科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２８　日現在

機関番号：　１２６０８
研究種目：　基盤研究(B)
研究期間：　2011～2013
課題番号：　２３３０００５４

研究課題名（和文）統計的アブダクションによる不確定性情報処理の研究

研究課題名（英文）Uncertainty informaion processing by statistical abduction

研究代表者
佐藤　泰介（Sato, Taisuke）

東京工業大学・情報理工学（系）研究科・教授

研究者番号：９０２７２６９０

交付決定額（研究期間全体）：（直接経費）　　10,700,000円、（間接経費）　　3,210,000円

研究成果の概要（和文）：統計的機械学習と論理的推論を融合した高水準の確率モデリング言語PRISMを改良し，ベイズ推論用の一般的MCMC法，Viterbi推論にもとづくVTおよびVTと変分ベイズ推論を融合したVB-VTと呼ばれるパラメータ学習法を開発，実装した。また確率の方程式を解くことにより確率文脈自由文法の接頭辞確率など確率の無限和が計算できるようにし，webのセッションログからのユーザの意図推定に応用した。

研究成果の概要（英文）：We have improved a logic-based modeling language PRISM which unifies statistical machine learning and logical inference by adding a general MCMC (Markov chain Monte Carlo) method, VT (Viterbi training) and VB-VT that extends VT with variational Bayes. We also enabled PRISM to calculate an infinite sum of probabilities through solving probability equations, which is applied to intention recognition of users from web log session data.

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：　確率論理　確率モデリング言語

１．研究開始当初の背景

DNA チップからのバイオデータ，email などのテキストデータ，あるいは Web ログデータなど多様なデータが日々大量にデータベースに蓄積されている．これらの生データはノイズ混じりで巨大であり，種々の不確定性を伴うために，論理的推論と統計的推論が組み合わせてデータマイニングを行うことにより論理的に一貫した最尤の情報を取り出すことが望まれていた．

一方統計的機械学習における確率モデリングと論理推論を融合した高度の機械学習技術の研究が欧米でも進展しつつあった．欧州では主としてイギリス，ベルギーの研究グループにより論理的な帰納推論を行う帰納的論理プログラミングに確率を導入した確率論理プログラミング（PLL, probabilistic logic learning）の研究が進められ，米国では主として西海岸の Stanford などのグループによりベイジアンネットワークに関係概念を取り込んだ確率関係モデルや近年注目を集めているマルコフ確率場に論理式を採り入れたマルコフ論理ネットワークなどのいわゆる統計関係学習（SRL, statistical relational learning) の研究が進められていた．

２．研究の目的

統計的機械学習と論理推論を統計的アブダクションの枠組みにもとづき融合し，欧米に於ける PILP や SRL の研究を先導する高度の記号的確率モデリング言語 PRISM およびその処理系を更に発展させることにより，汎用のベイズ推論や無限の確率和の計算など高度の不確定性情報処理を可能にする．

以下まず研究の基本的枠組みとなっている統計的アブダクションについて説明する．アブダクションは仮説推論とも言われ，演繹，帰納と並ぶ人間の論理推論機能の一つである．アブダクションを提唱した米国の哲学者 C.S.Peirce によるとそれは驚くべき観測事実 O に対し最良の説明 E を見出す推論であり，例えば道端で貝の化石を見つけた（O）とき，貝は年月が経つと化石になるという背景知識（KB）にもとづきここは昔海の底であった（E）と推論するのはアブダクションの例である．述語論理により定式化すると，アブダクションとは既存の知識を表す知識ベース KB，観測事実 O が与えられた時 E,KB が O を論理的に導くような E を推論することである．ここで E と KB は無矛盾でなければならない．しかし通常観測事実の説明は複数あり，我々はその内最良の説明を選択する必要がある．

統計的アブダクションはアブダクションに説明の確率を導入したものであり，説明を構成する単位となる abducible と呼ばれる基本的論理式に確率を割り当て説明の確率を計算することにより，最大確率を持つ説明を最良の説明として選択する．abducible の確率は観測事実から統計的に推定する．

我々は統計的アブダクションの一つの実現として，論理プログラムに基づいた確率モデリング言語 PRISM を以前から開発してきた．PRISM は KB として論理プログラム，説明として確率的選択を表す abducible である msw アトムの連言を採用している．

