DeepEye：一个基于深度学习的程序化交易识别与分类方法

徐广斌[1]，张伟[2]
1. 上海证券交易所资本市场研究所，上海 200120；2. 上海证券交易所产品创新中心，上海 200120

摘要
基于沪市A股交易数据，对A股市场程序化交易行为进行系统分析，构建程序化交易识别及分类特征指标体系，结合深度学习技术提出A股市场程序化交易的智能识别与分类方法——DeepEye，该方法可对程序化交易进行识别并分类。在真实交易行为数据集上的实验表明，所提出的方法在识别和分类上取得了较高的准确率，验证了将深度学习用于证券市场行为监管的可行性和有效性。该方法已辅助用于资本市场投资者画像及市场一线行为监管。

关键词
投资者行为；程序化交易；行为监管；深度学习；分类

中图分类号：TP393　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2018053

DeepEye: a deep learning-based method of recognition and classification of program trading

XU Guangbin[1], ZHANG Wei[2]
1. Capital Market Institute of Shanghai Stock Exchange, Shanghai 200120, China
2. Product Innovation Center of Shanghai Stock Exchange, Shanghai 200120, China

Abstract
Program trading behavior in A-share market has been systematically analyzed based on the Shanghai Stock Exchange's latest trading data and a feature indictor system has thus been built up for characterizing and classifying the program trading in the market. Furthermore, based on the deep learning technology, the A-share program trading intelligentized recognition and classification method, DeepEye, has been proposed, which enables program trading behavior in the market to be recognized and classified. The accuracy of the pilot implementation got about 70% which verified the feasibility and effectiveness of the new method. The proposed method can serve as an auxiliary measure to existing investor portraits and behavior supervision analysis for market regulation and can be a reference for improving the existing program trading regulatory rules.

Key words
investor behavior, program trading, behavior supervision, deep learning, classification

1 引言

程序化交易是指通过既定程序或特定软件，自动生成或执行交易指令的交易行为[1-4]。程序化交易操作策略复杂，运行高度自动化，对证券市场正常的交易秩序和交易安全造成了巨大的威胁[5]。如众所周知的美国股市“5·6闪崩”、骑士资本等事件，均是由程序化交易诱发或直接造成的。我国证券市场上程序化交易[6]发展较晚，但光大证券“8·16”等事故也充分反映出其可能造成的严重影响。因此近年来，加强程序化交易监管已成为市场诸方的一个共识。

2015年中国证券监督管理委员会和上海证券交易所、深圳证券交易所就程序化交易监管规则公开征求意见，其中首次提出A股程序化交易的量化标准（下文简称推荐规则）。推荐规则采取了“一刀切”式固定阈值方法来认定程序化交易，存在被技术性规避的局限。另一方面，程序化交易具有专业性和复杂性的特点，种类十分多样[7-9]。从订单功能上可将程序化交易大致分为：减少冲击类，如冰山订单、隐藏订单等算法；静默加减仓类，如大宗交易、保证成交量加权平均价格（guaranteed VWAP）算法等；交易执行类，如交易量加权平均价格（volume weighted average price，VWAP）、时间加权平均价格（time-weighted average price，TWAP）等算法；快速执行类，如市场直通（direct market access，DMA）及攻击算法等。从交易模式上则可分为做市策略、执行类策略（如订单拆分）、套利策略和方向性策略等。根据笔者近几年对沪市A股市场的研究，从整体上看，程序化交易对二级市场的流动性起到了积极的贡献作用，但不同类别的程序化交易对市场的影响作用差异显著，特别是在股市发生异常波动时，部分程序化交易会迅速消耗流动性，因此采用单一标准监管不同种类的策略，在科学性上显然有待商榷。

此外，由于境内外市场程序化交易发展差异巨大，境外的策略种类、认定标准和监管措施难以直接借鉴到国内市场，因此，如果不构建起A股市场的程序化交易识别和分类方法，要形成符合具体市场实际且行之有效的交易监管措施是十分困难的。

