基于互信息的最大相关最小冗余特征选择

一、特征选择定义：

在机器学习的实际应用中，特征数量往往较多，其中可能存在不相关的特征，特征之间也可能存在相互依赖，容易导致如下的后果：

- 特征个数越多，分析特征、训练模型所需的时间就越长。
- 特征个数越多，容易引起“维度灾难”，模型也会越复杂，其泛化能力会下降。

特征选择能剔除不相关(irrelevant)或冗余(redundant)的特征，从而达到减少特征个数，提高模型精确度，减少运行时间的目的。另一方面，选取出真正相关的特征简化了模型，使研究人员易于理解数据产生的过程。

特征选择的定义如下：

给定输入数据$\mathrm{D(N\ samples, M\ features)}$，$\mathrm{X} = \{x_i, \mathrm{i} = 1, \ldots, \mathrm{M}\}$，特征选择的目的就是从观测空间$R^M$中，找到一个最优子空间$R^m$，使其对应的m个特征能够最好的描述类别c。

二、选择特征方法：

2.1 最大相关最小冗余算法（mRMR）

特征选择的一般过程可用图 1 表示。首先从特征全集中产生出一个特征子集，然后用评价函数对该特征子集进行评价，评价的结果与停止准则进行比较，若评价结果比停止准则好就停止，否则就继续产生下一组特征子集，继续进行特征选择。选出来的特征子集一般还要验证其有效性。

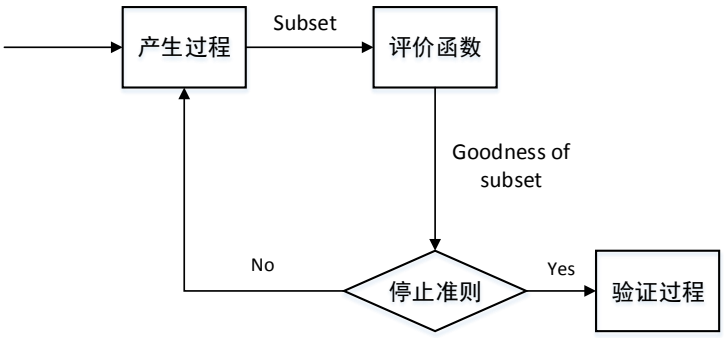

图 1 特征选择过程

综上所述，特征选择过程一般包括产生过程，评价函数，停止准则，验证过程，这 4 个

部分。其中，最重要的就是评价函数部分，它是评价一个特征子集好坏程度的准则。

通常情况下，最优的特征选择往往意味着最小化分类误差。在无监督且不指定分类器的前提下，最小化误差通常需要最大化目标类别c在子空间R^m数据分布的统计依赖性（statistical dependency），这也就是 Maximal dependency 方法。

实现 Max-Dependency 的常用方法是最大相关性（Maximal Relevance）特征选择：选择与目标类c具有最大相关性的特征，利用互信息来计算相关性的大小：

给定两个随机变量x, y以及概率密度p(x), p(y), p(x, y)，互信息定义为：

$$\mathrm{I(x;y)} = \int\int p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy$$

在 Max-Relevance 中，会选择互信息$\mathrm{I}(x_i;\mathrm{c})$最大的特征，这些特征与目标类c的相关性最大。所以在顺序搜索方法中，往往会选择与目标类别互信息最大的m个特征。但实际情况中，单个的最优特征组合起来并不会获得最好的分类性能，换句话说，m个最好的特征并不是最好的m个特征。所以一些研究者[1]试图找到一种方法，直接或者间接地去除特征间的冗余，即选择最小冗余的特征。因此，一种直观的思路就是，可以利用一个启发式的搜索方法，结合最大相关和最小冗余的特性，来选择特征，这样得到的特征性能更优[5]——Minimal-redundancy-maximal-relevant(mRMR)。接下来主要从原理上介绍 mRMR：

