基于采样特异性因子的实时异常检测

牛之贤 孙静宇 石淑萍
(太原理工大学计算机科学与技术学院 太原 030024)

摘 要 面向特异性的数据挖掘中,特异性因子是一个重要概念,但其计算时间复杂度过高。使用基于采样的方法定义特异性因子即采样特异性因子(Sampled Peculiarity Factor,SPF)可在不影响精度的情况下,提高运行效率。为提高基于SPF算法的异常检测效率,提出了基于SPF的学习采样频率算法,将SPF和最优采样频率结合起来提出了实时异常检测算法。在真实数据集上进行了实验,置信度为95%时,得到的最优采样频率序列为[1/32,1/16]。仿真实时异常实验表明该算法的误检率为2%。

关键词 采样特异性因子,采样频率,实时,异常检测
中图法分类号 TP181 **文献标识码** A

Real-time Anomaly Detection Based on Sampled Peculiarity Factor

NIU Zhi-xian SUN Jing-yu SHI Shu-ping
(College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan 030024, China)

Abstract Peculiarity factor is an important concept in the peculiarity-oriented mining, but its computation is too complex. Using sampling-based method to define the peculiarity factors called sampled peculiarity factor (SPF) can improve operational efficiency without affecting the accuracy. To improve the efficiency of anomaly detection algorithm based on the SPF, learning sampling frequency algorithm was proposed. Combined the optimal sampling frequency and SPF, real-time anomaly detection algorithm was proposed. Experiments use real data sets, take confidence as 95%, and the optimal sampling frequency sequence is between 1/32 and 1/16. Simulation results show that false detection rate of the algorithm is 2%.

Keywords Sample peculiarity factor(SPF), Sampling frequency, Real-time, Anomaly detection

1 引言

数据挖掘最基本的任务就是发现隐藏在数据集中的知识;知识的表示形式是多样的,例如:关联规则、分类、聚类、簇、惯例模式、趋势等等[1]。隐藏在大数据集中的少量偏差较大的数据中往往蕴含着重要信息,称这类数据为特异性数据,挖掘这类数据中的特异性规则显得越来越重要。这些异常往往是硬件故障、人为因素、干扰、噪音等,在金融欺诈、网络入侵检测、医学诊断、天气预报等领域研究这些数据有着现实的意义和价值[2]。

特异性数据有两个基本的性质,一是它们代表的是相对小数量的对象,二是在数据集中它们与其他对象不同,即量少且明显异于其他数据。面向特异性的挖掘旨在发现小的子集中的有趣特异性规则,包括两个任务:识别特异性数据和分析特异性数据。特异性数据识别是用一个分值标记每个数据,分值高的就认为是特异性数据;对于特异性数据分析,使用现有的数据挖掘方法分析特异性规则,深入分析数据集,发现其中隐藏的更多知识,用以指导实践。

在文献[1-3]中钟宁教授提出了面向特异性数据挖掘(Peculiarity-Oriented Mining,POM)方法,并给出了特异性因子(Peculiarity Factor,PF)的概念进行异常检测,当PF值大的数据点相对于正常数据偏差较大时,就认为它是特异性数据或离群点。然而PF能精确描述标准正态分布数据的特异性规则,但对于更为一般的分布显得无能为力。结合文献[4,5]提出划分子空间的概念,在子空间内PF进行异常检测仍然是有效的。这样,杨剑等人在文献[7,8]中提出了局部特异性因子(Local Peculiarity Factor,LPF)的概念,并使用LPF算法结合K临近算法进行异常检测,由于K的选择比较难以确定,算法的计算复杂度比较高,文献[9]从数理统计的角度分析采样方法,为采样方法进行异常检测提供了精度保证和质量度量,在数理统计中,可以用样本的方差来估计总体的方差,样本方差是总体的无偏估计。基于此,孙静宇等人在文献[10]中提出了采样特异性因子(Sample Peculiarity Factor,SPF)的概念,理论和实验都表明采样特异性因子在不影响计算精度的情况下,可节约计算时间。随着人们对欺诈检测、网络入侵、故障诊断等问题的关注,实时异常检测日益受到重视。本文提出了基于SPF的学习采样频率算法,并将采样特异性因子和采样频率结合起来,用于实时异常检测。仿真实