但是，前述的现有方法大多依据投资目的对程序化交易进行分类，在无法事前获知投资者意图的前提下，市场监管者要依据交易数据反推交易类型是很困难的。近来部分研究[10-12]从对市场定价影响的角度，探索通过分析订单流特征来对订单所属的投资者进行分类，如高频交易者、噪声交易者、基本面投资者和投机者等；也有相关研究[13-14]从投资者的交易记录中汇总出一些指标，再经过分类、聚类算法对投资者进行划分，然而这些方法的效果依赖于指标构建的好坏；从用户行为数据分析的角度看，有研究[15-16]对用户行为数据进行建模，如网站点击记录等，但这类数据涉及的属性相对简单、单一。整体而言，直接根据交易数据对A股程序化交易类型进行细分的相关研究迄今还比较缺乏。

鉴于深度学习[17-20]在音视频、文本、图像等[21-22]传统领域的表现，本文在系统分析证券交易行为数据的基础上构建了程序化交易特征指标体系，提出利用深度学习对A股程序化交易进行智能识别和分类的方法——DeepEye。本文提出的算法提取每日交易的基本的统计特征，并利用深度学习技术自行发掘深层次特征，简化了特征工程，同时避免了由专家设计指标时引入的偏差，提升了模型效果。

2 基于深度学习的程序化交易识别及分类

2.1 问题描述

由于A股市场上的程序化交易账号占比甚小，不足1%，而非程序化交易账户占到99%以上，故本文采用二阶段方式：首先，识别程序化交易账户，即给定一个账户的交易明细，判断该账户是否为程序化交易；然后，对程序化交易账户进行分类，即给定一个疑似程序化交易账户，判断该账户具体为9类程序化交易（见**表1**）中的哪一类。

2.2 方法概述

本文首先根据证券市场交易数据分析A股程序化交易的交易行为特征，构建了表征9类程序化交易行为的核心指标体系，并在此基础上，利用深度学习技术支持识别和分类的智能化。DeepEye主要的处理步骤如**图1**所示，包括：利用交易行为数据

表1 程序化交易账户类型分类

序号	类型	主要行为特征
1	篮子交易	满足推荐规则，有密集多标的报单
2	大单分拆	满足推荐规则，有个股密集报单
3	交易所交易基金（exchange traded fund，ETF）套利	满足推荐规则，有相当头寸的ETF反向交易
4	融券T+0（日内回转交易）	满足推荐规则，个股有相当头寸的信用买卖且无持仓
5	库存T+0	满足推荐规则，个股买卖头寸相当且有持仓
6	特殊账户	满足推荐规则，有大量订单申报但无成交的无成交账户、ETF基金账户和ETF做市商账户
7	混合对冲/调仓	满足推荐规则，整体有头寸相当的买卖
8	其他程序化交易	满足推荐规则，不属于上述分类
9	低频自动化	未达推荐规则，但持续进行低频的自动化报单

