一种具有动态邻域特点的自适应最近邻居算法

冯 骥 张 程 朱庆生

(重庆师范大学计算机与信息科学学院 重庆 401331)

(重庆大学软件理论与技术重庆市重点实验室 重庆 400044)

摘 要 传统的最近邻居算法主要分为k-最近邻居和逆最近邻居,然而二者均在邻域参数选择问题中饱受困扰。在这两种思想的基础上,提出一种具有动态邻域特点的最近邻居算法——自然邻居,并围绕其概念与特性形成了一套有效的方法。该算法从根本上克服了传统最近邻居思想在任意形状(如流型)数据集中参数选择的难题,摆脱了传统方法的参数依赖,并且取得了极佳的效果。自然邻居思想具有完善的理论模型和详细的实现算法,并且经验证其具有很强的鲁棒性和适应性。

关键词 最近邻居,自然邻居算法,动态邻域

中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.12.036

Adaptive Nearest Neighbor Algorithm with Dynamic Neighborhood

FENG Ji ZHANG Cheng ZHU Qing-sheng

(College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China)

(Chongqing Key Laboratory of Software Theory & Technology, Chongqing University, Chongqing 400044, China)

Abstract Traditional nearest neighbor algorithm includes k-nearest neighbor (KNN) and reverse nearest neighbor (RNN), and they have been proposed in the literature, but most of them are vulnerable to their parameter choice. In this paper, a novel algorithm of nearest neighbor was proposed, named natural neighbor (NaN). In contrast to KNN and RNN, it is a scale-free nearest neighbor, and it can be used in any dataset effectually, especially data on manifold. This article discussed the theoretical model and its detailed implementation algorithm of natural neighbor in a different field, and the related questions of NaN concepts were discussed by the experimental tests.

Keywords Nearest neighbor, Natural neighbor algorithm, Dynamic neighborhood

1 引言

近些年来,最近邻居的概念以及与其对应的邻居搜索算法在数据挖掘[1]、图像处理[2]、模式识别[3]等多个领域中有着广泛的应用基础,并取得了许多令人满意的结果。最近邻居概念中近邻关系的定义是其思想的基础,对该思想的各种方法起着决定性的作用。在众多的近邻关系中,应用最为广泛的无疑是k-最近邻居[4](k-Nearest Neighbor, KNN)和逆最近邻居[5](Reverse Nearest Neighbor, RNN)。然而,无论是KNN还是RNN,最近邻居的概念中始终存在着一个悬而未决的问题——如何选择大小合适的邻域。多数的最近邻居概念中关于邻域的定义都基于k-最近邻域和ε-最近邻域这两个根本的邻域概念,它们分别代表了两种邻域选择的思路:选择距离数据点 i 最近的 k 个数据点或者选择距离小于 ε 的点作为 i 的邻居。然而这两种技术都存在着参数难以确定的问题:参数的选择结果通常涉及人为干预,仅适用于实验中的特定数据集,不具备良好的普适性。更为重要的是,该参数的选择对算法的效率起着决定性的作用,基于最近邻居算法的效果因所选参数的不同而有明显差异,因此其取值仅能在一个极小的误差范围内浮动,稍微过大或过小都会严重影响计算结果,使其向不可知的方向严重偏离。

本文提出的自然邻居思想结合了KNN和RNN思想的优点,定义了一种新的邻居关系——自然邻居(Natural Neighbor)。自然邻居在k-最近邻居的查找过程中结合了逆最近邻居的概念与特性,使得其摆脱了传统方法中参数选择的难题,在自然邻居的查找过程中自适应地完成邻居关系的构建,同时获得具有数据集特征信息的自然邻居特征值和自然邻居邻域图。基于自然邻居思想构建的自然邻居关系和其具象化的自然邻居邻域图可以在不需要选取预知参数的情况下保留传统KNN的优点,同时在不规则分布、流形分布等分

到稿日期:2016-11-09 返修日期:2017-02-17 本文受重庆市自然科学基金(cstc2013jcyjA40049),重庆师范大学基金项目(17XLB003)资助。

冯 骥(1986-),男,博士,讲师,主要研究方向为机器学习和数据挖掘,E-mail:jifeng@cqnu.edu.cn;张 程(1977-)男,博士后,主要研究方向为社区网络和移动自组网络;朱庆生(1956-),男,博士,教授,主要研究方向为现代服务业、数字农业、数据挖掘。

布情况下获得极佳的效果。自然邻居查找过程中获得的自然邻居特征值不仅能够反映数据的分布规律，还能应用于传统的 KNN 和 RNN 思想中，作为自适应的邻域参数并取得极佳的结果。

本文的主要工作和贡献如下：

1）创新性地提出了自然邻居的概念，消除了传统最近邻居概念中邻域参数的选择问题，探讨了解决历史难题的一个新的切入点；

2）对自然邻居思想及该思想中所涉及到的各种概念进行了规范定义与特性分析，探寻了其正确性与适用范围，完善了自然邻居相关的理论模型的研究；

3）针对不同的应用领域提出了两种自然邻居搜索算法，并简单探讨了自然邻居思想的具体应用。

文中最后通过实验对文中的概念与算法进行了验证。

2　相关工作

2.1　k-最近邻居

k-最近邻居最先由 Stevens 提出[1]，他指出了由最近的 k 个点形成一个局部数据子集的邻域概念。k-最近邻方法于 20 世纪 50 年代早期被首次提出，然而由于受计算机运算水平的制约，直到 20 世纪 60 年代计算能力大大增强之后才流行起来，此后它被广泛用于模式识别、图像处理、数据挖掘、智能推荐等领域。

随着应用领域的不断扩展，KNN 算法在不同的数据集及应用领域中获得了良好的性能[6-7]。这些算法的改进主要是针对参数 k 的最优化选择[8-10]、距离度量方法的优化[11-12]两个方面。以上方法虽然大大提高了 KNN 算法的适应性和准确性，但是依然无法完全回避人工参数对 KNN 算法的制约问题。

