广西科学院学报
Journal of Guangxi Academy of Sciences
2017,33(1):19～24
Vol. 33, No. 1 February 2017

网络优先数字出版时间:2017-02-15 【DOI】10.13657/j.cnki.gxkxyxb.20170215.002
网络优先数字出版地址:http://www.cnki.net/kcms/detail/45.1075.N.20170215.1659.004.html

基于局部密度和纯度的自适应 k 近邻算法*
Adaptive k Neighbor Algorithm based on Local Density and Purity

张 兵,蒙祖强,沈亮亮,李虹利
ZHANG Bing, MENG Zuqiang, SHEN Liangliang, LI Hongli

(广西大学计算机与电子信息学院,广西南宁 530004)
(School of Computer, Electronics and Information in Guangxi University, Nanning, Guangxi, 530004, China)

摘要:【目的】针对 K 最近邻(K-Nearest Neighbor,KNN)算法中 k 值的选取通常是人为设定,而且通常是固定的缺点,研究如何更好地选取 k 值。**【方法】**引入 k 的可信度的概念,提出一种基于局部密度和纯度的自适应选取 k 值的方法,并将其引入到传统的 KNN 分类算法中。**【结果】**该算法合理的考虑了样本的局部密度、纯度与选取 k 值的关系,不仅解决了 k 值的选取问题,并且避免了固定 k 值对分类的影响。**【结论】**该算法是有效的,可以得到较高的准确率,但算法的时效性有待提高。

关键词: k 的可信度 自适应 k 值 KNN 分类
中图分类号: TP18 **文献标识码:** A **文章编号:** 1002-7378(2017)01-0019-06

Abstract:【Objective】Aiming at the selection of parameter k value(usually fixed) in KNN algorithm is usually set by users, we should study how to better select k values. **【Methods】**This paper introduces the concept of the credibility of k, and proposes an improved adaptive selection of k values based on the local density and purity, and introduces into the traditional KNN classification algorithm. **【Results】**The algorithm is reasonable to consider the relationship between the local density and purity and the selection of k values, which not only solves the problems of choosing k values, but also avoids the influence of fixed k value on classification. **【Conclusion】**The algorithm is effective and can get higher accuracy, and the timeliness is also enhanced.

Key words: credibility of k, adaptive k, KNN classification

0 引言

【研究意义】K 最近邻(K-Nearest Neighbor,KNN)算法是数据挖掘“十大经典算法”之一[1]。KNN 算法是在一组历史数据记录中,寻找一个或者若干个与当前记录最相似的历史记录的已知特征值,来预测当前记录的未知或者遗失特征值[2]。KNN 是基于统计的分类方法[3],如果待分类样本在特征空间中的 k 个最相似(即特征空间中最邻近)的样本中的大多数样本属于某一个类别,则该样本也属于这个类别。因此具有简单直观、无需先验统计知识、性能优越的特点,得到了广泛应用。但是传统的 KNN 算法是一种懒惰的学习方法,具有以下缺点:1)在样本较大以及特征属性较多时,分类的效率就大大降低;2)参数 k 值只能由经验设置,并且对某

收稿日期: 2016-12-20
作者简介: 张 兵(1991－),女,硕士研究生,主要从事数据挖掘和粗糙集研究,E-mail:1365026003@qq.com。
*国家自然科学基金项目(61363027)和广西自然科学基金项目(2015GXNSFAA139292)资助。

