软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(9):2320−2331 [doi: 10.13328/j.cnki.jos.004872]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种基于 *k* 近邻图的稀有类检测算法*

王 淞[1], 黄 浩[1], 余 果[2], 梁 楠[1], 王黎维[3], 孙月明[1]

[1](武汉大学 计算机学院,湖北 武汉 430072)

[2](武汉大学 中南医院,湖北 武汉 430072)

[3](武汉大学 国际软件学院,湖北 武汉 430072)

通讯作者: 黄浩, E-mail: haohuang@whu.edu.cn

摘 要: 稀有类检测的目标是为无类别标签的数据集中的每个类,特别是仅含少量数据样本的稀有类,寻找到至少一个数据样本以证明数据集中存在这些类.该技术在金融欺诈检测及网络入侵检测等现实问题中具有广泛的应用场景.但是,现有的稀有类检测算法往往存在以下问题:(1) 时间复杂度比较高;或(2) 对原始数据集需要一定的先验知识,如数据集中各类数据样本所占比例等.提出了一种基于 *k* 邻近图的无先验快速稀有类检测算法 KRED,通过利用稀有类数据样本在小范围内紧密分布所造成的与周边数据分布的不一致性来定位稀有类.为此,KRED 将给定数据集转化为 *k* 邻近图,并计算图中各顶点入度和边长的变化.最后,将以上变化最大的顶点对应的数据样本作为稀有类的候选样本.实验结果表明:KRED 有效提高了发现数据集中各个类的效率,明显缩短了算法运行所需时间.

关键词: 稀有类检测;*k* 邻近图;数据分布;变化系数;入度

中图法分类号: TP311

中文引用格式: 王淞,黄浩,余果,梁楠,王黎维,孙月明.一种基于 *k* 近邻图的稀有类检测算法.软件学报,2016,27(9):2320−2331. http://www.jos.org.cn/1000-9825/4872.htm

英文引用格式: Wang S, Huang H, Yu G, Liang N, Wang LW, Sun YM. Rare category detection algorithm based on *k*-nearest neighbor graphs. Ruan Jian Xue Bao/Journal of Software, 2016,27(9):2320−2331 (in Chinese). http://www.jos.org.cn/1000-9825/4872.htm

Rare Category Detection Algorithm Based on *k*-Nearest Neighbor Graphs

WANG Song[1], HUANG Hao[1], YU Guo[2], LIANG Nan[1], WANG Li-Wei[3], SUN Yue-Ming[1]

[1](Computer School, Wuhan University, Wuhan 430072, China)

[2](Zhongnan Hospital, Wuhan University, Wuhan 430072, China)

[3](Int'l School of Software, Wuhan University, Wuhan 430072, China)

Abstract: Rare category detection aims at finding at least one data example for each class in an unlabeled data set to prove the existence of these classes, especially the rare classes (a.k.a. rare categories) that have only a few data examples. It has various applications in the fields like financial fraud detection and network intrusion detection. Nevertheless, the existing approaches to this problem suffer either in terms of time complexity or the requirements for prior information about data sets (e.g., the proportion of data examples in each class). In this paper, a prior-free and efficient algorithm, called KRED is proposed for rare category detection. The algorithm explores the changes on local data distribution caused by the presence of the compact clusters of rare classes. To this end, it transforms a data set into a *k*-nearest neighbor graph, and investigates the variations in both edge lengths and in-degrees between the nodes. Finally, nodes with the

* 基金项目: 国家自然科学基金(61502347, 61272275, 61202033, 61070013, U1135005); 中央高校基本科研业务费专项资金(2042015kf0038); 武汉大学人才计划/引进人才科研启动经费

Foundation item: National Natural Science Foundation of China (61502347, 61272275, 61202033, 61070013, U1135005); Fundamental Research Funds for the Central Universities (2042015kf0038); Research Funds for Introduced Talents of Wuhan University

收稿时间: 2014-12-01; 修改时间: 2015-03-10; 采用时间: 2015-05-11

maximal variations are selected as the candidate data examples of rare classes. Experimental results show that KRED effectively improves the efficiency of discovering new classes in data sets, and notably reduces the execution time.

Key words: rare category detection; k-nearest neighbor graph; data distribution; variation coefficient; in-degree

稀有类检测旨在发现无类别标签数据集中存在哪些类,特别是哪些稀有类.这是因为这些稀有类虽然数据样本较少,但往往比占据数据集数据样本绝大多数的主要类更具有现实意义,更值得被进一步研究[1].例如:在海量金融交易记录中,有时隐藏着少量利用金融系统的漏洞或采取欺诈手段进行的不合法交易[16];在海量正常网络访问中,存在少量的恶意网络行为[17].除可以用于以上实际问题,稀有类检测还能从给定的无类别标签数据集获得少量已分类数据样本,因此,其检测结果能用于寻找已发现稀有类余下数据样本[19]、构造分类器[6]或者用于半监督的学习方法,如协同训练[10]与主动学习[11]等.因此,稀有类检测在实际应用和理论研究中都具有广泛应用场景和较高研究价值.

由于稀有类数据样本数过少且常常隐藏在主要类的数据分布中,如金融欺诈操作往往伪装成正常交易行为,因此,传统的聚类、分类技术往往较难快速、准确地检测出稀有类.不过,稀有类通常具有以下特征以助识别,例如形成紧实的簇[2,7−9]、与其周边区域间存在数据分布上的较大差异[2−5]等.因此,现有的稀有类检测算法的一般步骤是:首先分析数据集,将其中具有这些特征的数据样本选取出来作为稀有类的候选数据样本;然后,向贴标者(如具有领域知识的人类专家)问询它们的真实类别标签.一个好的稀有类检测算法应当既减少分析数据集时所需要的计算量,又减少贴标者的工作量,即,发现数据集全部类时贴标者的贴标次数.

