Multi-label Learning(MLL)

Classifier Chain

speaker Suhe Ye

College of Mathematics And Statistics, ShenZhen University

October 8, 2018

Outline

Outline

Multi-Label Object

Sunset

Clouds

Trees

Countryside

......

Multi-label Learning(MLL)

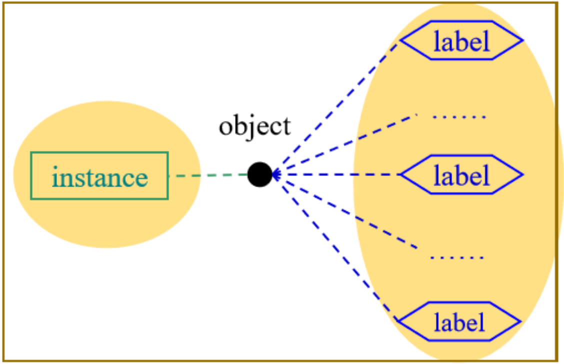


Figure: Multi-label Learning

MLL- Applications

- Text Categorization
- Automatic annotation for multimedia contents
  - Image, Audio, Video
- Bioinformatics
- World Wide Web
- Information Retrieval
- Directed marketing

Formal Definition of MLL

Settings

$\mathcal{X}$: d -dimensional feature space $\mathbb{R}^d$

$\mathcal{Y}$: label space with L labels $[1, 2, 3, \ldots, L]$

Inputs

$\mathcal{D}$: training set with N examples $(x_i, Y_i)|1 < i < N$

$x_i \in \mathcal{X}$ is a d -dimensional feature vector $(x_{i1}, x_{i2}, x_{i3}, \ldots, x_{id})^T$

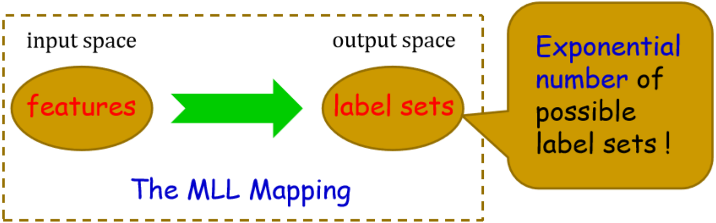
$Y_i \in \mathcal{Y}$ is the label set associated with x_i.

Outputs

h : multi-label predictor $\mathcal{X} \to 2^{\mathcal{Y}}$

Outline

The Major Challenge



q=5 ➔ 32 label sets

q=10 ➔ ~1k label sets

q=20 ➔ ~1M label sets

The Basic Philosophy

Exploiting Label Correlations

For example

An image labeled as lions and grassland would be likely annotated with label Africa

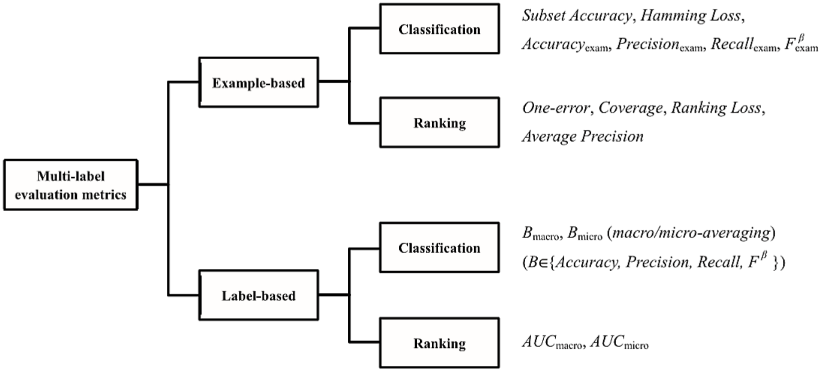
A document labeled as politics would be unlikely labeled as entertainment

