

基于视觉原理的分类算法

刘京鑫, 孙剑, 孟德宇

(西安交通大学理学院, 710049, 西安)

摘要: 从一种新的基于生物视觉原理的观点,提出了一种新的数据分类算法.将数据集看作图像,利用高斯导函数进行特征提取,并用提取出来的特征计算数据的局部结构,在此基础上设计各向异性感受野函数,最后根据各向异性的核函数构造出分类决策函数.在标准测试集上的实验表明:所提出的算法与支持向量机算法分类正确率相当,同时具有更高的训练速度;与 Parzen 窗分类算法相比,尽管训练速度相对较慢,但分类精度明显提高,很好地综合了分类算法对训练速度和分类精度的要求.

关键词: 数据挖掘;分类;视觉原理

中图分类号: O29 **文献标志码:** A **文章编号:** 0253-987X(2010)10-0116-04

New Classification Algorithm Based on Visual Principles

LIU Jingxin, SUN Jian, MENG Deyu

(School of Sciences, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: A novel data classification algorithm based on the biological visual principles is proposed. Viewing the data set as an image, the local structures of data image are extracted as the basis of an anisotropic receptive field function, and the decision function is designed. The experiments on standard datasets show that this algorithm is comparable to SVM in accuracy but with significantly higher training rate. Compared with the Parzen window classification algorithm, though with relatively slower in training rate, it achieves much higher accuracy. Thus, the proposed algorithm meets the requirements of both training rate and classification accuracy.

Keywords: data mining; classification; visual principles

数据挖掘是数据库知识发现(KDD)的主要过程,其目的在于从大量的数据中,抽取出潜在的、有价值的知识(模型或规则).数据挖掘中的典型问题包括聚类分析(聚类)、判别分析(分类)、回归分析(回归)、隐变量分析(可视化)等,其中分类问题是数据挖掘的基本问题之一.

分类问题通常从具有标号信息的数据集(训练集)中寻找到一个判别函数,再由这个函数对新来的未知标号的数据点进行判别,确定其标号(类别).由于分类问题在科学研究和工程技术领域有着广泛的应用,几十年来,人们对此有着广泛的关注.分类问

题是模式识别的基础,在计算机视觉、生物信息学等领域有着重要的应用,如手写字符识别、语音识别、人脸识别、基因序列分类等.

近几十年来,已经发展出许多分类方法.概括来说,目前被熟知并且广泛应用的分类方法主要包括传统的统计方法^[1]、判别分析方法^[2]、Boosting 算法^[3]、神经网络算法^[4]、支持向量机算法(SVM)^[5]等.

分类算法的设计主要关注如下两个基本问题,即分类的精确性与算法效率.一个好的分类算法,不仅需要具有较高的分类精度,同时也需要具有较高

的算法效率. 传统的基于统计或者判别分析的方法大多具有较低的运算代价,但算法精度不够高,而近代的神经网络算法、支持向量机算法和 Boosting 算法具有更高的分类精度,但也有更高的计算代价.

本文将通过模拟生物视觉系统的认知机制,进而提出一种新的分类算法,期望更好地解决分类算法的两个基本问题.

人类的视觉感知系统对外界视觉信息具有高效的表达和分类能力,它可以实时、准确地对观察到的客观世界进行识别和认知. 因此,通过对视觉系统的工作机制进行研究与建模,对于解决分类问题提供了一种可能的有效途径. 基于上述想法,我们所提出的分类算法的基本思想如下:首先,将数据点理解为进入视觉系统的光点,而数据集则构成光点图像;然后,通过对生物视觉系统的特征提取机制和感受野机制进行建模,建立面向数据点的各向异性感受野函数;最后,提出基于生物视觉原理的数据分类模型. 将该模型应用于标准测试数据集,实验结果表明:在所测数据集上,本文算法与 SVM 相比具有略高的分类正确率,同时具有更快的训练效率.

1 视觉感知原理

人类视觉系统主要分为视网膜、初级视皮层和高级视皮层. 在视网膜这一层次上,视觉信息处理主要有 3 个基本的步骤:①外部真实环境中物体的反射光通过视觉光学系统到达视网膜,并形成二维图像;②在初级视皮层上,未经处理的原始图像由感受野机制进行特征提取,提取出的视觉特征被传送到更高级的视觉皮层;③高级视觉皮层基于视觉特征对图像进行分析与理解.

