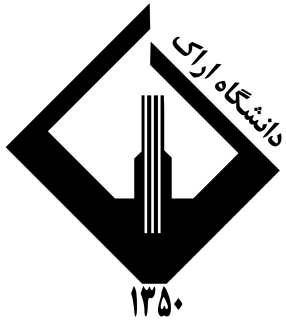


عنوان ارائه:

معیارهای ارزیابی مبتنی بر ارتباط برای دامنه‌های نامتعادل چند رده‌ای

Relevance-based Evaluation Metrics for Multi-class Imbalanced Domains

توسط: علیرضا صادقی نسب

استاد: دکتر حسین غفاریان

تاریخ ارائه: ۱۳۹۹/۹/۱۲

مقدمه

- اطلاعات مقاله

عنوان: *Relevance – based Evaluation Metrics for Multi – class Imbalanced Domains*

نویسندگان: *Paula Branco, Luis Torgo, and Rita P . Ribeiro*

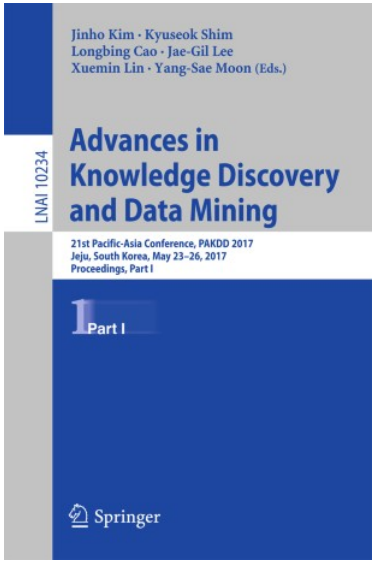
عنوان کتاب: *Pacific – Asia Conference on Knowledge Discovery and Data Mining*

سال چاپ: 2017

ناشر: *Springer*

تعداد ارجاع: 21

پیاده‌سازی: دارد (به صورت عمومی در دسترس است)



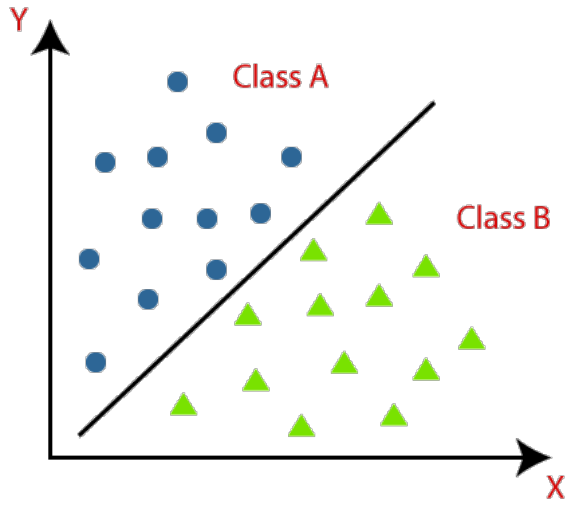
https: //github.com/paobranco/Relevance-basedMulticlassImbalanceMetrics.

فهرست مطالب

- مقدمه
- معرفی روش
- ارزیابی روش
- بررسی نقاط قوت و ضعف

مقدمه

- رده‌بندی یا طبقه‌بندی
  - ☑ یک نوع یادگیری تحت نظارت است
  - ☑ در فرآیند طبقه‌بندی، رده‌ای برای عناصر مشخص می‌شود که به آن‌ها تعلق داشته باشد و هنگامی که خروجی مقادیر متناهی و گسسته دارد، بهترین استفاده را داشته باشد
  - ☆ انواع رده‌بندی:
    - ☆ دو جمله‌ای
    - ☆ چند رده‌ای یا چندجمله‌ای



