



(12) 发明专利申请

(10) 申请公布号 CN 116935384 A

(43) 申请公布日 2023. 10. 24

(21) 申请号 202311196273.3

G06V 10/77 (2022.01)

(22) 申请日 2023.09.18

(71) 申请人 上海大学

地址 201900 上海市宝山区大场镇上大路
99号上海大学

(72) 发明人 李文君

(74) 专利代理机构 郑州知倍通知识产权代理事
务所(普通合伙) 41191

专利代理师 李玲玲

(51) Int. Cl.

G06V 20/69 (2022.01)

G01N 21/25 (2006.01)

G06V 20/10 (2022.01)

G06V 10/74 (2022.01)

G06V 10/764 (2022.01)

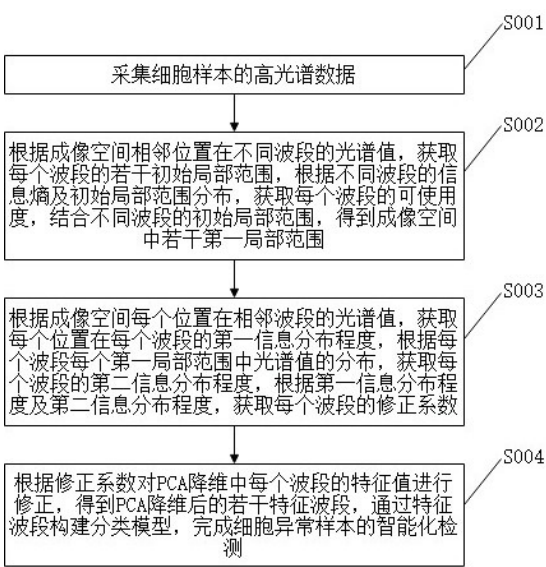
权利要求书4页 说明书10页 附图1页

(54) 发明名称

一种细胞异常样本智能化检测方法

(57) 摘要

本发明涉及光谱数据分析技术领域,提出了一种细胞异常样本智能化检测方法,包括:采集细胞样本的高光谱数据;根据成像空间相邻位置在不同波段的光谱值,获取每个波段的若干初始局部范围及可使用度,得到成像空间中若干第一局部范围;获取每个位置在每个波段的第一信息分布程度,以及每个波段的第二信息分布程度,得到每个波段的修正系数;根据修正系数对PCA降维中每个波段的特征值进行修正,得到若干特征波段,通过构建分类模型完成细胞异常样本的智能化检测。本发明旨在解决通过高光谱数据分析细胞过程中,由于对高光谱数据通过固定波段进行降维而未考虑信息分布而导致分类模型不准确影响检测结果的问题。



1. 一种细胞异常样本智能化检测方法,其特征在于,该方法包括以下步骤:

采集细胞样本的高光谱数据;

根据高光谱数据成像空间中相邻位置在不同波段的光谱值,以及不同波段的信息熵,获取每个波段的若干初始局部范围及可使用度;根据波段的可使用度对初始局部范围进行合并,得到成像空间中若干第一局部范围;其中,所述可使用度为波段对应的光谱数据在光谱空间中的重要程度;

根据第一局部范围及每个位置在不同波段的光谱值,获取每个位置在每个波段的第一信息分布程度及第二信息分布程度;根据每个位置在不同波段的光谱值、第一信息分布程度及第二信息分布程度,得到每个波段的修正系数;

根据修正系数对PCA降维中每个波段的特征值进行修正,得到PCA降维后的若干特征波段,通过特征波段构建分类模型,完成细胞异常样本的智能化检测。

2. 根据权利要求1所述的一种细胞异常样本智能化检测方法,其特征在于,所述每个波段的若干初始局部范围及可使用度,具体的获取方法为:

根据高光谱数据成像空间中相邻位置在不同波段的光谱值,得到每个波段的若干初始局部范围;根据任意两个波段的每个初始局部范围在成像空间中对应的位置,通过KM匹配获取任意两个波段的若干匹配局部范围对,以及每个波段每个初始局部范围在其他每个波段的匹配局部范围;

第*i*个波段的可使用度 γ_i 的计算方法为:

$$\gamma_i = \frac{C_i}{C_{max}} \times \frac{1}{M-1} \times \sum_{m=1, m \neq i}^M \left(\frac{1}{P_{i,m}} \times \sum_{p=1}^{P_{i,m}} \exp \left(- \left| 1 - \frac{S_{i,p}}{S_{i,p,m}} \right| \right) \right)$$

其中, C_i 表示第*i*个波段的信息熵值, C_{max} 表示所有波段的信息熵值的最大值, M 表示高光谱数据中波段的数量, $P_{i,m}$ 表示第*i*个波段与第*m*个波段中匹配局部范围对的数量, $S_{i,p}$ 表示第*i*个波段中第*p*个初始局部范围的面积, $S_{i,p,m}$ 表示第*i*个波段中第*p*个初始局部范围在第*m*个波段的匹配局部范围的面积, $| \cdot |$ 表示求绝对值, $\exp(\cdot)$ 表示以自然常数为底的指数函数;

获取每个波段的可使用度。

3. 根据权利要求2所述的一种细胞异常样本智能化检测方法,其特征在于,所述得到每个波段的若干初始局部范围,包括的具体方法为:

对于任意相邻两个位置在任意一个波段中的两个光谱值,获取两个光谱值的差值绝对值,若两个光谱值的差值绝对值小于相似阈值,将该相邻两个位置在该波段记为相似位置;对于成像空间中第*j*个位置在第*i*个波段的光谱值,获取第*j*个位置在第*i*个波段的相似位置,根据相似位置对应的光谱值继续获取八邻域范围的相似位置,直到不再获取相似位置停止,将在第*i*个波段提取到的若干相似位置,作为第*j*个位置在第*i*个波段的相似范围;

根据成像空间中相邻位置在每个波段的相似位置判断,对每个波段获取若干相似范围;根据相似范围及相似范围中光谱值及相似位置的分布,得到每个相似范围中每个位置的关联性;

