

一种基于自然最近邻的离群检测算法

朱庆生 唐 汇 冯 骥

(重庆大学计算机学院软件理论与技术重庆市重点实验室 重庆 400044)

摘 要 任何涉及 k 近邻求解问题的算法被应用于处理不同特征的数据集时,参数 k 值的选择都会明显影响算法的性能和结果。因而,如何选择 k 近邻算法中敏感参数 k 值一直是一个研究难点。提出了一种新的近邻关系——自然最近邻,它不需要设置参数 k ,每个节点的邻居是由算法自适应计算而形成的。针对离群点检测的特殊性,通过确定自然最近邻搜索算法的终止条件,提出一种基于自然最近邻的新的离群检测算法 ODb3N。实验表明,该算法不仅避免了 k 近邻中参数的选择问题,而且能够更有效地发现离群簇。

关键词 k 近邻,自然最近邻,离群检测,离群簇

中图分类号 TP301/TP391

文献标识码 A

Outlier Detection Algorithm Based on Natural Nearest Neighbor

ZHU Qing-sheng TANG Hui FENG Ji

(Key Lab of Software and Technology, College of Computer Science, Chongqing University, Chongqing 400044, China)

Abstract When the k -nearest neighbor method is used, it is difficult to choose an appropriate parameter k of the algorithm which affects obviously its efficiency and performance. Natural nearest neighbor proposed by us is a novel concept on nearest neighbor, in which each point's neighbors are formed by an adaptive algorithm without parameter k . In this paper, we proposed an outlier detection algorithm based on natural nearest neighbor (ODb3N) by means of modifying iteration stop condition. The experiments show that our method not only has the advantage of non-parameter, but also has the ability to discover both the outlier and the cluster of outliers.

Keywords k -nearest neighbor, Natural nearest neighbor, Outlier detection, Cluster of outliers

1 引言

离群数据挖掘通常是依据适当的评价标准找出数据集中与其他大部分数据不一样的数据点。而这些数据所包含的信息对于识别系统中的异常行为非常有用,这种异常的数据点被称为离群点。离群点的检测是数据挖掘中一个非常重要的领域,已经被广泛应用如网络入侵检测、移动通讯欺诈行为分析、新的疾病症状发现、天气预测等领域^[1]。

在离群检测研究方面,人们提出了很多算法。由于基于距离的算法无法识别数据集中的局部离群点, Breunig 提出一种基于密度的 LOF 算法^[2], 它把每个点的密度和它 Mints 邻域的平均密度比值作为该点的局部离群因子, 离群因子的值越大则离群程度越高。 Hautamaki 提出的 ODIN 算法^[3] 首先建立数据集的 KNN 图, 然后通过计算每个点的入度数检测离群点。 Angiulli^[4] 提出一种适用于定性属性的离群属性子空间获取的算法, 该算法能发现、解释在整个原始数据集上离群的数据对象。多数聚类的实现都设定了数据空间上的一种度量或相似性结构, 聚类算法的性能依赖于如何选择一种恰当的度量^[5]。

在利用 k NN 解决实际问题的研究方面, 也有很多研究成

果^[6]。哈工大 Yong Xu 等提出一种有效构造 K 近邻分类器的粗糙方法^[7]; 国防科大符永铨等提出分布式 KNN 搜索方法为网络服务器目标节点发现 k 个邻居以组织多级环^[8]。然而, 任何涉及 k 近邻的算法, 参数 k 值的选择都会明显影响算法产生的实际性能和检测结果。即使是采用同一个算法, 由于被处理的数据集特征不同, k 值选择也没有可借鉴性, 通常 k 值的选择都是依靠用户经验和大量实验来决定。我们提出自然最近邻的概念可以回避参数选择困难的问题。基于自然邻的基本思想并结合离群数据的特征, 本文进一步讨论了自然邻搜索算法的终止条件, 提出一种无参数的基于自然最近邻的离群检测算法。

2 自然最近邻

自然最近邻是我们提出的一种可以自适应形成的新的邻居关系, 其每一个数据节点的自然邻居的个数是由算法自适应计算产生的, 根据自然规律, 分布稠密的点拥有较多的自然邻居, 而分布稀疏的点拥有较少的自然邻居。

