文章编号:1004-1478(2011)03-0008-04

基于聚类融合的异常检测算法

苏晓珂, 王秉政

(郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450002)

摘要:针对任意形状聚类算法用于异常检测时参数设置困难的问题,提出一种基于聚类融合的异常检测算法:设置不同的半径阈值进行多次聚类,统计每次聚类中标记为异常的簇频率,将频率高的簇作为真正的异常.在UCI数据集上对该算法进行实验,结果表明:本算法可降低直接将小簇作为异常的高误报率,并且能提供给用户更为友好的操作.

关键词:异常检测;聚类融合;异常簇;任意形状聚类

中图分类号:TP391 **文献标志码**:A

An outlier detection algorithm based on clustering ensemble

SU Xiao-ke, WANG Bing-zheng

(*College of Comp. and Com. Eng., Zhengzhou Univ. of Light Ind., Zhengzhou* 450002, *China*)

Abstract: An outlier mining algorithm based on the clustering ensemble was presented in order to reduce the reliance for users and decrease the high false positive rate due to taking the small size clusters as the outliers directly. Outliers can be found according to the abnormal frequency of every record. The algorithm is able to provide the user a more friendly operation. The experimental results on the real-life datasets showed that the proposed algorithms are feasible and effective comparing with other classical algorithms and can be used for mixed dataset.

Key words: outlier detection; clustering ensemble; abnormal cluster; arbitrary shape clustering

0 引言

异常检测的目标是发现数据集中偏离大部分数据的数据,因为这些数据的偏离也许并非由随机因素产生,而是产生于完全不同的机制[1].在模式识别领域,异常检测可看做一个特殊的分类问题.作为无监督模式识别的一个重要分支,聚类具有不需要任何先验知识的特性.从聚类的角度看,异常是不在聚类中的点,也就是聚类的补,因此基于聚类的异常检测得到了广泛研究.文献[2-4]分别提出了几种基于聚类的异常检测方法,这些方法先采用特殊的聚类算法处理输入数据得到簇,再在簇的基础上检测异常.但文献[2]只能应用于纯数值属性数据集;文献[3]将小簇作为离群簇的定义不够准确,有些小簇可能是正常簇的边界;文献[4]只能直接应用于纯分类属性数据集.

本文拟提出一种基于聚类融合的异常检测算法:将任意形状聚类算法作为基础算法,对较大范

收稿日期:2011-04-19

基金项目:河南省科技攻关项目(092102210108);河南省教育厅自然科学基础研究计划项目(2010A520033);郑州轻工业学院博士科研基金资助项目

作者简介:苏晓珂(1979—),女,河南省巩义市人,郑州轻工业学院讲师,博士,主要研究方向为智能计算.

围内不同阈值下聚类得到的候选异常进行融合,统计对象作为候选异常的频率,频率高的候选异常被标记为真正的异常. 该算法对领域知识要求低,对参数依赖程度小.

1 聚类融合

融合方法在分类和回归中的使用已经比较成熟,近年来被引入到聚类领域. 受到在传感器融合和分类器融合方面成果的启发,A. L. Fred[5] 提出了一系列基于 Co-association 共生矩阵、用类似投票的方法融合聚类结果的聚类融合方法. 文献[6]正式提出聚类融合的概念:将多个对一组数据进行聚类的不同结果进行融合,而不使用数据原有的特征.

对于聚类融合方法的研究主要集中在2个方面:1)如何产生有效的聚类成员;2)如何设计共识函数,对聚类成员进行很好的合并. 文献[7]详细研究了聚类成员的差异性对聚类融合结果的影响和聚类融合的稳定性问题,聚类融合具体表达如下:

假设有包含 n 个对象的数据集 $X=\{x_1,x_2,\cdots,x_n\}$,对数据集 X 用 h 次聚类算法得到 h 个聚类结果,$H=\{C_1,C_2,\cdots,C_h\}$,其中 $C_k(k=1,2,\cdots,h)$ 为对第 k 次算法得到的聚类结果. 设计一种共识函数,对这 h 个聚类成员的聚类结果进行合并,得到一个最终的聚类结果[8].

聚类融合方法能得到比单一算法更为优越的结果[9].

1)鲁棒性:在各领域和数据集中的平均性能更为优越;2)适用性:能得到单一聚类方法难以达到的聚类结果;3)稳定性和确定性评估:噪声、孤立点和抽样方法对聚类结果的影响较小,同时聚类结果的不确定性可以从融合的分布情况上进行评估;4)并行性和可扩展性:能对数据子集进行并行聚类和合并,并能对分布式数据源或数据属性的聚类结果进行合并.

2 异常检测算法

用 N 表示记录总数,h 表示融合次数,$la[N]$ 为一向量,元素 $la[i]$($i\in 1,N$)表示在 h 次不同划分中对象 p_i 被标记为异常的次数. r 表示聚类阈值,上限和下限用 α,β 表示. 以任意形状聚类算法为基础算法[10],选择不同半径阈值,对数据集多次聚类,每次聚类结果依簇的尺寸升序排列,小簇作为候选异常,将多次检测结果进行融合,最终得到真正的异常.

