ピアレビューによる評価の一致度の統計的分析方法

峯尾　真一

一般財団法人　高度情報科学技術研究機構

mineo@rist.or.jp

A Methodology for Statistical Analysis of Peer Review Agreement

Shinichi Mineo

Research Organization for Information Science & Technology

概要

HPCI(革新的ハイパフォーマンス・コンピューティング・インフラ)の課題選定においては、公正な選定プロセスを実現するために、ピアレビューによる課題評価方式を採用している。ピアレビューは、複数の専門家による多面的な評価を行うものであり、評価点はレビュアーの観点により変化する。そこでレビュアーの評価がどの程度一致するかを統計的に分析する方法として、重み付きコーエンのκ係数を提案し、数値シミュレーションによる検証を実施した。

1　はじめに

HPCI(革新的ハイパフォーマンス・コンピューティング・インフラ)の課題選定においては、公正な選定プロセスを実現するためにピアレビューを実施している。基本的には対象課題と同じ分野の研究者を主とする複数のレビュアーに対して課題の審査を依頼し、その評価点を基に当該課題の評価を行っている。

ピアレビューは、複数の専門家による多面的な評価を行うものであり、評価点はレビュアーの観点により変化する。本論文ではレビュアーの評価の一致度を統計的に測定する指標を検討する。

2　一般的な評価の一致度とその問題点

2.1 一般的な評価の一致度

例として、レビュアーA~F の 6 人が課題 hp01~hp03 を 3 段階(1~3)で評価した結果を表１．に示す。この場合、課題 hp01 を A,B,C,D の 4 人が担当したことになる。

この表から評価者１（レビュアーA）と評価者２（他のレビュアーB~F）との評価結果の対を作成すると表２．の１０通りとなる。これをクロス集計表にしたものが表３．である。すなわち、例えば評価者１が「１」評価をした時に他の評価者が同じ評価をした回数が 2 回であったことが分かる。

さらにこの表から評価者１から見た一般的な評価の一致度を計算すると、１０通りの内、評価が一致したのは対角線上の 2+3+2=7 通りとなり、一致度は 7/10=70%となる。

表１．課題とレビュアーの対応例

	評価者１	評価者2				
課題	A	B	C	D	E	F
hp01	1	1	1	3		
hp02	2		1	2	2	2
hp03	3	2		3		3

表２．評価結果

case	評価者１	評価者２
1	1	1
2	1	1
3	1	3
4	2	1
5	2	2
6	2	2
7	2	2
8	3	2
9	3	3
10	3	3

表３．クロス集計表

	評価者２				
評価者１	評価	1	2	3	小計
	1	2		1	3
	2	1	3		4
	3		1	2	3
	小計	3	4	3	10

2.2 一致度計算の問題点

この様な一致度の計算には２つ問題点が存在する。第一に、評価点が順序尺度[1]であり、担当する課題も異なることから、完全に評価点が一致しなくても、近い評価点であればある程度は評価が一致していると考えられることである。

また第二に、計算された評価の一致度が、真のレビュアーの意志の外に、偶然に評価が一致した場合が含まれていることである。

3 新しい指標の導入

3.1 指標の選定

レビュアー毎に、同じ課題を担当した他の全てのレビュアーとの判断の一致度を重み付きコーエンの κ 係数により測定する。

コーエンの κ 係数とは、臨床心理学や社会心理学において人間の判断の一致度を測るために考案されたもの[2]であり、病理診断における医師の判断等の分析にも利用されている[3]。

3.2 重み付きコーエンの κ 係数の計算方法

そこで、このクロス集計表に対して、図２のような演算を行うことにより、 κ を計算する。ここで、Pow は、評価者１と評価者２の判断の一致率、Pew は、偶然による両者の判断の一致率を示す。

Pew は、率に直した評価数表 Po の小計の直積である表 Pe から計算される。例えば、表 Po において、一般的傾向として評価者１が「１」を付ける割合は、小計の 0.3、同じく評価者２が「１」を付ける割合は、小計の 0.3 であるから、両者が「１」を偶然に付ける割合は、0.3 x 0.3 = 0.09 となり表 Pe の(1,1)欄の値となる。

