1B4-OS-01a-3

機械学習を用いたナノデバイス出力パルス波形による生体識別技術の開発

Identification of Microorganisms using Machine Learning Based on Nanopore Sensing Output

吉田　剛　Takeshi Yoshida　鷲尾　隆　Takashi Washio　石井　陽　Akira Ishii　川合　知二　Tomoji Kawai
谷口　正輝　Masateru Taniguchi　筒井　真楠　Makusu Tsutsui　横田　一道　Kazumichi Yokota

大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research Osaka University

While interest on the identification of micro/nano-objects including microorganisms by nanopore sensing technique is now significantly increasing, its sensing ability is limited by the uncertainty coming from the small sensing volume and the heat noise. In this study, we applied machine learning techniques to the identification of microorganism based on the output signals of the nanopore sensors. We show that machine learning technique achieved high accuracy of the identification by effectively using the features of the output signals.

1. はじめに

ナノセンシング，微量計測，量子計測など先端センシングデバイス開発分野では，微細・微量な対象を計測するためのデバイスが次々と開発されつつあり，この分野で日本は世界をリードする立場にある．しかしながら，これら多くのデバイスは，計測系や計測対象が微小であるが故に，対象の部分的な情報のみを出力し，かつ出力が熱雑音などの影響を受けることが多い [Rosenstein 12]．そのために，実用的に高精度な計測結果を得るには，時空間や特徴空間上での出力情報に基づく適切な推定が必要となる．一方，統計処理やパターン認識を含めた広義の機械学習を中心とする技術は，このような推定や情報統合によって高精度，高信頼な結果を得ることに長じている．これらを背景として，先端センシングデバイス開発分野において，機械学習技術への期待が高まっている．筆者等は，内閣府が実施する革新的研究開発推進プログラム (ImPACT) の「進化を超える極微量物質の超迅速多項目センシングシステム」に携わっている [ImP 14]．その主要テーマの 1 つに，ナノ・マイクロポアによる細菌・ウイルスの高感度センサの開発がある．ナノ・マイクロポアセンサは，ナノ・マイクロスケールの穴を用いた微小対象物のセンサであるが，それ単独では上述したような計測上の制限を有する [Coulter 53, Luo 14, Kozak 11]．

このような制限を克服ないし軽減すべく機械学習技術を用いて，マイクロポアにより計測された複数種類の細菌・ウイルスからの出力データを解析し，その種類別個数分布の推定を行うことが本研究の目的である．我々は，出力パルス波形の種々の特徴を抽出して適切に組み合わせることで，対象試料中の細菌種別の個数分布を一定の精度で推定することができた．その手法と結果について報告する．

2. マイクロポアと測定実験

図 1 に示すように，ナノポアないしマイクロポアはナノからマイクロメートルスケールの穴を有する仕切り平板であり，電解質溶液で満たした空間の両側に電圧を加えることで一定の電流が流れる構造を持ち，微小な細菌やウイルスが穴を通過する瞬間を，その閉塞による電流減少パルスによって検出するセンサである．そのパルス波形には，細菌やウイルスの形状や特性が反映されるが，穴を通過する角度や姿勢などで波形に多様性を生じ，かつ電流が微小であるためノイズも多い [Goyal 15]．

