平成 29 年度 特別研究報告書要旨

機械学習における特徴量類似性と認識精度に関する研究

T140496　藤岡 優也

指導教員　三好 力 教授

1.　はじめに

機械学習における教師あり学習では人手によるラベル付きデータを多数学習に用いるほど識別率が高くなることが知られている.しかし,ラベル付きデータは一般的に高コストであり. このコストを削減し,識別機の性能を向上させることは機械学習において重大な課題の一つである.そこで本研究では特徴ベクトル間の類似性に着目し,少数のラベル付きデータから多数のラベルなしデータのクラスを特徴ベクトル間距離によって決定して訓練データとして用いる手法を検討し,①合成データの平均ベクトルからの距離の閾値が近いほど識別率が向上するのか②訓練データに加える合成データの数が多いほど識別率が向上するのかを検証する実験を行った.

2.　提案手法

本研究ではラベルなしデータのクラスを決定する手段として特徴ベクトル間の類似性を用いる.機械学習では主に特徴ベクトル間距離を用いるアルゴリズムが多く,本研究で用いる SVM でも距離が用いられていることから,類似度を特徴ベクトル間のユークリッド距離によって決定する. まず訓練用データとして少数のラベル付きデータを用意する.そのラベル付きデータからランダムで少数の特徴ベクトルを選び取り,その平均ベクトルからのユークリッド距離を全てのデータについて測定し,定めた距離内のもの全てを少数のラベル付きデータと同じラベルを付けて訓練データとする.使用するデータセットは野鳥の鳴き声データで,ミヤマオウム,キジカッコウ,ニュージーランドアオバズクの 3 種類それぞれ 60 セグメントのデータセットを使用した.なお,ミヤマオウムについては 3 種類以上の鳴き声,キジカッコウについては3種類程度の鳴き声,ニュージーランドアオバズクについては 1 種類の鳴き声のデータで構成されている.これらの鳴き声データを MFCC による特徴抽出を行い,910 次元ベクトルを作成し,特徴ベクトルとした.また,機械学習と認識精度の測定には線形 SVM を用いる.

3.　実験 1

提案手法によって決定した訓練データ全てを用いて 1 章①を検証する実験を行った.テストデータにはデータセットからランダムで 20 個用いて,距離には 5 段階の基準を設け,基準値以下であれば正例,以上であれば負例のラベルをつけて訓練データとして用いる. ,識別率の測定は各距離ごとに行い,30 回繰り返した場合の識別率を測定値とする.

4.　実験 2

1 章②を検討するため実験 1 の各距離に対して,正例データと負例データを比べて少ない方を 3 で割った数を測定回数とし,訓練データの数と認識精度の推移を訓練データ 6 個おきに正例データミヤマオウムとニュージーランドアオバズクの場合について測定した.テストデータに対する識別率を求め,30 回繰り返した場合の平均の値を測定値とする.

5.　実験結果と考察

実験 1 で得られた正例データニュージーランドアオバズクの場合の結果を図 1 に示す.横軸は距離,縦軸は識別率を表している.図 2,図 3 に実験 2 の結果を示す.図 2 は正例データニュージーランドアオバズク,図 3 は正例データミヤマオウムである,横軸は訓練データ数,縦軸は識別率を表している.

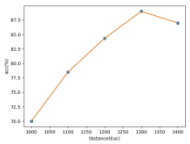
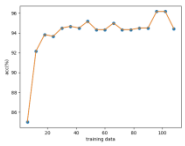


図 1:距離尺度と識別率



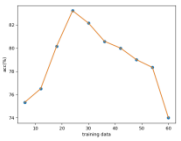


図 2:訓練データ数と識別率　図 3:訓練データ数と識別率

実験 1 の結果から,距離尺度と識別率の関係について,識別率が最高となる最適な距離が存在することが確認された. これは,取る距離が狭すぎる場合に多くの正例データを負例として学習させることになるからであると考えられる. 実験 2 の結果から,複数の鳴き方をする鳥の場合にはグラフは凸となったが,一種類の鳥の場合グラフはある一点から横並びとなった.これも鳴き声の種類の違いによるものであると考えられ,提案手法による訓練データとその数は多いほど良いということが示唆された.