Get Another Label? Improving Data Quality And Data Mining Using Multiple, Noisy Labelers

Victor S. Sheng

Foster Provost

Panagiotis G. Ipeirotis

Stern School, New York University

Presented by : Akash Abdu Jyothi

Background

- Obtaining expert labeling is an integral part of KDD (Knowledge Discovery in Databases) preprocessing
- Is it possible to obtain good data values (“labels”) relatively cheaply from multiple noisy sources (“labelers”)?
- Used as training labels for supervised modeling

Repeated Labeling…?

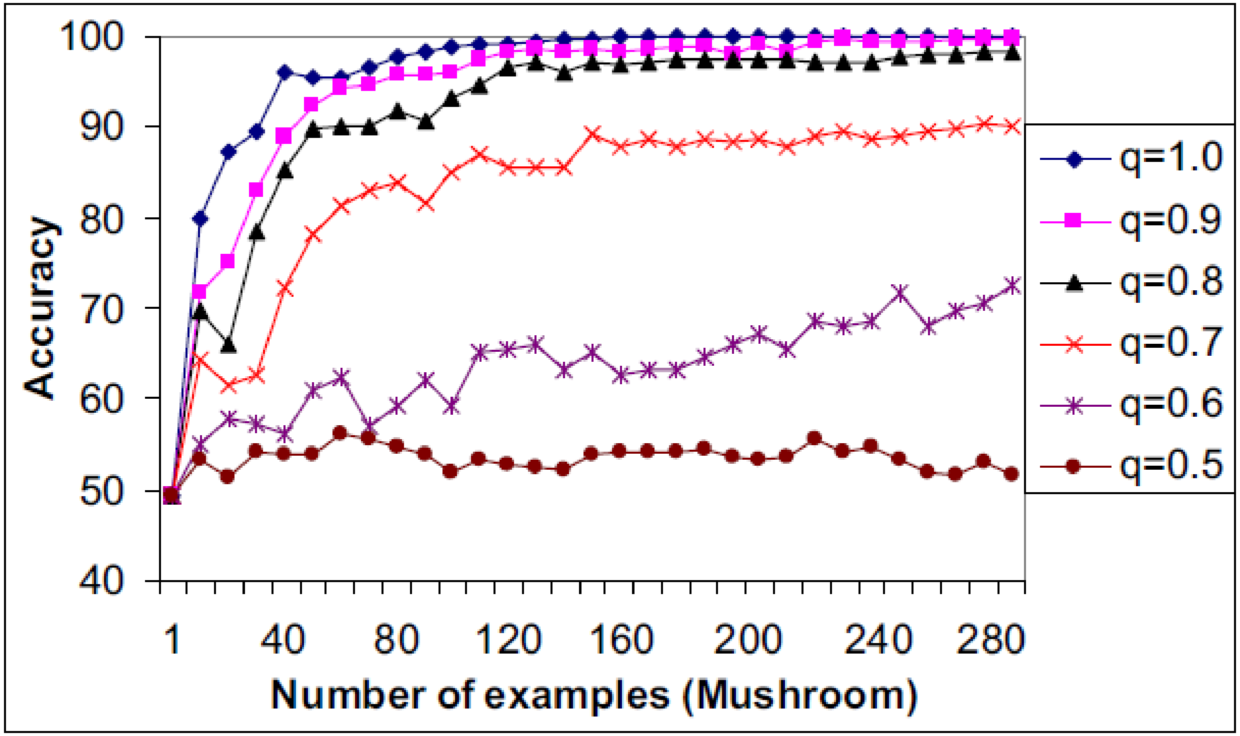
- Labels are imperfect
  - Raghu Ramakrishnan from his SIGKDD Innovation Award Lecture (2008)

    “the best you can expect are noisy labels”
  - Modeling tasks often require high quality labeling

- Outsourcing labeling tasks
  - Quality may be lower than expert labeling
  - But **low costs** can allow massive scale

Effect of Low Quality Labels

Learning curves under different quality levels(q) of training data for classification problem



Outline

- Data quality with repeated labeling
- Model quality with repeated labeling
- Summary and future work

Part 1 – Quality of Repeated Labeling

- **Problem** – supervised induction of a binary classification model

- Training example (x_i,y_i)
  - C_U – cost of procuring unlabeled “feature” portion
  - C_L – cost of labeling x_i with a label y_i

- **Assumptions**
  - C_U and C_L are constant for all examples
  - Labeler quality is constant regardless of the example
    - p_j is the probability that j^{th} labeler gets a label correct

