日本ソフトウェア科学会第 27 回大会 (2010 年度) 講演論文集

数値を含むデータの学習に対応したILPシステムの構築

鈴木 匠　大和田 勇人

機械学習の分野における数値データが含まれる問題に対して ILP(Inductive Logic Programming:帰納論理プログラミング) システムを用いた学習は困難であるとされている．それは，システム上で数値データが記号として扱われてしまうことが原因である．そこで本研究では，正事例のボトム節を用いた相対最小一般化に着目して論理式と数値範囲を示す不等式を算出することで，正事例を被覆する仮説の生成を行う．そして仮説生成において正事例を被覆しないリテラルを削除することで節長の短い仮説の生成を行う．また仮説生成時に負事例を被覆するかどうかを判定することで，正事例の学習において，より有益なルールを導くことができるように ILP システムを改良する．更にユーザにとって使いやすいインタフェースを持つシステムの構築を目指す．

1　はじめに

近年，機械学習の分野において，多量のデータから共通パターンを探し，規則を導くために ILP システム [1] がよく用いられている．しかし，ILP では一階述語論理表現を利用して仮説生成を行う時，正事例，負事例，背景知識が持つすべての述語の要素 (述語の引数) を定数 (記号) として扱うというアプローチを採っているため，数値を含むデータを処理する場合，数値も他のデータ同様に記号として扱われてしまう．数値を記号として扱える場合には問題にならないが，ある 2 つの要素が持つ数値の関係性を必要とする問題に対して ILP を適用することは適切ではない．

私たちは，[7] において制約論理プログラムを用いた不等式や点集合が示す領域 (凸包) を求めるプログラムを組み込むことで，正事例から得られる数値領域を算出し，仮説に追加できるように ILP システムを改良した．数値を得るために仮説生成において，初めに 2 つの正事例を別々に抽出してから仮説を生成し，残りの正事例集合から事例を 1 つずつ取り出して，数値範囲を拡張していったが，正事例をすべて被覆するように数値範囲の拡張を行ったため，仮説生成において過度の一般化が行われてしまった．また ILP システム Golem [2] における相対最小一般化 (Relational Least General Generalisation:RLGG) は初めに事例を説明する上で最低限必要な背景知識を羅列したボトム節を利用して，徐々に正事例集合を被覆するように構築された仮説を拡張していく．しかし，RLGG によって構築される仮説の節長は，被覆する正事例数により急激に増加してしまう問題がある．これらの問題は前研究においても考えられ，仮説生成に関して過度の一般化と仮説の節長に対する対策を考えないといけない．

そこで，本研究では溝口，大和田らによって開発された代表的な ILP システムの一つである GKS [5] に，正事例を被覆しないリテラルを削除することで節長の短い仮説を生成し，かつ数値範囲を求めるプログラムを仮説生成アルゴリズムに組み込むことで数値を含むデータからの学習において有益な仮説生成を行うことを目的とする．

ILP system for learning problem contained numerical data

Takumi Suzuki, 東京理科大学大学院 理工学研究科 経営工学専攻, Dept. of Industrial Administration, Graduate School of Science and Technology, Tokyo University of Science.

Hayato Ohwada, 東京理科大学 理工学部 経営工学科, Dept. of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science.

2 ILP(帰納論理プログラミング)

2.1 ILP

本研究の基盤となる ILP について説明を行う．一般的に現実問題に対して ILP を用いる時，常に正事例と負事例の両方が与えられるとは限らず，正事例のみと背景知識から学習を行わなければならない場合がある．例えば，数値を背景知識に持つ正事例のみ与えられた問題に対して，ILP を適用し学習を行うとする．しかし，既存の ILP システム GKS では一般的な仮説から特殊な仮説を生成するので，仮説を学習すると同時に正事例内の数値全てを含むような最小の領域を学習することが出来ないという問題がある．また，負事例を考慮に入れない場合，仮説において過度の一般化が行われてしまう問題がある．

