

機械学習によるリスク評価に基づくソフトウェアコスト見積り技術開発

笹川 文義¹⁾、花山 亨²⁾、川村 旭¹⁾、桑 照宣¹⁾、斉藤 孝広¹⁾、
有馬 淳¹⁾、楠本 博巳²⁾

1)株式会社富士通研究所、2)富士通株式会社

IT分野のソフトウェア開発プロジェクトでは、プロジェクト特性（リスク評価項目）は客観的に判断できる項目もあるが、プロジェクト要員の特性といった客観的な判断が困難で主観的な特性を扱わざるを得ないためにリスク（係数詳細項目）を客観的に判断するのが困難であった。このため専門家でなければプロジェクト特性からリスクをリスクから開発工数の見積りを行うのは難しいとされている。リスクの客観的基準を過去の完了プロジェクト実績に求め、過去実績から対象プロジェクトのリスクを予測する。プロジェクト特性を細分化して2値（Yes/No）で表現し、機械学習の技術を用いてプロマネ見積りアンケートサンプルデータ上で比較的高い精度で実績を再現できる。得られた多数のリスクを開発工数見積りエンジン（COCOMOの活用）への入力とすれば、開発工数の見積り値が得られる。これによりソフトウェアコスト見積りへの機械学習適用の妥当性を検証できた。

**Approach to the Technology that Supports Software Cost Estimation
by Risk Level Presumption based on the Machine Learning**
Fumiyoshi Sasagawa¹⁾, Toru Hanayama²⁾, Akira Kawamura¹⁾, Terunobu Kume¹⁾,
Takahiro Saito¹⁾, Jun Arima¹⁾ and Hiromi Kusumoto²⁾
1) FUJITSU LABORATORIES LTD. 2) FUJITSU LTD.

The technology and the evaluation improvement technique for achieving the software cost estimation model that excluded the vagueness of the project profiles as much as possible in cooperation with PM Office were investigated. The model has subdivided COCOMOII coefficient item, assumed binary (yes/no) input, and is expected to lead the development person-months effort for vague exclusion. Moreover, the risk level presumption technique to presume the evaluation of COCOMOII coefficient item of seven values from a binary input automatically was developed for this achievement. In addition, the evaluation improvement technique for consisting of the validity of the quality evaluation technology of the coefficient item by machine learning and the model, the accuracy assessments, and optimization approaches, etc. was developed. 63 project sample data is received from PM Office as the estimation specialist based on the completion projects, and the analysis of data, the improvement candidate of the result, the optimization model, the model has been extracted as trial.

1. はじめに

近年、技術の専門化、複雑化などにより、同じ開発量でも、開発物の難易度、プロジェクト(以下、“Pj”)の特性（業務経験者の有無、仕様の安定度など）によって、生産性が大きく変動する状況にある。このため、Pj 損益の予想が困難になっている。また、従来の経験とカンに頼るやり方は客観性、透明性が弱く、組織的な技術向上につながらない。IT分野のソフトウェア開発 Pj では Pj 特性、開発特性を考慮した工学的な見積り技術の確立が急務となっている。

この報告書では、ソフトウェア開発に関する見積りを、a)開発作業（設計・製造・テスト）と b)SES 作業（方式設計、環境設計・構築など）に分け、さらに a)開発作業を「規模見積り」と「リスク

係数見積りに分けて見積り技術を開発している（図1参照）。本研究は「リスク係数見積り」に関するものである。

リスク係数見積りに対応する見積り技術としては、世界的にCOCOMOⅡが最も有名である（図2参照）。COCOMOⅡは開発規模と、開発物やPjの特性に関する13項目を7段階評価で入力することで、開発工数、期間を出力するモデルで、実際の利用に際しては、各利用環境でのプロジェクトデータを集め、パラメータ調整することを前提としている。しかしながら、7段階の評価基準（以降「係数項目」と呼ぶ）が曖昧であり、主観によるバラツキが入りやすく、品質の高いプロジェクトデータを集めるためには、専門家による精緻なヒヤリングが欠かせないことが利用を妨げている大きな障害となっていた。データ収集に多大のコストがかかるため、実際の現場では利用されていない状況であった[1]。