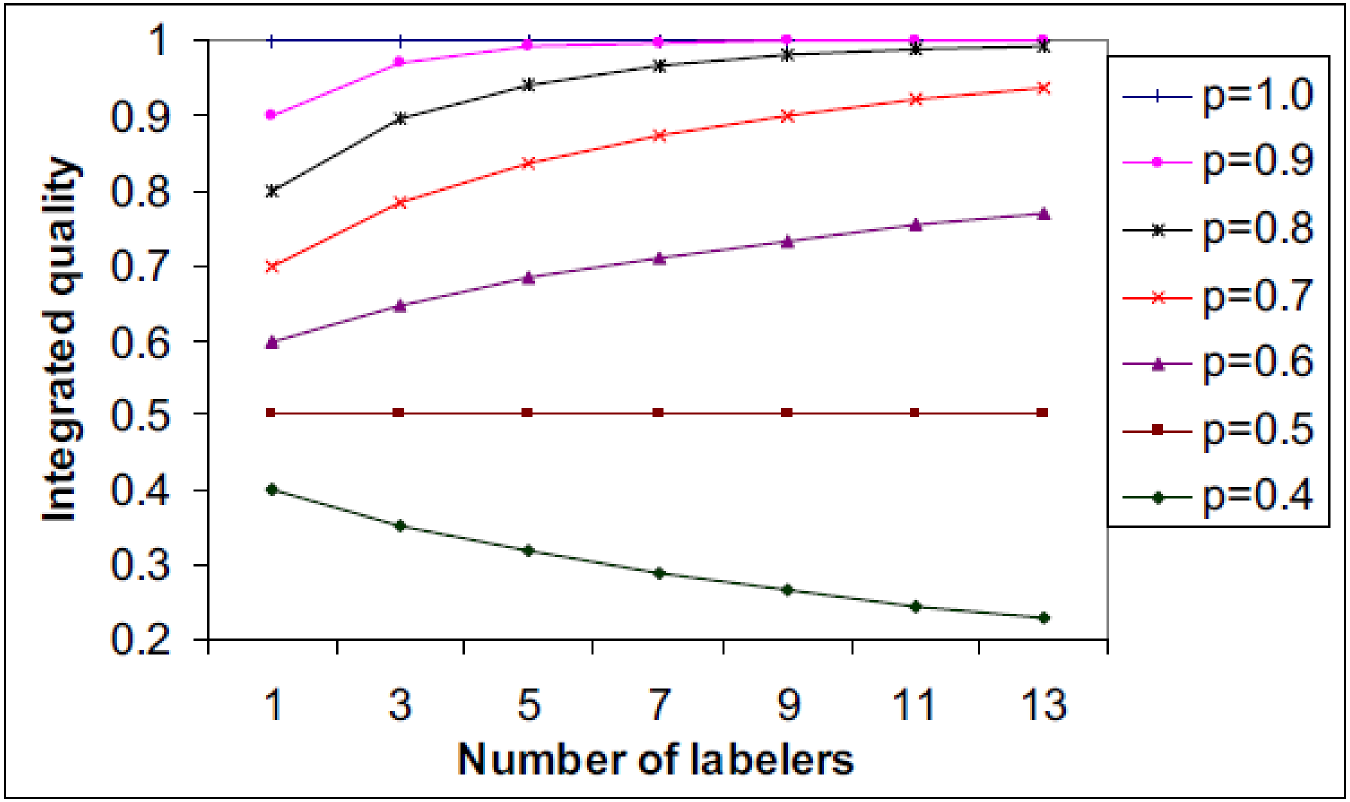
Majority voting – Uniform labeler Quality

- Using 2N+1 labelers of uniform quality i.e. $p_j = p$

- Integrated labeling quality q is the sum of probabilities where we have more correct than wrong answers

$$q = \sum_{i=0}^{N} \binom{2N+1}{i} . p^{2N+1-i} . (1-p)^{i}$$

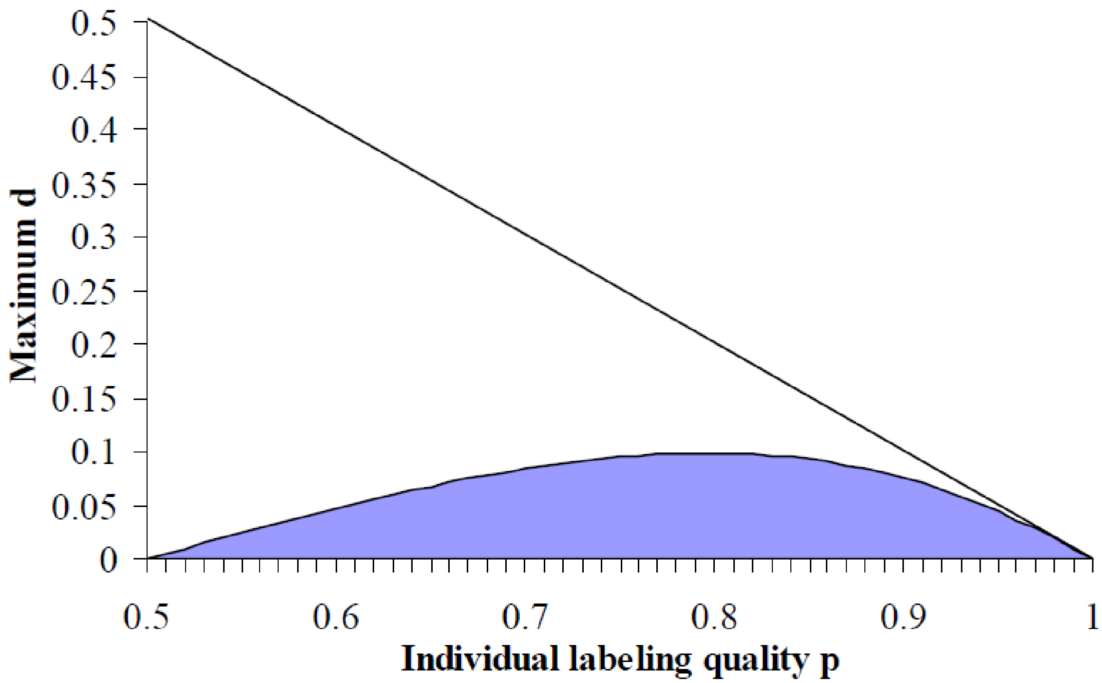
Majority voting – Uniform labeler Quality



The relationship between integrated labeling quality, individual quality, and the number of labelers

Majority voting – Different labeler Quality

- Special case of a group of three labelers with labeling qualities p-d, p and p+d



Repeated-labeling gives better quality than the best labeler (p+d) when d is below the curve

Uncertainty Preserving Labeling

- Majority voting – information about label uncertainty is lost!

