Spotkanie polskiej grupy badawczej systemów uczących się

Selekcja zmiennych w klasyfikacji wielo-etykietowej w oparciu o łańcuchy klasyfikatorów i sieć elastyczną

Paweł Teisseyre

Instytut Podstaw Informatyki PAN

Plan prezentacji

- Klasyfikacja z wieloma etykietami.
- Metoda CCnet: łańcuchy klasyfikatorów + sieć elastyczna.
- Eksperymenty.

Klasyfikacja z wieloma etykietami

Klasyfikacja z jedną etykietą

- Jedna zmienna odpowiedzi.
- Zbiór uczący: $(\mathbf{x}_i, y_i)$, $\mathbf{x}_i \in R^p$, $y_i \in \{0, 1\}$.

Klasyfikacja z wieloma etykietami

- Wiele zmiennych odpowiedzi.
- Zbiór uczący: $(\mathbf{x}_i, \mathbf{y}_i)$, $\mathbf{x}_i \in R^p$, $\mathbf{y}_i \in \{0, 1\}^K$.

Przykład: wielozachorowalność

BMI	Weight	Glucose	...	Diabetes	Hypotension	Liver disease	...
31	84	10	...	1	0	1	...
26	63	6	...	1	0	0	...
27	60	7	...	0	0	0	...

Zmienne x: charakterystyki pacjentów.

Etykiety y: wystąpienia chorób.

- Zadanie 1: przewidywanie które choroby wystąpią na podstawie pewnych charakterystyk pacjentów (PREDYCKJA).
- Zadanie 2: wyznaczenie które zmienne wpływają na występowanie poszczególnych chorób (SELEKCJA ZMIENNYCH).

Klasyfikacja z wieloma etykietami

Naturalne podejście:

1. Oszacowanie prawdopodobieństwa aposteriori:

$$p(\mathbf{y}|\mathbf{x})$$

2. Predykcja dla nowej obserwacji $\mathbf{x}_0$:

$$\hat{\mathbf{y}}(\mathbf{x}_0) = \arg \max_{\mathbf{y} \in \{0,1\}^K} \hat{p}(\mathbf{y}|\mathbf{x}_0),$$

gdzie $\hat{p}(\mathbf{y}|\mathbf{x}_0)$ to oszacowane prawdopodobieństwo aposteriori w $\mathbf{x}_0$.

Metoda CCnet

Estymacja prawdopodobieństwa aposteriori:

- Rozważamy rodzinę rozkładów: $\{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$.
- Estymujemy parametry używając metody NW:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \{-\frac{1}{n} \sum_{i=1}^{n} \log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta})\}.$$

- Wersja z regularyzacją:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \{-\frac{1}{n} \sum_{i=1}^{n} \log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta}) + \lambda_1 ||\boldsymbol{\theta}||_1 + \lambda_2 ||\boldsymbol{\theta}||_2^2\}.$$

Zaleta regularyzacji ℓ_1: część współrzędnych $\hat{\boldsymbol{\theta}}$ będzie równa 0 (selekcja zmiennych).

Metoda CCnet

Estymacja prawdopodobieństwa aposteriori:

- Rozważamy rodzinę rozkładów: $\{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$.
- Estymujemy parametry używając metody NW:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \{-\frac{1}{n} \sum_{i=1}^{n} \log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta})\}.$$

- Wersja z regularyzacją:

$$\hat{\theta} = \arg\min_{\theta} \{-\frac{1}{n} \sum_{i=1}^{n} \log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \theta) + \lambda_1 ||\theta||_1 + \lambda_2 ||\theta||_2^2\}.$$

Zaleta regularyzacji ℓ_1: część współrzędnych $\hat{\theta}$ będzie równa 0 (selekcja zmiennych).

Metoda CCnet

Estymacja prawdopodobieństwa aposteriori:

- Rozważamy rodzinę rozkładów: $\{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$.
- Estymujemy parametry używając metody NW:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \{-\frac{1}{n} \sum_{i=1}^{n} \log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta})\}.$$

- Wersja z regularyzacją:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \{-\frac{1}{n} \sum_{i=1}^{n} \log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta}) + \lambda_1 ||\boldsymbol{\theta}||_1 + \lambda_2 ||\boldsymbol{\theta}||_2^2\}.$$

Zaleta regularyzacji ℓ_1: część współrzędnych $\hat{\theta}$ będzie równa 0 (selekcja zmiennych).

