Computational Biology Group Seminar

Introduction to multi-label classification

Paweł Teisseyre

Institute of Computer Science, Polish Academy of Sciences
http://www.ipipan.waw.pl/staff/p.teisseyre/

Outline

- Single-label classification vs multi-label classification.
- Standard methods used in multi-label classification.
- Evaluation measures used in multi-label classification.
- Software and datasets.

Single-label classification

x_1	x_2	$\ldots$	x_p	y
1.0	2.2	$\ldots$	4.2	1
2.4	1.3	$\ldots$	3.1	1
0.9	1.4	$\ldots$	3.2	0
$\vdots$			$\vdots$	$\vdots$
1.7	3.5	$\ldots$	4.2	0
3.9	2.5	$\ldots$	4.1	?

Table: Single-label classification.

- $y \in \{0, 1\}$- target variable (label).
- $\mathbf{x} = (x_1, \ldots, x_p)^T$- vector of explanatory variables (features).

TASK: build a model which predicts y using $\mathbf{x}$.

Multi-label classification.

x_1	x_2	$\ldots$	x_p	y_1	y_2	$\ldots$	y_K
1.0	2.2	$\ldots$	4.2	1	0	$\ldots$	1
2.4	1.3	$\ldots$	3.1	1	0	$\ldots$	1
0.9	1.4	$\ldots$	3.2	0	0	$\ldots$	1
$\vdots$			$\vdots$	$\vdots$			$\vdots$
1.7	3.5	$\ldots$	4.2	0	1	$\ldots$	0
3.9	2.5	$\ldots$	4.1	?	?	$\ldots$	?

Table: Multi-label classification

- $\mathbf{y} = (y_1, \ldots, y_K)^T$- vector of target variables (labels).
- $\mathbf{x} = (x_1, \ldots, x_p)^T$- vector of explanatory variables (features).

TASK: build a model which predicts $\mathbf{y}$ using $\mathbf{x}$.

Multimorbidity (co-occurrence of two or more diseases)

	BMI	Weight	Glucose	Sex		Diabetes	Hypothension	Liver disease		Obesity
Id	**X1**	**X2**	**X3**	**X4**	**...**	**Y1**	**Y2**	**Y3**	**...**	**Y50**
1	31	94	10	M	...	1	0	0	...	1
2	26	63	6	F	...	1	0	0	...	1
3	27	60	7	M	...	0	0	0	...	0
...	...	...	...	...	...	...	...	...	...	...
1000	25	55	5	M	...	0	0	0	...	0

Task: predict occurrences of diseases based on patients' characteristics (e.g. BMI, Sex, etc.).

Text categorization

Text: *"The purpose of this study is to determine the best prediction of heart failure outcomes, resulting from two methods – standard epidemiologic analysis with logistic regression and knowledge discovery with supervised learning/data mining. Heart failure was chosen for this study as it exhibits higher prevalence and cost of treatment than most other hospitalized diseases. The prevalence of heart failure has exceeded 4 million cases in the U.S.. Findings of this study should be useful for the design of quality improvement initiatives, as particular aspects of patient comorbidity and treatment are found to be associated with mortality. This is also a proof of concept study, considering the feasibility of emerging health informatics methods of data mining in conjunction with or in lieu of traditional logistic regression methods of prediction. Findings may also support the design of decision support systems and quality improvement programming for other diseases."[Authors: Phillips K.T., Street W.N.]*

TASK: Assign tags (e.g. "statistics","medicine","computer science","psychology",etc.) to the above text.

Image annotation

Single-label classification: is the mountain in this picture?

$y = 1(yes)$.

Image annotation

Multi-label classification: what objects are in the picture?

$$\mathbf{y} = (sky, tree, snow, car)^T = (1, 1, 1, 0)^T.$$

Single-label classification

Standard approach:

1. **Training:** Estimate posterior probability:

$$p(y = 1|\mathbf{x}).$$

2. **Prediction:** Make prediction for some new observation $\mathbf{x}_0$:

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

where $\hat{p}(\mathbf{y}|\mathbf{x}_0)$ is estimated probability.

/We find a mode of the posterior distribution/

Only the first step is challenging.

