Active Learning Based on Fast Logistic Regression

Zhili Wu,[†] Serge Sharoff,[†] Katja Markert[‡]

Centre for Translation Studies,[†] School of Computing[‡]

{Z.Wu,S.Sharoff,K.Markert}@leeds.ac.uk

1. Highlights of Our Approach

1. Base Classifier: Fast Logistic Regression (Powered by Liblinear)
2. Probability Output of LR together with Random Queries
3. ALC and AUC Oriented Parameter Tuning
4. Applications to Text Genre Identification

2. Datasets for Development

Dataset	Domain	Feat. num.	Sparsity(%)	Train/Test num.
HIVA	Chemo-informatics	1617	90.88	21339
IBN_SINA	Handwriting recognition	92	80.67	10361
NOVA	Text Processing	16969	99.67	9733
ORANGE	Marketing	230	9.57	25000
SYLVA	Ecology	216	77.88	72626
ZEBRA	Embryology	154	0.04	30744

3. Datasets for Competition

Dataset	Feat. Type	Feat num.	Sparsity(%)	Missing(%)	Train num.	Test num.
A	mixed	92	79.02	0	17535	17535
B	mixed	250	46.89	25.76	25000	25000
C	mixed	851	8.6	0	25720	25720
D	binary	12000	99.67	0	10000	10000
E	continuous	154	0.04	0.0004	32252	32252
F	mixed	12	1.02	0	67628	67628

4. Performance and Rank of Our Team - "TEST" on Competition Sets

Dataset	AUC	Ebar	ALC	Rank
A	0.8831	0.0052	0.3472	11
B	0.6980	0.0044	0.3383	4
C	No Submission			
D	0.9623	0.0033	0.6576	4
E	0.7896	0.0044	0.4483	6
F	0.9796	0.0017	0.7007	7
Overall	Not Applicable			

5. First set of predictions given only one positive sample

1. A. Random predictions
2. B. Distances to the only sample
3. C. Clustering / One-class Classification (Costly, parameter-dependent)
4. D. Semi-supervised learning (Costly, hard to calibrate)

AUC of First Prediction

	ALEX	HIVA	IBN_SINA	NOVA	ORANGE	SYLVA	ZEBRA
A. Random	0.4953	**0.5106**	**0.5141**	**0.5073**	0.5033	0.5056	0.4924
B. Distance	**0.6366**	0.3820	0.1689	0.2606	0.5000	0.5000	0.5000

Observations: random guess at the beginning is not too bad.

