自主研究

ソフトウェア開発プロファイルデータの相関ルール分析

～ERPパッケージ導入実態調査結果への適用～

ソフトウェア開発プロファイルデータの相関ルール分析
－ ERP パッケージ導入実態調査結果への適用 －

奈良先端科学技術大学院大学 情報科学研究科　森崎 修司　木村 早苗　門田 暁人　松本 健一
財団法人 経済調査会 調査研究部 第三調査研究室

はじめに

ソフトウェアの生産性や品質の向上を目的として、ソフトウェアやその開発組織・プロジェクトの特性を表すデータ（ソフトウェア開発プロファイルデータ）を収集、分析する活動が盛んに行われている。国内では、独立行政法人情報処理推進機構／ソフトウェア・エンジニアリング・センター（IPA/SEC）の活動がよく知られており、収集されたデータの分析結果は、「ソフトウェア開発データ白書」として2005年以降毎年公表されている。2008年度版の同白書には、2000件を超えるプロジェクトの種別、ソフトウェア開発規模（ファンクションポイント, SLOC）、工数、工期、検出バグ数等の分析結果が掲載されており、各特性値の分布や標準値、特性値間の相関等を知ることができる。

ただし、ソフトウェア開発プロジェクトは、PMBOK（Project Management Body of Knowledge）などでも指摘されているように、何らかの点で相互に異なり、個別性が非常に高い。加えて、特性間の関係が複雑で、例えば、工数一つとっても、その決定要因は多数存在し、プロジェクト毎に異なる可能性が大きい。特性間の相関を調べるだけでは、普遍的な傾向やパターンを見出すことは難しく、プロジェクト管理に利用できるほど具体的な知見を得るのも容易ではない。

ソフトウェア開発プロファイルデータを分析する新たな技術として、相関ルール分析（アソシエーション分析）を適用する研究がすすめられている。相関ルール分析は、蓄積された大量のデータから、頻繁、かつ、同時に生起する事象を見つけ出す技術である。これを用いれば、例えば、工数の普遍的な決定要因は解明できなくても、（規模あたりの）工数が標準値より特に大きくなるという事象は、ソフトウェア開発においてどのような事象が発生した時に（プロジェクト特性値がどのような値をとる時に）発生する可能性が高くなるのかを知ることができる。二つの事象間の前後関係や因果関係を詳細に検討することで、工数の増大を防ぐための具体的な知見を得ることができる。

本稿では、財団法人経済調査会によって実施された「平成20年度　ERPパッケージに関する調査」によって得られた133社からの回答結果に対して相関ルール分析を適用した結果について述べる。

1．相関ルール分析

相関ルール分析はデータマイニング手法の基本的なものの1つであり、大量のデータの中から「AならばB」というルール（相関ルール）を見つけ出す技術である。相関ルールをA⇒Bと表記し、Aを前提部、Bを結論部と呼ぶ。

よく知られている適用例としては、小売店POSシステムの販売履歴データなどから、顧客の購買傾向やパターンを洗い出し、販売戦略に活用するというものである。ここで、説明を簡単にするため、販売履歴データとは、個々の販売記録をレコード（タプル）とするデータセットであり、販売記録には、販売日時等のコンテキスト情報と共に、商品ごとの販売数が記録されているとする。

同時購入される商品が知りたいのであれば、

前提部が命題「商品Xの販売数が0でない」、結論部が命題「商品Yの販売数が0でない」であるルールを抽出することになる。なお、前提部や結論部に命題の論理積を用いることができるので、3つ以上の商品の同時購入もルールとして抽出することができる。その結果、例えば、**「休日に"レジャーシート"を買う顧客は"おにぎり"と"お茶"も同時に買っている」といったルールが抽出され、レジャーシートをおにぎりやお茶のそばに配置して販売するといった戦略をたてることができる。**