假设数据集 $D = \{X_1, X_2, \dots, X_n\}$, 任意点 $\forall i \in D$ 。

定义 1 (最近邻居) $NN_r(i)$ 为点 i 的 r 最近邻居, 其中 r 的值为算法自动产生的, 满足 $|NN_r(i)| = r$ 。

到稿日期: 2013-05-21 返修日期: 2013-09-02 本文受国家自然科学基金(61272194, 61073058)资助。

朱庆生 教授, 博士生导师, 主要研究方向为面向服务的软件工程、离群检测与数据挖掘, E-mail: qszhu@cqu.edu.cn; 唐 汇 硕士生, 主要研究方向为离群检测与释义空间分析; 冯 骥 博士生, 主要研究方向为自然最近邻特性分析与应用。

定义 2(逆邻居) $RNN_r(i)$ 表示点 i 的逆最近邻居:

$$RNN_r(i) = \{j \in D | i \in NN_r(j), i \neq j\}$$

定义 3(自然最近邻) $NNN(i)$ 表示点 i 的自然最近邻:

$$NNN(i) = \{j \in D | j \in NN_r(i), j \in RNN_r(i)\}$$

由以上定义可以知,自然最近邻是一种对称的邻居关系,其中参数 r 的值只是在自然邻居搜索算法中的一个中间变量。我们最初给出的自然最近邻搜索算法的描述是以所有点至少出现在其它点的最近邻域 1 次为终止条件,如果这样计算,当存在一个远离数据集的离群点时,其算法搜索次数会极大增长,最坏情况是遍历搜索每个点的 $N-1$ 个邻居时才会结束。

针对离群检测应用研究的实际需要,本文分析并确定自然最近邻搜索算法的终止条件,当发现数据集中孤立点(即没有反向邻居的点)的个数不再变化时算法结束。实验证明,该算法可以形成一种非常接近于样本真实分布的邻域关系,并且同时具备了聚类能力,每个点和它的自然邻居属于同一个类别,不同类别的点之间不存在自然邻居关系,连接每个点的自然最近邻居,可以形成数据集上的自然最近邻域图,每个连通分支对应着一种聚类。

自然最近邻搜索算法的过程是从 $k=1$ 开始,依次选择每个数据点的前 k 最近邻居,然后记录每个数据点在其它点的 k 最近邻居中出现的次数,令 $k=k+1$,重复上一步直到数据集中未出现在其它点的邻域中的点的个数不再减少时结束,这样就能通过算法自适应形成每个点的最近邻居。自然最近邻搜索算法的具体过程如算法 1 所描述。

算法 1 自然最近邻搜索算法

输入:数据集 X

输出:集合 NNN 为集合 X 中每个点的自然最近邻域

$/* isStable(nb)$ 表示判断数组 nb 中 0 的个数是否稳定,如 0 的个数与上一步骤相比没有变化则返回 true,否则返回值为 false。get-
 $NNr(i, r)$ 函数为获取点 i 的第 r 个邻居。

算法描述:

1. $r=1, flag=0;$
2. $\forall i \in X, nb(i)=0, NN_r(i)=\emptyset$
 $RNN_r(i)=\emptyset, NNN(i)=\emptyset$
3. While ($flag==0$)
4. If $isStable(nb)$ Then $flag=1$
5. Else for $\forall i \in X$
6. $nnr(i)=getNN_r(i, r)$
7. $nb(nnr(i))=nb(nnr(i))+1$
8. $NN_r(i)=NN_r(i) \cup nnr(i)$
9. $RNN_r(nnr(i))=RNN_r(nnr(i)) \cup i$
10. End for
11. $r=r+1$
12. End if
13. End while
14. For $\forall i \in X$
15. $NNN(i)=NN_r(i) \cap RNN_r(i)$
16. End for

对于图 1 中的数据集,首先寻找每个点的第 1 个最近邻,如图 1(a)所示,箭头指向该点的第 1 个最近邻居点,记录下每个点出现在其它点的邻域中的次数。经过统计可知此时有 3 个点没有出现在其它点的邻域中,这 3 个点应于有向图 1(a)中的 3 个入度为 0 的点;继续寻找每个点的第 2 最近