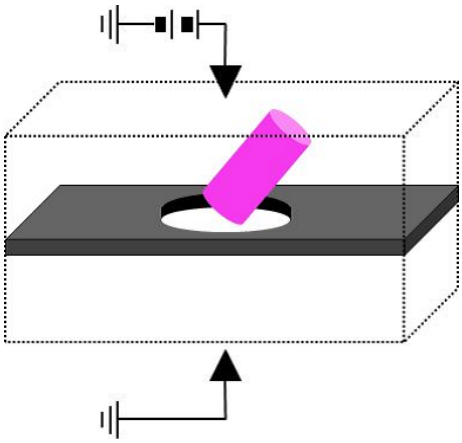

図 1: ナノ・マイクロポアセンサの概念図

本研究で対象とする実験においてはマイクロポアの直径は 3.0μm，ポアの深さ（仕切り平板の厚さ）は 50nm である [Yukimoto 13, Tsutsui 16]．試料として，大腸菌（Esherichia coli）と枯草菌（Bacillus Subtilis）の 2 種類の細菌を用いた．この 2 種の細菌は，僅かな大きさの違いや鞭毛の有無など特徴に差異はあるものの個体差も大きく，マイクロポアにより検出したパルスの大きさや形状が似ており識別が難しい．まず，大腸菌，枯草菌それぞれ 1 種のみをマイクロポアに通過させ，その際の電流値波形を 1000kHz サンプリングで測定し，パルス部分を含む電流値時系列データを得る．デバイスの実用化のためには高性能かつ廉価であることが必要であるから，精度を損なわない程度にサンプリング周波数を小さくすることが望ましい．そこで，一定の精度を保持したままどの程度までダウンサンプリングできるのかを調べるために，1000kHz の時系列データを元データとして，512kHz〜4kHz までダウンサンプリングした低周波数時系列データを作成する．これら時系列データからパルス部分のみを抽出し，大腸菌，枯草菌それぞれ

連絡先: tyoshida@ar.sanken.osaka-u.ac.jp

のパルスデータを作成する．このうちの一部を学習用データとし，残りについては，細菌 2 種のパルスデータをある一定の個数比にて混合し，細菌種別の個数分布推定のテスト用対象データとする．これらのパルスデータを解析し，個数分布推定の精度が，細菌種の個数比や電流測定値の周波数にどのように依存するのかを調べる．

3. 個数分布推定アルゴリズム

大腸菌（Esherichia coli）と枯草菌（Bacillus Subtilis）のパルスが混在しているデータを入力とし，このデータ内に含まれている大腸菌／枯草菌それぞれの個数を推定したい．そのために，まず初めにいずれの細菌種のパルスであるかがわかっている大腸菌データ N_E 個，枯草菌データ N_B 個のうちからそれぞれ RN_E，RN_B 個をランダムに取り出し，個数分布推定のための学習用データとする．ここで $0 < R < 1$ である．次に，残りの大腸菌データ $(1-R)N_E$ 個，枯草菌データ $(1-RN_B)$ 個のうちから 枯草菌 : 大腸菌 $= 1 : s$ となるように非復元抽出したものを混合して，大腸菌 n_E 個，枯草菌 n_B 個を含むデータを作成する．これを個数分布推定する対象のテスト用データとする．大腸菌／枯草菌それぞれの学習用データから取得したパラメータを元にして，2 種の細菌が混在したテスト用データに何個ずつの大腸菌／枯草菌が含まれているかを求める 2 値分類問題が，我々の解くべき問題である．ただし本研究においては，個別パルスを分類するのではなく，それぞれの細菌種が何個ずつあるのかという個数推定を行う．

3.1 パルス特徴量の定義

扱う対象データであるパルスデータは時系列の電流値データであり，大腸菌／枯草菌それぞれの何らかの特徴を反映したパルス波形であると考えられる．このパルス波形を特徴づける特徴量として以下の量を用いる．各種記号の定義を図 2 に示す．

(1) 波長 $\Delta t = t_e - t_s$

(2) ピーク波高 $|h| = |x_p - x_0|$

(3) ピーク位置比 $r = t_p - t_s / t_e - t_s$

(4) ピーク尖度 $k = (1/m)\sum_{t \in T_b}(t - \mathrm{ave}[t])^2$
電流値が $b|x_p - x_0|$ を横切る時刻の集合を T_b とし（$b = 0.3$ または 0.7），m を T_b に含まれる時刻 t の総数とする．ave$[t]$ は t の平均値である．すなわち，ピーク尖度はパルスピークの 30% または 70% の電流値をもつ時刻集合の分散を意味する．

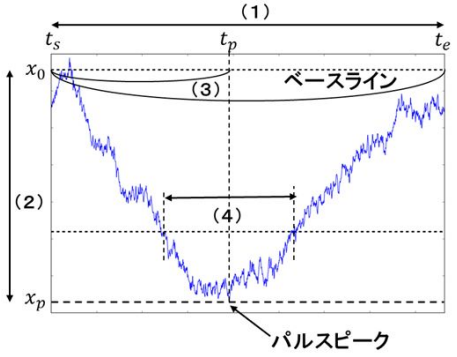

図 2: パルス波形の特徴量

3.2 カーネル密度推定と混合分布による尤度計算

細菌種別が同一であっても，細菌個々の個体差やノイズなどにより測定されるパルスの特徴量は同一の値にはならず，ある分布に従って計測される．この分布は未知であるが，測定される実データを元にしてカーネル密度推定法により以下のように推定する [Silverman 84]．パルス特徴ベクトル $\mathbf{x}$（$\mathbf{x}$ は，$\{\Delta t, |h|, r, k\}$ のうちいずれか d 個からなる）が測定される確率密度関数を以下のようにガウスカーネルを用いて定義する．i は細菌種を示す添字で，$i = E$（大腸菌）または B（枯草菌）とする．

