2A2-4

分類器学習における分類精度向上のための属性追加方式

Attribute addition method for classification accuracy improvement in classification machine learning

井芹史明[*1] 田中真樹[*2] 末田直道[*1]
Humiaki Iseri Masaki Tanaka Sueda Naomiti

[*1]大分大学大学院工学研究科
Graduate School of Engineering, Oita University

[*2]株式会社富士通九州システムエンジニアリング
FUJITSU KYUSHU SYSTEM ENGINEERING LIMITED

In data mining algorithms such as classification machines, a preprocessing of data is an important task. In this paper, we propose an attribute addition technique consisting of a data preprocessing technique that adds the attributes to contribute to the classification of data with two or more classification machines. We show effectiveness of the technique by comparative experiments that are cases of use and no use the technique.

1. はじめに

膨大なデータから有益な知識を抽出するデータマイニングにおいて，データを分類する技術に，決定木やニューラルネットワークなどのさまざまなアルゴリズムが提案されている．これらのアルゴリズムは分類器と呼ばれ，訓練データを与えられ，それによって分類器を構築していく．このとき訓練データにはノイズや，属性の記述が冗長であったり，属性が不足しているなどの分類器のクラス分類精度低下につながる要因が含まれることがある．これらの問題に対処するためデータの前処理といういタスクが重要になっている．既存のデータ前処理手法として，クラス分類に貢献しない属性を削除する属性選択や，クラス分類に貢献するような属性を生成し元データに追加する属性生成 [寺邊 00] などの手法が提案されている．また分類精度を向上させる技術として、Adaboosting 手法 [Witten00] などの複数の分類器を用いて精度向上を目指す集団学習と呼ばれる手法もある。

本稿では，集団学習や属性生成の考え方を取り入れ、複数の分類器を用いて，訓練データにクラス分類に貢献するような属性を追加するデータ前処理手法である属性追加手法を提案する．また属性追加手法を評価するため UCI の Machine Learning Repository の 10 種類のデータセットを用いて，属性追加手法を用いた場合と，用いなかった場合の分類精度の比較・評価を行う．

2. 属性追加

この節では，本稿で提案する属性追加の概要と処理手順を述べる．

2.1 概要

人間は個々の思考や能力が異なるため，同じ教師から学んだ生徒が同じ問題に対して同じ答えを出すとは限らない．これと同様に，分類器の生成アルゴリズムが異なれば，同じ訓練データから生成した分類器が同じ未知のデータに対して同じクラス分類結果を出すとは限らない．また生徒が一人で解くより，複数の生徒が共同で問題を解く方が正しい答えを導く可能性が高くなるように，複数の分類器が共同で未知のデータのクラス分類を行う方が正しいクラス分類結果を得られる可能性が高くなると思われる．そこである分類器と他の分類器が共同でクラス分類を行う分類精度が高い分類器の生成を目的とした．このような二つの分類器で構成する方法は集団学習の考え方を取り入れている部分である。

現状では，分類器同士が直接情報をやり取りする手段を持たないため，属性追加手法では図 1 のように訓練データを媒介として，他の分類器が出力したクラス分類結果を伝達することを考えている．

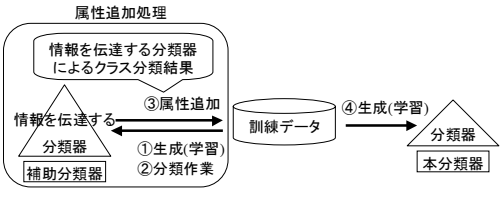

図 1: 属性追加の概要

情報を伝達したい分類器が出力したクラス分類結果は，訓練データに属性 (値) として追加することにより，生成する分類器に情報を伝達する．この訓練データに追加する属性が，追加される属性であり，この属性を追加するデータ前処理のことを本稿では属性追加と呼んでいる．属性追加を行うことにより，分類器生成アルゴリズムが他の分類器の意見 (クラス分類結果) を取り入れて分類器を生成する形式となるため，2つの分類器が共同でクラス分類を行う分類器の生成を行うことができる．本稿においては，情報を伝達する分類器のことを「補助分類器」とし，また最終的に生成する分類器を「本分類器」と呼ぶことにする．

2.2 属性追加の処理手順

次に属性追加の基本となる処理手順を以下に示す.

