6S-06

数値を含むデータから得られた識別パターンに基づく連関分類器とその評価

前田 健太郎 † 亀谷 由隆 †
† 名城大学理工学部情報工学科

1 はじめに

信用リスクの予測などで機械学習手法を適用する際，予測した根拠が得られると好ましい場面が多く存在する．そのような予測の根拠が読み取れる連関分類器 (associative classifier) を構築するために，本研究では最近提案された識別パターン発見手法である ECHO (Exhaustive Covering in Hybrid Domains)[1] を利用することを提案する．ECHO の特徴は次節で述べる．簡単のため，本研究では二値分類問題を対象にする．構築した分類器に対してランダムフォレスト（以下 RF）を比較対象の軸とする評価実験の結果について報告する．

2 数値を含むデータからの識別パターン発見

従来の識別パターン発見手法は入力属性の対象は離散値であり，入力属性に連続値が含まれている場合は事前に各属性を離散化する必要があった [2]．そこでは属性の組み合わせを考慮していないため情報損失が起こる．一方，ECHO では入力属性に連続値を含むデータでもパターン発見できる．決定木，回帰木なども入力属性に連続値を扱うことができるが貪欲法に基づくため網羅性がない．一方，ECHO は網羅的に複数のパターンを探索できる．

ECHO はパターン間の制約として飽和制約と最良カバー制約を使用する．飽和制約とはカバーする正事例の集合が同一である複数のパターンのうち最も特殊なパターンのみを残す制約である．最良カバー制約は「出力パターンはそれがカバーする正事例 t のいずれかにおいて，t をカバーするパターンの中で最大の関連度を持たなければいけない」という制約である．

次に，ECHO で利用する識別スコアとパラメータの導入を行う．まず，適合率（確信度）とはパターンを満たす事例のうち正例であるものの割合である．そして，正（負）の再現率（サポート）とは正（負）例のうちそのパターンを満たすものの割合である．この時，Support Differece（以下 SuppDiff）は（正の再現率）–（負の再現率）により定義される．また，F 値は適合率と正の再現率の調和平均，Lift は適合率を正例の割合で割ったものである．そして χ^2 値はパターンを満たすかどうか，正例であるかどうかという 2 つの真偽値の間で定義されるものである．我々の ECHO の定義において，ユーザは最小確信度 (minconf) や最小（正）サポート (minsup) を指定し，minconf 以上の確信度，minsup 以上の正サポートをもつパターンのみ出力させることができる．

3 分類器の実装

分類器は機械学習ライブラリ scikit-learn に準拠したモデルの形で構築した．scikit-learn 準拠モデルにすることで，RF などの既存モデルやグリッドサーチなどの評価モデルが容易に適用可能となる．scikit-learn 準拠モデルはクラス設計されており，fit 関数（学習用関数）と predict 関数（予測用関数）の実装を行った．fit 関数の実装では主に識別パターンの取得を行う．ECHO の標準出力とエラー出力をファイルに書き込むことでどのパターン群から予測を立てたのか根拠を残している．predict 関数では fit 関数で得られた識別パターンを用いてテストデータのクラス予測を行う．予測の流れを以下に示す．

1. 飽和制約ありの識別パターンを取得

2. テストデータにマッチする識別パターンの数
- **0 個** → ステップ 3 へ
- **1 個** → そのパターンのクラスを予測クラスに決定
- **2 個以上** → 予測スコアに基づき予測クラスを決定

3. 飽和制約を外した識別パターンを取得

4. テストデータにマッチする識別パターンの数
- **0 個** → デフォルトクラス（多数派クラス）を割り当て
- **1 個** → そのパターンのクラスを予測クラスに決定
- **2 個以上** → 予測スコアに基づき予測クラスを決定

なお，予測スコアは次節の ECHO 分類器のパラメータ調整によって確定する．

4 評価実験

4.1 実験条件

本実験において層化分割，交差検証，グリッドサーチ，RF，決定木，ワンホットエンコーディング，F 値算出は scikit-learn で提供される機能を利用する．

ECHO 分類器のパラメータ確定のために複数の組み合わせを試す予備実験を行った．検証用のデータは credit-german とする．予測スコアは前述の F 値，Lift，SupDiff，χ^2 値の 4 通り，minconf={0.5, 0.7}, minsup={1, 5, 10, 20} を指定した．予備実験で得られたパラメータを ECHO 分類器に適用し，10 分割層化交差検証を行うことで少数派クラスに対する F 値を求めた．なお，過学習防止と計算時間短縮のため 1 クラスの学習にかける時間の上限を 1 時間とする．上限時間 15 分の場合も試したが全体的にスコアが悪化したため 1 時間を本実験の設定とする．

比較対象には RF と決定木を採用し，比較のためデータの前処理とグリッドサーチによるパラメータ調整を行った．前処理段階では欠損値処理とワンホットエンコーディングを行っている．通常 RF は欠損値があっても学習可能であるが，scikit-learn の仕様上処理が必要である．欠損値の補完はデータの存在する値の中から任意の値を補完した．同様の理由で入力属性に離散値を含む場合，ワンホットエンコーディングを行い離散属性を数値化して学習させる． RF と決定木においてグリッドサーチしたパラメータと探索した値を以下に示す．RF は 1, 2, 3 全てを，決定木は 2, 3 のパラメータ調整を行った．n_estimators は生成する木の数，min_samples_leaf は葉が保持する最小のサポート数，class_weight はクラスの重み付みを考慮するかどうかである．

