


協調フィルタリングによるプロジェクトデータ分析

データ欠損とプロジェクト個別性に強いプロジェクト予測

奈良先端科学技術大学院大学
情報科学研究科
大杉 直樹, 松本 健一

第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005




ソフトウェアプロジェクト予測

- 過去プロジェクトのデータから予測モデルを作成する。
 - 線形モデル(重回帰分析), ニューラルネットなど
- 現行プロジェクトの実績値を予測モデルに当てはめ, 工数, バグ数, などを予測する。

予測モデル: $\text{試験工数} = 26.5 + \text{設計工数} \times 0.275$

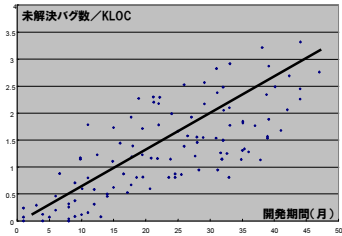
	設計工数	製造工数	基本設計 欠陥数	詳細設計 欠陥数	試験工数
現行プロジェクトX	50	20	3	10	40.25
過去プロジェクトA	45	18	2	9	39
過去プロジェクトB	55	22	3	11	44
過去プロジェクトC	10	10	4	5	30

第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005




従来モデルの問題点(1)

- 多様なソフトウェア開発プロジェクトを一つのモデルで表現することは難しい。



第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005




プロジェクトとは

- 組織の戦略計画を達成する手段。
- 通常の業務範囲内では対処できない要求に応える手段。
- 有期的
 - 明確な始まりと明確な終わりがある。
- 独自の(個別的)
 - 何らかの識別できる点において, 他とは異なる。
- 継続的, 反復的な「定常業務」とは異なる。

プロジェクトマネジメント知識体系ガイド(PMBOKガイド)–2000年版,
プロジェクトマネジメント協会, 2000.

第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005




従来モデルの問題点(2)

- データ欠損に対して脆弱である。
 - データ欠損を補う方法は開発されているが, 欠損率が30%を超えると, 予測精度は著しく低下する。

Kromrey, J., and Hines, C.: "Non-randomly missing data in multiple regression: An empirical comparison of common missing-data treatments," *Educational and Psychological Measurement*, 54, 3, pp.573-593 (1994).


第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005



データの欠損は避けられない

- 開発過程のデータ(リアルタイムに収集されるデータ)は, 取り直しがきかない。
- 分析目的や組織が異なれば, 収集データも異なる。
 - 収集データを共通化するために, 開発モデル, ツール, ドキュメントなどを全て共通化することは現実的ではない。
 - その時点での分析目的に合致しないデータを収集するだけの余裕があるとは限らない。

第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005



研究の目的とアプローチ

- 目的
 - プロジェクトの個性と収集データの欠損に強いプロジェクト予測方法の確立.
- アプローチ
 - モデル構築より類似プロジェクト検索
 - 協調フィルタリング技術の応用

第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005



モデル構築より類似プロジェクト検索

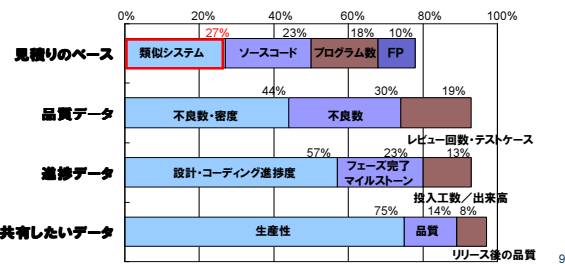
- 有能なプロジェクト管理者は、見積りや問題解決において、オールマイティなモデルを持っているわけではない.
- モデルよりも個々のデータを経験としてうまく活用している.
 - 過去に携わったプロジェクト群の中から、現行プロジェクトと似たプロジェクト(類似プロジェクト)を選び出す.
 - 類似プロジェクトにおける開発コスト、作業進捗、発生した問題とその解決策、などを、現行プロジェクトに(多少アレンジした上で)適用する.

第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005



受注ソフトウェア開発での計測

(社)情報サービス産業協会(JISA),「情報サービス産業における受注ソフトウェア開発の技術課題に関わるアンケート調査」, 2004年.



第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005



協調フィルタリング技術応用の代表例

- Amazon.com の書籍推薦システム (www.amazon.co.jp)
 - ユーザは、読み終えた本を5(好き)~1(嫌い)の5段階で評価する.
 - システムは、好みの傾向が似ているユーザが高く評価した本を推薦する.



第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005



Amazon.com の書籍評価予測

- ステップ1: 類似度計算
 - 推薦対象ユーザと他プロジェクト間の類似度を計算する.
 - 類似度の高い k (例えば $k=2$) 人のユーザを選ぶ.
- ステップ2: 予測値計算
 - 類似ユーザの評価を加重平均し、推薦対象のユーザの評価を予測する.

