A new method for mining disjunctive emerging patterns in high-dimensional datasets using hypergraphs

Supplementary Material – Characteristics of datasets used to assess the performance of our method after discretization

Renato Vimieiro and Pablo Moscato
Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine
The University of Newcastle
Hunter Medical Research Institute
Lot 1, Kookaburra Circuit, New Lambton Heights, NSW, 2305, Australia.
Email: {renato.vimieiro, pablo.moscato}@newcastle.edu.au

Characteristics of the datasets of the experiments to assess the performance of our method for mining disjunctive emerging patterns with hypergraphs

Table 1: Characteristics of dataset Colon for evaluating the performance of our method considering diverse discretization parameters. *Equal frequency* and *equal width* refers to the discretization method; *Bins* is the number of bins used to discretize the data; $\%\ ATT$ and $|A|$ are respectively the percentage and number of attributes randomly selected from the original set; $|F|$ is the number of features in the dataset, i.e., the size of the union of all attribute domains; $\mu|E|$ is the average edge length in hypergraphs generated for the dataset, $\sigma|E|$ is its standard deviation; $max(|E|)$ and $min(|E|)$ are the maximum and minimum edge lengths.

Colon — $(\|S^-\| = 22, \|S^+\| = 40)$								
	Bins	% ATT	$\|A\|$	$\|F\|$	$\mu\|E\|$	$\sigma\|E\|$	$max(\|E\|)$	$min(\|E\|)$
equal frequency	2	1	19	38	10.61	3.68	20	1
	2	2	39	78	21.02	6.95	40	1
	2	5	99	198	51.93	16.05	99	10
	2	10	199	398	103.83	32.72	198	20
	2	15	299	598	155.51	48.55	298	27
	2	25	499	998	258.05	81.72	496	41
	4	1	19	76	15.45	2.87	20	4
	4	2	39	156	30.91	4.88	40	10
	4	5	99	396	76.86	10.99	100	26
	4	10	199	796	153.69	22.01	199	54
	4	15	299	1196	230.54	32.51	299	78
	4	25	499	1996	384.09	53.95	498	122
equal width	2	1	20	40	6.01	3.73	18	1
	2	2	40	80	11.00	7.85	31	1
	2	5	100	200	24.61	18.88	75	1
	2	10	200	400	47.85	35.87	136	1
	2	15	300	600	71.30	53.45	199	1
	2	25	500	1000	121.41	88.81	341	2
	4	1	20	80	12.15	3.89	21	2
	4	2	40	160	23.52	7.51	41	4
	4	5	100	395	56.84	18.12	96	10
	4	10	200	794	112.70	35.19	189	18
	4	15	300	1191	168.19	52.71	283	25
	4	25	500	1984	281.17	86.01	471	41

Table 2: Characteristics of dataset ALL–AML for evaluating the performance of our method considering diverse parameters. *Equal frequency* and *equal width* refers to the discretization method; *Bins* is the number of bins used to discretize the data; $\%\ ATT$ and $|A|$ are respectively the percentage and number of attributes randomly selected from the original set; $|F|$ is the number of features in the dataset, i.e., the size of the union of all attribute domains; $\mu|E|$ is the average edge length in hypergraphs generated for the dataset, $\sigma|E|$ is its standard deviation; $max(|E|)$ and $min(|E|)$ are the maximum and minimum edge lengths.

ALL–AML — ($\|S^-\| = 11, \|S^+\| = 27$)								
	Bins	% ATT	$\|A\|$	$\|F\|$	$\mu\|E\|$	$\sigma\|E\|$	$max(\|E\|)$	$min(\|E\|)$
equal frequency	2	1	36	72	20.43	3.46	29	7
	2	2	71	142	39.13	5.46	53	18
	2	5	181	362	98.60	11.77	130	44
	2	10	341	682	185.67	18.38	236	113
	2	15	506	1012	273.25	24.36	337	171
	2	25	845	1690	456.24	38.50	571	297
	4	1	39	146	29.81	3.26	38	18
	4	2	79	291	58.54	4.95	70	40
	4	5	193	714	143.55	9.28	171	95
	4	10	381	1379	277.16	15.70	316	217
	4	15	577	2067	414.06	21.11	460	326
	4	25	963	3457	693.42	32.90	773	547
equal width	2	1	50	98	9.51	4.02	23	2
	2	2	100	195	20.32	8.12	49	4
	2	5	250	489	54.82	17.80	114	16
	2	10	500	979	105.88	31.87	219	40
	2	15	750	1473	163.58	46.99	316	62
	2	25	1250	2457	274.71	76.00	525	102
	4	1	50	172	18.70	4.52	32	8
	4	2	100	338	38.96	7.64	63	19
	4	5	250	876	105.09	16.07	154	61
	4	10	500	1762	207.77	30.23	292	117
	4	15	750	2657	316.43	44.25	441	195
	4	25	1250	4419	531.53	71.79	741	337

