The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

事例紹介:24 時間周期データに対する教師無し学習の適用

A pilot study: Application of machine learning to 24 hour period data – three case studies

後藤　勲[*1]
Isao Goto

[*1] 住友電気工業株式会社
Sumitomo Electric Industries, LTD

In this article, I'd like to introduce three case studies for which I tackled for three years as a pilot study. This pilot study was applied to three fields. 1. Energy consumption in houses. 2. Anomaly detection of photovoltaic power system at mega-solar system. 3. Sunlight combination type plant factory. In each case 24 hour period data was treated as a frame (but three data formats and types are different), and was analyzed in unsupervised learning. These analyses found out beneficial knowledge for each field. **Key words: k-means, Probabilistic Latent Semantic Analysis(PLSA), Bayesian network, Home energy management system, Photovoltaics, Sunlight combination type plant factory**

1. はじめに

本稿の目的は, 24 時間周期データに対して機械学習を適用した事例と, その結果の一部を報告することである. そのため, 背景, 方法論や結果の詳細については割愛している.

24時間周期データは世の中に多い. 太陽の動きと関連した事象, 例えば日射量, 気温や人の活動などもそれにあたる. これらのデータは, 一日の総量や平均値(±標準偏差), または連続的な時系列データとして取り扱われる場合が多いと思われるが, 今回は 24 時間周期を一つのフレームとして捉え, データを分析した. それぞれ紹介する事例の分析の目的は異なるが, このフレームに落とし込むことで, 同様の考え方で分析が可能となった.

本稿では, 実際に社内で取り扱った 3 つの事例を報告する. 一つは, Home Energy Management System(HEMS)から得られた世帯毎の消費電力データと属性データ, 太陽光発電(PV)の発電データ, そして, 太陽光利用型植物工場内の環境データと植物の吸水量データの 3 種類である.

今回, 主に使用した手法は, 教師無し学習の k-means と確率的潜在意味解析(PLSA)で, これらをデータと目的に応じて使い分けた. その他, 必要に応じてベイジアンネットワークを含む統計解析等を用いて結果を評価した.

2. 方法

2. 1 k-means

k-means はクラスタリング手法の一つで, 最も広く使用される手法のである[Lloyd 1982][Wu 2008]. 今回はその概要のみを記載する. k-means とは分析対象のデータ X を任意の k 個のクラスタに分割するアルゴリズムで, 目的関数である式(1)を最小化するクラスタ中心を見つける.

$$\varphi = \sum_{x_j \in X} \min_{i \in k} \|x_j - c_i\|^2 \cdots (1)$$

ここで, $x_j (j = 1, \dots, N)$は各データ, N はレコード数, $c_i (i = 1, \dots, k)$はクラスの重心を示す. アルゴリズムは以下の通りである.

1. 任意の k 個のクラスタ中心をランダムに選択
2. 全てのデータをx_jからもっとも近いクラスタに割当
3. 式(2)を用いてクラスタ毎に重心を算出

$$c_i = \frac{1}{|C_i| \sum_{x_j \in C_i} x_j} \cdots (2)$$

ここで, C_iはクラスタ i のデータ集合で$|C_i|$はクラスタC_iに含まれるデータ数である.

4. クラスタの変化がなくなるまで, 2, 3 を繰返し

計算は R とパッケージ stats 3.4.2 を用いた.

2. 2 確率的潜在意味解析(PLSA)

PLSA は自然言語処理における文書と単語の行列から文書の潜在的な意味を推定するために提案された分析手法の一つである[Hofmann 1999]. 本稿では, N 個の文書d_i(i = 1, ... N)と M 個の単語$w_j (j = 1, \dots M)$を目的に応じて読み替えて, PLSA を適用した[石垣 2011]. また潜在クラスタ数 k と仮定し, その変数をz_kと表し, その関係を以下の同時確率としてモデル化する.

$$p(x_i, y_j, z_k) = p(z_k)p(x_i|z_k)p(y_j|z_k) \cdots (3)$$

文章 i の単語 j の出現数をN_{ij}とすると, その尤度対数は,

$$L = \sum_i^n \sum_j^m \log \sum_k^K p(z_k)p(x_i|z_k)p(y_j|z_k) \cdots (4)$$

となる. このモデルを EM アルゴリズムで対数尤度を最大化する条件付き確率を推定する. 各条件付き確率に対して, 初期値を乱数で与えると, 式(3)の変形から潜在変数の条件付き確率は以下の式で計算できる.

$$p(z_k|x_i, y_i) = \frac{p(z_k)p(x_i|z_k)p(y_j|z_k)}{\sum_k^K p(z_k)p(x_i|z_k)p(y_j|z_k)} \cdots (5)$$

上記の計算後, それぞれd_iとw_jの最も所属確率が高いカテゴリに所属するとした.

