408

欠陥分類のための逐次TFCにおける効率的学習

北海道大学　○松尾　祥和　　新日本製鐵　小林　尊道
北海道大学　高氏　秀則　　北海道大学　金子　俊一

要旨

テスト特徴法（Test Feature Classification：以下TFC）では逐次的に与えられるデータを学習させ識別性能を向上させる学習アルゴリズム（Sucssessive TFC：以下sTFC）が提案されている．本論文では，識別に用いられる投票を基に，査定前のデータの中から学習に必要であるデータを選択することで査定コストを削減しつつ，効率的に学習させる手法として部分修正追加手法を提案する．そして検証実験によりその有効性を証明する．

1. 序論

欠陥の解析にパタン認識を導入する際には欠陥画像を検査員が査定をし，ラベルを付与し，欠陥の種類ごとに整理してパタン認識の教師データとして登録するという機械学習の準備作業が必ず必要となり，この作業は非常に負荷が高いことが知られている．パタン認識の精度向上のためには，非常に多くの査定されたデータ（以下査定データ）が必要になる．従来より査定データから学習に向いているデータを選択する手法は数多く提案されている[1]．また査定データ削減を考慮した手法も提案されている[2]，しかしこの手法は選択したデータを用いて教師なし学習を行うため，それまでの識別器による識別結果の信頼性が高いデータを選択することを提案している．また識別率を向上させる方法の中で，クラス情報を付与されていないデータを用いる手法として，半教師あり学習が提案されている[3]．この手法は，査定データいわゆる教師データと査定されていないデータを同時に加える手法である．

本研究では，逐次学習において学習の初期段階では，教師あり学習をし，その後教師なし学習へ切り替える自律的学習をすることで，査定データ使用の削減による査定コストの削減を最終目的とする．

本論文では，その初期段階である教師あり学習において，前TFCの投票結果を用いて，査定前にデータを選択することにより，学習に必要なデータを選択し，査定コストを削減する手法を提案する．さらに検証実験によりその有効性を証明する．

2. 部分修正追加

本研究では，逐次的に与えられるデータに対して学習を行うsTFCの拡張として，査定コストを削減するための新しい追加戦略を提案する．従来の追加戦略では，全追加では追加データが査定されていないデータであっても学習をすることができる．しかし，それ以外の追加戦略においては追加用データが査定されていることが前提である．査定データを作成することは，非常にコストのかかる作業である．今回提案する手法は，その査定データを作成するコストを削減することを目的として提案する手法である．本手法は，査定する前に学習に必要であるデータを選択し，査定するデータ数を減らすという手法である．これにより，査定前のデータの削減，それにより学習データの削減にもつながり，計算コストの削減も効果が期待できる．

部分修正追加では学習に必要であると言われているクラス間における境界の近くにあるデータを査定用として取り出し，査定し，それを学習に用いる．TFCにおいて，クラス間の境界線は線形的に形成することはできない．そこで，境界付近のデータとして，投票率を用いた選択方法を提案する．たとえばA, Bという2クラス問題のデータが存在するとき一度作成したTFCによって識別をすると，PTFの投票により，『A』クラスのデータかもしくは『B』クラスのデータに分類される．ここで，PTFの投票により分類されたときに，投票がどちらのクラスにも均衡しているデータという新しいクラスを作ると，3つに分類される．この投票でどちらのクラスにも均衡しているデータが，識別しづらいデータ，つまり境界付近のデータとなる．本手法では，このデータを学習に必要なデータとして選択して査定を行い，訓練データとする．

3. 査定前データ選択追加実験

3..1 事前準備実験

前節で査定前のデータ選択には前TFCの投票率を用いて選択をすると述べた．ここで，実際に境界付近のデータは投票率ではどの範囲を決定することが問題となる．そ

Table 1: 無害クラスへ投票した割合（Steel データ）

判定	無害クラスへ投票した割合（%）	疵種区分別頻度	
		有害	無害
有害	0〜0.05	368	0
	0.05〜0.1	202	0
	0.1〜0.15	169	0
	0.15〜0.2	175	0
	0.2〜0.25	155	7
	0.25〜0.3	160	8
	0.3〜0.35	151	13
	0.35〜0.4	135	26
	0.4〜0.45	125	35
	0.45〜0.5	88	51
無害	0.5〜0.55	80	70
	0.55〜0.6	61	96
	0.6〜0.65	39	117
	0.65〜0.7	36	178
	0.7〜0.75	22	226
	0.75〜0.8	15	285
	0.8〜0.85	9	363
	0.85〜0.9	7	504
	0.9〜0.95	2	597
	0.95〜1	0	256

こで事前準備実験として以下の実験を行った．使用するデータは Steel データを用いる．それぞれ使用する Steel データは 5000 個程度のデータを使用する．データをそれぞれ 10 分割ずつし，それぞれ cross-validation 法（交差検定）を用いて全てのデータにおいて投票率を算出し，投票率ごとの分布を示す．

表 1 に Steel データの投票分布を示す．この表から投票率がどちらも 3 割から 7 割の範囲のデータが誤識別が多くなっている．そこで片方のクラスへの投票率が 3 割から 7 割の範囲のデータを査定用データとして選択する．

3..2 全修正追加との比較

次に，全修正追加との比較実験を行う．データ総数は 6000 程度．評価データを 1200 程度とり，残りのデータを 8 分割し，逐次的にデータを追加しながら学習をさせていく．全修正追加では，すべてのデータを査定し，学習させる．部分修正追加では初期学習には与えられたデータをすべて学習に用い，2 回目以降の学習では，直前の TFC によりデータを選択し，追加学習を行った．データをランダムに入れ替え，5 回同じ実験を行ったときの平均値を示す．

図 1 に実験結果を示す．識別率には，全修正追加と部分修正追加に差がほとんどみられなかった．

訓練データの数を見てみると，全修正追加では 4800 個程度，部分修正追加では 1700 個程度のデータしか使っていない．全データのうちの 3 割強の訓練データで全修正追加と同等の識別率が得られた．

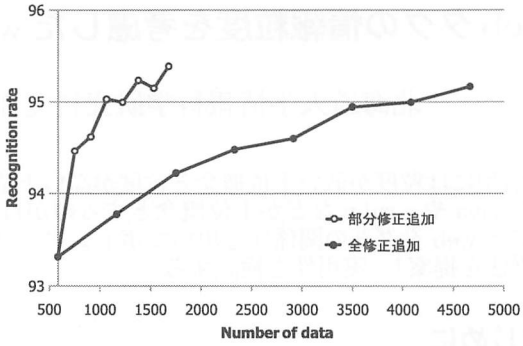

Fig. 1: 従来手法との比較（Steel データ）

4. 結論

査定前のデータに関して，前 TFC により識別を行い，投票率からデータの選択を行い査定データを削減する手法として部分修正追加を提案した．検証実験により，査定前のデータを効率よく削減することが示された．さらに他の手法において境界データが有効であるかを検証していきたい．

参考文献

[1] 原一之，中山謙二：“階層ニューラルネットワークにおける学習データ選択法”，情報処理学会第 51 回（平成 7 年後期），pp 41-42

[2] 岡谷一宏，吉田哲也：“共同学習における分類器の合意度を用いた追加データ選択法の提案”，The 23rd Annual Conference of the Japanese Society for Artificial Intelligence, https://kaigi.org/jsai/webprogram/pdf/168.pdf, 2009

[3] DengCai, Xiaofei He, Jiawei Han:“Semi-supervised Discriminant Analysis”, U.S National Science Foundation NFS IIS-05-13678