文章编号：2095-2163(2019)04-0040-06 中图分类号：TP391 文献标志码：A

基于自然最近邻的离群检测方法研究

李士果[1]，卢建云[2]，邓剑勋[2]
（1 中冶赛迪重庆信息技术有限公司 大数据事业部，重庆 401122；
2 重庆电子工程职业学院 人工智能与大数据学院，重庆 401331）

摘 要：在实际应用中，近邻技术具有简单、快速、高效的特点，受到研究人员的青睐。近来自然最近邻被提出并应用到离群检测和聚类中，鉴于自然最近邻消除了参数 k 设置的特点，本文将自然最近邻的概念应用到逆 k 最近邻、互 k 最近邻、共享 k 最近邻中，提出了自然逆最近邻、自然互最近邻和自然共享最近邻。并将提出的 3 种算法在离群点检测中进行了实验对比分析。实验结果表明自然逆最近邻和自然互最近邻能够有效发现局部和全局离群点。

关键词：近邻技术；离群点检测；自然最近邻；数据挖掘

Outlier detection methods based on natural nearest neighbors

LI Shiguo[1], LU Jianyun[2], DENG Jianxun[2]
(1 Big Data Business Section, CISDI Information Technology Co.,Ltd, Chongqing 401122, China;
2 School of AI and Big Data, Chongqing College of Electronic Engineering, Chongqing 401331, China)

【Abstract】 Near neighbor technique has been one of the important technologies in machine learning and data mining. It has the characteristics of simplicity, speed and efficiency in practical applications, which is favored by researchers. The existing near neighbor technologies include k-nearest neighbors, reverse k-nearest neighbors, mutual k-nearest neighbors, shared k-nearest neighbors, and natural nearest neighbors. In order to gain a comprehensive understanding of the characteristics of several near neighbor technologies, the mentioned near neighbor technologies are compared and analyzed. First, this article gives a comparative analysis of the near neighbor technologies and points out the special characteristics; Second, the concept of natural nearest neighbors is applied into other near neighbor technologies and natural reverse nearest neighbors, natural mutual nearest neighbors, and natural shared nearest neighbors are proposed; Finally, the proposed three algorithms are compared experimentally in outlier detection. The experimental results show that natural reverse nearest neighbors and natural mutual nearest neighbors can effectively discover local and global outliers.

【Key words】 nearest neighbors; outlier detection; natural nearest neighbors; data mining

0 引 言

近邻技术是数据挖掘和机器学习的重要技术之一，被广泛地应用到分类、聚类、异常检测、协同过滤、图像处理等研究领域。在实际应用中，近邻技术具有简单、快速、高效的特点，一直以来受到研究人员的重视。目前存在的近邻技术有 k 最近邻、互 k 最近邻、逆 k 最近邻、共享最近邻、自然最近邻。这些近邻技术各有特点，自提出以来都得到了广泛的应用。近来自然最近邻被应用到离群检测和聚类中，并取得了比较好的效果。鉴于自然最近邻能够消除参数 k 设置，本文将自然最近邻这一特点应用到其它近邻中，提出了自然逆最近邻、自然互最近邻、自然共享最近邻的概念，并给出了相应的算法描述。在离群检测应用中，对提出的自然逆最近邻、自然互最近邻、自然共享最近邻算法进行了实验对比分析，实验结果表明自然逆最近邻和自然互最近邻能够有效发现局部和全局离群点。

本文的主要贡献如下：

(1)结合自然最近邻消除参数 k 设置的特点，提出了无参的自然逆最近邻、自然互最近邻、自然共享最近邻的概念，并给出了相应的算法描述；

(2)在离群检测应用中，对提出的算法进行了实验对比分析，实验结果表明自然逆最近邻和自然互最近邻能够有效地检测局部和全局离群点。

基金项目：重庆市教委 2018 科技青年项目(KJQN201803109)。

作者简介：李士果(1983-)，女，硕士，工程师，CCF 会员，主要研究方向：大数据处理与分析；卢建云(1982-)，男，博士，副教授，CCF 会员，主要研究方向：数据挖掘、机器学习；邓剑勋(1978-)，男，博士，教授，主要研究方向：数字图像处理、机器学习。

