陆俊儒师兄论文修改

- 论文基本框架
  - 论文基本结构
    - abstract
    - introduction
    - related work
    - basic concept
      - SMOTE
      - SVM
      - SVM-RFE feature selection
        - 特征选择的方法，本文方法是基于该方法进行改进，即加入了对少数类误分样本的重采样，即下文中的特征选择
      - BRFE-PBKS-SVM
        - SVM-BRFE，用于特征选择
        - PBKS 重采样算法
      - experiment
        - evaluation criterions
          - 混淆矩阵定义等
          - 数据集描述
          - 交叉验证设置及实验结果
        - conclusion
  - 整体思路：基于高维不平衡问题的重采样和特征选择
  - 本文采样先特征选择，后重采样的问题，在执行过程中采用F1值作为标准
    - 特征选择（**根据SVM的W值的特征选择，加入了误分类的少数类样本**）
      - 特征选择利用了**支持向量是边界样本**的性质，并在训练过程中还增加了对**误分类少数类的过采样（由于SMOTE在高维空间失效，所以该过采样采用的是重采样的办法，即对样本进行复制）**，并选择F1值最高的W值，对其最低W值对应的特征进行删除
    - 重采样（**变换空间的SMOTE过采样**）
      - 特征选择过后的数据集，样本维数减少，因此对该数据集再进行过采样。
      - SMOTE在原始向量空间进行线性插值，但在SVM的决策平面上，重采样产生的样本可能并不是线性插值的结果了，因此，希望其在SVM的决策空间（希尔伯特空间）上进行线性插值
      - 由于该变换不是可逆变换，因此采用了近似原像的方式。
        - 关于在原始空间or希尔伯特空间采样的问题，是引用了某篇论文，可行性不需验证
      - 对采样率进行启发式搜索，并根据最后生成的采样率进行过采样，最后生成特征选择和重采样过后的数据集
- 老师的意见

- 1、关于高维的定义（特征数量和样本数量的对比，有引用文献）
  - **高维问题中，根据高维问题的定义，则少数类和多数类应该不同（可以思考P）**
- 2. 采样部分对于数据分类结果影响的原因分析

- 我们的思路
  - 扩充introduction和related work部分
    - introduction
      - 加入实例，以证明我们需要更多关注的少数类具有很重要的意义
      - 第二段中针对单方面可以提出一些方法，但指出没有人能同时解决两个问题等情况
      - 指出SVM在解决该类问题中的优越性（**建议放在介绍SVM之前或者之后，即basic concept里**）或许可以加入更多的参考文献以证明SVM的实用性
    - related work
      - 对related work部分的第一段，文中引用了文献，说明SMOTE在高维和低维空间的作用，需要更详细的解释该论文的**结论部分（比如在哪个空间更好用等）原因说明（为何采样办法在原始空间中不好用）**
      - 扩展SMOTE在高维空间的劣势，并给出理由
      - 主要是对先过采样后特征选择，或者只做其一的不好的方面
    - basic concept
      - 对希尔伯特空间稍加解释，这个可以加入到SVM部分
    - 语法部分
      - 句子语法
      - 建议imbalance 和unbalanced统一
  - 不需要修改的部分
    - basic concept
    - experiment part
      - 如果还有需要，可以扩充实验结果分析部分
      - 各种evaluate标准的意义等

- new idea
  - 可以在介绍高维问题里添加SVM，在介绍不平衡问题中添加SMOTE，然后分析两种问题产生的新问题，并指出单独的方法中的不足，然后提出融合方法，并指出其缺陷，提出本文方法。
  - 在融合方法中提出 里面有解决顺序的问题，并分析表明本文中采用的顺序的合理性
  - basic concept部分，SMOTE和SVM交换位置，以对应related work里的先高维后过采样。

- 论文的基本架构
  - 介绍高维问题和不平衡问题
    - 高维问题的解决方法：降维 **是否还需要介绍PCA，如果介绍的话，是否需要介绍其劣势。**
      - PCA
      - 特征选择
    - 不平衡问题：数据和算法，**是否还需要介绍算法类的算法==**
      - 数据层面的采样

        - 算法层面的代价敏感
    - 分析互相影响形成的新问题
    - 介绍SVM，并分析在解决高维问题和不平衡问题的优势
    - 改进SVM-RFE算法，对高维问题进行特征选择，即BRFE-PBKS-SVM算法，即融合了两种问题的解决算法
        - 特征选择和数据重采样
- 老师的修改意见8.5
    - 分析SVM在不平衡问题中的优势或者劣势
    - 扩充参考文献，到30篇左右
    - 英文文献的页数要到12页
- 10.7 修改意见
    - 本文的整体架构是改进了两个算法，首先是基于svm的包裹式特征选择，该特征选择算法是考虑了不平衡分布的，即在特征选择中加入了过采样成分，并根据最高的F1值来消除特征，然后是在希尔伯特空间下的不平衡算法，即在希尔伯特空间下的基于smote的边界采样算法，
    - 建议删去basic concept中的svm-rfe算法，