Metoda CCnet

Estymacja prawdopodobieństwa aposteriori:

- Rozważamy rodzinę rozkładów: $\{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$.
- Estymujemy parametry używając metody NW:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \{-\frac{1}{n}\sum_{i=1}^{n} \log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta})\}.$$

- Wersja z regularyzacją:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \{-\frac{1}{n}\sum_{i=1}^{n} \log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta}) + \lambda_1 ||\boldsymbol{\theta}||_1 + \lambda_2 ||\boldsymbol{\theta}||_2^2\}.$$

Zaleta regularyzacji ℓ_1: część współrzędnych $\hat{\boldsymbol{\theta}}$ będzie równa 0 (selekcja zmiennych).

Metoda CCnet

Estymacja prawdopodobieństwa aposteriori:

- Zamiast modelować $p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})$ bezpośrednio, używamy wzoru łańcuchowego:

$$p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}) = p(y_1|\mathbf{x},\boldsymbol{\theta}_1)\prod_{k=2}^{K} p(y_k|\mathbf{y}_{-k},\mathbf{x},\boldsymbol{\theta}_k),$$

gdzie: $\mathbf{y}_{-k} = (y_1,\ldots,y_{k-1})^T$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K)^T$.

Metoda CCnet

- Problem:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \{ -\frac{1}{n} \sum_{i=1}^{n} \log p(\mathbf{y}^{(i)} | \mathbf{x}^{(i)}, \boldsymbol{\theta}) + \lambda_1 ||\boldsymbol{\theta}||_1 + \lambda_2 ||\boldsymbol{\theta}||_2^2 \}.$$

- Rozwiązanie w postaci: $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_K)^T$, gdzie:

$$\hat{\theta}_k = \arg\min_{\boldsymbol{\theta}_k} \{ -\frac{1}{n} \sum_{i=1}^{n} \log p(y_k^{(i)} | \mathbf{y}_{-k}^{(i)}, \mathbf{x}^{(i)}, \boldsymbol{\theta}_k) + \lambda_1 ||\boldsymbol{\theta}_k||_1 + \lambda_2 ||\boldsymbol{\theta}_k||_2^2 \},$$

dla $k = 1, \ldots, K$.

Metoda CCnet

- Problem:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \{ -\frac{1}{n} \sum_{i=1}^{n} \log p(\mathbf{y}^{(i)} | \mathbf{x}^{(i)}, \boldsymbol{\theta}) + \lambda_1 ||\boldsymbol{\theta}||_1 + \lambda_2 ||\boldsymbol{\theta}||_2^2 \}.$$

- Rozwiązanie w postaci: $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_K)^T$, gdzie:

$$\hat{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta}_k} \{ -\frac{1}{n} \sum_{i=1}^{n} \log p(y_k^{(i)} | \mathbf{y}_{-k}^{(i)}, \mathbf{x}^{(i)}, \boldsymbol{\theta}_k) + \lambda_1 ||\boldsymbol{\theta}_k||_1 + \lambda_2 ||\boldsymbol{\theta}_k||_2^2 \},$$

dla $k = 1, \ldots, K$.

Metoda CCnet

- Zakładamy że prawdopodobieństwa warunkowe są w postaci:

$$p(y_k|\mathbf{z}_k, \boldsymbol{\theta}_k) = \frac{\exp(\boldsymbol{\theta}_k^T \mathbf{z}_k y_k)}{1 + \exp(\boldsymbol{\theta}_k^T \mathbf{z}_k)},$$

  gdzie: $\mathbf{z}_k = (\mathbf{y}_{-k}, \mathbf{x})^T$.

- Rozważamy parametry regularyzacji $\lambda_{1,k}$ and $\lambda_{2,k}$ niezależnie, dla każdego k.

- Rozwiązanie w postaci: $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_K)^T$, gdzie:

$$\hat{\theta}_k = \arg\min_{\theta_k} \{-\frac{1}{n}\sum_{i=1}^{n}[\theta_k^T \mathbf{z}_k^{(i)} y_k^{(i)} - \log(1+\exp(\theta_k^T \mathbf{z}_k^{(i)}))] + \lambda_{1,k}||\theta_k||_1 + \lambda_{2,k}||\theta_k||_2^2\},$$

  dla $k = 1, \ldots, K$.