＜属性追加の基本となる処理手順＞

Step1:追加属性の生成 補助分類器生成のための訓練データ集合 T_s にクラスが $c_1, c_2, \cdots, c_n$ の n 個存在する場合，属性値が「$true$」と「$false$」の 2 値の属性を n 個生成する．生成した各属性の属性名は，各クラス名「c_1」「c_2」…「c_n」をつける．

Step2:補助分類器の生成 補助分類器生成アルゴリズム AL_s から補助分類器生成用の訓練データ集合 T_s を用いて補助分類器 S を生成する．

連絡先: 井芹史明，大分大学大学院工学研究科，〒 860-157 大分県大分市大字旦野原 700 番地，TEL:097-554-7866，FAX:097-554-7866，Mail:crs1406@csis.oita-u.ac.jp

Step3:属性追加 (1) Step1 で生成した全ての属性を本分類器生成用の訓練データ集合 T_m の全ての訓練データに追加する．

Step3:属性追加 (2) 各訓練データに追加した属性の属性値は，本分類器生成用の各データを補助分類器 S に入力して得た出力により決める．その決定方法は，補助分類器 S の出力がクラス $c_i(1 \leq i \leq n)$ である場合，追加した属性の属性名が「c_i」である属性の属性値を「$true$」とし，それ以外の追加した属性の属性値は「$false$」とする．

Step4:本分類器の生成 本分類器生成アルゴリズム AL_m から Step3 で属性と属性値を追加した本分類器生成用の訓練データ集合 T_m を用いて本分類器 M を生成する．

Step1 では，属性追加に用いる補助分類器 S が出力した訓練データのクラス分類結果を訓練データに書き込むための属性を生成している．生成した各属性は，属性値により訓練データが属性名と同じ名前のクラスに属するかどうかを示すことができる属性となっている．例えば，属性名が「c_1」の属性の属性値が「true」の訓練データは，クラス「c_1」に属していることを示している．

Step2 では，属性追加に用いる補助分類器 S を生成している．

Step3(1) では，Step1 で生成した属性を本分類器生成用訓練データ集合 T_m の全ての訓練データに追加している．これは属性追加に用いる補助分類器 S が出力した各訓練データのクラス分類結果を各訓練データに書き込むためのものである．

Step3(2) では，図 2 のように属性追加で用いる補助分類器 S が出力した各訓練データのクラス分類結果に対応した属性値を設定している．このようにして，属性追加に用いる補助分類器 S が出力したクラス分類結果を訓練データ集合 T_m を媒介として生成する本分類器 M に伝えることを可能としている．

C1の値	C2の値	…	Cnの値	分類器Mの出力
TRUE	FALSE	…	FALSE	C1
FALSE	TRUE	…	FALSE	C2
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・
FALSE	FALSE	…	TRUE	Cn

図 2: 追加属性の属性値と分類器 M の出力の対応

2.3 観測データを用いた属性追加

図 3 の観測データを用いて属性追加の各ステップの詳細を説明する．この観測データの属性は「天候」「気温」「温度」「風」であり，クラスは「開催」と「中止」である．これを補助分類器用観測データ T_s と本分類器用観測データ T_m と半分に分割し説明を行う．

Step1 では，補助分類器 S が出力するクラス分類結果を訓練データに書き込むための属性を生成する．観測データはクラス「開催」とクラス「中止」と同じ名前の属性「開催」と属性「中止」を生成する．生成した属性は「true」と「false」の２値とする．そのため図のように，Step3(2) の補助分類器 S が訓練データがクラス「開催」に属すると判断した場合，訓練データに追加した属性「開催」の属性値は「true」に，属性「中止」の属性値は「false」に設定される．

Step2 では，観測データ T_S によって補助分類器 S を分類器生成アルゴリズム A_S によって生成する．

天候	気温	湿度	風	クラス
晴	85	85	なし	中止
晴	80	90	あり	中止
曇	83	86	なし	開催
雨	70	96	なし	開催
雨	68	80	なし	開催
雨	65	70	あり	中止
曇	64	65	あり	開催

補助分類器用観測データT_S

天候	気温	湿度	風	クラス
晴	72	95	なし	中止
晴	69	70	なし	開催
雨	75	80	なし	開催
晴	75	70	あり	開催
曇	72	90	あり	開催
曇	81	75	なし	開催
雨	71	91	あり	中止

