特征选择实验报告

1711290 李涵 信息安全

一、问题描述

传统监督学习主要是单标签学习，而实际情况往往更加复杂（如我们现在给定的数据集），含有多个标签，例如文本分类，就需要多标签学习。

而在进行机器学习的时候，当特征的维度超过一定界限后，分类器的性能随着特征维度的增加反而下降，因为这些高纬度特征中含有无关特征和冗余特征，因此需要进行特征选择，以去除特征中的无关特征和冗余特征。

二、解决方法

1.解决思路

对于多标签学习，我们利用BinaryRelevance（它基本上把每个标签当作单独的一个类分类问题）的方式，建立n个分类器，利用svm对数据进行拟合，构建分类器

对于很多很多的特征值，我们采用论文当中的算法进行特征选择（包括监督学习和半监督学习）

2.基本理论

多标签学习

多标签分类的策略可以分为三类：

- 一阶策略：忽略和其它标签的相关性，把多标签分解成多个独立的二分类问题
- 二阶策略：考虑标签之间的成对关联
- 高阶策略：考虑多个标签之间的关联

多标签分类评价指标：

Accuracy, Precision, Recall, F值：

$$Accuracy(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{|h(x^i) \cap y^i|}{|h(x^i) \cup y^i|}$$

$$Precision(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{|h(x^i) \cap y^i|}{|h(x^i)|}$$

$$Recall(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{|h(x^i) \cap y^i|}{|y^i|}$$

$$F^{\beta}(h) = \frac{(1+\beta^2) \cdot Precision(h) \cdot Recall(h)}{\beta^2 \cdot Precision(h) \cdot Recall(h)}$$

（1）把多标签问题转为其它学习场景，比如转为二分类，标签排序，多分类

（2）通过改编流行的学习算法去直接处理多标签数据，比如改编懒学习，决策树，核技巧

以下列举了几种：

二分类: 把多个标签分离开来，对于k个标签，建立k个数据集和k个二分类器来进行预测。使用one-vs-rest的方式，简单直接，但没有考虑标签之间的关联性，是一个一阶策略（first-order）

多分类： 把多标签学习问题转为多分类问题。把2^q个可能的标签集，映射成2^q个自然数。局限性在于预测的标签集是训练集中已经出现的，它没法泛化到未见过的标签集，且类别太大，低效

特征选择

特征选择的目的：

- 提高模型预测的准确率
- 减少模型训练的时间
- 降低储存的成本

特征选择的方法：

- Filter：过滤法，按照发散性或者相关性对各个特征进行评分，设定阈值或者待选择阈值的个数，选择特征。
- Wrapper：包装法，根据目标函数（通常是预测效果评分），每次选择若干特征，或者排除若干特征。
- Embedded：嵌入法，先使用某些机器学习的算法和模型进行训练，得到各个特征的权值系数，根据系数从大到小选择特征。类似于Filter方法，但是是通过训练来确定特征的优劣。其中包括：基于L1的特征选择、随机稀疏模型、基于树的特征选择

论文当中的特征选择方法：

对于特征矩阵X，标签矩阵Y（有的有标签，有的无标签），我们假定F矩阵，对于有标签的，F对应的值就是Y对应的值，对于无标签的，F对应的值是预测标签。我们要找到一个f使得下边的函数成立

$$\min_{f,F_l=Y_l} \sum_{i=1}^{n} loss(f(x_i), f_i) + \mu\Omega(f),$$

其中Ω是范数

其在最小二乘法上的应用函数为：

$$\min_{W,F,\mathbf{b},F_l=Y_l} \sum_{i=1}^{n} s_i \|W^T x_i + \mathbf{b} - f_i\|_2^2 + \mu\|W\|_F^2,$$

其中si表示一个训练数据点的得分，1为有标签，0为无标签

为了使模型更加有效，我们在正则化项（1、2范数）上施加稀疏特征选择模型：

$$\min_{W,F,\mathbf{b},F_l=Y_l} \sum_{i=1}^{n} s_i \|W^T x_i + \mathbf{b} - f_i\|_2^2 + \mu \|W\|_{2,1}.$$

其中 0≤fi≤1

设定对角矩阵S，其中Sii=si，得到：

$$\min_{W,F,\mathbf{b},F_l=Y_l} Tr((X^T W + \mathbf{1}b^T - F)^T S(X^T W + \mathbf{1}b^T - F)) + \mu \|W\|_{2,1},$$

