様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　２０　日現在

機関番号：　１４４０１
研究種目：　基盤研究(C)（一般）
研究期間：　2012～2015
課題番号：　２４５００１７２

研究課題名（和文）ベイジアンネットワークの構造学習で、離散と連続の属性が混在する場合

研究課題名（英文）Bayesian network structure learning when discrete and continuous variables are present.

研究代表者

鈴木　譲（Suzuki, Joe）

大阪大学・理学（系）研究科（研究院）・准教授

研究者番号：５０２１６３９７

交付決定額（研究期間全体）：（直接経費）　　3,900,000円

研究成果の概要（和文）：離散と連続の変数が混在している場合の、ベイジアンネットワークの構造学習について、検討した。従来は、連続変数は正規分布であるとか、離散変数の間に連続変数がおかれてはならないという制約があった。本研究では、そうした制約をおかず、一致性といって、サンプル数が大きくなった場合に、正しい構造を推定するという有効な性質を証明することに成功している。特に、相互情報量の推定や、独立性の検定などで、離散や連続という敷居をおかずに処理できる方式を提案することができた。そのような応用は、今後も数多く出てくるものと思われる。

研究成果の概要（英文）：We consider Bayesian network structure learning when discrete and continuous variables are present. The problem is rather hard and very few results are available. I particular, we had to assume that each continuous variable is Gaussian and no two discrete variable should be between a continuous variable. In this research, we mathematically prove consistency (the correct structure is estimated as the sample size increases). In particular, we proposed applications to independence testing and estimation of mutual information.

研究分野：　機械学習　情報理論

キーワード：　ベイジアンネットワーク　構造学習　事後確率最大　相互情報量の検定　独立性の検定

１．研究開始当初の背景

近年、データマイニングやパターン認識などの分野で、確率的知識の学習の研究がすすめられている。しかしながら、ベイジアンネットワークなどのグラフィカルモデルを仮定した学習は、数学的に難解であるばかりでなく、変数の個数に関して指数的に計算量が増えること、連続データからの学習に関してアプローチが確立していないことなどが、データサイエンスの現場への普及を妨げていると言われていた。

２．研究の目的

本研究では、実際例から、ベイジアンネットワーク(BN)の構造を学習する問題を検討した。BN は、確率変数に対応する属性を頂点で、その間の確率的因果関係を有向辺で表現した非巡回有向グラフとして、定義される。

従来は、BN に含まれる属性のすべてが離散、またはすべてが連続という非現実的な仮定のもとで検討されてきた。本研究では、ベイズ的に予測確率を計算して、事後確率最大の構造を求めるという、基本原理は変えずに離散や連続を仮定しない一般的な BN の構造学習の方法を確立することが目的となる。

既存の方法には、連続変数としてガウス分布を仮定すること、ナイーブベイズ的な方法を利用するために、離散変数の頂点の間に連続変数の頂点をおかないこと、さらには理論的な保証が無い、といった問題点があった。

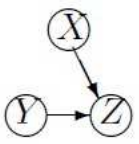
３．研究の方法

アルゴリズムを提案する、性能を保証するために有利な性質を数学的に証明する、実データを用いて動作を確認し、有効性を実証する。また、従来研究を調査すること、成果の各段階で議論することなども含まれる。