مقدمه

- رده‌بندی دوجمله‌ای

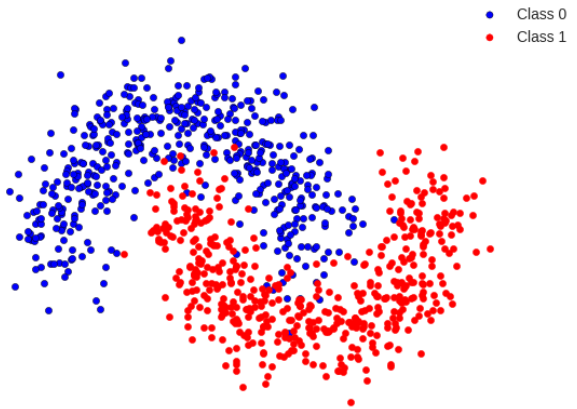
☑ در این نوع، عناصر یک مجموعه به دو گروه براساس قانون طبقه‌بندی، تقسیم می‌شوند

☑ مثال‌ها و کاربردها:

☆ آزمون بیماری: برای فهمیدن داشتن/نداشتن بیماری

☆ کنترل کیفیت در صنعت: تصمیم‌گیری در مورد برآورده شدن مشخصات

☆ بازیابی اطلاعات: تصمیم‌گیری در مورد بودن/نبودن یک صفحه در نتایج جستجو



مقدمه

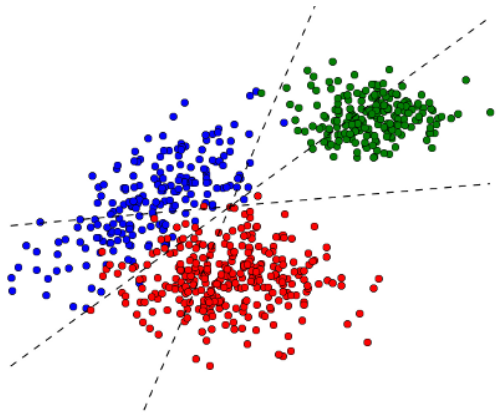
- رده‌بندی چندجمله‌ای

☑ در این نوع، عناصر یک مجموعه به تعداد بیش از دو گروه براساس قانون طبقه‌بندی، تقسیم می‌شوند

☑ مثال‌ها و کاربردها:

☆ دسته‌بندی حیوانات دانشنامه ویکیپدیا

☆ دسته‌بندی گواهینامه سنی فیلم‌های سایت imdb



مقدمه

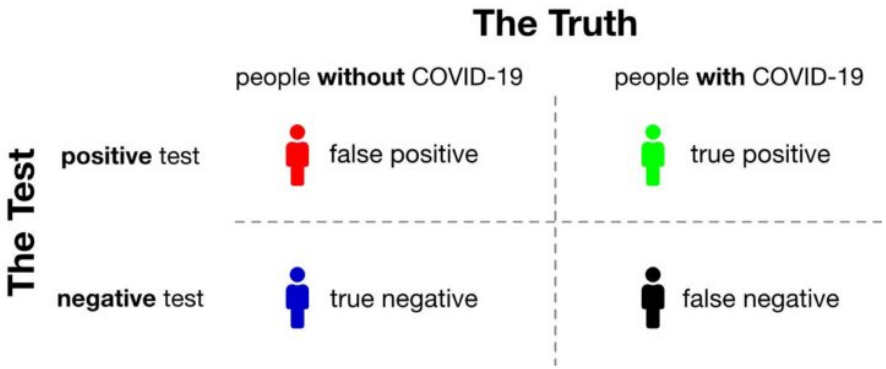
- ماتریس سردرگمی
  - ☑ جدولی است که برای شرح کارائی مدل رده‌بندی به کار می‌رود
  - ☑ هر ردیف در ماتریس سردرگمی نشانگر یک رده واقعی است
  - ☑ هر ستون در ماتریس سردرگمی نشانگر یک رده پیشبینی شده است

