金融・経済分析のためのテキストマイニング

# テキストマイニングと金融における評価指標

　金融テキストマイニングにおいて取り組まれているタスクの種類と評価について目的別に紹介．具体的には，予測タスク，関係抽出タスク，検索タスク，トピック分析，文書要約，極性分析の6つについて順に説明．

## 金融予測の評価指標

　金融における予測タスクは，株価や経済指標などの時系列データを予測の対象として，株価等のデータに加えてテキストデータを用いて，予測を行う．当該タスクは大きく分けると以下の2つにわけることができる．また，これら2つのタスクの延長として，運用テストによって評価を行う方法もある．

**【回帰問題と分類問題によるアプローチ】**

|  |  |  |
| --- | --- | --- |
|  | 回帰問題 | 分類問題 |
| 目的 | 将来の値そのものを予測（連続） | 将来の上昇or下降等を予測（離散） |
| 学習手法 | 特徴量作成（テキストマイニング）+ | |
| 機械学習手法：サポートベクター回帰  統計手法：線形回帰 | 機械学習手法：SVM，NN，RF  統計手法：ロジスティック回帰 |
| 評価指標 | 機械学習手法でよく用いられる手法を活用 | |
| 予測誤差（MSE，RMSE） | 正解率，適合率，再現率，F値 |

**【運用テストによる評価】**

　上記2つのタスクで予測を行った結果に基づいて取引行動を行い，運用テストの成績まで含めて検証する評価方法．

例）ポートフォリオ運用における執行コストを考慮したバックテスト

指標としては以下の通りだが，注意点としてはマーケットインパクトを考慮できないこと，全く同一の状況におけるデータは存在しないことがあげられる．

* 損益の計算

名称通り損益を計算し，目的に応じて最終的な損益，平均値などを求め比較する．

* 超過リターン

ベンチマークと比較した超過リターンで評価する．

* 最小リスク

標準偏差やそれをベースとした指標を基に評価する．

* 最大ドローダウン(MDD)

資産が最大のときと，その後の最小のときの最大価格差．

* シャープレシオ
* リスクリターン率

リスクに対するリターンの大きさであり，総損益を最大ドローダウンで割ることで算出．

## 金融テキストからの関係抽出・単語抽出

　金融テキストからの関係抽出・単語抽出タスクは基本的に自然言語処理と同様．ラベルが付与されたテキストを用いて，機械学習などの方法を用いて分類モデルを構築しその精度を評価する．また，関係抽出の一部である因果関係抽出において，因果関係を含むテキスト情報の抽出に取り組む研究もある．6章で紹介．

* 手法

入力：テキストから特徴量を抽出

手法：SVM,NN,RF,ロジスティック回帰等

出力：テキスト内における関係性ラベル

* 評価方法

分類タスク同様．適合率，再現率，F値等

## 金融テキストマイニングと検索タスク

　金融テキストマイニングにおける検索タスクは①類似文章の検索，②金融テキストから獲得した情報を用いた情報検索等があげられる．この手法は人手でラベルを付与したデータを用意する必要があるので，コストがかかる．具体的には，類似度を使用し，文書内の文章をインプットとして，似た文章が含まれた金融テキストを探し出すことで周辺情報の補完や，違った側面からの記述を見つけることができる．また7章では金融文書から取り出した因果関係を連ねた因果チェーンによる情報検索を紹介．

* 手法

BoW等によって文章をベクトル化してコサイン類似度を用いて類似度計算

Word2vecによってベクトル化してWord movers distanceやDoc2vecなどの文章間の類似度を計算する手法を用いて類似度計算

* 評価方法

ランキングの評価をする指標．Precision@N，Recall@N等，テキスト参照．

しかしこれらはラベルを付与したデータが必要⇒実際にはユーザに使ってもらいフォードバックを受ける

## 金融テキストのトピック分析

　文書の単語頻度を測定し，単語の共起性を確率的に分析して指定した数のトピックを求める分析．

* 手法

確率的アプローチによる潜在的意味解析pLSA，LDAなどのトピックモデルと呼ばれる手法（8章）

金融テキストは数値とテキストがセットとなったデータが多く存在するので，sLDAを使用

* 評価方法

人手で確認して判断

## 金融テキストと要約

　金融テキストにおける要約は，アナリストレポートの要約などが挙げられる．

* 手法

評価のためのデータセットが用意できるのであれば，SVM,RF,NNなどの教師あり学習手法が可能

翻訳タスクなどで用いられる深層学習手法の1つであるSequence-to-Sequenceモデルを用いた研究も多い．

* 評価方法

人間が作成した要約との一致度で評価．代表的な指標としてROUGE，BLEUが挙げられる．

やはり評価するためのデータ作成コストがかかる課題があるが，自動要約した文章を修正するなどの手間を省くのに使用可能．

## 極性分析の活用

金融分野における極性分析は経済動向を把握する上で有望な手法．表現の極性を分析し文書の情報を数値化することで，ポジネガを数値として表現する．金融テキストの主な方法は極性辞書によってテキスト極性値を推定し分析することが挙げられる．

* 手法

単語を特徴量として，文書がポジネガを学習，SVM,RF,NN，ロジスティック回帰が適用可能．また，極性値を利用して投資判断の基準とすることで運用テストによる評価・検証も可能．

* 評価方法

適合率，再現率，F値など．運用テストでは超過リターンなど．