Weakly supervised learning – focus on Active Learning

Vincent Lemaire
Preamble

Hope I'm not being too forward with my introduction, but I like your preface.

At The Book Club
What is this talk about?

- Machine learning from **big labeled data** is highly successful
  - Speech recognition, image understanding, natural language translation, …

- However, there are various applications where **massive labeled data is not available**
  - Medicine, robots, frauds, …

- In this talk I will discuss about **classification from limited information**
  1. Weak data (but we assume that we have a lot of them)
  2. Small data (possible without strong (prior) domain knowledge?)
Pour Facebook, l’intelligence artificielle sera économe en données ou ne sera pas

Julien Cadot - 09 décembre 2019 - Sciences

Focus on a particular target problem: Binary Supervised Classification

- Random Forest
- Kappa on Test

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#D</th>
<th>#Y</th>
<th>#num</th>
<th>#nom</th>
<th>maj (%)</th>
<th>min (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>musk</td>
<td>6581</td>
<td>2</td>
<td>166</td>
<td>1</td>
<td>84.55</td>
<td>15.45</td>
</tr>
</tbody>
</table>

- Large amount of labeled data yields better performances
- Estimation error decreases in order $1 / \sqrt{|L|}$
Focus on a particular target problem: Binary Supervised Classification

1. Unsupervised classification
2. Semi-Supervised classification
3. Supervised classification

Details to come on the first two
Unsupervised Classification

• Gathering labeled data is costly
• Try to use unlabeled data (only)
• Unsupervised Classification is typically clustering
• ‘Assumption’: each cluster corresponds to a class
Semi-Supervised Classification

- Use:
  - a large number of unlabeled samples
  - a small number of labeled samples
- Try to find a ‘boundary’ (for example using labels propagation) along the cluster structure

*to work well*
But not only…

Supervised

Semi Supervised

Active

Positive Unlabeled
Classification of Classification

Our target: High accuracy & low cost

Figure from [...]

unsupervised

Semi-supervised

supervised

low accuracy high

labeling cost
Wanting more labels or information!
But…

Insufficient quantity of labeled data

Insufficient subject-matter expertise to label data

- specific relevant expertise required
- become prohibitively expensive
  
  example in medical domain

Insufficient time to label and prepare data

- time spent in preparing data sets
- domain by nature rapidly evolves
  
  example in fraud detection or cybersecurity applications.

…

From wikipedia
Weak supervised learning

Taxonomy: an attempt

1. Geospiza magnirostris.
2. Geospiza fortis.
Strong versus Weak
Two aspects: Supervision, Labels

1 - Strong is strong…

many labeled examples with accurate labels
Strong versus Weak
Two aspects: Supervision, Labels

2 - Types of weak ‘learning’

- **Incomplete supervision:**
  - a small amount of labeled data
  - but sometimes abundant unlabeled data are available
  - only labels on a ‘positive class’

- **Inaccurate supervision:**
  - labels are not ‘guaranteed’ (some label information may suffer from errors)
  - labels are not ‘guaranteed’ (and are on ‘bag of examples’ (a set of keys))

- **Inexact supervision:**
  - labels are on ‘bag of examples’ (a set of keys)
Strong supervised learning

Weakly supervised learning
Strong supervised learning

Weakly supervised learning

True labels

Inaccurate labels (label noise, …)
Weakly supervised learning

Inaccurate Labels versus True labels

- Inaccurate or imprecise labels
  - labels are on ‘bag of examples’
  - labels are not ‘guaranteed’, noisy labels:
    - learning with label noise
      - use an algorithm robust to the label noise (if noise marginal)
      - try to model the labels and the noise (with assumption on the noise)
      - filter the noisy training set to have a clean one

- True labels but incomplete supervision (incomplete information)
  - Few labels are available
  - Only true labels on one class
  - Labels at or not at the right ‘proxy’

Zhi-Hua Zhou, 2017
“A Brief Introduction to Weakly Supervised Learning”

Strong supervised learning

Weakly supervised learning

3 criterion on labels:
• Quantity?
• Quality?
• Adapted?

true labels

Inaccurate label (label noise, …)
Strong supervised learning

- 3 criteria on labels:
  - Quantity?
  - Quality?
  - Adapted?