邻,从图 1(b)中可以明显看出,入度为 0 的点还剩 1 个;继续寻找每个点的第 3 最近邻居,如图 1(c)所示,统计得到仍然有 1 个入度为 0 的点,由于此时入度为 0 的点的个数与上一步骤相等,满足了算法的终止条件,因此认为整个邻域保持稳定,算法结束。到此,自适应形成的 r 值为 3,对应的 3 最近邻域如图 1(c)所示。其中互为近邻关系的点则为对方的自然最近邻居,对应图中具有双向连接的点则互为对方的自然最近邻居。连接每个点的自然最近邻居形成的自然最近邻域图如图 1(d)所示。

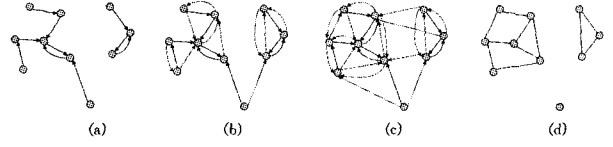


图 1 自然最近邻居的形成过程

通过观察图 1(d)可以发现,自然最近邻域图拥有良好的聚类能力,属于同一类的任意两点之间是连通的,不同类别的点之间是没有边相连的,因此每一个连通分支对应着一种聚类,孤立点则对应为数据集中的离群点。

3 基于自然最近邻的离群检测算法

多数现有算法都针对于离群点的检测,当数据集中存在离群簇时,这些算法的效果会受到很大影响。为了检测出离群簇,借鉴大小聚类和离群因子的概念。每个点的离群因子是由它与最近的大聚类的距离和它所在聚类的数据个数决定的。

假设数据集 $D = \{X_1, X_2, \dots, X_n\}$, 任意点 $\forall p \in D$, 设 $NNN(p)$ 为点 p 的自然最近邻居,令数据集 D 每一条记录作为图 G 的顶点,连接每个点与它的自然最近邻居,如此形成的图 G 则为数据集 D 对应的自然最近邻域图。

定义 4(聚类) 集合 $C = \{C_1, C_2, C_3, \dots, C_k\}$, 其中 $C_1, C_2, C_3, \dots, C_k$ 分别为图 G 的每个连通分支中的顶点集, $C_1 \cup C_2 \cup C_3 \cup \dots \cup C_k = D, C_i \cap C_j = \emptyset$ 。由于每个连通分支对应数据集中一个聚类,因此集合 C 就是数据集对应的聚类结果。

定义 5(大聚类) $LC = \{C_i | i \leq b\}, SC = \{C_j | j > b\}$, LC 表示数据集中的大聚类, SC 表示数据集中的小聚类, b 为 LC 和 SC 的边界。对集合 C 进行排序得到 $|C_1| \geq |C_2| \geq \dots \geq |C_k|$, 且边界参数 b 满足:

$$(|C_1| + |C_2| + \dots + |C_b|) \geq |D| * \alpha \quad (1)$$

$$\frac{|C_b|}{|C_{b+1}|} \geq \beta \quad (2)$$

考虑离群数据在整个数据集中的比率通常占到 20% 左右, α 的值设置为 80%, 式(1)的意义表示在有序的集合 C 中, 前 b 个聚类的数据总和大于整个数据集的 80%。式(2)表示第 b 个聚类数据总数大于第 $b+1$ 个聚类的 β 倍, 一般 β 值设置为 2。同时满足以上公式的 b 的最小值则为大小聚类的边界值, 有序集合 C 前 b 个聚类为大聚类, 剩下的则为小聚类。

定义 6(离群因子) 对于数据集中任意一点 p , $NLOF(p)$ 为点 p 的离群因子, 其中 $distance(p, C_i)$ 表示点 p 到聚类 C_i 中心的距离。

$$NLOF = \begin{cases} \frac{|D|}{|C_i|} * \min(\text{distance}(p, C_i)), p \in C_j, \\ C_j \in SC, C_i \in LC \\ \frac{|D|}{|C_i|} * \text{distance}(p, C_i), p \in C_i, C_i \in LC \end{cases} \quad (3)$$

式(3)表示每个点的离群因子是由该点所在聚类的总数和它与最近的大聚类的距离决定的,当存在2个边界比较模糊的聚类无法被算法区分的时候,同样也可以通过每个点与聚类中心的距离来识别离群点。算法2描述了具体的步骤。

算法2 基于自然最近邻的离群检测算法 ODb3N

输入:数据集 X

输出:top-n 离群点

算法步骤:

- Step 1 使用改进后的自然最近邻居算法得到每个点对应的自然最近邻域。
- Step 2 连接每个点与它的自然最近邻域,构造出数据集 X 对应的自然最近邻域图 G。
- Step 3 采用广度优先遍历图 G 得到 k 个连通分量,根据每个分量所包含顶点个数由大到小排序 $C = \{C_1, C_2, C_3, \dots, C_k\}$ 。
- Step 4 根据大小聚类定义把集合 C 分为大聚类集合 $LC = \{C_i | i \leq b\}$ 和小聚类集合 $SC = \{C_i | i > b\}$ 两部分。
- Step 5 根据离群因子的定义计算每个点的离群因子 NLOF。
- Step 6 按降序排列 NLOF,输出前 n 个离群点。

4 实验与分析

为了验证算法的有效性,我们基于 Matlab 进行了仿真实验,实验数据分别采用人工合成数据和真实数据集,并将其检测效果与传统的 kNN 算法和 LOF 算法进行对比。由于 LOF 算法的检测结果依赖于设置合适参数 MinPts,考虑 kNN 算法依赖于参数 k 的设置,不同的参数值会产生不同的结果,我们选取了经过多次实验得出的最好结果。

图 2(a)为人工合成的同时包含了离群点和离群簇的 2 维数据集,图 2(b)、图 2(c)、图 2(d)分别为 LOF 算法、kNN 算法和本文 ODb3N 算法在人工数据集上对应的检测结果。在实验中,本文算法 ODb3N 能够准确地识别数据集中的离群点以及离群簇;LOF 算法无法正确识别离群簇;kNN 算法结果最差,不仅无法识别离群簇,而且对于不同密度分布的点也无法识别。

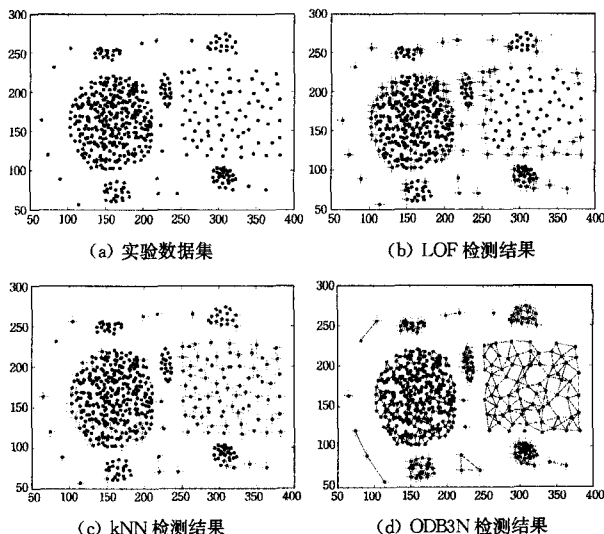


图2 人工数据集实验结果对比

为测试算法在真实数据集上的准确度,我们选取 UCI 标准数据集中 Iris Plants 数据和 Breast-cancer-Wisconsin 数据作为实验数据集。为评估算法结果,在输出 top-n 个离群点时,把其中正确的离群点数与数据集中总的离群点数量比值作为当前的离群点检测率。

Iris Plants 数据由 4 个属性组成,包含了 3 个分类即 Setosa, Versicolour, Virginica, 共 150 条数据,每一类平均占数据集的 30% 左右。我们的实验把 Setosa 类中的数据减少至 10 条作为离群点。图 3(a)为该数据集上的检测结果,横坐标表示输出的前 n 个离群点,纵坐标表示输出的正确的离群点占数据集中离群点总数的百分比。从图中可以看出,在输出离群度最大的前 11 个点时,本文算法刚好检测出全部的正确离群点,LOF 算法在输出前 15 个点的时候才能识别完全部的离群点,而 kNN 算法最多只能检测出 60% 的准确离群点。

Breast-cancer-Wisconsin 数据由 10 个属性组成,包含了 699 条诊断记录,分为 Benign 和 Malignant 2 个类别,实验中我们把其中 Malignant 类数据减少至 39 条作为离群数据。通过实验得到本文算法在输出 55 个离群点时达到最高的检测率为 97%,LOF 算法检测率为 15%,kNN 算法检测率为 71%。从图 3(b)中可以看出 LOF 算法效果明显要低于其它算法,由于实验中的离群点都是属于同 Malignant 类别的,具有簇的特征,而 LOF 算法检测的是局部离群点,因此在检测离群簇时效果会急剧降低。