PLSA の計算には, 国立研究開発法人産業技術総合研究所人工知能研究センターの開発した APOSTOOL3.0 を用いた.

2. 3 ベイジアンネットワーク(BN)

BN は対象とする確率変数のノードと変数同士の依存関係を確率的なネットワークとしてモデル化したものである. その確率ネットワークはグラフ構造として表現することが可能で, 視覚的に表

※本研究内容は住友電気工業株式会社の公式見解を示すものではありません.

連絡先:後藤　勲, 住友電気工業株式会社, 大阪市此花区島屋 1-1-3, 06-6466-8231, gotou-isao@sei.co.jp

現・理解し易く，グラフィカルモデルによる確率推論の手法を用いることができる．

BN の構築には，BayoNet6.2(現 BayoLink, NTT 数理システム)を用いた．

3. 結果

3.1 事例 1: HEMS データの活用-大規模 HEMS 情報基盤整備事業（経済産業省：平成 26 年度から平成 27 年度までの事業）

弊社は上記整備事業に参画し，事業内の企業コンソーシアム(i エネコンソーシアム)の枠組みで，平成 27 年 12 月から二か月間，三重県桑名市と四日市市周辺の 457 世帯にご協力いただき，一般家庭向けの節電実証を実施した．当整備事業内で匿名化された世帯毎の電力・属性データから，消費電力パターンと家族構成や節電機器の関係を分析した．

今回は PLSA を用いて世帯と消費電力量を同時に自動カテゴリ化する．同様の試みは存在する[Verdu 2006]が，同時カテゴリ化できないことや，結果をヒトが理解しづらいことなど課題があった．また一般的なクラスタリングでは，次元の呪いのため高次元データを取り扱えないことなども課題であった．そこで今回は PLSA を用いることで上記の課題を解決可能か検証し，その後，BN によりカテゴリと属性データ間の構造モデルを構築し，その関係性を検証した．

電力データは 30 分 1 コマのデータを平日 10 日分，属性データは世帯毎の家族/建屋デモグラ，設置機器等である．

今回は 24 時間周期をもつ消費電力量の時系列データを PLSA により世帯単位で時間帯毎の消費電力量を同時カテゴリ化した．潜在カテゴリ数(k)は赤池情報量基準(AIC)から決定することはできるが，本稿では k = 3 に固定し，10 回初期値を変えて計算した．本来ならばカテゴリ化後，その妥当性を検証すべきだが，正解データが存在しないため，カテゴリ化の結果と実際のカテゴリ毎の消費電力パターンの平均値を比較した．その後，潜在カテゴリと各世帯の属性情報との関係をベイジアンネットワークでモデル化し,カテゴリをエビデンスとしたときの変数（属性データ）の条件付き確率を比較した．

PLSA 適用結果を図 1 に示す．図 1 は各カテゴリの時間帯毎の電気を使うカテゴリ間と自身の時間帯毎の相対的な確率を示している．各カテゴリが異なる時間帯に電気を使用する確率が高いことが示されており，例えば，C03 は朝と夕方から夜にかけて電気を使う確率が高くなっている．図 2 は各カテゴリの時間帯毎の平均的な消費電力量を示している．それぞれが図 1 で示されたパターンと同様の特徴を示しており，カテゴリ毎の特徴を明確に分類できている．図 3 は BayoNet を用いて構造学習した結果で，この構造モデルからカテゴリをエビデンスとしたときの給湯