2.2 制約論理プログラム

[7] では正事例のみからの学習を対象として，ILP システムに数値範囲を求めるプログラムを追加するため制約論理プログラムを用いた．制約論理プログラムは「宣言性」という性質を持つことから，プログラミング処理による情報の流れを固定せずに，解くべき問題の記述に専念できる．また制約を与えることで，探索空間の縮小を図ることが出来る．制約論理プログラムを用いた理由は，ILP システム GKS において数値を含むデータを扱う場合に正事例内の数値が示す範囲の算出・学習を行うことが出来ないからである．そこで，背景知識内に数値が含まれる場合，数値が示す最小領域を求めるために凸包という概念を利用した．

凸包とは，平面上に与えられたすべての点を含むように出力される最小の凸多角形のことで，必要でない情報を多く含む空間を探索するのではなく，欲しい情報を含む空間のみを探索出来るという利点を持つ．凸包を採用する理由は 2 つある．1 つ目は「入力データに数値を与えれば，数値が示す最小の領域を求めることが出来る」こと，2 つ目は負事例には過度の一般化を防ぐ役割があるが，現実問題では，多くの負事例を与えることはなかなか難しく，正事例だけから学習を行わなければならない場合もある．そこで凸包を用いることにより「正事例だけからの学習が可能になる」ことである．GKS に組み込むプログラムは，[6] で Florence らによって作成された線形制約を用いた凸包プログラムを採用した．

2.3 関連研究

トップダウン探索を用いることが ILP システムにおいて主流であったが，トップダウン探索で用いられる枚挙法は，仮説を生成し，次に仮説に対して正負事例の被覆率を検査する．そのため枚挙法は非効率である．また Golem における仮説生成で利用される RLGG は，事例数が増加するほど仮説の節長が長くなってしまう．これらの問題に対して，Muggleton らは仮説生成において包摂順序に基づく Plotkin の RLGG の改良として，初期のボトム節の節長により生成される仮説の節長が制限される非対称相対最小一般化 (Asymmetric Relative Minimal Generalisation:ARMGs) を提案している [3]．ARMGs では，正事例集合から事例を 1 つ抽出し，現在の仮説において事例が被覆されないリテラルが無くなるまで削除する．RLGG とは異なり正負事例の被覆率を調べた後に仮説の生成を行うため，節構築操作が非対称である．次に ARMGs により得られた仮説に対して，負事例を被覆しないリテラルを探索・削除することで，負事例が被覆されない仮説を生成している．最終的に Progol で用いられるボトム節構築と ARMGs を組み合わせることで ILP システム ProGolem を構築している．しかし，このシステムでは初期仮説に初めに抽出した正事例のボトム節を格納してから仮説生成を行うため，[7] のように数値範囲を導くことは出来ない．

3 提案手法

3.1 アルゴリズム

本研究では，数値を含むデータからの学習に対応できるように ILP システム GKS の改良を目的としている．背景知識の中で与えられるデータから数値を取り出し，それらが示す数値範囲を学習するには，特殊な仮説から一般的な仮説を生成していかなければならないが [7] のように相対最小一般化を行うだけでは初期のボトム節の節長が長い場合，得られる仮説も節長が長くなる可能性がある．また，すべての正事

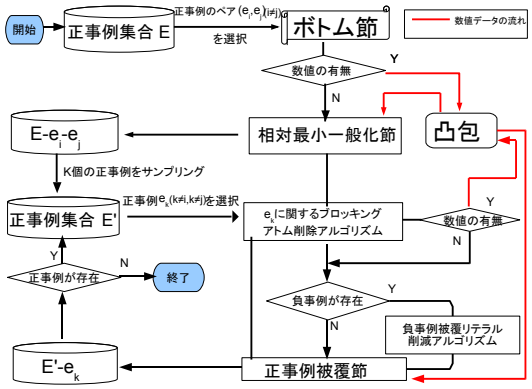

図 1　仮説生成アルゴリズム

例を被覆する一般的な仮説を生成していくため，仮説に対して過度の一般化が行われてしまう．そこで，本研究で構築したアルゴリズムを図 1 に示し，アルゴリズムの説明を行う．

