文章编号:1005-0523(2011)03-0083-05

# 基于LLE和LS_SVM的胃粘膜肿瘤细胞图像分类

甘 岚,吕文雅

(华东交通大学信息工程学院,江西 南昌 330013)

**摘要**:胃粘膜肿瘤细胞图像的复杂性,组织器官形状的不规则性以及不同细胞的差异性,使得采用一般的线性分类方法对其进行分类很困难,结合局部线性嵌入(LLE)在处理非线性数据及最小二乘支持向量机(LS_SVM)在处理小样本、高维数及泛化问题方面的优势,文章提出一种基于LLE+LS_SVM的胃粘膜肿瘤细胞图像分类方法,并采用LS_SVM的线性拟合误差来判断实验效果,最后比较本文方法与其他分类方法的优越性。实验结果表明,该方法在分类准确率和运行时间方面都有很大的优势。

**关键字**:LLE;LS_SVM;肿瘤细胞分类

**中图分类号**:TP181 **文献标识码**:A

由于胃粘膜肿瘤细胞图像的复杂性,组织器官形状的不规则性以及不同种类细胞的差异性,细胞的结构、形状、稀疏程度、排列形状等,都会有很大的差异。在对图像的分类过程中,会遇到各种各样的细胞图像,从这些繁复杂乱的细胞图像中提取细胞特征,并进行有效的分类是很困难的。目前,很多基于机器学习的分类方法应用到图像识别领域,但一般的线性分类方法应用于高维的胃粘膜肿瘤细胞图像时,存在严重的泛化问题[1]。因此,找出一种适合于胃粘膜肿瘤细胞图像且将特征提取和分类融合在一起的非线方法是很有必要的。

局部线性嵌入(LLE)是Sam T在Science[2]杂志上提出的一种非线性非监督的流形学习算法,起初主要是应用于人脸识别和文件中的文本识别领域。LLE具有时间复杂度低、参数少等优点,对于结构复杂、非高斯分布、含有较多冗余信息的胃粘膜肿瘤细胞图像来说,LLE的非线性以及低维嵌入特性非常有利于胃粘膜肿瘤细胞图像的特征降维,而利用LS_SVM的线性回归特性,拟合其线性回归误差[3-4],更加有利于图像的分类。所以本文借鉴两者的优点,提出一种LLE和LS_SVM相结合的肿瘤细胞图像分类方法,并通过不断的实验验证此方法在分类准确率和运行时间上的优势。

## 1 数据采集和预处理

胃粘膜肿瘤细胞图像内容丰富且结构复杂,如何有效的对其进行分类,是一个很复杂的问题,数据的采集和预处理对提高胃粘膜肿瘤细胞图像分类的准确率起着关键的作用。

### 1.1 数据采集

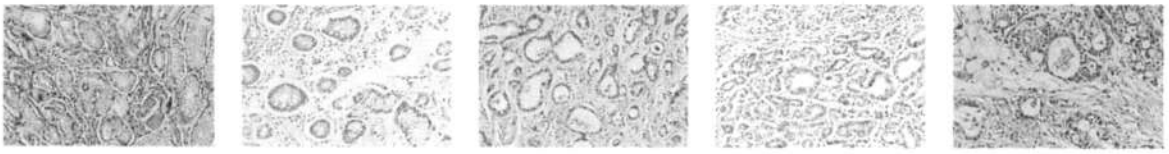
本文采集的图像为医院病理科的切片显微图像,胃粘膜肿瘤分为正常、癌变、增生3大类,增生图像又分为轻度增生、中度增生和重度增生3类,原始采集的五种典型图像如图1所示。

由上图可以看出,最初采集的图像,存在维数高、噪声严重、含有较多的冗余信息、细胞粘连严重、特征难以提取等问题。

**收稿日期**:2011-04-22

**基金项目**:江西省科技厅项目(20051B0104800)

**作者简介**:甘 岚(1964－),女,教授,研究方向为图像处理与模糊识别。



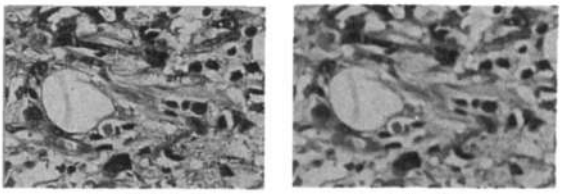
(a) 正常 (b) 轻度增生 (c) 中度增生 (d) 重度增生 (e) 癌变

