論 文

予測方法用の最良の属性の組み合わせを抽出するマスター細胞を用いる免疫アルゴリズム

An Immune Algorithm that Uses a Master Cell to Find the Best Combination of Components for a Prediction Method

鈴木 優基 [1]　Palacios Pawlovsky, Alberto *

[1] 桐蔭横浜大学 大学院工学研究科医用工学専攻
* 桐蔭横浜大学 医用工学部臨床工学科

（2019 年 3 月 16 日　受理）

Ⅰ．はじめに

平成 29 年の主な死因別死亡率の割合では、**図 1** に示すように全体の約 27.8% が、悪性新生物（癌）で 1 位であり、その次に多いのが心臓疾患である。その後に肺炎、脳卒中や腎不全などがみられる [1)]。腎不全は糖尿病に相関関係があるとされ、心疾患、脳血管疾患にも深く関わりがあるとされている。各疾患においては、様々な予防方法や診断方法が存在している。

疾患に罹患した際の死亡率を左右するのはその疾患を発見したタイミングであり、日々の健康診断などの健診によって疾患を発見できるかが重要となる。本論文では、医師が健診を行う際に、参考とすることができるような診断を高い精度で行う方法とそれを実装するソフトウェアの開発について述べる。

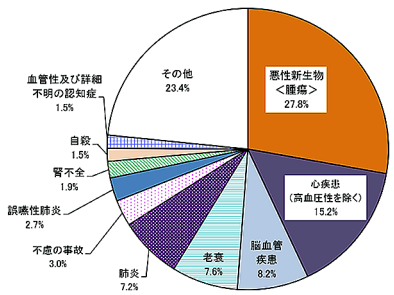

図 1　主な死因別死亡数の割合（平成 29 年）

最新の医療診断として、AI（Artificial Intelligence：人工知能）の一種の技術である機械学習が注目されており、患者の白血病タイプを特定するなどの例がある。膨大なデータから疾患を特定していくという作業が AI に適しており、精度の高いことも注目されている。本方法は予測（診断）に機械学習の分類方法の一つである k- 近傍法（k-Nearest Neighbor: k-NN）を使用している。また、予測に用いるデータの属性数が多い場合は、最良の属性の組み合わせを特定する必要があるため、組合せ最適化問題のアルゴリズムにおいて、問題に依存しないヒューリスティクスを導入した。自然界の免疫のシステムを参考にした免疫アルゴリズムを最良の組み合わせの検索に用いた。

* Palacios Pawlovsky, Alberto: Professor, Faculty of Biomedical Engineering, Toin University of Yokohama. 1614 Kurogane-cho, Aoba-ku, Yokohama 225–8503, Japan
[1] Suzuki Yuki: Graduate School of Engineering, Toin University of Yokohama

Ⅱ．予測に用いるデータ

1．乳癌再発生予測用のデータ

予測のためのデータとしては、対象者が多いデータや、倫理上の関係から匿名で集計されているデータ、一般公開されているデータが望ましいものとされる。オープンデータとして有名であることから、比較検討が容易であると考え、カリフォルニア大学が公開しているUCIデータベースの三つのデータを使用した。データは患者ごとに複数の項目（属性）で構成されている[2)]。

UCIデータとは米国のカリフォルニア大学のIrvineキャンパスの大学サイトで公開されているものを指す。乳癌再発生予測用のデータは米国の Wisconsin 大学の病院で取得され、提供されたものであり、乳癌に罹患して新たに検診を受けた198人の患者のデータである。その詳細を**表1**に示す。**表1**に含まれる10個の属性の平均値、標準誤差、最大値のほかに、腫瘍サイズ、癌のあるリンパ球の数、患者の区別番号（ID）、再発生か否かを示す記号、再発生までの時間などを記録した35項目のデータとなっている。同機関は、診断用基本データ、診断用データ、及び予後用データの3種類のデータを提供しているが、本研究では予後用データだけ使用した。

表1　乳癌データの属性（項目）

	項目
1	細胞の半径
2	組織の外観
3	周りの長さ
4	面積
5	滑らかさ
6	凝縮性
7	凹み
8	凹面側の固定参照点の数
9	左右対象
10	フラクタル次元