$$p_i(\mathbf{x}) = \frac{1}{RN_i}\sum_{j=1}^{RN_i}\frac{1}{(2\pi c^2|\mathbf{\Sigma}_i|)^{d/2}} \times \exp\left\{-\frac{(\mathbf{x}-\mu_{ij})^{\mathrm{T}}{\mathbf{\Sigma}_i}^{-1}(\mathbf{x}-\mu_{ij})}{2c^2}\right\}$$

μ_{ij} $(j = 1, \ldots, RN_i)$ は細菌種 i の学習用データそれぞれの特徴量の値であり，$\mathbf{\Sigma}_i$ はこの学習用データ特徴量の分散共分散行列である．c はガウスカーネル幅のパラメータで，今回は $c = 0.1$ とした．

求めたいものはテスト用データの細菌種ごとの個数の推定値である．テスト用データ総数を $N = n_E + n_B$ とする．

パルス $\mathbf{x}$ が測定される頻度は，細菌種 i において $n_i p_i(\mathbf{x})$ の頻度であるから，各種細菌が独立に混合されたデータにおける測定頻度は $n_E p_E(\mathbf{x}) + n_B p_B(\mathbf{x})$ となる．すなわち，複数種の細菌が混合したデータにおいてパルス特徴ベクトル $\mathbf{x}$ が測定される確率は

$$p(\mathbf{x}) = \frac{1}{N}[n_E p_E(\mathbf{x}) + n_B p_B(\mathbf{x})]$$

と表せる．これによりテスト用データの特徴量セット $\mathbf{D} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ に関する尤度関数は

$$\prod_{\mathbf{x}\in\mathbf{D}} p(\mathbf{x})$$

となり，これを最大化する $\mathbf{n} = [n_E, n_B]^{\mathrm{T}}$ が，最尤な細菌種別の個数分布となる．

$$\mathbf{n} = \arg\max_{\mathbf{n}} \prod_{\mathbf{x}\in\mathbf{D}} p(\mathbf{x})$$

この $\mathbf{n}$ は以下で述べるように EM アルゴリズムにより求める．

3.3 EM アルゴリズムによる反復解法

細菌種別の個数分布 $\mathbf{n} = [n_E, n_B]^{\mathrm{T}}$ を求めるために，以下に示す EM アルゴリズムの一種である Hasselblad 法を用いる [Frühwirth-Schnatter 06]．

初期値設定 $\mathbf{n} = [n_E, n_B]^{\mathrm{T}}$ の初期値として，${n_E}^{(0)} = {n_B}^{(0)} = N/2$ と設定する．

E ステップ 現在のパラメータ ${n_i}^{(t)}$（$i = E$ または B）を用いて，以下の負担率（特徴ベクトル $\mathbf{x}$ が観測された元で，その特徴ベクトルが細菌種 i のものである事後確率）を計算する．

$$\frac{{n_i}^{(t)} p_i(\mathbf{x})}{\sum_{l=E,B} {n_l}^{(t)} p_l(\mathbf{x})}$$

M ステップ 以下の反復式によりパラメータ ${n_i}^{(t)}$ を更新する．

$$
\begin{aligned}
{n_i}^{(t+1)} &= \sum_{\mathbf{x}\in\mathbf{D}} \frac{{n_i}^{(t)} p_i(\mathbf{x})}{\sum_{l=E,B} {n_l}^{(t)} p_l(\mathbf{x})} \\
&= {n_i}^{(t)} \sum_{\mathbf{x}\in\mathbf{D}} \frac{p_i(\mathbf{x})}{\sum_{l=E,B} {n_l}^{(t)} p_l(\mathbf{x})}
\end{aligned}
$$