2.1 算法描述

输入:数据集 D,异常簇包含记录数阈值 θ,异常频率阈值 λ,聚类阈值的上限 α 和下限 β

输出:异常集 OS

```
初始化 la[N] =0
j=0
while j<h do
  在范围[α,β]内随机选择阈值 r
  任意形状聚类,生成簇集合 C={C1,C2,…,Cw}
  l=0
  while l<w do
    if |Cl|<θ then
      Cl 被标记为候选异常
      ∀pi∈Cl(i∈[1,N])
      la[i-1]++
    end if
    l++
  end while
  j++
end while
i=0,OS=∅
while i<N do
  if la[i]>λ then
    pi 被标记为真正的异常,OS=OS∪{pi}
  end if
end while
```

任意形状聚类算法可描述为:初始时,簇集合 CS 为空,从数据集 D 中读入一个新对象,以这个对象构造一个新簇. 若已到数据集末尾,则聚类过程结束,否则读入新对象 p,寻找 p 的所有近邻,计算包含 p 近邻的簇个数 l. 若 $l=0$,即 p 没有近邻,以 p 构造一个新簇,继续读入下一对象. 若 l 个簇中存在可合并簇,将其合并;若不存在可合并簇,将 p 添加到具有最大相似度的簇中,并更新该簇,重复读入过程,直到 D 为空.

算法中的 j 表示融合次数,虽然输入参数有4个,但都能以粗粒度给出;而文献[10]中的任意形状聚类算法对半径阈值的取值非常敏感,必须经历多次试探才能得到最合适的阈值. 每次聚类,在范围[α,β]内随机选择阈值,然后调用任意形状聚类算法,形成的簇集合中包含对象数小于 θ 的簇中所有对象都作为候选异常,统计异常次数加1. 扫描向

量,对 h 次聚类结果进行融合,确定每个对象的最终状态.

2.2 效率分析

初始化各变量,计算复杂度为 $O(N)$. 任意形状聚类期望的时间复杂度为 $O(m \cdot N^2)$;每次聚类中标记候选异常对象和扫描向量,时间复杂度为 $O(w \cdot N)$;融合 h 次,时间复杂度为 $O(h \cdot (m \cdot N^2 + w \cdot N))$;最终需要扫描向量,时间复杂度为 $O(N)$. 由此可见,算法执行时间主要由任意形状聚类算法决定,期望的整体时间复杂度为 $O(h \cdot m \cdot N^2)$.

3 实验结果

检测率、假正率是衡量异常检测方法性能的2个指标. 检测率 DR(detection rate)表示被正确检测的异常记录数占所有异常总数的比例;假正率 FR(false positive rate)表示正常记录被检测为异常的记录数占整个正常记录数的比例. 对于异常检测,理想的结果是具有高的检测率和低的假正率[11].

在UCI机器学习数据集上进行实验[12],实验环境为:2.2 GHz Intel Pentium IV 处理器,512 MB 内存,Windows XP Professional 操作系统,编程语言使用 Visual C++ 6.0. 对所有的数据集融合10次.

3.1 Lymphography 淋巴系造影术数据集

为测试算法对分类属性数据集的检测性能,在淋巴系造影术数据集上进行测试. 此数据集包含148条记录,每条记录具有18个分类属性. 所有记录被分为4类,类1含2条记录,类2含81条记录,类3含61条记录,类4含4条记录. 类1与类4的记录只占整个数据集全部记录的4.05%,可看做异常记录,检测结果如表1所示.

表1 Lymphography 检测结果

λ	DR/%	FR/%
4	100.00	35.21
5	83.33	5.63
6	83.33	0.00
7	50.00	0.00
8	33.33	0.00

由表1看出,当异常频率阈值 $\lambda = 6$ 时,检测到的前5条记录均为真正的异常记录,因此误报率为0,检测率为83.33%,但若检测出全部6条记录,误报率将达到35.21%.

3.2 breast cancer 数值数据集

breast cancer 数据集有699条记录,其中良性记录458条,恶性记录241条,每条记录包含9个数值属性. 直观地判断,恶性与良性记录应有明显区别. 因此,随机选取483条记录构造分布不平衡的测试集,其中恶性记录有39条(8%),良性记录有444条(92%),期望能够将比例很小的那部分记录从测试集中检测出来. 检测结果如表2所示.

表2 breast cancer 数据集检测结果

λ	DR/%	FR/%
4	100.00	6.98
5	100.00	6.08
6	100.00	3.83
7	97.44	3.38
9	97.44	3.15

由表2看出,随着异常频率阈值的增大,检测率和误报率都逐渐降低. 原因在于异常频率阈值越大,被判断为异常的记录数就越少,检测率就越低,正常被判断为异常的记录数也越少,因此误报率也就越低. 当异常频率阈值 $\lambda = 6$ 时,检测效果最理想,检测率为100%,而误报率为3.83%.