最大相关原则指的是找到一个包含$|S|$个特征的特征集S，使得S中所有特征与类别之间的相关性最大化，最大化的条件为：

$$\max \mathrm{D(S,c)} \quad D = \frac{1}{|S|} \sum_{w_i \in S} I(w_i; c)$$

最小冗余原则指的是找到一个包含|S|个特征的特征集 S，使得 S 中的每个特征之间是互相最大不相似，即最小相似的，最小化的条件是：

$$\min \mathrm{R(s)} \quad R = \frac{1}{|S|^2} \sum_{w_i, w_j \in S} I(w_i; c)$$

最大相关最小冗余原则结合了上面两个原则，定义了运算Φ来同时优化D和R：

$$\max \Phi(D,R), \quad \Phi = D - R$$

事实上，使用上述的评价方式，利用增量搜索方式即可得到近似最优的特征。假设我们已经有了$S_{m-1}(m-1$ 个特征的集合)，目的是从集合$\{\mathrm{X} - S_{\mathrm{m-1}}\}$中找到第$m$个特征。相应的增量搜索算法最大化下式：

$$\max_{x_j \in X - S_{m-1}} [\mathrm{I}(x_j; \mathrm{c}) - \frac{1}{m} \sum_{x_i \in S_{m-1}} I(x_j;\ x_i)]$$

2.2 特征选择算法实现：

为了设计一个有效的算法，能够找到完备且紧凑的特征子集，采用了两步特征搜索选择方法。在第一个阶段，用 mRMR 增量选择算法先选择一个候选的特征集；第二个阶段用一个更复杂的方案，从候选特征集中选出紧凑的特征集，作为最终特征选择的结果。

（一） 选择候选特征集：

得到候选特征集的步骤如下：

1. 用 mRMR 增量选择法，从输入X中选择n个特征，得到一系列的特征集：$S_1 \in S_2 \in S_3 \in \cdots \in S_{n-1} \in S_n$。
2. 比较n个特征集，选择一个k的范围Ω，使得S_k对应的误差e_k相对一致，且较小。
3. 在Ω中，选择最小的分类误差$e^* = min\ e_k$，S_k即为最终得到的候选特征集。

综上所述，得到的候选特征集大小$n^* = \underset{k}{\operatorname{argmin}}\ e_k$。

（二） 选择紧凑（Compact）的特征集：

许多复杂的算法都可以用来从一个候选集S_{n^*}中搜索得到紧凑的特征集。在这个特征选择的算法中，使用封装器(Wrapper)来实现这一过程。

封装器实质上是一个分类器（如朴素贝叶斯分类器），它用选取的特征子集对样本集进行分类，分类的精度作为衡量特征子集好坏的标准。在第一个阶段中，我们已经使用 mRMR 算法找到了一个较小的候选特征集，所以在第二步中，大大的降低了分装器的运算复杂度。本文中，考虑封装器的两种选择方案——前向选择和后向选择：

1. 前向选择算法：特征子集 S 从空集开始，每次选择一个特征x_i加入特征子集S，使得特征函数J(S)最优。简单说就是，每次都选择一个使得评价函数的取值达到最优的特征加入，其实就是一种简单的贪心算法
2. 后向选择算法：从特征全集S_{n^*}开始，每次从特征集 S 中剔除一个特征x_i，使得剔除特征x_i后评价函数值达到最优。

2.3 主成分分析（PCA）算法

主成分分析是一种掌握事物主要矛盾的统计分析方法，它可以从多元事物中解析出主要影响因素，揭示事物的本质，简化复杂的问题。计算主成分的目的是将高维数据投影到较低维空间。PCA 是一种用原有变量的线性组合来表示事物主要方面的分析方法。