一个数据集分类效果较好的取值对于其他数据集可能并没有很好的分类效果。**【前人研究进展】**为了克服传统KNN算法的缺点，学者们提出了许多针对KNN的改进算法。这些算法大致可以分为两类。一类是通过优化或者降维来减少样本之间相关性的计算，以提高分类的效率。如胡元等[4]提出了一种基于区域划分的KNN文本快速分类算法，该算法是将训练样本集按空间分布划分成若干个区域，然后根据测试样本与各个区域之间的位置关系快速查找其k个近邻，大大降低了KNN算法的计算量。林啟锋等[5]提出结合同义向量聚合和特征多类别的改进KNN分类算法，该算法明显提高了文本分类效率和分类的精度。耿丽娟等[6]提出多层差分KNN算法，该算法对已知样本数据类域进行分层，大大降低了无效的计算量，提高了分类的准确性。另一类是通过改进参数k取值的方法，提高分类准确率。Sun和Huang[7]提出对测试样本用其最近邻样本的k值。首先对训练集中的每一个样本训练一个能将它正确分类的最小的k值，然后对测试样本计算最近邻，选用其最近邻的k值来作KNN分类。孙可等[8]提出引入稀疏学习理论，利用训练样本重构测试样本的方法。重构过程中使用样本间的相关性，利用投影变换矩阵w确定KNN算法中的k值。该算法的缺点是k值是全局的，即对所有的测试样本都用同一个k值，这对于分布不均匀的样本集是不合理的，会影响分类的准确率。杨柳等[9]提出一种自适应的大间隔近邻分类算法，该算法将自适应选择k值引入到大间隔近邻分类算法中，减少了k值的选择对分类性能的影响。黄少滨等[10]提出一种自适应最近邻的聚类融合算法，该算法能够根据数据分布的密度，为每一个数据点自动的选择合适的最近邻数。**【本研究切入点】**在上述研究的基础上，针对现有的KNN算法选取k近邻时采用固定k值的缺点，对每一个测试样本选取不同的k近邻，将自适应选取k值的方法引入到KNN算法中，提出了一种基于局部密度和纯度的自适应k近邻算法。本研究直接考虑样本间的相关性，在KNN算法中直接由数据本身驱动选取k值，避免了人为设定k值对分类准确率的影响。**【拟解决的关键问题】**对于不同的测试样本k的取值也不固定，通过计算测试样本局部密度和纯度来间接地控制k的取值，使得每个测试样本都由它的k(不固定)个相关的最近邻样本预测，分类的准确性得到提高。

1 KNN算法

KNN算法[11]是Cover和Hart于1968年提出的，该算法的思路：1)计算出待分类样本与已知类别的训练样本之间的距离或者相似度；2)找到距离或者相似度与待分类样本数据最近的k个邻居；3)根据这k个邻居所属的类别来判断待分类样本数据的类别。如果待分类样本数据的k个邻居中大多数属于某一个类别，那么待分类的测试样本也属于这个类别。

定义1(相似度) 向量空间模型下，我们把样本表示成向量，测试样本$X=\{X_1,X_2,X_3\cdots X_m\}$，$Y=\{Y_1,Y_2,Y_3\cdots Y_n\}$，$X_i=(X_{i1},X_{i2},X_{i3}\cdots X_{id})$，$Y_i=(Y_{i1},Y_{i2},Y_{i3}\cdots Y_{id})$，其中，$d$是样本属性的个数，$m$是测试样本$X$的个数，$n$是训练样本$Y$的个数。

我们采用欧氏距离来确定样本的相似性。欧氏距离的计算公式为

$$\text{distance}(X_i,Y_i)=\sqrt{\sum_{j=1}^{d}(X_{ij}-Y_{ij})^2}。\quad(1)$$

由定义1可知，样本X_i与Y_i的欧氏距离度量了两个样本的相似程度。当X_i与Y_i距离越小，两个样本的相似程度越高。

KNN算法流程：

步骤1：准备数据，设定k值；

步骤2：构造一个大小为k的按距离从大到小的优先级队列priorityQueue，存储最近邻的训练元组。从训练元组中选取前k个元组作为最近邻元组的初始值，用公式(1)分别计算测试元组到这k个元组的距离，然后将训练元组的序号和距离存入priorityQueue；

步骤3：遍历整个训练元组集，计算测试元组与当前训练元组的距离L，将L与优先级队列中的最大距离L_{max}进行比较。如果$L\geqslant L_{max}$，那么舍弃该训练元组，遍历下一个训练元组。如果$L<L_{max}$，则删除优先级队列中最大距离的元组，将当前训练元组存入priorityQueue中；