**图1 原始采集的典型图像**

**Fig.1 Original acquisition of typical images**

### 1.2 预处理

在模式识别中,预处理是图像分割和图像分类的前提,原始采集的胃粘膜肿瘤细胞图像是经过染色的彩色图像,图像的维数很高,且彩色信息对识别作用不大,因此图像的预处理主要是在图像分类之前对图像进行灰度化操作及一些去噪、增强等工作。灰度化一是可以降低图像的维数,二是可以去除图像多余的冗余信息,因此是每张图像必须采用的预处理方法,去噪、增强等方法则根据图像的不同有选择的进行。图2列出一幅原始图像灰度化操作之后的效果。



(a)原始图像 (b)灰度化图像

**图2 图像的灰度化**

**Fig.2 Grayscale of images**

## 2 算法

本文的算法是结合局部线性嵌入(LLE)及最小二乘支持向量机(LS_SVM)这两个核心算法的优点串联构成的。

### 2.1 LLE算法

LLE算法由$N$个输入实向量$\boldsymbol{X}_i$组成,$\boldsymbol{X}_i \in R^D$,$i \in [1, N]$。首先将流形分成许多相互连接的局部区域,每个区域被认为是线性空间,从而可以采用线性方法求出区域内每个点的线性组合性系数。这个系数在低维嵌入时保持不变,因此LLE方法是局部线性化方法[5-6],算法可概括为以下3步。

1) 为每个$\boldsymbol{X}_i$找到它的$K$个最近邻域$\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}, \ldots, \boldsymbol{X}_{ik}$;

2) 测量由每个$\boldsymbol{X}_i$的最近邻域组成的近似值造成的重构错误,计算重构权重,最小化重构误差[7];

3) 计算能保持最佳重构权重的局部几何低维嵌入。

### 2.2 LS_SVM算法

LS_SVM是在SVM的基础上进行改进的,其中把不等式约束改成等式约束,把偏差的一次方改为二次方,LS_SVM的线性回归和其核函数以及核参数选择与设置问题,是实验效果的关键。下面为常用的几种核函数[8]。

1) 多项式核函数:$K(x, x') = \{(x, x') + 1\}^{\gamma}$,此时得到的支持向量机是一个多项式分类器,$\gamma$为自主设定的参数。

2) 径向基(RBF)核函数:$K(x, x') = \exp(-\frac{|x - x'|^2}{2\sigma^2})$,每一个基函数的中心对应于一个支持向量机,得到的支持向量机为径向基函数分类器。式中:$x$为核函数的原始点;$x'$为核函数中心;$\sigma$为函数的宽度参数。

3) Sigmoid函数:$K(x, x') = \tanh(v(x, x') + c)$,这时SVM实现的就是一个多层感知器网络,式中:$v$和$c$为一个常数,这里选取径向基(RBF)核函数,通过设置不同的$\gamma$和$\sigma^2$参数来测试实验效果,其中$\gamma$和$\sigma^2$代表LS_SVM的回归参数。

### 2.3 基于LLE+LS_SVM方法的胃粘膜肿瘤细胞识别过程

将局部线性嵌入(LLE)及最小二乘支持向量机(LS_SVM)这两个核心算法串联构成基于LLE+

LS_SVM方法的胃粘膜肿瘤细胞识别算法,首先对灰度化图像采用LLE方法降维和聚类,基于LS_SVM的线性回归功能,再对降维后的图像采用LS_SVM进行线性拟合[9],下面给出胃粘膜肿瘤细胞识别的具体过程。

给定一个非正常胃粘膜显微图像训练集 $\boldsymbol{X}$ ,训练图像总数为 $N$ ,将其中属于癌变的归为一类,记为 $\boldsymbol{X}_1$ , $\boldsymbol{X}_1$ 类中的样本数目为 $N_1$ ,将其中属于增生的归为一类,记为 $\boldsymbol{X}_2$ , $\boldsymbol{X}_2$ 类中的样本数目为 $N_2$ ,即图像总的类别数为 $C=2$ ,且 $N=N_1+N_2$ ,各图像的高维特征维数为 $D=320\times240$ 。设任一癌变类图像为测试图像,下面为算法的具体实现。

1) $\forall x\in\boldsymbol{X}$ ,对测试样本任一数据 $x_i$ 与训练样本 $\boldsymbol{X}$ 集,并计算邻接点 $x_{i,j}$ 与当前点 $x_i$ 之间的距离 $\alpha$ ,其中: $x_i$ 为测试样 $x$ 的第 $i$ 个数据样本点;

