Multiclass Multilabel Classification with More Classes than Examples

Ohad Shamir
Weizmann Institute of Science
Joint work with Ofer Dekel, MSR

NIPS 2015 Extreme Classification Workshop
Label set is a **folksonomy** (a.k.a. collaborative tagging or social tagging)
Leonardo da Vinci

From Wikipedia, the free encyclopedia

"Da Vinci" redirects here. For other uses, see Da Vinci (disambiguation).

This is a Renaissance Florentine name. The name da Vinci is an indicator of birthplace, not a family name; this person is properly referred to by the given name Leonardo.

Leonardo di ser Piero da Vinci, more commonly Leonardo da Vinci, (Italian: [leoˈnarدو da (v)ˈvintʃi] ([listen]; 15 April 1452 – 2 May 1519) was an Italian polymath whose areas of interest included invention, painting, sculpting, architecture, science, music, mathematics, engineering, literature, anatomy, geology, astronomy, botany, writing, history, and cartography. He has been variously called the father of paleontology, ichnology, and architecture, and is widely considered one of the greatest painters of all time.[1] Sometimes credited with the inventions of the parachute, helicopter and tank,[2][3][4] his genius epitomized the Renaissance humanist ideal.

Many historians and scholars regard Leonardo as the prime exemplar of the "Universal Genius" or "Renaissance Man", an individual of "unquenchable curiosity" and "feverishly inventive imagination".[5]
Categories

1452 births / 1519 deaths / 15th century in science / ambassadors of the republic of Florence / Ballistic experts / Fabulists / giftedness / mathematics and culture / Italian inventors / Members of the Guild of Saint Luke / Tuscan painters / people persecuted under anti-homosexuality laws...
Problem Definition

- Multiclass multilabel classification
- $m$ training examples, $k$ categories
- $m, k \to \infty$ together
  - Possibly even $k > m$
- **Goal:** Categorize unseen instances
Supervised learning starts with binary classification ($k=2$) and extends to multiclass learning

- Theory: VC dimension $\rightarrow$ Natarajan dimension
- Algorithms: binary $\rightarrow$ multiclass

- Usually, assume $k = O(1)$

- Some exceptions
  - Hierarchy with prior knowledge on relationships – not always available
  - Additional assumptions (e.g. talk by Marius earlier)
• Classify the web based on Wikipedia categories
• Training set: All Wikipedia pages ($m = 4.2 \times 10^6$)
• Labels: All Wikipedia categories ($k = 1.1 \times 10^6$)
Challenges

• **Statistical problem**: Can’t get a large (or even moderate) sample from each class.

• **Computational problem**: Many classification algorithms will choke on millions of labels.
A bipartite graph derived from search engine logs: clicks encoded as weighted edges

Wikipedia pages are labeled web pages

Labels propagate along edges to other pages
Example


- Among them
  - “Renaissance artists” – good
  - “1452 births” – bad

- Observation: “1452 births” induces many false-positives (FP): best to remove it altogether from classifier output
  - (FP ⇒ TN, TP ⇒ FN)
Simple Label Pruning Approach

1. Split dataset to training and validation set
2. Use training set to build an initial classifier $h_{pre}$ (e.g. by propagating labels over click-graph)
3. Apply $h_{pre}$ to validation set, count FP and TP
4. $\forall j \in \{1, ..., k\}$, remove label $j$ if

$$\frac{FP_j}{TP_j} > \frac{1 - \gamma}{\gamma}$$

- Defines a new “pruned” classifier $h_{post}$
Explicitly minimizes empirical risk with respect to the $\gamma$-weighted loss:

$$\ell(h(x), y) = \sum_{j=1}^{k} \left[ \gamma \mathbb{I}(h_j(x) = 1, y_j = 0) + (1 - \gamma) \mathbb{I}(h_j(x) = 0, y_j = 1) \right]$$

FP (false positive)

FN (false negative)
Main Question

Would this actually reduce the risk?

$$\mathbb{E}_{(x,y)}[\ell(h_{post}(x),y)] < \mathbb{E}_{(x,y)}[\ell(h_{pre}(x),y)] \quad \text{- positive}$$
Baseline Approach

• Prove that uniformly for all labels $j$

\[
\frac{\hat{FP}_j}{\hat{TP}_j} \rightarrow \frac{FP_j}{TP_j}
\]

**Problem**: $m, k \rightarrow \infty$ together. Many classes only have a handful of examples.
Uniform Convergence Approach

