科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　２６　日現在

機関番号：　３３９１９
研究種目：　若手研究(B)
研究期間：　2012～2013
課題番号：　２４７００１４１

研究課題名（和文）識別パターン発見手法に基づく確率モデルの説明的分析手法の開発

研究課題名（英文）Explanatory Analysis of Probabilistic Graphical Models based on Discriminative Pattern Mining

研究代表者
亀谷　由隆（Kameya, Yoshitaka）

名城大学・理工学部・准教授

研究者番号：６０３６１７８９

交付決定額（研究期間全体）：（直接経費）　　2,000,000円、（間接経費）　　600,000円

研究成果の概要（和文）：本課題では説明可能性を備えた機械学習の実現に向けて，確率モデル，特に広く知られるベイジアンネットの説明的分析手法の開発を行った．この分析手法では，膨大な数の説明の候補の中から適切なものを選ぶために，データマイニング分野で研究されている識別パターン発見手法を利用する．本課題においては，説明の選択基準の洗練および説明探索の高速化については実用レベルの手法が確立できたといえる．一方，ベイジアンネット上での確率推論の高速化において実装上の課題が残ったが，本課題で構築したプロトタイプ実装を基に実装を継続することで説明的分析手法の実現への見通しが得られたものと考えている．

研究成果の概要（英文）：In this project, we have developed an explanatory analysis method for probabilistic models, Bayesian networks in paticular, to bring explanability to machine learning techniques. Our analysis aims to find an appropriate explanation for the observation from a huge number of possible ones. To do this in practical time, we built some sophisticated techniques for discriminative pattern mining based on a popular frequent pattern mining algorithm called FP-Growth. Finally, we have achieved to refine the selection criteria of explanations and to have a fast discriminative pattern mining algorithm. Although there remains a future work on optimizing probabilistic inference for our explanatory analysis, we have obtained a couple of new insights and prototype tools towards an implementation of our explanatory analysis method.

研究分野：　総合領域

科研費の分科・細目：　情報学・知能情報学

キーワード：　ベイジアンネット　説明的分析　識別パターン発見

１．研究開始当初の背景

近年大きく発展した機械学習では，医療データ等の重要なデータへの応用が増え，日常生活においても機械学習手法を使ったサービスが身近になってきており，それと同時に，医者等の専門家や一般消費者に学習・予測結果を納得してもらう（ブラックボックス化を避ける）ために，説明可能性を備えた機械学習技術の重要性が増してきているという背景があった．例えば，前者の場面では医者・マーケット分析者等の専門家が機械学習による解析結果を適切に理解し，意思決定する必要がある．また，後者のように一般消費者が機械学習システムの出力に直接触れる場合はその消費者への透明性は高い方がよい．そして，説明可能性を備えた情報システムの構築は 80 年代のエキスパートシステム研究以来の人工知能分野全体の大きな課題となっている．

説明可能性を備えた機械学習／情報システムの技術として，本研究では確率モデル，特にベイジアンネットを対象とした．ベイジアンネットは人工知能分野で広く普及しており，診断システム等で様々な応用例があることに加えて，研究代表者が長年携わってきた研究分野（論理型プログラミング言語による確率モデリング）と近い関係を持っていたからである．

２．研究の目的

上述のように，本研究では，確率モデル，特に広く普及しているベイジアンネットに基づく機械学習における説明可能性の向上を目指し，近年データマイニング分野で注目される識別パターンの発見手法に基づく効率的なベイジアンネットの説明的分析手法の開発を行うことを目的とした．

具体的には，主に 90 年代前半に行われていたベイジアンネットの証拠集合に基づく感度分析 (sensitivity analysis) 手法に対し，独自の識別パターン発見 (discriminative pattern mining) 手法を利用して大規模ベイジアンネット上での高速な説明の探索を行うことを考える．識別パターン発見は近年データマイニング分野で盛んに研究されており，興味ある観測 o に対して識別スコア，例えば F 値 F(o, e) の高い規則 e ⇒ o を効率的に探索するタスクである．

本課題で考えている説明的分析を Jensen らの例で説明する．この例では，Holmes 氏がある日自宅の芝生だけ濡れていることに気付き，その理由を考えているとする．図 1 は Holmes 氏の自宅の周辺状況を表現したベイジアンネットである．昨晩雨が降ったこと，自宅のスプリンクラーを止め忘れたことを yes, no の 2 値をとる確率変数 R, S で表し，同様に Holmes 氏，Watson 氏，Gibbon 夫人の家の芝生が濡れていることを 2 値の確率変数 H, W, G で表している．確率変数の依存関係は有向非循環グラフ（実線）で表現される．このとき我々の説明的分析では，観測 o={H=yes, W=no, G=no} に対し，「昨晩雨は降っておらず，スプリンクラーを止め忘れた」という説明 e = {R=no, S=yes} を提示する．