将相似范围中关联性大于关联阈值的位置进行保留,将相似范围中保留的部分记为初始局部范围;对每个相似范围获取初始局部范围,得到每个波段的若干初始局部范围。

4. 根据权利要求3所述的一种细胞异常样本智能化检测方法,其特征在于,所述得到每个相似范围中每个位置的关联性,包括的具体方法为:

对于第 j 个位置在第 i 个波段的相似范围,获取该相似范围的凸包结构,第 j 个位置与在第 i 个波段的相似范围内第 u 个位置的关联性 $\alpha_i(j, j_u)$ 的计算方法为:

$$\alpha_i(j, j_u) = \exp\left(-\left|1 - \frac{s_{j,i}}{s'_{j,i}}\right|\right) \times \frac{n_{j,i,u}}{N} \times \exp(-d_{j,i,u})$$

其中, $s_{j,i}$ 表示第 j 个位置在第 i 个波段的相似范围的面积, $s'_{j,i}$ 表示第 j 个位置在第 i 个波段的相似范围的凸包结构的面积, $n_{j,i,u}$ 表示第 j 个位置在第 i 个波段的相似范围中第 u 个位置在第 i 个波段的相似位置的数量, N 表示每个位置的邻域位置数量, $d_{j,i,u}$ 表示第 j 个位置在第 i 个波段的相似范围中第 u 个位置与第 j 个位置在第 i 个波段的相似范围的凸包结构的边缘位置的欧式距离的最小值, $|\cdot|$ 表示求绝对值, $\exp(\cdot)$ 表示以自然常数为底的指数函数;

获取第 j 个位置与在第 i 个波段的相似范围中每个位置的关联性;获取每个相似范围中每个位置的关联性。

5. 根据权利要求1所述的一种细胞异常样本智能化检测方法,其特征在于,所述得到成像空间中若干第一局部范围,包括的具体方法为:

将可使用度最大的前预设数量的波段进行提取,作为可使用波段,将所有可使用波段的初始局部范围在成像空间中对应的位置求并集,将合并后得到的若干局部范围,记为成像空间中的若干第一局部范围。

6. 根据权利要求1所述的一种细胞异常样本智能化检测方法,其特征在于,所述每个位置在每个波段的第一信息分布程度及第二信息分布程度,具体的获取方法为:

根据成像空间每个位置在相邻波段的光谱值,获取每个位置在每个波段的第一信息分布程度;将每个第一局部范围在每个波段对应的区域,记为每个第一范围在每个波段的局部区域,对每个局部区域通过形状上下文算法获取特征向量,根据特征向量对同一波段的若干局部区域进行聚类,得到每个波段的若干类别;

第 i 个波段的第二信息分布程度 λ_i 的计算方法为:

$$\lambda_i = \frac{Q - q_i}{Q} \times \frac{1}{q_i} \times \sum_{f=1}^{q_i} \left(\frac{g_{i,f}}{G - g_{i,f}} \times (A_{i,f} - \bar{A}_i)^2 \right)$$

其中, Q 表示第一局部范围的数量, q_i 表示第 i 个波段的类别数量, G 表示所有第一局部范围包括的位置数量, $g_{i,f}$ 表示第 i 个波段的第 f 个类别包括的所有局部区域对应的位置数量, $A_{i,f}$ 表示第 i 个波段的第 f 个类别包括的所有局部区域中的光谱值均值, $\bar{A}_i$ 表示第 i 个波段所有局部区域中的光谱值均值;

获取每个波段的第二信息分布程度,将每个波段的第二信息分布程度作为每个位置在

对应波段的第二信息分布程度。

7. 根据权利要求6所述的一种细胞异常样本智能化检测方法,其特征在于,所述每个位置在每个波段的第一信息分布程度,具体的获取方法为:

根据成像空间每个位置在每个波段的光谱值,得到每个位置的光谱序列及趋势序列,以及光谱序列中每个元素对应波段的斜率值;

第 j 个位置在第 i 个波段的第一信息分布程度 $\tau_{j,i}$ 的计算方法为:

$$\tau_{j,i} = |1 - \delta_{j,i}| \times \frac{1}{2B} \times \sum_{b=1}^B (|k_{j,i} - k_{j,i,b}^-| + |k_{j,i} - k_{j,i,b}^+|)$$

其中, $\delta_{j,i}$ 表示第 j 个位置的趋势序列在去除第 i 个波段对应元素前后的元素值方差比值, B 表示相邻波段范围数, $k_{j,i}$ 表示第 j 个位置的光谱序列中第 i 个波段对应元素的斜率值, $k_{j,i,b}^-$ 表示第 j 个位置的光谱序列中第 i 个波段前第 b 个波段对应元素的斜率值, $k_{j,i,b}^+$ 表示第 j 个位置的光谱序列中第 i 个波段后第 b 个波段对应元素的斜率值, $||$ 表示求绝对值;

获取每个位置在每个波段的第一信息分布程度。

8. 根据权利要求7所述的一种细胞异常样本智能化检测方法,其特征在于,所述得到每个位置的光谱序列及趋势序列,以及光谱序列中每个元素对应波段的斜率值,包括的具体方法为:

将第 j 个位置在每个波段的光谱值按波段顺序进行排列,得到第 j 个位置的光谱序列,对光谱序列中每个元素计算斜率值,得到第 j 个位置的光谱序列中每个元素对应波段的斜率值;获取每个位置的光谱序列及光谱序列中每个元素对应波段的斜率值;

对第 j 个位置的光谱序列通过DFA去趋势分析算法进行处理,得到的序列记为第 j 个位置的趋势后序列;对第 j 个位置的光谱序列减去第 j 个位置的趋势后序列得到的差值序列,记为第 j 个位置的趋势序列;获取每个位置的趋势序列。

9. 根据权利要求8所述的一种细胞异常样本智能化检测方法,其特征在于,所述得到每个波段的修正系数,包括的具体方法为:

对同一位置的所有第一信息分布程度进行线性归一化,得到的结果作为该位置在每个波段的第一信息程度,获取每个位置在每个波段的第一信息程度;将所有位置在所有波段的第二信息分布程度进行线性归一化,得到的结果作为每个位置在每个波段的第二信息程度;

