

近傍データ収集法を用いた SW-SVR の改良

鈴木雄也^{†1} 兼田千雅^{†2} 峰野博史^{†1}

モバイル通信技術の発達によって、温度や湿度等多様なセンサデータの収集、分析が可能となった。しかし、これらの微気象データは異なるデータ同士で複雑な相関を持っているだけでなく、時間変化とともにデータの特性も変化するため、従来の機械学習を適用することに課題があった。本論文では、当研究室で研究開発してきた時系列データの高精度な予測を実現する SW-SVR へ、予測対象データと類似したデータのみを利用する近傍データ収集法の適用を検討した。札幌、東京、浜松、那覇の 4 地域におけるアメダスの過去 3 年分のデータを用いて評価した結果、SVR に対し予測誤差を最大 78%、モデル構築時間を最大 57%削減できることを確認した。

SW-SVR improved by short-distance data collection method

YUYA SUZUKI^{†1} YUKIMASA KANEDA^{†2}
HIROSHI MINENO^{†3}

Mobile wireless sensor network has recently attracted considerable attention and it is possible to get various environmental parameters, such as temperature, humidity, and pressure, wherever any time. But these micrometeorological data have complex correlation among different data and the characteristics change according to the time elapsed. Thus, it is difficult to predict these meteorological data by using existing machine learning algorithms. In this paper, we propose a short-distance data collection method (SDC) to extract critical data automatically for high-accuracy micrometeorological data prediction. We applied it to SW-SVR, which is our previously proposed improved support vector regression algorithm, and revealed that the prediction accuracy was improved by maximum 78%, and the calculation time was reduced by maximum 57% in SW-SVR with SDC.

1. はじめに

近年モバイル通信技術の発達によって、スマートフォンやセンサデバイスが急速に普及している。温度や湿度、加速度等の時系列データを気軽に収集することが可能となり、これらを用いたビッグデータ活用が様々な分野で注目されている。また、分析結果を用いたアプリケーションの開発も活性化している。これまでにセンサデータを用いた予測・推定をするために機械学習を用いた研究がされてきたが、微気象データを始めとする時系列データは異なるデータ同士が複雑な相関を持っているだけでなく、時間経過とともにデータの特性が変化するため、ある期間の学習データで構築した予測モデルを長期的に運用すると予測誤差が時間経過とともに大きくなってしまいう課題があった。また、これらの課題を解決するために大量の学習データを学習させる方法等[1]があるが、大規模なデータの収集や、長時間にわたるモデルへの学習を定期的の実施しなければならず依然として、運用上大きな課題がある。

そこで本論文では、当研究室で研究開発してきた時系列データの高精度な予測を実現する SW-SVR[2]へ、予測対象データと類似したデータのみを収集する近傍データ収集法

の適用を検討する。SW-SVR は Support vector machine (SVM)[3]の応用アルゴリズムであり、何らかの規則変動のある時系列データに対して、適切な学習データ量を自動的に抽出し、予測精度が向上するよう自動的に予測モデルを構築し続けることのできるアルゴリズムである。一方で近傍データ収集法は、母集団と予測対象データとの標準偏差を基準とした距離を元に、予測対象と類似したデータのみを収集するアルゴリズムである。近傍データ収集法を SW-SVR へ適用することで、適切な学習データのみを用いることによる予測精度向上と、学習データ量の削減によるモデル構築時間向上が見込まれる。提案手法を気象庁の地域気象観測システムアメダスが提供している過去 3 年分の気象データを用いて、札幌、東京、浜松、那覇の各季節における 1 時間後の気温予測実験を実施し、予測誤差とモデル構築時間を評価することで、近傍データ収集法を用いた SW-SVR の性能を定量的に示す。

以下第 2 章で機械学習を用いた予測手法やアプリケーションの関連研究について紹介し、第 3 章で提案手法について述べる。第 4 章で近傍データ収集法のデータ収集性能の評価を示し、第 5 章で近傍データ収集法を SW-SVR に適用した場合の予測精度とモデル構築時間について述べる。最後に第 6 章で本論文のまとめを示す。

^{†1} 静岡大学大学院情報学研究科
Graduation school of informatics Shizuoka University

^{†2} 静岡大学大学院情報学部
Faculty of informatics Shizuoka University

2. 関連研究

近年、モバイルデバイスやセンサデータを対象にした機械学習を用いた将来予測や推定に関する研究は盛んに行われている。モバイルアプリケーションやセンサネットワークを用いた制御システム等への利用のために機械学習の基礎研究がされているが、これらの研究の目的を主に高精度な予測に関する研究と高速な学習モデル構築に関する研究の2つに区分して紹介する。

高精度な予測に関する研究は、データに前処理を加えたり、大規模なデータを学習させたりすることで高精度な予測を可能としている。前処理に関する研究に Adaboost [4] を応用した Boosting-SVM with Asymmetric Cost algorithm がある[5]。Adaboost を用いてデータの重みを逐次的に変化させて学習器を修正して利用することで、予測に最適な重み付けのされたデータを学習したモデルを利用できる。実験の結果、多様なオープンデータで高精度な予測ができることを示した。一方で、125 万件の大規模学習データを利用して高精度な Artificial neural network (ANN)による気温予測モデルを構築し評価を行った研究がある[1]。実験の結果大規模な学習データを用いてモデルを構築することで、Mean absolute error (MAE)の評価値で 1 時間後の気温を 0.525 度の誤差で予測することに成功した。他にも SVM を用いて土壌水分量や、気温を高精度に予測する研究などがある[6][7]。これらの研究は高精度な予測を実現しているが、モデル構築時間に関しては考慮されていない。そのため学習や前処理に大規模な計算コストが必要であるものが多く、システムでの運用には課題がある。

