seis-ml-api 中間レポート (再編)

情報科学類二年 江畑 拓哉 (201611350)

Contents

1 前書き **1**

2 seis-ml-api 概要 **1**
2.1 実験に用いるデータ 4

3 機械学習部分（時系列解析） **4**

4 機械学習部分（相関解析） **5**
4.1 決定木 5
4.2 ランダムフォレスト 5

5 データベース部分 **5**

6 参考文献 **6**

1 前書き

このレポートは “江畑拓哉個人” が作成した情報特別演習における中間レポートである。ここで記載されているものは、現在までの成果を出した進捗に限り、それまでの失敗した試行などについては記載されていない。

また既存のツールを用いて実験した結果や、機械学習やデータベースの細かいプログラムについては完成次第別紙に実験レポートに作成する。

最後にこのレポートに関する意見などに関しては混乱を避けるため、できうる限り直接ご指摘頂ければ幸いである。

2 seis-ml-api 概要

この api (と仮定する) は、この情報特別演習をグループで受講した我々（江畑, 栗本, 畑中）の最終的な目標である。我々はまず各々の学ぶ分野についての理解を深め、その成果を適宜持ち寄ってこれを作成する予定になっている。

Table 1: 主な役割について（但し互いに柔軟に協力し合う）

氏名	分野	内容
畑中	データベース	大規模データベースの作成
栗本	機械学習	機械学習の理解
江畑	上記２つの統合	大規模データと機械学習を利用する手法の学習

つまり、我々の情報特別演習のゴールは二段階あり、一段階はそれぞれの学習分野の習得、二段階はそれらを持ち寄ってこの seis-ml-api という何かの作成を行う、というものである。

問題となる api の詳細についてだが、機械学習が何を返すことができるものであるのかを全員が共有できていないため、詳細な情報を明記することはできない。しかし目標を理解しやすくするため、無理にこれについて個人的な見解に基づいた説明を行う。

