Automatic Coding of Open-ended Questions: Should You Double Code the Training Data?

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

This talk is based on joint work with my Ph.D. student Zhoushanyue (Sophie) He

Introduction

- Traditionally, text data collected from open-ended questions in surveys have been manually classified at great expense
- We consider automatic coding
- For automatic coding,
  - Train a statistical learning model on a small proportion of texts (manually coded)
  - Code the remainder automatically using the trained model

Turn text into n-gram variables

- Here: indicator variables of word occurrence

```
// Stata code
clear
input strL text
"Arie is Dutch!"
"Arie likes surveys"
"This survey is Dutch"
"Dutch people like Sinterklaas."
"Sinterklaas never took a survey"
end;

ngram text , threshold(2) stem prefix(_)
list
```

	text	_ari	_dutch	_like	_sinte~a	_survei	n_token
1.	Arie is Dutch!	1	1	0	0	0	3
2.	Arie likes surveys	1	0	1	0	1	3
3.	This survey is Dutch	0	1	0	0	1	4
4.	Dutch people like Sinterklaas.	0	1	1	1	0	4
5.	Sinterklaas never took a survey	0	0	0	1	1	5

Training data is subject to classification error

- The quality of automatic coding depends on the quality of the manually classified training data
- Human coders can make mistakes
  - human error
  - ambiguous texts are difficult to code
- To learn about the degree of coding disagreement a common practice is to double code texts

Goal

- If you already have double coded data, should you use it in statistical learning?
  - If so, how?
- If you have a fixed budget, should you single code or double code half the training data?

Strategies for addressing intercoder differences for statistical learning

- Replicate: Replicate double-coded text observations in the training data, once for each coding, regardless of whether the codes are the same or not
- Remove differences: Remove text observations coded differently by the two coders from the training data
- Majority vote: Ask a third coder to code only text observations coded differently by the two coders. The code is determined by majority vote
- Expert resolves: Ask an expert to code text observations coded differently by the two coders. Labels are determined by the expert.
  - Assume the expert is always correct (for now)

Example: training data by double coding strategy

				training data			
	Coder 1	**Coder 2**		**Rem. Diff.**	**Majority vote**	**Expert resolves**	**Re-plicate**
text1	C	C		C	C	C	C
text2	B	C			3rd coder: C	expert: D	B
text3	A	A		A	A	A	A
text1							C
text2							C
text3							A

Training data by double coding strategy

based on 3 observations manually coded by two coders into codes A,B,C,D

Cost of strategies varies

- Assume that the size of the training data is constant
- The cost of the strategies is not equal:
  - The cost of double coding is twice that of single coding
  - The cost of "remove differences" is twice that of single coding
  - The cost of "majority vote" is 2-3 times that of single coding
  - The cost of "expert resolves" is much higher assuming an expert costs much more than a regular coder

- Rather than keeping the size of the training data constant, we keep the cost constant

Fixed cost: size of training data varies

- For a fixed budget the size of the training data varies
- Single coding can afford to code N observations for the training data
  - For double coding, can afford N/2 observations
  - For "remove differences" can afford N/2 observations
  - For "majority vote" can afford less than N/2 observations
  - For "expert resolves" can afford much less than half the training data

Strategy	Number of texts coded under fixed budget
Single coding	N
Replicate	N/2
Remove difference	N/2
Majority vote	$\frac{N}{2+2p-p^2L/(L-1)}$
Expert resolves	$\frac{N}{2+2tp-tp^2L/(L-1)}$

where

- L is the number of outcome classes
- p is the probability of a coding error by the coder
- t is the cost of an expert relative to a regular coder (e.g. t=10 times as expensive)

Research Questions

1) Assuming a fixed budget, which of the 5 strategies maximizes prediction accuracy for statistical learning?

2) When the data are already double coded, which strategy maximizes prediction accuracy for statistical learning?

- Motivation: Sometimes the data may already be double coded; and the size of the training data is constant (unaffected by cost)

Experiment

- We simulated the 5 strategies on data sets (not all shown here)
- We assumed the codes available were true
- We simulated an error rate from p=0 to p=0.5

Coding Errors

- How to simulate coding errors?
- For 2 outcome classes, there is only a single incorrect code
  - When a coding error occurs, it is clear which code was chosen
- For multiple outcome classes, need to specify the coding error matrix for simulation
  - E.g. M1 : equal probability of misclassification

$$M_1 = \begin{pmatrix} 1-p & p/(L-1) & \dots & p/(L-1) \\ p/(L-1) & 1-p & \dots & p/(L-1) \\ \vdots & \vdots & \vdots & \vdots \\ p/(L-1) & p/(L-1) & \dots & 1-p \end{pmatrix}$$

where p is the probability of a coding error

Coding Errors for ordered outcomes

For ordered outcomes:

- M2: Misclassify into neighboring classes
- M3: Misclassify into higher classes
  - E.g. an optimistic coder may consider responses to be in more “optimistic" classes.
  - This has extra parameters for flexibility

$$M_2 = \begin{pmatrix} 1-p & p & 0 & 0 & \dots & 0 & 0 & 0 \\ p/2 & 1-p & p/2 & 0 & \dots & 0 & 0 & 0 \\ 0 & p/2 & 1-p & p/2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & 0 & \dots & 1-p & p/2 & 0 \\ 0 & 0 & 0 & 0 & \dots & p/2 & 1-p & p/2 \\ 0 & 0 & 0 & 0 & \dots & 0 & p & 1-p \end{pmatrix}$$

$$M_3 = \begin{pmatrix} 1-p & p(1-g_1) & \dots & p\prod_{i=1}^{L-3} g_i(1-g_{L-2}) & p\prod_{i=1}^{L-2} g_i \\ 0 & 1-p & \dots & p\prod_{i=1}^{L-4} g_i(1-g_{L-3}) & p\prod_{i=1}^{L-3} g_i \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & p(1-g_1) & pg_1 \\ 0 & 0 & \dots & 1-p & p \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

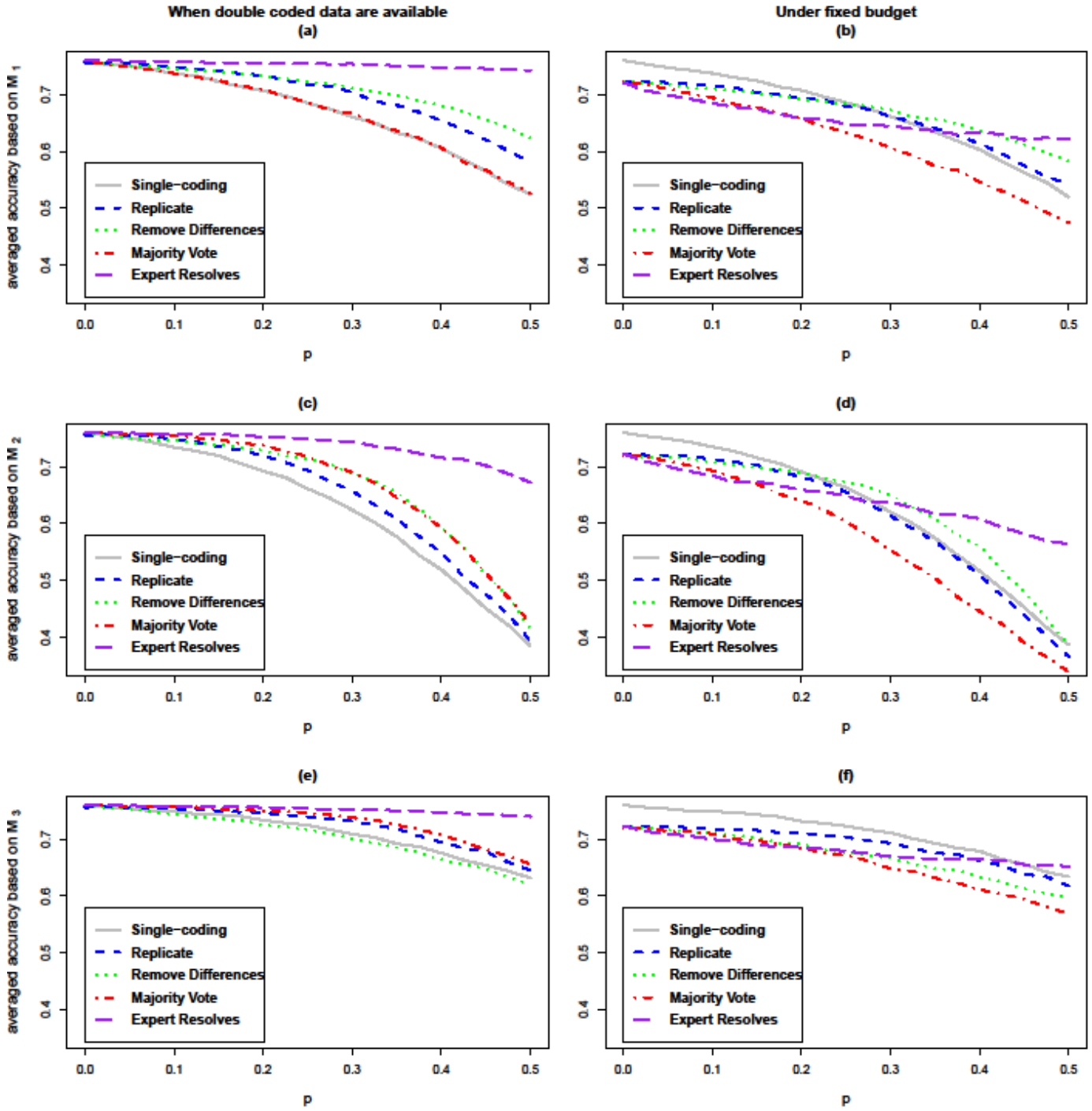
Data: Patient Joe