2) $\forall x\in\boldsymbol{X}$ ,如果存在 $p_i$ 个点满足 $d_{x_{ij},x_i}<\gamma$ ,那么这 $p_i$ 个点就可以作为最近邻域点,利用欧几里德距离计算测试样本 $x_i$ 与 $p_i$ 个癌类训练样本的最近邻距离 $\beta$ ;

3) $\forall x\in\boldsymbol{X}$ ,利用 $x_i$ 与 $\beta$ 构建癌类训练样本的权值矩阵 $\boldsymbol{w}_{i,j}$ ,计算 $x_i$ 两个不同邻域 $x_{ij}$ 和 $x_{il}$ 的协方差矩阵: $\boldsymbol{P}^i=(x_i-x_{ij})(x_i-x_{il})$ ;

4) $\forall x\in\boldsymbol{X}$ ,通过 $w_{i,j}$ 计算训练样本 $\boldsymbol{X}$ 的低维嵌入矩阵 $\boldsymbol{Y}$ : $\boldsymbol{Y}=\sum_{i=1}^{N}\left|y_i-\sum_{j=1}^{p_i}w_{ij}y_{ij}\right|^2$ ;

5) $\forall x\in\boldsymbol{X}$ ,利用矩阵 $\boldsymbol{Y}$ 计算 $x_i$ 与 $\beta$ 是否属于同一最优面,若是即可判别为癌变;

6) $\forall x\in\boldsymbol{X}$ ,假定任一测试样本为增生类图像,做与步骤1和步骤3同样的操作,得到增生类的低维嵌入矩阵 $\boldsymbol{Y}'$ ,利用 $\boldsymbol{Y}'$ 计算任一增生类训练样本与增生类最邻域距离是否属于同一最优平面,若是即可判别为增生类。

## 3　实验及其结果分析

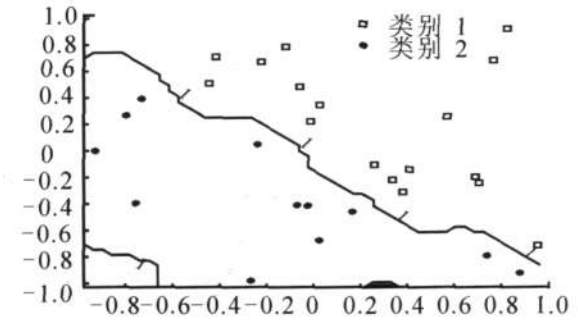
由于肿瘤细胞图像的复杂性,一次性分为5类比较困难,所以本文的实验目标是将其分为正常、癌变和增生3类。在实验过程中,选取45幅癌类胃粘膜图像和45幅增生类胃粘膜图像作为训练样本集,135幅癌类胃粘膜图像和28幅增生类胃粘膜图像作为测试样本。假定任一幅癌类图像作为测试图像。

### 3.1　基于LLE的图像降维和聚类

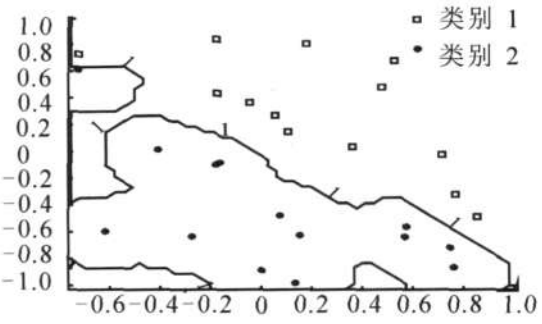
在用LLE进行降维和聚类时,分别选取不同的邻域数,来测试选取不同参数的实验效果,首先用LLE将3类样本的高维特征映射到低维的线性空间,然后再结合LS_SVM对其进行有效的分类和对其进行线性拟合。经过反复的实验可知,选取不同的 $K$ 邻域数,实验的效果是不同的,经过实验比对,选取邻域数 $K$=12时,LLE算法的降维效果是最好的。

### 3.2　基于LS_SVM的线性拟合

LS_SVM的线性回归和核函数及核参数的选择是线性拟合的关键,这里选取径向基(RBF)核函数作为核函数,并设置不同的 $\gamma$ 和 $\sigma^2$ 值来比对实验结果,图3为选取不同参数的实验结果。