通讯作者：卢建云 Email：lujianyun@cqu.edu.cn

收稿日期：2019-04-03

1　相关研究

最近邻概念直观、简单，认为一个对象与离其最近的对象具有相同的特点，最初被应用于基于示例的学习方法中。从概率的角度出发，可以通过多个对象投票的方式来确定被分类模式的标签，则引入了 k 最近邻的概念。k 最近邻有非常广泛的用途，例如模式分类、机器学习、数据挖掘等领域[1]。在模式分类中，k 最近邻作为分类器对待识别模式进行分类，考虑每个近邻的贡献不同，出现了基于权重的 kNN[2] 分类方法。在图像分割中，kNN 与贝叶斯网络结合，出现了贝叶斯 kNN 图像分割算法[3]，还有 kNN 作为启发的区域迭代的图像分割算法[4]。在数据挖掘中，k 最近邻与距离结合衡量数据集中对象的密度，被用来进行离群点检测。对于聚类应用，k 个最近邻能够形成子图，通过相似度度量函数，不断合并相似的子图，直到满足需要得到的子图数目，从而实现聚类。互 k 最近邻比 k 最近邻要严格，加入了一个限制条件，即要求 2 个对象 p 和 q 出现在对方的 k 最近邻域，则 p 和 q 为互 k 最近邻。通常，构建的互 k 最近邻图更加紧密，常被应用到去噪、分类中[5]。共享 k 最近邻进一步考虑了对象 p 和 q 邻域共有的近邻数目，采用量化的方法来度量两个对象的相似程度，常被应用到离群检测、聚类中[6]。逆 k 最近邻与 k 最近邻具有相反的概念，是考虑一个对象对其周围对象的影响，被应用于决策支持、资源定位与营销等研究领域[7]。目前，对逆 k 最近邻的研究主要集中在如何高效地进行搜索[8]。

最近提出的自然最近邻概念具有自适应的特点，不需要显示地指定参数 k 的值，而是通过迭代计算得到给定数据集的自适应 k 值。自然最近邻被应用到高维数据结构学习、离群检测、分类、聚类等场景[9]。自然最近邻在形成的过程中，带有密度信息和离群信息，能够被应用到局部密度估计和离群检测[10]。

2　近邻技术

本节中，给出了互 k 最近邻、逆 k 最近邻、共享 k 最近邻和自然最近邻的定义，并对各种近邻的特点进行了分析。

2.1　互 k 最近邻

在 k 最近邻的基础上，增加一个限制条件，即 2 个对象要互为 k 最近邻，这就是互 k 最近邻的思想。2 个对象互为 k 最近邻，恰恰反应了 2 个对象之间的紧密程度。相对于 k 最近邻反映的是单边关系，而互 k 最近邻描述的是一种双边关系。

定义 1　互 k 最近邻：给定包含 m 个对象的集合 $O=\{o_1,o_2,\cdots,o_m\}$，指定 k 的值，$0<k\leqslant m$，对象 o_i,o_j，$0<i,j\leqslant m$ 的 k 个最近邻表示为 $S_i=\{o_1,o_2,\cdots,o_k\}$，$S_j=\{o_1,o_2,\cdots,o_k\}$，$o_i\in S_j$ 并且 $o_j\in S_i$。

2.2　逆 k 最近邻

求解 k 最近邻时，一个对象总是返回 k 个最近邻，每个对象拥有最近邻的数目是不变的。在实际应用中，对于密度度量，处于稀疏空间的对象并不总是有 k 个最近邻。逆 k 最近邻是与 k 最近邻具有相反含义的一种最近邻概念。k 最近邻反应的是查询对象离某个被查询对象最近，而逆 k 最近邻描述的是被查询对象对数据集中其它查询对象的影响，被查询对象能够影响的对象数目可能有一个、多个或者是没有。

定义 2　逆 k 最近邻：给定包含 m 个对象的集合 $O=\{o_1,o_2,\cdots,o_m\}$，指定 k 的值，$0<k<m$，对象 o_i，$0<i\leqslant m$，$0\leqslant l\leqslant m$ 的逆 k 最近邻表示为 $RkNN_i=\{o_1,o_2,\cdots,o_l\}$，满足 $o_i\in kNN(o_l)$。