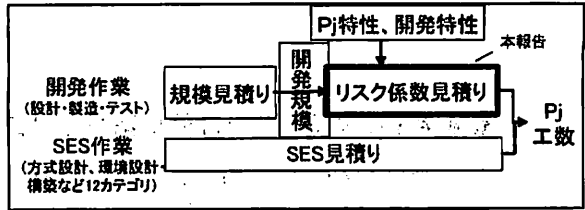


図1. 各種見積り技術の関係

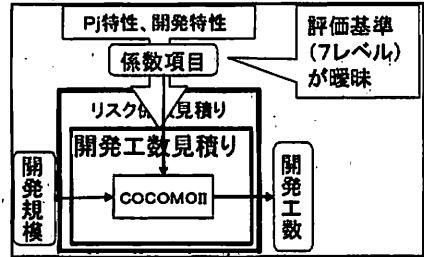


図2. 従来法

2. ソフトウェアコスト見積り技術へ向けて

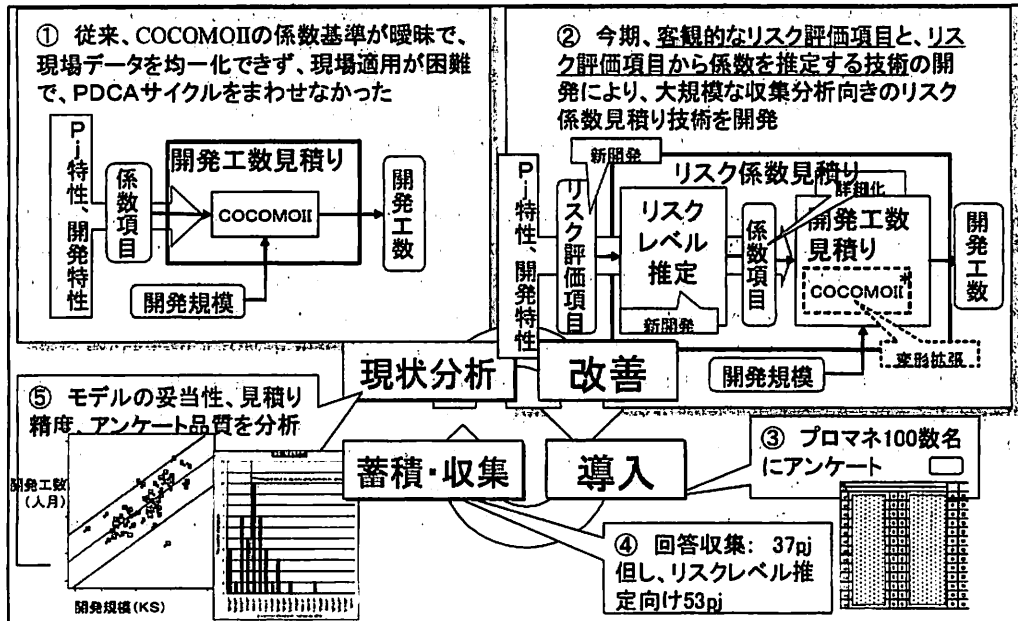


図3. PDCAサイクルの主なフェーズ

見積り技術の向上には、見積りモデルの精緻化と現場データによる評価フィードバックによるPDCAサイクルが欠かせないにもかかわらず、従来法では評価基準が曖昧なため現場データを均質化できず、品質の高い現場データの蓄積が困難であり、PDCAサイクルをうまくまわすことができていなかった（図3の①）。