- Solution…?
  1. Soft labels
     - Probabilistic label for each example
     - Difficult in practice – not all modeling techniques and software packages accommodate this
  2. Multiplied Examples (ME)
     - Create one replica of x_i with each unique label that is assigned
     - Assign weight (1/n) to each label based on the number of times it appears (n)
     - Can be incorporated into learning algorithms easily!

Part 2 - Repeated Labeling and Modeling

- How to improve classification by modifying dataset with noisy labels?

More examples

More labels per examples

Part 2 - Repeated Labeling and Modeling

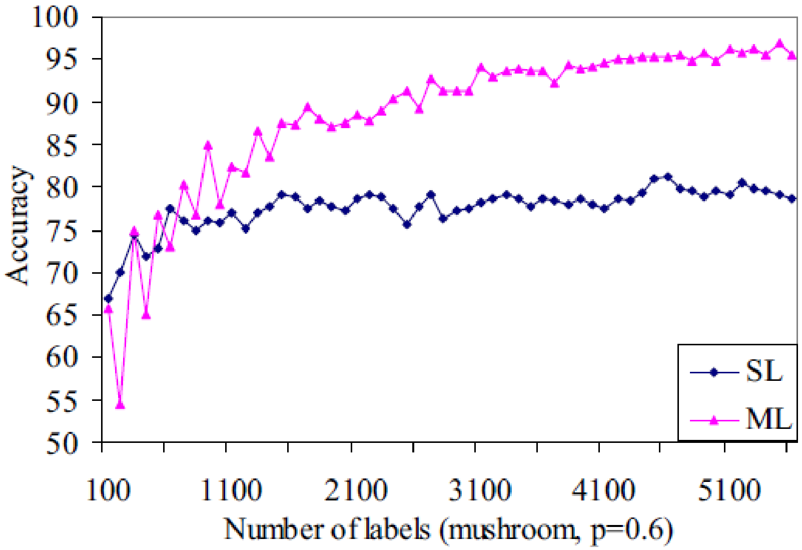
- 12 datasets selected for binary classification problem

Data Set	#Attributes	#Examples	Pos	Neg
bmg	41	2417	547	1840
expedia	41	3125	417	2708
kr-vs-kp	37	3196	1669	1527
mushroom	22	8124	4208	3916
qvc	41	2152	386	1766
sick	30	3772	231	3541
spambase	58	4601	1813	2788
splice	61	3190	1535	1655
thyroid	30	3772	291	3481
tic-tac-toe	10	958	332	626
travelocity	42	8598	1842	6756
waveform	41	5000	1692	3308

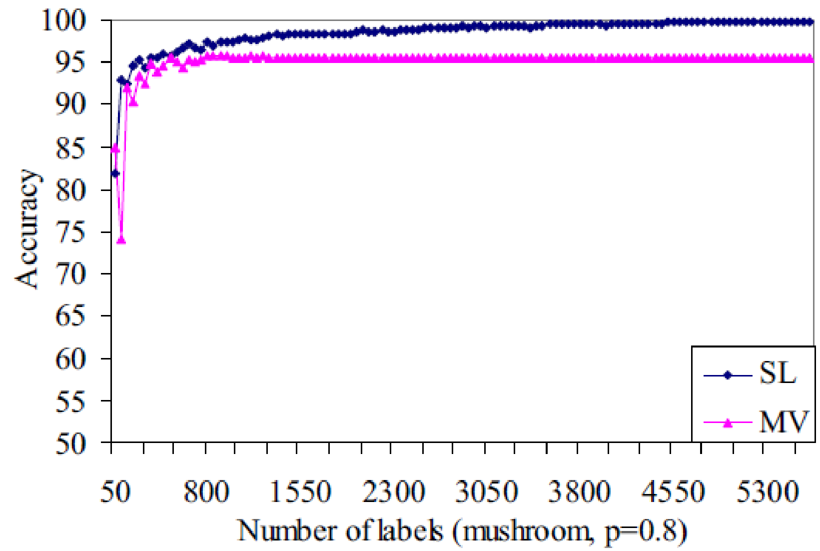
- J48 (decision tree) in WEKA used for the experiments
- 30% of examples held out in each case as test data

Round-robin Strategy, $C_U << C_L$

- Majority Voting (MV) acquires additional labels for the initial set of examples
- Single Labeling (SL) acquires new examples and their labels



