# Classification of open ended questions with multiple labels

Matthias Schonlau, Ph.D., University of Waterloo, Canada

Hyukjun Gweon, Ph.D., Western University, Canada

Dr. Marika Wenemark, Linköping University, Sweden

# All-that apply questions

- Answers to open-ended questions can be coded with a single label (category),  or with multiple labels (multiple categories)
- An example of an open-ended question intended to have multiple labels is:
  - Happy Data: "Name some positive things in your life, that are uplifting or make you happy: (you may write several things)“
  - “you may write several things” means it is a multi-label answer

# Data : Civil Disobedience Data

- “How important is it that citizens may engage in acts of civil disobedience when they oppose government actions?” (Not at all important 1 --- Very important 7), and then probed respondents’ comprehension with an open-ended question: “What ideas do you associate with the phrase `civil disobedience’? Please give examples.”
- Answers in German
- 768 observations

Behr, D., Braun, M., Kaczmirek, L., & Bandilla, W. (2014). Item comparability in cross-national surveys: results from asking probing questions in cross-national web surveys about attitudes towards civil disobedience. *Quality & Quantity, 48*(1), 127-148.

# What ideas do you associate with the phrase `civil disobedience'?

Some answers:

- To do something, that really is against the rules, because you believe the rules are not ok and should not be valid
- Fight against injustices
- Demonstrations
- Don't know
- Demonstrations etc.

Original German answers:

966. Etwas zu tun, das eigentlich gegen die Regeln verstößt, weil man der Meinung ist, die Regeln sind nicht ok und sollten nicht gelten
967. Gegen Ungerechtigkeiten zu kämpfen.
968. Demonstrationen
969. Weiß nicht
970. Demonstrationen etc.

# Immigrant data: some answers

- Question asked: “Which type of immigrants were you thinking of when you answered the question?

- Answers are classified into 14 labels
- Here “Islamic countries“ and “eastern Europe” always co-occur
- i.e. labels are correlated

| | positive | negative | neutral | general | islamic | eastern europe | asia | ex-jugoslavia | eu15 | latin am | subsahara | sintiroma | legalillegall | other |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mostly immigrants from poor countries like Africa, Albania, Russia, Turkey, the police statistics and my personal experiences confirm this unfortunately | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| Albanians, Russians, Turks | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Turks-Russians-Albanians | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Turks,Africans,Asians,Russians, etc. | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

Answers translated from German

# Multi-label Algorithms: BR

- Binary Relevance (BR)
- Predicts  labels independently from one another using whatever learning algorithm
  - Random Forest, Support Vector machines, Gradient Boosting…
- Default solution if you don’t know about multi-label algorithms

- Advantage: Easy
- Disadvantage:
  - Not as good in terms of evaluation measures because it does not utilize the correlation among labels

# Multi –label algorithms

- Binary Relevance (BR )  - Default
- Classifier chains (CC)
- Ensemble Classifier Chains (ECC)
- Label Powerset (LP)
- RAKEL
- …

# Multi-label Algorithms: CC

- CC fits each label sequentially using binary classifiers f()
  - f() is any learning algorithm
- All previously predicted labels are included as x-variables
  - For training, to estimate the function f() the actual predicted values 0 or 1 are included (not the probabilities y_hat)
  - For predicting, the y-values are replaced by their predicted values (probabilities)
  - To classify the predicted y's, a threshold of 0.5 is used

Prediction

$$y_1 = f(x_1, \ldots, x_p)$$

$$y_2 = f(x_1, \ldots, x_p, \hat{y}_1)$$

$$y_3 = f(x_1, \ldots, x_p, \hat{y}_1, \hat{y}_2)$$

$$\vdots$$

$$y_L = f(x_1, \ldots, x_p, \hat{y}_1, \ldots, \hat{y}_{L-1})$$

where f() and y_hat were estimated during training

# Multi-label Algorithms: CC

- Example with 3 y-variables: y1, y2, y3
  - Choose f()=SVM
- Put in random order, e.g. : y2, y3, y1
- Training:

$$y_2 = f_{SVM}(x_1, \ldots, x_p)$$

$$y_3 = f_{SVM}(x_1, \ldots, x_p, y_2)$$

$$y_1 = f_{SVM}(x_1, \ldots, x_p, y_2, y_3)$$

- Prediction:

$$y_2 = f_{SVM}(x_1, \ldots, x_p)$$

$$y_3 = f_{SVM}(x_1, \ldots, x_p, \hat{y}_2)$$

$$y_1 = f_{SVM}(x_1, \ldots, x_p, \hat{y}_2, \hat{y}_3)$$

# Multi-label Algorithms: CC

- Including previously predicted labels as covariates takes into account correlation among labels
- Unfortunately, the order in which the labels are predicted affects the performance
- Therefore order of the y’s is chosen at random

# Ensemble Chain classifier: ECC

- ECC to reduce the dependence of CC on the order of labels developed
- Use multiple CC:
  - Each CC has a random label order
  - Each CC is conducted on a bootstrap sample of size N
    - To reduce the computational burden for large data sets, the authors use a regular (not bootstrap) sample of a smaller size.
- How many CC?
  - The authors proposed 50 CC's for small data sets and 10 CC's for large data sets.

Read J, Pfahringer B, Holmes G, Frank E (2011) Classifier chains for multi-label classification. Machine Learning 85(3): 333-359

# Ensemble Chain classifier: ECC

- Calculate the average prediction w
- Example:
  - single observation
  - 10 rounds for the ensemble
  - 4 labels
  - W is the average prediction

| Round | Y1 | Y2 | Y3 | Y4 |
|---|---|---|---|---|
| 1 | 1 | 0 | 1 | 0 |
| 2 | 1 | 0 | 1 | 0 |
| 3 | 0 | 0 | 1 | 0 |
| 4 | 0 | 0 | 1 | 0 |
| 5 | 1 | 0 | 1 | 0 |
| 6 | 0 | 1 | 0 | 0 |
| 7 | 1 | 0 | 1 | 1 |
| 8 | 0 | 1 | 0 | 1 |
| 9 | 0 | 0 | 1 | 0 |
| 10 | 0 | 0 | 1 | 0 |
| w | 0.4 | 0.2 | 0.8 | 0.2 |

# Ensemble Chain classifier: ECC

- For prediction, the same threshold is used for all L y-variables

$$\hat{y}_j = \begin{cases} 1 & \text{if } w_j \geq t \\ 0 & \text{otherwise} \end{cases}$$

- Using a threshold of t=0.5 corresponds to majority vote.
- However, majority vote (t=0.5) leads to inferior results.

# Ensemble Chain classifier: ECC

- Set the threshold to match the observed number of labels in the training data:

$$t = \arg\min_{t} \left| LCARD(D) - \left( \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \sum_{j=1}^{L} I(w_j \geq \mathrm{t}) \right) \right|$$

  - Where LCARD(D) is the average number of labels per observation in the training data
  - *I* is the indicator function
  - $w_j$ is the fraction label j has been predicted in the test data
    - E.g. If in an ensemble with 10 chains Label j is predicted 3 times, then $w_j$=0.3

# Label Powerset learning (LP)

- LP transforms a multi-label classification into a multiclass classification
- Each unique label set is treated as a distinct label.
- Example:
  - Movie Labels: romantic, comedy, award-winning, science-fiction, violent
- Each combination observed in the training data becomes a label.

