广西科学院学报
Journal of Guangxi Academy of Sciences

2010,26(4):409~411
Vol. 26, No. 4 November 2010

基于欧式距离的最近邻改进算法*
Improved kNN Algorithm Based on Euclidean Distance

刘星毅,韦小铃
LIU Xing-yi, WEI Xiao-ling

(钦州学院,广西钦州 535000)
(Qinzhou University, Qinzhou, Guangxi, 535000, China)

摘要:依托欧拉距离,使用杂合距离算法改进 Minkowski 距离公式,使得最近邻算法能够针对不同实际需要计算两事例距离,适用到属性是混合型的情形,也能避免时序列中出现的错误计算问题。

关键词:邻算法 欧式距离 Minkowski 距离
中图法分类号:TP311 **文献标识码**:A **文章编号**:1002-7378(2010)04-0409-03

Abstract: Based on Euclidean distance, a hybrid method was employed by making k Nearest Neighbor (kNN) algorithm available to calculate the distance between two instances. The proposed method can be applied for the case with all kinds of data and avoid the mistake computation in the time series data.

Key words: nearest neighbor algorithm, Euclidean distance, Minkowski distance

最近邻(NN)算法于 1967 年在文献[1]首次提出,由于容易理解,操作简单,效果明显,无论在科研还是实际生活中都具有广泛应用,目前已经被嵌入一些常见的软件中,例如,SAS 等,美国人口普查部和加拿大统计署甚至已经把这种方法作为默认的统计方法。最近邻算法的高分类准确率以及多项式的运行时间,吸引了大量研究者进行进一步求精的兴趣,例如,文献[2]通过哈希技术把最近邻算法的复杂度降低到了次线性程度,使之能轻易处理特大型的数据集。在填充缺失数据方面,最近邻方法是最热门的冷卡方法。最近邻算法中唯一需要注意的也是最重要的部分就是如何有效地计算两个事例之间的距离。实际应用中,两个事例的距离通常使用 Minkowski 距离来计算[3]。Minkowski 参数的选取对结果有着重要的影响[4,5]。同一个数据集如果选取不同的 Minkowski 参数,结果会有十分大的差异[1]。虽然最近邻方法延伸出一个扩展版本:kNN 算法[6]使这个问题得到了一定的缓解,但是没有得到根本的解决,kNN 算法还是要进行 Minkowski 参数选取,只是选取的个数有所减少而已。此外,Minkowski 距离公式在处理各种属性时只有连续型属性的效果明显,处理离散型、连续性或者各种类型属性同时存在的情况效果不很好[7,8]。Minkowski 距离公式在一些特殊数据中会出现与现实不符的问题,比如在时序研究中经常有两个事例几乎是平行的情况,但是用 Minkowski 距离公式计算距离会是一个大于 0 的数值。还有,如果两事例有缺失数据,如何计算它们之间的距离,目前为止尚未发现有文献讨论。因此,本文依托欧拉距离,使用杂合距离算法改进 Minkowski 距离公式,使得最近邻算法能够针对不同实际需要计算两事例距离,适用到属性是混合型的情形,也能避免时序列中出现的错误计算问题。

