Vol.2009-MPS-75 No.17
2009/9/10

境界領域の容易化を目的としたデータ前処理手法の提案

木村　幸代[†1]　　渡　邉　真　也[†2]

本論文では，識別問題に対する新たなデータ前処理アプローチとして，学習用データの分布に基づくクラスタリングを利用する方法を提案しその有効性について検証を行う．本アプローチでは，クラスごとのクラスタリングに基づくクラス分割（サブクラス化）を行い．特徴空間でのまとまりの強いクラスを生成する．クラス数は増加するものの，クラスごとの分割は容易になるため，結果として対象問題の識別の容易化につながると期待している．本論文では数値実験として，2 種類のテスト問題に対して提案手法を適用しその有効性について検証を行った．

A proposal of a data preprocessing approach to simplify a shape of boundary region

Yukiyo Kimura[†1] and Shinya Watanabe[†2]

This paper proposes a new data preprocessing approach, which utilize clustering method to increase the density of data class. This approach performs clustering method based on the distribution of each class and divide into subclass. By the use of this approach, the total number of class is increased, but the difficulty of classification can be reduced since it becomes easy to classify for each class. In this paper, we examine the characteristics and effectiveness of the proposed approach through two different test problems.

†1 室蘭工業大学大学院
Graduate School of Muroran Institute of Technology
†2 室蘭工業大学
Department of Information and Electronic Engineering, Muroran Institute of Technology
∗1 現在，情報工学科
Presently with Department of Computer Schience & Systems Engineering

1. はじめに

ニューラルネットや SVM(Support Vector Machine) といった識別手法そのものではなく，データの前処理を施すことにより識別精度を向上させようとする研究が近年，注目を集めている[1)–6)]．

これらの多くは，画像といった実問題を対象にした場合に生じるノイズ除去[2)]，クラス間でのデータの偏りの解消[3)]，データの属性次元を縮約[4)] といったデータの信頼性向上，データ空間の領域削減を目的としている．一方，その数は非常に限られているもののデータ集合の分布特性を付加情報として識別精度向上のために積極的に利用する研究[5)] も報告されている．しかしながら，識別手法やその分析に比べ前処理に関する関心は高くないのが現状である[6)]．

そこで本研究では，識別精度向上のための新たな前処理アプローチとして，学習データのクラス分布形状に基づくサブクラス化を提案する．提案手法では，まずクラスごとにクラスタリングを行い，その結果に基づくクラスの分割（サブクラス化）を行う．提案するサブクラス化により対象となる識別クラスは多数クラス化するものの，各クラスごとの分布密度は向上しクラス分類自体は容易化することができるため，結果として問題の単純化を期待することができる．

本論文では，UCI レポジトリ[7)] から入手した 2 つのデータセットに対して提案手法を適用し，その有用性について検証を行った．

2. データ前処理に関する先行研究

対象データの前処理による識別性能向上を図る研究は，識別手法の改良に比べ多くないものの，実計測データに対するノイズ除去，特徴抽出はその必要性から古くから研究が行われており，特に画像に関するノイズ除去については Wavelet 処理を利用したもの等多数存在する[2)]．また，データ集合サイズの縮約を目的とした事例選択，特徴選択に関する研究も目的に応じて様々な手法が提案されている[1)]，

しかしながら，これらの手法はデータの識別境界面の複雑性については全く考慮していないため，本質的に境界が複雑である問題の容易化を実現することはできない．

そこで，本論文ではデータ境界面の複雑性をより直接的なアプローチで解消することを目的とした，新たなデータ前処理アプローチの提案を行う．

© 2009 Information Processing Society of Japan

Vol.2009-MPS-75 No.17
2009/9/10

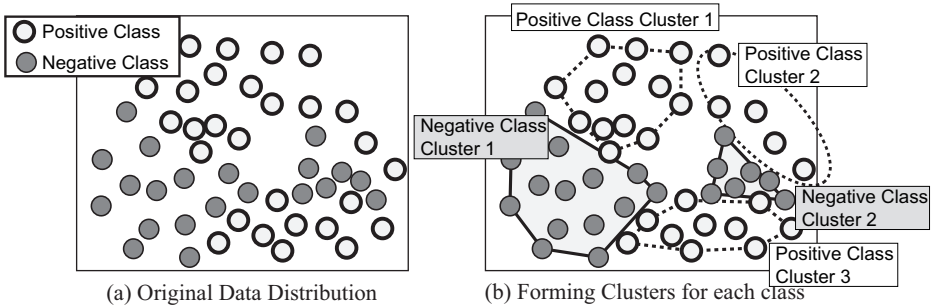

図 1 The procedure of the splitting approach.