感受野机制是生物视觉系统进行视觉信息识别的重要机制. 通常,影响一个神经元响应的视觉细胞区域称为该神经元的感受野. 视觉细胞的感受野定义在更低层视觉皮层或视网膜层次上,只有在视觉感受野区域内的视觉细胞的活性才会影响所关注视觉细胞的响应. 通过视觉感受野机制,生物视觉系统构造成为复杂的多层网络系统^[6]. 某视觉细胞的感受野区域内的细胞对该视觉细胞响应的影响是不同的,感受野内不同细胞对所关注视觉细胞响应的影响建模为感受野函数^[7-8]. 在数学上,感受野可以描述为用一个被称之为感受野函数的非平凡定义域刻画. 视网膜上视觉细胞感受野函数的基本形式包括高斯函数^[9]、多高斯函数^[10]、高斯导数^[11]、Gabor 函数^[12]等.

提取感受野机制的视觉特征,可以建立感受野内细胞对所关注细胞的神经元输入与感受野函数的卷积的模型. 因此,研究的重点之一就是要设计能很好反映数据结构的感受野函数.

在感受野机制的基础之上,我们对分类问题进行了数据建模及算法设计,具体思想如下:首先,把数据转换成抽象的数据图像,这可以用 δ 函数来实现;其次,利用感受野机制,模拟视网膜特征提取方式,即用感受野函数对数据图像进行滤波,滤波的过程就是观察和认知数据图像的过程;最后,基于视觉图像的特征,提出数据分类模型.

2 视觉分类算法

2.1 算法流程

本文将所提出的基于视觉特征提取的分类算法称为视觉分类机(VCM),其基本思想是:把分类问题看成一个视认知问题,通过模拟视网膜的滤波与特征提取机制设计数据分类模型,通过视觉尺度选择设定滤波器参数,由此形成一个全新的分类算法. 特别地,VCM 将数据看作图像,利用感受野函数进行特征提取,用提取出来的特征计算数据的局部结构,在此基础上设计各向异性的核函数(感受野函数),最后根据各向异性的核函数设计分类模型和算法,进而对数据进行分类. VCM 的流程如图 1 所示,其具体步骤如下:

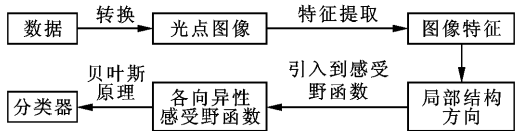


图 1 视觉分类算法流程

(1)由数据集(样本点) $X=\{x_1,x_2,\cdots,x_N\}$ 形成(虚拟)数据图像. 这一步可由 δ 函数实现,每一个数据点对应一个单位光强,数据图像中任一像素点的像素值如下

$$p(x)=\frac{1}{N^++N^-}\left[\sum_{i=1}^{N^+}\delta(x-x_i^+)+\sum_{i=1}^{N^-}-\delta(x-x_i^-)\right]$$

式中: N^+ 为正类样本点的个数; N^- 为负类样本点的个数; $N=N^++N^-$; x_i^+ 为正类样本点; x_i^- 为负类样本点.

如图 2a 所示,以二维数据集为例,生成了两类数据点. 由这个数据集形成的数据图像,可由图 2b 表示.

(2)应用高斯导数感受野函数对数据图像进行特征提取(提取出来的特征是每个数据点的梯度)

$$\mathbf{g}(\mathbf{x}_i) = p(\mathbf{x}_i) * \nabla G(\mathbf{x}_i, \sigma)$$

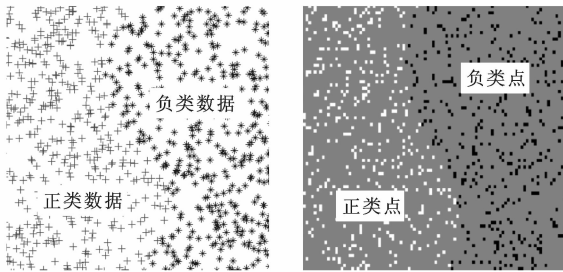
式中: $\nabla G(\mathbf{x}, \sigma)$ 为高斯导数感受野函数. $\|\mathbf{g}(\mathbf{x}_i)\|$ 实际上刻画了数据点位于分类边界的可能性. 如图 2c 所示, 位于分类边界点的梯度模值要比位于类内部点的梯度模值大得多.