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

مقدمه

- تعریف برخی مفاهیم پایه با مثال تشخیص بیماری
  - ☆ مثبت واقعی (TP): مدل تشخیص بیماری داده و فرد آن بیماری را دارد
  - ☆ منفی واقعی (TN): مدل تشخیص صحت داده و فرد نیز آن بیماری را ندارد
  - ☆ مثبت کاذب (FP): مدل تشخیص بیماری داده ولی فرد آن بیماری را ندارد
  - ☆ منفی کاذب (FN): مدل تشخیص بیماری نداده ولی فرد آن بیماری را دارد



مقدمه

- برای رده i، معیارهای زیر را می‌توان تعریف کرد:

$$recall_i = \frac{t_{p_i}}{t_i}$$

$$precision_i = \frac{t_{p_i}}{p_i}$$

$$F_{\beta i} = \frac{(1+\beta^2) precision_i \,.\, recall_i}{\beta^2 \,.\, precision_i + recall_i}$$

مقدمه

Metric	Description	Definition
$AvAcc$	Classes average accuracy.	$\frac{1}{C}\sum_{i=1}^{C}\frac{tp_i+tn_i}{tp_i+tn_i+fp_i+fn_i}$
$MAvG$	Geometric average of recall in each class [15].	$\sqrt[C]{\prod_{i=1}^{C} recall_i}$
Rec_M	Arithmetic Macro-average of recall in each class.	$\frac{1}{C}\sum_{i=1}^{C} recall_i$
$Prec_M$	Arithmetic Macro-average of precision in each class.	$\frac{1}{C}\sum_{i=1}^{C} precision_i$
Rec_μ	Arithmetic Micro-average of recall in each class.	$\sum_{i=1}^{C} tp_i / \sum_{i=1}^{C} t_i$
$Prec_\mu$	Arithmetic Micro-average of precision in each class.	$\sum_{i=1}^{C} tp_i / \sum_{i=1}^{C} p_i$
$F_{\beta M}$	Mean F_β measure evaluated with Macro-averaged precision and recall [14].	$\frac{(1+\beta^2)\cdot Prec_M \cdot Rec_M}{\beta^2 \cdot Prec_M + Rec_M}$
$F_{\beta \mu}$	Mean F_β measure evaluated with Micro-averaged precision and recall [14].	$\frac{(1+\beta^2)\cdot Prec_\mu \cdot Rec_\mu}{\beta^2 \cdot Prec_\mu + Rec_\mu}$
AvF_β	Extension for any value of β of the definition for F_1 measure to multi-class [4].	$\frac{1}{C}\sum_{i=1}^{C}\frac{(1+\beta^2)\cdot precision_i \cdot recall_i}{\beta^2 \cdot precision_i + recall_i}$
CBA	Class Balance Accuracy [12].	$\frac{\sum_{i=1}^{C}\frac{mat_{i,i}}{max\left(\sum_{j=1}^{C} mat_{i,j}, \sum_{j=1}^{C} mat_{j,i}\right)}}{C}$
MCC	Matthews Correlation Coefficient introduced for two-class problems and extended to multi-class [11, 6].	$\frac{X}{YZ}$, where $X=\sum_{k,l,m=1}^{C}(mat_{k,k}mat_{m,l}-mat_{l,k}mat_{k,m})$ $Y=\sqrt{\sum_{k=1}^{C}\left(\sum_{l=1}^{C} mat_{l,k}\right)\left(\sum_{\substack{f,g=1\\f\neq k}}^{C} mat_{g,f}\right)}$ $Z=\sqrt{\sum_{k=1}^{C}\left(\sum_{l=1}^{C} mat_{k,l}\right)\left(\sum_{\substack{f,g=1\\f\neq k}}^{C} mat_{f,g}\right)}$
RCI	Relative Classifier Information [13]	$\frac{H_d-H_o}{H_d}$, where $H_d=-\sum_{i=1}^{C}\left(\frac{\sum_{l=1}^{C} mat_{i,l}}{C} log \frac{\sum_{l=1}^{C} mat_{i,l}}{C}\right)$ $H_o=\sum_{j=1}^{C}\left(\frac{\sum_{k=1}^{C} mat_{k,j}}{C} H_{oj}\right)$ and $H_{oj}=-\sum_{i=1}^{C}\left(\frac{mat_{i,j}}{\sum_{k=1}^{C} mat_{k,j}} log \frac{mat_{i,j}}{\sum_{k=1}^{C} mat_{k,j}}\right)$
CEN	Confusion Entropy [16].	$\sum_{j=1}^{C}(P_j CEN_j)$, where $P_j=\frac{\sum_{k=1}^{C} mat_{j,k}+mat_{k,j}}{2*\sum_{k,l=1}^{C} mat_{k,l}}$, $CEN_j=-\sum_{\substack{k=1\\k\neq j}}^{C}(P_{j,k}^{j}\log_{2(C-1)}(P_{j,k}^{j})+P_{k,j}^{j}\log_{2(C-1)}(P_{k,j}^{j}))$ $P_{i,i}^{i}=0,\quad P_{i,j}^{i}=mat_{i,j}/\left(\sum_{k=1}^{C}(mat_{i,k}+mat_{k,i})\right), i\neq j$

