別紙 1

論文の内容の要旨

論文題目	Exact Learning of Augmented Naive Bayes Classifiers （Augmented Naive Bayes 分類器の厳密学習）
学位申請者	菅原　聖太

ベイジアンネットワークは，離散確率変数をノードとし，ノード間の条件付き従属関係を非循環有向グラフ（Directed Acyclic Graph: DAG）で表し，同時確率分布を各ノードの親ノード集合を所与とした条件付き確率パラメータの積に分解する，確率的グラフィカルモデルである．ベイジアンネットワークにおける一つのノードを目的変数とし，その他のノードを説明変数としたベイジアンネットワーク分類器（Bayesian Network Classifier: BNC）は，離散変数を扱う分類器として知られている．

一般にベイジアンネットワークのDAG構造はデータから推定する必要があり，この問題をベイジアンネットワークの構造学習と呼ぶ．構造学習では，候補構造から最適な学習スコアを持つ構造を探索するスコアベースアプローチが従来から行われてきた．本論では候補構造に制約を課さずに学習したBNCを General Bayesian Network（GBN）と呼ぶ．一般にスコアベースアプローチでは，漸近一致性を有する，構造の周辺尤度（Marginal Likelihood: ML）を学習スコアとして用いる．MLを用いると，全変数の同時確率分布をモデル化する生成モデルとしてBNCを学習できる．

従来研究では，BNCの構造学習スコアとして，生成モデルではなく，説明変数を所与とした目的変数の条件付き確率分布をモデル化する識別モデルのためのスコアを用いるべきだと主張されてきた．そのような学習スコアとして，説明変数を所与とした目的変数の条件付き対数尤度（Conditional Log Likelihood: CLL）が提案された．しかし，CLLを最大にするパラメータ推定式は閉形式で表せないため，構造の探索に効率的なアルゴリズムを用いることができず，学習時間が膨大になってしまう．これを解決するため，構造探索も効率にできるようCLLを線形近似したapproximated CLL（aCLL）が提案されている．また，CLLをスコアとして，貪欲法のHill-Climbingアルゴリズムを用いて構造を探索する手法も提案されている．これらの近似手法で学習したBNCの方が，MLで学習したBNCよりも分類精度が高いことが報告されている．

しかし，ML最大化よりCLL最大化の方がなぜ良いかという理由については未だ明らかにされていない．MLは推定構造に対して漸近一致性が保証されており，サンプルサイズが大きい時にMLの分類精度がCLLに劣るのは奇異である．また，BNCのMLは閉形式で表せるためCLLより計算効率がよく，MLを大域的に最大化する構造を探索する厳密学習を効率的に行える．先行研究の比較実験では，MLを局所的に最大化する構造を探索する近似学習を行なっているため，探索精度の悪さが影響したのかもしれない．

そこで，本研究ではまずMLによる厳密学習とCLLによる近似学習によって得られたBNCの分類精度を比較する．結果として，ML最大化によるBNCは厳密 学習することで，大きく精度が向上することがわかった．特にサンプルサイズが大きいときに，最も分類精度が高いことが示された．しかし，厳密学習ではサンプルサイズが小さくなるとMLを最大化するBNCの分類精度が低くなり，最も単純な構造をもつNaive Bayesよりも低い場合もあった．特に，目的変数の親変数が多く子変数が少ないような構造を学習する場合に分類精度が低くなっていることがわかった．その理由は，目的変数の親変数が多いと，パラメータ数が指数的に増えるため，一つのパラメータ学習のためのサンプルサイズが小さくなり，推定精度が悪くなってしまうからである．

この問題を緩和するため，本論では，目的変数が親変数を持たず，説明変数が必ず目的変数の子となるAugmented Naive Bayes（ANB）構造を制約としたBNCの厳密学習を提案する．ANBはこれまで識別モデルとして扱われてきたため，MLを最大化して学習することはなかった．本論の提案は，識別モデルの学習ではなく，生成モデルとしてのGBNの学習に，目的変数の親変数が増えないようにANB構造を制約とすることである．また，本論では，全説明変数が分類に影響を及ぼし，全説明変数が目的変数と隣接しているという仮定のもとで，厳密学習したANBは漸近的に真の構造と全く同じ分類確率を表現することを証明する．しかし，全説明変数が分類に影響を及ぼすという仮定は一般には成り立たない．この仮定を不要にするため，提案手法では構造学習の前に変数選択を適用する．真に分類に影響を及ぼす変数の選択を漸近的に保証するような従来手法は計算量が大きい．そこで本論ではベイズファクターを用いることで，真に分類に影響を及ぼす変数の選択を漸近的に保証する変数選択手法を提案する．また，数値実験により，この手法で変数選択をした後に厳密学習したANBの優位性を示す．

厳密学習は変数の増加に対して計算量が指数的に増加するため，提案手法では数十変数の学習が限界である．そこで，本論文では大規模変数をもつBNCを学習できる手法を提案する．因果モデルの研究分野では，条件付き独立性検定（CIテスト）とエッジの方向付けによる計算効率の高い構造学習法が提案されており，制約ベースアプローチと呼ぶ．従来研究では，CIテストにベイズファクターを用いることで真の構造への漸近一致性を有しつつ1000変数以上の構造学習を実現している．本論では，大規模なANBを学習するため，CIテストにベイズファクターを用いた制約ベースアプローチ手法を提案する．また，提案手法がANB構造について漸近的にパラメータ数を最小にして真の同時確率分布を表現することを示す．実験により，大規模構造学習において提案手法が有用であることを示す．

