F-038

機械学習における特徴量類似性と認識精度に関する検討
Study on feature similarity and recognition accuracy in machine learning

藤岡 優也‡　　三好 力‡
Yuya Fujioka　Tsutomu Miyoshi

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[1]。しかし、ラベル付きデータは専門家の知識や人の手間などがかかり、高コストである。このコストを削減し、識別器の性能を向上させることは機会学習において重大な課題の一つである。ラベル付きデータが高コストである一方で、ラベルなしデータの場合は低コストで大量に獲得できる場合が多い。たとえば鳥の鳴き声の音声データのであれば、森の中に録音機を設置するだけで容易に獲得できる。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する。この方法に対して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③識別率が向上する距離尺度はあるのか④識別率が向上する特徴抽出法はあるのか等の検討が求められる。③に対しては特徴量間の距離尺度についてバタチャリヤ距離の平方根[2]などの多くの距離が提案されているが、本研究では最も広く使われており、直感的に理解がしやすいユークリッド距離を距離尺度として用いる。④については広く使われており、事前実験で性能が高かったメル周波数ケプストラム係数(Mel-Frequency Cepstrum Coefficients : 以下 MFCC とする)を用いている[3][4]。

本研究では類似性を測る尺度とこの方法によって決定した訓練データの数の関係性をニュージーランドに生息する野鳥の鳴き後のデータを例に、①を調べるため、特徴量の類似性として最適な距離が存在するのかどうかを検討する実験と、②に対応する訓練データの数と識別率の推移の関係を検討するための実験を行った。

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、本研究で扱う SVM においても距離が用いられていることから、類似性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サポートベクターマシン(SVM)[5]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の平均値を測定値とする。

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。

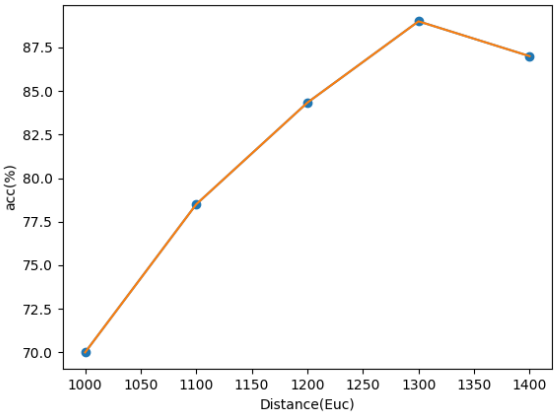

図 1:距離尺度と識別率

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

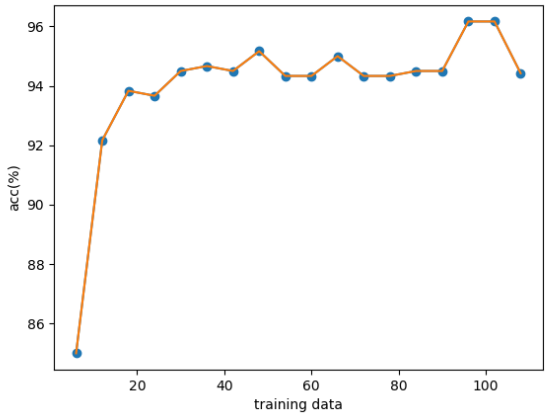

図 2:訓練データ数と識別率(1)

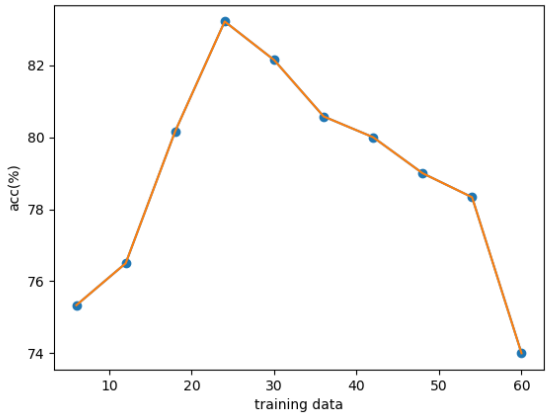

図 3:訓練データ数と識別率(2)