一方で、高速な学習を実現するために SVM の最適化問題の高速化や、学習データの削減、並列分散処理を利用している研究がある。SVM の最適化問題の高速化について、問題を計算幾何学問題に帰着し、コアセットを用いた計算幾何学問題の高速近似解法を流用する Core Vector Machines (CVM)がある[8]。実験の結果、データ数が大きくなるにつれて SVM と同等の予測誤差を示しながら CVM の方が高速にモデルの構築が可能であることを示した。また研究[1]と同様の学習データ 125 万件から、ランダムサンプリングで学習データを一部抽出し、学習データの全体数を削減することでモデル構築時間の高速化をした研究がある[9]。ランダムサンプリングによって学習データを減らしつつ、パラメータチューニングによって最適化された SVM を用いることで研究[1]に示された ANN モデルの予測精度と同等の予測精度を示すことに成功した。他にも学習に不要と予想されるデータを削除する手法を考案し、モデル構築を高速化した研究[10]や、Twister[11]の並列分散処理を利用して、モデル構築時間を削減した研究もある[12]。これらの研究は学習の高速化についてのみ検討しており、予測

精度は従来の SVM と同等か僅かに低い値を示している。そのため、高精度な予測精度を求められるアプリケーションへの利用には課題がある。

以上のことから高精度な予測や高速なモデル構築についての研究が盛んに行われているが、実用化については依然として課題が存在する。そこで本研究では予測結果を用いたモバイルアプリケーションやモデル予測制御システムへの応用を想定し、予測対象に対して必要となる学習データのみを収集できる近傍データ収集法を SW-SVR に適用することで、予測精度の向上とモデル構築時間削減の両立を目指す。

3. 提案手法

3.1 Sliding window-based support vector regression

Sliding window-based support vector regression (SW-SVR) は、時系列予測に特化した Support vector regression (SVR) の改良アルゴリズムである。SVR は SVM を回帰アルゴリズムに応用した機械学習の一種であり、現在知られている多くの手法の中で最も認知性能が優れた学習モデルのひとつである。SVM はカーネル関数を用いて各データ点を高次元へ写像し、線形分離可能な問題へと変換する。その後各データ点との距離が最大となるようなマージン最大化超平面を用いて分類を行うことで、汎化性能の高い学習器を構築できる。一方で SVM の計算量は $O(n^2) \sim O(n^3)$ であるため、大規模なデータの学習には時間がかかるという課題がある。そのためシステム上で運用するためには学習データを適切な量に設定しなければならない。

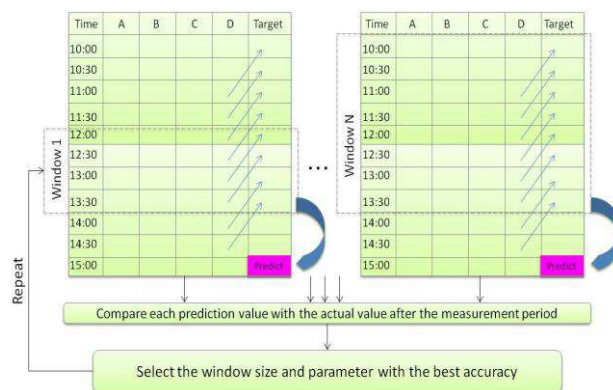


図 1 SW-SVR の概要図

SW-SVR の概要図を図 1 に示す。SW-SVR は母集団の中から予測対象データに対して適切なデータのみを選択してモデルを構築し、時間変化によってデータの特徴が変化し学習器の予測誤差が閾値を上回った場合に再構築をするというサイクルを繰り返す、時系列データ予測に特化したアルゴリズムである。SW-SVR を利用することで、大規模なデータを用いた汎用モデルを作らなくても、対象データに適した部分データの利用と定期的なモデルの再構築によっ

て、短時間で高精度な時系列データ予測が可能となる。

SW-SVR はデータ入力部、部分集合抽出部、モデル構築部、モデル再構築部の 4 つから構成される。データ入力部では一定の学習データ X を入力する。部分集合抽出部では入力された学習データ X をウィンドウサイズ N に分割する。ウィンドウサイズの分割はデータ量や季節、時間帯等様々な方法が考えられる。モデル構築部では N 個の分割された学習データを用いてそれぞれ学習器を構築し、最も予測精度の高いものを選択する。モデル再構築部では、予測モデルの予測値を計測周期毎に発生するセンサデータと比較し、定めた閾値を超えるかどうかを判定する。気象データのような時系列データは時間経過とともに特性が変化するため、特に SW-SVR のようなある予測時刻に特化したモデルを構築した場合、時間経過とともに学習器の予測誤差が大きくなる傾向がある。予測誤差が一定値を上回った場合は、最新のセンサデータを高精度に予測できる学習データパターンを再探索し、最新データに特化したモデルを再度構築する。以上の 4 サイクルを繰り返すことで、時系列データの特性変化に対応した予測モデルを構築し続けることが可能となる。

