システム創生学特別演習 4C「近未来金融システムの創成」第 12 回講義レポート

第 12 回は、東京大学大学院 工学系研究科の和泉潔教授から「テキストマイニングの可能性」という題で講義があった。本日の講義は金融と技術の各論の最終回である第 4 回で、テキストマイニングなどの AI 技術を資産運用に活用する事例などを先生にお話しいただいた。

代替データへのシフト

GDP や既存の経済指標はデータが発表されるまで 2,3 ヶ月かかることが多い。だが、クレジットカードの決済データを用いれば、より詳細に早く経済動向を読み取れる。決済データではなくても、新宿への来街者数を調べれば消費動向の予測に使うことができる。最近はこのような代替データ(非伝統的データ)が使われるようになっており、代替データを提供するビジネスや調査も増加している。今までは経済動向の把握のためにクレジットカードや POS データといった構造化データを用いていたが、これからは消費に直接的には関わってはいないが分析すれば経済状況がわかるかもしれない非構造化データを活用する方向性にシフトしてくるのではないか。

非構造化データの活用で期待されているのは人工知能や機械学習などの技術である。人工知能系の国際会議では今まで経済や金融に関わる発表はされていなかったが、ここ数年で金融をテーマにした研究会やワークショップが立ち上がり始めた。海外だけでなく、国内でもその傾向はある。人工知能学会の金融情報学研究会の聴講者数は増加しており、発表件数も金融機関とアカデミアの共同論文の数が特に増えている。また、アカデミアだけでなく海外の金融機関ではデータ分析チームが立ち上がっている。役割が綺麗に分かれているアメフト型のチームが特徴的で、大学の教授を招聘するなどアカデミアとの連携も活発だ。

人工知能技術が金融に与えた影響

画像・音声データや大規模な数値データ、テキストデータなどの新たな種類のデータを取り扱えるようになり、金融に関する予測や要約・検知を行えるようになった。例えば画像・音声データでは、アメリカの Orbital Insight 社が石油タンクの蓋に映し出される影を分析して石油の備蓄量を弾き出している。それをエネルギー関連企業や政府、投資家などへ提供し石油に関する需給ステータスを知る助けにしている。また、テキストデータにおいては Insight360 というサービスが企業に関するテキストデータや数値データなどを分析し ESG に関する 14 のトピッ

クの指数を生成している。テキストマイニングの手法を用いることでバイアスがかかりやすいESG を踏まえた企業価値の数値化を可能にしている。

人工知能技術のコモディティ化

金融データマイニングには 3 つのトレンドがある。一つ目は人工知能技術のコモディティ化だ。今までは、テキストマイニングをビジネスに用いようとしても、膨大なデータを用意する必要があり、かつ高性能の計算機で長時間計算しなければいけないためハードルが高かった。だが、学習済みモデルの公開によってこれらの問題が解消しつつある。例えば、Google BERT という汎用言語表現モデルは、単語列感の距離(類似度)を計算する。

そもそも、金融テキストマイニングの枠組みは、まず自然言語データの特徴量を数値に変換する前処理から始まる。数値に変換された X を入力して、株価や為替のリターンやボラリティを予想するのだ。ここで重要なのは、学習のアルゴリズムを難しくすることではなく、自然言語を数値にいかに綺麗に変換するかである。いらない単語を数えないようにして、重要な単語を弾き出さなければ意味がないからだ。この部分に働くのが BERT である 。今までは文章中の単語の前後関係などを無視して出現回数だけで考える Bag-of-wards を用いていた。だが、これには文構造や係り受け、類語などの単語間の関係を無視してしまう問題が生じていた。だが、BERT は自然言語における単語系をベクトルとして表現する。しかも、意味を考慮して分析しする。例えば「業績予想を上方修正する」という文があり、もう一方の文が「業績予想を上方修正はない」という否定的な言葉が使われているものだったら、この二つの間の距離は遠いと判断する。一方、「売り上げ実績が好調である」という異なる単語を使ったほぼ同じ意味の文章との距離では近いと判断する。そして近年、この BERT を用いて金融テキストマイニングの予測をする研究が行われた。中国系の企業の株価予想を BERT で分析すると言う内容だが、残念ながら予測結果はあまりよくなかった。文章がポジティブな内容なのかネガティブな内容なのという判断の精度は高いが、BERT を用いた研究の決定版はまだない。