目前为止,根据是否需要先验知识,如类别个数[1]和各个类的数据占数据集的比例[2−4],稀有类检测算法可以分为基于先验知识的稀有检测算法和无先验知识的稀有类检测算法两类.对于前者,现有算法可归纳为如下几种:基于模型的方法[1]、基于边界度的方法[2]、基于密度差异的方法[3,4].但是在实际应用中,由于并不是每一个用户对其所使用的数据集都具有先验知识,因此该类算法在适用范围上具有一定的局限性.另外,无先验知识的算法[5,7,8]虽然摆脱了需要先验知识这一局限性,但它们在数据分析阶段需要的时间开销并不理想,通常具有较高的时间复杂度.

为了避免现有算法的局限性,本文提出了一种更为高效的无先验稀有类检测算法——基于 k 邻近图的稀有类检测算法 KRED(k-nearest neighbor graph based rare category detection).该算法通过利用稀有类数据样本在小范围内集中紧密出现所造成的局部数据分布的突变来锁定稀有类的可能分布区域.具体来说,KRED 首先使用自动选取的 k 值将给定数据集转化为 k 近邻图,然后,利用变化系数 Vc(即,结合入度变化和邻接边长变化的参数)来确定局部数据分布变化最大的数据点,问询其类别标签,从而发现稀有类.同现有的无先验方法相比,我们提出的算法不需要像半参密度估计(semi-parametric density estimation)[5]或层次均值平移方法(hierarchical mean shift)[7]那样复杂的计算开销,因此能够更快速有效地解决稀有类检测问题.实验证明:KRED 算法明显减少了算法运行所需要的时间,并有效地减少了发现数据集全部类所需要的贴标次数.

本文第 1 节回顾现有相关工作.第 2 节介绍变化系数 Vc 等相关概念.第 3 节介绍 KRED 算法.第 4 节给出实验结果与分析.最后,第 5 节总结全文,并给出未来工作方向.

1 相关工作

本节将回顾稀有类检测的相关工作,根据是否利用了用户对给定数据集的先验知识,将稀有类检测方法分为两种进行介绍:(1) 基于先验知识的稀有类检测方法;(2) 无先验知识的稀有类检测方法.

1.1 基于先验知识的稀有类检测方法

基于先验知识的稀有类检测是指在用户对所使用的不平衡数据集具有先验知识的情况下,通过利用这些先验知识帮助用户为该数据集中的每个稀有类找出至少一个数据样本,从而发现稀有类.常用的先验知识包括两种:(1) 数据集中类的个数;(2) 每个类的数据样本在数据集中的大致比例.

在现有文献中,基于先验知识的稀有类检测问题的解决方案分为 3 类:(1) 基于模型的方法;(2) 基于边界度

的方法;(3) 基于密度差异的方法.

1.1.1 基于模型的方法

Interleave 算法[1]是一种典型的基于模型的稀有类检测算法,该算法假设数据集中 m 个类构成 m 个高斯分布,并利用 EM 算法训练得到数据集的高斯混合模型.在每个高斯分布中,隶属度最低的数据点被认为最不不满足该模型,且将被选择出来给贴标者贴类别标签.而高斯混合模型也将因为有新的有类别标签的数据产生而被更新,并在新的一轮迭代中选出最不满足新模型的数据点用于贴标,直到为每个类发现一个数据样本.该算法的时间复杂度为 $O(dn^2)$,其中,d 为维度,n 为数据集中数据样本的数目.需要指出的是:由于该算法假设稀有类和主要类分别构成不同高斯分布,使得该算法往往在稀有类和主要类线性可分的情况下才能有较好的表现[6].

1.1.2 基于边界度的方法

RADAR 算法[2]和 CATION 算法[18]利用了各个类的数据分布在边界区域比内部区域更稀疏的特点,通过计算数据点反向 k 近邻的变化来衡量数据点的边界度.为了检测稀有类的边界,该算法利用稀有类数据占总体数据的比例来控制计算边界度过程中数据区域选取的大小,使得稀有类的边界点具有更高的边界度.最终,该算法按照边界度从高到低的顺序选取数据点并询问其类别.但是,主要类中的噪音或者不合适的数据选取区域的大小,有时会较为明显地降低该算法识别稀有类边界区域的准确率.

1.1.3 基于密度差异的方法

NNDM 算法[3]和 GRADE 算法[4]通过稀有类与周边数据之间的密度差异来发现稀有类,它们假设主要类的概率分布函数是局部平滑的,而每个稀有类则聚集于一个小区域内.因此,稀有类出现时,其出现区域的局部密度会发生剧烈变化.那么,在局部密度变化较大的区域选择数据样本问询,会有较高的几率发现稀有类.为此,这两个算法以各个数据点为球心划超球(超球的大小由稀有类类的样本占数据集的比例大小决定),并将超球内点的数目作为各个数据点的局部密度,并考察其与周边数据点之间密度差异.但是,该算法需要花费较多询问次数才能发现那些与周边数据密度差异不大的稀有类.

1.2 无先验知识的稀有类检测方法

目前,只有少数的现有方法研究了无先验知识的情况,即:当用户对待处理的数据集完全没有先验知识时,如何进行稀有类检测.而在实际应用中,用户对手中数据集没有先验知识的情况更为常见,换言之,无先验知识的稀有类检测具有更为普遍的应用场景.现有的无先验知识的稀有类检测技术可以划分为 3 大类:(1) 基于离群程度的稀有类检测技术;(2) 基于密度变化率的稀有类检测技术;(3) 基于近邻关系的稀有类检测技术.

1.2.1 基于离群程度的方法

HMS 算法[7]假设稀有类在小范围内集中分布且远离主要类.为此,该算法通过层次聚类在不同粒度上寻找分布紧实且距离其他数据样本较远的聚类,认为他们是可能存在的稀有类,继而向贴标者询问聚类中代表点的类别标签来确定该聚类的类别.不过,由于需要在多层次上完成聚类和分析操作,该算法的时间复杂度一般不低于 $O(Mdn^2)$,其中,M 是均值平移(mean-shift)迭代的次数.