# Label Powerset learning (LP)

- Example Movie Labels: romantic, comedy, award-winning
- Normal coding of training data

Training data

| | romantic | comedy | award winning |
|---|---|---|---|
| romantic comedy | 1 | 1 | 0 |
| comedy | 0 | 1 | 0 |
| award winning | 0 | 0 | 1 |
| ramantic comedy | 1 | 1 | 0 |
| award winning romantic comedy | 1 | 1 | 1 |
| award winning | 0 | 0 | 1 |
| comedy | 0 | 1 | 0 |
| award winning romantic comedy | 1 | 1 | 1 |
| award winning comedy | 0 | 1 | 1 |
| comedy | 0 | 1 | 0 |

# Label Powerset learning (LP)

- LP coding of training data
- The three labels have 2^3=8 possible combinations
- 5 combinations are observed in the training data
- Each of them becomes a label (y1..y5)

Training data

| | comedy | award winning | romantic comedy | award winning romantic comedy | award winning comedy |
|---|---|---|---|---|---|
| | y1 | y2 | y3 | y4 | y5 |
| romantic comedy | 0 | 0 | 1 | 0 | 0 |
| comedy | 1 | 0 | 0 | 0 | 0 |
| award winning | 0 | 1 | 0 | 0 | 0 |
| ramantic comedy | 0 | 0 | 1 | 0 | 0 |
| award winning romantic comedy | 0 | 0 | 0 | 1 | 0 |
| award winning | 0 | 1 | 0 | 0 | 0 |
| comedy | 1 | 0 | 0 | 0 | 0 |
| award winning romantic comedy | 0 | 0 | 0 | 1 | 0 |
| award winning comedy | 0 | 0 | 0 | 0 | 1 |
| comedy | 1 | 0 | 0 | 0 | 0 |

# Label Powerset learning (LP)

Disadvantages:

- Labelsets not seen in the training data cannot be predicted
- The number of classes increases exponentially as a function of L
  - $2^L$ Possible combinations of labels

Advantages:

- Correlations among labels are accounted for by excluding combinations not seen in the training data

Tsoumakas G, Katakis I (2007) Multi-label classification: An overview. International Journal of Data Warehousing and Mining 3:1-13.

# RAkEL: Applying LP to label subsets

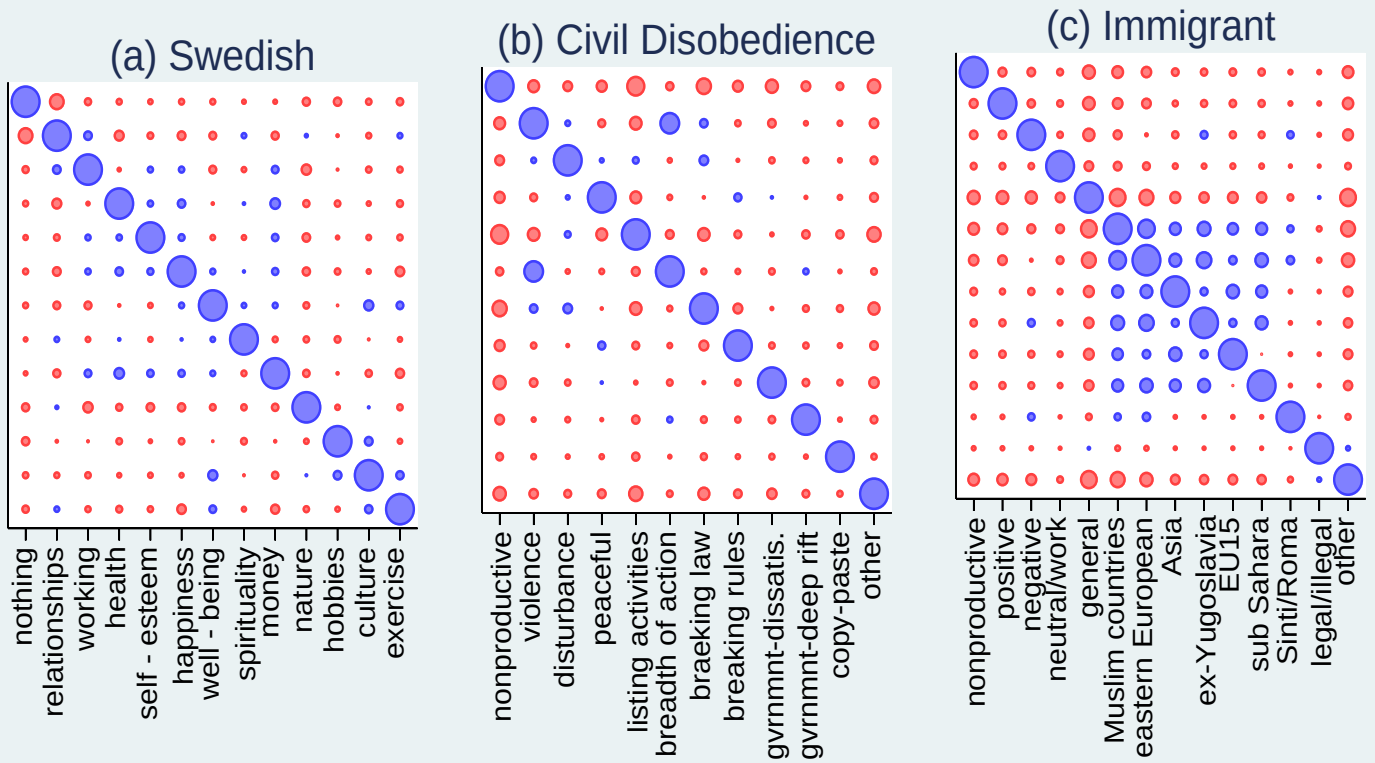
- Break the L labels into random subsets of approximately equal size k
- Use the LP method of constructing labels for each subset
- Construct m models for each subset separately
- Predict by majority vote each subset and combine the result

