1，迁移学习

将已有的技术和模型运用到不同的业务场景，对公司开展不同的业务可能有帮助。比如可以把信息安全行业的自然语言处理模型应用于银行业等其他行业

附件1：什么是迁移学习

2， 异常检测

异常检测算法的目标是发现与大部分其他对象不同的对象（数据的离群点），这类算法常用于欺诈检测和入侵检测。针对我们 log 大数据而言，可以利用异常检测算法发现其中的异常日志。

技术手段：异常检测算法通常包括

- 基于模型的技术
- 基于密度的技术
- 基于聚类的技术

我觉得我们都可以尝试。

以基于模型的技术而言，主要思路是利用统计学方法，为数据创建一个概率分布模型，然后考虑我们要检测的对象在多大程度上符合该模型。

以基于密度的技术而言，离群点是在低密度区域中出现的对象。我们可以先定义密度，定义完密度之后我们再从低密度区域中寻找异常 log。

以基于聚类的技术而言，可以先对数据进行聚类创建模型，然后我们可以度量去掉某个点对聚类中心影响的程度。也就是说，异常点是对聚类中心扭曲最大的点。

异常检测总结性的论文3篇
附件3
附件4
附件5

3， 特征工程

特征工程的子问题

1. 机器学习中的特征（Feature）

在机器学习和模式识别中，特征是在观测现象中的一种独立、可测量的属性。选择信息量大的、有差别性的、独立的特征是模式识别、分类和回归问题的关键一步。
最初的原始特征数据集可能太大，或者信息冗余，因此在机器学习的应用中，一个初始步骤就是选择特征的子集，或构建一套新的特征集，减少功能来促进算法的学习，提高泛化能力和可解释性。
在表格数据中，观测数据或实例（对应表格的一行）由不同的变量或者属性（表格的一列）构成，这里属性其实就是特征。但是与属性一词不同的是，特征是对于分析和解决问题有用、有意义的属性。
在机器视觉中，一幅图像是一个观测，但是特征可能是图中的一条线；在自然语言处理中，一个文本是一个观测，但是其中的段落或者词频可能才是一种特征；在语音识别中，一段语音是一个观测，但是一个词或者音素才是一种特征。

2. 特征的重要性（Feature Importance）

你可以客观的评价特征的实用性。判别特征的重要性是对特征进行选择的预先指标，特征根据重要性被分配分数，然后根据分数不同进行排序，其中高分的特征被选择出来放入训练数据集。
如果与因变量（预测的事物）高度相关，则这个特征可能很重要，其中相关系数和独立变量方法是常用的方法。
在构建模型的过程中，一些复杂的预测模型会在算法内部进行特征重要性的评价和选择，如多元自适应回归样条法(Multivariate Adaptive Regression Splines, MARS)、随机森林(Random Forest)、梯度提升机(Gradient Boosted Machines)。这些模型在模型准备阶段会进行变量重要性的确定。

3. 特征提取（Feature Extraction）

一些观测数据如果直接建模，其原始状态的数据太多。像图像、音频和文本数据，如果将其看做是表格数据，那么其中包含了数以千计的属性。
特征提取是自动地对原始观测降维，使其特征集合小到可以进行建模的过程。
对于表格式数据，可以使用主元素分析(Principal Component Analysis)、聚类等映射方法；对于图像数据，可以进行线(line)或边缘(edge)的提取；根据相应的领域，图像、视频和音频数据可以有很多数字信号处理的方法对其进行处理。

4. 特征选择（Feature Selection）

不同的特征对模型的准确度的影响不同，有些特征与要解决的问题不相关，有些特征是冗余信息，这些特征都应该被移除掉。
特征选择是自动地选择出对于问题最重要的那些特征子集的过程。
特征选择算法可以使用评分的方法来进行排序；还有些方法通过反复试验来搜索出特征子集，自动地创建并评估模型以得到客观的、预测效果最好的特征子集；还有一些方法，将特征选择作为模型的附加功能，像逐步回归法(Stepwise regression)就是一个在模型构建过程中自动进行特征选择的算法。

5. 特征构建（Feature Construction）

特征重要性和选择是告诉使用者特征的客观特性，但这些工作之后，需要你人工进行特征的构建。
特征构建需要花费大量的时间对实际样本数据进行处理，思考数据的结构，和如何将特征数据输入给预测算法。
对于表格数据，特征构建意味着将特征进行混合或组合以得到新的特征，或通过对

特征进行分解或切分来构造新的特征；对于文本数据，特征够自己按意味着设计出针对特定问题的文本指标；对于图像数据，这意味着自动过滤，得到相关的结构。

6. 特征学习（Feature Learning）

特征学习是在原始数据中自动识别和使用特征。
现代深度学习方法在特征学习领域有很多成功案例，比如自编码器和受限玻尔兹曼机。它们以无监督或半监督的方式实现自动的学习抽象的特征表示（压缩形式），其结果用于支撑像语音识别、图像分类、物体识别和其他领域的先进成果。
抽象的特征表达可以自动得到，但是你无法理解和利用这些学习得到的结果，只有黑盒的方式才可以使用这些特征。你不可能轻易懂得如何创造和那些效果很好的特征相似或相异的特征。这个技能是很难的，但同时它也是很有魅力的，很重要的。

