［報 告］計算科学・計算機科学人材育成のためのスーパーコンピュータ無償提供利用報告

東北大学大学院データサイエンスプログラムにおける疑似プロジェクト演習

中尾光之・瀬川悦生・山田和範
東北大学大学院情報科学研究科

1. 概要

2016 年 7 月 27 日から 29 日の 3 日間において，東北大学大学院情報科学研究科が主催するデータサイエンスプログラムにて，講義の一環として DSP Training camp II および Big data challenge を行いました．この実習では，修士課程の学生がトレーニングを受ける対象，博士課程の学生がトレイナーとして，現実社会に存在する，ビッグデータを用いた疑似的な研究プロジェクトに取り組みました．

授業のテーマのひとつがビッグデータであり，大量のデータを扱う必要がありましたので，サイバーサイエンスセンターによって提供されるスーパーコンピュータ無償提供制度を利用させていただきました．

実際の実習では，受講者 18 名が 3 つのチームに分かれ，計算言語学，生命科学，経済学に関連する問題を，提供していただいたスパコンをはじめとする計算機を駆使することで解決しました．今回，実習で扱った研究テーマは答えが用意されていない発見的なものでしたが，チームによっては，これまでに明らかにされていなかったような結果を出すことができ，実習が終わった現在においても，さらに発展的に研究を進めています．

以下では，その内容について簡単にご報告いたします．

2. 取り組んだ問題

取り組んだ問題は以下の 3 つです．

A) 言語学の問題：1 年間にわたり取得された日本語のツイッターの時系列データから，1 日毎における社会的な雰囲気を抽出し，ベクトル化，その社会的な雰囲気が，日本の主要会社の株価，日経平均 225 のリーディングインジケーターとなり得るかどうかを検証しました．また，リーディングインジケーターとなり得る場合，その時系列データを利用して，株価の変動の予測器を構築することとしました．

B) 生命科学の問題：タンパク質立体構造中にディスオーダー領域と呼ばれる，水中で他の確固たる構造部位より変動が大きい部位があります．この部位は何らかの生物学的な機能を

C) 有する場合が多く，タンパク質立体構造科学的に重要な領域です．この領域を，これまでに得られた既存のデータから予測する予測器を，深層学習法を用いて構築しました．

D) 経済学の問題：株価の変動を意味するボラティリティという指標があります．ボラティリティが大きいということは，株式取引を行う際に，その分，損をする可能性は大きくなるものの大きなキャピタルゲインを得ることもできることを意味します．その点において，ボラティリティを精度よく予測するモデルは重要です．この問題では，これまでに提案されているいくつかのボラティリティのモデルを日経平均に適用し，どれが最も良くボラティリティを説明するのかを検証しました．

3. 実習の様子および成果

各チームとも，チーム内で役割分担を決め，効率的に仕事を進めました．特に扱うデータ量が大きかったのは，言語学の問題を扱ったチーム A でした．圧縮されたツイッターのデータだけで約 700GB です．その処理にはスパコンの 1 ノードをフル活用したにもかかわらず，約 21 時間もの計算時間が必要でした．このような処理はサイバーサイエンスセンターからスパコンを提供していただかなければ為し得なかったことでありましたので，今回使わせていただいて良かったと思います．その他のチームも慣れないスパコンのジョブ管理システムの使い方に戸惑いながらも，何とか使いこなし，与えられた問題のみならず各チームが発展的な研究成果を出しました．

また，得られた全ての研究成果は 2016 年 8 月 9 日に東北大学とオハイオ州ケースウエスタンリザーブ大学との共同ワークショップにて発表しました．

実習の様子. 各々の端末からスパコンに SSH アクセスをしています.

4. 最後に

本実習は，サイバーサイエンスセンターのスーパーコンピュータを利用することではじめて実現することができました．また，実習中には技術的なサポートを頂きましたことを感謝申し上げます．