本分類器用観測データT_m

図 3: 観測データ

開催の値	中止の値	補助分類器Sの判断
TRUE	FALSE	開催
FALSE	TRUE	中止

図 4: 生成した属性の属性値と補助分類器 S の判断の対応

Step3(1) では，図 5 のように Step1 で生成した属性を観測データ T_m に追加する．つまり属性「開催」と属性「中止」を追加する．この時点では，追加した各属性の属性値は決まっていないため欠落値「?」である．

気候	気温	湿度	風	クラス
晴	72	95	なし	中止
晴	69	70	なし	開催
雨	75	80	なし	開催
…	…	…	…	…
雨	71	91	あり	中止

Step3(1) ↓　　追加属性

気候	気温	湿度	風	開催	中止	クラス
晴	72	95	なし	?	?	中止
晴	69	70	なし	?	?	開催
雨	75	80	なし	?	?	開催
…	…	…	…	…	…	…
雨	71	91	あり	?	?	中止

図 5: 属性追加 (1)

Step(2) では，生成した補助分類器 S から得た出力により，Step3(1) で追加した各属性の属性値の設定を行っている．先に説明したように入力がきて分類器が「開催」に分類すると，属性「開催」の属性値は「?」が「true」となり，属性「中止」の属性値は「?」が「false」となる．このようにして，全ての観測データ T_m に分類結果より属性「開催」と「中止」の属性値を決定していく．

3. 実験

本節では今回の提案手法である属性追加の比較・評価を行うための実験の内容を示す．実験にはニュージーランドのワイカイト大学で開発されているデータマイニングソフト Weka を使用する．Weka で用意されている分類器生成アルゴリズムを使用し，また分類精度も Weka によって算出する．また実験データには UCI の Machine Learning Repository に収められているサンプルデータを Weka で扱うことのできるファイル形式 (arff 形式) に変換された実験データを無作為に１０個用いた．図 6 に実験データを示す．

実験データ名	属性数	クラス数	データ数
audiology	69	24	226
soybean	35	19	683
vowel	13	11	990
zoo	17	18	101
autos	25	7	205
vehicle	18	4	846
lymph	18	4	148
sonar	60	2	208
credit-g	20	2	1000
vote	16	2	435

図 6: 実験データ

3.1 実験内容

次に実験内容を示す．属性追加の評価方法を 2 パターン用意し，それぞれの方法で提案手法である属性追加を施した前と施した後を比較する内容となっている．

＜評価方法 1 ＞ (図 7)

Step1 実験データを $\frac{1}{2}$ に分割する．

Step2 半分を補助分類器用訓練データ集合 T_s に設定し，半分を本分類器用訓練データ集合 T_m に設定する．

Step3 属性追加処理を行う．(属性追加を施さない場合はこのステップを除く)

Step4 属性追加を施した本分類器用訓練データ集合を用いて交差検定法を行い本分類器の分類精度を得る．

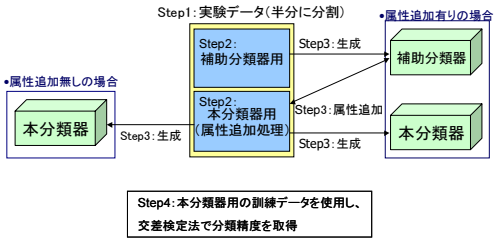

図 7: 評価方法 1

＜評価方法 2 ＞ (図 8)

Step1 実験データをデータを順に $T_1, T_2, \cdots, T_{10}$ の 10 セットに分割する．

Step2 1 セットのデータ T_i(初期値 $i = 1$) をサンプルデータに設定，残り 9 セットを補助分類器用訓練データ集合 T_s，また本分類器用訓練データ集合 T_m に設定する．

Step3 属性追加処理を行う．(属性追加を施さない場合はこのステップを除く)

Step4 本分類器にテストデータを入力し，分類精度を得る．

Step5 サンプルデータを T_{i+1} に設定し，Step2 へ戻る．

Step6 $i <= 10$ になるまで Step2〜Step5 を繰り返す．

Step7 10 回のループで得た分類精度の平均をとり，本分類器の分類精度を得る．

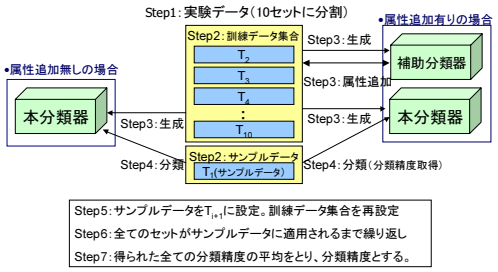

図 8: 評価方法 2

また属性追加手法の処理手順で使用する補助分類器の生成アルゴリズム AL_s と本分類器の生成アルゴリズム AL_m は以下の 3 種類とする．

- 決定木 (DT)
- ニューラルネットワーク (NN)
- サポートベクターマシン (SVM)