3. クラスタリングに基づくデータ前処理

本論文では，クラス間の複雑なデータ分布を解消するためにクラスタリングを用いたデータ前処理アプローチを提案する．提案手法はクラス分布形状の複雑化要因を多クラス化により取り除くというアイデアに基づいている．

3.1 提案手法の概要

提案手法の手順を概念的に示したものを**図 1** に示す．図 1 では 2 次元 2 クラス問題の例を対象にしており，色つき（濃い灰色）クラスと白丸クラスそれぞれに対してクラスタリングを行いサブクラス化している様子を示している[⋆1]．

図 1 から分かるように提案手法では各クラスごとにクラスタリングを行い，その結果に基づくサブクラス化を実現している．クラスタリング手法としては Affinity Propagation[8] を使用し，極力クラスタ数が増加しすぎないようデータ間の近接度合い設定を行った．提案手法の操作により，クラスタ数は増加するものの識別を複雑化している要因が多クラス化により間接的に取り除かれることとなり，結果として識別性能の向上を期待することができる．

3.2 サブクラス化データに対する結果からの最終識別結果の決定

本研究で組み入れるクラス分割を用いた場合には，各分類器からの結果はサブクラス化したデータに対する結果となり，元のクラスに対する識別結果となっていない．そのため，サブクラス化したデータに対する集計結果から，何らかの方法を用いて最終的な元のクラスに対する識別結果を示す必要がある．

⋆1 ここでは白丸クラスのデータのうち，黒丸の凸領域を浸食しているものを違反データと呼ぶ．

表 1 Databases.

Name	Heart-disease	Liver-disorders
# classes	2	2
# Attributes	13	6
# Instances	297	326

本論文では，多クラス問題に対応した SVM として OAO(One Against One) を用いている[9]．OAO では 2 クラスに対する分類を全ての組み合わせに対して行うため k(k-1)/2 個の SVM を実装し，各 SVM の多数決により最終的な結果が決定する．

サブクラス化データに対する結果からの元問題に対する結果を決定する方法には様々な方法が考えられるが，本論文では各 SVM の多数決により最多得票を得たサブクラスが属する元クラスを最終的な結果として出力するものとした．

4. 数 値 実 験

提案するクラス分割の適用により，対象となる識別クラスは多数クラス化するため必要となる計算量は増加する．また，過剰な多数クラス化は単に計算量の増加だけでなく，クラス数増加による性能悪化をもたらす危険性がある．

本論文では，提案手法の有効性および上記の問題点についての検証を行うため UCI レポジトリ[7] から入手した 2 クラス問題である Heart-disease（心臓疾患データ），Liver-disorders（肝臓病患者のデータ）を用いた．両問題は，2 クラス問題としての難易度が比較的高い問題として取り上げた．

また，手法の効果を視覚的に確認するためにカーネル次元削減法（Kernel dimension reduction：KDR）[10] を用いたデータ分布の視覚化を行った．KDR は，線形判別法と同様にクラス分類が既知である高次元データから，クラスの情報をできるだけ保持するような低次元空間を求めることを目的としており，特徴空間 X に含まれる応答変数（クラス）Y の情報をすべて保持するような説明変数空間の低次元空間（有効部分空間）S を求める．その特徴は，カーネル法を利用しているため非線形の問題にも適応できる点，問題に対しモデルや制約をなるべく置かずにセミパラメトリックなアプローチをとる点である．

4.1 対象テスト問題

Heart-disease（心臓疾患データ），Liver-disorders（肝臓病患者のデータ）の特徴を**表 1** に，KDR により 2 次元化したデータ分布を**図 2** に示す．