ただし、「販売数が0でない」商品の組み合わせは数多くあり、抽出されるルール数が膨大となることは容易に想像できる。抽出されたルール全てに従って商品の配置を決めることは現実的ではない。相関ルール分析では、ルールの重要さを表す3つの指標が提案されており、それぞれの下限値を設定することで、抽出ルール数を制御するのが一般的である。3つの指標は次の通り。

支持度＝前提部と結論部が同時に真となる場合数（レコード数）／全場合数（全レコード数）

信頼度＝前提部と結論部が同時に真となる場合数（レコード数）／前提部が真となる場合数（レコード数）

リフト値＝信頼度／結論部が真となる場合数（レコード数）／全場合数（全レコード数））

支持度は、ルールの出現頻度を表す指標であり、信頼度とリフト値は、前提部と結論部の関連の強さを表す指標である。信頼度が大きいほど、前提部が真の場合に結論部も真となりやすいことを表し、リフト値が大きいほど、結論部が真の場合に前提部も真となりやすいことを表す。例えば、販売記録数（レコード数）が20、商品Xが販売された（場合数が10、商品Yが販売された場合数が8、商品Xと商品Yが同時に販売された場合数が6だとすると、「商品Xが販売されるならば商品Yも販売される」というルールの支持度は0.3（＝6/20）、信頼度は0.6（=6/10）、リフト値は1.5（=0.6/（8/20））となる。

なお、相関ルールは、「AならばB」と表現されるが、一般には、前提部Aと結論部Bの共起関係を示しているに過ぎない。抽出されたルールを利用する場合は、AとBの間の前後関係や因果関係に留意する必要がある。

2.相関ルール抽出支援ツール NEEDLE

2.1 プロファイルデータ分析に向けた機能

本稿では、相関ルールの抽出に、文部科学省EASE（Empirical Approach to Software Engineering）プロジェクト[1]で開発されたNEEDLEを用いる。NEEDLEにおけるルール抽出プロセスを説明する前に、従来システムにはないNEEDLE固有の機能を紹介する。

従来の相関ルール分析の対象は、名義尺度または順序尺度に基づくデータ（質的データ）である。特に、１章で例示したように「商品が販売されたか（購入したか）どうか」といった２値データや、性別、年齢層、満足度（５段階評価）などといった、カテゴリ数の少ないデータである場合が多い。これに対して、本稿で対象とする「ソフトウェアやその開発組織・プロジェクトの特性を表すデータ（ソフトウェア開発プロファイルデータ）」には、「業種」や「ERPパッケージ製品名」といった多数のカテゴリを持つ名義尺度や順序尺度に基づくデータだけでなく、「契約金額」や「１人月の基準時間」といった間隔尺度や比例尺度に基づくデータも含まれる。NEEDLEでは、そうしたデータを対象とした相関ルール分析を可能とするため、次のような機能が追加されている。

[1] http://www.empirical.jp/top.html

[追加機能１] 前提部・結論部への論理和・否定の導入

相関ルールA⇒Bの前提部Aと結論部Bにおいて、命題の論理和を用いることができる。但し、論理和として記述できるのは、名義尺度もしくは順序尺度に基づくデータのカテゴリ結合を目的とした場合に限る。これにより、例えば、「導入・運用しているERPパッケージはGLOVIA/SUMMITもしくはGLOVIA-Cである。」といった命題を前提部や結合部で用いることが可能になる。なお、論理和の記述範囲を限定しているのは、組合せ爆発による計算量の増大を防ぎ、現実的な時間でルール抽出を実現するためである。

更に、前提部Aと結論部Bにおいて、命題の否定も可能とする。これにより、例えば、「業種は「その他」以外である」といった命題を前提部や結合部で用いることが可能になる（同様の命題を論理和でも表現可能であるが、結合すべきカテゴリが多数となり煩雑になる場合がある）。