2. 1. PDCAサイクルの確立

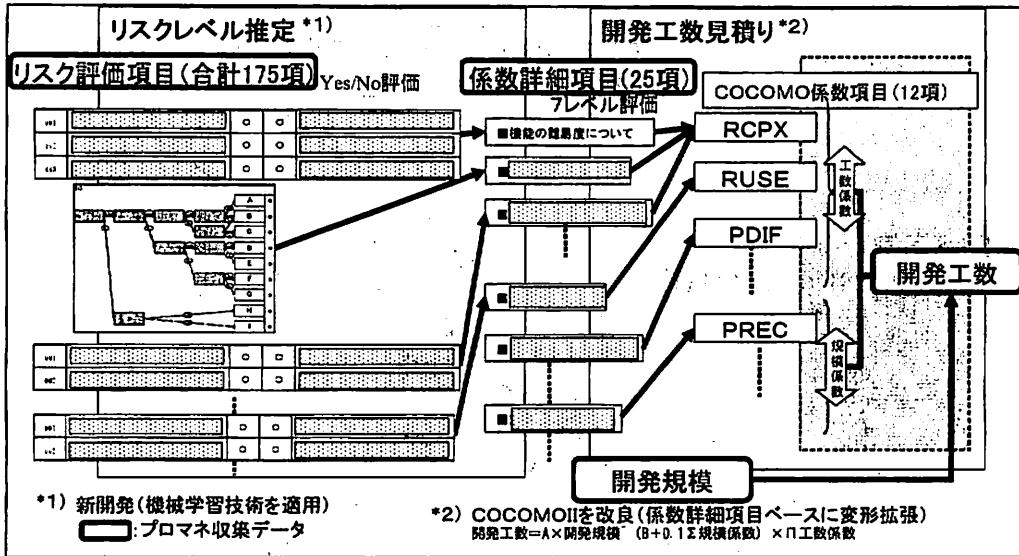


図4. リスクレベル推定と開発工数見積りによる見積りモデル

これに対し我々は、曖昧性を極力排除するため、開発物やプロジェクトの特性に関わると考えられ、かつ、客観的な事象として2値(yes/no)で答えられる評価項目を失敗Pj事例などからPMオフィスのプロマネらが抽出し、そこから従来の特性評価(国際標準的な見積り手法COCOMO IIの7レベル)を機械的に推定する技術を開発した(図3の②)。つまり今回、COCOMO IIのインプットである係数項目の代わりに、新たに曖昧性の少ない2値の評価項目(以降、「リスク評価項目」と呼ぶ)をインプットとし、COCOMO IIのインプットである係数項目を機械的に推定(以降、「リスクレベル推定」)する方式を開発した。

モデルの具体化(数値レベルでの詳細化)ならびに評価に向けた現場データ収集のために、PMオフィスから、プロマネ100数名に対し、過去Pjに関するデータの提供を呼びかけた(図3の③)。プロマネを対象とした理由は、開発物やプロジェクト特性に対する判断、係数項目のリスクレベルの判断について最も適格な判断を下せることが期待できたことを見込まれたためである。また、過去Pjが対象になったのは、開発量、開発工数の実績値を得る必要があったためである。過去Pjデータは全63Pj分が集まっている。うち、開発工数などの欠損のため、すべてのモデル評価に利用できたPj数は46Pjとなっている(図3の④)。

収集されたデータをもとに、モデルの妥当性、モデルの精度、ならびに各評価項目自身の品質評価を行った(図3の⑤)。その結果、本手法の現時点での全体精度はまだ不十分であるが、方向性は妥当であり、このままPDCAをまわしていくことで技術が確立する見込みを得ることができ