1.2.2 基于密度变化率的方法

SEDER 算法[5]利用稀有类在主要类中出现造成的主要类中局部密度的变化来发现稀有类.为此,该算法首先利用半参密度估计方法(semi-parametric density estimation)获得数据集中各个数据样本上的密度变化率;继而选取密度变化率最大的数据样本作为稀有类的候选数据样本,并向贴标者询问其确切的类别标签.每次询问后,为了避免在已发现的类中重复的选择数据样本,询问点周边一定范围内数据样本将被加入免责列表,避免再次被选到.由于使用了半参密度估计来估算数据集各处的密度变化率,SEDER 算法的时间复杂度高达 $O(d^2n^2)$.

1.2.3 基于近邻关系的方法

CLOVER 算法[8]根据不同类型数据样本近邻关系上的区别来区分不同类型的数据.具体来说,该算法利用稀有类数据样本与奇异点数据样本之间的互 k 近邻(mutual k-nearest neighbor,简称 MkNN)个数上的差异来区分这两类数据,同时,利用稀有类数据样本与主要类数据样本 k 近邻(k-nearest neighbor,简称 kNN)之间距离上的差异来区分稀有类和主要类.由于需要计算数据样本之间的近邻关系,该算法的时间复杂度大约为 $O(n^2)$,而在使

用 kd 树[14]等 k 近邻查找技术后,其算法时间复杂度可降至为 $O(dn^{2-1/d})$.

2 相关概念

本文所提出的KRED算法是通过考察k近邻图中入度和边长的变化系数来发现数据分布的局部突变情况,从而定位稀有类的.因此,在介绍 KRED 算法之前,我们首先介绍与 k 近邻图和变化系数有关的重要概念.

定义一个 k 近邻图 $G=(V,E)$为一个加权有向图,其中每个节点 $p\in V$ 表示一个数据样本,每条边 $e\in E$ 表示边的两个端点是近邻关系,边的权值为两个数据样本间的欧氏距离.从每个节点出发,有 k 条有向边指向其 k 个距离最近的节点.

选用 k 近邻图的主要原因在于:它的结构不受数据样本的局部密度或数据样本之间绝对距离的影响,而只与各局部区域数据样本之间的相对位置关系有关,因而能够动态且较为真实地表达数据样本间的相对位置关系.由于以上优点,k 近邻图常被用于聚类问题中,用来发现数据集中不同密度区域的聚类[12,13].而在本文,则利用 k 近邻图的这一优点帮助我们分析数据集中不同密度区域数据分布的局部突变情况,从而定位发现这些密度区域中的稀有类.下面给出要使用到的与 k 近邻图有关的两个重要定义,分别是入度和边长集合.

定义 1(入度(in-degree)). 给定一个 k 近邻图 $G=(V,E)$和其中的一个节点 $p\in V$,其入度记为 $deg(p)$,等于指向节点 p 的边的数量.

对于在给定数据集上构造的 k 近邻图中,其中,不同节点的入度随节点所对应数据样本数据分布的不同而变化.具体来说,我们可以观察到以下现象:

现象 1:在数据分布局部平滑的区域,数据样本对应的节点通常具有相近的入度.

图 1 较直观地解释了这个性质.从图中我们可以观察到:数据分布局部平滑区域中,各数据样本会在相似的方向和位置上找到其 k 近邻.那么,各数据样本也会成为相似位置和方向上其他数据样本的 k 近邻,使得各样本成为其他数据样本的 k 近邻样本的次数相近.换言之,该区域中节点具有相近的入度.

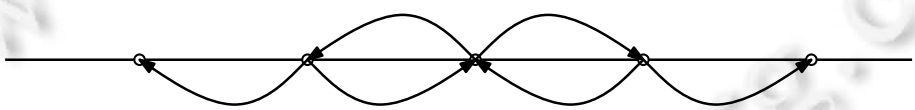


Fig.1 A 2NN graph of data points with even distribution

图 1 平滑分布的数据点的 k 近邻图(k=2)

现象 2:在数据分布局部不平滑的区域,局部密度较高的数据样本对应的节点比之周边邻接节点往往具有更高的入度.

图 2 较为直观地解释了这个性质.如图所示:左半边的节点分布密度较低,右半边的节点分布密度较高.在密度较高部分的节点,特别是高密度部分边缘的节点,其入度会比低密度一侧的节点要高.这是因为与高密度区域邻接的低密度区域的节点一般会去选择高密度区域的节点作为其近邻节点.反之,高密度区域的节点一般会在高密度区域内部选择近邻节点,而非找与其邻接的低密度区域的节点.这样一来,就会导致高密度区域的数据样本所对应的节点一般具有更高的入度.

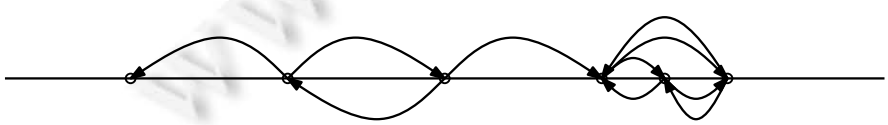


Fig.2 A 2NN graph of data points with uneven distribution

图 2 非平滑分布数据点的 k 近邻图(k=2)

定义 2(边长集合(edge length set)). 给定一个 k 近邻图 $G=(V,E)$,节点 $p\in V$ 是其中的一个节点,其边长集合记为 $EL(p)$,是以节点 p 为顶点的所有边的长度集合.

一个节点的边长集合就是在 k 近邻图中节点的出边与入边的集合.由于在 k 近邻图中每个节点都有 k 个出边,因此节点的边长集合的特性主要由入边决定.与入度的性质类似,边长集合的性质也与节点的分布情况有关.在数据分布平滑的区域,一个节点的边长集合中各个边之间的长度变化一般并不大.而在数据分布不平滑的区域,节点之间的距离往往会产生较大的变动,这也导致在这些区域中节点的边长集合中各个边之间的长度变化比较大.

另外需要注意的是:在这里,我们将一个节点的出边和入边的集合当做边长集合,而非单独考虑出边或入边.这样做的目的是为了让边长集合能够更加全面地反映节点周围密度的变化情况,而尽量避免所反映密度变化过于局部.

