Supplemental Material to 'Learning classifier chains using matrix regularization: application to multimorbidity prediction'

Paweł Teisseyre

In the following we give formal definitions of the evaluation measures used in the experiments. In the paper we consider example-based measures which are calculated for each test example and then averaged across the test set. Below we recall definitions of the considered measures (definitions are given for one instance in the test set). Let $\mathbf{y} = (y_1, \ldots, y_K)$ be a vector of true labels and $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_K)$ be the vector of predicted labels for some instance $\mathbf{x}$ in the test set. Let $|A|$ be the cardinality of set A, $I(A)$ be an indicator function and $\bar{A}$ be a complement of A.

Subset accuracy is defined as

$$\text{Subset accuracy}(\mathbf{y}, \hat{\mathbf{y}}) = I(\mathbf{y} = \hat{\mathbf{y}}),$$

which is directly related to subset loss. Subset accuracy measures the correctness of joint prediction for all labels. It can be regarded as a multi-label counterpart of the traditional accuracy metric.

Hamming measure, defined as

$$\text{Hamming measure}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{K} \sum_{k=1}^{K} I(y_k = \hat{y}_k)$$

is an average number of correct predictions.

F measure, defined as

$$\text{F measure}(\mathbf{y}, \hat{\mathbf{y}}) = 2 \cdot \frac{\text{Prec}(\mathbf{y}, \hat{\mathbf{y}}) \cdot \text{Recall}(\mathbf{y}, \hat{\mathbf{y}})}{\text{Prec}(\mathbf{y}, \hat{\mathbf{y}}) + \text{Recall}(\mathbf{y}, \hat{\mathbf{y}})},$$

is a harmonic mean of precision and recall which are defined as

$$\text{Prec}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{|\{i : y_i = 1, \hat{y}_i = 1\}|}{|\{i : \hat{y}_i = 1\}|}, \quad \text{Recall}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{|\{i : y_i = 1, \hat{y}_i = 1\}|}{|\{i : y_i = 1\}|}.$$

Precision measures how many labels are correctly predicted as relevant with respect to all labels predicted as relevant. Recall is a fraction of labels correctly predicted as relevant with respect to all relevant ones.

Ranking measure is defined as follows. Let $f(\mathbf{x}, y_k)$ be a real-valued function that returns the confidence of y_k being a proper label of $\mathbf{x}$ (i.e. confidence of $y_k = 1$). We first define ranking loss as

$$\text{Ranking loss}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{|T| \cdot |\bar{T}|} |\{(y_i, y_j) \in T \times \bar{T} : f(\mathbf{x}, y_i) < f(\mathbf{x}, y_j)\}|,$$

where $T = \{i : y_i = 1\}$, $\bar{T} = \{i : y_i = 0\}$ are sets of relevant and irrelevant labels, respectively. It measures how many irrelevant labels are ranked higher than relevant ones. Ranking measure is

$$\text{Ranking measure}(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \text{Ranking loss}(\mathbf{y}, \hat{\mathbf{y}}).$$

Table 1: MIMIC-III. List of features (part 1).

	Description
1	Gender
2	Age
3	Arterial BP [Systolic]
4	Arterial BP Mean
5	BSA
6	BSA - English
7	BSA - Metric
8	Braden Score
9	CaO2
10	Compliance (40-60ml)
11	Eye Opening
12	FiO2 Set
13	GCS Total
14	Heart Rate
15	High Insp. Pressure
16	High Resp. Rate
17	LCW
18	LCWI
19	LVSW
20	LVSWI
21	Low Exhaled Min Vol
22	Mean Airway Pressure
23	Minute Volume(Obser)
24	Motor Response
25	NBP [Systolic]
26	NBP Mean
27	O2 Flow (lpm)
28	PEEP Set
29	Peak Insp. Pressure
30	Plateau Pressure
31	Pressure Support
32	Previous Weight
33	Previous WeightF
34	Resp Rate (Spont)
35	Resp Rate (Total)
36	Respiratory Rate
37	Respiratory Rate Set
38	SVI
39	SVRI
40	Sensitivity-Vent
41	SpO2
42	Stroke Volume
43	Temperature C (calc)
44	Temperature F
45	Tidal Volume (Obser)
46	Tidal Volume (Set)
47	Tidal Volume (Spont)
48	Verbal Response
49	Weight Change
50	Admit Wt

