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

Machine learning from numerical data in ILP using constraint logic programs

鈴木 匠[*1] 松井 藤五郎[*2] 大和田 勇人[*2]
Takumi Suzuki Tohgorou Matsui Hayato Ohwada

[*1]東京理科大学 大学院 理工学研究科 経営工学専攻
Department of Industrial Administration, Graduate School of Science and Technology, Tokyo University of Science

[*2]同 理工学部 経営工学科
Department of Industrial Administration, Faculty of Science and Technology

In recent years, study uses ILP as the method of data mining.However, machine learning from numerical data using ILP is difficult, because numerical data is recognized symbolic data in ILP. Therefore, this paper proposes a method of learning both logical expression and range of numerical data by generalizing given positive examples and using a concept called "convex hulls". To compute convex hulls is to get the smallest field represented by numerical value. This method is to implement program which computed convex hulls using a linear solver the CLP(Q) library in ILP system "GKS". Accordingly, with this method learning from numerical data is possible under conditions, which are two attributes including numerical value and three different types of numerical value exist in given examples.

1. はじめに

近年、機械学習の分野において、複雑でより現実的な問題を扱うために、最も表現力に富んだ手法として、一階述語論理を知識表現言語とした帰納論理プログラミング (Inductive Logic Programming:ILP)[1] が用いられている。しかし、ILP では、一階述語論理表現を利用して仮説を生成する時、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とっているため、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ってしまうという問題がある。数値を記号として扱える場合には問題にならないが、ある要素とある要素が持つ数値の関係性を調べたい時など実世界の問題に対して ILP を適用する場合、適切とは言えない。

また、数値的制約などの様々な制約を扱い、問題解決を行う制約論理プログラムというプログラミング言語がある。制約論理プログラムは、論理プログラムの中に制約を解消するメカニズムを組み込んだものであるので、背景知識を論理プログラムで記述できる ILP と親和性が高い。その制約論理プログラムを使用して、不等式や点集合が示す領域 (凸包) を求める研究 [2] が行われている。凸包を求めることは、正事例しか与えられていない問題において、データ内に含まれる数値が示す範囲を学習する際にすべての正事例を含む最小の領域が求められるので有効である。

よって、本研究では、制約論理プログラムに基づいた凸包という概念を溝口、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[3] に組み込むことで数値を含むデータからの学習を可能にすることを目的としている。

連絡先: 鈴木 匠，東京理科大学大学院 理工学研究科 経営工学専攻，千葉県野田市山崎 2641，04(7124)1501，j7409615@ed.noda.tus.ac.jp
なお 4 月 1 日現在、松井の所属は [とうごろう機械学習研究所] に変更

2. 帰納論理プログラミング

本研究の基盤となる帰納論理プログラミング (ILP) について説明を行う。現実問題に ILP を用いると、常に正事例と負事例の両方を得られるとは限らず、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すべてを含むような最小の領域を学習することができないという問題がある。

3. 制約論理プログラム

本研究で数値から得られる最小の領域を学習するために用いる制約論理プログラムについて説明を行う。制約論理プログラムは特徴として、「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しないため、解くべき問題の記述に専念できること、また制約を与えることにより、探索空間の縮小を図ることができる。本研究では、ILP システム GKS に制約論理プログラムを取り入れる。それは、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を求めて、それを学習をすることができないからである。そこで、対象とする問題の正事例内の数値が示す範囲を学習するため、本研究では凸包という概念を利用する。

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、「必要でない情報も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できる。」という利点がある。本研究で GKS に凸包を採用する理由には、2 つある。1 つ目は、凸包の利点を踏まえた上で、「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できる。」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。

本研究では、数値の示す範囲を学習する凸包プログラムは、[2] 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する。

4. 提案手法

4.1 アルゴリズム

本研究では、GKS を用いた数値を含むデータからの学習を目的としている。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。そこで、構築したシステムは、既存の GKS で採用されているトップダウン探索ではなく、ボトムアップ探索の概念を採用し、一般的な仮説を生成する。