Order Of Correlations

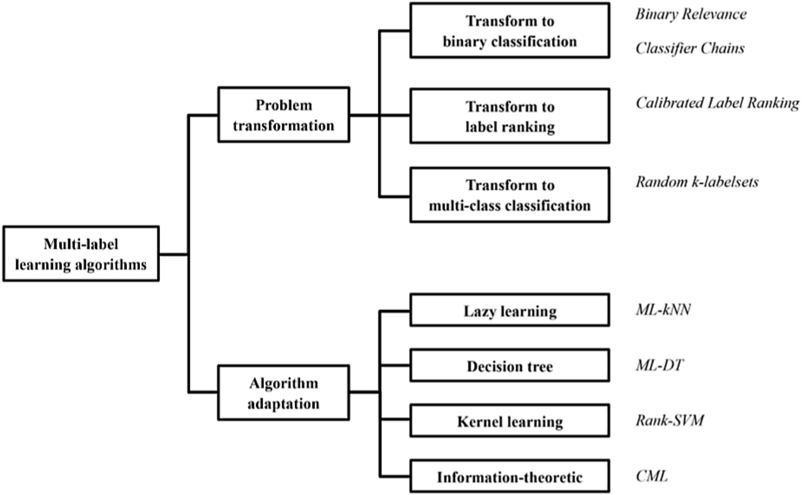
- First-Order Strategy : Tackle MLL problem in a label-by-label style, ignore the co-existence of other labels.
- Second-Order Strategy: Tackle MLL problem by considering pairwise relations between labels.
- High-Order Strategy : Tackle MLL problem by considering high-order relations between labels.

Outline

The Major Evaluation Metrics



The Major Learning Algorithm



Outline

Basic Idea

The Classifiers Chain model(CC) involves $|L|$ binary classifiers as in BM. Classifiers are linked along a chain where each classifier deals with the binary relevance problem associated with label $l_j \in L$. The feature space of each classifier in the chain is extended with the 0/1 label associations of all previous classifiers. Table 1 shows a simple example of CC model with the input $x_1 = [1, 0, 1, 0, 0, 1]$. Finally, it forms a classifiers chain.

h_i	y
h_1 :[1, 0, 1, 0, 0, 1]	1
h_2 :[1, 0, 1, 0, 0, 1, 1]	0
h_3 :[1, 0, 1, 0, 0, 1, 1, 0]	1
h_4 :[1, 0, 1, 0, 0, 1, 1, 0, 1]	0
h_5 :[1, 0, 1, 0, 0, 1, 1, 0, 1, 0]	0
h_6 :[1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0]	1

- Pros: more appropriate for realistic correlations.
- Cons: high model complexity, less scalable.

Outline

Framework

1. The conditional entropy builds a complete label-correlated graph, according the maximum conditional entropy principle of the deleted loop, we obtain a directed acyclic graph, i.e. the initial network.
2. The initial network is sorted by topology, and the initial order is obtained. According to algorithm 1, the optimal parent label sets is obtained.
3. To get the optimal Bayesian networks(BN) structure according to Algorithm 2.
4. Topological sorting of BN labels, training and testing.

Conditional Entropy

If $H(Y|X=x)$ is the entropy of the discrete random variable Y conditioned on the discrete random variable X taking a certain value x, then $H(Y|X)$ is the result of averaging $H(Y|X=x)$ over all possible values x that X may take.

$$\begin{aligned} H(L_j|L_i) &= \sum_{l_i \in L_i} p(l_i) H(L_j|L_i = l_i) \\ &= -\sum_{l_i \in L_i} p(l_i) \sum_{l_j \in L_j} p(l_j|l_i) log(l_j|l_i) \\ &= -\sum_{l_i \in L_i} \sum_{l_j \in L_j} p(l_i, l_j) logp(l_j|l_i) \\ &= -\sum_{l_i \in L_i} \sum_{l_j \in L_j} p(l_i, l_j) log \frac{p(l_i, l_j)}{p(l_i)} \end{aligned} \tag{1}$$

conditional entropy is measured correlation between labels.