2．心臓病発生予測用のデータ

心臓病の発生の予測の検討には、クリーブランドクリニック財団で臨床科学データとして取得され、米国のカリフォルニア大学のIrvineキャンパスの大学サイトに公開されているデータを用いた。患者のデータは14項目であるが、第14項目に記録されているデータは心臓病なのか否かを表しているため、予測には使用しないこととし、残りの13項目を検討に用いた。

このデータには患者303人の記録があるが、欠損がみられるデータ6個を不適切なデータと判断し取り除き、計297名のデータを使用した。検討に用いた13項目を**表2**に示す。

表2　心臓病データの属性（項目）

	項目
1	年齢
2	性別
3	胸痛
4	安静時血圧
5	血清コレステロール（mg / dl）
6	空腹時血糖> 120（mg / dl）
7	安静時心電図検査結果
8	最大心拍数
9	労作時狭心症
10	運動によって誘発される ST
11	ST 勾配
12	血管数
13	サラセミア

3．糖尿病発生予測用のデータ

糖尿病の発生の予測を対象とした検討で使用したデータは、米国に居住するピマ・インディアンを対象としたUCIの糖尿病データである（**表3**）。

各々の人のデータは、8項目で構成されて記録されている。全768人のデータのうち、欠損がみられるデータを除外した393人のデ

表3　糖尿病データの属性（項目）

	項目
1	妊娠した回数
2	血漿グルコース
3	拡張期血圧
4	三頭筋皮下脂肪厚さ
5	ボディマス指数
6	糖尿病血統機能
7	年齢
8	2 時間の血清インスリン

ータを検討に使用した。

Ⅲ．予測用の方法

1．k-NN 法の概要

k-NN 法は既に分類の分かる（予測用）データを使って、未知の（診断したい）データの分類を予測する。このため、予測したいデータから予測用データの１つ１つのデータとの類似度を測り、昇順整列した類似度で、k 個の予測用データを用いて未知のデータの種類（診断結果）を予測する方法である。**図２**にその２次元での概念図を示している。

k-NN 法の予測精度を測るために、疾患ごとのデータを予測用と検証用に分ける。予測用と検証用のデータは両方ともすでに結果が判明している臨床データである。検証用データを未知データ（**図２**では「?」の印のデータ）とし、予測用データとの類似度を測って k-NN 法で分類させる。k-NN 法の予測結果が実際の結果と当たった割合は、k-NN 法の正答率となる。

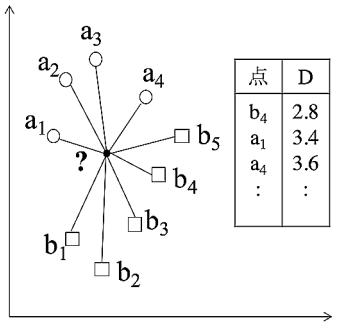

図２　k-NN 法の概念図

2．k-NN 法の詳細

（1）類似度の指標として用いる距離

k-NN 法で用いるデータは、ランダムに予測用データと検証用データに分ける。予測を行うときは類似度をデータ間の距離として計算する。距離の概念は多種多様に存在する。

本方法で採用した距離は、ユークリッド距離、マンハッタン距離、チェビシェフ距離、キャンベラ距離、ソーレンセン距離、マハラノビス距離の６つである。**表４**に示す６つの距離の計算式を実装して、k-NN 法の処理過程に用いた。

表４　類似度の計測に用いた距離

距離	計算式
Euclid	$d_{Euc}=\sqrt{\sum_{i=1}^{n}\lvert P_i-Q_i\rvert^2}$
Manhattan	$d_{Man}=\sum_{i=1}^{n}\lvert P_i-Q_i\rvert$
Chebyshev	$d_{Che}=\max_i\lvert P_i-Q_i\rvert$
Canberra	$d_{Can}=\sum_{i=1}^{n}\frac{\lvert P_i-Q_i\rvert}{P_i+Q_i}$
Sorensen	$d_{Sor}=\frac{\sum_{i=1}^{n}\lvert P_i-Q_i\rvert}{\sum_{i=1}^{n}(P_i+Q_i)}$
Mahalanobis	$d_{Mah}=\sqrt{\sum_{i=1}^{n}\frac{(P_i-Q_i)^2}{s_i^2}}$

