数p和k的孤立点，记做$Outs(p,k)$。

这个定义与Knorr和Ng在文[4]中提出的孤立点的定义是相容的，而且也可以推广到从统计学角度对孤立点的定义[5]。在定义一个孤立点$Outs(p,k)$时，p和k的值可以由用户自己设定，样本数据集D中孤立点的数量可以由k决定，也可以由样本数据集长度的一个分数fr来确定，其中$p=fr|D|$。

该孤立点检测算法的基本思想是：首先利用偏差抽样技术计算样本密度值，然后对数据密度较低的区域进行孤立点检测。假设在算法中考查的数据点是基于欧几里得距离的点。

为了简化算法的描述，定义一个描述数据点个数的函数$N_D(o,k)$，它是指样本数据集D中离数据点o的距离小于k的数据点的个数。显然，如果$N_D(o,k)<p$，那么，$o\in D$是$Oust(p,k)$的一个孤立点。

设D是一个有$|D|=n$个数据点的样本数据集，数据维数为d，设f是数据集D的密度估算函数，给定输入的参数k和p，对于每一个数据点o，可以计算出以o为圆心，以k为半径的球内的点的个数。因此，可以通过下面给出的公式(4)估算出$N_D(o,k)$的值，其中$N_D(oo,k)$近似等于$N_D(o,k)$，

$$N_D(o,k)=\int_{\text{为圆心}k\text{为半径的球}} f(x_1,\cdots,x_d)\mathrm{d}x_1\cdots\mathrm{d}x_d \tag{4}$$

利用公式(4)，对样本数据集进行第一次扫描就可以估算出可能的孤立点的数目，但是在此过程中保留了比数据点o具有更小期望值的邻近点，它们是相似孤立点，可以针对这些相似孤立点用相同的方法进行第二次扫描，从而确定它们是否为真正的孤立点。

3.2 时间复杂度分析

对样本数据集进行一次扫描时，因为每个点只能被读取一次，所以发现孤立点所用的时间与数据集的长度呈线性关系。为每一点$o\in D$计算出$N_D(o,k)$，通过核心密度估算理论[6,7]可知，用密度估算函数f对样本长度为b，维数为d的数据集D进行抽样所需的时间为$O(bd)$。如果相似孤立点均为真正的孤立点且存储空间足够大，那么发现孤立点的总的运行时间是$O(dmn)$，其中d为数据集的维数，m为孤立点的数目，n为数据点的总数。如果使用高效数据结构(如，CF树)来存储相似孤立点，总的运行时间可以减少到$O(d\log mn)$。

该算法对数据集总的扫描次数为$2+\lceil m/M\rceil$，其中m是相似孤立点的数目，M是存储空间的大小。在实际应用中，由于孤立点是数据集中的罕见事件，并且存储空间也可以容纳下所有的相似孤立点，因此，一般需要对数据集进行三次扫描才能够完成孤立点检测的任务。但与通过简单随机抽样获取样本数据集并进行孤立点检测相比较而言，这种方法在时间复杂度上要小得多。

结束语 在现实数据集中，包括地理空间数据集和二维数据集，均可能包含一个很大的簇和一系列孤立点，使用基于偏差抽样的孤立点检测算法进行数据处理，能快速检测到孤立点。我们以文[8]中的实验数据集作为实验对象，用本文提出的算法对数据集进行了实验。实验结果表明，该算法在效率上要优于文[8]中提出的基于距离的孤立点检测算法。

在大规模数据集上进行数据挖掘任务时，文中提出的密

度偏差抽样技术可以根据用户的要求快速地从一个大规模数据集中抽取出准确性很高的偏差样本，对孤立点检测算法的快速高效执行提供了充分的保证。

本文证明了密度偏差抽样技术可以应用于孤立点检测这项具体的数据挖掘任务，同时，探索密度偏差抽样技术在其他数据挖掘技术如分类、决策树、关联规则等方面的应用也具有积极的指导作用，这是我们下一步将要研究的课题。

参 考 文 献

1 Han J, Kamber M. Data Mining: Concepts and Techniques. Copyright by Morgan Kaufmann Publishers, Inc. 2001

2 Palmer C R, Faloutsos C. Density biased sampling: An improved method for data mining and clustering. In: Proc. Of the ACM SIGMOD'2000, 2000

3 Guha S, Rastogi R, Shim K. CRUE: An Efficient Clustering Algorithm for Large Database. In: Proc. ACM SIGMOD, June 1998. 73～84

4 Knorr E, Ng R. Algorithms for Mining Distance Based Outliers in Large Databases. In: Proc. Very Large Data Bases Conf., Aug. 1998. 392～403

