南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Department of Computer Science and Engineering

Advanced Computer Science Experiment	
ASSESSMENT FORM	
List of team students	11612527 曾政然 11611403 张林琛 11612601 何治成
Project Title	面向金融研究报告的数据处理
The Stage of Project Inspection	The first project inspection (何治成) The second project inspection (张林琛) The finally project inspection (曾政然)
Supervisor Name	骆宗伟
Inspector Name	王琦

Project Report:

一、 问题背景与定义

金融作为一个高速变化、竞争激烈的行业，随着大数据、云计算等新兴技术的涌现，我国金融信息化建设的步伐正在加快，金融行业的从业人员面临的压力也越发巨大。在投资行为中，如何在有限的时间内对获取到的信息进行有效的处理，使之成为对投资有利的情报，以此来提高投资收益率是投资时需要思考的一个重要问题；而对于投资人而言，金融数据分析的结果很大程度的影响着他们的投资决策。如此，一个能自动进行金融数据处理与分析的程序会对金融行业的从业人员作出进一步分析或投资人的投资行为有所帮助。

二、 预期结果

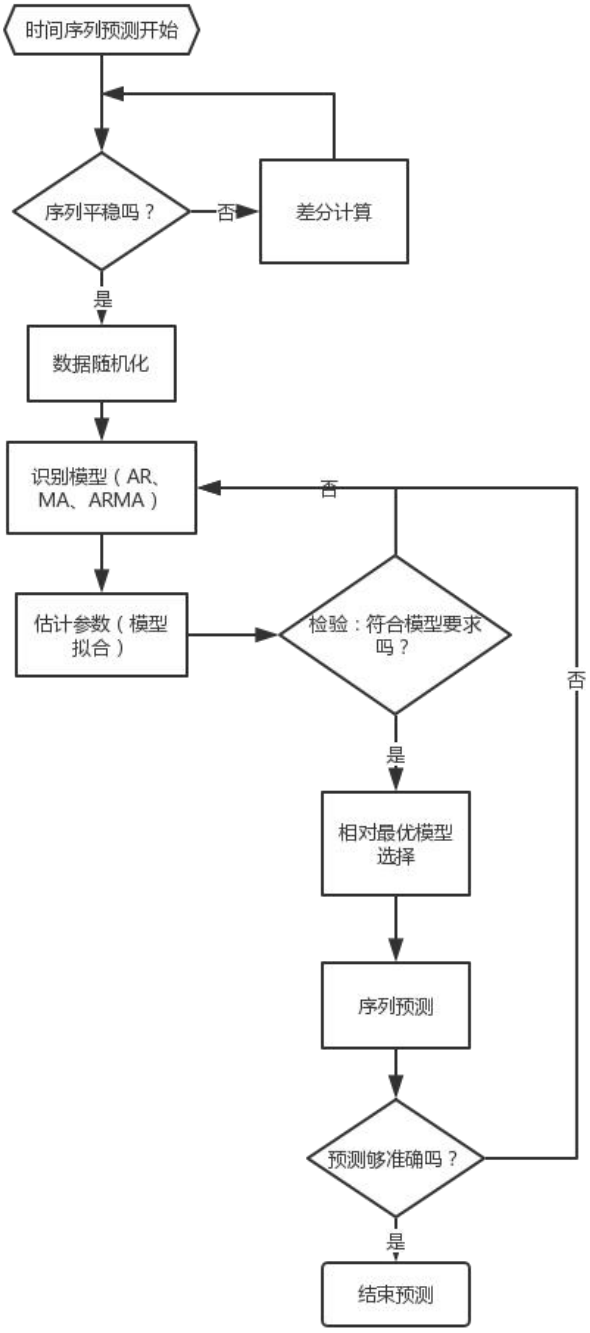
本项目将参照证券分析师的分析过程，实现一个自动化的证券分析系统。首先应确定分析过程的 TISM 模型，并根据 TISM 模型进行分析系统的数学模型构建，最终得到一个类似证券分析师的证券分析系统。

三、 研究方法

1. 时间序列研究方法

用于对输入数据进行预处理以及对分析系统进行长时间维度的检验。

- 金融数据的时间序列处理的一般步骤



- 时间序列概念

时间序列是指将一个现象的某一个统计指标在不同时间上的各个数值收集起来，并按照时间的先后顺序排列而形成的序列。

- 时间序列的预处理
  - 平稳化

    平稳化序列（stationary series），去除基本上不存在趋势的序列。将时间序列的趋势项和季节项都去掉，只留下随机项（随机化，排除干扰）。
- 模型拟合

  使用 ARMA 模型(自回归滑动平均模型，Auto-Regressive and Moving Average Model），是研究时间序列的最常用方法
- 模型检测
- 模型优化

2. TISM（总体解释结构模型，Total Interpretative Structural Modeling）

- ISM（解释结构模型，Interpretative Structural Modeling）

  解释结构模型法是用于分析教育技术研究中复杂要素间关联结构的一种专门研究方法，作用为通过利用系统要素之间已知的零乱关系，揭示出系统的内部结构。解释结构模型法的具体操作是利用图形和矩阵描述出各种已知的关系，然后通过矩阵运算推导出系统结构的关系

- TISM

  ISM 仅仅能定性地推导出系统中各个因素的层次关系，但通过估值和检测，改进后的 TISM 能够在一定程度上反映出各个因素间影响的强弱关系，为时间序列分析提供多维度参考

