Active Learning with Rationales

Mustafa Bilgic

*In collaboration with my PhD students*
Bio

• Education
  • BS, University of Texas at Austin, 2000
    • Recommender systems and explanations
    • w/ Raymond Mooney
  • PhD, University of Maryland at College Park, 2010
    • Active learning and statistical relational learning
    • w/ Lise Getoor

• Currently
  • Associate Professor of Computer Science
  • Director of the Machine Learning Laboratory
  • Director of the Masters in Artificial Intelligence Program
Research Interests

- Machine learning
- Probabilistic graphical models
- Recommender systems
- Active machine learning
Recent Projects

- Active learning
- Active inference
- Learning with rationales
- Filter bubbles in news recommender systems
- Deep learning for biological image analysis
- Active evaluation
- Human-like classification
Recent Projects

• Active learning
• Active inference
• Learning with rationales
• Filter bubbles in news recommender systems
• Deep learning for biological image analysis
• Active evaluation
• Human-like classification
Machine Learning Background
Machine Learning

- Supervised learning
- Unsupervised learning
- Reinforcement learning
Machine Learning

- Supervised learning
- Unsupervised learning
- Reinforcement learning
Supervised learning

- Train a predictive model on instances of data
- The model is a function that maps data to a target $f: X \rightarrow Y$
- Humans provide the *supervision* on instances

Examples:
- **Products** → "rating"
- **Credit card transactions** → "fraud or legitimate"
- **Speech recording** → "transcription of speech"
1. Feature-based representation

   • Each instance is a vector of features
     • A patient: symptoms, laboratory test results, ...
     • A loan application: income, credit score, ...
     • A document: a bag of words
     • An image: scale-invariant feature transform (SIFT)
     • Gene sequence: n-grams, ...

   • Vector-based classifiers
     • Naïve Bayes, logistic regression, decision trees, support vector machines, neural networks, ...
The $X$ in $f:X \rightarrow Y$

2. Similarity-based representation
   • Pair-wise similarity among the instances
     • How similar are these images, documents, gene sequences, ...?
   • Similarity-based classifiers
     • Nearest neighbor, support vector machines
**The X in f:X→Y**

3. Image, text, sequence, “raw” data
   - Let the classifier learn the “features”
   - Neural networks with several hidden layers
     - a.k.a. deep learning
   - Examples
     - Convolutional neural networks for image analysis
     - Long Short-Term Memory networks for text analysis
The $Y$ in $f: X \rightarrow Y$

- The target variable
  - Patients: the diagnosis
  - Loan application: the decision
  - Document: the category
  - Image: the person

- Often, it is hard to obtain, because it might require
  - Expertise
  - Manual labor
  - Laboratory tests
Bias-Variance Trade-off

• The more assumptions a model makes, the less data it needs
  • Naïve Bayes typically requires less data than logistic regression
• The fewer the assumptions a model makes, the more data it needs
  • Deep learning with millions of parameters
  • GPT-3 has 175 billion parameters
1. Active Learning
Active Learning

• The $X$ is plenty; the $Y$ is scarce

• $X; Y$
  • Images; annotations
  • Speech; transcription
  • Text; translation
  • Review; sentiment
  • News; category
  • ...

How to choose few, but useful instances for labeling?

Active Learning (AL)
Active learning algorithm

Training Data
Features

<table>
<thead>
<tr>
<th>f_1</th>
<th>f_2</th>
<th>f_3</th>
<th>...</th>
<th>...</th>
<th>...</th>
<th>f_n</th>
<th>labels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Active Learner

Analysis

Unlabeled Instances

HOW?

instance: x^*

Unlabeled Instance

instance: x^*

Label: y^*

Predictive Model

Human Expert

Training Data

f_1 f_2 f_3 ... ... ... f_n labels

instance: x^*

Label: y^*
Active learning strategies

• Common utility-based active learning algorithms:
  • Query-by-Committee [Seung, Opper & Sompolinsky, COLT’92]
  • Uncertainty Sampling [Lewis & Gale, SIGIR’94]
  • Variance Reduction [Cohn, Ghahramani & Jordan, JAIR’96]
  • Bias Reduction [Cohn, NIPS’97]
  • Expected Error Reduction [Roy & McCallum, ICML’01]
  • And many more...
Ask the learner “why”