经过一系列的公式推导，我们可以得到：

$$XHSHX^T W + \mu DW = XHSHF,$$

W就是我们需要得到的特征选择矩阵，其中D一个对角矩阵，Dii为wi的2范数，所以

$$W = (XHSHX^T + \mu D)^{-1} XHSHF.$$

3.算法流程

（1）输入features和labels以及迭代系数μ

（2）计算S矩阵，如果有标签，则Sii=1，无标签，Sii=0

（3）随机初始化W矩阵

（4）根据上述计算D矩阵，W t+1矩阵，

（5）计算b：$b_{t+1} = \frac{1}{m} F^T S\mathbf{1} - \frac{1}{m} W^T X S\mathbf{1}$；

（6）计算F t+1：$\widetilde{F}_{t+1} = X^T W + \mathbf{1}b^T$ 并调整F，如果F值大于1，则调整为1，若F值小于0，则调整为0

（7）重复（4）~（6）直到收敛

三、实验分析

1.实验数据

数据集1：音乐情感的多标签分类，有593行数据，每行有72个特征值，以及6个标签。音乐中情感的自动检测被建模为多标签分类任务，其中一段音乐可能属于多个类。

数据集2：蛋白质的多标签分类，有662行数据，每行有1186个特征值，以及27个标签

2.实验设计

首先要对读取到的数据进行随机打乱，再对features进行标准化处理，能够使得算法的收敛速度更快

设定迭代系数k，按照算法流程进行计算，最终得到F和W，要尝试不同的k，观察收敛速度，选取较优的k值

利用W计算2范式，根据设定的选取率（1/6，2/6，……，6/6）得到选择选取的特征的（利用headpq.nlargest函数）

对于监督学习，我们完全信任Y，所以用经过特征选择的X和Y对svm进行拟合，再对测试集进行测试；而对于半监督学习，有标签的我们用Y的值，没有标签的我们用F的值，来进行拟合以及测试。