对于第 j 个位置与第 $j+1$ 个位置,获取两个位置的光谱序列的皮尔逊相关系数,获取两个位置的光谱序列中相同次序的元素的差值绝对值,将两个位置的光谱序列中第一个元素的差值绝对值与所有元素的差值绝对值的最大值的比值,与皮尔逊相关系数的乘积,作为两个位置的光谱相似度;

将第 j 个位置与其他所有位置的光谱相似度的均值,作为第 j 个位置的整体参考权重,将1减去整体参考权重的差值作为第 j 个位置的局部参考权重,将局部参考权重作为第一信息程度的权重,整体参考权重作为第二信息程度的权重,对同一波段下的第一信息程度及第

二信息程度进行加权求和得到第 j 个位置在该波段的综合信息程度；

获取每个位置在每个波段的综合信息程度，将所有位置在同一波段的综合信息程度的均值，作为该波段的修正系数，得到每个波段的修正系数。

10. 根据权利要求1所述的一种细胞异常样本智能化检测方法，其特征在于，所述得到PCA降维后的若干特征波段，包括的具体方法为：

在对高光谱数据进行PCA降维过程中，获取到协方差矩阵中每个波段的特征值后，将每个波段的修正系数与特征值的乘积作为每个波段的修正后的特征值；根据修正后的特征值完成对高光谱数据的PCA降维，得到若干特征波段。

一种细胞异常样本智能化检测方法

技术领域

[0001] 本发明涉及光谱数据分析技术领域,具体涉及一种细胞异常样本智能化检测方法。

背景技术

[0002] 高光谱成像技术是一种能够获取连续多个波长范围内的光谱信息的成像方法。它结合了光谱学和成像技术,可以提供丰富的光谱数据以及空间分辨率。在高光谱成像中,将不同波长的光分解并捕获,形成高维的光谱数据,以获取样本在不同光谱特征下的反射、吸收或荧光等信息。细胞异常样本的智能检测是生命科学和医学领域中的重要任务之一,传统的细胞异常样本检测是根据传统的化学分析方法,经过复杂的前处理进行检测,而高光谱成像技术可以进行无损无接触检测的优点,利用高光谱成像技术来对细胞异常样本进行鉴别得到了广泛的应用。

[0003] 对于细胞异常样本的检测,往往是通过构建分类模型来进行识别,而在构建过程中,由于高光谱数据中包含数百个波段,而每个波段中成像空间和光谱信息有较强的关联性,进而造成了很大的冗余性,若将所有的波段数据用于构建分类模型,会由于数据量大,且对特征提取识别造成很大的问题,进而得到错误的分类模型;因此通常会对采集的高光谱数据进行PCA降维处理,在进行PCA降维过程中,会使用固定的前几个波段主要成分来代替高光谱数据,这会造成构建分类模型的结果取决于这几个主要成分的好坏;同时PCA降维过程中,是根据主成分之间的方差大小来表示信息的多少,因此会由于没有考虑噪声的影响以及高光谱数据中不同波段的信息含量而造成信息的大量丢失,进而影响分类模型的构建,导致细胞样本的异常检测结果不准确。

发明内容

[0004] 本发明提供一种细胞异常样本智能化检测方法,以解决现有的通过高光谱数据分析细胞过程中,由于对高光谱数据通过固定波段进行降维而未考虑信息分布而导致分类模型不准确影响检测结果的问题,所采用的技术方案具体如下:

本发明一个实施例提供了一种细胞异常样本智能化检测方法,该方法包括以下步骤:

采集细胞样本的高光谱数据;

根据高光谱数据成像空间中相邻位置在不同波段的光谱值,以及不同波段的信息熵,获取每个波段的若干初始局部范围及可使用度;根据波段的可使用度对初始局部范围进行合并,得到成像空间中若干第一局部范围;其中,所述可使用度为波段对应的光谱数据在光谱空间中的重要程度;

根据第一局部范围及每个位置在不同波段的光谱值,获取每个位置在每个波段的第一信息分布程度及第二信息分布程度;根据每个位置在不同波段的光谱值、第一信息分布程度及第二信息分布程度,得到每个波段的修正系数;

根据修正系数对PCA降维中每个波段的特征值进行修正,得到PCA降维后的若干特征波段,通过特征波段构建分类模型,完成细胞异常样本的智能化检测。

[0005] 进一步的,所述每个波段的若干初始局部范围及可使用度,具体的获取方法为:

根据高光谱数据成像空间中相邻位置在不同波段的光谱值,得到每个波段的若干初始局部范围;根据任意两个波段的每个初始局部范围在成像空间中对应的位置,通过KM匹配获取任意两个波段的若干匹配局部范围对,以及每个波段每个初始局部范围在其他每个波段的匹配局部范围;

第*i*个波段的可使用度 γ_i 的计算方法为:

$$\gamma_i = \frac{C_i}{C_{max}} \times \frac{1}{M-1} \times \sum_{m=1, m \neq i}^M \left(\frac{1}{P_{i,m}} \times \sum_{p=1}^{P_{i,m}} \exp \left(- \left| 1 - \frac{S_{i,p}}{S_{i,p,m}} \right| \right) \right)$$

其中, C_i 表示第*i*个波段的信息熵值, C_{max} 表示所有波段的信息熵值的最大值, M 表示高光谱数据中波段的数量, $P_{i,m}$ 表示第*i*个波段与第*m*个波段中匹配局部范围对的数量, $S_{i,p}$ 表示第*i*个波段中第*p*个初始局部范围的面积, $S_{i,p,m}$ 表示第*i*个波段中第*p*个初始局部范围在第*m*个波段的匹配局部范围的面积, $||$ 表示求绝对值, $\exp()$ 表示以自然常数为底的指数函数;

获取每个波段的可使用度。

[0006] 进一步的,所述得到每个波段的若干初始局部范围,包括的具体方法为:

对于任意相邻两个位置在任意一个波段中的两个光谱值,获取两个光谱值的差值绝对值,若两个光谱值的差值绝对值小于相似阈值,将该相邻两个位置在该波段记为相似位置;对于成像空间中第*j*个位置在第*i*个波段的光谱值,获取第*j*个位置在第*i*个波段的相似位置,根据相似位置对应的光谱值继续获取八邻域范围的相似位置,直到不再获取相似位置停止,将在第*i*个波段提取到的若干相似位置,作为第*j*个位置在第*i*个波段的相似范围;

根据成像空间中相邻位置在每个波段的相似位置判断,对每个波段获取若干相似范围;根据相似范围及相似范围中光谱值及相似位置的分布,得到每个相似范围中每个位置的关联性;

将相似范围中关联性大于关联阈值的位置进行保留,将相似范围中保留的部分记为初始局部范围;对每个相似范围获取初始局部范围,得到每个波段的若干初始局部范围。

[0007] 进一步的,所述得到每个相似范围中每个位置的关联性,包括的具体方法为:

对于第*j*个位置在第*i*个波段的相似范围,获取该相似范围的凸包结构,第*j*个位置与在第*i*个波段的相似范围内第*u*个位置的关联性 $\alpha_i(j, j_u)$ 的计算方法为:

$$\alpha_i(j, j_u) = \exp \left(- \left| 1 - \frac{s_{j,i}}{s'_{j,i}} \right| \right) \times \frac{n_{j,i,u}}{N} \times \exp(-d_{j,i,u})$$

其中, $s_{j,i}$ 表示第*j*个位置在第*i*个波段的相似范围的面积, $s'_{j,i}$ 表示第*j*个位置在第*i*个波段的相似范围的凸包结构的面积, $n_{j,i,u}$ 表示第*j*个位置在第*i*个波段的相似范围中第*u*

个位置在第*i*个波段的相似位置的数量,*N*表示每个位置的邻域位置数量, $d_{j,i,u}$ 表示第*j*个位置在第*i*个波段的相似范围中第u个位置与第*j*个位置在第*i*个波段的相似范围的凸包结构的边缘位置的欧式距离的最小值, $||$ 表示求绝对值, $\exp()$ 表示以自然常数为底的指数函数;

获取第*j*个位置与在第*i*个波段的相似范围中每个位置的关联性;获取每个相似范围中每个位置的关联性。

[0008] 进一步的,所述得到成像空间中若干第一局部范围,包括的具体方法为:

将可使用度最大的前预设数量的波段进行提取,作为可使用波段,将所有可使用波段的初始局部范围在成像空间中对应的位置求并集,将合并后得到的若干局部范围,记为成像空间中的若干第一局部范围。

[0009] 进一步的,所述每个位置在每个波段的第一信息分布程度及第二信息分布程度,具体的获取方法为:

根据成像空间每个位置在相邻波段的光谱值,获取每个位置在每个波段的第一信息分布程度;将每个第一局部范围在每个波段对应的区域,记为每个第一范围在每个波段的局部区域,对每个局部区域通过形状上下文算法获取特征向量,根据特征向量对同一波段的若干局部区域进行聚类,得到每个波段的若干类别;

第*i*个波段的第二信息分布程度 λ_i 的计算方法为:

$$\lambda_i = \frac{Q - q_i}{Q} \times \frac{1}{q_i} \times \sum_{f=1}^{q_i} \left(\frac{g_{i,f}}{G - g_{i,f}} \times (A_{i,f} - \bar{A}_i)^2 \right)$$

其中,*Q*表示第一局部范围的数量,*q_i*表示第*i*个波段的类别数量,*G*表示所有第一局部范围包括的位置数量, $g_{i,f}$ 表示第*i*个波段的第*f*个类别包括的所有局部区域对应的位置数量, $A_{i,f}$ 表示第*i*个波段的第*f*个类别包括的所有局部区域中的光谱值均值, $\bar{A}_i$ 表示第*i*个波段所有局部区域中的光谱值均值;

获取每个波段的第二信息分布程度,将每个波段的第二信息分布程度作为每个位置在对应波段的第二信息分布程度。

[0010] 进一步的,所述每个位置在每个波段的第一信息分布程度,具体的获取方法为:

根据成像空间每个位置在每个波段的光谱值,得到每个位置的光谱序列及趋势序列,以及光谱序列中每个元素对应波段的斜率值;

第*j*个位置在第*i*个波段的第一信息分布程度 $\tau_{j,i}$ 的计算方法为:

$$\tau_{j,i} = |1 - \delta_{j,i}| \times \frac{1}{2B} \times \sum_{b=1}^B (|k_{j,i} - k_{j,i,b}^-| + |k_{j,i} - k_{j,i,b}^+|)$$

其中, $\delta_{j,i}$ 表示第*j*个位置的趋势序列在去除第*i*个波段对应元素前后的元素值方差比值,*B*表示相邻波段范围数, $k_{j,i}$ 表示第*j*个位置的光谱序列中第*i*个波段对应元素的斜率值, $k_{j,i,b}^-$ 表示第*j*个位置的光谱序列中第*i*个波段前第*b*个波段对应元素的斜率值, $k_{j,i,b}^+$ 表示

第 j 个位置的光谱序列中第 i 个波段后第 b 个波段对应元素的斜率值, $| |$ 表示求绝对值;

获取每个位置在每个波段的第一信息分布程度。

[0011] 进一步的,所述得到每个位置的光谱序列及趋势序列,以及光谱序列中每个元素对应波段的斜率值,包括的具体方法为:

将第 j 个位置在每个波段的光谱值按波段顺序进行排列,得到第 j 个位置的光谱序列,对光谱序列中每个元素计算斜率值,得到第 j 个位置的光谱序列中每个元素对应波段的斜率值;获取每个位置的光谱序列及光谱序列中每个元素对应波段的斜率值;

对第 j 个位置的光谱序列通过DFA去趋势分析算法进行处理,得到的序列记为第 j 个位置的趋势后序列;对第 j 个位置的光谱序列减去第 j 个位置的趋势后序列得到的差值序列,记为第 j 个位置的趋势序列;获取每个位置的趋势序列。