（2）予測計測：データの分割

疾患を予測するために採用した k-NN 法は、新たなデータを分類するときに、そのデータを分類の分かる既存のデータと比較して、その類似度によって分類を予測する方法である。上記にも述べたように、使えるデータを予測用データと検証用データの２つに分ける必要がある。分類に使用するデータの割合は自由に決められるが、詳細な検討を可能にしながら処理時間を抑える設定として 10%、20%、30%、40%、50%、60%、70%、80% および 90% の９つの分類用データの大きさで k-NN 法の予測の正答率を測定した。

（3）データの加工：標準化

公開データはそのまま使用できるが、正答率の向上のために、データの加工が必要である場合がある。その加工方法が標準化である。本検討で使用した標準化方法は２種類である。

その一つ目は、データセット内の最小値を元の値から差し引き、その結果を最大値と最小値の差で除する方法である。これによってデータの範囲は０と１の間となる。

二つ目の標準化方法は、各値から平均値を差し引き、その結果を標準偏差で除算するも

のである。これによってデータの平均が0になり、標準偏差は1となる。元の公開データを含めると、合計で3種類のデータを使用して、検討を行った。

Ⅳ．免疫アルゴリズム

1．免疫アルゴリズムの概要

k-NN法を使用した疾患に関する予測の正答率の改良に関しては様々な方法が考えられる。その改良の対象として第一に考えられるのは臨床データの属性（項目）の選択である。k-NN法は臨床データの検査項目の類似度を測るために通常ではすべての項目を使用して予測を行う。しかし、使用する項目を取捨選択することで正答率に大きな影響を与えるということが分かっており、項目の取捨選択に関しての様々なアプローチがある。また、すべての項目の組み合わせ（全網羅）も考えられるが、項目数が多いと、組み合わせ数は多大になり、処理時間が膨大となるため、事実上全項目の組み合わせの試み（検証）は不可能である。

最適な項目の組み合わせではなく現実的に検索可能な最良の組み合わせを求める遺伝的アルゴリズムや免疫アルゴリズム（IA法）などの多種多様な近似方法（ヒューリスティックアルゴリズム）が考案されてきている[3)]。本アプローチは最良の項目の組み合わせの検索に、免疫アルゴリズムを採用した。

本免疫アルゴリズムでは、患者の臨床データの項目使用の有無を指定するものを細胞に対応させ、**図3**に示す様に項目使用の有無を0か1かで表す。同図内の5つの0と1で構成されている1次元行列は、免疫細胞を二つ表している。その細胞の評価は一つずつではなく、複数を同時に行い、その集合は細胞群と呼ばれる。

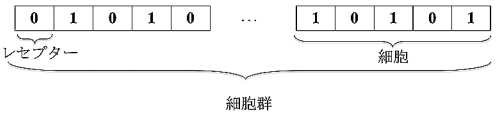

図3 免疫アルゴリズムの細胞と細胞群の概念

2．記憶細胞の生成

免疫システムとは生体内に侵入した異物（抗原）を認識し、抗体をつくりだし素早く排除するものである。また、過去に排除した抗原に対しては、その特徴を記憶しており、より早く排除するとされている。

免疫アルゴリズムは生体における免疫機能を模倣し、工学的にモデル化したシステムであると定義されるため、免疫のシステムをどのようにとらえ、どう表現するかによってそのアルゴリズムの構造は様々なものになる。

本免疫アルゴリズムでは、初期の（項目の組み合わせの情報を持つ）細胞群がランダムに生成される。その後k-NN法での評価が行われ、その中の最良の正答率を持つ細胞を記憶細胞として保存する。その処理の概念図を**図4**に示す。

実装した免疫アルゴリズムの処理の流れを**図5**に示す。まず、免疫候補細胞の初期細胞

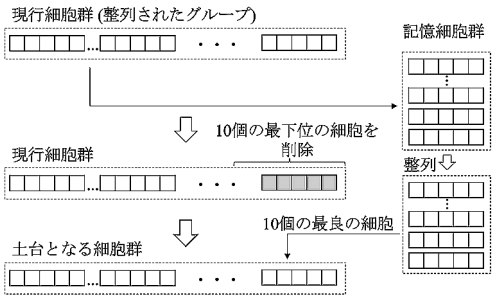

図4 記憶細胞の選定と免疫細胞の入れ替え

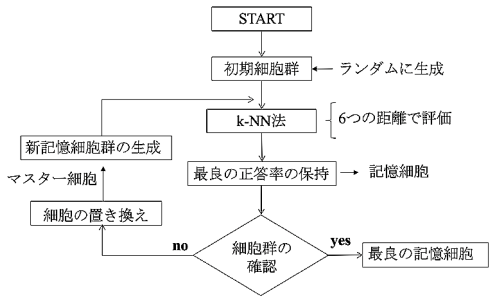

図5 免疫アルゴリズムに基づいた検索過程

群はランダムに生成される。その後、6つの距離によりk-NN法で各々の細胞（項目の組み合わせ）の評価が行われ、その中の最良の正答率を持つ細胞を記憶細胞として保存する。

