基金分类——基于文本数据挖掘

1. 引言

本文旨在提供一种文本数据挖掘方法以实现基金自动化分类。**分类任务共 4 个，包括投资方式维度一级、二级分类和投资标的维度一级、二级分类，详见表 1。**

表 1 基金分类体系

分类维度	一级分类	二级分类
投资方式	主动型	主动型
	指数型	被动指数型
		指数增强型
投资标的	QDII 型	QDII 股票型
		QDII 债券型
		QDII 混合型
		QDII 另类型
	股票型	股票型
	债券型	纯债型
		混合债券型
		理财债券型
	货币型	货币型
	混合型	FOF
		量化对冲型
		偏债混合型
		偏股混合型
		平衡混合型
		灵活混合型
		其他混合型
	其他型	商品型
		REITS

2. 数据

2. 1 数据描述

本文选取了 12672 个基金包括’fund_name’，’investment_target’，’investment_scope’，’investment_strategy’，’risk_return_character’，’comparison_criterion’,’tracking_benchmark’共 7 个特征的文本表述。**相应的基金类型由人工标注，称为标签（type_name_x, stype_name_x, type_nmae_y, stype_nmae_y）。**

2. 2 数据处理

1. 对人工标注错误的标签进行了修正。

2. 删除了多个特征缺失的基金。

3. 分类原则

1. 人工校核成本尽可能低（准确率尽可能高，判别率尽可能高）。

2. 大样本采用机器学习判别，小样本利用关键词提取。

3. 区分度低采用机器学习判别，区分度高利用关键词提取（逻辑复杂或模糊）。

4. 模型方法

考虑到关键词匹配和机器学习的优缺点（见表 2），单一方法的分类错误较高，本文通过组合两种方法对基金进行分类。图 1 显示了两种方法的基金分类流程。

表 2 关键词匹配和机器学习的优缺点

方法	优点	缺点
基于关键词匹配	高效，计算成本小	匹配逻辑的设计复杂； 泛化能力差
基于机器学习模型判定	处理复杂逻辑的能力； 泛化能力； 具有概率输出能力	训练较复杂； 计算成本大；

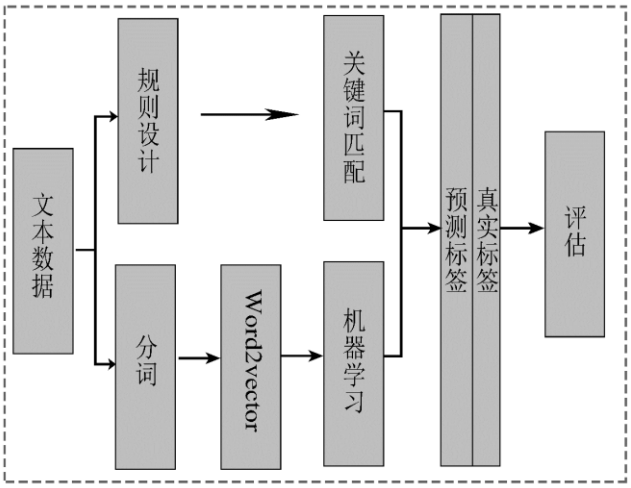

图 1 关键词匹配和机器学习分类流程

具体地说，对于某一维度的基金一级分类，本文的一般分类流程如下：首先根据由关键词设定的判定规则优先确定基金类别，进而对剩余基金进行机器学习判定。对于可直接应用机器学习的分类任务，则无需进行关键词匹配。二级分类由一级分类结果决定，分类顺序与一级分类相似。更多细节见附录 A、B。

4.1 关键词匹配

关键词匹配方法利用某一类型基金的专有标志词、高频词或表述规则进行匹配以确定类别。关键词匹配包括以下细节：

1. 判别特征、字段设计（特征对的结果的影响较大）。
2. 判别顺序设计（特征差异）。
3. 判别逻辑设计（关键词类间尽可能独立，区分度尽可能高）。

4.2 机器学习

机器学习对基金进行分类的主要包括 5 个步骤：

1. 利用分词模型将文本分成若干词，得到每个基金对应的词列表；
2. 利用了 word2vec 模型对金融领域语料库进行训练。考虑到时间成本和模型构建的复杂性，本文利用了 https://github.com/Embedding/Chinese-Word-Vectors 已训练得到的 50 万词向量库，该词库对本项目的词覆盖率为 80%；
3. 利用 word2vec 模型将词转化为向量，并平均该基金的所有词向量得到一个 300 维词向量；
4. 利用 PCA 降维；
5. 结合相应标签利用模型训练和测试。

