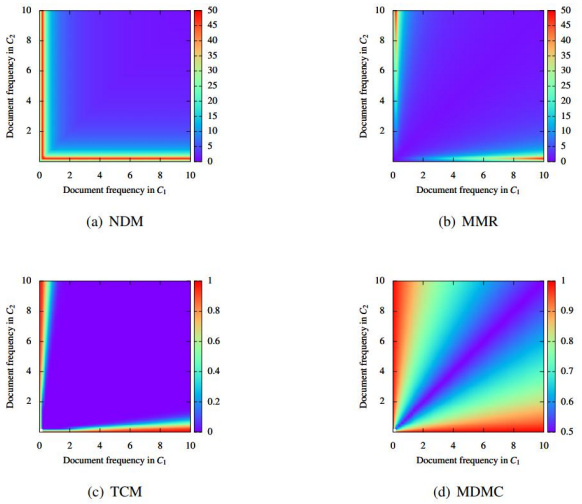
Max-difference maximization criterion: A feature selection method for text categorization

Lingbin JIN, Li ZHANG, Lei ZHAO

Frontiers of Computer Science, DOI: 10.1007/s11704-022-2154-x

Problems & Ideas

- Problems of approaches based on ACC2:
  - ACC2 incorrectly equally treats terms with the same document rate difference but different discrimination.
  - Existing improved methods (NDM, MMR and TCM) based on ACC2 may confuse the importance of rare and sparse terms on account of challenge for parameter selection.
- Ideas: A new weight without parameter for ACC2 by applying category information occupancies is proposed in MDMC.



(a) NDM (b) MMR

(c) TCM (d) MDMC

Weight distribution of a term obtained by four metrics (NDM, MMR, TCM and MDMC) with two classes C_1 and C_2 .

Main Results

- Results:
  - It can be seen that MDMC is capable of catching more discriminant terms without any parameter than other filter ones regardless of classifier, and shows its superiority over other dimensionality reduction methods (ISCA, PCA and NMF).

Comparison of average Micro-F1 (%) for filter algorithms

Classifier	Dataset	OR	MI	MMR	NDM	RDC	EFS	TCM	TRDC	TNDM	TMMR	TTCM	MDMC
SVM	K1a	47.28	58.23	72.35	68.12	64.25	69.32	70.44	68.40	73.38	74.65	74.33	**75.94**
	K1b	79.41	82.49	92.47	87.26	86.80	93.62	92.39	89.03	92.70	93.80	93.85	**95.15**
	La2	53.49	43.48	71.13	63.69	68.30	76.26	69.33	70.82	71.74	76.42	74.43	**78.31**
LR	K1a	48.11	58.54	72.50	68.79	64.31	69.33	71.08	68.32	73.65	74.64	74.63	**75.67**
	K1b	81.17	82.06	92.91	88.02	87.15	93.78	93.03	89.24	93.10	93.90	94.23	**95.46**
	La2	54.93	44.85	72.26	65.93	68.88	76.66	70.78	71.31	73.00	77.41	76.15	**78.73**

Comparison of average Macro-F1 (%) for filter algorithms

Classifier	Dataset	OR	MI	MMR	NDM	RDC	EFS	TCM	TRDC	TNDM	TMMR	TTCM	MDMC
SVM	K1a	28.94	42.15	55.39	49.97	47.64	52.94	52.68	52.24	56.22	57.53	57.19	**60.54**
	K1b	52.97	68.83	84.89	73.41	76.48	86.83	79.84	80.25	80.76	84.15	83.71	**90.62**
	La2	37.57	32.66	61.29	52.81	62.85	69.56	58.97	65.68	63.23	70.50	68.30	**73.63**
LR	K1a	28.95	41.85	55.29	50.16	47.49	52.56	52.97	52.07	55.90	57.27	57.62	**59.83**
	K1b	53.38	68.69	85.33	74.34	76.97	87.18	80.61	81.02	81.57	84.48	83.96	**91.06**
	La2	38.49	35.17	62.77	55.24	63.84	70.36	60.92	66.34	64.83	71.95	70.49	**74.26**

Comparison of max Micro-F1 (%) for more methods

Classifier	Dataset	ISCA	PCA	NMF	MDMC
SVM	K1a	81.67	82.52	74.87	**87.14**
	K1b	94.94	94.44	92.65	**98.29**
	La2	83.35	85.79	77.88	**87.77**
LR	K1a	82.95	84.66	73.93	**86.84**
	K1b	97.61	96.15	92.31	**98.42**
	La2	84.68	87.61	80.03	**89.40**

Comparison of max Macro-F1 (%) for more methods

Classifier	Dataset	ISCA	PCA	NMF	MDMC
SVM	K1a	66.68	70.48	51.55	**76.08**
	K1b	96.71	87.97	76.17	**97.27**
	La2	80.52	83.56	71.86	**85.58**
LR	K1a	67.81	72.21	47.51	**74.53**
	K1b	96.01	90.49	74.14	**97.13**
	La2	82.04	85.51	73.99	**87.50**