

Multi-label Classification via Feature-aware Implicit Label Space Encoding

Zijia Lin^{1,2}

Guiguang Ding²

Mingqing Hu³

Jianmin Wang²

LINZIJIA07@TSINGHUA.ORG.CN

DINGGG@TSINGHUA.EDU.CN

HUMINGQING@ICT.AC.CN

JIMWANG@TSINGHUA.EDU.CN

¹Department of Computer Science and Technology, Tsinghua University, Beijing, P.R. China

²School of Software, Tsinghua University, Beijing, P.R. China

³Institute of Computing Technology, Chinese Academy of Sciences, Beijing, P.R. China

Note: This supplementary material is the appendix of our paper, providing demonstrations of lemmas and supplementary experimental results.

1. Appendix

1.1. Demonstrations for lemmas

To demonstrate Lemma 1 in the paper, firstly we need to derive the following lemma.

Lemma 0. *The eigenvalues of $\Delta = X(X^T X)^{-1} X^T$ will always be either 0 or 1.*

Proof. Assuming the rank of the matrix X is M , i.e. $\text{rank}(X_{N \times F}) = M$ s.t. $1 \leq M \leq \min(N, F)$, there always exist M linearly independent column vectors and row vectors in X . With $\Delta X = X(X^T X)^{-1} X^T X = X$ and $\text{rank}(X) = M$, it is evident that 1 is an eigenvalue of Δ and occurs at least M times, since we can always find M linearly independent $\mathbf{x}_{N \times 1}$ in column vectors of X satisfying $\Delta \mathbf{x} = \mathbf{x}$. Moreover, as $\text{rank}(X) = M$, there exist $(N - M)$ free variables in the variable vector $\mathbf{v}_{N \times 1}$ of the system of linear equations $X^T \mathbf{v} = \mathbf{0}$, and thus $(N - M)$ linearly independent $\mathbf{v}_{N \times 1}$ can always be found to ensure $X^T \mathbf{v} = \mathbf{0}$ and then $\Delta \mathbf{v} = X(X^T X)^{-1} X^T \mathbf{v} = \mathbf{0} = 0 \times \mathbf{v}$. Hence 0 is an eigenvalue of Δ and occurs at least $(N - M)$ times. Since the number of eigenvalues of $\Delta_{N \times N}$ is less than or equal to N and the total number of occurrences of the found eigenvalues (i.e. 0 and 1) is greater than or equal to $(N - M) + M = N$, it is evident that the eigenvalues of Δ will always be either 0 or 1. \square

Lemma 1. *For any given matrix $C_{N \times L}$ satisfying $C^T C = I$, $\text{Tr}[C^T \Delta C]$ has an upper bound being $\min(L, \text{rank}(\Delta))$.*

Proof. To derive the upper bound of $\text{Tr}[C^T \Delta C]$, we intro-

duce the following optimization problem.

$$\max_C \text{Tr}[C^T \Delta C], \quad \text{s.t.} \quad C^T C = I$$

And it can be further decomposed into L optimization sub-problems, one for each column $C_{:,i}$ of C , as shown in the following formula.

$$\begin{aligned} \max_{C_{:,i}} \quad & C_{:,i}^T \Delta C_{:,i} \\ \text{s.t.} \quad & C_{:,i}^T C_{:,i} = 1, C_{:,j}^T C_{:,i} = 0 \ (\forall j < i) \end{aligned}$$

With the method of Lagrange multipliers, it can be derived that the optimal $C_{:,i}$ should satisfy the following optimality condition.

$$\Delta C_{:,i} = \lambda_i C_{:,i}$$

where λ_i is an introduced Lagrange multiplier and will also be the optimal value of the optimization sub-problem. Then optimization for the introduced objective function can be transformed to an eigenvalue problem. Specifically, the optimal C will consist of the normalized eigenvectors of Δ corresponding to the top L largest eigenvalues. And the maximal value of $\text{Tr}[C^T \Delta C]$ is the sum of the top L largest eigenvalues of Δ . According to Lemma 0, it will be less than or equal to L . Moreover, as Δ is a real symmetric matrix, the number of non-zero eigenvalues (i.e. 1) will be equal to its rank, and thus the maximal value of $\text{Tr}[C^T \Delta C]$ will always be less than or equal to $\text{rank}(\Delta)$. Therefore, for any given matrix $C_{N \times L}$ satisfying $C^T C = I$, $\text{Tr}[C^T \Delta C]$ has an upper bound being $\min(L, \text{rank}(\Delta))$. \square

1.2. Supplementary Experimental Results

1.2.1. EXPERIMENTAL RESULTS OF LSDR WITH MORE DETAILS

Detailed experimental results of the proposed FaIE and other algorithms on all datasets are presented in Table 1 - 8 in this document, in terms of mean *label-based macroF1*

and *example-based accuracy* with the corresponding standard errors over 5 runs. From the detailed experimental results, we can draw consistent observations as those presented in the paper.