- Training Data
  - Features
    - $f_1$, $f_2$, $f_3$, ..., $f_n$
  - Labels
  - Unlabeled Instances

- Active Learner
  - Trains
  - Predictive Model

- Analysis

- Instance: $x^*$
  - Ask the learner explain why it chose $x^*$
  - Label: $y^*$

- Human Expert
Uncertainty sampling

• Queries instances about which the classifier is most uncertain how to label

• E.g., entropy as an uncertainty measure

\[ x^* = \arg\max_{x^{(i)} \in \mathcal{U}} - \sum_{y \in Y} P_\theta(y|x^{(i)}) \log (P_\theta(y|x^{(i)})) \]

Ask the learner why it is uncertain about \( x^* \)
Evidence-based framework

We discovered two reasons for model’s uncertainty on instances

Conflicting-evidence uncertainty:

Insufficient-evidence uncertainty:

Traditional uncertainty sampling:
Does not consider the reasons for uncertainty, as long as $E_{-1}(X) \approx E_{+1}(X)$
Datasets & measures

Eight datasets:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Domain</th>
<th>Size</th>
<th>Minority class %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spambase</td>
<td>Email. classification</td>
<td>4,601</td>
<td>39.4%</td>
</tr>
<tr>
<td>Ibn Sina</td>
<td>Handwriting recognition</td>
<td>20,722</td>
<td>37.8%</td>
</tr>
<tr>
<td>Calif. Housing</td>
<td>Social</td>
<td>20,640</td>
<td>29%</td>
</tr>
<tr>
<td>Nova</td>
<td>Text processing</td>
<td>19466</td>
<td>28.4%</td>
</tr>
<tr>
<td>Sick</td>
<td>Medical</td>
<td>3,772</td>
<td>6.1%</td>
</tr>
<tr>
<td>Zebra</td>
<td>Embryology</td>
<td>61,488</td>
<td>4.6%</td>
</tr>
<tr>
<td>LetterO</td>
<td>Letter recog.</td>
<td>20,000</td>
<td>4%</td>
</tr>
<tr>
<td>Hiva</td>
<td>Chemo-inform.</td>
<td>42,678</td>
<td>3.5%</td>
</tr>
</tbody>
</table>

Performance measures:

- **AUC**: All datasets
- **Accuracy**: Medium-imbalanced datasets
- **F1**: Highly-imbalanced datasets
How to interpret the results?

Performance measure, e.g., Accuracy, AUC, F1

The higher the better!

E.g., Accuracy, AUC, F1

Budget, e.g., Number of instances
Results – Ibn Sina dataset

AUC vs Number of Instances

- Random sampling
- Traditional uncertainty sampling
- Conflicting-evidence
- Insufficient-evidence

77% savings
2. Learning with Rationales
Ask the humans “why”

Ask human to provide a rationale for her classification of documents

Human Expert

Instance: x*
Label: y*
Rationale: R*

Active Learner

Training Data
Features

Unlabeled Instances

instance: x*

Rationale: R*

Predictive Model

Analysis

Label: y*

f₁ f₂ f₃ ... ... ... fₙ labels
2. Learning with Rationales
2.a. Text Classification
The approach

Document 1
This is a **great** movie.

Document 2
The plot was great, but the performance of actors was **terrible**. **Avoid it.**

Document 3
I’ve seen this at an outdoor cinema; great atmosphere. The movie was **terrific**.