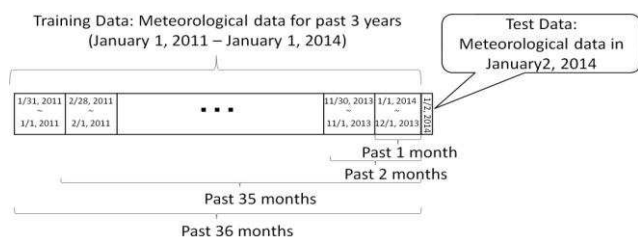


図 2 データの分割方法

例として、図 2 にデータ量を基準に過去何件分のデータを利用することが適切であるかを探索するアルゴリズムを適用した場合の動作を示す。学習データ量 X が 2011 年 1 月 1 日から 2014 年 1 月 1 日までの 36 ヶ月分で、 N が 36 に設定されたとする。予測対象が 2014 年 1 月 2 日の場合、予測対象日から直近 1 ヶ月毎に $X/36$, $2X/36 \dots 35X/36$, X といったようにデータを 36 分割する。その後並列処理を用いて 36 種類の予測モデルを構築し、最も予測誤差の少ない予測モデルをモデル構築部で選択する。その後モデルを運用し続けながら周期的に発生するセンサデータと予測値を比較し、予測誤差が一定値を上回った場合に再構築を行う。

SW-SVR は何らかの規則変動のある時系列データに対して、適切な学習データ量を自動的に抽出し、予測精度が向上するよう自動的に予測モデルを構築し続けることのできるアルゴリズムである。学習の探索に必要なウィンドウサイズの取り方は様々な方法があるが、どのように学習データの部分集合を抽出するかについての検討は十分できていなかった。例えば過去何件分のデータを利用することが適切であるかを探索するアルゴリズムの場合、季節の変わり目等の

予測対象データと直近のデータが異なる特性を持っているタイミングでは、適切な学習データ量を選択できない可能性がある。一方で季節や時間帯でウィンドウサイズを設定する場合、地域や年次によって季節変動のタイミングが異なるため、一意にウィンドウの分割幅を定めることに大きな課題がある。そこでデータの特性やタイミングに関わらず常に対象データと類似した学習データを選択できる近傍データ収集法を提案し、SW-SVR への適用を検討する。

3.2 近傍データ収集法

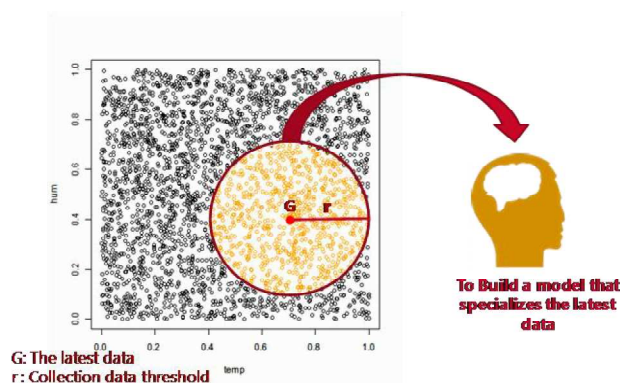


図 3 近傍データ収集法の概要図

近傍データ収集法(SDC: Short-distance data collection method)の概要図を図 3 に示す。近傍データ収集法は存在するセンサデータの中で最新のデータ G を中心として、半径 r から作られる円の範囲に収まるデータを収集するアルゴリズムである。本アルゴリズムによって r を基準として、 G に近傍したデータのみを母集団の中から収集できると考えている。 r の求め方を式(1)に示す。

$$r = \sqrt{\frac{\sum_{i=1}^N (S_i - G)^2}{N}} \quad (1)$$

N は母集団のデータ数、 S_i は母集団の各データ、 G は最新データを示す。 r はユークリッド距離[13]を用いて最新データ G と G を除いた各データとの標準偏差から求められる。式(1)は標準偏差を基準に構築されていることから、母集団の分布が正規分布で表される場合、最新データに近いデータの約 2/3 が利用する学習データとして収集されたため、全体データから約 1/3 のデータを削減できる。そのため SVR の計算量を $O(n^2)$ 、入力した学習データ全てがモデル構築時にサポートベクトル (SV) として利用されると仮定すると、近傍データ収集法の適用によってモデル構築時間を $1-(2/3)^2=5/9$ に削減できる。

次に近傍データ収集法を SW-SVR の部分集合抽出部に利用する方法を述べる。時系列データを予測する場合、大量のデータを学習させるよりも、対象データと類似したデータのみを学習させるほうが高い予測精度を示すことが期待できる。例えば夏のデータを予測するとき、冬のデータはノイズとなりうるため、夏のデータに近いデータのみを学

習させたほうが高精度な予測ができる．そのため、対象データと類似したデータのみを収集できる近傍データ収集法を SW-SVR の部分集合抽出部に利用することで、高精度な予測モデルの構築が可能であると考えられる．従来の SW-SVR ではどのように適切なデータを選択するかの検証が不十分であったため、データの特性や利用するタイミングによって手法の優位性に差異があった．一方で近傍データ収集法は、最新データに対してデータとの距離がどれだけ近いかを基準にデータを収集するため、データの特性やタイミングに関わらず対象データと類似したデータを収集できる．そのため近傍データ収集法を SW-SVR へ適用することで、データの特性やタイミングに依存せずに適切な学習データ量を選択できると考えられる．

4. 近傍データ収集法の基礎評価

4.1 実験目的

SW-SVR への適用を検討する近傍データ収集法のデータ収集性能(DCP: Data collection performance)を評価する．データ収集性能は、対象データとどれだけ近いデータを収集できるかを示す指標である．データ収集性能を式(2)に示す．

$$DCP = \frac{\sum_{i=1}^N d(C_i, G)}{N} \quad (2)$$

データ収集性能は収集されたデータと対象となるデータとのユークリッド距離の平均値で評価する．ここで N は母集団のデータ数、 C_i は収集された各データ、 G は対象データで $d(C_i, G)$ はユークリッド距離を表す．データ収集性能は値が 0 に近いほど対象データと近いデータを収集できたことを意味する．