２次の重み W は Po と Pe の各要素の同位置に掛ける値を示し、対角線上の完全一致に対して「１」、それから離れるに従い 0.75、0 としている。

すなわち κ =0.5(50%)は、評価の近さに重みを付けて足し合わせ、かつ偶然による両者の判断の一致の影響を引いたものと考えられる。

m(=3)段階評価の例（評価数）

	評価者２			
評価者１	評価	1	2	3	小計
	1	2		1	3
	2	1	3		4
	3		1	2	3
	小計	3	4	3	10

⇩ 評価数を率に直す

Po	評価者２			
評価者１	評価	1	2	3	小計
	1	0.2		0.1	0.3
	2	0.1	0.3		0.4
	3		0.1	0.2	0.3
	小計	0.3	0.4	0.3	1

⇩

Pe	評価者２			
評価者１	評価	1	2	3	
	1	0.09	0.12	0.09	0.3
	2	0.12	0.16	0.12	0.4
	3	0.09	0.12	0.09	0.3
		0.3	0.4	0.3	1

例えば(1,1)は 0.3*0.3=0.09

* W

２次の重み(W)

W	評価者２		
評価者１	評価	1	2	3
	1	1	0.75	0
	2	0.75	1	0.75
	3	0	0.75	1

$$W_{i,j} = 1 - \left(\frac{i-j}{m-1}\right)^2$$

Po*W	評価者２		
評価者１	評価	1	2	3
	1	0.2	0	0
	2	0.08	0.3	0
	3	0	0.08	0.2

Pow=Σ(要素)=0.85

Pe*W	評価者２		
評価者１	評価	1	2	3
	1	0.09	0.09	0
	2	0.09	0.16	0.09
	3	0	0.09	0.09

Pew=Σ(要素)=0.7

$$\kappa = \frac{Pow - Pew}{1 - Pew} = 0.5$$

図２．コーエンの κ の計算方法

3.3 コーエンのκ係数の数値シミュレーション

コーエンのκ係数の測定結果がどの様な意味を持つかを、乱数を用いた数値シミュレーションの採点試行により確認する。

採点試行の条件は次の通りとする。（アルゴリズムを図３に示す）

- 評価対（あるレビュアーの評価点と、同じ課題を評価した他のレビュアーの評価点の対）：20
- 合意率（判断が完全に一致する確率）ｐ：0.0-1.0(0.1 毎に採点試行)
- 完全一致以外は無作為に 1-5 と採点
- 採点試行当たりのレビュアー数：50,000(累計の評価対は 1,000,000 となる)

採点試行の結果を表４に示す。この結果から、合意率とκ係数の平均値はほぼ等しいことが分かる。すなわち、コーエンのκ係数は、帰納法的な解釈をすれば、判断が完全に一致する確率を示すと言える。

3.4 コーエンのκ係数による分析の問題点

本係数の特長は、観測された判断の一致率から偶然による判断の一致の影響を引いて補正するという考え方である。

表４．採点試行の結果

合意率	試行回数	平均	最大	最小	標準偏差
0	50000	0.000803	0.768786	-0.88571	0.216653
0.1	50000	0.094692	0.842932	-0.75029	0.222777
0.2	50000	0.189505	0.916045	-0.83894	0.227046
0.3	50000	0.286439	0.964455	-0.71558	0.225743
0.4	50000	0.384386	1	-0.69975	0.221374
0.5	50000	0.483529	1	-0.51198	0.213634
0.6	50000	0.58307	1	-0.38243	0.198945
0.7	50000	0.687119	1	-0.28247	0.179708
0.8	50000	0.789092	1	-0.15055	0.153498
0.9	50000	0.893682	1	0.038462	0.112842
1	50000	1	1	1	1.59E-16

この偶然による判断の一致率は、評価者毎の評価点の分布が、その評価者の一般的な評価傾向（例えば評価点１を付ける割合）であるという前提から算出されている。しかし、ここで扱っている問題では、評価者２は、一人の人間ではなく、評価者１と同じ課題を審査したレビュアーの総体となっている。

もし一般的な評価傾向が、個人の人間に由来するものであった場合、かならずしも適切な計算方法とは言えない。また、評価する課題が少ない場合も、一般的な評価傾向を計算する場合の誤差が大きくなると考えられる。