2.3　共享 k 最近邻

互 k 最近邻描述了 2 个对象之间的相近或紧密程度。这种度量关系比较简单，只有紧密、不紧密，或者相近、不相近这样的二元关系。共享 k 最近邻突破了这种二元关系，采用一种可以量化的度量方法，即衡量 2 个对象共有最近邻的数目，这种量化度量方式很好地表达了相对程度。

定义 3　共享 k 最近邻：给定包含 m 个对象的集合 $O=\{o_1,o_2,\cdots,o_m\}$，指定 k 的值，$0<k<m$，对象 o_i,o_j，$0<i,j\leqslant m$ 的 k 个最近邻表示为 $S_i=\{s_1,s_2,\cdots,s_k\}$，$S_j=\{s_1,s_2,\cdots,s_k\}$，$o_i$ 与 o_j 的共享 k 最近邻为 $SkNN=S_i\cap S_j$。

2.4　自然最近邻

在 k 近邻技术中，参数 k 是需要设置的，对于不同的应用场景，参数 k 的设置也不尽相同，往往通过经验分析得到合适的参数值。为了减轻参数 k 对最近邻计算结果的影响，邹咸林等人提出了自然最近邻概念[9]。求解自然最近邻时，不需要指定参数 k 或者邻域半径 ε，其是一种无尺度的最近邻概念。求解自然最近邻的核心思想是设置计算的终止条件，整个计算过程是对给定数据集的一个自适应过程，当迭代计算收敛时，得到数据集中每个对象的自然最近邻。自然最近邻数目是一种量化的度量方法，能够反映数据集疏密分布情况。求解自然最近邻的方式可以有多种，表现为迭代计算的终止条件

不同。

定义 4 自然最近邻:给定包含 m 个对象的集合 $O=\{o_1,o_2,\cdots,o_m\}$,对于对象 $o_i,0<i\leqslant m$,若有对象 $o_j,0<j\leqslant m$ 的最近邻路径经历 o_i,且当 O 中最离群的对象都有最近邻路径到达时,则称 o_j 为 o_i 的自然最近邻。

3 自然最近邻在其它近邻中的应用

本节中,给出了自然逆最近邻、自然互最近邻和自然共享最近邻的定义和算法描述。自然最近邻概念的要点是"数据集中最离群的数据对象都至少有一条路径到达",这个特点使得每个对象具有不固定尺度的近邻数目,也减轻了参数 k 对最终近邻数目的影响。

3.1 自然最近邻定义

定义 5 自然逆最近邻:给定包含 n 个对象的集合 $S=\{s_1,s_2,\cdots,s_n\}$,当任意对象 $s_i,0<i\leqslant n$ 都有逆最近邻时,s_i 的逆最近邻称为自然逆最近邻。

定义 6 自然互最近邻:给定包含 n 个对象的集合 $S=\{s_1,s_2,\cdots,s_n\}$,当任意对象 $s_i,0<i\leqslant n$ 都有互最近邻时,s_i 的互最近邻称为自然互最近邻。

定义 7 自然共享最近邻:给定包含 n 个对象的集合 $S=\{s_1,s_2,\cdots,s_n\}$,指定共享最近邻数目 m,当任意对象 $s_i,s_j,0<i,j\leqslant n,i\neq j$,存在 $s_i\cap s_j\geqslant m$ 时,s_j 称为 s_i 的自然共享最近邻。

在定义 7 中,自然共享最近邻不是指2个对象p和q共有的近邻对象,而是指 p 和 q 的一种关系,如果 p 和 q 共有的近邻数目大于等于 m,则 p 和 q 是一种自然共享最近邻的关系。下面给出定义 5 到定义 7 的算法实现描述。

3.2 自然最近邻算法描述

算法 1 自然逆最近邻算法(NRNN)

输入:数据集 D 包含 n 个对象

输出:D 中每个对象的自然逆最近邻数目

算法描述如下:

初始化变量 $r=1$,向量 $nrnn(i)=0,0<i\leqslant n$;

```
while r
  for each p in D
     计算 p 的第 r 最近邻 q;
     nrnn(q) = nrnn(q) + 1;
    if all( nrnn(i) ≠ 0)
       r = 0;
    else
    r=r+1
  end
end
```