步骤4：遍历结束，计算priorityQueue中k个元组的多数类，并将其作为测试元组的类别；

步骤5：测试元组集测试完毕后计算分类准确率，继续设定k值重新训练，最后取分类准确率最高的k值。

针对传统的k近邻算法中k需要人工指定的缺点，孙可等[8]提出引入稀疏学习理论，利用训练样本

重构待分类样本的方法,寻找投影变换矩阵 W,然后利用得到的 W 确定待分类样本分类所需的 k 值,进而进行 k 近邻分类,完成了 k 值是根据样本特性自动选定的功能。但该算法也有不足之处:k 值是固定的,也就是说对于每一个待分类样本使用的 k 值都一样。但是在实际应用中,这种采取固定 k 值的方法经常是不合理的。

如图1所示,当 $k=5$ 时,根据 k 近邻算法可以得到2个样本空间(图1中两个实线圆圈)。其中待分类样本1的 k 个最近邻样本选取比较合理,但对于待分类样本2而言,实际上令 $k=3$ 更为合理(如虚线圆圈所示),因为虚线圆圈中样本的密度明显比实线圆圈中样本的密度大,如果把它们也作为最近邻样本,很可能会影响到分类准确率。因此本研究认为对于不同的测试样本,k 值的选取应该根据样本的密度来决定。为了更准确地学习 k 值,还应当考虑类别的纯度和 k 的取值之间的关系。

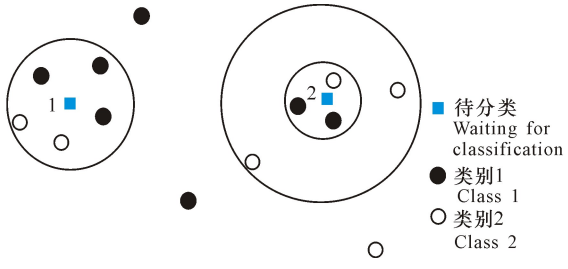

图1　$k=5$ 时测试样本的分类情况[12]

Fig. 1　An example of KNN classification task with $k=5$[12]

2　改进的 KNN 算法

2.1　相关概念

定义2　设 k 代表近邻样本数,k_i 代表近邻样本中最大类别的样本个数,d 为待测样本与第 k 个近邻样本的距离。称 $F(k)$ 为待测样本的 k 的可信度,计算公式如下:

$$F(k)=k/d+k_i/k。\tag{2}$$

如果 $F(k)$ 越大,k 的可信度越高。公式(2)由两部分组成:第一,距离待测样本单位长度上样本的个数(也就是样本的密度),如果密度越大,k 的可信度越高;第二,待测样本的 k 近邻样本中最大类别所占的比重(也就是纯度),如果纯度越大,k 的可信度越高。k 的可信度越高,待测样本的近邻样本数的取值越好。

为了使 KNN 算法中 k 值不再固定,本研究从 k 值的选取进行改进。算法的思路:对待测样本的每一个近邻样本计算 k 的可信度,利用可信度来控制 k 的取值。

2.2　算法的步骤

基于局部密度和纯度的自适应 k 近邻算法的步骤如下:

输入:测试样本 $X=\{X_1,X_2,X_3\cdots X_m\}$,训练样本 $Y=\{Y_1,Y_2,Y_3\cdots Y_n\}$。

输出:测试样本 X 的类别属性。

步骤1:用公式(1)计算出测试样本 X_i 与每一个训练样本 Y_i 的距离 distance;

步骤2:对每一个测试样本 X_i 按照距离 distance 大小排序,然后依次保存在数组 juli[]中;

步骤3:对于任意的测试样本 $X_i(i\in[1,m])$

```
//设 h 的初始值为-1
for j = 1 :10
{ if  juli[j]= = 0 && juli [j+1]= = 0
    j++
else if juli[j]= = 0 && juli [j+1]! = 0
    break
else   //设 max 为 F(j)的最大值, h 是 max 对应的 k 值
  { if max > F(j)
      { max = F(j)
h= j }
else
      j++}
}
        if(h = = -1)  k[i] = j;
else      k[i] = h;
```

得到每一个待测试样本 X_i 对应的的 k 值 $k[i]$;