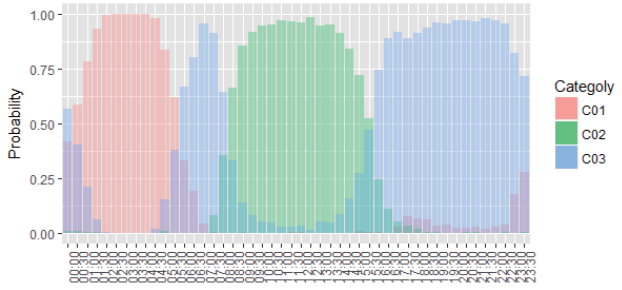

図 1 カテゴリ毎の時間帯別電気使用確率

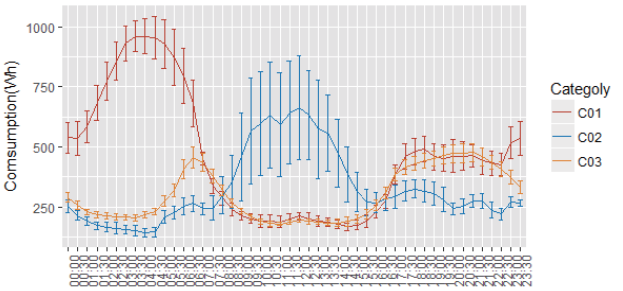

図 2 カテゴリ別の時間帯別平均消費電力量

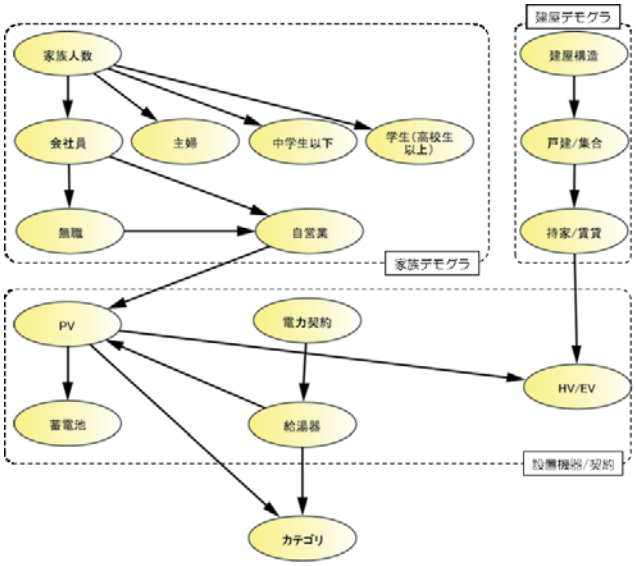

図 3 構築されたベイジアンネットワークモデル

器と PV の所有確率(事後確率)を表 1 に示す．C01 は，深夜から早朝に電力を使用する確率が高いが，その原因がエコキュートによる夜間電力を用いた電力の消費であることが予想できる．また，C2 は他のカテゴリ PV 設置率が高いことは日中に発電した電力を積極的に使用していることと一致すると考えられる．

表 1 カテゴリ毎の省エネ機器設置確率

	カテゴリ		事後確率	
エビデンス	所属世帯数（世帯）	電力消費パターン	エコキュート設置率(%)	PV 所有率 (%)
C01 = 1	83	深夜-早朝	76	52
C02 = 1	17	日中	31	60
C03 = 1	346	朝,夕方-夜	30	17
全体	446	-	38	25

3.2 事例 2: 太陽光発電の異常検知

今後のエネルギーミックスの問題からも PV の有効活用は今後重要であるが，その課題の一つとして異常検知が挙げられる．異常の検知のため，まず「異常」を定義することが必要である．そして，さらに分析のための長期間のデータや別途照度計等のデータが必要になってくる．上記の課題を解決するため，ストリング単位のデータを 1 フレーム 24 時間周期と捉え，ストリング毎の発電パターンをクラスタリングした．その後，現地調査によりクラスタ毎の異常原因を特定した．

データは，当社のストリング監視システムから取得したストリング単位の発電データで，国内のあるメガソーラーサイトの 456 ストリングから得た．データのサンプリングは，1 分単位で計測可能

