Klasyfikacja wieloetykietowa (Mutlilabel classification)

Paweł Teisseyre

SER 2014

Overview

1. Introduction
   - Examples
   - Loss functions
2. Methods
   - Binary Relevance
   - Label Powerset
   - Classifier chains
   - Ising Model
3. Experiments on real datasets in R.

Binary classification:

X_1	X_2	...	X_p	Y
1.0	2.2	...	4.2	1
2.4	1.3	...	3.1	1
0.9	1.4	...	3.2	0
⋮			⋮	⋮
1.7	3.5	...	4.2	0
3.9	2.5	...	4.1	?

Tabela : Binary classification.

- We consider one target variable.
- Prediction: predict y for a new instance $\mathbf{x}$ (2 possible values).

Multilabel binary classification:

X_1	X_2	...	X_p	Y_1	Y_2	...	Y_K
1.0	2.2	...	4.2	1	0	...	1
2.4	1.3	...	3.1	1	0	...	1
0.9	1.4	...	3.2	0	0	...	1
⋮			⋮	⋮			⋮
1.7	3.5	...	4.2	0	1	...	0
3.9	2.5	...	4.1	?	?	...	?

Tabela : Multilabel classification.

- We consider many target variables simultaneously.
- Prediction: predict **vector y** for a new instance **x** (2^K possible values).

Example: (image annotation)

- Y_1: grass (presence (1) or absence (0)),
- Y_2: snow (presence (1) or absence (0)),
- Y_3: rocks (presence (1) or absence (0)),
- Y_4: sky (presence (1) or absence (0)).

Other examples: Target variables $Y_1, \dots, Y_K$ can refer to:

- Text categorization: different topics (policy, war, Wladimir Putin, research, biology).
- Ecology: presence or absence of species in the ecosystem.
- Medicine: presence or absence of diseases.

Some remarks:

- It is worthwile to take into account dependences between targets.
- In real data examples: p, n, K can be large.

Hamming loss:

- $$L_H(\mathbf{y}, h(\mathbf{x})) := \frac{1}{K} \sum_{k=1}^{K} 1[y_k \neq h_k(\mathbf{x})]. \tag{1}$$
- Fraction of labels whose relevance is incorrectly predicted.
- Risk (1) minimizer is obtained by:

$$h_H^*(\mathbf{x}) = (h_{H_1}(\mathbf{x}), \ldots, h_{H_K}(\mathbf{x})),$$

(marginal mode) where:

$$h_{H_k}^*(\mathbf{x}) = \arg \max_{y \in \{0,1\}} P(Y_k = y|\mathbf{x}).$$

Subset 0/1 loss:

- $$L_s(\mathbf{y}, h(\mathbf{x})) := 1[\mathbf{y} = h(\mathbf{x})]. \tag{2}$$
- It generalizes the well-known 0/1 loss from the conventional to the multi-label setting.
- Risk (2) minimizer is obtained by:

$$h_s^*(\mathbf{x}) = \arg \max_{\mathbf{y} \in \{0,1\}^K} P(\mathbf{y}|\mathbf{x})$$

(mode of the joint distribution).

Binary relevance (BR)

- Train a separate binary classifier $h_k(\cdot)$ for each label $k = 1, \ldots, K$ using e.g. logistic regression or decission tree.
- Learning is performed independently for each label, ignoring all other labels.
- Well-tailored for Hamming loss minimization.
- Not suitable for 0/1 subset loss.

Methods optimal w.r.t. subset 0/1 loss

We describe some methods that are optimal w.r.t. subset 0/1 loss.

- Label powerset (LP),
- Classifier chains (CC),
- Ising Model.

The last two methods seek to estimate the joint distribution $P(Y_1, \ldots, Y_K|x)$:

Label Powerset (LP)

- Proposed by Tsoumakas and Katakis (2007).
- **This approach reduces the MLC problem to multi-class classification, considering each label subset as a distinct meta-class.**
- The number of these meta-classes may become large (2^K).
- Since prediction of the most probable meta-class is equivalent to prediction of the mode of the joint label distribution, LP is tailored for the subset 0/1 loss.
- Main drawback: large number of classes produced by this reduction and very few training examples for each class.

Label Powerset (LP)

X_1	Y_1	Y_2
1	0	0
2	0	0
3	1	0
4	1	0
5	1	1

Tabela : Before reducion.

X_1	Y
1	1
2	1
3	2
4	2
5	3

Tabela : After reducion.

Classifier chains (CC)

- Proposed by Dembczynski et al (2010).
- **Product rule of probability:**

$$P(Y_1,\ldots,Y_K|\mathbf{x}) = \prod_{k=1}^{K} P(Y_k|Y_1,\ldots,Y_{k-1},\mathbf{x}). \qquad (3)$$

- To estimate (3) we build binary classification models in which:
  - Y_k is class variable,
  - $\mathbf{x}, Y_1,\ldots,Y_{k-1}$ are attributes,

  $k = 1,\ldots,K$.

