DOI:10.3969/j.issn.1673-4785.201307015
网络出版地址:http://www.cnki.net/kcms/doi/10.3969/j.issn.1673-4785.201307015.html

融合邻域信息的 k-近邻分类

林耀进[1],李进金[1,2],陈锦坤[2],马周明[2]
(1.闽南师范大学 计算机科学与工程系,福建 漳州 363000; 2. 闽南师范大学 数学与统计学院,福建 漳州 363000)

摘　要:距离度量是影响 k-近邻(KNN)法分类精度的重要因素之一。提出一种融合邻域信息的 k-近邻算法。首先,定义了样本邻域的概念,并根据邻域的影响提出 2 条相应准则;然后,在计算测试样本与训练样本的距离时,综合考虑了样本邻域所带来的影响。该算法不仅可以更加精确地刻画样本之间的距离,而且一定程度上增强了 KNN 的稳定性。该方法在 UCI 标准数据集上进行了测试,结果表明,性能优于或与其他相关的分类器相当, 并且在噪声扰动下具有较强的鲁棒性。

关键词:k-近邻; 邻域信息;分类学习;距离测量;噪音干扰
中图分类号: TP181　　**文献标志码**:A　　**文章编号**:1673-4785(2014)02-0240-04

中文引用格式:林耀进,李进金,陈锦坤, 等. 融合邻域信息的 k-近邻分类[J]. 智能系统学报, 2014, 9(2): 240-243.
英文引用格式:LIN Yaojin, LI Jinjin, CHEN Jinkun, et al. K-nearest neighbor classification algorithm fusing neighborhood information[J]. CAAI Transactions on Intelligent Systems, 2014, 9(2): 240-243.

K-nearest neighbor classification algorithm fusing neighborhood information

LIN Yaojin[1], LI Jinjin[1,2], CHEN Jinkun[2], MA Zhouming[2]
(1. Department of Computer Science and Engineering, Zhangzhou 363000, China; 2. School of Mathematics and Statistics, Zhangzhou 363000, China)

Abstract: Distance measurement is one of the important factors which affect the classification accuracy of the k nearest neighbor (KNN) algorithm. In this paper, an improved k nearest neighbor algorithm fusing neighborhood information is presented. Firstly, the concept of the instance neighborhood is defined and two criterions are presented according to neighborhood influence; then, the influence of the instance neighborhood is comprehensively considered when the distance between the testing instances and the training instances is computed. This algorithm can characterize the distance among instances more precisely, and enhance the stability of the KNN to some extent. This presented method was tested on the UCI datasets, and the results showed that this proposed technique is better than or equal to other classifiers, and it is more robust under the noise disturbance.

Keywords: k-nearest neighbor; neighborhood information; classification learning; distance measurement; noise disturbance

k-近邻法是一种非常简单有效的分类算法,广泛应用于数据挖掘和模式识别的各个领域[1-3]。其基本思想是通过计算寻找训练集中距离待分类样本最近的 k 个邻居,然后基于它们的类别信息,依据投票的原则对待分类样本的类别进行判定。k-近邻算法的分类精度很大程度受影响于样本之间距离的度量。

近几年,出现了许多改进的距离度量方法以提高 k-近邻算法的分类性能,主要分为局部距离和全局距离两大类。在传统的全局距离度量方面,针对异构特征,提出了相应的距离度量方法,如:值差度量(value difference metric, VDM)、修正的值差度量(modified value difference metric, MVDM) 和异构欧

收稿日期:2013-06-22.　**网络出版日期**:2014-03-31.
基金项目:国家自然科学基金资助项目(61303131,61379021);福建省自然科学基金资助项目(2013J01028,2012D141);福建省 A 类科技资助项目(JA12220)
通信作者:林耀进. E-mail:zzlinyaojin@163.com.

几里德—重叠度量(heterogeneous euclidean-overlap metric, HEOM)等[4-5]。另外,许多学者考虑了样本之间的权重以增强样本之间的相似性。Hu 等[6]提出一种通过梯度下降的方法估计样本之间的权重进行改进 KNN 的分类算法;Wang 等[7]提出一种简单的自适应距离度量来估算样本的权重。同时,一些学者通过属性加权或属性选择途径改进距离度量[8-9]。在局部距离度量方面,许多方法利用局部自适应距离处理全局优化问题,如:ADAMENN 中的自适应距离,WAKNN 中的权重校正度量方法及 DANN 中的差异化自适应度量方法[10-11]。

上述方法虽能有效地度量样本之间的距离,但基本上都是从单一的距离进行考虑,存在着以下缺点:1)并未考虑样本之间的邻域结构;2)易受噪声的影响;3)不能处理多模态分布问题。因此,本文受推荐中的用户群体影响概念的启发[12],提出了一种融合样本邻域信息的 k-近邻分类算法。