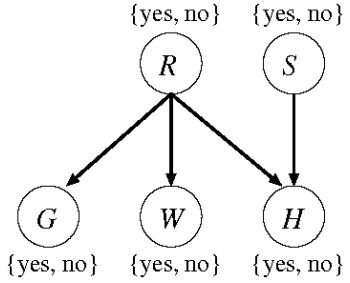

図 1

このように観測 o に対して尤もらしい説明 e を求めることができるならば，上述の説明可能性を備えた機械学習／情報システムに一歩近づくことができると思われる．しかし現実的には，考え得る説明 e の数は原因変数の数に対して指数的であり，大規模なベイジアンネットでは実時間で列挙しきれないという計算上の問題がある．そこで本課題では，本研究では独自の識別パターン発見手法を利用して大規模ベイジアンネット上での高速な説明の探索を行うことを目指した．

３．研究の方法

当初予定していた部分課題は以下の４つである．

(1) 説明の選択基準の洗練
(2) 説明の探索の高速化
(3) 確率推論の高速化
(4) 実証

部分課題(1)における説明の選択基準に関しては，近年進展しつつある explanation-aware computing という研究分野での議論，分類器の評価手法，特徴選択手法，識別パターン発見手法に関する文献を調査した．

そして，部分課題 (2) では，独自の識別パターン発見の利用を考え，その改良を目指した．具体的には，観測 o に対する説明 e の選択基準のスコアにおいて上位 k 個の説明を効率よく出力するパターン発見アルゴリズムを考案することを目指した．

また，部分課題 (1) の対象である説明の選択基準の多くでは確率 p(e | o) や p(o | e) を用いている．ベイジアンネットの説明的分析ではこれらの確率値はベイジアンネットから推論される．従って，効率のよい説明的分析を行うためには，ベイジアンネット上での高速な確率推論が必要であり，これを部分課題 (3) とした．

部分課題（4）では一つの説明分析の例として，確率モデルに基づくクラスタ分析結果の説明を本課題で開発した説明的分析手法で求めることを考えた．クラスタ分析はデータマイニングの基本手法の一つとして知られ，類似事例をグループ化することにより分析対象のデータにおける部分的な傾向を抽出する．このクラスタ分析の実用上の問題点として，「得られたクラスタ（事例のグループ）が理解できない」というものがある．この問題を汎用の枠組みで解決する方法として本課題の説明的分析手法を利用することを目指した．

４．研究成果

まず，部分課題(1)の「説明の選択基準の洗練」では説明の探索の基本手法となるパターン発見アルゴリズム RP-Growth （国際会議 SDM-12 で発表）を実装した結果，F 値（説明探索の場合に置き換えると e を説明，o を観測としたときの p(e | o) と p(o | e) の調和平均 F(e, o)）を用いた場合に効率性と自然さのバランスが良くなることが分かった．例えば，従来の分類器の評価手法，特徴選択手法で広く用いられているχ^2 値や Yuan et al．らによって提案された一般化ベイズ因子(generalized Bayes factor, GBF)では p(o | e) を重視するために上界が大きく評価され，分枝限定法による枝刈りが効きにくいことが実験により分かっている．また，生産性制約（productivity constraint）と呼ばれるパターン間の制約（説明探索の場合に置き換えると，説明 e, e’に対し e ⊂ e’であるにも関わらず F(e, o) ≧ F(e’, o) であるなら e’は非生産的であるとして削除）を用いると簡潔で自然なパターンの集合が得られることが分かった． また，説明の選択基準に対する一般的性質として双単調性（dual-monotonicity, p(e | o) に対して単調増加し，p(e | ¬o) に対して単調減少する)を見つけ，双単調性を満たす選択基準では説明空間上の探索において分枝限定法に基づく枝刈りが可能であることを示した（国際学会 DaWaK-13 で発表）．