(3)利用梯度特征计算数据的局部结构(结构张量), 有

$$\mathbf{T}(\mathbf{x}_i) = \mathbf{g}(\mathbf{x}_i) \mathbf{g}^T(\mathbf{x}_i)$$

对任意点 $\mathbf{x}$, 其结构张量 $\mathbf{H}(\mathbf{x})$ 定义为

$$\mathbf{H}(\mathbf{x}) = \sum_{i=1}^N G(\mathbf{x} - \mathbf{x}_i, \sigma) \mathbf{T}(\mathbf{x}_i)$$



(a)数据点集

(b)数据图像

(c)梯度示意图

(d)分类边界示意图

图2 视觉分类示意图

由于以上 $\mathbf{g}(\mathbf{x}_i)$ 所反映的边界特性, 如此构造的结构张量亦能体现出分类问题的边界特点.

(4)利用局部结构设计各向异性高斯感受野函数如下

$$K(\mathbf{x}; \mathbf{x}_i, \sigma^*, \lambda^*) = e^{-\lambda(x-x_i)^T \mathbf{H}(\mathbf{x}_i)(x-x_i)}$$

各向异性的感受野函数进一步反映了数据集的边界信息. 如图 2d 所示, 椭圆为各向异性感受野函数的等高线. 边界区域形状为椭圆, 并且该椭圆沿着垂直于分类面的方向最窄, 而内部区域形状接近于圆形. 结构张量 $\mathbf{H}$ 的设计使得 K 为各向异性, 且沿着垂直于分类面方向的方差小于平行于分类面方向的方差. 这样设计的目的是: 在分类边界区域, 降低跨类

的训练数据对测试数据分类决策的影响.

(5)定义 VCM 分类器

$$f(\mathbf{x}) = \text{sgn}\left\{\sum_i y_i K(\mathbf{x}; \mathbf{x}_i, \sigma^*, \lambda^*)\right\}$$

如图 2d 所示, S 型曲线为分类边界.

以上步骤的基本原理可利用图 2 在二维数据情形下得以直观地理解, 而其计算步骤亦可很自然地在高维数据上得以实现.

在算法的训练阶段, 计算梯度的时间复杂度为 $o(n^2 + nd^2)$, 计算局部结构张量的时间复杂度为 $o(n)$, 因此算法总的训练的时间复杂度为 $o(n^2)$. 支持向量机方法训练阶段的时间复杂度为 $o(n^3)$, 因此本文算法在训练阶段的时间复杂度要优于支持向量机方法.

2.2 算法的参数学习

算法中有两个参数 σ^* 和 λ^* 需要确定. 我们将参数 σ^* 的选择问题理解为视觉认知问题: 在所选尺度下能把数据的全局特征和局部特征看得最清晰. 据此原理, 采用本项目组之前研发的最优尺度先验公式^[13]

$$\sigma^* = \underset{\alpha}{\operatorname{argmax}} \|d_{\text{far}} - d_{\text{near}}\|^2 = \left(\frac{d_{\text{far}}^2 - d_{\text{near}}^2}{2(\ln d_{\text{far}}^2 - \ln d_{\text{near}}^2)} \right)^{1/2}$$

随机从数据集中选取 N 个样本点, 数据集的总样本数目为 l , 将所选样本点与数据集合的所有点对的距离进行排序. 第 $(1-\alpha) * (N(l-1))$ 个距离为 d_{far} , 反映数据集的全局结构, 第 $\alpha * (N(l-1))$ 个距离为 d_{near} , 反映数据集的局部结构. 此公式已经应用到了 SVM 算法的模型选择中, 也取得了非常好的效果.

采用交叉验证方法来确定 λ^* . 具体地, 可以采用 5 倍交叉验证方法, 即把训练集随机分成 5 份, 选择里面的 4 份作为新的训练集, 剩下的 1 份作为新的测试集, 依次做 5 次测试, 把 5 次测试的误差的平均值作为最终的测试误差, 以此方式, 选择一个使测试误差最小的 λ 作为最终的 λ^* .

3 实验

为了说明以上算法的有效性, 本文对数据挖掘和机器学习领域中通常采用的 11 个标准数据集进行了实验. 这些数据集来自 <http://ida.first.gmd.de/raetsch/data/benchmarks.htm>, 数据集基本信息如表 1 所示, 本文方法、SVM、Parzen 窗分类算法(PWC)的测试结果见表 2.

表 1 11 组标准数据集的基本信息

数据集	维数	训练数据个数	测试数据个数
Banana	2	400	4 900
Breast-cancer	9	200	77
Diabetis	8	468	300
Flare-solar	9	666	400
German	20	700	300
Heart	13	170	100
Image	18	1 300	1 010
Thyroid	5	140	75
Titanic	3	150	2 051
Twonorm	20	400	7 000
Waveform	21	400	4 000

在训练时间上,本文算法比 SVM 的训练时间少,大约只有 SVM 的 14%左右,VCM 的平均分类误差为 15.99%,SVM 的平均分类误差为 16.03%,VCM 的分类误差要比 SVM 的低.与 PWC 相比,本文方法虽然训练时间更长,但分类误差比 PWC 显著低,在精度上具有很大的优势.