2.2　逆最近邻居

在 k-最近邻居出现后，该领域提出了一种与最近邻居具有相反含义的邻居概念——逆最近邻居，该概念在决策支持、营销和资源定位等领域得到了较好的应用，同时也是数据和知识工程领域中的一个热点问题。

Korn 和 Muthukrishnan[5] 于 2000 年首次引入了逆最近邻居 RNN 的概念，并对 RNN 搜索算法做了奠基性的工作。他们定义了一个对象 i 的逆最近邻居：以查询对象 i 为最近邻居的对象，这些对象到数据集中其他对象之间的距离都不小于到查询对象 i 之间的距离。逆最近邻居的推广形式为逆 k-最近邻居[13]（Reverse k Nearest Neighbors，RkNN），点 p 为查询点 q 的一个逆 k-最近邻居的条件是：查询点 q 是点 p 的一个 k-最近邻居点。近年来，逆最近邻居的研究主要集中在查找问题上，如何快速、准确地获得数据的逆最近邻居是当前的热门研究方向[14]。

3　自然邻居

3.1　自然的邻居关系

无论是 k-最近邻居还是逆最近邻居，其核心思想“邻居”始终无法摆脱邻域参数的制约。传统的方法在解决该问题时主要将思路集中在邻域参数的选择方法上，虽然改进算法能在一定程度上解决参数选择的难题，但都因参数选择方法的局限性、方法的复杂性或新参数的引进而无法对问题给出较完善的解决方案。因此我们尝试一种新的思路，即从最基本的概念“邻居”入手，通过新的“自然邻居”定义彻底解决最近邻居概念中邻域参数的选择问题。

我们尝试用社会学中的思想具体阐述数据间的关系问题。基本的社会学思想认为“人生不能无群”，这里的“群”就是“社会”。社会学家研究时常常分析群体，如社会组织、宗教组织、政治组织及商业组织。大到社会，小到各种组织，维系其相互关系的就是各种各样的人际关系。我们用“信任”这种简单的人际关系进行更为直观的分析。信任关系可以局限于单方面，也就是“A 信任 B”和“A 被 B 所信任”两种模式。然而我们认为，单方面的信任是一种不稳定的人际关系。真正意义上的信任关系应该被定义为一个双方面的相互关系，即“A 信任 B”且“A 被 B 所信任”。我们认为，在社会体系中，与单方面的信任关系相比，这种相互的信任关系作为自然稳定的人际关系，隐含着更丰富、更准确的关系特征信息，能够更好地对社会体系、社会结构等问题进行直观的描述，具有进一步研究的价值。

基于上述逻辑体系，我们可以更好地由 k-最近邻居、逆最近邻居引申出本文所提出的新概念——自然邻居。假设取 k 值为 5，即每个人寻找出自己最信任的 5 个人，k-最近邻居的思想认为，我最信任的 5 个人就是我的“k-最近邻居”；而着眼于逆最近邻居的思想，如果我能排在他人最信任的 5 个人的名单中，那么他就是我的“逆最近邻居”。毫无疑问，这两种思想受制于邻域参数 $k=5$，使得我们对数据整体的把握只能依赖于先验知识，且所产生的信任关系都只是基于单方面的概念。我们在这两种邻居概念的基础上提出了自然邻居的概念：首先让每个人都存在于他人的信任名单中，此时整个信任体系处于一个稳定的关系结构，在这种稳定的关系结构中，与我相互信任的人才是我的“自然邻居”。

自然邻居的思想摆脱传统思想中邻域参数制约的要点就在于稳定的关系结构概念的提出。在研究中发现，k-最近邻居的思想能反映出一定的关系信息，在 k-最近邻居查找过程中的逆最近邻居信息则能够对 k-最近邻居思想中 k 值的选取起到重要的参考作用。在之前的例子中，这种思路表现为：一个和谐稳定的社会关系可以归结于一个简单直观的判断条件——每个正常人都必须获得来自于他人的信任。基于这种思路，我们提出自然邻居思想中的稳定关系结构：在自然邻居算法寻找邻居的过程中，当数据集中的所有数据点都至少有一个逆邻居时，数据集的自然邻居结构则为稳定的关系结构。

本文所提出的自然邻居中的“自然”的思想主要体现在以下 3 个方面。

（1）自然邻居关系的定义，本文提出的相互之间的信任关系定义更符合自然的关系结构。

(2)稳定的结构关系构建过程,本文借助稳定关系结构的思想,将邻居搜索终止条件的决定权交由数据完成。在整个构建过程不存在人为参与的干扰,形成自然、准确的邻居关系。

(3)自然邻居的数目,本文中获得数据集的数据点的自然邻居是一个被动的过程,且每个数据点的自然邻居数目不受任何强制约定的限制,互不相同且相互独立,体现了一种无约束的自然状态。

3.2 自然邻居思想

与现有的各种最近邻居方法相比,自然邻居是一种无尺度的概念。在传统的 k-最近邻居思想中,首先由先验知识给出具体的 k 值,然后查找数据集 X 中每个数据的前 k 个邻居。自然邻居以传统的 k-最近邻居方法为基础,在查找过程中对 k 值进行递增并观测所有点的逆邻居,以确定是否达到搜索稳定状态。当算法自动执行到搜索稳定状态时,获取数据集的自然邻居关系,同时得到自然邻居特征值与自然邻居邻域图。

为了准确地描述自然邻居思想,首先对自然邻居思想中所提出的概念进行定义。若无特别说明,后文中涉及到的数据集皆为数据规模为 n 的数据集 $X=\{x_1,x_2,x_3,\cdots,x_n\}$。

定义 1(搜索稳定状态,Search Stable State) 给定数据集 X,依次取 $k=1,2,3,\cdots,n$,对数据集 X 进行 k-最近邻居的查找。在这个查找过程中,当数据集 X 中所有的数据点都至少有一个互为最近邻居时停止循环,我们称在查找深度 $depth=\lambda$ 时,数据集 X 所处的状态为搜索稳定状态,即

$$(\forall x_i)(\exists x_j)(r\in N)\wedge(x_i\neq x_j)\rightarrow(x_i\in KNN_r(x_j))\wedge(x_j\in KNN_r(x_i))$$