1.2.2. PARAMETER ANALYSIS ON OTHER DATASETS

Experimental results of parameter analysis for α in the proposed FaIE on *CAL500*, *mediamill* and *ESPGame* are respectively illustrated in Fig. 1, 2 and 3 in this document. For each dataset, we present the variances of multi-label classification performance (the left sub-figure) and the value fluctuations of predictability and recoverability in the objective function (the right sub-figure) as α varies in $\{10^{-2}, 10^{-1}, \dots, 10^4, 10^5\}$ in a run of FaIE with $L/K = 10\%$. From these supplementary illustrations of other datasets, we can draw the following consistent observations with those presented in the paper concerning *delicious*. 1) On each dataset, the performance of FaIE, in terms of both *label-based macroF1* and *example-based accuracy*, firstly increases and then decreases as α varies from 10^{-2} to 10^5 . It further validates the reasonableness of jointly considering recoverability and predictability, as a good trade-off between both can yield a superior performance. 2) On each dataset, as α increases, the value of recoverability in the objective function will decrease, while that of predictability will increase and finally converge to the theoretical upper bound, *i.e.* $\min(L, \text{rank}(\Delta))$, as guaranteed by Lemma 1.

Table 1. Label-based macroF1 on delicious: means (and standard errors) over 5 runs

L/K		10%	20%	30%	40%	50%
BR	L-SVM	0.0790 (0.0026)				
	L-RR	0.0308 (0.0005)				
CS		0.0057 (0.0001)	0.0172 (0.0005)	0.0341 (0.0004)	0.0366 (0.0008)	0.0320 (0.0004)
PLST		0.0196 (0.0004)	0.0214 (0.0005)	0.0217 (0.0005)	0.0217 (0.0005)	0.0217 (0.0005)
CPLST		0.0240 (0.0003)	0.0244 (0.0002)	0.0244 (0.0002)	0.0244 (0.0002)	0.0244 (0.0002)
MLC-BMaD		0.0180 (0.0006)	0.0214 (0.0006)	0.0265 (0.0006)	0.0277 (0.0006)	0.0288 (0.0006)
ML-CSSP		0.0139 (0.0011)	0.0197 (0.0012)	0.0230 (0.0010)	0.0253 (0.0004)	0.0266 (0.0007)
R-FaIE (\approx PLST)		0.0197 (0.0004)	0.0213 (0.0005)	0.0216 (0.0004)	0.0215 (0.0004)	0.0216 (0.0004)
P-LinearFaIE (\sim CPLST)		0.0369 (0.0007)	0.0407 (0.0007)	0.0412 (0.0007)	0.0412 (0.0007)	0.0412 (0.0007)
OP-FaIE		0.0472 (0.0011)	0.0515 (0.0010)	0.0526 (0.0015)	0.0527 (0.0014)	0.0514 (0.0010)
LinearFaIE		0.0411 (0.0006)	0.0431 (0.0008)	0.0434 (0.0007)	0.0435 (0.0007)	0.0435 (0.0007)
FaIE		0.0544 (0.0010)	0.0591 (0.0012)	0.0602 (0.0015)	0.0603 (0.0012)	0.0586 (0.0012)
kernel-CPLST		0.0341 (0.0001)	0.0341 (0.0001)	0.0341 (0.0001)	0.0341 (0.0001)	0.0341 (0.0001)
kernel-FaIE		0.0566 (0.0011)	0.0688 (0.0013)	0.0726 (0.0019)	0.0744 (0.0019)	0.0750 (0.0017)

Table 2. Example-based accuracy on delicious: means (and standard errors) over 5 runs

