1D3-2

2クラス分類の為の遺伝的プログラミングを用いた特徴量変換手法の提案

A method of multiple feature construction for two symbolic classification problems using Genetic Programming

白石 駿英[*1] Toshihide SHIRAISHI　吉田 倫也[*1] Tomoya YOSHIDA　山本 詩子[*2] Utako YAMAMOTO　廣安 知之[*2] Tomoyuki HIROYASU

[*1]同志社大学大学院生命医科学研究科 Graduate School of Life and Medical Sciences, Doshisha University

[*2]同志社大学生命医科学部 Faculty of Life and Medical Sciences, Doshisha University

In this paper, a feature transformation method for two-class classification using genetic programming (GP) is proposed. GP derives a transformation formula to improve the classification accuracy of SVM. In this paper, we propose a weight function to evaluate converted feature space and the proposed function is used as an evaluate function of GP. In the proposed function, the ideal two-class distribution of items is assumed and the distance between the actual and ideal distributions is calculated. The weight is imposed to these distances. To examine the effectiveness of the proposed function, numerical experiment is performed. As the result, the classification accuracy of the proposed method is better than that of the previous method.

1. はじめに

近年，機械学習による識別問題が盛んに研究されてきた．機械学習における識別問題に関して，精度の高い識別器を構築することにより，識別精度を高めるといった報告が多数為されている [G.E.Hinton 06][V. N. Vapnik 98][F. Otero 03]．具体的な識別器としてニューラルネットワーク (Neural Network:NN)，決定木 (Decision Tree) や SVM(Support Vecter Machine) といった精度の高い様々な手法が提案されている．応用分野としては，金属製品の異常検知問題や癌の腫瘍の良性・悪性の識別等があり，幅広い分野で識別問題の解決手法として用いられている．

しかし，より良い識別精度を達成する為には，識別器の判別の能力だけでなく，識別器が判別する特徴量空間を考慮する必要がある．データが複雑な特徴量分布を持つ場合，識別器の識別能力は下がってしまうことが考えられる．また，NN や SVM ではパラメータを変更することにより，最適な識別線を探索することが可能である．SVM ではグリッドサーチにより，自動で最適な識別線を求めるパラメータを探索できる [F. Friendrichs 05]．しかし，その探索は関数の種類に制約があり，柔軟なものであるとは言えない．

そこで，本稿では識別の精度を高める特徴量空間を構築するアプローチとして遺伝的プログラミング (Genetic Programing:GP[Koza 92]) を用いた特徴量変換方法である GPMFC(Genetic Programing Multiple Feature construction)[Zhang 12] における特徴量変換手法を検討する．具体的には，与えられた特徴量空間を決定木やルール学習等の識別器の識別精度を高める目的で新しい空間に演算により変換する変換式を構築する．その際，GP の遺伝的操作を分類に最適な特徴量空間を得る為に行う．今回，先行研究で用いられている評価関数よりも精度の高い特徴量空間が構築可能である評価関数を提案した．評価実験により，既存手法と提案手法によって構築した特徴量空間における識別率を比較した．

連絡先: 白石　駿英，同志社大学大学院生命医科学研究科，京都府　京田辺市多々羅都谷 1-3，0774-65-6130，sshiraishi@mis.doshisha.ac.jp

2. GPMFC

GPMFC は最適化手法である遺伝的プログラミング (GP) を用いて分類最適な特徴量を得る為の変換式を構築する手法である．また，ユーザは特徴量に関する事前知識に関係なくこのアルゴリズムにより分類最適な特徴量を得ることができる．GPMFC では GP の評価，選択，交叉，突然変異といった遺伝的操作を繰り返し行うことにより特徴量変換式の最適化を図る．以下に GPMFC のアルゴリズムを示す．

