



证券研究报告·金融工程专题报告

“逐鹿” Alpha专题报告（十六）：基于Graph Embedding的行业因子向量化

分析师：丁鲁明

dingluming@csc.com.cn

SAC编号：S1440515020001

分析师：王超

wangchaodcq@csc.com.cn

SAC 编号：S1440522120002

发布日期：2023年9月5日

核心观点

传统方法在处理行业因子信息时通常采用one-hot将行业转变为哑变量进行处理（对于某些特殊模型，可以采用category变量）。这两者均无法表示行业之间的内在关联关系。且哑变量为稀疏矩阵，信息密度较低，增加了计算开销。

本文利用行业相关性矩阵构建图结构，采用Graph Embedding的算法Node2vec将行业因子进行向量化处理，得到的行业因子能够有效的包括行业内在相关性。

提纲

01

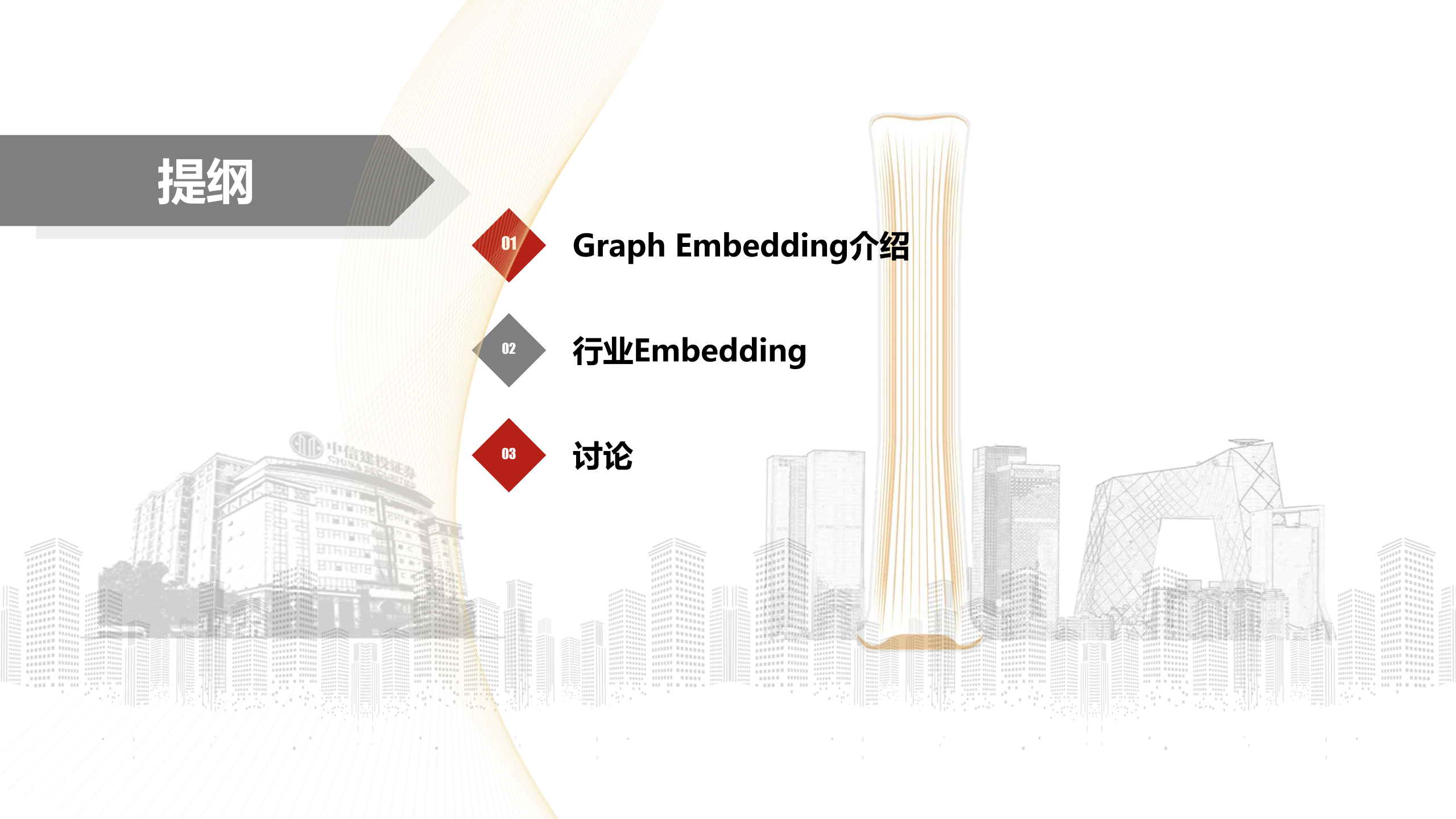
Graph Embedding介绍

02

行业Embedding

03

讨论



PART 1

Graph Embedding介绍

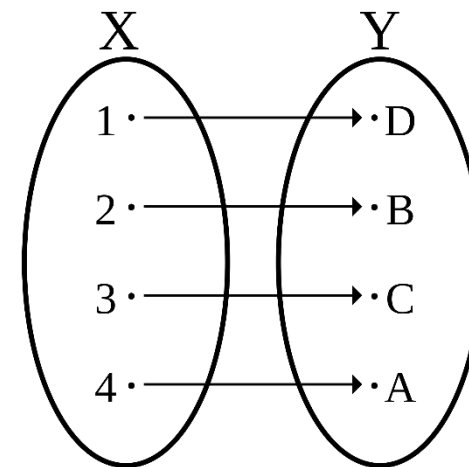
Embedding

嵌入 (Embedding) 是指一个数学结构经映射包含到另一个结构中。某个物件X称为嵌入到另一个物件Y中，是指有一个保持结构的单射 $f: X \rightarrow Y$ ，这个映射 f 就给出了一个嵌入