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どうかを検討するために行った実験 1 の結果から、距離が近いほど識別率が向上するという予想に反して、距離が近いものに正例のラベルをつけると認識精度が上がるわけではなく、認識精度が最高となる最適な距離が存在することがわかった。これは、距離が近すぎる場合には実際のラベルが正例の場合でも負例であるとして学習させており、距離が遠すぎる場合であっても逆のことが言えるからであると考えられる。本研究では少数のラベル付きデータの平均ベクトルからの距離によって訓練データを決定している。今回の実験ではミヤマオウム、キジカッコウは複数の種類の鳴き方のデータが混ざり合っており、データが広く分布していると考えられることから、平均ではなく分散や標準偏差などを加味した方法も検討することが今後の課題として挙げられる。

実験 2 は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ために行った。図 3 を見ると凸型のグラフに、図 2 を見るとある一点から横並びのグラフになっていることがわかる。この実験では加える合成データの数が多いほど識別率が良くなるという結果が推測できるが、この推測に反してミヤマオウム、ニュージーランドアオバズクの両方について距離の大小と訓練データ数による認識精度の向上を確認することはできない。これは、ミヤマオウムの場合鳴き声データは 3 種類以上の鳴き声データで構成されており、距離の大小のみによってラベル付けを行うことは非常に困難であるためだと考えられる。一方でニュージーランドアオバズクの場合は 1 種類のみの鳴き声で構成されており、こちらの場合はある一点を境にして訓練データ数を増やしていってもほぼ認識精度はほぼ横ばいになっていることがわかる。複数の鳴き方をする鳥に関してはその鳴き方ごとにラベル付けを行うことが望ましいことが考えられることから、提案手法による訓練データの決定によって訓練データ数を無数に増やしていっても識別率の向上は見込めないことが推測できるが、識別率が下がることはないため、訓練データの数は多数あればよいということが示唆された。

6. おわりに

本研究では、教師あり学習において、学習に多数の訓練データを用いるほど識別率が高くなることが知られていることから、少数のラベル付きデータの平均ベクトルからの距離が近いものをラベル付きデータと類似したデータとみなして訓練データとして用いる手法を提案し、実験を行った。このとき距離は近ければ良いというわけではなく、識別率が最高となる最適な距離が存在することが確認できた。また、訓練データの数を変動させて行った実験から、ミヤマオウムなどの複数の種類の鳴き声が混ざっている鳴き方をする鳥に関しては、鳴き声ごとにラベルをつけることが望ましいため、有用な結果を得ることができなかったが、鳴き方の種類が一種類のみのニュージーランドアオバズクの場合はこの手法によって訓練データ数を増やしていくことによる認識精度の向上は見込めなかったが、認識精度が下がるということは起こらなかったため、訓練データの数は多いほど良いという結論に至る。

今後の研究では鳴き方の種類が増えると鳴き声が広く分布していることが考えられることから平均ベクトルではなく分散や標準偏差を加味した距離の決定や、鳴き声の種類ごとにラベルをつけて実験を行う必要があると考えられる。

参考文献

[1] 栗津妙華, 福尾真実, 高田雅美, 城和貴” 多フォント漢字認識手法における各カテゴリと必要教師データ数の分析”, 研究報告数理モデル化と問題解決, 2014-MPS-97,pp1-6(2014)
[2] 峯松 信明,志甫 淳,村上 隆夫,丸山 和孝,広瀬 啓吉,“音声の構造的表像とその距離尺度”,電子情報通信学会技術研究報告.SP,音声 105(98), 9-12(2005)
[3] 宇津呂 武仁, 渡邉 友裕, 中川 聖一, 小玉 康広,西崎 博光,”機械学習を用いた複数の大語彙音声認識モデルの出力の混合”,情報処理学会研究報告音声言語情報処理,2002-SLP-045,pp95-100(2002)
[4] 栗原 一貴, 佐々木 洋子, 緒方 淳, 後藤 真孝” 音声区間自動検出技術を用いた変速再生方式による映像の高速鑑賞システムの検討”, 研究報告ヒューマンコンピュータインタラクション, 2002-SLP-045,pp95-100(2003)
[5] Corinna Cortes,Vladimir Vapnik,Support-vector networks, Machine Learning,Vol. 20,pp273-297(1995).

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.