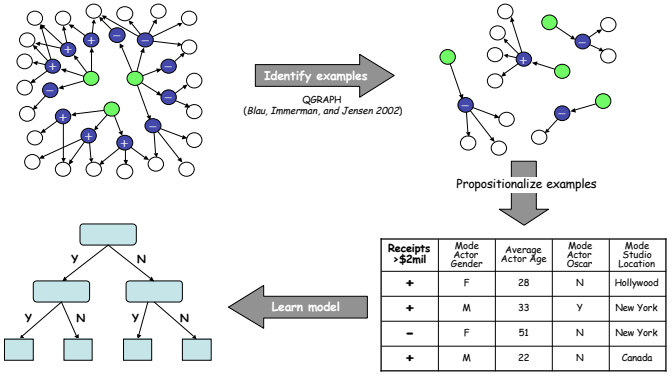
Learning Relational Probability Trees

Jennifer Neville, David Jensen, Lisa Friedland & Michael Hay
Knowledge Discovery Laboratory, University of Massachusetts Amherst

Extending Trees to a Relational Setting

- Heterogeneous data instances
  - Models need to consider relational neighborhoods which vary in size
  - Makes direct application of conventional techniques difficult
- Non-independent instances
  - Greatly complicates the statistics of both learning and inference
    - Jensen and Neville ICML2002, Jensen, Neville and Hay ICML2003



Receipts >$2mil	Mode Actor Gender	Average Actor Age	Mode Actor Oscar	Mode Studio Location
+	F	28	N	Hollywood
+	M	33	Y	New York
–	F	51	N	New York
+	M	22	N	Canada

Relational Probability Trees (RPTs)

Input

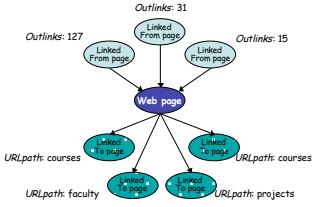
- Collection of subgraphs
- Each contains a single target object to be classified, other objects and links in subgraph form relational neighborhood

Output

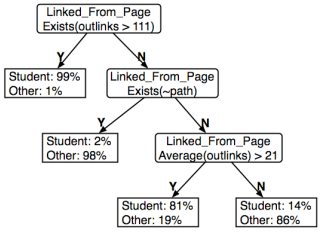
- RPT model: conditional probability distribution over target class label
- RPT represents a series of questions to ask about a subgraph

Learning Algorithm

- Recursive partitioning algorithm
- Searches binary relational feature space
  - Aggregation functions map a set of values into a single value
    - Avg/Mode, Count, Proportion, Degree
  - Chi-square feature scores measure association with class
- Bonferroni-adjusted p-value cutoff stops tree growth
- Randomization tests adjust for feature selection biases

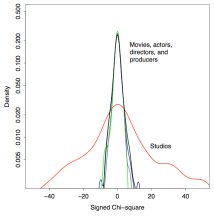


LinkedToPage: Mode(*URLpath*)=courses
LinkedFromPage: Count(outlinks>25)>1
LinkedFromPage: Degree>2

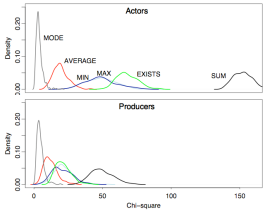


Feature Selection Biases

- Linkage and autocorrelation increase variance of feature scores
  - Increases probability of selecting random features



- Degree disparity and aggregation increase bias of feature scores
  - Increases probability of selecting surrogate degree features



- Novel randomization tests account for relational data characteristics and provide a method for accurate hypothesis testing
  - Retain relational structure (e.g. autocorrelation, degree disparity)
  - Randomize attribute values before aggregation

Empirical Evaluation

Four algorithms

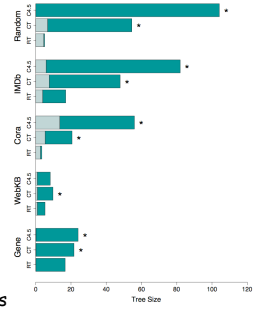
- Conventional test RPT (CT)
- Randomization test RPT (RT)
- C4.5 with flattened features used in RPT
- Relational Bayes classifier (RBC)

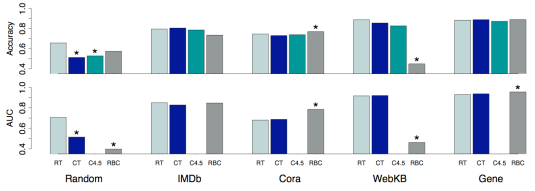
Five datasets

- Random IMDb, IMDb, Cora, WebKB, Gene

Performance measurements

- Accuracy, area under ROC curve (AUC)
- Tree size, weighted proportion of degree features





Conclusions

- RPTs built using randomization tests (RTs) are significantly smaller than other models and achieve equivalent, or better, performance
  - CTs and C4.5 select surrogates for degree and have unnecessary complexity
  - RBC models perform poorly when degree is only feature correlated with class

Acknowledgements

This research is supported by DARPA and NSF under contract numbers F30602-01-2-0566 and EIA9983215, respectively. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of DARPA, NSF, or the U.S. Government.