Vol.2016-ARC-219 No.38
Vol.2016-SLDM-175 No.38
Vol.2016-EMB-40 No.38
2016/3/25

センサプログラミングにおける機械学習支援システム

倉片謙太郎[†1]　早川栄一[†2]

拓殖大学　大学院　工学研究科[†1]　拓殖大学　工学部　情報工学科[†2]

概要：近年，性能の高いマイコンボードや複数のセンサを用いたセンサプログラミングが流行している．複数のセンサから得られた値をシステムに利用する場合，閾値を設けるなどをし，アルゴリズム的に処理をしてきた．しかし，この方法では，閾値算出のためのルールを考える必要があり，ルール同士の干渉を確認することが難しい．そこで本研究では，この問題を解決するべく，センサプログラミングに機械学習を提案し，その支援システムの開発を行う．本研究では，Jubatus と MongoDB を組み合わせたセンサプログラミングを提案し，その利用を Web アプリケーションを用いて支援する．Web アプリケーション上ではセンサ値の修正や，Jubatus のチューニングを行うことができ，コーディングの手間を減らし，Jubatus の設定ファイルを出力することができる．

キーワード：機械学習，　センサプログラミング，組込みシステム

Development of a machine learning support system on sensor programming

Kentaro　KURAKATA and Eiichi HAYAKAWA

Takushoku University　Graduate School[†1], Department of Computer Science[†2]

1. 緒言

近年，性能の高い安価なマイコンボードとして，Raspberry Pi や Intel Edison 等が流行しており，それらと複数のセンサを用いた組込みシステムの開発[1]や教育もある．センサから得られた値をシステムに利用する場合，閾値を設けるなどをし，アルゴリズム的に処理してきた．しかし，この手法では，複数のセンサ同士のルールの干渉を確認することが難しい．本研究ではこの問題を解決するべく，機械学習を提案し，その支援環境を開発する．機械学習には次の 2 種類が存在する．

(1)　バッチ機械学習

(2)　オンライン機械学習

オンライン機械学習の一つとして，オンライン機械学習向け分散フレームワークである Jubatus[2]がある．これは，分散したデータを常に素早く深く分析することができる．しかし，機械学習の要であるチューニング作業が非常に面倒である．また，チューニングに必要なデータ解析はユーザが行う必要があり，機械学習の他に，統計データ処理に関する知識が必要となる．

そこで本研究では，Jubatus を用いたセンサプログラミングにおける機械学習支援を行う．センサプログラミングに関する知識や，データベース，統計に関する知識がないユーザでも機械学習を利用できる支援環境を開発する．

2. 問題分析

2.1. 機械学習

機械学習とは，データから反復的にデータのパターンを学習し，そこに潜むパターンを見つけ出すことである．パターンに従って，予測値を算出することができる．機械学習には次の 2 種類の手法がある．

(1)　教師あり機械学習

(2)　教師なし機械学習

教師あり機械学習とは，データと結果をペアにしたものを教師データとし，それを学習させる．解答となるデータ群を与えておくことで，機械学習の予測に利用する手法である．教師なし機械学習は，回答となるデータ群を持たずに，データの偏りなどから機械学習の予測を行う手法である．

機械学習を用いることで，データを機械学習にかけることで，結果となる値を予測するため，センサ同士のルールを考える必要がなくなるメリットがある．

本節では 2 つの予備実験を通して，システムの問題を抽出する．

Vol.2016-ARC-219 No.38
Vol.2016-SLDM-175 No.38
Vol.2016-EMB-40 No.38
2016/3/25

2.2. Kaggle と scikit-learn を用いたバッチ学習

バッチ機械学習の特性を理解するために，scikit-learn[3]と Kaggle[4]を利用したプログラミングを行った．Kaggle に投稿されているタイタニックの乗客データから，タイタニックの乗客を使った生存者の推定モデルの生成を行った．生成したモデルを評価するために，教師データを学習用と評価用の 2 つに分割した．しかし，どの部分を教師データにするかによって，機械学習の予測結果は異なる問題がある．そこで，K-交差検証を行った．この検証を行うことで，データセットのみで予測値の正答率を判断することができるため，データセットでシステム評価を行うことができる．これを行うことにより，scikit-learn で教師ありバッチ機械学習を行う際は，教師データを一度にすべて読み込み，その後予測したいデータを一度に読み込まなければならない．