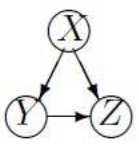
４．研究成果

4.1 離散や連続を仮定しない予測確率の提案と、一致性の証明

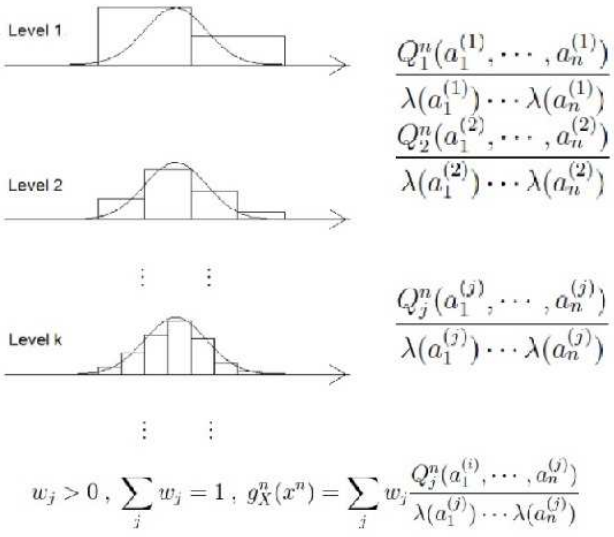
構造学習は、局所スコアの積が高くなるように分布を因子の積に分解することと同値になる。たとえば、下図左では分子の 3 個の因子の局所スコアを分母の因子の局所スコアで割ったものが、極大スコアとなる。このスコアに構造の事前確率をかけたものを構造どうし比較する。下図右も同じ事前確率を持てば、P(X)P(Y)と P(X,Y)に対応するスコアを計算して、どちらの構造の事後確率が高いかが決まる。



$$\frac{P(X)P(Y)P(XYZ)}{P(XY)}$$



$$P(XYZ)$$



$$\frac{Q_1^n(a_1^{(1)},\cdots,a_n^{(1)})}{\lambda(a_1^{(1)})\cdots\lambda(a_n^{(1)})}$$

$$\frac{Q_2^n(a_1^{(2)},\cdots,a_n^{(2)})}{\lambda(a_1^{(2)})\cdots\lambda(a_n^{(2)})}$$

$$\frac{Q_j^n(a_1^{(j)},\cdots,a_n^{(j)})}{\lambda(a_1^{(j)})\cdots\lambda(a_n^{(j)})}$$

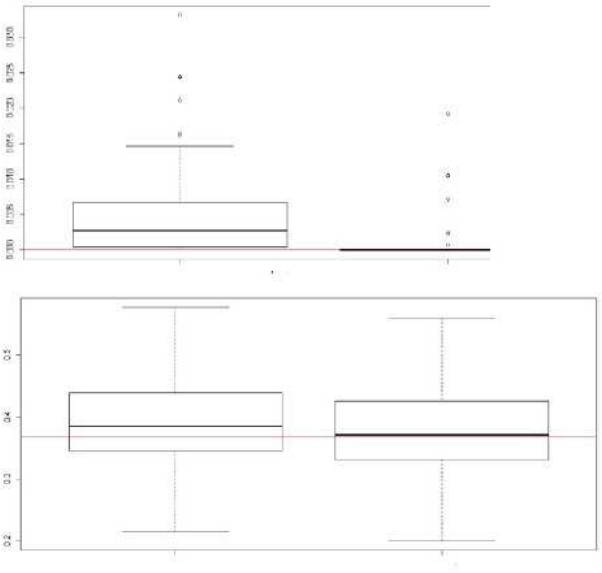
$$w_j>0\,,\ \sum_j w_j=1\,,\ g_X^n(x^n)=\sum_j w_j\frac{Q_j^n(a_1^{(i)},\cdots,a_n^{(j)})}{\lambda(a_1^{(j)})\cdots\lambda(a_n^{(j)})}$$

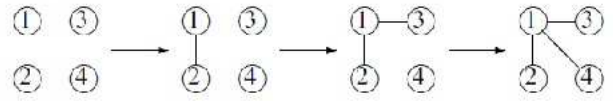
したがって、局所スコアの計算を離散のものから連続でも使えるように一般化することが、当初からの方針であった。本研究では、連続量を離散化したヒストグラムを複数構築し、それらのスコアを幅(ルベーグ測度とよばれる)で割ったものに重み付けした値をその因子の密度関数の局所スコアとした。

その結果、一定の正規条件のもとで、サンプル数が増加するとともに正しい構造を学習する性質(一致性)を数学的に証明することができた。

4.2 離散や連続を仮定しない独立性検定と、性能の保証

4.1 の局所スコアの計算方法は、種々の問題に応用できる。たとえば、相互情報量を推定して、その値が零であれば独立、正であれば独立ではないという方法を提案した。既存の相互情報量の推定は必ず正の値をとるので、そのような独立性の検定に適用することができない。本研究で提案した相互情報量は、そのような検定に有効であることがわかった(下図)。