• Algorithm implicitly chooses a hypothesis from a certain hypothesis class
  – Pruning rules on top of fixed predictor $h_{pre}$

• Prove uniform convergence by bounding VC dimension / Rademacher complexity

• Conclude that if empirical risk decreases, the risk decreases as well
Uniform Convergence Fails

- Unfortunately, no uniform convergence...
- ... and even no algorithm/data-dependent convergence!

\[
\mathbb{E}[R(h_{post}) - \hat{R}(h_{post})] \geq \\
\sum_{j=1}^{k} Pr(j \text{ pruned}) (TP_j - FP_j) \\
= \sum_{j=1}^{k} Pr(\hat{FP}_j > \hat{TP}_j) (TP_j - FP_j)
\]

Weak correlation in \( m \approx k \) regime
Prove **directly** that risk decreases

Important (but mild) assumption: Each example labeled by \( \leq s \) labels

**Step 1:** Risk of \( h_{post} \) is **concentrated.** For all \( \epsilon \),

\[
\Pr \left( \left| R(h_{post}) - \mathbb{E}R(h_{post}) \right| \right)
\]
Part 2: Enough to prove $R(h_{pre}) - \mathbb{E}R(h_{post}) > 0$

Assuming for $\gamma = \frac{1}{2}$ for simplicity, can be shown that

$$R(h_{pre}) - \mathbb{E}R(h_{post}) > \text{pos} - \mathcal{O}\left(\sqrt{\frac{1}{m} \left\| (FP_j + TP_j)_j \right\|_{1/2}}\right)$$

where $\|w\|_{1/2} = (\sum_j \sqrt{w_j})^2$
A Less Obvious Approach

\[ \sum_{j:FP_j \geq TP_j} (FP_j - TP_j) \]

For probability vector, always at most \( k \)

Smaller the more non-uniform is the distribution

\[ R(h_{pre}) - E_R(h_{post}) \]

\[ > \ pos - O \left( \sqrt{\frac{\| (FP_j + TP_j)_j \|_1}{m}}^{1/2} \right) \]

where \( \| w \|_{1/2} = (\sum_j \sqrt{w_j})^2 \)
Wikipedia Power-Law: $r = 1.6$
Wikipedia Power-Law: $r = 1.6$

\[ R(h_{\text{pre}}) - \mathbb{E}R(h_{\text{post}}) > \text{pos} - \mathcal{O}\left(\sqrt{\frac{k^{0.4}}{m}}\right) \]
Experiment

Click graph on the entire web (based on search engine logs)
Experiment

Categories from Wikipedia pages propagated twice through graph
Experiment

Train/test split of Wikipedia pages
How good are propagated categories from training set in predicting categories at test set pages?
Experiment

![Graph showing comparison of different methods in terms of ratio with best loss against parameter γ. The graph includes three lines: blue for our algorithm, magenta for original, and black for random. The x-axis represents different values of γ from 0.02 to 0.1, while the y-axis represents the ratio with best loss ranging from 0 to 1.]
Another less obvious approach

\[ R(h_{pre}) - \mathbb{E}R(h_{post}) \]

\[ = \sum_{j=1}^{k} Pr(j \text{ pruned}) (FP_j - TP_j) \]

\[ = \sum_{j=1}^{k} Pr(\hat{FP}_j > \hat{TP}_j) (FP_j - TP_j) \]

Weak but positive correlation, even if only few examples per label

For **large** \( k \), sum will tend to be positive
Different Application: Crowdsourcing

(Dekel and S., 2009)
Different Application: Crowdsourcing

(Dekel and S., 2009)
Different Application: Crowdsourcing

(Dekel and S., 2009)
Different Application: Crowdsourcing

(Dekel and S., 2009)
• How can we improve crowdsourced data?
• Standard approach: Repeated labeling, but expensive
• A bootstrap approach:
  – Learn predictor from data of all workers
  – Throw away examples labeled by workers disagreeing a lot with the predictor
  – Re-train on remaining examples
• Works! (Under certain assumptions)
• Challenge: Workers often label only a handful of examples
# examples/worker might be small, but many workers...
Different Application: Crowdsourcing

# examples/worker might be small, but many workers...
Different Application: Crowdsourcing

# examples/worker might be small, but many workers...
Conclusions

• # classes $\rightarrow \infty$ violates assumptions of most multiclass analyses
  – Often based on generalizations of binary classification

• Possible approach
  – Avoid standard analysis
  – “Extreme X” can be a blessing rather than a curse

• Other applications? More complex learning algorithms (e.g. substitution)?
Thanks!