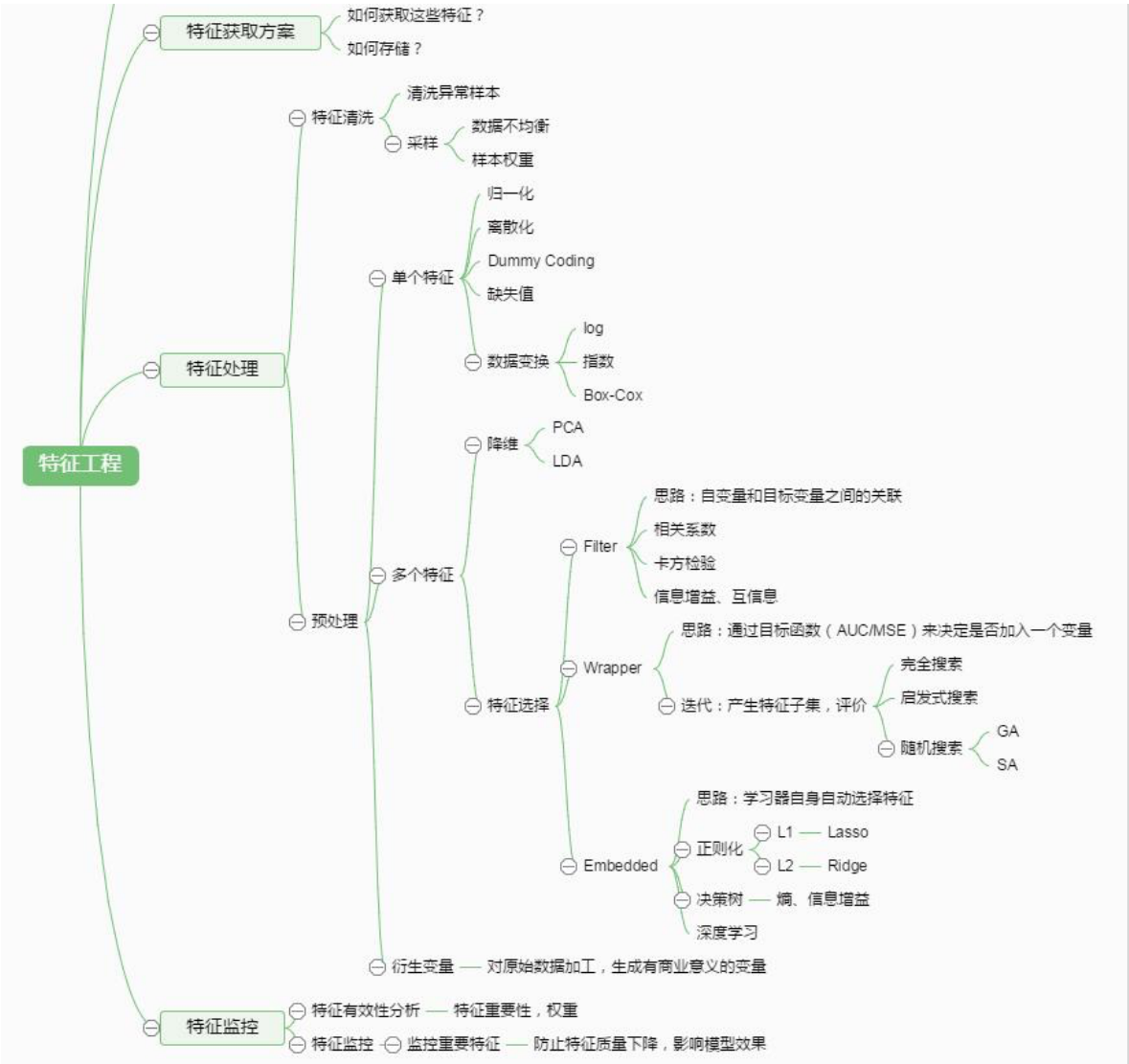
特征工程的流程

机器学习中数据的转换过程：

- 选择数据：收集整合数据，将数据规划化为一个数据集
- 预处理数据：对数据进行清洗、格式化、采样
- 转换数据：特征工程所在
- 对数据建模：构建模型、评估模型、调整模型

特征工程的迭代过程：

- 对特征进行头脑风暴：深入分析问题，观察数据特点，参考其他问题的有关特征工程的方法并应用到自己问题中
- 特征的设计：你可以自动提取特征，手动构造特征，或将两者相结合
- 特征选择：使用不同的特征重要性评分方法或特征选择方法
- 评估模型：利用所选择的特征对测试数据进行预测，评估模型准确性



我自己的一篇论文，关于特征选择
附件：6

4，自然语言处理中的特征工程技术

附件：7

同时，自然语言处理比较热的技术
1. 深度学习
2. 增强学习
3. 知识图谱

附近8： 基于知识图谱的中外自然语言处理研究的对比分析（http://manu44.magtech.com.cn/Jwk_infotech_wk3/article/2014/1003-3513/1003-3513-30-12-51.html#outline_anchor_21）