

一种基于多标记的局部离群点检测算法

钱景辉, 梁 栋

(南京工业大学 计算机科学与技术学院, 江苏 南京 211816)

摘 要: 基于密度的局部离群点检测算法(LOF)不适用于解决高维度、多义性的数据集检测. 通过对 LOF 算法的分析,提出了一种基于多标记学习(Multi-Label Learning, MLL)的局部离群点检测算法 MLL-LOF(a local outlier factor based on multi-label learning). 该算法采用 MLL 框架,首先将真实对象数据拆分成多示例包形式,然后运用退化策略及相应的权重调整,计算最终离群点因子,判别离群点. 并运用实际企业的监控数据将 MLL-LOF 算法与其他经典的局部离群点检测算法进行了对比实验,结果表明提出的 MLL-LOF 算法检测的精准率、召回率、F1 值以及时间效率均优于传统的局部离群点检测算法.

关键词: 机器学习;局部离群点;多标记;类别权重

中图分类号: TP31

文献标识码: A

文章编号: 1000-7180(2017)10-0110-05

Local Outlier Detection Algorithm Based on Multi-Label Learning

QIAN Jing-hui, LIANG Dong

(College of Computer Science and Technology, Nanjing Tech University, Nanjing 211816, China)

Abstract: The density-based local outlier detection algorithm (LOF) is not suitable for detecting the data set which is high dimension and polysemous. In this paper, we propose a local outlier detection algorithm based on multi-label learning(MLL-LOF). The main ideal of the MLL-LOF algorithm is as follows: Firstly, the real object data is divided into multi-instance by using an MLL framework, then the MLL-LOF calculates the final outlier factor and detects outliers by using degradation strategy and weight adjustment. We compare the MLL-LOF algorithm with other classical local outlier detection algorithms by using actual data set which is comes from enterprise monitoring. Experimental results show that the precision, recall, F1 and time efficiency of the MLL-LOF algorithm are superior to the traditional local outlier detection algorithm.

Key words: machine learning; local outlier; multi-label learning framework; class weights

1 引言

21 世纪以来,离群点检测作为机器学习的一个重要分支,备受关注. 离群点检测又名异常检测,是找出其行为不同于预期对象的一个过程. 离群点检测广泛应用在实际生活中,如医疗处理、公共安全、工业损毁检测、图像处理和入侵检测等.

自离群点被提出以来,也有许多检测离群点的算法应运而生. 这些算法中包括基于统计学的方法、基于深度的方法、基于距离^[1]的方法、基于聚类的方法、基于偏离的方法等.

21 世纪初, Breuing 等人^[2]提出了一种基于密度的离群点检测算法(local outlier detection algorithm, LOF),引进了局部离群点的概念.

传统的基于密度^[2]的局部离群点检测算法将对象属性杂糅在一个特征向量中,通过计算该对象和其邻近的相对密度,来判断该对象是否是离群点. 但是由于把对象的所有特征都杂糅在了一个特征向量中,使得该对象的关键信息和部分特征难以精确地体现出来,以至对是否是离群点的判断有一定的误

差.另外如果对象属性向量的维度过高,也使得计算变得复杂,这也是传统的基于密度的局部离群点检测算法的局限所在.

本文提出一种基于多标记^[3]学习框架(Multi-Label Learning framework)的局部离群点检测算法.

2 多标记学习 (MLL)

真实世界的对象往往不只具有一个语义^[4],而是具有多义性的,例如,图1这幅图像,既可以认为它描述的是“老虎”,也可以认为它属于“水边”“森林”;图2这个网页,可以认为它属于介绍詹姆斯这个人,也可以认为它属于“体育”这个类别.由于类似这样多义性的对象不只具有单一的语义,所以只考虑单一语义的学习框架很难取得良好的效果,因此我们引入了多标记学习的概念.



图1 一幅图像

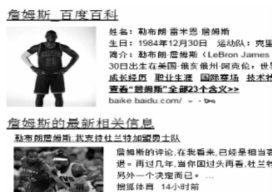


图2 一个网页

多标记学习是机器学习的一个新领域,已经成为机器学习的研究热点之一.多标记学习概念源于文档分类(text categorization)中遇到的多义性问题^[5],在多标记学习框架下,每个对象由单个示例表示并对应多个类别标记.

