CF-013

ランキング学習を用いた顕微授精時の精子選別尺度獲得の検討

Investigation for acquisition of sperm selection criterion in Intracytoplasmic Sperm Injection by using learning to rank

佐々木 勇人 † 岸 大輝 † 山本 みずき ‡
竹島 徹平 ‡ 湯村 寧 ‡ 濱上 知樹 †
Hayato SASAKI Daiki KISHI Mizuki YAMAMOTO
Teppei TAKESHIMA Yasushi YUMURA Tomoki HAMAGAMI

1. はじめに

顕微授精の成功率を高めるためには精子選別の質を高めることが重要である．しかしながら精子選別作業は胚培養士の経験知に大きく依存しており，精子選別の尺度は統一されていない．このことは，精子選別作業の再現性や信頼性の担保を困難にしている．

そこで本稿では精子選別作業の支援及び信頼性確保を目的として，精子選別尺度を機械学習により獲得する．具体的には精子画像に対するランキング学習により胚培養士の知見に基づいた精子選別尺度を獲得する．加えて,「胚培養士間で共通の選別尺度」と「各培養士に特有の選別尺度」を切り分けるための分析を行う．

両者の切り分けが実現されれば，精子のどんな特徴に着目して選別を行っているのかを一般化できるだけでなく，培養士個人やクリニックごとの選別尺度にどのような傾向があるのかを分析する一助となる．更に，培養士ごとの選別尺度の差異が定量化されることにより，顕微授精における精子の選別方法と着床率・受精率との関連性をより定量的に評価可能になると期待される．

2. ランキング学習

ランキング学習とはインスタンス集合 $\mathcal{X}$ に対して順位付けを行う学習である．このランキング学習は教師ラベルの与え方によってポイントワイズ，ペアワイズ，リストワイズの 3 種類に大別され，一般的にポイントワイズ，ペアワイズ，リストワイズの順に性能が高くなる．一方で同じ順でラベル付与のコストも高くなる．本稿では性能とラベル付与コストのバランスからペアワイズ手法を扱う．

まず，$\mathcal{X}$ 中から任意に選択した 2 つのインスタンス $(\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{X} \times \mathcal{X}$ がある尺度 $F^* : \mathcal{X} \to \mathbb{R}$ によって比較可能であり，$\boldsymbol{x}$ が $\boldsymbol{x}'$ よりも高いランク付けがなされることを $\boldsymbol{x} \succ \boldsymbol{x}'$ と表記する．すなわち

$$F^*(\boldsymbol{x}) > F^*(\boldsymbol{x}') \Rightarrow \boldsymbol{x} \succ \boldsymbol{x}' \tag{1}$$

とする．このようにインスタンス同士の比較が尺度 F^* に基づいて行われていると仮定すると，ランキング学習は F^* を推定する学習問題として定式化される．

ペアワイズ手法では以下の損失 $\mathcal{R}$ を最小化する．

$$\mathcal{R}[F] = \frac{1}{|\mathcal{D}|^2} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{D} \times \mathcal{D}} C(P_{ij}, F(\boldsymbol{x}_i) - F(\boldsymbol{x}_j)) \tag{2}$$

ただし，$\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^N$ は訓練データである．また，$P_{ij} := Pr(\boldsymbol{x}_i \succ \boldsymbol{x}_j)$ は教師信号として働く．更に，$C : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ は損失関数であり，$\hat{P}_{ij} := sigmoid(F(\boldsymbol{x}_i) - F(\boldsymbol{x}_j))$ としたときの P_{ij} と $\hat{P}_{ij}$ とのクロスエントロピーが使われることが多い．

ペアワイズ手法と分類される手法には LambdaRank[1] や LambdaMART[2] などがあり，ニューラルネットワークや勾配ブースティングがベースの学習モデルが多い．これらの手法は，式 (2) に関する勾配降下法という点で共通している．

3. 精子選別尺度の獲得

ランキング学習は教師あり学習であり，訓練用のデータセットが必要になる．しかしながら精子選別尺度獲得のための訓練用データセットは存在しないため，新たにデータを収集する必要がある．

そこで本研究では，精子に対する 5 段階のグレード情報 (1 から 5) を T 人の胚培養士から収集して教師データを作成した．なお 5 段階のグレード情報以外に，判断ができない画像に対しては「その他」という特殊なラベルがついている (0 に対応)．したがって，精子画像の特徴 $\boldsymbol{x}_i$ に対して $y_i^t \in \{0, 1, \ldots, 5\} (t = 1, 2, \ldots, T)$ という複数のラベルが付与される．