Step.1 初期化 (Initialization)
初期個体群として，ランダムに集団数だけの個体（木構造の変換式）を生成する．

Step.2 評価 (Evaluation)
各個体の適合度を評価関数によって計算する．各個体による変換式によって特徴量を変換し，変換した特徴量において分布の重なるデータの範囲を求める．この区間にある特徴量の数を求め，これを評価値とする．遺伝的操作 (選択) の際に評価値が低い木構造程が高い評価となる．図.1 にデータの変換と評価値の概要図を示す．

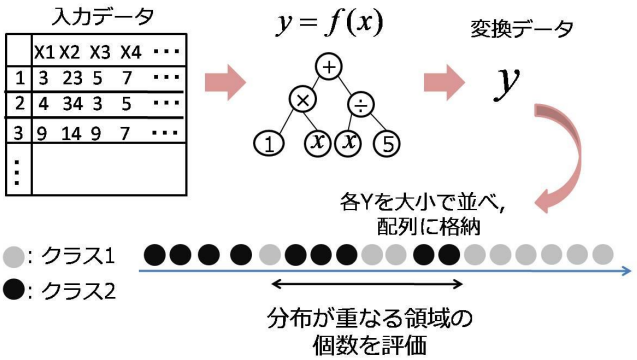

図 1: 変換式の評価

また，図.2 に特徴量変換後のヒストグラムのイメージ図を示す．評価関数はこの図のように視覚的に見ると，2 クラス分布が両側に分かれているものを評価する．

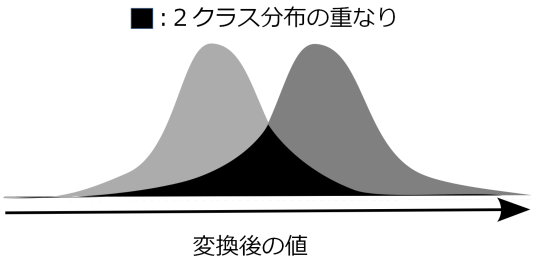

図 2: 1 次元特徴量ヒストグラム

Step.3 選択 (Selection)
集団の中から適合度を基準として， 次世代に残す個体を集団数だけ選択する．

Step.4 交叉 (Crossover)
集団の中からランダムに選択した 2 個体を対象とし，それぞれランダムに交叉点を選んだ後，枝を差し替え交配する．

Step.5 突然変異 (Mutation)
対象となる個体からランダムに突然変異点を選択し，突然変異点以降の木をランダムに作成した突然変異木と入れ替える．

Step.6 終了条件 (Terminal Criterion)
予め決めておいた終了条件に到達するまで Step 2〜Step 5 を繰り返す．主な終了条件として探索世代数や目的の個体に達したか否かなどが挙げられる．GPMFC では評価値が 0 となり，データの属性が完全に分離することを目的として最適化を行う．

3. 属性の分布を考慮した評価関数の検討

3.1 先行研究の評価関数における問題点

前章の先行研究における GPMFC の評価関数には問題点が存在する．それは，1 次元に変換した特徴量平面において，2 クラス分布が重なる領域での頻度のみを考慮することである．図.2 のような 2 クラスの重なりを表す領域の中の分布の善し悪しを既存手法では評価することができない，その為，識別を向上させる平面を得ることができるとは限らず，既存手法は最適な評価関数とは言えない．ここで，図.3 に 2 クラス分類精度の低くなる平面と精度の高くなる平面について示す．右側の図のような特徴量分布を持つ平面がより良い評価を受けるように評価関数を考慮する必要がある．

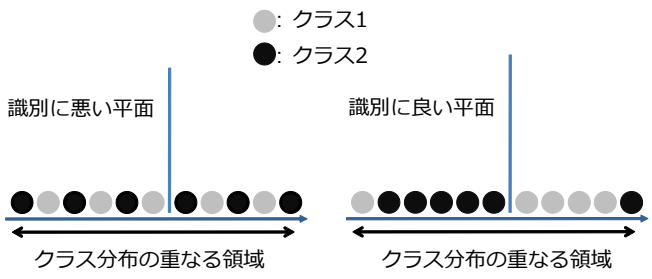

図 3: 1 次元 2 クラス分布の比較

このように 2 クラス分布の重なる領域において，識別に不利な平面は 2 クラスデータが交互に配置される場合である．それに対して，識別に有利な平面では 2 クラスデータは分離中心を境としてある程度局在するように分布することが考えられる．そこで，本稿では理想的な 2 クラス分布を仮定し，実際に得られる特徴量分布との誤差をペナルティーとして評価する評価関数を提案する．次章でその評価手法について述べる．