由于这个是多标签学习，普通的svm并不能满足，所以需要调用sklearn.multiclass当中的OneVsRestClassifier类来进行多标签学习。

在评估多标签学习和特征选择的性能好坏的时候，我们调用clf.score来评判学习性能。

3.实验结果

数据集1的监督学习：

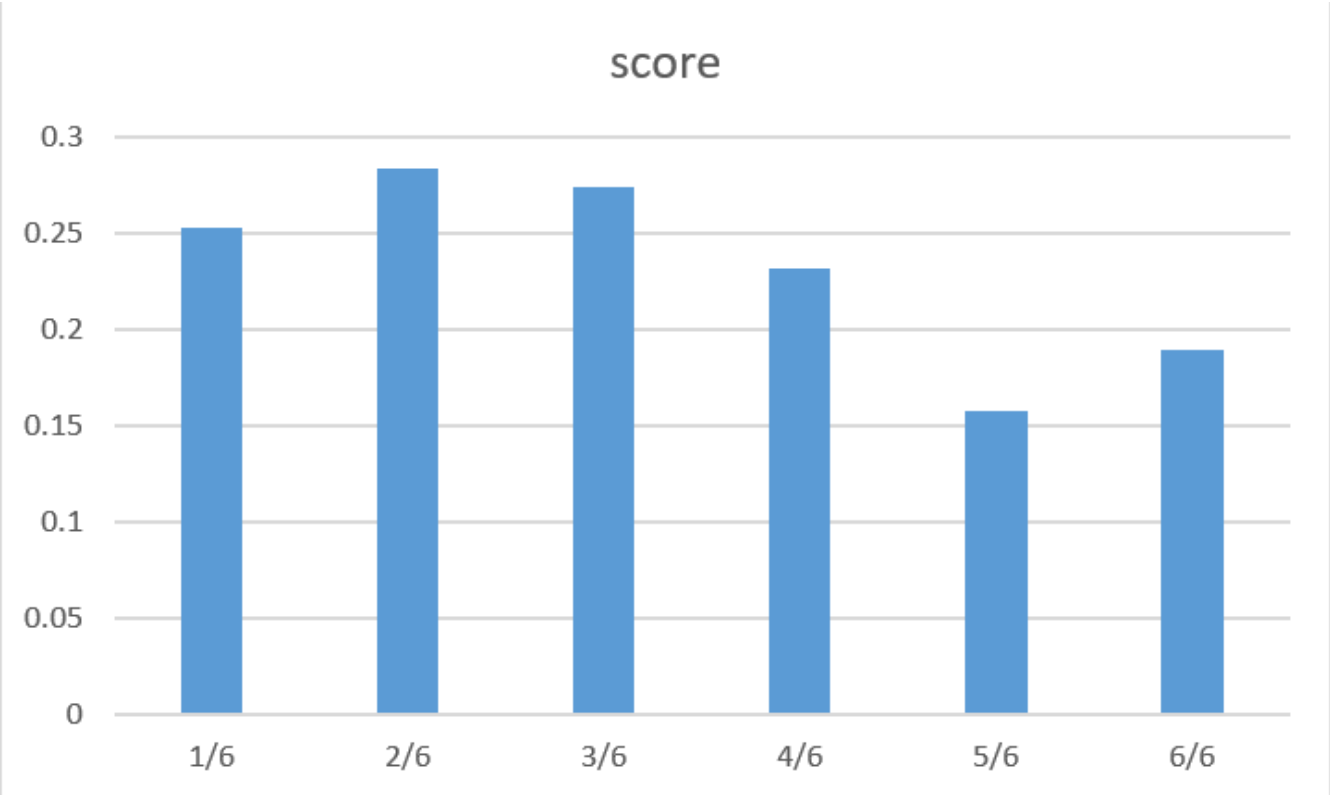
k=1的时候收敛速度很快，所以就让k=1

通过特征选择算法，我们获取的各个特征值的重要性排序：

[2, 17, 0, 67, 65, 3, 18, 16, 22, 1, 4, 26, 19, 64, 57, 27, 39, 58, 42, 55, 68, 21, 60, 59, 69, 5, 23, 47, 44, 62, 28, 25, 29, 41, 35, 36, 54, 43, 31, 20, 7, 61, 46, 9, 53, 15, 63, 49, 30, 32, 56, 40, 70, 12, 6, 66, 24, 50, 52, 33, 45, 10, 34, 38, 37, 71, 51, 14, 8, 13, 11, 48]

初始W的不同会使得重要性有小幅度的波动，但大致趋势是不变的

选择特征比率为1/6，2/6，3/6，4/6，5/6，6/6时，分类器拟合之后的得分分别为（运行100次取平均值）：0.252632、0.284211、0.273684、0.231579、0.157895、0.189474

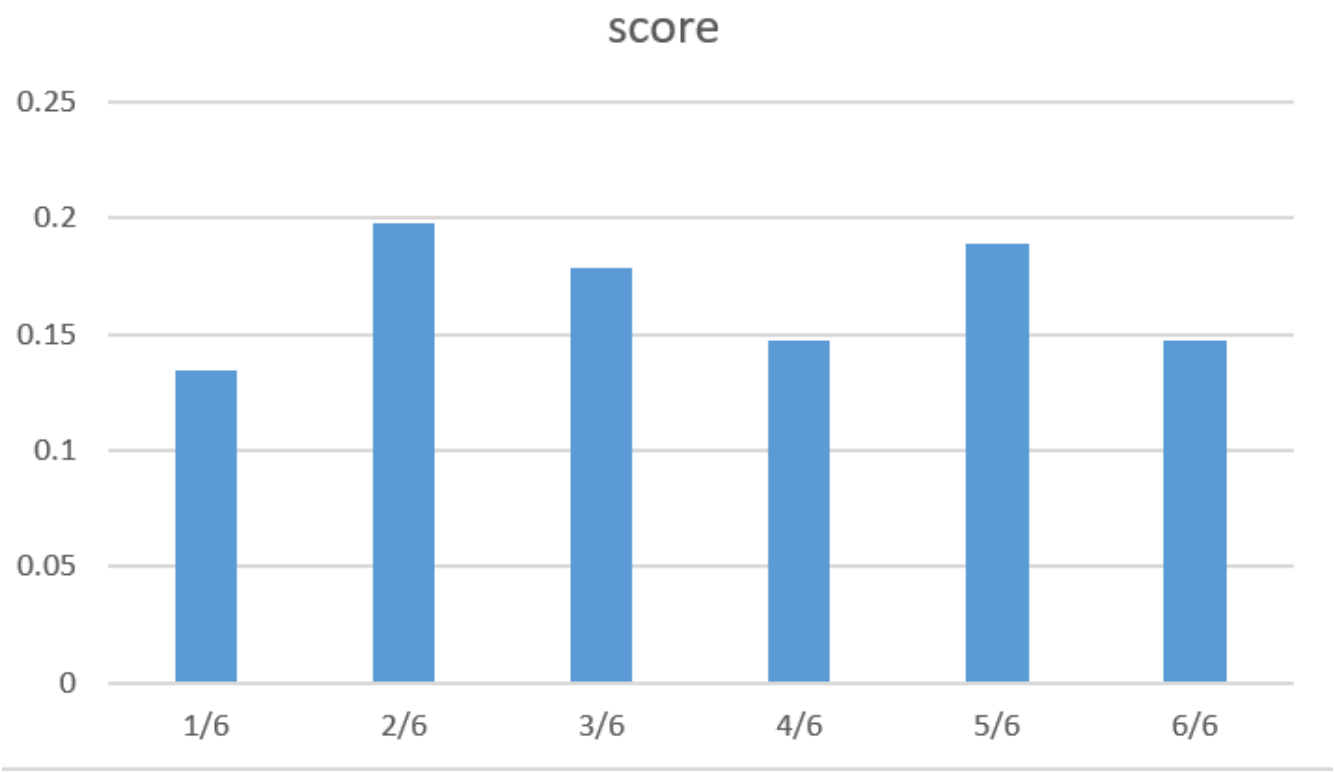


在特征选择比率为2/6时，效果最好，大致有一个先增后减的趋势，但是每次实验的结果都有波动，不一定完全符合。

实验结果表明，在进行特征选择之后，能够过滤掉一些无关特征或冗余特征，能够提高机器学习算法的性能，但是选择的特征过少，一些有用的特征值也被筛选掉，也会导致算法性能的下降，所以特征选区的比率就很重要。