到稿日期:2012-05-27 返修日期:2012-09-23 本文受山西省青年科技研究基金项目(200821024)资助。
牛之贤(1963-),女,硕士,副教授,主要研究方向为数据挖掘,E-mail:niuzx@163.com;孙静宇(1975-),男,博士,讲师,主要研究方向为数据挖掘、机器学习。

验表明该算法误检率为2%。

2 采样特异性因子及采样频率学习

文献[10]给出了采样特异性因子的概念，并提出了基于SPF的异常检测算法来计算每个数据点的采样特异性因子值，在理论研究和实验中表明SPF值大的数据点偏离正常数据较大，这些点亦即离群点或特异性数据。在单维实例中，将其称为属性集SPF，在多维实例中称为记录级SPF。本章通过选择适当的采样方法，利用折半最优的训练方法学习采样频率，将SPF和最优采样频率结合起来提出实时异常检测算法。

2.1 特异性因子

在文献[10]中给出了其定义和基于采样特异性因子的异常检测算法，见定义1。

定义1 假设 $T=\{C_1,C_2,\cdots,C_n\}$ 是包含 n 个数据点的采样集合，而 $S=\{S_1,S_2,\cdots,S_t\}$ 是来自 T 的一个采样，且每一数据点包含 m 个属性：$A_1,A_2,\cdots,A_m$，那么任意属性值 C_{ij} 的属性级采样特异性因子可由式(1)确定：

$$SPF(C_{ij})=\sum_{C_l\in S}D(C_{ij},C_{lj})=\sum_{C_l\in S}|C_{ij}-C_{lj}|^{\alpha} \tag{1}$$

式中，$D(x,y)$ 是一个距离函数，由参数 α 确定。设 β_j 是属性 A_j 的权重，那么，任意点 C_i 的记录级采样特异性因子可由式(2)确定：

$$SPF(C_i)=\sum_{S_i\in S}\sqrt{\sum_{j=1}^{m}\beta_j(SPF(C_{ij})-SPF(C_{lj}))^2} \tag{2}$$

上式可简化为：

$$SPF(C_i)=\sum_{j=1}^{m}\beta_j SPF(C_{ij}) \tag{3}$$

2.2 学习最优采样频率

基于SPF的异常检测，计算每个数据点的SPF值，理论上认为SPF值较大的数据点更有可能是异常点。为提高基于SPF的异常检测算法性能，进一步研究采用适当的采样方法在数据集中学习最优采样频率，以获得采样子集。本文实验使用UCI上的Lymphography数据集和Mammography数据集；Lymphography数据集中数据量较少，为保证采样的公平随机性和等概率性，采用随机采样方法。考虑到数据集的整体分布，照顾到各种类型的数据点，对于Mammography数据集使用均匀采样方法，在数据集上每隔 m 个数据点抽取一个数据点组成采样子集。采用多次采样求均值的方法来提高计算精度。

在基于SPF的异常检测算法中，要通过采样频率确定采样子集，通过研究数据集学习较优的采样频率，为进行实时异常检测提供强有力的理论支持。

采样频率优劣的评价，仍使用文献[5,6]中的方法，假设数据集由真实的Outlier和正常数据组成。算法预测出的Outlier可分为两种情况，即TP(True Positives)和FP(False Positives)。而预测出的正常数据也可分为两情况，即FN(False Negatives)和TN(True Negatives)。那么，检准率(Detection Rate)可定义为 $DR=TP/(TP+FN)$，假警率(False Alarm Rate)定义为 $FR=FP/(FP+TN)$。ROC曲线(Receive Operating Characteristic Curve)是以检准率作为X轴，以假警率作为Y轴而绘制的曲线，AUC(Area Under the Curve)是指ROC曲线下面的面积。理想情况时，ROC曲线是0%的假警率和100%检准率，但实际中只能近似达到。AUC是一种评估异常检测算法的定量近似方法。AUC越接近1，越说明接近理想的ROC，算法的效果越好。对于基于SPF的异常检测来说，希望在不影响精度的情况下，定义如下最优采样频率序列。