其中,查找深度 $depth$ 从 1 增长至 λ。

上述定义运用互为邻居对 k-最近邻居查找进行约束,其本质在于在最近邻居查找的过程中,用逆最近邻居控制查找过程的迭代,其搜索稳定状态时的查找深度 k 接近 k-最近邻居思想中 k 的最优选择。然而数据集中的噪声数据会对查找深度的取值产生极大的影响,因此我们在当前概念分析中仅针对去噪数据集,对可能含有噪声数据的广义数据集的处理将在后文的搜索算法中进行详解。

定义 2(自然邻居,Natural Neighbor) 在数据集 X 处于搜索稳定状态时,对于任意 $x,y\in X$,x,y 互为自然邻居,当且仅当点 x 与点 y 互为邻居。即:

$$x_j\in NN(x_i)\Leftrightarrow(x_i\in KNN_\lambda(x_j))\wedge(x_j\in KNN_\lambda(x_i))$$

与传统的最近邻居相比,自然邻居加强了互为邻居的观念,且数据点的邻居数也并非由算法统一指定,包含了更丰富的关系信息。

定义 3(自然邻居特征值,Natural Neighbor Eigenvalue) 当数据集 X 处于搜索稳定状态时,查找深度 $depth$ 即为数据集 X 的自然邻居特征值 λ。

$$\lambda\triangleq r_{r\in N}\{r\mid(\forall x_i)(\exists x_j)(r\in N)\wedge(x_i\neq x_j)\rightarrow(x_i\in KNN_r(x_j))\wedge(x_j\in KNN_r(x_i))\}$$

该定义对自然邻居特征值进行了形式化的描述,其中 λ 被定义为自然稳定状态时查找深度 r 的值,因此定义的后半部分即为自然稳定状态的定义。自然邻居特征值作为查找深度的最大值,反映了数据集的分布规律,同时也可以作为传统 KNN 方法中的邻居参数 k 的选取参考。

定义 4(自然邻居邻域图,Natural Neighborhood Graph) 当数据集 X 处于搜索稳定状态时,由数据集 X 中自然邻居关系构成的邻域图为数据集 X 的自然邻居邻域图。在自然邻居邻域图 $G=(V,E)$ 中,顶点集 V 中每一个顶点 v_i 对应数据集 X 中的数据点 x_i,而图中两个顶点 v_i 和 v_j 之间存在一条边,当且仅当 x_i 和 x_j 互为自然邻居,即:

$$e(v_i,v_j)\in E\Leftrightarrow x_j\in NN(x_i)$$

自然邻居邻域图是自然邻居关系的具象描述,更为直观、准确地反映了数据间的自然邻居关系。自然邻居邻域图在不同的应用领域可以有不同的变种,我们将在后文进行讨论。

自然邻居的查找过程可以用下面的算法进行描述。

算法 1 自然邻居查找算法

```
输入:数据集 X
输出:自然邻居特征值 λ,自然邻居邻域图边集 NaN_Edge
1. r=1,flag=0,NaN_Edge=∅
2. 创建数据集 X 对应的 k-d 树 T
3. ∀ x_i∈X,NaN_Num(x_i)=0
4. While flag==0 do
   For all x_i∈X do
      knn_r(x_i)=findKNN(x_i,r,T)
      KNN_r(x_i)=KNN_r(x_i)∪{knn_r(x_i)}
      If x_i∈KNN_r(knn_r(x_i))&&{knn_r(x_i),x_i}∉NaN_Edge
      then
      NaN_Edge=NaN_Edge∪{x_i,knn_r(x_i)}
      NaN_Num(x_i)=NaN_Num(x_i)+1
      NaN_Num(knn_r(x_i))=NaN_Num(knn_r(x_i))+1
      End If
   End For
   If all (NaN_Num(x_i))≠0
        flag=1
   End If
   r=r+1
5. End While
6. λ=r-1
7. Return λ,NaN_Edge
```

在算法 1 中,λ 为自然邻居特征值,NaN_Edge 为自然邻居邻域图边集,其中每一个元素代表了自然邻居邻域图中的边的两个顶点。通过自然邻居邻域图边集 NaN_Edge 可以进一步得到每个数据点的邻居数以及完整的自然邻居邻域图,或者除了加权自然邻居邻域图之外的其他几种自适应自然邻居邻域图。如果需要得到加权自然邻居邻域图,则需要在算法的 NaN_Edge 的构造过程中记录每条边生成时的搜索循环次数 r,并将其作为该边的权值保存在 NaN_Edge 中。k-d 树 T 的构造可以加快最近邻居的查找速度,使得函数 $findKNN(x_i,r,T)$ 能更高效地找到数据点 x_i 的第 r 个邻居。

通常情况下，该算法的整体时间复杂度为 $O(N*\log N)$，最差情况为 $O(N^2*\log N)$。首先，k-d 树 T 构造阶段的时间复杂度为 $O(N*\log N)$。在此之后，对于每一个搜索循环 r，自然邻居搜索的时间复杂度为 $O(N*\log N)$，因此整个自然邻居搜索阶段的时间复杂度为 $O(\lambda*N*\log N)$。自然邻居特征值 λ 的取值范围为 $2\leqslant\lambda<N$，在没有离群点的情况下，其通常取值 6 或 7，随着维度的增高，λ 的值会变得更大，但通常情况下能保持在 30 以内。

4　自然邻居搜索算法

4.1　鲁棒的自然邻居搜索算法

数据噪声指在一组数据中无法解释的数据变动，即一些不与其他数据相一致的数据。数据集中的噪声数据会对自然邻居思想的算法产生难以估量的影响，主要表现为在搜索稳定状态中查找深度($depth$)的递增过程。举个简单的例子，在无噪声的数据集 X 中加入处于无穷远的噪声数据 $noise$ 而构成新的数据集 X'，此时数据集大小变为 $n+1$。显然，只有当查找深度 $depth=n$ 时数据集 X' 才能达到搜索稳定状态，而查找深度从 λ 到 $n+1$ 的过程中生成的冗余自然邻居关系使得所有的有用信息都被掩埋。

