综合实训 机器学习之特征提炼 任务一

年级	**2018**级	专业	软件工程
学号	18342115	姓名	杨玲

一、任务要求

- 任务一：计算每一种特征与标签的相关程度，并按照相关程度对特征进行排序，输出特征顺序。
- 截止时间：10.23
- 提交内容：
  - 一份可读的代码
  - 对于每一个数据集，输出特征的相关程度排序
- 工作量：代码量不大，主要是熟悉数据和机器学习基础

二、相关基础知识

1. 特征选择的意义

特征选择是特征工程中的重要一环，其主要目的是从所有特征中选出相关特征 (*relevant feature*)，或者说在不引起重要信息丢失的前提下去除掉无关特征 (*irrelevant feature*) 和冗余特征 (*redundant feature*)。进行特征选择的好处主要有以下几种：

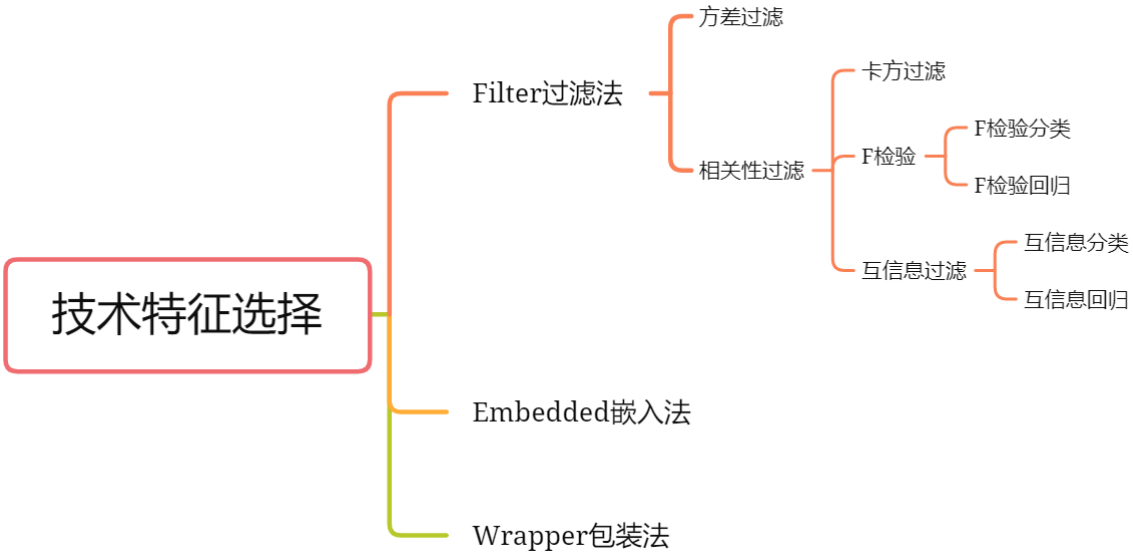
- 降低过拟合风险，提升模型效果
- 提高训练速度，降低运算开销
- 更少的特征通常意味着更好的可解释性

2. 特征选择方法

（**1**）业务特征选择

特征选择之前一定要理解数据中特征的含义，可以先从业务上就能剔除一些不必要的特征，然后再进行技术上的特征选择。

（2）技术特征选择



- 过滤法
  - 方差过滤
    - 首先需要去除样本中方差为0的特征。
  - 相关性过滤
    - 相关性主要是评判特征之间以及特征和标签之间的相关性，
    - 去除特征之间的相关性——因为诸如线性回归之类的模型训练时特征之间相关产生共线性的问题而影响模型效果。
    - 去除与标签不相关的特征——因为如果特征与标签无关，那只会白白浪费我们的计算内存，可能还会给模型带来噪声。
- 嵌入法
  - 嵌入法是一种让算法自己决定使用那些特征的方法，即特征选择和算法训练同时进行。
  - 在使用嵌入法时，我们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，根据权值系数从大到小选择特征。这些权值系数往往代表了特征对于模型的某种贡献或某种重要性，我们就可以基于这种贡献的评估，找出对模型建立最有用的特征。
  - 相比于过滤法，嵌入法的结果会更加精确到模型的效用本身，对于提高模型效力有更好的效果。并且，由于考虑特征对模型的贡献，因此无关的特征(需要相关性过滤的特征)和无区分度的特征(需要方差过滤的特征)都会因为缺乏对模型的贡献而被删除掉，可谓是过滤法的进化版。
- 包装法

- 包装法也是一个特征选择和算法训练同时进行的方法，与嵌入法十分相似，它也是依赖于算法自身的选择来完成特征选择。
- 但不同的是，包装法往往使用一个目标函数作为黑盒来帮助我们选取特征，而不是自己输入某个评估指标或统计量的阈值。包装法在初始特征集上训练评估器，并且获得每个特征的重要性。然后，从当前的一组特征中修剪最不重要的特征。在修剪的集合上递归地重复该过程，直到最终到达所需数量的要选择的特征。
- 区别于过滤法和嵌入法的一次训练解决所有问题，包装法要使用特征子集进行多次训练，因此它所需要的计算成本是最高的。