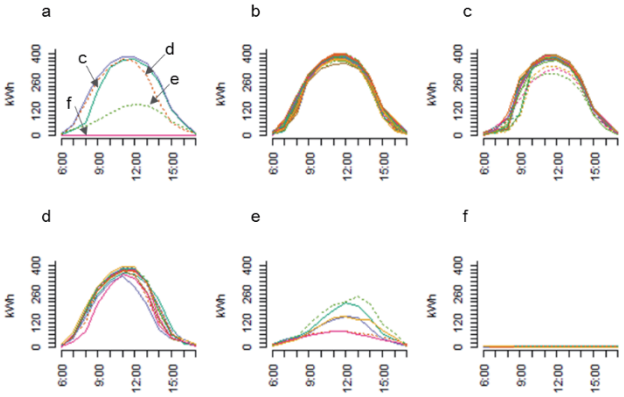

図 4 クラスタリング結果と各クラスタのストリング発電データ

だが, 雲等の影響を除去するため, データを 60 分の平均値とした. さらに, データを晴天時の一日分, そして発電が期待される時間帯を 6-18 時と限定すると 12 次元データとなり, 適用範囲が ~8 次元程度とされている k-means の採用は難しいが, 全ストリングの発電パターンは太陽の動きに依存し, 全てが基本的にはベル型(同じ形)となるため k-means でのクラスタリングが十分可能と考えた. k の決定については, 変化させ AIC 等で決定する方法はあるが, 今回は,実務者の意見を取り入れ, k=5 とした. これによりヒトの感覚に近い異常の定義となった(k は 2-10 まで変化させた). これらの手続きで各クラスタの特徴から異常を定義した.

図4a は結果の各クラスタの重心である. 図 4 b-f は各クラスタの所属ストリング毎の発電データで, b が最も総発電量が大きく異常はないと判断した. c と d はそれぞれ午前, 午後の発電量が低くなっているおり, e は全体的に発電量が低く, f はデータが 0 に近かった. これらの結果を現場調査結果と突合すると, c と d はそれぞれ位置的な問題で物体の影となりこれらの異常が発生していることが分かった. また f はデータ取得エラーとなっており, e は日中, 影の影響がみられる型とシステムによる異常が見つかった. 本手法により異常有りと判断されたストリングは実務者の目視でも異常と判断された. また, 同様の手法で別の二つのメガソーラーサイトでも同様の結果が得られている(Data not shown).

3.3 事例 3: 温室におけるトマトの給水量予測モデル

作物(今回はトマト)の吸水量は, その生長状態や健康状態の定量化に重要な情報であるが, 環境によってもその量が変動するため, その予測は難しい. これまでの研究から一日の平均室温の積算値(積算温度)が一日の吸水量と相関することはわかっているが, 社内の複数の検証実験では精度は, 平均絶対パーセント誤差(MAPE) ≒ 0.2-0.3 で精度的には十分ではなかった (Data not shown).

本稿で用いるデータは, 千葉大学柏の葉キャンパスの太陽光利用型植物工場(以下, 温室とする)でトマト栽培から得られた環境データ(湿度: Hum., 光粒子: Ill., 飽差: Sat. Def., 室温: Temp.)と, 当社の製品ニューサンドポニックスより得られた給水量データである. 今回は給水量 ≒ 吸水量と考える. データの粒度は環境データが 1 時間単位, 吸水量データが 24 時間単位であり, 単位を吸水量に合わせる必要がある. しかし, 環境データを 1 日の統計値としてしまうと, 環境の変化パターンを無視することとなり, 重要な現象を見逃す可能性もある. 今回はこの課題を解決するために, 以下の手法(図 5)を検討し, その有効性を検証した.