Classifier chains (CC)

Prediction:

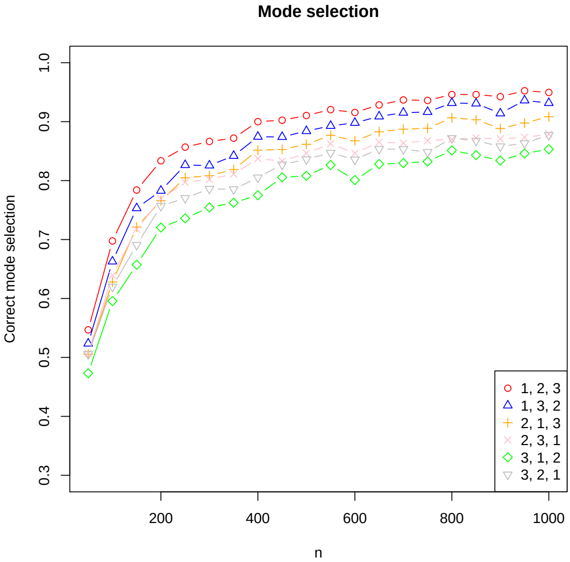
- Exhaustive search requires 2^K operations.
- Greedy search:
  - Find: $y_1^* = \arg\max_{y\in\{0,1\}} P(Y_1 = y|\mathbf{x})$.
  - Find: $y_2^* = \arg\max_{y\in\{0,1\}} P(Y_2 = y|y_1^*, \mathbf{x})$.
  - Find: $y_3^* = \arg\max_{y\in\{0,1\}} P(Y_3 = y|y_1^*, y_2^*, \mathbf{x})$.
  - ...

  requires K operations.
- Other possibilities: beam search (Kumar et al. 2013).

Classifier chains (CC)

- Theoretically, the result of the product rule does not depend on the order of the variables. Practically, however, two different classifier chains will produce different results.
- Example:
  - We generate data from CC.
  - Ordering in data generation: Y_1, Y_2, Y_3.
  - Paramters: $\beta_k = (0.3, \ldots, 0.3)'$, $\alpha_k = (0.5, \ldots, 0.5)'$.
  - Test set: 50 observations, number of simulations: 50.

Classifier chains (CC)



Ensembles of classifier chains (ECC)

- Proposed by Read (2009).
- Average the multi-label predictions of CC over a (randomly chosen) set of permutations.

Ising Model

- Ising model: Ising model with covariates:

$$P(y_1,\ldots,y_K|\mathbf{x}) = \frac{1}{Z(\theta(\mathbf{x}))}\exp\left[\sum_j \theta_{jj}(\mathbf{x})y_j + \sum_{j<k}\theta_{jk}(\mathbf{x})y_j y_k\right].$$

- the natural choice: $\theta_{ij}(\mathbf{x}) = \theta_{ij0} + \theta'_{ij}\mathbf{x}$.
- $Z(\theta)$ is the normalization function ensuring that 2^K probabilities sum up to 1.
- We assume: $\theta_{jk} = \theta_{kj}$.
- There are $K(K+1)/2$ parameters.

Ising Model with covariates

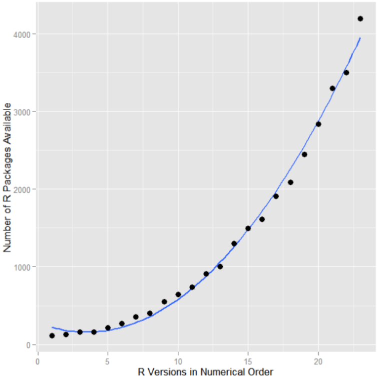
- Ising model is associated with LOGISTIC REGRESSION:

$$\log\left[\frac{P(y_j=1|\mathbf{y}_{-j},\mathbf{x})}{P(y_j=0|\mathbf{y}_{-j},\mathbf{x})}\right]=\theta_{jj0}+\theta'_{jj}\mathbf{x}+\sum_{k:k\neq j}[\theta_{jk0}+\theta'_{jk}\mathbf{x}]y_k, \quad (4)$$

  where $\mathbf{y}_{-j}=(y_1,\ldots,y_{j-1},y_{j+1},\ldots,y_K)$.

- Ising Model is an example of Markov Network:
  - $\theta_{ij}=0$: no edge between y_i and y_j.
  - $\theta_{ij}=0 \iff y_i \perp\!\!\!\perp y_j|\mathbf{y}_{-\{i,j\}},\mathbf{x}$.

- MULAN- java library for multilabel learning, `http://mulan.sourceforge.net/`
- Multilabel learning in R...???



source: r4stats.com

Emotions dataset ($n = 593$, $p = 72$, $K = 6$):

Method	1-01	1-Hamming	Recall	Precision
BR	0.21	0.78	0.47	0.53
PCC	0.28	0.77	0.64	0.62
ISING	0.21	0.70	0.51	0.53

Flags dataset ($n = 194$, $p = 19$, $K = 7$):

Method	1-01	1-Hamming	Recall	Precision
BR	0.06	0.68	0.71	0.65
PCC	0.14	0.66	0.64	0.62
ISING	0.14	0.69	0.67	0.68

Mediamil dataset top 10 ($n = 978,\ p = 1449,\ K = 45$):

Method	1-01	1-Hamming	Recall	Precision
BR	0.17	0.82	0.59	0.74
PCC	0.21	0.81	0.59	0.69
ISING	0.20	0.81	0.54	0.69

Scene dataset ($n = 2407,\ p = 294,\ K = 6$):

Method	1-01	1-Hamming	Recall	Precision
BR	0.43	0.88	0.49	0.49
PCC	0.65	0.89	0.69	0.71
ISING	0.52	0.85	0.55	0.58

Yeast dataset top 10 ($n = 2417$, $p = 103$, $K = 14$):

Method	1-01	1-Hamming	Recall	Precision
BR	0.06	0.60	0.63	0.50
PCC	0.20	0.72	0.66	0.64
ISING	0.06	0.51	0.59	0.43

Dziękuje za uwagę!!!

Literature

- Dembczynski, et. al. (2010), On label dependence and loss minimization in multi-label classification, Machine Learning.
- Tsoumakas, Katakis, (2007), Multi-label classification: An overview, International Journal of Data Warehousing and Mining.