1. n 個の正事例 (e_1,e_2, $\cdots$ e_n) を集めて，正事例集合 E を形成する．
2. 正事例集合 E からランダムに正事例の組み合わせ (e_i,e_j ただし i≠j) を取り出し，それぞれのボトム節 $\perp_e$ を生成する．ボトム節とは，1 つの正事例を説明するすべての無矛盾する仮説の中で最も弱い仮説のことである．
3. ボトム節を構成する背景知識に数値があれば，凸包プログラムに数値をパラメータとして渡し，数値範囲を学習する．
4. e_i,e_j のボトム節から相対最小一般化節を求めた後，正事例 (e_i,e_j) を正事例集合 E から取り除き，正事例集合 E を更新する．相対最小一般化節とは，ボトム節を構成する背景知識の引数を比較し，同じ値ならばそのまま残し，異なる値ならば定数を変数に置き換えることで求められる一般化節である．
5. 正事例集合 E から K 個の正事例をサンプリングし E′ とする．そして E′ から 1 つ正事例 (e_k ただし k≠i,k≠j) をランダムに取り出し，正事例 e_k のボトム節を生成する．
6. 4 で求めた相対最小一般化節のリテラルに対して 5 で取りだした正事例が被覆されるか調べ，正事例を被覆しないリテラル (ブロッキングアトム：第 3.2 節参照) を削除する．
7. e_k を説明するための背景知識に数値があれば，凸包プログラムに数値と前の数値範囲をパラメータとして渡し，新しい数値範囲を学習する．
8. 負事例が与えられている場合，負事例削減アルゴリズム (第 3.3 節参照) を実行する．
9. e_k を正事例集合 E′ から削除し，正事例集合を更新する．
10. K 個のサンプリングされた正事例集合 E′ が空になるまで，5~9 を繰り返し行い，正事例を被覆する最適な仮説と正事例が示す数値範囲を求める．

3.2　ブロッキングアトム

前研究では事例数が増加することで生成される仮説の節長は増加してしまう可能性があった．そこで，本研究では事例数が増加しても仮説の節長は初期生成のボトム節の節長で制限されつつ，正事例を被覆する仮説を生成するために，ブロッキングアトムを考慮に入れる．ブロッキングアトムに関しては [3] に定義されている仮説において正事例を被覆しないリテラルのことである．本アルゴリズムにおける流れを図 2 に示す．生成される正事例集合からランダムに K 個の正事例をサンプリングし，それらを対象として，現在の仮説を構成するリテラルが事例を被覆するかことが出来るかを調べる．もしブロッキングアトムが存在した時，仮説からブロッキングアトムが無くなるまで削除する．これにより，節長を短くしつつ，正事例を被覆する仮説の生成を行っている．

3.3　負事例削減アルゴリズム

現実問題から規則を導く場合，負事例を多く与えることは難しく，正事例のみからの学習が必要であることが多い．また，[7] では凸包により数値範囲を求めるためには正事例から数値を取り出さないといけない．しかし，数値範囲を導く代わりにすべての正事例を被覆するように仮説を構築していくため，得られた仮説は負事例を被覆する場合がある．そこで，本研究では負事例が少しでも与えられている場合，構築

入力：事例から得られた相対最小一般化節 $rlgg(e_i, e_j)$, 正事例 e_k
出力：正事例被覆節 $\overrightarrow{C}$

1. $\overrightarrow{C}$ には $rlgg(e_i, e_j)$ =h $\leftarrow$ b_1, ...,b_n を格納する．
2. $\overrightarrow{C}$ から e_k に関するブロッキングアトム b_i を削除する．
3. $\overrightarrow{C}$ からヘッド部やボディ部 $b_j(1< j <$i) に存在する変数を含まない原子文を削除する．
4. $\overrightarrow{C}$ のボディ部における e_k に関するブロッキングアトム b_i がなくなるまで，2 と 3 を繰り返し行い，正事例被覆節 $\overrightarrow{C}$ を求める．