二つ目の公開されている学習済みモデルの例は金融専門の極性辞書だ。これは和泉研で公開されているもので 19,629 語もの金融用語を金融の文脈においてポジティブかネガティブかを数値化している。例えば、「人手不足」や「価格競争」などの単語は消費者にとっては好ましいことだが、企業にとっては好ましくないことなのでマイナスの極性値が付けられている。このスコアを学習するためにニューラルネットワークを用いた。過去 10 年間のトムソン・ロイターの記事を入れて、翌日の日経平均株価のリターンを予測させる学習を行なった。この時、各単語にリンクの値がスコアの値になるようにした。この極性辞書を用いてアナリストのレポートテキストを分析し株価の変動を予測させ、二週間後の株価と相関があるかどうかをテストし

た。結果、一般的な極性辞書と比較すると金融専門の極性辞書を用いた予測の方が精度は高かった。

経済的因果の分析

２番目の金融データマイニングの潮流は経済的因果の分析だ。経済や金融では基本的に因果の分析を行う。例えば、イギリスの EU 離脱がどのような経済影響を与えるか等だ。だが、自然科学には基本方程式があるが社会現象や経済現象にはない。そのため、経済現象を分析するには人間が認識して予測する必要がある。そこで、この因果分析を機械で行うことはできないかと試みた。テキストマイニングを用いて 2012 年 10 月から 2018 年 5 月までの決算短信のテキストを分析し、約 100 万個の因果関係のデータを生成した。決算短信には「経営成績に関する分析」や「事業などのリスク」など原因と結果を表しているテキストが多くある。このようなテキストから因果関係を分析した。

この研究では二つの分析アプローチを用いている。一つ目は、単語やフレーズの頻度と順序の情報を使うパターン認識アプローチだ。二つ目は、単語や分の構造情報を用いる自然言語処理アプローチである。

自然言語処理アプローチでは係り受け解析を行い、何が主語になってどの単語にかかっているのかを分析した。この時、手がかり表現という「~により」や「〜をもたらし」といった、いかにも原因の結果を表すテキストを事前にリストアップする。これらを含む文章の中で原因と結果を表す情報はどこかを自然言語処理と機械学習によって箇所の推定を自動的に行なうのだ。原因とその結果起きたことを一つのペアにし、決算短信の中からこのようなペアを 100 万個抽出した。

この因果のペアを用いれば、どんな波及効果や経済的な影響が出るのか因果関係が連鎖的にわかる。例えば、「感染症」というテキストを入力すると、感染症が原因となる波及効果がヒットする。出てきた因果をクリックすれば、今度はその結果が原因となり起こりうる波及効果がヒットし、因果の関係がチェーンのように連鎖的に表示されるのだ。例えば、「小麦価格」からの波及効果を調べると、第一段階で因果が出てくる企業は日清製粉や山崎製パンなどいかにも小麦と繋がりそうな会社名とその波及効果が出てくる。第二段階では少し小麦粉と関係する会社、第三段階ではもう少し間接的に小麦粉とつながる会社と波及効果が出てくる。実際に 2018 年 7 月に小麦の国際価格が上昇した時に、検索結果に出てきた企業がどのような影響を受けたのか調べた研究がある。これらの会社のリターンの絶対値の平均のグラフを見ると、第一段階・第二段階の会社はニュースが出て二週間近く経つとその影響は消える。だが、最も間接

的に影響がある第三段階の会社は 1 ヶ月近くそのインパクトが続いていた。すると、「インダイレクトな方が普及し尽くすのに時間がかかる」と言う仮説が考えられる。

他の応用例として企業と企業の株価の関係を調べた研究がある。従来はサプライチェーンを用いて企業間の関係を調べる手法が多かった。そこでこの研究では因果チェーンを用いて企業と企業の結びつきの強さを調べた。例えば、A という企業が B の何かしらの原因に結びつけば、その企業間の関係は強いとみなした。逆に因果の関係が薄いものは関係が薄いとみなした。興味深いのは、同じ会社のペアであっても因果の向きによって関係の強弱が異なることである。つまり、A は B に対して強い関係を示していたとしても B は A に対しては関係が弱いと言うケースが見られた。