4 结 论

本文尝试从一种新的基于生物视觉原理的观点,通过对生物视觉系统的感受野机制的建模,提出了一种新的数据分类算法.在标准测试集上的初步实验表明:所提出的算法与 SVM 的分类正确率相

表 2 SVM、PWC 及 VCM 在 11 组标准数据集上的测试结果

数据集	训练时间/s			分类误差/%		
	SVM	PWC	VCM	SVM	PWC	VCM
Banana	4 501.41	7.83	771.36	11.53±0.66	10.81±0.51	10.92±0.47
Breast-cancer	773.46	6.63	200.28	26.04±4.74	24.82±4.07	24.29±4.32
Diabetis	7 830.93	13.68	1 090.82	23.79±1.80	25.84±1.81	24.43±1.76
Flare-solar	20 419.71	24.23	3 877.20	32.43±1.82	35.01±1.72	34.12±1.93
German	24 397.02	41.00	2 619.53	23.61±2.07	25.27±2.39	24.04±2.28
Heart	538.11	4.30	146.96	15.95±3.26	17.22±3.51	15.93±3.66
Image	1 346 476.00	129.70	9 144.44	2.96±0.60	3.62±0.63	2.68±0.54
Thyroid	368.43	3.13	97.02	4.80±2.19	4.35±2.34	4.23±2.20
Titanic	403.59	3.53	110.02	22.42±1.02	22.31±1.00	22.90±1.24
Twonorm	449.84	15.63	843.13	2.96±0.23	2.67±0.39	2.48±0.17
Waveform	4 586.25	18.75	845.67	9.88±0.44	10.64±0.98	9.88±0.61
Mean	12 825.00	24.40	1 795.13	16.03±1.71	16.60±1.76	15.99±1.74

当,但具有更高的训练速度;与 Parzen 窗分类算法相比,尽管训练速度相对较慢,但分类精度明显提高.因此,本文方法很好地综合了分类算法对训练速度和分类精度的要求,具有较好的应用前景.

参考文献:

[1] RAJAN J, RAYNER P. Time series classification using the Volterra connection model and Bayes decision theory[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway, NJ, USA: IEEE, 1993:27-30.

[2] FUKUNAGA K. Introduction to statistical pattern recognition[M]. San Diego, USA: Academic Press,

1990:30-33.

[3] RATSCH G, MIKA S, MULLER S B. Constructing boosting algorithms from SVMs: an application to one-class classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(9):1184-1199.

[4] EGMONT PETERSEN M, DE RIDDER D, HANDELS H. Image processing with neural networks: a review[J]. Pattern Recognition, 2002, 35(10):2279-2301.

[5] CORTES C, VAPNIK V. Support-vector networks [J]. Machine Learning, 1995, 3(3):273-297.

[6] HUBEL D H, WIESEL T N. Receptive fields, binocular interaction and functional architecture in the cat's

- visual cortex[J]. Journal of Physiology, 1962, 160(1):106-154.
- [7] HOWELL A J, HOWELL A J, BUXTON H, et al. Receptive field functions for face recognition[C]//Proc 2nd Int Workshop Parallel Modelling of Neural Operators for Pattern Recognition. Faro, Portugal: University of Algarve, 1995:83-92.
- [8] DAUGMAN J G. Two-dimensional spectral analysis of cortical receptive field profiles[J]. Vision Research, 1980, 20(10):847-856.
- [9] LINDBERG T. Scale-space for discrete signals [J]. PAMI, 1990, 12(3):234-254.
- [10] SCENIAK M P, RINGACH D L, HAWKEN M J, et al. Contrast's effect on spatial summation by macaque V1 neurons[J]. Nature Neuroscience, 1999, 2(8): 733-739.
- [11] SHEPP L A. Radon-Nikodym derivatives of Gaussian measures [J]. The Annals of Mathematical Statistics, 1966, 37(2):321-354.
- [12] JAIN A K, FARROKHNI F. Unsupervised texture segmentation using Gabor filter [J]. Pattern Recognition, 1991, 24(12):1167-1186.
- [13] XU Zongben, DAI Mingwei, MENG Deyu. A fast heuristic strategy for model selection of support vector machines [J]. IEEE Transactions on Systems, Man and Cybernetics: part B, 2009, 39(5):1292-1307.

(编辑 杜秀杰)