[0012] 进一步的,所述得到每个波段的修正系数,包括的具体方法为:

对同一位置的所有第一信息分布程度进行线性归一化,得到的结果作为该位置在每个波段的第一信息程度,获取每个位置在每个波段的第一信息程度;将所有位置在所有波段的第二信息分布程度进行线性归一化,得到的结果作为每个位置在每个波段的第二信息程度;

对于第 j 个位置与第 $j+1$ 个位置,获取两个位置的光谱序列的皮尔逊相关系数,获取两个位置的光谱序列中相同次序的元素的差值绝对值,将两个位置的光谱序列中第一个元素的差值绝对值与所有元素的差值绝对值的最大值的比值,与皮尔逊相关系数的乘积,作为两个位置的光谱相似度;

将第 j 个位置与其他所有位置的光谱相似度的均值,作为第 j 个位置的整体参考权重,将1减去整体参考权重的差值作为第 j 个位置的局部参考权重,将局部参考权重作为第一信息程度的权重,整体参考权重作为第二信息程度的权重,对同一波段下的第一信息程度及第二信息程度进行加权求和得到第 j 个位置在该波段的综合信息程度;

获取每个位置在每个波段的综合信息程度,将所有位置在同一波段的综合信息程度的均值,作为该波段的修正系数,得到每个波段的修正系数。

[0013] 进一步的,所述得到PCA降维后的若干特征波段,包括的具体方法为:

在对高光谱数据进行PCA降维过程中,获取到协方差矩阵中每个波段的特征值后,将每个波段的修正系数与特征值的乘积作为每个波段的修正后的特征值;根据修正后的特征值完成对高光谱数据的PCA降维,得到若干特征波段。

[0014] 本发明的有益效果是:本发明通过高光谱成像技术对细胞异常样本进行分类模型的构建,从而保证对细胞样本无损无接触,同时通过对高光谱数据进行自适应的PCA降维,提高细胞样本异常检测结果的准确性;其中通过获取高光谱数据每个波段的修正系数,使得包含信息且可以反映单独某种细胞结构的波段,更容易被作为主成分进行降维,从而作为特征波段被获取,保证降维后的高光谱数据可以保留有效信息;修正系数获取过程中,通过成像空间中每个位置在不同波段中的光谱值,获取第一局部范围来反映连通的某个细胞结构区域,进而根据同一位置在不同波段的整体及局部的趋势变化,以及第一局部范围在波段的信息分布,获取修正系数,保证最终PCA降维过程中可以保留更多有效信息,提高分

类模型构建的准确性,进而提高细胞样本的智能化异常检测结果的准确性。

附图说明

[0015] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动性的前提下,还可以根据这些附图获得其他的附图。

[0016] 图1为本发明一个实施例所提供的一种细胞异常样本智能化检测方法流程示意图。

具体实施方式

[0017] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0018] 请参阅图1,其示出了本发明一个实施例所提供的一种细胞异常样本智能化检测方法流程图,该方法包括以下步骤:

步骤S001、采集细胞样本的高光谱数据。

[0019] 本实施例的目的是通过高光谱成像技术对细胞样本进行智能化的异常检测,因此首先需要获取细胞样本的高光谱数据;由于是对细胞样本进行光谱分析,本实施例通过显微高光谱成像系统进行高光谱数据获取,其中显微高光谱成像系统包括光学显微系统、AOTF分光计、CCD及支架等,高光谱成像技术为公知技术,本实施例不再赘述;本实施例通过获取100份细胞切片来作为所有的细胞样本进行分析,其中每个细胞切片包含若干细胞样本,对每个细胞切片得到对应的高光谱数据,后续均以任意一个高光谱数据进行分析。

[0020] 至此,获取到了细胞样本的高光谱数据。

[0021] 步骤S002、根据成像空间相邻位置在不同波段的光谱值,获取每个波段的若干初始局部范围,根据不同波段的信息熵及初始局部范围分布,获取每个波段的可使用度,结合不同波段的初始局部范围,得到成像空间中若干第一局部范围。

[0022] 需要说明的是,通过对细胞样本的高光谱数据进行自适应的PCA降维获取特征波段,而本实施例所期望的特征波段需要同时满足两个条件,其一是特征波段在光谱空间中包含重要组成部分的信息,也就是在光谱空间中特征波段对应的光谱数据较为重要,其二是特征波段在成像空间中包含的信息差异性较大,也就是在成像空间中特征波段对应的光谱数据与其他光谱数据存在较大差异。在光谱空间中包含较多信息,且局部分布和整体分布均包含较多信息;而对于成像空间,特征波段之间差异应较大,不同特征波段仅对应成像空间上的一部分信息,例如不同特征波段分别反映细胞壁边缘、细胞质部分以及细胞核部分的信息,保证所有特征波段在成像空间上不会存在信息冗余。

[0023] 进一步需要说明的是,对于成像空间中每个位置,每个位置属于细胞样本的某个区域,相同区域的位置在不同波段上的光谱值存在一定的相近性,因此需要根据这种相近性对每个波段进行初始局部范围获取,即根据光谱值进行初始局部范围的区域划分;再根

据不同波段包含的信息程度,量化得到每个波段的可使用度,进而得到对于成像空间中的第一局部范围划分,第一局部范围可以反映同一类别的一个细胞结构区域。

[0024] 具体的,对于任意相邻两个位置在任意一个波段中的两个光谱值,获取两个光谱值的差值绝对值,预设一个相似阈值,本实施例相似阈值采用10进行叙述,若两个光谱值的差值绝对值小于相似阈值,则将该相邻两个位置在该波段记为相似位置,其中相邻本实施例定义为八邻域相邻;以成像空间中第 j 个位置在第 i 个波段的光谱值为例,获取第 j 个位置在第 i 个波段的相似位置,并根据相似位置对应的光谱值继续获取八邻域范围的相似位置,直到不再获取相似位置停止,将此时在第 i 个波段提取到的若干相似位置,作为第 j 个位置在第 i 个波段的相似范围;对该相似范围获取凸包结构,凸包结构获取为公知技术,本实施例不再赘述;则第 j 个位置与在第 i 个波段的相似范围内第 u 个位置的关联性 $\alpha_i(j, j_u)$ 的计算方法为:

$$\alpha_i(j, j_u) = \exp\left(-\left|1 - \frac{s_{j,i}}{s'_{j,i}}\right|\right) \times \frac{n_{j,i,u}}{N} \times \exp(-d_{j,i,u})$$