PCA 主要用于数据降维，对于一系列例子的特征组成的多维向量，多维向量里的某些元素本身没有区分性，比如某个元素在所有例子中都为 1，或者与 1 差距不大，那么这个元素本身就没有区分性，用它做特征来区分，贡献会非常小。所以我们的目的是找那些变化大的元素，即方差大的那些维，而去除掉那些变化不大的维，从而使特征留下的都是主要的成分，同时使得计算量也大大降低。

在 MATLAB 中有 PCA 的函数 princomp(X)，对 n 行 n 列的数据集 X 做完主成分分析以后会返回主成分系数，X 的每行表示一个样本的观测值，每一列表示特征变量。

返回的第一个参数 COEFF 是一个 p 行 p 列的矩阵，每一列包含一个主成分函数，列是按主成分变量递减顺序排列，也就是说，COEFF 是 X 矩阵所对应的协方差矩阵 V 的所有特征向量组成的矩阵，即变换矩阵或投影矩阵，COEFF 每列对应一个特征值的特征向量，列的排列顺序是按特征值的大小递减排序的。

返回的 SCORE 是对主成分的打分，也就是说原来 X 矩阵在主成分空间的表示。SCORE 每行对应样本观测值，每列对应一个主成分（变量），它的行和列的数目和 X 的行列数目相同。

返回的 latent 是一个向量，它是 X 所对应的协方差矩阵的特征值向量。

三、实验设计：

我们首先编写了 mRMR 算法的 C++代码，从特征集中选择候选子集，并保存对应特征的

index。然后，我们用选到的特征构建 SVM 分类器，统计分类准确率与所选特征数目的关系，确定特征选择算法的有效性。

同时我们也进行了 PCA 和 SVM 的实验组合，从一维开始一直到主成分含量为 100%的最小维度，进行分类实验。

这个实验中我们使用的数据集为 NCI data 和 Lung Cancer data（见程序附件）

四、实验结果：

4.1 lung_s3 数据集实验结果：

Lung 数据集是肺癌数据集，其中样本数目：73，特征数目：325，类别数目：7

实验中，我们首先使用 mRMR 算法，提取不同数量的特征（0,1,5,10...325），然后将提取到的特征以及对应的类别输入到 SVM 分类器中，用 SVM 进行分类，统计交叉验证的错误率与特征数量的关系，实验结果见下图：

左图为从 325 维特征中，按照 mRAR 算法，用增量搜索法提取出了不同维数的特征后，用 SVM 分类器测得的交叉验证分类误差，由于特征数量过多，右图画出了特征数量在 0-50 范围内的 FeatureNumber-ClassificationError 曲线。

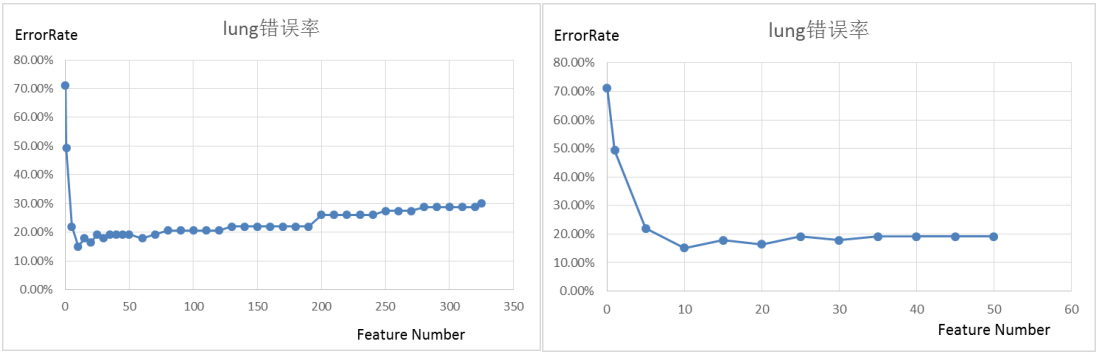

图 2 左：lung 提取不同数量的特征后，错误率变化　右：左图特征数目较小时的结果

从图中不难发现，当特征数量为 10，也就是从 325 维特征中，提取与目标类别 c 互信息较大的十维，可以获的较好的分类性能。之后随着特征数目的增大，有可能出现过拟合等现象，反而使分类误差升高了。可见，这种特征选择的方法能够有效的选择较好的特征，在数据量很大的情况下，可以通过这种方法筛选出少量有效的特征，从而减小复杂度，提高分类任务的准确度。

