2C1-3

帰納論理を用いたネットワークデータからの知識獲得

Knowledge discovery from data of network type using inductive logic programming

金城敬太[*1] 友寄隆智[*2] 相澤彰子[*3] 古川康一[*2]
Kinjo Keita Tomoyose Takatoshi Aizawa Akiko Furukawa Koichi

[*1] 総合研究大学院大学・JMR 生活総合研究所 Graduate University for Advanced Studies, Japan Consumer Marketing Research Institute
[*2] 慶應義塾大学大学院 Keio University, Graduate School
[*3] 国立情報学研究所 National Institute of Informatics

This research aims at analyzing network structured data with time variance using induction logic programming. Our preliminary experiment shows that the proposed method successfully extracts a structural change and the cause for the mining industry.

1. 背景

今日、ネットワーク構造が多くの注目を集めている. その理由として、これまで社会的に観測されてきたネットワークにおけるベキ則が単純なシミュレーションによって説明できるようになってきたことや、ソーシャルネットワークやブログなどの発達でネットワーク構造を持ったデータが手に入り易くなったことが挙げられる. ここで、伝統的な社会ネットワーク分析の研究では、関係の有無を調べて共通の構造を観察するといったアプローチも多く見られた. しかし、現実のネットワークを詳細に分析する場合には、次の三点について考える必要がある. 関係には様々な種類が存在していること、そして個々のノードの持つ属性情報も関係を形成する重要な要因となっていること、加えて関係構造は動的に変化しており、時間についても考慮に入れる必要があるということである.

これらをふまえ本研究では、時間的に変化するネットワーク構造から変化点を抽出し、さらに変化要因の特定を行う問題を設定した. その過程では、ノードやリンクの属性に基づき帰納論理プログラミングを適用する. 本研究では、提案手法の有効性を確認するため、産業連関データを対象としてネットワーク分析を行った. 産業については、これまでグラフ構造を用いた分析[市橋 01]や産業連関の時系列を扱う研究、Burt をはじめとした構造的空隙(くうげき)の研究などいくつかあるものの[Burt 88]、動的要素や属性などを含めた統合的な立場からの分析は少なかった. 結論としては、構造の変化やその原因の特定、また制約の抽出に成功し、主に鉱業の分野での変化を特定できた.

2. 問題設定

問題設定を行う. 問題は二つある. まず、時間的に変化のするネットワークデータを用意し、変化をした時点はどこであるかを特定すること、次に、その要因はどこにあるのかを探索することである. 具体的には、ネットワーク上のノードに属性情報を追加した上で変化抽出を行うことで、変化した構造や変化しない制約条件はどのようなものかを推論する. このような分析は、1. の議論でも示したように現在、ネットワーク構造分析ではあまり扱われておらず、産業間の分析例において自動的に構造を抽出可能であったことから意義があると考える.

連絡先:総合研究大学院大学, 国立情報学研究所
〒101－8430 東京都千代田区一ツ橋 2－1－2,
E-mail: kinjo@nii.ac.jp

3. 提案手法

提案手法について概説する. はじめにネットワーク分析を行うために、関係性をもつデータをある基準をもとにバイナリ表現された隣接行列データに変換する(3.1)、次に変換された隣接行列をもとにして、変化点を抽出する(3.2)、さらに詳細な分析を行うために作成したデータにノードの属性を追加した上で述語表現に変換する(3.3). 最後に変換した述語に対し、帰納論理プログラミングを用いることでノードの属性も含めた変化を抽出し、また一貫性制約の抽出を行い、各時点で変化してない構造とその差を探る(3.4).

本手法のポイントは次の三点に集約できる. まず、動的に変化するネットワークからの構造の抽出、さらに制約を抽出する手順をしめしたこと、また属性や複雑な関係構造を容易に組み込む方法を提案したことである. 利点として解釈が容易であることも挙げられる. なお、本研究のデメリットとしては、ネットワーク分析と異なり、隣接行列の行列計算などマクロな情報を容易には扱えないこと、またあるルールを学習するのに述語の組み合わせをすべてチェックする必要があり、計算コストが高くなる場合があることが挙げられる. この場合、前者は属性としてデータに組み込んで分析を行うこと、後者は知識を組み込み仮説空間を小さくする、仮説の深度を浅くすることなどが考えられる.

