5V-02

クラスタリングを用いた時系列回帰モデル化手法の提案

高橋 佑里子† 鈴木 成人‡ 山本 拓司‡ 福田 裕幸‡ 小口 正人†
†お茶の水女子大学 ‡富士通研究所

1 はじめに

近年のクラウドサービスにおいて物理サーバ (Physical Machine: PM) の CPU 使用率は低く，そのパフォーマンスを十分に発揮できない状態が続いている．これを改善すべく，事業者では，サーバを仮想化することで使用率を向上させ，PM 数を削減する取り組みが行われている．この取り組みでは，PM が自身の CPU 資源を超えた CPU を割り当てられるオーバコミット状態に陥ることで，仮想サーバ (Virtual Machine: VM) の性能が低下する可能性がある．そのため，図 1 のようにあらゆる VM の CPU 使用率を予測し，値が上昇する前に VM を別の PM へマイグレーションする等の制御を行う必要がある [1]．しかし，現状の CPU 使用率の予測モデルは汎用性が低く，特定の VM に対しては高い精度で予測できる一方，その他の VM に対する予測精度は低くなるという課題がある [2]．

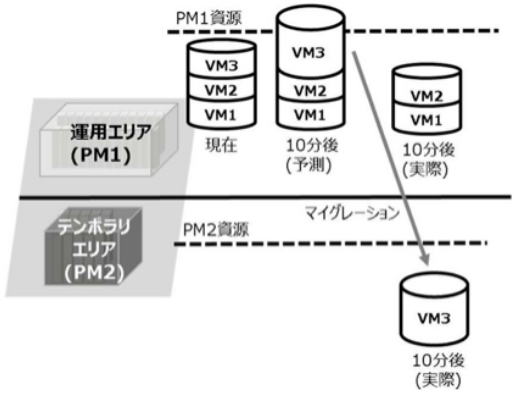

図 1: VM 制御のイメージ

本研究では，VM の CPU 使用率の汎用的な深層学習予測モデルの生成に向けて，時系列データの回帰モデル化手法を提案する．深層学習モデルは学習時間が長いため，なるべく少ない再学習での精度向上が求められる．そこで，データを何らかの方法で適切に選定することで，少ないデータでの再学習で既存モデルの回帰精度向上の施策を検討した．長い時系列データのままでは，概形で特徴が判断されるため良い選定を行うことが難しい．方法を模索した結果，時系列データを学習に必要な長さに細かく分割し，それらをクラスタリングした結果を元にデータを選定することで，通常よりも少ない再学習でのモデル回帰精度向上に成功した．

2 関連技術

本研究では，深層学習ライブラリとして TFLearn[3] を使用し，モデルには RNN を長期依存が可能かつ計算量が比較的少なくなるよう改良した GRU を採用した．モデルの再学習には，学習済みのモデルの上層部を一部再学習させるファインチューニングという手法を用いた．

3 実験

本実験では，VM の CPU 使用率時系列データセットとして，Bitbrains IT Services Inc.[4][5] が公開しているものを使用した．前処理として，2000 個のデータセットを平滑化，正規化処理を行った後，階層型クラスタリングを行った．結果は図 2 のようになった．() 内の数字は，データの個数である．この結果から，学習元データセット (以下 A とする) として赤色部分右側の山に分類されたものの中から 103 個，A と似ていないターゲットデータセット (以下 B とする) として赤色部分左側の山から 103 個のデータを選んだ．

まず，A,B を学習に必要な長さに細かく分割し，それらを k-means 法クラスタリングにより 10 種類に分類した．結果は図 3 のようになった．

この結果に基づいて，A で事前学習したモデルに B を使用してファインチューニングを 2 種類の方法で行った．方法 1 は，多く分類されたクラスタのデータを使用してファインチューニングを行うというもので，図 3 でクラスタ No.1 に B 全体の 18.0%，No.2 に B 全体の 16.9%，No.7 に B 全体の 48.2%と，B のデータが多く分類されたため，上記 3 つのクラスタのデータを使用して行った．方法 2 は，B の各クラスタのデー

Proposal of Time Series Regression Modeling Method Using Clustering
†Yuriko Takahashi
‡Shigeto Suzuki
‡Takuji Yamamoto
‡Hiroyuki Fukuda
†Masato Oguchi
†Ochanomizu University
‡FUJITSU LABORATORIES LTD.