训练和测试过程如下：

为避免类不平衡引起的模型训练，本文从每个基金类别中选取相同数量的样本用于训练，其他样本作为测试；为避免训练样本的随机性，本文重复 30 次训练，最终测试结果通过平均得到。另外本文测试了常用的 21 个分类模型，最终选用的模型为 ExtraTrees。

4.2.1 分词

由于中文词具有无边界性，因此相对于英文分词，中文分词更复杂。本文应用结巴中文文本分词以保证下一步的词向量映射。结巴分词基于概率语言模型，它通过一个预先准备好的词典生成一棵 trie 树，并且计算字典中的词的词频。当处理一个需要分词的句子的时候，它会生成一个 DAG（有向无环图）来记录每一种可能的分词的情况。这样一个 DAG 就是一个字典，字典的键是一个词起始的位置，而字典的值则是由一个词可能的结尾的位置组成的列表。

对于 DAG 中每种可能的分词，需要基于预先准备好的词典计算概率。然后结巴会通过从右向左的方向计算概率最大的路径。这一条概率最大的路径为最大可能性的分词。对于句子中出现词典里没有的词的情况，结巴采用 HMM（隐马尔科夫模型）和 Viterbi 算法（维特比算法）来进行分词。

对于不在词典中的字，它们会有四种可能的状态：B（起始字），M（中间字），E（结束字）和 S（单字）。这些状态表明了这个字在词中的位置。而分词的过程主要就是基于这些状态。结巴分词的作者通过对大量文本的训练得到了三个概率表，并利用 Viterbi 算法来计算一个字最有可能的状态，并依据这种一个句子每个词的状态构成的状态链来进行分词。

4. 2. 2 Word2Vector

Word2Vector 是 Google 在 2013 年推出的一个 NLP 工具，它的特点是将所有的词向量化，进而度量词之间的关系，挖掘词之间的联系。Word2Vector 包括 CBOW（Continuous Bag-of-Words 和 Skip-Gram 两种模型，前者根据中心词周围的上下文单词来预测该词的词向量，后者则根据中心词预测周围上下文的词的概率分布。本文选用的词向量基于 Skip-Gram 模型训练得到, 图 2 显示了 Skip-Gram 模型结构。

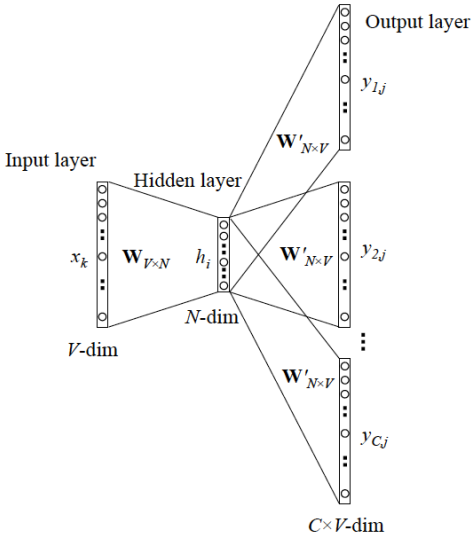

图 2 Skip-Gram 模型结构

假设词典索引集D的大小为$|D|$，且记 $D=\{1,2,\ldots|D|\}$。给定一个长度为T的文本序列，其中第t个词记为$w^{(t)}$. 当窗口大小为m时，Skip-Gram 模型要求最大化任一中心词生成距离不超过m个词的背景词的总概率

$$\prod_{t=1}^{T}\prod_{-m\le j\le m,j\neq 0,1\le t+j\le|T|}P\left(w^{(t+j)}\middle|w^{(t)}\right)$$

相应的似然函数为，

$$\sum_{t=1}^{T}\sum_{-m\le j\le m,j\neq 0,1\le t+j\le|T|} logP\left(w^{(t+j)}\middle|w^{(t)}\right)$$