3.1 ネットワークデータの作成

まず、ネットワークデータの作成法について述べる. 通常、ネットワーク構造をもったデータは二種類あり、重み付きグラフであるか重みなしグラフであるかに分かれる. ただ、単純にネットワーク分析を行う場合においては、その関係構造に着目をするという意味で、抽象化して重みなしで考えることが多い. この場合、重みのあるグラフに対して、基準値を定めることで、バイナリの隣接行列、すなわち重みなしグラフを作成することが出来る.
本研究でも、この方法をとった.

3.2 変化点の検出

次に 3.1 で作成したネットワークデータを時間に沿って複数個用意する. その上で、ネットワーク構造が変化した点を調べる. ネットワーク構造の変化点を調べるには、大きく二つの方法がある. 目的変数がある場合と目的変数がない場合である. 目的変数がある場合というのは、ネットワーク構造をもとにして説明を行いたいノードが持つ属性(例えば、産業における利益率など)、またはネットワーク構造全体がもつひとつの特徴量などが存在している場合を指す. このようなとき、変化点はこの特徴量の二時点の変化(ベクトルの場合は、各時点同士の距離)をもとにあ

る規準値以上の点を探ることで抽出できる. 一方、後者の場合は、ネットワーク構造をもつデータのみが与えられているときで、このような場合は隣接行列などから各ノードのもつ中心性などの特徴量を計算した後に距離を計算して変化を出すほか、ネットワーク同士の距離や QAP 相関から変化を探る方法などが考えられる. 本研究は、前者の目的変数がある場合に相当する.

3.3 データの選定と変換

続いて、データの選定と変換について述べる. データの選定というのは、次の 3.4 以降での処理でノードに付随する変数(属性)を必要なものだけ選ぶ作業である. この作業についても下記の 2 つの場合がある. まず、各ノードに目的変数がある場合、また全体に対して目的変数がある場合である. 前者の場合、ノードの目的変数の量とノードの特徴量〈例えば、マクロなネットワークの特徴量;中心性ベクトルなど〉との相関により、関連するもの選定する. 後者の場合、距離変化と目的変数の変化の相関から考えることができる. 例えば、ネットワーク全体の取引量などが存在した場合、その時間変化〈2 時点間変化〉と中心性ベクトルの時間変化の相関により選定することができる. なお、目的変数がない場合は、このような作業はできない.

以上の作業後、隣接行列およびノードの特徴行列を述語形式にデータを変換する. ノードの特徴行列と隣接行列は、それぞれインデックス(id)と、ノードの属性(node_attribute)、リンクの属性(link_attribute)および時間(time)で構成され

node(id, time,node_attribute). link(id1,id2, time,link_attribute).

と記述される. link は id1 から id2 への有向リンクを表す. ここで 1で述べたように関係に複数の種類がある場合、link_attribute で関係を分けることで表現することが可能である. なお、述語論理と関係データについての関連研究は 6 で述べる.

3.4 帰納論理プログラミング

最後に本研究では、3.3 で作成したデータを、帰納論理プログラミング上に組み込んでルールを獲得する. 帰納論理プログラミングとは、関係データマイニング(リレイショナルデータマイニング)を行う一階述語論理ベースの機械学習器である[古川 01]. 抽出できるルールとしては、分類ルール(判別分析)や頻出パタンの検出(相関ルール)などが挙げられる. 特徴としては、述語論理によって知識を表現することができるため表現力が豊かであることと背景知識を加えることが出来る点がある. 本研究ではこれらのふまえ下記の 3 点を行う. まず構造に関する知識を組みこむこと、次に本研究の目的である変化した構造について抽出すること、最後に変化前後における一貫性制約条件の違いについての抽出することである.

①構造に関する知識の導入

一般的に、ルールについては制限をおかず、グラフマイニングのように頻出するものを探索していくことが有効である. ただし、得たいルールに関する知識がある程度ある場合は、背景知識を組み込むことで、探索を制限することや、より詳細な知識の獲得に有効となる. 例えば、時間関係の知識を取り入れる場合、

before(X,Y,Z):-link(T1,X,Y),link(T2,X,Z),T2 is T1－1.

というような情報を仮説に加えることで link に関して1区間のみの前後関係のみを調べることも可能となる. 同様にある変数の大小関係を組み込むことも可能である.

②変化構造の抽出

変化構造の抽出について、本研究では2つのパタンを考える. まず変化点特定後、変化した前後で判別するルールを学習する. こうすることにより、どのような関係が変化したかがわかり、その変化により全体としての変化が起きたと特定できる. また、最も変化の大きかったノードに対して、その時間を通じての共通ルールを抽出する操作も行う. これにより、全体としての構造変化が起きた原因となるローカルな変化原因の特定を行う.

