基于降维和机器学习的地下金属目标的稳健高效分类（李彤师兄小论文）

- Abstract
  - 地下目标的分类由于噪声的存在以及数据的高维原因，无法达到精确且高效的分类。
  - 本文内容：研究电磁感应探测领域下，高效且稳健的基于数据的对地下目标的材质、形状进行分类的方法。
  - 研究了基于11种降维方式与3种分类算法组合的33种分类策略。分类算法包含：ANN，L-SVM（线性支持向量机），GNB（高斯朴素贝叶斯）。降维方式：LASSO（L1正则化），GA-SVM（遗传支持向量机），PCC（皮尔逊系数），MI（互信息），mRMRP（最大相关最小冗余皮尔逊相关），mRMRMI（最大相关最小冗余互信息），SF（统计特征），PCA，KPCA（核主成分分析），LLE（局部线性嵌入），SDAE（堆叠去噪自动编码器）
  - 评估分类策略的：精确度，降维后的特征数量，特征类型的重要性和时间消耗。
  - 结果：基于材质的分类中，基于KPCA的ANN算法的精确度最高。LLE在基于材质和形状的分类中，提高机器学习的鲁棒性，降低了ANN分类的时间消耗
  - 研究的意义：对比研究为地下金属目标分类提供了一种稳健有效的基于数据的策略，降维方法和机器学习模型的交叉组合策略为寻找最优机器学习模型提供了一种途径。
- Introduction
  - 大的背景，区分地下目标的材质以及形状是一个需要解决的问题。
  - 现存的地下目标的分类主要分为基于模型的分类与基于数据的分类方法。基于模型的方法首先需要通过反演得到金属目标的属性然后进行分类，因此分类的准确率取决于正向模型的合理性和反演算法的稳定性。基于数据的方法不涉及物理理论，直接利用获取到的响应数据进行分类，比较容易应用。
  - 基于数据的方法是建立从观测数据提取的特征到金属目标属性的映射关系，其中有两个重要的影响因素：映射关系和输入的特征。对于映射关系，目前有很多的分类的机器学习模型，以及模型匹配的方法（不太懂这种方法的应用）。关于模型输入的特征，一个比较常见的问题就是处理输入数据的高维问题。
  - 时域电磁探测产生的数据可能因为时间信道、接收器元件的数量和空间采样密度导致特征数量比较大。但是并不是所有的数据对分类都是有用的，这造成了特征的冗余，另外数据量的增大也会需要更大的计算能力，造成更大的计算开销。另外大规模的特征造成的维数灾难，会增加数据和结果之间关系的复杂性。
  - 介绍降维：机器学习中，特征中主要解释映射关系的其一个相对较小的子集，因此需要在观测数据中提取和选择有区别的特征，实现特征降维。
    - 降维分为两种：特征提取、特征选择。特征选择是选择一部分重要的特征，特征提取是从原始维度的空间提取新的特征集。