顕微授精に用いられる精子画像はオープンデータセットではなく，大規模なサンプルサイズを確保することは難しい．ランキング学習ではニューラルネットワークやブースティングが用いられるが，データ数の少なさに起因する過学習を回避するために，精子選別尺度の学習ではアンサンブル学習の一種である勾配ブースティングを用いる．

勾配ブースティングでは学習器 F が弱学習器 $h \in \mathcal{H}$ の加法モデルで表現される．すなわち

$$F(\boldsymbol{x}) = \sum_{s=1}^{S} \beta_s h_s(\boldsymbol{x}) \tag{3}$$

と表現される．ここで，$\beta_s \geq 0$ は弱学習器 h_s の信頼度である．

ブースティングステップ s で追加される弱学習器 h_s は，$\{\partial \mathcal{R}[F] / \partial F|_{F=F(\boldsymbol{x}_i)}\}_{i=1}^N$ を最もよく近似する $h \in \mathcal{H}$ が選択される．以降，弱学習器が近似するターゲットを $g_i = \partial \mathcal{R}[F] / \partial F|_{F=F(\boldsymbol{x}_i)}$ とする．また，$\boldsymbol{g} := (g_1, g_2, \ldots, g_N)$ とする．

†横浜国立大学
‡横浜市立大学附属市民総合医療センター

4. 胚培養士間で共通の精子選別尺度

胚培養士 t の精子選別尺度を F^{t*} とする．また胚培養士間で共通する尺度を F^{0*} とし，F^{t*} を次のように表す．

$$F^{t*}(\boldsymbol{x}) = F^{0*}(\boldsymbol{x}) + f^{t*}(\boldsymbol{x}) \tag{4}$$

このとき F^{0*} を共通尺度とすると f^{t*} は胚培養士 t 特有の尺度を表す．式 (4) で表されるモデルは文献 [3] で用いられている．

精子選別作業の再現性や信頼性を担保するためには，いかにして共通尺度 F^{0*} と各胚培養士特有の尺度 f^{t*} を切り分けるかが重要である．

まず，推定値 F^0 と f^t の和で F^t を表現した場合，いずれの関数も $\mathcal{X}$ から $\mathbb{R}$ への写像であるため，F^0 で重要視される特徴と f^t で重要視される特徴の比較が可能である．

また，共通尺度 F^{0*} や各培養士に特有な尺度 f^{t*} はブースティングの加法モデルで推定する．したがって，F^{t*} を両者の和で表現するモデルはブースティングとの相性が良い．ブースティングでは逐次的に弱学習器が追加される．すなわち，ブースティングを用いた場合，共通尺度を獲得しうる損失関数の元で最適化をしたあとで培養士 t に特化した損失関数 $\mathcal{R}^t$ の最適化をすることで，F^0 と f^t を獲得可能である．

一方で，式 (4) を勾配ブースティングで最適化することは，共通尺度と特有尺度を切り分けるための必要条件ではあるが，十分条件ではない．何をもって共通尺度とするかが重要である．そこで本稿では，共通尺度を獲得するための方法として以下の 2 手法を分析する．

なお，勾配ブースティングでは損失関数の勾配が得られれば損失関数そのものは用いる必要がない．したがって，共通尺度を得るための勾配決定方法に着目する．

4.1 マルチタスク学習による共通尺度の獲得

本稿では胚培養士 t に対応するランキング学習をタスク t とよぶ．各タスク t には式 (2) で示した損失関数が設定されている．

マルチタスク学習では，タスク t に対する損失 $\mathcal{R}^t(t = 1, 2, \ldots, T)$ の最小化を同時に行うことで F^0 を推定する．具体的には $\mathcal{R}[F] = \sum_{t=1}^{T} \mathcal{R}^t[F]$ を最小化する．

勾配ブースティングでは損失に対する勾配 $\partial\mathcal{R}[F]/\partial F$ を用いて強学習器 F を更新するため，各タスクの損失和を最小化する場合には，

$$\boldsymbol{g}^t := (\ldots, \frac{\partial \mathcal{R}^t[F]}{\partial F}|_{F=F^0(\boldsymbol{x}_i)}, \ldots) \tag{5}$$

を用いた勾配和

$$\boldsymbol{g}^0 = \sum_{t=1}^{T} \boldsymbol{g}^t \tag{6}$$

を近似する弱学習器を逐次的に獲得する．

4.2 多目的最適化による共通尺度の獲得

上記のマルチタスク学習では $\boldsymbol{g}^0 = \sum_{t=1}^{T} \boldsymbol{g}^t$ を用いた勾配降下を行っていた．しかしながら，この手法は各ブースティングステップで $\mathcal{R}^t(t = 1, 2, \ldots, T)$ すべてを小さくする保証はない．場合によってはあるタスク a の損失 $\mathcal{R}^a$ を小さくする一方でタスク b の損失 $\mathcal{R}^b$ を大きくする可能性がある．これを回避するのが MGDA(Multiple Gradient Descent Algorithm)[4] である．MGDA は deep learning を用いたマルチタスク学習にも利用されており，その有効性が確認されている [5].