PCA 实验中，得到下图:

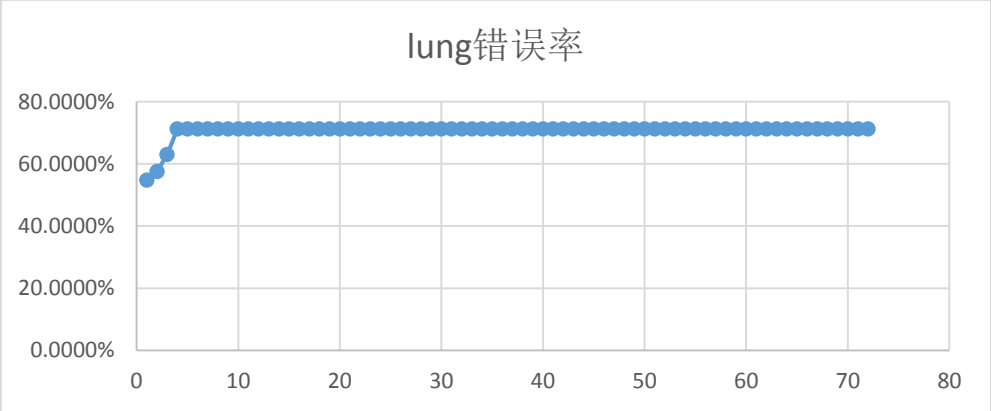

图 3：lung 经过 PCA 后提取不同维度的特征后，错误率变化

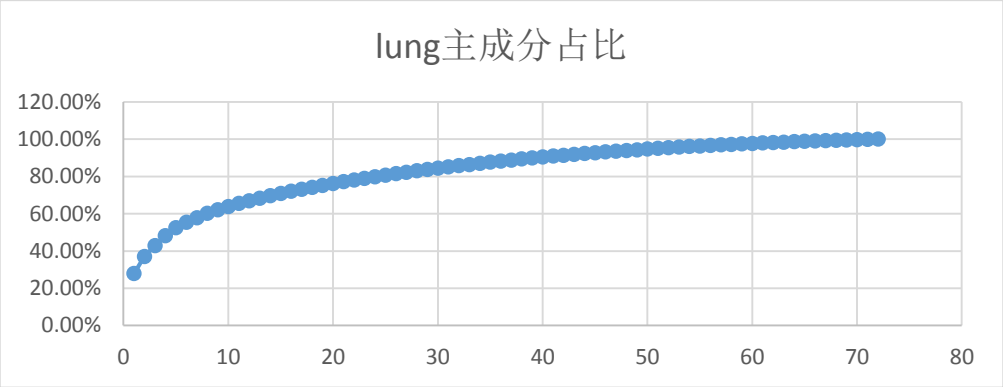

图 4：lung 在 PCA 变化后主成分占比和特征维度的关系

4.2 nci9_s3 数据集实验结果：

nci9 数据集是基因数据集，其中样本数目：60，特征数目：9712，类别数目：9
实验过程与上述的 lung 数据集相同，在此不再赘述。实验结果如下图所示：

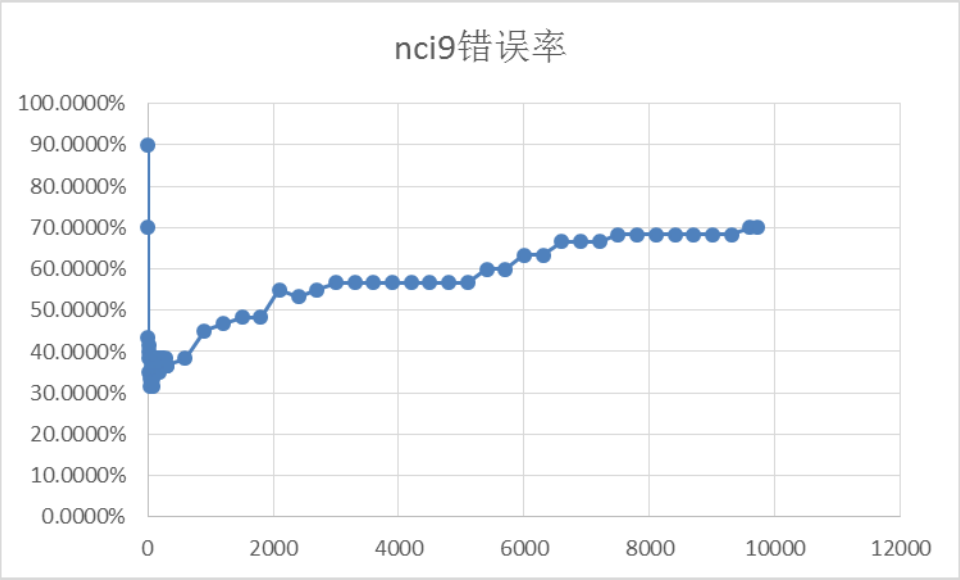