定义1 对于给定数据集 $D = \{(x_1, Y_1), (x_2, Y_2), \dots, (x_3, Y_3)\}$,目标是学得 $f: X \rightarrow 2^Y$. 其中, $x_i \in X$ 为一个示例, $Y_i \subseteq Y$ 为 x_i 的一组合适类别标记 $\{y_{i1}, y_{i2}, \dots, y_{il_i}\}$, $y_{ik} \in Y (d = 1, 2, \dots, l)$, l_i 为 Y_i 中所含标记类别个数. 其中 X 表示示例空间, Y 表示类别标记空间.

多标记学习框架如图3所示.

在现在的世界,每一个数据都具有多义性和复杂性等特性,一个对象的内部存在着许多不同类型的特征和内在属性,传统单一的离群点检测算法恰

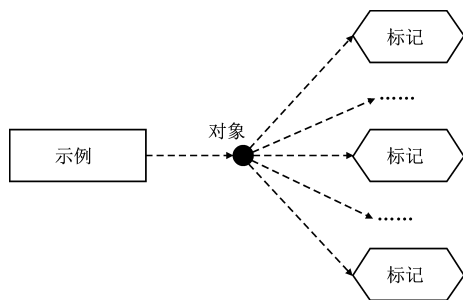


图3 多标记学习框架

恰忽略了这一点,而 MLL (Multi-Label Learning framework, 多标记学习) 框架正好可以很好地利用解决这一问题,鉴于上述分析本文提出了 MLL-LOF 算法,根据机器学习多标记的思想来分类,结合传统的离群点检测算法,在离群点检测这个领域取得更精确的结果,并将其应用到工业控制系统的异常检测之中.

3 MLL-LOF 算法

3.1 相关定义

定义2 对象 o 的 k -距离(记为 $dist_k(o)$). 对于给定数据集 D 和自然数 k ,使得 o 与另一个对象 $p \in D$ 之间的距离(记为 $dist(o, p)$),满足条件:

(1) 至少有 k 个对象 $o' \in D - \{o\}$,使得 $dist(o, o') \leq dist(o, p)$.

(2) 至少有 $k-1$ 个对象 $o'' \in D - \{o\}$,使得 $dist(o, o'') \leq dist(o, p)$.

换言之, $dist_k(o)$ 是 o 与其第 k 个最近邻之间的距离.

定义3 对象 o 的 k -距离邻域. o 的 k -邻域包含其到 o 的距离不大于 $dist_k(o)$ 的所有对象,记为 $N_k(o) = \{o' \mid o' \in D, dist(o, o') \leq dist_k(o)\}$

(1)

定义4 对象 o' 到 o 的可达距离 $reachdist_k(o \leftarrow o') = \max\{dist_k(o), dist(o, o')\}$. 对于两个对象 o 和 o' ,如果 $dist(o, o') \geq dist_k(o)$,则从 o' 到 o 的可达距离是 $dist(o, o')$, 否则是 $dist_k(o)$. 记为 $reachdist_k(o \leftarrow o') = \max\{dist_k(o), dist(o, o')\}$

(2)

定义5 局部可达密度($lrd_k(o)$). 对象 o 的局部可达密度是对象 o 与它的 $dist_k(o)$ 的平均可达距离的倒数. 计算公式为

$$lrd_k(o) = \frac{\|N_k(o)\|}{\sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)} \quad (3)$$

定义 6 局部离群点因子($LOF_k(o)$). 局部离群点因子是 o 的可达密度与 o 的 k - 最邻近的可达密度之比的平均值. 对象 o 的局部可达密度低, 并且 o 的 k - 最邻近的局部可达密度越高, 局部离群点因子越高. 计算公式为

$$LOF_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{lrd_k(o')}{lrd_k(o)}}{\|N_k(o)\|}$$
$$= \sum_{o' \in N_k(o)} lrd_k(o') * \sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o) \quad (4)$$

定义 7 多示例包集合. 根据文献[6] 中的定义, 给定集合 $X, \forall X_i \in X$, 称 X_i 为一个包, 包内含有多个示例 X_{ij} . 则定义 X 为多示例包集合.

