学 位 論 文 の 要 旨

少量データ下における機械学習及び関連手法の開発
（Developing Machine Learning and Related Methods with Small Data）

氏 名　　大野 侑亮

近年，人工知能分野の研究は，計算機の性能向上とともに，パラメータやデータサイズの大規模化が進んでいる．しかし，大規模化が進むほど計算機やデータ収集に対してコストが高くなるため，一部の研究室や企業しか扱うことが出来なくなる可能性がある．また，機械学習を適用したい問題は，世の中に多数存在しているが，現在の深層学習では，大量にデータを用意しないと汎化能力を確保しにくい問題がある．したがって，汎化能力を考えた少量データ下における機械学習では，根拠のしっかりとした情報をなるべく多く引き出すことが求められている．

本研究では，少量データの例として，腎臓にある糸球体上皮細胞の SEM 画像を使用する．糸球体上皮細胞は腎疾患により，形態の変性が起こると考えられているが，主観の入った評価が多く，これを客観的に捉えている文献は少ない．また，このデータの問題点として，変性のおおよその傾向をつかむことが出来るが，全体に対して当てはまる特徴を挙げることが非常に困難であり，人が評価する際は，形状を観察して総合的に判断していると考えられる．よって，この曖昧な評価基準を機械学習によって再現することで，腎疾患により変性している事実を人の目を介さず判断することで客観性を担保したい．そのために，高精度で判別可能な機械学習モデルが必要となる．

したがって，本論文では，糸球体上皮細胞の足突起画像の少量データセットを用いて，腎疾患によって糸球体上皮細胞の足突起が形態の変性を起こすことを客観的にとらえる医学的課題とこれらの変性を機械学習で捉えるために，少量データ下における学習手法を開発する工学的課題の解決を目的とし，関連する 3 つの手法の開発を行った．将来的には，これらの手法を組み合わせることで，さらなる高精度化を目指すが，本論文では，個々の手法の開発まで行った．

まず，前処理を目的とした曖昧な境界を持つ物体の領域抽出方法の開発を行った．我々が扱ったデータセットでは，興味の対象が存在する関心領域とその外側の非関心領域があり，これらの差が小さく境界線が判別しにくいといった問題があった．少量データ下においては，特徴の核心を突いた学習を行う必要があるため，これらの区別をつける必要がある．本手法では，白飛びや黒つぶれをしている要素の除外を目的とした輝度に関する要素，非関心領域に平らな領域が多いことに注目した輝度勾配に関する要素，画像中央に関心領域があり，外側に非関心領域が広がる可能性が高いことに注目した楕円フィルターの 3 要

素を根幹とし，開発を行った．手動で抽出した領域に対して，重なり度合いを評価した一致率によって，抽出精度を評価した結果，提案手法では 0.796 の一致率を得られた．

次に，少量データ下における機械学習では，教師データの少なさがしばしば問題になることに注目し，これらの拡張を目的として，少量のデータセットより元の画像群の特徴分布から大きく逸脱しない教師データとして扱えるような偽画像の生成手法の開発を行った．偽画像生成においては，クラス判別を取り入れた ACGANs をベースに開発を行った．ACGANs では，クラスの判別がつかなくなる Mode Collapse などが起きやすいといった欠点がある．提案手法である制約付き ACGANs では，2 つの並行な生成器と判別器を用いた．1 つ目のフローには，通常の GANs にある潜在変数から偽画像を生成する機能を，2 つ目のフローには，ACGANs でも扱っているクラスラベルから，新たに取り入れた形状ラベルを生成する機能をつけた．1 つ目のフローから 2 つ目のフローへ中間層の出力を加算することで，学習の方向性に制約を付けた．画像から算出した 5 つの指標（平均，分散，歪度，尖度，複雑度）に対して，教師データの分布の 1σ 区間に収まったものを教師データに近い画像が得られたと判断する評価を行ったところ，従来手法である ACGANs と比較し，135%高い精度で，本物に近い特徴を持つ偽画像を生成できた．