この反復式は，前小節で導入した尤度関数を未定乗数法により最適化することで得られるものであり，収束が保証されていることが証明されている．

収束判定 以下の判定に基づいて計算を終了する．本実験においては $\alpha = 0.1$ とした．

$$
\max_{i=E,B} \left(\left| {n_i}^{(t)} - {n_i}^{(t-1)} \right| \right) \leq \alpha
$$

4. 実験結果とまとめ

個数推定実験は以下のように交差検定により行った．2. 節で述べたようにパルスデータを学習用／テスト用の 2 つにランダムに分け，個数推定を行う．一度個数推定が終わったら改めてパルスデータ全体からランダムに学習用／テスト用データに切り分けて再度，個数推定を行う．このようにパルスデータをランダムに学習用／テスト用に切り分けることを 50 回行い，そのそれぞれで個数推定を行い交差検定とする．

以下のように，50 回の試行それぞれで得た重み付き平均相対誤差を用いる．

$$
\begin{aligned}
&\frac{1}{50} \sum_{j=1}^{50} \left\{ \sum_{i=E,B} \left(\frac{n_i}{N} \cdot \frac{|n_i - \hat{n}_i^{(j)}|}{n_i} \right) \right\} \\
&= \frac{1}{50} \sum_{j=1}^{50} \left\{ \sum_{i=E,B} \frac{|n_i - \hat{n}_i^{(j)}|}{N} \right\}
\end{aligned}
$$

$\hat{n}_i^{(j)}$ は j 回目の試行におけるテスト用データに対する細菌種 i の推定個数である．

結果の一例として，1MHz サンプリングデータに対し特徴量組として「尖度 k（30% ）」,「波高 $|h|$」を用いて推定した結果を表 1 に示す．数値は全て 50 回の試行の平均値である．この例が示しているのと同様に，全般的な傾向として混合比 s が小さい時には小数の細菌種を過大評価する傾向にあるが，概ね 10〜16%程度の誤差で大腸菌と枯草菌の個数を推定可能であることがわかる．

次に，異なる特徴量組み合わせに対して重み付き平均相対誤差を比較する．表 1 に示した 7 通りの混合比に対する誤差の平均値を指標とし，サンプリング周波数に対する依存性を図 3 に示す．この結果によると，高サンプリング周波数において

表 1: 混合比 s の推定精度への影響（特徴量は k と $|h|$）

混合比 s	n_E	$\hat{n}_E$	n_B	$\hat{n}_B$	誤差
0.1	15	27.45	146	133.55	0.16
0.2	29	37.22	146	137.78	0.12
0.3	44	56.19	146	133.81	0.14
0.35	51	61.12	146	135.88	0.12
0.4	58	65.42	146	138.58	0.11
0.45	66	68.01	146	143.99	0.10
0.5	73	70.19	146	148.81	0.10

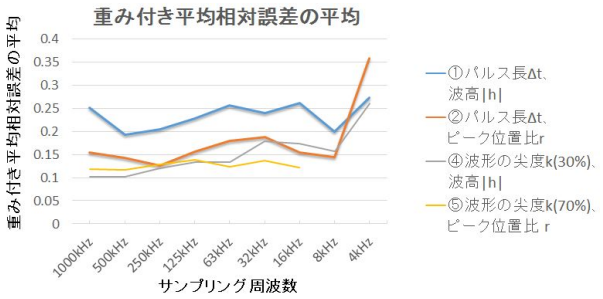

図 3: サンプリング周波数と特徴量の推定精度への影響

「尖度 k（30% ）と波高 $|h|$」の組み合わせが最もよい識別精度を示し，誤差 10〜13% 程度であった．中程度のサンプリング周波数においては「尖度（70% ）とピーク位置比」が最良で，誤差 12〜13% 程度，低サンプリング周波数では「波長とピーク位置比」が最良の組み合わせであり，誤差 14〜15% 程度であることがわかった．また，4kHz データは共通して精度が悪く，データとして適切ではないことがわかった．

以上のように小数細菌を過大評価してしまうという傾向はあるものの，各サンプリング周波数ごとに特徴量を適切に組み合わせることで，低サンプリング周波数データであっても一定の精度で細菌種別の個数分布推定を実現することができた．