Weakly supervised learning

- True labels
- Labels at the right 'proxy'

Inaccurate label (label noise, ...)

- Labels not at the right 'proxy'
Inexact supervision

- concerns about the situation where some supervision information is given, but not as exact as desired or at the right proxy or labels* are on subsets of the data
- example 1: is there an protest?
  - detect people, how many people, distance between people, …

*but that could be noisy and may conflict
*general: multiple noisy labeling functions can conflict and have dependencies
Inexact supervision

- concerns about the situation where some supervision information is given, but not as exact as desired or at the right proxy or labels* are on subsets of the data
- transfer learning
- multi-instance learning
- build ‘labels” (Snuba, Snorkel, …)

*but that could be noisy and may conflict
*general: multiple noisy labeling functions can conflict and have dependencies
Strong supervised learning

- Weakly supervised learning

3 criterion on labels:
- Quantity?
- Quality?
- Adapted?

true labels

labels at the right ‘proxy’

Inaccurate label (label noise, …)

labels not at the right ‘proxy’
Strong supervised learning

- 3 criterion on labels:
  - Quantity?
  - Quality?
  - Adapted?

Weakly supervised learning

- True labels
- Labels at the right 'proxy'

- Inaccurate label (label noise, …)
- Labels not at the right 'proxy'

- Multi instances learning
- Build labels at the right 'proxy' (snuba…)

- Move to right 'proxy'
  - Transfer learning (domain adaptation)
Strong supervised learning

3 criterion on labels:
- Quantity?
- Quality?
- Adapted?

Weakly supervised learning

true labels

Innacurate label (label noise, …)

labels at the right ‘proxy’
- very few labels
- few labels
- few but more labels

move to right ‘proxy’
- transfert learning (domain adaptation)
- build labels at the right ‘proxy’ (snuba…)

labels not at the right ‘proxy’
- multi instances learning

very few labels
few labels
few but more labels

3 criterion on labels:
- Quantity?
- Quality?
- Adapted?
Strong supervised learning

Weakly supervised learning

3 criteria on labels:
- Quantity?
- Quality?
- Adapted?

true labels

Inaccurate label
(label noise, …)

labels at the right
‘proxy’

labels not at the right
‘proxy’

very few labels

few labels

few but more labels

move to right ‘proxy’

transfert learning
(domain adaptation)

multi instances
learning

build labels at
the right ‘proxy’
(snuba…)

few but more labels

|\(|L|\)
Strong supervised learning

Weakly supervised learning

3 criteria on labels:
- Quantity?
- Quality?
- Adapted?

true labels

Inaccurate label (label noise, …)

labels at the right ‘proxy’

labels not at the right ‘proxy’

very few labels

few labels

few but more labels

very few labels

few labels

few but more labels

move to right ‘proxy’

transfert learning (domain adaptation)

multi instances learning

build labels at the right ‘proxy’ (snuba…)

labels at the right ‘proxy’

3 criteria on labels:
- Quantity?
- Quality?
- Adapted?

active learning (oracle)

semi supervised learning (SSL) (|L| + |U|)

self training

co-training (et extension)

semi supervised learning (SSL) (|L| + |U|)
Active Learning (principle)

Definition:
• Active Component: ask queries to an oracle
• Improve the performance of a classifier
• Minimizing the cost of obtaining labeled data

Conclusion:
• Active Learning optimizes a performance which is induced by a classifier through selecting the most beneficial unlabeled instances to be labeled by an oracle to build the training basis.
Semi-supervised learning

Semi-supervised learning attempts to automatically exploit unlabeled data in addition to labeled data to improve learning performance, where no “human” intervention is assumed.

