Positive and Unlabeled Relational Classification through Label Frequency Estimation

Jessa Bekker and Jesse Davis

Computer Science Department, KU Leuven, Belgium
firstname.lastname@cs.kuleuven.be

Abstract. Many applications, such as knowledge base completion and patient data, only have access to positive examples but lack negative examples which are required by standard ILP techniques and suffer under the closed-world assumption. The corresponding propositional problem is known as Positive and Unlabeled (PU) learning. In this field, it is known that using the label frequency (the fraction of true positive examples that are labeled) makes learning easier. This notion has not been explored yet in the relational domain. The goal of this work is twofold: 1) to explore if using the label frequency would also be useful when working with relational data and 2) to propose a method for estimating the label frequency from relational PU data. Our experiments confirm the usefulness of knowing the label frequency and of our estimate.

1 Introduction

ILP traditionally requires positive and negative examples to learn a theory. However, in many applications, it is only possible to acquire positive examples. A common solution to this issue is to make the closed-world assumption and assume that unlabeled examples belong to the negative class. In reality, this assumption is often incorrect, for example: diabetics often go undiagnosed (Claesen et al., 2015) and people do not bookmark all interesting pages. Considering unlabeled cases as negative is therefore suboptimal. To cope with this, several score functions have been proposed that use only positive examples (Muggleton, 1996; McCreath and Sharma; Schoenmackers et al., 2010).

In propositional settings, having training data with only positive and unlabeled examples is known as Positive and Unlabeled (PU) learning. It has been noted that if the class distribution is known, then learning in this setting is greatly simplified. Specifically, knowing the class prior allows calculating the label frequency, which is the probability of a positive example being labeled. The label frequency is crucial as it enables converting standard score functions into PU score functions that can incorporate information about the unlabeled data (Elkan and Noto, 2008). Following this insight, several methods have been proposed to estimate the label frequency from PU data (du Plessis et al., 2015; Jain et al., 2016; Ramaswamy et al., 2016). As far as we know, this notion has not been exploited in relational settings. We propose a method to estimate the

label frequency from relational PU data and a way to use this frequency when learning a relational classifier.

Our work goes beyond the existing ILP score functions that only use positive examples to incorporate unlabeled examples in the model evaluation process. Our main contributions are: 1) Investigating the helpfulness of the label frequency in relational PU learning by adjusting relational decision trees to incorporate the label frequency; 2) Proposing a method for estimating the label frequency in relational data through tree induction; and 3) Evaluating our approach experimentally.

2 PU Learning and Using the Label Frequency

In PU Learning, the labeled positive examples are commonly assumed to be 'selected completely at random' (Elkan and Noto, 2008). This means that the probability $c = \Pr(s = 1|y = 1)$ for a positive example to be labeled is constant: equal for every positive example. c is called the *label frequency*. Many propositional PU learners exploit c to simplify the learning (Denis et al., 2005; Zhang and Lee, 2005). To the best of our knowledge, using the label frequency in relational PU learning has not been investigated yet. We briefly review some of the propositional methods that can be adjusted to the relational domain.

The most basic method is to learn a probabilistic model which considers unlabeled as negative and adjust the predicted probability: $\Pr(y = 1|x) = \frac{1}{c}\Pr(s = 1|x)$ (Zhang and Lee, 2005). In the experiments, we apply this approach to the first-order logical decision tree learner TILDE (Blockeel and De Raedt, 1998).

Another method is to adapt learning algorithms that make decisions based on counts of positive and negative examples (Elkan and Noto, 2008). The positive and negative counts P and N can be obtained with $P = L/c$ and $N = T - P$. Decision trees, for example, assign classes to leaves and score splits based on the positive/negative counts in the potential subsets (Denis et al., 2005). This idea can be applied straightforwardly to relational learners based on counts, like TILDE or Aleph (Srinivasan).

3 Label Frequency Estimation

To estimate the label frequency in relational PU data, we will use the insights of a propositional label frequency estimator. We first review the original method and then propose a relational version.

3.1 Label Frequency Estimation in Propositional PU data

The propositional estimator is TI*c*E (Bekker and Davis, under review). It is based on two main insights: 1) a subset of the data naturally provides a lower bound on the label frequency, and 2) the lower bound of a large enough positive subset approximates the real label frequency. TI*c*E uses decision tree induction

to find likely positive subsets and estimates the label frequency by taking the maximum of the lower bounds implied by all the subsets in the tree.

The label frequency is the same in subsets of the data because of the 'selected completely at random' assumption, therefore it can be estimated in a subset of the data. Clearly, the true number of positive examples P in a subset cannot exceed the total number of examples in that subset T. This naively implies a lower bound: $c = L/P \geq L/T$. To take stochasticity into account, this bound is corrected with confidence $1 - \delta$ using the one-sided Chebyshev inequality which introduces an error term based on the subset size:

$$\Pr\left(c \leq \frac{L}{T} - \frac{1}{2}\sqrt{\frac{1-\delta}{\delta T}}\right) \leq \delta \tag{1}$$

The higher the ratio of positive examples in the subset, the closer the bound gets to the actual label frequency. The ratio of positive examples is unknown, but directly proportional to the ratio of labeled examples. Therefore, TI*c*E aims to find subsets of the data with a high proportion of labeled examples using decision tree induction. To this end, it uses the *max-bepp* score (Blockeel et al., 2005) that chooses the split that provides the subset with the largest proportion of labels. To avoid overfitting, i.e. finding subsets where $L/P > c$, k folds are used to induce the tree and estimate the label frequency on different datasets.