L/K		10%	20%	30%	40%	50%
BR	L-SVM	0.1419 (0.0019)				
	L-RR	0.0810 (0.0019)				
CS		0.0260 (0.0004)	0.0534 (0.0005)	0.0854 (0.0007)	0.0932 (0.0004)	0.0927 (0.0005)
PLST		0.0734 (0.0019)	0.0762 (0.0021)	0.0768 (0.0022)	0.0769 (0.0022)	0.0768 (0.0022)
CPLST		0.0787 (0.0021)	0.0787 (0.0021)	0.0788 (0.0021)	0.0788 (0.0021)	0.0788 (0.0021)
MLC-BMaD		0.0525 (0.0043)	0.0682 (0.0027)	0.0690 (0.0013)	0.0703 (0.0019)	0.0705 (0.0018)
ML-CSSP		0.0555 (0.0034)	0.0659 (0.0024)	0.0724 (0.0027)	0.0729 (0.0012)	0.0760 (0.0017)
R-FaIE (\approx PLST)		0.0736 (0.0020)	0.0764 (0.0020)	0.0771 (0.0022)	0.0772 (0.0023)	0.0771 (0.0022)
P-LinearFaIE (\sim CPLST)		0.0892 (0.0022)	0.0971 (0.0022)	0.0987 (0.0024)	0.0987 (0.0024)	0.0987 (0.0024)
OP-FaIE		0.1073 (0.0030)	0.1083 (0.0028)	0.1084 (0.0029)	0.1082 (0.0027)	0.1083 (0.0018)
LinearFaIE		0.0984 (0.0022)	0.1058 (0.0026)	0.1061 (0.0027)	0.1061 (0.0027)	0.1061 (0.0027)
FaIE		0.1198 (0.0027)	0.1207 (0.0027)	0.1203 (0.0026)	0.1202 (0.0026)	0.1116 (0.0023)
kernel-CPLST		0.1048 (0.0025)	0.1048 (0.0025)	0.1048 (0.0025)	0.1048 (0.0025)	0.1048 (0.0025)
kernel-FaIE		0.1448 (0.0028)	0.1496 (0.0029)	0.1506 (0.0032)	0.1507 (0.0031)	0.1508 (0.0030)

Table 3. Label-based macroF1 on CAL500: means (and standard errors) over 5 runs

L/K		10%	20%	30%	40%	50%
BR	L-SVM	0.1397 (0.0015)				
	L-RR	0.0569 (0.0025)				
CS		0.0677 (0.0020)	0.0820 (0.0007)	0.0906 (0.0018)	0.0976 (0.0020)	0.1142 (0.0040)
PLST		0.0604 (0.0027)	0.0605 (0.0027)	0.0606 (0.0028)	0.0609 (0.0028)	0.0608 (0.0028)
CPLST		0.0640 (0.0045)	0.0643 (0.0046)	0.0644 (0.0045)	0.0645 (0.0045)	0.0645 (0.0045)
MLC-BMaD		0.0485 (0.0029)	0.0444 (0.0056)	0.0420 (0.0083)	0.0472 (0.0068)	0.0468 (0.0025)
ML-CSSP		0.0453 (0.0027)	0.0498 (0.0017)	0.0507 (0.0023)	0.0528 (0.0027)	0.0543 (0.0025)
R-FaIE (\approx PLST)		0.0596 (0.0022)	0.0600 (0.0022)	0.0600 (0.0021)	0.0601 (0.0022)	0.0600 (0.0021)
P-LinearFaIE (\sim CPLST)		0.0800 (0.0019)	0.0976 (0.0021)	0.1000 (0.0021)	0.0998 (0.0023)	0.0998 (0.0023)
OP-FaIE		0.1034 (0.0031)	0.1080 (0.0028)	0.1088 (0.0032)	-	-
LinearFaIE		0.1062 (0.0026)	0.1107 (0.0040)	0.1105 (0.0042)	0.1105 (0.0043)	0.1105 (0.0043)
FaIE		0.1199 (0.0026)	0.1245 (0.0035)	0.1260 (0.0038)	0.1249 (0.0041)	0.1245 (0.0041)
kernel-CPLST		0.0754 (0.0023)	0.0774 (0.0015)	0.0774 (0.0015)	0.0774 (0.0015)	0.0774 (0.0015)
kernel-FaIE		0.1178 (0.0040)	0.1243 (0.0043)	0.1250 (0.0037)	0.1290 (0.0039)	0.1291 (0.0040)