最大似然函数等价于最小化损失函数，

$$-\frac{1}{T}\sum_{t=1}^{T}\sum_{-m\le j\le m,j\neq 0,1\le t+j\le|T|} logP\left(w^{(t+j)}\middle|w^{(t)}\right)$$

$\boldsymbol{V}$和$\boldsymbol{U}$分别表示中心词和背景词的向量，也就是说，对于索引为i的词，它作为中心词和背景词时的向量表示分别为$\boldsymbol{v_i}$和$\boldsymbol{u_i}$。而要训练的模型参数就是词典中所有词的这两种向量。为了将模型参数植入损失函数，需要使用模型参数表达损失函数中的给定中心词生成背景词的条件概率。给定中心词，假设生成各个背景词是相互独立的，那么对于中心词w_c，背景词w_b, b, c 为这两个词在词典中的索引。那么给定中心词w_b生成背景词 w_c的条件概率可以通过 softmax 函数定义为

$$P\left(w^{(t+j)}\middle|w^{(t)}\right)=\frac{exp\left(\boldsymbol{u_b^T v_c}\right)}{\sum exp(\boldsymbol{u_i^T v_c})}$$

通过微分可以得到上述条件概率的梯度

$$\frac{\partial logP\left(w^{(t+j)}\middle|w^{(t)}\right)}{\partial v_c}=\boldsymbol{u_b}-\sum_{j\in D}\frac{exp(\boldsymbol{u_b^T v_c})}{\sum exp(\boldsymbol{u_i^T v_c})}\boldsymbol{u_j}$$

利用梯度下降法或随机梯度下降法来进行迭代求解，最终求得使得损失函数最小时，词典中所有词的中心词和背景词的词向量$\boldsymbol{v_i}$和$\boldsymbol{u_i}$，$i=1,2,\ldots|D|$。

4. 2. 3 Extremely randomized trees（ExtraTrees）

本文测试了常用的 20 个分类模型，多个分类任务的结果表明，AdaBoost、Extra−Trees 等基于决策树的模型具有较好的表现，图 3 显示了偏股、偏债二分类任务中验证集的评估结果。本文选用 Extra−Trees 作为最终的分类器，该模型与随机森林十分相似，均由若干决策树构成。前者相较于随机森林的主要区别在于：