2.3. Jubatus を利用したセンシングデバイス作成

一般的にセンサプログラミングに Jubatus を利用する際のプロセス確認のために，Jubatus，CDS セルを用いた在室確認システムの開発を行った． CDS セルを用いたセンシングデバイスを図 1 に示す. マイコンボードには Raspberry Pi 2 を用いて行い，Python でコーディングを行った．教師データとして，在室時の CDS セルの値，不在時のセルの値のデータ収集を行った．今回は，在室時，不在時の値が取得できるので，多値分類器を使用した．在室時，不在時と別のスクリプトを動かすことで，2 種類の結果を持つ実験を行った．これらの値を MongoDB に格納した．チューニングを行う際，センサ値に関するグラフや表を Jupyter notebook を用いて表示した. システム検証のために, CDS セルから適当な値を MongoDB に格納し，その値を 1 秒ごとに取り出し，機械学習の予測を行った．それらから，Jubatus をセンサプログラミングに利用するためには，センサプログラミングの他にグラフ描画や統計量の取得が必要であることがわかった．また，実験ミスなどによるノイズデータが教師データに含まれている場合，正しい機械学習が行えないことがわかった．

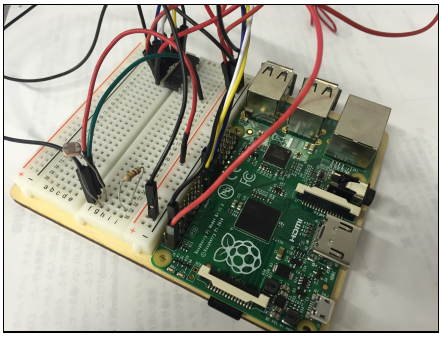

図 1 CDS セルを用いたセンシングデバイス

2.3. システムへの要求

この二つの実験により分かったシステムへの要求は次のとおりである．

(1) 容易な操作環境の提供

(2) 統計情報の計算，表示

(3) データセットと実験結果との関連づけ

(4) データのグラフ化とノイズデータの抽出と再実験や比較の容易性

3. 設計方針

本システムの設計方針は次のとおりである．

(1) Web アプリケーションとして提供する

ユーザにとって操作がしやすく，環境を選ばない Web でアプリケーションを提供する. 対象となる IoT は，センサデータをネットワークを介してやりとりすることから，このデータをサーバに保存することにより，データへのアクセスや操作を容易にする

(2) センサデータの DB への保存と実験結果へ関連付け

機械学習では，データセットやアルゴリズムによって結果が異なることから，同一のデータを用いて再実験する必要がある．そこで，センサデータをオーバヘッドの少ない DB に格納し，再実験を容易にする．

(3) センサデータに関連する情報の自動的な付加

センサデータは，通常の Web データとは異なり，ハードウェアに関する情報や時間情報が重要になる. 例えば, 使用する A/D コンバータによってダイナミックレンジが異なるため，使用す

Vol.2016-ARC-219 No.38
Vol.2016-SLDM-175 No.38
Vol.2016-EMB-40 No.38
2016/3/25

る A/D コンバータのダイナミックレンジの情報が必要である．また，使用するデバイスによって処理速度が異なる問題がある．そこで，次のようなセンサデータの付加すべき情報を，データと併せて管理して，機械学習時の結果の検証に用いる．

A) A/D コンバータ情報
B) 使用デバイス情報
C) タイムスタンプ
D) センサ値の統計量

(4) センサデータのグラフ化とノイズデータへの容易な操作

教師データを収集する際に，実験ミスによりノイズデータを取得してしまうことがある．ノイズデータを含んだ教師データでは正しい機械学習を行うことができないため，ノイズデータを発見，削除する必要がある．

これらの情報を統計データとともにグラフ化し，Web アプリケーション上で確認，操作することができる環境を開発する．さらに，ノイズデータに対してはグラフィカルにデータの削除処理を可能にする．

4. 設計

4.1. システムフロー

システムフローを図 2 に示す．本研究では，Web アプリケーション上でユーザがセンサデータを表やグラフを用いて Jubatus のチューニングを行い，Jubatus の設定ファイルを出力することができる．また，ユーザが表計算ソフトを使用するなどして，データを活用できる，データの視認性を向上させるために，MongoDB 内のデータを CSV に変換し，保存する．次に本研究のシステムフローを示す．例として教師有り機械学習を行うものとする．

(1) ユーザはセンシングデバイスで教師データを取得する
(2) 教師データを MongoDB に格納する
(3) MongoDB のデータを CSV に変換する
(4) ユーザは，Web アプリケーションを用いて教師データの操作を行う
(5) アルゴリズム，重みの設定を行う
(6) Jubatus の設定ファイルを出力する
(7) 本研究のライブラリを用いて，Jubatus による機械学習を行う

