Hans Journal of Data Mining 数据挖掘, 2017, 7(2), 46-50
Published Online April 2017 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2017.72005

Hans汉斯

The Classification of Osteosarcoma Based on Relative Transformation

Xianfa Cai[1], Shan Hu[2], Jie Li[1]

[1]Medical Information Engineering School, Guangdong Pharmaceutical University, Guangzhou Guangdong
[2]Computer Center of Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou Guangdong
Email: cxianfa@126.com

Received: Apr. 6th, 2017; accepted: Apr. 27th, 2017; published: Apr. 30th, 2017

Abstract

As a common disease in department of orthopedics, osteosarcoma is a malignant tumor with high malignancy and poor prognosis. Because the disease often occurs in young people and is very harmful, therefore, early detection, early diagnosis and early treatment are key to the treatment of osteosarcoma. In this paper, local classifier based the nearest neighbor is introduced into the classification of osteosarcoma data, which greatly improves the classification of the automatic and effect. However when dealing with the sparse, noisy and imbalance data, it cannot guarantee to obtain good performance. Based on the relative cognitive law, this paper proposes a feasible strategy called relative local mean center classifier by using the relative transformation to local mean center classifier. The relative space is constructed which may be more line with people's intuition. It should be indicated that relative transformation can improve the distinguishing ability among data points and diminish the impact of noise on classification. The experimental result shows that relative local mean center classifier has a very good classification effect, and can effectively assist clinicians.

Keywords

k Nearest Neighbors Classifier, Local Mean Center Classifier, Relative Transformation, Relative Local Mean Center Classifier

基于相对变换的骨肉瘤分类算法

蔡先发[1]，胡　珊[2]，李　洁[1]

[1]广东药科大学医药信息工程学院，广东 广州
[2]中山大学中山医学院计算机中心，广东 广州
Email: cxianfa@126.com

收稿日期：2017年4月6日；录用日期：2017年4月27日；发布日期：2017年4月30日

文章引用：蔡先发，胡珊，李洁．基于相对变换的骨肉瘤分类算法[J]．数据挖掘, 2017, 7(2): 46-50.
https://doi.org/10.12677/hjdm.2017.72005

摘　要

作为一种常见的骨科疾病，骨肉瘤属于恶性程度甚高、预后极差且转移较快的骨原发性恶性肿瘤。由于该病多发于青少年且危害很大，因此，早期发现、早期诊断和早期治疗便成为治疗骨肉瘤的关键。将机器学习中的基于近邻的局部分类器引入到骨肉瘤的数据分类中来，极大的提高了分类的自动性以及效果。然而由于骨肉瘤数据可能存在稀疏、噪声和非平衡等问题，如此算法的效果往往不佳。本文根据认知的相对性规律提出了基于相对变换的局部均值分类算法，通过相对变换将数据的原始空间变换到相对空间，在相对的空间中度量数据的相似性更符合人们的直觉，从而提高了数据之间的可区分性，同时在一定条件下相对变换还能抑制噪声的影响。实验结果表明，相对局部均值算法具有非常好的分类效果，可以有效地辅助临床医生。

关键词

k近邻分类器，局部均值算法，相对变换，相对局部均值算法

1. 引言

作为一种常见的骨科疾病，骨肉瘤属于恶性程度甚高、预后极差且转移较快的骨原发性恶性肿瘤。由于该病多发于青少年且危害很大，因此，早期发现、早期诊断和早期治疗便成为治疗骨肉瘤的关键。在机器学习，计算机视觉，图像处理等领域中，将事物按照一定的特征或者规律进行分类是非常重要的一个步骤。将机器学习中的分类器引入到骨肉瘤的数据分类中来，极大的提高了分类的自动性以及效果。过去的数十年间，产生了大量的分类算法，经典的比如 k 近邻算法(k nearest neighbors, KNN)及其各种变体[1] [2] [3] [4] [5]。由于理论上极其简单，也不需要对数据的分布做任何的假设，且只需要一个参数，因此 KNN 成为非常实用并且高效的分类器。然而，KNN 在分类的时候，由于将每个样本同等看待，因此，当近邻间的信息不可以忽略并且在高密度区域外变得越来越大时，KNN 分类器的效果往往比较差。另外一方面，当遇到非平衡数据的情况下，比如一类的数据明显比另一类多且分界线明显倾向于数据比较少的类的时候，KNN 算法的分类效果也比较差。为了解决这个问题，Y. Mitani 等人[6]设计了局部均值中心分类算法(Local mean center classifier, LMC)，即将每一类的 k 个近邻计算其局部均值中心，然后将待分样本分给距离该中心比较近的类；Boyu Li 等提出了一种基于局部概率中心的分类算法(Local probability center classifier, LPC) [7]，该方法关注于寻找最优分类面两侧具有代表意义的局部概率中心；P. Vincent 等提出了一种局部超平面的分类器算法(K-local hyperplane nearest neighbor classifier, HKNN) [8]。