图 5 nci9 提取不同数量的特征后，错误率变化

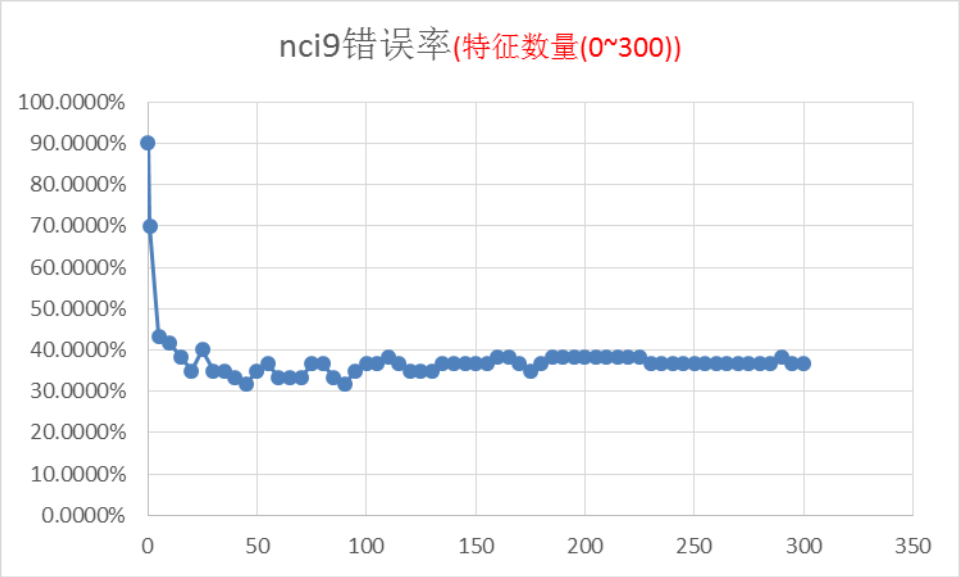

图 6 nci9 提取特征数目 0~300 时的错误率变化

与 lung_s3 不同的是，nci9_s3 的特征数目很多，在开始实验时，我们预计大概需要提取的特征数目应该≥300 个，才可以达到比较好的分类效果。而真正实验时，我们发现，虽然 nci9 数据集有 9712 个特征，但我们只需要提取其中的 50 个左右的特征便可以较好的完成分类任务，之后随着特征数目的增加，分类器的准确率反而会下降。分析觉得，可能是由于样本的数目较小，特征数目太多，过多的特征反而会更容易产生过拟合现象，导致性能下降。