万物皆可Embedding (word embdding、item embedding、graph embedding)

经典的Word Embedding算法包括： Word2vec、 GloVe、 FastText等

大语言Embedding模型： BERT、 EloMo、 GPT等



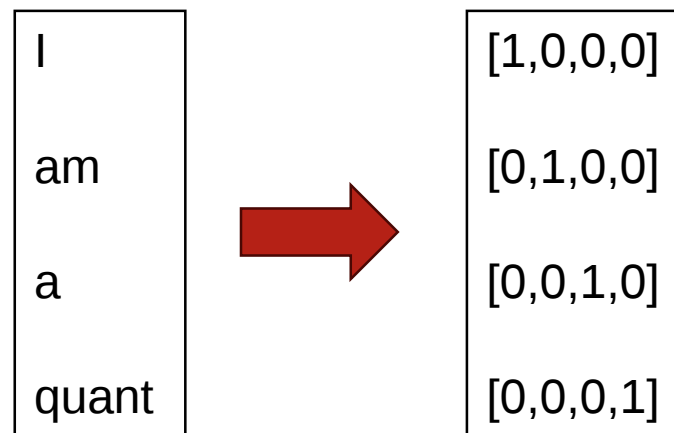
One-hot

独热编码 (one-hot) : 用N位状态寄存器来对N个状态进行编码

在NLP中, 将单词转换为独热向量

缺点:

- 维数灾难
- 语义鸿沟



Word2vec

Word2vec: 将单词转换为连续向量

基于连续向量，计算单词相似度：

model.most_similar('贵州茅台')：

('洋河股份', 0.9215401411056519)

('酒鬼酒', 0.9041466116905212)

('600519', 0.899872362613678)

('海天味业', 0.8987340331077576)

('口子窖', 0.8912389874458313)

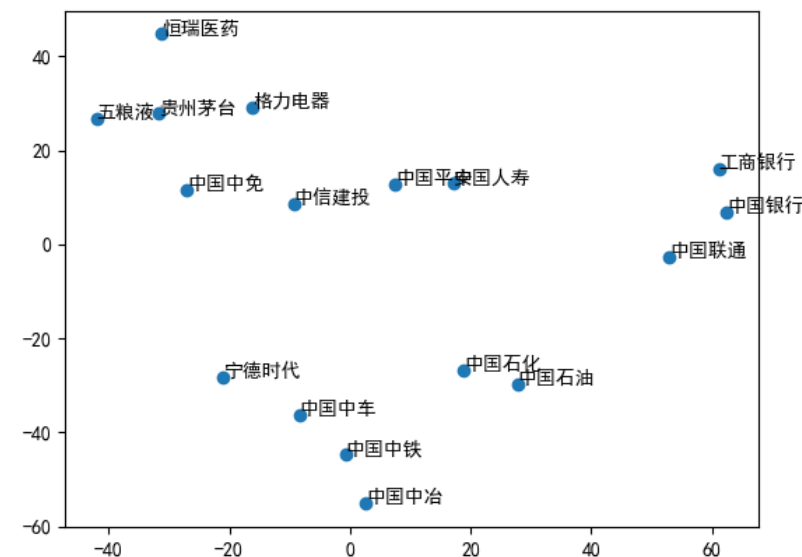
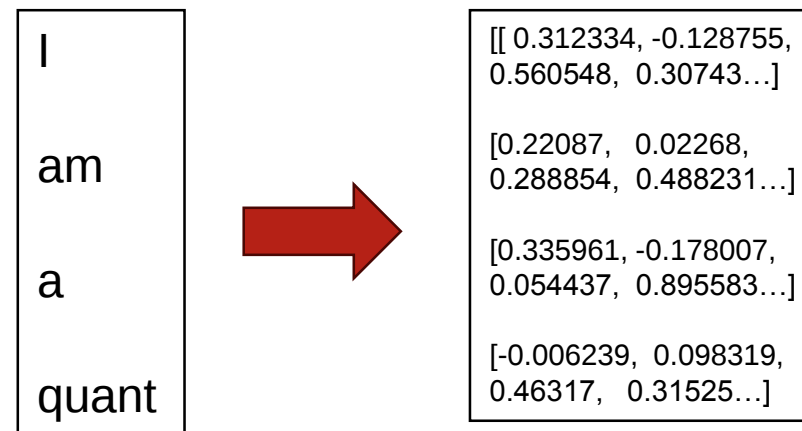
('顺鑫农业', 0.8851066827774048)

('五粮液', 0.8838584423065186)

('老白干酒', 0.8835573196411133)

('山西汾酒', 0.8828251361846924)

('古井贡酒', 0.8824747204780579)