3. 事件研究法

用于对分析系统进行检验。事件研究法（Event Study）是金融学领域的经典研究方法之一，是一种经验性财务研究技术，应用这种技术可以使观察者评估某一特定事件对公司股价的影响。

四、 系统结构设计

1. 因子分析

根据 Jeffrey C .Hooke 的 Security Analysis and Business Valuation on Wall Street: A Comprehensive Guide to Today's Valuation Methods (Wiley Finance Book 458) 2nd，我们总结了证券分析过程中需要考虑的十个主要影响因子。

System Enablers

Variable	Name
E1	宏观经济形势
E2	相关股市
E3	行业分析
E4	政府政策
E5	竞争
E6	企业状况
E7	历史财务状况
E8	财务预测
E9	估值
E10	安全边际

1.1 宏观经济形势

进行一次证券分析的首要任务是确定当前的宏观经济形势，要确定某些关键指标，如未来国民生产总值增长率，未来的利率和汇率等，这一步主要由经济学家考虑，而分析师要判断当前的经济形势会对企业造成何种影响。（p63-66）

1.2 相关股市

企业所在的股市也会对最终的分析结果造成很重要的影响，分析师要判断和预测股市的大致情况，判断当前股市是过热还是过冷，从而做出合理的投资决定。（p63-66）

1.3 行业分析

企业所处的行业也是进行证券分析的一个因子，分析师需要判断行业的趋势，确定该行业是属于成长期，成熟期或衰退期，同时还要考虑政府的政策，有无技术革新，主要原材料价格变动和客户需求变化等，最终预测行业的前景。（p72-97）

1.4 政府政策

政府的政策对公司的运营有着很大的影响，可能会作用于宏观经济，行业和

特定的公司。分析师需要格外留意。（p80-81）

1.5 竞争优势

对行业内竞争的分析也是证券分析的一个重要过程，分析师需要尽可能的分析每个市场参与者的优势与劣势，分析每个参与者的市场地位，获利能力，技术能力，产品质量，进而评估企业在行业中的地位。（p95-97）

1.6 企业状况

企业的经营状况也是证券分析过程的一个要点，企业的生命周期，研发投入，组织架构还有政府管制都是分析师所需要考虑的。（p99-117）

1.7 历史财务状况

对公司当前和历史的财务分析是分析师做出预测和估值的基础，分析师需要判断公司以往的财务状况是否健康，同时评估公司当前的盈利能力。（p116-137）

1.8 财务预测

财务预测是分析师做出决策的一个重要依据，分析师在一定的假设前提下，通过宏观经济形势，历史财务数据，行业趋势等对公司未来财务状况做出合理的预测。（p182-194）

1.9 估值

书中提供了多种估值方法，如内在价值法，相对价值法，收购价值法，杠杆收购法，技术法等，每种方法都有各自的局限，因此书中也建议采用多种方法对企业估值。（p196-252）

1.10 安全边际

安全边际原理在证券分析中占有重要的地位，因为经济和金融预测具有内在的不确定性，作为防御措施，应该在市场价格和分析师认定的价值之间设立保护性的缓冲区。（p69-70）

2. 解释逻辑

我们运用整体解释模型（TISM）来分析和解释各个因子间的逻辑影响关联。近年来，TISM 在运营管理学领域的运用越来越广，许多先期的研究都表明，在复杂

系统的解释过程中 TISM 比别的解释模型更有优势。

在 TISM 中，我们需要咨询相关领域的专业人士得到各个因子间的影响关系，为此我们向学校金融系的老师发了一份调查问卷，并等到了如下结果。

Structural self-interaction matrix of enablers (SSIM)

	E10	E9	E8	E7	E6	E5	E4	E3	E2	E1
E1	V	O	V	O	O	O	A	V	V	X
E2	V	V	O	O	O	O	A	O	X	
E3	V	V	V	V	V	V	A	X		
E4	O	O	O	V	V	V	X			
E5	O	V	V	X	A	X				
E6	O	V	V	V	X					
E7	O	V	V	X						
E8	O	V	X							
E9	O	X								
E10	X									

在得到了各因素之间的相互影响关系图后，我们进一步的根据命题的传递原理得到各因子间的完整关系图，并最终得到它们之间相互影响关系的二进制矩阵图。并根据矩阵图对因子进行分级。