Table 4. Example-based accuracy on CAL500: means (and standard errors) over 5 runs

L/K		10%	20%	30%	40%	50%
BR	L-SVM	0.2436 (0.0046)				
	L-RR	0.1995 (0.0044)				
CS		0.1130 (0.0013)	0.1299 (0.0024)	0.1626 (0.0036)	0.1904 (0.0018)	0.1835 (0.0077)
PLST		0.2099 (0.0059)	0.2103 (0.0062)	0.2103 (0.0061)	0.2106 (0.0059)	0.2104 (0.0060)
CPLST		0.2003 (0.0061)	0.2007 (0.0063)	0.2009 (0.0062)	0.2010 (0.0062)	0.2010 (0.0062)
MLC-BMaD		0.1286 (0.0256)	0.1215 (0.0296)	0.1194 (0.0312)	0.1244 (0.0314)	0.1255 (0.0261)
ML-CSSP		0.1806 (0.0063)	0.1880 (0.0062)	0.1913 (0.0053)	0.1958 (0.0071)	0.1966 (0.0059)
R-FaIE (\approx PLST)		0.2099 (0.0046)	0.2109 (0.0041)	0.2109 (0.0042)	0.2109 (0.0042)	0.2106 (0.0040)
P-LinearFaIE (\sim CPLST)		0.2093 (0.0040)	0.2210 (0.0020)	0.2205 (0.0015)	0.2206 (0.0012)	0.2206 (0.0012)
OP-FaIE		0.2283 (0.0017)	0.2262 (0.0016)	0.2251 (0.0015)	-	-
LinearFaIE		0.2329 (0.0030)	0.2300 (0.0019)	0.2301 (0.0024)	0.2299 (0.0025)	0.2300 (0.0025)
FaIE		0.2413 (0.0014)	0.2414 (0.0019)	0.2417 (0.0020)	0.2384 (0.0023)	0.2379 (0.0024)
kernel-CPLST		0.2139 (0.0039)	0.2148 (0.0030)	0.2148 (0.0030)	0.2148 (0.0030)	0.2148 (0.0030)
kernel-FaIE		0.2429 (0.0040)	0.2443 (0.0051)	0.2422 (0.0040)	0.2430 (0.0048)	0.2419 (0.0043)

Table 5. Label-based macroF1 on mediamill: means (and standard errors) over 5 runs

L/K		10%	20%	30%	40%	50%
BR	L-SVM	0.0554 (0.0016)				
	L-RR	0.0454 (0.0004)				
CS		0.0056 (0.0004)	0.0145 (0.0004)	0.0134 (0.0003)	0.0311 (0.0006)	0.0274 (0.0005)
PLST		0.0419 (0.0005)	0.0435 (0.0003)	0.0436 (0.0004)	0.0436 (0.0005)	0.0436 (0.0004)
CPLST		0.0433 (0.0005)	0.0440 (0.0004)	0.0440 (0.0003)	0.0440 (0.0003)	0.0440 (0.0003)
MLC-BMaD		0.0412 (0.0004)	0.0426 (0.0004)	0.0425 (0.0003)	0.0423 (0.0004)	0.0425 (0.0001)
ML-CSSP		0.0343 (0.0007)	0.0406 (0.0024)	0.0428 (0.0013)	0.0416 (0.0025)	0.0437 (0.0005)
R-FaIE (\approx PLST)		0.0419 (0.0004)	0.0438 (0.0004)	0.0441 (0.0005)	0.0440 (0.0005)	0.0440 (0.0005)
P-LinearFaIE (\sim CPLST)		0.0425 (0.0005)	0.0444 (0.0004)	0.0453 (0.0005)	0.0452 (0.0005)	0.0452 (0.0005)
OP-FaIE		0.0450 (0.0005)	0.0469 (0.0004)	0.0472 (0.0003)	0.0474 (0.0003)	0.0471 (0.0003)
LinearFaIE		0.0444 (0.0002)	0.0463 (0.0004)	0.0464 (0.0005)	0.0461 (0.0005)	0.0461 (0.0005)
FaIE		0.0570 (0.0003)	0.0595 (0.0005)	0.0609 (0.0007)	0.0602 (0.0006)	0.0607 (0.0007)
kernel-CPLST		0.0492 (0.0007)	0.0521 (0.0011)	0.0521 (0.0011)	0.0521 (0.0011)	0.0521 (0.0011)
kernel-FaIE		0.0682 (0.0008)	0.0805 (0.0007)	0.0852 (0.0006)	0.0857 (0.0021)	0.0855 (0.0012)