考虑到生活中存在大量稀疏，噪声和非平衡数据，这些将极大地影响到分类器的性能。本文根据认知的相对性规律提出了基于相对变换的局部均值分类算法，通过相对变换将数据的原始空间变换到相对空间，在相对的空间中度量数据的相似性更符合人们的直觉，从而提高了数据之间的可区分性，同时在一定条件下相对变换还能抑制噪声的影响。基于相对变换的局部均值分类算法的主要优点如下：1) 将善于区分噪声，稀疏和非平衡数据的相对变换引入到分类器中来，极大地提升了分类器的性能；2) 将基于相对变换的局部均值分类器应用到骨肉瘤的分类中来，表明机器学习在临床辅助方面具有一定的作用。

2. 局部均值算法

作为 KNN 分类器的改进版本，由 Y. Mitani 等人于 2006 年提出的局部均值算法是一个局部的、懒惰的、非参数的分类器[6]。LMC 算法的主要思想是要计算每个类的局部几何中心作为测试样本的最近邻，这一几何中心称为局部平均向量。具体说来，LMC 方法中首先在训练样本的每个类中选出 k 个距离测试样本最近的样本，然后用在每个类别中选取的 k 样本计算出每个类的局部平均向量。由于每个类都计算出了一个局部平均向量，因此这些局部平均向量可以视为是每个类的一个类代表点。最后，该方法将测试样本分类到与该测试样本最近的局部平均向量所属的类别。局部均值分类器的具体算法描述如下[9]：

步骤 1：在每个类中选出与测试样本最近的 k 个样本；

步聚 2：对于每个类，由选取的 k 个最近邻样本，计算出针对每个类的局部平均向量；

步骤 3：对于每个类，计算测试样本与局部平均向量之间的欧氏距离；

步聚 4：根据测试样本与每个类的局部平均向量计算出的距离，将测试样本分类到具有最小欧氏距离的类别中去。

3. 相对变换

相比机器而言，人类在区分稀疏、噪音以及非平衡数据方面具有一种与生俱来的本领，这点值得机器向人类学习。当前的识别人脸、基因分类的机器学习方法常常需要数百甚至是上千的样本做训练，而人类视觉识别仅仅需要少量的样本就可以[10] [11] [12]。这是由于人类在区分事物的时候一方面根据它们各自的属性，另外一方面是受团队中的成员影响，并且事物本身的属性又常常被周围的环境影响和改变着。经验表明，人类的感知具有相对性，如图 1 所示。在图 1 中的两个虽然实际上是一样大的圆 x 和 y [12]，常常会误认为圆 x 比圆 y 大。原因是与周围的圆相比，圆 x 显得比较大，而圆 y 则明显小于其周围的圆，最终会误认为圆 x 大于圆 y 。

由于相对变换并不是等距变换，而是一种具有放大作用的变换，因此更加容易凸显数据间的拓扑结构，并因此提高了数据之间的可区分性，如图 2 所示。在原始空间中由于 $d(x_3,x_1)=d(x_3,x_4)$，x_1 和 x_4 等概率成为 x_3 的最近邻，而在相对空间中由于 $d(y_3,y_1)<d(y_3,y_4)$，则很容易判断相比 y_4，y_1 与 y_3 更近，如此也更符合人类的直觉。因此相对变换并不是简单的线性变换，通过相对变换，原始空间中不好区分的数据在相对空间中则容易区分开来，从而提高了数据间的可区分性。此外，相对变换还容易识别出孤立点，如图 2 所示。从直观上看，图 2(a)中的 x_4 很有可能是孤立点，但是由于在原始空间中满足了 $d(x_3,x_1)=d(x_3,x_4)$，这使 x_1 和 x_4 拥有等概率成为 x_3 的近邻，这也与我们的直觉不一致。而在相对空间中由于 $d(y_3,y_1)<d(y_3,y_4)$，这意味着孤立点 y_4 更加远离正常数据点，从而提高了数据间的区分性。