入手Pjデータ数		46 (リスクレベル推定向け 63)	
リスク評価項目(合計175項)			
係数詳細項目	リスク評価項目の数	係数詳細項目	リスク評価項目の数
機能	13	緊急経路	5
ドキュメント	*	系統経路	3
システム接続	9	プロマネ	12
専門性	8	ハード信頼性	5
セキュリティ	9	ソフト信頼性	5
複雑性	9	開発基礎	8
教育性	6	開発参加	11
再利用性	16	チーム作り	10
高品質度	5	目標共有	8
業務普及性	*	要員構成	*
顧客要件	8	品質管理	6
仕様確定	8	リスク	8
業務経路	5		

*: 特殊形式(デジジョントツリ形式など)

表1. プロマネ見積りサンプルデータの要約

た。また、ユーザのリスク評価を効率化するための評価項目の改善方針を具体的に示すことができた。

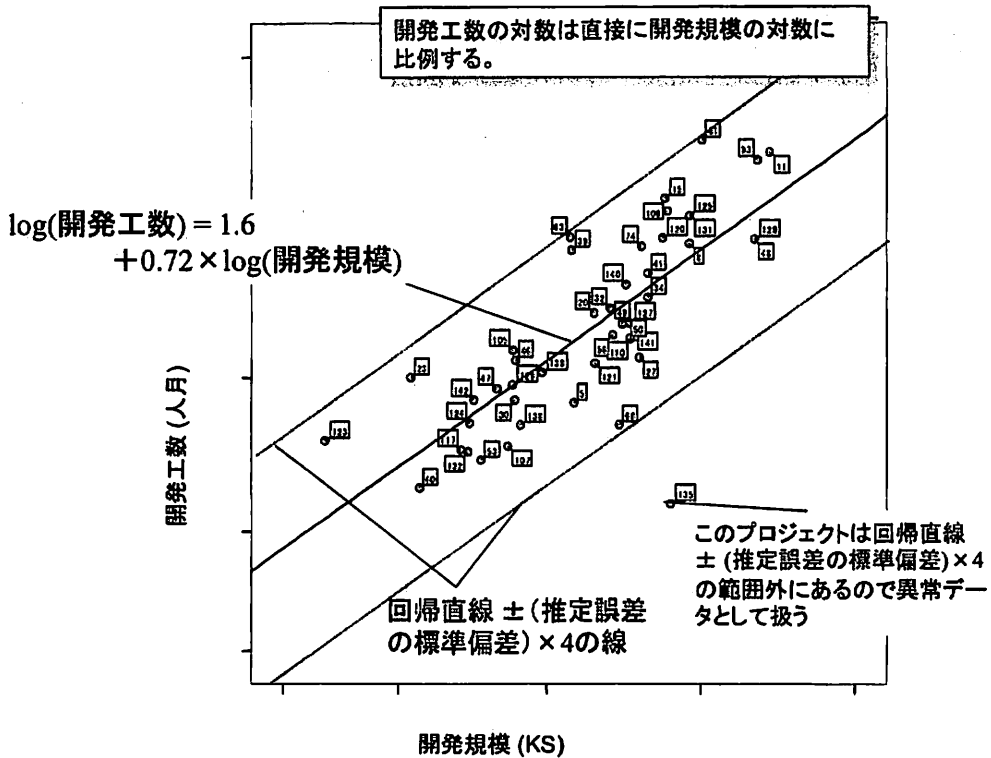


図 5. 開発規模と開発工数の対数線形性

このように、本報告では、現状分析からはじめ、見積りモデルの改善、現場に導入し、データを収集、そしてこの改善モデルの現状分析を行うところまで行うことができ、PDCAサイクルを一巡させることができた。

2. 2. リスクに基づく見積りモデル

今回開発した見積りモデルは、リスク評価項目（全 175 項目）と開発規模から開発工数を導くものである（図 4 参照）。本モデルは、「リスク評価項目」から「係数詳細項目」（25 項）の各 7 レベル値を推定する「リスクレベル推定」部分と、係数詳細項目と開発規模から開発工数を導く「開発工数見積り」部分の 2 つに分かれる。