Table 6. Example-based accuracy on mediamill: means (and standard errors) over 5 runs

L/K		10%	20%	30%	40%	50%
BR	L-SVM	0.3130 (0.0251)				
	L-RR	0.4173 (0.0030)				
CS		0.0103 (0.0009)	0.0296 (0.0005)	0.0343 (0.0011)	0.1403 (0.0022)	0.1357 (0.0045)
PLST		0.4117 (0.0032)	0.4144 (0.0026)	0.4142 (0.0026)	0.4141 (0.0026)	0.4143 (0.0025)
CPLST		0.4142 (0.0030)	0.4148 (0.0032)	0.4149 (0.0031)	0.4149 (0.0031)	0.4148 (0.0031)
MLC-BMaD		0.4027 (0.0039)	0.4037 (0.0043)	0.4027 (0.0034)	0.4027 (0.0043)	0.4027 (0.0037)
ML-CSSP		0.3378 (0.0361)	0.3954 (0.0134)	0.4058 (0.0071)	0.4045 (0.0122)	0.4148 (0.0039)
R-FaIE (\approx PLST)		0.4121 (0.0028)	0.4150 (0.0020)	0.4150 (0.0022)	0.4150 (0.0022)	0.4151 (0.0022)
P-LinearFaIE (\sim CPLST)		0.4135 (0.0033)	0.4145 (0.0025)	0.4151 (0.0022)	0.4155 (0.0021)	0.4156 (0.0022)
OP-FaIE		0.4163 (0.0026)	0.4185 (0.0025)	0.4182 (0.0021)	0.4177 (0.0022)	0.4179 (0.0021)
LinearFaIE		0.4154 (0.0031)	0.4172 (0.0024)	0.4160 (0.0026)	0.4153 (0.0027)	0.4154 (0.0027)
FaIE		0.4327 (0.0024)	0.4339 (0.0027)	0.4338 (0.0031)	0.4334 (0.0029)	0.4336 (0.0026)
kernel-CPLST		0.4228 (0.0029)	0.4234 (0.0026)	0.4240 (0.0025)	0.4240 (0.0025)	0.4240 (0.0025)
kernel-FaIE		0.4401 (0.0017)	0.4440 (0.0032)	0.4447 (0.0028)	0.4423 (0.0045)	0.4385 (0.0047)

Table 7. Label-based macroF1 on ESPGame: means (and standard errors) over 5 runs

L/K		5%	10%	15%	20%	25%
BR	L-SVM	0.0213 (0.0017)				
	L-RR	0.0014 (0.0002)				
CS		0.0004 (0.0000)	0.0006 (0.0000)	0.0013 (0.0001)	0.0013 (0.0001)	0.0018 (0.0001)
PLST		0.0008 (0.0000)	0.0008 (0.0000)	0.0008 (0.0000)	0.0008 (0.0000)	0.0008 (0.0000)
CPLST		0.0008 (0.0000)	0.0008 (0.0000)	0.0008 (0.0000)	0.0008 (0.0000)	0.0008 (0.0000)
MLC-BMaD		0.0009 (0.0000)	0.0009 (0.0000)	0.0010 (0.0001)	0.0009 (0.0000)	0.0010 (0.0001)
ML-CSSP		0.0008 (0.0001)	0.0005 (0.0000)	0.0007 (0.0001)	0.0007 (0.0001)	0.0008 (0.0001)
R-FaIE (\approx PLST)		0.0008 (0.0000)	0.0008 (0.0000)	0.0008 (0.0000)	0.0008 (0.0000)	0.0008 (0.0000)
P-LinearFaIE (\sim CPLST)		0.0015 (0.0002)	0.0018 (0.0003)	0.0020 (0.0003)	0.0021 (0.0003)	0.0021 (0.0003)
OP-FaIE		0.0019 (0.0003)	0.0024 (0.0005)	0.0024 (0.0005)	0.0024 (0.0004)	0.0026 (0.0004)
LinearFaIE		0.0021 (0.0003)	0.0022 (0.0003)	0.0023 (0.0002)	0.0024 (0.0002)	0.0024 (0.0003)
FaIE		0.0023 (0.0001)	0.0028 (0.0003)	0.0028 (0.0003)	0.0029 (0.0004)	0.0028 (0.0004)
kernel-CPLST		0.0009 (0.0000)	0.0009 (0.0000)	0.0009 (0.0000)	0.0009 (0.0000)	0.0009 (0.0000)
kernel-FaIE		0.0038 (0.0001)	0.0045 (0.0001)	0.0054 (0.0004)	0.0058 (0.0004)	0.0062 (0.0005)