(a) LSSVM+RBF, $\gamma$=10, $\sigma^2$=0.2

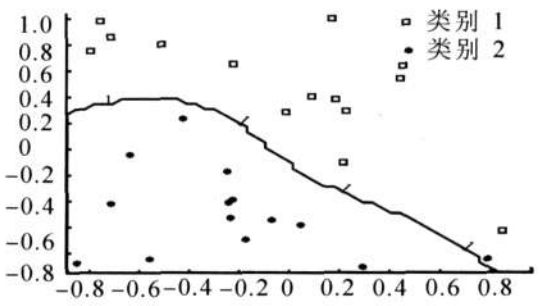



(b) LSSVM+RBF, $\gamma$=15, $\sigma^2$=0.1

由图3的实验结果可知，选取不同的核参数，对实验结果影响很大，当选择 $\gamma=1$，$\sigma^2=0.9$ 时，拟合效果最明显。

### 3.3 基于LS_SVM的线性评估

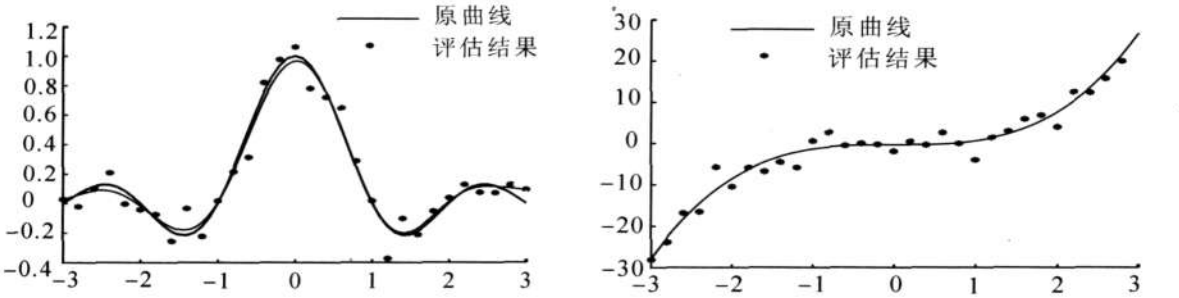
基于LS_SVM的线性回归特性，对实验结果进行线性评估，以进一步验证该方法的有效性。图4为选取不同参数对其进行线性评估的结果。




（c）LSSVM+RBF，$\gamma=1$，$\sigma^2=0.9$

**图3 LS_SVM的线性拟合结果**

**Fig.3 linear fitting results of LS_SVM**




（a）LSSVM+RBF，$\gamma=10$，$\sigma^2=0.3$　　（b）LSSVM+RBF，$\gamma=1$，$\sigma^2=1$

**图4 LS_SVM+RBF的线性估计**

**Fig.4 Linear estimation of LS_SVM + RBF**

由图4(a)可知，*点大部分落在曲线的波峰和波谷处，只有少数几个点偏离了波峰和波谷，说明该方法的聚类效果比较好。由图5(b)可知，*点几乎全部落在曲线上，只有几个点偏离，说明该方法有较高的分类准确率。

### 3.4 实验结果分析

LLE选取不同的邻域数，LS_SVM选取不同的核参数，将他们综合起来，针对分类准确率和分类运行时间以及与其它分类方法的优越性进行比较，分别做了一系列的实验比对，分类准确率由表1表示，分类运行时间由表2表示。与其它分类方法的优越性比较由表3表示，表1列出了选取不同的邻域数和不同的核参数时该方法的准确率，表2列出了选取不同的邻域数和不同的核参数时该方法的运行时间。表3列出了选取不同数量训练样本和测试样本时的分类准确率和分类运行时间。

**表1 不同参数的分类准确率**

**Tab.1 Accuracy of different parameters**

| 核参数 | 邻域数/个 | 训练/个 | 测试/个 | 准确率/% |
|---|---|---|---|---|
| $\gamma=10$，$\sigma^2=0.2$ | $K=8$ | 100 | 80 | 80.39 |
| $\gamma=15$，$\sigma^2=0.1$ | $K=10$ | 100 | 80 | 71.08 |
| $\gamma=1$，$\sigma^2=0.9$ | $K=12$ | 100 | 80 | 87.58 |

