Estimating Accuracy from Unlabeled Data

A Probabilistic Logic Approach

Emmanouil Antonios Platanios e.a.platanios@cs.cmu.edu
Hoifung Poon hoifung@microsoft.com
Tom Mitchell tom.mitchell@cs.cmu.edu
Eric Horvitz horvitz@microsoft.com

Problem

Using unlabeled data we can measure **consistency** but not **correctness**

Consistency

Does this implication hold? ↓ If yes, under what conditions?

Correctness

SETTING

There exist **binary functions** that we do not know:

"Pittsburgh" $X \sim \mathcal{D}$ $\xrightarrow{f_1}$ Is city? $\xrightarrow{f_2}$ Is fish? $\xrightarrow{f_3}$ Is bird?

Instead, we have a set of approximations (i.e., classifiers) to those functions and **we want to know how accurate they are:**

- $\hat{f}_1^1$ Using orthographic features *(e.g., ends with "-burgh")*
- $\hat{f}_2^1$ Using the context *(e.g., appears after "lives in")*
- ...

LOGICAL CONSTRAINTS

The function outputs may be constrained:

Is bird? ---- **MUTUAL EXCLUSION** ---- Is fish?
If something is a bird it cannot also be a fish

Is bird? —— **SUBSUMPTION** → Is animal?
If something is a bird it must also be an animal

Can we use this information to better predict the accuracy of the classifiers?

Intuitive example:
If classifier 1 says that **X** is a bird and classifier 2 says that **X** is a fish, then at least one of them **must be wrong**

Approach

CLASSIFIER OUTPUTS

"Shark"	Classifier 1	Is animal?	99%
		Is fish?	95%
		Is bird?	5%
	Classifier 2	Is animal?	98%
		Is fish?	93%
		Is bird?	7%
	...		
"Sparrow"	Classifier 1	Is animal?	95%
		Is fish?	10%
		Is bird?	26%
	Classifier 2	Is animal?	95%
		Is fish?	2%
		Is bird?	84%
	...		

LOGICAL CONSTRAINTS

Is animal?
Is fish? ------- Is bird?
...

GROUNDING

Inputs: Predicted probability for each *classifier-sample-concept*

Outputs: Set of *sample-concept* classification pairs and *concept-classifier* error-rate pairs that are not directly constrained to be 0 or 1 from the constraints

GROUND PREDICATES

Observed	Unobserved	
$\hat{f}_1^{\texttt{animal}}(shark) = 0.99$	$e_1^{\texttt{animal}}$	Error rates of classifier 1 for various concepts
$\hat{f}_1^{\texttt{fish}}(shark) = 0.95$	$e_1^{\texttt{fish}}$	
$\hat{f}_1^{\texttt{bird}}(shark) = 0.05$	$e_1^{\texttt{bird}}$	
$\hat{f}_1^{\texttt{animal}}(sparrow) = 0.95$	...	
$\hat{f}_1^{\texttt{fish}}(sparrow) = 0.10$	$f^{\texttt{animal}}(shark)$	Final (combined) predictions aggregated across classifiers
$\hat{f}_1^{\texttt{bird}}(sparrow) = 0.26$	$f^{\texttt{fish}}(shark)$	
...	$f^{\texttt{bird}}(shark)$	
$\mathrm{SUB}(\texttt{animal},\texttt{fish}) = 1$	$f^{\texttt{animal}}(sparrow)$	
$\mathrm{SUB}(\texttt{animal},\texttt{bird}) = 1$	$f^{\texttt{fish}}(sparrow)$	
$\mathrm{ME}(\texttt{fish},\texttt{bird}) = 1$	$f^{\texttt{bird}}(sparrow)$	

GROUND RULES

$$\mathrm{SUB}(\texttt{animal},\texttt{fish}) \wedge \neg\hat{f}_1^{\texttt{animal}}(shark) \wedge f^{\texttt{fish}}(shark) \rightarrow e_1^{\texttt{animal}}$$
$$\dots$$
$$\mathrm{ME}(\texttt{fish},\texttt{bird}) \wedge \hat{f}_1^{\texttt{fish}}(sparrow) \wedge f^{\texttt{bird}}(sparrow) \rightarrow e_1^{\texttt{fish}}$$

PROBABILISTIC INFERENCE

Inputs: Ground predicates and rules

Step 1: Create a Markov Random Field (MRF)
Step 2: Perform probabilistic inference to obtain the most likely values for the unobserved ground predicates. Inference is performed using a modified version of the Probabilistic Soft Logic (PSL) framework that uses a custom stochastic version of the consensus Alternating Direction Method of Multipliers (ADMM).