プロマネから収集したデータ項目は、この「リスクレベル推定」と「開発工数見積り」のそれぞれのインプット、アウトプットになる。リスクレベル推定は、リスク評価項目の 2 値ベクトルと各係数詳細項目のレベルの関係をデータから学習する方式となっており、これによりリスク評価項目を変える場合や、見積りを利用する領域毎に評価レベルが異なるような場合でも、アルゴリズムの変更無しに、データを集めるだけで柔軟に対処することができる。今回学習方式としては、近年の学習方式として最も優れた結果を多く出している SVM (support vector machine) というアルゴリズムを採用している [2]。

開発工数見積りは、COCOMO II をベースにし、データに基づいて変形拡張している。まず、

COCOMO II の係数項目(12 項)を詳細係数項目 (25 項)に詳細化している。開発工数の見積りは大きく 2 つの方法を試している。1 つは、詳細係数項目から係数項目の値を求め、COCOMO II モデルを調整し開発工数を得る方式で、詳細係数項目から係数項目のマッピングは単純に詳細係数項目の平均値を対応する係数項目の値としている。もう 1 つは COCOMO II モデルを拡張し、詳細係数項目と開発規模から直接開発工数を導くものとなっている。

3. プロマネ見積りサンプルデータによる見積りモデル評価

収集されたデータをもとに、新しく開発したモデルの精度評価、および、そもそもの収集データを構成しているリスク評価項目の品質評価について述べる。

3. 1. リスクレベル推定

正確な見積りを行なうためには、プロジェクトの生産性に影響をあたえると思われる様々なリスクを考慮する必要がある。今回、実施したアンケートにおいては、それらのリスクを係数詳細項目として分類し、各々の危険度を評価したリスクレベル、そのリスクレベルに影響を及ぼすと考えられる質問(リスク

評価項目)に対する回答が得られている。しかしながら、ここで付与されるリスクレベルは回答者の主観に基づくものであり、全アンケート結果を考慮した統一的観点に基づくものにはなっていない。この問題を解決するには、リスク評価項目の回答結果から統一的観点に基づいて算出されるリスクレベルを回答者に示唆する機能が必要となる。そこで、

リスク評価項目への回答結果からリスクレベル推定を行なう実験を行なった。

本ツールは SVM(Support Vector Machine)と呼ばれる機械学習手法に基づく分類エンジンである

[1]。機械学習においては正解を必要とするが、今回のアンケート結果を正解とみなした。正解データが少ないので、今回はクローズドテストを行なうことにした。クローズドテストとは、正解データとテストデータを同一のものを使用したテストであり、正解データが十分に収集された理想状態における性能を見積ることができる。推定結果の一位正解率を係数詳細項目毎に表 2 にまとめる。一位正解率とは、ツールによって最も可能性の高いと判断されたリスクレベルが正解かどうかを評価するものである。結果として、項目により正解率に大きく差が出ているが、平均 73.7%の精度となった。

係数詳細項目	最大正解率	正解率	係数詳細項目	最大正解率	正解率
1.機能	0.93	0.86	12.開発経験	0.81	0.74
2.他システム接続	0.89	0.70	13.基礎経験	0.60	0.60
3.専門性	0.92	0.72	14.プロマネ	0.88	0.81
4.セキュリティ	0.86	0.76	15.ハード信頼性	0.73	0.73
5.複雑性	0.91	0.82	16.ソフト信頼性	0.72	0.61
6.許容性	0.73	0.58	17.開発基礎	0.88	0.76
7.再利用	1.00	0.98	18.顧客参加	0.95	0.90
8.基礎品質	0.80	0.68	19.チーム作り	0.82	0.73
9.顧客要件	0.90	0.75	20.目標共有	0.87	0.80
10.仕様確定	0.78	0.65	21.品質管理	0.95	0.78
11.業務経験	0.66	0.57	22.リスク	0.89	0.69