部分課題（2）については，まず上述の RP-Growth の実装を完了した．RP-Growth は著名な頻出パターン発見アルゴリズムである FP-Growth に基づいており，深さ優先探索による省メモリ性，FP-tree と呼ばれるデータ構造（長いパターンを探索していく際に縮約可能な条件付きデータベースと見なせる）を引き継いでいる．そして，RP-Growth では FP-Growth で暗黙のうちに利用されていた接尾探索木を意図的に使用することにより，上述の生産性制約と分枝限定法に基づく枝刈りが可能であることが示されている．このような探索の工夫により，大規模データに対しても RP-Growth は良好な探索性能を示している．更に，効率的な飽和パターン発見手法として著名な LCM アルゴリズムに対して，この接尾探索木を用いた生産性制約による枝刈りの仕組みを安全に導入可能であることが理論的に示されている（LCM が実施する prefix 保存閉包拡張と対称的な操作を施すと接尾探索木が形成される）．そして標準的データセットを用いたベンチマーク評価によって，従来の接頭探索木上での深さ優先探索，幅優先探索および反復深化法と比較しても安定して効率的であることが示した．これらの結果は国際学会 DaWaK-13 で発表済みである．

部分課題（3）の確率推論の高速化については，まず既存の高速な厳密確率推論アルゴリズムの検討を行った．具体的には，接合木(junction tree) アルゴリズム等，ベイジアンネットの確率推論アルゴリズムの詳細を文献調査し，これらのアルゴリズムをベイジアンネットの説明的分析にどのように適用するかを検討した．接合木アルゴリズムにおいてベイジアンネットから接合木を構築する過程では，実装の容易さを考慮して，三角グラフを経由して接合木を構築するグラフ理論的な手法ではなく，変数消去法(variable elimination)を経由して構築する方法を採り，構築アルゴリズムのプロトタイプ実装を行った．接合木上での確率推論アルゴリズムは著名なベイジアンネット構築ソフトウェア Hugin (www.hugin.com)で用いられるメッセージ伝播（message passing)方式を採り，こちらもプロトタイプ実装を行った．しかし，当初予定していた，僅かに異なる説明 e, e’間での部分推論結果を共有することによる説明的分析の高速化は実現されておらず，今後の課題として残っている．

部分課題（4）については，Naive Bayes モデルという名で知られる，ベイジアンネットの特殊形の確率モデルを用いたクラスタ分析結果に対する説明的分析を試みた．基本となる分析手法は本課題に先立つ 2011 年に国際会議 ICTAI-11 で発表していたが，この方法は幅優先探索に基づくパターン発見手法である Apriori 法を元にしたものであった．本課題においては，まず，この分析手法をメモリ効率の良い深さ優先探索に基づく RP-Growth をベースとしたものに変更した．更に，ICTAI-11 で発表した手法では個々の説明を AND 式の形に限定し，選択基準（F 値）の上位 k 個にランクされる説明の集合を提示していたが，説明式を従来の AND 式から選言標準形（disjunctive normal form）の AND-OR 式に拡張することで個々の説明の選択基準値が向上し，得られた説明も直観に合っていることが確認された．この AND-OR 式を効率的に探索するために，(i) まず RP-Growth に基づく方法で AND 式の説明集合を求め，(ii) 得られた AND 式の中で共通部分の多いものを OR で組み合わせる，という作業を行う．ステップ(i)，(ii) 共に説明の選択基準の上位 k, k’個を求めている．特

にステップ（ii）における探索において，説明選択基準の双単調性を利用した，従来とは逆向きの汎化方向での分枝限定法に基づく枝刈りを行う点が新しいと評価している．この工夫が説明式の双方向探索につながると期待される．

以上のように，双単調性やそれを用いた双方向の分枝限定法への可能性等，新しい知見を得ることができたことを含め，本課題の各部分課題について発展させることができたと考えている．その一方で，部分課題 (3)の進展が当初予定より遅れ，説明的分析手法全体の実装には至らなかった．本課題終了後も，これまでに構築してきたプロトタイプ実装を基に説明的分析手法の実装を継続し，今後の研究発表に繋げていきたい．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計 3 件）

① Y. Kameya, H. Asaoka, Depth-First Traversal over a Mirrored Space for Non-redundant Discriminative Itemsets, The 15th International Conference on Data Warehousing and Knowledge Discovery (DaWaK-13), 2013, pp.196-208.

② 小島諒介，亀谷由隆，佐藤泰介，Naive Bayes モデルを用いた効率的なクラスタラベリング手法，人工知能学会第 88 回人工知能基本問題研究会，2013, pp.19-24.

③ Y. Kameya, T. Sato, RP-Growth: Top-k Mining of Relevant Patterns with Minimum Support Raising, The 2012 SIAM International Conference on Data Mining (SDM-12), 2012, pp.816-827.

６．研究組織

(1)研究代表者

亀谷　由隆（KAMEYA, Yoshitaka）
名城大学・理工学部・准教授
研究者番号：60361789