资料来源：中信建投

Graph Embedding

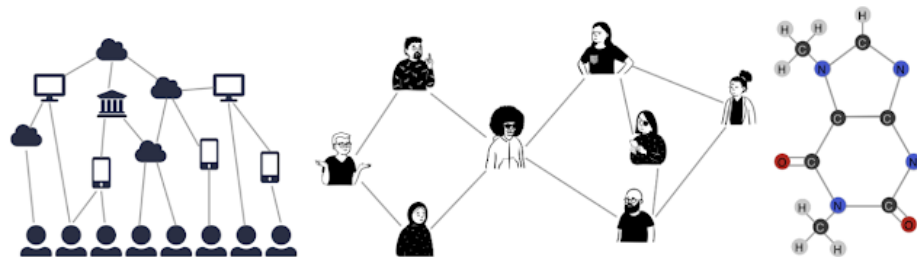
图数据结构用来表达物体之间的关联关系

图的定义：

V代表顶点集合，E代表边

邻接矩阵：

$$G = (V, E)$$



$$A_{ij} = \begin{cases} 1 & \text{if } \{v_i, v_j\} \in E \text{ and } i \neq j, \\ 0 & \text{otherwise} \end{cases}$$

常见的图嵌入算法包括DeepWalk, LINE, Node2Vec, SDNE, Struct2Vec, GraphSAGE等

图算法能够有效的计算节点之间的关系

图嵌入（Graph Embedding）和图神经网络（Graph Neural Network, GNN）两者互相关联，embedding生成的向量可以作为GNN的输入，同时某些embedding需要用到GNN的算法。

DeepWalk

在图中采用随机游走生成结点序列，利用Word2Vec (SkipGram) 算法进行无监督训练，生成embedding结果

Algorithm 1 DEEPWALK(G, w, d, γ, t)

Input: graph $G(V, E)$

 window size w

 embedding size d

 walks per vertex γ

 walk length t

Output: matrix of vertex representations $\Phi \in \mathbb{R}^{|V| \times d}$

1: Initialization: Sample Φ from $\mathcal{U}^{|V| \times d}$

2: Build a binary Tree T from V

3: **for** $i = 0$ to γ **do**

4: $\mathcal{O} = \text{Shuffle}(V)$

5: **for each** $v_i \in \mathcal{O}$ **do**

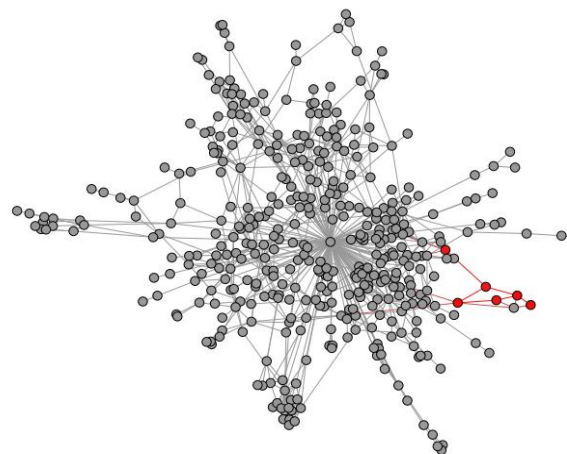
6: $\mathcal{W}_{v_i} = \text{RandomWalk}(G, v_i, t)$

7: SkipGram($\Phi, \mathcal{W}_{v_i}, w$)

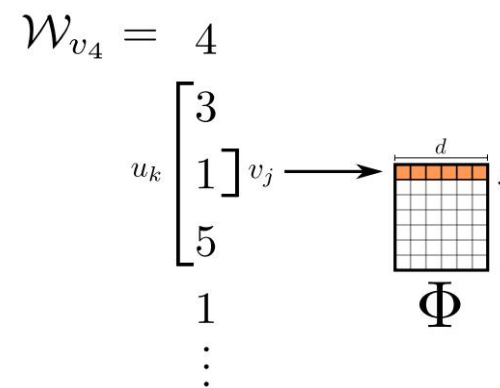
8: **end for**

9: **end for**

资料来源: DeepWalk: online learning of social representations, 中信建投



(a) Random walk generation.



(b) Representation mapping.

资料来源: DeepWalk: online learning of social representations, 中信建投

优点：算法简单高效

缺点：没有考虑边的属性，无法跳出有向图中出度为0的节点

Node2Vec

Node2vec是在DeepWalk的基础上，引入参数p和q，将随机游走序列调节为宽度优先搜索和深度优先搜索的序列生成

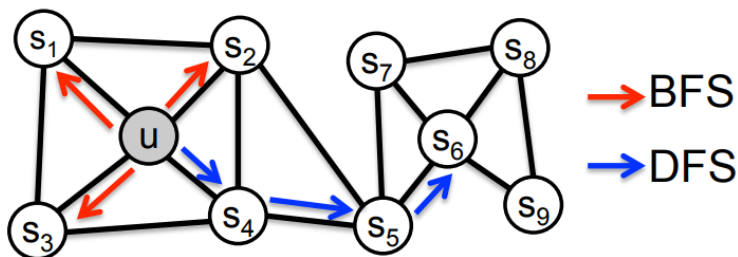
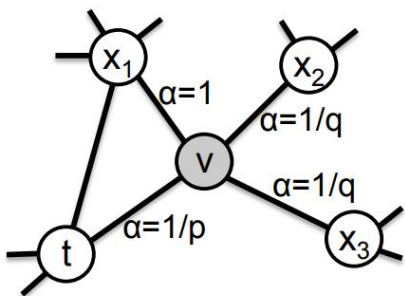


Figure 1: BFS and DFS search strategies from node u ($k = 3$).



