Knowledge Discovery in Relational Databases

Mandana Hamidi
Arezoo Rajabi
Sogol Balali

Prof: Arash Termehchy
Final project CS540, winter 2015

Overview:

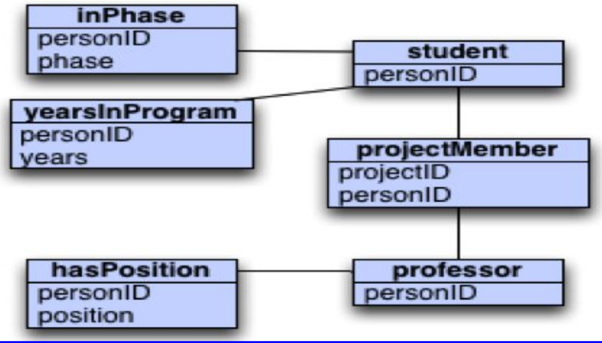
- Goal: Learning new concepts from structured database

- Why: Is an important problem with many applications in data management and machine learning.

- Solutions: Applying machine learning methods such as FOIL, TILDE, Mixture Model Membership..

Problem Statement:

Structured Data set (UW_CSE dataset)



Target concept:
advisedBy(student , professor)

Positive example:	**Negative Example:**
advisedBy(Jose, Arash)	advisedBy(Jose, Tom)
advisedBy(Vahid, Arash)	advisedBy(Arash, Jose)
…	…

Relational Learning Algorithm
(FOIL, TILDE, Mixture Model Membership)

New Concept (Definition)

advisedby(A,B): - publication(C,B) , publication(C,A).

"A student is advised by a professor if they have a common publication."

Dataset

UW-CSE Dataset

- Consists of information about the University of Washington Department of Computer Science and Engineering.
- Consists of 12 relations, 2673 tuples, and 113 positive examples

J.Picado, A.Termehchy, and A.Fern. Schema Independence of Relational Learning Algorithms, ACM SIGMOD Workshop on Big Uncertain Data, 2014.

Evaluation Criteria

- Accuracy:

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

- Precision: refers to as positive predictive value

$$\text{Precision} = \frac{tp}{tp + fp}$$

- Recall: refers to as the true positive rate or sensitivity

$$\text{Recall} = \frac{tp}{tp + fn}$$

- F-measure: harmonic mean of precision and recall

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Tp (True Positive): measures the proportion of actual positives which are correctly identified.
Tn (True Negative):measures the proportion of negatives which are correctly identified.
Fp (False Positive): indicates a given condition has been fulfilled, when it actually has not been fulfilled.
Fn (False Negative):indicates that a condition failed, while it actually was successful.

FOIL (First Order Inductive Learner)

- Top-down (general to specific) approach originally applied to first-order logic (Quinlan, 1990).
- A greedy algorithm that learns rules to distinguish positive examples from negative ones.
- Repeatedly searches for the current best rule and removes all the positive examples covered by the rule until all the positive examples in the data set are covered.

J.Picado, A.Termehchy, and A.Fern. Schema Independence of Relational Learning Algorithms, ACM SIGMOD Workshop on Big Uncertain Data, 2014.

FOIL (continue)

- Tries to maximize the gain of adding literal p to rule r
- P: the set of positive examples
- N: the set of negative examples
- When p is added to r, then there are P* positive and N* negative examples satisfying the new rule

$$gain(p) = |P^*| \left(\log \frac{|P^*|}{|P^*| + |N^*|} - \log \frac{|P|}{|P| + |N|} \right)$$