Table 3: Characteristics of dataset Leukemia for evaluating the performance of our method considering diverse parameters. *Equal frequency* and *equal width* refers to the discretization method; *Bins* is the number of bins used to discretize the data; $\%\ ATT$ and $|A|$ are respectively the percentage and number of attributes randomly selected from the original set; $|F|$ is the number of features in the dataset, i.e., the size of the union of all attribute domains; $\mu|E|$ is the average edge length in hypergraphs generated for the dataset, $\sigma|E|$ is its standard deviation; $max(|E|)$ and $min(|E|)$ are the maximum and minimum edge lengths.

Leukemia — ($\|S^-\| = 11, \|S^+\| = 27$)								
	Bins	% ATT	$\|A\|$	$\|F\|$	$\mu\|E\|$	$\sigma\|E\|$	$max(\|E\|)$	$min(\|E\|)$
equal frequency	2	1	71	142	39.05	5.09	53	24
	2	2	142	284	76.55	8.95	101	45
	2	5	356	712	190.47	19.37	237	135
	2	10	713	1426	381.11	39.85	485	254
	2	15	1069	2138	569.94	61.14	739	362
	2	25	1782	3564	949.47	98.64	1229	599
	4	1	71	284	56.63	4.07	68	43
	4	2	142	568	112.18	6.58	132	86
	4	5	356	1424	279.73	13.43	312	223
	4	10	713	2852	558.37	27.48	622	448
	4	15	1069	4276	836.43	41.85	922	655
	4	25	1782	7128	1394.40	69.09	1537	1104
equal width	2	1	71	142	25.90	5.08	43	13
	2	2	142	284	49.81	9.41	78	22
	2	5	356	712	127.39	21.83	195	56
	2	10	712	1424	266.99	43.12	386	117
	2	15	1069	2138	398.26	64.21	575	181
	2	25	1782	3564	663.83	110.02	967	301
	4	1	71	277	44.81	5.11	60	30
	4	2	142	555	88.16	8.87	111	63
	4	5	356	1393	221.75	19.62	265	158
	4	10	712	2804	454.20	37.49	548	316
	4	15	1069	4203	678.45	56.33	824	461
	4	25	1782	6999	1128.52	93.56	1352	776

Table 4: Characteristics of dataset Lymphoma for evaluating the performance of our method considering diverse parameters. *Equal frequency* and *equal width* refers to the discretization method; *Bins* is the number of bins used to discretize the data; $\%\ ATT$ and $|A|$ are respectively the percentage and number of attributes randomly selected from the original set; $|F|$ is the number of features in the dataset, i.e., the size of the union of all attribute domains; $\mu|E|$ is the average edge length in hypergraphs generated for the dataset, $\sigma|E|$ is its standard deviation; $max(|E|)$ and $min(|E|)$ are the maximum and minimum edge lengths.

	Lymphoma — ($\|S^-\| = 50, \|S^+\| = 46$)							
	Bins	% ATT	$\|A\|$	$\|F\|$	$\mu\|E\|$	$\sigma\|E\|$	$max(\|E\|)$	$min(\|E\|)$
equal frequency	2	1	40	80	22.90	3.94	36	9
	2	2	80	160	44.33	6.64	66	24
	2	5	201	402	110.01	13.93	152	59
	2	10	402	804	218.09	26.07	297	126
	2	15	603	1206	326.01	36.73	437	194
	2	25	1006	2012	542.16	58.45	724	343
	4	1	40	160	32.36	2.96	40	20
	4	2	80	320	63.60	4.65	77	43
	4	5	201	804	158.18	9.67	186	119
	4	10	402	1608	314.78	17.71	359	235
	4	15	603	2412	471.23	24.82	529	359
	4	25	1006	4024	784.69	39.42	869	610
equal width	2	1	40	80	19.42	3.88	31	7
	2	2	80	160	36.31	6.33	54	16
	2	5	201	402	87.14	12.46	127	45
	2	10	402	804	176.04	22.48	253	101
	2	15	603	1206	262.09	32.71	376	150
	2	25	1006	2012	435.40	53.99	624	252
	4	1	40	160	27.02	3.50	37	14
	4	2	80	320	52.47	5.90	72	29
	4	5	201	803	130.15	12.06	171	89
	4	10	402	1605	258.43	22.50	334	175
	4	15	603	2408	387.95	32.75	498	263
	4	25	1006	4019	645.62	53.35	820	449