【基金业数字化转型专题】建信基金：深度学习辅助设置客户标签

【编者按】为深入贯彻落实党的二十大精神，引导基金行业机构践行《证券期货业科技发展“十四五”规划》，共促基金行业数字化转型，按照中国证监会总体工作部署，于2022年11月开展“证券期货业数字化转型主题宣传月”活动。通过开展“证券期货业数字化转型主题宣传月”活动，搭建交流平台，展现数字化转型成果案例，激发金融科技创新活力，营造金融科技长效发展新生态。该篇为“证券期货业数字化转型主题宣传月”系列宣传之十八。

深度学习辅助设置客户标签
——建信基金

一、背景

为了更好地服务客户，维护金融业的安全与稳定，基金行业监管愈趋严格，数据报送工作也越来越重要。建信基金作为金融行业内首批成立的银行系基金公司，除完成证监会要求的FISP报送、CISP资管报送、基金申赎及基金投资人结构日报表，基金业协会要求的资管业务运行月报、场外债券投资交易明细表、公募资金来源表等基金行业监管相关的报送工作外，还需要完成人民银行要求的金融机构资管产品数据报送、银行业金融机构黄金市场业务监测表、人民银行系统重要性，银保监会要求的综合化经营自查附属机构内部交易报送等银行业监管相关报送。其中证监会FISP报送、人民银行金融机构资管产品数据报送、人民银行系统重要性、基金业协会公募资金来源表、附属机构内部交易报送工作，需要业务人员每日手动填充新注册的对公客户的7项属性标签（以下简称“打标签”）。

为完成打标签工作，业务人员投入了越来越多的精力。一方面，业务人员需要结合业务经验才能正确填写标签，另一方面，有些属性包含的标签类型数量较大，如“FISP 分类”有 40 个以上标签可选项，需要业务人员查找与比对之后才能选出正确的一项。随着公司业务的不断发展，新注册的对公客户数量越来越多，业务人员手动填写标签的工作压力不断增大。

2021 年《证券期货业科技发展“十四五”规划》（以下简称《“十四五”规划》）正式发布[1]，强调了“推进行业数字化转型发展”与“数据让监管更加智慧”两大主题。建信基金从数字化经营角度对监管报送业务中的打标签工作进行了分析，尝试通过人工智能算法对标签进行预填充，减少业务人员在填写标签上花费的时间和精力，提升员工工作效率和工作体验。

对新注册客户打标签，本质是把客户分配到该标签对应类别中。通过建立机器学习模型，可以对客户的类别进行预测，并完成自动填写，供业务人员审核或修正。建信基金算法团队在对两种深度学习算法——文本卷积神经网络[2][3]和基于注意力机制的双向长短期记忆网络[4]进行探索后，对报送工作中需人工填写的 7 个属性进行了预测，均获得了极高的准确率，已在建信基金统一报送平台上线使用。

二、两种深度学习模型

下面简介两种深度学习模型的基本原理。

（一）TextCNN 模型

在短文本分类领域常用文本卷积神经网络（后面简称 TextCNN）来完成分类任务。参考句子分类的卷积神经网络 TextCNN 网络结构[2] ，本文模型如图 1 所示。

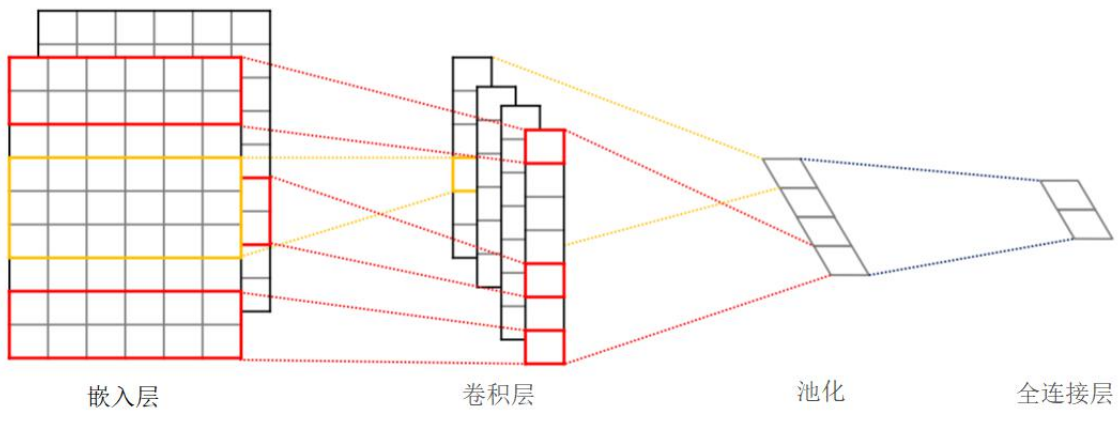

图 1 TextCNN 网络

该网络结构主要包括嵌入层、卷积层、池化、全连接层四部分。

TextCNN 先使用预训练的词向量作为嵌入层,然后在卷积层使用一维卷积提取特征，再通过池化函数捕获最重要的特征，在全连接层建立特征到类别的全连接，将输出结果进行归一化转换后，可得到每个类别标签的概率。