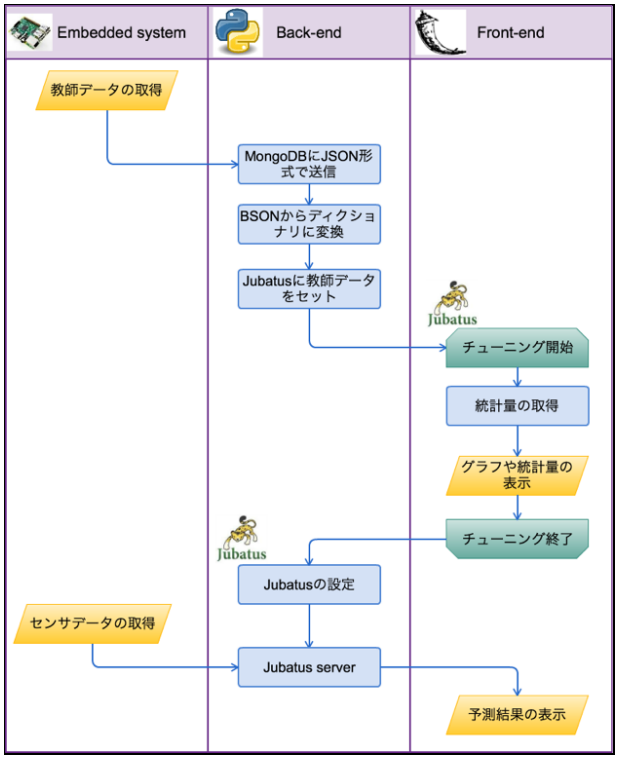

図 2 システムフロー

4.2. センサプログラミング

本研究では，教師有り機械学習の利用を前提としている．これは，機械学習の約 70%は教師有り機械学習である[5]．センサプログラミングにおいては，多値分類器異常値検知を用いて，センサ値の判別を行う．教師データを取得することができる状況では，教師なし機械学習で用いられる異常値検知よりも，教師あり異常値検知（多値分類器）のほうが，精度が高いとされている．また，教師有り機械学習は，教師データを作成することから，教師なし機械学習より手間がかかるからである．

本研究では，機械学習フレームワークに Jubatus，データベースに MongoDB を使用する．Jubatus はスケーラブラルであり，アルゴリズムやパラメータを設定ファイルで切り替えることが可能であり，実験がやりやすいというメリットがある．また，MongoDB はデータベースへのデータの追加のオーバヘッドが少ないことから選択した．

ユーザがセンサデータを格納したら，Web アプリケーションにより，Jubatus の設定ファイルを出力する．設定ファイル内のパラメータ選択を Web からできるようにすることで，設定ファイルの JSON を直接編集する必要がなくなる．また，教師デー

Vol.2016-ARC-219 No.38
Vol.2016-SLDM-175 No.38
Vol.2016-EMB-40 No.38
2016/3/25

タや分類データのデータセットを同時に指定することで，データセットと学習結果とを関連づけて扱う事が可能になる．センサデータのセットは基本的に増加していくことから，処理時間が伸びる傾向にあるが，処理に使うデータを属性付けして扱うことにより，有効な結果が得られた場合の利用環境を記録することが可能になった．

5. 実装

5.1. フロントエンド

作成した Web アプリケーショントップページを図 3 に，グラフ表示画面を図 4 に示す．

本 Web アプリケーションは Python の軽量 Web アプリケーションフレームワークである Flask を用いて開発を行った．Flask は，データベース抽象化レイヤや，フォームの検証等の機能をもっていない．そのため，本研究で用いる MongoDB の利用には，Python ライブラリの PyMongo を使用している．

統計量の表示には，jQuery プラグインの jquery.csv2table.js を用いる．統計量表示画面では，統計量が表となり，表示される．表のインデックスをクリックすることで，そのインデックスをもとに昇順，降順の並替えを行うことができる．

グラフの表示には，JavaScript ライブラリの HIGHCHARTS を使用している．HIGHCHARTS を使用することで，マウスを任意のグラフ値にスクロールすると，センサ値と ID 番号が表示される．

本 Web アプリケーション開発にあたり，サンプルデータとして，センサ値を 5 種類それぞれ 10000 件ずつ使用した．すべてのセンサ値の情報を HIGHCHARTS で Web ブラウザ上に表示すると，データ量が多いためにグラフ描画ができずにブラウザがクラッシュしてしまう問題が発生した．一般的にセンサ値は順次到着し，分類や教師データとして用いられる．そこで，そこで，全センサ値を一つのグラフに描画するのではなく，1 種類のセンサ値をグラフ上に描画する対策を行った．グラフを読み込む前に，MongoDB からセンサの名前情報だけを取り出し，描画したいグラフをユーザに選択させる方式を採用した．