定义2 设频率级数 $\{r_1,r_2,\cdots,r_n,\cdots,r_N\}$，$f(r)$ 为 r 对应的使用异常检测算法得到的AUC值，$f_{\min}$ 为算法对应的最小AUC值，若有 $f(r_n)>f(r_N)>f_{\min}$，则 $\{r_n,\cdots,r_N\}$ 为最优采样频率序列。

为求得 $\{r_n,\cdots,r_N\}$ 最优采样频率序列，从分析数据服从的分布着手，由数理统计的思路可知，某些统计量在大样本条件下近似分布，这样将问题的本质归于正态总体的情形。

设 $(X_1,X_2,\cdots,X_n)^T$ 为总体 X 的样本，$EX=\mu$ 未知，当 n 较大时，求总体均值 μ 的置信度为 $1-\alpha$ 的置信区间。当 $n\to\infty$ 时，

$$\frac{\overline{X}-\mu}{S_n/\sqrt{n}}\xrightarrow{L}N(0,1)$$

对于给定的置信度，可求得 $u_{\frac{\alpha}{2}}$，使

$$P\{\frac{\overline{X}-\mu}{S_n/\sqrt{n}}<u_{\frac{\alpha}{2}}\}\approx 1-\alpha$$

即有

$$P\{\overline{X}-u_{\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}<\mu<\overline{X}+u_{\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\}\approx 1-\alpha$$

故均值 μ 的置信度为 $1-\alpha$ 的置信区间为

$$(\overline{X}-u_{\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}},\overline{X}+u_{\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}) \tag{4}$$

结合式(4)就通过分析AUC的均值和给定置信区间，得到AUC的取值范围，这样就可以在给定置信区间下得到最优采样频率。

实验结果与基于PF异常检测算法比较，设 $f(n)$ 为使用基于PF算法计算得到的AUC值，将实验的精度误差控制在千分之一之内，当相邻两个采样频率得到的AUC值小于千分之一时，就认为在此次学习的过程中找到了合适的采样频率，有式(5)：

$$\begin{aligned} &f(r)-f(n)>0.001*f(n)\\ &f(r)-f(r-1)<0.001 \end{aligned} \tag{5}$$

本文使用折半学习法，在数据集中学习最优采样频率。由采样特异性因子和采样频率，可得基于SPF的学习采样频率算法。

算法1 基于SPF的学习采样频率算法

输入：数据集 $T=\{C_1,C_2,\cdots,C_n\}$，置信度 $1-\alpha$，初始采样频率 r；

输出：采样频率范围。

步骤：

1. 用采样频率 r 在数据集中采样，获得采样数据集，对于每个 r 采样50次，使用式(1)、式(2)计算每个数据点的SPF，计算 r 对应的 f(r) 均值；
2. r 折半减小，利用式(4)、式(5)判断；
3. 输出采样频率范围。

这样得到一个最优的采样频率范围，可以根据不同的精度要求获取采样子集。实时情况下，计算实时数据的采样特

异性因子，通过排名判断是否为异常，并将其添加到已有数据集中，通过实时比照的方法进行实时异常检测。

2.3 基于 SPF 的实时异常检测算法

在原始数据集中学习最优采样频率，可提高计算 SPF 的效率，这样基于 SPF 的异常检测在精度影响不大的情况下，计算速度明显较快，适合用于实时要求比较高的场合，进而提出基于 SPF 实时异常检测算法，其步骤如下。