定义 8 权重比例集合. 根据业内专家给出的数据集的标记集合 $L = \{l_1, l_2, \dots, l_m\}$, 以及对应标记的权重集合 $W = \{\omega_1, \omega_2, \dots, \omega_m\}$, 则对应标记的权重比例集合符合公式:

$$P = \{ \frac{|\omega_1|}{\sum_{i=1}^m |\omega_i|}, \frac{|\omega_2|}{\sum_{i=1}^m |\omega_i|}, \frac{|\omega_3|}{\sum_{i=1}^m |\omega_i|}, \dots, \frac{|\omega_m|}{\sum_{i=1}^m |\omega_i|} \} \quad (5)$$

定义 9 最终局部离群点因子($LOF_k(X_i)$). 最终局部离群点因子, 是根据专家给出的标记集合 L , 以及权重集合 W , 综合计算的结果. 计算公式为

$$FLOF_k(X_i) = \sum_{i=1}^m p_i * LOF_k(X_{ij})$$
$$= \sum_{i=1}^m \frac{|\omega_i|}{\sum_{j=1}^m |\omega_j|} * LOF_k(X_{ij}) \quad (6)$$

式中, $X_{ij} \in X_i, p_i \in P; \omega_i \in W; P$ 表示权重比例集合; W 表示权重集合; m 表示数据集所含标记的类别个数.

3.2 算法思想

随着科学技术的快速发展, 数据集愈趋复杂. 这给传统的以单一数据属性为基础的局部离群点检测算法带来了诸多挑战. 由于现实世界中数据对象属性多义性的特点, 在传统的局部离群点检测算法中往往忽视了数据属性的特性, 从而造成检测结果和预期结果存在误差.

运用 MLL 技术, 将原有的复杂数据对象, 根据业内专家给出的标记集合以及权重集合, 把不同对象根据标记集合拆分成多示例包的形式. 既在保留了数据属性共性的同时, 还体现了数据属性的特性,

避免了传统算法中由于忽略数据属性特性而造成的检测误差. 在通过 MLL 学习框架进行集合拆分后, MLL-LOF 算法采用退化策略^[7], 将单示例多标记问题转化成传统的单示例单标记问题来解决, 将通过多标记框架拆分成不同的多示例包之后的数据对象, 根据 lof 算法中的公式(3)和公式(4), 计算拆分过后的局部可达密度和局部离群点因子. 通过 MLL 学习框架的拆分以及退化策略的应用, 不仅有效地解决了数据复杂的多义性问题, 同时极大减少了数据属性的维度, 简化了计算过程.

在实际生活中, 不同类别标记的属性, 它们所带来的影响也是不同的, 所以 MLL-LOF 算法在计算出拆分过后的局部可达密度以及局部离群点因子后, 加入了定义 8 的权重调节集合, 在业内专家给出标记集合的同时也确定对应标记的权重集合. 在定义 9 中, 根据不同标记的不同权重, 计算出某一数据对象最终的局部离群点因子, 进行判断离群点. 通过权重调节, 使 MLL-LOF 算法对真实的对义性数据的适应性更强, 提高了局部离群点检测的准确性.

MLL-LOF 算法总共有三个阶段构成. 初始化阶段: 信息设置和数据采集; 第一阶段: 根据 MLL 学习框架将数据分割成多示例包的形式; 第二阶段: 根据 LOF 算法计算局部离群点因子; 第三阶段: 结合权重调整策略计算最终离群点因子, 并判别离群点. 具体如图 4 所示.

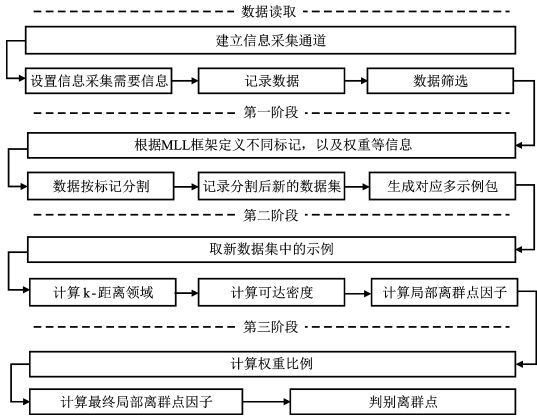


图 4 算法流程图

3.3 算法描述

算法描述如下:

- 输入:
- X : 对象数据集
 - L : 标记集合
 - n, m, s : 数据集总数, 标记总数,

示例中特征属性个数

W :权重集合

P :权重比例集合

输出:

$FLOF$:数据集合的最终局部离群点因子集合

Y :离群点集合

步骤:

1. for $\forall X_i \in X$ do
2. $MLLBag(X, L)$, 根据不同标记生成多示例包;
3. end
4. if ($W == \text{null}$)
5. $init(W)$ 初始化权重系数
6. end
7. for $\forall X_i \in X$ do
8. for $\forall X_{ij} \in X_i$ do
9. $brd_k(X_{ij})$ 计算示例块中的可达密度
10. $LOF_k(X_{ij})$ 计算示例块中的局部离群点因子
11. end
12. $FLOF_k(X_i)$ 根据权重计算最终局部离群点因子
13. end
14. $GetField(X)$ 计算离群点域
15. $Diff(Y)$ 判别离群点
16. End

4 算法仿真实验及分析

本文利用该安监局平台下属化工企业的真实数据集上的采样数据,对 MLL-LOF 算法的有效性和性能进行分析. 并与传统的基于密度的离群点检测算法(LOF)进行比较. 算法仿真运行环境 CPU 为 Intel(R) Core(TM) i7-4700MQ 2.40 GHz,内存为 8 GB;操作系统 Windows 7. 仿真代码用 Python 编写,开发工具为 IDLE.

某安监局平台下属化工企业有多个存储危险化学品原料的储罐,平台对化工企业储罐的状态进行监控,当某些状态发生变化超过阈值时,有及时的报警功能. 但由于算法原因,有时并不能及时地检测出哪些储罐的状态变化波动超过阈值,以至增加事故发生的危险性. 为了解决这个难题. 经过一系列的调查研究,我们将 MLL-LOF 算法应用到该平台下,对处理这类问题,有了一定的效果.

该化工企业具有 70 个化学储料罐,每 2 小时通过 OPC 将数据传输到服务器数据库中,我们从数据库中提取 2016 年 1 月到 2016 年 7 月的所有储罐属性作为数据源. 数据属性有温度、压力、水位等信息. 经过专家审定对不同的属性划分不同的标记和权重. 部分数据如表 1 所示.

表 1 初始数据表				
特征属性	1	...	$n-1$	n
1 101	0.918 1	...	8.1	17.999 7
1 102	0.915 6	...	14.4	13.218 8
1 103	0.815 8	...	7	28.343 8
1 104	0.791 3	...	6.1	11.843 8
1 105	0.791 3	...	6.12	-41
...	...	...	...	...
1 201	0.900 6	...	5.28	2.625
1 202	0.879	...	6.6	24.75
1 203	0.874 2	...	6.04	54

标记拆分成多个特征向量的形式,如标记物理状态属性、化学状态属性等. 针对第一条初始数据 $X_{1101} = (0.918\ 1, 14.765, 18.4, 17, 0.008, 1, 17.999\ 7, 1\ 500, 0, 8.1, 35, 0)$,按照各种属性可以分为三个特征向量 $X_{1001_1} = (0.918\ 1, 14.765, 18.4, 17, 0.008)$ 为化学属性, $X_{1001_2} = (1, 17.999\ 7, 1\ 500, 0, 8.1)$ 为物理属性, $X_{1001_3} = (35, 0)$ 为材质属性. 在安全方面考虑,化学属性最为重要,因此化学属性权重最高,物理属性居中,材质属性最低. 对应权重为,化学属性:5,物理属性:3,材质属性:1.

初始所有数据有 176 400 条,其中每个储罐拥有 2 520 条数据,现在我们通过随机采样的方式选取一个储罐 1605 号储罐,里面所放物料为甲醇. 我们在 1 605 号储罐的 2 520 条数据中抽取 700 条数据进行仿真检测. 对于 1 605 号储罐抽取数据共有 26 个离群点,设置系数 k 为 20. 分别运用 MLL-LOF、DBLOD、DLOF 以及 LOF 算法进行仿真检测,结果如图 5 所示.