■: 正解率比0.8以上の係数詳細項目

表 2. リスクレベル推定結果

この正解率のパラツキは、主にリスク評価項目数に依存していると考えられる。全く同じ回答パターンであるが、付与されたリスクレベルが異なるプロジェクトが存在する場合、そのどちらが実際の正解であるかは判定できないので確実に精度が低下するが、リスク評価項目数が少ない場合には、この事態が発生する確率が高くなると考えられるからである。そこで、新たな評価基準として、最大正解率の概念を導入し、ツールの推定結果が最大正解率にどの程度近いかという観点に基づいて項目毎に統一的評価を行なうことにする。

最大正解率とは、一つの回答パターンにおける最も数の多いリスクレベルを選択した時の正解率である。最大正解率は、正解データにおける回答パターンに全く重複がない場合や、重複があっても同じ回答パターンにおいては必ず一つのリスクレベルが付与されている場合には最大値 1 を取り、矛盾するデータ（同じ回答パターンに複数のリスクレベルが付与されているデータ）が増えるに従って減少する。表 2 には、係数詳細項目毎に、この最大正解率も記載した。最大正解率とツールの正解率の比（正解率比）が、0.8 を切っている項目は 4 項目のみであり、平均 0.878 とかなり高い水準を達成した。

リスクレベル推定実験において最大正解率という概念を導入し、推定精度の評価を行なったが、最大正解率は係数詳細項目毎の適切さの評価にもなっている。同じ回答パターンにおいても付与されるリスクレベルが異なるという事象は、リスク評価項目が不足しており、リスク評価項目にない要因によってリスクレベルの差異が生じたと考えることができるからである。そこで、この観点に基づく係数詳細項目の評価を行なうことにする。ただし、最大正解率によってあとどの程度のリスク評価項目（質問）が必要なかを判定することは難しい。この判定を行なうために、新たにエントロピーの概念を導入する。

エントロピーは以下の式で定義されるスカラー量であり、完全に矛盾のない状態、すなわち、同じ回答パターンにも関わらず付与されているリスクレベルが異なるデータセットが存在しない状態を達成するためには、あと何ビットの情報が必要かを示唆する。今回のアンケートにおいてはリスク評価項目の回答は 2 値であるので、この値は不足質問数に相当することになる。ただし、この評価は最大正解率 100% を達成するためのものであるため、現状でも高い最大正解率を達成している係数詳細項目にさらにリスク評価項目を質問として追加する意義は薄いかもしい。また、例えばエントロピーが 1 より大きく 2 より小さい場合には追加すべき項目数は 2 となるが、ここで必要とされる評価項目は現状の項目数では弁別できないデータを弁別するためのものでなければならず、とにかく 2 つのリスク評価項目を追加すれば、必ず最大正解率 100% を達成できるわけでない。その意味で、エントロピーによって算出される不足数はあくまでも目安として利用すべきである。

$$H = - \sum_i \sum_j p_{ij} \log_2 p_{ij}$$

係数詳細項目	エントロピー	平均	係数詳細項目	エントロピー	平均
1.機能	1.13944	0.087649	12.開発経緯	1.68341	0.336681
2.他システム接続	1.23227	0.136919	13.技術規模	1.34774	0.449246
3.専門性	0.845816	0.140986	14.プロマネ	1.6687	0.138889
4.セキュリティ	1.26471	0.140523	15.ハード信頼性	1.43505	0.287009
5.複雑性	0.881613	0.097957	16.ソフト信頼性	1.77619	0.355237
6.許容性	1.87041	0.311735	17.開発基盤	1.3038	0.162975
7.再利用	0	0	18.顧客参加	0	0
8.基礎品質	1.50187	0.312374	19.チーム作り	0.77897	0.077897
9.顧客要件	0.892216	0.124027	20.目標共有	0.992488	0.124061
10.仕様確定	2.43929	0.304911	21.品質管理	0.212412	0.035402
11.業務経験	2.06752	0.413505	22.リスク	0.499736	0.052467

