**学习报告**

我认为之前阅读的文章有所不足，特搜寻了国外的论文以及国内的证券研究报告来补充自己的知识以及寻找更多可行的线索，文献综述如下：

**Mining corporate annual reports for intelligent detection of financial statement fraud - A comparative study of machine learning methods, Petr Hajek, Roberto Henriques, Knowledge-Based Systems, 2017：**这篇文章是一篇综述性质比较强的文章，里面详细的介绍了人工智能方法在公司造假方面的识别作用方式，列举了一些主要的金融变量并对他们做了实证分析。

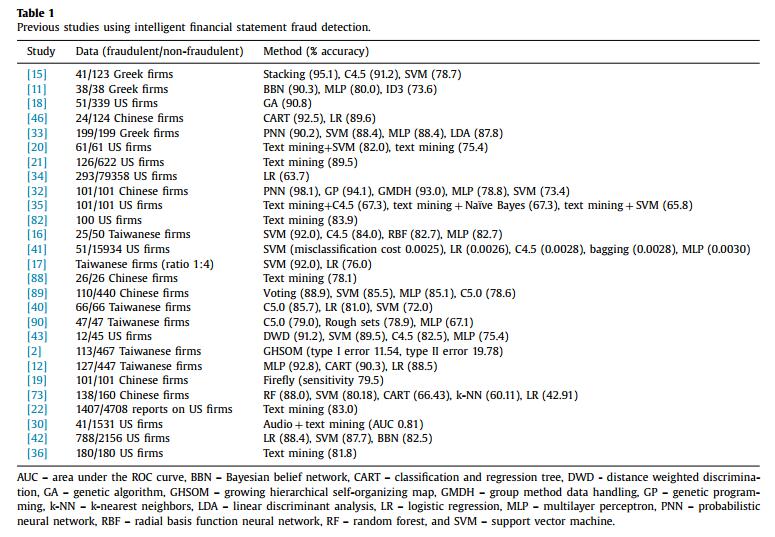
**中国上市公司财务造假预测模型\_钱苹：**预测模型结果表明经营应计项指数、现金销售率、股票换手率波动率、股权集中度、机构投资者持股比率、其他应收款比例、公司当年是否处于亏损状况、是否进行了再融资和股市是否处于熊市等变量显著地反映了中国财务造假公司特征。

**数据挖掘在上市公司财务造假识别中的应用：**这篇研究报告使用了许多数据挖掘的模型对中国造假的样本进行实证，寻找共同的识别变量以及比较他们的表现。审计信息、前一年是否亏损、其他应收款占流动资产比例、前五大股东股权集中度等具有显著预测意义。

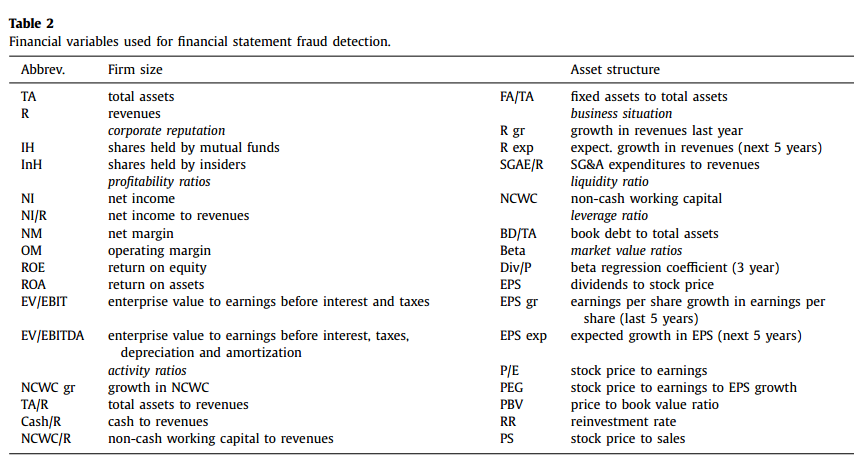
造假动机：防止ST；吸引资金；过渡危机…

造假手段：虚增交易；虚增资产；提前确认收入、虚增收入；利用过渡性科目、调节利润；隐瞒或不及时披露重大事项

检测模型(method)：传统方法（Mscore，Fscore，Cscore）；人工智能方法：



影响因子总结：profitability, activity, asset structure, liquidity, business situation, leverage, market value; insider holdings, reinvestment ratios… (listed below in detail); linguistic variables extracted from firm-related textual documents(by text mining)…



Researches related to Text mining:

Recent studies have begun to concentrate on linguistic variables extracted from firm-related

textual documents that may contain misleading statements. Text mining is reported highly useful for financial statement fraud because large amounts of textual data is associated with this type of fraud [9]. Humpherys et al. [35] performed a textual analysis of the MD&A (management discussion and analysis) section of Form 10-Ks which public companies must file annually with the SEC. Humpherys et al. [35] achieved up to 67.3% accuracy by using measures of linguistic characteristics such as lexical diversity and syntactic complexity. Sentiment analysis and part-of-speech features were used by [36], demonstrating that both positive and negative sentiment is more pronounced in fraudulent reports. Minhas and Hussain [37] compared bag-of-words and sentiment analysis to show that they perform similarly in predicting financial statement fraud. Additionally, non-verbal vocal cues have shown promising results in financial statement fraud detection [30,38]. Using common pre-processing and statistical methods, the unstructured data can be transformed into quantitative variables. This enables the subsequent use of classification methods.