Table 8. Example-based accuracy on ESPGame: means (and standard errors) over 5 runs

L/K		5%	10%	15%	20%	25%
BR	L-SVM	0.0726 (0.0018)				
	L-RR	0.0452 (0.0019)				
CS		0.0166 (0.0005)	0.0213 (0.0007)	0.0445 (0.0012)	0.0467 (0.0013)	0.0505 (0.0080)
PLST		0.0457 (0.0022)	0.0457 (0.0022)	0.0457 (0.0022)	0.0457 (0.0022)	0.0457 (0.0022)
CPLST		0.0469 (0.0017)	0.0470 (0.0017)	0.0470 (0.0017)	0.0470 (0.0017)	0.0470 (0.0017)
MLC-BMaD		0.0450 (0.0019)	0.0449 (0.0019)	0.0458 (0.0012)	0.0446 (0.0017)	0.0446 (0.0018)
ML-CSSP		0.0432 (0.0024)	0.0280 (0.0040)	0.0357 (0.0055)	0.0321 (0.0051)	0.0397 (0.0042)
R-FaIE (\approx PLST)		0.0455 (0.0019)	0.0455 (0.0019)	0.0453 (0.0020)	0.0455 (0.0019)	0.0453 (0.0020)
P-LinearFaIE (\sim CPLST)		0.0394 (0.0008)	0.0491 (0.0017)	0.0554 (0.0014)	0.0581 (0.0009)	0.0581 (0.0013)
OP-FaIE		0.0593 (0.0011)	0.0596 (0.0009)	0.0593 (0.0012)	0.0591 (0.0012)	0.0595 (0.0009)
LinearFaIE		0.0556 (0.0013)	0.0638 (0.0036)	0.0639 (0.0041)	0.0659 (0.0026)	0.0670 (0.0015)
FaIE		0.0641 (0.0019)	0.0640 (0.0019)	0.0666 (0.0035)	0.0669 (0.0028)	0.0690 (0.0027)
kernel-CPLST		0.0488 (0.0019)	0.0483 (0.0020)	0.0488 (0.0023)	0.0483 (0.0020)	0.0488 (0.0023)
kernel-FaIE		0.0831 (0.0025)	0.0834 (0.0026)	0.0837 (0.0024)	0.0840 (0.0026)	0.0841 (0.0025)

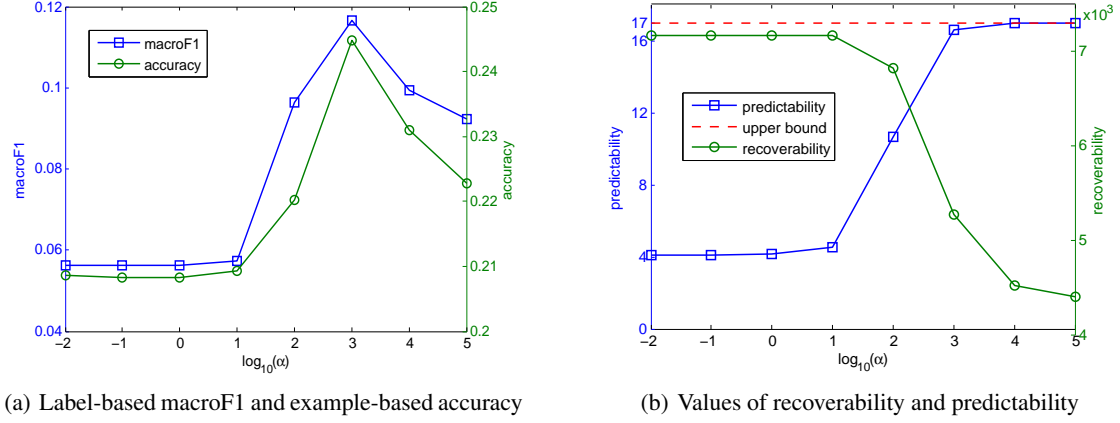


Figure 1. Effects of α in FaLE on the performance of multi-label classification (sub-figure 1(a)) and the values of recoverability and predictability (sub-figure 1(b)) on *CAL500*, with $L/K = 10\%$ and the theoretical upper bound $\min(L, \text{rank}(\Delta)) = 17$.

