ファジィ測度を使用したアクティブラーニングによる物体識別

○木村慶豪　濱上知樹　（横浜国立大学）

Object Classification using Fuzzy Measured Active Learning

∗Keigo Kimura and Tomoki Hamagami (Yokohama National University)

Abstract– In ICSI, a kind of ART, embryologists select sperm with good shape and motility from microscopic movies. This is a time-consuming and labor-intensive task, and there is a demand for systems that support sperm detection and evaluation. In general, a large amount of labeled data is required for image classification learning, but the data that can be annotated is limited due to the high level of expertise required to make decisions. Therefore, we adopt active learning and proposed the Query Strategy that uses fuzzy measure, which represents the certainty of belonging to a class, in order to efficiently retrieve queries based on the features of data with large individual differences within a class and adjacent classes.

Key Words: Active Learning, Fuzzy Measure, Assisted Reproductive Technology

1　はじめに

様々な分野で画像データが集積されるに伴い，研究や医療分野で扱われる専門的な画像中の物体識別を機械学習により行い，専門家の判断の支援をする需要が増大している。

機械学習による画像中の物体識別では物体のクラスをラベル付けするアノテーションを必要とする。しかし，専門的な画像のアノテーションには高度な知識を必要とするため作成できるラベル付きデータが限られる。また，同じ分野で使用された専門画像でも画像特徴がカメラの距離に起因する被写体の大きさ，ライティング，解像度を原因として変化するため撮影環境に依存しない学習データを作成できない。このため，画像に応じて専門家が自らアノテーションを行う必要がある。

アクティブラーニングは機械学習モデルがユーザに対して逐次的にデータのクラスを問い合わせる手法であり，少ないデータでモデルの精度を改善する学習手法である。このため専門的な画像の物体識別ではアクティブラーニングが最も有効に働く。

アクティブラーニングで機械学習モデルが問い合わせるデータは推論が不確かなデータであることを意味しており予測した決定境界の改善を目的とする。しかし，判別が難しい専門画像をアクティブラーニングで効率的に学習するとき，2 種類の不確かさを考える必要がある。1 つ目は複数のクラスで迷う不確かさ，2 つ目はどのクラスにも属さないような不確かさである。従来の不確かさの尺度では 2 つの不確かさを混同して考えるため一方の不確かさに追加されるデータが偏ったときに学習の効率が低下する場合がある。

ゆえに明示的に 2 種類の不確かさを別々に定義することが有効である。そのために本研究では各クラスへの帰属の確からしさを表すファジィ測度を用いたアクティブラーニングを提案する。

この手法の活用分野の 1 つに顕微授精がある。顕微授精とは生殖補助医療の一種であり採取された精子から正常な形状かつ運動性に優れた精子を選定し，直接卵子に注入する手法を指す。この作業は胚培養士によって行われるが顕微鏡中から精子を発見し評価を行うため多大な時間と専門的な技術を必要とする。機械学習では顕微鏡動画像を使用するが物体の判別の難しさ，画像が不鮮明さを原因として精子と他の細胞との判別が難しい。そこでアクティブラーニングを用いて胚培養士と機械学習モデルが対話的にデータを加えて顕微鏡中の物体識別学習を行う。その後機械学習モデルが顕微鏡動画像から精子を検出することで胚培養士の精子の発見と評価を支援できると考えている。

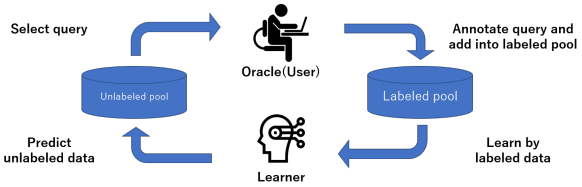

Fig. 1: Pool Based Active Learning

2　アクティブラーニング

アクティブラーニングはモデルの性能向上に寄与すると考えられるデータを神託 (Oracle) に問い合わせる半教師あり学習の一種である。モデルの性能向上に有効なデータはデータセット中の少数であり，そのようなデータを優先的にラベル付けすることで少ない訓練データ数で十分な汎化性能を出せる仮定に基づいている。アクティブラーニングの手法の 1 つに Pool Based がある。この手法はラベルが付与されていないデータから性能向上に有効とされるデータを問い合わせる手法であり，Fig.1 のようにデータの追加とモデルの学習が繰り返し行われる。