3.2 提案手法

GPMFC の特徴量変換における評価方法として，事前に理想的な 2 クラス分布に対する知識を与えるような方法を提案する．この手法ではまず，理想的な 1 次元データに変換される特徴量の分布をその値の大きさの順位の大小によって決定する．次に，1 次元特徴量変換データをソートして，変換データと理想的なデータの属性を比較する．属性に誤りが生じた場合はそのデータの位置を考慮したペナルティー (重み) を付加する．重みは 2 つの属性の分離中心から離れるほど重く設定することにより，2 つの理想的な属性の分離状態により近くすることが可能となる．図.4 に評価関数の概要図を示す．この図では，両属性に 2 次の重み関数を設定した場合について示している．次章の評価実験においても同様に，2 次の重み関数を用いている．

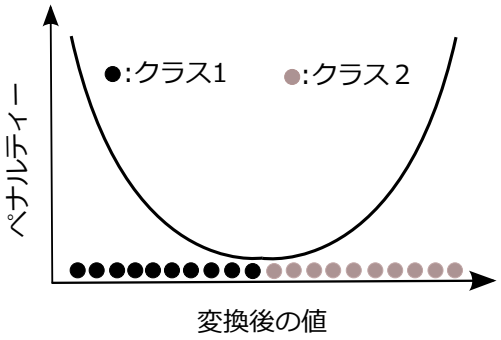

図 4: 重み関数概要図

4. 評価実験

提案手法と先行研究の手法を比較する評価実験を行い，提案手法の評価関数の有効性を評価する．また，変換前のデータよりも識別結果が向上するかも併せて評価を行う．

4.1 実験方法

実験では学習データとして Breast Cancer Wisconsin の乳癌データ [Merz C.J. 98](データ数 420) を用いる．また，識別の検証には 5fold-crossvalidation を用い，学習データを 336 サンプルと未学習データを 84 サンプルとする．変換したデータを SVM を用いて識別し，識別結果を上記の 2 つの手法について比較する．また，得られた特徴量平面がどのような分布を持つかも併せて比較する．

4.2 実験パラメーター

表 1 に今回の実験で用いる GP のパラメータを示す．このパラメータは提案手法と先行研究の手法において共通である．

表 1: GP のパラメータ

パラメータ	値
個体数	1000
世代数	100
交叉率	0.6
突然変異率	0.3
特徴量数	3
関数ノードの種類	+，−，×，÷
交叉方法	一点交叉
選択方法	トーナメント選択

4.3 実験結果

図.7 に提案手法と先行研究の手法において，識別率の比較を行った結果を示す．なお，この実験では提案手法と既存手法の比較をしやすくする為に識別率の低いデータを用いた．その為，前述の Breast Cancer Wisconsin の乳癌データから特に識別率の低い特徴量を 3 種類抽出した．このデータの SVM による識別率は RBF カーネルを用い 70.95%である．まず，実際にデータを変換した際の一次元の特徴量分布を，図.5 に既存手法，図.6 に提案手法について示す．既存手法においてはデータの分布が重なる領域において，2 つのクラス分布が入り交じっているが，提案手法においては，2 クラスの領域がある程度分かれた分布を得た．