- معیارهایی که در پژوهش، مورد بررسی قرار گرفته‌اند:

☑ روش‌های مبتی بر *recall*، بر روی برچسب‌های *true* تمرکز دارند

☑ روش‌های مبتنی بر *precision*، برچسب رده‌های پیشبینی شده را در نظر می‌گیرند

☑ روش‌های عمومی، هر دو دیدگاه را تجمیع کرده و به صورت یک مقدار واحد درمی‌آورند

مقدمه

- مسائل نامتعادل: کاربر تمایز متفاوتی در رده‌های مسئله دارد
- راهکارهای نامبرده شده، برای این گونه دامنه‌ها مناسب نیستند زیرا منعکس کننده ترجیحات کاربر نیستند و می‌توانند نتایج گمراه‌کننده داشته باشند
- مسائل نامتعادل چند رده‌ای را می‌توان به سه گروه چند اقلیتی، چند اکثریتی و کامل تقسیم کرد
- در حالت چند اقلیتی، یک رده به طور قابل توجهی بیشتر از میانگین تعداد نمونه همه رده‌ها، نمونه دارد
- در حالت چند اکثریتی، یک رده به طور قابل توجهی کمتر از بقیه است
- در حالت کامل، چندین رده به طور قابل توجهی بزرگتر از چندین رده‌ای هستند که کوچکتر از میانگین هستند

مقدمه

- در جدول زیر، ۳ نمونه به ترتیب برای مثال‌های چند اقلیتی، چند اکثریتی و کامل به همراه معیارهای $rec_i, prec_i \ and \ F_{1i}$ به تصویر کشیده شده‌اند
- در نمونه اول، معیارهای رده اکثریت، در نمونه دوم، معیارهای رده اقلیت و در نمونه سوم، معیارهای رده اقلیت، گمراه‌کننده هستند

Case 1

trues \ preds	c_1	c_2	c_3
c_1	5	0	0
c_2	0	10	0
c_3	0	300	0

Class	rec_i	$prec_i$	F_{1i}
c_1	1	1	1
c_2	1	0.032	0.063
c_3	0	n. def.	n. def.