学習に有効とされるデータの選び方を Query Strategy と呼ぶ。タスクや学習済みのラベル付きデータセットの数に応じて最も有効となる Query Strategy は変化する。

3　提案手法

ラベルなしデータは十分に存在すること，撮影環境に応じて画像の特徴が変化する課題に着目して学習する画像データ毎にモデルを作成し，精度改善に有効なデータを逐次的に加えるアクティブラーニングを用いた学習モデルを採用し，必要となるデータの特性からファジィ測度を用いた Query Strategy を提案する。

モデルの精度改善に必要なデータを考察する。少量のデータセットで学習したとき，識別が困難となるデータは決定境界近傍や未学習の領域のデータである。提

案手法はそのような 2 種類の識別な困難なデータを追加して学習器の推定する密度分布を改善することを目的とする。前者はクラス間の密度分布の分離に寄与し，後者はクラスの密度分布の拡大に寄与する。

代表的な Query Strategy に Uncertainty Sampling がある。この手法はラベルなしデータを学習器で推論した事後確率 $P(y|\boldsymbol{x})$ を用いて各データに対する推論の不確実度を計算する [1)]。不確実度が高いデータを判別が難しいデータとみなして Query とする。しかし，確率測度を用いた Uncertainty Sampling は密度分布の分離と拡大に寄与するデータを区別できない。例えば 2 つのデータが存在し，一方はすべてのクラスに属していると判断したデータ，もう一方はどのクラスにも属していないと判断したデータとする。この場合，確率測度では同じ確率を返す。このように異なる不確かさを確率測度では混同するため Query が一方のデータに偏る。このため，効率的に学習ができない場合があると予想される。

そこで，2 種類の精度改善に寄与するデータを明示的に Query とするためにファジィ測度を用いた Query Strategy を提案する。各クラスへのファジィ測度 $\mathbb{G}(\boldsymbol{x})$ は式 1 のように計算される。

$$\mathbb{G}(\boldsymbol{x}) = \{g_1(\boldsymbol{x}), g_2(\boldsymbol{x}) \ldots, g_K(\boldsymbol{x})\} \tag{1}$$

ファジィ測度は確率測度と異なり 1 に正規化されない。このため，全てのクラスに属するようなデータに対しては高くなるように，どのクラスにも属さないようなデータに対しては低くなるようにファジィ測度が算出される。前者は分布の分離に寄与するデータであり，後者は分布の拡大に寄与するデータであるためファジィ測度によって 2 種類の精度改善に寄与するデータを明示的に取り出すことが可能となる。本手法では分布の分離に寄与するデータ $\boldsymbol{x}^*_{sep}$ と拡大に寄与するデータ $\boldsymbol{x}^*_{exp}$ を (2) 式で定義する。

$$\begin{aligned} \boldsymbol{x}^*_{sep} &= \arg\max_x \left\{ \sum_i g_i(x) - \max_i g_i(x) \right\} \\ \boldsymbol{x}^*_{exp} &= \arg\min_x \{\max_i g_i(x)\} \end{aligned} \tag{2}$$

分布の分離に寄与するデータの指標として各クラスへのファジィ測度の合計値から最大のファジィ測度を引いた値を使用する。値が大きいほど複数のクラスへの帰属の確かさが大きいことを意味しており，判断が難しいデータといえる。また，分布の拡大に寄与するデータの指標として最大となるクラスへのファジィ測度を使用する。最大ファジィ測度が小さいほどどのクラスにも属していないと判断しているため未学習の領域に存在するデータであるといえる。本手法では式 2 での 2 種類の不確実度に従って Query を取り出す。

4 実験

4.1 実験 1

画像識別学習においてファジィ測度を用いた Query Strategy が分布の分離と拡大に寄与するデータを明示的に与えることで既存の戦略と比較して効率的にモデルの改善を行えるかを確認した。