真实情况的数据集通常都会含有一定的噪声，在面对绝大多数情况时，数据集通常都未进行去噪的预处理。为了提高算法的普适性，对搜索稳定状态的定义进行如下修改。

定义 5(鲁棒的搜索稳定状态)　给定数据集 X，依次取 $k=1,2,3,\cdots,n$，对数据集 X 进行 k-最近邻居的查找。在查找过程中，统计未获得逆邻居的点的个数，记为 $zeroNumber_k$，并判定以下两个条件：

1) $zeroNumber_k=0$

2) $zeroNumber_k-zeroNumber_{k+k/2}=0$

若 $zeroNumber$ 在查找深度 $depth=k$ 时满足上述任一条件，则称在查找深度 $depth=k$ 时，数据集 X 所处的状态为搜索稳定状态。

上述定义中，条件 1)即为定义 1 中搜索稳定状态的定义，当数据集中不存在噪声或存在多个类似噪声使得其聚为噪声簇时，搜索过程满足该条件并达到搜索稳定状态。在有一个或多个相互独立的噪声时，条件 2)可以很好地消除噪声带来的不良影响。令 $X_{zero}(r)$ 为查找深度为 r 时数据集中逆邻居数目为零的点的集合，即 $X_{zero}(r)=\{x\mid rnb_r(x)=0,x\in X\}$。当 $X_{zero}(k)=X_{zero}(k+k/2)$ 时，即在连续的 $k/2$ 次最近邻居查找后没有更多的点获得逆最近邻居，则在查找深度为 k 时数据集已经到达搜索稳定状态，此时 $X_{zero}(k)$ 为数据集中的独立噪声点。

鲁棒的自然邻居改进算法不仅保存了自然邻居方法无参数、流形适应等特征，而且进一步加强了算法的鲁棒性。该算法通过自然邻居邻域图可以直接对数据集进行聚类和离群检测，具体的效果可以在实验部分进行简单验证。

4.2　基于最小生成树的自然邻居特征值准确估量算法

除了自然邻居思想中的特性应用以外，自然邻居特征值作为数据集自适应的邻居参数，可以作为传统的 KNN 方法中的邻居数 k 值，使得传统的 KNN 方法摆脱参数选择的困扰，并取得极佳的实验结果。该邻居领域并不关心数据集中的自然邻居关系，因此如何利用自然邻居的概念快速地获取自然邻居特征值成为了主要目标。基于上述需求，我们提出了基于最小生成树(MST)的自然邻居特征值估量算法，该方法能回避全局的自然邻居查找，通过最小生成树快速获得自然邻居特征值的估计值，同时保证其误差处于可接受的范围内。

定义 6(自然邻居邻域图的关键点与关键边)　前文提到 $X_{zero}(r)$ 为当查找深度为 r 时数据集中逆邻居数目为零的点的集合，则自然邻居邻域图的关键点 v_{key} 满足：

$$\{v_{key}\mid v_{key}\in X_{zero}(\lambda-1)-X_{zero}(\lambda)\}$$

自然邻居邻域图中所有与自然邻居邻域图关键点相连接的边为自然邻居邻域图关键边 e_{key}。

直观地说，通过对自然邻居特性的分析可以得出，当且仅当查找深度到达自然邻居特征值时，自然邻居邻域图关键点 v_{key} 才能获得逆邻居，且此时尚未获得逆邻居的点皆为噪声数据。称其为自然邻居邻域图关键点，也正是因为这一类数据具有决定自然邻居特征值的作用。

通过证明得到，自然邻居邻域图的关键点和关键边具有以下性质。

推论 1

1)自然邻居邻域图中一定存在至少一个自然邻居邻域图的关键点和关键边。

2)在加权邻域图中，自然邻居特征值 λ 与 e_{key} 的权值相等。

3)关键边至少有一个顶点的度为 1。

4)如果关键边的另一个顶点的度大于 1，则经过该顶点的所有边中，关键边的长度最长。

同时构造数据集的自然邻居邻域图和最小生成树，结合推论 1，可以得到以下结论。

推论 2

1)自然邻居邻域图的关键点在最小生成树中一定也是叶节点。

2)自然邻居邻域图的关键点在最小生成树中所在的边，与其另一点的所有边相比，长度也最长。

3)噪声点一定是最小生成树中的叶节点。

4)除噪声点和自然邻居邻域图关键点以外，其他的最小生成树叶节点在加权自然邻居邻域图中的边的最大权值一定小于关键边的权值。

基于推论 1 中的思想，在自然邻域图中，自然邻居特征值的计算问题等价于获取数据集 N 的加权自然邻居邻域图的关键边的权值。

定理 1(加权自然邻居邻域图中权值的计算)　令 $neighborSort(x,y)$ 为点 x 在点 y 的最近邻居列表中所处的位置，

即 x 是 y 的第 $neighborSort(x,y)$ 个邻居。自然邻居邻域图中任意边 $e=\{x_1,x_2\}$，记其近邻差 $\Delta E(e)=\max(neighborSort(x_1,x_2),neighborSort(x_2,x_1))$。加权自然邻居邻域图中边 e 的权重 $weight(e)=\Delta E(e)$。