- generative models
- low-density separation
- graph-based methods
- heuristic approaches
  - self training
  - co-training
  - …

« Semi-Supervised Learning », Chapelle et al. The MIT Press 2010
Self training

Idea: Train, predict, re-train using classifier’s best predictions, repeat

1-NN good case
Self training

Idea: Train, predict, re-train using classifier’s best predictions, repeat

1-NN bad case
Co-Training

- Each instance has “two- (independent)-views”
- Each view should provide a “good classifier”
- Each view teach the other view (by providing labeled instances)

Strong supervised learning

- active learning (oracle)
- semi supervised learning (SSL) \( |L| + |U| \)
- self training
- co-training (et extension)

Weakly supervised learning

- true labels
- labels at the right 'proxy'
- very few labels
- few labels
- few but more labels

- Innacurate label (label noise, …)
- labels not at the right 'proxy'
- move to right 'proxy'
- transfert learning (domain adaptation)
- multi instances learning

3 criterion on labels:
- Quantity ?
- Quality ?
- Adapted ?
**Strong supervised learning**

- **Weakly supervised learning**
  - True labels
  - Labels at the right ‘proxy’
    - Very few labels
    - Few labels
    - Few but more labels
  - Labels not at the right ‘proxy’
    - Innaccurate label (label noise, …)
    - Multi instances learning
    - Move to right ‘proxy’
      - Transfer learning (domain adaptation)

**Criteria on labels:**
- Quantity?
- Quality?
- Adapted?

**Methods:**
- Active learning (oracle)
- Semi supervised learning (SSL) (|L| + |U|)
- Self training
- Co-training (et extension)
- Usual SSL
- Transductive learning (TL)
- Inductive Learning (IL)
- Positive unlabeled learning (PUL)
- IL
- TL

**Types of learning:**
- Transductive learning (TL)
- Inductive Learning (IL)
- Active learning (oracle)
- Semi supervised learning (SSL)
- Self training
- Co-training (et extension)
- Usual SSL
- Positive unlabeled learning (PUL)
- IL
- TL
Strong supervised learning

- 3 criterion on labels:
  - Quantity?
  - Quality?
  - Adapted?

Weakly supervised learning

- True labels
- Innacurate label (label noise, …)

- Labels at the right ‘proxy’
  - Few labels
  - Few but more labels
  - Very few labels

- Labels not at the right ‘proxy’
  - Multi instances learning
  - Move to right ‘proxy’
    - Transfer learning (domain adaptation)

Very few labels

Few labels

Few but more labels

Active learning (oracle)

Semi supervised learning (SSL) (|L| + |U|)

Self training

Co-training (et extension)

Usual SSL

Positive unlabeled learning (PUL)

Inductive Learning (PUL)

Transductive Learning (TL)

Inductive Learning (IL)

Transductive learning (TL)

Inductive Learning (IL)

Active learning (oracle)

Semi supervised learning (SSL) (|L| + |U|)

Self training

Co-training (et extension)

Usual SSL

Positive unlabeled learning (PUL)

Inductive Learning (PUL)

Transductive Learning (TL)

Inductive Learning (IL)
Weakly supervised learning

Strong supervised learning

3 criteria on labels:
- Quantity?
- Quality?
- Adapted?

Inaccurate labels (label noise, …)

True labels

labels at the right 'proxy'

- very few labels
- few labels
- few but more labels

Move to right 'proxy'

Transfer learning (domain adaptation)

Multi instances learning

Build labels at the right 'proxy' (snuba…)

37
Ok but more formalism required…

Example

Model Label Noise

(1) Random Classification Noise (RCN):
\[ \rho_Y(X) = P(\bar{Y}|Y, X) = P(\bar{Y}|Y); \rho_{+1}(X) = \rho_{-1}(X) = \rho. \]

(2) Class-Dependent Noise (CCN):
\[ \rho_Y(X) = P(\bar{Y}|Y, X) = P(\bar{Y}|Y); \rho_{+1}(X) = \rho_{+1}, \rho_{-1}(X) = \rho_{-1}. \]

(3) Instance- and Label-Dependent Noise (ILN):
\[ \rho_Y(X) = P(\bar{Y}|Y, X). \]
Hope I'm not being too forward with my introduction, but I like your preface.

At The Book Club
Active learning – outline

- Topic 1: Selection Strategies (or not)
- Topic 2: Evaluation of Pool-based Active Learning
- Topic 3: Software Framework
- Application: Sorting Robot
Topic 1: Selection Strategies (or not)
Active Learning

From Education...

C. Bonwell and J. Eison [1]: In active learning, students participate in the process and students participate when they are doing something besides passively listening. It is a model of instruction or an education action that gives the responsibility of learning to learners themselves.