1. RandomForest 应用的是 Bagging 模型，ExtraTrees 使用的所有的样本，只是特征是随机选取的，因为分裂是随机的，所以在某种程度上比随机森林得到的结果更加好。

2. 随机森林是在一个随机子集内得到最佳分叉属性，而 ExtraTrees 是完全随机的得到分叉值，从而实现对决策树进行分叉的。

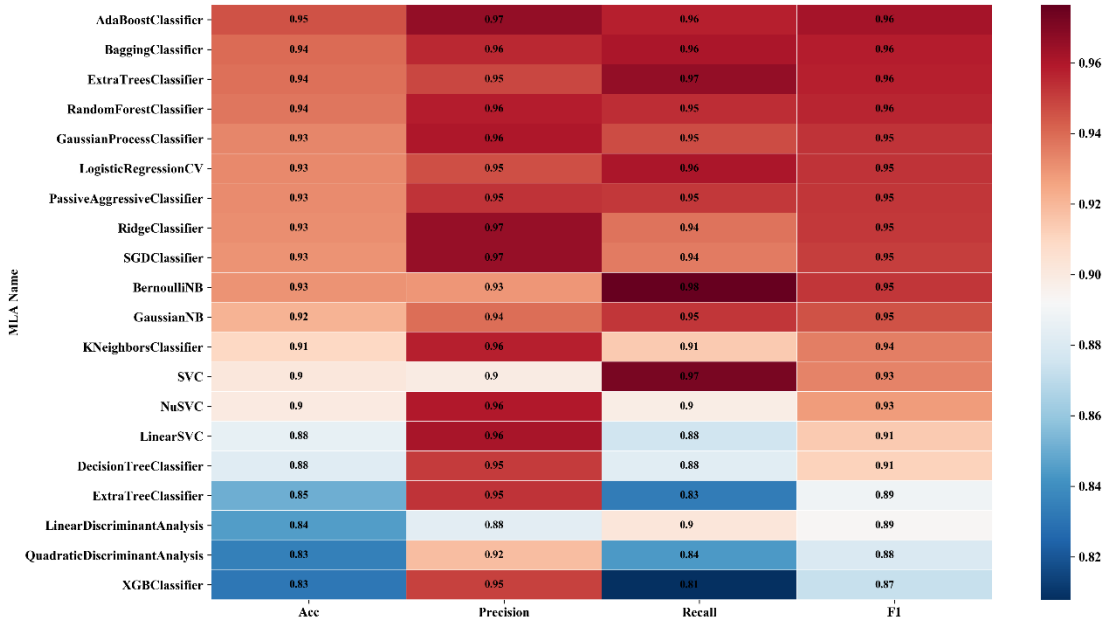

图 3 偏股、偏债二分类任务中分类器的验证集评估结果

5. 评价指标

TP（True Positive），**TN**（True Positive）：预测答案正确；**FP**（False Positive）：错将其他类预测为本类；**FN**（False Negative）：本类标签预测为其他类标。

1. 准确率（Accuracy）：分类正确的样本占总样本的比例。

$$acc = \frac{TP + TN}{TP + TN + FP + FN}$$

2. 精确率（Precision）：对于某一实际类别，预测正确的样本占比。

$$precision = \frac{TP}{TP + FP}$$

3. 召回率（Recall）：对于某一预测类别，预测正确的样本占比。

$$recall = \frac{TP}{TP + FN}$$

4. F1−score：兼顾精确率和召回率，两者的调和平均。

$$F1_{score} = \frac{2precision \times recall}{precision + recall}$$

5. 混淆矩阵。

6. 投资方式评估

6. 1 一级分类评估

表 3 投资方式一级分类混淆矩阵

	主动型	指数型
主动型, 9534	9521	13
指数型, 1863	4	1859

表 4 投资方式一级分类细分指标

	Precision	**Recall**	**F1**
主动型	0. 99958	0. 998636	0. 999108
指数型	0. 993056	0. 997853	0. 995448

表 5 投资方式一级分类总指标

	Indicator
Acc	0. 998508
Precision	0. 993056
Recall	0. 997853
F1	0. 995448

6. 2 二级分类评估

表 6 投资方式二级分类混淆矩阵

	主动型	指数增强型	被动指数型
主动型, 9534	9521	2	11
指数增强型, 187	2	177	8
被动指数型, 1676	2	6	1668

表 7 投资方式二级分类细分指标

	Precision	**Recall**	**F1**
主动型	0. 99958	0. 998636	0. 999108
指数增强型	0. 956757	0. 946524	0. 951613
被动指数型	0. 988737	0. 995227	0. 991971

表 8 投资方式二级分类总指标

	Indicator
Acc	0. 99728
Precision	0. 997283
Recall	0. 99728
F1	0. 997279

7. 投资标的评估

7. 1 二级分类评估

表 9 投资标的一级分类混淆矩阵

	QDII 基金	债券型基金	其他基金	混合型基金	股票型基金	货币型基金
QDII 基金, 336	320	0	1	7	8	0
债券型基金, 3921	0	3920	0	1	0	0
其他基金, 40	0	0	40	0	0	0
混合型基金, 4417	4	0	0	4408	4	1
股票型基金, 1947	26	0	3	1	1917	0
货币型基金, 731	0	51	0	0	0	680

表 10 投资标的一级分类细分指标

	Precision	Recall	F1
QDII 基金	0. 914286	0. 952381	0. 932945
债券型基金	0. 987157	0. 999745	0. 993411
其他基金	0. 909091	1	0. 952381
混合型基金	0. 997962	0. 997962	0. 997962
股票型基金	0. 993779	0. 984592	0. 989164
货币型基金	0. 998532	0. 930233	0. 963173