	Description
51	Daily Weight
52	ALT
53	AST
54	Albumin (>3.2)
55	Alk. Phosphate
56	Amylase
57	Arterial Base Excess
58	Arterial CO2(Calc)
59	Arterial PaCO2
60	Arterial PaO2
61	Arterial pH
62	BUN (6-20)
63	CPK
64	CPK/MB
65	Calcium (8.4-10.2)
66	Carbon Dioxide
67	Chloride (100-112)
68	Creatinine (0-1.3)
69	Differential-Bands
70	Differential-Basos
71	Differential-Eos
72	Differential-Lymphs
73	Differential-Monos
74	Differential-Polys
75	Fibrinogen (150-400)
76	Fingerstick Glucose
77	Glucose (70-105)
78	Hematocrit
79	Hemoglobin
80	INR (2-4 ref. range)
81	Ionized Calcium
82	Lactic Acid(0.5-2.0)
83	Magnesium (1.6-2.6)
84	Mixed Venous O2% Sat
85	PT(11-13.5)
86	PTT(22-35)
87	Phosphorous(2.7-4.5)
88	Platelets
89	Potassium (3.5-5.3)
90	RBC
91	SaO2
92	Sodium (135-148)
93	Total Bili (0-1.5)
94	Vancomycin/Random
95	Venous pH
96	WBC (4-11,000)
97	Admit Ht
98	Art.pH
99	WBC (4-11,000)
100	Arterial BP [Diastolic]

	Description
101	NBP [Diastolic]
102	PAP [Diastolic]
103	LDH
104	BUN
105	PT
106	Albumin
107	Calcium
108	Chloride
109	Creatinine
110	Glucose
111	INR
112	Magnesium
113	PTT
114	Phosphorous
115	Potassium
116	Sodium
117	Total Bili
118	WBC
119	Fibrinogen
120	ABP Alarm [Low]
121	HR Alarm [Low]
122	NBP Alarm [Low]
123	Resp Alarm [Low]
124	SpO2 Alarm [Low]
125	ABP Alarm [High]
126	CVP Alarm [High]
127	HR Alarm [High]
128	NBP Alarm [High]
129	Resp Alarm [High]
130	SpO2 Alarm [High]
131	Apnea Time Interval
132	RSBI (<200)
133	Spon RR (Mech.)
134	Spon. Vt (L) (Mech.)
135	RSBI (<100)
136	Lactic Acid
137	Tank A psi.
138	Tank B psi.
139	Ve High
140	Heart rate Alarm - High
141	Heart Rate Alarm - Low
142	Arterial Blood Pressure systolic
143	Arterial Blood Pressure diastolic
144	Arterial Blood Pressure mean
145	Arterial Blood Pressure Alarm - Low
146	Arterial Blood Pressure Alarm - High
147	Non Invasive Blood Pressure systolic
148	Non Invasive Blood Pressure diastolic
149	Non Invasive Blood Pressure mean
150	O2 saturation pulseoxymetry

Table 2: MIMIC-III. List of features (part 1).