...to Machine Learning:

Settles [2, p.5]: Active learning systems attempt to overcome the labeling bottleneck by asking queries in the form of unlabeled instances to be labeled by an oracle. In this way, the active learner aims to achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data.


Active Learning  From Education . . . . to Machine Learning:
Active Learning

Setting
• Some information is costly (some not)
• Active learner controls selection process

Objective
• Select the most valuable information
• Baseline: Random selection

Historical Remarks
• Optimal experimental design
• Learning with queries/query synthesis
• Selective sampling
Selective Data Acquisition Tasks

Active Learning Scenarios

- Query synthesis: example generated upon query
- Pool U of unlabeled data: static, repeated access
- Stream: sequential arrival, no repeated access

Type of Selected Information

- Active label acquisition
- Active feature (value) acquisition
- Active class selection, also denoted Active class-conditional example acquisition
- \ldots
Selective Data Acquisition Tasks

A short diverticula

Active Sampling (inductive learning)
main assumption: obtaining an unlabeled instance is not free

Selective Sampling (transductive learning)
main assumption: obtaining an unlabeled instance is free

Combine?
Link with "Active class selection"?

from pool of unlabelled data

this talk
Definition of Active Learning

Definition:
- Active Component: ask queries to an oracle
- Improve the performance of a classifier
- Minimizing the cost of obtaining labeled data

Conclusion:
Active Learning optimizes a performance which is induced by a classifier through selecting the most beneficial unlabeled instances to be labeled by an oracle to build the training basis.
What factors influence the decision?

- Density (improve the classifier, where decisions are important)
- Decision boundary (be specific, where change is expected)
- Label density (explore unexplored regions)
Random sampling

- Also called passive sampling
- Selects instances randomly for labeling
- Competitive approach
- Standard baseline
- Free of heuristics
- Performs very well on the ‘banana dataset’
Uncertainty sampling

Idea
• Select those instances where we are least certain about the label

Approach
• 3 labels preselected
• Linear classifier
• Use distance to the decision boundary as uncertainty measure

Uncertainty sampling

- easy to implement
- fast

- no exploration (often combined with random sampling)
- impact not considered (density weighted extensions exist)
- problem with complex structures (performance can be even worse than random)

Pure exploitation, does not explore
Can get stuck in regions with high Bayesian error
Ensemble-Based Strategy


Idea
Use disagreement between base classifiers

Approach
1. Get an initial set of labels
2. Split that set into (overlapping) subsets
3. On each subset, train a different base-classifier
4. Repeat until stop
5. On each unlabeled instance do
6. Apply all base-classifiers
7. Request label, if base-classifiers disagree
8. Update all base-classifiers
9. Go to step 4
Expected Error Reduction

- Simulates the acquisition of each label candidate and each possible outcome (class)
- Calculates the generalization error of the simulated new model
- Chooses the label with lowest generalization error

\[ x^* = \arg\min_x \sum_{i \in \{1, \ldots, C\}} P_{\theta}(y_i \mid x) \left( \sum_{x' \in \mathcal{U}} 1 - P_{\theta^+(x,y_i)}(\hat{y} \mid x') \right) \]

+ decision theoretic model
  - long execution time (closed form solutions for specific classifiers, approximations for speed up)
Probabilistic Active Learning

- Models the true posterior as being Beta-distributed
  - variance of posterior is correlated with the number of local observations
  - thereby omit the complex simulation of expected error reduction
- Calculates the performance improvement of the model
  + decision theoretic model
  + fast w.r.t. expected error reduction
  - local number of labels required

"Optimized probabilistic active learning (OPAL) for fast, non-myopic, cost-sensitive active classification", Georg Krempl, Daniel Kottke, and Vincent Lemaire. In Machine Learning, 100(2), 2015.
DUAL

- combination of density weighted uncertainty sampling and standard (uniform) uncertainty sampling
- adaptive weights

4DS

- Uses four different scores for a classifier based on Gaussian mixtures (CMM):
  - distance, density, diversity, distribution
  - automatically weighted