図 2 におけるクラス間の重なり具合から分かるように，対象問題のクラス分離は非常に

© 2009 Information Processing Society of Japan

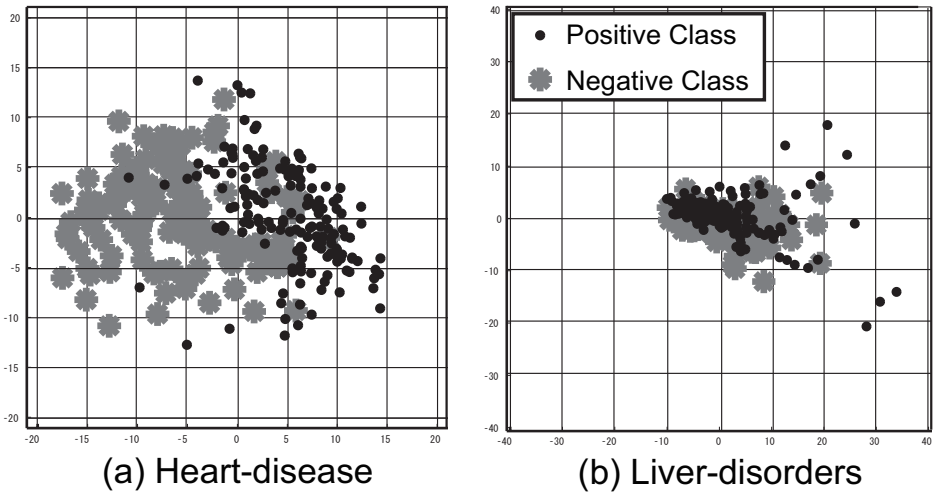

図 2 The two-dimensional distribution map of two datasets.

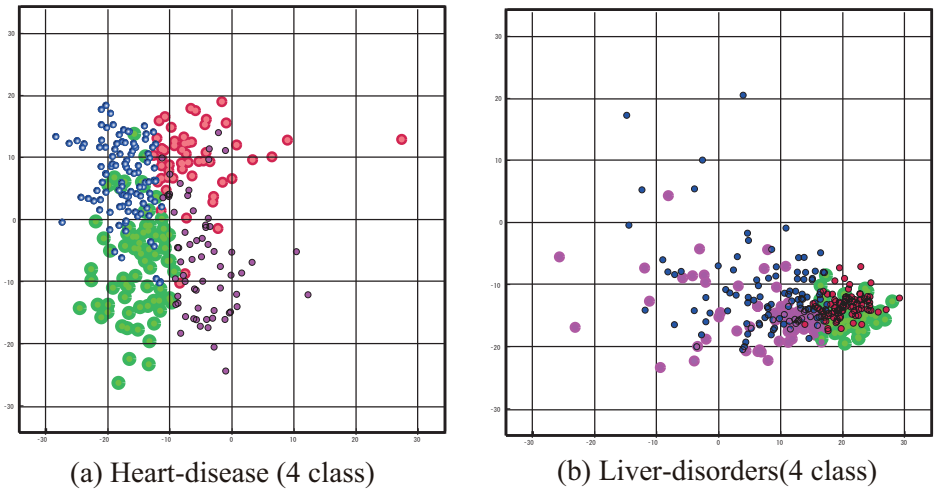

図 3 The two-dimensional distribution map of sub-class results.

困難であることが分かる．

4.2 評価指標

本研究では，得られた解に対して様々な角度から評価を行うため，特徴の異なる以下の 4 つの評価指標を用いた．

- 正解率（Accuracy rate）[11)]
- 感度（Sensitivity）[11)]
- 特異度（Specificity）[11)]
- F 値 (F-measure)[12)]
- サポートベクターの数 #SV(Number of Support Vetor)

正解率はポジティブ，ネガティブ全てのデータに対する全体の正解率を表している．感度はポジティブを発見する能力であり，特異度は逆に非ポジティブをポジティブだと誤らない能力を表している．また，F 値は感度と特異度の両方の観点から総合的な評価値を示している．サポートベクターの数は，SVM において分離平面を形成するのに用いられるベクトル点の数であり，一般的に少ないほど汎用性が高いとされている．

4.3 実験結果

上述の 2 種類の問題に対して適用実験を行った．SVM パッケージとしては，広くこの分野で利用されている LIBSVM を用いた[13)]．また，カーネル関数としては RBF カーネル（Radial Basis Function Kernel）を使用し，カーネルパラメータについては各クラス数の場合に対して事前実験より得られた最良値を設定した．

表 2 Classification performance of experiments.