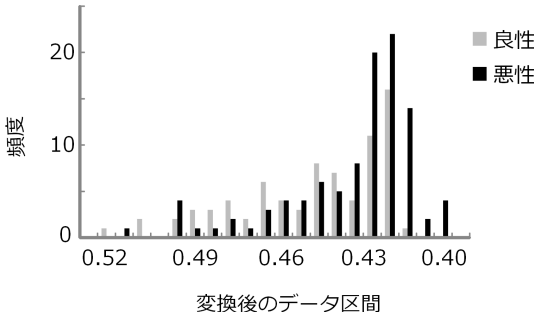

図 5: 特徴量ヒストグラム (既存手法)

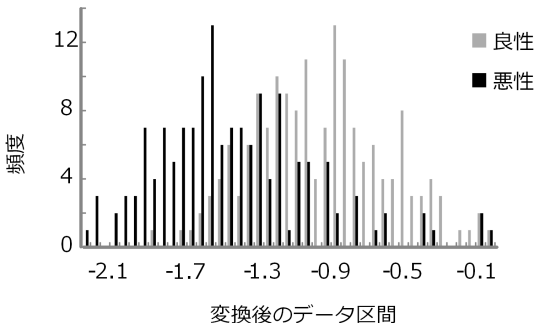

図 6: 特徴量ヒストグラム (提案手法)

次に，識別率について示す．識別率は GPMFC10 試行の平均値を算出して示した．提案手法による平均の識別率は 73.34%となり，従来手法の平均識別率 65.58%を上回り，かつ変換前よりも 2.39%識別率が向上した．この結果により提案手法の有効性が示唆された．

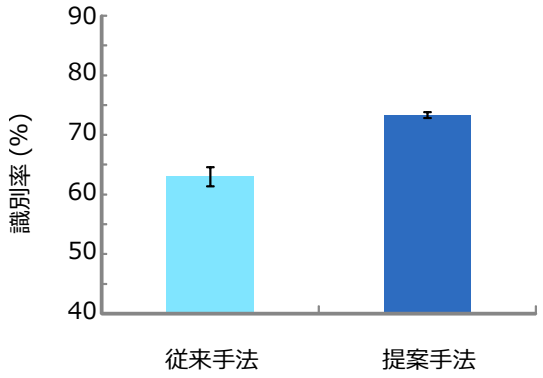

図 7: 識別率の比較 (10 試行平均)

5. まとめ

今回の実験から，GPMFC の評価関数には単純な分布を考慮する手法よりも，事前知識を含む評価関数が適していることが示された．しかし，学習サンプル数が同数必要であるなどの課題を解決する必要がある．また，多クラスの分類問題について，GPMFC を用いた特徴量変換手法を検討する必要がある．

一方で，この手法がどの種類の識別器と相性が良いのかということも検討項目として挙げられる．今回は 2 クラス分類において精度の高い SVM を選んだが，今後は NN や決定木において，この手法を適応し，精度の比較検討を行っていく．

参考文献

[G.E.Hinton 06] G.E.Hinton and R.R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. Science, Vol. 313, No. 5786, pp. 504-507, 2006.

[V. N. Vapnik 98] V. N. Vapnik. Statistical Learning Theory. Wiley, 1998.

[F. Otero 03] F. Otero, M. Silva, A. Freitas, and J. Nievola, Genetic programming for attribute construction in data mining, in Genetic Programming (Lecture Notes in Computer Science, vol. 2610.) Berlin, Germany: Springer, 2003, pp. 101-121.

[F. Friendrichs 05] Frauke Friedrichs and Christian, Evolutionary tuning of multiple SVM parameters,Trends in Neurocomputing: 12th European Symposium on Artificial Neural Networks 2004,Volume 64, March 2005, Pages 107-117.

[Zhang 12] K.Neshatian,Mengjie Zhang,and P.Andreae, A filter approach to multiple feature construction for symbolic learning classifiers using genetic programming : Evolutionary Computation, IEEE Transactions on, Vol. 16, No. 5, pp.645-661,oct. 2012.

[Merz C.J. 98] Merz C.J. Blake and C.L. , Uci repository of ma- chinelearning databases, 1998.

[Koza 92] J.Koza: Genetic programming,on the programming of computers by means of natural selection, MIT Press, 1992