Case 2

trues \ preds	c_1	c_2	c_3
c_1	1	0	3
c_2	0	100	0
c_3	0	0	200

Class	rec_i	$prec_i$	F_{1i}
c_1	0.25	1	0.4
c_2	1	1	1
c_3	1	0.985	0.993

Case 3

trues \ preds	c_1	c_2	c_3	c_4
c_1	1	3	0	0
c_2	9	1	0	0
c_3	0	0	100	0
c_4	0	0	0	200

Class	rec_i	$prec_i$	F_{1i}
c_1	0.25	0.1	0.14
c_2	0.1	0.25	0.14
c_3	1	1	1
c_4	1	1	1

مقدمه

* جدول زیر، مقادیر به دست آمده از الگوریتم‌های پیشین را برای سه نمونه مذکور، نشان می‌دهد. ستون سوم میزان تطابق با تنظیمات کاربر را نشان می‌دهد

* معیارها نتوانستند تنظیمات کاربر را به درستی نشان دهند و مقادیری بیشتر یا کمتر از حد تخمین زده شده را ارائه کرده‌اند

Metric	Case 1			Case 2			Case 3		
	N.Val.(%)	Value	Ac.	N.Val.(%)	Value	Ac.	N.Val.(%)	Value	Ac.
$AvAcc$	36.5	0.365	×	99.3	0.993	×	98.1	0.981	×
$MAvG$	0.0	0.000	×	63.0	0.630	✓	39.8	0.398	✓
Rec_M	66.7	0.667	✓	75.0	0.750	×	58.8	0.588	×
$Prec_M$	not defined		×	99.5	0.995	✓	58.8	0.588	×
Rec_μ	4.8	0.048	×	99.0	0.990	×	96.2	0.962	×
$Prec_\mu$	4.8	0.048	×	99.0	0.990	✓	96.2	0.962	×
F_{1M}	not defined		×	85.5	0.855	×	58.8	0.588	×
$F_{1\mu}$	4.8	0.048	×	99.0	0.990	×	96.2	0.962	×
AvF_1	not defined		×	79.8	0.798	×	57.1	0.571	×
CBA	34.4	0.344	×	74.5	0.745	✓	55.0	0.550	×
MCC	65.1	0.301	✓	98.9	0.978	×	96.2	0.923	×
RCI	36.8	0.368	×	92.6	0.926	×	97.9	0.979	×
CEN	97.8	0.022	✓	98.1	0.019	×	98.5	0.015	×

معرفی روش

- روش پیشنهادی مبتنی بر این فرض است که رده‌ها، ارتباط متفاوتی برای کاربر دارند
- ایده اصلی این است که هنگام ارزیابی عملکرد مدل‌ها، از مقادیر ارتباطی جهت وزن‌دهی رده‌ها بهره‌برداری شود. این مقدار ارتباطی توسط تابعی به نام $\phi()$ به دست می‌آید
- در روش پیشنهادی، فرض بر این است که کاربر به هر رده مسئله، یک امتیاز اختصاص می‌دهد

$$Rec^{\phi} = \frac{1}{\sum_{i=1}^{C} \phi(i)} \sum_{i=1}^{C} \phi(i) \cdot recall_i \qquad Prec^{\phi} = \frac{1}{\sum_{i=1}^{C} \phi(i)} \sum_{i=1}^{C} \phi(i) \cdot precision_i$$

$$F_{\beta}^{\phi} = \frac{(1+\beta^2) \cdot Prec^{\phi} \cdot Rec^{\phi}}{(\beta^2 \cdot Prec^{\phi}) + Rec^{\phi}} \qquad CBA^{\phi} = \sum_{i=1}^{C} \phi(i) \cdot \frac{mat_{i,i}}{max\left(\sum_{j=1}^{C} mat_{i,j}, \sum_{j=1}^{C} mat_{j,i}\right)}$$

$$AvF_{\beta}^{\phi} = \frac{1}{\sum_{i=1}^{C} \phi(i)} \sum_{i=1}^{C} \frac{\phi(i) \cdot (1+\beta^2) \cdot precision_i \cdot recall_i}{(\beta^2 \cdot precision_i) + recall_i} = \frac{1}{\sum_{i=1}^{C} \phi(i)} \sum_{i=1}^{C} \frac{\phi(i) \cdot (1+\beta^2) \cdot tp_i}{\beta^2 \cdot t_i + p_i}$$

