**实验报告**

1. **问题描述**

通过降维对已知的SARS病毒siRNA各碱基位的碱基数据进行预处理，提取特征，以达到由此剔除数据中的噪声并提升机器学习方法的性能。

1. **数据**

通过搜集获取共163个现有的SARS病毒siRNA的各碱基位碱基及干扰效率。

1. **模型描述**

采用降维方法中的主成分分析方法（PCA）。

算法：

1. 对原始数据进行标准化处理

假设原始观测数据矩阵为



可按如下方法进行标准化：

 ，

其中，





1. 计算相关系数矩阵

相关系数矩阵：



其中：



1. 计算相关系数矩阵R的特征值及特征向量

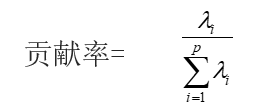
特征值：



特征向量：



1. 选择重要的主成分



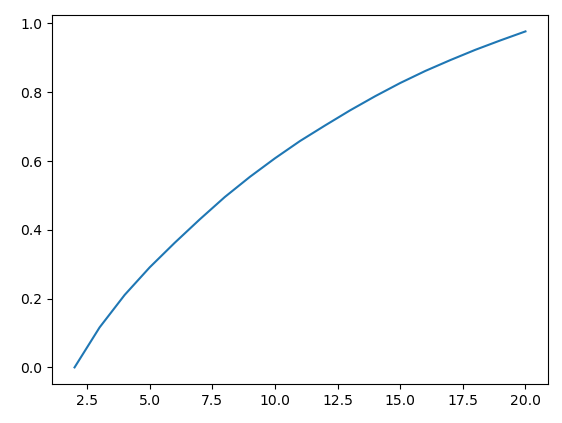
* 贡献率越大，说明该主成分所包含的原始变量的信息越多。主成分个数k的选择，主要根据主成分的累计贡献率决定，一般要求累计贡献率达到85%以上，这样才能包括原始变量的绝大多数信息。
* ⑤ 计算主成分得分
* 根据标准化的原始数据，按照各个样品，分别代入主成分分析表达式，即得主成分得分。具体形式如下图所示：
* 
* 其中，
* 

**四、实验结果**

实例1：将siRNA的各碱基位上的碱基按以下规则进行编码，

A,U,C,G分别对应1,2,3,4。将编码后的siRNA数据进行降维处理。

此过程中累计贡献率的变化趋势图如下：



**分别设置维度从1到19，统计对应的累计贡献率如下：**

|  |  |
| --- | --- |
| **降维后维数** | **累计贡献率** |
| **1** | **0.0000** |
| **2** | **0.1163** |
| **3** | **0.2107** |
| **4** | **0.2909** |
| **5** | **0.3624** |
| **6** | **0.4305** |
| **7** | **0.4955** |
| **8** | **0.5540** |
| **9** | **0.6080** |
| **10** | **0.6582** |
| **11** | **0.7034** |
| **12** | **0.7474** |
| **13** | **0.7885** |
| **14** | **0.8269** |
| **15** | **0.8619** |
| **16** | **0.8935** |
| **17** | **0.9235** |
| **18** | **0.9508** |
| **19** | **0.9770** |

**可以看到当降维至15个特征时，累计贡献率首次超过85%。因此，我们选取15作为降维维数。此时前15个主成分的贡献率分别为**

|  |  |
| --- | --- |
| **主成分序号** | **对应贡献率** |
| **1** | **0.1163** |
| **2** | **0.0944** |
| **3** | **0.0802** |
| **4** | **0.0715** |
| **5** | **0.0681** |
| **6** | **0.0650** |
| **7** | **0.0585** |
| **8** | **0.0539** |
| **9** | **0.0503** |
| **10** | **0.0452** |
| **11** | **0.0440** |
| **12** | **0.0411** |
| **13** | **0.0384** |
| **14** | **0.0350** |
| **15** | **0.0316** |

**降维后的数据一览**

**[[ 0.82365416 0.44524947 -1.59218897 ... 0.41348857 0.09536857**

**-0.41362305]**

**[ 0.20176385 0.39386032 0.4087873 ... -0.15335167 0.68403615**

**-1.1678849 ]**

**[-1.46233617 2.12736036 -1.04702745 ... -1.22963349 -1.04956133**

**-0.02937413]**

**...**

**[-0.71894248 -0.94548476 -0.75130871 ... 0.08346268 -1.3808468**

**0.55449653]**

**[-0.15805021 -0.87117995 -0.61978036 ... 1.36212848 0.27474476**

**0.97372189]**

**[-0.73246544 -0.98623003 0.16211647 ... 2.30569116 -1.1508089**

**0.63518086]]，具体见pca1.txt文件。**

**实例二：将siRNA的各碱基位上的碱基按以下规则进行编码，**

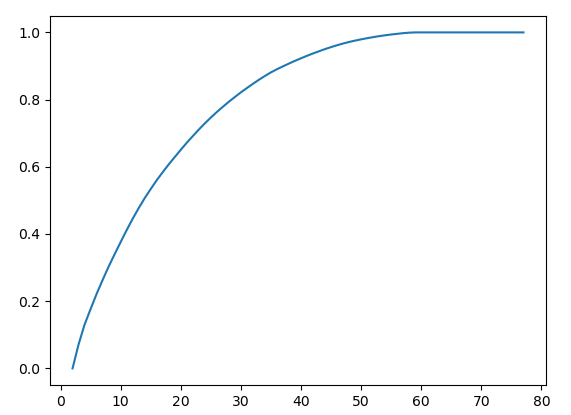
**A碱基的表示：用19个0-1数字表示各碱基位，其中对siRNA中A出现的位数置1，其余位数置0。**