同时，由推论 1 和推论 2 可以得出，对于最小生成树中与叶节点相连的边，如果其另一父节点所连接的边的长度均小于该点，则此边为自然最近邻域图中的关键边。

基于推论 2 中的思想，提出一种最小生成树中选取候选边 $e=\{x,y\}$ 的条件：

1）x 为叶节点；

2）所有与 y 相连的点中，e 的长度最短。

所有符合该条件的边组成侯选边集 $E_{candidate}$，最小生成树的关键边一定包含于其中。此时，可以通过定理 1 计算该候选集中边的近邻差，并通过一定的算法获得自然邻居特征值的准确估计。该估计值的准确性在于：虽然最小生成树的关键边与自然邻居邻域图的关键边不完全是一一对应的关系，但在数据集分布呈现流形、螺旋形等特征时，自然邻居特征值通常较小，此时可将定理 1 所得的自然邻居特征值作为准确值；仅当数据集的自然邻居特征值较大且存在个别特殊关键点时该估计值可能有较小的偏差，可以在最后运行一定的修正运算以获得更为准确的估计。在大部分情况下，最小生成树的关键边同时也是自然邻居邻域图的关键边。

基于上述特点，本文提出了基于最小生成树的自然邻居特征值估量算法，在获得数据集的最小生成树的基础上对叶子节点进行分析处理，进而更高效地对自然邻居特征值的近似值进行计算。

算法 2 基于最小生成树的自然邻居特征值估量算法

输入：数据集 X

输出：自然邻居特征值 λ

1. 创建数据集 X 对应的最小生成树 T
2. 选取 T 中的侯选边构成侯选边集 $E_{candidate}=\{e\}$
3. $\forall e\in E_{candidate}$，$\Delta N=calculateN(e)$
4. 对 ΔN 进行排序，$\Delta N_1\geqslant\Delta N_2\geqslant\cdots\geqslant\Delta N_k$，令 $|\Delta N_i|$ 为近邻差的值为 ΔN_i 的边的数量，flag=0
5. $|\Delta N_m|=\max|\Delta N_i|$
6. if m==k，step10；
7. $rep=|\Delta N_{m+1}|-|\Delta N_m|$，if rep<m/2，m=m+1，step6；
8. if flag==1，step10；
9. $\Delta N_i=$ recalculate N(e) (i=m，m+1，…，k)，flag=1，step6；
10. $\lambda=\Delta N_m$
11. Return λ

对于数据集 X，$calculateN(e)$ 参照定理 1 的方法计算边 e 的两个顶点之间的近邻差，$recalculateN(e)$ 是对边 e 所对应的叶节点进行近邻差的修正计算，通过修正计算找到准确的自然邻居邻域图中关键边所对应的近邻差。前文已分析出最小生成树的关键边与自然邻居邻域图的关键边不完全符合，因此为了保证检测的准确性，算法后部分使用的 $recalculateN(e)$ 时包含了多个 $calculateN(e)$ 循环，具体在算法中表现为在获得边 e 中叶节点 x 与其前 ΔN_k 个邻居所构成的边集后，计算该边集中所有边的权值并返回其中的最小权值，以此来降低仅使用最小生成树进行权值计算方时产生的偏差。

该算法可以分为 3 个部分：构造最小生成树（step1）；通过计算最小生成树关键边侯选集的近邻差估计自然邻居特征值（step2－step8）；重新计算准确自然最近邻居关键边候选集的近邻差，获得自然邻居特征值（step9－step10）。侯选边集中近邻差大于特征值的边只有两种情况：1）叶节点为离群点；2）叶节点为关键点的非自然邻居邻域图关键边。通常，若特征值较小（$\lambda\leqslant 6$），则在 step9 重新计算近邻差之前获得的近邻差与实际的特征值基本一致；当特征值较大时，算法最终获得的特征值估计值通常略大于实际的特征值，但其仍然会远小于离群点所产生的近邻差。因此，通过算法 2 的 3 个步骤，得到的自然邻居邻域图特征值估计值即便无法保证与准确值完全一致，也可以将误差控制在可接受的极小范围内。

4.3 自然邻居算法在聚类分析、离群挖掘、机器学习等方面的应用探索

自然邻居算法由于其概念的独特性，在聚类分析、离群挖掘、机器学习等方面有着先天的优势。传统的基于邻居思想的方法往往由于邻域选择不当产生错误。如在流形数据的处理中，若邻域值选择过大则很容易产生短路边错误，且一旦有短路边出现，就会导致距离矩阵发生改变，使低维嵌入结构不同；如果选择邻域值太小，则可能使流形出现大量不连通区域，导致无法得到满意的结果；选择“正好合适的邻域”则因不同数据集的独特特征成为一个难以判断的困境。而摆脱了参数困扰的自然邻居算法一方面由于其自适应性可以直接得到极佳的聚类或者分类效果，另一方面自然邻居特征值可以为邻域值的选择提供有力的支持。

在自然邻居思想的研究过程中，我们对其进行了多方面的尝试和改进[15-17]，并最终形成本文中完善的自然邻居思想体系。凭借自然邻居思想的参数自适应性和邻域动态性，在当前的研究过程中尝试将自然邻居思想用于聚类分析、分类分析、离群点检测等多个领域，以期取得更令人满意的实验效果。

5 实验结果

本文独创性地提出了自然邻居思想，围绕自然邻居概念提出了自然邻居特征值、自然邻居邻域图等一系列概念，并对其理论概念和性质进行了定义和验证。本节将对自然邻居思想中的相关问题进行简要阐述，并在具体实验中进行验证。

5.1 自然邻居动态邻域

为了更好地展示自然邻居思想的特点，本实验对两个人工数据集进行了自然邻居搜索，并在图中标识了其各个点的自然邻居数目。数据集 Glass 和 Wine 均为 UCI 标准数据集，其中 Glass 中点的数目为 214，数据集的维度为 9；Wine 则由 173 个 13 维的数据点组成。

从图 1 可以看到，在自然邻居思想中，每个点的邻居数目

是不固定的。自然邻居的查找过程不需要邻域参数的限定，且这种动态的邻域思想能够更准确地反映数据点之间的关系，使得每个数据点根据自己的环境特性找到合适的邻居。具体来说，具有更多邻居的点附近的数据更为密集，而边缘点则具有较少的自然邻居。