この api では以下の抽象的なチャートに基づいた設計を行う予定である。

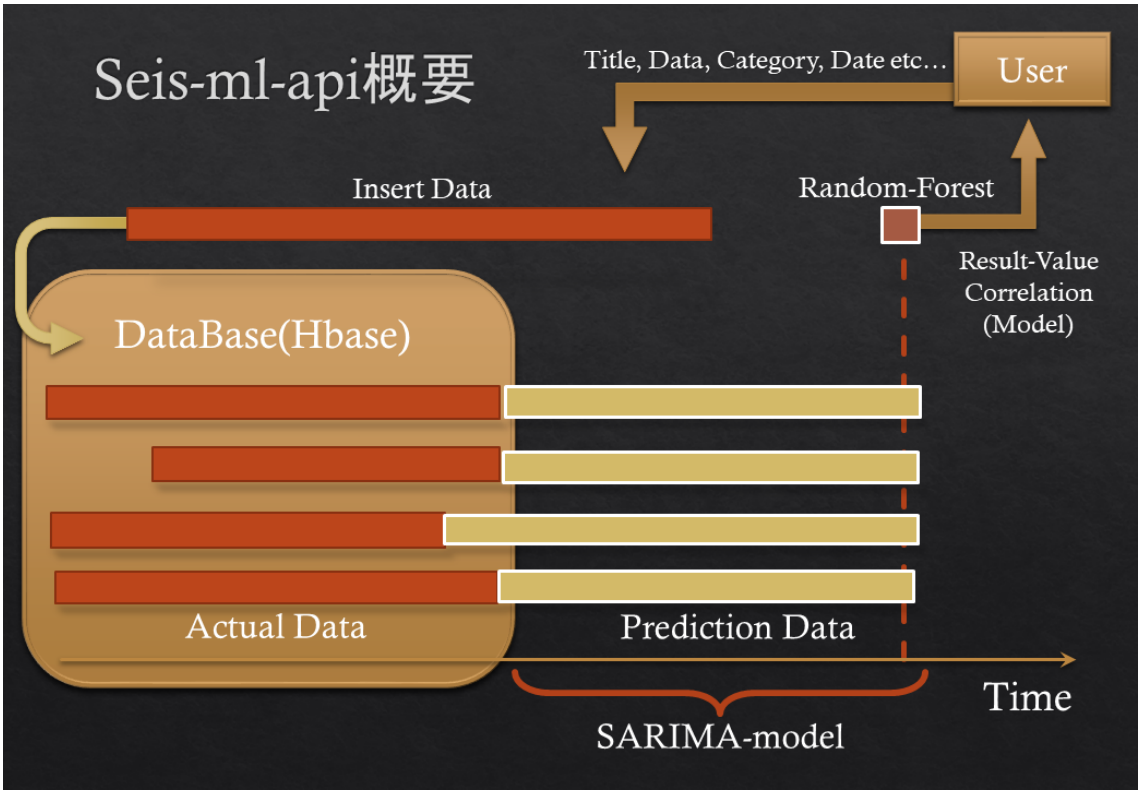

Figure 1: seis-ml-api の抽象的チャート

ここで注意しなければならないのは、この案に関してはデータベースに Hbase を利用するという点以外に関しては江畑が独断で学習した成果によって作成されたものであり、機械学習分野のこれからの成果次第では作成されるものも、またこの制作物の利用法も十

分に変わり得るということである。
その前提に於いて、現在までのこの物の利用価値について説明するならば、恐らくこのapi は “大量の時系列データを利用した価格予測 api” と呼べるものなのではないだろうか。ユーザから時系列データを受取り、データベースに保存されたデータと関連付けて機械学習を行い、ユーザから指定された時点の予測値や相関の強いデータの名称を返す、というのが現時点での江畑の理解である。
そしてこの構成要素は、大きく分けて機械学習とデータベースの二つである。
データベースに関しては畑中氏が作成している、大規模データを扱うことのできる HBase を用いる予定である（詳細は彼と彼の担当の教員に尋ねて頂きたい）。江畑はこのデータベースに入手したデータ、それを予測したデータ、後述する SARIMA 予測をする際に用いるモデルなどを入れるつもりである。
機械学習部分では、まずデータベースから問題となる時系列データについて関連性のあるデータを読み出し、それらについて時系列予測を行う。そしてそれらを元にして問題となった時系列データとの相関を調べ、いわば欠損値となっているユーザから指定された時点での値を求める。そしてその際に求まる相関の強かった関連データの名称も合わせてユーザに返すことがこの部分での実装予定のものである。

これだけでは抽象的で理解が難しいため、江畑の考えている動作の流れを以下に示す。
例えばユーザから、現在までの時系列データがそれの所属するカテゴリ付きで渡されたとする。まずデータベース側から、指定されたカテゴリに関するデータが機械学習側に渡される。機械学習側ではそれらをユーザから指定された時点（例えば一週間後）まで時系列予測する。次に、ユーザから渡されたデータに対して先ほど予測したデータ群との相関を求めていく。その相関を用いて指定された時点の値を予測する。結果としてその値と相関の強かったデータの名称群を返す。最後にユーザから渡されたデータをデータベース側に保存して一連の動作は終了となる。

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
- 欲しい時点についての情報

そして返すデータは、

- 予測値

- 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

ここでは、初期にデータベースに入っているデータとして挙げられ、なおかつ実験段階に於いて使用できると考えられるデータの入手元について紹介する。江畑個人の考えとしては、データの相関を求める関係上、これらの入手元から日本国内の経済についてのデータを集めて行きたいと考えている。

- 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予定である。

- 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と考えている。

3 機械学習部分（時系列解析）

この章に関する内容は全て江畑個人の報告であり、他のメンバーの活動に何ら影響を与えるものではない。

実験に関しては、別紙にまとめて示す。(仮決定のこの部分のみの実験データは同フォルダの report.ipynb である)