(a) $p = 0.6$, $\#examples = 100$, for MV



(b) $p = 0.8$, $\#examples = 50$, for MV

Round-robin Strategy, General Costs

- Define data acquisition cost

$$C_D = C_U * T_r + C_L * N_L$$

T_r - Number of new unlabeled samples collected

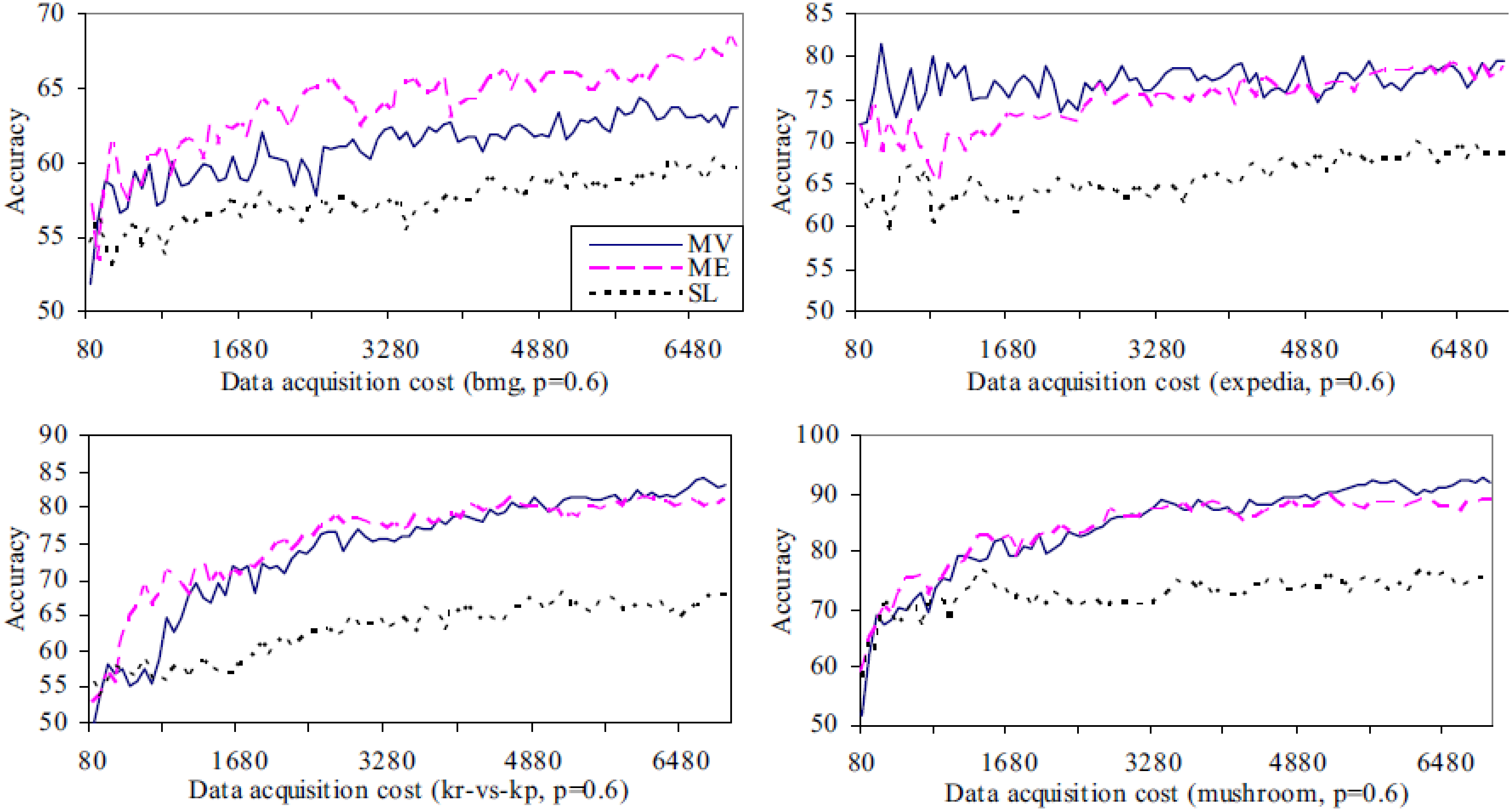
N_L - Number of samples to be labeled

- $N_L = T_r$ for single labeling, $N_L > T_r$ for repeated labeling

- New repeated labeling strategy – for every new example acquired repeated labeling acquires a fixed number of labels k, i.e. $N_L = k * T_r$