资料来源: node2vec: Scalable Feature Learning for Networks, 中信建投

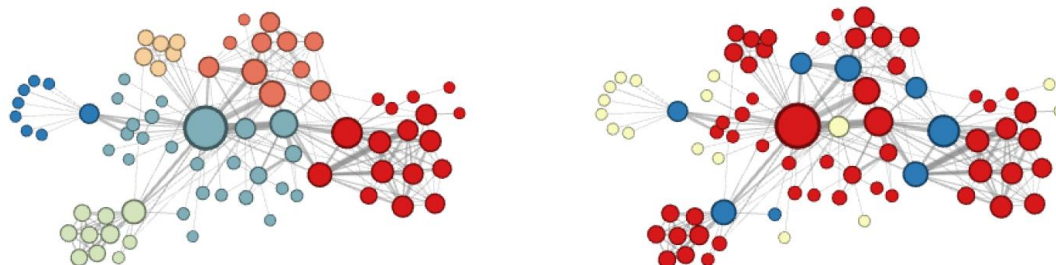
优点：可以学习同质图和同构图结构

缺点：超参p, q需要调节

$$P(c_i = x \mid c_{i-1} = v) = \begin{cases} \frac{\pi_{vx}}{Z} & \text{if } (v, x) \in E \\ 0 & \text{otherwise} \end{cases}$$

$$\alpha_{pq}(t, x) = \begin{cases} \frac{1}{p} & \text{if } d_{tx} = 0 \\ 1 & \text{if } d_{tx} = 1 \\ \frac{1}{q} & \text{if } d_{tx} = 2 \end{cases}$$