其中, $s_{j,i}$ 表示第 j 个位置在第 i 个波段的相似范围的面积, $s'_{j,i}$ 表示第 j 个位置在第 i 个波段的相似范围的凸包结构的面积, $n_{j,i,u}$ 表示第 j 个位置在第 i 个波段的相似范围中第 u 个位置在第 i 个波段的相似位置的数量, N 表示每个位置的邻域位置数量,本实施例采用 $N=8$ 进行叙述, $d_{j,i,u}$ 表示第 j 个位置在第 i 个波段的相似范围中第 u 个位置与第 j 个位置在第 i 个波段的相似范围的凸包结构的边缘位置的欧式距离的最小值,即相似范围中该位置与凸包结构中距离最近的边缘位置的欧式距离, $||$ 表示求绝对值, $\exp()$ 表示以自然常数为底的指数函数,本实施例通过 $\exp(-)$ 函数来呈现反比例关系及归一化处理,实施者可根据实际情况设置反比例函数及归一化函数;

首先对相似范围进行圆润程度的量化,通过相似范围的面积与凸包结构的面积获取比值,根据比值与1的差异来判断圆润程度,差异越小,相似范围越接近凸包结构,越为圆润,相似范围越形近细胞结构,对相似范围内的其他位置获取关联性的可信度越高,得到的关联性也就越大;通过对相似范围内位置的相似位置数量的量化,以及与凸包结构边缘位置欧式距离的量化,相似位置数量越多,关联性越大,同时利用距离来对相似位置数量的量化值进行限制,与边缘位置欧式距离越小,对凸包构成越重要,避免相似范围内部的位置的相似位置数量较大而获取过大的关联性;同时通过归一化函数保证关联性计算过程中三个部分的量纲同一,进而最终得到关联性。

[0025] 进一步的,按照上述方法获取第 j 个位置与在第 i 个波段的相似范围中每个位置的关联性,预设一个关联阈值,本实施例关联阈值采用0.65进行叙述,将相似范围中关联性大于关联阈值的位置进行保留,将相似范围中保留的部分记为初始局部范围;需要说明的是,相似范围的获取对于起始位置的获取是任意的,因此相似范围中每个位置的关联性是基于相似范围的,而不是基于起始位置的,即对于任意一个波段,最终都可以根据相似位置判断得到若干相似范围,同时相似范围获取初始局部范围也与起始位置无关,本实施例中第 j 个

位置仅为举例说明;因此,按照上述方法对每个波段分别获取若干相似范围,并对每个相似范围获取初始局部范围,则得到了每个波段的若干初始局部范围。

[0026] 进一步的,对于任意两个波段的若干初始局部范围,每个初始局部范围在成像空间中均对应若干位置,因此对两个波段的初始局部范围进行KM匹配,将一个波段的每个初始局部范围作为二分图中左侧节点,另一个波段的每个初始局部范围作为二分图中右侧节点,左右两侧节点之间的边值采用对应的两个初始局部范围在成像空间中对应的相同位置的数量,通过KM匹配得到若干匹配局部范围对,将匹配局部范围对中的两个初始局部范围作为各自的匹配局部范围;按照上述方法获取任意两个波段的若干匹配局部范围对,以及每个波段每个初始局部范围在其他每个波段的匹配局部范围,需要说明的是,存在某个波段的某个初始局部范围在其他某个波段不存在匹配局部范围的情况,该情况不影响后续处理,本实施例不再赘述。

[0027] 进一步的,以第*i*个波段为例,第*i*个波段的可使用度 γ_i 的计算方法为:

$$\gamma_i = \frac{C_i}{C_{max}} \times \frac{1}{M-1} \times \sum_{m=1, m \neq i}^M \left(\frac{1}{P_{i,m}} \times \sum_{p=1}^{P_{i,m}} \exp\left(-\left|1 - \frac{S_{i,p}}{S_{i,p,m}}\right|\right) \right)$$

其中, C_i 表示第*i*个波段的信息熵值, C_{max} 表示所有波段的信息熵值的最大值,信息熵计算为公知技术,本实施例不再赘述; M 表示高光谱数据中波段的数量, $P_{i,m}$ 表示第*i*个波段与第*m*个波段中匹配局部范围对的数量, $S_{i,p}$ 表示第*i*个波段中第*p*个初始局部范围的面积, $S_{i,p,m}$ 表示第*i*个波段中第*p*个初始局部范围在第*m*个波段的匹配局部范围的面积, $||$ 表示求绝对值, $\exp()$ 表示以自然常数为底的指数函数,本实施例通过 $\exp(-)$ 函数来呈现反比例关系及归一化处理,实施者可根据实际情况设置反比例函数及归一化函数;

首先,可使用度可以表征波段对应的光谱数据在光谱空间中的重要程度,波段的可使用度越大,在光谱空间中该波段对应的光谱数据就越重要,后续被用确定为可使用波段的概率就会越大;通过信息熵值来量化可使用度,信息熵越大则可使用度越大;其次通过该波段与其他每个波段在匹配局部范围对中初始局部范围的面积差异来量化可使用度,面积差异越小,可使用度越大;首先对该波段与任意一个其他波段中每个匹配局部范围对中两个初始局部范围的面积进行比值计算,比值越接近1,面积差异越小;再对与其他所有波段的面积差异量化值求均值,最终得到可使用度,其中对与其他所有波段的面积差异量化值求均值的计算过程中,由于不包括第*i*个波段,因此通过 $\frac{1}{M-1}$ 来进行均值计算;按照上述方法获取每个波段的可使用度。

