様式 C-19

科学研究費補助金研究成果報告書

平成２３年　４月８日現在

機関番号：１２６０８
研究種目：基盤研究（Ｂ）
研究期間：２００８～２０１０
課題番号：２０３００００５３
研究課題名（和文）　確率知識モデルによる不確定性推論の研究
研究課題名（英文）　Uncertainty inference by probabilistic models

研究代表者
佐藤　泰介（TAISUKE SATO）
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

研究成果の概要（和文）：

現在データマイニングなど多量且つノイズを含むデータからの知識獲得のため、機械学習に於けるグラフィカルモデルや、統計的自然言語処理に於ける確率文法などの技術が各分野で使われているが、素性を利用するため関係をうまく表現できないなどさまざま制約があり、複雑な確率モデリングを困難にしている。我々は確率と論理を融合し、論理型言語に基づく確率モデリング言語 PRISM を開発した。　PRISM は確率モデルの記述と計算・学習を完全に分離することにより、論理の記述力を活かした効率的な確率モデリングを可能にしている。

研究成果の概要（英文）：

Currently techniques from machine learning and statistical natural language processing are popular in various fields including data-mining and bioinformatics. However they are feature-based and it is difficult to capture interdependent relationships in real data. We have developed a logic-based modeling language PRISM which unifies logical semantics and statistical parameter learning. It separates model description by logical formulas from their probability computation and parameter learning, thereby enabling an expressive yet efficient complex probabilistic modeling.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	3,300,000	990,000	4,290,000
２００９年度	3,300,000	990,000	4,290,000
２０１０年度	3,500,000	1,050,000	4,550,000
年度			
年度			
総　計	10,100,000	3,030,000	13,130,000

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：学習と知識獲得

１．研究開始当初の背景

近年バイオデータや Wikipedia などデータ・知識の大量集積が行われており、データマイニング等による有効利用が諸方面から期待されている。しかしながら同時にそれらは誤りのあるデータであったり、不完全な知識であったりすることも多い。このような不確定

性を含むデータに対する情報処理モデルとしては以前から人工知能におけるベイズネット、音声認識・バイオインフォマティクスにおける隠れマルコフモデル、統計的自然言語処理における確率文脈自由文法などが典型的なものとしてあるが、どれも個体や関係が表現できず論理的知識の処理能力を欠いていた。２０００年代初頭からこのような欠点を補い、統計的手法と論理概念を結合する事により C4．5 や SVM のような素性のみに頼る既存のデータマイニング手法を越える、関係に基づく不確定性情報処理を目指す動きが欧米を中心に高まって来た。代表例として主としてアメリカ西海岸で発展しつつある SRL（統計関係学習）と西ヨーロッパを中心に発展しつつある PLL （確率論理学習）がある。

SRL はベイズネットの研究から発展したものてある。ベイズネットが命題論理レベルの表現であるのに対し、論理変数や関係表現を採り入れる事により、より複雑な確率事象を容易に取り扱えるようにしている。但し確率推論や学習はベイズネットのアルゴリズムを流用している。SRL の枠組ですでに関係データベースへの確的的拡張や Web データのリンク解析などが行われている。一方 PLL は帰納推論をコンピュータ上に実現しようとする帰納論理プログラミングの研究に確率を採り入れる形で発展して来たものである。論理的表現に強い一方、統計的処理は未発達であり、ベイズネットの統計処理と組み合わせる方式が提唱されている。

我々はこれまで SRL や PLL の動きと同調しつつ、述語論理による知識表現と最尤推定によるパラメータ学習を分布意味論と呼ばれる確率的プログラム意味論の下に統合した確率知識モデル用言語 PRISM を独自に開発して来た。PRISM は第一階述語論理に基づく確率的プログラムにより確率モデルを記述、実行、学習するが、Turing 完全な計算能力と観測データからのパラメータ学習能力を備えているという意味で世界最初の学習能力を持ったプログラミング言語である。PRISM は上記の３大モデル（ベイズネット、隠れマルコフモデル、確率文脈自由文法）だけでなく、確率文脈自由グラフ文法などこれまで取扱われなかった確率モデルも統一して扱う事ができる汎用の確率計算機構と EM アルゴリズムに基づく汎用のパラメータ学習機構を備えており、その記述の柔軟性と計算の効率性が評価されて音楽やバイオインフォマティクスなど諸方面に適用されている。

PRISM は汎用の確率知識モデリング用言語であるが、一方でデータの確率的生成過程を Horn 節と呼ばれる含意の形をした論理式から成るプログラムで記述することを要請している。そのため定義可能な分布のクラスがいわゆる生成的（generative）なモデルに限定されており、またプログラムにも種々の数学的条件が課されている。更にプログラムが書けないと分布が定義出来ないことから、例えばバイオインフォマティクスにおける遺伝子ネットワーク推定問題のように、対象データの知識が不完全な状況では、プログラムを書けず何も推論出来ない場合があるという問題があった。