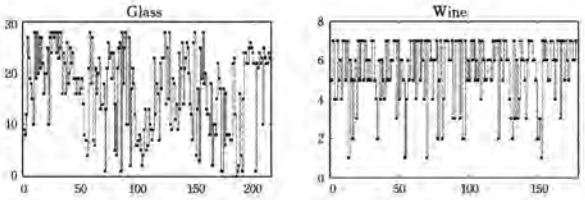

图 1　数据点的自然邻居动态邻域

5.2　自然邻居邻域图与 k-最近邻居邻域图的比较

邻域图是自然邻居思想在数据集上最为直观的反映，它将数据集中数据点之间的自然邻居关系用边的形式展现在图中，因而可以通过图中连通性的分析探寻数据间隐含的信息。在数据挖掘领域，许多算法通过对数据集的 k-最近邻居邻域图进行分析与变换，进而完成对数据的分类、聚类以及离群挖掘等任务。

与 k-最近邻居邻域图相比，自然邻居邻域图具有以下优点：

1）自然邻居邻域图是自适应邻域图，在图的构建过程中，自然邻居邻域图不需要额外参数的帮助。

2）自然邻居邻域图反映了数据集中数据点的自然邻居关系，因此图中每个顶点的度是由自然邻居的数目决定的，其大小取决于对应的数据点周围数据的分布情况，并非统一的值。

3）在自然邻居邻域图中，处于密集区域的数据点具有更多的邻居，对应的顶点度较大，边缘点对应的顶点的度较小，而噪声点对应的顶点的度为零。

4）对于不同分布规律的数据集，自然邻居特征值各不相同，因而自然邻居邻域图具有不同的密度；而 k-最近邻居邻域图的密度由参数决定，通常情况下无法适应数据集的变化。

接下来，选取 3 个具有不同特性的人工数据集[18]，通过实验对比自然邻居邻域图与 k-最近邻居邻域图的差别。这 3 个数据集为 data-c-cc-nu-n，data-uc-cc-nu-n 和 data-c-cv-nu-n，分别代表了某种形态的不均匀分布规律。

图 2 从上到下依次为数据集的数据分布、邻域参数 $k=N*1\%$ 时的 k-最近邻居邻域图、$k=N*2\%$ 时的 k-最近邻居邻域图和自然邻居邻域图。可以直观地看到，当人为确定邻域参数时，构造所得的邻域图的抗噪性极差，因而建立在该邻域图上的后续数据挖掘需要更为精密的分析才能得到理想的结果；并且参数的选取也具有极大的不确定性，不同数据集上的最优邻域参数不尽相同，且该最优值仅存在于一个极小的区间，略大或者略小均会严重影响所获得的邻域图的质量。而自然邻居邻域图摆脱了传统方法对参数的依赖，能够自适应地得到令人满意的邻域图，该邻域图可直观准确地反映数据集中数据的基本关系。与此同时，自然邻居邻域图同时也具有很强的抗噪性，在邻域图的构建过程中可找到噪声点并进行标示。

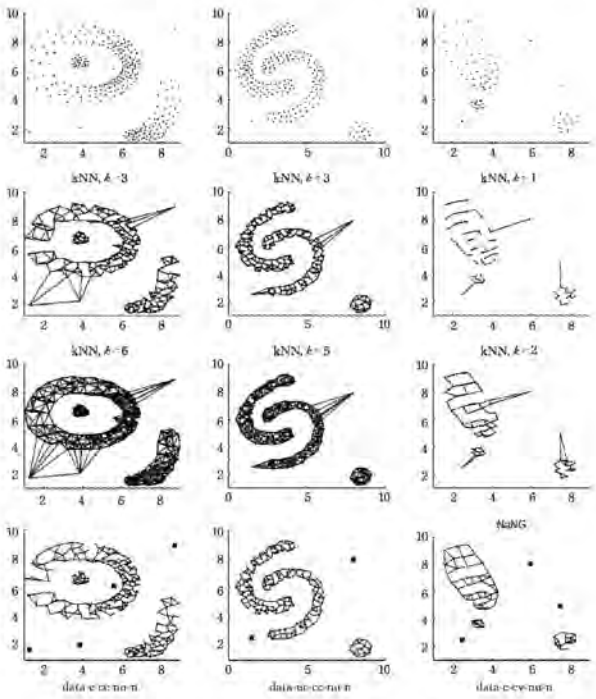

图 2　自然邻居邻域图与不同参数取值的 k-最近邻居邻域图

在与不同参数的 k-最近邻居邻域图的对比中可以看出，自然邻居邻域图很好地展示了数据的基本分布规律，完成了良好的聚类和离群挖掘，并且能将自然邻居特征值保持在合适的范围内。自然邻居邻域图在针对噪声数据时，能很好地排除噪声数据造成的干扰；针对螺旋形数据时，很好地保持了数据的规律，没有产生过载或者短边的情况。

5.3　自然邻居特征值的稳定性

本文提出的自然邻居思想共包含 4 个基本概念：搜索稳定状态、自然邻居、自然邻居特征值、自然邻居邻域图。自然邻居特征值在自然邻居思想及各领域运用方面起着至关重要的作用，是自然邻居思想的核心指标，其稳定性决定了自然邻居思想中搜索稳定状态、自然邻居关系及自然邻居邻域图的稳定性。因此只需要验证自然邻居特征值的稳定性，即可通过推导得到其他概念的稳定性。

本次实验分别采用本文提出的非鲁棒（原始算法）和鲁棒（改进算法）的两种自然邻居搜索算法，对均匀分布和高斯分布下不同数据规模的数据集分别进行 200 次实验，实验结果如表 1 所列。

表 1　基于最近邻居思想的自然邻居特征值的重复实验的平均值

	500	600	700	800	900	1000
原始算法 均匀分布	6.0750	6.3150	6.4750	6.4250	6.6050	6.5650
改进算法 均匀分布	6.0250	6.3400	6.2500	6.2900	6.5700	6.6450
原始算法 高斯分布	21.2350	18.8250	22.4750	19.7600	22.3950	20.7350
改进算法 高斯分布	10.0150	10.3300	10.1150	10.3400	10.3500	10.1700