■:曖昧性が高い要検討グループ

表 3. エントロピーによる不足リスク評価項目数評価（エントロピー：完全な推定を行うために必要とする最小評価項目数、平均：現状における 1 評価項目辺りの曖昧性）

ここで、 p_{ij} は回答パターン i において、リスクレベル j をとる確率である。なお、 $p_{ij} = 0$ のとき、 $p_{ij} \log_2(p_{ij}) = 0$ と定義する。この式より明らかに、各回答パターンに各々唯一のリスクレベルが対応する場合のエントロピーは 0 となり、矛盾データの数が增大するにつれて値が大きくなる事が分かる。

また、この値を項目数で割った平均エントロピーは、個々のリスク評価項目の不確定さを平均的に評価する評価値としてみることが出来る。係数詳細項目毎にエントロピー及び平均エントロピーを表 3 にまとめる。なお、平均エントロピーをみると、値の高い項目と低い項目の間には大きなギャップがあることが分かり、高エントロピーグループにおいてはリスク評価項目を増やす以外にも、現状のリスク評価項目がリスクレベルの決定要因をうまく聞いているかについても検討を行なう必要があるといえる。

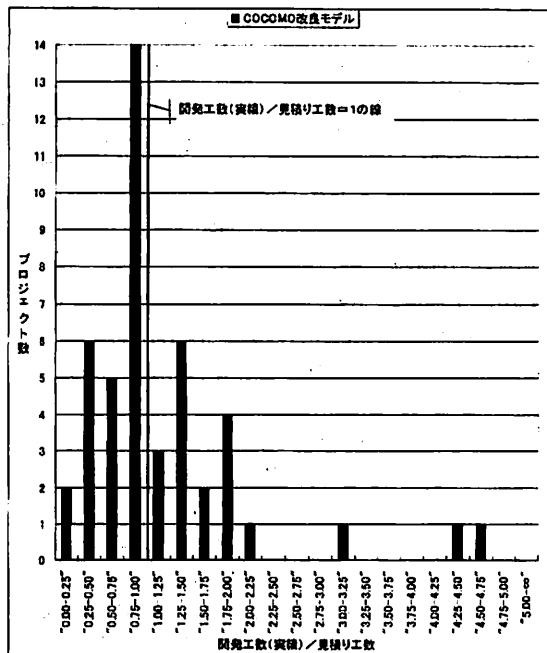


図 6. 開発工数(実績) / 見積り工数のヒストグラム (COCOMO 改良モデル)

3. 2. 開発工数見積り

今回のソフトウェア開発工数の見積りモデルの基盤として採用した COCOMOII のモデルのパラメータは、モデルへの入力項目のレベル判定基準とともに、入力されたデータを統計的に処理した上で、複数の見積りの専門家が協議して最終調節を行うこととなっている。ここでは同様に、アンケート結果のデータを用いた統計処理によって見積りモデルのパラメータの調節と精度評価を行う [3] [4]。

評価は異常データを取り除いたデータを元に、次の 3 つの見積りモデルを評価対象とし比較した。

a) COCOMO II モデルの妥当性を検討するための比較対象として最も単純な開発規模のみを因子とする”線形モデル”、b) 純粋な COCOMO II モデルとして COCOMO 係数項目を因子とする”COCOMO モデル”、c) COCOMO II の式の構造だけを活かし、係数詳細項目の主成分を因子とする”COCOMO 改良モデル”である。以下、それぞれの詳細について結果を記す。

アンケートデータ中、開発規模及び開発工数(実績)が揃っている 46 Pj を対象として、開発規模の対数のみを説明変数とし、開発工数の対数を従属変数とする線形回帰分析を行った。