[追加機能２] 尺度水準の変換

間隔尺度や比例尺度に基づくデータ（量的データ）を順序尺度に基づくデータに変換する（カテゴリ数は変数定義ファイルで与えられている）。更に、ルール抽出においては、最下位カテゴリと最上位カテゴリそれぞれを、隣接するカテゴリと順次結合していく。これにより、例えば、「ERPパッケージ規模が大きい（対象データにおいて上位を占める）⇒１人月の基準時間が大きい」といったルールの抽出が可能となるだけでなく、抽出ルールの信頼度やリフト値が最大となるようなカテゴリ分けが自動的に得られる。

[追加機能３] 結論部におけるデータ指定

ルール抽出において、結論部に現れるデータ（変数）を指定することができる。例えば、「生産性」と指定すると、([追加機能２]と相まって）対象データにおいて生産性が上位あるいは下位を占める場合を特定するようなルールのみを抽出することができる。これにより、ルール抽出に要する計算時間が大幅に短縮されると共に、抽出されたルールに解釈を与えたり、その有用性を判断したりするための分析者の工数も小さくなる。

[追加機能４] 結論部への統計量の導入

結論部に量的データの統計量（平均値と標準偏差）を持つルールを抽出することができる。例えば、「業種が建設⇒１人月の基準時間の平均がa、標準偏差がb」といったルールの抽出が可能である。NEEDLEでは、このようなルールを「量的ルール」と呼んでいる。

「量的ルール」に対しては、リフト値に該当する指標を次のように新たに定義している。

基準化平均＝量的ルールにおける当該量的データの平均a／全場合（全レコード）における当該量的データの平均

基準化標準偏差＝量的ルールにおける当該量的データの標準偏差b／全場合（全レコード）における当該量的データの標準偏差

例えば、ある量的ルールの基準化平均が2だとすると、その量的ルールの前提部が成り立つ場合、当該量的データの平均値が通常の２倍になることになる。なお、量的ルールの抽出においては、平均値、標準偏差それぞれの上限と下限を指定することができる。

2.2 ルール抽出プロセス

NEEDLEにおけるルール抽出プロセスを**図１**に示す。NEEDLEへの入力は、「対象データ」、「変数定義ファイル」、「ルール抽出指示ファイル」の３つであり、すべてCSV形式のテキストファイルである。

対象データは、分析対象となるデータであ

り、１行が１レコードであり、ソフトウェア開発に携わる組織やプロジェクトに対応する。各レコードは、組織やプロジェクトの特性を表す変数で構成される。

変数定義ファイルには、対象データのレコードを構成する変数それぞれについて、尺度水準（名義尺度、順序尺度、間隔尺度、比例尺度）、間隔尺度や比例尺度に基づくデータ（量的データ）の場合には質的データへの変換におけるカテゴリ分けの方法とカテゴリ数、欠損値の補完方法、などが定められている。

ルール抽出指示ファイルには、抽出ルール数を制御するための指標（支持度，信頼度，リフト値，基準化平均，基準化標準偏差）それぞれの上限値、下限値などが定められている。