（二）Bi-LSTM + Attention 模型

长短期记忆网络 LSTM，是一种循环神经网络模型。双向长短期记忆网络 Bi-LSTM 能更好地捕获句子中上下文的信息。而基于注意力机制的双向长短期记忆网络（以下称为 Bi-LSTM + Attention）在关系分类[4]的实验中获得比较显著的效果。本文选用的 Bi-LSTM + Attention 模型如图 2 所示。

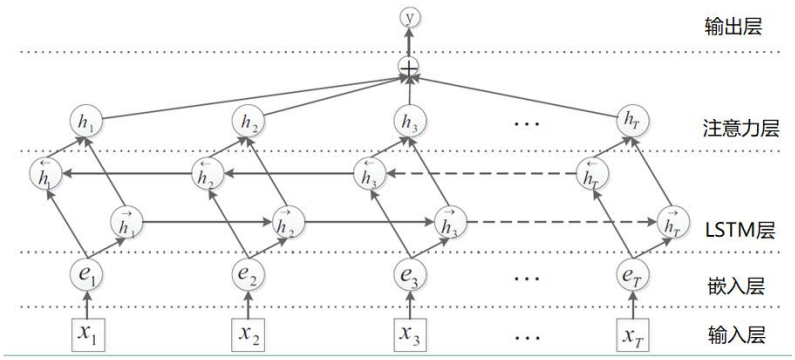

图 2　Bi-LSTM+Attention 模型

三、业务分析与模型构建

对公客户在基金公司注册成功后，就成为基金公司的新客户（以下简称“客户”)。根据监管要求,每一个新客户的加入,在报送时需对其补充“客户类型”、“FISP 分类”、“人行类型”、“内部交易类型”、“资金来源”“公募来源投向”、“人行重要性分类”共 7 个属性标签，而其中每一属性的填充，均需要从监管要求的标准标签集中选出一个标签。

比如，注册名称为“×银理财稳享固收精选 2 个月定开 3 号理财产品”的客户，设置 “客户类型”属性为“产品客户”,“FISP 分类”属性为“银行子公司公募理财”,“人行类型”属性为“银行非保本理财”,“人行重要

性分类”属性为“特定目的载体”，“内部交易类型”属性为“非内部交易客户”，“资金来源”属性为“其他机构”，“公募来源投向”属性为 “除上述类型外的其他机构投资者”。而名称为“北京××有限公司”的客户，则设置“客户类型”属性为“机构客户”，“FISP 分类”属性为“境内非金融机构”标签，“人行类型”属性为“非金融企业”，“人行重要性分类”属性为“非金融企业”，这些属性标签与前者明显不同。

为实现智能化设置标签的目标，需要用模型预测出每个属性选择哪个标签是最适合的。通过对客户数据考察发现，客户名称是体现客户特点的核心因素，对于每个属性选择标签起重要作用。对客户名称建模，并从一个标准标签集中筛选出最合适的一个标签，填充到该客户某一属性上，是自然语言处理领域文本分类技术的一个典型应用场景。