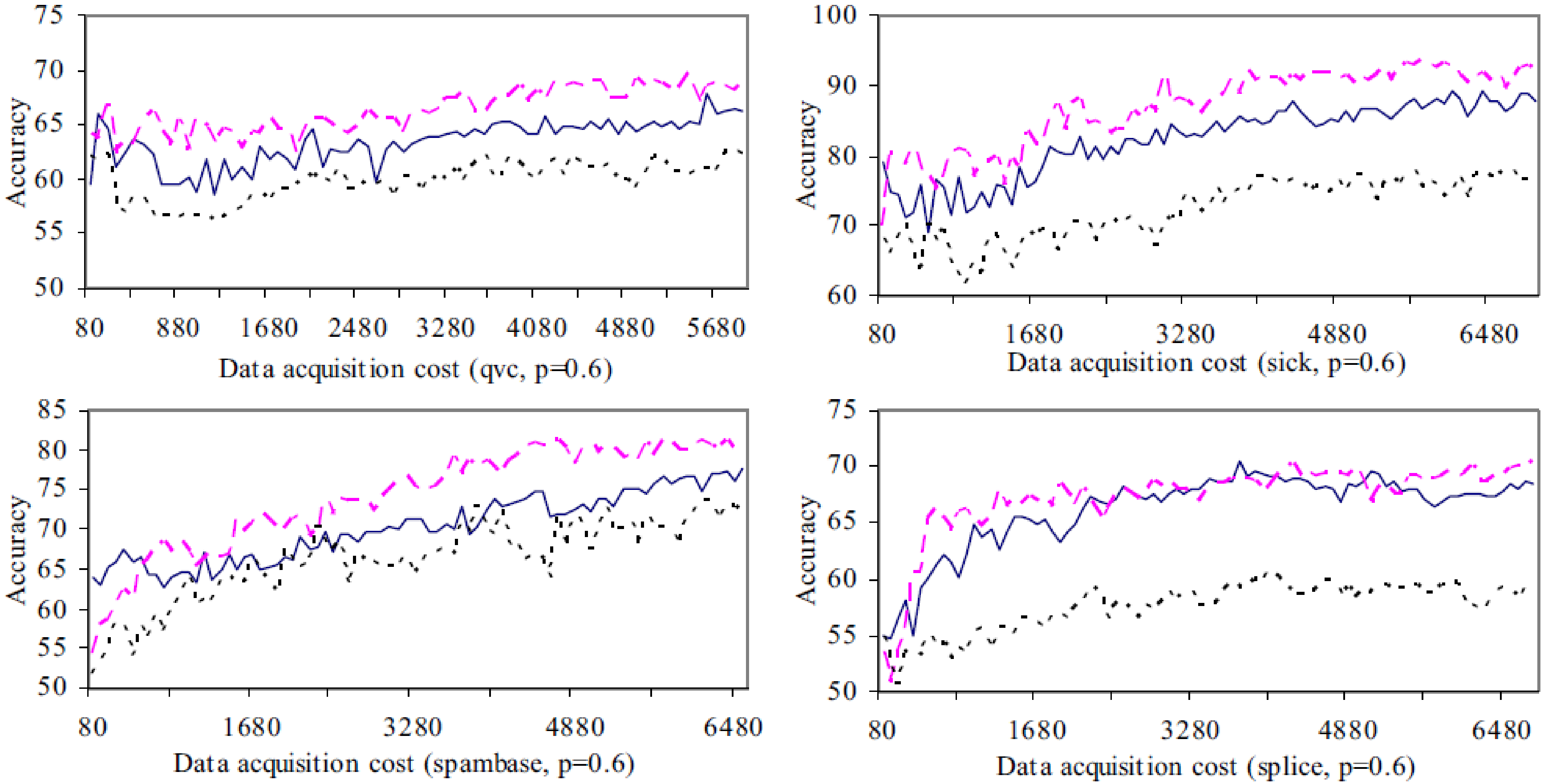
- Cost ratio ρ is defined as C_U/CL

Round-robin Strategy, General Costs



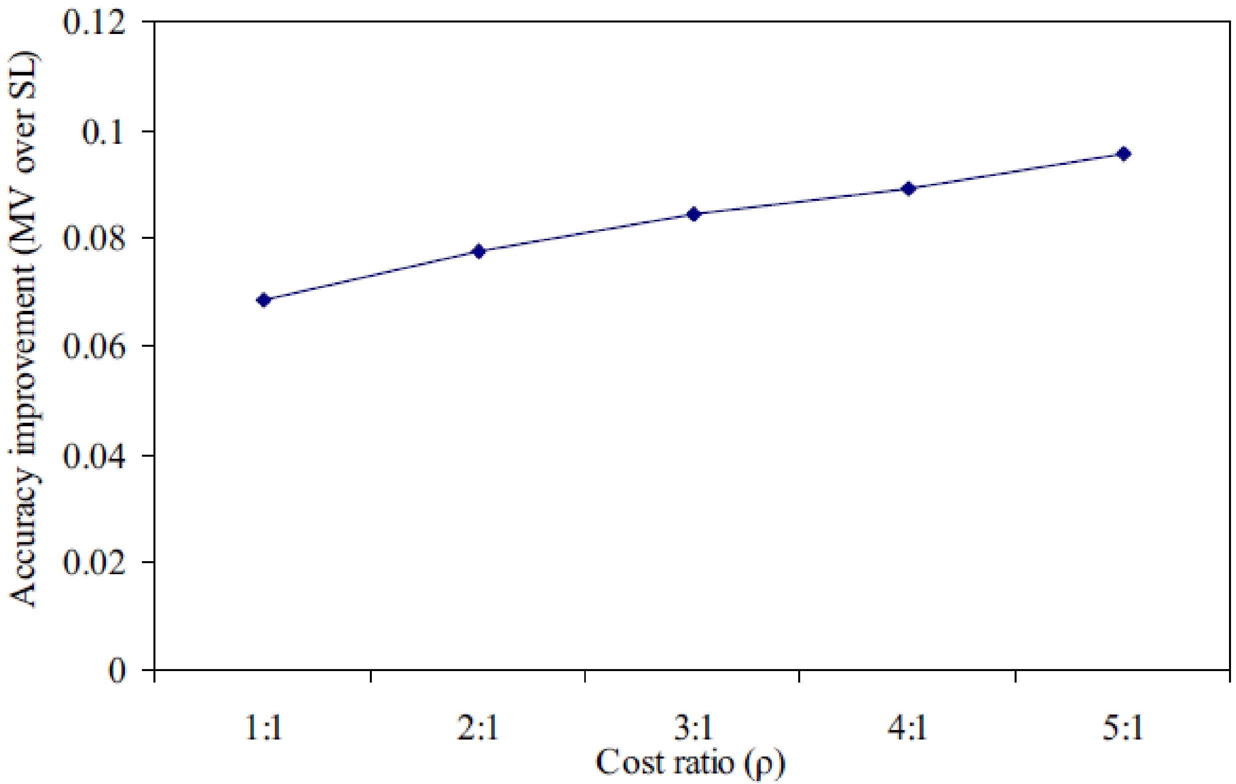
Increase in model accuracy vs data acquisition cost (ρ = 3, k = 5)