第 1 步　数据集预处理：借鉴分类问题，将数据集分成两类数据：正常数据集(Normal Dataset)和异常数据集(Anomalous Dataset)。

第 2 步　基于采样频率折半学习的方法得到一个基于正常数据集的采样频率，获取采样子集，计算各数据点的 SPF 值。该采样子集中只含有正常数据，相当于入侵检测的“白名单”，异常数据集就相当于“黑名单”。

第 3 步　实时处理时，将当前数据加入采样子集，计算 SPF 值，使用排名方法，若 SPF 排在最前面，则其为异常数据，进行异常处理，并将其放入“黑名单”中，否则，放在“白名单”中。

假设数据点为 C_n，有 $g(C_n)=0$，表示数据点 C_n 为正常数据，反之，$g(C_n)=1$，表示数据点 C_n 为异常数据，需要进行异常处理。详见算法 2。

算法 2　基于 SPF 的实时异常检测算法

输入：数据集 $T=\{C_1,C_2,\cdots,C_n\}$，采样频率 r，实时数据 C_{ni}

输出：$g(C_{ni})$的值

步骤：

1. 数据集 T 分类：正常数据集 N 和异常数据集 A；
2. 用采样频率 r 在正常数据集 N 中采样，获得采样子集 S 并计算每个数据点的 SPF，$S=S\cup C_{ni}$，使用式(1)、式(2)计算实时数据的 SPF 值；
3. 降序排列数据点的 SPF 值，计算 $g(C_{ni})$；
4. 输出 $g(C_{ni})$的值，返回第 2 步继续执行。

总之，使用采样频率获取采样子集，实时计算当前数据点的 SPF，由于异常数据具有多样性，判断时仅与正常数据，即“白名单”比照，就可以确定是否为异常，明显节约了时间，满足实时情况的要求。

3 实验

3.1 实验准备

实验数据使用 UCI 上的 Lymphography 数据集和 Mammography 数据集；实验环境，普通的 PC 机，内存 1G，硬盘 120G，内存类型 ddr2，硬盘转速 7200 转；操作系统 Windows XP，MATLAB 7.0。

3.2 实验结果分析

本文将基于 SPF 的异常检测算法与基于 PF、LPF 的异常检测算法进行对比实验，使用 Lymphography 数据集和 Mammography 数据集进行实验。Lymphography 数据集中包含标记 1，2，3，4 的 148 个实例，使用 18 个连续属性和一个离散属性来描述每一条记录。仅有 6 条记录(占 4.1%)标记为 1 或 4142 条记录标记为 2 或 3，把标记为 1 或 4 的记录认为是异常点，标记为 2 或 3 的认为是正常数据。Mammography 数据集由标记为 1 的 10923 条记录和标记为 2 的 260 条记录组成，每条记录用 6 个连续属性来表征，把标记为 2 的记录视为异常点，其他的为正常数据。

基于 Mammography 数据集的 3 种算法的实验结果为：基于 PF 算法的实验结果为：计算 PF 用时 163.485000s，AUC=0.8705；基于 LPF 算法取 $k=2000$，计算 LPF 用时 144.109000s，AUC=0.8633；基于 SPF 算法取 $r=0.05$，实验结果为：计算 SPF 用时 7.515000s，AUC=0.8731，三者对应的 ROC 曲线如图 1 所示，在采样频率线性递减的过程中比较 3 种算法用时和对应的 AUC 值如表 1 所列。

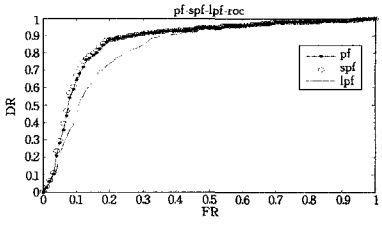

图 1　Mammography 数据集上 3 种算法对应的 ROC

表 1　Mammography 数据集基于 3 种算法的实验结果

	采样频率 r	时间 t/s	AUC 值
PF	1	183.375000	0.8701
LPF	1	123.325000	0.8689
SPF	0.5	90.390000	0.8694
	0.45	76.016000	0.8694
	0.4	69.687000	0.8702
	0.35	59.922000	0.8702
	0.3	50.734000	0.8702
	0.25	38.281000	0.8704
	0.2	30.656000	0.8693
	0.15	21.235000	0.8689
	0.1	14.594000	0.8709
	0.05	7.734000	0.8690