معرفی روش

- روش‌های تخمین ارتباط

- چارچوب ارزیابی (اسلاید قبل)، کاملاً به دسترس بودن اطلاعات دامنه در رابطه با رده‌ها بستگی دارد اما این اطلاعات می‌توانند در سطوح متفاوتی، وجود داشته باشند

  - غیررسمی: هنگامی مطرح می‌شود که هیچ کمیتی در مورد اهمیت هر کلاس، وجود نداشته باشد. غالباً فقط گفته می‌شود که "طبقات اقلیت، مهترین هستند". در این حالت، مشکلات جدی برای ارزیابی عملکرد ایجاد می‌شود

  - غیررسمی متوسط: در این حالت، اطلاعات بیشتری (اما محدود) وجود دارد. فرض بر این شده است که در این حالت، کاربر نظم جزئی از رده‌ها با توجه به اهمیتشان را ارائه می‌دهد

  - رسمی متوسط: مانند حالت بالا است اما نظم کامل رده‌ها ارائه شده است

  - رسمی: اطلاعات تابع ارتباطی تکمیل است. این حالت در دامنه‌های واقعی، مرسوم نیست

معرفی روش

- روش‌های تخمین ارتباط (ادامه)

- اگر کاربر به طور کامل، تابع ارتباط را مشخص کند (سطح رسمی)، آنگاه نیازی به ارائه راهکار تخمینی نداریم

- در حالت غیررسمی، از آنجایی که کاربر هیچ تنظیماتی را ارائه نکرده است، راهکار تخمینی صرفا براساس میزان تکرر رده خواهد بود:

$$\hat{\phi}(i) = \frac{1/t_i}{\sum_{i=1}^{C} 1/t_i}$$

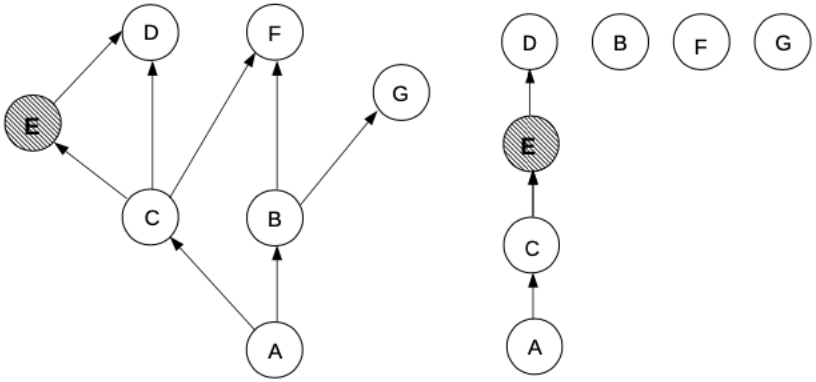
- در حالت غیررسمی متوسط، از نظم جزئی و مدل نظم جزئی محلی استفاده شده است. تابع امتیاز هر گره در این مدل به صورت $Rank(x) = |S| + 1 + \frac{|U|}{2}$ تعریف شده است و در نهایت، تابع ارتباط مشخص می‌شود:

$$\hat{\phi}(i) = \frac{Rank(i)}{\max_{\forall i \in C} Rank(i)}$$

معرفی روش

- روش‌های تخمین ارتباط (ادامه)

☑ یک مجموعه منظم شده جزئی و مدل محلی آن را به عنوان مثال می‌توان به صورت زیر نشان داد:



$$Rank(E) = |S| + 1 + \frac{|U|}{2} = 2 + 1 + 1.5 = 4.5$$

معرفی روش

- روش‌های تخمین ارتباط (ادامه)