これら 3 種類を補助分類器生成アルゴリズムと本分類器生成アルゴリズムに組み合わせて設定し実験を行う．組み合わせは以下の 6 組の組み合わせである．これは実験方法 1 と方法 2 ともに同様である．

＜実験を行った補助分類器と本分類器の組み合わせ＞

組み合わせ 1 補助分類器 (NN)，本分類器 (DT)

組み合わせ 2 補助分類器 (NN)，本分類器 (SVM)

組み合わせ 3 補助分類器 (DT)，本分類器 (NN)

組み合わせ 4 補助分類器 (DT)，本分類器 (SVM)

組み合わせ 5 補助分類器 (SVM)，本分類器 (DT)

組み合わせ 6 補助分類器 (SVM)，本分類器 (NN)

3.2 実験結果

今回は前節に示した補助分類器と本分類器の組み合わせの中からスペースの関係上二つを絞り，実験結果を示すことにする．

＜評価方法 1 ＞

評価方法 1 で補助分類器をニューラルネットワーク，本分類器を決定木に設定した場合の実験結果を図 9 に示す．つまり決定木だけでクラス分類した場合と，ニューラルネットワークと共同でクラス分類を行った場合を比較したものである．

…精度向上例　…精度低下例

実験データ名	属性追加前	属性追加後	向上値
audiology	73.4513%	77.8761%	4.4248%
soybean	88.3041%	94.7368%	6.4327%
vowel	63.2323%	81.0101%	17.7778%
zoo	88.2353%	88.2353%	0.0000%
autos	73.7864%	67.9612%	−5.8252%
vehicle	69.0307%	82.5059%	13.4752%
lymph	60.8108%	87.8378%	27.0270%
sonar	59.6154%	64.4231%	4.8077%
credit-g	69.4000%	87.6000%	18.2000%
vote	95.4128%	97.2477%	1.8349%

図 9: 評価方法 1：DT の属性追加前と後の分類精度の比較

次に評価方法1で補助分類器を決定木，本分類器をニューラルネットワークに設定した場合の実験結果を図 10 に示す．つまりニューラルネットワークだけでクラス分類した場合と，決定木と共同でクラス分類を行った場合を比較したものである．

実験データ名	属性追加前	属性追加後	向上値
audiology	73.4513%	75.2212%	1.7699%
soybean	92.9825%	94.1520%	1.1695%
vowel	80.8081%	79.5960%	−1.2121%
zoo	98.0392%	98.0392%	0.0000%
autos	66.0194%	79.5960%	13.5766%
vehicle	78.0142%	85.3428%	7.3286%
lymph	81.0811%	78.3784%	−2.7027%
sonar	80.7692%	78.8482%	−1.9210%
credit-g	70.2000%	72.4000%	2.2000%
vote	91.7431%	94.4954%	2.7523%

図 10: 評価方法 1：NN の属性追加前と後の分類精度の比較

＜評価方法 2 ＞

次に評価方法 2 で評価方法 1 と同様の分類器の組み合わせで実験を行った時の結果を示す．評価方法 2 で補助分類器をニューラルネットワーク，本分類器を決定木に設定した場合の実験結果を図 11 に示す．

実験データ	属性追加前	属性追加後	向上値
audiology	73.1028%	82.3518%	9.2490%
autos	33.6809%	42.5238%	8.8429%
credit-g	72.2000%	72.2000%	0.0000%
lymph	79.0476%	85.1429%	6.0953%
sonar	63.0476%	66.4286%	3.3810%
soybean	83.0669%	90.0405%	6.9736%
vehicle	73.6541%	82.2759%	8.6218%
vote	97.0243%	95.6448%	−1.3795%
vowel	52.2222%	66.9697%	14.7475%
zoo	92.2727%	95.1818%	2.9091%