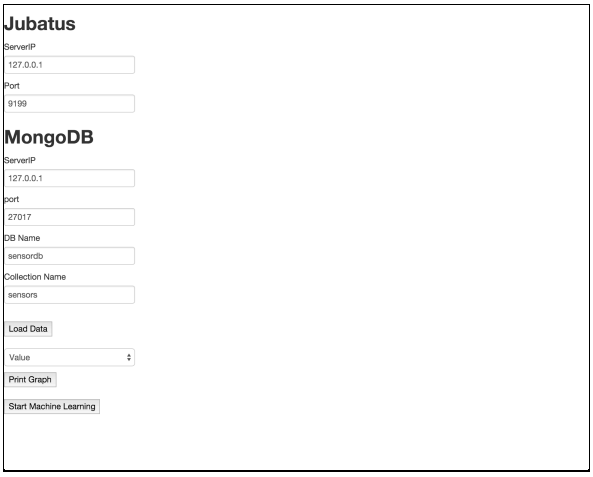

図 3 Web アプリケーショントップページ

5.2. サーバサイド

Pandas を用いて値をデータフレームに変換する．データフレームに変換を行うことで，データの削除，並替え，追加を行うことができる．値をデータフレームに変換後，CSV に変換し，オリジナル教師データとして保存を行う．CSV に変換する理由として次の 2 点が挙げられる．

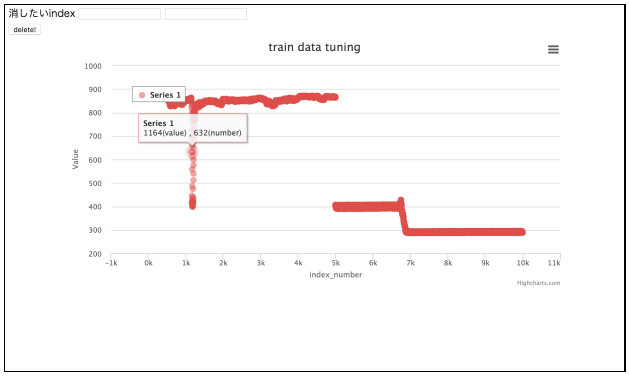

図 4 グラフ表示画面

(1) DB に関する知識が浅いユーザでも，表計算ソフトを用いて値の確認を行うことができる

(2) 視認性が良い

DB に保存されている状態では，ユーザが実験したセンサ値の情報を有効に活用できない．機械学習を利用しない場合でも，CSV に変換し，保存されていることで，表計算ソフトで利用する，データの送信を行う等，使用するユーザが自分好みの使い方をすることができるメリットがある．

CSV に変換された値を複製し，編集用ファイルを作成する．理由として，ユーザが Web アプリケーションを用

Vol.2016-ARC-219 No.38
Vol.2016-SLDM-175 No.38
Vol.2016-EMB-40 No.38
2016/3/25

いて行うデータ操作が，正しいデータ操作であるか確証がないからである．グラフを用いたデータ操作は，ユーザ自身が不要なデータかどうか判断するので，ユーザがデータ操作をミスした場合，いつでもオリジナルなデータに戻れる必要がある．これを編集ファイルとする．

編集ファイルを読み込み，グラフを生成する．ユーザはグラフに表示されている ID をフォームに入力し，削除するデータを決定する．Pandas を用いて，CSV をデータフレームに変換を行う． ID によるデータ編集を行い，同名ファイル（編集ファイル）で CSV に変換して保存を行う．

データの操作が終了したら，編集した CSV ファイルを元のデータベースに上書きし，教師データを MongoDB にセットする．

6. 評価と考察

システムの有効性を得るために， CDS セルを用いて在室状態の結果を Jubatus の多値分類器（Classifier）である jubaclassifier を使用し，Twitter に投稿するセンシングデバイスの作成を行った．本 Web アプリケーションを利用した結果，次の 2 点のメリットが明らかになった．

(1) MongoDB に教師データを格納することで，再チューニング時に教師データを再取得する必要がない
(2) 統計量の表示を行なうことで，チューニングの指標になる

従来ならば，統計量，グラフをユーザ自身がセンサ値から作成していたが，本 Web アプリケーションの利用により，センサ値から自動生成されるので統計量やグラフに関するコーディング量が減った．

また，評価の結果から，アルゴリズムの選定に時間がかかることがわかった．Jubatus に実装されている全てのアルゴリズムの結果を確認し，適切なアルゴリズムを選定する作業には時間を要するという問題が見つかった．