1 改进的最近邻算法

1.1 算法的模型

选用最小-最大规范化对连续数据进行线性变换。假定 $\min_A$ 和 $\max_A$ 分别为属性 A 的最小值和最大值,最小-最大规范化通过计算

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A)$$

收稿日期:2010-09-16
修回日期:2010-10-11

作者简介:刘星毅(1972-),男,硕士,副教授,主要从事数据库技术和计算机网络研究。

*广西自然科学基金项目(桂科自 0899018),广西教育厅科研项目(200808MS062)资助。

$+new\_\min_A$，

将属性A的值v映射到区间$[new_\min_A, new_\max_A]$中的v'。

把现实应用中的属性分为五类：连续型，对称二进制型，非对称二进制型，无序离散型和有序离散型。

连续型是应用中最为常见的类型，比如工资属性，属于实数范畴且有序，这种类型采用Minkowski距离公式最为合理，但是，时序列中计算出错问题经常出现在连续型属性中，所以，定义：$d(i,j)=\sqrt{\sum_{k=1}^{n}((A_{i,k}-A_{j,k})-(\bar{A}_i-\bar{A}_j))^2}$，其中$n$代表在事例$i$和$j$中有$n$个连续属性，$A_{i,k}$是事例$i$第$k$个属性的属性值，$\bar{A}_i$是事例$i$中$n$个连续属性的平均值。这样设置就可以避免时序列中计算出错问题。

二进制类型属性就是属性值只有两个取值的属性，分为对称二进制和非对称二进制。对称二进制属性两个取值分布是均匀的，计算两个事例的距离时，对两个值的权值可以取相同。非对称二进制两个取值分布是不均匀的，计算两个事例的距离时，两个取值概率不同，权值也应该不同。所以根据相依表来计算二进制属性的距离，表1中q代表事例i和事例j的取值为"1"的个数，依次类推，可以得出对称二进制距离的公式为：$d(i,j)=\frac{r+s}{q+r+s+t}$，非对称二进制的距离公式为：$d(i,j)=\frac{r+s}{q+r+s}$。

表1 两事例二进制属性相依表

		事例i		
事例j		1	0	总计
	1	q	r	$q+r$
	0	s	t	$s+t$
	总计	$q+s$	$r+t$	$q+r+s+t$

无序离散型属性和有序离散型属性都是离散型属性。如果在总共p个无序离散型属性的数据集中，两个事例中有相同属性值得个数是m，则它们之间距离可以定义为：$d(i,j)=\frac{p-m}{p}$。一些离散属性的属性值是有顺序的，例如属性"排名"，"1"和"2"是有顺序的，但是又与连续属性不同，因为这些有序离散属性之间没有1～2之间的数据存在，因此，计算这样属性的距离也综合无序离散属性和连续属性的特点。首先，把每个属性值r_i进行这样的转换：$O_i=\frac{r_i-1}{\mathrm{Max}-1}$，其中Max这个属性中最大的属性值，$O_i$是最终的转换结果。按照这样方式转换后，就可以使用前面提出的计算两个连续属性的距离公式对这两有序离散属性的距离进行计算了。

一个数据集经常出现以上这五类属性的几种或者全部，所以综合以上方法可得出

$$d(i,j)=\frac{\sum_{k=1}^{n}\delta_{ij}^{f}d_{ij}^{f}}{\sum_{k=1}^{n}\delta_{ij}^{f}}, \quad (1)$$

其中δ_{ij}^{f}代表事例i和j是否有缺失现象，如果有则为0，否则为1。f为在五类属性中第f类属性，n是属性个数。

1.2 算法的步骤

算法的步骤如下。

步骤1 对所有数据进行数据规范；

步骤2 对每一个要找最近邻的事例，根据公式(1)计算它与数据集中其他事例的距离；

步骤3 对得到的所有距离排序，选取K个最小距离；

步骤4 如事例是离散属性，取这K个距离中的最大类为此事例的值；如事例是连续属性，取这K个距离中的中位数为此事例的值。

2 改进算法的实例验证

采用UCI数据集Hepatitis和Water-Treatment中没有缺失的事例，通过kNN算法，对改进的最近邻算法(NNH)与常见的曼哈顿距离算法(NNM)和欧式距离算法(NNO)进行分类测试，比较各种算法分类准确率。数据集Hepatitis含155个事例，每个事例有19个条件属性和类标签，是两类分类问题，并且条件属性中既有离散型属性又有连续属性。数据集Water-Treatment有527个事例，每个事例有38个属性(其中含有离散型属性和连续型属性)和一个多类决策属性。k取值从1到10。