由图 1、表 1 可知，基于 SPF 异常检测算法在不同采样率时占用的 CPU 时间均小，当采样率超过 0.4 时，AUC 值超过了由基于 PF 异常检测算法得到的 AUC 值；而采样率为 0.05 时，AUC 值只比基于 PF 异常检测算法低不到 1%，而比基于 LPF 异常检测算法低不到 3%，这时的 CPU 占用时间分别为这两种算法的 1/20 和 1/30[10]。

由上面的实验可知，基于 SPF 的异常检测算法在不影响精度的情况，可节约 CPU 时间。进一步分析数据集时使用算法 1 学习最优采样频率，本文取置信度为 95%时，并以不同的 r 对应的 AUC 均值和标准差，来求置信度为 95%对应的 r 置信区间，Lymphography 数据集上的实验结果显示 AUC 置信区间为[0.9487，0.97509]，对应 r 的置信区间为[1/32，1/16]，即为所求的最优采样频率范围，如图 2 所示。

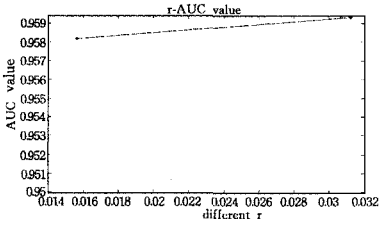

图 2　Lymphography 数据集上 r 的置信区间

Mammography 数据集上的实验结果显示 AUC 置信区间为[0.85762，0.87502]，对应 r 的置信区间为[1/64，1/2]，希望得到的 r 尽可能地小，可以记[1/64，1/32]为所求的最优采样频率范围，如图 3 所示。

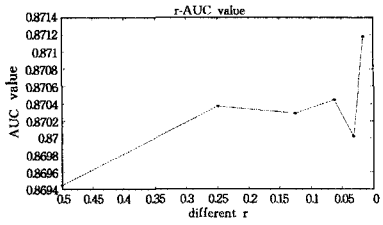

图 3 Mammography 数据集上 r 的置信区间

可见，在数据集规模较大时，在频率折半的过程中精度影响不大的情况下，计算时间也在折半减少，且在采样频率为[1/32,1/16]时算法较稳定。为提高精度，细化采样频率，Lymphography 数据集和 Mammography 数据集上基于 SPF 的采样频率在[1/32,1/16]之间以线性递减。采用多次采样求均值的方法，求不同的 r 对应的 AUC，实验结果如图 4、图 5 所示。

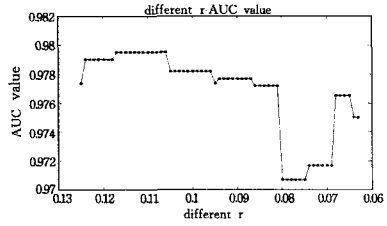

图 4 Lymphography 数据集上不同的 r 对应的 AUC

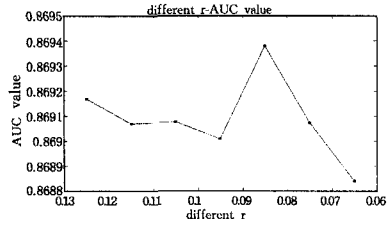

图 5 Mammography 数据集上不同的 r 对应的 AUC

为了得到最优采样频率，本文使用基于 SPF 算法进行仿真实验，实验中同样使用 Lymphography 和 Mammography 两个数据集。将数据集分为两部分，前一部分作为原始数据集，后一部分作为测试数据集。