- Advantage:
  - the number of possible labels does not grow exponentially as in the LP method.
  - Takes into account correlations within a label subset
- Disadvantage: cannot predict new label combinations within a subset

Tsoumakas G, Vlahavas I (2007) Random k-labelsets: An ensemble method for multilabel classification. In: Procedures of the 18th European Conference on Machine Learning (ECML 2007), Springer, Berlin, pp 406-417.

# Multi-label

- Plot of bivariate label correlations of the three data sets.
  - Red : negative correlations
  - Blue: positive correlations
- Each number on the x/y-axis corresponds to one of the labels.
- Swedish Data have relatively small correlations




Graph produced with “corrplot”. Available at www.schonlau.net/stata

# Multi-label

- Hamming Loss
  - i.e. percentage of labels classified incorrectly

| | BR | ECC | RAKEL |
|---|---|---|---|
| **Happy** | 0.0506 | 0.0538 | 0.0532 |
| **Civil** | 0.0600 | 0.0612 | 0.0618 |
| **Immigrant** | 0.0439 | 0.0420 | 0.0426 |

- Not much difference

# Evaluation under 0/1 loss

- Definition 0/1 Loss:
  - If all labels are predicted correctly, loss=0
  - If one or more labels are predicted incorrectly, loss=1

| | BR | ECC | RAKEL |
|---|---|---|---|
| Happy | 0.4430 | 0.4491 | 0.4529 |
| Civil | 0.5238 | 0.4646 | 0.4776 |
| Immigrant | 0.4741 | 0.3575 | 0.3850 |

- In absolute terms the error is large; 0/1 loss is unforgiving
- For “Civil Disobedience” and “Immigrant”, ECC and RAKEL result is substantially better performance (smaller loss) as compared to BR.

# Discussion

- Weak bivariate label correlations in the Happy data, and stronger bivariate label correlations in the immigrant and civil disobedience data.
- For the data with stronger correlations:
  - 0/1 loss:  both multi-label methods performed substantially better than BR
  - Hamming loss: little effect on accuracy. However, most labels are zero, so average accuracy cannot improve much.
  - (CS audience: Did not evaluate F because this is not widely known outside of CS)

# Conclusion

- automatic classification of open-ended questions that allow multiple answers may benefit from using multi-label algorithms for 0/1 loss
- The degree of correlations among the labels may be a useful prognostic tool

ECC and RAkEL are implemented in Stata.
Contact me for details.

# THE END

Contact info:
schonlau@uwaterloo.ca
www.schonlau.net

I gratefully acknowledge funding from the Social Sciences and Humanities Research Council (SSHRC) of Canada