まず, 24 時間(フレーム)毎に分割した各環境データを k-means によりクラスタリングし, パターンを抽出した. 今回は k を

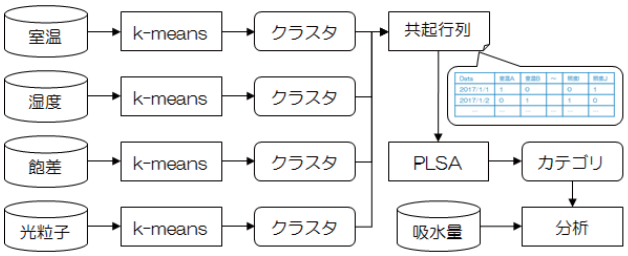

図 5 データ分析の流れ

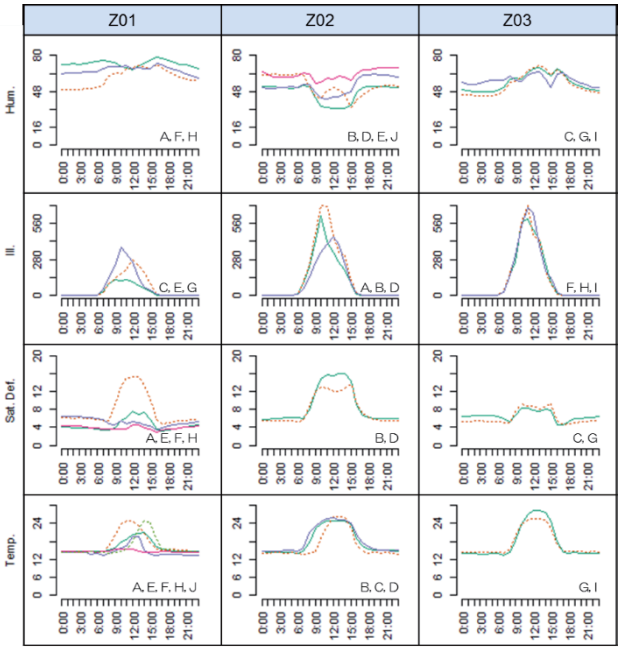

図 6 カテゴリ毎の所属環境パターン

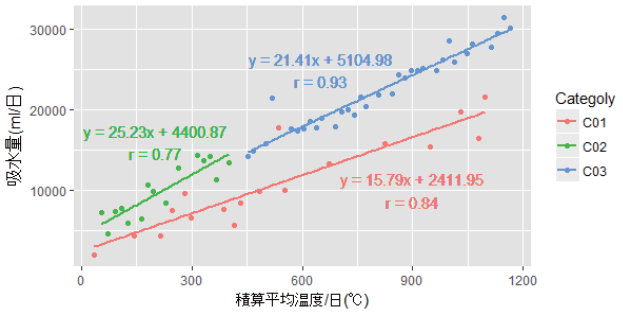

図 7 カテゴリ別の積算平均温度と吸水量

Gap 統計量とクラスタ内分散により決定した. またクラスタは各環境データに A から順にアルファベットでクラスタ名を付けた. その後, クラスタリング結果から共起行列を作成し, PLSA により潜在カテゴリを抽出した. PLSA のカテゴリ数に関しては, 温室の作業従事者と共に検討し 3 とした. この手続きにより, 文章と単語と同様に, 日付と環境パターンで同時カテゴリ化が可能になった.

図 6 は各環境データのパターン毎(図中の A-J)の所属カテゴリを示している. 各カテゴリで異なるパターンを持つことがわかる. さらに, 一日の平均温度の積算値と吸水量の散布図をカテゴリ毎に分類したものを図 7 に示す. この図から, カテゴリ毎にモデルが異なっていることが予想される. さらに, 全データを用いた場合とカテゴリ毎に分け, 単回帰/重回帰による給水量予測モデル

を検討した. 交差検証(5-fold, 1000 回実施)時の平均絶対パーセント誤差(MAPE)を表 2 に示す. カテゴリ化と平均温度の積算値との回帰が最も MAPE が低くなった.

表 2 吸水量予測精度の比較

MAPE ± SD	単回帰	重回帰
全データ	0.22±0.07	0.22±0.06
C03 のみ	0.04±0.02	0.05±0.02

4. まとめと今後の課題

4.1 まとめ

本稿では 3 つの事例における高次元データを 24 時間というフレームに落とし込みデータ分析を進めた. 分析方法を主として教師無し学習を用い, データから特徴を抽出し, 抽出されたデータを他のデータと組み合わせることでビジネス活動に有用な知見を得ることができた.

4.2 高次元データから 24 時間周期データへ

今回, 時間と共に変化する時系列データを, 目的に合わせデータを変形/修正/圧縮し, 24 時間というフレームに当てはめて分析した. このフレームに当てはめることで, データの取り扱いやすくなり, その後の分析が容易になった. 例えば, 事例 1 では, 480 次元×約 450 世帯のデータを BN に適用でき, 事例 3 でも無数にある環境パターンの組み合わせも, データからあり得る組合せを抽出することで, 回帰分析による精度の高いモデリングが可能になった.