基于入度和边长集合的定义,我们提出变化系数的概念来衡量局部数据分布的变化程度:

定义 3(变化系数(variation coefficient)). 给定一个 k 近邻图 $G=(V,E)$和一个节点 $p\in V$,点 p 的变化系数,记为 $Vc(p)$,定义为

$$Vc(p)=\max V(p)\times std(EL(p)),$$

其中,

$$\max V(p)=\frac{\deg(p)}{\min_{q\in p\cup kNN(p)}\deg(q)},$$

$$std(EL(p))=\sqrt{\frac{\sum_{l\in EL(p)}\left(l-\frac{\sum_{\ell\in EL(p)}\ell}{|EL(p)|}\right)^2}{|EL(p)|-1}}.$$

$kNN(p)$表示 p 的 k 个最近邻近节点,l 为节点 p 的边长集合 $EL(p)$中的某一边长.

稀有类的出现,往往会导致数据样本所对应节点的局部数据分布产生变化.根据上文的定义得知,一个节点的入度和边长集合都与节点周围数据的分布情况有关.在节点的局部数据分布产生变化时,节点的入度和边长集合都会表现出不同的特性.我们通过捕捉这些特性来发现节点周围数据分布的变化程度,进而发现稀有类.

具体来说,max$V(p)$反映了一个节点与其周围节点入度的差别.根据现象 1 和现象 2 可知:在数据分布平滑的区域,节点之间的入度差异并不大;而在数据分布产生变化的区域,节点之间的入度往往会产生比较大的变化.max$V(p)$的作用就是为了发现由于数据分布的变化导致的节点入度变化.$std(EL(p))$反映了节点边长集合的标准差.根据边长集合的定义,数据分布均匀的区域,节点的边长集合的标准差一般会比较小;而数据分布产生变化的区域,节点的边长集合的标准差会比较大.$std(EL(p))$的作用,就是发现由于数据分布变化而导致的节点边长集合的标准差的变化.通过将两者组合起来,得到变化系数 Vc,我们可以有效衡量一个节点周围局部数据分布的变化程度.

3 KRED 算法描述

在稀有类检测问题中不难发现:相对于紧密成簇的稀有类,样本数量占绝大多数的主要类通常广泛分布于特征空间中.与现有文献[1−5]的假设类似,我们假设主要类的数据样本是局部平滑分布的.基于这个假设,若稀有类在主要类中的某一局部区域集中出现时,则它们边界区域会产生明显的数据分布的变化.因此,通过发现和定位这些数据分布上明显变化,则可以有较大概率检测到稀有类.

本文所提出的 KRED 算法正是使用变化系数 Vc 来测定局部数据分布的变化程度,以完成稀有类检测.该系数基于 k 近邻图得出,能够帮助我们评估节点入度和边长的变化情况.通过采用自动选取 k 值的方法来找到合适的 k 值用于变化系数 Vc 的计算,将 k 值带入 KRED 算法中得出每个数据节点的 Vc 值;然后,选取有最大 Vc 值的数据样本供专家进行标注,从而找出稀有类.我们的目标是:通过尽可能少的问询次数,为所有类别找到至少一个数据样本.

在接下来的介绍中,我们首先对将要用到的符号进行定义,然后介绍为 k 近邻图自动选取 k 值的方法,给出在 k 近邻图上完成稀有类检测的步骤,最后分析整个 KRED 算法的时间复杂度.

3.1 符号定义

我们将一个未标注数据样本类别标签数据集记作 $S=\{x_1,x_2,...,x_n\}$,其第 i 个数据样本记作 $x_i=(x_{i1},x_{i2},...,x_{id})$.这里,$1\leqslant i\leqslant n$,$d$ 为数据集维度.这些数据样本分别来自于 m 个不同类,类别标签记为 $y_i\in\{1,2,...,m\}$,其中包含主要类和稀有类.表 1 列出了本文中运用的所有符号.

Table 1 Symbols and descriptions

表 1 本文运用符号及其描述

符号	描述
S	原始未标注数据集
n	S 中的数据样本总量
d	S 特征空间的维数
m	S 中样本类别总数
x_i	S 中第 i 个数据样本
x_{ij}	x_i 的第 j 个特征
k	用于构建 k 近邻图的 k 值
cov	S 的协方差矩阵
eig	cov 的特征向量

3.2 k值的自动选取

在构建 k 近邻图时,k 值的选取是至关重要的,直接影响 KRED 算法在分析入度和边长集合时的覆盖范围,从而决定变化系数 Vc 的大小.为了反映局部区域数据分布的变化,首先,k 值的选择不能过大,过大则不利于分析发现稀有类(本身数据样本就很少)边界所在小范围局部区域数据分布变化;相反,k 值选择也不宜太小,太小则会使得各数据样本上只有极少的边(如 k 取 1 时,部分数据样本可能只有一条边),因而会导致各数据样本的边长集合过小,不利于通过分析边长之间标准差来反映和分析真实的局部数据分布变化.

一个好的 k 值应当:(1) 尽可能地反应稀有类和主要类边界所在小范围局部区域数据分布变化,使得该区域的变化系数 Vc 的值足够大;(2) 同时,又尽可能地使得主要类内部其他数据样本(即,数据分布局部平滑的区域)的变化系数 Vc 的值尽量小.这样才能凸显稀有类边界区域的 Vc 值,从而帮助发现稀有类.基于上一段中的讨论,第 1 个要求可以通过选取一个尽可能小的 k 值来满足.而如何满足第 2 个要求,我们有以下讨论:根据现象 1,在主要类内部数据平滑分布的区域,每个数据节点往往具有大致相同的入度.这意味着无论 k 的取值大小,这些区域的 $\max V(p)$均趋近于 1.因此,当局部数据分布平滑时,Vc 的值大小主要取决于各点边长集合标准差 $std(EL(p))$,且为了满足第 2 个要求,我们所选的 k 值应当使得此时的 $std(EL(p))$尽量小.