How do we use \(<x, y, r>\) for supervised learning?
Datasets & experimental setup

Four text classification datasets:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th># instances</th>
<th># Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMDB</td>
<td>Sentiment analysis of movie reviews</td>
<td>25,000</td>
<td>27,272</td>
</tr>
<tr>
<td>NOVA</td>
<td>20 Newsgroups dataset: Email classification</td>
<td>12,977</td>
<td>16.969</td>
</tr>
<tr>
<td>SRAA</td>
<td>UseNet articles: Aviation vs. Auto</td>
<td>48,812</td>
<td>31,883</td>
</tr>
<tr>
<td>WvsH</td>
<td>20 Newsgroups dataset: Windows vs. Hardware</td>
<td>1176</td>
<td>4,026</td>
</tr>
</tbody>
</table>

Three classifiers:
- Multinomial naïve Bayes (MNB)
- Logistic regression (LR)
- Support vector machines (SVM)

Two data representations:
- Binary
- Tf-idf
Results – SRAA dataset with MNB

- AUC vs. Number of documents graph
- Learning with rationales + labels (tfidf)
- Learning with labels (tfidf)

88% savings
2. Learning with Rationales
2.b. Anomalous Flight Detection
Collaboration w/ NASA

GOAL: **effectively** train a model to identify operationally significant (OS) anomalies using **less time** of experts.

**OS**: operationally significant
**NOS**: not operationally significant

MKAD (Multiple Kernel Anomaly Detection)

Unsupervised learning

Flights data

Subject Matter Expert

Statistical anomalies
Flights data

ORIGINAL FEATURES
• Latitude
• Longitude
• Altitude
• Ground speed
• Horizontal separation
• Vertical separation
• Aircraft size
• Turn-to-final (TTF) parameters:
  • Maximum overshoot
  • Speed at TTF
  • Distance at TTF
  • Angle at TTF
  • Altitude difference at TTF
• Nearest neighboring (NN) flight info:
  • NN flight on same runway
  • NN flight on parallel runway
  • NN flight part of the same flow
Rationales

“Loss of separation”
- Horizontal separation < 3 miles AND
  Vertical separation < 1000 ft AND nearest neighboring flight is not on parallel runways and not part of the same flow

“Large overshoot”
- Maximum overshoot is greater than a threshold based on values of flights with positive labels

“Unusual flight path”
- Overall deviation from expected (average) trajectory of all landing flights on that runway

More complex rationales
Active learning framework

Features

<table>
<thead>
<tr>
<th>Flights</th>
<th>f_1</th>
<th>f_2</th>
<th>f_3</th>
<th>\ldots</th>
<th>f_{n-1}</th>
<th>f_n</th>
</tr>
</thead>
<tbody>
<tr>
<td>flight 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>flight 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\ldots</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>flight m</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

labels

Active Learner

Unlabeled flights

Analysis

Predictive Model

Subject Matter Expert

flight: x*

Label: y*

rationale

Loss of separation

OS
Selecting informative flights

Active learning strategy: Most-likely positive strategy

Objective function: \( \mathbf{x}^* = \arg \max_{\mathbf{x} \in \mathcal{U}} P_\theta(\hat{\mathbf{y}}^+ | \mathbf{x}) \)
Including rationales into learning

Including rationales improves performance over learning with labels only

**Precision@5**

Number of labeled flights

- **Most-Likely Positive**
- **Most-Likely Positive (with rationales)**
3. Fresh, still in the oven, projects
Projects currently in the oven

- Active evaluation
  - Curate a dataset for only evaluation purposes
- Human-like classification
  - Given a case, skim all features but focus on what is most important for that case
Collaboration Opportunities

• We develop methods
  • Active learning, learning with rationales, active evaluation, human-like classification, etc.

• Collaboration opportunities
  1. Application areas
     • If you have problems/datasets where these methods might be applicable (not enough labeled data, experts provide rationales, human-like and interpretable decision making, etc.), I’d be very happy to discuss them and work with you
       AND/OR
  2. Foundational work
     • If you also work on these areas, I’d be very happy to talk to you about potential collaboration opportunities

• Please see the next slide for my contact info
Contact

Email: mbilgic@iit.edu
Lab: http://ml.cs.iit.edu
Thank you