- 1758 responses to an open-ended question in the LISS panel
- “Joe’s doctor told him that he would need to return in two weeks to find out whether his condition had improved. But when Joe asked the receptionist for an appointment, he was told that it would be over a month before the next available appointment. What should Joe do?”
- Answers classified into one of 4 classes
  - Proactive
  - Somewhat proactive
  - Passive
  - Counterproductive

Simulation: Patient Joe

- Double coded data available:
  - "Expert resolves" leads to highest accuracy
  - 2nd best strategy depends on coding matrix

- Fixed budget:
  - If coding error is low-ish, "single coding" is best



Application to real data

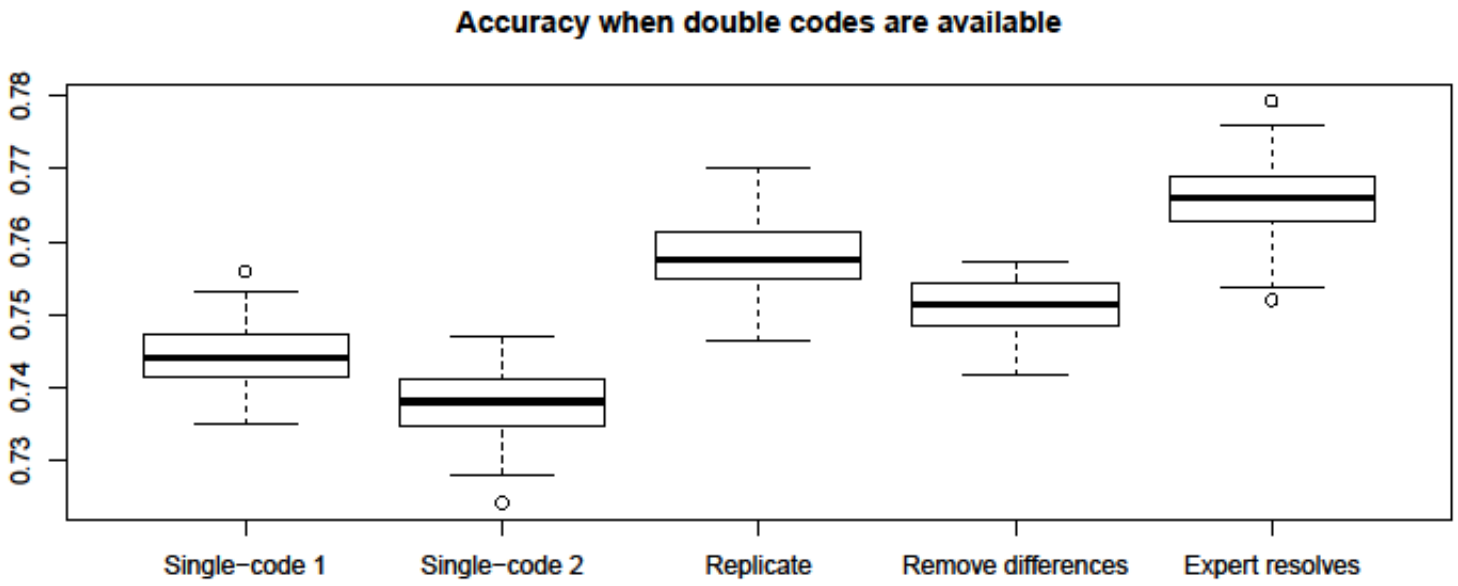
- In a simulation we can inject error to the “gold standard” classification
- Don’t need the actual human classifications, just the gold standard
- Observe dominant strategy as a function of human error rate p