	Description		Description		Description
151	Hematocrit (serum)	201	Gait/Transferring	251	Inspiratory Ratio
152	Chloride (serum)	202	Mental status	252	Troponin-T
153	Glucose (serum)	203	16 Gauge Dressing Occlusive	253	CK-MB
154	Sodium (serum)	204	Potassium (serum)	254	Richmond-RAS Scale
155	GCS - Eye Opening	205	HCO3 (serum)	255	18 Gauge placed in the field
156	Riker-SAS Scale	206	Platelet Count	256	CHG Bath
157	Temperature Fahrenheit	207	Prothrombin time	257	Insulin pump
158	O2 Saturation Alarm - High	208	Smoking Cessation Info	258	Arterial O2 pressure
159	O2 Saturation Alarm - Low	209	Central Venous Pressure	259	Arterial O2 Saturation
160	Pain Level	210	Minute Volume Alarm - Low	260	Arterial CO2 Pressure
161	O2 Flow	211	Minute Volume Alarm - High	261	PH (Arterial)
162	GCS - Verbal Response	212	PEEP set	262	Cuff Pressure
163	GCS - Motor Response	213	Non-Invasive BP Alarm - High	263	Spont Vt
164	Braden Sensory Perception	214	Non-Invasive BP Alarm - Low	264	Spont RR
165	Braden Moisture	215	Inspired O2 Fraction	265	Any fear in relationships
166	Braden Activity	216	Ventilator Type	266	ETOH
167	Braden Mobility	217	Ventilator Mode	267	Recreational drug use
168	Braden Nutrition	218	Paw High	268	Currently experiencing pain
169	Braden Friction/Shear	219	Vti High	269	Dialysis patient
170	Resp Alarm - High	220	Fspn High	270	Incentive Spirometry
171	Resp Alarm - Low	221	Apnea Interval	271	Alkaline Phosphate
172	Parameters Checked	222	Tidal Volume (set)	272	Total Bilirubin
173	Pain Level Response	223	Tidal Volume (observed)	273	TCO2 (calc) Arterial
174	Alarms On	224	Tidal Volume (spontaneous)	274	Baseline pain level
175	See chart for initial assessment	225	Minute Volume	275	20 Gauge placed in outside facility
176	Eye Care	226	Respiratory Rate (Set)	276	Subglottal Suctioning
177	Skin Care	227	Respiratory Rate (spontaneous)	277	High risk (>51) interventions
178	Back Care	228	Respiratory Rate (Total)	278	18 Gauge Dressing Occlusive
179	Cough/Deep Breath	229	Total PEEP Level	279	20 Gauge Dressing Occlusive
180	Arterial Line Zero/Calibrate	230	PSV Level	280	Ventilator Tank #1
181	Called Out	231	Inspiratory Time	281	Ventilator Tank #2
182	Bed Bath	232	Is the spokesperson the HCP	282	20 Gauge placed in the field
183	Calcium non-ionized	233	Unable to assess psychological	283	Discharge needs
184	Glucose finger stick	234	Harm by partner or close relation	284	Goal Richmond-RAS Scale
185	Arterial Line outside facility	235	Social work consult	285	ST Segment Monitoring On
186	Multi Lumen outside facility	236	Visual / hearing deficit	286	CAM-ICU MS Change
187	16 Gauge placed in outside facility	237	Self ADL	287	22 Gauge placed in outside facility
188	SpO2 Desat Limit	238	History of slips / falls	288	22 Gauge Dressing Occlusive
189	Marital Status	239	Intravenous access	289	Potassium (whole blood)
190	Admission Weight (Kg)	240	Difficulty swallowing	290	PH (Venous)
191	Admission Weight (lbs.)	241	Special diet	291	PH (dipstick)
192	Religion	242	Unintentional weight loss >10 lbs.	292	Impaired Skin Odor #1
193	Language	243	CK (CPK)	293	Unable to assess cognitive / perceptual
194	Anion gap	244	Differential-Neuts	294	Unable to assess activity / mobility
195	Arterial Line Dressing Occlusive	245	18 Gauge placed in outside facility	295	Unable to assess nutrition / education
196	Multi Lumen Dressing Occlusive	246	No wallet / money	296	Unable to assess teaching / learning needs
197	History of falling (within 3 mnths)	247	Sexuality / reproductive problems	297	Lipase
198	Secondary diagnosis	248	Height	298	ICU Consent Signed
199	Ambulatory aid	249	Height (cm)	299	Sodium (whole blood)
200	IV/Saline lock	250	Expiratory Ratio	300	Specific Gravity (urine)

Table 3: MIMIC-III. List of features (part 3).

	Description
301	CVP Alarm - High
302	CVP Alarm - Low
303	Unable to assess pain

	Description
304	Multi Lumen Zero/Calibrate
305	Chloride (whole blood)
306	Glucose (whole blood)

	Description
307	Hematocrit (whole blood - calc)
308	Unable to assess habits

Table 4: Subset accuracy for benchmark datasets, for $T = 20$.

	BR+l21	parCC	CC+l1	BR+l1	ACC-ABS	ACC-SUM
cal500	0.016	**0.028**	0.020	0.018	0.008	0.012
nuswide	0.312	**0.316**	0.305	0.305	0.305	0.305
medical	0.709	**0.710**	0.653	0.653	0.613	0.522
music	0.236	**0.267**	0.054	0.054	0.152	0.181
yeast	0.091	**0.144**	0.032	0.031	0.115	0.110
scene	0.023	0.023	0.000	0.000	0.041	**0.092**
mediamill	0.115	**0.150**	0.105	0.093	0.111	0.106
flags	0.144	0.211	0.191	0.144	**0.257**	0.252
Science	0.168	0.172	0.160	0.159	0.181	**0.192**
Reference	0.230	0.227	0.151	0.145	0.245	**0.249**
Health	0.308	0.314	0.309	0.310	0.319	**0.323**
avg. rank	3.6	**5.0**	2.4	1.9	4.1	4.1

Table 5: F measure for benchmark datasets, for $T = 20$.