Round-robin Strategy, General Costs



Increase in model accuracy vs data acquisition cost (ρ = 3, k = 5)

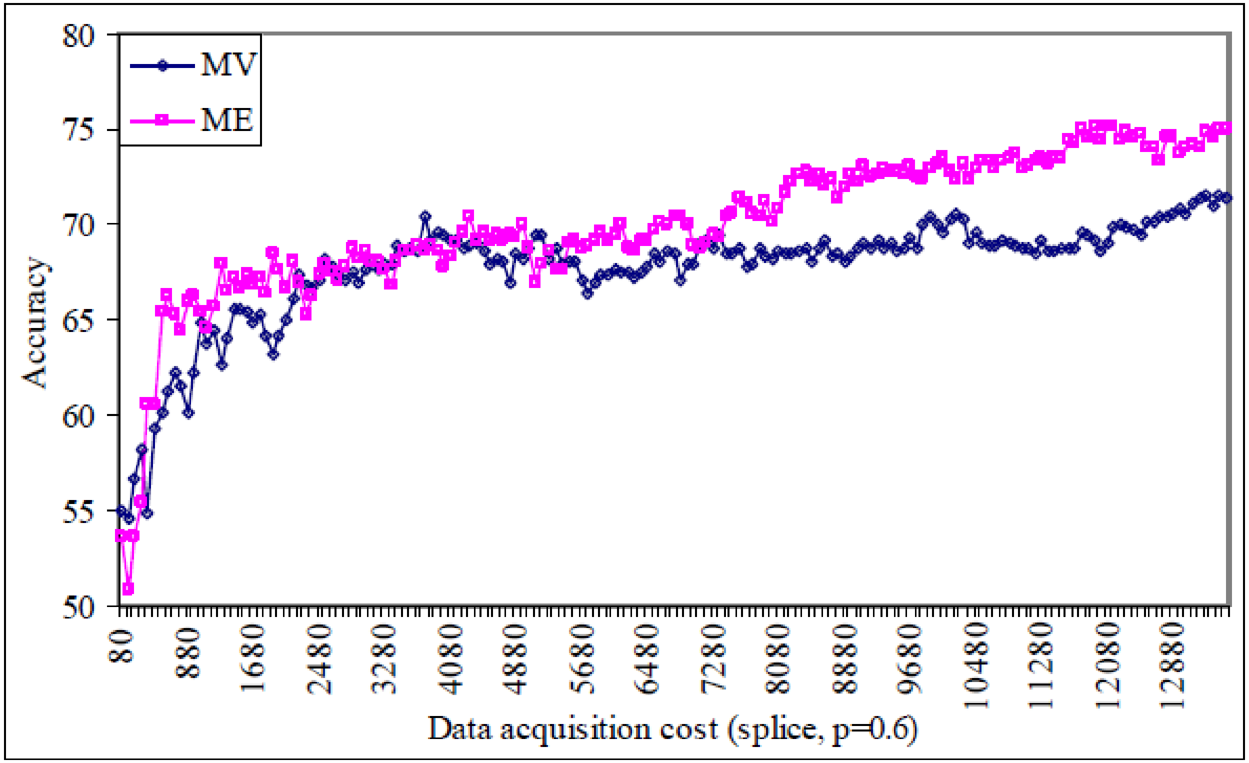
Round-robin Strategy, General Costs



Average improvement per unit cost of repeated-labeling with majority voting over single labeling

Round-robin Strategy, General Costs

Uncertainty-preserving repeated labeling performs at least as well as majority vote



The learning curves of MV and ME with p = 0.6, ρ = 3, k = 5, using the splice dataset