5 Barnett Y, Lewis T. Outliers in Statistical Data. John Wiley & Sons, 1994

6 Scott D. Multivariate Density Estimation: Theory, Practice and Visualization. Wiley and Sons, 1992

7 Wand M P, Jones M C. Kernel Smoothing. Monographs on Statistics and Applied Probability, Chapman and Hall, 1995

8 Knorr E, Ng R. Algorithms for Mining Distance-based Outliers in Large Datasets. In: Proc. 1998 Int. Conf. Very Large Data Base(VLDB98), New York, 1998(8):392～403

（上接第199页）

较好的初始温度和下降系数为：$T_0=5$和$r=0.9$。从整体上看SA-MMI-L算法具有较好的学习效果。

表4 500个训练样本时的计算量

V	SA-BDE 算法		GA-MDL		SA-MMI-L	
	循环次数	时间	循环次数	时间	循环次数	时间
1	1	1	1	1	1	1
5	30	1	30	1	30	1
10	2483	1	1345	1	1124	1
15	41783	4	12753	1	9637	1
20	81308	8	61635	5	45684	4
25	162593	16	84138	7	58852	5
27	190949	19	121937	11	81351	7
29	220835	22	152018	15	121756	11
31	312638	31	220957	21	173564	16
33	371904	37	278769	26	214683	19
35	414473	41	318546	29	235429	22
37	568256	56	369354	34	307538	27

表5 1000个训练样本时的计算量

V	SA-BDE 算法		GA-MDL		SA-MMI-L	
	循环次数	时间	循环次数	时间	循环次数	时间
1	1	1	1	1	1	1
5	135	1	87	1	65	1
10	7592	1	5247	1	3867	1
15	73756	7	13793	1	11354	1
20	111536	9	51785	5	43659	4
25	164565	17	81178	7	77860	7
27	363745	35	109935	11	94681	9
29	561085	58	153426	15	138213	13
31	648362	64	269928	27	238156	22
33	719620	73	348667	35	338907	33
35	873247	87	418546	41	407679	40
37	1175585	109	574356	57	527640	51

从表3～5可以看出，当样本数目较小时，三种算法所用时间基本相同，当样本数量增大时，SA-MMI-L算法和GA-MDL算法速度较快，SA-MMI-L算法较GA-MDL算法稍好。

结论 近年来，基于相互信息熵理论应用于贝叶斯网络结构学习得到了广泛关注，附加网络复杂度约束的最大熵记分原则能够很好地用于贝叶斯网络结构的评估，最大相互信息保证了变量间信息量最大，附加的约束函数使得网络拓扑结构不过于复杂，是逼近程度和复杂度之间取折衷的最合适的评价方法；模拟退火算法是传统的智能优化方法，通过设计不同的优化函数、邻近值产生机制、温度下降策略以及算法终止和结束条件，可以将模拟退火算法应用于贝叶斯网络的结构学习，实验表明该算法在计算复杂度和学习精度上都具有较好效果。

参 考 文 献

1 Heckerman D. Bayesian Networks for Data Mining. Data Mining and Knowledge Discovery, 1997, 1:79～119

2 Cooper G F, Herskovits E. A Bayesian Method for the induction of probabilistic networks from data. Machine Learning, 1992, 9: 309～347

3 Suzuki J. A Construction of Bayesian Networks from Databases Based on an MDL Principle. In: Proc. of the 9th Conf. on Uncertainty in Artificial Intelligence. Washington D. C., 1993. 266～273

4 Lam W, Bacchus F. Using Causal Information and Local Measures to Learning Bayesian Networks. In: Proc. of the 9th Conf. on Uncertainty in Artificial Intelligence. Washington D. C., 1993. 243～250

5 Suzuki J. Learning Bayesian belief networks based on the MDL principle: an efficient algorithm using the branch and bound technique. In: Proc. of the Intl. Conf. on Machine Learning, Bally, Italy, 1996

6 Suzuki J. Learning Bayesian Belief Networks Based on the MDL Principle: An Efficient Algorithm Using the Branch and Bound Technique TIEICE. IEICE Transactions on Communications/Electronics/Information and Systems, 1998. 1～12

7 Suzuki J. Learning Bayesian Belief networks Based on the Minium Description Length Priciple: Basic Properties. IEEE Transactions on Fundamentals, 1999, E82-A(9)

8 Chickering D, Geiger D, Heckerman D. Learning bayesian networks: Search methods and experimental results. In: Proc. of the fifth conf. on Artificial Intelligence and Statistics, IEEE Press, 1995. 112～128

9 Larranaga P, et al. Structure Learning of Bayesian Networks by Genetic Algorithms: Aperformance analysis of Control parameters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18b(9): 912～926

10 Wong M L, Lam W, Leung K-S. Using Evolutionary Programming and Minimum Description Length Principle for Data Mining of Bayesian Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21(2): 174～178

11 李刚. 知识发现的图模型方法. 中国科学院软件研究所：[博士学位论文]. 2001