为了达到以上效果,一种直观且行之有效的方法是:找到数据样本的主要分布方向,并在该方向上选取各数据样本最近的点对作为各点的 k 近邻.也就是说:如果数据样本主要分布方向是一维的(其他非主要方向或维度对点对间距离计算基本无贡献或者贡献很小),例如图 3 所示,则可令 $k=2\times1$,即,选取该方向各点前后两个点(如对 p_3 来说,选取 p_2 和 p_4)作为各点 k 近邻,而如若引入过多的 k 近邻则失去考察局部区域数据分布变化的意义,且会使得边长集合标准差 $std(EL(p))$有增大的趋势;同理,如果数据样本主要分布方向是二维的(两个主要分布方向或维度之间是相互垂直的),则可令 $k=2\times2$,即,选取各点在这两个方向前后左右 4 个点作为各点 k 近邻;类似地,推广到 d 维数据集上,亦只需找到数据样本主要分布方向,然后可令 $k=2\times c$,这里,$c\leqslant d$ 为主要分布方向的维数.

为了找到 d 维数据集中数据样本的主要分布方向,一种被广泛使用的方法[8,19]是对数据集做主成分分析(principal component analysis,简称 PCA).PCA 的主要过程分为两步:计算给定数据集的协方差矩阵;计算该协方差矩阵的特征值集合.特征值集合中,各特征值的大小体现着各数据分布维度对原始数据集的表达能力,表明各数据分布维度是否为主成分,也即是否为数据样本的主要分布方向.

综上,我们的自动选取 k 值方法的具体步骤可以总结为:

1) 计算给定数据集 S 的协方差矩阵 cov;

2) 计算 cov 的特征值集合 eig;

3) 令 K=2,对 *eig* 采取 K-mean 聚类以将 *eig* 自动划分为特征值较大和较小的两部分;

4) 选取聚类结果中特征值较大的部分,统计其中特征值的个数 c;

5) 返回 k=2×c.

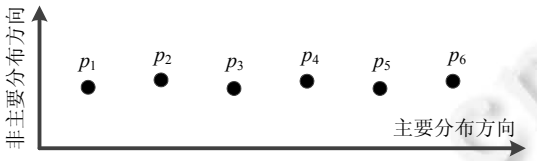


Fig.3 An example of k selection in locally even data distribution

图 3 局部数据分布平滑时选 k 近邻示例图

3.3 KRED算法

为了评估局部数据分布的变化程度,我们提出了基于 k 近邻图的变化系数 Vc 作为衡量标准.而 KRED 算法(见算法 1)的核心思想是:将数据集转化为 k 近邻图,然后,通过计算变化系数 Vc 找到局部数据分布变化最明显的区域从而发现稀有类.该算法以一个未标注数据集 S 和在该数据集上自动选取的 k 为输入,并输出其选取的可能来自稀有类的数据样本及它们真实类别标签.

算法 1. 基于 k 近邻图的稀有类检测算法(KRED).

输入:未标注数据集 S,自动选取的 k 值;

输出:所选数据样本集合 I、所选数据样本真实类别标签集合 L.

1: 构造 S 的 k 近邻图;

2: $\forall x_i \in S$,根据定义 3 计算 $Vc(x_i)$;

3: **while** 可用询问次数大于 0 **do**

4: 询问 $x=\text{argmax}_{x\in S}(Vc(x))$的类别标签 y_x;

5: $I=I\cup x$, $L=L\cup y_x$;

6: $\forall p\in S$,若 p 与 x 是近邻关系,即,有邻边存在,则令 $Vc(p)=-\infty$;

7: **end while**

KRED 算法具体步骤如下:首先,在数据集 S 上构造 k 近邻图;继而计算每个数据样本的 Vc 值;然后,在步骤 3~步骤 7 的循环中,算法将找出当前具有最大 Vc 的数据样本 x,并向专家询问其真实类别标签;同时,为了避免在同一区域重复选取数据样本,与 x 是近邻关系的数据点的 Vc 值将被置为$-\infty$;当可用询问次数未用完时,可继续询问余下数据样本中 Vc 值最大者的真实类别标签;否则循环结束,算法停止.

3.4 复杂度分析

在自动选取 k 的过程中,计算协方差矩阵和其特征值分别需要 $O(nd^2)$和 $O(d^3)$,其中,n 为数据样本数量,d 为数据维度;K-mean 聚类算法需要 $O(2dt)$,其中,t 是算法的迭代次数;通过 *kd* 树[14],我们可以在 $O(dn^{2-1/d})$的时间复杂度内构造出 k 近邻图并完成 k 近邻查找;最后,计算数据点的 Vc 值需要 $O(nk)$的时间,且 $k\leqslant 2d<<n$.综上,KRED 算法的时间复杂度为 $O(dn^{2-1/d})$.与现有的无先验稀有类检测算法相比,我们的算法时间复杂度显著低于 HMS 算法和 SEDER 算法,同 CLOVER 算法的时间复杂度大致持平.而实验结果表明:KRED 算法在运行时间上优于以上算法,且在稀有类检测准确率上具有优势.

4 实验结果与分析

首先介绍实验用到的数据集,然后分别从 KRED 算法的稀有类检测准确率和运行时间、k 值选取对 KRED 算法结果的影响等方面评估 KRED 算法.实验结果证明:KRED 算法在稀有类检测性能和时间效率上优于现有的无先验稀有类检测算法,且自动选取的 k 值能够帮助 KRED 算法有效地提高稀有类检测性能.

4.1 实验设置

我们运用了 UCI 数据库[15]中 8 个数据集来测试我们的算法,分别是 Glass,Ecoli,Statlog,Yeast,Abalone,Shuttle,Wine Quality 和 Page Block.其中,Statlog 是一个子样本集合,它的数据来自其他所有集合.子样本集 Statlog 能够很好地模拟稀有类数据集在现实数据中的情况,即,构造出一个不平衡数据集.按照文献[7]中的标准,我们调整了 Statlog 数据集,将其中的最大类设置为具有 256 个样本,其他类的大小依次减半,直至最小类含有 8 个数据样本.表 2 详细说明了这些数据集的有关特征,其中,n 是数据样本个数,d 是维数,m 是类别个数,Largest 和 Smallest 分别代表最大类和最小类占该数据集的比例大小.同时,为不失一般性,以上数据集都做了标准化处理[8],使得数据样本在各个维度上均值为 0,方差为 1.此外,本文的算法编写和编译是在 MATLAB7.9 中实现,实验环境为 Intel Core 2 Quo 2.8 GHz CPU,2GB 内存.