実験ではアメダスが提供している東京の秋季気象データを用いて、近傍データ収集法と K-means clustering [14] のデータ収集性能を比較した．K-means clustering は代表的な非階層的クラスタリング手法の一つであり、 K の数に標本をクラスタ分割できる．その後対象データがどのクラスタに所属しているかを評価することで、対象データと類似したデータを収集できるアルゴリズムである．本アルゴリズムと近傍データ収集法を評価することで既存手法に比べてどれだけデータ収集性能が高いかを定量的に示すことができる．アメダスの気象データに関しては第 5 章で詳細に述べるが、実験では過去 3 年分 (2011/1/1～2013/9/30) の気象データからデータ量を 1,000 件～250,000 件までの範囲で変化させた時に、それぞれの場合で 2013/10/1 00:00 のデータと近似したデータをどれだけ収集できたかを評価した．

収集するデータ量を揃えるため、K-means clustering と近傍データ収集法にそれぞれパラメータを設定する．データ収集性能はデータ数に依存して値が決定するため、公平に評価するためにはアルゴリズム毎で収集されたデータ量が近似していることが望ましい．実験に利用する気象データ

の分布は正規分布に従っていると考えられるため、近傍データ収集法はデータの約 2/3 を収集すると予想できる．そのため、K-means clustering よりも多くのデータを収集すると考えられるので、 r に重み付けをすることで収集されるデータ量を削減し、アルゴリズムの評価を公平に行えるようにした．また、K-means clustering もクラスタ分類数 K の値によって収集するデータ量が変化するため、 K の値を複数パターン試行し、近傍データ収集法と近似した収集データ量になる条件を探した．なお、K-means clustering の繰り返し回数を意味するパラメータ C は 30 に設定した．

図 4 に近傍データ収集法と K-means clustering が収集したデータ量を示す．図 4 から、近傍データ収集法の r を 0.6 倍した場合と、 K が 5 の場合の K-means clustering との収集したデータ量が近似していることがわかる．データ収集性能はデータの母数に依存するため、提案手法の評価は母数の類似したものと士で行うことが望ましい．よって、データ収集性能の評価には r を 0.6 倍した場合の近傍データ収集法と K が 5 の場合の K-means clustering を利用した．

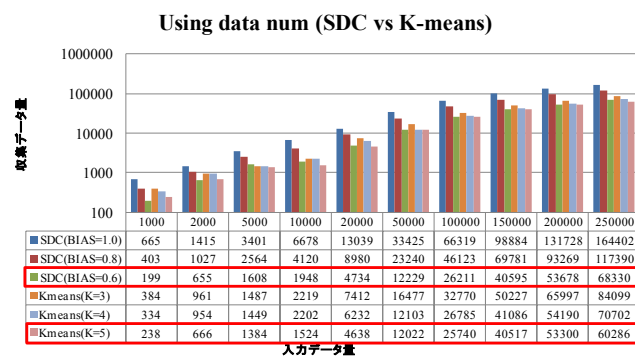


図 4 収集されたデータ量

4.2 実験結果

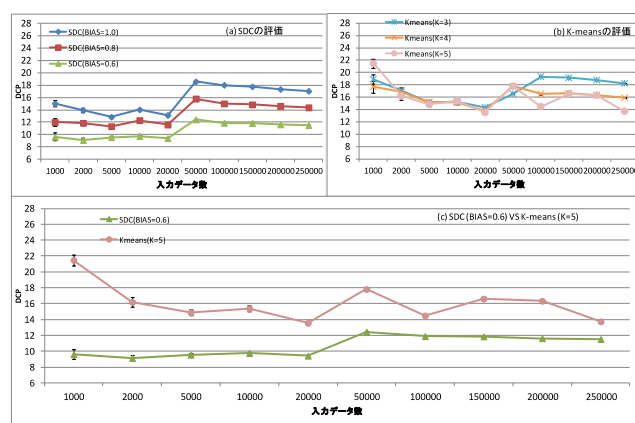


図 5 近傍データ収集法と K-means のデータ収集性能

図 5 に近傍データ収集法と K-means clustering のデータ収集性能を示す．図 5 の(a)は重みに対する近傍データ収集法、(b)は K に対する K-means clustering におけるデータ収集性能を示している．各点に付加されているエラーバーは 95%

信頼区間を表す。(a)から近傍データ収集法の重みが小さいほど、データ収集性能が高いことがわかった。これは設定した重みが小さいほど r の値が小さくなるため、対象データとの距離がより近いものだけを収集できるからである。近傍データ収集法は r の重み付けによって、収集されるデータの条件を調整できるため、想定するアプリケーションによって r の値を設定すればよい。

一方で、(b)から K-means clustering ではデータ量によってデータ収集性能が最も良いパラメータに差があることがわかった。このことから K-means clustering ではデータ量に応じて適切なパラメータをその都度グリッドサーチ等を用いて探索しなければならないことがわかる。

(c)は r を 0.6 倍した場合の近傍データ収集法と K が 5 の場合の K-means clustering の結果を示している。グラフから、データ量にかかわらず近傍データ収集法のほうが K-means clustering よりも高いデータ収集性能をもつことがわかる。以上のことから近傍データ収集法は従来の代表的なクラスタリング手法である K-means clustering と比較して、対象データと類似したデータを高精度に収集できることを示した。