Outline

Bayesian information Criterion Scoring Function

$$S(G) = L_D(G) - \frac{Dim_G}{2} logN \tag{2}$$

$$L_D(G) = -N \sum_{i=1}^{L} H(I_i | PA(I_i)) \tag{3}$$

$$S(G) = \sum_{i=1}^{L} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} log \frac{N_{ijk}}{N_{ij}} - \sum_{i=1}^{L} q_i (r_i - 1) logN \tag{4}$$

learning the optimal parent label sets

Algorithm 1 learning the optimal parent label sets

input: D:training data set containing L labels $(l_1, l_2, l_3, \ldots, l_L)$
initial order: $l_1, l_2, l_3, \ldots, l_L$
output: PA: optimal parent label sets PA of each label

```
PA = ∅;
for i = 1 to L do
    Pred_i = [l_1, l_2, ..., l_{i-1}, l_{i+1}, ..., l_L];
    PA(l_i) = ∅;
    maxBIC = BIC(l_i|PA(l_i));
    while j ≤ L ⋃ N_{PA(l_i)} ≤ log_2 N do
        let l_z be the label in Pred_i PA(l_i) that maximizes
        BIC(l_i|PA(l_i) ⋃ l_z);
        BIC_new = BIC(l_i|PA(l_i) ⋃ l_z);
        if BIC > maxBIC then
            PA(l_i) = PA(l_i) ⋃ l_z;
        end if
    end while
    PA = PA ⋃ PA(l_i)
end for
return PA;
```

Learning Bayesian networks structure

Algorithm 2 learning the optimal Baysian networks structure

input: G: A directed graph built by best parent label sets
output: G′:the optimal Baysian networks(a directed acyclic graph)

```
while G has circles do
    CG=circle
    maxweight=-inf;
    for i = 1 to size(CG, 1) do
        for j = 1 to size(CG, 1) do
            if CG(i, j) > maxweight then
                maxweight = CG(i, j);
                row=i;
                column=j;
            end if
        end for
    end for
    CG(row,column)=0;
end while
G′=G;
return G′;
```

Outline

Table: characteristics of data sets

data sets	Instances	Labels	Features	Domain
emotions	593	6	72	music
scene	2407	6	294	image
flags	194	7	19	image
yeast	2417	14	103	biology
art	5000	26	462	text
genbase	662	27	1185	biology
education	5000	33	550	text
science	5000	40	743	txet
medical	978	45	1149	text
enron	1702	53	1001	text
CAL500	502	174	68	music

hammingloss

datasets	BR	LR	CC	GCC	LPLC	BNCC
emotions	0.1808 ± 0.0065	$\mathbf{0.1795 \pm 0.0103}$	0.1827 ± 0.0094	0.3163 ± 0.0351	0.3032 ± 0.0236	0.1815 ± 0.0038
flags	0.2780 ± 0.0127	0.2681 ± 0.0105	0.2668 ± 0.0153	0.3220 ± 0.0336	0.4411 ± 0.0393	$\mathbf{0.2645 \pm 0.0020}$
scene	0.0787 ± 0.0035	0.0775 ± 0.0035	0.0800 ± 0.0038	0.1030 ± 0.0091	0.1046 ± 0.0076	$\mathbf{0.0746 \pm 0.0019}$
yeast	0.1948 ± 0.0042	0.1981 ± 0.0037	0.1934 ± 0.0041	$\mathbf{0.1884 \pm 0.0228}$	0.2267 ± 0.0088	0.1908 ± 0.0021
medical	0.0094 ± 0.0006	0.0249 ± 0.0012	0.0094 ± 0.0006	0.7745 ± 0.0404	0.0168 ± 0.0021	$\mathbf{0.0092 \pm 0.0004}$
enron	0.0506 ± 0.0028	0.0518 ± 0.0028	0.0501 ± 0.0026	$\mathbf{0.0467 \pm 0.0019}$	0.0686 ± 0.0076	0.0499 ± 0.0005
genbase	0.0006 ± 0.0007	0.0265 ± 0.0090	0.0009 ± 0.0008	0.0007 ± 0.0006	0.0032 ± 0.0020	$\mathbf{0.0005 \pm 0.0001}$
science	$\mathbf{0.0330 \pm 0.0011}$	0.0347 ± 0.0007	0.0337 ± 0.0007	0.0330 ± 0.0016	0.0415 ± 0.0076	0.0334 ± 0.0007
art	$\mathbf{0.0523 \pm 0.0007}$	0.0570 ± 0.0008	0.0539 ± 0.0012	0.0650 ± 0.0146	0.1014 ± 0.0196	0.0539 ± 0.0003
CAL500	$\mathbf{0.1376 \pm 0.0043}$	0.1464 ± 0.0033	0.1888 ± 0.0159	0.1429 ± 0.0043	0.3692 ± 0.0581	0.1794 ± 0.0020
education	$\mathbf{0.0376 \pm 0.0009}$	0.0415 ± 0.0007	0.0385 ± 0.0004	0.0434 ± 0.0102	0.0482 ± 0.0055	0.0380 ± 0.0006