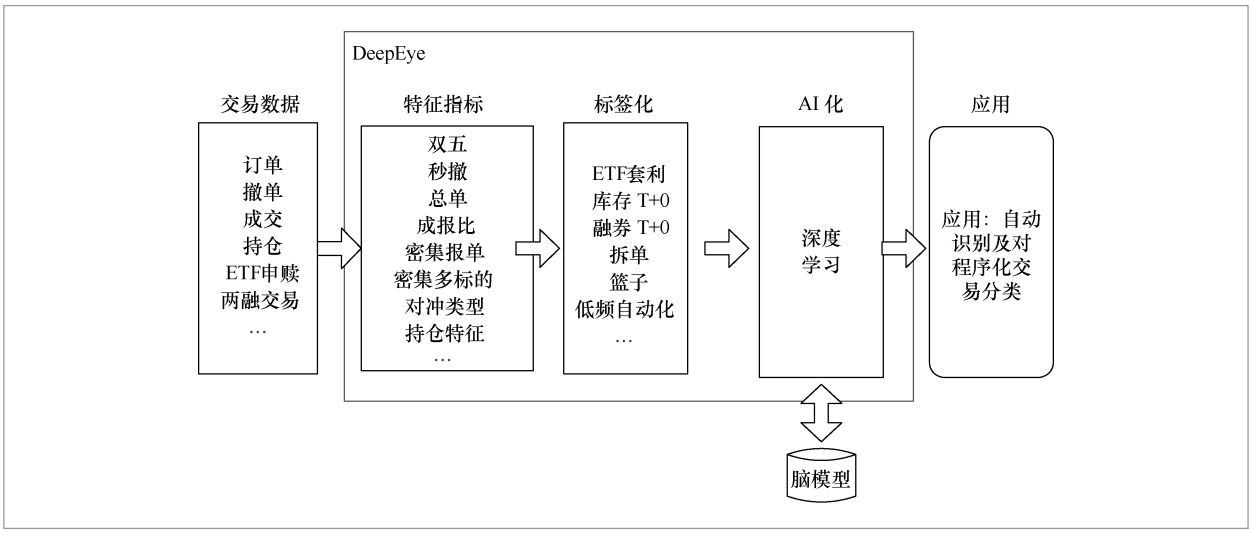

图1 DeepEye 主要处理流程

进行核心指标的计算，并根据类别识别规则，给目标账户（以推荐规则筛选）打上标签，将指标向量作为非类别属性，对应的标签向量作为类别属性（即Label属性）；构建深度学习模型进行训练，并用测试集进行验证，如果模型训练达到预期目标，则保存训练好的模型（含状态和权重）。在具体应用时，DeepEye预先加载保存的模型，并根据交易数据计算各账户的特征指标值，然后将其输入模型，即可识别其所属的类型。

2.3 数据预处理：特征指标体系

为支持上述处理流程的标签化处理，笔者系统梳理了A股市场上特征明显的程序化交易类型，结合调研、文献检索、交易数据分析以及监管经验，将A股市场的程序化交易分为套利交易、高频交易、风控交易和辅助交易几种模式。

具体地，从使用程序化交易的主要市场参与者的角度来看，基金、券商自营和合格境外投资者（qualified foreign institutional investors，QFII）主要使用拆单和算法交易（如VWAP、TWAP）降低市场冲击成本，同时也进行ETF套利、期现套利、对冲等交易，其中部分券商自营（金融工程部门）使用种类多样的其他交易策略；大户、超大户主要使用ETF套利、期现套利和一些绝对收益策略，部分使用批量委托、拆单、自动报撤单和统计套利；少量散户主要使用统计套利，或利用客户端工具中相对简单的自动化策略进行委托，或编写简单算法交易脚本实现自动化交易。

从交易行为特征上，T+0（日内回转交易）者主要表现为在同一证券品种或与衍生品之间（如信用账户融资融券与普通账户证券现货之间）进行头寸相反的一次或多次买卖，其中对于库存股T+0，上述行为主要表现在同一股票上且盘后具有持仓；套利交易者表现为在相关品种（如ETF成分股和ETF之间）上进行快速集中的对冲交易；混合对冲类型是综合考虑股票、融资融券和ETF申赎在所有标的上买卖两个方向头寸的对冲情况，这包含了跨品种套利或以调仓为目的的程序交易行为；执行类算法交易者主要执行拆单或篮子交易订单，一般不进行对冲交易，也不进行回转交易，但进行快速集中的单个或多个证券的交易；推荐规则主要依据报单速度和报单次数来界定，忽略了以低频（如分钟级别）自动化执行的算法交易，本文将此类交易行为也纳入程序化交易范畴。前几类程序化交易者都属于对冲型交易，使用对冲或回转的方式来保持日初和日终的风险敞口基本不变，而算法交易（含低频自动化）者则属于单向看多或看空。因此，可以先根据对冲比例，将程序化交易划分为对冲型、部分对冲型和非对冲型，然后再根据各类程序化交易的典型特征，进一步细分类型。为排除疑似伞形账户（如沪股通账户）的干扰，笔者也将伞形账户作为单独一类的非程序化账户纳入特征分析。