所以，本文对客户 7 个属性分别建立了 7 个独立的分类子模型，预测各自对应的标签结果。每个模型训练与预测过程如图 3 所示。

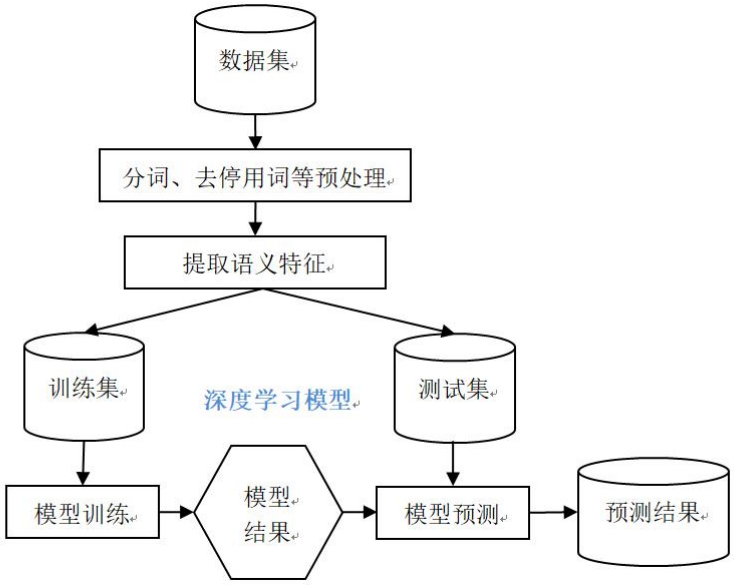

图 3 模型流程图

在算法处理过程中，首先使用开源分词工具结巴分词，把客户名称切成有意义的词条序列，并去除助词、标点等无意义词条。

下一步通过词嵌入模型 word2vec[5]进行语义抽取，每个词条的语义用同一维度的向量表示出来，客户名称就从词条序列转换成计算机可运算的语义数字矩阵。

最后用深度学习模型进行训练和预测。以 TextCNN 网络为例，前一步得到语义数字矩阵作为 TextCNN 网络的嵌入层输入数据，经过网络模型计算后，最终在全连接层后，预测出每个候选标签的概率，并选择概率最高的标签为某一属性最终的填充结果。

本文参考句子分类实验的模型结构和参数[3]，选择卷积过滤器窗口大小为 2、3、4、5，用以提取词之间多元语义信息。此外，卷积模式设计也重点考虑了文本首尾的边界特征的有效提取。比如名称为“××博时组合"的客户，因为“组合”是名称的结尾，则应标记为 “产品客户”，而名称为“北京××有限公司”的客户，是以“有限公司”为结尾的，则更可能归属于“机构客户”。

对 Bi-LSTM + Attention 模型也是类似的，客户名称的词条序列转化为语义向量序列 X1,X2,…，XT 作为输入层，经过模型运算后，预测出概率最高的标签作为最终结果。

四、实验结果

实验选用了 2019 年 6 月份到 2022 年 6 月份业务部门手工标记的客户标签数据，随机划分得到训练集约 4 万条和测试集约 1 万条数据。

在客户 7 个属性的预测任务上，两种模型的实验准确率如表 1 所示。

表 1 7 个属性的预测准确率

属性\模型	TextCNN(％)	Bi-LSTM + Attention(%)
客户类型	99.23	99.09
FISP 分类	91.44	91.20
人行类型	93.37	93.13
内部交易类型	99.62	99.52
资金来源	96.97	96.85
公募来源投向	95.96	95.81
人行重要性分类	94.89	94.57

TextCNN 模型和 LSTM+ATTENTION 模型都获得了高准确率的实验效果，而机器实测运行效率上，前者是后者的十倍以上，所以生产应用上更倾向

TextCNN 模型。

五、给公司带来的效益

（一）降低工作难度和工作量

在 TextCNN 模型上线之前，按照数据报送业务的操作流程，业务人员需要对新客户的 7 个属性依次进行手工选择，为每个属性选择正确的标签作为属性值。业务人员为每个属性选择标签时，主要依赖其个人经验作为判断依据，从一个候选的标签列表中寻找出正确的标签作为属性值，做出判断的难度较大，很多情况下需要依赖互联网等工具反复搜索和汇总信息后才能最终确认，操作过程枯燥乏味，且当信息不全时还有产生误选的风险。另外，由于纯手工操作的效率低，当出现新用户量暴增的情况时，打标签操作会对业务人员的工作负荷和数据报送业务的按时完成带来不小的压力。

为了解决这些问题，公司的数据报送业务流程进行了升级，加入了 TextCNN 模型辅助设置客户标签的功能。TextCNN 模型对新用户的 7 个属性都能预先给出高准确率的预测结果，自动设置属性标签，而业务人员的操作方式也从原来的手工选择标签，转变为标签的审核与修正，简化了操作步骤，并显著降低了业务人员的工作难度和工作量。

TextCNN 模型上线后，选择一个月的实际数据统计，如表 2 所示，各个属性均保持了稳定的高准确率，平均准确率约 97.85%，仅有不超过 3%的标签需要进行人工修正，业务人员的工作量得到了显著降低。

表 2 上线后 1 个月实测 7 个属性的准确率(%)

日期\属性	客户类型	FISP 标签	人行类型	内部交易类型	资金来源	公募来源投向	人行重要性
2022/9/1	100	93.88	96.94	100	98.98	97.96	94.9
2022/9/2	100	97.03	93.07	100	100	92.08	98.02
2022/9/5	97.56	91.46	95.12	100	98.78	95.12	97.56
2022/9/6	98.95	89.47	94.74	100	98.95	90.53	94.74
2022/9/7	100	95.05	97.03	100	98.02	89.11	96.04
2022/9/8	100	95.04	97.16	100	98.58	94.33	96.45
2022/9/9	99.7	96.1	99.4	100	99.85	98.65	99.55
2022/9/13	99.31	99.31	95.14	100	99.31	90.97	97.22
2022/9/14	100	96.12	100	100	100	96.12	99.03