本研究で提案するシステムの構成を図 1 に示し、システムの動き (1∼ 8) を説明する。

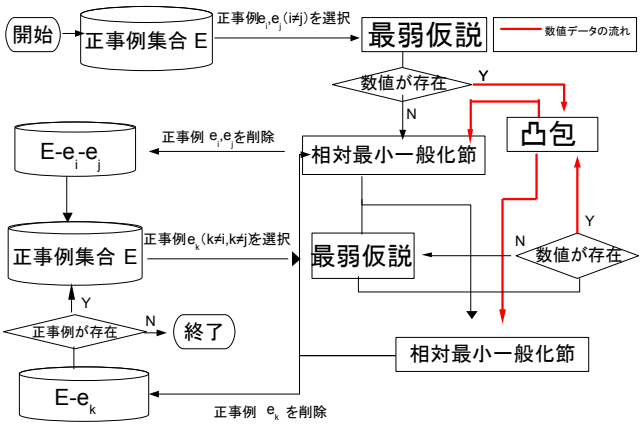

図 1: システム構成

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を 2 つ (e_i,e_j ただし i≠ j) 取り出し、それぞれの最弱仮説を生成する。最弱仮説とは、一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. 最弱仮説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 の最弱仮説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、最弱仮説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. また 1 つ正事例 (e_k ただし k≠ $i, k \neq j$) をランダムに正事例集合 E から取り出し、正事例 e_k の最弱仮説を生成する。
6. 生成した e_k の最弱仮説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。
7. 4 で求めた相対最小一般化節と 5 で新たな正事例から得られる最弱仮説からさらに一般化した相対最小一般化節を求め、取り出した正事例 e_k を正事例集合 E から取り除き、正事例集合 E を更新する。
8. 正事例集合 E が空になるまで、5,6,7 を繰り返し行い、すべての正事例を説明する最も一般的な相対最小一般化節とすべての正事例が示す数値範囲を求める。

以上の 1 から 8 を行うことで、数値を含むデータからの相対最小一般化した節と数値範囲、両方の学習が可能となる。

4.2 実行例

本研究では、数値を含むデータとして、表 1 に示す「ゴルフのプレイ条件に関する問題」[4] を用いる。

表 1: ゴルフのプレイ条件に関する問題

day	class	outlook	windy	temp(°F)	humidity(%)
day1	play_golf	sunny	TRUE	75	70
day5	play_golf	sunny	FALSE	69	70
day6	play_golf	overcast	TRUE	72	90
day7	play_golf	overcast	FALSE	83	78
day8	play_golf	overcast	TRUE	64	65
day9	play_golf	overcast	FALSE	81	75
day12	play_golf	rain	FALSE	75	80
day13	play_golf	rain	FALSE	68	80
day14	play_golf	rain	FALSE	70	96

この問題は、正事例 (play_golf)、背景知識 (outlook(天候)、windy(風の有無)、temp(温度 [° F])、humidity(湿度 [%])) を持つ。背景知識の outlook の値には、sunny(晴れ)、overcast(曇り)、rain(雨) があり、windy の値には TRUE(風が有る)、FALSE(風が無い) がある。データ内の day の欄に day2、day3、day4、day10、day11 が存在しない理由はそれらが負事例であり、本研究では、データに数値を持つ正事例のみからの学習を行うため、省略している。

アルゴリズムの 6 において正事例集合から取り出した 2 つの正事例 (day5、day7 とする) と新たに取り出した正事例 (day8) から凸包を生成するまでの実行例を示す。初めに取り出した 2 つの正事例の背景知識に数値が存在すると判断し、学習した数値範囲は以下のようになる。

```
Y=<rat(78,1),
Y>=rat(70,1),
X=rat(-107,2)+rat(7,4)*Y
```

X に temp の値を、Y に humidity の値をとる。不等式中の rat(A,B) は, 有理数 (A/B) を表現する関数である。正事例 day8 の背景知識 (outlook,windy,temp,humidity) から数値を引数に持つ背景知識 (temp,humidity) を取り出す。そして、凸包のプログラムにパラメータとして、不等式 [Y=<rat(78,1),Y>=rat(70,1),X=rat(-107,2)+rat(7,4)*Y] と day8 の持つ数値 [X=64,Y=65] を与える。その結果、元は 2 つの正事例 (day5、day7) を含む範囲を示す不等式であったが、3 つの正事例 (day5,day7,day8) を含む範囲に更新され、新しい範囲を示す不等式が次のように出力される。

```
X-rat(19,13)*Y=<rat(-31,1),
X-Y>=rat(-1,1),
X-rat(7,4)*Y>=rat(-107,2)
```