examaccuracy

datasets	BR	LR	CC	GCC	LPLC	BNCC
emotions	0.5627 ± 0.0180	0.5518 ± 0.0217	0.5830 ± 0.0233	0.4024 ± 0.0490	0.4053 ± 0.0279	$\mathbf{0.5837 \pm 0.0057}$
flags	0.5770 ± 0.0220	0.5703 ± 0.0204	0.5842 ± 0.0158	0.5067 ± 0.0440	0.5103 ± 0.0329	$\mathbf{0.5864 \pm 0.0058}$
scene	0.7163 ± 0.0144	0.6975 ± 0.0094	0.7361 ± 0.0128	0.6914 ± 0.0289	0.6451 ± 0.0255	$\mathbf{0.7414 \pm 0.0046}$
yeast	0.5088 ± 0.0083	0.4778 ± 0.0085	0.5251 ± 0.0094	0.2062 ± 0.0102	0.5246 ± 0.0193	$\mathbf{0.5350 \pm 0.0035}$
medical	0.7610 ± 0.0138	0.1186 ± 0.0413	0.7757 ± 0.0140	$\mathbf{0.8197 \pm 0.0416}$	0.6076 ± 0.0435	0.7844 ± 0.0014
enron	0.4416 ± 0.0267	0.3599 ± 0.0397	0.4499 ± 0.0222	$\mathbf{0.4626 \pm 0.0211}$	0.0608 ± 0.0763	0.4517 ± 0.0046
genbase	0.9919 ± 0.0104	0.3695 ± 0.2038	0.9899 ± 0.0089	0.9917 ± 0.0085	0.9715 ± 0.0149	$\mathbf{0.9933 \pm 0.0023}$
science	0.3135 ± 0.0177	0.1150 ± 0.0088	0.3175 ± 0.0123	$\mathbf{0.3671 \pm 0.0206}$	0.2570 ± 0.0193	0.3254 ± 0.0137
art	0.2990 ± 0.0081	0.1456 ± 0.0143	0.3114 ± 0.0065	$\mathbf{0.3713 \pm 0.0730}$	$\mathbf{0.0521 \pm 0.0462}$	0.3128 ± 0.0055
CAL500	0.2051 ± 0.0083	0.0909 ± 0.0102	0.2021 ± 0.0192	0.2106 ± 0.0158	$\mathbf{0.2376 \pm 0.0137}$	0.2054 ± 0.0034
education	0.3171 ± 0.0104	0.1092 ± 0.0104	0.3260 ± 0.0070	$\mathbf{0.3596 \pm 0.0571}$	0.3159 ± 0.0262	0.3292 ± 0.0066