为模型化该认知规律，以原始数据空间 $X=[x_1,x_2,\cdots,x_n]^{\mathrm{T}}, x_i\in R^D$ 中的每个数据点作为基向量来构造新的空间，这样任意点 x_i 到其他所有点的距离就构成该点在新空间中的坐标，这个过程称之为相对变换。

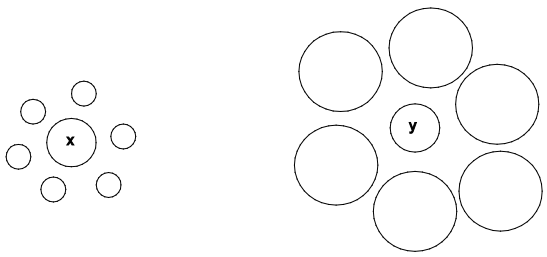

Figure 1. Human visual perception is relative
图 1. 人类视觉的相对性

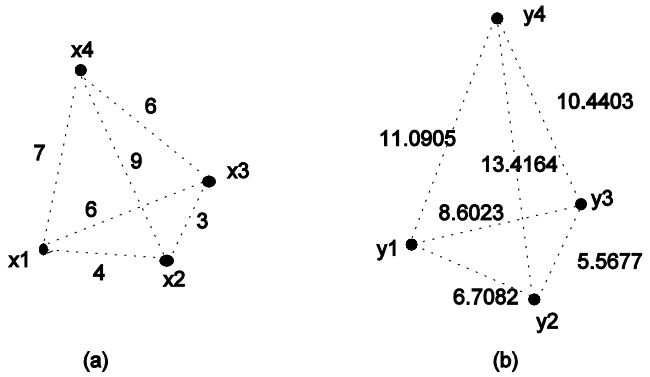

Figure 2. Function of the relative transformation on noisy data where (a) is the original space and (b) is the relative space
图 2. 相对变换能抑制噪声的影响，(a) 原始空间，(b) 构造的相对空间

4. 相对变换的局部均值分类算法

算法 RLMC(x,X,k)

/* x 为测试样本，X 为训练样本集，k 为在每个类中选取的最近邻样本数 */

步骤 1：在每个类 ω_j 中选出与测试样本 x 最近的 k 个样本，由 $X_k\left(x,\omega_j\right)$ 表示；

步骤 2：用下面的方法构建相对空间：

$$x^r\left(q,k\right)=f^r\left(\left(q,k\right)\cup\left\{q\right\}\right) \tag{4-1}$$

步聚 3：对于每个类 ω_j，在相对空间中，由选取的 k 个最近邻样本，按如下方法计算出针对每个类 ω_j 的相对局部平均向量：

$$\overline{x}_{rj}=\frac{1}{k}\sum_{t=1}^{k}x_t,\ x_t\in x^r\left(q,k\right) \tag{4-2}$$

步骤 4：对于每个类 ω_j，计算测试样本 x 与相对局部平均向量 $\overline{x}_{rj}$ 之间的相对欧氏距离：

$$d_{rj}=\left\|x-\overline{x}_{rj}\right\| \tag{4-3}$$

步聚 5：根据测试样本 x 与每个类 ω_j 的相对局部平均向量 $\overline{x}_{rj}$ 计算出的距离 d_{rj}，将测试样本 x 分类到具有最小相对欧氏距离 $d_{r\min}$ 的类别 ω_τ。

5. 结论

实验中选用正常人长骨 CR 图像和长骨骨肉瘤图像，图像格式均为 DICOM 格式。有效样本共计 110 例，其中骨肉瘤患者为 58 例，正常人 52 例。由于这些图像来源于不同的机器，图像的分别率会有所不同，因此不同图像的感兴趣区域的分别率差异有可能成数量级变化，而该差异对纹理特征的提取结果会有较大的影响。因此对图像进行了预处理，以降低分辨率差异造成的不良影响。