1　k-近邻分类法

k-近邻分类法是一种非常简单有效的用于分类学习和函数逼近的算法。给定由 n 个样本标签对组成的数据集 D, $D=\{(x_1,c_1),(x_2,c_2),\cdots,(x_n,c_n)\}$,其分类的任务在于获取映射函数 f, 使得能正确预测无标签样本。设 $N_k(x)$ 为测试样本 x 的 k-近邻集合,k-近邻分类法在于通过测试样本 x 的 k-近邻进行大众数投票进行确定 x 的标签,其公式为

$$c=\underset{c_i}{\operatorname{argmax}}\sum_{c_i\in C}\sum_{x_j\in N_k(x)} I(c_j=c_i) \tag{1}$$

式中:c_j 为样本 x_j 的类标签,$I(\cdot)$ 为指示函数,当 c_i 与 c_j 一样时,$I(c_j=c_i)=1$,否则,$I(c_j=c_i)=0$。

2　融合邻域信息的 k-近邻分类算法

2.1　邻域信息的影响

传统的 k-近邻分类算法本质上是利用样本个体与个体之间的距离(寻找对测试样本影响最大的前 k 个近似邻居)来预测测试样本的类标签,该预测只是简单地考虑样本个体之间的相似性,而忽略了样本的邻域信息。因此,在计算样本个体距离时不仅要考虑样本个体之间的距离,也要考虑样本邻域信息产生的距离。图 1 清楚地描述了样本邻域信息产生的影响作用,从图 1 可以看出,虽然样本 x_1 与 x_2 之间的距离与样本 x_1 与 x_3 之间的距离相等,即 $d(x_1,x_2)=d(x_1,x_3)$,但是样本 x_1 的邻域信息与 x_3 的邻域信息之间包含更多的大量的共同邻居,则从认识论的角度出发,$d(x_1,x_3)\geqslant d(x_1,x_2)$。根据以上分析,可以得出准则 1。

准则 1　考虑样本邻域信息的影响能更加精确地刻画样本之间的距离。

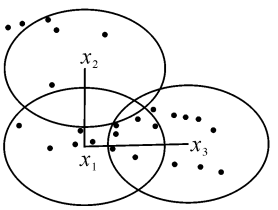

图 1　样本邻域图

Fig.1　Neighborhood of sample

2.2　样本邻域的定义

据 2.1 节分析可知,考虑样本邻域之间的距离可以更加精确地刻画样本之间的距离 Δ, 因此,寻找样本邻域对提高 k-近邻分类算法具有重要的影响。本节中给出样本的度量空间及样本邻域的定义。

定义 1[13]　给定一个 m 维的样本空间 Ω, Δ: $X^m\times X^m\to X$, 称 Δ 是 X^m 上的一个度量,如果 Δ 满足:1) $\Delta(x_1,x_2)\geqslant 0$,$\Delta(x_1,x_2)=0$, 当且仅当 $x_1=x_2$, $\forall x_1,x_2\in X^m$; 2) $\Delta(x_1,x_2)=\Delta(x_2,x_1)$, $\forall x_1,x_2\in X^m$; 3) $\Delta(x_1,x_3)\leqslant\Delta(x_1,x_2)+\Delta(x_2,x_3)$, $\forall x_1,x_2,x_3\in X^m$; 称 $\langle\Omega,\Delta\rangle$ 为度量空间。

在 N 维欧氏空间中,给定任意 2 点 $x_i=(x_{i1},x_{i2},\cdots,x_{iN})$ 和 $x_j=(x_{j1},x_{j2},\cdots,x_{jN})$, 其距离为

$$\Delta(x_i,x_j)=\left(\sum_{l=1}^{N}(x_{il}-x_{jl})^2\right)^{\frac{1}{2}} \tag{2}$$

另外,为了处理异构特征,许多学者提出了多种距离函数。如 VDM 、HVDM 和 HEOM。其中,VDM 定义为 $\mathrm{VDM}(x_i,x_j)=\sum_{l=1}^{N}d_l(x_{il}-x_{jl})$, 且 $d_l(x_{il},x_{jl})=(P(x_{il},x_{jl}))^2$, $P(x_l)$ 表示样本 x 在特征 l 下的分布概率。

定义 2　给定样本空间上的非空有限集合 $X=\{x_1,x_2,\cdots,x_m\}$, 对于 X 上的任意样本 x_i, 定义其 δ 邻域为 $\delta(x_i)=\{x\mid x\in X,\Delta(x,x_i)\leqslant\delta\}$, 其中, $0\leqslant\delta\leqslant 1$。$\delta(x_i)$ 称为样本 x_i 相应的 δ 邻域。