从表 1 可以得到以下结论：

1）在任何对等或者不对等的数据规模下，均匀分布和高

斯分布的自然邻居特征值都具有极大的差异性；

2)在有规律的数据集中，两种自然邻居搜索算法均能很好地保证自然邻居特征值的稳定性；

3)在无噪声点的情况下，鲁棒算法能够发现局部离群点，并提前结束自然邻居搜索。

两种算法在均匀分布中的特征值基本一致，而在高斯分布中的差别较大，其主要原因在于高斯分布的特殊性。高斯分布因存在 3σ 区域的小概率事件，使得部分数据处于离群点的状态，因此非鲁棒算法的特征值大于鲁棒算法的特征值，且稳定性低于鲁棒算法。由此可见，本文中鲁棒的自然邻居搜索算法适用性更广，性能更好。

接下来进一步研究在多次重复实验的情况下，自然邻居特征值在不同规模的均匀分布和高斯分布数据集下的变化情况。

在本实验中，人工生成均匀分布数据的规模从 500 到 1000 不断增加，对每个数据规模重复 200 次实验。通过对实验结果(见图 3)的观察，我们认为特征值基本稳定于 6，其稳定性和数值并没有随着数据集规模的变化而产生变化。

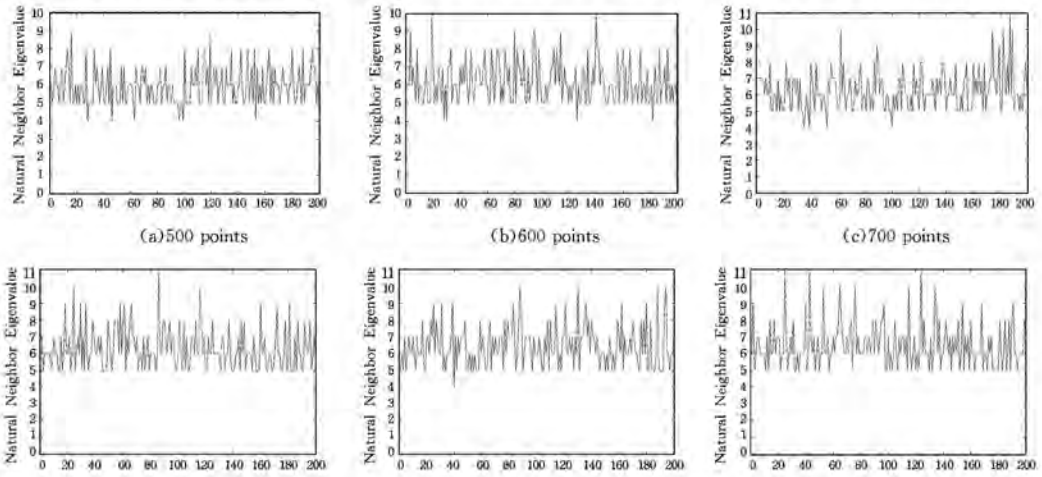

图 3 均匀分布下自然邻居特征值的稳定性

在本实验中，人工生成高斯分布数据集的规模从 500 到 1000 不断增加，对每个数据规模重复 200 次实验。通过对实验结果(见图 4)的观察，我们认为特征值基本稳定于 10，其稳定性和数值同样没有随着数据集规模的变化而产生变化。

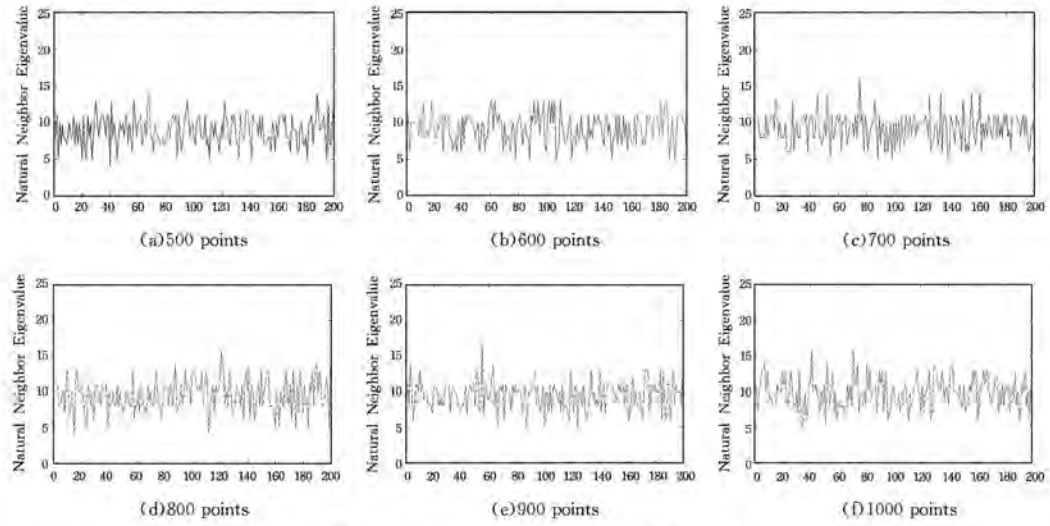

图 4 高斯分布下自然邻居特征值的稳定性

通过对以上实验以及实验结果的分析，我们验证了自然邻居特征值的稳定性，进而获得了搜索稳定状态、自然邻居关系和自然邻居邻域图的稳定性验证。

结束语 KNN 和 RNN 算法及其改进算法通常都需要领域知识预先取得邻域参数，本文则采用一种不同的思维方式来确定最近邻居，用自然邻居思想对邻居的概念进行了全新的定义。

本文研究了基于自然邻居的最近邻居思想，详细定义了搜索稳定状态、自然邻居、自然邻居特征值、自然邻居邻域图等基本概念，构建并完善了基于自然邻居思想的理论模型。在此基础上，本文详细阐述了鲁棒的自然邻居搜索算法和基于最小生成树自然邻居特征值的准确估量算法，概述了自然邻居思想在聚类分析、离群挖掘、机器学习等方面的应用，并且通过实验展现了自然邻居思想和多样化的自然邻居邻域图，验证了自然邻居特征值的稳定性。在今后的工作中，我们将进一步探讨并完善相关概念，探寻更有效的搜索算法以降

低当前算法的复杂度，尝试将自然邻居思想应用于图像处理、大数据等热门领域，将自然邻居思想发展得更为成熟。

参 考 文 献

[1] AMBERT K H, COHEN A M. k-Information gain scaled nearest neighbors: a novel approach to classifying protein-protein interaction-related documents[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2012, 9(1): 305-310.