Dataset	*#class*	Accuracy	Sensitivity	Specificity	F-measure	*#SV*
Heart Disease	2	0.835	0.881	0.781	0.852	122
	4	0.855	0.810	0.894	0.838	176
	8	0.825	0.766	0.875	0.802	213
Liver-disorders	2	0.745	0.850	0.600	0.794	191
	4	0.704	0.790	0.586	0.756	206
	10	0.701	0.593	0.780	0.625	205

Heart Disease に対しては 4，8 クラスにサブクラス化した場合，Liver-disorders に対しては 4，10 クラスにサブクラス化した場合の結果について**表 2** に示す．また，Heart Disease および Liver-disorders における 4 クラスの場合の KDR での 2 次元縮約の結果を**図 3** に示す．

表 2 から分かるように，項目の一部において提案するサブクラス化の結果が勝っている

© 2009 Information Processing Society of Japan

ものの全般的にはサブクラス化により性能が僅かに劣化しているのが分かる．一方，元々の2クラスデータに対する2次元縮約結果（図2）に比べ，サブクラス化に対する結果はクラスごとの分布がより明確に分離されていることが分かる．このことは，識別器の工夫により従来の2クラスの場合よりも優れた識別が可能であることを示唆していると思われる．

5. 結　　論

本論文では，新たなデータ前処理アプローチとしてクラスタリングを用いたクラス分割を提案しその有効性について検証を行った．提案する手法では，サブクラス化によりクラスごとの分布の密集度を向上させることによる分離の容易化を試みたものの数値実験ではその有用性を確認することはできなかった．しかしながら，2次元縮約の結果では通常の場合よりもクラスごとの分布がより明確な結果が得られており，識別器および手法の改良により通常の場合を上回る性能向上の実現が期待できる．

今後は，本論文で提案するクラスタリングをベースに識別平面の容易化を強く意識した手法を考案，検証していく予定である．

参 考 文 献

1) Dorian Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann Pub, 1999.
2) QiLi, Tao Li, Shenghuo Zhu, and C.Kambhamettu. Improving medical/biological data classification performance by wavelet preprocessing. pp. 657–660, 2002.
3) Jerzy Stefanowski and Szymon Wilk. Selective pre-processing of imbalanced data for improving classification performance. In *DaWaK '08: Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery*, pp. 283–292, 2008.
4) Xiuju Fu and Lipo Wang. Data dimensionality reduction with application to improving classification performance and explaining concepts of data sets. *Int. J. Bus. Intell. Data Min.*, Vol.1, No.1, pp. 65–87, 2005.
5) Paraskevas Orfanidis and DavidJ. Russomanno. Preprocessing enhancements to improve data mining algorithms. *IJBIDM*, Vol.3, No.2, pp. 196–211, 2008.
6) SvenF. Crone, Stefan Lessmann, and Robert Stahlbock. The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research*, Vol. 173, No.3, pp. 781–800, 2006.
7) A.Asuncion and D.J. Newman. UCI machine learning repository, 2007.
8) BrendanJ. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, Vol. 315, pp. 972–976, 2007.
9) Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, Vol.13, No.2, pp. 415–425, 2002.
10) Kenji Fukumizu, FrancisR. Bach, and MichaelI. Jordan. Kernel dimension reduction in regression. In *Technical Report 715, Department of Statistics, University of California, Berkeley*, 2006.
11) M.Doumpos, C.Zopounidis, and V.Golfinopoulou. Additive support vector machines for pattern classification. *IEEE Transactions on SMC, Part B*, Vol.37, No.3, pp. 540–550, 2007.
12) Tom Fawcett. ROC graphs: notes and practical considerations for data mining researchers, 2003.
13) Chin-Wei Hsu, Chin-Changa, and Chin-Jen Lin. LIBSVM:`http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

© 2009 Information Processing Society of Japan