通过上面两个实验，不难发现，一个好的特征提取算法，不仅可以有效的降低计算很多特征所需要的资源（存储空间以及运算复杂度），还可以有效的提升分类器的性能，避免过拟合现象。

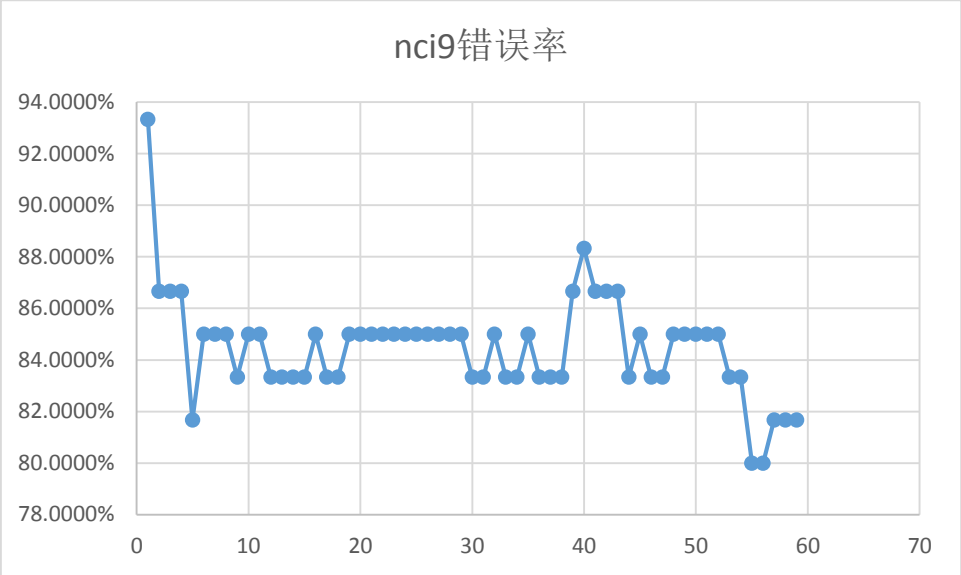

图 7：nci9 经过 PCA 后提取不同维度的特征后，错误率变化

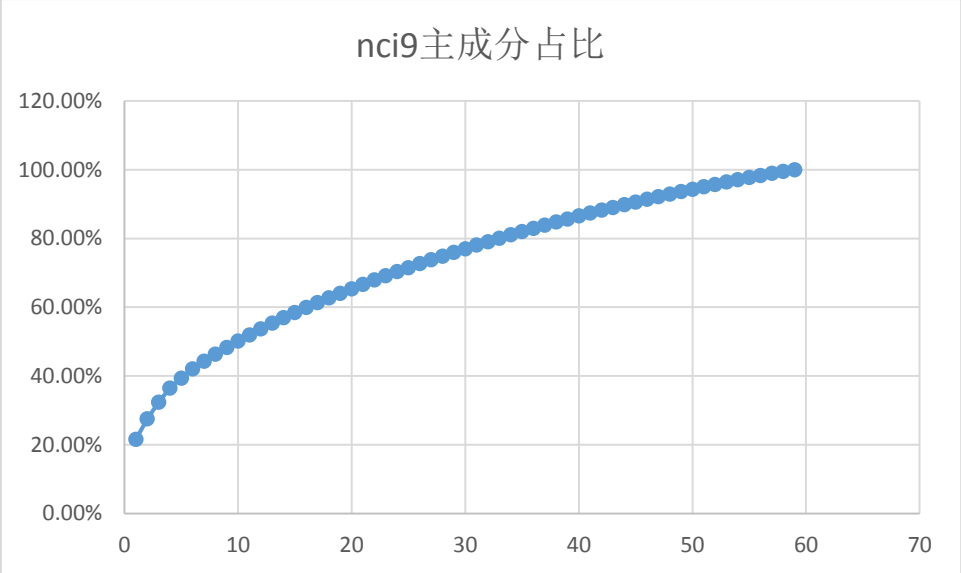

图 8：nci9 经过 PCA 后提取不同维度的特征后，错误率变化

由这两个数据集在 PCA 后的分类结果来看，呈现出特征维度越高，错误率越低的趋势，但是这样的趋势并没有什么意义，因为整体错误率太高，这可能跟数据集样本数目太少有关系。总的来说，PCA 后的分类效果没有 mRMR 好，这和主成分分析特征选择算法中相关矩阵只能衡量变量之间线性关系的局限性有一定关系，考虑了特征之间互信息的 mRMR 法在特征维数降低的同时也保证了信息的最大可能保留，分类效果更好。

针对 PCA 在主成分含量为 100%时的分类准确率也很低的情况，是因为全部的特征含有很多的冗余信息，这在一定程度上会降低分类准确率。

五、总结：

特征选择在模式识别当中占有重要的地位。从分类的角度看，模式识别是把具体事物归到某一类的过程也就是先用一定数量的样本，根据它们之间的相似性进行分类器设计，而后

用所设计的分类器对待识别的样本进行分类决策。分类过程既可以在原始数据空间中进行，也可以对原始数据进行变换，将数据映像到最能反映分类本质的特征空间中进行。相比而言，后者使得决策机器的设计更为容易，它通过更为稳定的特征表示，提高决策机器的性能，删去多余或不相关的信息，并且更加容易发现研究对象之间的固有联系。因而，特征是决定样本之间的相似性和分类器设计的关键[6]。在分类目的决定之后，如何找到合适的特征是认知与识别的核心问题[7][8]。但是，由于在很多实际问题当中，常常不容易找到那些最重要的特征，或者受条件限制不能对它们进行测量，这使得特征选择和提取的任务复杂化，从而成为构造模式识别系统、提高决策精度的最困难的任务之一。

本文介绍了一种两步特征选择算法：第一步基于滤波评价策略的特征选择算法——mRMR，利用特征与目标类别的互信息大小，通过增量搜索法得到候选特征集；第二步基于嵌入式评价策略，用 SVM 分类器，观察所选特征的分类性能，得到特征选择的结果。最终也取得了较好的性能。本文还将 mRMR 和 PCA 进行了对比，实验结果显示 mRMR 可以取得更好的结果。

六、参考文献：