本研究を通し，センシングデバイスで得られた値を機械学習に通すことで得られる予測値は有効な値であることが推測される．センサ値から機械学習を行うことで，複雑なアルゴリズムを組む必要はない．本研究を利用せずに Jubatus を利用する場合，チューニングに第三のツール（統計量やグラフを表示するツール）が必須となる．これらのツールを利用せずに，一つのツールだけで解決できる本研究は，センサプログラミング初心者にも利用しやすい環境であると言える．また，ノイズデータの削除を行うことで，実験ミスを排除したデータで教師学習を行うことができるので，より精度の高い機械学習が可能となる．

7. 関連研究

Raspberry Pi 等のマイコンボードを利用すると，ネットワークに接続することが容易であり，ネットワークに接続した組込みシステムを利用し問題を解決する Internet of things（以下 IoT）も流行し，IoT の分野ではセンサと機械学習を組み合わせた研究[6]が行われている．そして，それらの開発を支援するためのシステムとして Preferred Infrastructure の Sedue Predictor[7]がある．このシステムの特徴は，学習データの管理，学習した結果を可視化する機能がある．しかし，チューニングは手入力で行う必要がある．また，オープンソースでないことや，対象データがテキストデータという点は，IoT の分野では欠点である．

Jubatus 以外の機械学習のオープンソースフレームワークとして，次の 4 つを挙げる．

(1) Chainer[8]
(2) Mahout[9]
(3) scikit-leran[10]

Chainer は，深層学習ができ，音声認識，画像分類で成果を出している．しかし，これらのフレームワークを用いてセンサデータを用いることは難しい．理由として，バッチ機械学習であるからである．バッチ機械学習では，センサ値を蓄積したものを機械学習に利用するため，リアルタイム処理が必要なセンサプログラミングでは利用は難しい．センサプログラミングでは，リアルタイム処理が可能なオンライン機械学習が必要である．

Mahout は，バッチ機械学習が利用できる Java のライブラリである[11].

scikit-learn は，バッチ機械学習が利用できる Python のライブラリである．

これらより，Python で使用でき，オンライン機械学習をサポートしているフレームワークとして，Jubatus だけである．

8. 結言

本研究は，機械学習やアルゴリズム，データベースに関する知識や，センサプログラミング以外に必要な，グラフ描画，統計量計算を行うことができる．また，センサ値のノイズデータを削除することができるため，ノイズデータのない機械学習が

Vol.2016-ARC-219 No.38
Vol.2016-SLDM-175 No.38
Vol.2016-EMB-40 No.38
2016/3/25

可能である．Jubatus を使用する際，本研究で作成したライブラリを使用することで，Jubatus に関する知識が浅くても利用が可能になった．

また，本研究の課題として，5 万件以上のデータを HIGHCHARTS でグラフ表示すると，ブラウザがクラッシュしてしまう問題が発生した．機械学習では，多くの教師データがあるほど，より精度の高い機械学習を行うことができる．センサ値のデータは多くなりがちなので，これに対応しなければならない．

さらに,アルゴリズムの選定の自動化が挙げられる。今回は、すべてのアルゴリズムを実行し，保存するという解決策をとったが，そこからいかに正しいアルゴリズムを選択するかはユーザが行うが，正しいアルゴリズムを選定する作業には時間がかかってしまう．その解決策として，ユーザが正しいと判断したアルゴリズムと実行結果を保存しておき，そのデータに機械学習を行うことで，ある程度アルゴリズムを予測する手法を検討する．

参考文献

[1]小型・低価格デバイスを用いたデジタルサイネージ表示システムのプロトタイピング．三島和宏，桜田武嗣，萩原洋一．研究報告デジタルコンテンツクリエーション(DCC).
[2] Jubatus
http://jubat.us/ja/
[3] scikit-learn
http://scikit-learn.org/stable/
[4] Kaggle
https://www.kaggle.com/
[5] sas
http://www.sas.com/ja_jp/insights/analytics/machine-learning.html
[6] ビッグデータのリアルタイム分析基盤「Jubatus」を活用し、センサデータ機械学習検証システムを構築–ビニールハウスのデータ異常検知の自動化を実現–. 住友精密工業株式会社，株式会社 Preferred Infrastructure，株式会社ブリスコラ，2014/02/19
[7] Sedue Predictor．Preferred Infrastructure. https://preferred.jp/product/predictor/，2015/07/15
[8] Chainer http://chainer.org/
[9] Mahout http://mahout.apache.org/
[10] scikit-learn http://scikit-learn.org/stable/
[11] エンタープライズジン Mahout 使って分析しちゃいました http://enterprisezine.jp/dbonline/detail/4727