Metoda CCnet

- Zakładamy że prawdopodobieństwa warunkowe są w postaci:

$$p(y_k|\mathbf{z}_k, \boldsymbol{\theta}_k) = \frac{\exp(\boldsymbol{\theta}_k^T \mathbf{z}_k y_k)}{1 + \exp(\boldsymbol{\theta}_k^T \mathbf{z}_k)},$$

gdzie: $\mathbf{z}_k = (\mathbf{y}_{-k}, \mathbf{x})^T$.

- Rozważamy parametry regularyzacji $\lambda_{1,k}$ and $\lambda_{2,k}$ niezależnie, dla każdego k.

- Rozwiązanie w postaci: $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_K)^T$, gdzie:

$$\hat{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta}_k}\{-\frac{1}{n}\sum_{i=1}^{n}[\boldsymbol{\theta}_k^T \mathbf{z}_k^{(i)} y_k^{(i)} - \log(1+\exp(\boldsymbol{\theta}_k^T \mathbf{z}_k^{(i)}))] + \lambda_{1,k}||\boldsymbol{\theta}_k||_1 + \lambda_{2,k}||\boldsymbol{\theta}_k||_2^2\},$$

dla $k = 1, \ldots, K$.

Metoda CCnet

- Rozwiązanie w postaci: $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_K)^T$, gdzie:

$$\hat{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta}_k}\{-\frac{1}{n}\sum_{i=1}^{n}[\boldsymbol{\theta}_k^T \mathbf{z}_k^{(i)} y_k^{(i)} - \log(1+\exp(\boldsymbol{\theta}_k^T \mathbf{z}_k^{(i)}))] + \lambda_{1,k}||\boldsymbol{\theta}_k||_1 + \lambda_{2,k}||\boldsymbol{\theta}_k||_2^2\},$$

dla $k = 1, \ldots, K$.

- K problemów optymalizacji wypukłej: do rozwiązania można użyć algorytmu CCD (funkcja `glmnet` w `R`).

Metoda CCnet

- Rozwiązanie w postaci: $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_K)^T$, gdzie:

$$\hat{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta}_k} \{-\frac{1}{n}\sum_{i=1}^{n}[\boldsymbol{\theta}_k^T \mathbf{z}_k^{(i)} y_k^{(i)} - \log(1+\exp(\boldsymbol{\theta}_k^T \mathbf{z}_k^{(i)}))] + \lambda_{1,k}||\boldsymbol{\theta}_k||_1 + \lambda_{2,k}||\boldsymbol{\theta}_k||_2^2\},$$

dla $k = 1, \ldots, K$.

$$\Downarrow$$

- K problemów optymalizacji wypukłej: do rozwiązania można użyć algorytmu CCD (funkcja `glmnet` w `R`).

Wybór parametru regularyzacji

- Przyjmujemy $\lambda_{1,k} = \alpha\lambda_k$, $\lambda_{2,k} = (1-\alpha)\lambda_k$, gdzie: $\alpha \in [0,1]$.
- Dla ustalonego α znajdujemy optymalną wartość λ_k używając kryterium GIC (Generalized Information Criterion):

$$\hat{\lambda}_k = \arg\min_{\lambda_k}\{-\frac{1}{n}\sum_{i=1}^{n}\log p(y_k^{(i)}|\mathbf{z}_k^{(i)}, \hat{\boldsymbol{\theta}}_k) + a(n)\cdot df\},$$

gdzie:

- $df := |\{r : \hat{\boldsymbol{\theta}}_{k,r} \neq 0\}|$,
- $a(n) = \log(n)$ (BIC) lub $a(n) = 2$ (AIC).

Wyniki teoretyczne

- Stabilność CCnet ze względu na wybraną funkcję straty: niewielka zmiana zbioru treningowego nie wpływa znacząco na wartość funkcji straty dla CCnet.
- Oszacowanie błędu generalizacji dla CCnet, dla wybranej funkcji straty: używamy pomysłu opisanego w pracy Bousquet & Elisseeff (JMLR 2002), który pozwala udowodnić oszacowanie błędu generalizacji używając stabilności.