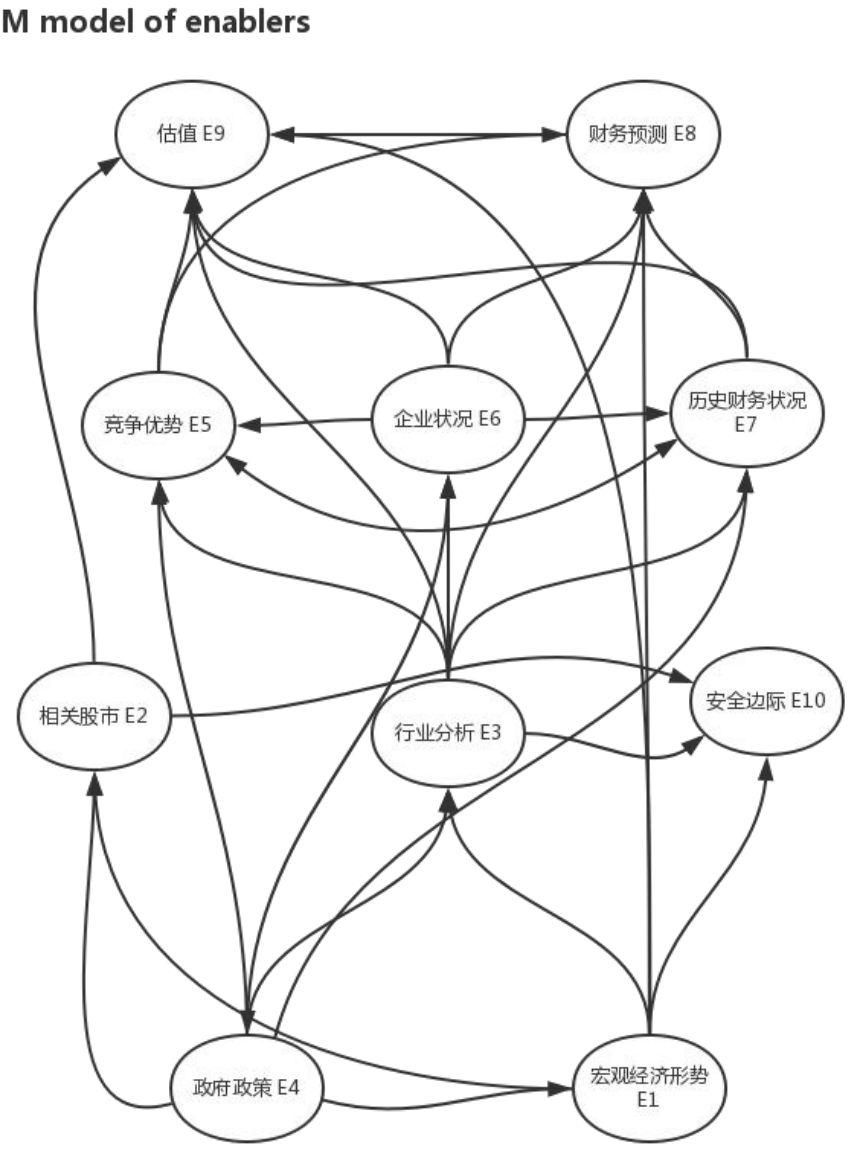
Final binary matrix-enablers

0	E10	E9	E8	E7	E6	E5	E4	E3	E2	E1	Driving Power
E1	1	1	1	0	1	1	0	1	1	-	7
E2	1	1	0	0	0	0	0	0	-	0	1
E3	1	1	1	1	1	1	0	-	0	0	5
E4	0	1	1	1	1	1	-	1	1	1	7
E5	0	1	1	1	0	-	0	0	0	0	2
E6	0	1	1	1	-	1	0	0	0	0	4
E7	0	1	1	-	1	1	0	0	0	0	4
E8	0	1	-	0	0	0	0	0	0	0	1
E9	0	-	0	0	0	0	0	0	0	0	0
E10	-	0	0	0	0	0	0	0	0	0	0
Dependence	3	8	6	4	4	5	0	2	2	1	

Level matrix of enablers

Variable	Level
E9,E8	level 1
E5,E6,E7	level 2
E2,E3,E10	level 3
E4,E1	level 4

最终的到我们所需的 TISM 模型。



3. 数学模型的构建

在上一步中得到分析系统的 TISM 模型后，我们需要着手把各因子间的相互影响权重确定下来，为此我们可以运用层次分析法，通过 TISM 中的影响传递关系和传递距离，确定初始权重，然后便可通过层次分析法得出最终的因子间相互影响的权重。

Final binary matrix-enablers

0	E10	E9	E8	E7	E6	E5	E4	E3	E2	E1
E1	0.14	0.04	0.06	0	0.12	0.08	0	0.5	0.5	-
E2	0.43	0.08	0	0	0	0	0	0	-	0
E3	0.43	0.08	0.12	0.18	0.25	0.17	0	-	0	0
E4	0	0.04	0.06	0.09	0.12	0.08	-	0.5	0.5	1
E5	0	0.15	0.25	0.36	0	-	0	0	0	0
E6	0	0.15	0.25	0.36	-	0.33	0	0	0	0
E7	0	0.15	0.25	-	0.5	0.33	0	0	0	0
E8	0	0.31	-	0	0	0	0	0	0	0
E9	0	-	0	0	0	0	0	0	0	0
E10	-	0	0	0	0	0	0	0	0	0

4. 财务预测

财务预测在很大程度上决定了估值的准确度和最终评级结果的好坏，而 Jeffrey C .Hooke 在书中建议我们通过时间序列的方法并综合因果关系判断对未来的财务报表进行预测。

因此我们先通过时间序列方法的 ARMA 模型对历史财务报表中净利润，自由现金流，股东权益等科目进行预测，然后再通过 TISM 模型的影响因子对预测模型进行修正，以求尽可能准确的预测出来年的财务数据。

5. 估值

Jeffrey C .Hooke 在书中提供了五种不同的估值方法，分别为内在价值法，相对价值法，收购价值法，杠杆收购法和技术法。在项目中，由于收购价值的数据经常缺失，因此我们只使用内在价值法，相对价值法和技术法，三种方法综合对企业未来的股价进行估值。

其中内在价值法需要通过财务预测得到来年的自由现金流和股息的年化增长率，并且通过无风险利率和权益风险溢价计算理想的回报率。

$$P = \frac{D}{k - g}$$

其中 P 为企业内在价值，D 为企业的自由现金流，k 为股东要求的年化回报率，g 为每年的增长率。

相对价值法通过企业所在行业的平均市盈率和平均市净率计算出企业相对于同行业的企业的合理价值。每股市价=平均市盈率 x 每股净利润或平均市净率 x 每股净资产。

技术法，即通过研究先前的股价交易模式，预测出股票未来的价格，时间序列法便是一直技术法，我们也将通过该方法对股价进行预测。

在获得估值后，判断估值与现价的偏离率情况，若偏离率大于等于安全边际，则股票的评级为买入或卖出，若偏离率大于等于 1/2 安全边际但小于安全边际，则股票的评级为增持或减持，其余为中性评级。