Table 2 Properties of the real data sets

表 2 真实数据集的相关属性

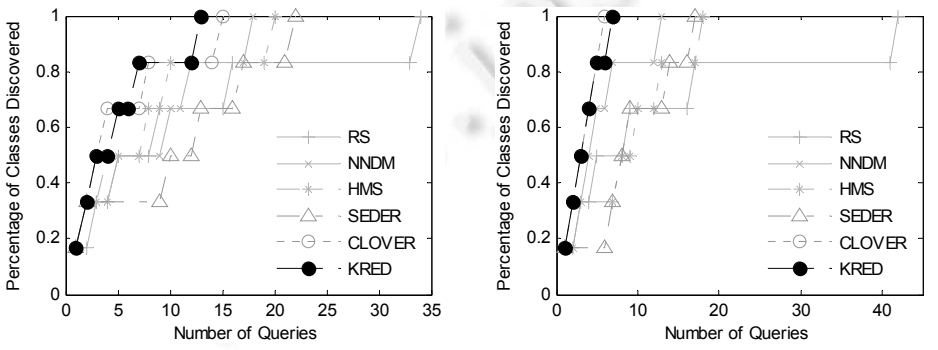
Data set	n	d	m	Largest (%)	Smallest (%)
Glass	214	9	6	35.51	4.21
Ecoli	336	7	6	42.56	2.68
Statlog	512	19	7	50.00	1.56
Yeast	1481	8	10	31.68	0.33
Abalone	4177	7	20	16.50	0.34
Shuttle	4515	9	7	75.53	0.13
Wine quality	4898	11	6	44.88	0.41
Page block	5473	10	5	89.77	0.51

4.2 KRED算法性能评估

本节中,我们将 KRED 算法和现有的基于先验知识的稀有类检测算法 NNDM[3]、无先验知识的稀有类检测算法 HMS[5]、SEDER[7]、CLOVER[8]以及随机采样方法(random sampling,简称 RS)在 8 个测试数据集上进行比较,以证明 KRED 算法能够利用尽量少的询问次数发现数据集中的所有类,且拥有较高的时间效率.

(1) 图 4 表现了各算法每从数据集中发现一个新类时所需要询问次数.从图中可以看出:KRED 发现测试数据集中所有类所需要的询问次数明显少于 NNDM 算法、HMS 算法和 SEDER 算法的询问次数,与 CLOVER 算法的询问次数基本持平,且在大多数情况下略少于 CLOVER 算法.

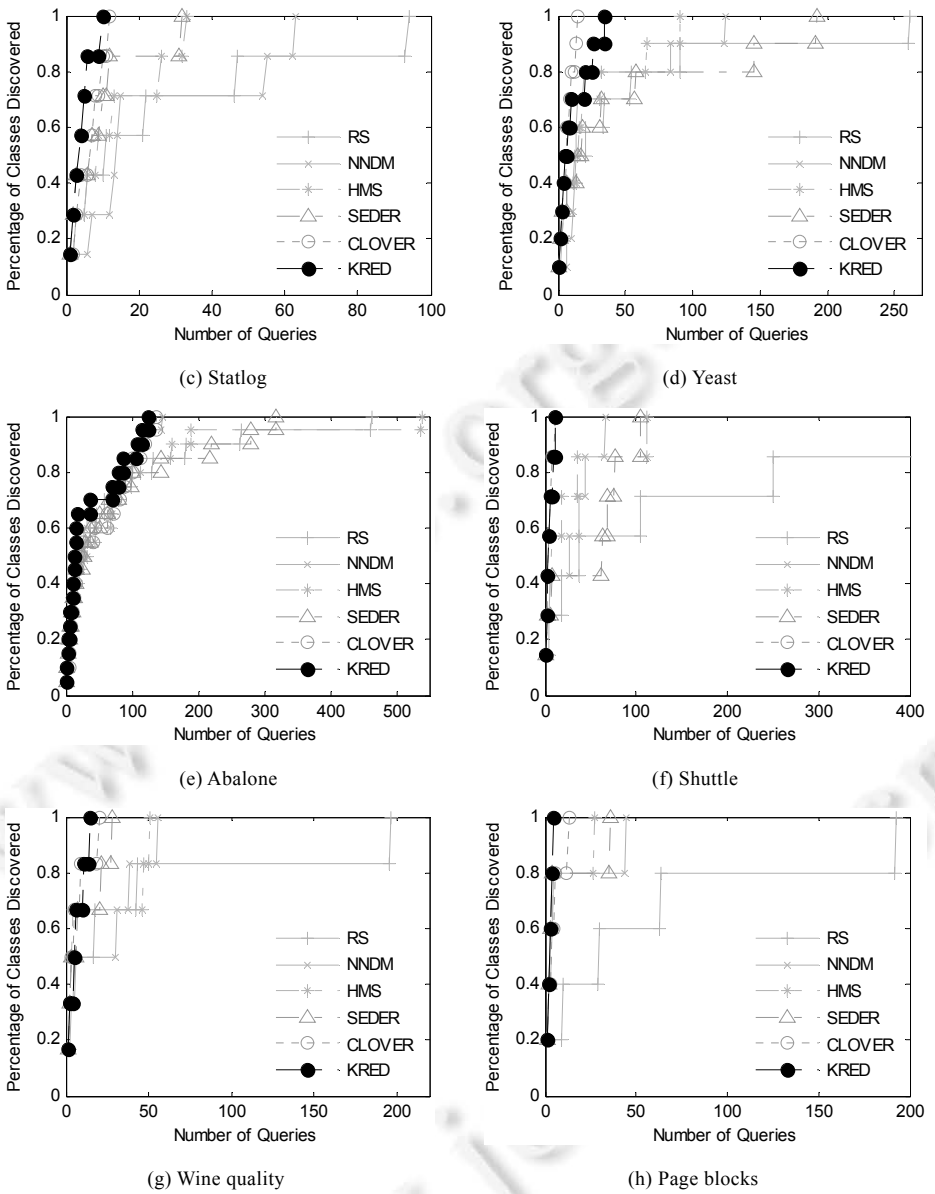
(2) 我们记录了各个算法的运行时间,见表 3,其中,数值单位为秒.因为随机采样方法没有对数据进行分析,因此,这里并未将该方法列在时间复杂度表中进行比较.从表 3 中我们发现:KRED 算法的时间开销随着数据样本数量的增加而增加,但是仍然远小于 SEDER 和 HMS 算法,也略优于 NNDM 算法和 CLOVER 算法.