定义 3　给定样本空间上的非空有限集合 $X=\{x_1,x_2,\cdots,x_m\}$, 对于 X 上的任意样本 x_i 及 x_j, 根据 VDM 公式,定义样本 x_i 及 x_j 之间的邻域距离为

$$n(x_i,x_j)=\sum_{l=1}^{N}\left(\frac{|\delta_l(x_i)|}{|X|}-\frac{|\delta_l(x_j)|}{|X|}\right)^2 \tag{3}$$

性质 1[13]　给定一个度量空间 $\langle\Omega,\Delta\rangle$, 一个非空有限样本集合 $X=\{x_1,x_2,\cdots,x_m\}$。如果 $\delta_1\leqslant\delta_2$, 则有

1) $\forall x_i\in X$:$\delta_1(x_i)\geqslant\delta_2(x_i)$;

2) $\bigcup_{i=1}^{m}\delta(x_i)=X$。

根据定义 3 及性质 1,随着样本距离 δ 的增大,样本邻域中所包含的对象数量随着增加,样本之间的区分度将降低。如图 2 所示,随着距离的增大,原来不属于同一邻域的样本 x_1、x_2、x_3 变成处于同一邻域:即样本 x_1、x_2、x_3 在图 1 中不属于同一邻域,而在图 2 中则处于同一邻域。根据以上分析,可以得出准则 2。

准则 2 选择样本邻域的大小影响着样本之间距离的精确刻画。

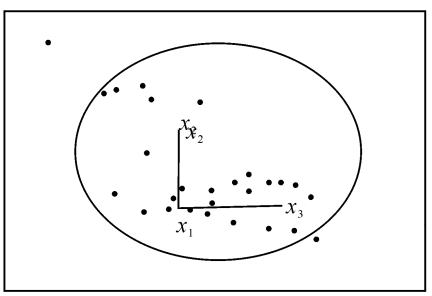

图 2 δ 减小后的样本邻域图

Fig.2 Neighborhood of sample with smaller δ

3 算法设计

在对样本邻域影响分析的基础上,利用欧式距离计算样本之间的距离,用改进的 VDM 计算样本邻域之间的距离,设计融合样本邻域信息的 k-近邻分类算法如下:

算法 1 融合样本邻域信息的 k-近邻分类算法(FK3N)。

输入:数据集 D, 测试样本 x, 距离阈值 δ, 近邻个数 k;

输出:测试样本 x 的标签 $c(x)$。

1)初始化 $c(x)=\varphi$;

2)根据式(2)获取测试样本与训练样本的 $k/2$ 个近邻 R_1;

3)根据式(3)获取测试样本邻域与训练样本邻域的 $k/2$ 个近邻 R_2;

4)获得测试样本 x 的融合近邻集合 $R=R_1\cup R_2$ 后,即测试样本 x 的 k 近邻 $N_k(x)$;

5)根据式(1)获得测试样本 x 的类标签 $c(x)$。

4 实验结果及分析

为了验证所提出算法的有效性,从 UCI 数据集中挑选了 6 组数据。其中,为了验证算法的适用性,其数据集的类别从 2 类到 3 类,特征数从 5~60,具体描述信息见表 1。同时,进行 2 组实验,第 1 组实验与经典的分类算法 KNN、NEC[13]、CART、LSVM 进行比较;另一组检测在噪声数据影响下,本文所提出的 FK3N 与其他分类器的比较。

表 1 数据描述

Table 1 Data set description

数据集	Instances	特征数	类别
Heart	270	13	2
ICU	200	20	3
Rice	104	5	2
Sonar	208	60	2
Wdbc	569	30	2
wpbc	198	33	2

实验 1 为了验证 FK3N 的分类性能,在本实验中,与其他流行的分类算法进行了比较,如表 2。

表 2 不同分类器的分类精度比较

Table 2 The comparison of classification accuracy with different classifiers

数据集	KNN	NEC	CART	LSVM	FK3N
Heart	82.59±6.06	80.00±6.10	74.07±6.30	83.33±5.31	**84.44±6.00**
ICU	92.61±2.25	86.29±17.85	79.40±31.64	92.56±4.27	**93.61±3.12**
Rice	81.69±10.30	83.07±10.96	82.07±11.70	78.16±8.10	82.98±10.90
Sonar	72.62±7.05	82.74±5.48	72.07±13.94	77.86±7.05	79.31±5.59
Wdbc	96.67±2.09	95.79±2.37	90.50±4.55	97.73±2.48	97.01±2.04
Wpbc	76.26±5.89	78.26±7.24	70.63±7.54	77.37±7.73	76.79±7.96
平均	83.74	84.35	78.12	84.50	85.69