$$\pi_{vx} = \alpha_{pq}(t, x) \cdot w_{vx}$$



资料来源: node2vec: Scalable Feature Learning for Networks, 中信建投

PART 2

行业Embedding



中信建投证券
CHINA SECURITIES

行业因子

传统方法在处理行业因子信息时通常采用one-hot将行业转变为哑变量进行处理（对于某些特殊模型，可以采用category变量）

两者均无法表示行业之间的内在关联关系

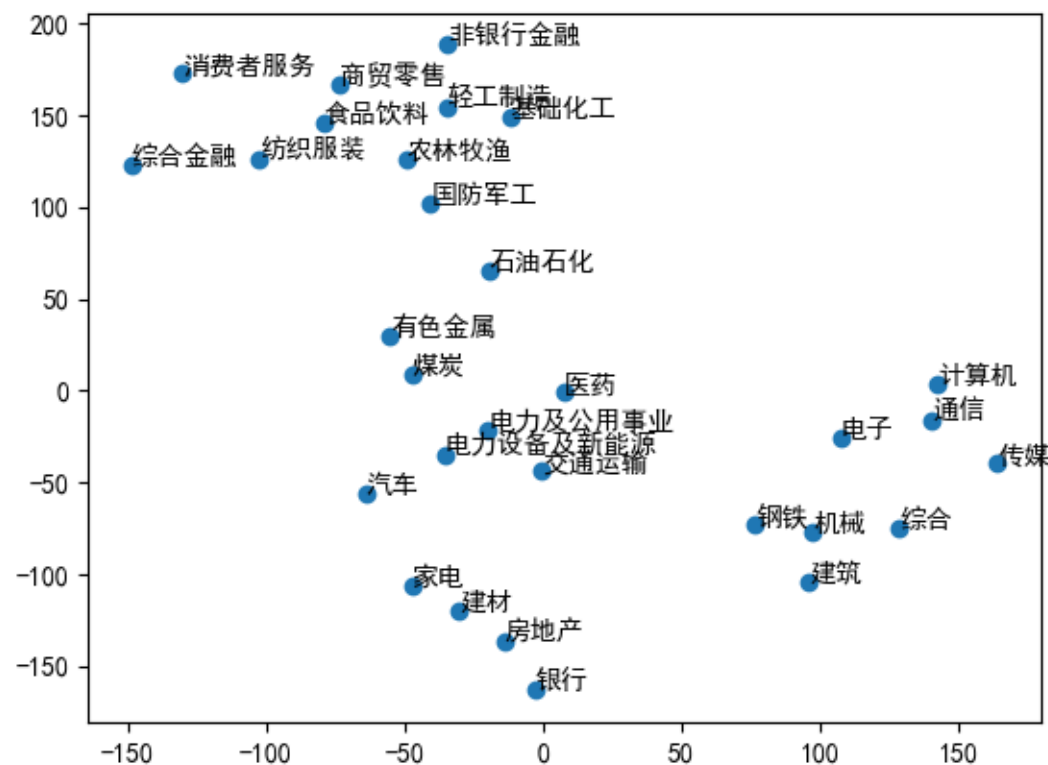
且哑变量为稀疏矩阵，信息密度较低，增加了计算开销

行业因子Word Embedding

直接采用Word Embedding提取行业名称的向量

能够反映行业之间的关系，但与二级市场的行业关系存在一定的区别，此关系主要来自文本资料，映射到二级市场会发生一定的改变

单次训练，无法回测



资料来源：中信建投

Node2Vec

构造有权完全无向图，指数收益率相关性矩阵作为邻接矩阵

节点为行业，边为行业之间的相关性

使用Node2vec 对行业因子进行embedding

Node2vec

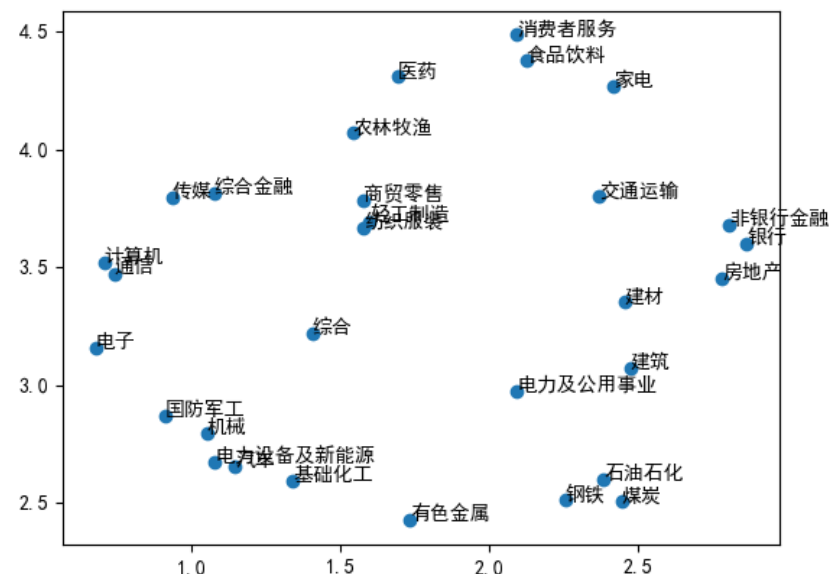
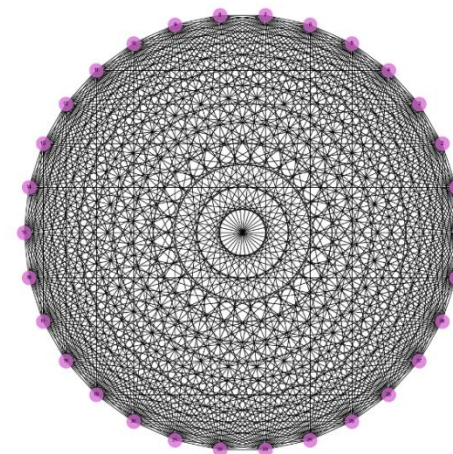
walk_length: 10

num_walks: 1e6

window_size: 5

embed_size: 30

每个行业均可用30维向量表示，
利用TSNE将向量压缩到2维可视化展示



资料来源: WIND, 中信建投

PART 3

讨论



中信建投证券
CHINA SECURITIES

讨论

相比于行业哑变量，行业graph embedding向量信息密度更高，且向量之间内在的包含了行业之间的相关性关系。

Embedding向量可以作为各类机器学习和深度学习模型的输入，更好的提高预测效果。

风险提示

- 本报告中所有数据结果是基于历史统计结果的展示，未来有可能发生风格切换导致因子失效的风险。模型运行存在一定的随机性，初始化随机数种子会对结果产生影响，单次运行结果可能会有一定偏差。历史数据的区间选择会对结果产生一定的影响。模型参数的不同会影响最终结果。模型对计算资源要求较高，运算量不足会导致结果存在一定的欠拟合风险。本文所有模型结果均来自历史数据，模型存在统计误差，不保证模型未来的有效性，对投资不构成任何建议。

附录

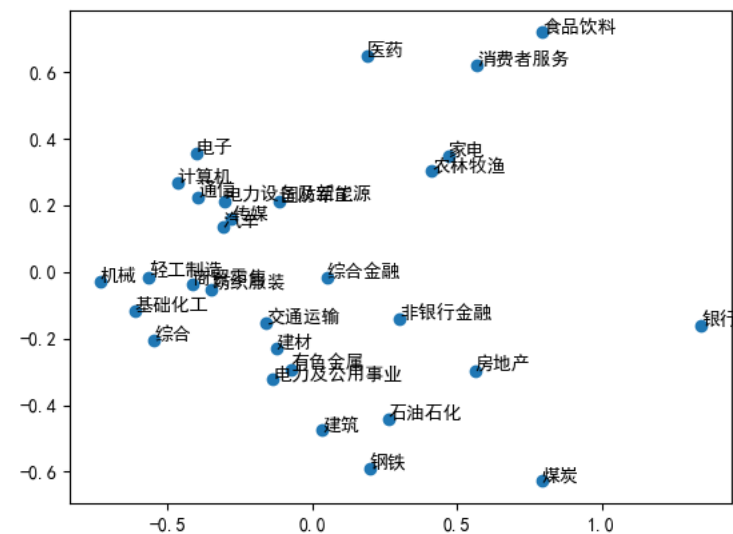
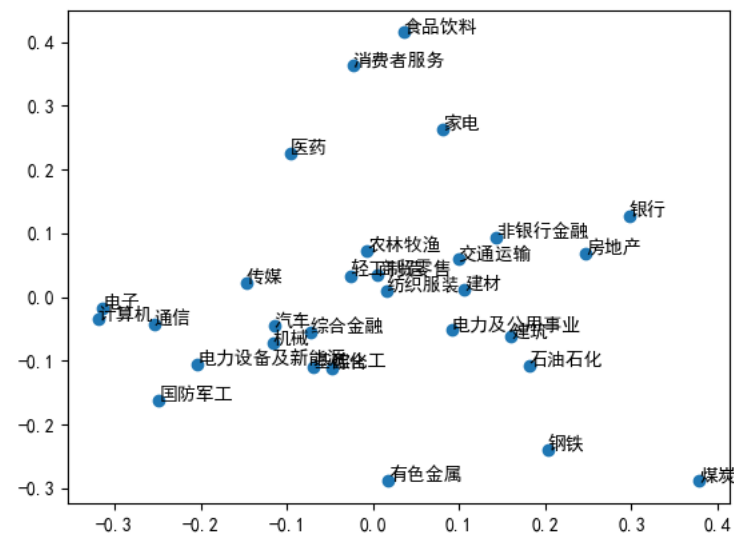