根据上述各类型在订单层面的交易特征，并结合推荐规则，可将A股程序化交易行为细分为**表1**的9种类型。

在对各类交易行为进行分析的基础上，笔者梳理出了表征各类型的12项核心特征指标（见**表2**），包括推荐规则的“一个交易日内出现5次以上每秒申报5笔”“1 s内完成申报并撤销申报，且日内出现3次以上”“日内申报2 000笔以上”以及对冲类型、交易及持仓等关键性特征。其中指标1至指标3在标签化处理时，按是否达到推荐规则来赋值（“1/0”代表“真/伪”）。

表 2 程序化交易特征指标体系

序号	指标	特征说明	
1	日内申报5笔/s的发生次数	对应推荐规则	
2	日内1 s申报并撤销的发生次数	对应推荐规则	
3	日内申报笔数	对应推荐规则	
4	对冲类型	个股对冲(1)	成交个股的买卖量对冲，即|买量−卖量|/(买量+卖量)<0.1
		信用交易对冲(2)	不属于前一类，且|股票净买额−融券净卖额|/(股票净买额+融券净卖额)<0.1
		ETF对冲(3)	不属于前两类，且|股票净买额−ETF净卖额|/(股票净买额+ETF净卖额)<0.1
		混合对冲(4)	不属于前三类，且|股票净买额−融券净卖额|/(股票净买额+融券净卖额)<0.1
		非对冲(0)	其他情况
5	成交申报比	成交量/申报量，无申报、显著低(<0.1)、低(0.1~0.5)、正常(>0.5)	
6	持仓多标的	持仓标的数量达15	
7	个股买卖不平衡显著性	成交个股的|买量−卖量|/(买量+卖量)，无成交、显著低(<0.1)、中(0.1~0.9)、显著高(>0.9)	
8	低频自动化特征	个股申报有连续60个1 min步长滑动窗口(窗口大小为2 min)内均有申报	
9	个股密集报单	有个股报单笔数/s达10 笔及以上	
10	密集多标的	有申报标的/s达10 只及以上	
11	交易多标的	成交标的数量达10 只及以上	
12	无/低持仓	无成交、无持仓或持仓量/成交量<0.1、其他	

2.4 基于循环神经网络的账户识别与分类

笔者采用深度神经网络实现程序化交易账户的识别与分类。通过输入训练数据集，即目标账号的12项特征指标数值和标签化结果(即标签向量)，训练深度神经网络。通过多轮的学习，网络模型将记录包含隐层中上下文信息(张量及权重)，如果经验证集检验，模型达到要求，则学习结束，学习好的模型可以保存下来供后续应用使用。这里的标签向量包括：用于识别是否为程序化交易的二分类标签；用于分类程序化交易账户的多分类标签。分别对应程序化交易账户识别以及程序化交易账户的类型分类。

具体地，采用循环神经网络(recurrent neural network，RNN)作为学习模型，在设计模型时，配置使用线性修正单元(rectified linear unit，ReLU)作为中间层的激活函数，softmax作为输出层的激活函数，预测值与真实类别的交叉熵作为损失函数，并采用RMSprop优化器对模型参数进行优化(如图2所示)。对于从目标账户中区分程序化交易和非程序化交易账户的任务，模型输出的是程序化交易的概率；对于识别后的程序化交易账户的类型进行细分，输出为细分的9类标签的对应概率。具体算法如下。