4.1.1 データセット

実験では MNIST 手書き文字分類データセットを使用し，60000 のデータをアクティブラーニングのための Pool とした。また，10000 のデータをテストデータとした。MNIST を使用したのはクラス間で類似したデータが存在し，同一のクラス内で形状に差が存在するため，顕微鏡動画像データセットの特徴と類似しているためである。実験では Pool を Unlabeled Pool，Labeled Pool に分割し，Labeled Pool を用いて学習する。学習後，Query Strategy に従い Unlabeled Pool から Query を作成し，Labeled Pool に追加する。これはラベルが与えられていないデータを学習器が推論してモデルの精度改善に向上するとされるデータを問い合わせる作業に対応している。

4.1.2 学習器の設定

学習器は各クラスに 1 つの勾配ブースティング決定木を作成し，ファジィ測度 $\mathbb{G}(\boldsymbol{x}) = \{g_1(\boldsymbol{x}), g_2(\boldsymbol{x}) \ldots, g_K(\boldsymbol{x})\}$ を推論する。各クラスへのファジィ測度 $g_i(\boldsymbol{x})$ は独立して計算される。学習では決定木に対応するクラスのデータが与えられたときには 1 となるように，対応しないクラスのデータが与えられたときには 0 となるように学習する。勾配ブースティング決定木の深さを 2，ブースティングの回数を 50 とした。確率測度を用いる Query Strategy では $\mathbb{G}(\boldsymbol{x})$ にソフトマックス関数を適用し，確率測度とした。

4.1.3 比較した Query Strategy

実験では以下に示す 4 つの Query Strategy を比較した。

- FuzzyMeasuredUncertaintySampling(Proposed)
- Entropy
- LeastCertainty
- RandomSampling(BaseLine)

提案手法ではファジィ測度をもとにモデルが予測する密度分布の分離と拡大の尺度を計算し，上位のデータを 1:1 の割合で Query とする。

Entropy では事後確率 $P(y|\boldsymbol{x})$ をもとにエントロピーが最大となるデータを Query とする。

$$\boldsymbol{x}^*_{ENT} = \arg\max_{\boldsymbol{x}} \left\{ -\sum_i P(y_i|\boldsymbol{x}) \log P(y_i|\boldsymbol{x}) \right\} \tag{3}$$

Least Certainty では最大となるクラス y^* の事後確率 $P(y^*|\boldsymbol{x})$ が最小となるデータを Query とする。

$$\boldsymbol{x}^*_{LC} = \arg\min_{\boldsymbol{x}} P(y^*|\boldsymbol{x}) \tag{4}$$

ベースラインである RandomSampling では Unlabeled Pool からランダムに Query とする。

4.1.4　実験手順

最初に与えるラベル付きデータを 100，アクティブラーニングのイテレーションごとに追加する Query の大きさを 100 とした。最初に与えるラベル付きデータは Query Strategy に依らず共通とした。その後テストデータで Accuracy を測定した。測定後，Unlabeled Pool からサンプリングした 1000 のデータを学習器に推論させ，Query Strategy に従い Query を作成し Labeled Pool に追加した。追加した Labeled Pool で再度学習を行いテストデータで Accuracy を測定した。Query の追加と再学習および Accuracy の測定は合計 20 回行った。

その後，Query の追加回数を横軸，最初に与えるデータを変更して実験を複数回行ったときの Accuracy の平均値を縦軸に取りグラフにプロットした。これは最初に与えるデータによって予測する決定境界が変更し，取り出す Query に変化が生じるためである。

4.1.5　実験結果

Fig.2 に実験結果を示す。20 回の Query での Accuracy は提案手法が 0.868，Entropy が 0.839，LeastCertainty が 0.853，RandomSampling が 0.835 となり，提案手法がベースラインの 4%，LeastCertainty と比較して 2%効率的に精度向上に寄与するデータを取り出せたと示された。特に，ラベル付きデータが少ない 1,2 回目の Query で提案手法の精度向上の傾きが大きい。これは提案手法によって取り出された Query がより精度向上に寄与していることを示している。