- در حالت رسمی متوسط، از نظم کامل رده‌های مسئله استفاده شده است. در این حالت نیز از همان تابع امتیازدهی استفاده شده است با این تفاوت که دیگر گره غیرقابل مقایسه وجود نخواهد داشت و در نتیجه: $Rank(x) = |S| + 1$

- تابع ارتباط در حالت رسمی متوسط مانند حالت غیررسمی متوسط خواهد بود و بازه آن از $\frac{1}{C}$ تا 1 خواهد بود

- با توجه به اینکه در شرایطی، معیارهای $recall_i$ یا $precision_i$ مقدار غیرقابل تعریف را نشان می‌دهند؛ بنابراین در این حالات، از AvF_1^{ϕ} بهره برده شده است

$$AvF_{\beta}^{\phi} = \frac{1}{\sum_{i=1}^{C}\phi(i)} \sum_{i=1}^{C} \frac{\phi(i)\cdot(1+\beta^2)\cdot precision_i \cdot recall_i}{(\beta^2\cdot precision_i)+recall_i} = \frac{1}{\sum_{i=1}^{C}\phi(i)} \sum_{i=1}^{C} \frac{\phi(i)\cdot(1+\beta^2)\cdot tp_i}{\beta^2\cdot t_i+p_i}$$

ارزیابی روش

- در جدول زیر، میزان ارتباط تعریف شده کاربر و ارتباط استنباط شده از روش‌های تخمینی گفته شده، نشان داده شده است

	Case 1				Case 2				Case 3				
	$\phi(c_1)$	$\phi(c_2)$	$\phi(c_3)$	order	$\phi(c_1)$	$\phi(c_2)$	$\phi(c_3)$	order	$\phi(c_1)$	$\phi(c_2)$	$\phi(c_3)$	$\phi(c_4)$	order
PREV	0.66	0.33	0.01		0.94	0.04	0.02		0.64	0.32	0.03	0.02	
PO	1	1	0.4	$c_3 < c_1$ $c_3 < c_2$	1	0.5	0.5	$c_3 < c_1$ $c_2 < c_1$	1	0.86	0.57	0.42	$c_3 < c_1$ $c_4 < c_1$ $c_4 < c_2$
TO	1	0.67	0.33	$c_3 < c_2 < c_1$	1	0.67	0.33	$c_3 < c_2 < c_1$	1	0.75	0.5	0.25	$c_4 < c_3 < c_2 < c_1$
ϕ	1	0.9	0.1		1	0.2	0.1		1	0.9	0.2	0.1	

- جدول زیر نیز کارائی به دست آمده براساس معیارهای گوناگون را نشان می‌دهد

Metric	Case			Metric	Case			Metric	Case		
	1	2	3		1	2	3		1	2	3
$AvAcc$	36.5	99.3	98.1	Rec^{PREV}	**98.9**	**29.2**	**24**	Rec^{TO}	**83.3**	**62.5**	**43**
$MAvG$	0	**63**	**39.8**	$Prec^{PREV}$	**67.7**	**100**	**17.8**	$Prec^{TO}$	**61.3**	**99.8**	**41.5**
Rec_M	**66.7**	75	58.8	F_1^{PREV}	**80.4**	**45.3**	**20.4**	F_1^{TO}	**70.6**	**76.9**	**42.2**
$Prec_M$	34.4 [a]	**99.5**	58.8	AvF_1^{PREV}	**68**	**43.4**	**17.8**	AvF_1^{TO}	**52.1**	**69.9**	**40**
Rec_μ	4.8	99	96.2	CBA^{PREV}	**67**	**29.2**	**13.7**	CBA^{TO}	**51.1**	**62.3**	**37**
$Prec_\mu$	4.8	**99**	96.2	Rec^{PO}	**83.3**	**62.5**	**46.7**	Rec^{ϕ}	**95**	**42.3**	**29.1**
F_{1M}	45.4 [a]	85.5	58.8	$Prec^{PO}$	**51.6**	**99.6**	**46**	$Prec^{\phi}$	**54.2**	**99.9**	**28.4**
$F_{1\mu}$	4.8	99	96.2	F_1^{PO}	**63.7**	**76.8**	**46.4**	F_1^{ϕ}	**69**	**59.4**	**28.7**
AvF_1	35.4 [a]	79.8	57.1	AvF_1^{PO}	44.3	**69.8**	**44.3**	AvF_1^{ϕ}	**52.8**	**53.8**	**26**
CBA	34.4	**74.5**	55	CBA^{PO}	43	**62.1**	**41.5**	CBA^{ϕ}	**51.5**	**42.2**	**22.3**
MCC	**65.1**	98.9	96.2								
RCI	36.8	92.6	97.9								
CEN	**97.8**	98.1	98.5								