从图1可以看出，在不同k的选取和不同的数据集中，NNH算法不管在两类分类问题还是多类分类问题的分类准确率评估中的准确率高于NNM算法和NNO算法，填充效果好于NNM算法和NNO算法。

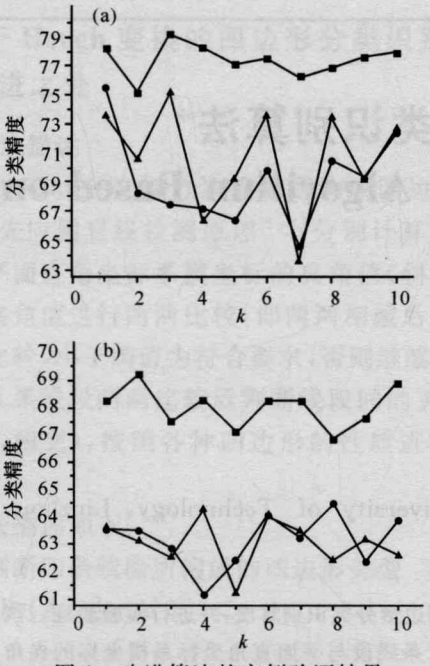

图 1 改进算法的实例验证结果

(a)两类分类问题;(b) 多类分类问题。

■:NNH,●:NNO,▲:NNM。

3 结束语

本文对实际生活中广泛应用的最邻近算法进行改进,并应用实例进行验证。改进的算法除了仍然具有容易理解,操作简单,效果明显的特点外,还能解决原算法经常出现的两事例距离计算问题,时序列中计算出现错误问题,无法适用到属性混合情形的问题,以及实现两事例有缺失数据时的距离计算。

参考文献:

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

[2] Vassilis Athitsos, Michalis Potamias, Panagiotis Papapetrou, Nearest Neighbor Retrieval Using Distance-Based Hashing[C]. ICDE, 2008:327-336.

[3] Han J, Kamber M. Data Mining: concepts and techniques: 2nd edition[M]. Morgan Kaufmann Publications, 2006.

[4] Little R, Rubin D. Statistical analysis with missing Data[M]. Wiley, 2002.

[5] 刘星毅. GBNN-填充缺失属性值算法[J]. 微计算机信息,2007,23(15):246-248.

[6] Yang Tao, Cao Longbing, Zhang Chengqi. A novel prototype reduction method for the K-Nearest neighbor algorithm with K>=1[M]. PAKDD, 2010: 89-100.

[7] 杨涛,骆嘉伟,王艳,等. 基于马氏距离的缺失值填充算法[J]. 计算机应用,2005,25(12):2868-2871.

[8] 何晓群. 多元统计分析[M]. 北京:中国人民大学出版社,2004.

(责任编辑:邓大玉)

(上接第 408 页)

[4] Chen Liang. IEEE an ant colony algorithm for text clustering[C]. International Conference of Computing, Control and Industrial Engineering, 2010:249-252.

[5] Mao Xinyan, Sun Binjie, Zhang Ying, et al. Color image segmentation method based on region growing and ant colony clustering[C]. IEEE WRI Global Congress of Intelligent Systems, 2009:173-177.

[6] Pal M, Foody G M. Feature selection for classification of hyperspectral data by SVM[J]. Geoscience and Remote Sensing, IEEE Transactions on, 2010, 48(5): 2297-2307.

[7] Liao Liang, Wang Dongyun, Wang Fengge, et al. A fast kernel-based clustering algorithm with application in MRI image segmentation[C]. IEEE International Conference of Intelligent Computing and Intelligent Systems, 2009:405-410.

[8] Jiang Quansheng, Jia Minping. Novel hybrid clustering algorithm incorporating artificial immunity into fuzzy kernel clustering for pattern recognition[C]. Control Conference, 2007:592-596.

[9] 王国胜. 核函数的性质及其构造方法[J]. 计算机科学,2006,33(6):172-178.

(责任编辑:尹 闯)