	BR+l21	parCC	CC+l1	BR+l1	ACC-ABS	ACC-SUM
cal500	0.723	0.711	**0.729**	0.726	0.708	0.708
nuswide	**0.058**	0.033	0.000	0.000	0.001	0.000
medical	0.612	**0.612**	0.574	0.574	0.518	0.446
music	0.492	**0.565**	0.172	0.172	0.382	0.444
yeast	**0.564**	0.558	0.485	0.475	0.546	0.545
scene	0.024	0.024	0.000	0.000	0.043	**0.103**
mediamill	**0.588**	0.584	0.536	0.520	0.542	0.533
flags	0.678	0.688	0.687	0.683	**0.699**	0.698
Science	0.511	0.514	0.508	0.507	0.516	**0.521**
Reference	0.652	0.650	0.621	0.619	0.658	**0.658**
Health	0.518	0.526	0.522	0.521	**0.534**	0.533
avg. rank	4.0	**4.4**	2.6	2.0	4.1	3.9

Table 6: Hamming measure for benchmark datasets, for $T = 20$.

	BR+l21	parCC	CC+l1	BR+l1	ACC-ABS	ACC-SUM
cal500	**0.625**	0.616	0.619	0.623	0.606	0.607
nuswide	**0.886**	0.883	0.877	0.877	0.877	0.877
medical	0.962	**0.963**	0.957	0.957	0.952	0.943
music	**0.788**	0.778	0.716	0.716	0.742	0.740
yeast	**0.734**	0.718	0.710	0.707	0.718	0.715
scene	0.825	0.825	0.821	0.821	0.827	**0.837**
mediamill	**0.809**	0.809	0.793	0.785	0.791	0.789
flags	0.733	0.737	0.731	0.735	**0.739**	0.738
Science	0.872	0.872	0.871	0.870	0.873	**0.874**
Reference	0.900	0.899	0.891	0.891	0.901	**0.901**
Health	**0.825**	0.815	0.811	0.814	0.814	0.815
avg. rank	**4.9**	4.3	2.2	2.0	3.8	3.7

Table 7: Subset accuracy for benchmark datasets, for $T = 10$.

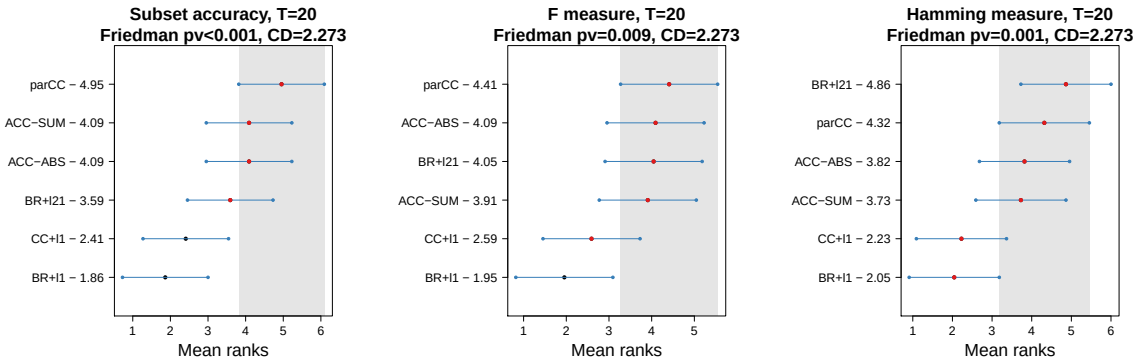
	BR+l21	parCC	CC+l1	BR+l1	ACC-ABS	ACC-SUM
cal500	0.008	**0.024**	0.002	0.006	0.022	0.020
nuswide	0.305	**0.305**	0.305	0.305	0.305	0.305
medical	0.446	0.446	0.213	0.213	**0.529**	0.505
music	0.189	**0.235**	0.012	0.012	0.010	0.024
yeast	0.035	**0.105**	0.031	0.031	0.043	0.067
scene	**0.002**	0.002	0.000	0.000	0.000	0.000
mediamill	0.097	**0.109**	0.094	0.094	0.093	0.104
flags	0.155	0.196	0.150	0.150	0.227	**0.237**
Science	0.158	0.157	0.152	0.152	0.161	**0.164**
Reference	0.161	0.163	0.118	0.120	0.201	**0.218**
Health	0.308	0.310	0.308	0.308	0.310	**0.316**
avg. rank	3.6	**5.0**	1.9	2.0	3.7	4.7