训练：

输入：指标X，标签y'，模型参数W(随机初始化)。

输出：模型参数W。

循环：

计算预测值：y=softmax(rnn(X))。

计算损失：loss=cross-entropy(y,y')。

计算参数更新方向：d=rmsprop(loss,W)。

更新参数：W=W+d。

预测：

输入：指标X，模型参数W。

输出：y=softmax(rnn(X))。

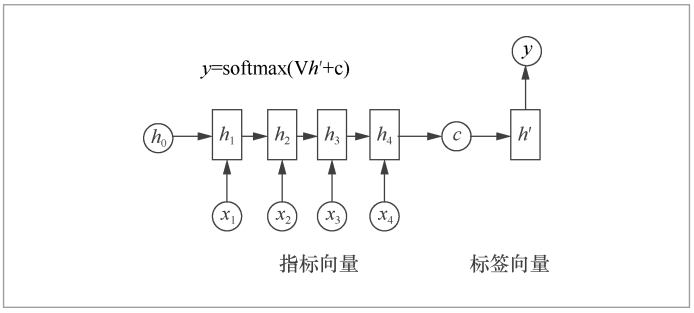

图 2 程序化交易分类模型示意

3 数据与实验

3.1 数据来源与处理

为反映A股程序化交易在市场稳定阶段的最新发展状况，本文选取上海证券交易所（以下简称上证）A股市场2017年上半年的119个连续交易日作为研究样本区间进行相关实证。本文数据主要采用上海证券交易所数据仓库的上证A股的逐笔委托、撤单数据、成交数据、融资融券及ETF交易数据和持仓数据。研究目标时间段为日内9:30—15:00的盘中竞价阶段。

样本区间活跃账户共3 739万户，对全市场数量占比超过99%的个人投资者账户（即A账户）采用抽样方法进行实验，具体如下：所有个人投资者账户按0.01的抽样率随机抽样，其他类型账号保留，抽样后的账号共65.6万户，占总活跃账户的1.75%，其中个人投资者账户37.2万户，机构账户28.4万户；程序化交易账号6 596户，其中各类型占比如**图3**所示。

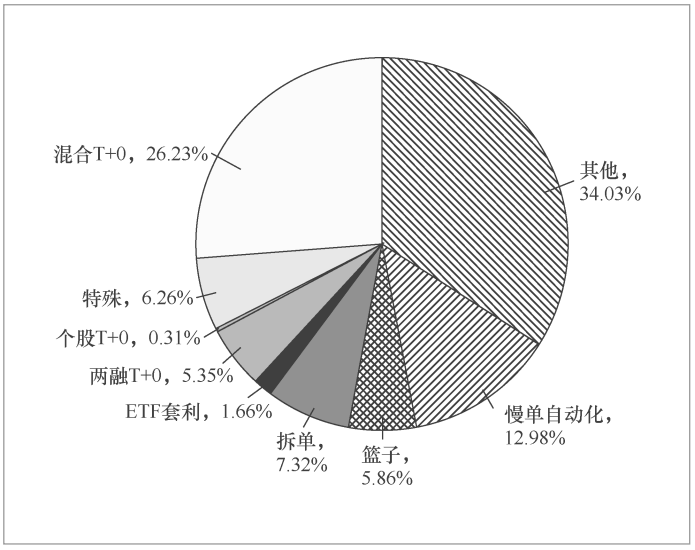

图 3 程序化交易类别比例

3.2 实验结果

实验采用TensorFlow1.0作为深度学习后端，使用Keras 2.0.9提供高级别接口，开发语言为Python 3.6，开发和实验环境为Fedora 22 Linux 。设备平台为惠普Z840工作站。

首先是程序化交易识别，模型的输入为**表2**中的指标1、2、3、8，输出为是否为程序化交易账户。经最低申报笔数筛选和伞形特征账户筛除，样本期间内共有20.7万个待分类账户，按9:1比例划分训练集和验证集，18.6万户为训练集，2.1万户为验证集，鉴于实验条件，以5 120个为一个批次大小（batch size）进行学习，学习结果为属于程序化/非程序化交易账户的概率，选择概率大的来确定类别。在具体应用中，也可通过调节界定阈值来放松或收紧监管范围。经验证，训练后的模型识别准确率为