4.3 得られた知見

本手法によりビジネスに有用となる知見を得ることができた. 例えば, 事例 1 からは, 消費電力量のパターンから世帯の設置機器等を推定できる可能性を示した. これは今後普及が進むスマートメーターデータの活用することで 新たなサービスが拡がる可能性を示している. 一方でこれらのデータの管理(特にセキュリティー)面の重要性も示している.

また, 事例 2 は, ある条件のもとでは, 簡単なクラスタリング手法で異常検知に有用な知見が得られた. 今回想定したユースケースでは(例えば, 一日一回の異常検知), 本手法も十分に有効であった. また, 今回得られた結果を教師データとして用いれば, 教師有り学習による, より賢い異常検知も可能かもしれない. さらに今回の手法は, 別途照度計の設置や教師データを必要としないこと, 晴天時一日分のデータで良いことが 利点として挙げられる. しかし, 本手法を用いても, 時定数の短い瞬時的な発電異常の発見は難しいと考えられるため, 網羅的な異常検知にはさらに研究が必要となる.

IoT(Internet of Things)技術の進歩で, 農業ビッグデータも蓄積されつつある. 例えば, 今回用いた植物工場内の環境データもその一つである. 事例 3 では環境データを分析し, 温室内の環境状態を離散化(カテゴリ化)できた. さらに給水量データと組み合わせることで, 精度の高い給水量予測モデルを構築できた. 精度の高さから, このモデルはある程度環境要因による給水量への影響を相殺できたと考えられる. 植物の生長状態が「吸水量」と関係すると考えられるため, 「給水量 ≒ 吸水量」と環境データから, 植物の生長状態や健康状態を推定可能になるかもしれない. また, 本稿における事例は特定の温室の一つの作物についての結果だが, 同様枠組みにより, 温室/作物に限定せず給水量モデルの構築は可能と考えられる.

4.4 今後の課題

3 つの事例の共通した課題は, k の決定である. 今回はヒトの感覚やわかりやすさを重視し決定した. 実際, 結果は作業従事者には理解しやすかったが, このような主観的な方法を用いることで, 見逃してしまう現象もあるかもしれない. 今後は k を変化させ, それぞれどのようなパターンが生成されるのかを検証していく必要がありかもしれない.

参考文献

[Lloyd 1982] S. Lloyd: Least squares quantization in PCM, IEEE Trans Inform Theory, IT-28, pp. 129-137, 1982

[Wu 2008] Xindong Wu, et al.: Top 10 algorithms in data mining, Knowl. Inf. Syst., Vol. 14, pp. 1-37, 2008

[Hofmann 1999] Hofmann, T.: Probabilistic latent semantic analysis , Proc. of Uncertainty in Artificial Intelligence , pp289-296, 1999

[石垣 2011] 石垣司, 竹中毅, 本村陽一: 百貨店 ID 付き POS データからのカテゴリ別状況依存的変数間関係の自動抽出法, オペレーションズ・リサーチ, Vol. 56, No. 2, pp.77-83, 2011

[本村 2003] 本村陽一: ベイジアンネットソフトウエア BayoNet, 計測と制御, Vol. 42, pp. 693-694, 2003

[Verdu 2006] S.V. Verdu, M.O. Garcia, C. Senabre, A.G. Marin, F.J.G. Franco: Classification, Filtering, and Identification of Electrical Customer Load Patterns Through the Use of Self-Organizing Maps, Vol. 21, pp. 1672 – 1682, 2006

謝辞

経済産業省:大規模 HEMS 情報基盤整備事業にご協力いただいた桑名市, 四日市市の 450 世帯の皆様, KDDI 株式会社をはじめとする i エネコンソーシアム参加企業, 特に CCC マーケティング株式会社, 株式会社ヤマダ電機に感謝いたします.

本成果の一部は経済産業省の委託業務「平成29年度新エネルギー等の保安規制高度化事業(電気施設保安技術高度化の評価・検証事業)」である.

本成果の一部は国立大学法人千葉大学との共同研究の一部である.

最後に, 本研究にあたりご指導いただいた, 産業技術総合研究所本村陽一首席研究員に感謝いたします.