[0028] 进一步的,预设一个可使用数量,本实施例可使用数量采用5进行叙述,将可使用度最大的前五个波段进行提取,将这些波段作为可使用波段,将可使用波段的所有初始局部范围进行合并,即是对每个初始局部范围在成像空间中对应的位置求并集,将合并后得到的若干局部范围,记为成像空间中的若干第一局部范围,即由于是在成像空间中进行位置并集计算,因此最终得到的第一局部范围也是基于成像空间得到的。

[0029] 至此,得到了成像空间中若干第一局部范围。

[0030] 步骤S003、根据成像空间每个位置在相邻波段的光谱值,获取每个位置在每个波

段的第一信息分布程度,根据每个波段每个第一局部范围中光谱值的分布,获取每个波段的第二信息分布程度,根据第一信息分布程度及第二信息分布程度,获取每个波段的修正系数。

[0031] 需要说明的是,获取到第一局部范围后,则需要对成像空间中每个位置进行信息程度的分析,通过每个位置在不同波段的第一信息分布程度,以及每个波段根据第一局部范围得到的信息分布程度,完成对每个波段的修正系数的量化,进而得到修正系数;其中第一信息分布程度根据每个位置在每个波段的光谱值的变化来获取,通过趋势变化及方差贡献量化得到第一信息分布程度,方差贡献越大,同时趋势变化越大,该位置在该波段的第一信息分布程度就越大;而第二信息分布程度则根据不同波段中第一局部范围中光谱值分布来量化,通过对不同波段中第一局部范围根据对应的光谱值进行聚类,根据聚类结果进行量化,类别数量越少,表征信息种类越少,越能反映单种细胞结构信息,则波段的第二信息分布程度越大。

[0032] 具体的,以第 j 个位置为例,将第 j 个位置在每个波段的光谱值按波段顺序进行排列,得到第 j 个位置的光谱序列,对光谱序列中每个元素计算斜率值(纵坐标为光谱值,横坐标为波长),则得到了第 j 个位置的光谱序列中每个元素对应波段的斜率值,其中第一个元素的斜率值设置为第二个元素的斜率值;对光谱序列通过DFA去趋势分析算法进行处理,得到的序列记为趋势后序列,DFA去趋势分析算法为公知技术,本实施例不再赘述;对光谱序列减去趋势后序列得到的差值序列(对应次序的元素相减),记为第 j 个位置的趋势序列;则第 j 个位置在第 i 个波段的第一信息分布程度 $\tau_{j,i}$ 的计算方法为:

$$\tau_{j,i} = |1 - \delta_{j,i}| \times \frac{1}{2B} \times \sum_{b=1}^B (|k_{j,i} - k_{j,i,b}^-| + |k_{j,i} - k_{j,i,b}^+|)$$

其中, $\delta_{j,i}$ 表示第 j 个位置的趋势序列在去除第 i 个波段对应元素前后的元素值方差比值,其中比值由小值比大值得到,通过计算趋势序列的元素值方差,以及趋势序列中去除第 i 个波段对应元素后的元素值方差,得到元素值方差比值; B 表示相邻波段范围数,本实施例采用 $B = 5$ 进行叙述,需要说明的是,若波段对应元素靠近光谱序列边界,无法获取完整的相邻波段范围,则通过二次线性插值对光谱序列进行填充,并对填充后的光谱序列计算斜率值,其中原光谱序列中得到的斜率值不变,新光谱序列中第一个元素的斜率值设置为第二个元素的斜率值; $k_{j,i}$ 表示第 j 个位置的光谱序列中第 i 个波段对应元素的斜率值, $k_{j,i,b}^-$ 表示第 j 个位置的光谱序列中第 i 个波段前第 b 个波段对应元素的斜率值, $k_{j,i,b}^+$ 表示第 j 个位置的光谱序列中第 i 个波段后第 b 个波段对应元素的斜率值, $| \cdot |$ 表示求绝对值;

首先通过趋势序列中方差贡献度进行第一信息分布程度量化,去除该波段对应元素前后的方差比值,比值与1的差异越大,去除前后方差变化越大,方差贡献度越大,第一信息分布程度越大;同时对相邻波段范围内的斜率差异进行分析,斜率差异越大,该波段对于光谱序列的趋势变化贡献越大,第一信息分布程度越大;通过方差贡献度反映整体趋势分布,通过局部斜率差异反映局部趋势分布;按照上述方法获取每个位置在每个波段的第一

信息分布程度。

[0033] 进一步的,以第*i*个波段为例,获取每个第一局部范围在第*i*个波段中对应的区域,记为每个第一局部范围在第*i*个波段的局部区域,对每个局部区域通过形状上下文算法获取特征向量,其中特征向量维度数设置为同一维度数,形状上下文算法为公知技术,本实施例不再赘述;对第*i*个波段的若干局部区域进行DBSCAN聚类,聚类距离采用局部区域的特征向量之间的DTW距离,得到第*i*个波段的若干类别,则第*i*个波段的第二信息分布程度 λ_i 的计算方法为:

$$\lambda_i = \frac{Q - q_i}{Q} \times \frac{1}{q_i} \times \sum_{f=1}^{q_i} \left(\frac{g_{i,f}}{G - g_{i,f}} \times (A_{i,f} - \bar{A}_i)^2 \right)$$

其中, Q 表示第一局部范围的数量, q_i 表示第*i*个波段的类别数量, G 表示所有第一局部范围包括的位置数量, $g_{i,f}$ 表示第*i*个波段的第*f*个类别包括的所有局部区域对应的位置数量, $A_{i,f}$ 表示第*i*个波段的第*f*个类别包括的所有局部区域中的光谱值均值, $\bar{A}_i$ 表示第*i*个波段所有局部区域中的光谱值均值;

首先通过类别数量来反映波段包含的信息程度,类别数量越少,越能反映单种细胞结构信息,波段的第二信息分布程度越大;再通过波段包含的类别中单个类别与其他类别的差异性来反映第二信息分布程度,类别中位置数量越多,且类别中光谱值均值与整体光谱值均值差异越大,类别差异越大,越能反映某种细胞结构信息,第二信息分布程度越大;按照上述方法获取每个波段的第二信息分布程度,将每个波段的第二信息分布程度作为每个位置在对应波段的第二信息分布程度。