Evaluating an Associative Classifier that Uses Discriminative Patterns from Hybrid Transactions

† Kentaro Maeda
† Yoshitaka Kameya
Department of Information Engineering, Faculty of Science and Technology, Meijo University (†)

表 1: ECHO 分類器と既存モデルの比較 (F 値)

分類器					データセット					
					hybrid		numeric		symbolic	
	学習スコア	予測スコア	minconf	minsup	credit-g	heart-c	breast-w	diabetes	mushroom	vote
ECHO	χ^2 値	χ^2 値	0.5	1	0.498	0.726	0.903	0.664	0.913	0.932
ECHO	F 値	Lift	0.7	10	0.521	0.706	0.884	0.640	0.939	0.770
ランダムフォレスト					0.594	0.807	0.956	0.695	0.950	0.948
決定木					0.579	0.793	0.919	0.639	0.941	0.939

表 2: ECHO 分類器と既存モデルの比較 (実行時間 (秒))

分類器					データセット					
					hybrid		numeric		symbolic	
	学習スコア	予測スコア	minconf	minsup	credit-g	heart-c	breast-w	diabetes	mushroom	vote
ECHO	χ^2 値	χ^2 値	0.5	1	144083	144049	710	144071	328	70
ECHO	F 値	Lift	0.7	10	114159	87949	734	144026	470	73
ランダムフォレスト					2530	1266	1258	2226	5238	1137
決定木					50	12	4	4	73	3

1 n_estimators : [50, 100, 200, 500, 1000]
2 min_samples_leaf : [1, 2, 5, 10, 20]
3 class_weight : ["balanced", None]

RF と決定木においては 10 分割層化交差検証に上記のグリッドサーチを入れ子にすることで F 値の算出を行った．

実験データセットは UCI 機械学習リポジトリから選んでおり，離散属性と連続属性の両方を含むデータ (以下 hybrid) である credit-german, heart-cleveland，連続属性のみのデータ (以下 numeric) である breast-w, diabetes，離散属性のみのデータ (以下 symbolic) である mushroom, vote を用いた．いずれも二値分類データである．

- credit-german : 信用リスク/クラス: bad, good
- heart-cleveland : 心臓病/<50, >50_1
- breast-w : 乳癌/benign, malignat
- diabetes : 糖尿病/tested_negative, tested_positive
- mushroom : きのこ/edible, posonous
- vote : 議会投票/democrat, republican

なお実験には Weka の公開サイト (`https://www.cs.waikato.ac.nz/ml/weka/`) で提供されている ARFF 形式のファイルをトランザクション形式に変換したものを利用している．

4.2 実験結果

予備実験の結果，学習スコア χ^2 値，予測スコア χ^2 値, minconf=0.5, minsup=1 と，学習スコア Lift，予測スコア Lift, minconf=0.7, minsup=10 の 2 通りを ECHO に適用した．実験結果を表 1 と表 2 に示す．表 1 は F 値に関する結果，表 2 は実行時間に関する結果である．本実験では学習から予測，パラメータ調整にかかるすべての時間を実行時間と判断し合計の時間を示している．

表 1 の F 値について，χ^2 値のスコアは credit-german データ以外のデータに対しては決定木を上回るまたは同等のスコアを出している．Lift のスコアは他の ECHO パラメータより良い F 値を出したが，全体として決定木より劣っている．表 2 の実行時間について，ECHO はデータセットが大きくなると探索時間も指数的に増え多大な時間がかかる．credit-german においてはパラメータにもよるが交差検証全体で 30 時間から 40 時間の実行時間を要した．データ数が少ない時 RF より早い実行時間になったのは，グリッドサーチの有無の影響である．説明可能性について，ECHO 実行で得られた識別パターンの一部を図 1 に示す．これは credit-german データの bad（信用リスク高）クラスに対する識別パターンである．checking_status='<0'（口座の預金状態）や other_parties='none'（共同申請者や保証人の有無）などから，ユーザが直接的にわかりやすいものを得られたと言える．

22.50≤age<60.50, checking_status='<0'
605.0≤credit_amount<10330, 11.50≤duration
foreign_worker='yes', 2.500≤installment_commitment
other_parties='none'

図 1: ECHO 実行で得られた識別パターンの一部

5 おわりに

本研究では，数値を含むデータから得られた識別パターンを生成する ECHO を利用した分類器を構築し，その評価を行った．今後の課題として，データセットとパラメータの相性を考慮して ECHO 分類器にグリッドサーチ等でのパラメータ調整を実施すれば，ECHO 分類器の F 値を本実験より向上できると考える．また説明可能性について，RF および決定木と比較する必要がある．

参 考 文 献

[1] 亀谷由隆: 数値を含むデータからの効率的なパターン発見に向けて，FIT-18 予稿集 (2018).

[2] F. Thabtah: A Review of Associative Classification Mining, Knowledge Engineering Review, Vol. 22, No. 1, pp. 37–65 (2007).