Eksperymenty

Porównanie metod:

1. *BRlogit*[1],
2. *BRtree*,
3. *BRnet*, dla $\alpha = 0, 0.5, 1$[2],
4. *CClogit*[3],
5. *CCtree*,
6. *CCnet*, dla $\alpha = 0, 0.5, 1$.

---

[1] Dembczynski et. al. 2012

[2] Liu 2015

[3] Kumar et. al. 2013; Montanes 2014; Dembczynski et. al. 2012

Eksperymenty

Miary oceny:

- Dokładność zbioru:

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

- Miara Hamminga:

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

- Liczba wybranych zmiennych.
- Czas budowy modelu.

Eksperymenty

Dataset	CClogit	CCtree	CCnet ($\alpha = 1$)	CCnet ($\alpha = 0$)	CCnet ($\alpha = 0.5$)	BRlogit	BRtree	BRnet ($\alpha = 1$)	BRnet ($\alpha = 0$)	BRnet ($\alpha = 0.5$)
music	0.215	0.221	0.275	0.267	**0.282**	0.186	0.191	0.257	0.253	0.220
yeast	0.214	0.168	0.184	**0.226**	0.192	0.156	0.048	0.123	0.163	0.115
scene	0.473	0.467	0.629	**0.639**	0.592	0.385	0.337	0.416	0.543	0.356
birds	0.349	0.375	0.532	0.535	**0.538**	0.332	0.375	0.535	0.535	0.538
flags	**0.227**	0.139	0.216	0.196	0.196	0.124	0.072	0.165	0.139	0.144
medical	0.181	0.690	**0.760**	0.218	0.697	0.180	0.634	0.752	0.218	0.667
cal500	0.008	0.006	**0.020**	0.014	0.020	0.012	0.004	0.008	0.010	0.010
genbase	**0.989**	0.985	0.986	0.029	0.986	0.989	0.985	0.989	0.029	0.986
mediamill	0.223	0.096	0.197	**0.200**	0.197	0.192	0.057	0.155	0.160	0.156
enron	0.037	0.170	0.139	**0.210**	0.136	0.038	0.130	0.095	0.202	0.090
bookmarks	0.108	0.287	**0.754**	0.741	0.754	0.067	0.292	0.754	0.739	0.754
bibtex	0.361	0.404	0.780	**0.788**	0.777	0.359	0.414	0.780	0.787	0.777
avg rank	5.269	4.077	**7.769**	**7.769**	7.577	3.500	2.538	5.808	5.769	4.923

Tabela : Dokładność zbioru. Parametr λ wybrany za pomocą BIC. Pogrubione liczby odpowiadają najlepszej metodzie.

Eksperymenty

Dataset	CClogit	CCtree	CCnet ($\alpha = 1$)	CCnet ($\alpha = 0$)	CCnet ($\alpha = 0.5$)	BRlogit	BRtree	BRnet ($\alpha = 1$)	BRnet ($\alpha = 0$)	BRnet ($\alpha = 0.5$)
music	0.749	0.720	0.775	0.782	0.778	0.765	0.734	0.794	**0.796**	0.783
yeast	0.722	0.659	0.724	0.731	0.730	0.741	0.631	0.744	**0.744**	0.742
scene	0.843	0.838	0.884	0.900	0.873	0.846	0.831	0.887	**0.901**	0.878
birds	0.809	0.864	0.923	0.915	0.922	0.804	0.863	**0.924**	0.914	0.922
flags	0.706	0.658	**0.733**	0.719	0.728	0.716	0.669	0.731	0.725	0.721
medical	0.774	0.956	**0.968**	0.906	0.962	0.773	0.955	0.967	0.906	0.959
cal500	0.588	0.545	0.616	0.608	**0.616**	0.596	0.541	0.615	0.600	0.616
genbase	0.999	0.998	0.999	0.901	0.999	0.999	0.998	**0.999**	0.901	0.999
mediamill	0.825	0.703	0.816	0.820	0.816	**0.833**	0.688	0.822	0.825	0.823
enron	0.589	0.780	0.806	**0.834**	0.809	0.605	0.772	0.816	0.834	0.815
bookmarks	0.719	0.770	**0.969**	0.967	0.969	0.684	0.822	0.969	0.967	0.969
bibtex	0.895	0.902	0.975	0.976	0.975	0.898	0.913	0.975	**0.976**	0.975
avg rank	3.654	2.654	6.654	6.308	6.308	4.423	2.500	**8.192**	7.231	7.077

Tabela : Miara Hamminga. Parametr λ wybrany za pomocą BIC. Pogrubione liczby odpowiadają najlepszej metodzie.