FOIL Results

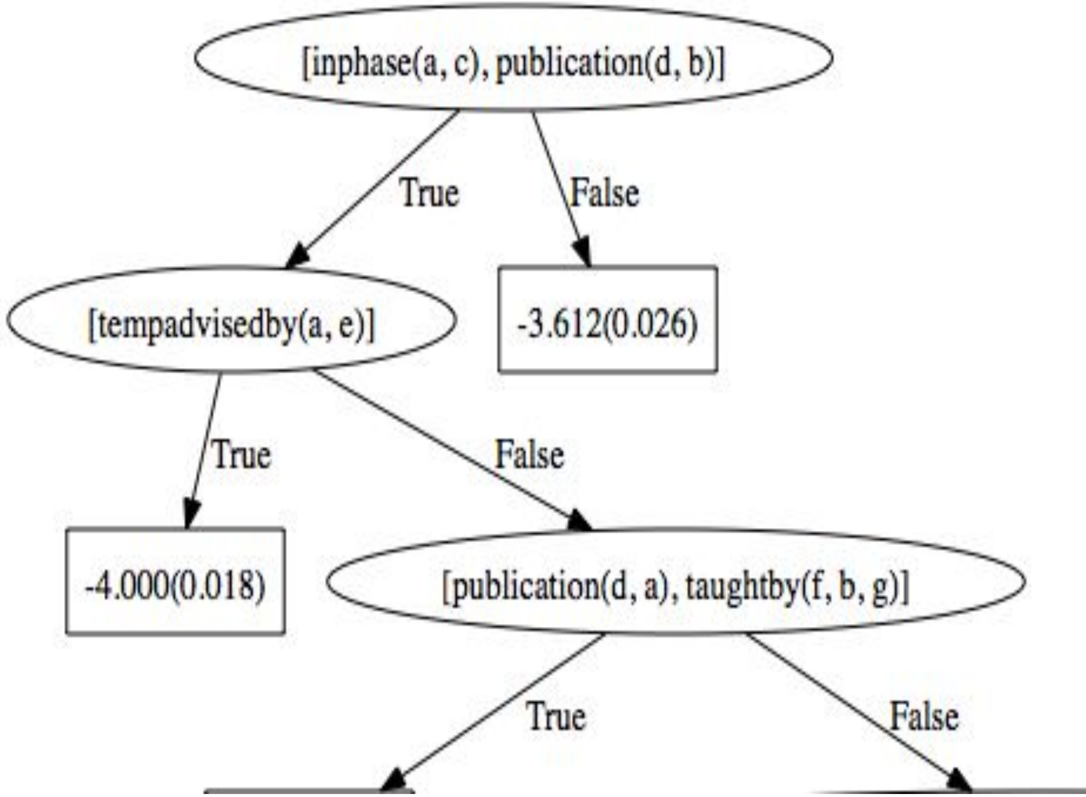
	Total clauses constructed	Accuracy	Precision	Recall	F-measure	Time Taken Train (seconds)
Train 1	90000	0.638	0.468	0.629	0.537	13.194
Train 2	170000	0.850	0.704	0.95	0.809	27.794
Train 3	110000	0.741	0.583	0.778	0.666	20.323
Train 4	90000	0.667	0.5	0.848	0.629	14.229
Train 5	390000	0.771	0.857	0.375	0.522	57.715

FOIL Results: learned rules

	Learned Rules	Pos Cover	Neg Cover
Train 1	rule 1: advisedby(A,B): - publication(C,B) , publication(C,A). rule 2: advisedby(A,B): -inphase(A, post_quals),publication(C,B). rule 4:advisedby(A,B):-inphase(A, post_generals), ta(C,A,D), publication(E,B)	rule1 :27 rule2: 29 rule4:18	rule 1: 1 rule2: 13 rule4: 9
Train 2	rule 1: advisedby(A,B): - hasposition(B,faculty), inphase(A, post_generals) rule 2: advisedby(A,B): -inphase(A, post_quals),publication(C,B),publication(D,B), diff(D,C).	rule1: 40 rule2: 29	rule1: 36 rule2: 14
Train 3	rule1: advisedby(A,B): -inphase(A, post_generals). rule2: advisedby(A,B): -inphase(A, post_quals),publication(C,B).	rule1: 51 rule2: 35	rule1: 49 rule2: 25
Train 4	rule1: advisedby(A,B): -inphase(A, post_generals). rule 2: advisedby(A,B): -inphase(A, post_quals),publication(C,B),publication(D,B), diff(D,C).	rule1: 38 rule2: 26	rule1: 32 rule2: 19
Train 5	rule 2: advisedby(A,B): -inphase(A, post_generals), ta(C,A,D). rule 6: advisedby(A,B): -yearsinprogram(A, year_4), publication(C,A) rule 8: advisedby(A,B): -inphase(A, post_generals), publication(C,B),publication(C,A)	rule 2: 19 rule 6: 14 rule 8: 30	rule 2: 16 rule 6: 7 rule 8: 2

TILDE (Top-down Induction of Logical Decision Trees):

- A first order logic extension of the C4.5 decision tree algorithm.
- **Internal nodes**: of the tree are logical predicates
- **Leaf nodes**: regression values



H. Blockeel and L. De Raedt. Top-down induction of first-order logical decision trees. *Artif. Intell.*, 101 (1-2):285–297, May 1998. ISSN 0004-3702.