[0034] 进一步的,对同一位置的所有第一信息分布程度进行线性归一化,得到的结果作为该位置在每个波段的第一信息程度,获取每个位置在每个波段的第一信息程度;将所有位置在所有波段的第二信息分布程度进行线性归一化,得到的结果作为每个位置在每个波段的第二信息程度;获取每个位置的光谱序列,以第*j*个位置与第*j+1*个位置为例,获取两个位置的光谱序列的皮尔逊相关系数,获取两个位置的光谱序列中相同次序的元素的差值绝对值,将两个位置的光谱序列中第一个元素的差值绝对值与所有元素的差值绝对值的最大值的比值,与皮尔逊相关系数的乘积,作为两个位置的光谱相似度;将第*j*个位置与其他所有位置的光谱相似度的均值,作为第*j*个位置的整体参考权重,将1减去整体参考权重的差值作为第*j*个位置的局部参考权重,将局部参考权重作为第一信息程度的权重,整体参考权重作为第二信息程度的权重,对同一波段下的第一信息程度及第二信息程度进行加权求和得到第*j*个位置在该波段的综合信息程度。

[0035] 进一步的,按照上述方法获取每个位置在每个波段的综合信息程度,将所有位置在同一波段的综合信息程度的均值,作为该波段的修正系数,则得到了每个波段的修正系数。

[0036] 至此,获取到了每个波段的修正系数。

[0037] 步骤S004、根据修正系数对PCA降维中每个波段的特征值进行修正,得到PCA降维后的若干特征波段,通过特征波段构建分类模型,完成细胞异常样本的智能化检测。

[0038] 需要说明的是,现有技术中通过PCA对高光谱数据进行降维分析的过程中,是根据每个波段的特征值进行分析得到若干特征波段,而通常前几个波段的特征值较大,特征波段也多为前几个波段,这会导致降维后的高光谱数据出现信息丢失的问题;而获取到修正系数后,通过修正系数对特征值进行调整,从而保证得到的特征波段能够最大程度反映高光谱数据所包含的信息,进而提高细胞样本的异常检测结果的准确性。

[0039] 具体的,在对高光谱数据进行PCA降维过程中,获取到协方差矩阵中每个波段的特征值后,将每个波段的修正系数与特征值的乘积作为每个波段的修正后的特征值;根据修正后的特征值完成对高光谱数据的PCA降维,得到若干特征波段,其中PCA降维为公知技术,本实施例不再赘述。

[0040] 进一步的,获取每个细胞切片对应的高光谱数据的特征波段,则可以获取到所有细胞样本的特征波段后,将所有细胞样本及其特征波段分为两类,一类作为训练集,一类作为预测集,训练集与预测集的细胞样本数量比本实施例设置为3:1,通过细胞学专家或技术人员对所有细胞样本的特征波段进行人工标注,标注每个细胞样本对应的细胞类型,例如正常细胞及各类异常细胞等,本实施例通过支持向量机及神经网络来分别构建分类模型,采用交叉熵函数作为损失函数;通过训练集对分类模型完成训练,对两个分类模型分别输入预测集,计算预测误差均方根,将预测误差均方根最小的分类模型作为最终采用的分类模型,根据该分类模型可以实现对细胞样本的智能化异常检测。

[0041] 至此,通过对细胞样本利用高光谱成像技术,实现了细胞样本的智能化异常检测。

[0042] 以上所述仅为本发明的较佳实施例而已,并不用以限制本发明,凡在本发明的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

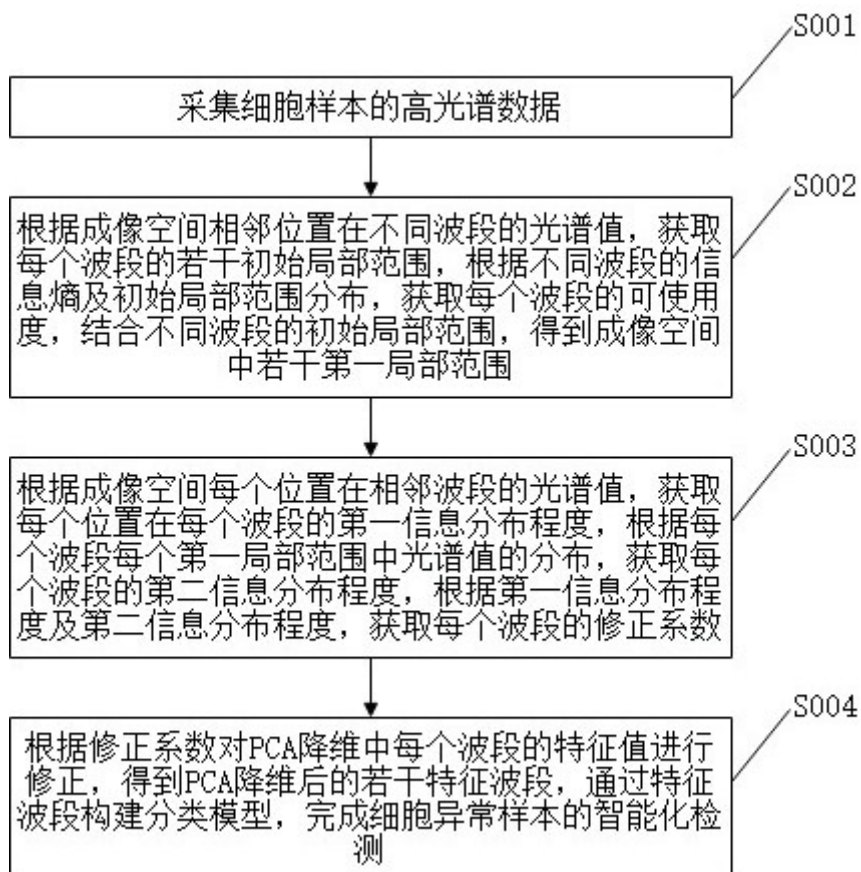


图 1