2022/9/15	95.32	95.32	97.45	100	99.57	93.19	98.3
2022/9/16	100	96.94	97.38	100	100	98.69	99.56
2022/9/19	99.6	97.59	99.6	100	99.6	96.79	99.2
2022/9/20	100	90.41	97.26	100	100	93.84	96.58
2022/9/21	100	98.21	98.81	100	100	95.83	97.62
2022/9/22	98.41	93.65	98.94	100	100	95.24	97.88
2022/9/23	100	98.08	99.04	100	100	99.04	98.56
2022/9/26	100	97.16	96.59	100	100	98.3	96.59
2022/9/27	97.78	96.67	99.44	100	100	97.22	99.44
2022/9/28	100	99.22	99.61	100	100	95.69	99.61
2022/9/29	99.65	96.88	96.88	99.31	99.31	96.88	96.53
2022/9/30	98.27	97.11	95.95	100	100	95.38	98.84

（二）提升工作效率

由于深度学习模型已给出高准确率的预测结果，业务人员仅需关注在结果的审核和少量错误数据的修正上，节省了97%以上的操作时间，提升了工作效率，为数据报送业务每日按时顺利完成提供了强大的技术支持，并避免了新用户量大幅上涨可能引起人工操作时间暴涨的情况，降低了运营风险。

比如，根据TextCNN模型上线后实际操作日志统计，假设新用户每个属性点选标签平均需要4秒钟，则原本1小时以上的标签手工填写工作，可缩减为几分钟内完成。即使偶尔出现新客户暴涨的情况，业务人员的操作时间依然能够控制在很小的时间范围内，轻松完成，如表3所示。

表3 上线后1个月实测人工操作成本对比

日期	上线前 人工点击量	上线后 人工点击量	上线前点击 耗时(分钟)	上线后点击 耗时(分钟)
2022/9/1	770	19	51	1
2022/9/2	707	20	47	1
2022/9/5	630	22	42	1
2022/9/6	1050	49	70	3
2022/9/7	5999	212	400	14
2022/9/8	1561	41	104	3
2022/9/9	4760	46	317	3

2022/9/13	1078	29	72	2
2022/9/14	798	10	53	1
2022/9/15	1778	53	119	4
2022/9/16	1722	18	115	1
2022/9/19	1855	20	124	1
2022/9/20	1113	35	74	2
2022/9/21	1232	17	82	1
2022/9/22	1491	34	99	2
2022/9/23	1575	12	105	1
2022/9/26	2604	42	174	3
2022/9/27	1365	18	91	1
2022/9/28	1876	16	125	1
2022/9/29	2212	46	147	3
2022/9/30	1470	30	98	2

（三）提升报送数据的质量

采用深度学习 TextCNN 模型辅助设置客户标签的新业务模式，相比原来纯手工模式，还有助于提升各个属性的准确率，从而进一步提升报送数据的质量。

这是因为，一方面 TextCNN 模型的预测结果已有很高的准确率，再由业务人员结合自身经验进行人工审核和修正后，进一步提升了数据质量；另一方面，当业务人员对新客户的信息掌握不足、存在偏差等特殊情况下，在某些属性上容易出现人工判断错误的风险时，TextCNN 模型的预测结果有可能提供出正确的候选，从而帮助业务人员突破依赖个人经验进行操作的限制，避免误选属性标签。

比如，仅凭个人经验，“招商财富”容易被误认为是“招商银行”的子公司，但实际上“招商财富”是“招商基金”的子公司。类似这样的情况，深度学习 TextCNN 模型的预测结果可以给出正确的候选标签供业务人员参考和审核，从而一定程度上减少了误选的风险。

综上所述，高准确率的深度学习模型在数据报送业务上的应用，不仅降低了业务人员的操作难度，更显著降低了人工的工作量和操作时间，提升了工作效率和工作体验，还有助于进一步提升报送数据的质量，降低了运营风险，为行业数字化转型进行了有益的探索实践。

参考文献

[1] 罗逸姝.新华社客户端官方帐号.《证券期货业科技发展“十四五”规划》发布[N]. 百度百家号.2021-10-22

[2] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746 – 1751, Doha, Qatar. Association for Computational Linguistics.

[3] Ye Zhang and Byron Wallace. 2017. A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 253 – 263.

[4] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 207 – 212, Berlin, Germany. Association for Computational Linguistics.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. ICLR Workshop, 2013a.