1. 词汇、句法复杂度
2. 情绪和词性特征
3. “词袋”模型
4. 演讲等音像资料的线索

**我的看法**

舆情方面：舆情的情绪基调（负面舆情占比…）、微信传播指数WCI等，基于假设：新闻记者具备更敏锐的嗅觉以及消息不会毫无根据（000006 2016-01负面新闻导致股票大幅下跌；603322 2017-03；000631 2016-01等）

正式报告方面：可以参照前人的研究，暂时没有想到什么创新点，这方面我碰到了一个困难。

研究目标的角度：不一定非要是造假，也可以是公司某种能力的一个预测，比如盈利能力、偿债能力，因为不同的投资者可能会关注一个公司的某一个方面，也可以作财务困境或者破产风险的预测（例如大量债务即将到期但是偿债能力很差就可以预测具有较大的财务困境）

传统方法方面：根据前人的研究选出比较显著的指标以外，可以构建一系列index来评估成长能力、营运能力、现金流量、偿债能力、盈利能力与收益质量等等作为目标变量，或者是选取其中某些具有代表性的指标作为目标变量，然后看看前人研究中对公司造假显著的变量是不是对这些目标变量有一定的预测作用。潜在的造假可能跟这些index之间有一定的关联，可以通过index来联合预测（比如盈利能力很高但是其他应收款占流动资产比率高、审计信息不好），在对公司情况进行预测的同时，根据预测结果来判断公司是否有利润下降风险等。

**可能的方向**

1. 分析舆情情绪等预测公司未来股价变化（要考虑时间周期以及影响发生后的评述性舆情）
2. 像研究报告里面所述一样，自变量是审计意见、前一年是否亏损、股权集中度等，因变量是公司是否出现造假；我们用过往数据来训练这个系统，然后把这个系统应用到现在的股票市场（比如A股市场）筛选出有造假风险的公司，再来人工判定或者想别的办法判定（比如舆情方面等）
3. 我从wind上面下载了一个evaluator模板，我有这样一个新的想法：用某种方式来估计营业收入等等预测所需核心假设的上行或下行变化来对这个公司的未来报表做出估计。举例：我预测由于市场供求关系变化等等，生产的产品的相对价格上升，营业收入相对于营业成本增加，利润率增加，预测公司利好；这种变化分行业和个别公司；我想，是否能通过分析公开文件，如政府报告或者市场上的研究报告的语气来估计某行业产品价格浮动趋势或者原材料价格浮动趋势等等？或者还可以通过其他途径，暂时还没想到。
4. 目前关联交易还是比较成问题，能不能通过大数据的办法，发现隐藏的关联交易？比如调查某次交易的双方公司实际控股人的关联性的方式来研究关联交易
5. 一般有着较高利润率但是市值并不是特别高的企业容易发生利润的不可持续的情况，无论是造假还是提前确认收入、虚增收入等等；还有逆行业趋势的公司也是如此，如行业以及行业龙头的利润率都有所下降但是该公司利润率走高，能不能通过分析公司对外公开文件或者一些专利信息等来判断这个公司是不是具有独特的竞争优势来解释如此高的利润率，并让投资者相信至少在最近一段时间内利润是可以持续的
6. 根据股东减持公司股票以及股权质押等结合现金流量表的情况来分析公司现金流情况、是否存在较大的违约风险。（单单考虑股东减持公司股票以及股权质押对股票价格波动的预测往往是不够的，可能还要结合现金流量表等其他因子来考虑，可能要用以前的出现了债务违约的公司数据进行学习）

**一个具体研究办法**

1. 主干思想：根据研究报告**数据挖掘在上市公司财务造假识别中的应用**中提到的几种模型（多层神经网络、SVM、决策树），三种任选一个或者结合，用历年被发现的造假公司作为训练集，针对现在市场中的股票进行预测造假的可能性，观察结果：

路径1：如果大多数公司造假的可能性不大，只有少部分公司的预测造假可能性大，这个时候用下面第二点的细节分析来从另一个角度解释这个结果

路径2：如果预测造假可能性大的公司数量比较多，就用下面第二点的细节分析来进一步缩小范围（可以分类，分作三类，细节分析预测造假的、细节分析预测未造假的、细节分析得到的结果不明朗的）

无论是哪一条路径，根据分类结果，对每一类做回归，回归因子参考文献综述里面的前两篇文章，分辨观察显著性等等，再做类似于那篇英文论文的分析（相当于把原文章的分类“是否造假”改为“预测是否造假”，启发来自于2sls），之后可以跟造假、未造假公司的结果进行对照。

1. 细节分析：
2. 竞争优势分析：如专利、上下游（成本、售价）优势、企业规模等，可以是财务的，也可以是非财务的，更可以做结合，但我认为只有当完全没有竞争优势的情况下才能比较精确的预测高收益的不可持续性
3. 情绪分析：分析公司公告、报告的情绪等（借鉴**Mining corporate annual reports for intelligent detection of financial statement fraud - A comparative study of machine learning methods**提到的方法）；舆情的情绪；公司热度（如微信指数等）
4. 内控分析：可以通过挖掘内控报告里面的差异（深度and广度对比，差异主要在财务和非财务标准的制定方面）来为内控作出辅助意见，帮助我们挑选存在潜在内控问题的公司

……

其他：

为了证明细节分析对公司造假的标志性，可以研究历史造假公司的数据，也可以通过举例说明，甚至可以不说明（因为原因是一个简单的逻辑链，讲清楚逻辑应该也可行？）

内控报告可以从金融界网站搜索到，行业间差异可能会比较大，同行业也要考虑公司历史因素，可能不是特别能说明问题，但是内控的评估过于复杂浩大，如果这方面不是很显著的话，我觉得内控部分可能比较难下手

注：两个表格均截取自Mining corporate annual reports for intelligent detection of financial statement fraud - A comparative study of machine learning methods, Petr Hajek, Roberto Henriques, Knowledge-Based Systems, 2017