行业因子Graph Embedding

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
0	-0.077	0.168	-0.014	-0.227	-0.008	0.019	-0.111	-0.042	0.019	0.012	-0.094	-0.098	0.156	0.036	0.108	0.106	-0.047	0.088	0.058	-0.080	0.158	-0.173	-0.092	0.095	0.064	0.053	0.165	0.020	0.136	-0.314
1	0.053	0.019	-0.082	-0.153	-0.068	0.007	-0.180	-0.069	-0.084	0.043	-0.052	-0.009	0.099	0.126	-0.019	-0.058	0.136	0.153	0.117	-0.226	-0.031	-0.127	-0.249	-0.055	0.045	0.056	0.284	-0.029	-0.038	-0.265
2	-0.005	0.039	-0.172	-0.077	-0.160	0.110	-0.089	-0.141	-0.097	0.088	-0.020	0.070	0.158	-0.125	0.150	-0.008	0.191	0.079	-0.109	-0.132	-0.112	0.037	-0.071	-0.117	-0.082	-0.021	0.023	0.070	-0.013	-0.322
3	-0.192	0.101	-0.110	0.038	0.157	-0.154	-0.041	0.099	0.015	-0.178	0.022	0.077	0.062	0.040	0.088	-0.177	0.176	0.032	0.084	-0.051	-0.044	-0.212	-0.011	-0.022	-0.109	-0.019	0.149	0.052	-0.038	-0.271
4	0.056	-0.036	-0.059	0.060	0.018	-0.027	-0.276	-0.011	-0.101	-0.061	-0.109	0.165	0.174	0.050	-0.030	-0.069	0.010	0.107	-0.099	-0.146	0.123	0.004	-0.226	-0.044	0.052	-0.068	0.137	0.097	-0.081	-0.290
5	-0.007	-0.076	0.043	-0.023	-0.159	0.220	-0.105	0.005	-0.158	-0.005	0.094	0.010	0.171	0.034	0.140	-0.089	0.085	0.000	-0.026	0.035	0.133	-0.208	0.134	0.152	-0.049	-0.002	0.009	0.015	-0.152	-0.268
6	0.030	0.028	-0.153	-0.057	0.064	0.012	-0.093	-0.054	0.133	-0.173	-0.023	-0.130	-0.015	0.074	-0.133	-0.054	-0.028	-0.062	0.048	-0.091	-0.022	-0.144	-0.031	-0.115	0.175	-0.023	-0.004	0.093	-0.182	-0.371
7	-0.028	-0.020	-0.006	-0.047	-0.062	0.009	-0.146	-0.042	-0.081	-0.189	-0.015	0.055	-0.204	-0.056	-0.021	-0.030	-0.070	-0.003	-0.120	-0.089	-0.034	0.045	0.112	0.025	0.071	0.196	0.159	-0.101	-0.022	-0.372
8	-0.225	-0.066	-0.170	0.000	-0.123	0.029	0.139	0.157	-0.021	0.131	-0.116	0.045	-0.020	-0.009	-0.070	-0.204	0.028	0.040	-0.007	0.063	0.172	0.049	0.016	0.007	0.092	0.112	0.039	0.085	0.001	-0.312
9	0.118	-0.031	0.032	-0.136	0.009	0.019	0.149	0.161	-0.047	-0.081	-0.035	0.052	0.250	-0.171	0.136	-0.105	0.025	-0.123	-0.027	0.050	-0.013	0.084	0.008	0.024	0.150	-0.067	0.137	-0.102	-0.094	-0.309
10	-0.034	0.198	0.167	0.101	-0.246	-0.030	0.143	-0.021	-0.208	-0.113	0.012	0.007	0.197	-0.038	0.094	-0.066	0.104	-0.142	-0.031	0.013	-0.003	0.014	-0.052	0.007	-0.019	0.073	-0.043	0.085	-0.054	-0.311
11	-0.055	-0.043	0.035	-0.199	-0.089	-0.053	0.031	0.158	-0.190	0.029	-0.020	0.009	0.055	-0.011	0.039	0.044	0.080	-0.304	-0.212	-0.065	0.021	-0.018	0.022	0.013	-0.104	-0.128	0.065	0.192	-0.086	-0.260
12	-0.008	0.124	-0.026	0.012	0.059	-0.034	0.183	-0.070	-0.228	0.101	0.091	0.012	0.059	0.091	0.213	-0.011	-0.088	-0.018	-0.043	-0.194	0.074	0.071	0.027	0.066	-0.004	-0.042	-0.035	-0.026	-0.198	-0.326
13	-0.002	-0.061	-0.091	-0.076	0.008	-0.097	0.098	-0.014	0.091	0.021	0.029	-0.022	-0.016	-0.026	0.129	-0.015	-0.054	0.082	-0.220	0.207	0.007	-0.114	-0.160	0.010	-0.061	0.204	0.144	0.134	-0.246	-0.222
14	-0.044	-0.116	0.161	-0.054	0.096	0.091	0.252	-0.043	0.031	-0.150	0.099	0.035	-0.061	0.117	0.045	0.114	0.190	0.104	-0.125	0.076	0.162	0.168	0.022	-0.031	0.008	0.062	-0.009	0.096	0.075	-0.340