    - 引用一些在地下目标探测过程中采样降维技术的文献，提到建立合适的特征集可以提高分类的性能。
    - 在基于数据的地下目标分类中，降维是重要的数据处理方法，这直接影响到分类的效率与准确率。
    - 为了进一步提高基于数据的方法的效率和鲁棒性，有必要研究不同降维方法对地下金属目标分类的影响。
- 本文研究了常见的分类算法以及不同的降维方法在基于数据的地下目标探测中的性能表现。
- 建立仿真平台加以验证，探测的内容：目标的形状和材质，基于TDEM方法采用近似椭球体模型作为正演模型产生数据，然后对6种特征选择方法以及5种特征提取方法在3种分类模型上的表现加以验证。
    - 特选择方法：LASSO（L1正则化），GA-SVM（遗传支持向量机），PCC（皮尔逊系数），MI（互信息），mRMRP（最大相关最小冗余皮尔逊相关），mRMRMI（最大相关最小冗余互信息）
    - 特征提取方法：SF（统计特征），PCA，KPCA（核主成分分析），LLE（局部线性嵌入），SDAE（堆叠去噪自动编码器）
    - 分类模型：ANN，L-SVM（线性支持向量机），GNB（高斯朴素贝叶斯）
    - 评估指标：在采用留出法的数据集上比较准确率与混淆矩阵。
    - 对通过不同降维方法获得的特征，进行相关性分析和冗余分析。
    - 对降维方法对分类器时间消耗的影响进行分析，以及对所选特征在分类中的重要性进行分析。
- 贡献：
    - 研究了33种分类策略，此外，利用归一化特征类型重要性(NFTI)系数来分析所选特征类型的重要性，并提出互信息来评估输入的冗余性和相关性。
    - 分析了在不同的信噪比下，不同降维方法下分类器的平均精度。局部线性嵌入(LLE)提高了机器学习分类器的鲁棒性和人工神经网络的分类效率（为什么重点提这点）。比较结果表明，分类模型的性能对噪声水平敏感，适当的降维方法(如LLE和KPCA)可以显著增强模型对噪声的鲁棒性。
    - 对比研究证明，在基于材料的分类和基于形状的分类中，人工神经网络分类器辅助核主成分分析(KPCA)特征提取方法具有最佳性能。本研究提出的分类策略提高了地下金属目标探测的预测精度。
- 论文的组织：
    - 第二部分介绍TEM方法以及通用物理模型与近似椭球体正演模型
    - 第三部分和第四部分介绍6种特征选择方法和5种特征提取方法以及3种分类模型，相关性分析以及冗余分析
    - 第五部分介绍仿真设计与评价标准
    - 第六部分介绍仿真结果
    - 第七部分介绍discussion，第八部分介绍conclusion
- 2.系统模型

- 收集目标区域的观测数据：发射线圈中的瞬态电流产生一个脉冲磁场，金属目标感应产生涡流，涡流产生二次场，二次场随时间衰减，最后在接收线圈接收二次场感应出的电压。采用正交偶极子模型来刻画目标属性和瞬变电磁响应之间的关系。
- 偶极子模型的前提：当金属目标和探测器之间的距离大于目标的大小的时候就可以将二次场等价为由偶极子产生的磁场。
- 在TDEM中，目标的响应和二次场的导数成正比，特征响应矩阵是磁极化率张量的负导数，与目标的属性紧密相关。
- 通用模型的主轴极化率有三个，对于我们研究的轴对称物体，其主轴极化率可以进一步化简，一个为平行于对称轴的，另外两个相同都为垂直于对称轴的响应。
- 时域中，金属目标的脉冲响应可以看作无数个指数的和，为了方便计算，用由最少数量的参数定义的经验函数来重写脉冲响应，该函数复制了脉冲响应的特点。该函数有四个拟合参数。球形物体的特征响应被描述为时间衰减曲线的三个不同阶段，椭球体的特征响应可以近似为球体的特征响应

- 3.降维
  - 瞬变电磁系统中随着数据量的增大，会存在无关的、冗余的以及带有噪声的数据，因此降维可以节省计算时间以及存储空间，降维去除无关、冗余以及噪声的特征，使数据质量提高，可以提高分类的效率和泛化能力。降维主要分为特征选择与特征提取两种方式。
  - 特征选择：特征选择是指在特定的评价标准和搜索策略下，选择特征集中最重要的部分。
    - 包装器方法：采用遗传算法启发式搜索特征，采用svm对候选解进行验证
    - 嵌入式方法：嵌入式方法在训练学习模型的同时进行特征选择，用损失函数指导搜索，速度要比包装器的方法快，如：LASSO利用L1范数惩罚来正则化线性回归系数，达到稀疏特征选择的目的。
    - 过滤器方法：过滤器根据评估标准评估特征的优劣，基于是否考虑特征之间的依赖性，分为二元和多元方法。（这种方法没接触过，还不是很理解）
  - 特征提取：从原始数据集中提取新的特征，从而改变特征。
    - 统计特征：计算响应信号的统计特征：包括平均值、方差、标准差、最小值、最大值、偏斜度。
    - 主成分分析（未了解）
    - 核主成分分析（未了解）
    - 局部线性嵌入（未了解）
    - 堆叠去噪自动编码器（未了解）
- 4.分类模型
  - 分为生成式和判别式
    - 生成式：GNB（高斯朴素贝叶斯）GNB假设特征服从高斯分布。
    - 判别式：
      - 支持向量机：采用线性核函数