“Let us know your decision: Pool-based active training of a generative classifier with the selection strategy 4DS”, Tobias Reitmaier and Bernhard Sick, in Information Sciences, 2013
One-by-one vs. Batch Acquisition

- **Definition:**
  - One-by-one: subsequently selecting instances
  - Batch: selects a specific number of labeling candidates for labeling at one time

- **Batch-Acquisition:**
  - Problem: most approaches would select very similar instances
  - Approach: diversity score

*Fig. D.g – Résultats complets pour le jeu de données “Glass”*
**Strategy vs. Classifiers?**

<table>
<thead>
<tr>
<th>Stratégie</th>
<th>RF</th>
<th>SVM</th>
<th>5NN</th>
<th>GNB</th>
<th>C4.5</th>
<th>RL</th>
<th>VFDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Margin</td>
<td>[A]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entropy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[B]</td>
</tr>
<tr>
<td>SGmulti</td>
<td></td>
<td></td>
<td>[C]</td>
<td></td>
<td></td>
<td></td>
<td>[C]</td>
</tr>
<tr>
<td>ATU</td>
<td></td>
<td>[A, C]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OER</td>
<td>[A]</td>
<td>[A]</td>
<td>[A]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A new age?
Combining all of these (heuristics strategies)?

“Active Learning by Learning”, Hsu et al. in AAAI 2015.

Active Learning By Learning (ALBL) algorithm is a meta active learn algorithm designed to solve this problem. ALBL considers multiple existing active learning algorithms and adaptively learns a querying strategy based on the performance of these algorithms.

- Strategy 1: ask most confused question
- Strategy 2: ask most frequent question
- Strategy 3: ask most helpful question

Do you use a fixed strategy in practice?
Rather fixing a strategy learning a strategy?
Combining all of these (heuristics strategies)?

“Active Learning by Learning”, Hsu et al. in AAAI 2015.

Strategy 1: ask most confused question → uncertainty

Strategy 2: ask most frequent question → representative

Strategy 3: ask most helpful question → exp.-err. Reduction

Choosing one single strategy is non-trivial
Combining all of these (heuristics strategies)?

“Active Learning by Learning”, Hsu et al. in AAAI 2015.

K strategies: $A_1, A_2, \ldots, A_K$

- try one strategy
- “goodness” of strategy as reward

K bandit machines: $B_1, B_2, \ldots, B_K$

- try one bandit machine
- “luckiness” of machine as reward

Two issues: try and reward

One well-known probabilistic bandit learner (EXP4.P)

**UNCERTAIN** Best

**PSDS** Best

**QUIRE** Best

Vehicle

Sonar

Diabetes
Rather fixing a strategy learning a strategy?


"Learning active learning: an evaluation", L. Desreumaux, V. Lemaire submitted to Intelligent Data Analysis (IDA) 2020
Where are we?

Which method used (or recommend) in an industrial context?

B: Learn how to combine strategies

“Active Learning by Learning”, Hsu et al. in AAAI 2015.

Active Learning By Learning (ALBL) algorithm is a meta active learn algorithm designed to solve this problem. ALBL considers multiple existing active learning algorithms and adaptively learns a querying strategy based on the performance of these algorithms.

Claim : A ≤ B

A: Strategies

So many (heuristics) strategies suggested in the literature:

- random
- uncertainty
- error reduction
- density based
- ...

C: Learn (and transfer) a strategy


Claim : A ≤ B ≤ C

D. Pereira-Santos et al., «Empirical investigation of active learning strategies», Neurocomputing, 2019

Best(A): RF+Margin
What did we compare?