五、 实证检验

按照上一节中的系统结构设计，我们用 python 实现了一个初步的证券分析系统，系统需要使用者自行给出 TISM 中的部分影响因子的数值，其数值分布在-2~2 之间，代表着该因子的投资评级，这要求使用者有一定的金融知识和研报阅读能力。同时还需提供企业近几年的历史财务报表，以进行未来的财务预测。

我们选取了上证市场的前十二个股票进行分析，通过其 13 年至 17 年的财务报表数据，我们对各公司今年的财务数据进行了预测，并进行估值。估值结果在下表中。

股票代码	名称	17年末价格	当前价格	涨跌幅度	估值价格	偏差率	评级
000001	上证指数	3296.38	2596.84	-21.22%			
600000	浦发银行	12.59	10.49	-16.68%	13.48	7.07%	增持
600004	白云机场	14.7	10	-31.97%	7.7	-47.62%	卖出
600006	东风汽车	5.85	3.9	-33.33%	5.54	-5.3%	中性
600007	中国国贸	17.14	14.45	-15.69%	13.71	-20.01%	卖出
600008	首创股份	5.14	3.54	-31.13%	4.96	-3.5%	中性
600009	上海机场	45.01	48.52	7.79%	43.79	-2.71%	中性
600010	包钢股份	2.46	1.58	-35.77%	1.64	-33.33%	卖出
600011	华能国际	6.17	6.7	8.59%	6.98	13.13%	增持
600012	皖通高速	11	5.87	-46.64%	10.05	-8.64%	减持
600015	华夏银行	9	7.79	-13.44%	10.85	20.56%	买入
600016	民生银行	8.39	6.09	-27.41%	8.49	1.19%	中性

其中评级为正面的股票的平均涨跌幅度为-7.17%，明显优于上证指数，评级为负面的股票平均涨跌幅度为-32.52%，表现不及上证指数，而评级为中性的股票平均涨跌幅度为-21.02%，可见对于上述 11 个股票，该分析系统具备一定的价值分析能力，根据系统评级投资将损失更小。

六、 机器学习的应用

在前文中我们已经了解了什么数据会对一只股票的评级有比较重要的影响，当我们获取到这些数据后，我们可以使用机器学习中的分类模型根据以往的分类经验来对当前的股票进行分类评级。

1. 数据获取

项目中所使用的数据来自东方财富网的 Choice 金融终端。根据前文中的 TISM 模型，我们选用了包括年成交额，总市值，市盈率，市净率，每股自由现金流。。。共 18 组数据。囊括了沪深两市 2000 余只个股从 2013 年到 2018 年共 5 年的数据。

2. 数据预处理

在获取数据后，我们对所有数据进行了标记，首先找出个股和上证指数下一年的涨跌幅，然后比较其中的差值。

对于 SVM 数据，若差值为负则标记该样本为-1，若为正则标记该样本为 1。

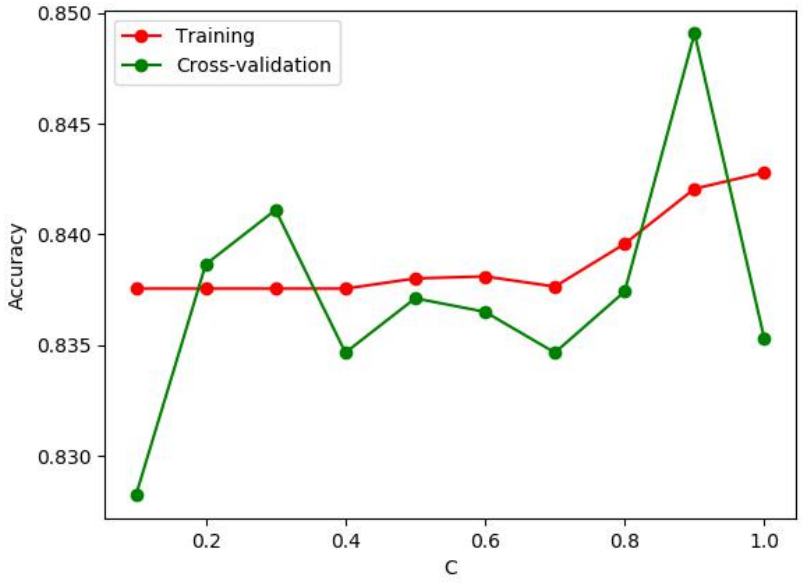
对于 MLP 数据，若差值为 60 以上则标记为 2，若差值为 60 至 30 以上则标记为 1，若差值为 30 至-30 则标记为 0，若差值为-30 至-60 则标记为-1，若样本为-60 以下则标记为-2.