        - 神经网络：隐藏层采用一层隐藏单元数为50，激活函数采用tanh，采用BFGS优化模型的权重和偏置。
- 仿真验证：
    - A.数据预处理
        - 建立仿真平台，基于椭球模型，应用降维方法验证分类的表现，金属目标的形状分为两种，金属类型分为3种，生成数据时添加噪声模拟真实场景，对数据进行归一化，以提高基于梯度的优化算法的表现。
    - B.分类表现评估的框架
        - 为了防止double-dipping（怎么翻译），通过两种方式进行验证，分别为：留出法，k折交叉验证法
    - C.相关性和冗余性分析
        - 候选输入特征集中存在与分类问题无关的以及冗余的特征，为了量化输入特征集中的这种相关性以及冗余性，利用皮尔逊系数以及互信息来评估特征和标签之间的线性以及非线性关系。相关性矩阵、冗余性矩阵以及二者差值的矩阵用来量化特征集的质量。
    - D.分类误差评估标准
        - 采用精确度和混淆矩阵去量化分类表现
    - E.基于过滤器的特征选择中所选特征数的评估
        - 特征选择的数量是一个关键参数，为了减少由不同分类器引起的性能偏差，在基于过滤的特征选择方法中，利用三个分类器的平均精度来评估所选择的特征数。三种分类器采用三折交叉验证法。
    - F.分类中特征类型重要性的评估
        - 特征选择方法中，响应信号在不同时间阶段的重要性可以通过分析所选特征的分布来量化，本文采用归一化特征类型重要性(NFTI)系数来量化特征选择的特征类型。
        - 具体来说，就是统计与其时间阶段相对应的所选特征的数量。然后，所选的比上该阶段所有的特征数量进行标准化。最后，我们得到了每种特征选择方法中三种特征类型的NFTI系数。
- 结果
    - A.降维方法和分类模型的比较：以留出法评估33中分类策略的表现
        - 基于精确度的评估
            - 无论降维方法是什么，在对材质和形状分类时，ANN分类器表现优于其他两种分类器。
            - 基于形状的分类精确度要高于基于材质的分类的精确度
            - 11种降维方法中，特征提取的方法LLE在GNB分类器以及L-SVM分类器中对分类精确度有一个较大的提高。
            - 与原始数据相比，大多数的降维方法对分类的性能均有一定提升，具体的：
                - 基于材质的分类中，特征提取方法KPCA+ANN取得了最高的精确度0.99，特征选择方法PCC+ANN和MI+ANN取得了最高的精确度0.98