Table 8: F measure for benchmark datasets, for $T = 10$.

	BR+l21	parCC	CC+l1	BR+l1	ACC-ABS	ACC-SUM
cal500	0.726	0.717	0.724	0.719	0.730	**0.732**
nuswide	**0.017**	0.002	0.000	0.000	0.000	0.000
medical	0.293	0.293	0.000	0.000	**0.430**	0.417
music	0.421	**0.498**	0.065	0.065	0.067	0.107
yeast	0.520	**0.531**	0.475	0.475	0.500	0.516
scene	**0.002**	0.002	0.000	0.000	0.000	0.000
mediamill	**0.550**	0.538	0.521	0.521	0.520	0.511
flags	**0.700**	0.698	0.690	0.691	0.696	0.698
Science	0.506	0.507	0.504	0.504	0.508	**0.509**
Reference	0.625	0.626	0.609	0.610	0.641	**0.647**
Health	0.519	0.523	0.522	0.522	0.524	**0.528**
avg. rank	4.4	**4.5**	2.0	2.0	3.8	4.3

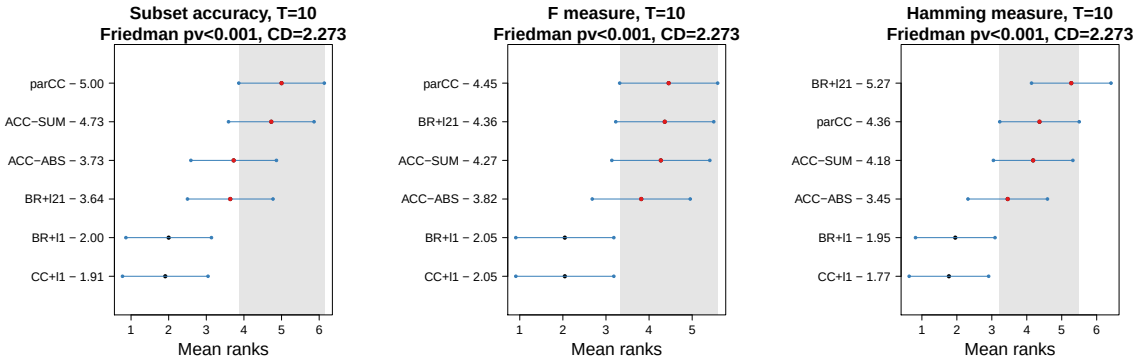
Table 9: Hamming measure for benchmark datasets, for $T = 10$.

	BR+l21	parCC	CC+l1	BR+l1	ACC-ABS	ACC-SUM
cal500	**0.623**	0.616	0.616	0.618	0.614	0.619
nuswide	**0.880**	0.878	0.877	0.877	0.877	0.877
medical	0.933	0.933	0.906	0.906	**0.943**	0.941
music	**0.773**	0.769	0.697	0.697	0.701	0.702
yeast	**0.722**	0.719	0.707	0.707	0.714	0.713
scene	**0.821**	0.821	0.821	0.821	0.821	0.821
mediamill	**0.795**	0.794	0.779	0.779	0.781	0.788
flags	**0.745**	0.744	0.730	0.731	0.736	0.739
Science	0.870	0.870	0.870	0.870	0.871	**0.871**
Reference	0.892	0.893	0.888	0.888	0.897	**0.898**
Health	**0.818**	0.812	0.809	0.809	0.810	0.812
avg. rank	**5.3**	4.4	1.8	2.0	3.5	4.2



(a) (b) (c)

Figure 1: Results of Friedman and pairwise tests, for $T = 20$.



(a) (b) (c)

Figure 2: Results of Friedman and pairwise tests, for $T = 10$.