examF

datasets	BR	LR	CC	GCC	LPLC	BNCC
emotions	0.6763 ± 0.0169	0.6584 ± 0.0201	$\mathbf{0.6957 \pm 0.0235}$	0.4796 ± 0.0504	0.4907 ± 0.0344	0.6952 ± 0.0044
flags	0.7046 ± 0.0218	0.6899 ± 0.0217	0.7073 ± 0.0133	0.6418 ± 0.0374	0.6573 ± 0.0256	$\mathbf{0.7149 \pm 0.0048}$
scene	0.7493 ± 0.0129	0.7270 ± 0.0087	0.7601 ± 0.0127	0.7028 ± 0.0290	0.6595 ± 0.0251	$\mathbf{0.7633 \pm 0.0037}$
yeast	0.6434 ± 0.0087	0.6205 ± 0.0088	0.6550 ± 0.0103	0.6231 ± 0.0247	0.6348 ± 0.0185	$\mathbf{0.6632 \pm 0.0040}$
medical	0.7993 ± 0.0143	$NaN \pm NaN$	0.8107 ± 0.0129	0.7012 ± 0.0437	0.6425 ± 0.0417	$\mathbf{0.8184 \pm 0.0018}$
enron	0.5787 ± 0.0250	0.4900 ± 0.0449	0.5830 ± 0.0230	0.5717 ± 0.0198	0.0764 ± 0.0951	$\mathbf{0.5856 \pm 0.0060}$
genbase	0.9953 ± 0.0067	0.3792 ± 0.2051	0.9941 ± 0.0059	0.9942 ± 0.0064	0.9802 ± 0.0108	$\mathbf{0.9962 \pm 0.0017}$
science	0.350 ± 30.0175	0.1292 ± 0.0099	0.3521 ± 0.0140	$\mathbf{0.3841 \pm 0.0214}$	0.2750 ± 0.0214	0.3613 ± 0.0147
art	0.3343 ± 0.0093	0.1670 ± 0.0167	0.3474 ± 0.0063	$\mathbf{0.4045 \pm 0.0558}$	0.0603 ± 0.0564	0.3481 ± 0.0070
CAL500	0.3401 ± 0.0115	0.1672 ± 0.0174	0.3347 ± 0.0266	0.3412 ± 0.0205	$\mathbf{0.3796 \pm 0.0169}$	0.3413 ± 0.0050
education	0.3512 ± 0.0107	0.1261 ± 0.0113	0.3607 ± 0.0079	$\mathbf{0.3876 \pm 0.0420}$	0.3457 ± 0.0259	0.3623 ± 0.0077

macroF

datasets	BR	LR	CC	GCC	LPLC	BNCC
emotions	0.6558 ± 0.0210	0.6571 ± 0.0155	0.6724 ± 0.0201	0.4776 ± 0.0687	0.5120 ± 0.0408	$\mathbf{0.6736 \pm 0.0075}$
flags	0.5912 ± 0.0351	$\mathbf{0.6449 \pm 0.0303}$	0.6012 ± 0.0240	0.4361 ± 0.0489	0.6026 ± 0.0295	0.6219 ± 0.0214
scene	0.7809 ± 0.0091	0.7750 ± 0.0086	0.7769 ± 0.0109	0.7137 ± 0.0254	0.6851 ± 0.0220	$\mathbf{0.7891 \pm 0.0051}$
yeast	0.3355 ± 0.0051	0.3118 ± 0.0040	0.3628 ± 0.0079	0.3535 ± 0.0156	$\mathbf{0.4501 \pm 0.0207}$	0.3717 ± 0.0034
medical	0.3385 ± 0.0158	0.0837 ± 0.0272	0.3375 ± 0.0124	0.3185 ± 0.0289	0.1918 ± 0.0223	$\mathbf{0.3401 \pm 0.0110}$
enron	0.1940 ± 0.0230	0.1607 ± 0.0261	$\mathbf{0.1942 \pm 0.0205}$	0.1462 ± 0.0137	0.0071 ± 0.0126	0.1852 ± 0.0070
genbase	0.6550 ± 0.0435	0.3073 ± 0.1402	0.6537 ± 0.0697	0.6387 ± 0.0655	0.5801 ± 0.0438	$\mathbf{0.6582 \pm 0.0023}$
science	0.2073 ± 0.0246	0.1301 ± 0.0137	$\mathbf{0.2147 \pm 0.0155}$	0.1743 ± 0.0172	0.1712 ± 0.0172	0.2068 ± 0.0075
art	0.2113 ± 0.0048	0.1326 ± 0.0156	$\mathbf{0.2212 \pm 0.0150}$	0.1932 ± 0.0171	0.0067 ± 0.0089	0.2185 ± 0.0067
CAL500	0.0613 ± 0.0038	0.0376 ± 0.0034	0.1008 ± 0.0110	0.0552 ± 0.0070	$\mathbf{0.2194 \pm 0.0105}$	0.1162 ± 0.0037
education	0.1901 ± 0.0115	0.1240 ± 0.0093	0.1883 ± 0.0155	0.1660 ± 0.0232	0.1606 ± 0.0185	$\mathbf{0.1901 \pm 0.0109}$