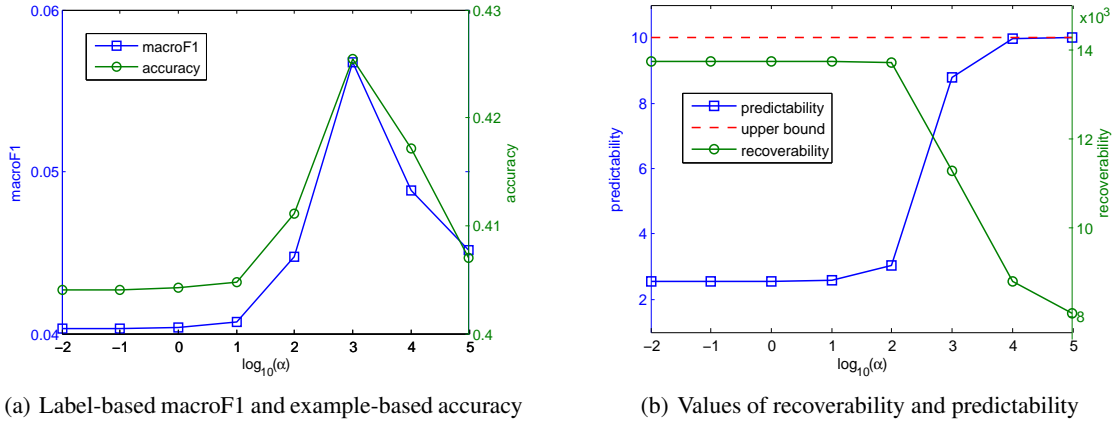


Figure 2. Effects of α in FaLE on the performance of multi-label classification (sub-figure 2(a)) and the values of recoverability and predictability (sub-figure 2(b)) on *mediamill*, with $L/K = 10\%$ and the theoretical upper bound $\min(L, \text{rank}(\Delta)) = 10$.

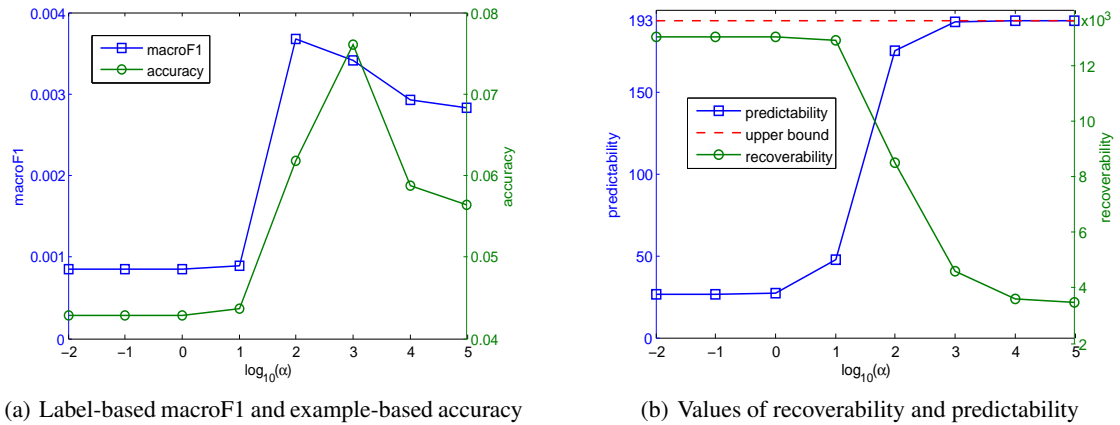


Figure 3. Effects of α in FaLE on the performance of multi-label classification (sub-figure 3(a)) and the values of recoverability and predictability (sub-figure 3(b)) on *ESPGame*, with $L/K = 10\%$ and the theoretical upper bound $\min(L, \text{rank}(\Delta)) = 193$.