在 Lymphography 数据集中将前 100 条数据作为原始数据集，后 48 条数据作为测试数据集，将测试数据集动态加入原始数据集，判断其是否为异常。实验结果与原数据中的标志位的比对显示，该算法的误检率为 2%，且所有异常数据均无误地被检测出来。误检仅是将正常数据判断为异常，经分析，误检的数据其某一维属性值明显异于其他，属于单维属性的异常点。在 Mammography 数据集上的实验也得到相似的结论。

结束语 由于基于采样特异性因子的异常检测在不影响精度的情况下可明显节约 CPU 时间，计算复杂度较低，因此本文通过优化采样方法，在训练集中折半学习采样频率，提出基于 SPF 的学习采样频率算法，得到最优采样频率范围，在 UCI 真实数据集上进行的实验表明采样频率为[1/32,1/16]，算法性能稳定。将 SPF 和最优采样频率结合起来提出了实时异常检测算法，该算法通过学习到的最优采样频率来获取采样子集，在实时处理中，计算当前数据的 SPF 值，通过排名方式判断其是否为异常，并进行异常处理。仿真实验表明，该算法误检率为 2%。

在未来的工作中，将进一步研究采样方法，分析数据集，提炼其数据分布，引进重采样或重点采样的方法。同时探索使用机器学习的方法学习采样频率和实时分析合理的评价准则。基于 SPF 的实时异常检测算法可以实现分布式实时异常检测，通过分布式处理提高算法的实时检测效率。

参 考 文 献

[1] Ohshima M, Zhong Ning, Yao Y Y, et al. Relational peculiarity oriented mining[J]. Data Mining and Knowledge Discovery, 2007, 15: 249-273

[2] Zhong Ning, Yao Y Y, Ohshima M, et al. Interestingness peculiarity, and multi-database mining[C]// Proceedings of the 2001 IEEE International Conference on Data Mining. 2001: 566-573

[3] Zhong Ning, Ohshima M, Ohsuga S. Peculiarity oriented mining and its application for knowledge discovery in amino-acid data [C]// Advances in Knowledge Discovery and Data Mining. 2001, 2035: 260-269

[4] 薛安荣. 空间离群点数据挖掘[D]. 镇江：江苏大学，2008

[5] 薛安荣，鞠世光，何伟华，等. 局部离群点挖掘算法研究[J]. 计算机学报，2007，30(8)：1456-1466

[6] Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey [J]. ACM Computing Survey, 2009, 41(3): 1-54

[7] Yang Jian, Zhong Ning, Yao Y Y, et al. Local peculiarity factor and its application in outlier detection[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Nevada, USA: the ACM, 2008: 776-784

[8] Yang Jian, Zhong Ning, Yao Y Y, et al. Peculiarity analysis for classifications[C]// Proceedings of the 2009 IEEE International Conference on Data Mining. Washington, DC, USA: IEEE Computer Society, 2009: 607-616

[9] Wu Ming-xi, Jermaine C. Outlier Detection by Sampling with Accuracy Guarantees[C]// Proceedings of the 2006 IEEE International Conference on Data. Washington, DC, USA: IEEE Computer Society, 2006

[10] 孙静宇. 基于 CBR 的协同 Web 搜索研究[D]. 太原：太原理工大学，2010

[11] Lazarevic A, Kumar V. Feature bagging for outlier detection[C]// Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2005: 157-166

[12] Ramaswamy S, Rastogi R, Kyuseok S. Efficient algorithms for mining outliers from large data sets[C]// Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. 2000: 427-438

[13] Bay S D, Mark S. Mining distance-based outliers in near linear time with randomization and a simple Pruning rule[C]// Proc. of the ACM SIGMOD Int'1 Conf. on Knowledge Diseovery and Data Mining. 2003: 29-38

[14] Angiulli F, Pizzuti C. Fast outlier detection in high dimensional spaces[C]// Proceedings of the Sixth European Conference on the Principles of Data Mining and Knowledge Discovery. 2002: 15-26