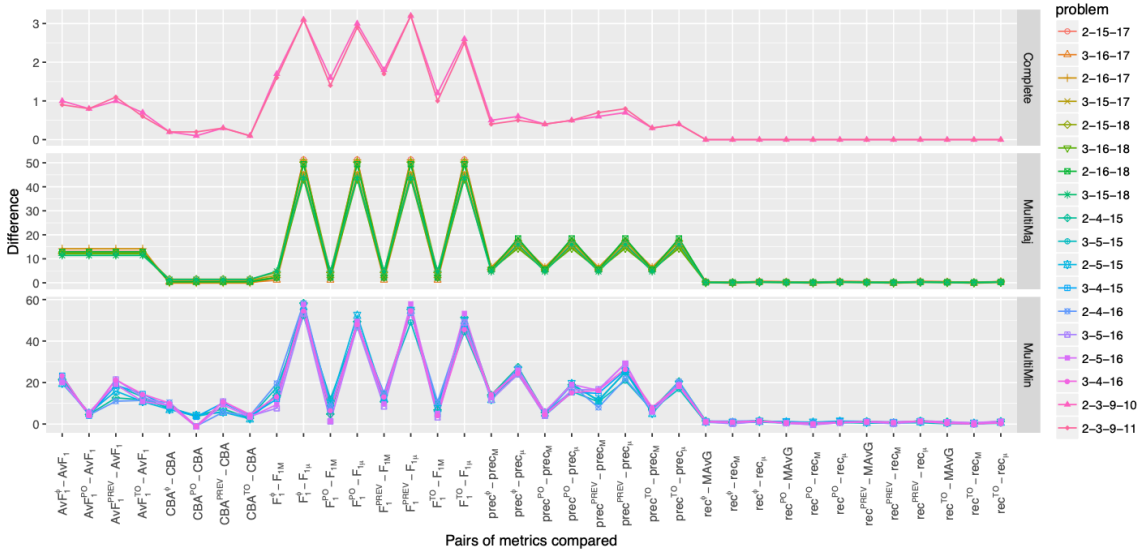
ارزیابی روش

- توانایی فرق‌گذاری

☑ سناریوهای چند اقلیتی، چند اکثریتی و کامل آزموده شده است

☑ تعداد رده‌ها در این سناریوها به ترتیب ۳، ۳ و ۴ بوده است

☆ شکل زیر تفاوت درصد فرق‌گذاری جفت معیارها (معیار براساس ارتباط و معیار موجود) را نشان می‌دهد:



بررسی نقاط قوت و ضعف

- نقاط قوت
  - معیار مبتنی بر ارتباط ارائه شده برای دامنه‌های نامتعادل مناسبتر هستند زیرا تنظیمات کاربر را نیز بازتاب می‌کنند
  - افزایش توانایی فرق‌گذاری نسبت به همان معیار بدون در نظر گرفتن تابع ارتباط‌دهی
  - پیاده‌سازی و انتشار عمومی کدها و ارزیابی‌ها
- نقاط ضعف
  - نتایج به دست آمده صرفاً مبتنی بر مثال است

با تشکر از توجه شما