3. 模型训练与参数选择

我们使用了 python 的 sklearn 工具包来帮助我们构建与训练分类模型，在本项目中我们分别用 SVM 和 MLP 两种不同的分类模型对历年的股票进行了分类，并使用下一年的数据对对预测结果进行了检验。

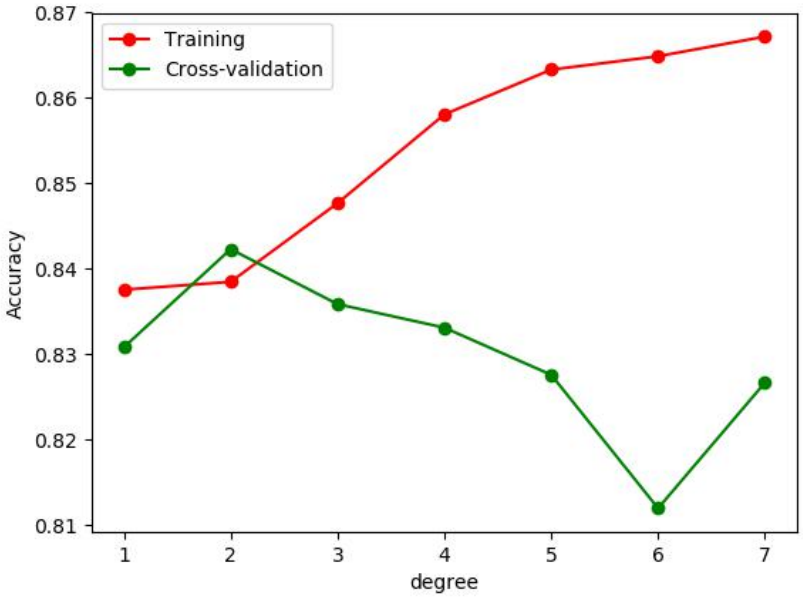
我们希望通过训练能得到一个通过个股现在的数据预测个股未来评级的模型，为此我们需要以个股 2017 年的财务数据为训练样本，以个股 2018 年的表现作为样本标记，进而得到一个通过 2017 年财务数据预测 2018 年市场表现的分类模型，并期望这个模型在 2018 年时仍然有一定的准确度，以便于我们能根据个股 2018 年的财务数据通过该分类模型，对个股 2019 年的表现进行预测分类。

在训练过程中我们使用交叉验证以确定能使模型准确度最高的参数。



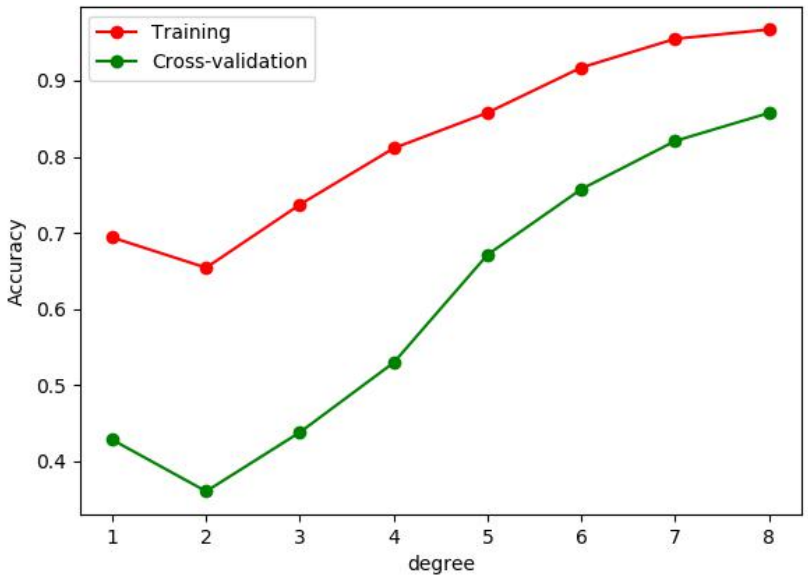
SVM 模型惩罚系数

由图中所见，当 SVM 模型的惩罚系数为 0.9 左右时，模型的准确度最高，当惩罚系数接近 1 时，模型出现过拟合，准确度下降。因此我们选择 0.9 作为模型的惩罚系数。



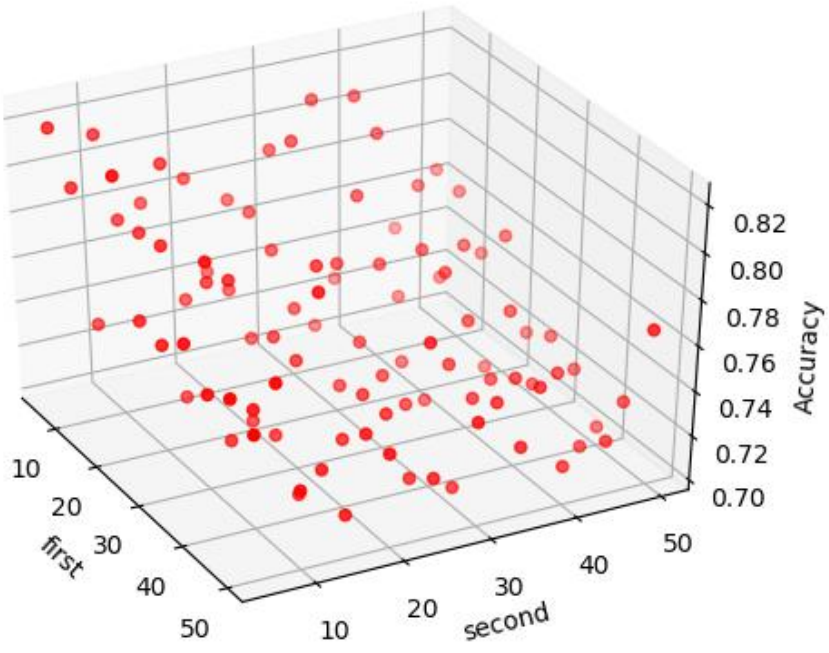
SVM-ploy 模型多项式级数

由图中所见当 SVM-ploy 模型的多项式级数为 2 左右时，模型的准确度最高，当多项式级数大于 2 时，模型出现过拟合，准确度下降。因此我们选择 2 作为模型的多项式级数。