Single-label classification

Logistic regression - training:

- Target variable: $y = 1$ (presence of heart disease), $y = 0$ (absence of disease)
- Feature: *Age*.
- It is assumed that:

$$p(y = 1|Age) = \frac{\exp(\beta_0 + \beta_1 \cdot Age)}{1 + exp(\beta_0 + \beta_1 \cdot Age)}.$$

- Parameters β_0, β_1 are estimated using training data.

Single-label classification

Logistic regression - Prediction:

- Estimated parameters: $\hat{\beta}_0 = -3.521$, $\hat{\beta}_1 = 0.064$.
- Q: What is the estimated probability of occurrence of heart disease for a patient in age 75?
- A:

$$P(y = 1|Age = 75) = \frac{\exp(-3.521 + 0.0641 \cdot 75)}{1 + \exp(-3.521 + 0.064 \cdot 55)} \approx 0.78.$$

- Prediction $\hat{y} = 1$ as $P(y = 1|Age = 75) > P(y = 0|Age = 75)$.

Multi-label classification

Standard approach:

1. **Training:** Estimate posterior probability:

$$p(\mathbf{y}|\mathbf{x}) = p(y_1, y_2, \ldots, y_K|\mathbf{x}).$$

2. **Prediction:** Make Prediction for some new observation $\mathbf{x}_0$:

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

where $\hat{p}(\mathbf{y}|\mathbf{x}_0)$ is estimated probability.

/We find a mode of the multivariate posterior distribution/

Both steps are challenging.

Standard methods used in multi-label classification

1. Binary Relevance (BR).
2. Classifier chains (CC).
3. Label Powerset (LP).
4. Ising model (IS).

Binary relevance (BR)

- Build K independent single-label classifiers:

$$y_1 \sim x_1, \ldots, x_p,$$
$$y_2 \sim x_1, \ldots, x_p,$$
$$\ldots$$
$$y_K \sim x_1, \ldots, x_p.$$

- This allows to estimate posterior probabilities $p(y_1 = 1|\mathbf{x}), \ldots, p(y_K = 1|\mathbf{x})$.
- Prediction is performed independently for each $k = 1, \ldots, K$:

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

Binary relevance (BR)

- The method allows to estimate the multivariate posterior probability if the labels are conditionally independent: $p(y_1, \ldots, y_k|\mathbf{x}) = \prod_{k=1}^{K} p(y_k = 1|\mathbf{x})$.

Pros:

- Training is straightforward.
- We can use any single-label classifier (e.g. logistic regression or decision tree).
- Prediction is straightforward.

Cons:

- We do not take into account possible dependencies between labels!

Classifier chains (CC)

- The method utilizes the product rule of probability:

$$p(y_1, \ldots, y_K|\mathbf{x}) = p(y_1|\mathbf{x}) \prod_{k=2}^{K} p(y_k|y_1, \ldots, y_{k-1}, \mathbf{x}).$$

- To estimate the conditional probabilities we build K single-label classifiers:

$$y_1 \sim x_1, \ldots, x_p,$$
$$y_2 \sim x_1, \ldots, x_p, y_1,$$
$$y_3 \sim x_1, \ldots, x_p, y_1, y_2,$$
$$\ldots$$
$$y_K \sim x_1, \ldots, x_p, y_1, \ldots, y_{K-1}.$$

Classifier chains

Pros:

- Training is straightforward.
- We can use any single-label classifier (e.g. logistic regression or decision tree).
- Prediction is easy if the greedy search is applied.

Cons:

- Order of fitting the models may influence the results.

  /Solution: Ensemble of classifier chains=family of models built for different orderings of labels/

Label Powerset (LP)

- This approach reduces the mutli-label problem to single-label, multi-class problem, considering each labelset as a distinct meta-class.
- The number of labelsets may become large (2^K).
- Prediction of the most probable labelset is equivalent to prediction of the mode of the joint posterior distribution.

Label Powerset (LP)

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

Table: Before reduction.

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

Table: After reduction.

Label Powerset (LP)

Pros:

- Effective and simple approach.
- It takes into account dependencies between labels.

Cons:

- Large number of classes produced by this reduction.
- Very few training examples for each class.
- It cannot predict unseen labelsets that may also lead to a tendency to overfit the training data.