別紙２

論文審査の結果の要旨

学位申請者氏名　　菅原　聖太

審査委員主査　　植野　真臣

委員　　大濱　靖匡

委員　　岡本　吉央

委員　　村松　正和

委員　　八木　秀樹

委員　　印＊

委員　　印＊

（＊自筆署名の場合に限り、押印省略可）

本論文では，予測精度の高いベイジアンネットワーク分類器および大規模ベイジアンネットワーク分類器の構造学習の研究に取り組んでいる．ベイジアンネットワーク分類器に対しては，小サンプルサイズにおいて，目的変数の親変数数の増加によって分類精度が著しく低下する問題がある．本研究では，目的変数から説明変数に強制的にエッジを引いたAugmented Naive Bayes制約を課して構造を厳密学習することで，小サンプルサイズでも高精度な分類手法を提案している．大規模構造学習に対しては，変数数の増加による計算時間の指数関数的増加が問題点として残されている．本研究では制約ベースアプローチにベイズファクターを適用することで，漸近的に真の構造の学習を保証しつつ，1000変数以上の分類器の学習が可能な手法を提案している．

本論文は全6章より構成されており，第3章，第4章，第5章に学位申請者が行った研究が述べられている．以下，本論文の構成に沿って，各章の内容を要約する．

第1章で，本研究の目的，先行研究，得られた研究成果を要約している．

第2章はベイジアンネットワークに関する性質や定理を紹介している．1節では，まず，ベイジアンネットワークの定義を述べている．次に，ベイジアンネットワークにおける条件付き独立性を表現する有向分離を定義し，重要な性質であるI-mapの定義を行っている．続いて，ベイジアンネットワークのパラメータ推定について述べている．パラメータの事前分布として共役事前分布であるディリクレ分布を仮定し，パラメータの期待事後確率推定量を閉形式で表せることを紹介している．次に，構造学習法の一つであるスコアベースアプローチを紹介している．スコアの有用な性質として漸近一致性や分解可能性を定義し，その二つの性質をもつBDeu(Bayesian Dirichlet equivalent uniform)スコアを紹介している．2節では，ベイジアンネットワーク分類器の先行研究について述べている．BDeuよりも条件付き対数尤度(Conditional log likelihood: CLL)スコアを用い

た方が分類精度の高い構造が学習できると報告している先行研究や，CLLに関する先行研究を紹介している．

第3章では，BDeuを用いた厳密学習とCLLを用いた近似学習の分類精度比較実験を行っている．サンプルサイズが大きいときはBDeuの厳密学習はCLLの近似学習と比較して必ずしも低いわけではないという結果を報告している．一方で，サンプルサイズが小さい時において，BDeuで厳密学習した構造では目的変数の親変数数が過多になり，分類精度が著しく低下するという結果を報告している．目的変数の親変数数が増えると，パラメータ数が指数的に増えるため，一つのパラメータ学習のためのサンプルサイズが小さくなり，推定精度が悪くなってしまうことを原因として考察している．

第4章では，上記の分類精度低下の問題を解決するため，目的変数から説明変数へのエッジが引かれたAugmented Naive Bayes分類器の厳密学習を提案している．続いて，提案手法のアルゴリズムとして，BDeuの分解可能性を利用した動的計画法を提案し，その詳細を記述している．次に，提案手法がもつ有用な性質を二つ証明している．一つ目は，提案手法で学習した構造が漸近的にパラメータ数最小のI-map ANBに概収束することである．二つ目は，目的変数が全ての説明変数と隣接しているという仮定のもとで，提案手法で学習した構造が真の構造と分類等価な構造に概収束するという性質である．これらの性質によって，提案手法が真の同時確率分布を漸近的に推定できることを述べている．提案手法の二つ目の性質は，全説明変数が目的変数のマルコフブランケットという仮定をしているが，この仮定が不自然であるため，変数選択手法を提案している．実験により，変数選択手法を適用したANBの厳密学習が従来手法よりも統計的有意に高い分類精度を持つことを報告している．

第5章では，大規模ANBを学習できるアルゴリズムを提案している．まず，第4章まで述べていたスコアベースアプローチでは最先端手法を用いても60変数程度が限界であることを述べている．次に，大規模構造学習を可能にする手法として，完全グラフから条件付き独立性検定（conditional independence test: CIテスト）によってエッジ削除して構造を学習する制約ベースアプローチを紹介している．続いて，制約ベースアプローチの先行研究では，CIテストに統計検定や条件付き相互情報量を用いているため，漸近的に真の独立性を検出する保証がないという問題点を述べている．本研究では，漸近的に真の独立性を検出できるベイズファクターを制約ベースアプローチのCIテストに用いることで，構造推定の一致性を保証しつつ大規模なANBを学習できる手法を提案している．実験により，提案手法は従来手法より有意に分類精度が高いことを示し，1000変数以上の構造を学習できることを報告している．

第6章では，本研究で得られた成果の要約と今後の研究の展望を述べている．

本論文は，ベイジアンネットワーク分類器をデータ生成モデルから構成する手法とその厳密解を求める手法を提案し、従来手法より分類精度向上とアルゴリズムの高速化に関する成果を示したものであり，その内容は高く評価できる．よって，本論文は博士（工学）の学位論文として十分な価値を有するものと認める．