③一貫性制約の抽出

次に、一貫性制約条件の抽出について説明する. ある構造が変化した要因を制約条件の変化として捉える. 一貫性制約は、もとはデータベースの分野でできてきた用語で、あるデータベースにおいて「違反をしてはいけない」規則をあらかじめ記述しておくことで、データベースの一貫性を保つという役割を果たす. 例えば、動物のデータベースであれば、「ある動物は哺乳類と爬虫類に同時に分類されてはならない」などが一貫性制約となる. 帰納論理プログラミングにおいて、このような制約は

false:-A,…,notB,… .

という形式で扱うことができる. ここで扱われている not というのは、ある条件が成り立たないことを理由として、その否定が成り立つことを表現している. 具体的な制約の抽出方法は、予め設定した制約が成り立つかどうかをチェックしていき、すべてのデータに対して成り立たない場合、一貫性制約として抽出するというものである. このような研究の例では例えば Alipio によるモンテカルロ法を用いて効率的に抽出する研究がある[Alipio 96]. なお、制約が変化したかどうかをみるためには、変化した前後の制約を抽出し、その後差分を計算することで得られる.

4. 分析

3 で提案した手法を用いて、産業連関表についてデータを分析した. 分析の目的は、産業の利益率の変化からそれらを規定する構造とその要因の抽出といった新たな知見の獲得である.

まず、データについて簡単に説明を行い、次に分析結果と考察を述べる.

4.1 データの説明

産業連関表は、経済学者レオンチェフにより開発されたもので、産業間の投入と産出の構造を集計しており、生産波及効果などを把握するのに使用されている. 日本では総務省および各省庁で 5 年おきに更新されており、部門についても13部門,32部門,104 部門等が存在する. 本研究では、正式なものとして直近で公開されている平成 2 年、平成 7 年、平成 12 年の 3 年間分のデータ,32 部門を用いた. 産業間の関係については、レオンチェフ逆行列を用いている. レオンチェフ逆行列とは、ある産業部門が単位1増加した場合に各部門に及ぼす値を示しており、本研究で扱う産業間の影響度として妥当である. 値は A を投入係数としたとき $(I-A)^{-1}$ で計算される. このレオンチェフ逆行列に対し、全体平均をとり、それよりも大きいものに対して関係があるとし、バイナリデータの隣接行列を作成した. なお、本研究においては、バイナリデータであるため、関係の属性を扱わなかったが、関係が複数ある場合や関係の階層などがある場合は、3. 3で述べたように関係の属性を容易に組み込むことが出来る.

次に目的変数となる売上高営業利益率を、各ノードに対して割り当てた(財務省「法人企業統計」より各 23 部門,他は平均を使用し計算). また、組み込む属性としては、ノードが持つマクロ

なネットワークの特徴量である中心性、産業間の関係性の固定としてのクラスター係数、産業内の寡占を表現するハーフィンダール指数などが考えられる. 32 部門で産業をくくった場合のハーフィンダール指数は計算ができないため本研究では用いてない. 中心性は次数と betweenness が利益率と相関が高かったため使用した.

4.2 結果と考察

以上のデータを3で述べた方法により帰納論理プログラミングにより分析した結果を述べる. なお本研究では Aleph[Srinivasan 99]をもとに構築した.

まず、変化構造の抽出について述べる. 変化した時点を売上高営業利益率のベクトルの時系列変化をおった結果、平成 7 年と平成 12 年の間での変化が最も高かったため、ここを変化点とした. 次に、全体で変化した構造について述べる. 上記の期間での変化が大きかった 10 産業を正の事例として、そのルールを抽出した. 結果、運輸(21:数値はすべて id)、金融・保険(23)というように運輸から投入、金融保険から投入がある産業が抽出された(accuracy=1). 具体的には、鉱業(2), 鉄鋼(8),窯業・土石製品(9)であった. これらの資金の流れは、これらの産業が設備投資などを背景としていると考えられ妥当性もある.

link(T,21,X),link(T,23,X),link(T,21,23).

また、各産業に対して、最も変化の大きかった部門に対して、時間を通じてのルールを抽出してルールを抽出したところ、非鉄金属(10),金属製品(11)において以下のルールが抽出された(accuracy=1).

Syuueki_chenge(X) :-

link(h7, X, 11), node(h2, X, link-low, between-high).

すなわち、平成 7 年時に金属製品(11)とのリンクがあり、さらに平成 2 年において次数の中心性が低く、一方で beweenness が高かったことと解釈ができ、このような状況下にある産業において今後収益率が向上する可能性を示唆している.