Ising Model (IS)

Ising model [1]

$$p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}) = \frac{1}{N(\mathbf{x})} \exp\left[\sum_{k=1}^{K} a_k^T \mathbf{x} y_k + \sum_{k<l} \beta_{k,l} y_k y_l\right], \qquad (1)$$

where:

$$N(\mathbf{x}) = \sum_{\mathbf{y}\in\{0,1\}^K} \exp\left[\sum_{k=1}^{K} a_k^T \mathbf{x} y_k + \sum_{k<l} \beta_{k,l} y_k y_l\right] \qquad (2)$$

and $\boldsymbol{\theta} = (a_1, \ldots, a_K, \beta_{1,2}, \beta_{1,3}, \ldots, \beta_{K-1,K})^T$.

[1] E. Ising, Beitrag zur Theorie des Ferromagnetismus, Zeitschrift für Physik, **1925**

Ising model (IS)

Pros:

- Natural generalization of logistic regression.
- Easy interpretation of the coefficients.

Cons:

- Large number of parameters.
- Direct estimation using maximum likelihood is difficult.
- Prediction can be difficult for large number of labels (2^K labelsets!).

Standard methods

Popular taxonomy of multi-label methods:

- Problem transformation methods, e.g. BR, CC, LP.

  (transforms the multi-label problem into one or more single-label problems)

- Algorithm adaptation methods, e.g. Ising model, MLKNN.

  (extends a single-label algorithm in order to directly deal with multi-label data)

Evaluation measures

True label vector:

$$\mathbf{y} = (y_1, \ldots, y_K)^T.$$

Predicted label vector:

$$\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_K)^T.$$

We would like to assess the quality of our predictive model using some performance measure $PM(\mathbf{y}, \hat{\mathbf{y}})$.

- PM should be large if the prediction is good.
- PM should be small if the prediction is poor.

Evaluation measures

Example:

$$\mathbf{y} = (1, 1, 1, 0, 0)^T.$$

$$\hat{\mathbf{y}} = (1, 1, 1, 1, 0)^T.$$

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

  $Hamming(\mathbf{y}, \hat{\mathbf{y}}) = \frac{4}{5}$.
- Subset Accuracy (0-1): $Subset(\mathbf{y}, \hat{\mathbf{y}}) = I(\mathbf{y} = \hat{\mathbf{y}})$.

  $Subset(\mathbf{y}, \hat{\mathbf{y}}) = 0$.

Evaluation measures

Example:

$$\mathbf{y} = (1, 1, 1, 0, 0)^T.$$

$$\hat{\mathbf{y}} = (1, 1, 1, 1, 0)^T.$$

- Precision: $Precision(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\#\{k:\hat{y}_k=1, y_k=1\}}{\#\{k:\hat{y}_k=1\}}$.

  $Precision(\mathbf{y}, \hat{\mathbf{y}}) = \frac{3}{4}$.

- Recall: $Recall(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\#\{k:\hat{y}_k=1, y_k=1\}}{\#\{k:y_k=1\}}$.

  $Recall(\mathbf{y}, \hat{\mathbf{y}}) = \frac{3}{3} = 1$.

/Usually the above measures are averaged over all examples in the test data./

Software and datasets

Software:

- MULAN (Java), `http://mulan.sourceforge.net`,
- MEKA (Java), `http://meka.sourceforge.net`,
- mlr (R), `https://mlr-org.github.io/mlr-tutorial/devel/html/index.html`,
- mldr (R), `https://cran.r-project.org/web/packages/mldr/vignettes/mldr.pdf`,
- SciKit Learn (Python), `http://scikit-learn.org/stable/index.html`.

Datasets:

- `http://mulan.sourceforge.net/datasets.html`
- `http://meka.sourceforge.net/#datasets`

References

1. E. Gibaja, S. Ventura, *A tutorial on multilabel learning*, ACM Comput. Surv., 2015.
2. Jesse Read, Tutorial, `https://jmread.github.io/talks/Tutorial-MLC-Porto.pdf`.
3. K. Dembczynski, Tutorial, `http://www.cs.put.poznan.pl/kdembczynski/pdf/phd_studies_2015/mlc.pdf`.
4. K. Dembczynski, et. al., On label dependence and loss minimization in multi-label classification, Machine Learning, 2010.