２．研究の目的

本研究では柔軟で高度の確率情報処理を実現するため、上記の PRISM の問題点に対処しつつ PRISM の計算能力および学習能力を強化する。

３．研究の方法

具体的には以下の研究を行う。
PRISM では背反性条件と呼ばれる条件をプログラムに課している。背反性条件はプログラムの実行過程が複数ある場合、それらが確率的に背反である事を要請している。この条件は PRISM のモデル記述における一番きつい制約となっており、廃止によりモデリングの自由度を飛躍的に上げる事が出来る。一方背反性条件は確率計算における P(A ∨ B) = P(A) + P(B)の成立を保証し、効率的確率計算を可能にしている条件である。従ってその廃止は意味論的には問題ないものの、計算効率の低下を招く恐れがある。我々はここに BDD（２分決定グラフ）を使って背反性を回復しつつ効率的な確率計算の実現を目指す。BDD による確率計算、パラメータ学習はこれまで世界的に見ても余り手が付けられていない未知の領域であり、BDD への encoding 法、確率計算のための BDD の構造化、パラメータ学習アルゴリズムの導出など数多くの課題がある。また PRISM には統計的モデリングに不可欠のパラメータ学習の機能があるが、学習の際多量のデータに対して、全解探索を行うため、メモリが不足がちであった。この問題解決するため、近似的な確率計算や近似的なパラメータ学習を開発実装する。

４．研究成果

2008 年度は PRISM プログラムに課されている背反性条件を除くため、BDD による確率計算および統計パラメータの学習のアルゴリズムの開発を行った。BDD はブール式をコンパクトに表現する有向グラフである。我々は確率的命題変数からなる任意のブール式を表

すBDDに対し、隠れマルコフモデルの確率計算で使われる内側確率および外側確率の概念を適用することにより、ダイナミックプログラミングに基づいて効率的に確率計算を行う一般化内側・外側確率アルゴリズムを導出した。さらにこのアルゴリズムを使い、BDDから確率的命題変数に付随する確率を推定するBDD-EMアルゴリズムと呼ばれるEM学習アルゴリズムを導出した。BDDとEM学習アルゴリズムを結びつけた研究は今までになく、BDD が広く使われている現状からして、BDD-EM アルゴリズムは広い応用が期待される。また、PRISM による統計的アブダクション（統計的仮説推論）に使われる論理式を確定節から一般の節に拡大するため、独立の確率的命題変数からなる一般の節を考え、そのような節の集合を条件部に持つ条件付き分布についてその性質を調べた。その結果、このような条件付き分布により任意の離散結合分布が表現できるとこと、およびPRISMプログラムにより表現される（無限の）分布もある数学的条件のもとに表現できることを証明した。

2009年度では前年度末に導出したBDD-EMアルゴリズムをバイオインフォマティクスの一分野である代謝経路のモデリングに適用し、有効性を確認した。BDD-EMアルゴリズムは本研究プロジェクトで開発している確率知識モデリング言語であるPRISMに於ける背反性制約を除くための中核的アルゴリズムであり、任意の率命題論理で記述された確率モデルのパラメータを学習できる。具体的には大腸菌の代謝経路をアブダクション（仮説推論）により実験データから論理的に推論し、可能な代謝経路（仮説）を得た後、それらを確率的命題論理により記述しBDD-EMアルゴリズムに与えることにより、代謝経路の各ステップを表す命題が真である確率を推論した。この確率にもとづき、候補となる代謝経路の確率を計算して最尤のものから順位付けを行ったところ、専門家の目から見て妥当な順に付けを得ることができた。この結果は将来バイオインフォマティクスなどの科学実験を行う際、半自動的に仮説を発見し評価する科学発見の自動化に繋がるものである。

PRISM の学習機能強化に関しては、統計的推論で重要な機能を果たすベイズ推論を取り入れる枠組みとして変分ベイズ法を採用し、PRISM の実行メカニズムである命題化計算と組み合わせることにより汎用の変分ベイズ法を実現できることを証明し、PRISM に実装した。このような汎用の変分ベイズ法はこれまでなく、変分ベイズ法の適用範囲を広げるものである。また変分ベイズ法により確率モデリングで重要なモデルの構造探索於けるモデル選択基準である周辺尤度を効率良く計算できるので、PRISM による確率モデルの構造探索が可能になった。