MGDA は各ブースティングステップで全タスクの損失を小さくするような $\boldsymbol{g}^0$ を算出する．更に，この手法はパレート最適な解 (より厳密には pareto stationary な解) への収束が保証されている．

MGDA における具体的な $\boldsymbol{g}^0$ は次式を満たすように算出される．

$$\|\boldsymbol{g}^0\|_2 = \min_{\alpha_1,\ldots,\alpha_T} \left\{ \|\sum_{t=1}^{T} \alpha^t \boldsymbol{g}^t\|_2 \mid \sum_{t=1}^{T} \alpha^t = 1, \alpha^t \geq 0 \, \forall t \right\} \tag{7}$$

すなわち，$\boldsymbol{g}^t, t = 1, \ldots, T$ に関する凸包上のベクトルで，ノルムが最も小さくなるベクトル $\boldsymbol{g}^0$ を用いて勾配降下を行う．

この MGDA により F がパレート最適になると $\|\boldsymbol{g}^0\|_2 = 0$ となる．このとき，$\boldsymbol{g}^t, t = 1, \ldots, T$ は釣り合っており，どの方向に勾配降下をしてもいずれかの $\mathcal{R}^t$ は大きくなる．

5. 実験

前章に示した 2 つの手法により共通尺度を獲得し，その性能を比較する．本研究では，共通尺度と培養士に特有の選別尺度の切り分け方を重要視している．特に，共通尺度そのものの性能，及び，共通尺度と特有尺度の非類似性が重要となる．そこで

- 共通尺度 F^0 によるランキングの性能
- 共通尺度 F^0 と培養士 t 特有の選別尺度 f^t との類似性
- f^t 及び $f^{t'}$ $(t \neq t')$ の類似性

の 3 つに関して分析を行った．

5.1 データセット

2 クリニック計 6 人 (各クリニック 3 人ずつ) の胚培養士の協力を得てグレード情報を収集した．精巣内精子採取術 (TESE) の検体を撮影した顕微動画像中の各精子に対してグレードを付与している．また，各精子は頭部を中心に正方形に切り抜いた動画像と対応づいており，それらの動画像に対して 6 人の胚培養士によるグレードが与えられている．

動画像中には 222 の精子が存在しており，それらの精子の動画像を訓練用とテスト用に 4:1 で分割した．なお，精子動画像のフレーム数は一定ではなく，最小フレーム数が 15, 最大フレーム数が 444 である．結果として訓練データセットの全フレーム数は 19129，テストデータセットの全フレーム数は 4572 となった．

更に，各フレームを回転させて頭部の向きが全フレームで一致するようにした．ただし，TESE の検体には動いている精子がほとんど存在しないため，精子動画

表 1 勾配ブースティングのハイパーパラメータ

イテレーション数 (共通尺度)	1000
学習率 (共通尺度)	0.1
イテレーション数 (特有尺度)	1000
学習率 (特有尺度)	0.01
サンプリング回数 M	10
弱学習器	回帰木 (最大深さ 3)

像の先頭フレームの向きを手作業で揃えたあとで，以降のフレームも同じ角度だけ回転させている．

5.2 実験設定

勾配ブースティングによりランキング学習を行う．勾配ブースティングの諸元は表 1 の通りである．表が示すように，共通尺度の学習を行った後に特有尺度の学習を行う．

なお，損失関数には 2 章で述べたクロスエントロピーを用いている．このとき, 胚培養士 t に関して

$$P_{ij}^t = \begin{cases} 1 & \text{if } (y_i^t > y_j^t) \\ \frac{1}{2} & \text{if } (y_i^t = y_j^t) \\ 0 & \text{if } (y_i^t < y_j^t) \end{cases} \tag{8}$$

として，P_{ij}^t と $\hat{P}_{ij}^t = sigmoid(F^t(\boldsymbol{x}_i) - F^t(\boldsymbol{x}_j))$ のクロスエントロピーを算出する．