在算法 1 中,第 5 行给出了算法的终止条件,即数据集 D 中的每个对象都至少有一个逆最近邻。在实际应用中,对于包含一些相对离群对象的数据集,算法的收敛性变的很差。因此,在算法的具体实现中,加入了另外一个终止条件,即在迭代过程中数据集中没有逆最近邻的数据对象的数目连续两次没有变化,则算法停止。该终止条件对算法的本质并没有影响,因为算法的目的是统计数据集中对象的逆最近邻数目,逆最近邻数目反映了数据对象的密度分布信息。算法提前终止说明数据集中的一些对象处于相对稀疏的区域,这对数据集整体的密度分布并没有影响。在下面 2 个算法的具体实现中,也加入了类似的终止条件。

算法 2 自然互最近邻算法(NMNN)

输入:数据集 D 包含 n 个对象

输出:D 中每个对象的自然互最近邻数目

算法描述如下:

初始化变量 $r=1$,向量 $\mathrm{nmnn}(i)=0,0<i\leqslant n$;

```
while r
  for each p in D
     计算 p 的第 r 最近邻 q;
  end
  for each p,q in D
      if ismember(p,rNN(q)) and ismember(q,
rNN(p))
          nmnn(p) = nmnn(p) + 1
          nmnn(q) = nmnn(q) + 1
    if all( nmnn(i) ≠ 0)
      r = 0;
    else
       r = r + 1;
  end
end
```

在算法 2 中,第 7 行的 if 条件中函数 ismember (p,rNN(q))表示对象 p 是对象 q 的 r 最近邻,函数 ismember (q,rNN(p)) 表示 q 是 p 的 r 最近邻,当这 2 个条件同时成立,p 和 q 为互最近邻。第 10 行是算法的终止条件,即数据集中的每个对象都至少有一个互最近邻。

算法 3 自然共享最近邻算法(NSNN)

输入:数据集 D 包含 n 个对象

输出:D 中每个对象的自然共享最近邻数目

初始化变量 $r=m$,向量 $nsnn(i)=0,0<i\leqslant n$;

```
while r
  for each p in D
    计算 p 的第 r 最近邻 q;
  end
  for each p,q in D
    snn = intersect(rNN(p),rNN(q))
    if length(snn) > m
      nsnn(p) = nmnn(p) + 1
      nsnn(q) = nmnn(q) + 1
    if all(nsnn(i) ≠ 0)
      r = 0;
    else
      r = r + 1;
end
end
```

在算法 3 中，第 7 行函数 *intersect*(*r*NN(p)，*r*NN(q)) 为计算对象 p 和对象 q 的共享最近邻数目，第 8 行判断该数目是否大于阈值 m，如果成立，p 和 q 为自然共享最近邻。第 11 行是算法的终止条件，即数据集中每个对象至少有另外一个对象与其的共享最近邻数目是大于等于 m 的。