MLP 模型分类距离

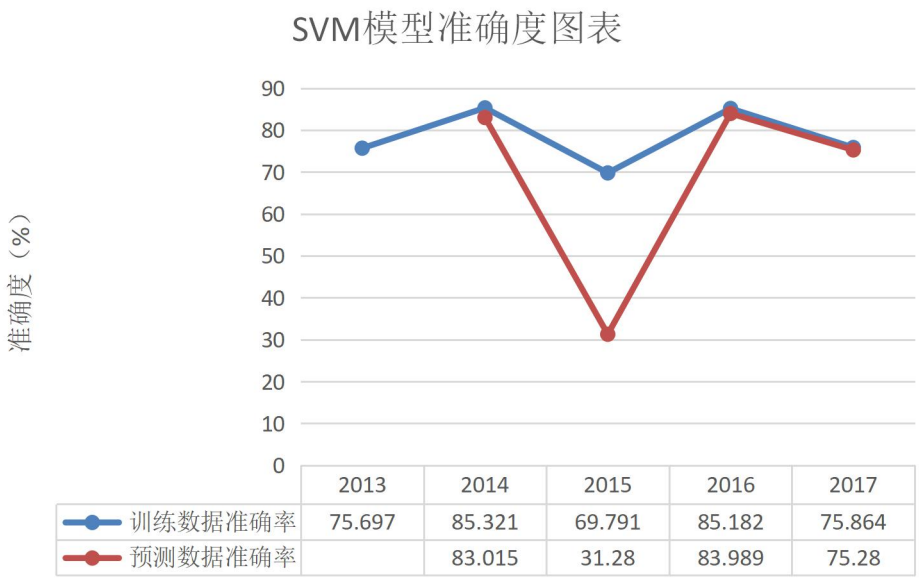
由图中所见，当分类的涨跌幅差越大时（图中 8 表示当涨跌幅差大于 80%时，样本被标记为 2 类），MLP 模型的分类准确度越高，但由于涨跌幅差距越大，样本就越集中，因此我们选择与 SVM 模型准确度相近的 60%作为分类标准。



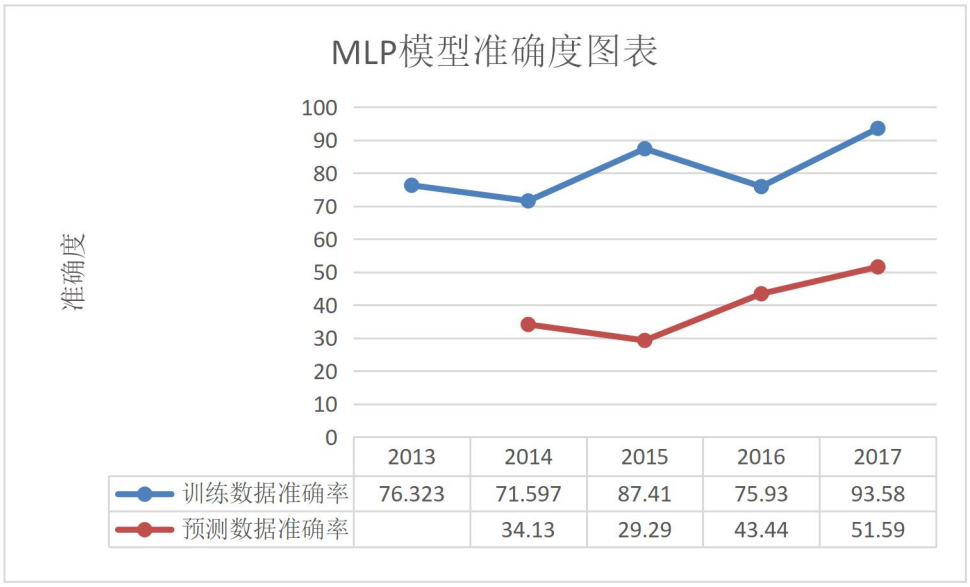
MLP 模型宽度

Lippmann 在 1987 年的论文“An introduction to computing with neural nets ”中有一个理论发现，它表明具有两个隐藏层的 MLP 足以创建任何所需形状的分类区域，因此我们设定 MLP 模型的层数为 2。由图中所见，当两层的 MLP 模型的宽度分别为 20 和 5 时，MLP 模型的分类准确度越高，因此我们设定与 MLP 模型的宽度为 20,和 5。