5. 近傍データ収集法適用による SW-SVR の性能評価

5.1 概要

近傍データ収集法を SW-SVR に適用したときの予測精度とモデル構築時間を評価する。ここで近傍データ収集法を適用した SW-SVR を SW-SVR (SDC)と定義する。近傍データ収集法は対象データの予測に適したデータのみを収集できるため、予測精度向上とモデル構築時間削減の両方を実現できると考えられる。実験では従来手法である SVR と SW-SVR、そして SW-SVR (SDC)の 3 種類を、予測精度とモデル構築時間の 2 つについてそれぞれ評価する。

5.2 評価環境

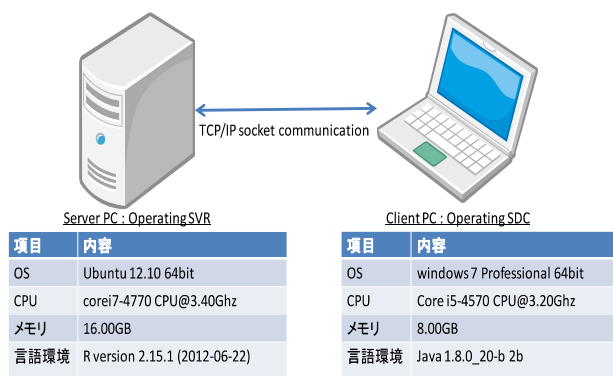


図 6 実験アーキテクチャとマシンスペック

図 6 に実験に用いた各計算機の性能と、実験アーキテクチャを示す。SVR のモデル構築や予測はサーバ PC で行い、データの分割や近傍データ収集法の動作はクライアント PC で動作させた。また、クライアント・サーバ間の通信は

TCP/IP のソケット通信で行う。そのためモデル構築時間は、クライアント PC の処理時間とサーバ PC の処理時間、通信にかかる時間の 3 つから構成される。

表 1 に SW-SVR の各パラメータ設定値を示す。設定項目のうち Window size と再構築閾値以外は SVR に用いたパラメータを表しており、評価学習器 3 種類全てで共通である。これらの値は R 言語の SVM 用パッケージ e1071[15]で設定されているデフォルトパラメータである。Window size は近傍データ収集法を適用しない SW-SVR のみに関係するパラメータで、学習データの分割数を示している。Window size が 8 に設定されている理由は実験環境のクライアント PC のコア数が 8 だからである。再構築閾値は、紹介している気象予測に関する研究で構築した各予測モデルの予測誤差を RMSE で評価すると約 0.5~1.0 であったことから 1.0 に設定した。また、実験では十分な学習データを確保するために近傍データ収集法の r への重み付けは行わなかった。

表 1 SW-SVR のパラメータ

項目	設定値
Method	Epsilon - SVR
カーネル関数	Radial Basis Function (RBF)
Cost Parameter (C)	1.0
Hyper parameter of RBF (γ)	0.125
Tube parameter (ϵ)	0.1
Window size	8
再構築閾値	1.0

5.3 評価データ

評価データに気象庁の提供するアメダスのオープンデータを利用し、札幌、東京、浜松、那覇の 4 地域についてそれぞれの四季 3 年分における 1 時間後の気温の予測を実施した。評価に用いた 4 地域は平均気温や日照時間などそれぞれ異なる気象特性を持っている。また、日本には四季があり季節によっても気象特性が異なる。そのため、気象特性が異なる 4 地域に対して 4 季節分の評価を実施することで、多様な条件下で提案手法を評価できると考えた。

表 2 説明変数と目的変数

項目	内容
説明変数	気温、相対湿度、気圧、平均風速、最大風速、日照時間
目的変数	1 時間後の気温
計測周期	10 分毎

表 2 に利用する説明変数と目的変数、データの計測周期を示す。説明変数には気温や湿度等 6 種類のセンサデータ、目的変数には 1 時間後の気温を用いた。また、データの計測周期は 10 分に 1 回である。これらのアメダスが提供しているオープンデータから欠損箇所を除いた部分を実験に利用した。

表 3 に各季節におけるデータの学習期間と評価期間を示す．実験では 4 地域の四季に対してそれぞれ約 3 年分の学習データから 1,000 件～100,000 件の範囲で学習量を変化させた場合の評価対象学習器の予測誤差とモデル構築時間を評価した．予測誤差には 2 乗平均平方誤差 (RMSE: Root mean square error) を評価指標に利用した．RMSE を式(3)に示す．

表 3 学習期間と評価期間

季節	学習期間	評価期間
春	2011/1/1～2013/3/31	2013/4/1～2013/5/1
夏	2011/1/1～2013/6/30	2013/7/1～2013/8/1
秋	2011/1/1～2013/9/30	2013/10/1～2013/11/1
冬	2011/1/1～2013/12/31	2014/1/1～2014/2/1

$$RMSE = \sqrt{\frac{\sum_{i=1}^N \bar{e}_i^2}{N}} \quad (3)$$

RMSE は個々の予測誤差を 2 乗した後に，全体の期間平均を算出し平方根をとる予測誤差を示す指標である．ここで N は評価期間の標本数， $\bar{e}_i$ は予測値と実測値の差を表している．RMSE は値が 0 に近いほど予測精度が高いことを意味しており，予測誤差の 60%～70%は $\pm RMSE$ の範囲に収まる．

5.4 実験結果

5.4.1 予測誤差

札幌，東京，浜松，那覇の 4 地域について各季節の予測誤差について述べる．評価学習器は SVR，SW-SVR と SW-SVR (SDC) である．それぞれの季節，地域に対して約 3 年分の気象データから学習データ量を 1,000 件～100,000 件まで変化させた場合の予測誤差を評価した．