- **RF + Random**

- **RF + Margin**

- **Deep QL**
## Results

<table>
<thead>
<tr>
<th>Jeu de données</th>
<th>Rnd/LR</th>
<th>Margin/LR</th>
<th>LAL/LR</th>
<th>Rnd/RF</th>
<th>Margin/RF</th>
<th>LAL/RF</th>
<th>maj (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>adult</td>
<td>77.93</td>
<td>78.91</td>
<td>78.97</td>
<td>80.17</td>
<td>81.27</td>
<td>81.21</td>
<td>76.06</td>
</tr>
<tr>
<td>banana</td>
<td>53.03</td>
<td>57.39</td>
<td>53.12</td>
<td>80.24</td>
<td>73.81</td>
<td>73.58</td>
<td>55.16</td>
</tr>
<tr>
<td>bank-marketing-full</td>
<td>86.85</td>
<td>87.62</td>
<td>87.72</td>
<td>88.19</td>
<td>88.34</td>
<td>88.49</td>
<td>88.30</td>
</tr>
<tr>
<td>climate-simulation-craches</td>
<td>87.22</td>
<td>89.13</td>
<td>88.62</td>
<td>91.15</td>
<td>91.14</td>
<td>91.13</td>
<td>91.48</td>
</tr>
<tr>
<td>cecg-eye-state</td>
<td>56.08</td>
<td>55.32</td>
<td>56.11</td>
<td>65.53</td>
<td>67.58</td>
<td>64.42</td>
<td>55.12</td>
</tr>
<tr>
<td>hiva</td>
<td>64.43</td>
<td>70.84</td>
<td>71.80</td>
<td>96.32</td>
<td>96.47</td>
<td>96.44</td>
<td>96.50</td>
</tr>
<tr>
<td>ibn-sina</td>
<td>84.77</td>
<td>88.58</td>
<td>88.90</td>
<td>90.53</td>
<td>93.41</td>
<td>92.75</td>
<td>76.18</td>
</tr>
<tr>
<td>magic</td>
<td>76.49</td>
<td>77.93</td>
<td>77.64</td>
<td>78.05</td>
<td>80.79</td>
<td>79.68</td>
<td>65.23</td>
</tr>
<tr>
<td>musk</td>
<td>83.73</td>
<td>82.34</td>
<td>81.95</td>
<td>89.55</td>
<td>96.18</td>
<td>95.35</td>
<td>84.55</td>
</tr>
<tr>
<td>nomao</td>
<td>89.45</td>
<td>91.43</td>
<td>91.37</td>
<td>89.41</td>
<td>92.32</td>
<td>92.07</td>
<td>69.40</td>
</tr>
<tr>
<td>orange-fraud</td>
<td>76.70</td>
<td>81.74</td>
<td>74.26</td>
<td>89.15</td>
<td>90.66</td>
<td>90.48</td>
<td>63.75</td>
</tr>
<tr>
<td>ozone-onehr</td>
<td>92.90</td>
<td>94.26</td>
<td>95.06</td>
<td>96.61</td>
<td>96.83</td>
<td>96.89</td>
<td>97.11</td>
</tr>
<tr>
<td>qsar-biodegradation</td>
<td>80.98</td>
<td>82.62</td>
<td>83.53</td>
<td>80.34</td>
<td>82.76</td>
<td>82.40</td>
<td>66.35</td>
</tr>
<tr>
<td>seismic-bumps</td>
<td>90.87</td>
<td>92.59</td>
<td>92.14</td>
<td>92.48</td>
<td>92.92</td>
<td>93.02</td>
<td>93.41</td>
</tr>
<tr>
<td>skin-segmentation</td>
<td>77.05</td>
<td>82.69</td>
<td>83.21</td>
<td>91.51</td>
<td>95.70</td>
<td>95.77</td>
<td>71.51</td>
</tr>
<tr>
<td>statlog-german-credit</td>
<td>70.76</td>
<td>72.12</td>
<td>72.34</td>
<td>72.25</td>
<td>72.93</td>
<td>72.78</td>
<td>70.00</td>
</tr>
<tr>
<td>thoracic-surgery</td>
<td>83.76</td>
<td>83.93</td>
<td>82.72</td>
<td>83.51</td>
<td>84.41</td>
<td>84.18</td>
<td>85.11</td>
</tr>
<tr>
<td>thyroid-hypothyroid</td>
<td>97.21</td>
<td>97.99</td>
<td>97.97</td>
<td>97.75</td>
<td>98.77</td>
<td>98.71</td>
<td>95.43</td>
</tr>
<tr>
<td>wilt</td>
<td>93.53</td>
<td>95.18</td>
<td>92.87</td>
<td>94.86</td>
<td>97.23</td>
<td>97.02</td>
<td>94.67</td>
</tr>
<tr>
<td>zebra</td>
<td>86.40</td>
<td>90.31</td>
<td>91.36</td>
<td>94.71</td>
<td>95.54</td>
<td>95.25</td>
<td>95.42</td>
</tr>
<tr>
<td>Moyenne</td>
<td>80.51</td>
<td>82.65</td>
<td>82.08</td>
<td>87.12</td>
<td>88.45</td>
<td>88.08</td>
<td>79.53</td>
</tr>
</tbody>
</table>