これら３つの入力を用いて、ルール抽出は次に示す４つのステップで進められる。

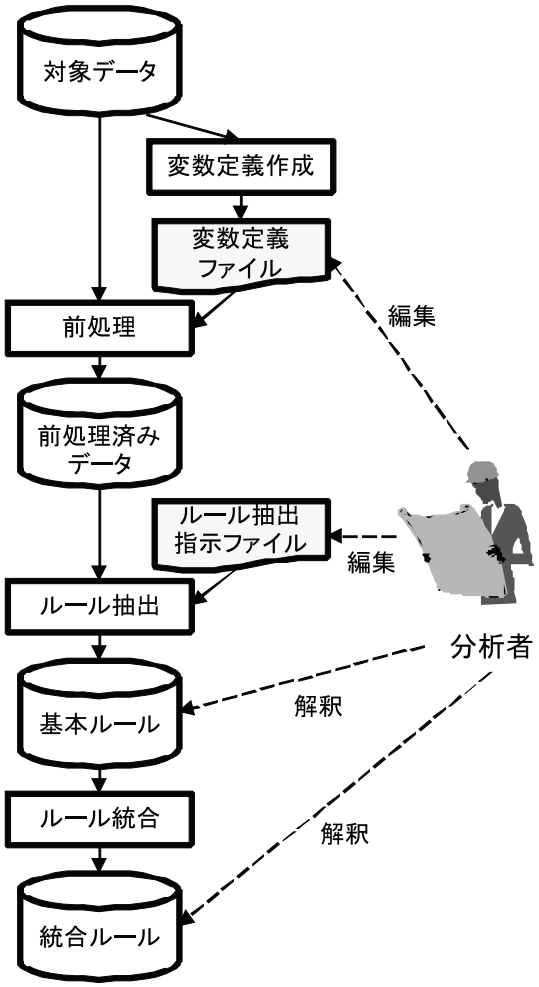

図１NEEDLEにおけるルール抽出プロセス

Step 1　変数定義作成

対象データを解析し、変数の尺度水準などを可能な限り自動判別し、「変数定義ファイル」のひな型を生成する。分析者（ルール抽出者）は、このひな型を編集し、変数定義ファイルとして完成させる。

Step 2　前処理

変数定義ファイルに基づき、対象データにおける欠損値の補完、尺度水準の変換、などを行い、「前処理済みデータ」を生成する。

Step 3　ルール抽出

ルール抽出指示ファイルに基づき、前処理済みデータから基本ルール（前提部・結論部に論理和を含まないルール）を抽出し、支持度等の指標値を算出する。抽出された基本ルールは「基本ルールファイル」に保存される。

Step 4　ルール統合

前提部および結論部に含まれる変数のカテゴリ結合を目的として、基本ルールファイル中のルールの統合を行い、支持度等の指標値を算出する。統合されたルールは統合ルール（前提部・結論部に論理和を含むルール）として「統合ルールファイル」に保存される。

分析者は、こうして作成された基本ルール、および、統合ルールファイルに対して、各種指標値を参考にしながら解釈を与え、分析目的に沿ったルールを選び、活用することになる。

3．ソフトウェア開発プロファイルデータへの適用

3.1 対象データ

財団法人経済調査会によって実施された「平成20年度　ERPパッケージに関する調査」によって得られた133社からの回答結果が対象データである。調査票はⅠ～Ⅴに分かれており、それぞれの調査概要、設問数、回答数を**表1**に示す。調査票Ⅰで、ERPパッケージ導入のコンテキストを、調査票Ⅱ～Ⅴで、要件定義に始まりバージョンアップまで、ERP導入作業を工程順に調査する構成となっている。同調査は、ERPパッケージを導入している企業に対して、そのライフサイクルコストの観点から実態を明らかにし、ERPパッケージの導入作業別（調査票Ⅱ，Ⅲ，Ⅳ，Ⅴの別）にそれぞれの作業工数やコストを予測するための根拠となる情報を収集することを目的として実施された。

3.2 抽出された主なルール

３．１で述べた調査結果にNEEDLEを適用した。分析の観点はいくつもあるが、本調査は、ユーザ企業に対して行われたものであり、ユーザ企業にとってERPパッケージを導入する際のパートナー企業の存在は大変大きいものである。そこで、本稿では、ERPパッケージ導入において、どのような条件が成り立つ場合にパートナー企業の支援を受ける確率が高くなるのか、また、パートナー企業の支援を受けた場合に、ERPパッケージ導入プロジェクトがどのような特徴を持つ可能性が高くなるのか、を導入作業別に見ることとした。

具体的には、抽出条件を、

- ルール結論部に質問項目「パートナー企業の有無」への回答を含む
- 支持度0.05以上
- リフト値1.3以上

としてERPパッケージ導入作業別に抽出を行った。

表1　ERPパッケージ調査の概要

調査票	調査概要	設問数	回答数（うち自由記述数）
Ⅰ	回答組織の概要とERPパッケージの導入状況	6	38 (4)
Ⅱ	要件定義～パッケージ選定の作業状況	11	75 (21)
Ⅲ	構築・設定～運用テスト・導入教育の作業状況	21	112 (24)
Ⅳ	保守・運用の作業状況	18	134 (34)
Ⅴ	バージョンアップの作業状況	11	83 (20)