4. **模型检验**

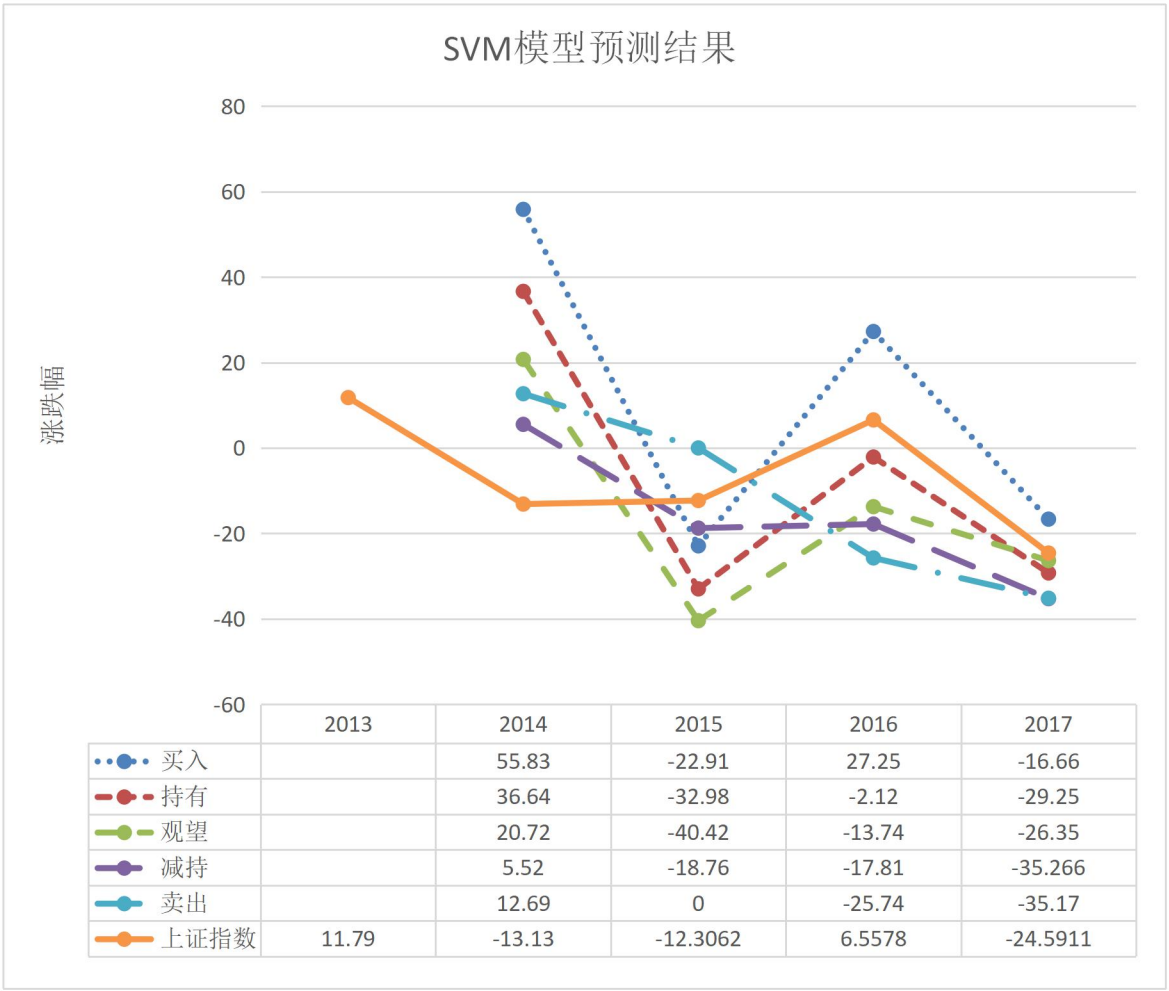


上图为 SVM 模型的训练结果，可见 SVM 模型对于证券分类有着很好的准确度，除了 2015 年由于牛市的原因模型无法准确预测，其余的年份，模型对于下一年个股与大盘涨跌幅的差异判断都能有 70%至 80%的准确度。



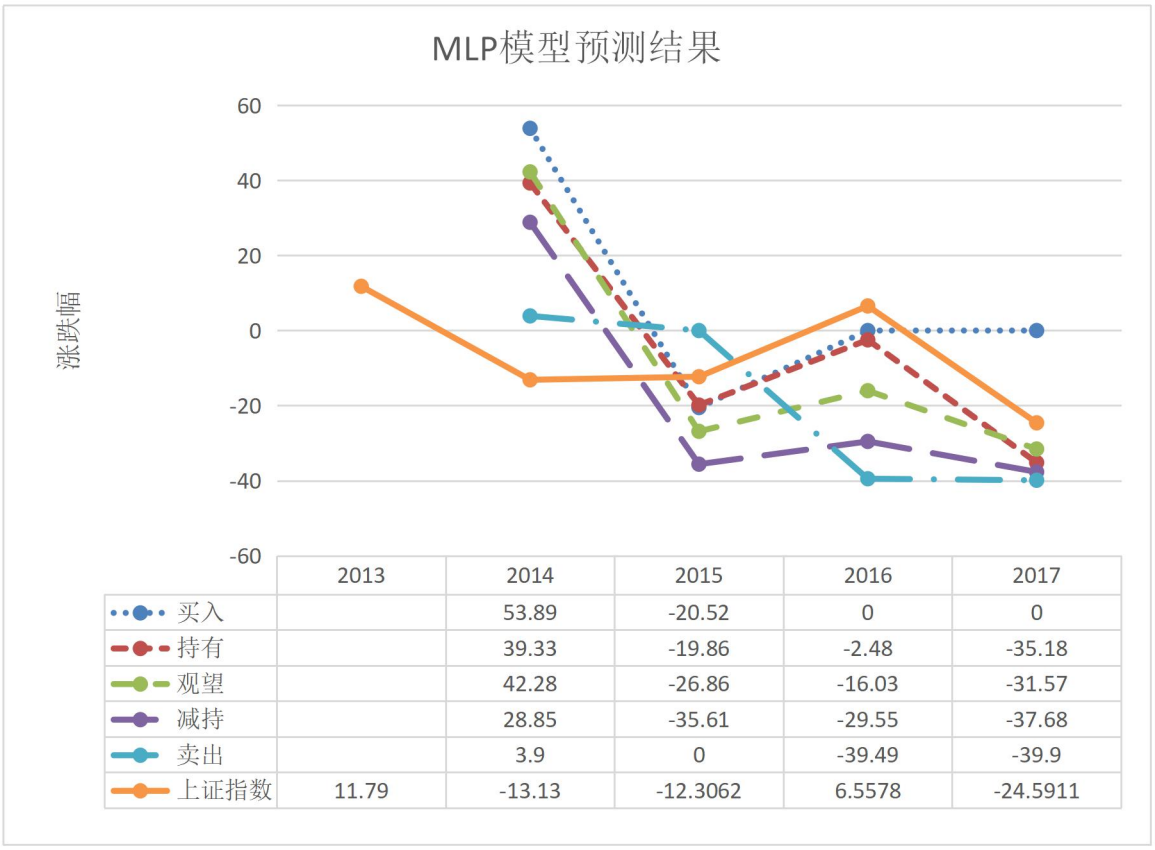
上图为 MLP 模型的训练结果，可见与 SVM 模型相比，MLP 模型的预测准确度较低，但训练的准确度相当。这其中主要原因是 MLP 模型需要把样本分