これを用いて、TOPIX に組み込まれている 500 社を 1 ヶ月毎に前月から株価が上がった銘柄順に並べ、上から 20%ずつ 5 つのグループに分けた。そして因果チェーンによる企業間の結びつきを使って、前の月に株価がよく上がった上位 20%の企業と因果関係の深いものは何か、また株価が下がった下位 20%の企業と因果関係が深いものを分析しリストアップした。そして、上位 20%の企業と因果関係が強い企業は次の月の株価が上昇すると予想して株を買い、下位 20%の企業と強い関係を示す企業は次の月に株価が下落すると予想して株を売るというポートフォリオを組んだ。この運用結果を見ると基本的にかなり安定していた。だが、同様にポートフォリオ生成をサプライチェーンによる企業間の結びつきを用いて行なうとあまり良い結果は得られなかった。

この結果を踏まえて、個人投資家に向けた因果チェーンの応用も考えられる。例えば、ニュースの内容の波及効果と閉経情報の掲示だ。少子高齢化が株価にどのような影響を与えるかなどを調べたり、昨日株価が下がった原因は何かを検索できるようになる。当たり前な因果関係は経済新聞には書かれていない。つまり、この技術は経済新聞と一般投資家にある業間を埋められるのではないかと考えられる。また、質問応答システムという応用例もある。例えば、「今回の株価下落要因は何？」と聞いたらその答えを言ってくれるサービスも考えられるだろう。

過学習の克服

金融データマイニングの三つ目の潮流は過学習の克服だ。現状ではインデックスに比べると、AI 投信はあまり成績がよくない。この問題点は、過去のデータを基に投資判断する AI が過去にない性質の相場環境の大きな変化にうまく対応できなかったことにある。現在はもう「やってみました」ではなく、信頼できて実務で常に使えるのかということが試されている時期にある。AI が誤った投資判断をしてしまった例には、フェイクニュースによって売り判断をしてし

まいダウが突然 140 ドル以上急落してしまったことがある。人間であればフェイクだとわかるニュースでも機械は過去のデータから組み合わせて予測しているため、過去のデータが役に立たない分野は苦手なのだ。これが過学習の問題である。金融データはノイズが多く、情報密度が不足していて、非定常かつ因果関係を特定しにくい。これらの点をデータ分析とシミュレーションの統合で克服しようとしている。

データ拡張はもともと画像処理の分野で使われ始めた技術で、学習用データが少ないときに基データを回転させたり上下左右反転させたりしてデータを水増しすることである。だが、こういったデータ拡張は金融データでは行えない。そのため、AlphaGo が行った自己対戦をして(深層)強化学習を金融の世界でも行うことができないか試した研究が行われた。過去の実際の株価データを使ってプログラム同士で株の注文を入れ取引のシミュレーションを行った。シミュレーションの中には株価が上がるシナリオや暴落するシナリオなど様々な市場シナリオを再現できる。このようなデータ拡張を行った方が精度が向上した。

まとめ

AI 技術を用いる資産運用は現状、全ての状況で勝てる万能な AI プログラムは開発されるのは難しい。そのため、複数の AI プログラムと組み合わせたり、金融分野の常識を大規模なテキストデータから自動作成したり、自己対戦シミュレーションによる相手のトレード戦略の学習を行うことが今後の発展に繋がるだろう。

海外ではすでに金融機関が AI を用いたデータ分析を行っている。日本でも金融と情報を結びつける人材を育てることが望まれる。

Q&A

Q. 現状の市場（特に日本の株式市場）においては所謂外国人の取引によるインパクトが大きいと思う。 このような状況の中では英語でのテキストマイニングのほうが良い結果が出るということはないか？
A.英語のテキストマイニングはチャンスがある。すでに IBM は英語での因果チェーンを分析している。

Q. 原因の種類によって結果が現れるまでの時間が異なるため、今日のテキストマイニングの結果と、明日の株価のペアで学習させるのは難しいのではないか？

A. 波及効果が現れるまでの時間についてはまだ情報はない。今は 1 ヶ月あれば市場への影響が出尽くすのではと予想しているだけであるため、因果関係のタイムラグについては今後の研究課題である。