三、实验过程和结果

1. 总体思路

因为嵌入法和包装法都是特征选择和算法训练同时进行，在任务一的条件下，就暂时不考虑，所以主要考虑过滤法。又由于F检验过滤只能捕捉线性相关性，且要求数据服从正态分布，不太符合条件，所以我主要考虑了卡方过滤和互信息过滤。

2.实验内容

主要是调用sklearn库的feature selection来进行，具体详见代码

3.实验结果

（1）各表具体结果(以2-AS为例)

- 卡方过滤

卡方过滤计算结果：

Number	Feature Name	Scores	p-values	Order
1	max_degree	84.24496541755356	0.6227595454605492	16
2	fail_node_degree	137652.39013174846	0.0	9
3	fail_neber_degree	245929.23296959396	0.0	8
4	fail_degree_sum	381967.09943477967	0.0	7
5	max_load	825490.8955629405	0.0	5
6	big_load_num	502.53502620013404	5.785267663700783e-59	15
7	fail_load_sum	117875020.48519863	0.0	2
8	fail_num	102804.11088544445	0.0	10
9	first_round_fail	96384.35890574704	0.0	11
10	neber_fail_num	81444.51501433809	0.0	12
11	fail_round	2348.966603222667	0.0	14
12	subgraph_num	35682.54168055037	0.0	13
13	fail_node_load	91974111.34551983	0.0	3
14	load_change	255382384.20271373	0.0	1
15	degree_change	686726.305230291	0.0	6
16	fail_neber_load	28740557.387447417	0.0	4

- 互信息分类

互信息分类计算结果：

Number	Feature Name	Scores	Order
1	max_degree	0.4959089972928732	16
2	fail_node_degree	1.0381089751640178	13
3	fail_neber_degree	1.5157696415972435	5
4	fail_degree_sum	1.3628676856728195	9
5	max_load	1.3616243807851887	10
6	big_load_num	0.7473408730328481	14
7	fail_load_sum	1.6431472158887002	2
8	fail_num	1.3763053404912955	7
9	first_round_fail	1.3670777731444255	8
10	neber_fail_num	1.3158935259943383	12
11	fail_round	0.7201366986150464	15
12	subgraph_num	1.463023830680343	6
13	fail_node_load	1.5459127328845645	4
14	load_change	1.8971085221531334	1
15	degree_change	1.3296214618432431	11
16	fail_neber_load	1.641046679250258	3

- 标准化互信息分类

标准化互信息分类计算结果：

Number	Feature Name	Scores	Order
1	max_degree	0.2582647840260078	16
2	fail_node_degree	0.386436120009427	13
3	fail_neber_degree	0.6375940584227631	2
4	fail_degree_sum	0.45684037299582836	11
5	max_load	0.47128246958707637	10
6	big_load_num	0.3569291808920301	14
7	fail_load_sum	0.556339663719416	7
8	fail_num	0.6101596719525884	5
9	first_round_fail	0.6106712601112297	4
10	neber_fail_num	0.5904978933457751	6
11	fail_round	0.3361580643600254	15
12	subgraph_num	0.6348915461963823	3
13	fail_node_load	0.5553593111589533	8
14	load_change	0.4984630961003943	9
15	degree_change	0.454093156263777	12
16	fail_neber_load	0.7054391511800481	1

- 最大互信息数(MIC)

最大互信息系数(MIC)计算结果:

Number	Feature Name	Scores	Order
1	max_degree	0.42101223352174477	16
2	fail_node_degree	0.554293549399555	14
3	fail_neber_degree	0.7481053446689994	8
4	fail_degree_sum	0.676584546560325	13
5	max_load	0.6968257343045013	11
6	big_load_num	0.5475173153954597	15
7	fail_load_sum	0.8288181194905814	3
8	fail_num	0.7548001183168939	6
9	first_round_fail	0.7517882678809957	7
10	neber_fail_num	0.7313917390023617	9
11	fail_round	0.722524503052896	10
12	subgraph_num	0.8441285836171308	2
13	fail_node_load	0.7880875685010161	5
14	load_change	0.8671541132824024	1
15	degree_change	0.6813405332777978	12
16	fail_neber_load	0.8183902428522624	4

（**2**）各表结果总和

- 2-AS

Number	Feature Name	chi2	mutual_info	normalized_mutual_info	Maximal Information Coefficient(MIC)
1	max_degree	16	16	16	16
2	fail_node_degree	9	13	13	14
3	fail_neber_degree	8	5	2	8
4	fail_degree_sum	7	10	11	13
5	max_load	5	9	10	11
6	big_load_num	15	14	14	15
7	fail_load_sum	2	3	7	3
8	fail_num	10	7	5	6
9	first_round_fail	11	8	4	7
10	neber_fail_num	12	12	6	9
11	fail_round	14	15	15	10
12	subgraph_num	13	6	3	2
13	fail_node_load	3	4	8	5
14	load_change	1	1	9	1
15	degree_change	6	11	12	12
16	fail_neber_load	4	2	1	4

