seis-ml-api 中間レポート

情報科学類二年 江畑 拓哉 (201611350)

Contents

1 前書き

このレポートは情報特別演習における中間レポートとして編集したものだ。但しここにおける中間レポートの意味は、今後の方針を明確にするための現在までに仮決定された事項をまとめるということにある。即ちここに書かれるものは、これまでに実験した、学習した内容ではなく、それを総括して最終的に今後使用する可能性のあるものに限られる。

更に既存のツールを用いたツールで実験した結果や機械学習の詳細な導出については別紙に実験レポートとして作成する。

2 seis-ml-api 概要

この api（と仮定する）は、この情報特別演習を受けた我々の追加目標であり、私の課題である。我々は 3 人でグループを組んでいるが、それぞれにテーマを設けて研究を行っており、私は二人の研究成果をまとめる他に、それを活用して独自にこの api を作成することになっている。

この api の目指すものは未確定部分が多いものの (詳しい目標については発案者に質問いただきたい) 現在私が作成を行っているものは、”時系列データ分析とランダムフォレストを利用した、相関に基づく時系列予測 ”とでも言うべきものである。そしてこの抽象的なチャートについては以下の通りである。

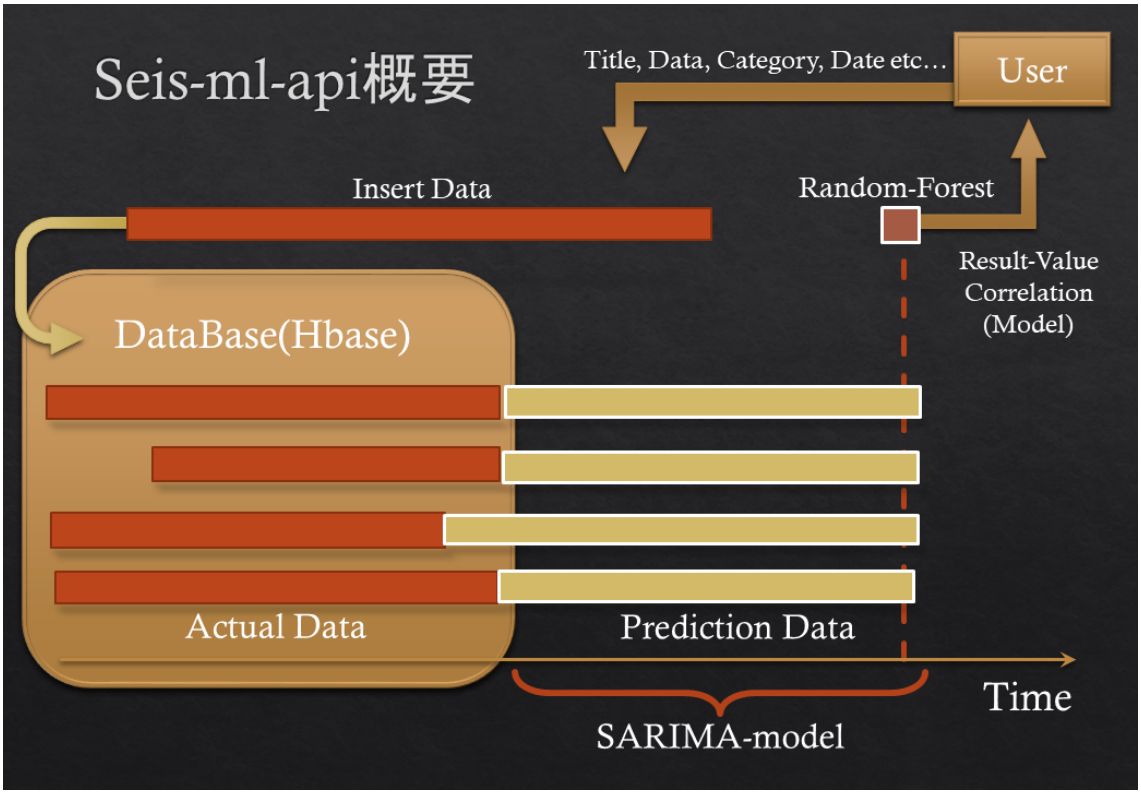

Figure 1: seis-ml-api の抽象的チャート

例えば、ユーザーから 2015 年までの株価データの入った ”(csv) ファイル（タイトル）” と、 ”それの属するカテゴリ” 、 そして必要なデータの 2017 年 12 月といった ”時間データ” 転送されたとする。まずデータベースに登録されている指定されたカテゴリに属するデータについて、 SARIMA モデルを用いて時系列解析を行う。そして指定された時間データまでの値を予測する。次にこれらの予測データに加え与えられたファイルを用いてランダムフォレストによる回帰解析を行い、指定された時間データの値を欠損値補完する。そして得られた ”補完データ” と ”相関についてのデータ (相関の高い順 3 つのタイトルなど) ” をユーザに返し、ユーザから与えられたファイルは指定されたカテゴリに基づいてデータベースに保存する。