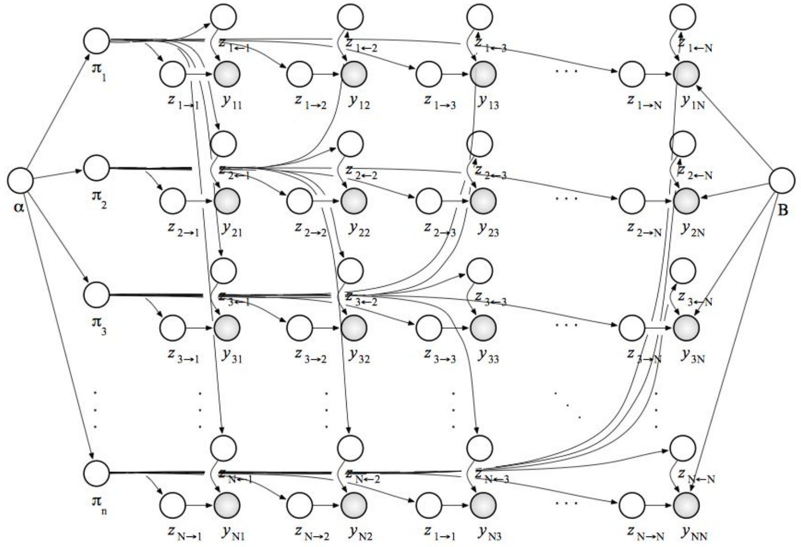
Example of leanred tree by TILDE

TILDE Results

	Total clauses constructed	Accuracy	Precision	Recall	F-measure	Time Taken Train
Train1	90000	0.8667	1.00	0.685714	0.813559	13.620 s
Train2	170000	0.7729	0.704	0.550000	0.709677	15 m, 54.916 s
Train3	110000	0.9267	1.00	0.777778	0.875000	27 m,13.039 s
Train4	90000	0.7232	0.575000	0.696970	0.630137	5 m, 11.928 s
Train5	390000	0.7462	0.733333	0.375	0.709677	16 m, 5.381s

Mixed Membership Blockmodel

- A blockmodel for group detection in graphs
- Overlapping between groups is allowed
- Number of K is required



Airoldi, Edoardo M., David M. Blei, Stephen E. Fienberg, and Eric P. Xing. "Mixed membership stochastic blockmodels." In *Advances in Neural Information Processing Systems*, pp. 33-40. 2009.

Mixed Membership Blockmodel

Matrix definition:

- It’s defined based on similarities among student
- Phase, Year, # Courses with same Prof., # papers with same Prof.
- Probability of existence of a link between two nodes $\propto$ Similarity among them

Mixed Membership Blockmodel Results

	Precision	Recall	F-measure	Time Taken Train
Train1	0.17	0.88	0.29	<1sec
Train2	0.23	0.76	0.35	< 1sec
Train3	0.3	0.6	0.4	< 1sec
Train4	0.17	0.87	0.29	<1sec
Train5	0.23	0.84	0.44	<1sec

Mixed Membership Blockmodel

Why it does not work well:

- It's unsupervised method
- The used matrices are unweighted and constructed by a probability proportional to the nodes' similarities
- We did not use any training dataset

Summary

- In this project we compared three different machine learning algorithms for learning relational concepts on relational database.
- TILDE algorithm performs better than FOIL, because it constructs very complicated trees.
- Mixed membership blockmodel was introduced as an unsupervised method to find relevant objects.

Question?

Thanks

FOIL Algorithm

Basic algorithm for instances with discrete-valued features:

```
Let $A$={} (set of rule antecedents)
Let $N$ be the set of negative examples
Let $P$ the current set of uncovered positive examples
Until $N$ is empty do
    For every feature-value pair (literal) ($F_i$=$V_{ij}$) calculate
        Gain($F_i$=$V_{ij}$, $P$, $N$)
    Pick literal, $L$, with highest gain.
    Add $L$ to $A$.
    Remove from $N$ any examples that do not satisfy $L$.
    Remove from $P$ any examples that do not satisfy $L$.
Return the rule: $A_1$ Ù$A_2$ Ù… Ù$A_n$ → Positive
```