步骤4:对于任意的测试样本 X_i 计算出 k 个元组的多数类,并将其作为测试元组的类别。

2.3　算法的实现

下面用一个例子说明算法的实现过程。

例1　表1是 Glass 数据集的一部分,随机抽出15条数据(X_1-X_{15})为训练集,后两条(X_{16} 和 X_{17})为测试数据,前9个属性为条件属性,最后一个属性为类别。

下面用本文的算法来判断 X_{16} 和 X_{17} 的类别。

表1 训练集与测试样本

Table 1 Training set and testing samples

数据 Data	条件属性 Condition attribute									类别 Class
	A	B	C	D	E	F	G	H	I	
X_1	1.52101	13.64	4.49	1.10	71.78	0.06	8.75	0.00	0.00	1
X_2	1.51761	13.89	3.60	1.36	72.73	0.48	7.83	0.00	0.00	1
X_3	1.51618	13.53	3.55	1.54	72.99	0.39	7.78	0.00	0.00	1
X_4	1.51766	13.21	3.69	1.29	72.61	0.57	8.22	0.00	0.00	1
X_5	1.51742	13.27	3.62	1.24	73.08	0.55	8.07	0.00	0.00	1
X_6	1.51574	14.86	3.67	1.74	71.87	0.16	7.36	0.00	0.12	2
X_7	1.51848	13.64	3.87	1.27	71.96	0.54	8.32	0.00	0.32	2
X_8	1.51593	13.09	3.59	1.52	73.10	0.67	7.83	0.00	0.00	2
X_9	1.51631	13.54	3.57	1.57	72.87	0.61	7.89	0.00	0.00	2
X_{10}	1.51596	13.02	3.56	1.54	73.11	0.72	7.90	0.00	0.00	2
X_{11}	1.51665	13.14	3.45	1.76	72.48	0.60	8.38	0.00	0.17	3
X_{12}	1.52127	14.32	3.90	0.83	71.50	0.00	9.49	0.00	0.00	3
X_{13}	1.51779	13.64	3.65	0.65	73.00	0.06	8.93	0.00	0.00	3
X_{14}	1.51610	13.42	3.40	1.22	72.69	0.59	8.32	0.00	0.00	3
X_{15}	1.51694	12.86	3.58	1.31	72.61	0.61	8.79	0.00	0.00	3
X_{16}	1.51832	13.33	3.34	1.5	72.14	0.56	8.99	0.00	0.00	3
X_{17}	1.51743	13.30	3.60	1.14	73.09	0.58	8.17	0.00	0.00	1

步骤1:计算测试样本 X_{16} 与训练样本间的距离。由公式(1)得到 X_{16} 与训练样本 X_1-X_{15} 的欧式距离分别为{2.049,2.114,2.3,1.013,1.902,5.394,1.034,2.4,1.8,2.3,0.615,2.771,1.978,0.866,0.595}。

步骤2:按距离大小排序:{0.595,0.615,0.866,1.013,1.034,1.8,1.902,1.978,2.049,2.114,2.3,2.3,2.4,2.771,5.394}

步骤3:按公式(2)计算测试样本 X_{16} 对应的 k 的可信度,选取出可信度最大的 k 值来选取近邻样本。则

$F(1)=1/0.595+1/1=2.68$,

$F(2)=2/0.615+2/2=4.25$,

$F(3)=3/0.866+3/3=4.46$,

$F(4)=4/1.013+3/4=4.69$,

$F(5)=5/1.034+3/5=5.44$,

$F(6)=6/1.8+3/6=3.83$,

$F(7)=7/1.902+3/7=4.109$,

$F(8)=8/1.978+4/8=4.54$,

$F(9)=9/2.049+4/9=4.84$,

$F(10)=10/2.114+4/10=5.13$。

比较求得的各个 k 的可信度,可判断出当 $k=5$ 时,可信度最大。所以测试样本 X_{16} 对应的 k 值为5。

步骤4:根据上一步求得的 k 值,可判断测试样本 X_{16} 的类别为3,得到正确的类别判断。

同理,计算出 X_{17} 对应的 k 值为1,判断出测试样本 X_{17} 类别为1,得到正确的类别判断。

对于测试样本 X_{16} 和 X_{17},根据传统的KNN算法,所有测试样本的近邻数是一样的,如果 k 的取值都是5,那么对于 X_{16} 没什么影响,可以得到正确的分类结果。但是对于 X_{17},可以得到5个近邻:X_5,X_{14},X_4,X_8,X_9,其中2个为类别1,2个为类别2,1个为类别3,可能会得到错误的分类结果。