図 11: 評価方法 2：DT の属性追加前と後の分類精度の比較

次に評価方法 2 で補助分類器を決定木，本分類器をニューラルネットワークに設定した場合の実験結果を図 12 に示す．

実験データ	属性追加前	属性追加後	向上値
audiology	82.3518%	79.7036%	−2.6482%
autos	42.5238%	34.8571%	−7.6667%
credit-g	72.0000%	70.4000%	−1.6000%
lymph	85.8572%	83.9048%	−1.9524%
sonar	66.4286%	63.5238%	−2.9048%
soybean	88.1607%	88.4697%	0.3090%
vehicle	82.6316%	75.0700%	−7.5616%
vote	95.6448%	95.4228%	−0.2220%
vowel	67.2728%	52.8283%	−14.4445%
zoo	96.0909%	93.2727%	−2.8182%

図 12: 評価方法 2：NN の属性追加前と後の分類精度の比較

3.3 考察

評価方法1については，補助分類器 NN，本分類器 DT で行った場合実験データ autos 以外の実験データで属性追加後の分類精度向上が見られる．また補助分類器 DT，本分類器 NN で行った場合においても実験データ vowel，lymph，sonar 以外の実験データで属性追加後の分類精度向上が見られる．このことより評価方法 1 については属性追加のデータ前処理としての有用性が示されている．

評価方法 2 については，補助分類器 NN，本分類器 DT で行った場合実験データ vote 以外の実験データで属性追加後の分類精度向上が見られる．しかし評価方法 2 において補助分類器に決定木，本分類器にニューラルネットワークを設定し属性追加を行った場合，実験データ soybean 以外のすべての実験データで分類精度の低下がみられた．

補助分類器 DT、本分類器 NN の組み合わせを考えた場合、もともと決定木はニューラルネットワークの分類精度に比べその精度は低い結果となる．また精度の悪い補助分類器から属性追加により精度の悪い情報が本分類器に伝えられると本分類器の精度も悪くなることは当然と考えられる．つまり補助分類器にニューラルネットワークより精度の低い決定木を設定したことで，本分類器のニューラルネットワークの分類精度が低下した．評価方法 2 では，補助分類器用訓練データ集合と本分類器用訓練データ集合を同じものに設定しおり，同じ訓練データ集合で学習しているため，本分類器に対して，先に学習した補助分類器のクラス分類結果の影響が高くなると考えられる．そのためニューラルネットワークの分類精度が大幅な低下がみられた．評価方法 1 では，補助分類器用訓練データ集合と，本分類器用訓練データ集合とを分けて設定しているため，本分類器への補助分類器の影響が少なく，適度に補助分類器の情報を取り入れ分類精度向上が行われていると考えられる．

本稿の実験で，属性追加手法を適用するにあたって，補助分類器と本分類器の訓練データ集合に同じデータを用いると，補助分類器のクラス分類結果が強く影響し，補助分類器の精度が低いことで精度低下を招く恐れがあることがわかった．しかし，補助分類器と本分類器の訓練データ集合を分けて分類器を生成することで，補助分類器の精度が低い場合であっても影響が弱く，さらに本分類器に精度向上をもたらす作用があることがわかった．つまり提案手法である属性追加手法のデータ前処理としての有用性はあることが証明された．

4. おわりに

本稿では，データ前処理手法として，属性追加手法を提案し，属性追加手法適用前と適用後の分類精度の比較を行い，属性追加の有用性の検証を行った．今回属性追加手法では補助分類器に影響され本分類器の分類精度が低下するという結果を得た．しかし補助分類器と本分類器の各訓練データ集合を補助分類器の影響を考え設定を行うことで属性追加手法のデータ前処理としての有用性は示される．今後の課題としては，他のデータ前処理手法との比較を行い属性追加手法の有効性などを検証する必要がある．また属性追加手法は複数の学習法を用いて構成されているため，集団学習の分野にあたる．このため集団学習のアルゴリズムである AdaBoosting 手法などと比較し，属性追加手法の有効性の検証を行って行きたい．

参考文献

[寺邊 00] 寺邊 正大，片井 修，椹木 哲夫，鷲尾 隆，元田 浩: 相関ルールにもとづく属性生成手法: 人工知能学会誌，Vol.15，No1，pp187-197(2000)．

[元田 97] 元田 浩，鷲尾 隆: 機械学習とデータマイニング: 人工知能学会誌，Vol.12，No7，pp11-18(1997)．

[Witten00] lan H.Witten, Eibe Frank: Data Mining-Practical Machine Learning Tools and Techniques with Java Implementations:Morgan Kaufmann Publishers(2000)．