由表1可知，当领域数选取12，核参数选取 $\gamma=1$，$\sigma^2=0.9$ 时，此时分类识别的准确率最高，为87.58%。

**表2 不同参数的运行时间**

**Tab.2 Running time of different parameters**

| 训练样本/个 | 测试/个 | 核参数 | 邻域/个 | LLE+LS_SVM/s |
|---|---|---|---|---|
| 50 | 30 | $\gamma=10$，$\sigma^2=0.2$ | 17 | 68.97 |
| 100 | 50 | $\gamma=15$，$\sigma^2=0.1$ | 15 | 14.25 |
| 150 | 80 | $\gamma=8$，$\sigma^2=0.2$ | 13 | 39.56 |
| 200 | 110 | $\gamma=4$，$\sigma^2=0.8$ | 10 | 27.32 |
| 300 | 200 | $\gamma=1$，$\sigma^2=0.9$ | 12 | 9.79 |

由表2可知，当领域数选取12，核参数选取 $\gamma=1$，$\sigma^2=0.9$ 时，分类识别运行时间受样本集大小影响最小，运行速度最快。

**表3　不同分类方法的比较**

**Tab.3　Comparison of different classification methods**

| 分类方法 | 训练样本/个 | 测试样本/个 | 分类准确率/% | 分类运行时间/s |
|---|---|---|---|---|
| PCA | 50 | 130 | 72.9 | 43.8 |
| LLE | 50 | 130 | 73.3 | 42.1 |
| PCA+LDA | 50 | 130 | 80.1 | 30.4 |
| LLE+LDA | 50 | 130 | 81.6 | 26.1 |
| LLE+LS_SVM | 50 | 130 | 87.3 | 12.7 |

由表3可知，本文分类方法在分类准确率和分类运行时间两方面都比其它分类方法有很大的优越性。

## 4　结束语

基于LLE解决非线性数据与LS_SVM解决高维数据、线性拟合的优势，将它们结合应用于胃粘膜肿瘤细胞图像的分类过程中，实验结果表明，该方法在分类准确率和运行时间方面都有很大的优势。但是该方法一次将肿瘤细胞分为5类的识别率不高，因此本文初步将肿瘤细胞图像分为正常、肿瘤和增生3大类，实验证明效果较好。在下一步工作中，可再次采用该方法将增生分为轻度增生、中都增生和重度增生3类，以完成对胃粘膜肿瘤细胞图像5个类别的分类识别工作。

**参考文献：**

[1] LANGAN W. Imporved PCA+LD aapplies to gastric cancer image classification [J]. Journal of Computational Information Systems，2010，6(14)：4867-4875.

[2] SAM T ROWEIS，LAWRENCE K S. Nonlinear dimensionality reduction by lo-cally linear embedding [J]. Science，2000，290(5500)：2323-2326.

[3] VAPNIKVN. The Nature of Statistical Learning Theory [M]. NewYork：Springer，1995.

[4] PYUNG K，SEHUN R. Three-dimensional inspection of ball grid array using laser vision system [J]. IEEE Transactions on Electronics Packaging Manufacturing，1999，22(2)：151-155.

[5] KWANG I K. Support vector machines for texture classification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2002，124(11)：1542-1550.

[6] 候越先，吴静怡，何丕廉. 基于局域主方向重构的适应性非线性维数约减[J]. 计算机应用，2006，26(4)：895-897.

[7] 文贵华，包丽，丁月华. 局部线性嵌入算法中参数的选取[J]. 计算机应用研究，2007，10(2)：60-62.

[8] 崔世林，樊京. 最小二乘支持向量机及其在故障诊断中的应用[J]. 微计算机信息，2006，22(6)：214-216.

[9] 彭代强，林幼权. 基于AdaBoost算法的加权二乘向量回归机[J]. 计算机应用，2010，30(3)：776-778.

# Classification of Gastric Cancer Cells Based on LLE and LS_SVM

Gan Lan，Lv Wenya

（School of Information Engineering，East China Jiaotong University，Nanchang 330013，China）

**Abstract**：It is difficult to recognize gastric tumor cell images by the the linear classification methods for the complexity of gastric tumor cell images，the irregular shape of tissues and organs and the differentiation of different cells. As nonlinear classification methods，local linear embedding（LLE）can well deal with nonlinear data and least squares support vector machine（LS_SVM）can well resolve small sample size，high dimension and generalization issues. A classification method is proposed in this paper based on LLE and LS_SVM. The linear fitting function is used to fit its linear errors，the linear fitting error is used to determine the results，finally superiority of method in this paper is compared with other classification methods.It is proved by the experiment results that this method has a significant advantage in classification accuracy and running time.

**Key words**：locally linear embedding；least square support vector machine；tumor cell classfication