次に抽出された一貫性制約条件についてのべる. 今回、上記であげた鉱業(2)について、平成 12 年においてレオンチェフ逆行列で高い値を誇った上位5つ、石油・石炭製品(7)、電力・ガス・熱供給(18)、窯業・土石製品(8)、非鉄金属(10)、鉄鋼(9)、6(化学製品)の部門間での関係についての制約の抽出を行い、変化の前後で制約の差分をとった. 結果として

false:-link(T,2,2),link(T,2,7).

という制約の差分が抽出された. すなわち、石油・石炭製品と鉱業の制約関係が生じおり、制約が変化していることが特定された. なお、not に関する制約の差分は得られなかった.

5. 関連研究

この節では関連した研究について述べる. ネットワーク構造、もしくはグラフ構造と帰納論理プログラミングについての考察はこれまでも何度かなされており[猪口 04] [Ketkar 06] [金城 08]、例えば計算速度などの利点から考察がなされている. ただ応用や個々の利点を活かした研究は少ない. また、ネットワークの時系列モデリングについては鹿島らの研究[鹿島 07]や社会ネットワークでの研究[井上 05]などがある. また、関係データマイニングは確率的な関係学習としても研究が数多くされている[Lise 07].

応用については、金融で須田らが銀行の与信に関しての研究を行っている[須田 07]ほか、企業間のネットワークに関する研究[湯川 07]があるなど社会科学の分野でも徐々に応用が研究されてきている.

6. まとめと今後

以上、時間変化のあるネットワークから変化をした時点を特定し、構造変化やその要因はどこにあるのかを探ることを目的として試論を行った. 結果として、変化した要因として産業構造の変化、また個別産業の変化ではその要因の特定、さらに制約の変化についても意義のある結果を抽出することが出来た.

今後は、変数・時間変化を増やすことを検討しているほか、計算量の問題、また発展としてはアブダクションなどと組み合わせ、目的に合わせた新たな関係の提案などへ結びつけることなどが課題である.

参考文献

[Alipio 96]Alipio Jorge, Pavel Brazdil, Integrity Constraints in ILP using a Monte Carlo approach,Proceedings of the 6th International Workshop on Inductive Logic Programming 1996.

[Burt 88] Burt:The stability of American Markets. American Journal of Sociology, 94, 356-395. 1988

[Gulati 98]Gulati:R. Alliances and networks,Strategic Management Journal, 19, 293-317, 1998.

[Ketkar 06] Nikhil S.Ketlar LawrenceB.Holder Diane J.Cook comparison of Graph-based and logic-based multi-relational data mining SIGKDDI volume7,Issue2 2006

[Lise 07]Lise Getoor etc,Introduction to Statistical Relational Learning (Adaptive Computation and Maching Learning) 2007

[Srinivasan.99]A.Srinivasan,:The Aleph Manual. Available at :http://www.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/, 1999

[相原 06]相原基大, 秋庭太:企業者ネットワークに関する経験的研究の現状と展望,北海道大学,經濟學研究, 56(1): 57-75.2006

[市橋 01]市橋勝:連関構造データによる産業ネットワークの把握『地域経済研究』第 12 号,2001.3

[井上 05]井上寛:社会ネットワークの変動,『ネットワークダイナミクス』,勁草書房,2005

[猪口 04]猪口明博 グラフマイニングと ILP システムの比較考察 the 18th annual conference of the Japanese society for artificial intelligence 2004

[鹿島 07]鹿島久嗣,安部直樹,ネットワーク構造の確率的な時変モデルに基づく教師ありリンク予測,人工知能学会論文誌,Vol.22,No2,2007

[金城 08]金城敬太,相澤彰子,古川康一:帰納論理プログラミングによる定性ネットワーク分析 2007 年度新領域融合プロジェクトによる研究会, 2008

[須田 07]須田侑子,服部正純:ネットワーク分析からみた国際的な銀行与信関係の発展 日本銀行ワーキングペーパーシリーズ 2007.9

[古川 01]古川康一,尾崎知伸,植野研:帰納論理プログラミング共立出版,2001

[湯川 04]湯川 抗:企業間ネットワークからみたネット企業のクラスターと企業戦略湯川 抗, 富士通総研、『研究レポート』,2004

[渡邊 05]渡邊剛,小坂武:日本における企業間関係の社会ネットワーク分析 渡邊剛,小坂武 東京理科大学経営情報学会 2005 年春季全国研究発表大会, pp.356-359.