つまり必要となるデータは、

- タイトル
- 時系列データ
- そのデータの属するカテゴリ
- 欲しい時間データ

そして返すデータは、

- 補完データ
- 相関についてのデータ

ということになる。

2.1 実験に用いるデータ

今回実験に用いるデータについて明確にするため、ここで今後活用して行くデータの入手先を明確にしておく。これからまず手に入れるであろうデータのカテゴリは、日本の経済系にあるデータで、例えば日経平均株価がそれに当たる。そして挑戦できるならばそれに加えて日本の人口のデータも検討してみたいと考えている。

- google finance
  日本のデータを csv で入手することは困難だが、海外のデータは容易に手に入る。
- Quandl
  ほとんどすべてのデータはここで手に入る。但し、どうやら長期間のデータは乏しいようである。主にはこちらから得た株価データを用いて分析を行う。
- 総務省統計データ
  あまりめぼしいデータはないが、ゼロというわけではないため活用していきたい。

3 機械学習部分（時系列解析）

導出に関しては別紙にまとめて示す。

時系列解析に用いるモデルは季節的自己回帰和分移動平均モデルこと SARIMA モデルを用いる。

SARIMA とは、 Seasonal ARIMA モデルであり、これは季節性のあるモデルと ARIMA モデルを組み合わせたものである。

Seasonal 部分は、ARIMA モデルをある周期区間に適用したもので、これを下記の ARIMA モデルと組み合わせることで、ある程度の周期性を持たせることができる。例えば、毎年同じような経営方針の会社があれば、この Seasonal 部分が組み込まれることで、より現実的な値を返すことができるだろう（そしてこのような周期性はおおよそどのような国でも見受けられるようであることを別紙に記載する実験により予測された）。

ARIMA モデルとは、 Autoregressive Integrated Moving Average Model である。

Autoregressive とは、重回帰のようなアルゴリズムを用いた自己回帰法である。例えば $a- > b- > c$ という遷移があれば、 $b- > c- > d$ であると考えられるようなものである。

Moving Average とは、移動平均という意味であるが、これは簡単には説明しづらいものがあるため、抽象的な考え方で説明を行いたいが、その前にこれが移動平均と呼ばれる所以を説明する。ある区間の平均値を求めた際に、その平均値に対して計算に用いる要素は区間の中心ではなく、区間の最新部分である。つまりは平均値を格納する位置を移動しているといったことが移動平均と呼ばれる所以の一つなのだ。そして、MA モデルの重大な特徴として、このモデルはある時系列データそのものではなく、先程求めた移動平均とそれが格納されている時系列データの要素との差、つまり移動平均誤差について所謂回帰の議論を行うというものである。

4 機械学習部分（ランダムフォレスト）

導出と詳細の式に関しては別紙にまとめて示す。

上の機械学習部分に関して、求めたい時系列タイトルを除いた関連データの時系列解析が終わったところで、呼び出されることになるこのアルゴリズムは、複数の低級な決定木を多く生やすことでデータの分析をするもので、今回はこれを回帰の欠損補完に用いることとする。

ランダムフォレストの説明に移る前に決定木、特にその回帰について説明する。

4.1 決定木

決定木とは、複数の説明変数を持つデータセットに対して、もっとも議論のデータセットを分割できるように分け、分けた要素について再帰的に同様の処理を行うという分類・回帰法であり、これによってどのような条件でどのような目的値がふさわしいのかを求めることができる。この分割における最適化を用いる方法には、例えばエントロピーやジニ係数、尤離度（逸脱度）などの所謂 ”分離度 ” を測る式を用いる。