4　实验与结果

本节将上述提出的 3 种自然最近邻算法在离群检测应用中进行了实验对比分析。

4.1　实验数据集

本次实验采用了 2 个人工合成数据集 DS1 和 DS2，分布包含 6 个、11 个离群点。实验数据集的具体信息见表 1。

表 1　实验数据集的基本信息

Tab. 1　Basic information of experimental dataset

数据集名称	维度	大小	离群点数目
DS1	2	74	6
DS2	2	122	11

4.2　评价指标

采用精确率(Precision)、召回率(Recall)和 F-Measure 对 3 种算法的实验结果进行评价，3 种评价指标的解释如下：

精确率=(表示模型预测为所有正样本数量中真正为正样本的比例)；

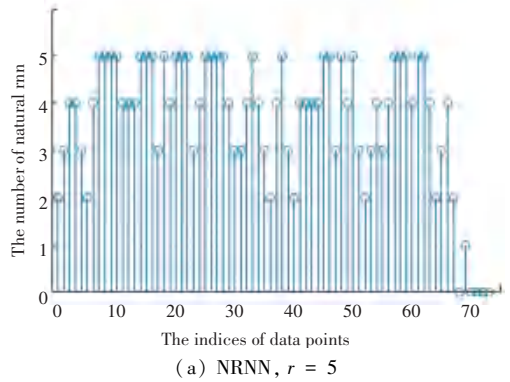
召回率=(表示模型准确预测为正样本的数量占所有正样本数量的比例)；

$$F-Measure=\frac{(a^2+1)(\text{精确率}*\text{召回率})}{a^2(\text{精确率}+\text{召回率})}$$

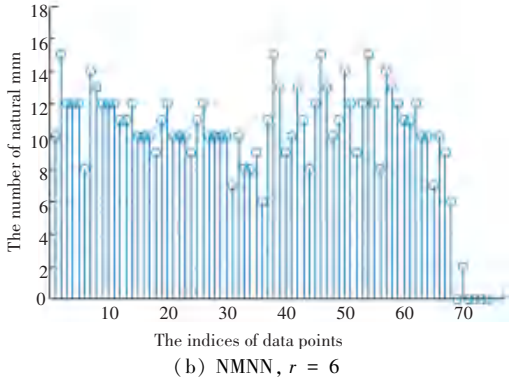
当 $a=1$ 时，$F1=2*(\text{精确率}*\text{召回率})/(\text{精确率}+\text{召回率})$。

4.3　实验和结果

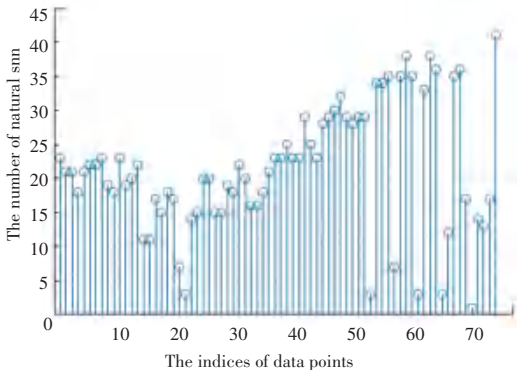
通过 2 个实验对 3 种自然最近邻算法 NRNN、NMNN 和 NSNN 进行对比分析。第一个实验分析 3 种自然最近邻数目在 DS1 和 DS2 数据集上的统计分布情况，如图 1 和图 2 所示。统计 3 种自然最近邻数目分布的目的，是通过分布反映数据集中对象所在区域的疏密情况。通常情况下，拥有很多自然最近邻的对象处于数据集中比较密集的区域，具有很少自然最近邻的对象位于数据集中相对稀疏的区域，这与利用密度进行离群检测的方法吻合，通常离群点是位于数据集中相对稀疏的区域。