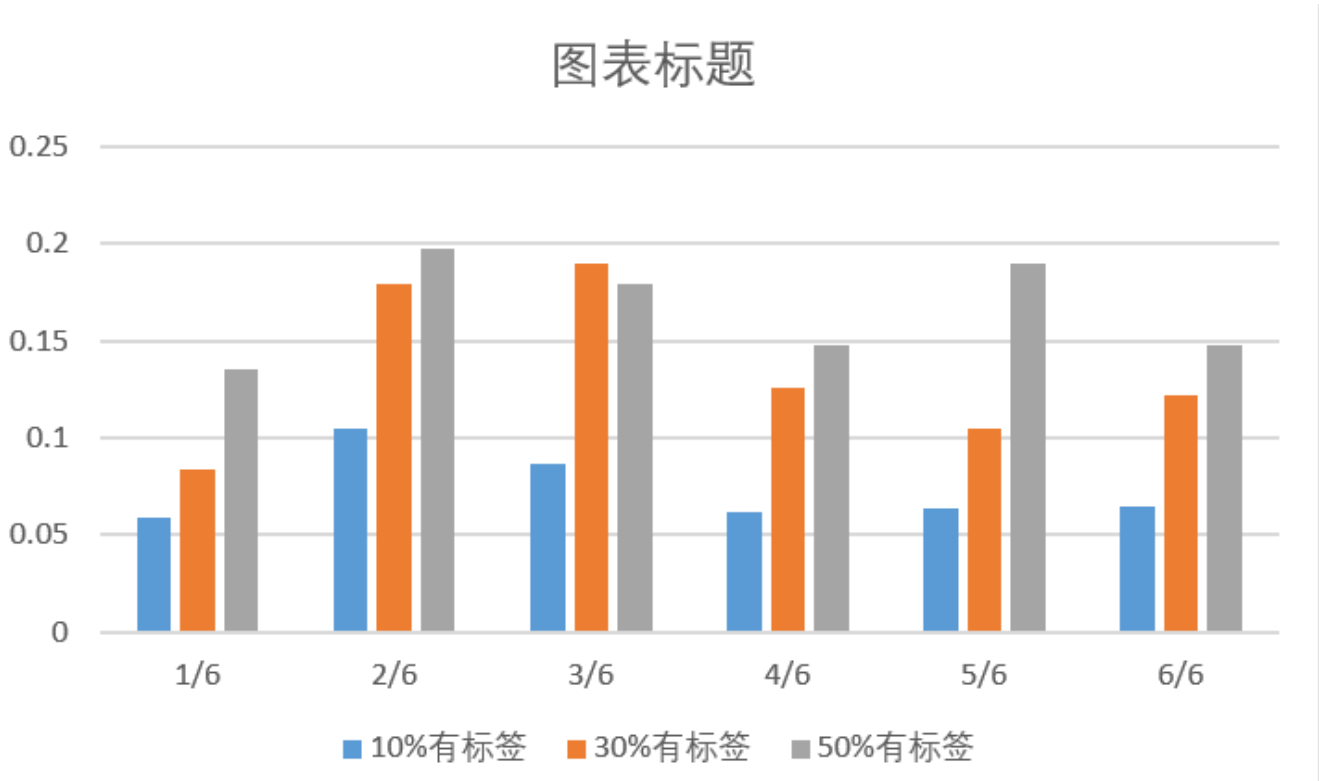
数据集2的半监督学习：

有标签比例为50%时选择特征比率为1/6，2/6，3/6，4/6，5/6，6/6时，分类器拟合之后的得分分别为：0.134947、0.197684、0.178947、0.147368、0.189474、0.147368



在特征选择比率为2/6时，效果最好，半监督学习的特征选择对于算法性能的影响，也和上述监督学习的类似，大致有一个先增后减的趋势，每次结果可能有波动。

在有标签比例为0.1，0.3，0.5，0.7时的分类器性能如下图所示：



由图可知，0.1，0.3时分类器性能都有下降，0.1时下降幅度非常大，收敛速度也很明显慢了，所以在进行半监督学习的时候，有标签的最好比无标签的要多，这样分类器的性能会比较好。