

生体信号/センサ信号/為替などの一連データを読む!

時系列データのAI未来予測入門

足立 悠

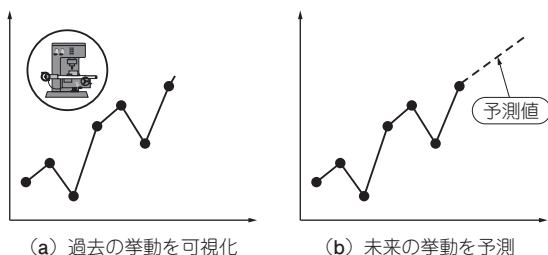


図1 時系列データの主な用途は可視化と今後の挙動や推移の予測

Browse Through: 83 Data Sets

Default Task	Name	Data Types	Default Task
Classification (58) Regression (29) Clustering (29) Other (10)	UCI Absenteeism at work	Multivariate, Time-Series	Classification, Clustering
Attribute Type Categorical (0) Numerical (71) Mixed (7)	UCI Activities of Daily Living (ADLs) Recognition Using Binary Sensors	Multivariate, Sequential, Time-Series	Classification, Clustering
Data Type - Undo Multivariate (333) Univariate (20) Sequential (66) Time-Series (23)	UCI Activity recognition from smartphone-mounted Accelerometer	Sequential, Time-Series	Classification, Clustering
IGBT (48) Domain Theory (23) Other (21)	UCI Activity Recognition system based on Multisensor data fusion (AReM)	Multivariate, Sequential, Time-Series	Classification

図2 UCI Machine Learning Repositoryの時系列データ

時系列データが読めるということ

時系列データは時間的順序を追って一定間隔ごとに観察され、しかも相互に統計的依存関係が認められるようなデータの系列のことです⁽¹⁾。例えば温度や加速度などを計測するセンサ・データ、電力使用量データ、商品売上データ、購買履歴データ、人口推移データなどがあります。

時系列データには、可視化することによって、過去から現在の状態を把握することの他に、さまざまな使い道があります。最近では機械学習やディープ・ラーニングを適用することによって、過去の状態から規則性を見出し、未来の状態を予測することにも使われています。

より具体的には、センサ・データから機器の故障や異常の兆候を予測(図1)して計画外の停止を回避したり、購買履歴データから顧客の嗜好を予測して商品を推薦したりすることに役立てられています。

予測に必要な時系列データ

● 集め方

時系列データを使って未来の状態を予測するためには、まずデータを収集しなければなりません。このとき、何を予測したいかによって収集すべきデータの種類の異なります。そして、データの種類によっては自力で収集できるものもあれば、ウェブなどを介して有

料または無料で収集できるものもあります。

すぐに時系列データを入手したいのであれば、まずはUCI Machine Learning Repositoryへアクセスするとよいでしょう⁽²⁾。このサイトでは、機械学習の練習用データセットが豊富に提供されています。図2ではTime-Series(時系列)を選択しました。

この他、国や自治体が公開しているオープンデータ・カタログ・サイト、電力各社が公開しているでんき予報サイトからもデータを入手できます。積極的に活用していきましょう。

● 予測に必要なデータのタイプ

未来の状態を予測したいとき、教師あり学習のアルゴリズムを使って予測モデルを作成するため、次の2つを準備しなければなりません。

- 結果データ(教師データまたは目的変数)
- 結果を説明するデータ(説明変数)

機器の故障や異常の兆候を予測したいのであれば、結果データとして機器の故障を記録したイベント・ログが、結果を説明するデータとして機器から収集し蓄積したセンサ・ログが必要です(図3)。

● 課題：異常は減多に起こらないためデータを集めにくい

しかし現実には、機器の故障や異常は減多に起こらないため(モノによっては1年に数回程度)、結果データの件数は非常に少なく、収集は難しいものです。予測モデルを作りたくとも、結果データがなければ機器