観測データ G と KB が与えられた時 PRISM は SLD 探索により E,KB |- G となる説明 E = msw1 ∧ ・・・ ∧ mswk をすべて探し，E の確率をその成分の msw アトムの確率の積として，G の確率をその説明の確率の和として計算する．なお SLD 探索に於いては探索結果を保存するテーブリング技法により探索の重複を防いでいる．また G の最大確率の説明は HMM（隠れマルコフモデル）の Viterbi アルゴリズムを一般化したアルゴリズムにより求める．

PRISM は動的プログラミング使った確率計算により，BN（ベイジアンネットワーク），HMM，PCFG （確率文脈自由文法）など既存の各種確率モデルを効率的に（既存のアルゴリズムと同等の時間計算量で）計算し，パラメータ学習を行うだけでなく，その汎用性により確率左隅文法や確率文脈自由グラフ文法など記述や実装の複雑さ故殆ど手が着けられたことのない確率モデルも容易に扱えることが以前の記述実験により示されている．

統計的機械学習ではデータの確率分布をベイジアンネットワークなどの確率モデルにより学習し，得られた確率分布から判別や予測などの情報処理を行う．他方統計的自然言語処理では PCFG に見られるように文法規則に確率を加えた確率文法によりデータのノイズや文法の曖昧性を処理する手法が発達している．前者はベイズ推論など確率分布の扱いに優れ，後者は構文解析など規則性の扱いに優れている．PRISM では両者を融合することにより，個体や関係が表現できず背景知識の処理能力を欠いていた従来の統計的機械学習の欠点を取り除いている．

３．研究の方法

PRISM の機械学習としての枠組みは生成的確率モデリングであり，ベイズ推論との適合性が良い．また生成的確率モデリングは次に述べる判別的確率モデリングに比して，結果の理解可能性が高いという利点を備えている．一方機械学習の応用として重要な判別問題においては，生成的確率モデルは判別的確率モデルに比較して，判別性能が劣る場合が多い．

そこで研究方法としては，PRISM におけるベイズ推論を実現させつつ，判別問題に対する判別性能を向上させる取り組みを行う．ま

た実問題に用いてプランニングに於ける意図推定など生成的モデリングによる説明可能性の問題を追及する.

４．研究成果

[H23 年度]
PRISM 処理系に H23 年夏に人工知能国際会議で発表したベイズ推論用の一般的 MCMC 法を実装した．全面的に C 言語による実装であり，Prolog による実装に対し 10 倍以上のサンプリング速度を実現することができた．その結果 PRISM は推論法としては今までの最尤法，変分ベイズ法に加えて MCMC 法による Vibterbi 解の推定，および対数周辺尤度の推定が可能になり，他に例のない豊富な種類のベイズ推論をユーザに提供するモデリング言語なった．また計算量の多いベイズ推論とは別に計算量の少ない VT(Viterbi training)に基づいたパラメータ学習を試験的に実装した.

[H24 年度]
PRISM のパラメータ学習法はすでに最尤推定に基づく EM 学習，事前分布として Dirichlet 分布を取り入れた MAP（maximum a posterirori）学習，変分ベイズに基づく VB（variational Bayes）学習が利用可能であるが，それらはすべて生成モデルの（周辺）尤度に基づくもので，必ずしも判別問題に於ける判別精度と結びつくものではなかった．そこで判別問題における判別精度の向上を目指し，前年度予備的実装を行った VT 学習を処理系に組み込み公開した.
VT が最大化する目的関数は尤度とは異なり隠れ変数を考慮し，判別問題に優れる識別的モデルのパラメータ学習法に類似している．その結果経験的に尤度に基づいたパラメータ学習に比べしばしば生成モデルの判別精度を向上させることが知られている．しかしながらこれまでの VT 学習は個別モデルに対して開発されており，PRISM のような汎用のプログラミング言語に基づく確率モデルに適用できる VT 学習法は存在しなかった．我々は PRISM の確率計算の中間データ構造である説明グラフ上で動作する VT 学習法を PRISM の意味論である分布意味論より数学的に導出し，幾つかの標準的モデルで良好に動作することを確認した．さらに VT のベイズ化を進め，VT と VB を融合した VTVB 学習の導出にも成功し，PRISM 処理系に組み込んで公開した.