67%、召回率为60%。

其次是程序化交易分类，样本期间内共有6 596个程序化交易账户待分类，指标1至指标12作为输入，9类分类标签作为输出，按9:1比例划分训练集和验证集。以512个为一个批次大小进行学习。见**表3**，训练后的模型分类准确率超过70%。

最后，将模型在样本期数据上进行应用，得到样本期A股程序化交易账户共含1.08万户，占活跃总账号的0.029%，个人账户（A账户）程序化交易占比0.01%，机构账户程序化交易占比2.50%。

本文提出的方法在识别和分类上取得了较高的准确率，可以识别现有方法无法识别的、占比近13%的低频自动化程序化交易，还支持类型细分和模糊识别，基本验证了用于证券市场行为监管的可行性和有效性。

4 结束语

本文提出程序化交易的智能识别与分类方法DeepEye，验证了将深度学习用于投资者行为监管的可行性和有效性。借助DeepEye识别并确定程序化交易的类别后，可以使用现有的账户分析技术进行更为深入的分析，比如分析各类别的交易特征、持仓特征、账户特征等，可辅助用于一线监管中的投资者行为分析和投资者画像分析；分类规则和特征指标体系可作为进一步完善程序化交易监管制度的参考；方法经适应性修改后还可应用于期权、期货等其他交易市场。该方法与现有程序化交易的识别方法完全兼容，覆盖类型范围更广，而且支持类型细分和模糊识别，现已作为投资者画像模块的重要组成部分，应用于上海证券交易所新一代市场监察系统中。

表 3 实验结果

分类	准确率	召回率
识别	67%	60%
分类	72%	—

需要指出的是，本文的方法具有一定的局限性。一是基于人工智能的方法，预测结果对模型、算法、参数的敏感性高，可解释性差，且有一定不确定性，需结合大数据验证和人工决策进行应用。二是对程序化交易账户的筛选处理采用了比较严格的标准，难以对特征不明显的行为进行处理。三是部分特征指标阈值的合理性还需在不同市场阶段和板块上进一步论证。四是暂时缺乏衍生品市场数据，难以对跨市场的程序化交易行为进行分析。

后续将使用不同市场阶段数据进一步进行大样本验证、模型调优，并结合衍生品市场数据，扩大分析范围。

参考文献:

[1] 刘逖. 市场微观结构与交易机制设计高级指南[M]. 上海：上海人民出版社，2012：569-583.
LIU T. Market microstructure and trading mechanism advanced guideline [M]. Shanghai: Shanghai People's Press, 2012: 569-583.

[2] ALDRIDGE I. High frequency trading[M]. New Jersey: John Wiley & Sons, Inc., 2010: 7-35.

[3] SEYFERT R. Bugs, predations or manipulations? Incompatible epistemic regimes of high-frequency trading[J]. Economy and Society, 2016, 45(2): 251-277.

[4] 叶伟. 我国资本市场程序化交易的风险控制策略[J]. 证券市场导报，2014(8)：46-52.
YE W. The risk control strategies of program trading of Chinese capital market[J]. Securities Market Herald, 2014(8): 46-52.

[5] 熊熊，袁海亮，张维，等. 程序化交易及其

风险分析[J]. 电子科技大学学报(社科版), 2011, 13(3): 32-39.
XIONG X, YUAN H L, ZHANG W, et al. Program Trading overview and risk analysis[J]. Journal of University Electronics Science and Technology of China, 2011, 13(3): 32-39.

[6] 彭蕾. 中国证券市场程序化交易研究[D]. 成都: 西南财经大学, 2005: 4-17.
PENG L. The research on program trading of the Chinese securities market[D]. Chengdu: Southwest University of Finance and Economics Press, 2005: 4-17.