(a) Glass (b) Ecoli

Fig.4 Performance comparison results on real data sets

图 4 算法在真实数据集上的性能结果比较



(c) Statlog (d) Yeast

(e) Abalone (f) Shuttle

(g) Wine quality (h) Page blocks

Fig.4 Performance comparison results on real data sets (Contiuned)

图 4 算法在真实数据集上的性能结果比较(续)

Table 3 Runtime performance of each algorithm (s)

表 3 算法运行时间比较 (s)

Data set	NNDM	HMS	SEDER	CLOVER	KRED
Glass	0.09	4.54	0.71	0.16	0.06
Ecoli	0.11	9.98	1.04	0.18	0.09
Statlog	0.25	30.85	13.48	0.49	0.19
Yeast	5.18	284.33	25.42	1.35	0.63
Abalone	42.71	3789.75	166.98	7.26	3.55
Shuttle	26.34	2572.93	292.35	7.91	3.91
Wine quality	19.88	8124.98	515.76	10.21	4.77
Page block	15.61	7828.76	524.22	13.05	5.72

4.3 k值选取的影响

通过实验可以观察到,k 值的选取对计算 Vc 有十分显著的影响.图 5 记录了当选取不同的 k 值(k=1, 2,…,10)以及其对应的询问次数变化情况.这里,我们设置 k 值最大为 10,原因如下:首先,检测局部数据分布情况不需要考虑太大的数据范围,即,不需要构建太大的 kNN 图;其次,由于真实数据集中主成分分析的结果一般为 2 或 3,因此根据自动选取 k 的算法,我们得出的 k 不会大于 10.图 5 中,x 轴代表 k 值大小,y 轴代表检测出所有类别需要的询问次数,实心点表示我们为该数据集自动选出的 k 值.值得注意的是,同其他的 k 值相比,算法得到的 k 一般会使 KRED 算法的具有更好的效果,尽管在 Yeast 数据集中,k=1 似乎会产生最好的结果.

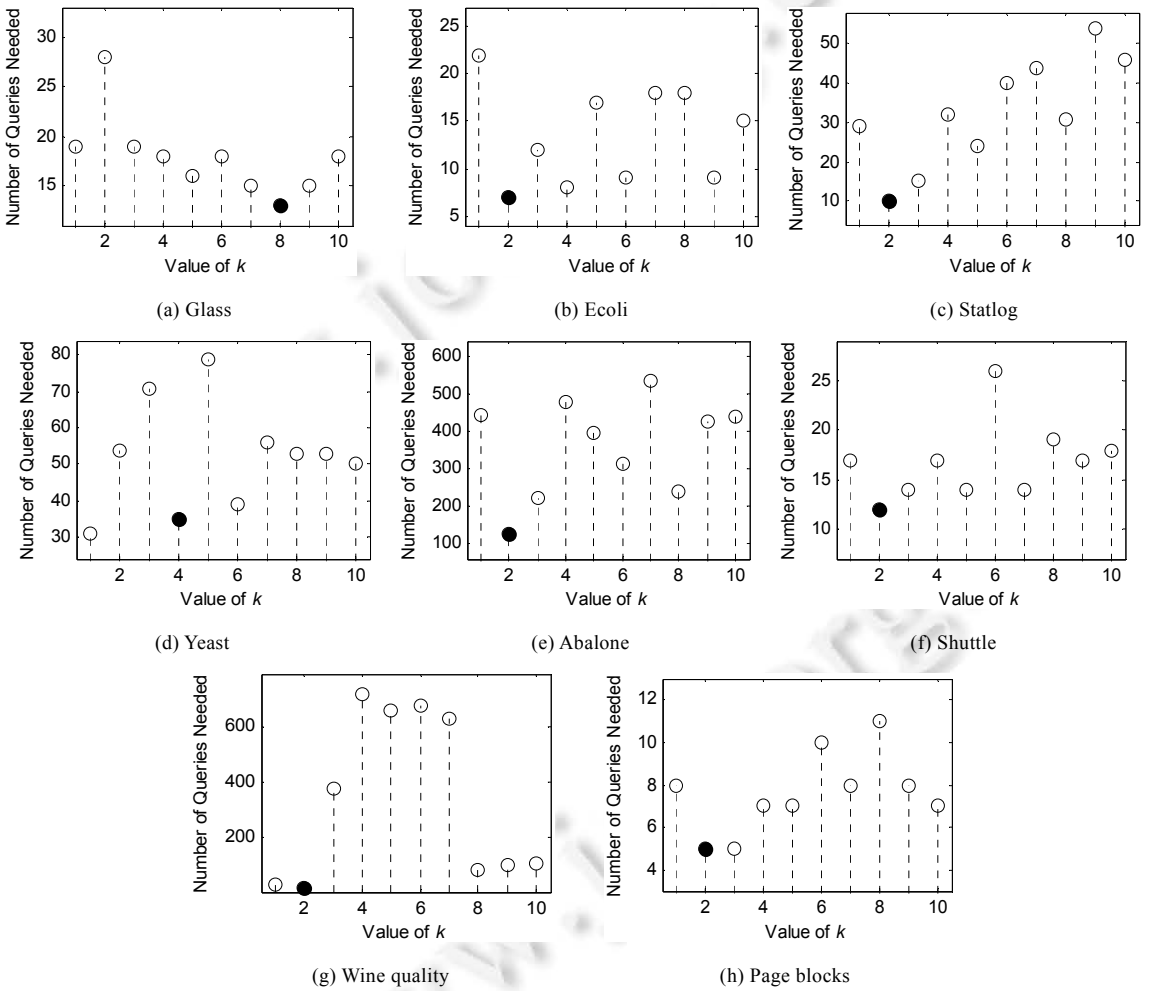


Fig.5 KRED performance vs. varying k

图 5 k 值对 KRED 的影响

5 结束语

本文利用检测数据集中数据样本分布的局部突变的方法来进行稀有类检测.在自动选取 k 值并构建出 k 近邻图后,通过变化系数 Vc 来衡量数据样本分布的变化情况,并选出具有最大 Vc 的数据点,询问其类别标签来发现稀有类的数据样本.