5. おわりに

我々は，試料中に含まれる大腸菌／枯草菌の個数分布推定を目的として，マイクロポアを通過した大腸菌／枯草菌が出力したパルス波形解析を行った．カーネル密度推定法と混合分布を用いた尤度計算を行い，これを EM アルゴリズムの一種である Hasselblad 法により最適化することで個数推定を行った．この推定精度が大腸菌／枯草菌の個数比にどのように依存するか，一定精度を維持したうえでサンプリング周波数をどの程度まで小さくできるか，高精度で個数推定するためにパルス波形を特徴づけるのに適した特徴量およびその組み合わせは何かについて調べた．その結果，以下の知見を得た．

- 細菌の個数比に大きな差がある場合には小数細菌を過大評価してしまうという傾向がある．2 種細菌の個数が同程度になるにつれ精度は良くなる．
- 高サンプリング周波数（125-1000kHz 程度）において，「尖度 k（30% ）と波高 $|h|$」の特徴量組み合わせが最良であり，誤差は 10〜13% 程度であった．
- 中程度サンプリング周波数（16-125kHz 程度）では「尖度（70% ）とピーク位置比」が最良の組み合わせであり，誤差 12〜13% 程度であった．
- 低サンプリング周波数（8-16kHz 程度）では「波長とピーク位置比」が最良の組み合わせであり，誤差 14〜15% 程度であった．
- 4kHz サンプリングデータはどの特徴量組み合わせでも誤差が大きく，データとして不適切であった．

このように，細菌の個数差が大きい場合には小数細菌を過大評価してしまうという傾向はあるものの，各サンプリング周波数ごとに適切な特徴量の組み合わせを選ぶことで，10〜15% 程

度の一定の精度で細菌種別の個数分布推定を実現することができた

今後は，小数細菌を過大評価してしまう問題の解決と，より高精度の推定を実現することを課題とし，本研究を継続していく予定である．

参考文献

[Coulter 53] Coulter, W. H.: Means for Counting Particles Suspended in a Fluid., US patent 2,656,508 (1953)

[Frühwirth-Schnatter 06] Frühwirth-Schnatter, S.: *Finite Mixture and Markov Switching Models*, Springer, New York (2006)

[Goyal 15] Goyal, G., Mulero, R., Ali, J., Darvish, A., and Kim, M. J.: Low Aspect-ratio Micropores for Single-Particle and Single-Cell Analysis., *Electrophoresis*, Vol. 36, pp. 1164–1171 (2015)

[ImP 14] 進化を超える極微量物質の超迅速多項目センシングシステム, http://www.jst.go.jp/impact/program09.html (2014)

[Kozak 11] Kozak, D., Anderson, W., Vogel, R., and Trau, M.: Advances in Resistive Pulse Sensors: Devices Bringing the Void Between Molecular and Microscopic Detection., *Nano Today*, Vol. 6, pp. 531–545 (2011)

[Luo 14] Luo, L., German, S. R., Lan, W. J., Holden, D. A., Mega, T. L., and White, H. S.: Resistive-Pulse Analysis of Nanoparticles., *Annual Rev. Anal. Chem*, Vol. 7, pp. 513–535 (2014)

[Rosenstein 12] Rosenstein, J. K., Wanunua, M., Merchant, C. A., Drndic, M., and Shepard, K. L.: Integrated nanopore sensing platform with sub-microsecond temporal resolution, *Nature Methods*, pp. 487–492 (2012)

[Silverman 84] Silverman, B. W.: *Density Estimation for Statistics and Data Analysis*, Chapman & Hall/CRC, London (1984)

[Tsutsui 16] Tsutsui, M., He, Y., Yokota, K., Arima, A., Hongo, S., Taniguchi, M., Washio, T., and Kawai, T.: Particle Trajectory-Dependent Ionic Current Blockade in Low-Aspect-Ratio Pores, *ACS Nano*, Vol. 10, pp. 803–809 (2016)

[Yukimoto 13] Yukimoto, N., Tsutsui, M., He, Y., Shintaku, H., Tanaka, S., Kawano, S., Kawai, T., and Taniguchi, M.: Tracking single-particle dynamics via combined optical and electrical sensing., *Sci. Rep.*, Vol. 3, No. 1855 (2013)