ペアワイズ手法では異なるインスタンスを比較する必要があり，本実験では精子動画像の比較が必要となる．一方で精子動画像に映る精子はほとんど動かない．そこで，各ブースティングステップにおいて，各精子動画像の 1 フレームをランダムサンプリングして勾配計算を行った (確率的勾配法に対応)．更に，学習を安定させるために，ランダムサンプリングによる勾配算出を M 回行い，そうして得られた M 個の勾配ベクトルを連結させている (ミニバッチ勾配降下法に対応)．すなわち，訓練データセット中の精子動画数を N としたときに，勾配ベクトル $\boldsymbol{g}^0$ は $N \times M$ 次元となる．

各フレームの特徴としては Hisgrams of Oriented Gradients(HOG) を用いている．フレームサイズは精子頭部が収まるように定めており，64×64 ピクセルとした．このフレームに関して，4×4 ピクセル, 8×8 ピクセル, 16×16 ピクセルの各セルサイズで HOG を抽出し，それら 3 つの HOG を連結することで特徴量 $\boldsymbol{x}$ とした．なお，いずれの HOG もセル内のビン数は 8 であり，ブロックサイズは 1×1 セルである．最終的な $\boldsymbol{x}$ の次元は 2728 次元となった．

5.3 実験結果

5.3.1 共通尺度の性能分析

ランキング学習によって獲得される共通尺度を NDCG(Normalized Discounted Cumulative Gain) で評価した．NDCG は真のグレードの分布に左右されるため，比較指標として精子の動画に対してランダムに順位を割り当てた場合の NDCG を算出している．表 2 にマルチタスク学習及び多目的最適化によって獲得される共通尺度の NDCG を示す．

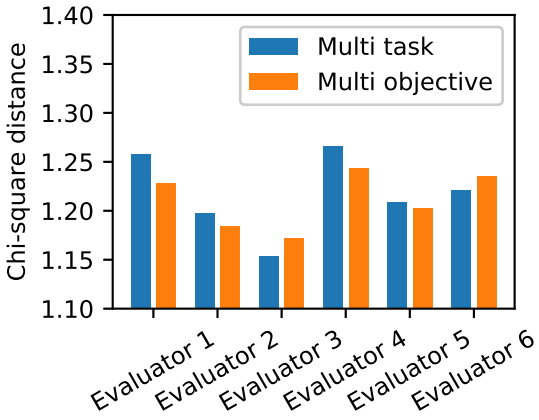

図 1 共通尺度及び特有尺度の特徴重要度間のカイ二乗距離

結果から分かるように，いずれの培養士のグレードに関しても多目的最適化によるランキングの NDCG が最も高いことが分かる．一方でマルチタスク学習のなかで胚培養士 1 に関する NDCG がランダムランキングの NDCG よりも低くなってしまっている．

5.3.2 共通尺度と培養士個人に特有の選別尺度との差異

共通尺度 F^0 と特有尺度 f^t の差異を，両者の特徴重要度のカイ二乗距離によって比較する．ブースティングモデルの特徴重要度 $\boldsymbol{\gamma}$ は弱学習器 h_s の特徴重要度を $\boldsymbol{\gamma}_s$ としたときに以下のように算出した．

$$\boldsymbol{\gamma} = \sum_{s=1}^{S} \beta^s \boldsymbol{\gamma}_s \tag{9}$$

図 1 にマルチタスク学習及び多目的最適化で得られた共通尺度・特有尺度の差異を示す．カイ二乗距離の大小をマルチタスク学習と多目的最適化とで比較すると，大小関係に一貫性が無いことが分かる．

いずれかの学習手法が共通尺度と特有尺度をうまく分離できているとすると，カイ二乗距離は大きくなるはずである．この実験ではカイ二乗距離の大小に明確な差異があらわれておらず，両尺度の分離に関して 2 種法の明確な差異は現れなかった．

5.3.3 各胚培養士に特有の選別尺度の比較

培養士 t 及び t' $(t \neq t')$ に特有の選別尺度 f^t, $f^{t'}$ を特徴重要度のカイ二乗距離で比較する．このとき ${}_TC_2$ 通りのカイ二乗距離が計算される．

図 2 にマルチタスク学習及び多目的最適化で得られた ${}_TC_2$ 個のカイ二乗距離を示す．カイ二乗距離の中央値を両手法で比較すると，マルチタスク学習の方が大

表 2 マルチタスク学習及び多目的最適化で学習したランキング関数に対する NDCG@10

	培養士 1	培養士 2	培養士 3	培養士 4	培養士 5	培養士 6
ランダムランキング	0.635	0.638	0.583	0.350	0.506	0.422
多目的最適化	**0.665**	**0.767**	**0.734**	**0.594**	**0.745**	**0.611**
マルチタスク学習	0.611	0.752	0.692	0.568	0.712	0.590