[2] BRIAN M, CAROLINA G, GERT L. Contextual Object Localization With Multiple Kernel Nearest Neighbor [J]. IEEE Transactions on Image Processing, 2011, 20(2): 570-585.

[3] SALVADOR G, JOAQUIN D, JOSE R C. Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study [J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2012, 34(3): 417-435.

[4] STEVENS S S. Mathematics, Measurement, and Psychophysics [M]// Handbook of Experimental Psychology. 1951: 1-49.

[5] KORN F, MUTHUKRISHNAN S. Influence Sets Based on Reverse Nearest Neighbor Queries [J]. ACM SIGMOD Record, 2000, 29(2): 201-212.

[6] WANG J, NESKOVIC P, COOPER L. Improving Nearest Neighbor rule with a simple adaptive distance measure [J]. Pattern Recognition Letter, 2007, 28(2): 207-213.

[7] SALVADOR G, JOAQUIN D, JOSE R C. Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(3): 417-435.

[8] GHOSH A K. On Optimum Choice of k in Nearest Neighbor classification [J]. Computational Statistics & Data Analysis, 2006, 50(11): 3113-3123.

[9] HASTIE T, TIBSHIRANI R. Discriminant Adaptive Nearest neighbor Classification [J]. IEEE Transactions Pattern Analysis and Machine Intelligence, 1996, 18(6): 607-616.

[10] GHOSH A K, ANIL K. On Nearest Neighbor Classification sing Adaptive Choice of k [J]. Computational & Graphical Statistics, 2007, 16(2): 482-502.

[11] DOMENICONI C, PENG J, GUNOPULOS D. Locally Adaptive metric Nearest-Neighbor Classification [J]. IEEE Transactions Pattern Analysis and Machine Intelligence, 2002, 24(9): 1281-1285.

[12] BHATTACHARYA G, GHOSH K, CHOWDHURY A S. An affinity-based new local distance function and similarity measure for kNN algorithm [J]. Pattern Recognition Letters, 2012, 33(3): 356-363.

[13] YIU M L, MAMOULIS N. Reverse Nearest Neighbors Search in Ad Hoc Subspaces [J]. IEEE Trans. on Knowledge and Data Engineering, 2007, 19(3): 412-426.

[14] WANG S S, CHAI S, LV Q N. A Pruning Based Continuous RkNN Query Algorithm for Large k [J]. Chinese Journal of Electronics, 2012, 21(3): 523-552.

[15] ZHANG Y. Study on Classification algorithm based on natural nearest neighbor [D]. Chongqing: Chongqing University, 2015. (in Chinese)
张莹. 基于自然最近邻居的分类算法研究[D]. 重庆：重庆大学，2015.

[16] HUANG J L. Study on non-parametric clustering based on natural nearest neighborhood [D]. Chongqing: Chongqing University, 2014. (in Chinese)
黄金龙. 基于自然最近邻的无参聚类算法研究[D]. 重庆：重庆大学，2014.

[17] TANG H. An outlier detection algorithm based on natural nearest neighbor [D]. Chongqing: Chongqing University, 2014. (in Chinese)
唐汇. 基于自然最近邻居的离群检测算法研究[D]. 重庆：重庆大学，2014.

[18] İNKAYA T, KAYALIGIL S, ÖZDEMIREL N E. An adaptive neighbourhood construction algorithm based on density and connectivity [J]. Pattern Recognition Letters, 2015, 52: 17-24.

(上接第 182 页)

Advanced defensive metrics for nba basketball[C]// Proceedings of the 9th MIT Sloan Sports Analytics Conference. Boston, MA, USA, 2015: 1-8.

[14] WIENS J, GUHA BALAKRISHNAN J B, GUTTAG J. To Crash or Not To Crash: A quantitative look at the relationship between offensive rebounding and transition defense in the NBA [C]// Proceedings of the 7th MIT Sloan Sports Analytics Conference. Boston, MA, USA, 2013: 1-7.

[15] REN L, DU Y, MA S, et al. Visual analytics towards big data [J]. Journal of Software, 2014, 25(9): 1909-1936. (in Chinese).
任磊，杜一，马帅，等. 大数据可视分析综述[J]. 软件学报，2014，25(9)：1909-1936.

[16] HERMAN I, MELANCON G, MARSHALL M S. Graph visualization and navigation in information visualization: A survey [J]. IEEE Transaction on Visualization and Computer Graphics, 2000, 6(1): 24-43.

[17] ZHANG X, YUAN X R. Treemap visualization[J]. Journal of Computer-Aided Design & Computer Graphics, 2012, 24(9): 1113-1124.

[18] BALZER M, DEUSSEN O. Voronoi Treemaps [C]// IEEE Symposium on Information Visualization. Los Alamitos: IEEE, 2005: 49-56.

[19] GOU L, ZHANG X. Treenetviz: Revealing patterns of networks over tree structures[J]. IEEE Transaction on Visualization and Computer Graphics, 2011, 17(12): 2449-2458.

[20] ZHANG T, YU J, LIAO B, et al. The Construction and Analysis of Pass Network Graph Based on GraphX[J]. Journal of Computer Research and Development, 2016, 53(12): 2729-2752. (in Chinese)
张陶，于炯，廖彬，等. 基于 GraphX 的传球网络构建及分析研究[J]. 计算机研究与发展，2016，53(12)：2729-2752.