トップエスイー修了制作

複数のソフトウェアメトリクスにおける外れ値の分析

日本電気株式会社　日下裕貴　y-hinoshita@cj.jp.nec.com

開発における問題点

ソフトウェアの品質管理や品質向上を目的としたソフトウェアメトリクス(以下メトリクス)の測定において, 大部分の傾向とは異なるデータである外れ値はソースコードの構造上の問題を含む可能性があるため検出・分析すべきである.
しかし, データの相関分析を優先する一般の開発現場では外れ値の検出漏れが発生し得る.

手法・ツールの適用による解決

各種メトリクスの測定結果に対して箱ひげ図によりそれぞれ外れ値を検出.
統計解析ツールRにより無相関であると判断したメトリクス間において, データの相関分析時には検出されない外れ値の有無の調査とその評価を実施.

目標の設定

【仮説】外れ値の検出は**相関があるメトリクス間**において他の値から大きく外れた値を対象とする方法が**実際の開発現場では一般的**である

- **無相関のメトリクス間では外れ値**となるが, **相関があるメトリクス間では外れ値とならない**データの存在を調査
- 本調査で外れ値が検出された場合, その原因となるソースコードの構造上における問題を分析

分析手順

箱ひげ図による外れ値検出

箱ひげ図による各種メトリクスの外れ値

メトリクス	5数要約					外れ値(件数)
	最小値	第1四分位数	中央値	第2四分位数	最大値	
CBO	0	2	3.5	8	16	5
DIT	0	0	1	1	2	5
LCOM-CK	0	0	0.3	0.6	1	0
NOC	0	0	0	0	0	10
NPM	1	4	8	13	25	10
RFC	3	14	27	48	91	8
Ce	0	1	1	3	5	15
WMC	1	8	18	39	85	6

箱ひげ図により検出された外れ値(相関があるメトリクス間における外れ値は除く)が無相関であると判定されたメトリクスの組に存在するか調査

箱ひげ図による外れ値検出

各種メトリクス間の相関係数

	CBO	DIT	NOC	NPM	RFC	Ce	WMC
CBO	1	−0.31	−	0.32	0.49	0.05	0.31
DIT	−0.31	1	−	0.07	−0.00	0.11	0.07
NOC	−	−	−	−	−	−	−
NPM	0.32	0.07	−	1	0.79	0.34	0.71
RFC	0.49	−0.00	−	0.79	1	0.29	0.80
Ce	0.05	0.11	−	0.34	0.29	1	0.29
WMC	0.31	0.07	−	71	0.80	0.29	1

各種メトリクス間のp値と有意水準

メトリクスの組	p値	有意水準
CBO, Ce	0.63	0.27
DIT, NPM	0.54	0.27
DIT, RFC	0.97	0.20
DIT, Ce	0.316	0.312
DIT, WMC	0.51	0.27

分析結果

- 相関があるメトリクス間で外れ値とならないソースコードの中に他のメトリクスの観点で見ると外れ値となるソースコードが存在

無相関における外れ値の件数

メトリクスの組	件数
DITとCBO共に外れ値	1
DITとCe共に外れ値	3

- 本外れ値となるソースコードは結合クラスや依存クラスが共に深い階層で存在

各ソースコードにおけるメトリクスのデータ

ソースコード(ファイル名)	DIT	CBO	Ce
DataObjectFactory	3	23	1
ResponseListImpl	4	5	16
TwitterException	3	10	65
JSONException	3	0	46

まとめ

- 無相関のメトリクス間におけるデータを分析
- 本分析では, 相関があるメトリクス間で外れ値とならないソースコードの中に, **他のメトリクスの観点で見ると外れ値となるソースコードが存在し得る**ことが分かった
- 開発者は相関があるメトリクス間において検出される外れ値以外に, 各種メトリクスにおける外れ値について分析することが望ましい
- 今後の課題の1つとして, 本分析ではプロダクトデータを対象としたが, 開発プロセスにおけるデータについても本稿と同様の結論となるかの分析が挙げられる