

- 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 $k=2 \times c$.

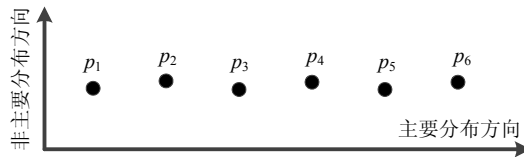


Fig.3 An example of k selection in locally even data distribution

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 V_c 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 V_c 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 $V_c(x_i)$;
- 3: **while** 可用询问次数大于 0 **do**
- 4: 询问 $x = \operatorname{argmax}_{x \in S}(V_c(x))$ 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 $V_c(p) = -\infty$;
- 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 V_c 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 V_c 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 V_c 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 V_c 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 V_c 值需要 $O(nk)$ 的时间,且 $k \leq 2d \ll n$. 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 Duo 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.

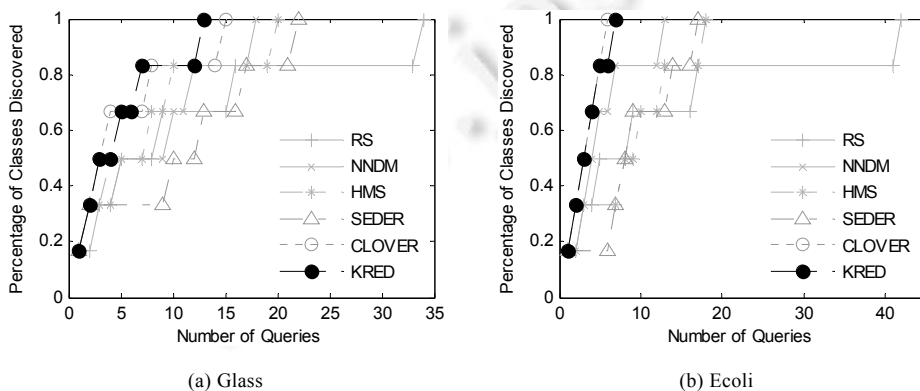


Fig.4 Performance comparison results on real data sets

图 4 算法在真实数据集上的性能结果比较

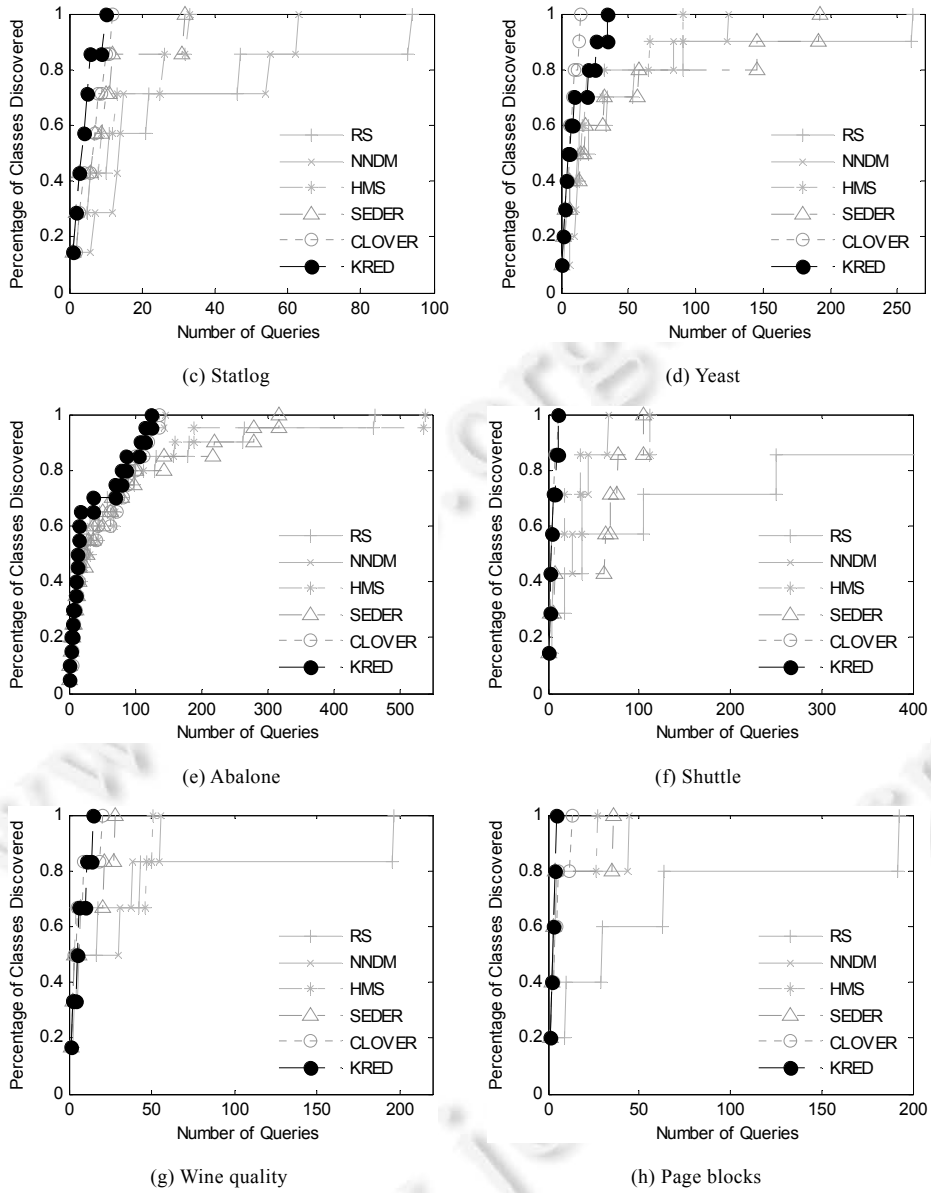


Fig.4 Performance comparison results on real data sets (Continued)

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 V_c 有十分显著的影响.图 5 记录了当选取不同的 k 值($k=1, 2, \dots, 10$) 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 k NN 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.