最後に腎疾患による糸球体上皮細胞の変性を客観的に捉えることを目的とし，人間の評価基準を再現することで高精度に判別可能な少量データ下における機械学習モデルの二段階学習の開発を行った．提案手法では，二段階の学習を行い，一段階目では，噛み合わせ部分を捉えた特徴領域から抽出した特徴量で学習を行う．続いて，二段階目で，画像全体から取得した局所画像に対して，一段階目で学習した事前知識を用いて，スコアの算出を行い，このスコアから計算した統計量を再度学習する．提案手法を 3 クラスの問題に対して適用した結果，予測精度として 78.5%を達成した．これにより医学的課題であった腎疾患による糸球体上皮細胞の変性を客観的に捉えることは十分に達成したと考えられる．

学 位 論 文 の 要 旨

少量データ下における機械学習及び関連手法の開発
（Developing Machine Learning and Related Methods with Small Data）

氏 名　　大野 侑亮

In recent years, research in the field of artificial intelligence has been increasing in scale in terms of parameters and data size, along with improvements in computer performance. However, the larger the scale, the higher the cost of computers and data collection becomes, which means that only a few laboratories and companies may be able to handle the technology. In addition, there are many problems in the world for which machine learning can be applied, but with current deep learning, it is difficult to secure generalization capability without a large amount of data. Therefore, machine learning with small data is required to extract as much well-founded information as possible, considering the generalization capability.

In this study, we use SEM images of glomerular epithelial cells in the kidney as an example of small data. Although glomerular epithelial cells are thought to undergo morphological degeneration due to renal disease, there is little objective information in the literature on this phenomenon, as it is often evaluated subjectively. One problem with this data is that although it is possible to obtain an approximate trend of degeneration, it is extremely difficult to identify characteristics that apply to the entirety of the cells, and it is thought that when people evaluate the cells, they make a comprehensive judgment based on observation of the shape. Therefore, we would like to reproduce this ambiguous evaluation criterion by machine learning to ensure objectivity by judging the fact of degeneration due to renal disease without human eye intervention. To achieve this, a machine learning model that can discriminate with high accuracy is necessary.

Therefore, this paper aims to solve the medical problem of objectively capturing the morphological degeneration of glomerular epithelial cell foot processes caused by renal disease using a small dataset of glomerular epithelial cell foot process images, and the engineering problem of developing a learning method under small data to capture such degeneration by machine learning. We have developed three related methods to solve these problems. In the future, we plan to combine these methods to achieve even higher accuracy.

First, we developed a method for extracting regions of interest in objects with ambiguous boundaries for preprocessing. In our dataset, there are regions containing

objects of interest and non-regions, and the difference between these regions is small, making it difficult to distinguish their boundaries. To learn the core features, it is necessary to accurately identify these regions, especially when the amount of data is limited. Our method includes three elements: a luminance element to exclude areas with whiteout or blackout, a luminance gradient element focusing on the fact that non-regions have many flat regions, and an ellipse filter that focuses on the fact that regions of interest are generally located in the center of the image while non-regions tend to be found outside of it. We evaluated the accuracy of the extraction by comparing it to manually extracted regions and measuring the degree of overlap. We found that our method achieved an agreement rate of 0.796.

Next, we developed a method for generating fake images from small data that can be used as additional training data without deviating significantly from the feature distribution of the original image set. This is important because small data is often a limitation in machine learning. The method is based on ACGANs, which have the ability to incorporate class discrimination. However, ACGANs are known to have a problem with mode collapse, which leads to indistinguishability between classes. To address this, we proposed a method that uses two parallel generators and discriminators: the first flow generates fake images from latent variables using a standard GAN, and the second flow generates new shape labels from the class labels used in ACGANs. The direction of learning is constrained by adding the output of the intermediate layer from the first flow to the second flow. We evaluated the fake images using five indices (mean, variance, skewness, kurtosis, and complexity) and determined that an image is similar to the real data if it falls within a 1σ interval of the distribution of the real data. As a result, we have succeeded in generating fake images with features closer to real images than ACGANs by using our method. This is a 135% improvement in accuracy over conventional methods.

Finally, we developed a two-step learning method for machine learning models under small data to objectively detect glomerular epithelial cell degeneration caused by renal disease, which can reproduce human evaluation criteria and discriminate with high accuracy. In the first step, features extracted from the feature region that captures the interdigitating part are used for learning. In the second step, scores are calculated using the prior knowledge learned in the first step on local images obtained from the entire image, and these scores are learned again. We applied the proposed method to three classes of problems and achieved a prediction accuracy of 78.5%. This result is considered sufficient to objectively detect glomerular epithelial cell degeneration caused by renal disease, which has been a medical problem.