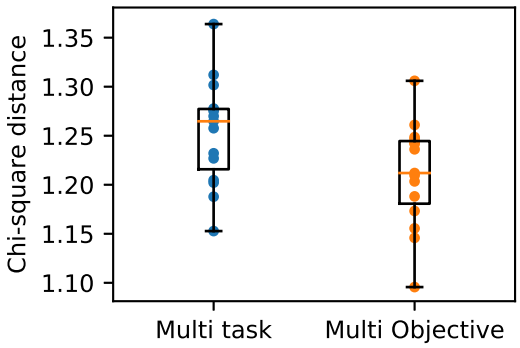

図 2 特有尺度 f^t 及び $f^{t'}$ の特徴重要度間のカイ二乗距離の分布

きいことが分かる．このことは，マルチタスク学習のほうが特有尺度間の差異が大きいことを意味している．共通尺度と特有尺度の分離度が大きいと，特有尺度はタスクによって大きく異なることが予想される．したがって本実験結果はマルチタスク学習のほうが共通尺度と特有尺度の分離度が大きいことを示唆している．

5.4 考察

マルチタスク学習では全タスクの重要度を等しいものと仮定する．したがって，あるタスクの損失を増加させるかわりに全タスクの平均損失を下げる可能性がある．一方で MGDA を用いた多目的最適化では，常に全タスクの損失を下げるように最適化する．NDCG によりランキング性能を比較した表 2 の結果を見ると，マルチタスク学習では特定のタスク (胚培養士 1) の性能が下がっていることが確認された．

MGDA を用いた多目的最適化は，$\boldsymbol{g}^t (t = 1, 2, \ldots, T)$ が釣り合うようなパレート最適解が得られる．このとき，$\boldsymbol{g}^t$ はそれぞれコサイン類似度が低くなっている．したがって，パレート最適解を取得した後にタスク特有の学習を始めるということは，他タスクとの類似度が低い $\boldsymbol{g}^t$ に対して弱学習器を訓練することになる．これによって，f^t がタスクごとに異なるように学習されるはずである．

しかしながら，図 1 の結果は MGDA の優位性を示しておらず，図 2 の結果ではマルチタスク学習のほうが両尺度を分離できることが示唆されている．マルチタスク学習のほうが両尺度を適切に分離できるのか，若しくは，MGDA の尺度分離を阻害する何らかの要因があるのか，さらなる検討が必要である．

6. おわりに

本稿では胚培養士の精子選別尺度をランキング学習により推定するとともに，胚培養士間で共通の選別尺度と胚培養士個人に特有の選別尺度を切り分ける方法を検討した．

ある胚培養士の選別尺度が，共通尺度と特有尺度の和で表現されるモデルを考えた場合，両尺度における特徴重要度を算出可能である．この分離モデルは勾配ブースティングにおける加法モデルで表現可能であり，共通尺度を獲得したのちに各タスクにファインチューニングするだけで済む．

一方で，共通尺度と特有尺度を切り分けるためには，共通尺度の推定方法が特に重要である．マルチタスク学習及び多目的最適化の 2 手法で共通尺度の推定を行う実験を行ったところ，マルチタスク学習のほうが両尺度の分離度が大きいことが示唆された．

この結果は予想と異なっており，今後は MGDA を用いた多目的最適化において両尺度の分離度が大きくならなかった原因について分析を進める．

参考文献

[1] Christopher J Burges, Robert Ragno, and Quoc V Le. “Learning to rank with nonsmooth cost functions”. In *Advances in neural information processing systems*, pp. 193–200, 2007.

[2] Christopher JC Burges“From ranknet to lambdarank to lambdamart: An overview”. *Learning*, Vol. 11, No. 23-581, p. 81, 2010.

[3] Olivier Chapelle, Pannagadatta Shivaswamy, Srinivas Vadrevu, Kilian Weinberger, Ya Zhang, and Belle Tseng“Boosted multi-task learning”. *Machine learning*, Vol. 85, No. 1-2, pp. 149–173, 2011.

[4] Jean-Antoine Désidéri“Multiple-gradient descent algorithm (MGDA) for multiobjective optimization”. *Comptes Rendus Mathematique*, Vol. 350, No. 5-6, pp. 313–318, 2012.

[5] Ozan Sener and Vladlen Koltun. “Multi-task learning as multi-objective optimization”. In *Advances in Neural Information Processing Systems*, pp. 527–538, 2018.