	書籍1	書籍2	書籍3	書籍4	予測結果
推薦対象ユーザX	5(大好き)	4(好き)	2(嫌い)	3(普通)	4.52
ユーザA	類似度: 0.94 (好き)	4(好き)	2(嫌い)	(欠損値)	5(大好き)
ユーザB	類似度: 0.87 (損値)	4(好き)	2(嫌い)	4(好き)	4(好き)
ユーザC	類似度: -0.87 (大嫌い)	1(大嫌い)	(欠損値)	5(大好き)	1(嫌い)

第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005



協調フィルタリングを用いた予測

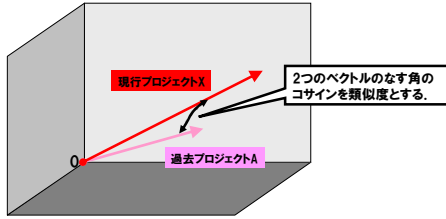
- ステップ1: 類似度計算
 - 現行プロジェクトと過去プロジェクト間の類似度を計算する.
 - 類似度の高い k (例えば $k=2$) 個のプロジェクトを選ぶ.
- ステップ2: 予測値計算
 - 過去プロジェクトの工数を加重平均し、現行プロジェクトの特性(工数、残存バグ数、など)を予測する.

	設計工数	製造工数	基本設計欠損数	詳細設計欠損数	予測結果
現行プロジェクトX	50	20	3	10	40.0
過去プロジェクトA	45	18	2	(欠損値)	36
過去プロジェクトB	22	3	11	44	44
過去プロジェクトC	10	10	(欠損値)	5	30

第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005



類似度計算



第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005



ケーススタディ

- 協調フィルタリングを用いた方法と従来法(重回帰分析)とで、試験工数の予測精度を比較した。
- 1081個のプロジェクトから無作為に半数を選び(予測に用いる)過去プロジェクトとし、残りを(予測対象の)実行プロジェクトとした。

	設計工数	試験工数
プロジェクトA	80	40
プロジェクトB	70	10
...

過去プロジェクト

	設計工数	試験工数
プロジェクトA	80	???
プロジェクトC	90	???
...

実行プロジェクト

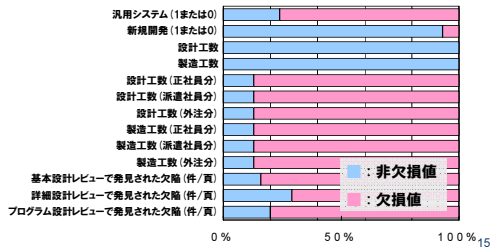
予測値と実測値を比較

第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005



ケーススタディで用いたデータ

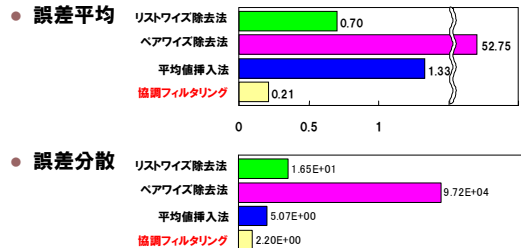
- プロジェクト数:1081 (予測対象分 540)
- データ欠損率:平均60%



第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005



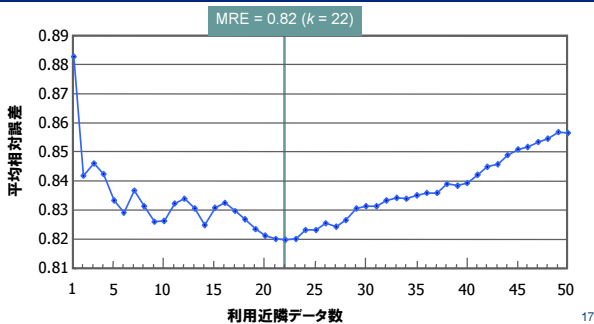
予測精度



第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005



利用近傍データ数と平均相対誤差



第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005



まとめ

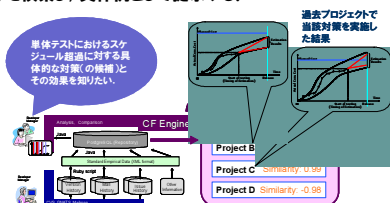
- 協調フィルタリング技術を用いて、プロジェクト特性を予測する方法を考案した。
 - 多様なソフトウェア開発プロジェクトを表現できるとするモデルを構築したりはしない。
 - 類似プロジェクトの情報を利用する。
 - データ欠損に強い。
- 欠損率60%のデータを用いたケーススタディでは、従来法よりも高い予測精度を示した。
- 正反対のプロジェクトの情報を利用することも可能である。

第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005



P³システム

- Progressive Project Profiling System
 - ソフトウェア開発プロジェクトの状況(特徴)をできるだけ正確, 詳細, リアルタイムに把握する.
 - 類似プロジェクトを検索し, 具体例として提示する.



第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005



参考文献

- 柿元健, 角田雅照, 大杉直樹, 門田暁人, 松本健一, “協調フィルタリングに基づく工数見積りロボラス性評価”, ソフトウェア工学の基礎XI, 日本ソフトウェア科学会 FOSE2004, pp.73-84, Nov. 2004.
- N. Ohsugi, M. Tsunoda, A. Monden, and K. Matsumoto, “Effort estimation based on collaborative filtering,” *Proc. of 5th International Conference on Product Focused Software Process Improvement (Profes2004)*, Kyoto, Japan, F. Bomarius and H. Iida (ed.), Lecture Notes in Computer Science, Vol.3009, pp.274-286, Apr. 2004.

第5回エンピリカルソフトウェア工学研究会, Feb. 8, 2005