day5、day7、day8 の数値が上の不等式を満たすことから、すべての正事例を含む凸包が求められていることがわかる。

5. 実証

本研究で構築したシステムを問題 (表 1) に適用することで、数値を含むデータから学習が行えているかどうかを実証した。

5.1 実証環境

本研究で構築したシステムは、Prolog プログラムの処理系に SICStus Prolog 4.0.4 を用いて実証を行った。

5.2 実証結果

「ゴルフのプレイ条件に関する問題」に本システムを適用し、出力された学習結果を図 2 に示す。

```
%%% LGG ~CONVEX HULL~%%%[X=temp,Y=humidity]
play_golf(A) :-
 outlook(A,B), windy(A,C), temp(A,D), humidity(A,E),
 X-rat(2,3)*Y=<rat(31,1), X-rat(6,5)*Y=<rat(-9,1),
 X-rat(6,31)*Y>=rat(1594,31), X-rat(11,5)*Y=<rat(-79,1),
 X+rat(13,18)*Y=<rat(418,3).
```

図 2: 学習結果

図 2 で示す学習結果として出力された節は、一般化節と数値を含むデータから得られる不等式を結合した結果である。ゴルフを行う A という日は、天候 B が sunny、overcast、rain のどれかであり、風の有無 C に true、false のどちらかをとり、温度が D[° F]、湿度が E[%] であり、また述語 temp(X) と humidity(Y) の値は、次に示す 5 つの不等式

- `X-rat(6,31)*Y>=rat(1594,31)`
- `X+rat(13,18)*Y=<rat(418,3)`
- `X-rat(11,5)*Y=<rat(-79,1)`
- `X-rat(6,5)*Y=<rat(-9,1)`
- `X-rat(2,3)*Y=<rat(31,1)`

をすべて満たす範囲にあることを意味している。

上記の不等式に上から①,②,③,④,⑤を割り振る。そして、ゴルフのプレイ条件に関する問題に本システムを適用した結果、正事例が持つ数値の示す範囲が学習されているかどうかを確認するため、X 座標に temp(温度)、Y 座標に humidity(湿度) をとった X-Y 座標平面に、正事例を緑点で、求められた不等式を赤線で書き表すと、図 3 のようになる。

①~ ⑤の不等式が示す範囲を青線で記述した。青線で示される部分から、表 1 で記したすべての正事例を含む領域を最小の凸多角形で示していることがわかる。これより数値を含むデータを対象とする時、本研究で構築したシステムを用いることで、すべての正事例から得られる一般化された論理式 (相対最小一般化した節) と数値範囲の両方を学習できたと言える。

6. 結論

本研究は、ILP システム GKS を用いて数値を含むデータから学習するために、凸包プログラムを GKS に組み込み、仮説生成に関しては与えられた事例を一般化していくことで論理式と数値範囲を学習する手法を提案し、システムを構築した。「ゴルフのプレイ条件に関する問題」に構築したシステムを用いて実証を行うことで、正事例の中に引数に数値を持つ属性が 2 つ、かつ値の異なる数値が最低でも 3 点あれば、既存の

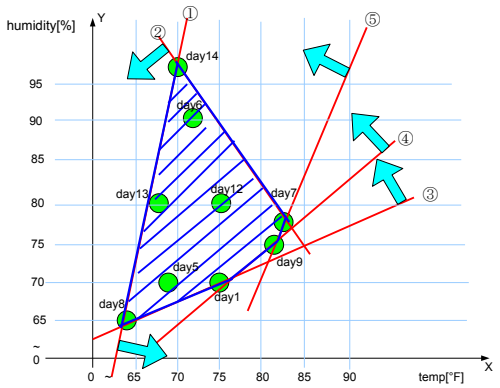

図 3: 学習された数値範囲

GKS では出来なかった数値を含むデータからの数値範囲と論理式両方の学習が本システムにおいて可能であることを確認できた。

参考文献

[1] 古川, 尾崎, 上野：帰納論理プログラミング, 共立出版.(2001).

[2] Florence Benoy, Andy King, Frederic Mesnard: Computing convex hulls with a linear solver,TPLP 5(1-2),259-271.(2005).

[3] Mizoguchi,F.,Owada,H: Constrained Relative Least General Generalization for Inducing Constraint Logic Programs. New Generation Computing,Vol.13, pp.335-368.(1995).

[4] J.Ross Quinlan: C4.5:PROGRAMS FOR MACHINE LEARNING ,Morgan Kaufmann Publishers.(1993).