通过上述例子的演算结果可以看出,对于不同的测试样本选择合适的 k 值,能够更好地判断出测试样本的类别。通过综合考虑测试样本的局部密度以及最大类所占的比重,可以使测试样本选择可信度高的 k 值,使得测试样本的 k 值是通过学习样本的相关性得到的,而不是人为设定的,对于不同的测试样本选取的 k 值也不固定,从而提高了分类的准确率。

3 实例验证

3.1 数据来源

本研究选用的7个数据集全部来自UCI数据库(http://archive.ics.uci.edu/ml/datasets.html),考虑到选取数据集的一般性,我们选取了二类数据集和多类数据集。实验采用eclipse软件,在PC机上进行编程操作。数据集信息统计如表2所示。

3.2 评价方法及指标

实验采用五折交叉验证法(Cross validation)进

行分类准确度评价。数据集被随机均分成 5 个子集,本研究提出的算法在每个数据集上运行 5 次,每次取一个子集作为测试集,其余的 4 个子集作为训练集,然后取 5 次实验结果的平均值作为该数据集的分类结果。

表 2　数据集信息统计

Table 2　Data set information statistics

数据集 Data set	样本数 Number of samples	属性数 Number of attributes	类别数 Number of classes
Zoo	101	17	7
Ionosphere	351	35	2
Iris	150	4	3
Letter	20 000	17	26
Soybean	307	36	19
Wine	178	13	3
Glass	214	10	7

本实验采用的评价指标是分类的正确率,正确率越高表明分类的效果越好。分类正确率的计算公式如下:accuracy $= n_{\text{correct}}/n$,其中 n_{correct} 为测试样本中正确分类的个数,n 为测试样本的总数。

将每组数据使用本文提出的算法和传统的 KNN 算法、AdaNN 算法[7] 进行比较,此处把传统的 KNN 算法称为算法 1,把 AdaNN 算法称为算法 2,把基于局部密度和纯度的自适应 k 近邻算法称为算法 3。在相同的条件下比较 3 种算法的分类准确率,为使所得结果更加准确,我们采用的是五折交叉法。令算法 1 中的 $k=5$,而算法 2 和算法 3 的 k 值是由数据驱动产生,无需事先处理。

3.3　结果与分析

如表 3 所示,表中加粗的数据表示每一个数据集对应的最好的分类准确率。对表 3 进行分析,可以发现:

1)对于表 3 中的同一组数据,算法 3 得到的分类准确率相比于算法 1 和算法 2 要高,说明本文提出的自适应 k 值的选取方法是可行的,能够得到较高的分类准确率。

表 3　分类准确率比较

Table 3　Accuracy comparison

数据集 Data set	算法 1 Algorithm 1	算法 2 Algorithm 2	算法 3 Algorithm 3
Zoo	0.881 2	**0.960 4**	**0.960 4**
Ionosphere	0.837 6	0.854 7	**0.871 8**
Iris	0.96	0.96	**0.973 3**
Letter	0.951	0.955 6	**0.957 65**
Soybean	0.814 3	0.853 4	**0.876 2**
Wine	0.719 1	0.752 8	**0.758 4**
Glass	0.686 9	0.715 0	**0.719 6**

注:表中加粗的数据表示每一个数据集对应的最好的分类准确率

Note: The bold data in the table represents the best classification accuracy for each data set