図 7～図 10 に札幌，東京，浜松，那覇の予測誤差をそれぞれ示す．グラフは季節，地域毎に 1,000 件，2,000 件，5,000 件，10,000 件，20,000 件，50,000 件，100,000 件の 7 つの学習データ量パターンを用いて予測を行ったときの予測誤差を示している．実験の結果，実験項目 112 項目のうち 102 項目で SW-SVR (SDC) の RMSE が最小であった．

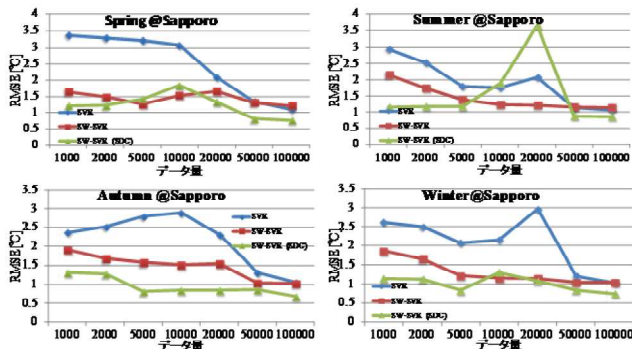


図 7 予測誤差(札幌)

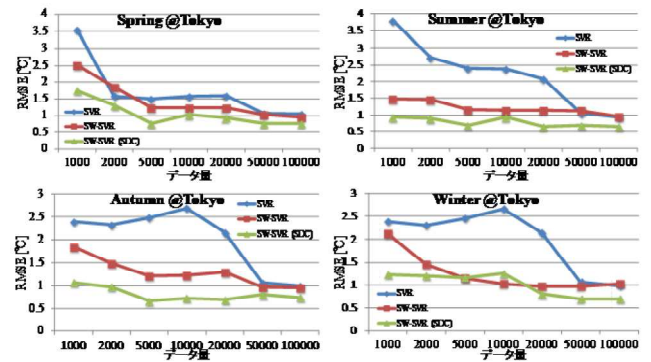


図 8 予測誤差(東京)

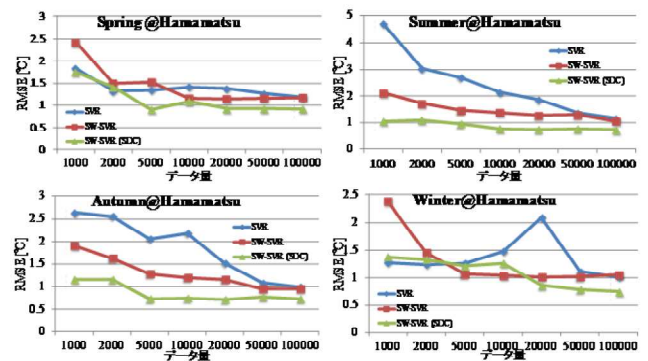


図 9 予測誤差(浜松)

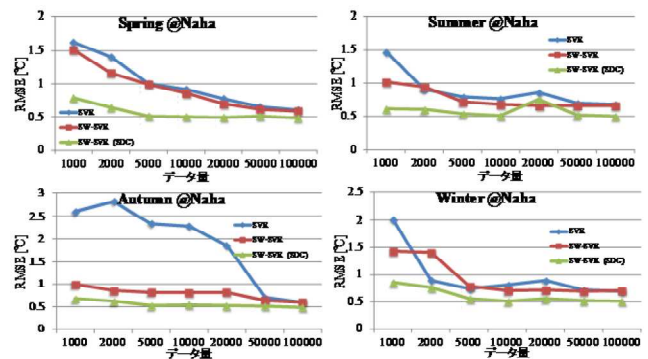


図 10 予測誤差(那覇)

SVR と比較して SW-SVR (SDC) は最大 78%，平均 40% の予測誤差を削減し，一方で SW-SVR と比べて最大 50%，平均 24% の予測誤差削減を達成した．SW-SVR (SDC) と SW-SVR の比較結果から，近傍データ収集法を SW-SVR の部分集合抽出部へ適用することは有効であることがわかった．特に SW-SVR の特性である少量の学習データでも高精度な予測が可能であるという特徴を，近傍データ収集法の適用によってより改善できることがわかった．また SVR との比較結果から，SVR と SW-SVR との予測誤差に差がなくなる 50,000 件や 100,000 件でも SW-SVR (SDC) は SVR より小さい予測誤差で 1 時間後の気温を予測可能であることがわかった．以上のことから SW-SVR に近傍データ収集法を

適用することで、地域や季節、学習データ量にかかわらず、予測誤差を削減できることを示した。

5.4.2 モデル構築時間

SVR, SW-SVR, SW-SVR (SDC)それぞれのモデル構築時間について論じる。SW-SVR は、入力する学習データ X に対して $X/8, 2X/8 \dots 7X/8, X$ とデータを 8 分割し並列で評価を行うため、モデルの構築時間は SVR が学習データ X を全て学習した場合のモデル構築時間と同じになる。そのため、モデル構築時間の評価では SVR と SW-SVR は同等であるとみなし評価した。

実験では図 6 に示す実験アーキテクチャで、モデル構築時間の実測値と理論値を比較し評価する。そのため、実測値のモデル構築時間は近傍データ収集法を行うクライアント PC の処理時間、SVR のモデル構築を行うサーバ PC の処理時間と、クライアント・サーバ間の通信時間の和から求められる。一方で理論値のモデル構築時間は、SVR の計算量を $O(n^2)$ として、SVR が入力された学習データをすべて SV に利用することを前提に求められる。また、実験では入力した学習データ量とモデル構築時間との関係性について考察することが目的であるため、評価期間には代表値として東京秋のデータのみを用いた。