Selective Repeated Labeling

Do not use

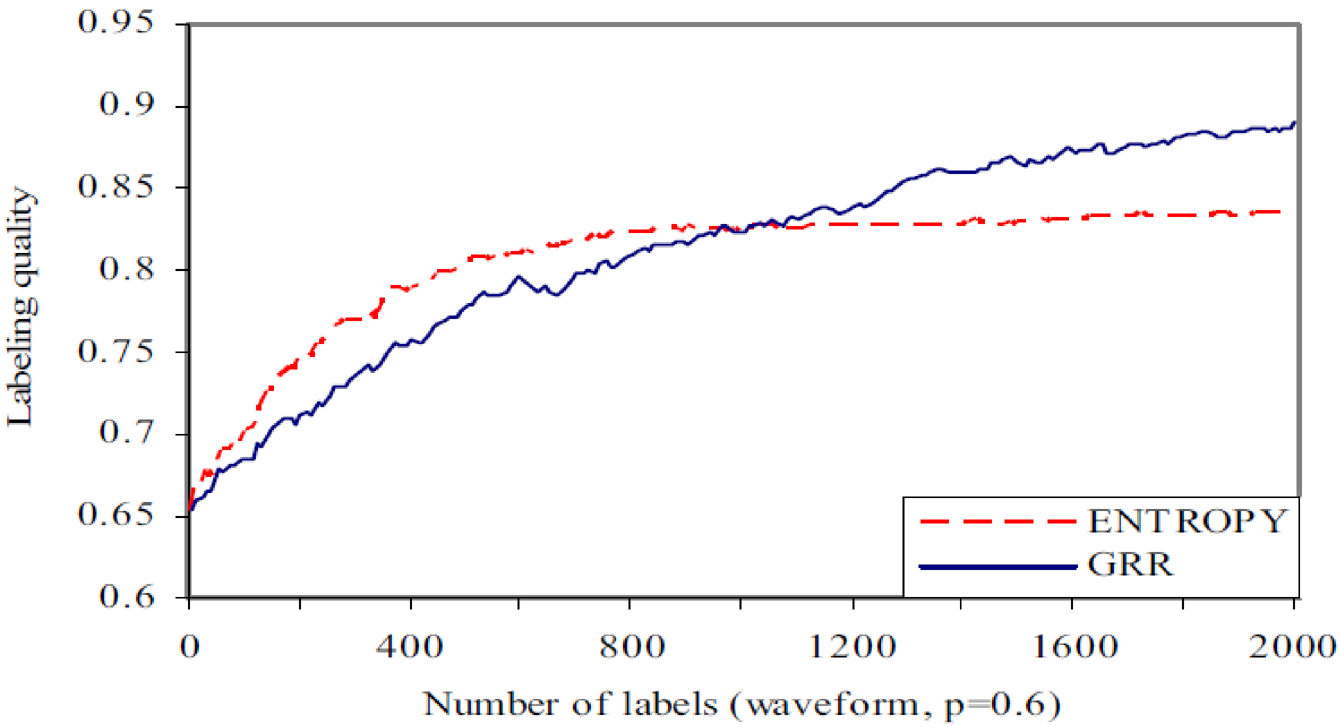
- Use entropy measure to choose examples for further labeling
  - A small set of examples are chosen many times
  - More pure but incorrect examples are never visited

- Entropy is scale invariant
  - (3+, 2-) has the same entropy as (600+, 400-)

- Fundamental problem : Entropy is not for uncertainty, but for mixture

Selective Repeated Labeling

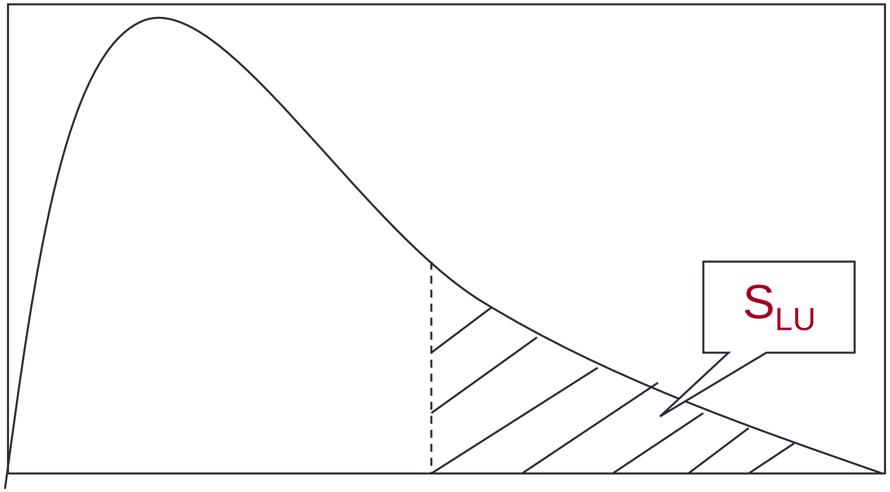
- Generalized round-robin repeated labeling outperforms entropy based selective repeated labeling



Estimating Label Uncertainty (LU)

- We compute a Bayesian estimate of the uncertainty in the class of the example
- Prior distribution over the true label is assumed to be uniform in the interval [0, 1]
- Posterior probability thus follows a Beta distribution $B(L_{pos} + 1, L_{neg} + 1)$
- Tail probability below a labeling decision threshold (0.5) is chosen as the measure of uncertainty

Beta probability density function



Estimating Model Uncertainty (MU)

- We apply traditional active learning score ignoring the current multiset of labels
- Learn a set (m) of models each of which predicts the probability of a class membership, yielding the uncertainty score:

$$S_{MU} = 0.5 - \left| \frac{1}{m} \sum_{i=1}^{m} \Pr(+|x, H_i) - 0.5 \right|$$

- $\Pr(+|x, H)$ is the probability of classifying he example x into $+$ by the learned model H
- In our experiments, $m = 10$ and model is set to random forest (WEKA)