        - 基于形状的分类中， 特征提取方法KPCA+ANN，特征提取方法LLE+ANN、PCC+ANN 、MI +ANN取得了最高的精确度0.99，而特征提取方法SF、SDAE降低了分类的精确度，并且分类器GNB的精确度较低。
    - 基于混淆矩阵的评估
        - 基于LLE+L-SVM的混淆矩阵，在基于材质和基于形状的分类表现：基于材质的分类中，对铝的分类要比其余两种材质更准确，对镍的分类效果较差。LLE降维之后，对镍的分类准确率提高了而且对钢的分类准确率也提高了。基于形状的分类中，长椭球体的分类更加准确，LLE降维之后，扁椭球体的分类精确率提高了，长椭球体的分类准确率也提高了。
- B.降维方法对特征降噪的影响
    - 研究了降维方法在不同信噪比下的表现，为了降低不同分类器的性能偏差，利用三种分类器的平均表现评估降维方法。
    - 结果表现随着信噪比的增大，精确度逐渐提高，这表明分类性能受噪声的影响。
    - 具体的信噪比下基于形状的分类的精确度高于基于材质的分类精确度，所以基于形状的分类器更加鲁棒。
    - 大多数降维方法不会降低基于材质与基于形状的分类的表现
        - 具体而言，与原始数据集相比，通过特征选择方法获得的每个子集，包括最大似然比、最大似然比、主成分分析、最小似然比、遗传SVM和LASSO，具有同等的平均精度
        - LLE、KPCA和主成分分析等特征提取方法极大地提高了基于材料的分类和基于形状的分类的平均精度。
        - 特别当信噪比大于或等于25dB时，LLE是基于材质分类的最佳降维方法。同时在基于形状的分类中，LLE在不同信噪比下的平均准确率优于其他降维方法
        - 但是，特征提取方法(如SF和SDAE)，在基于材料和基于形状的分类中会降低分类器的平均精度。
- C.确定分类中选择的特征类型
    - 所选的特征的数量在基于过滤器的特征选择方法（mRMRP，MRMI，PCC和MI）的性能中起着重要的作用，通过调整每种方法的参数，以获得一系列选定的子集，所选特征数的范围从5到400，间隔为5，然后每个选择的特征分别由三个分类器通过重复的三重交叉验证来训练。
    - 基于材质的分类中mRMRMI是实现最高平均精度的最快的特征选择方法。当特征数量小于150时，随着特征的增多，精确度逐渐增长，当选择的特征数量大于150时，精确度逐步趋于稳定。
    - 基于形状的分类中，mRMRMI和MI以比大多数特征选择方法更小的选择特征数(特征数:100)实现了稳定的平均精度。
    - 基于过滤的特征选择方法的最优参数是平均精度最高的特征数，瞬变电磁系统中的衰减响应信号可分为三个阶段，即早期阶段、中期阶段和晚期阶段，统计了基于材料的分类和基于形状的分类的NFTI（归一化特征类型重要性系数）结果。

结果表明：基于材质的分类中，后期的时间响应信号在大多数特征选择算法选择的特征子集中占较大的比例，而早期的时间响应信号几乎没有被选择。在基于形状的分类中，大多数特征选择选择算法选择的三个时期的特征所占的比例非常接近。而LASSO算法较其他的特征选择算法具有较低的精确度，它选择了较少的中期的特征响应信号。

- D.降维方法对输入的冗余性和相关性的影响
  - 采用相关性系数、冗余性系数以及二者的差值用来量化不同的数据集，基于材质的分类中，11种降维方法降低了冗余性，增强了相关性，差值要比原始数据的差值大。相比于特征选择方法，特征提取方法在降低数据的冗余性上效果更好。
- E.降维方法对分类模型效率的影响
  - 进一步分析在基于材料的分类和基于形状的分类中达到最高精度的人工神经网络模型，分析ANN采用降维方法后的具体时间消耗。
  - 输入特征的数量对分类模型的效率影响较大，因此先统计了基于材质与基于形状的分类中降维之后的特征数量，然后分析了ANN采用降维方法后的训练时间与预测时间。
  - 结果显示降维方法降低了训练时间和分类时间，而且训练时间远大于分类时间。特征选择方法LASSO 方法在基于材质和形状具有最短的构建时间，特征提取方法SF基于材质分类时具有最短的构建时间，LLE在基于形状的分类时具有最短的构建时间。