- **the choice of model is decisive**
- **using margin sampling with this model allows a significant performance improvement.**
- **LAL: a good active learning strategy has been learned**
- **but the learned strategy is no**
- **better than margin sampling**
- **and not always better than random**
- **hard to beat the majority vote in case of very imbalanced problems**
Learning Curves

- Plot performance against the number of requested labels
- Expected behavior:
  - Performance increases with number of labels
  - Convergence: after \( \infty \) label requests, all strategies should have the same performance
Evaluation

How to interpret the results of a learning curve?
• converging as fast as possible
• converging to the highest overall value

How to summarize results from a learning curve?
• Table at specific time points (early, mid, late)
• Area under the learning curve, mean (depends on stopping point)
• deficiency
• data utilization rate
• comparison of score differences
Area Under the Learning Curve (AULC)

- AULC above that of a random-sampling learner
- Calculated for maximum budget, thus sensitive to budget
- Negative value indicates worse-than-random performance
- Note: all strategies should pass through the same |L| values

“Active learning to maximize area under the roc curve”, Matt Culver, Deng Kun, and Stephen Scott, in Sixth International Conference on Data Mining (ICDM’06)
Deficiency


**Data Utilization Rate (DUR)**

- The minimum number of samples needed to reach a target accuracy, divided by the number of samples needed by a random sampling learner
- Indication of efficiency for selecting of data
- Sensitive to choice of target accuracy, ignores performance changes at other points

“Active learning to maximize area under the roc curve”, Matt Culver, Deng Kun, and Stephen Scott, In Sixth International Conference on Data Mining (ICDM’06),

![Graph showing performance vs. number of labels](image)
Evaluation

The evaluation methodology should be
1. reliable
   - robust to varying seeds or shuffling data
   - reproducible (well-described, availability of data)
2. realistic
   - valid assumptions for real applications
3. comparable
   - development of a standardized active learning evaluation gold standard to compare algorithms without reimplementing

How many repetitions are required?

Comparison of algorithms using 5-fold cross validation

- Which values to compare?
  - not across label acquisitions (highly correlated) but across multiple repetitions
  - at which point in time?
- Statistical tests
  - t-Test cmp. mean (assumes that mean is normal distributed)
  - Wilcoxon Signed Rank Test cmp. tendency (parameter-free test)
- always present results with statistical significance and effect size
Parameters

• tuning instances should be considered in the number of acquisitions
• how many instances should be used for tuning? (many classifiers are sensitive to the number of instances)
• normally, no instances for supervised parameter tuning available
• tuning parallel to sampling may be complicated
• no test set!
"Learning with few examples: an empirical study on leading classifiers", Christophe Salperwyck and Vincent Lemaire, in International Joint Conference on Neural Networks (IJCNN), 2011
Real applications often are more challenging

- Often highly specialized (hard to transfer approaches to related domains)
- Imperfect labelers (experts might be wrong)
- In real-world only one shot (mean results are not representative)
- Labels are not always available (in time and space)
- Performance guarantees (cmp. random sampling)
- Assess online performance of an actively trained classifier
- Different costs for different annotations or classes
- Impossible to tune the ‘user’ parameters of the classifier
- Ground truth might not be available (no test set…)

Thanks

• Daniel Kottke (University of Kassel, Germany) – slides AL
• Georg Kreml (University of Utrecht, Netherlands) – slides AL
• Alexis Bondu (Orange Labs, France) – beta test of this talk
• Antoine Cornuéjols (AgroParisTech, France) – beta test of this talk and discussion to create the “weakly supervised taxonomy”
• Pierre Nodet (Orange Labs, France) – beta test of this talk and discussion to create the “weakly supervised taxonomy”
• Bruno Kauffman (Orange Labs, France) – beta test of this talk