成 5 个类别，而 SVM 模型只需要分成 2 个。



	2013	2014	2015	2016	2017
买入		55.83	-22.91	27.25	-16.66
持有		36.64	-32.98	-2.12	-29.25
观望		20.72	-40.42	-13.74	-26.35
减持		5.52	-18.76	-17.81	-35.266
卖出		12.69	0	-25.74	-35.17
上证指数	11.79	-13.13	-12.3062	6.5578	-24.5911

上图为 SVM 模型的预测结果图，我们分别通过，2014，2015，2016，2017 年的数据分别对个股 2015,2016,2017 及 2018 年表现进行了预测分类，而其中 5 个不同类别的平均涨跌幅如图中所示。总体来说，分类模型对沪深两市的个股进行了有效分类，不同级别的个股涨跌幅有较大差异，而其中买入级别的个股的总体表现也要优于大盘，具有一定的投资参考价值。



	2013	2014	2015	2016	2017
买入		53.89	-20.52	0	0
持有		39.33	-19.86	-2.48	-35.18
观望		42.28	-26.86	-16.03	-31.57
减持		28.85	-35.61	-29.55	-37.68
卖出		3.9	0	-39.49	-39.9
上证指数	11.79	-13.13	-12.3062	6.5578	-24.5911

上图为 MLP 模型的预测结果图，可见该模型的分类效果要比 SVM 模型差很多，不同级别的个股涨跌幅差异不明显，分类界限模糊，而其中买入级别的个股的总体表现也无法战胜大盘，没有太大的投资参考价值。

七、 存在的问题

目前的分析系统还不完善，模型的还没有经过严谨的验证，模型的部分影响因子需要使用者依靠自己的主观经验给出，而这些因子很大程度上决定了评级的结果，使得分析系统的结果存在很大的差异性。同时真正的分析人员在分析不同的企业时还会考虑许多特异性因子，使得分析过程更加复杂，这些都是在本分析系统中所未能实现，也是我们需要继续思考的方向。而其中的机器学习模型的准确性还可以继续提高，还可以尝试多种不同的分类模型，初始样本的标签也仍需要更加深刻的考虑。

八、 Reference

[1] Jeffrey C. Hooke, Security Analysis and Business Valuation on Wall Street: A Comprehensive Guide to Today's Valuation Methods, 2 edition, 2010.

[2] Benjamin Graham, David Dodd, Security Analysis, 6 edition, 1934.

[3] Dubey, R., Gunasekaran, A., Sushil, & Singh, T. (2015). Building theory of sustainable manufacturing using total interpretive structural modelling. International Journal of Systems Science: Operations & Logistics, 2(4), 231-247.

[4] Shibin, K. T., et al. "Frugal innovation for supply chain sustainability in SMEs: multi-method research design." *Production Planning & Control* 29.11 (2018): 908-927.

[5] 杨明.中国股市股票价值分析[J].现代商业,2008(23):26.

[6] 黄小康. 证券分析师投资评级信息含量及影响因素实证研究[D].浙江大学,2008.

[7] Lippmann R P . An introduction to computing with neural nets[M]. ACM, 1988.

[8] http://sklearn.apachecn.org/

Components	Comments	Max.	Marks awarded
Report content		30	
Report structure		30	
Project management		20	
Communications		20	

The marks are given and signed by the Supervisor

Student Name : Total Marks out of 100:

Student Name : Total Marks out of 100:

Student Name : Total Marks out of 100:

Assessed by Supervisor:
(Signature)

The marks are given and signed by the Inspector

Student Name : Total Marks out of 100:

Student Name : Total Marks out of 100:

Student Name : Total Marks out of 100:

Assessed by Inspector:
(Signature)

Guideline for assessing project

1. Report content: the work, description of the work, difficulty of the work, how well the work was done, original objectives and how they have been met, etc.
2. Report structure: clarity, correctness, completeness, including format, section structures, references, presentation, language, etc.
3. Project management: including independence, proactiveness, time management,

documentation, etc. (marked by the supervisor only)

4. Communications: project presentation (concise and clear, to the point, within the given time, etc.), weekly meetings with the supervisor, etc.