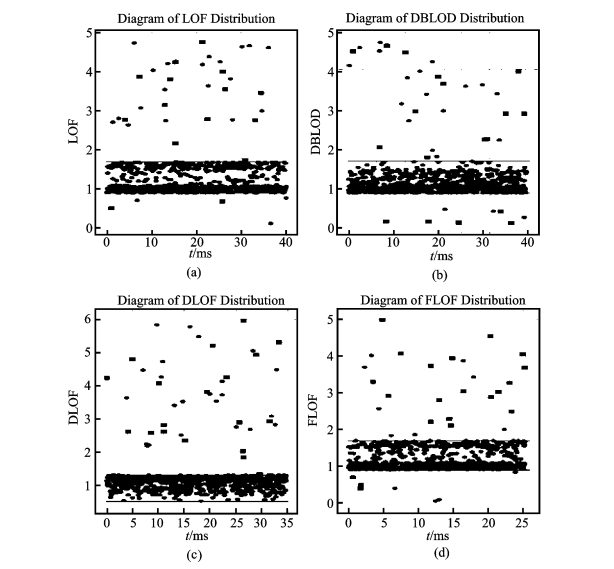


图 5 仿真对比图

如图 5 所示,图中有分割线,内部为正常数据.通过图 5 得知,传统的 LOF 算法检测出离群点有 43 个,正确个数为 15 个,用时 39.973 1 ms. DBLOD 算法检测出 39 个离群点,正确个数为 18 个,用时 31.459 8 ms,DLOF 算法检测出离群点有 42 个,正确个数为 19 个,用时 34.746 1 ms,MLL-LOF 算法公检测出离群点 37 个,正确个数为 21 个,用时 25.387 6 ms.

为了检测算法的有效性,采用精确率(Precision)、召回率(Recall)以及 F1 值来进行验证评测.

(1)Precision=提取出的正确信息条/提取出的信息条数.

(2)Recall = 提取出的正确信息条数/样本中的信息条数.

(3)F1 值=正确率 * 召回率 * 2/(正确率+召回率).

评测结果见表 2. 由于 LOF、DBLOD、DLOF、算法在针对局部数据块分析中,DBLOD 采用先分区,再单独计算偏离度;DLOF 算法通过信息熵,计算加权距离,在得到可达密度. 以上算法都忽略了实际数据多义性的差异. MLL-LOF 算法充分考虑了数据的复杂性、多义性特点,在工业数据检测中有着更好的结果.

表 2 实验结果对比表				
算法	LOF	DBLOD	DLOF	MLL-LOF
精确率	0.348 8	0.461 5	0.452 4	0.567 6
召回率	0.576 9	0.692 3	0.730 8	0.807 7
F1 值	0.434 7	0.553 9	0.558 8	0.666 7
平均用时/ms	0.057 1	0.044 9	0.049 6	0.036 3

5 结束语

离群点检测是一个非常有应用价值的研究领域,近年来受到学者和企业越来越多的关注,但由于数据的复杂性、多义性,至今仍然没有通用、高效的方法来解决这个问题. 本文通过对离群点检测的研

究提出了一种基于多标记(Multi-Label Learning)的离群点检测方法,能有效地解决数据复杂性、多义性的问题,也提高了离群点检测的准确率和效率. 然而 MLL-LOF 算法也存在着一些不足之处,对于标记的分类,以及权重的把握可能对算法的结果造成一定的影响. 如何能够准确的定义标记、权重,采用智能化的方法来进行动态的权重分配都是我们以后要探索的问题.

参考文献:

[1] Sun Huanliang, Bao Yubin, Yu Ge, et al. An algorithm based onpartition for outlier detection [J] . Journal of Software, 2006, 17(5):1009-1016.

[2] Breunig M , Kriegel H, Ng R, et al. Lof: Identifying density-based local outliers [C] // Proc. SIGMOD Conf. IEEE, 2000:93-104.

[3] Zhou Z H, Zhang M L, Huang S J, et al. Multi-instance multi-label learning[J]. Artificial Intelligence, 2012, 176(1):2291-2320.

[4] Boutell M R, Luo J, Shen X, et al. Learning multi-label scene classification[J]. Pattern recognition, 2004, 37(9): 1757-1771.

[5] Schapire R E, Singer Y. BoosTexter: a boosting-based system for text categorization[J]. Machine Learning, 2000, 39(2/3): 135-168.

[6] Dietterich T G, Lathrop R H, Lozano-Pérez T. Solving the multiple-instance problem with axis-parallel rectangles[J]. ArtificialIntelligence, 1997, 89(1/2): 31-71.

[7] Zhou Z H, Zhang M L, Huang S J, et al. Multi-instance multi-label learning[J]. Artificial Intelligence, 2012, 176(1): 2291-2320.

作者简介:

钱景辉 男,(1978 -),博士,讲师. 研究方向为计算机控制自动化.

梁 栋(通讯作者) 男,(1991 -),硕士研究生. 研究方向为机器学习、离群点检测. E-mail:js_liangdong@163. com.