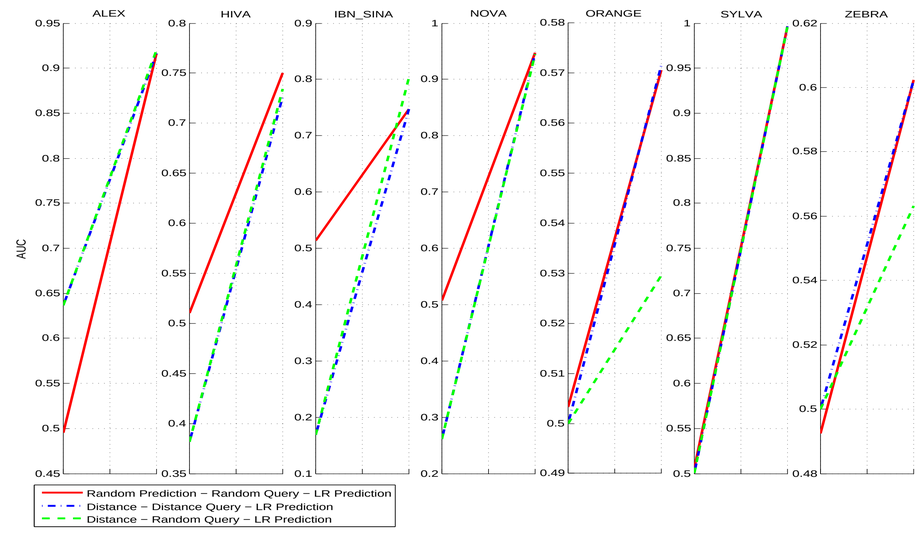
6. Subsequent Queries and Models

6.1 L_2-Logistic Regression Model on Queried Samples

$$\min_w \quad \frac{1}{2} w^T w + C \sum_{i=1}^{l} log(1 + e^{-y_i w^T x_i}).$$

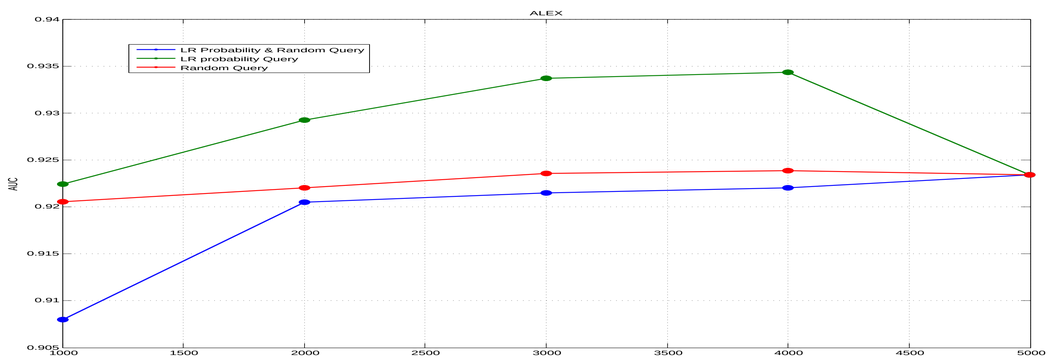
A fast implementation based on trust region Newton method is available in Liblinear.
Its probability outputs are used as uncertainty scores for query.
Parameter C is tuned to maximize the AUC score in 4-fold cross validation.

6.2 Initial Query and LR Model

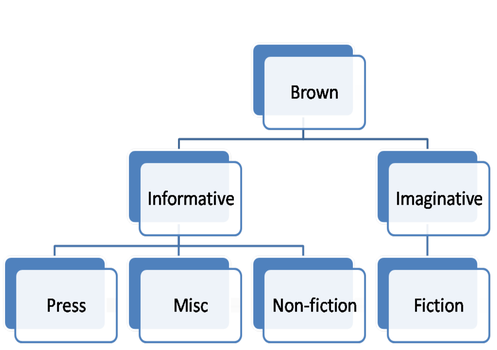


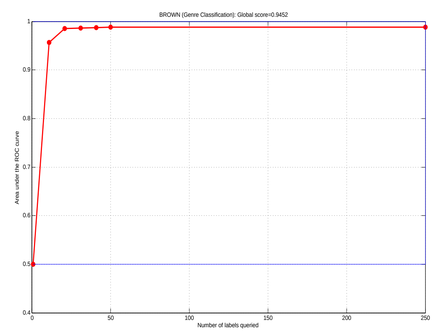
First of the five linear-step queries. Except for ALEX, the random guess with random query becomes a good baseline.

6.3 Combining Random Query and LR Probability in Query



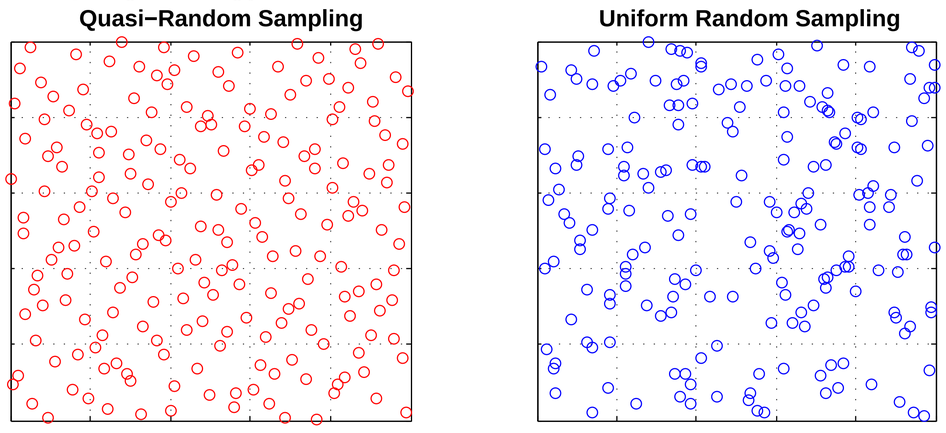
7. Applications to Text Genre Identification





8. Conclusions & Discussions

- **Fast LR, Random & LR Probability Queries**
- **Question 1:** For the first prediction and query, how to (efficiently) outperform random guess?
- **Question 2:** How to extend to multi-class/structural outputs?
- **Question 3** Submiting/retrieving queries on client machine (e.g. web service api, XML format) rather than checking webpage each time.
- Other sampling strategy?



9. Acknowledgements

We would like to thank the organizers of this competition. We are also grateful to the support of University of Leeds, and also Google Inc for their Google Research Awards programme.