図 11 に東京秋における各学習器のモデル構築時間の実測値と、モデル構築時間の理論値を示す。実験の結果、全ての学習データ量で、提案手法のほうが既存手法よりも高速にモデルの構築が可能であることが確認された。特にデータ数が大きくなるほど従来手法との計算時間の差が大きくなり、250,000 件の場合は SVR と比較して 57%のモデル構築時間削減を確認した。加えて、実測値と理論値のグラフ特性が類似していることから、提案手法の仮説が正しかったことを確認できる。

次に、近傍データ収集法の適用によってモデル構築時間が削減できた原因について考察する。図 12 に近傍データ収集法が削減した学習データの割合を示す。図 12 から、全ての学習データの場合で学習データが約 33%削減されていることが確認できる。これは近傍データ収集法の半径 r を標準偏差の公式に基づいて求めることで、全体の学習データのうち約 2/3 を収集したからである。そのため、SVR の用いる SV の数が削減し、モデル構築時間を短縮できたと考えられる。

図 13 に理論値と実測値のモデル構築時間削減率を示す。グラフから 50,000 件以上の学習データを用いた場合、モデル構築時間の削減率が理論値と実測値でほぼ一致することがわかる。一方で 50,000 件未満の場合、実測値は理論値と比較してモデル構築時間を削減できないことがわかった。近傍データ収集法の適用により学習データは全体の 2/3 に削減されるため、SVR の計算量 $O(n^2)$ からモデル構築時間の減少率は 5/9 となる。一方で理論値の計算には SVR の計算量だけを考慮して算出していることから、実測値に含ま

れる近傍データ収集法の処理時間や、クライアント・サーバ間のソケット通信がボトルネックとなり、50,000 件未満の学習データを用いる場合は、理論値ほどモデル構築時間を削減できなかったと考察できる。また 100,000 件以上の場合に実測値が理論値を上回っている理由は、SVR が入力された学習データを全て SV に利用することを前提に計算しているからである。実際には SVR のチューブパラメータ ϵ 内に存在するデータは学習に用いないため、SVR は前提条件よりも少ない量の SV を利用する。そのため理論値に比べて実測値の計算時間削減率が高くなったと考えられる。

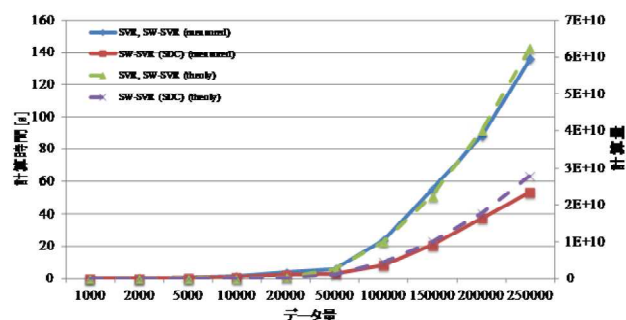


図 11 モデル構築時間の理論値と実測値

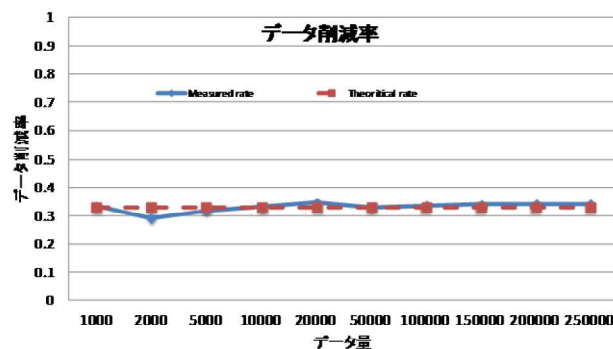


図 12 近傍データ収集法による学習データ削減率

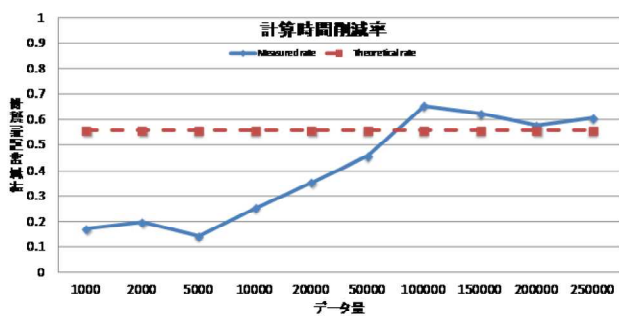


図 13 モデル構築時間削減率

図 14 に学習データ量毎の近傍データ収集法のモデル構築時間と SVR のモデル構築時間の内訳を示す。提案手法の

モデル構築時間は近傍データ収集法の処理時間と SVR のモデル構築時間の和となる。また、SVR のモデル構築時間にはクライアント・サーバ間の通信時間が含まれている。グラフから、近傍データ収集法の処理時間は全体のモデル構築時間に対して非常に小さいことがわかった。特に学習データ量が増えるに従って SVR のモデル構築時間と近傍データ収集法の処理時間の差は大きくなる。例えば、学習データ量 1,000 件の場合、全体のモデル構築時間のうち近傍データ収集法のモデル構築時間の割合は約 14.8%であるのに対して、100,000 件の場合は 3.1%, 250,000 件の場合は 0.7%であった。以上のことから、近傍データ収集法を SW-SVR に適用することは、モデル構築時間には大きく影響しないことがわかる。よって学習データ量が小さいときに理論値に比べて実測のモデル構築時間の削減率が少ない理由は、クライアント・サーバ間の通信のボトルネックが原因であると考察できる。この課題は SVR のモデル構築や予測を同一の環境下で行えるように変更し、通信が必要なシステムアーキテクチャに変更することで解決できると考えている。