RESULTS

	ERROR RATES		PREDICTIONS		
Classifier 1	Is animal?	1%	Shark	Is animal?	1%
	Is fish?	5%		Is fish?	5%
	Is bird?	57%		Is bird?	57%
Classifier 2	Is animal?	1%	Sparrow	Is animal?	1%
	Is fish?	2%		Is fish?	2%
	Is bird?	9%		Is bird?	9%

LOGICAL RULES

Ensemble	Identifiability	Mutual Exclusion
$\hat{f}_j^d(X) \wedge \neg e_j^d \rightarrow f^d(X)$	$\hat{f}_j^d(X) \rightarrow f^d(X)$	$\mathrm{ME}(d_1,d_2) \wedge \hat{f}_j^{d_1}(X) \wedge f^{d_2}(X) \rightarrow e_j^{d_1}$
$\neg\hat{f}_j^d(X) \wedge \neg e_j^d \rightarrow \neg f^d(X)$	$\neg\hat{f}_j^d(X) \rightarrow \neg f^d(X)$	**Subsumption**
$\hat{f}_j^d(X) \wedge e_j^d \rightarrow \neg f^d(X)$		$\mathrm{SUB}(d_1,d_2) \wedge \neg\hat{f}_j^{d_1}(X) \wedge f^{d_2}(X) \rightarrow e_j^{d_1}$
$\neg\hat{f}_j^d(X) \wedge e_j^d \rightarrow f^d(X)$		

Needed because replacing all error rates with 1 minus their value would result in the ensemble rules still being satisfied

Intuitively: If two approximations produce the same prediction, then that prediction is more likely to be correct → **consistency implies correctness**

Experiments

DATASETS

1. Never-Ending Language Learning (NELL)

Task: *Predict whether a noun phrase (NP) represents a concept (e.g., "city").*

Classifiers: 6 logistic regression classifiers using different features.

Constraints:

Animal
Vertebrate ---------- Invertebrate
Bird --- Fish --- Mammal   Arthropod --- Mollusk

Location
City ------- Country   River ------- Lake

Size: 553,940 NPs | 11 concepts or 7 concepts (leafs of the above tree)

We also experimented with an unconstrained version, called "uNELL", which contained 20,000 NPs per concept, 15 concepts, and 4 classifiers. This is so we can compare our results to those of previous work.

2. Neuroscience

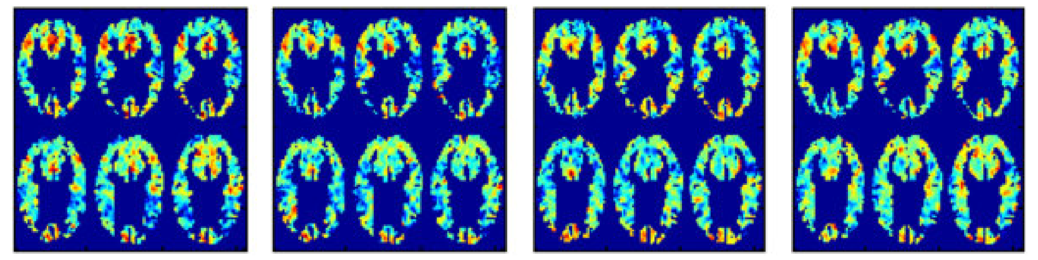
Task: *Predict which of two 40 second long story passages corresponds to an unlabeled 40 second time series of fMRI neural activity.*

Classifiers: 8 classifiers using different text features.

Size: 1,000 samples for **11 brain regions**

	Passage #1	Passage #2
	...They were hoping for a reason to fight Malfoy...	...Harry had heard Fred and George Wesley complain...
- Number of letters per word	...4 4 6 3 1 6 2 5 6...	...5 3 5 4 3 6 7 8...
- Parts of speech	...PRP VBD VBG IN DT NN TO VB NNP...	...NNP VBD VBN NNP CC NNP NNP VB...
- ...		

Which passage corresponds to this fMRI recording?



RESULTS

$\text{MAD}_{\text{error rank}}$ → Measures the correctness of the ranking of the classifiers based on correctness

$\text{MAD}_{\text{error}}$ → Measures the correctness of the error rate estimates of the classifiers

$\text{AUC}_{\text{target}}$ → Measures the accuracy of the final predictions.

	$\text{MAD}_{\text{error rank}}$	$\text{MAD}_{\text{error}}$	$\text{AUC}_{\text{target}}$	$\text{MAD}_{\text{error rank}}$	$\text{MAD}_{\text{error}}$	$\text{AUC}_{\text{target}}$
MAJ	7.71	0.238	0.372	7.54	0.303	0.447
AR-2	12.0	0.261	0.378	10.8	0.350	0.455
AR	11.4	0.260	0.374	11.1	0.350	0.477
BEE	6.00	0.231	0.314	5.69	0.291	0.368
CBEE	6.00	0.232	0.314	5.69	0.291	0.368
HCBEE	5.03	0.229	0.452	5.14	0.324	0.462
LEE	**3.71**	**0.152**	**0.508**	**4.77**	**0.180**	**0.615**

We also outperform the state-of-the-art in the unconstrained datasets