[7] 陈梦根. 算法交易的兴起及最新研究进展[J]. 证券市场导报, 2013(9): 11-17.
CHEN M G. Algorithmic trading's rising and advances[J]. Securities Market Herald , 2013(9): 11-17.

[8] 蓝海平. 高频交易的技术特征、发展趋势及挑战[J]. 证券市场导报, 2014(4): 59-64.
LAN H P. HFT: the technique feature, developments and challenges[J]. Securities Market Herald, 2014(4): 59-64.

[9] 郭朋. 国外高频交易的发展现状及启示[J]. 证券市场导报, 2012(7): 56-61.
GUO P. Development of high frequency trading and its implication[J]. Securities Market Herald, 2012(7): 56-61.

[10] YANG S Y, QIAO Q F, BELING P A, et al. Gaussian process-based algorithmic trading strategy identification[J]. Quantitative Finance, 2015, 15(10): 1683-1702.

[11] QIAO Q F, BELING P A. Decision analytics and machine learning in economic and financial systems[J]. Environment Systems and Decisions, 2016, 36(2): 109-113.

[12] YANG S Y, QIAO Q F, BELING P A. Algorithmic trading behavior identification using reward learning method[C]//The 2014 International Joint Conference on Neural Networks, July 6-11, 2014, Beijing, China. Red Hook: Curran Associates, 2014: 3807-3414.

[13] 张鸿萍. 基于时间序列交易数据的服装电商客户分类研究[J]. 现代管理, 2017, 7(6): 481-492.
ZHANG H P. Research for customer classification of clothing E-business based on time series transaction data[J]. Modern Management, 2017, 7(6): 481-492.

[14] 毛瑞, 费宇. 基于交易数据的客户分类研究[J]. 中国证券期货, 2012(1): 22-23.
MAO R, FEI Y. The study of customer classification based on trading data[J]. Securities & Futures of China, 2012(1): 22-23.

[15] WANG G, ZHANG X, TANG S, et al. Clickstream user behavior models[J]. ACM Transactions on the Web, 2017, 11(4): 1-37.

[16] BENSON A R, KUMAR R, TOMKINS A. Modeling user consumption sequences[C]// The 25th International Conference on World Wide Web. International World Wide Web Conferences, April 11-15, 2016, Montréal, Canada. New York: ACM Press, 2016: 519-529.

[17] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[18] 马世龙, 乌尼日其其格, 李小平, 等. 大数据与深度学习综述[J]. 智能系统学报, 2016, 11(6): 728-742.
MA S L, WUNIRI Q Q G, LI X P, et al. Deep learning with big data: state of the art and development[J]. CAAI Transactions on Intelligent Systems, 2016, 11(6): 728-742.

[19] SCHMIDHUBER J. Deep learning in neural networks: an overview[J]. Neural Networks, 2015, 61(1): 85-117.

[20] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. Computer Science, 2014, arXiv:1409.3215.

[21] 孙志远, 鲁成祥, 史忠植, 等. 深度学习研究与进展[J]. 计算机科学, 2016, 43(2): 1-8.
SUN Z Y, LU C X, SHI Z Z, et al. Research and advances on deep learning[J]. Computer Science, 2016,

43(2): 1-8.

[22] KARNOWSKI T P, AREL I, ROSE D C. Deep spatiotemporal feature learning with application to image classification[C]//The 9th International Conference on Machine Learning and Applications, December 12, 2010, Washington, DC, USA. Piscataway: IEEE Computer Society, 2010: 883-888.

作者简介

徐广斌(1976-),男,博士,上海证券交易所资本市场研究所高级工程师、业务主管,主要研究方向为证券信息技术、大数据、金融计算。

张伟(1989-),男,就职于上海证券交易所产品创新中心,主要研究方向为金融工程。

收稿日期:2018-06-25

2018053-9