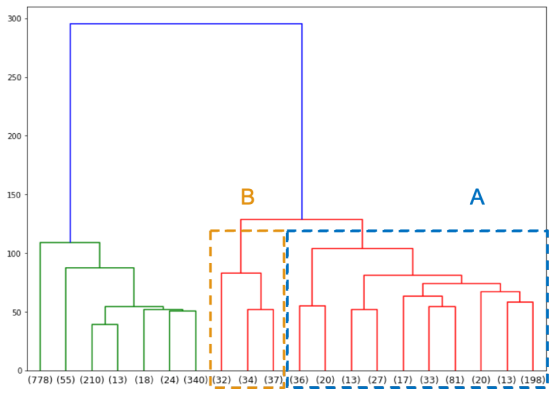

図 2: データセット全体の階層型クラスタリング結果

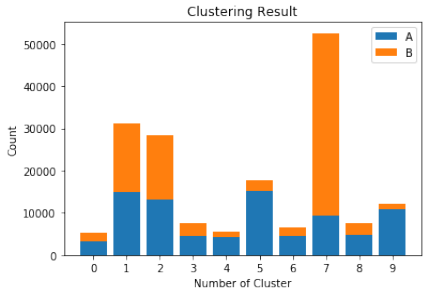

図 3: A,B の k-means 法クラスタリング結果の内訳

タを一定の割合で使用してファインチューニングを行うというもので，割合を 20%刻みで段階的に変更して行った．

そして，A で事前学習したファインチューニングを行う前のモデル，方法 1,2 でファインチューニング行った後のモデル，および B そのもので学習したモデルのそれぞれで B を予測し，それらの回帰精度を比較することで，適切なファインチューニングの方法を検討した．精度の評価指標には，RMSE の平均値を用いた．この値が小さいほど，精度が良いということになる．

4 実験結果

実験結果は図 4 のようになった．方法 1(緑色) の No.1, No.2, No.7 の結果からは，ターゲットデータセットであっても偏ったデータを使用してファインチューニングを行うことで，精度が低下するということが読み取れる．また，方法 2(赤色) の結果からは，ターゲットデータセットを 40%以上一定の割合で使用してファインチューニングを行うことで，ターゲットデータセットすべてを使用して学習したモデルと同等な精度にまで向上するということが読み取れる．

再学習で使用するデータ数と精度の向上という点で考えると，方法 2 の 40%の場合が最も良いということが分かる．方法 1 の No.1,2,7(B 全体の 83.1%) の場合も精度は向上しているが，方法 2 の結果と比較すると，データ数と精度の両面で劣る．これより，クラスタ間のデータの分布傾向の違いを使ってファインチューニングを行うことが重要であるということが分かる．

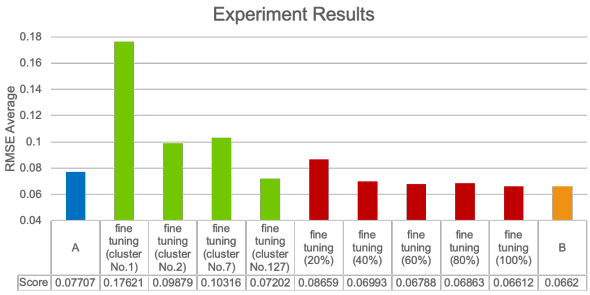

図 4: 実験結果

5 まとめと今後の予定

モデルの汎用化に向けて，ターゲットデータセットでの少ない再学習で回帰精度を向上させる方法を検討した．実験の結果，学習に必要な長さで細かく分割したターゲットデータセットを分類し，各クラスタのデータを 40%ずつ使用してファインチューニングを行うという方法により，モデルの汎用化に向けた少ない再学習で精度が向上することを確認した．

今後は，ファインチューニングを行う範囲を変えたり，新たなデータの選定方法を考えたりしながら，モデルの汎用化に向けてさらに実験を進めていきたい．回帰の次の段階として，予測の精度向上に向けた取り組みも進めていきたいと考えている．

謝辞

本研究の一部はお茶の水女子大学と富士通研究所との共同研究契約に基づくものである．

参考文献

[1] 児玉宏喜, 鈴木成人, 福田裕幸, 吉田英司: マイグレーションを利用したデータセンタの高効率運用手法の提案とオーバコミット時における VM の性能評価, 情報処理学会論文誌 (2018)

[2] 鈴木成人, 児玉宏喜, 遠藤浩史, 福田裕幸: JIT モデリングによるサーバ負荷予測手法の検討と評価, 電子情報通信学会ソサイエティ大会 (2018)

[3] TFLearn: http://tflearn.org

[4] Siqi Shen, Vincent van Beek, Alexandru Iosup: Statistical Characterization of Business-Critical Workloads Hosted in Cloud Datacenters, CCGrid (2015)

[5] http://gwa.ewi.tudelft.nl/datasets/gwa-t-12-bitbrains