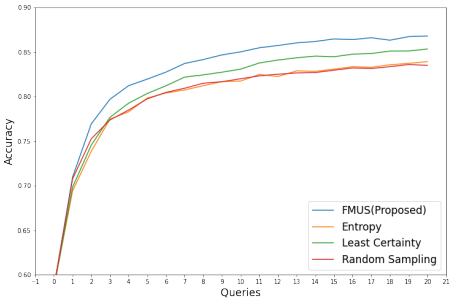

Fig. 2: Learning curves between Query Strategies

各 Query Strategy によって Labeled Pool に追加されたデータと推論結果の傾向を確認するために 1 回目の Query となったデータを比較する。Query Strategy 毎に取り出された不確実度スコアの上位 10 データと学習器による推論結果のヒストグラムを Fig.3 に示す。ここで，推論結果はすべてファジィ測度で示しており，また提案する Query Strategy では分布の拡大に寄与するデータの指標，分離に寄与するデータの指標それぞれから上位 5 つのデータを取り出している。Entropy や LeastCertainty を見ると全てのクラスに対して低いファジィ測度を示すデータを優先して Query としている。このようなデータは学習していない特徴空間に存在するデータと考えられるため予測する密度分布の拡大に寄与する。Entropy や LeastCertainty では，全てのデータに対して同じ確率を与えた時に最も不確実なデータとなるが，そのようなデータはすべてのクラスに属するように見えるデータとどのクラスにも属していないように見えるデータの 2 種類である。ラベル付きデータが少ないときは学習していない特徴空間が広いため，後者の密度分布の拡大に寄与しやすいデータが多くなる。ゆえにそのようなデータが優先的に Query となり，反対に分布の分離に寄与するデータが Query とならなかったと予想される。

このように画像データのような特徴空間の次元数が多いデータを扱うとき，少ないラベル付きデータでは学習していない特徴空間のデータが多い。そのようなデータに対して確率測度で推論するとき全てのクラスに同じ確率を与えるため不確実度スコアが最も高くなる。学習していない特徴空間が存在しなくなるまで Query とするため，確率測度を用いた Query Strategy では決定境界近傍のデータを Query とする回数が少なくなる。

一方で，提案手法では全てのクラスに対して低いファジィ測度を示すデータの他に複数のクラスに対して高いファジィ測度を示すデータを Query としている。提案手法が最も効率よくモデルの精度の改善を行えていることを考慮すると本データセットの学習においては決定境界近傍のデータと学習していない領域に存在するデータの両方を加えることがより効率的な学習に必要と考えられる。

なお，RandomSampling で精度改善の効率が相対的に悪かった理由は明瞭に識別できるデータを加えているからである。ただ 1 つのクラスのみ高いデータはモデルが明確な決定境界を持っていることを意味しているため，加えても密度分布の改善に寄与しない。結果として改善に有効なデータが少なくなったため，精度向上が遅くなったと考えられる。

4.2　実験 2

実験 1 でファジィ測度を用いて予測する密度分布の分離と拡大に寄与するデータを Query とすることで Uncertainty Sampling と比較して効率的にモデルの精度改善ができると示された。

しかし，学習器に推論結果に依存する Query Strategy はラベル付きデータが少ないときは決定境界が信頼できないため不確実度スコアが高いデータがモデルの精度向上に寄与しない場合がある [2)]。そこで，最も不確実度スコアの高いデータを Query とするのではなく，Unlabeled Pool から不確実度スコアが比較的高いデータの集合を取り出し，その集合からランダムにデータを取り出す工夫を提案する。この手法はラベル付きデータが少ないときの信頼できない決定境界に起因する課題を緩和するために行う。また，提案手法の予測分布の拡大に寄与するデータに関する Query にも同様の工夫を行う。分布の拡大に寄与するデータの指標が大きい場合，データが外れ値の場合も含まれるためであり，外れ値が集中した場合に学習の効率が低下すると考えられるからである。

4.2.1　ランダム性の導入

Uncertainty Sampling や提案手法では不確実度スコアの上位 n サンプルを Query としていたが上位 $n \times 5$ サンプルからランダムに n サンプルを取得する工夫を加えた。この工夫により学習器の推論結果に過剰に依存することを避け，ラベル付きデータが少ないときに探索的に精度向上に寄与するデータを Query とすることを期待する。