**其余U,C,G表示方法同上。共19×4=76个0-1数字表示。**

**将编码后的siRNA数据进行降维处理。**

**分别设置维度从1到76，统计对应的累计贡献率如下：**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **降维后维数** | **累计贡献率** | **降维后维数** | **累计贡献率** | **降维后维数** | **累计贡献率** | **降维后维数** | | **累计贡献率** |
| **1** | **0.0000** | **20** | **0.6716** | **39** | **0.9229** | | **58** | **1.0000** |
| **2** | **0.0712** | **21** | **0.6913** | **40** | **0.9304** | | **59** | **1.0000** |
| **3** | **0.1303** | **22** | **0.7107** | **41** | **0.9374** | | **60** | **1.0000** |
| **4** | **0.1766** | **23** | **0.7291** | **42** | **0.9439** | | **61** | **1.0000** |
| **5** | **0.2216** | **24** | **0.7466** | **43** | **0.9503** | | **62** | **1.0000** |
| **6** | **0.2630** | **25** | **0.7630** | **44** | **0.9563** | | **63** | **1.0000** |
| **7** | **0.3028** | **26** | **0.7786** | **45** | **0.9618** | | **64** | **1.0000** |
| **8** | **0.3399** | **27** | **0.7936** | **46** | **0.9669** | | **65** | **1.0000** |
| **9** | **0.3759** | **28** | **0.8077** | **47** | **0.9716** | | **66** | **1.0000** |
| **10** | **0.4115** | **29** | **0.8216** | **48** | **0.9756** | | **67** | **1.0000** |
| **11** | **0.4453** | **30** | **0.8344** | **49** | **0.9793** | | **68** | **1.0000** |
| **12** | **0.4769** | **31** | **0.8469** | **50** | **0.9827** | | **69** | **1.0000** |
| **13** | **0.5067** | **32** | **0.8591** | **51** | **0.9859** | | **70** | **1.0000** |
| **14** | **0.5338** | **33** | **0.8703** | **52** | **0.9888** | | **71** | **1.0000** |
| **15** | **0.5600** | **34** | **0.8810** | **53** | **0.9914** | | **72** | **1.0000** |
| **16** | **0.5841** | **35** | **0.8903** | **54** | **0.9937** | | **73** | **1.0000** |
| **17** | **0.6072** | **36** | **0.8990** | **55** | **0.9958** | | **74** | **1.0000** |
| **18** | **0.6290** | **37** | **0.9073** | **56** | **0.9978** | | **75** | **1.0000** |
| **19** | **0.6506** | **38** | **0.9154** | **57** | **0.9992** | | **76** | **1.0000** |



**可以看到当降维至32个特征时，累计贡献率首次超过85%。因此，我们选取32作为降维维数。**

**此时前32个主成分的贡献率分别为**

|  |  |  |  |
| --- | --- | --- | --- |
| **主成分序号** | **对应贡献率** | **主成分序号** | **对应贡献率** |
| **1** | **0.0712** | **17** | **0.0218** |
| **2** | **0.0591** | **18** | **0.0216** |
| **3** | **0.0462** | **19** | **0.0210** |
| **4** | **0.0450** | **20** | **0.0197** |
| **5** | **0.0414** | **21** | **0.0194** |
| **6** | **0.0398** | **22** | **0.0184** |
| **7** | **0.0371** | **23** | **0.0175** |
| **8** | **0.0360** | **24** | **0.0165** |
| **9** | **0.0356** | **25** | **0.0156** |
| **10** | **0.0338** | **26** | **0.0150** |
| **11** | **0.0316** | **27** | **0.0141** |
| **12** | **0.0297** | **28** | **0.0138** |
| **13** | **0.0272** | **29** | **0.0129** |
| **14** | **0.0262** | **30** | **0.0125** |
| **15** | **0.0241** | **31** | **0.0121** |
| **16** | **0.0231** | **32** | **0.0113** |

**降维后的数据为[[ 0.16278315 -0.44833842 1.1948647 ... -0.10378984 0.19239059**

**0.66334766]**

**[ 0.4754666 -0.6444048 0.23277724 ... 0.26555434 0.11255354**

**-0.03045301]**

**[-0.83695745 -0.46818015 1.0282485 ... -0.38420537 1.1056609**

**-0.4781293 ]**

**...**

**[-1.9825766 -0.24465185 0.48535264 ... -0.4615752 0.26967672**

**-0.07986056]**

**[ 0.62812316 0.24451917 0.17529452 ... 0.00235551 0.6169548**

**0.17006147]**

**[-0.8170177 0.09001049 -0.7232217 ... 0.5245523 -0.1470552**

**0.11764796]]，具体见pca2.txt文件。**

**总结：降维技术能有效地降低输入特征的维数，从而减少开销。尤其在信息存在冗余表达的情况下（如实例2），能有效去除数据中的噪音，提升机器学习系统的性能。**