microF

datasets	BR	LR	CC	GCC	LPLC	BNCC
emotions	0.6872 ± 0.0165	0.6831 ± 0.0191	0.6960 ± 0.0189	0.5093 ± 0.0663	0.5312 ± 0.0400	$\mathbf{0.6979 \pm 0.0068}$
flags	0.7176 ± 0.0206	0.7233 ± 0.0164	0.7236 ± 0.0140	0.6607 ± 0.0330	0.6741 ± 0.0282	$\mathbf{0.7310 \pm 0.0059}$
scene	0.7723 ± 0.0104	0.7692 ± 0.0099	0.7697 ± 0.0107	0.7030 ± 0.0265	0.6892 ± 0.0221	$\mathbf{0.7826 \pm 0.0045}$
yeast	0.6403 ± 0.0070	0.6174 ± 0.0074	0.6516 ± 0.0080	0.6434 ± 0.0186	0.6468 ± 0.0161	$\mathbf{0.6601 \pm 0.0034}$
medical	0.8218 ± 0.0112	0.1919 ± 0.0625	0.8209 ± 0.0104	0.8202 ± 0.0304	0.6785 ± 0.0380	$\mathbf{0.8262 \pm 0.0065}$
enron	0.5580 ± 0.0217	0.4810 ± 0.0344	0.5614 ± 0.0194	$\mathbf{0.5765 \pm 0.0167}$	0.0638 ± 0.0823	0.5626 ± 0.0058
genbase	0.9934 ± 0.0078	0.5530 ± 0.2397	0.9906 ± 0.0081	0.9926 ± 0.0069	0.9657 ± 0.0188	$\mathbf{0.9941 \pm 0.0016}$
science	0.4016 ± 0.0178	0.1955 ± 0.0142	0.3966 ± 0.0108	$\mathbf{0.4156 \pm 0.0192}$	0.3148 ± 0.0250	0.4044 ± 0.0130
art	0.3822 ± 0.0081	0.2222 ± 0.0212	0.3877 ± 0.0118	$\mathbf{0.3952 \pm 0.0491}$	0.0577 ± 0.0545	0.3858 ± 0.0064
CAL500	0.3338 ± 0.0114	0.1652 ± 0.0158	0.3322 ± 0.0257	0.3380 ± 0.0210	$\mathbf{0.3813 \pm 0.0170}$	0.3365 ± 0.0045
education	0.4206 ± 0.0107	0.1920 ± 0.0152	0.4217 ± 0.0085	0.4234 ± 0.0456	0.3855 ± 0.0281	$\mathbf{0.4243 \pm 0.0104}$

Thank you