与以前的无先验稀有类检测方法相比,KRED 方法效率更高,算法时间开销较低.此外,通过自动选取 k 值的

方法,我们有效地提高了数据集中各个类的发现效率,并显著减少了发现数据集中全部类所需要的问询次数.通过在大量真实数据集上进行对比实验,我们证明 KRED 是十分有效的稀有类检测方法.

本文所提的 KRED 算法在考虑全部维度的情况下有较好的检测效果,但是现实数据往往存在这样的情况:数据仅在部分维度下呈紧密聚集状态,对于这样的子空间内稀有类的检测,KRED 算法还不能达到较好的效果.因此,我们下一步的工作将结合数据的特征维度选取与对数据本身类别的考量来改进算法,使之成为能检测出部分维度上稀有类的算法.

References:

[1] Pelleg D, Moore A. Active learning for anomaly and rare-category detection. In: Proc. of the NIPS 2004. 2004. 1073−1080. http://papers.nips.cc/paper/2554-active-learning-for-anomaly-and-rare-category-detection.pdf

[2] Huang H, He QM, He JF, Ma LH. RADAR: Rare category detection via computation of boundary degree. In: Proc. of the PAKDD 2011. 2011. 258−269. [doi: 10.1007/978-3-642-20847-8_22]

[3] He JR, Carbonell J. Nearest-Neighbor-Based active learning for rare category detection. In: Proc. of the NIPS 2007. 2007. 633−640. http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2007_51.pdf

[4] He JR, Liu Y, Lawrence R. Graph-Based rare category detection. In: Proc. of the ICDM 2008. 2008. 833−838. [doi: 10.1109/ICDM.2008.122]

[5] He JR, Carbonell J. Prior-Free rare category detection. In: Proc. of the SDM 2009. 2009. 155−163. [doi: 10.1137/1. 9781611972795.14]

[6] He JR, Tong HH, Carbonell J. Rare category characterization. In: Proc. of the ICDM 2010. 2010. 226−235. [doi: 10.1109/ICDM.2010.154]

[7] Vatturi P, Wong WK. Category detection using hierarchical mean shift. In: Proc. of the KDD 2009. 2009. 847−856. [doi: 10.1145/1557019.1557112]

[8] Huang H, He QM, He JF, Ma LH. CLOVER: A faster prior-free approach to rare-category detection. Knowledge and Information Systems, 2013,35(3):713−736. [doi: 10.1007/s10115-012-0530-9]

[9] Huang H, Wang SP, Ma LH. An enhanced category detection based on active learning. In: Proc. of the ISKE 2010. 2010. 224−227. [doi: 10.1109/ISKE.2010.5680880]

[10] Blum A, Mitchell T. Combining labeled and unlabeled data with co-train. In: Proc. of the COLT '98. 1998. 92−100. [doi: 10.1145/279943.279962]

[11] Jain P, Kapoor A. Active learning for large multi-class problems. In: Proc. of the CVPR 2009. 2009. 762−769. [doi: 10.1109/CVPR.2009.5206651]

[12] Karypis G, Han EH, Kumar V. CHAMELEON: Hierarchical clustering using dynamic modeling. Computer, 1999,32(8):68−75. [doi: 10.1109/2.781637]

[13] Franti P, Virmajoki O, Hautamaki V. Fast agglomerative clustering using a *k*-nearest neighbor graph. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2006,28(11):1875−1881. [doi: 10.1109/TPAMI.2006.227]

[14] Moore A. A tutorial on kd-trees. University of Cambridge Computer Laboratory Technical Report, 1991. http://www.autonlab.org/autonweb/documents/papers/moore-tutorial.pdf

[15] Frank A, Asuncion A. UCI machine learning repository. 2010. http://archive.ics.uci.edu/ml

[16] Bay S, Kumaraswamy K, Anderle M, Kumar R, Steier D. Large scale detection of irregularities in accounting data. In: Proc. of the ICDM 2006. 2006. 75−86. [doi: 10.1109/ICDM.2006.93]

[17] Stokes J, Platt J, Kravis J, Shilman M. Aladin: Active learning of anomalies to detect intrusions. Microsoft Research Technical Report, 2008. http://research.microsoft.com/en-us/um/people/jstokes/aladintechreport.pdf

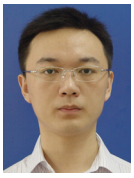
[18] Huang H, He QM, Chen Q, Qian F, He JF, Ma LH. CATION: Rare category detection algorithm based on weighted boundary degree. Ruan Jian Xue Bao/Journal of Software, 2012,23(5):1195−1206 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4104.htm [doi: 10.3724/SP.J.1001.2012.04104]

[19] Huang H, Kevin Chiew, Gao YJ, He QM, Li Q. Rare category exploration. Expert System with Applications, 2014,41(9):4197−4210. [doi: 10.1016/j.eswa.2013.12.039]

附中文参考文献:

[18] 黄浩,何钦铭,陈奇,钱烽,何江峰,马连航.基于加权边界度的稀有类检测算法.软件学报,2012,23(5):1195−1206. http://www.jos.org.cn/1000-9825/4104.htm [doi:10.3724/SP.J.1001.2012.04104]

王淞(1991−),男,湖北武汉人,博士生,主要研究领域为数据挖掘.



黄浩(1986−),男,博士,副教授,CCF 会员,主要研究领域为数据挖掘.

余果(1986−),女,硕士,主要研究领域为数据管理与分析.

梁楠(1993−),男,本科生,主要研究领域为数据挖掘.

王黎维(1981−),女,博士,副教授,主要研究领域为数据质量,数据溯源,科学工作流.

孙月明(1992−),女,硕士生,主要研究领域为数据挖掘.