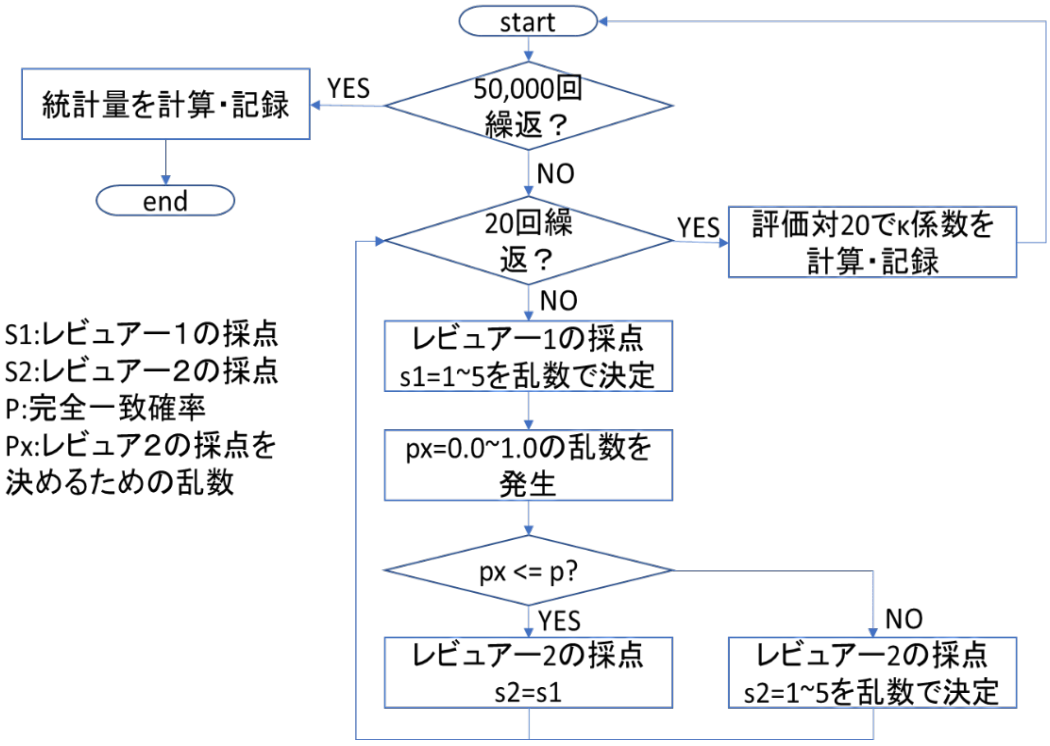

図３．採点試行のアルゴリズム

4　指標の活用方法に関する考察

このκ係数は、3.4 で述べた問題があるため、個別のレビュアーの特性分析に利用することは必ずしも適切ではない。しかし、レビュアー全員のκ係数の平均値や分散は、評価点の平均や分散と共に評価項目の特性分析に活用できると考える。

すなわち、もしκ係数（評価の一致度）の平均値が高く、かつ評価点の平均が極端に高いか低い場合、この評価項目には何らかのバイアスが存在する可能性を示す。

また逆にκ係数（評価の一致度）が低く、かつ評価点の分散が大きい場合、その評価項目自体の適切性を見直す必要があると言える。

5　まとめ

重み付きコーエンのκ係数という指標により、ピアレビューの評価の一致度を測定する方法を検討し、その指標を用いて評価項目の特性分析を行う可能性を述べた。

今後は、実際の採点データに対して本解析を実施し、HPCI におけるピアレビューの実績データを集積するとともに、公正な選定プロセスを維持する一助にして行きたいと考える。

謝辞

本論文の作成にあたり的確な助言を頂いた一般財団法人高度情報科学技術研究機構の皆様に感謝します。

参考文献

[1] S.S.Stevens, On the Theory of Scales of Measurement, SCIENCE Vol.103, No. 2684, June 7,1946

[2] Jacob Cohen, A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement Vol. XX, No.1,1960

[3] Harold L.Kundel,Marcia Polansky, Measurement of Observer Agreement, Radiology Vol. 228, 303-308, August 2003