2)对于不同的数据集,无论样本个数的大小,属性个数的多少,算法 3 都能得到较理想的结果,说明本文提出的基于局部密度和纯度的自适应 k 近邻算法是可行的。

4　结论

本研究提出了一种基于局部密度和纯度的自适应 k 近邻算法,有效地解决了传统 KNN 分类算法的两个缺陷:k 值需要事先给定且固定;未考虑样本之间的相关性。该算法的思想是根据测试样本的局部密度和最大类别的纯度来计算 k 的可信度,选取 k 的最大的可信度来选择测试样本的近邻,实现了 k 值的选取由数据本身驱动,无需人为设定。因此,本算法可以用于无法通过经验或者需要长时间实验选取 k 值的情况,大幅度减少选取 k 值的时间。最后通过对 UCI 标准数据集实验验证表明,与已有的两种算法对比,本文的方法可以得到较高的分类准确率。虽然取得了一些有益的结果,但是在对不平衡数据集如何更好的选取 k 值和搜索近邻问题上仍然需要进一步深入研究和探讨。

参考文献:

[1] WU X D, KUMAR V, QUINLAN J R, et al. Top 10 algorithms in data mining[J]. Knowledge and Information Systems, 2008, 14(1): 1-37.

[2] 俞蓓,王军,叶施仁. 基于近邻方法的高维数据可视化聚类发现[J]. 计算机研究与发展, 2000, 37(6): 714-720.
YU B, WANG J, YE S R. Visual clustering for high dimensional data based on nearest neighbor[J]. Journal of Computer Research and Development, 2000, 37(6): 714-720.

[3] MITCHELL H B, SCHAEFER P A. A "soft" K-nearest neighbor voting scheme[J]. International Journal of Intelligent Systems, 2001, 16(4): 459-468.

[4] 胡元,石冰. 基于区域划分的 kNN 文本快速分类算法研究[J]. 计算机科学, 2012, 39(10): 182-186.
HU Y, SHI B. Fast kNN text classification algorithm based on area division[J]. Computer Science, 2012, 39(10): 182-186.

[5] 林啟锋,蒙祖强,陈秋莲. 结合同义向量聚合和特征多类别的 KNN 分类算法[J]. 计算机科学, 2013, 40(12): 55-58.
LIN Q F, MENG Z Q, CHEN Q L. KNN text categorization algorithm based on semantic-vector-combination and multiclass of feature[J]. Computer Science, 2013,

40(12):55-58.

[6] 耿丽娟,李星毅. 用于大数据分类的 KNN 算法研究[J]. 计算机应用研究,2014,31(5):1342-1344.
GENG L J,LI X Y. Improvements of KNN algorithm for big data classification[J]. Application Research of Computers,2014,31(5):1342-1344.

[7] SUN S L,HUANG R Q. An adaptive k-nearest neighbor algorithm[C]//Proceedings of 2010 seventh international conference on fuzzy systems and knowledge discovery. Piscataway,Shandong:IEEE,2010:91-94.

[8] 孙可,龚永红,邓振云. 一种高效的 K 值自适应的 SA-KNN 算法[J]. 计算机工程与科学,2015,37(10):1965-1970.
SUN K,GONG Y H,DENG Z Y. An efficient SA-KNN algorithm with adaptive K value[J]. Computer Engineering & Science,2015,37(10):1965-1970.

[9] 杨柳,于剑,景丽萍. 一种自适应的大间隔近邻分类算法[J]. 计算机研究与发展,2013,50(11):2269-2277.
YANG L,YU J,JING L P. An adaptive large margin nearest neighbor classification algorithm[J]. Journal of Computer Research and Development,2013,50(11):2269-2277.

[10] 黄少滨,李建,刘刚. 一种基于自适应最近邻的聚类融合方法[J]. 计算机工程与应用,2012,48(19):157-162.
HUANG S B,LI J,LIU G. Clustering ensemble algorithm based on adaptive nearest neighbors[J]. Computer Engineering and Applications,2012,48(19):157-162.

[11] LIU Y,CHEN G S. KNN algorithm improving based on cloud model[C]//Proceedings of 2010 2nd international conference on advanced computer control (ICACC). Changsha:IEEE,2010:63-66.

[12] 邓振云,龚永红,孙可,等. 基于局部相关性的 kNN 分类算法[J]. 广西师范大学学报:自然科学版,2016,34(1):52-58.
DENG Z Y,GONG Y H,SUN K,et al. A kNN classification algorithm based on local correlation[J]. Journal of Guangxi Normal University:Natural Science Edition,2016,34(1):52-58.

(责任编辑:陆 雁)