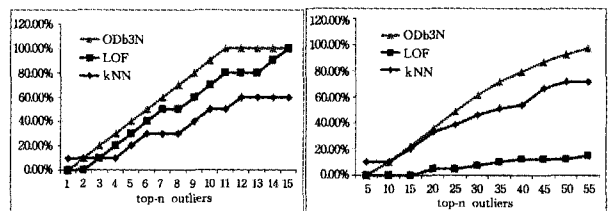


图3 真实数据集实验结果对比

实验对比发现:ODb3N 算法在包含了离群簇的人工数据集和真实数据集上的检测准确度是最高的,同时 kNN 和 LOF 算法检测结果是在设置不同参数经过多次实验后取得的,而 ODb3N 算法却无需参数。从上述结论可以得出本文算法是明显优于其它算法的。

结束语 本文引入自然最近邻的思想,提出一种基于自然最近邻的离群检测算法,成功避免了传统 k 近邻算法中的参数选择问题。仿真实验结果证明:在包含离群簇的数据集中,本文算法与传统的 kNN、LOF 算法比较,也明显具有更好的效果和性能。

参考文献

- [1] Gogoi P, Borah B, Bhattacharyya D K. Outlier identification using symmetric neighborhoods[J]. Procedia Technology, 2012, 6: 239-246
- [2] Breunig M M, Kriegel H P, et al. LOF: identifying density-based local outliers[J]. Proc. of 2000 ACM SIGMOD international conference on Management of data. ACM Sigmod Record, 2000, 29(2): 93-104

(下转第 305 页)

式中, s 为一次查询中检索到的相关图像的数目, u 为一次检索过程中检索到的不相关的图像数目。图 3 为平均查准率直方图, 从图中可以看到本文所提出的方法较文献[4, 5]中的方法有更好的检索效果。另一种衡量查询结果的方法是查准率-查全率曲线。查全率的定义为:

$$recall = \frac{s}{s+v} \quad (13)$$

v 为图像库中和检索图像相关但未被检索到的图像数目。查准率和查全率这两个指标越高, 说明检索性能越好, 但从这两个指标的定义可以看出, 它们之间相互矛盾: 查准率越高则查全率越低, 查全率越高则查准率越低。在相同查全率条件下, 查准率越高性能越好。图 4 显示了所有查询图像平均查准率和平均查全率绘制的曲线, 从图中的曲线可以看出, 本文的算法比其他两种方法效果要好一些。

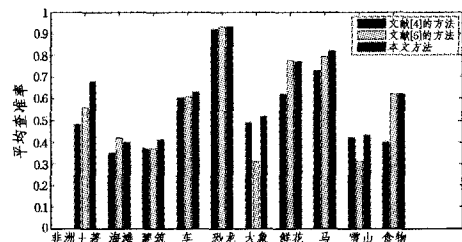


图 3 平均查准率对比

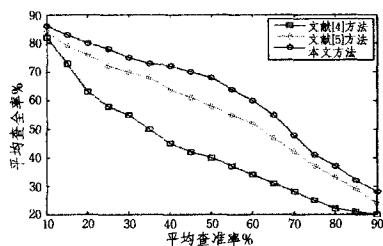


图 4 平均查准率-查全率曲线

结束语 针对基于内容的图像检索方法中提取出来的特征向量不稳定以及对局部噪声敏感的问题, 本文提出一种用稀疏低秩描述图像集特征的图像检索方法。对图像集进行检索时, 首先提取图像的颜色、纹理、形状特征, 然后通过稀疏低秩描述的方法获得稳定特征, 在此图像集描述的基础上进行图像检索。实验表明本文方法具有较好的检索性能, 可将原来不稳定的特征转换为稳定的特征, 并且在保持了全局子空间最小结构的同时, 降低了局部噪声对于检索结果的影响。

参考文献

- [1] Datta R, Joshi D, Li J, et al. Image retrieval: ideas, influences, and trends of the new age[J]. ACM Computing Surveys, 2008, 40(2): 1-60
- [2] Han C H, SimKwee Bo. Real-time face detection using AdaBoost algorithm[C] // Control, Automation and Systems ICCAS, 2008