ある一定まで既存記憶細胞の中身に基づいてマスター細胞を作成する。このマスター細胞により新免疫細胞群の生成が行われる。この処理は、100の細胞の細胞群、100の細胞群のグループを生成するように設定し、上位の記憶細胞10個を出力するように行われる。

Ⅴ．マスター細胞の生成

本検討で考案したマスター細胞とは、免疫アルゴリズムの記憶細胞群に保存された細胞（項目の組み合わせ）の集団を対象に、すべての項目の要素を受け継ぐように新しく生成した一つの細胞である。マスター細胞を作成する際には、記憶細胞群に含まれる上位10個の細胞だけを対象に、それらすべての出現頻度の高い特徴を参照する。マスター細胞に最上位の記憶細胞の特徴を受け継がせるために、属性を選択する値の平均値を計算して、その値が0.5を超えればマスター細胞のその位置の値を1に、超えない場合は0に設定するようにした。**図6**には、記憶細胞群の中の五つの細胞を対象とした例を示している。同じ位置に所属する項目の1と0の数字を比較すると、項目の位置ごとに出現頻度の高い方の数字（属性を用いるか否かを決定する値）がマスター細胞に引き継がれていることが確認できる。

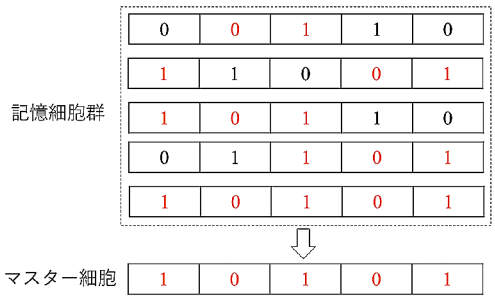

図6　マスター細胞の生成

Ⅵ．予測の正答率の評価と比較

1．免疫アルゴリズムの効果

データの全ての項目を使用したk-NN法で得られた最大平均正答率をベースライン（比較基準）として、本方法の結果と比較する。

条件を同じにするため、試行回数を100回として比較を行った。**図7**に、各疾患の最大平均予測正答率の比較を示す。白色の棒を本方法の結果、青色（灰色）の棒をベースラインの結果として示している。乳癌データ、心臓病データ、糖尿病データのすべてにおいて、正答率の上昇が確認できた。その上昇の範囲は0.5%～3%となっている。乳癌データ、糖尿病データにおいては正答率の上昇は0.5%～1%であるが、心臓病データに関しては3%の上昇が見られ、本方法は有効であると言える。

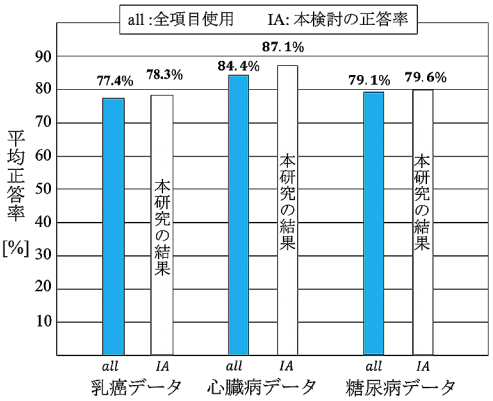

図7　ベースラインと本方法の平均正答率比較

2．他の研究成果との比較

海外で発表されている、同じ臨床データを対象としたさまざまな研究成果がある。これらの研究では、試行回数は10回と設定されているため、条件を合わせるために本方法の試行回数10回で得た結果を比較に用いる。