[1] T.M. Cover, “The Best Two Independent Measurements Are Not the Two Best”, IEEE Trans. Systems, Man, and Cybernetics, vol.4, pp. 116-117, 1974.

[2] T. Cover and J. Thomas, Elements of Informayion Theory. New York: Wiley, 1991.

[3] A.K. Jain, R.P.W. Duin, and J. Mao, “Statistical Pattern Recognition: A Review,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 2, pp.153-158, Feb.1997.

[4] A. Webb, Statistical Pattern Recognition. Arnold, 1999.

[5] Hanchuan Peng, Fuhui Long, and Chris Ding, “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp.1226-1238, AUGUST 2005.

[6] De Wang, Feiping Nie, and Heng Huang, “Feature Selection via Global Redundancy Minimization,” IEEE Trans on knowledge and data engineering, vol.27, no. 10, Oct 2015

[7] D. Wang, F. Nie, and H. Huang, “Unsupervised feature selectionvia unified trace ratio formulation and k-means clustering(track).” in Proc. Eur. Conf. Mach. Learning Knowl. Discovery Databases,2014, pp. 306–321.

[8] D. Wang, L. Yang, Z. Fu, and J. Xia, “Prediction of thermophilicprotein with pseudo amino acid composition: An approach from combined feature selection and reduction,” Protein Peptide Lett., vol. 18, no. 7, pp. 684–689, 2011.