15	0.210	-0.052	0.131	0.005	0.064	-0.149	0.072	-0.104	-0.019	0.084	-0.003	-0.053	-0.165	0.105	0.065	-0.146	0.071	0.029	-0.053	0.129	0.133	0.022	0.150	-0.055	0.002	0.075	0.036	0.022	0.148	-0.391
16	-0.098	-0.183	0.003	0.046	0.093	-0.094	-0.034	-0.029	0.055	0.030	-0.015	-0.042	0.042	-0.177	0.133	0.034	0.069	-0.043	0.159	-0.064	0.074	0.036	-0.011	0.122	0.059	0.204	0.077	0.268	-0.162	-0.256
17	0.052	0.038	-0.098	0.162	0.001	-0.016	0.097	0.080	-0.062	-0.028	0.143	0.021	-0.026	0.040	0.007	0.183	0.188	-0.061	-0.089	0.227	0.079	0.091	0.066	0.199	0.054	0.064	0.015	0.100	0.051	-0.353
18	0.107	-0.040	0.069	0.059	0.001	-0.003	0.180	-0.042	0.073	-0.059	0.140	-0.076	-0.080	0.174	-0.003	-0.027	0.085	0.078	-0.063	0.146	0.045	0.184	0.066	0.140	-0.086	0.081	0.035	0.070	0.154	-0.399
19	-0.084	-0.187	0.018	-0.021	-0.119	-0.215	-0.003	0.167	-0.011	-0.006	0.152	-0.098	0.044	-0.044	-0.026	0.105	-0.007	0.100	0.093	0.040	-0.031	0.026	-0.073	-0.001	-0.181	-0.006	-0.049	-0.067	-0.056	-0.407
20	0.076	0.086	0.120	-0.014	0.152	0.132	-0.125	-0.040	0.138	0.086	-0.025	-0.083	-0.159	0.172	-0.094	-0.072	-0.049	0.118	0.107	0.005	0.042	0.058	-0.008	-0.074	-0.033	0.038	0.277	0.034	0.046	-0.371
21	0.125	0.111	0.113	0.056	0.173	0.214	0.028	0.093	0.005	0.055	-0.075	-0.073	-0.111	0.046	-0.078	-0.038	-0.120	0.035	0.051	0.041	-0.034	-0.004	-0.036	-0.104	-0.132	0.000	0.242	0.147	-0.006	-0.338
22	-0.090	0.076	0.101	-0.033	0.045	-0.016	-0.178	-0.068	0.134	0.086	0.079	-0.063	-0.164	0.061	-0.059	-0.088	0.058	0.040	0.038	0.086	0.122	0.057	0.012	-0.159	0.130	-0.052	0.285	-0.035	-0.178	-0.306
23	-0.104	-0.103	0.143	-0.006	-0.038	0.002	0.016	0.045	0.174	0.064	-0.009	0.197	-0.083	0.208	0.163	0.041	-0.001	0.000	0.056	0.021	-0.177	-0.092	-0.082	0.023	0.141	-0.011	0.058	0.073	0.020	-0.347
24	-0.078	-0.128	-0.028	0.072	-0.063	0.002	0.184	-0.099	-0.341	-0.060	-0.002	-0.093	0.015	-0.044	0.104	-0.035	-0.070	-0.033	-0.060	0.138	-0.101	-0.003	0.056	0.050	0.077	-0.199	0.177	0.172	-0.011	-0.243
25	-0.040	0.001	0.113	-0.022	-0.020	-0.131	0.195	-0.022	-0.082	0.072	-0.032	0.049	0.108	-0.148	-0.189	0.025	-0.003	0.110	-0.099	-0.039	-0.115	-0.172	0.188	0.119	0.082	-0.093	0.104	0.169	-0.137	-0.270
26	-0.123	-0.007	0.074	0.006	0.081	-0.004	-0.005	0.254	-0.196	0.056	0.087	-0.213	0.102	-0.050	0.030	-0.038	0.035	0.082	-0.222	0.033	-0.174	-0.028	0.108	0.028	0.178	0.044	-0.004	0.197	-0.061	-0.235
27	-0.212	-0.114	0.074	0.020	0.212	-0.004	0.016	-0.062	-0.119	0.094	-0.201	-0.082	0.091	-0.058	-0.103	0.054	0.109	-0.033	-0.090	0.182	-0.113	-0.059	-0.009	0.052	0.016	0.035	-0.059	-0.056	-0.147	-0.328
28	-0.074	0.067	0.028	-0.027	0.025	0.043	0.029	0.069	-0.073	0.023	0.025	-0.012	-0.194	-0.262	0.058	-0.091	0.158	0.028	-0.031	-0.047	0.030	-0.129	-0.249	0.177	0.042	-0.135	-0.030	0.063	-0.137	-0.278
29	-0.110	-0.112	0.051	0.263	0.053	0.062	0.127	0.101	0.090	0.085	-0.054	-0.136	0.014	-0.127	0.026	0.069	0.054	-0.046	-0.185	-0.125	0.026	-0.248	-0.060	-0.067	-0.007	-0.004	0.195	-0.043	-0.060	-0.265