図8に、各疾患の最大平均予測正答率の比較を示す。試行回数10回の結果では、乳癌

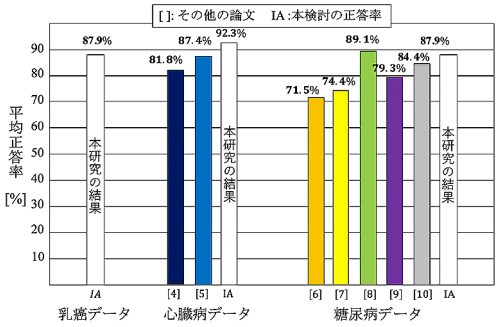

図8　他の技術論文の結果との比較

データを対象とした技術論文はなかったため表示していない。糖尿病データでは、本方法は高い最大平均正答率 87.9% を出したが、最大の正答率を記録している論文 [8] に約 1% 届かない結果となった。心臓病データにおいては、本方法は 92.3% の平均正答率であり、最大となっていることが分かる。

3．処理時間

本方法での免疫アルゴリズムの処理（最良の組み合わせの検索）に要した時間を**表5**に示す。乳癌データでは約 5 日、心臓病データで約 23 日、糖尿病データで約 76 日の処理時間を必要とした。項目数の多い乳癌データよりも、患者数が多い糖尿病データでの処理に時間がかかっていることが分かる。

表5　免疫アルゴリズムの処理時間

データ	処理時間(分)
乳癌データ	8399.10
心臓病データ	32435.26
糖尿病データ	109424.94

Ⅶ．まとめと今後の展望

k-NN 法に良い影響を与えるヒューリスティックアルゴリズムの検討は長く行われてきており、本方法の免疫アルゴリズムにおいては Manhattan 距離を使用した場合に最も良い結果が得られることが多かった。しかし、Manhattan 距離以外にも Euclid 距離、Chebyshev 距離、Sorensen 距離、Mahalanobis 距離などでも良い予測の平均正答率が得られており、予測を Manhattan 距離にのみ限るような設定はしない方が良いと考える。またそれに関連して、本方法の検討過程においては Canberra 距離の計算を使用して良い平均正答率をみられるというようなことはなかった。これを参考にして、免疫アルゴリズムでの検討に限っては Canberra 距離を採用しないという決定をすることによって、処理時間の短縮ができると考える。

また、新たな細胞群の生成に用いられる交叉率と突然変異率の変更の検討が必要であると考えられ、突然変異率を 1% 以下にする検討を行うことができる。

交叉率に関しても最良と思われる 60% の設定を使用して検討を行ったが、他の設定での検討は行っていないため、交叉率を 60% 以上と定め、10% 間隔ずつ上昇させて検討を行うことができると考える。

【参考文献】

1) https://www.mhlw.go.jp/toukei/saikin/hw/jinkou/geppo/nengai17/dl/kekka.pdf
2) http://archive.ics.uci.edu/ml/datasets.html
3) A. Palacios P., “An Immune Algorithm with an Evolutionary Scheme for Component Selection for the kNN Method,” Proc. of the IEEE Congress on Evolutionary Computation (CEC 2018), pp.2554–2560, July, 2018.
4) S.Ozsen and S. Gunes, “Attribute weighting via genetic algorithms for attribute weighted artificial immune system (AWAIS) and its application to heart disease and live disorders problems,” Expert Systems with Applications, No.36, pp.386–392, 2009.
5) G. Dudek, “An Artificial Immune System for Classification with Local Feature Se-

lection," IEEE Trans. on Evolutionary Computation, Vol.16, No.6, pp.847–860, December 2012.
6) A. Secker and A. A. Freitas, "WAIRS: improving classification accuracy by weighting attributes in the AIRS classifier," in Proceedings of the IEEE Congress on Evolutionary Computation, pp.3759–3765, September 2007.
7) A. Sharma and D. Sharma, "Clonal Selection Algorithm for Classification," Proceedings of the International Conference on Artificial Immune Systems (ICARIS), pp.361–370, 2011.
8) M. Saidi, M. A. Chikh and N. Settouti, "Automatic Identification of Diabetes Diseases using a Modified Artificial Immune Recognition System 2," Proceedings of the Third International Conference on Computer Science and its Applications, CEUR workshop proceedings, Vol.825, paper 20, 2011.
9) M. S. Uzer, N. Yilmaz and O. Inan, "Feature Selection Method Based on Artificial Bee Colony Algorithm and Support Vector Machines for Medical Datasets Classification," Scientific World Journal, Vol.2013, Article ID 419187, 2013.
10) Kemal Polat, "Intelligent Recognition of Diabetes Disease via FCM Based Attribute Weighting," International Journal of Computer and Information Engineering, Vol.10, No.4, pp.783–787, 2016.