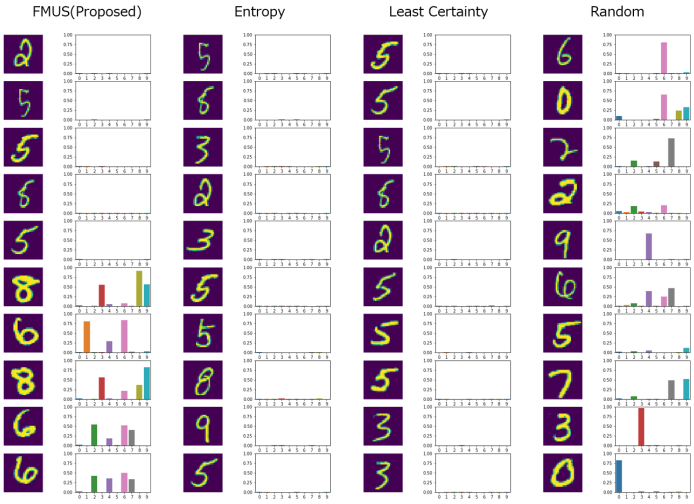

Fig. 3: Added data from each Query Strategy

4.2.2 実験設定

データセットは実験 1 と同様に MNIST を使用した。提案手法についてランダム性を導入しないモデル，ランダム性を導入したモデルを比較した。ベースラインとして Random Sampling も比較した。ランダム性を導入することでラベル付きデータが少ないときにおける精度向上の速度改善を目的としている。

実験は最初に与えるラベル付きデータを 100，追加する Query のサイズを 100 とし，アクティブラーニングの手順は実験 1 と同様に行った。

4.2.3 実験結果

ランダム性の導入有無による比較結果をそれぞれ Fig.4 に示す。

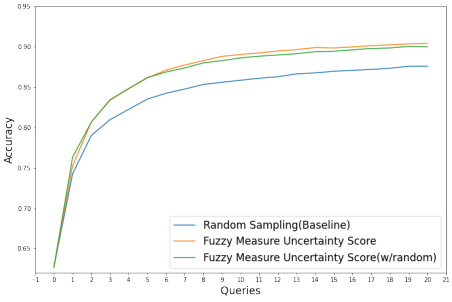

Fig. 4: Learning Curves of Fuzzy Measured Uncertainty Sampling

Fig.4 では 1 回目の Query においてわずかに傾きが増加し，学習が進行するにしたがってランダム性がないモデルの精度が上回っていたと確認できたが学習効率にはほとんど影響を及ぼさなかった。しかし，ラベル付きデータが少ない 1 回目のイテレーションでは効率が改善できていることから，学習初期においては探索的機構を導入し，進行に伴って段階的に機構を除くことでより学習効率の高い Query Strategy になると予想される。

5 おわりに

顕微鏡動画像中の物体の識別をアクティブラーニングにより学習しようとした場合，物体の判別の難しさから追加するデータは予測する密度分布の分離に寄与する決定境界近傍のデータと密度分布の拡大に寄与するデータを加える必要があり，ファジィ測度を使用してモデルの精度向上に寄与する 2 種類のデータを明示的に取り出す Query Strategy を提案した。

MNIST の手書き文字認識タスクにおいて既存の Uncertainty Sampling と比較して効率的に精度を改善できることを示した。反対に既存の確率測度による Uncertainty Sampling では画像データの特徴空間の次元数が多い場合に学習していない領域に存在するデータが多いために拡大に寄与するデータを優先的に Query とし，決定境界近傍に存在するような密度分布の分離に寄与するデータが提示されない課題が明らかになった。

今回の発表でアクティブラーニングを用いてモデルを少ないラベル付きデータで効率的に学習するとき，密度分布の分離と拡大に寄与するデータが必要と確認できたがどちらのデータをより重視しているのかは課題として残った。モデルが必要とするデータの傾向を定量的に表現し適応的に追加するデータの種類を決定するとさらに効率的な学習が可能と考えている。

また，学習器の推論結果に依存する Query Strategy ではラベル付きデータが少ないときに精度向上に有効なデータを Query とすることが難しい課題に対して不確実度スコアから Query の候補を限定してランダムに取り出す工夫を導入したが，精度向上にはほとんど影響を及ぼさないと実験的に示された。

参考文献

1) Burr Settles：Active Learning Literature Survey, Computer Sciences Technical Report 1648 (2009)
2) Davi Pereira-Santos, Ricardo Bastos Cavalcante Prudêncio, André C.P.L.F. de Carvalho：Empirical investigation of active learning strategies, Neurocomputing, 326-327 号 15/27 (2019)