(a) NRNN, $r=5$



(b) NMNN, $r=6$

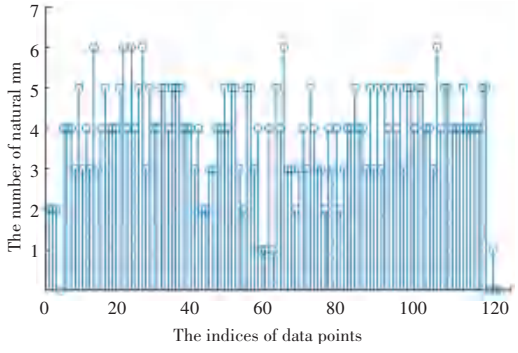


(c) SNN>9, $r=14$

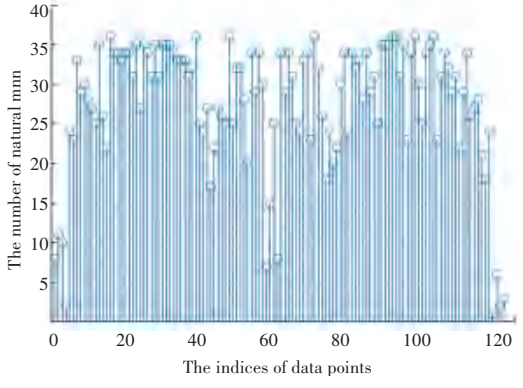
图 1　NRNN、NMNN 和 NSNN 在 DS1 数据集上的分布

Fig. 1　Distribution of NRNN、NMNN and NSNN on DS1 dataset

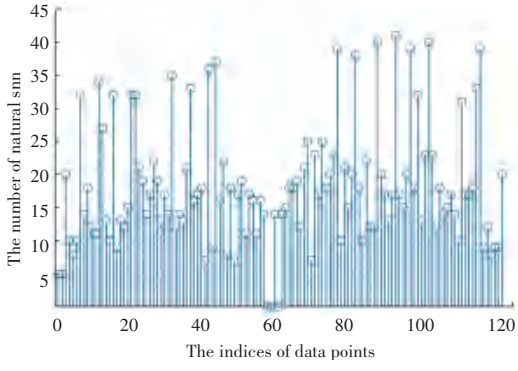
图 1 显示了 NRNN、NMNN 和 NSNN 在 DS1 数据集上的统计分布。由此可以看出,当 r 值为 5 时,NRNN 算法收敛,此时 DS1 中不是每个数据对象都至少有一个自然逆最近邻。算法收敛是因为在连续 2 次迭代的过程中,没有自然逆最近邻的对象数目没有变化。同理,在 r 值为 6 时,NMNN 算法收敛。NSNN 算法设置的共享最近邻数目阈值为 9,当 r 值为 14 时,DS1 中每个对象都至少有一个自然共享最近邻。



(a) NRNN, r = 6



(b) NMNN, r = 9



(c) SNN>9, r = 15

图 2 NRNN、NMNN 和 NSNN 在 DS2 数据集上的分布

Fig. 2 The distribution of NRNN、NMNN and NSNN on DS2 dataset

图 2 显示了 NRNN、NMNN 和 NSNN 3 种算法在 DS2 数据集上的统计分布。可以看出,NRNN 算法在 r = 6 时收敛,收敛时,数据集中仍有几个对象没有自然逆最近邻。当 r 取值为 9 时,NMNN 算法收敛,此时数据集中的所有对象都至少有一个自然互最近邻。对于 NSNN 算法,共享最近邻数目阈值设置为 9,当 r 值为 15 时,NSNN 算法收敛,此时,数据集中仍有几个对象没有自然共享最近邻。

离群点通常处于数据集中稀疏的区域,根据自然最近邻数目在数据集上的分布,通过设置阈值找到拥有相对少的自然最近邻数目的对象,从而实现离群检测。实验对比结果见表 2 和表 3。

表 2 NRNN,NMNN 和 NSNN 在 DS1 数据集上的实验对比结果

Tab. 2 The experimental results comparison of NRNN、NMNN and NSNN on DS1 dataset %

	NRNN			NMNN			NSNN		
	=0	<2	<3	=0	<5	<10	<5	<10	<15
精确率	100	100	46.2	100	100	28.6	25	16.7	14.3
召回率	83.3	100	100	83.3	100	100	16.7	16.7	33.3
$F1$	90.9	100	62.2	90.9	100	44.5	20	16.7	20

表 3 NRNN,NMNN 和 NSNN 在 DS2 数据集上的实验对比结果

Tab. 3 The experimental results comparison of NRNN、NMNN and NSNN on DS2 dataset %

	NRNN			NMNN			NSNN		
	=0	<2	<3	=5	<10	<17	<5	<8	<11
精确率	100	100	64.7	100	100	100	100	62.5	40.9
召回率	36.4	72.7	100	36.4	72.7	100	27.3	45.5	81.8
$F1$	53.4	84.2	78.6	53.4	84.2	100	42.9	52.7	54.5

从表 2 可以看出, NRNN 和 NMNN 能够取得 100%的精确率和召回率,NRNN 在阈值小于 2 时,$F1$ 指标为 100%,NMNN 在阈值小于 5 时, $F1$ 指标为 100%。通过整体对比,NRNN 略优于 NMNN。NSNN 在精确率、召回率和 $F1$ 指标上的结果都相对比较低。

分析表 3 得出,NMNN 在阈值小于 17 时,取得了 100%的精确率和召回率,整体来看,NMNN 略优于 NRNN。NSNN 在阈值小于 5 时,取得了 100%的精确率,但只有 27.3%的召回率,说明挖掘出的离群点很少。NSNN 阈值小于 11 时,取得了 81.8%的召回率。

总体来看,NMNN 和 NRNN 在离群检测应用中取得了比较好的效果,NSNN 在 DS2 数据集上可以实现离群检测,但在 DS1 数据集上表现不太理想,这与数据集中数据分布的相对密度有关。NRNN 倾向于发现处于稀疏区域和处于数据集外边缘的对

(下转第 50 页)