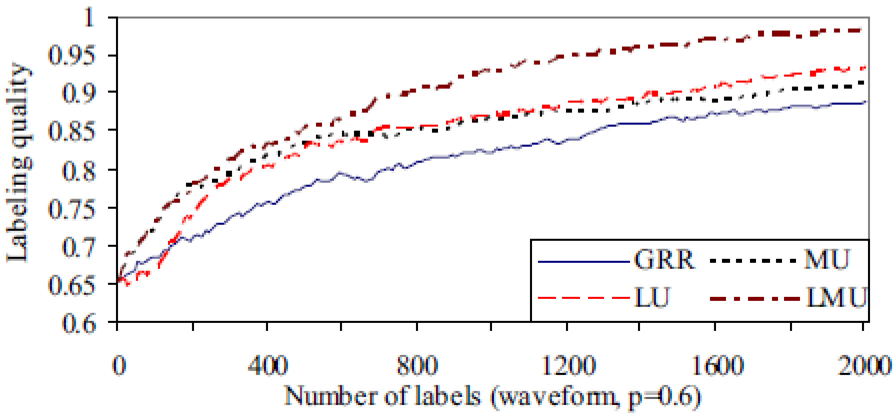
Combining Label and Model Uncertainties (LMU)

- Finally we combine label and model uncertainty scores to get the best of both worlds

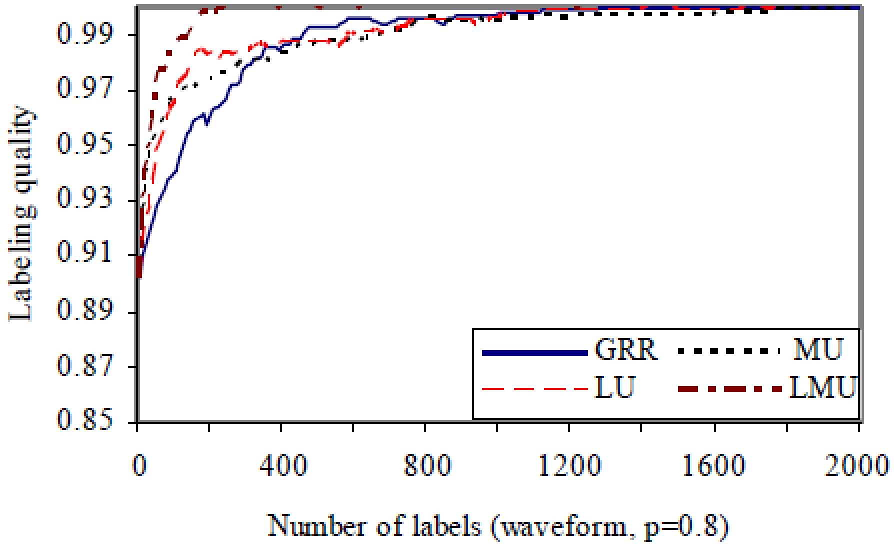
$$S_{LMU} = \sqrt{S_{LU} \times S_{MU}}$$

Experiment Results

- In high noise setting ($p = 0.6$), MU performs well – learned models can help to choose good examples to relabel!
- LMU dominates throughout



(a) $p = 0.6$



(b) $p = 0.8$

Experiment Results

Average accuracies for noisy setting, $p = 0.6$

Data Set	GRR	MU	LU	LMU
bmg	*62.97*	**71.90**	64.82	68.93
expedia	*80.61*	84.72	81.72	**85.01**
kr-vs-kp	76.75	*76.71*	81.25	**82.55**
mushroom	*89.07*	94.17	92.56	**95.52**
qvc	*64.67*	**76.12**	66.88	74.54
sick	*88.50*	93.72	91.06	**93.75**
spambase	*72.79*	79.52	77.04	**80.69**
splice	69.76	*68.16*	**73.23**	73.06
thyroid	*89.54*	93.59	92.12	**93.97**
tic-tac-toe	*59.59*	62.87	61.96	**62.91**
travelocity	*64.29*	**73.94**	67.18	72.31
waveform	*65.34*	69.88	66.36	**70.24**
average	*73.65*	78.77	76.35	**79.46**

Summary of Results

- Repeated labeling can improve **data quality** and **model quality** (but not always)
- Repeated labeling can be preferable to single labeling when labels aren't particularly cheap
- When labels are relatively cheap, repeated labeling can do much better
- Round-robin repeated labeling does well
- Selective repeated labeling performs better

Future Work

- Estimating labelers' quality by observing assigned labels could allow for more sophisticated selective repeated-labelling strategies.
- Study of labeling quality variation with labeler payment.
- Here we introduced noise to the labels. Using real labelers should give a better understanding of the effects of repeated labeling.
- We compared repeated labeling vs fixed labeling, a hybrid process of combining both based on the expected benefit of either methods could provide better data quality.


THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
DANKSCHEEN
TASHAKKUR ATU
YAQHANYELAY
SUKSAMA
EKHMET
GRAZIE
MEHRBANI
BİYAN
SHUKRIA
TINGKI
BOLZÏN
MERCI
PALDIES
KOMAPSUMNIDA
MAAKE
GOZAIMASHITA
EFCHARISTO
JUSPAXAR
SPASSIBO
SNACHALHUYA
NUHUN
CHALYU
WABEEJA
MAITEKA
YUSPAGARATAM
HUI
DHANYABAAD
ANIHA
ATTO
MERSI
SPASIBO
DENKAUJA
NENACHALHYA
UNALCHEESH
GUI
HATUR
EKOJU
SIKOMO
MAKETAI
SANCO
MERASTAWHY
GAEJTHO
LAH
TAVTAPUCH
MEDAWAGSE
AGUYJE
FAKAAUE
BAIIKA
MINMONCHAR

KEEP
CALM

AND

RAISE YOUR HAND TO

ASK A QUESTION