2010年度ではBDD-EMアルゴリズムを更に拡張したshared BDD-EMアルゴリズムを導出し実装した。shared BDD-EM アルゴリズムは、複数のBDDを構造化したshared-BDDを使い、従来BDDを使った確率計算で仮定されていた単一データに対する確率計算を、統計的学習では必須の数百、数千の iid（独立同分布）データに対する確率計算に拡張したもので、複数のデータに対する確率計算の共有を行いつつ、パラメータ学習を行うものである。計算機実験の結果、その有効性が確認された。また論理に基づいた確率モデリングにおいては確率モデルは命題論理式に変換されるが、その際命題変数を順序付けした上でorder-encoding と呼ばれる手法を使うことにより、より効率的に変換できることを理論的且つ実験的に確認した。一方PRISMの推論機能強化に関しては、統計的推論で重要な機能を果たすViterbi推論の実験的実装を行った。Viterbi 推論は最尤の解を与える統計パラメータを推論するが、EMアルゴリズムに比べて、計算が簡単になり、より多量のデータを扱えるようになるという特徴がある。我々は PCFG などの確率文法に於いてこのようにして得た統計パラメータが EM アルゴリズムより得た統計パラメータより高い精度で正解構文木を与えることを初めて実験的に確認した。

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１２件）

1. Sato, T., Ishihata, M. and Inoue, K., Constraint-based probabilistic modeling for statistical abduction, Machine Learning, 2010, 査読有

2. 佐藤泰介, 湊真一, ベイジアンネットワークと離散構造処理系, 人工知能学会誌, Vol.25, No.6, pp.796-802, 2010, 査読無

3. Sneyers, J., Meert, W., Vennekens, J., Kameya, Y. and Sato, T., CHR(PRISM)-based probabilistic logic learning, Theory and Practiace of Logic Programming, Vol.10, pp.433-447, 2010, 査読有

4. 佐藤泰介,

統計的アブダクション,
人工知能学会誌, Vol.25, No.3, pp.400-407, 2010, 査読無

5. 石畠正和, 亀谷由隆, 佐藤泰介, 湊真一,
BDD 上の命題化計算に基づく EM アルゴリズム,
人工知能学会論文誌, Vol.25, No.3, pp.475-484, 2010, 査読有

6. Sneyers, J., Meert, W., Vennekens, J., Kameya, Y. and Sato, T.,
CHR(PRISM)-based probabilistic logic learning,
Theory and Practice of Logic Programming, Vol.10, No.4-6, pp.433-447, 2010, 査読有

7. Sato, T., Kameya, Y., Kurihara, K.,
Variational Bayes via propositionalized probability computation in PRISM,
Annals of Mathematics and Artificial Intelligence, Vol.54, No.1-3, pp.135-158, 2009, 査読有

8. 佐藤泰介,
記号的統計モデリングの世界を探る,
コンピュータソフトウェア, Vol.25, No.3, pp.33-36, 2008, 査読無

9. Sato, T.,
A glimpse of symbolic-statistical modeling by PRISM,
Journal of Intelligent Information Systems, Vol.31, No.2, pp.161-176, 2008, 査読有

10. Sato, T. and Kameya, Y.,
New advances in logic-based probabilistic modeling by PRISM,
In Probabilistic Inductive Logic Programming, LNCS 4911, Springer, pp.118-155, 2008, 査読有

11. Zhou, N.-F., Sato, T. and Shen, Y.-D.,
Linear tabling strategies and optimization,
Theory and Practice of Logic Programming, Vol.8, No.1, pp.81-109, 2008, 査読有

〔学会発表〕(計6件)
1. Ishihata M.(Sato, T.),
An EM algorithm on BDDs with order encoding for logic-based probabilistic models,
ACML-2010, 2010/11/09, Tokyo.

2. Kameya, Y.(Sato, T.),
A Bayesian hybrid approach to unsupervised time series discretization,
TAAI-2010, 2010/11/18, Hsinchu.

3. Zhou, N.-F. (Sato, T.),
Mode-directed tabling for dynamic programming, machine learning, and constraint solving,
ICTAI-2010, 2010/11/19, Hsinchu.

4. Inoue, K.(Sato, T.),
Evaluating abductive hypotheses using an EM algorithm on BDDs,
IJCAI-2009, 2009/07/15, Pasadena.

5. Sato, T.,
Generative modeling by PRISM,
ICLP-2009, 2009/07/14, Pasadena.

6. Sato, T.,
Logic-based probabilistic modeling,
WoLLIC-2009, 2009/07/21, Tokyo.

〔その他〕
ホームページ等
http://sato-www.cs.titech.ac.jp

６．研究組織
(1)研究代表者
佐藤　泰介 (TAISUKE SATO)
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
亀谷　由隆 (YOSHITAKA KAMEYA)
東京工業大学・大学院情報理工学研究科・助教
研究者番号：６０３６１７８９