抽出されたルールのうち、「契約金額根拠」の項目を含むものの一部を**表2**に示す。調査票Ⅲから得られたルールが**表2**に含まれていないのは、上記条件を満足するルールが得られなかったためである。なお、支持度0.05以上という条件で抽出していることから、**表2**のルールで表わされるような関係は、全回答の5%以上で見られることになる。更に、リフト値1.3以上ということから、ルールの前提部が真となる場合には、そうでない場合と比べて、ルールの結論部が真となる確率（パートナー企業の支援を受ける確率）は1.3倍以上となる。すなわち、比較的多くの企業に当てはまり、パートナー企業の支援を受ける確率を格段に高くするルールと言える。なかでも、保守・運用工程における**表2**の項番6、7のルールは、支持度がそれぞれ0.214、0.286と非常に高く、ERPパッケージ導入の多くの場合で見られる関係であることがわかる。また、要件定義～パッケージ選定工程における**表2**の項番1～4のルールは、リフト値が1.530と高く、当該ルールの前提部は、パートナー企業の支援を受ける典型的な状況や強い条件と考えることができる。

なお、ルールの抽出において信頼度は考慮していないが、**表2**に列挙したルールの大半は信頼度1.000であり、低いものでも0.955である。このことは、前提部が真となる場合、ほとんど

表2　得られた主な相関ルール

項番	対象調査票 作業工程）	相関ルール	支持度	信頼度	リフト値
1	Ⅱ 要件定義～ パッケージ選定	（契約金額根拠＝パートナー見積）＆（成果物＝候補選定結果）→（パートナー企業の支援＝あり）	0.109	1.000	1.530
2		（契約金額根拠＝パートナー見積）＆（時間/人月＝160以上）→（パートナー企業の支援＝あり）	0.158	1.000	1.530
3		（契約金額根拠＝過去実績）＆（成果物＝作業報告書）＆（成果物＝システム要件定義書）→（パートナー企業の支援＝あり）	0.119	1.000	1.530
4		（契約金額根拠＝過去実績）＆（成果物＝作業報告書）＆（成果物＝業務要件定義書）→（パートナー企業の支援＝あり）	0.119	1.000	1.530
5	Ⅳ 保守・運用	（作業内容＝保守契約）＆（含バージョンアップ費用＝含まれない）＆（契約形態＝請負）＆（契約金額根拠＝パートナー見積）→（パートナー企業の支援＝あり）	0.112	1.000	1.400
6		（契約形態＝請負）＆（契約金額根拠＝パートナー見積）＆（成果物＝作業報告書）→（パートナー企業の支援＝あり）	0.214	0.955	1.336
7		（契約金額根拠＝パートナー見積）＆（成果物＝作業報告書）→（パートナー企業の支援＝あり）	0.286	0.966	1.352
8	Ⅴ バージョンアップ	（作業内容＝現状調査）＆（契約金額根拠＝パートナー見積）＆（成果物＝作業報告書）→（パートナー企業の支援＝あり）	0.175	1.000	1.500
9		（作業内容＝現状調査）＆（契約金額根拠＝パートナー見積）＆（成果物＝機能一覧表）→（パートナー企業の支援＝あり）	0.123	1.000	1.500
10		（作業内容＝現状調査）＆（作業内容＝業務検証）＆（契約金額根拠＝パートナー見積）＆（成果物＝テスト結果報告書）→（パートナー企業の支援＝あり）	0.105	1.000	1.500
11		（作業内容＝現状調査）＆（作業内容＝業務検証）＆（契約金額根拠＝パートナー見積）＆（成果物容＝機能一覧表）→（パートナー企業の支援＝あり）	0.105	1.000	1.500

の場合に結論部も真になる（パートナー企業の支援を受ける）ことを意味しており、このことからも、当該ルールの前提部は、パートナー企業の支援を受ける典型的な状況や強い条件と考えることができる。