[H25 年度]
一般に生成的確率モデルはロジステック回帰などの判別的確率モデルに比較し，識別問題における性能が劣ると言われている．H25 年度はこの問題を克服するため，PRISM のダイナミックプログラミングに基づく確率計算機能を活かしつつ，判別的確率モデルを記述し学習する D-PRISM を開発実装した.
D-PRISM は PRISM と同様に論理式によりモデルを記述するが，PRISM とは異なり，論理式に与える確率を実数の重みに一般化してあり，そのため正規化操作により確率分布を定義する．定義可能な判別的確率モデルは良く知られているロジステック回帰，条件付きマルコフ確率場などから最先端のマルコフ確率場文脈自由文法など広範囲に渡り，実験的に対応する生成的確率モデルに対して識別性能が向上することを確認した.
この D-PRISM および従来の PRISM により統計的機械学習の代表的モデルクラスである生成的確率モデルと判別的確率モデルを論理的に記述し，ダイナミックプログラミングを用いた確率の厳密計算により効率的にパラメータ学習することが可能になったと言える.
他方確率モデリングではマルコフ連鎖に於ける状態到達確率のような確率の無限和を計算する必要性が多々生じる．確率の無限和が計算出来ると先に述べたマルコフ連鎖を使った確率的遷移システムの検証や確率文法における接頭辞の確率計算，および接頭辞確率にもとづくプランニング認識など種々の応用が可能になる．従来確率の無限和を計算する試みは個別モデルに対して試みられて来たが，PRISM における説明グラフにループを許すことによりこのような確率の無限和を一般的に計算できることを証明した.
また実際に確率文脈自由文法の接頭辞の確率計算に適用した．更にインターネットの Web のログデータに接頭辞の確率計算を応用することにより，不完全な Web のログデータからもユーザの意図の推定が可能になることを実験的に示した.

以上の継続的研究開発の結果 PRISM は統計関係学習（SRL）あるいは確率論理学習(PLL)と呼ばれる関係概念を基本に据えた不確定性情報処理の研究の世界的研究開発競争に於いて，先駆者であると同時に先端的機能を提供する先導的モデリング言語としての位置を保ち続けている.

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）

〔雑誌論文〕（計 4 件）

１．小島諒介，佐藤泰介: アクセスログ分における接頭部分列からのプラン認識．人工知能学会論文誌，Vol.29，No.3，pp.301－310，2014． 査読有

2．Sato, T. and Kubota, K.: Viterbi training in PRISM. Theory and Practice of Logic Programming, FirstView Article / Jan. 2014, pp 1 - 22 DOI:10.1017/

S1471068413000677. 査読有

3. 石畠正和, 佐藤泰介: 命題化確率計算に基づく MCMC ベイズ推定. 人工知能学会論文誌, Vol.28, No.2, pp.230-242, 2013. 査読有

4. Sato, T. and Meyer, P.: Infinite probability computation by cyclic explanation graphs. Theory and Practice of Logic Programming, FirstView Article / Nov. 2013, pp 1 - 29 DOI:10.1017/S1471068413000562. 査読有

〔学会発表〕(計 5 件)

1. Sato, T., Kubota, K. and Kameya, Y.: Logic-based Approach to Generatively Defined Discriminative Modeling, Proceedings of the 23rd International Conference on Inductive Logic Programming(ILP 2013), Rio de Janeiro, Brazil, Aug. 29, 2013.

2. Kameya, Y. and Sato, T.: RP-growth: Top-k mining of relevant patterns with minimum support raising. Proceedings of the 2012 SIAM International Conference on Data Mining (SDM-2012), pp.816-827, Anaheim, California, USA, Apr. 26, 2012.

3. Sato, T. and Meyer, P.: Tabling for infinite probability computation. The 28th International Conference on Logic Programming, (ICLP-2012), Technical Communications, Budapest, Hungary, Sept. 7, 2012.

4. Ishihata, M. and Sato, T.: Bayesian inference for statistical abduction using Markov chain Monte Carlo. Proceedings of the 3rd Asian Conference on Machine Learning (ACML-2011), JMLR Workshop and Conference Proceedings, Vol.20, pp.81-96, Taoyuan, Taiwan, Nov. 14, 2011.

5. Ishihata, M., Sato, T. and Minato, S.: Compiling Bayesian Networks for Parameter Learning based on Shared BDDs. Proceedings of the 24th Australasian Joint Conference on Artificial Intelligence (AI-2011), LNAI 7106, Springer, pp.203-212, Western Australia, Australia, Dec. 8, 2011.

〔図書〕(計 0 件)

〔産業財産権〕
○出願状況(計 0 件)

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況(計 0 件)

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
佐藤 泰介 (SATO TAISUKE)
東京工業大学・大学院情報理工学研究科・教授
研究者番号：90272690

(2)研究分担者
( )

研究者番号：

(3)連携研究者
亀谷 由隆 (KAMEYA YOSHITAKA)
名城大学・理工学部情報工学科・准教授
研究者番号：60361789