図 2　仮説生成アルゴリズム

した仮説に対して，負事例を被覆しないように節を縮小するアルゴリズムを組み込む．本研究において，QG/GA [4] における負事例削減アルゴリズムを採用する．

4　実装

本研究で改良した ILP システム GKS は Prolog プログラム処理系として SICStus Prolog VC9 4.1.2 を用いている．また改良した GKS を java から呼び出し，実行することで，ユーザが java 上から仮説生成を行えるようにしている．SICStus Prolog，java(version 1.6.0_14)，MySQL Server 5.0，Apach Tomcat6.0 すべてを連携させることで実行結果から得られる仮説を MySQL データベースに格納し，ユーザが求めたい仮説をデータベースから取り出すことで，Tomcat により HTML 画面で表示することが出来るようになっている．各々の連携図は図 3 に示す．

5　結論と今後の展望

本研究は，ILP システム GKS を用いて数値を含むデータから有益のある仮説を生成するために，凸包プログラムを組み込むことで数値範囲の学習を行い，また正事例集合から正事例のペアをランダムに抽出し，それらの相対最小一般化節に対して残りの正事例が被覆されるかどうかを調べ削除することで，正事例を被覆しながら節長の短い仮説の生成を行う方法を組

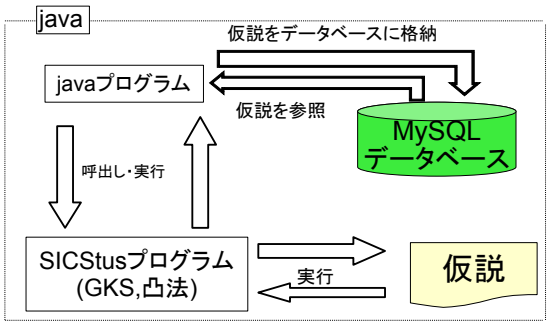

図 3　ソフトウェア間の連携図

み込むアルゴリズムを提案し，実装した．更に負事例が存在する場合，仮説が負事例を被覆しないように負事例削減アルゴリズムを組み込んだ．本研究で提案したアルゴリズムにより，数値を含むデータを扱う場合に節長が短くかつ数値範囲を含んだ有益な仮説を生成できると考えられる．現在のアルゴリズムでは数値生成を行うために正事例のペアをランダムに生成・抽出を行っているが数値範囲を生成する上で近い値を持つ数値データからサンプリングし，仮説生成が行えれば現在より容易に良い仮説を得ることが出来ると考えられる．そこで，扱う問題ごとに数値範囲は異なるため難しいかもしれないが，数値領域に対して近い値を持つデータのペアを探し出し，そこから仮説を生成する方法を考える必要がある．また実際にどのくらいの処理速度で仮説が生成できるか調べる必要がある．

また本研究では構築する ILP システムをユーザが容易に利用出来るように java や MySQL，Tomcat と連携している．実装しているシステムに関してまだユーザが使用するには程遠いため，仮説を生成・表示するだけでなくシステムのインタフェースについても考えなければならない．

参 考 文 献

[1] 古川，尾崎，植野：帰納論理プログラミング，共立出版,2001.
[2] Muggleton,S.,Feng,C.: Efficient induction of logic programs. In: Muggleton, S.(ed) Inductive Logic Programming, pp. 281-298.Academic Press, London,1992.
[3] Muggleton,S.,Santos,J.,and Tamaddoni-Nezhad,A.: ProGolem:A System Based on Relative Minimal Generalisation,ILP 2009,LNCS 5989, pp.131-148, 2010.
[4] Muggleton,S.H.,Tamaddoni-Nezhad,A.: QG/GA:A stochastic search for Progol. Machine Learning 70(2-3),123-133(2007),doi:10.1007/s10994-007-5029-3.
[5] Mizoguchi,F.,Ohwada,H: Constrained Relative Least General Generalization for Inducing Constraint Logic Programs. New Generation Computing,Vol.13,pp.335-368,1995.
[6] Benoy,F.,King,A.,Mesnard,F.: Computing convex hulls with a linear solver.TPLP 5(1-2),pp.259-271,2005.
[7] 鈴木 匠，松井 藤五郎，大和田 勇人: ILP における制約論理プログラムに基づいた数値を含むデータからの学習，第 23 回人工知能学会，高松，2009.