- International Conference on Seoul, Korea, 2008; 1892-1895
- [3] Konstantinidis K, Gasteratos A, Andreadis I. Image retrieval based on fuzzy color histogram processing[J]. Optics Communications, 2005, 248(15): 375-386
- [4] Zakariya S M, Ali R, Ahmad N. Combining visual features of an image at different precision value of unsupervised content based image retrieval[C] // Computational Intelligence and Computing Research, 2010 IEEE International Conference, 2010; 1-4
- [5] Hiremath P S, Pujari J. Content based image retrieval using color, texture and shape features[C] // 15th International Conference on Advanced Computing and Communications, IEEE Computer Society, 2007; 780-784
- [6] Chen G, Peng R, et al. Pallet Recognition and Localization Method for Vision Guided Forklift[C] // Wireless Communications, Networking and Mobile Computing, 2012 8th International Conference on Shanghai, China, 2012
- [7] Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on feature distribution[J]. Pattern Recognition, 1996, 29(1): 51-59
- [8] 王向阳, 李东明, 杨红颖. 基于 Zernike 色度分布矩的彩色图像检索算法[J]. 模式识别与人工智能, 2012, 25(2): 313-317
- [9] Wee C H, Paramesran R. On the computational aspects of Zernike moments[J]. Image and Vision Computing, 2007, 25(6): 967-980
- [10] Lv X, Chen G, Wang Z C, et al. Grassmannian Manifolds Discriminant Analysis Based On Low-Rank Representation for Image Set Matching[J]. Chinese Conference on Pattern Recognition, 2012(321): 17-24
- [11] Liu G, Lin Z, Yu Y. Robust subspace segmentation by low-rank representation[C] // 27th International Conference on Machine Learning, Haifa; National Science Foundation, 2010; 663-670
- [12] Zhuang L S, Gao H Y, Lin Z C, et al. Non-negative low rank and sparse graph for semi-supervised learning[C] // Computer Vision and Pattern Recognition, IEEE Computer Society, 2012
- [13] Lin Z, Chen M, Wu L, et al. The augmented Lagrange multiplier method for exact recovery of a corrupted low-rank matrices[J]. Mathematical Programming, 2009
- [14] Cai J F, Candes E J, Shen Z. A singular value thresholding algorithm for matrix completion[J]. SIAM Journal on Optimization, 2010, 20(4): 1956-1982
- [15] Hale E T, Yin W, Zhang Y. Fixed-point continuation for l1-minimization: methodology and convergence[J]. SIAM Journal on Optimization, 2008(19): 1107-1130
- [16] Nguyen G P, Worrington M, Smeulders A W M. Similarity learning via dissimilarity space in CBIR[C] // 8th ACM International Workshop on Multimedia Information Retrieval, 2006; 107-115
- [17] Li J, Wang J Z. Real-time computerized annotation of pictures[C] // ACM Multimedia Conference, 2006; 911-920
- [18] 黄丽雯, 汪鑫, 王涛. 一种基于形状特征的颅颌面 X 片图像检索方法[J]. 重庆理工大学学报: 自然科学版, 2013, 27(6): 72-75

(上接第 278 页)

- [3] Hautamaki V, Karkkainen I. Outlier detection using k-nearest neighbor graph[C] // Proc. 17th IEEE Int. Conf. on Pattern Recognition, 2004, 3: 430-433
- [4] Angiulli F, Palopoli L. Detecting outlying properties of exceptional objects[J]. ACM Transaction on Database Systems, 2009, 34(1): 62-74
- [5] Richard J, Chris C. Fuzzy-rough nearest neighbor classification and prediction[J]. Theoretical Computer Science, 2011, 412

- (42): 5871-5884
- [6] Pandya D H, Upadhyay S H, Harsha S P. Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-kNN[J]. Expert Systems with Applications, 2013, 40(10): 4137-4145
- [7] Xu Yong, Zhu Qi, et al. Coarse to fine K nearest neighbor classifier[J]. Pattern Recognition Letters, 2013, 34(9): 980-986
- [8] 符永铨, 王意洁. DKNNS: 面向延迟敏感型应用的可扩展精确分布式 K 近邻搜索算法研究[J]. 中国科学(信息科学), 2012, 42(5): 561-577