资料来源：WIND, 中信建投

行业因子PCA



资料来源: WIND, 中信建投

分析师介绍

丁鲁明：同济大学金融数学硕士，中国准精算师，现任中信建投证券研究发展部执行总经理，金融工程团队、大类资产配置与基金研究团队首席分析师，中信建投证券基金投顾业务决策委员会成员，上海证券交易所定期专家交流组成员。13年证券从业，创立国内“量化基本面”投研体系，继承并深入研究经济经典长波体系中的康波周期理论并积极应用于实务，多次对资本市场重大趋势及拐点给出精准预判，对资产配置与经济周期运行具备深刻理解与认知。多次荣获团队荣誉：新财富最佳分析师2009第4、2012第4、2013第1、2014第3等；水晶球最佳分析师2009第1、2013第1等；Wind金牌分析师2018年第2、2019年第2等、2020年第4等。

王 超：南京大学粒子物理博士，曾担任基金公司研究员，券商研究员，有丰富的研究和投资经验，2021年加入中信建投，主要负责量化多因子选股。

评级说明

投资评级标准		评级	说明
报告中投资建议涉及的评级标准为报告发布日后6个月内的相对市场表现，也即报告发布日后的6个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A股市场以沪深300指数作为基准；新三板市场以三板成指为基准；香港市场以恒生指数作为基准；美国市场以标普 500 指数为基准。	股票评级	买入	相对涨幅15%以上
		增持	相对涨幅5%—15%
		中性	相对涨幅-5%—5%之间
		减持	相对跌幅5%—15%
		卖出	相对跌幅15%以上
	行业评级	强于大市	相对涨幅10%以上
		中性	相对涨幅-10-10%之间
		弱于大市	相对跌幅10%以上

分析师声明

本报告署名分析师在此声明：（i）以勤勉的职业态度、专业审慎的研究方法，使用合法合规的信息，独立、客观地出具本报告，结论不受任何第三方的授意或影响。（ii）本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

法律主体说明

本报告由中信建投证券股份有限公司及/或其附属机构（以下合称“中信建投”）制作，由中信建投证券股份有限公司在中华人民共和国（仅为本报告目的，不包括香港、澳门、台湾）提供。中信建投证券股份有限公司具有中国证监会许可的投资咨询业务资格，本报告署名分析师所持中国证券业协会授予的证券投资咨询执业资格证书编号已披露在报告首页。

在遵守适用的法律法规情况下，本报告亦可能由中信建投（国际）证券有限公司在香港提供。本报告作者所持香港证监会牌照的中央编号已披露在报告首页。

一般性声明

本报告由中信建投制作。发送本报告不构成任何合同或承诺的基础，不因接收者收到本报告而视其为中信建投客户。

本报告的信息均来源于中信建投认为可靠的公开资料，但中信建投对这些信息的准确性及完整性不作任何保证。本报告所载观点、评估和预测仅反映本报告出具日该分析师的判断，该等观点、评估和预测可能在不发出通知的情况下有所变更，亦有可能因使用不同假设和标准或者采用不同分析方法而与中信建投其他部门、人员口头或书面表达的意见不同或相反。本报告所引证券或其他金融工具的过往业绩不代表其未来表现。报告中所含任何具有预测性质的内容皆基于相应的假设条件，而任何假设条件都可能随时发生变化并影响实际投资收益。中信建投不承诺、不保证本报告所含具有预测性质的内容必然得以实现。

本报告内容的全部或部分均不构成投资建议。本报告所包含的观点、建议并未考虑报告接收人在财务状况、投资目的、风险偏好等方面的具体情况，报告接收者应当独立评估本报告所含信息，基于自身投资目标、需求、市场机会、风险及其他因素自主做出决策并自行承担投资风险。中信建投建议所有投资者应就任何潜在投资向其税务、会计或法律顾问咨询。不论报告接收者是否根据本报告做出投资决策，中信建投都不对该等投资决策提供任何形式的担保，亦不以任何形式分享投资收益或者分担投资损失。中信建投不对使用本报告所产生的任何直接或间接损失承担责任。

在法律法规及监管规定允许的范围内，中信建投可能持有并交易本报告中提公司的股份或其他财产权益，也可能在过去12个月、目前或者将来为本报告中所提公司提供或者争取为其提供投资银行、做市交易、财务顾问或其他金融服务。本报告内容真实、准确、完整地反映了署名分析师的观点，分析师的薪酬无论过去、现在或未来都不会直接或间接与其所撰写报告中的具体观点相联系，分析师亦不会因撰写本报告而获取不当利益。

本报告为中信建投所有。未经中信建投事先书面许可，任何机构和/或个人不得以任何形式转发、翻版、复制、发布或引用本报告全部或部分内容，亦不得从未经中信建投书面授权的任何机构、个人或其运营的媒体平台接收、翻版、复制或引用本报告全部或部分内容。版权所有，违者必究。

中信建投证券研究发展部

北京
东城区朝内大街2号凯恒中心B
座12层
电话：(8610) 8513-0588
联系人：李祉瑶
邮箱：lizhiyao@csc.com.cn

上海
浦东新区浦东南路528号南塔2103室
电话：(8621) 6882-1612
联系人：翁起帆
邮箱：wengqifan@csc.com.cn

深圳
福田区福中三路与鹏程一路交汇处广电金融中心
35楼
电话：(86755) 8252-1369
联系人：曹莹
邮箱：caoying@csc.com.cn

中信建投（国际）

香港
中环交易广场2期18楼
电话：(852) 3465-5600
联系人：刘泓麟
邮箱：charleneliu@csci.hk