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

$$q = \sum_{i=0}^{N} \binom{2N + 1}{i} \cdot p^{2N+1-i} \cdot (1 - p)^i$$
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 example
Part 2 - Repeated Labeling and Modeling

- 12 datasets selected for binary classification problem

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

- 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

\[ C_D = C_U \times T_r + C_L \times N_L \]

- For single labeling, \( N_L = T_r \)
- For repeated labeling, \( N_L > T_r \)

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

- Cost ratio \( \rho \) is defined as \( C_U/CL \)
Round-robin Strategy, General Costs

Increase in model accuracy vs data acquisition cost ($\rho = 3$, $k = 5$)
Round-robin Strategy, General Costs

Increase in model accuracy vs data acquisition cost ($\rho = 3, k = 5$)
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 $p = 0.6$, $\rho = 3$, $k = 5$, 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
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 the 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
## Experiment Results

Average accuracies for noisy setting, $p = 0.6$

<table>
<thead>
<tr>
<th>Data Set</th>
<th>GRR</th>
<th>MU</th>
<th>LU</th>
<th>LMU</th>
</tr>
</thead>
<tbody>
<tr>
<td>bmg</td>
<td>62.97</td>
<td>71.90</td>
<td>64.82</td>
<td>68.93</td>
</tr>
<tr>
<td>expedia</td>
<td>80.61</td>
<td>84.72</td>
<td>81.72</td>
<td>85.01</td>
</tr>
<tr>
<td>kr-vs-kp</td>
<td>76.75</td>
<td>76.71</td>
<td>81.25</td>
<td>82.55</td>
</tr>
<tr>
<td>mushroom</td>
<td>89.07</td>
<td>94.17</td>
<td>92.56</td>
<td>95.52</td>
</tr>
<tr>
<td>qvc</td>
<td>64.67</td>
<td>76.12</td>
<td>66.88</td>
<td>74.54</td>
</tr>
<tr>
<td>sick</td>
<td>88.50</td>
<td>93.72</td>
<td>91.06</td>
<td>93.75</td>
</tr>
<tr>
<td>spambase</td>
<td>72.79</td>
<td>79.52</td>
<td>77.04</td>
<td>80.69</td>
</tr>
<tr>
<td>splice</td>
<td>69.76</td>
<td>68.16</td>
<td>73.23</td>
<td>73.06</td>
</tr>
<tr>
<td>thyroid</td>
<td>89.54</td>
<td>93.59</td>
<td>92.12</td>
<td>93.97</td>
</tr>
<tr>
<td>tic-tac-toe</td>
<td>59.59</td>
<td>62.87</td>
<td>61.96</td>
<td>62.91</td>
</tr>
<tr>
<td>travelocity</td>
<td>64.29</td>
<td>73.94</td>
<td>67.18</td>
<td>72.31</td>
</tr>
<tr>
<td>waveform</td>
<td>65.34</td>
<td>69.88</td>
<td>66.36</td>
<td>70.24</td>
</tr>
<tr>
<td>average</td>
<td>73.65</td>
<td>78.77</td>
<td>76.35</td>
<td>79.46</td>
</tr>
</tbody>
</table>
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
KEEP CALM
AND
RAISE YOUR HAND TO
ASK A QUESTION