- 2-AS_cla

Number	Feature Name	chi2	mutual_info	normalized_mutual_info	Maximal Information Coefficient(MIC)
1	max_degree	16	15	9	15
2	fail_node_degree	9	13	12	13
3	fail_neber_degree	8	4	4	5
4	fail_degree_sum	7	6	11	6
5	max_load	6	11	15	11
6	big_load_num	15	14	7	14
7	fail_load_sum	2	3	13	3
8	fail_num	10	8	3	10
9	first_round_fail	11	5	1	8
10	neber_fail_num	12	9	2	9
11	fail_round	14	16	8	16
12	subgraph_num	13	12	5	12
13	fail_node_load	3	10	14	4
14	load_change	1	1	16	1
15	degree_change	5	7	10	7
16	fail_neber_load	4	2	6	2

- 2-BA

Number	Feature Name	chi2	mutual_info	normalized_mutual_info	Maximal Information Coefficient(MIC)
1	max_degree	16	16	16	16
2	fail_node_degree	12	13	15	13
3	fail_neber_degree	8	5	5	6
4	fail_degree_sum	7	6	6	8
5	max_load	6	15	13	15
6	big_load_num	15	14	14	14
7	fail_load_sum	2	9	11	11
8	fail_num	9	2	1	3
9	first_round_fail	10	3	2	2
10	neber_fail_num	11	7	4	7
11	fail_round	14	11	9	4
12	subgraph_num	13	12	10	10
13	fail_node_load	3	10	12	12
14	load_change	1	1	8	1
15	degree_change	5	8	7	9
16	fail_neber_load	4	4	3	5

- 2-BA_cla

Number	Feature Name	chi2	mutual_info	normalized_mutual_info	Maximal Information Coefficient(MIC)
1	max_degree	16	16	12	16
2	fail_node_degree	12	12	11	11
3	fail_neber_degree	8	7	5	7
4	fail_degree_sum	7	5	9	5
5	max_load	6	14	16	13
6	big_load_num	14	13	7	14
7	fail_load_sum	2	9	13	8
8	fail_num	9	2	2	3
9	first_round_fail	10	3	3	4
10	neber_fail_num	11	8	4	9
11	fail_round	15	15	6	15
12	subgraph_num	13	11	1	12
13	fail_node_load	3	10	15	10
14	load_change	1	1	14	1
15	degree_change	5	6	10	6
16	fail_neber_load	4	4	8	2

- 2-USAirlines

Number	Feature Name	chi2	mutual_info	normalized_mutual_info	Maximal Information Coefficient(MIC)
1	max_degree	15	16	13	15
2	fail_node_degree	9	14	16	16
3	fail_neber_degree	5	7	3	6
4	fail_degree_sum	8	11	14	12
5	max_load	7	6	12	11
6	big_load_num	16	15	10	14
7	fail_load_sum	2	3	9	4
8	fail_num	10	10	4	6
9	first_round_fail	11	9	4	6
10	neber_fail_num	12	8	6	6
11	fail_round	14	13	7	6
12	subgraph_num	13	2	1	1
13	fail_node_load	3	4	8	5
14	load_change	1	1	11	2
15	degree_change	6	12	15	13
16	fail_neber_load	4	5	2	3

- 2-USAirlines_cla

Number	Feature Name	chi2	mutual_info	normalized_mutual_info	Maximal Information Coefficient(MIC)
1	max_degree	15	16	9	16
2	fail_node_degree	9	15	16	15
3	fail_neber_degree	5	5	3	7
4	fail_degree_sum	7	11	13	12
5	max_load	8	7	14	6
6	big_load_num	16	14	8	14
7	fail_load_sum	2	2	10	2
8	fail_num	10	8	4	9
9	first_round_fail	11	10	4	9
10	neber_fail_num	12	9	6	8
11	fail_round	14	13	1	11
12	subgraph_num	13	6	7	5
13	fail_node_load	3	3	11	3
14	load_change	1	1	12	1
15	degree_change	6	12	15	13
16	fail_neber_load	4	4	2	4

四、分析与总结

从实验结果可以看出，不同的过滤方法得出的结果是不同的，毕竟每种方法的原理都不同，且都涉及到不同调整方法的超参数。以上卡方过滤和互信息过滤的方法的区别，我猜测可能是因为卡方过滤对于出现次数较少的特征更容易给出高分。例如某一个特征就出现过一次在分类正确的数据中，则该特征会得到相对高的分数，而互信息则给分较低。其主要原因是互信息在外部乘上了一个该类型出现的概率值，从而打压了出现较少特征的分数。

一般来说，过滤法更快速，但更粗糙；包装法和嵌入法更精确，比较适合具体到算法去调整，但计算量比较大，运行时间长。

- 当数据量很大的时候，优先使用方差过滤和互信息法调整，再上其他特征选择方法。
- 使用逻辑回归时，优先使用嵌入法。
- 使用支持向量机时，优先使用包装法。

- 不知从何开始时，一般从过滤法开始，再看具体数据具体分析。

五、参考文档

机器学习如何计算特征的重要性_简介机器学习中的特征工程

机器学习如何计算特征的重要性_机器学习之特征工程