为验证我们提出的方法的有效性，在骨肉瘤数据集上进行实验。实验中将本文提出的相对局部均值分类器与几种懒惰、非参数的方法进行比较，他们分别是：KNN、HKNN、LMC 和 LPC。从表 1 中的测试结果可以看到，在骨肉瘤数据上，与其它 4 个算法相比，RLMC 算法具有最好的表现，说明该方法具有相当好的分类性能。说明通过相对变换构造的相对空间，在相对空间里面选择近邻更加符合人们的直觉，从而提高了数据的分类效果。

Table 1. Average classification errors for osteosarcoma data
表 1. 骨肉瘤数据集上各算法的平均分类错误率

数据	KNN	HKNN	LMC	LPC	RLMC
骨肉瘤	2.54 ± 0.025	52.72 ± 0.13	47.27 ± 0.012	47.15 ± 0.05	**0.36 ± 0.49**

6. 结束语以及展望

本文根据认知的相对性规律提出了基于相对变换的局部均值分类算法，通过相对变换将数据的原始空间变换到相对空间，在相对的空间中度量数据的相似性更符合人们的直觉，从而提高了数据之间的可区分性，同时在一定条件下相对变换还能抑制噪声的影响。基于相对变换的骨肉瘤分类算法具有非常好的分类效果，可以有效地辅助临床医生。同时，实验表明，将认知规律结合当前的分类器可以有效地提升分类性能，未来将探索更多的认知规律并且将它们应用到分类器中来。

基金项目

广东省自然科学资金(2015A030310267, 2016A030310300)资助。

参考文献 (References)

[1] Gao, Y., Pan, J., Ji, G. and Yang, Z. (2012) A Novel Two-Level nearest Neighbor Classification Algorithm Using an Adaptive Distance Metric. *Knowledge-Based Systems*, **26**, 103-110. https://doi.org/10.1016/j.knosys.2011.07.010

[2] Karl, S.N. and Truong, Q.N. (2009) An Adaptable k-Nearest Neighbors Algorithm for MMSE Image Interpolation. *IEEE Transactions on Image Processing*, **18**, 1976-1987. https://doi.org/10.1109/TIP.2009.2023706

[3] Keller, J.M., Gray, M.R. and Givens, J.A. (1985) A Fuzzy k-NN Neighbor Algorithm. *IEEE Transactions on System, Man and Cybernetics*, **15**, 580-585. https://doi.org/10.1109/TSMC.1985.6313426

[4] Denoeux, T. (1995) A k-Nearest Neighbor Classification Rule Based on Dempster Shafer Theory. *IEEE Transactions on Systems, Man and Cybernetics*, **25**, 804-813. https://doi.org/10.1109/21.376493

[5] Wang, H. (2006) Nearest Neighbors by Neighborhood Counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**, 942-953. https://doi.org/10.1109/TPAMI.2006.126

[6] Mitani, Y. and Hamamoto, Y. (2006) A Local Mean-Based Nonparametric Classifier. *Pattern Recognition Letters*, **27**, 1151-1159. https://doi.org/10.1016/j.patrec.2005.12.016

[7] Li, B., Chen, Y.W. and Chen, Y.Q. (2008) The Nearest Neighbor Algorithm of Local Probability Centers. *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, **38**, 141-154. https://doi.org/10.1109/TSMCB.2007.908363

[8] Vincent, P. and Bengio, Y. (2002) K-Local Hyperplane and Convex Distance nearest Neighbor Algorithms. *Advances in Neural Information Processing Systems*, **14**, 985-992.

[9] 李海生. 基于证据理论的分类方法研究[D]: [博士学位论文]. 广州: 华南理工大学, 2013.

[10] 蔡先发. 基于图的半监督算法及其应用研究[D]: [博士学位论文]. 广州: 华南理工大学, 2013.

[11] Li, D.Y., Liu, C.Y., Du, Y. and Han, X. (2004) Artificial Intelligence with Uncertainty. *Journal of Software*, **15**, 1583-1594. (In Chinese with English Abstract). http://www.jos.org.cn/1000-9825/15/1583.htm

[12] Wen, G., Jiang, L.J. and Wen, J. (2009) Relative Transformation-Based Neighborhood Optimization for Isometric Embedding. *Neurocomputing*, **72**, 1205-1213. https://doi.org/10.1016/j.neucom.2008.02.009