- discussion
  - 在瞬变电磁系统中，地下金属目标的特性是基于由时域电磁传感器产生的高维观测数据来估计的。
  - 本文中，降维方法以及不用的分类模型用来研究基于数据的地下金属目标的分类性能。还讨论了其他可控变量，例如所选特征的数量、特征类型，用于指导时域电磁传感器的数据采集参数设置
  - 11种分类方法和3种分类器进行研究，结果表明KPCA特征提取方法结合ANN在对地下目标的材质和形状进行分类时取得了较好的效果。ANN分类器的表现要优于其他两种分类器，这表明ANN是一种有效的分类模型。
  - 11种降维方法中，特征提取方法LLE和大部分的分类器结合具有较高的精确度。统计特征方法在以往的研究中被证明是一种有效的特征提取方法，不过和其他的特征提取方法相比，其精确度不高。
  - 虽然GA-SVM 能够避免局部最优，不过GA方法依赖一些参数，这种特征选择方法不优于过滤式的方法。混淆矩阵的结果表明降维方法能够提高分类器的表现。
  - 为进一步研究噪声对精度的影响，我们比较了不同信噪比下3种分类器采用降维方法的平均精度，结果表明分类模型对噪声敏感。不过可以适当采用降维方法提高分类的鲁棒性。特别是特征提取方法LLE在基于材质的分类和基于形状的分类中，提高了机器学习分类的鲁棒性。
  - 尽管特征选择的方法降低了输入的维度，但是提高的精确度是有限的。
  - 四种基于过滤器的特征选择方法的特征选择数的比较结果表明，与线性标准(即MRMRMRP和PCC)相比，非线性标准(即MRMRMRMI和MI)以较小的特征选择数实

现了稳定的平均精度，这表明特征和标签变量(形状类型和材料类型)之间的关系是非线性的。

- 此外，当所选特征数超过150时，基于材质的分类的平均准确率略有下降，表明特征选择是提高分类性能的有效策略。
- 通过特征类型的分析，发现后期的响应信号在基于材质和形状的分类中，与其他特征相比是比较重要的特征，这些结果为时域电磁传感器的数据采集参数设置提供了重要的指导。
- 提出了一种基于互信息的数据质量量化方案，冗余性和相关性的分析表明原始的输入是高度冗余的，部分特征提取方法在降低输入的冗余性提高数据的相关性上表现较优。
- GNB的表现不佳，因为它假设每对特征之间的条件独立性。GNB的这假设导致了分类准确性的降低，这与先前研究中的相应结果是一致的。
- 进一步比较了11种降维方法在ANN分类器上的表现，结果表明，降维方法大约降低了50%的输入特征的数量，并且降低了ANN的训练时间。LLE降低了基于材质和形状的分类时间消耗。
- 利用这些方法可以使机器学习的策略更加有效。

- conclusion
  - 本文基于11种降维方法，3种分类模型研究了33种分类策略，去寻找鲁棒的高效的对地下目标材质和形状分类方法基于EMI电磁模型。
  - 建立仿真平台验证提出的分类策略，在所有的分类策略中，基于特征提取方法的KPCA+ANN在对地下目标的材质和形状进行分类时取得了较好的效果，优于先前研究的基于模型的分类方法，同时表明ANN分类器优于其他两种分类器。
  - 提出了一种基于互信息的输入冗余度和相关性评估方案，证明了降维方法的必要性。比较的结果表明噪声的影响是不可忽略的，分类的准确度随着信噪比的降低而降低。降维方法可以提高分类的鲁棒性。
  - 性能分析表明降维方法可以有效降低ANN的时间消耗，LLE方法降低了50%的特征输入，降低了ANN分类器的构建时间和分类时间。
  - 比较的结果表明KPCA+ANN是有效的基于数据的分类策略，在计算能力有限的便携设备的地下目标的分类应用中具有一定指导意义。
  - 降维方法和机器学习模型的交叉组合策略为寻找地下目标检测的最优机器学习模型提供了一种途径。
  - 我们将致力于优化分类模型，进一步提高分类策略在低信噪比下对地下金属目标探测的鲁棒性。