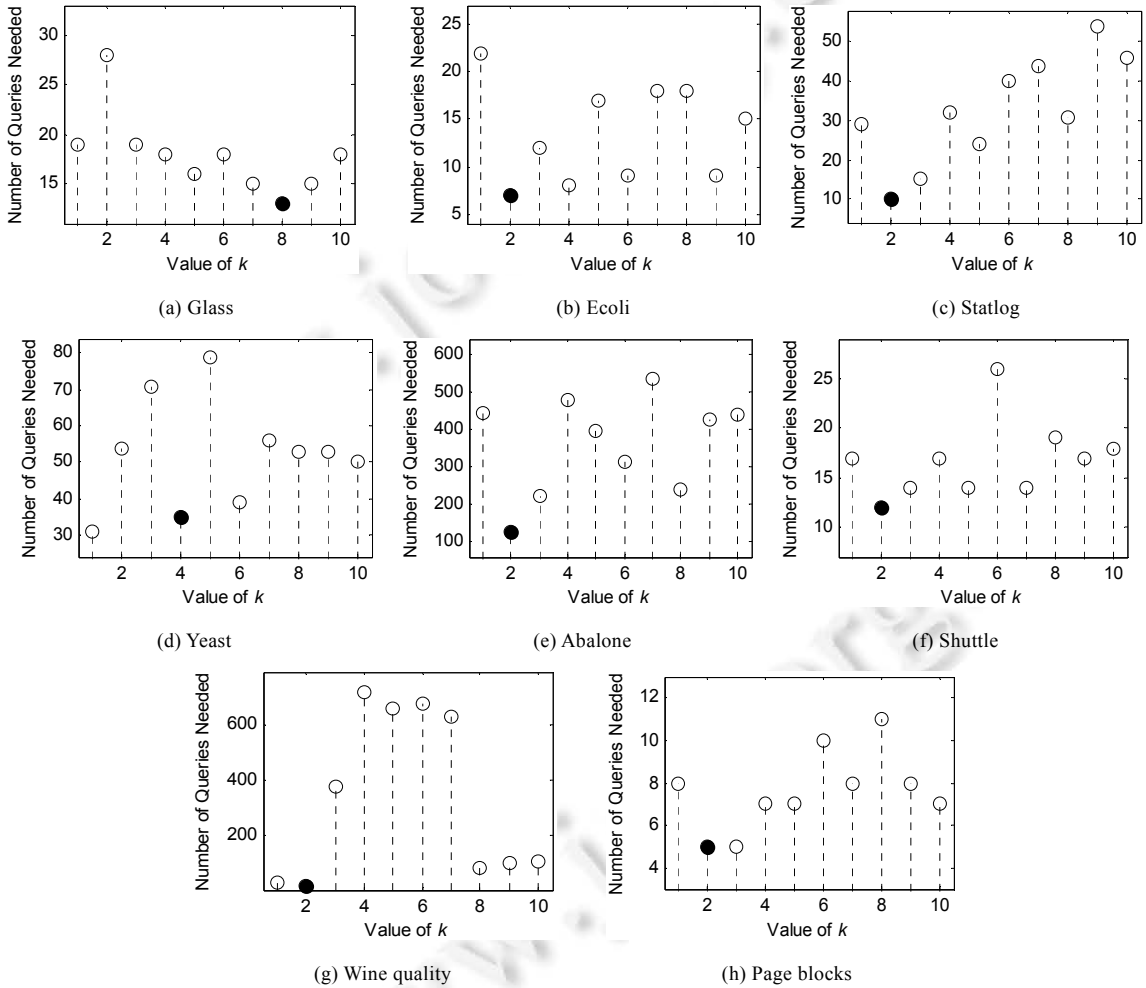


Fig.5 KRED performance vs. varying k

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 V_c 来衡量数据样本分布的变化情况,并选出具有最大 V_c 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, Virtajoki 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:

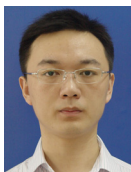
- [18] 黄浩,何钦铭,陈奇,钱烽,何江峰,马连航.基于加权边界度的稀有类检测算法.软件学报,2012,23(5):1195-1206. <http://www.jos.org.cn/1000-9825/4104.htm> [doi:10.3724/SP.J.1001.2012.04104]



王淞(1991-),男,湖北武汉人,博士生,主要研究领域为数据挖掘.



梁楠(1993-),男,本科生,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.



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 workflows.



余果(1986-),女,硕士,主要研究领域为数据管理与分析.



孙月明(1992-),女,硕士生,主要研究领域为数据挖掘.

www.jos.org.cn