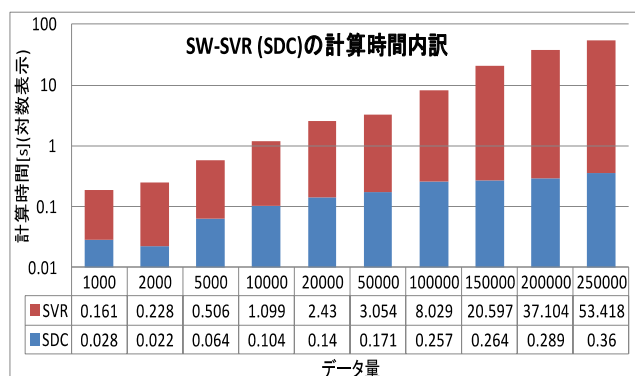


図 14 SW-SVR + SDC のモデル構築時間内訳

6. まとめ

本研究ではSW-SVRに近傍データ収集法の適用を検討した。実験の結果、近傍データ収集法を適用することで予測精度を向上させ、モデル構築時間を削減できることを示した。モバイルデバイスやネットワークの発達によって様々なセンサ機器が普及している背景から、機械学習を用いたモバイルアプリケーションや予測制御システムの開発が注目されている。そのため、これらのシステムに適用可能な高精度かつ高速な機械学習アルゴリズムの開発が必要である。近傍データ収集法を SW-SVR に適用することにより、従来手法よりも高精度かつ高速な機械学習アルゴリズムとなるため、今後必要とされる様々なモバイルアプリケーションや予測制御システムへの適用が期待される。

今後の課題は、実際の予測制御システムに本提案手法を適用した場合の、手法の優位性を確認することである。筆者らは現在、センサネットワークを用いた農業向け液中窒

素濃度予測制御システムを構築し、現場実証実験中である。また、近傍データ収集法にマハラノビス距離[16]等のユークリッド距離以外の距離指標を利用の検討や、SW-SVRの並列処理で複数の学習モデルを構築し適切なモデルを選択するという動作に、近傍データ収集法を導入することを進める。加えて、今回の実験で予測精度の向上が見込まれなかった箇所に関する分析と改良を進めていく予定である。

謝辞 本研究は、総務省 SCOPE 地域 ICT 振興型研究開発（2013-2014、ならびに文科省科研費挑戦的萌芽研究 26660198（2014-2015）によって実施したものである。

参考文献

- 1) Smith, A. Gerrit, H. and Ronald, M.: Artificial neural networks for automated year-round air temperature prediction, Computers and Electronics in Agriculture 68.1, pp.52-61(2009).
- 2) Suzuki, Y. Ibayashi, H. Kaneda, Y. and Mineno, H.: Proposal to Sliding Window-based Support Vector Regression, Procedia Computer Science, 35, pp.1615-1624 (2014).
- 3) Cortes, C. and Vapnik, V.: Support vector networks, Machine learning 20.3 pp.273-297 (1995).
- 4) Freund, Y. and Robert, E. S.: A decision-theoretic generalization of on-line learning and an application to boosting, Computational learning theory (1995).
- 5) Wang, B. X. and Nathalie J.: Boosting support vector machines for imbalanced data sets, Knowledge and Information Systems 25.1, pp.1-20 (2010).
- 6) Gill, M. Asefa, T. Kembrowski, M. and McKee, M.: Soil moisture prediction using support vector machines, Journal of the American Water Resources Association, Vol.42, pp.1033-1046 (2006).
- 7) Hiroyuki, M. and Daisuke, K. Application of support vector regression to air temperature forecasting for short-term load forecasting, Neural Networks, International Joint Conference on IEEE (2007).
- 8) Tsang, W. James, K. and Pak-Ming, C.: Core vector machines: Fast SVM training on very large data sets, Journal of Machine Learning Research, pp.363-392 (2005).
- 9) Chevalier, F. Gerrit, H. Ronald, M. and Paz, J. A.: Support vector regression with reduced training sets for air temperature prediction: a comparison with artificial neural networks, Neural Computing and Applications 20.1, pp.151-159 (2011).
- 10) Laskov, P. Gehl, C. Kruger, S. and Muller, R. K.: Incremental support vector learning: Analysis, implementation and applications, The Journal of Machine Learning Research 7, pp 1909-1936 (2006).
- 11) Matsumoto, M. and Takuji, N.: Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator, ACM Transactions on Modeling and Computer Simulation 8.1, pp.3-30 (1998).
- 12) Sun, Z. and Fox, G.: Study on parallel SVM based on MapReduce, International Conference on Parallel and Distributed Processing Techniques and Applications, pp.16-19 (2012).
- 13) Danielsson, P. E.: Euclidean distance mapping, Computer Graphics and image processing 14.3, pp.227-248 (1980).
- 14) Hartigan, A. and Manchek, W.: Algorithm AS 136: A k-means clustering algorithm, Applied statistics, pp.100-108 (1979).
- 15) Meyer, D.: Support vector machines: The interface to libsvm in package E1071 (2004).
- 16) De, M. Roy. Delphine, J. R. and Désiré, L. M.: The mahalanobis distance, Chemometrics and intelligent laboratory systems 50.1 pp.1-18 (2000).