3.3 考察

要件定義～パッケージ選定作業に関するルール（調査票Ⅱから得られたルール。**表2**の項番1～4）をみると、「パートナー見積り」を契約金額根拠としている場合には、「候補選定結果」が成果物であり、1人月あたりの時間数は160時間／月以上と見積もられていることがわかる。一方、過去の実績を契約金額根拠としている場合には、業務要件定義書やシステム要件定義などのプロダクトが成果物であり、作業報告書も合わせて納品されていることがわかる。

システム構築・設定～運用テスト・導入教育に関するルール（調査票Ⅲから得られたルール）は、いずれもリフト値が1.3未満であり、**表2**には含まれていない。少なくとも今回の調査項目には、パートナー企業の支援の有無に大きく影響を与える要因は含まれていなかったと言える。

保守・運用作業に関するルール（調査票Ⅳから得られたルール。**表2**の項番5～7）をみると、パートナー企業の見積りを契約金額根拠としていること、成果物を作業報告書としていること、契約金額にバージョンアップ費用が含まれないことがわかる。保守・運用においては、ベンダ主導で契約金額が決まることが多く、また、作業報告書による報告が中心と言えるようである。なお、**表2**には示していないが、調査票Ⅳからは、パートナー見積り以外が契約金額根拠となるような場合のルールは数件しか得られなかった。保守・運用作業の委託においては、多くの組織で、パートナー企業の見積りに基づいて契約金額が決まるというのが現状のようである。

バージョンアップ作業に関するルール（調査票Ⅴから得られたルール。**表2**の項番8～11）をみると、パートナーの見積りを契約金額根拠とし、現状調査、業務検証を作業内容とし、機能一覧表、テスト結果報告書を成果物としていることがわかる。**表2**には示していないが、調査票Ⅴからは、パートナー見積り以外が契約金額根拠となるようなルールは発見されなかった。バージョンアップ作業の委託においても、多くの組織で、パートナー企業の見積りに基づ

いて契約金額が決まるというのが現状のようである。

以上をまとめると、ERPパッケージ導入には次のような実態があると推察される。

- 開発の上流（要件定義～パッケージ選定）においては、契約金額根拠として過去実績、パートナー見積りの両方が存在する。過去実績を根拠とする場合には、成果物がプロダクト（システム自体）に関するものである場合が多い。これらは、パートナー選定や契約金額設定において選択肢が多いことの現れとも考えられる。
- システム構築・設定～運用テスト・導入教育においては、パートナー企業の支援は作業に対する強い要因とはなっていない。
- 保守･運用作業においては、バージョンアップ作業を含まない契約とし、契約金額根拠はパートナー見積り、成果物を作業報告書とする等、パートナー企業主導の様子が伺える。
- バージョンアップにおいては、現状調査や業務検証を作業内容とし、保守・運用作業と同様にパートナー企業主導の様子が伺える。また、過去実績が契約金額根拠とされることはほとんどない。

4．まとめ

本稿では、財団法人経済調査会によって実施された「平成20年度　ERPパッケージに関する調査」によって得られた133社からの回答結果に対して相関ルール抽出支援ツールNEEDLEを適用した結果について述べた。

分析対象としたデータ数は134件であり、一般的な相関ルール分析におけるデータ数に比べると１ケタか２ケタ少ない数である。それでも、ERPパッケージの導入においてパートナー企業の支援を受ける典型的な状況や強い条件、また、パートナー企業の支援を受けた場合に、ERPパッケージ導入プロジェクトが持つことになる特徴をいくつか指摘することができた。これは、ソフトウェアやその開発組織・プロジェクトの特性を表すデータ（ソフトウェア開発プロファイルデータ）を分析対象とするための独自の機能がNEEDLEに追加されているためである。

本稿で紹介した分析手法は、ソフトウェア開発に関する同様の調査結果や蓄積データに広く適用できるものである。