その結果の線形回帰モデルから得られる見積り工数の対数の値から、推定誤差の標準偏差の 4 倍以上離れた値の開発工数(実績)を持つ Pj を異常データとした。

図 5 に Pj 毎の開発規模-開発工数(実績)のプロット結果と、線形分析結果のモデル及び、推定誤差の標準偏差の 4 倍の範囲を示す。異常データを除くデータ全体を用いてパラメータ調節を行った後の各モデルの見積り精度を表 4 にまとめた。ここで、” $PRED(.34) = 33\%$ ” という表記は、「見積り工数の $\pm 3.4\%$ 開発の範囲内に、開発工数(実績)が納まっていた割合が、 3.3% であった」ということを意味している。

また、モデル評価のためのテストデータとして、パラメータ調節に用いたデータを用いた。即ち、

クローズドテストを行った。
COCOMO モデル及び、COCOMO改良モデルの見積り精度が、開発規模のみを説明変数とする線形モデルよりも良いことから、両モデルの基盤とした COCOMO II の見積り式の構造の妥当性が確認された。

見積りエンジン	PRED (. 3 4)	PRED (. 2 0)
線形モデル	33%	22%
COCOMO モデル	42%	27%
COCOMO 改良モデル	51%	36%

表 4. データ全体を用いた開発工数見積りエンジンの性能

また、アンケート結果のデータを直接利用することが可能な COCOMO 改良モデルの方が、COCOMO モデルよりも良い見積り精度が得られた。

調節後の COCOMO 改良モデルの見積り式：

$$\begin{aligned}
 (\text{見積り工数}) &= 5.61 \times (\text{開発規模})^E \\
 &\quad \times (\text{工数係数グループの第1成分})^{0.43} \\
 &\quad \times (\text{工数係数グループの第2成分})^{-0.83} \\
 &\quad \times (\text{工数係数グループの第3成分})^{0.0} \\
 &\quad \times (\text{工数係数グループの第4成分})^{0.56} \\
 &\quad \times (\text{工数係数グループの第5成分})^{0.65} \\
 &\quad \times (\text{工数係数グループの第6成分})^{-0.69}
 \end{aligned}$$

$$\begin{aligned}
 E &= 0.58 + 0.195 \times (\text{規模係数グループの第1成分}) \\
 &\quad + 0.104 \times (\text{規模係数グループの第2成分}) \\
 &\quad - 0.092 \times (\text{規模係数グループの第3成分})
 \end{aligned}$$

調節後のモデルの見積り精度を評価するために、開発工数（実績）が見積り工数に対してどのように分布するかを調査すると図 6 のグラフのようになった。

4. まとめと今後

プロマネ見積りアンケートの妥当性の分析評価がプロマネコスト見積りサンプルデータに基づいて機械学習によって実施された。さらに COCOMO II に従って開発工数（人月）を計算する開発工数見積りモデルがプロマネ見積りサンプルデータに基づいて得られた。効果的に見積り技術を高精度化するために PDCA サイクルを回すための新技術の妥当性を強固なものにした。プロマネ見積りデータの蓄積には現場プロジェクトマネジャーの積極的なデータの投入を促進するためのメカニズムの開発が必要である。このメカニズムによって PDCA サイクルは見積り技術をより有効にできる。

5. 参考文献

- [1] 久保宏志、富士通における品質保証の実際、日科技連、1989
- [2] Cristianini, Nello and Shawe-Taylor, John、サポートベクターマシン入門、共立出版、2005
- [3] B.W. Boehm, C. Abts, A.W. Brown, S. Chulani, B.K. Clark, E. Horowitz, R. Madachy, D. Reifer and B. Steece, *Software Cost Estimation with COCOMO II*, Prentice HALL, New Jersey, 2000.
- [4] B.W. Boehm, B. Clark, E. Horowitz, C. Westland, R. Madachy and R. Selby, "Cost models for future software life cycle processes: COCOMO 2.0", *Annals of Software Engineering* 1, Springer, Amsterdam, 1995, pp. 57-94.