4.3 Chow-Liu アルゴリズムと欠損値を含む場合の構造の学習

Chow-Liu アルゴリズムといって、相互情報量の大きい 2 変数の対から辺を結んで、最後に森を生成するアルゴリズムがある(上図)。本研究では、4.2 で提案した相互情報量の推定量を用いて、この方法を実現する(離散も連続も区別しない)ことを提案した。また、その方式が、欠損値が存在する場合でも事後確率が最大になることを証明した。

4.4 分岐限定法

ベイジアンネットワークの構造学習では計算量が膨大になる。本研究では記述長最小の意味では保証しながら、計算量を削減する方式を以前に提案している。今回は、記述超最小ではなく、事後確率最大の方式を提案した。

4.5 線形回帰モデルにおける Hannan-Quinn の命題の証明

情報量基準で、BIC は一致性があり、AIC だと一致性がないということは、よく知られている。情報量基準は、データがモデルにどれだけ適合するかということのスコアと、モデルの複雑さに関するスコアの和のスコアを最小にするもので、AIC,BIC はその重み付けだけが異なっている。Hannan-Quinn といって、一致性を満足する境界の重み付けを利用するものであり、その適用は従来自己回帰、分類に限られていた。本研究では、そのことが線形回帰にも適用可能であることを証明した。

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）

Joe Suzuki, "An Estimator of Mutual Information and its Application to Independence Testing", Entropy 2016, 18(4), 109; doi:10.3390/e18040109 (2016 年 3 月)

Joe Suzuki, "Consistency of Learning Bayesian Network Structures with Continuous Variables: An Information Theoretic Approach", Entropy 2015, 17(8): 5752-5770, doi:10.3390/e17085752 (2015 年 8 月)

Takanori Ayano and Joe Suzuki, "On d-Asymptotics for High-Dimensional Discriminant Analysis with Different Variance-Covariance Matrices", IEICE TRANSACTIONS on Information and Systems, E95-D(12): 3106-3108 (2012 年 12 月)

Joe Suzuki, "The Hannan-Quinn Proposition for Linear Regression", International Journal of Statistics and Probability, 1(2),doi: 10.5539/ijsp.v1n2p179 (2012 年 11 月)

〔学会発表〕（計 34 件）

Joe Suzuki, "Structure Learning of Bayesian Networks with p Nodes from n Samples when n<<p"（招待講演), Workshop on Probabilistic Graphical Models, 統計数理研究所 (2016 年 3 月)

鈴木譲「確率的グラフィカルモデルにおける構造学習」(招待講演), 数学協働プログラム, 電気通信大学 (2015 年 3 月)

Joe Suzuki, "The MDL principle for arbitrary data: either discrete or continuous or none of them"(招 待 講 演), The Sixth Workshop on Information Theoretic Methods in Science and Engineering (WITMSE), 東京大学 (2013 年 8 月)

Joe Suzuki, "The Chow-Liu algorithm based on the MDL when discrete and continuous variables are present", The fourth Workshop on Algorithmic issue for Inference in Graphical Models, Paris (2014 年 9 月)

鈴木譲「Cover 先生の研究室で思い出に残っているテーマ: 定常エルゴードな系列に対してのユニバーサルな予測」(招待講演), 電子情報通信学会情報理論研究会, 別府(2012 年 12 月)

〔図書〕（計 2 件）

Joe Suzuki and Maomi Ueno (Editors): The Second Workshop on Advanced Methodologies for Bayesian Networks. Springer Lecture Notes on Artificial Intelligence Volume 9505 (2015 年 11 月)

植野、鈴木他「確率的グラフィカルモデル」(2016 年 8 月)

〔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
鈴木 譲（Suzuki, Joe　　）
大阪大学・大学院理学研究科・准教授
研究者番号：50216397

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
鷲尾 隆　(Washio, Takashi)
大阪大学・産業科学研究所・教授
研究者番号： 00192815

狩野 裕　(Kano, Yutaka)
大阪大学・大学院基礎工学研究科・教授
研究者番号：20201436