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 モデルを用いる予定である。

SARIMA モデルとは、三つの要素が重ね合わさったモデルである。

まず、AR という部分は Autoregressive を表し、これは自己回帰を意味する。自分の以前の観測データに対して重回帰分析を行うものである。例えば、 $a->b->c$ という遷移があれば、 $b->c->d$ といったことを考えられることに似ている。

MA というのは Moving average 、つまり移動平均を意味している。移動平均とは、ある区間 [a, b] の平均値と b 或いは a などを比較する際に用いられる言葉のようで、平均値を平均の位置ではない別の位置の値と比較することを意味している。

ARIMA というのは、以上の 2 つを組み合わせるという意味である。

そして S というのは、Seasonal、季節性という意味で、ある期間の周期性を用いるとい

うことである。これは例えば毎年同じような活動をするものに対して非常に有効な手段であるようで、 ARIMA モデルの拡張の 1 つとして広く認められているようである。

4 機械学習部分（相関解析）

この章に関する内容は全て江畑個人の報告であり、他のメンバーの活動に何ら影響を与えるものではない。

実験に関しては別紙にまとめて示す予定である。

概要で紹介したように、関連データについての時系列解析が終わった後に行う処理がこの相関関係を解析する部分である。江畑はここではランダムフォレストの回帰を用いた解析のうちの 1 つ、欠損値補完を行いたいと考えている。ランダムフォレストの大まかなアルゴリズムは以下で紹介する決定木の低いものを多く生やすことでデータの分析を行うもので、特に今回は回帰木を用いる。

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、最も議論のデータセットを分割できるように境界を設け、そこで分割されたそれぞれのデータセットに対して同様の処理を繰り返していくことで、データの特徴を抽出していく機械学習の手法の 1 つである。データセットの分割に用いられる指標として、尤離度（逸脱度）やジニ係数、エントロピーなどを挙げることができる。またここで用いる決定木の高さとは、あるデータに対してどの程度分割処理を行ったか、というものである。そして分割数が多いものは高い決定木、分割数か少ないものは低い決定木と呼ぶこととする。また当然のことながら、決定木は低ければ大まかな予測が可能であり、高い場合には精度は上がるものの、過学習を起こす可能性もある。

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に抽出して集めたデータセットを複数作り、それぞれに低い決定木を用いた学習を行い、結果を集計することで元のデータセットの分析を行うという仕組みのことである。今回の回帰を用いた欠損補完においては、それぞれの決定木が求めた値の平均を取ることで欠損値補完を行う。そして、決定木で学習しなかった残りのデータを用いて説明変数の重要度を分析する。

5 データベース部分

この章に関する部分のうちデータベースの選択、作成に関しては畑中の貢献によるものであり、江畑は何も関与していない。データベースの利用方法については江畑が独自に行ったものであり、他のメンバーとは共有していない事項である。

データベースの作成部分に関しては 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詳細な設計については江畑の理解が追いつくものでもなく、機械学習の手法次第では挿入するデータに大きな変化がある可能性があるが、畑中の報告を伺う限りでは数 TB 程度の完全分散システムにするとのことであった。

データベースの利用については、Python と Clojure での Hbase の利用方法についての学習を行った。しかし前者はしっかりとしたモジュールがあったが、後者は自信をもって選択できるものがなかった。そのため、この言語の特性を活かして Java からの利用を目指し、できるならば自作の独自のクエリ（例えばｃｓｖ読み込み）などを実装したモジュールを作成したいと考えている。

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] [18] [14] [17] [1]
- RandomForest について [8] [20] [7] [11] [4] [10] [6] [3] [15] [2]
- 決定木について [19]
- データベースについて [13] [9] [5]
- その他 [16]

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] 重穂 中村. "論文に於ける「だ」と「である」の選択条件に関する試行的考察". In: (Dec. 2009). URL: https://eprints.lib.hokudai.ac.jp/dspace/handle/2115/45684.

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

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

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

[20] 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.

Emacs 25.2.1 (Org mode 9.0.9)