3.3 A_1 入侵检测数据集

Kddcup99 数据集中的每条记录包含7个分类属性和34个数值属性. 选择与文献[11]中相同的数据子集 A_1,包含38 841条正常记录和1 618条攻击记录,攻击记录占4%,其中 DoS 攻击占98.39%,U2R 攻击占0.06%,R2L 攻击占0.37%,Probe 攻击占1.11%,其他攻击占0.07%. 检测结果如表3所示.

表3 A_1 数据集检测结果

λ	DR/%	FR/%
1	99.01	1.97
2	98.95	1.90
3	98.76	1.88
4	28.37	0.00
8	22.25	0.00

仔细分析表3可以看出,最优情况下异常频率阈值为1时,检测率可达99.01%,而假正率仅为1.97%,说明了 A_1 中的正常记录很好地聚集在一起,并且异常记录远离正常记录.

3.4 算法对比

将本文提出的算法同文献[11]中的 TOD,CBOD 和文献[13]中的 ODBUG 算法,文献[3]中的 UAD,文献[14]中的 Cluster 在3种数据集上的实验结果进行对比,比较各种算法的检测率与假正率,

从而测试这些算法的异常检测效果.结果见表4.

表4　相关算法对比　　%

算法	Lymphography		breast cancer		A_1	
	DR	*FR*	*DR*	*FR*	*DR*	*FR*
本文	83.33	0.00	100.00	3.83	99.01	1.97
TOD	100.00	2.82	97.44	2.70	98.76	7.80
CBOD	100.00	2.11	100.00	4.05	98.39	5.30
ODBUG	100.00	3.52	100.00	3.83	98.45	0.14
UAD					88.00	8.14
Cluster					93.00	10.00

由表4可以看出,本文算法在 A_1 数据集上的检测效果明显优于其他2种算法,取得了最高的检测率和较低的假正率;在 breast cancer 数据集上的检测效果与 ODBUG 相当,检测率达到100%,而假正率仅为3.83%;而在 Lymphography 纯分类属性数据集上的检测效果较差.

4　结论

本文基于数据集中的正常记录占绝大部分、异常记录会偏离正常记录、正常记录与异常记录会聚集在不同类中的思想,提出了一种基于聚类融合的异常检测算法,对较大范围内不同阈值下得到的候选异常进行融合,识别真正的异常.在医疗、网络入侵检测等真实数据集上的测试结果表明,本算法很好地检测到了数据集中的异常,并能提供给用户更为友好的操作.

参考文献:

[1] Patcha A, Park J M. An overview of anomaly detection techniques: Existing solutions and latest technological trends[J]. Comp Networks, 2007, 51(12): 3448.

[2] Jiang M F, Tseng S S, Su C M. Two-phase clustering process for outliers detection[J]. Computational Statistics and Data Analysis, 2001, 36(3): 351.

[3] Portnoy L, Eskin E, Stolfo S. Intrusion detection with unlabeled data using clustering[C]//Proc of the ACM Workshop on Data Mining Applied to Security, Philadelphia: PA, 2001: 5-8.

[4] He Z, Xu X, Deng S. Discovering cluster-based local outliers[J]. Pattern Recognition Letters, 2003, 24(9-10): 1651.

[5] Fred A L. Finding consistent clusters in data partitions[C]//Procs of the Second Int Workshop on Multiple Classifier Syst Lecture Notes in Comp Sci, London: Springer-Verlag, 2001: 309-318.

[6] Strehl A, Ghosh J. Cluster ensembles—A knowledge reuse framework for combining multiple partitions[J]. J of Machine Learning Research, 2003, 3(3): 583.

[7] Kuncheva L I, Vetrov D P. Evaluation of stability of *K*-means cluster ensembles with respect to random initialization[J]. IEEE Trans Pattern Analysis and Machine Intelligence, 2006, 28(11): 1798.

[8] 蒋盛益.基于投票机制的融合聚类算法[J].小型微型计算机系统, 2007, 28(2): 306.

[9] Topchy A, Jain A, Punch W. A mixture model for clustering ensembles[C]//Proc of the 4th SIAM Int Conf on Data Mining, Orlando: Florida, 2004: 379-390.

[10] 苏晓珂,兰洋,程耀东,等.可处理混合属性的任意形状聚类[J].计算机工程与应用, 2010, 46(34): 136.

[11] 蒋盛益,姜灵敏.一种高效异常检测方法[J].计算机工程, 2007, 33(7): 166.

[12] Merz J, Merphy P. UCI repository of machine learning databases [DB/OL]. (1999-01-01)[2010-12-10]. http://www.ics.uci.edu/~mlearn/MLRRepository.html.

[13] 蒋盛益,李庆华,赵延喜.一种两阶段异常检测方法[J].小型微型计算机系统, 2005, 26(7): 1237.

[14] Eskin E, Arnold A, Prerau M. A geometric framework for unsupervised anomaly detection: detecting intrusions in unlabeled data[C]//Proc of the Appli of Data Mining in Comp Security Advances in Infor Security, Norwell: Kluwer Academic Publishers, 2002: 272.