後に言う、高い・低い決定木とはこの分割の回数を意味し、高い決定木であればあるほ

どより細かい分類・回帰が可能となり、低い決定木では大まかな概要を掴むことができる。

4.2 ランダムフォレスト

ランダムフォレストとは、与えられたデータセットの中から任意のデータを抽出して集めた複数のデータセットについて低い決定木を並行して行い、それによって求められた結果の集合に基づいて元のデータセットの分析をするというものである。今回の回帰を用いた欠損補完においては、決定木のうちの回帰木を用いて求まった値の平均を取りそれを参考にすることで欠損補完を行う。そして、その結果と計算に用いられなかった残りのデータを用いることで説明変数の重要度を探ることができる。

5 データベース部分

データベース部分に関しては Apach HBase を用いた大規模スケールの箱を作る予定である。大規模データベースの中身の設計については割愛する（データの安全保障や機械学習の制度などの問題から、データを増やし過ぎることが厳しい）が、 hbase のクエリと csv のファイルとつなげる部分や、同一データの判定などの機能を作成する必要があるためデータベースとの接合に関しては未だ展望が見えていないのが現実である。

6 参考文献

以下にそれぞれで用いた参考文献を示す。なお、これらの文献は今後より深く読み進めていく予定である。

- SARIMA モデルについて [12] [17] [14] [16] [1]
- RandomForest について [8] [19] [7] [11] [4] [10] [6] [3] [15] [2]
- 決定木について [18]
- データベースについて [13] [9] [5]

References

[1] Carolia Garcia-Martos Andres M. Alonso. *Time Series Analysis*. URL: `http://www.etsii.upm.es/ingor/estadistica/Carol/TSAtema4petten.pdf`.

[2] Teppei Baba. 機械学習ハッカソン：ランダムフォレスト. SlideShare. URL: https://www.slideshare.net/teppeibaba5/ss-37143977.

[3] Leo Breiman. "Random Forests". In: *Mach. Learn.* 45.1 (Oct. 2001), pp. 5–32. ISSN: 0885-6125. DOI: 10.1023/A:1010933404324. URL: http://dx.doi.org/10.1023/A:1010933404324.

[4] Hemant Ishwaren Fei Tang. "Random Forest Missing Data Algolithms". In: (Jan. 2017). URL: https://arxiv.org/pdf/1701.05305.pdf.

[5] *HBase Tutorial*. tutorials point. URL: https://www.tutorialspoint.com/hbase/index.htm.

[6] *Imputation with Random Forests*. Cross Validated. URL: https://stats.stackexchange.com/questions/49270/imputation-with-random-forests.

[7] *Imputing missing values before building an estimator*. scikit learn. URL: http://scikit-learn.org/stable/auto_examples/missing_values.html.

[8] Satoshi Kato. *Imputation of Missing Values using Random Forest*. SlideShare. URL: https://www.slideshare.net/kato_kohaku/imputation-of-missing-values-using-random-forest?ref=http://kato-kohaku-0.hatenablog.com/entry/2016/05/01/155908.

[9] Christopher Miles. *All Your HBase Are Belong to Clojure*. Jan. 2012. URL: https://twitch.nervestaple.com/2012/01/12/clojure-hbase/.

[10] "Random Forests". In: (). URL: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf.

[11] *rfImpute*. RDocumentation. URL: https://www.rdocumentation.org/packages/randomForest/versions/4.6-12/topics/rfImpute.

[12] George Athanasopoulos Rob J Hyndman. *Forecasting: principles and practice*. Oct. 2017. URL: https://www.otexts.org/fpp/8.

[13] David Santiago. *clojure-hbase*. June 2014. URL: https://github.com/davidsantiago/clojure-hbase.

[14] The Pennsylvania State University. *Seasonal ARIMA models*. 2017. URL: https://onlinecourses.science.psu.edu/stat510/node/50.

[15] *What is the proper way to use rfImpute? (Imputation by Random Forest in R)*. Cross Validated. URL: https://stats.stackexchange.com/questions/226803/what-is-the-proper-way-to-use-rfimpute-imputation-by-random-forest-in-r?rq=1.

[16] 時系列解析. URL: https://upo-net.ouj.ac.jp/tokei/contents/sub_contents/c01_06_00.xml.

[17] 時系列解析__理論編. Logics of Blue. June 2017.

[18] 決定木. Matlab. URL: https://jp.mathworks.com/help/stats/classification-trees-and-regression-trees.html#bsw6a62.

[19] 石岡恒憲. *Random Forest* を用いた欠測データの補完とその応用. 大学入試センター研究開発部, Nov. 2010. URL: http://www.rd.dnc.ac.jp/~tunenori/doc/jjasRf2010slide.pdf.