其中将 FK3N、KNN 涉及到的参数 k 设置为 10,将 FK3N、NEL 涉及到的参数 delta 设置为 0.1。为了显示标注 FK3N 在 6 个 UCI 数据集上的分类精度,在 FK3N 中加粗的代表分类精度最高,下划线代表分类精度第 2。另外,在表 2 最后一行显示不同分类器的平均分类精度。从表 2 可以看出,FK3N 虽然只在 2 个数据集上取得最优的分类效果,但是在其他 3 个数据集上取得第 2(或并列)的分类精度。另外,在平均分类精度可以看出,FK3N 取得最高的平均分类精度,比 LSVM 还高出 1.19%。因此,从本实验可

以看出,与其他流行的分类器相比,说明了本文提出的 FK3N 算法具有较为优越的分类性能。

实验 2　为了考察 FK3N 的稳定性,在训练数据的属性中注入噪声。首先生成一个服从标准正态分布的 $m \times n$ (m 为样本数, n 为属性数)的噪声数据,然后乘以系数 a 后加入原始训练数据中。本文设 a 的值为 0.1。从表 3 可以看出,与其他分类器相比,在存在噪声情况下,FK3N 在多个数据集上的分类精度取得良好的结果。

表 3　噪声数据下不同分类器的分类精度比较

Table 3　The comparison of classification accuracy with different classifiers under noisy data

数据集	KNN	NEC	CART	LSVM	FK3N
Heart	82.22±6.49	80.37±4.96	77.78±7.61	83.33±6.11	82.96±5.30
ICU	92.61±2.25	87.29±18.03	84.19±29.92	91.55±5.44	**92.61±2.24**
Rice	81.80±6.61	81.78±7.96	77.05±10.69	77.05±3.84	**81.98±8.96**
Sonar	71.64±11.21	78.28±7.20	69.21±11.35	77.38±6.98	77.52±9.11
Wdbc	94.96±2.49	94.56±3.35	93.16±3.74	97.03±2.01	94.68±2.89
Wpbc	73.29±7.42	74.66±11.25	71.11±9.89	76.32±5.61	74.79±8.51
平均	82.71	82.82	78.75	83.78	84.09

5　结束语

本文提出了一种 FK3N 分类算法。首先,从度量空间角度定义了样本邻域信息,分析了样本的邻域对能否精确地计算样本之间的距离具有重要的影响,提出了 2 条符合实际情况的准则;然后在计算样本个体之间的距离时,综合考虑了样本邻域之间的相似性;最后提出了一种获取最近邻的计算方法。在多个公开 UCI 数据集上的实验结果表明,本文方法在原始数据和噪声数据上分类性能优于或相当于其他相关分类器。

参考文献:

[1] COVER T, HART P. Nearest neighbor pattern classification [J]. IEEE Transactions on Information Theory, 1967 (13): 21-27.

[2] 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.

[3] 吕锋, 杜妮, 文成林. 一种模糊-证据 kNN 分类方法[J]. 电子学报, 2012, 40(12): 2930-2935.

[4] WANG H. Nearest neighbors by neighborhood counting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005,28 (6): 942-953.

[5] WILSON D R, MARTINEZ T R. Improve heterogeneous distance functions [J]. Journal of Artificial Intelligence Research, 1997 (6): 1-34.

[6] HU Q H, ZHU P F, YANG Y B, et al. Large-margin nearest neighbor classifiers via sample weight learning[J]. Neurocomputing, 2011, 74 (4): 656-660.

[7] WANG J, NESKOVIC , COOPER L.N. Improving nearest neighbor rule with a simple adaptive distance measure[J]. Pattern Recognition Letters, 2007, 28 : 207-213.

[8] KOHAVI R, LANGLEY P, YUNG Y. The utility of feature weighting in nearest neighbor algorithms [C]//Proceedings of the Ninth European Conference on Machine Learning. [S. l.], 1997.

[9] SUN Y J. Iterative RELIEF for feature weighting: algorithms, theories, and applications [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29 (6): 1035-1051

[10] MU Y, DING W, TAO D C. Local discriminative distance metrics ensemble learning[J]. Pattern Recognition, 2013, 46 (8): 2337-2349.

[11] SONG Y, HUANG J, ZHOU D, et al. IKNN: informative k-nearest neighbor pattern classification[C]//PKDD 2007. [S.l.], 2007: 248-264.

[12] 林耀进, 胡学钢, 李慧宗. 基于用户群体影响的协同过滤推荐算法 [J], 情报学报, 2013, 32(3): 299-350.

[13] HU Q H, YU D R, XIE Z X. Neighborhood classifiers[J]. Expert Systems with Applications, 2008, 34 (2): 866-876.

作者简介:

林耀进,男,1980 年生,讲师,主要研究方向为数据挖掘、粒计算。

李进金,男,1960 年生,教授,博士生导师,主要研究方向为粗糙集理论及应用。