表 11 投资标的一级分类总指标

	Indicator
Acc	0. 990607
Precision	0. 990785
Recall	0. 990607
F1	0. 990582

7. 2 二级分类评估

表 12 投资标的二级分类混淆矩阵

	FOF	QDII 债券型基金	QDII 另类投资基金	QDII 混合型基金	QDII 股票型基金	偏债混合型	偏股混合型	商品型基金	平衡混合型	混合债券型	灵活配置型	理财债券型	纯债型	股票型	货币型	量化对冲型
FOF, 156	155	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
QDII 债券型基金, 98	0	98	0	0	0	0	0	0	0	0	0	0	0	0	0	0
QDII 另类投资基金, 23	0	0	9	9	4	0	0	1	0	0	0	0	0	0	0	0
QDII 混合型基金, 55	1	0	1	47	2	0	3	0	0	0	1	0	0	0	0	0
QDII 股票型基金, 160	1	0	3	1	146	0	0	0	1	0	0	0	0	8	0	0
偏债混合型，618	0	0	0	0	0	615	3	0	0	0	0	0	0	0	0	0
偏股混合型，1268	0	0	0	0	0	12	1249	0	2	0	0	0	0	3	1	1
商品型基金，40	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0
平衡混合型，29	0	0	0	0	0	15	11	0	3	0	0	0	0	0	0	0
混合债券型，921	0	0	0	0	0	0	0	0	0	920	0	0	1	0	0	0

灵活配置型，2316	0	0	0	1	0	25	49	0	4	0	2235	0	0	0	0	2
理财债券型，62	0	0	0	0	0	0	0	0	0	0	0	61	1	0	0	0
纯债型, 2938	0	0	0	0	0	0	1	0	0	13	0	0	2924	0	0	0
股票型, 1947	0	0	0	0	26	1	0	3	0	0	0	0	0	1917	0	0
货币型，731	0	0	0	0	0	0	0	0	0	0	0	51	0	0	680	0
量化对冲型，30	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	27

表 13 投资标的二级分类细分指标

	Precision	Recall	F1
FOF	0. 987261	0. 99359	0. 990415
QDII 债券型基金	1	1	1
QDII 另类投资基金	0. 692308	0. 391304	0. 5
QDII 混合型基金	0. 770492	0. 854545	0. 810345
QDII 股票型基金	0. 820225	0. 9125	0. 863905
偏债混合型	0. 920659	0. 995146	0. 956454
偏股混合型	0. 949088	0. 985016	0. 966718
商品型基金	0. 909091	1	0. 952381
平衡混合型	0. 3	0. 103448	0. 153846
混合债券型	0. 986066	0. 998914	0. 992449
灵活配置型	0. 999553	0. 965026	0. 981986
理财债券型	0. 544643	0. 983871	0. 701149
纯债型	0. 999316	0. 995235	0. 997271
股票型	0. 993779	0. 984592	0. 989164
货币型	0. 998532	0. 930233	0. 963173
量化对冲型	0. 9	0. 9	0. 9

表 14 投资标的二级分类总指标

	Indicator
Acc	0. 97665
Precision	0. 978207
Recall	0. 97665
F1	0. 976563

A. 投资方式分类流程

A. 1 一级分类流程

1. 机器学习判定（主动型、指数型二分类模型）。

A. 2 二级分类流程

A. 2. 1 指数型二级分类流程

1. 关键词提取指数增强型；

2. 剩余样本为被动指数型。

B. 投资标的分类流程

B. 1 一级分类流程

1. 关键词提取 QDII 型；

2. 关键词提取其他型（尽管其他类型存在二级分类，但由于 REITS 类型仅包含一个样本，本文对其他型二级分类不做区分）；

3. 机器学习判定（股票、债券、货币、混合型四分类模型）。

B. 2 二级分类流程

B. 2. 1 QDII 二级分类流程

1. 关键词匹配判定（由于仅利用关键词匹配方法，因此在 QDII 二级分类中存在不能判别的样本）。

B. 2. 2 债券型二级分类流程

1. 关键词提取理财债券型；

2. 机器学习判定（纯债型、混合债券型二分类模型）。

B. 2. 3 混合型二级分类流程

1. 关键词提取 FOF；

2. 关键词提取量化对冲型；

3. 关键词提取灵活配置型；

4. 关键词提取平衡混合型；

5. 机器学习判定（偏债偏股二分类模型）。