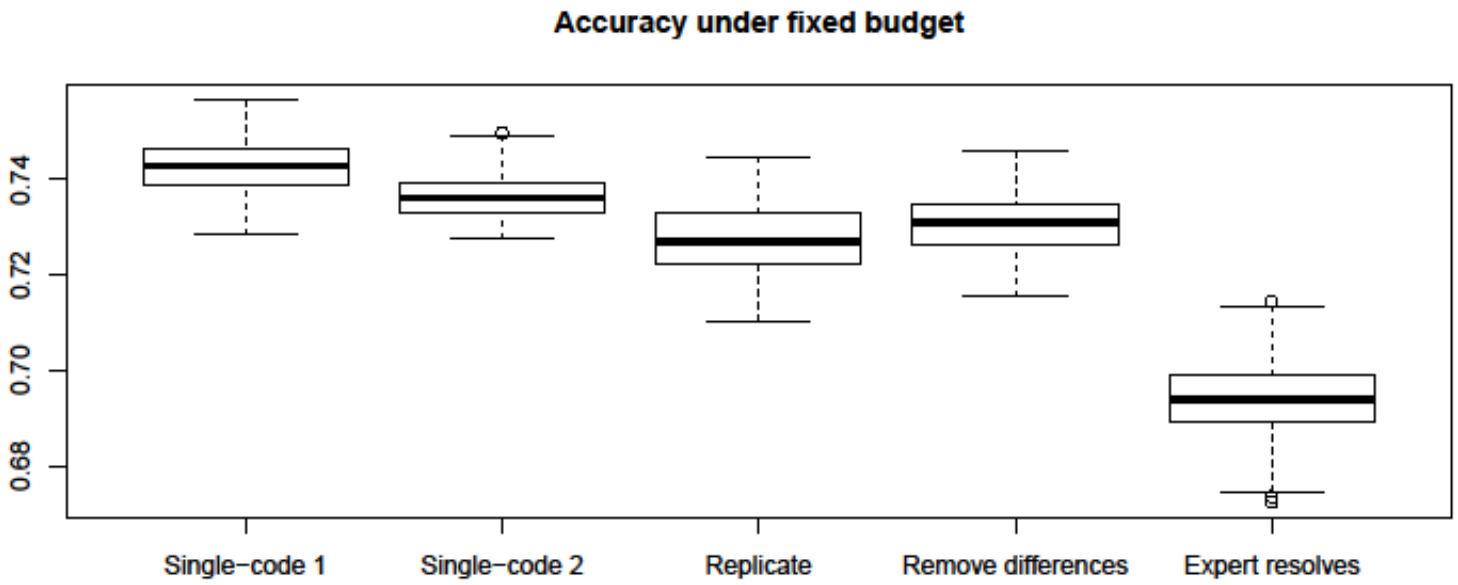
- In real data, we use the human classifications
- The human error rate is whatever it is

Case study: Patient Joe

- Double codes available: Expert resolves best, single coding worst
- Fixed budget: findings are reversed

- Note: Real data: the strategy “majority vote” could not be tested





Recommendation

(for the purpose of statistical learning only)

Fixed budget:

- Use single coding (unless coders' error is large, e.g. >20-35%)
- If coders make more error than 20-35%, this probably should be addressed before thinking about automatic classification

Data are already double coded:

- Never use single coding
- Use experts resolve differences
- If unavailable, for unordered outcomes (M1) use "replicate" or "remove differences"
  - Exception: With 2 outcome classes "remove differences" is not competitive

Contact: schonlau@uwaterloo.ca

References

This talk:

- He, Z, Schonlau, M. Automatic Coding of Text Answers to Open-ended Questions: Should you Double Code the Training Data? Social Science Computer Review. Online first: https://doi.org/10.1177/0894439319846622 . (To appear in print)
- He, Z, Schonlau M. Automatic Coding of Open-ended Questions Into Multiple Classes: Whether and How to Use Double Coded data for prediction. Survey Research Methods. (revision invited)

Open-ended questions:

- Schonlau, M., Couper M. Semi-automated categorization of open-ended questions. Survey Research Methods. Aug 2016, 10(2), 143-152.
- McLauchlan, C, Schonlau, M. Are Final Comments in Web Survey Panels Associated with Next-Wave Attrition? Survey Research Methods, Dec 2016, 10(3), 211-224.
- Gweon, H., Schonlau, M., Kaczmirek L., Blohm, M., Steiner, S. Three Methods for Occupation Coding Based on Statistical Learning. Journal of Official Statistics. 2017, 33 (1), 101-122.
- Schonlau, M, Gweon H, Wenemark, M. Automatic classification of open-ended questions: check-all-that-apply questions, Social Science Computer Review. First published online August 20, 2019. https://doi.org/10.1177/0894439319869210

Software:

- Schonlau, M., Guenther, N. Sucholutsky, I. Text mining using ngram variables. The Stata Journal. Dec 2017, 17(4), 866-881.