Eksperymenty

Dataset	CClogit	CCtree	CCnet ($\alpha = 1$)	CCnet ($\alpha = 0$)	CCnet ($\alpha = 0.5$)	BRlogit	BRtree	BRnet ($\alpha = 1$)	BRnet ($\alpha = 0$)	BRnet ($\alpha = 0.5$)
music	71	69	36	71	45	71	71	**34**	71	39
yeast	103	95	70	103	85	103	103	**54**	103	59
scene	294	177	161	294	212	294	190	**160**	294	190
birds	260	34	44	260	**30**	260	33	45	260	30
flags	19	19	16	19	16	19	19	**6**	19	8
medical	1449	80	33	1200	47	1449	75	**31**	1200	45
cal500	68	68	5	68	4	68	68	2	68	**1**
genbase	1186	28	**13**	97	26	1186	29	14	97	26
mediamill	120	81	75	120	96	120	94	**72**	120	91
enron	1001	303	86	1001	96	1001	374	**78**	1001	86
bookmarks	2150	450	**56**	2150	92	2150	453	57	2150	93
bibtex	1836	171	**108**	1836	133	1836	169	109	1836	131
avg rank	9	5	**2**	8	4	9	5	**2**	8	3

Tabela : Liczba wybranych zmiennych. Parametr λ wybrany za pomocą BIC. Pogrubione liczby odpowiadają najlepszej metodzie.

Eksperymenty

Dataset	CClogit	CCtree	CCnet ($\alpha = 1$)	CCnet ($\alpha = 0$)	CCnet ($\alpha = 0.5$)	BRlogit	BRtree	BRnet ($\alpha = 1$)	BRnet ($\alpha = 0$)	BRnet ($\alpha = 0.5$
music	0.96	0.59	0.55	0.94	0.47	**0.25**	0.54	0.61	0.74	0.56
yeast	3.57	5.69	3.06	4.32	3.75	**1.85**	7.50	3.33	4.43	3.38
scene	23.98	8.25	**2.08**	7.97	3.08	20.44	8.45	2.39	8.05	3.64
birds	7.94	2.18	1.18	3.22	1.29	7.68	2.18	**0.94**	3.72	1.43
flags	0.05	0.09	0.24	0.40	0.28	**0.04**	0.09	0.26	0.34	0.29
medical	1426.40	12.49	**5.45**	19.89	5.63	1419.75	11.40	5.59	20.54	5.48
cal500	0.24	0.91	0.66	1.05	0.71	**0.21**	1.31	0.63	1.04	0.67
genbase	1254.63	6.35	2.97	3.16	2.86	1257.29	6.03	**2.60**	3.31	2.78
mediamill	18.16	28.22	6.60	20.53	8.82	9.37	35.58	**6.07**	20.50	7.72
enron	197.43	25.30	8.19	44.82	10.37	188.95	28.81	**7.95**	42.51	11.43
bookmarks	3370.57	246.90	50.34	391.62	58.44	3100.20	249.42	**47.61**	385.67	58.53
bibtex	4911.14	209.11	70.49	525.61	82.10	5019.62	210.31	**67.61**	520.32	77.41
avg rank	7.46	6.62	2.77	7.62	4.15	5.77	6.85	**2.54**	7.38	3.85

Tabela : Czas budowy modelu. Parametr λ wybrany za pomocą BIC. Pogrubione liczby odpowiadają najlepszej metodzie.

Eksperymenty

Wnioski:

1. CCnet (z dowolną α) osiąga większą dokładność zbioru niż inne metody.
2. Wartość α nie ma bardzo dużego wpływu na dokładność i miarę Hamminga. Wartość $\alpha > 0$ jest zalecana ze względu na selekcję zmiennych.
3. BRnet osiąga największe wartości miary Hamminga.
4. Kara lasso (BRnet, $\alpha = 1$ oraz CCnet, $\alpha = 1$) pozwala na wybór najmniejszej liczby zmiennych.
5. Najmniejsze czasy dopasowania modelu obserwujemy dla kary lasso (BRnet, $\alpha = 1$ oraz CCnet, $\alpha = 1$).

Dziękuje za uwagę!