3.2 Label Frequency Estimation in Relational PU Data

We propose TI*c*ER (Tree Induction for c Estimation in Relational data). The main difference with TI*c*E is that it learns a first order logical decision tree using TILDE (Blockeel and De Raedt, 1998). Each internal node splits on a formula: the examples that satisfy the formula go to the left, the others right. Each node in the tree, therefore, specifies a subset of the data, and each subset implies a lower bound on the label frequency through equation 1. The estimate for the label frequency is the maximal lower bound implied by the subsets.

Ideally, the splits are chosen such that subsets with high proportions of labeled examples are found. However, in this preliminary work, we did not adapt TILDE to use a score function to enable this.

To prevent overfitting, k folds are used to induce the tree and estimate the label frequency on different datasets. With relational data, extra care should be taken that the data in different folds are not related to each other.

4 Experiments

We aim to evaluate if knowing the label frequency makes learning from relational PU data easier and if TI*c*ER provides a good estimate for the label frequency.

Datasets We evaluate our approach on 4 commonly used datasets for relational classification (Table 1). All datasets are available on Alchemy[1], except for

[1] `http://alchemy.cs.washington.edu/data/`

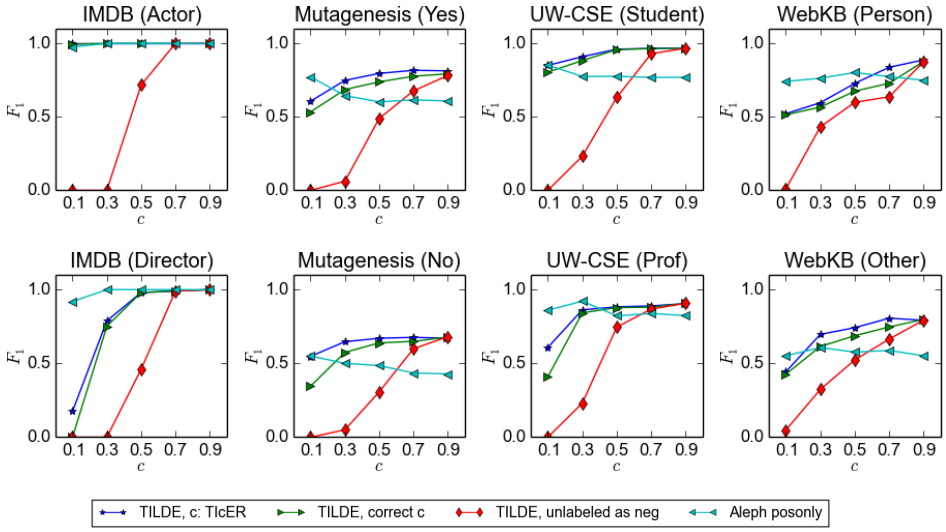

Fig. 1. PU Classifier Comparison. Clearly, taking unlabeled examples as negative is suboptimal. Using the label frequency and posonly seem to work well in different cases. Posonly cannot handle disjunctive concepts well, which is demonstrated by WebKB (Other) which contains webpages from departments, courses, and research projects. It also has difficulties with the overlapping classes in Mutagenesis. Interestingly, TI*c*ER estimate outperforms the exact label frequency. The estimate is an underestimate, which seems to result in more robust classifiers, but this should be further investigated.

Mutagenesis.[2] The datasets were converted to PU datasets by selecting some of the positive examples at random to be labeled. The labeling was done with frequencies $c \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. For each label frequency, all experiments were executed 5 times with different labelings.

Methods To evaluate if knowing the label frequency makes learning easier, we compared four PU classifiers: 1) TILDE adjusted with the TI*c*ER label frequency estimate, 2) TILDE adjusted with the true label frequency, 3) TILDE taking unlabeled examples as negative (i.e. $c = 1$), and 4) Aleph posonly[3]: Muggleton (1996)'s approach. Standard settings were used for all classifiers, except for requiring minimum two example per rule in posonly. k-fold cross validation was applied for validation, i.e. $k - 1$ folds were used for learning the classifier and the other fold to evaluate it. TI*c*ER also needs folds for estimation, it used 1 fold for inducing a tree and the other $k - 2$ folds for bounding the label frequency. The classifiers are compared using the F_1 score and the average absolute error of the estimated label frequency is reported.

Results The label frequency indeed makes learning from PU data easier. It often outperforms posonly, in particular when the positive concept is conjunctive

[2] `http://www.cs.ox.ac.uk/activities/machlearn/mutagenesis.html`
[3] `http://www.cs.ox.ac.uk/activities/machinelearning/Aleph/aleph`

Table 1. Datasets

Datasets	#Examples	Class 1 (#)	Class 2 (#)	# Folds
IMDB	268	Actor (236)	Director (32)	5
Mutagenesis	230	Yes (138)	No (92)	5
UW-CSE	278	Student (216)	Professor (62)	5
WebKB	922	Person (590)	Other (332)	4

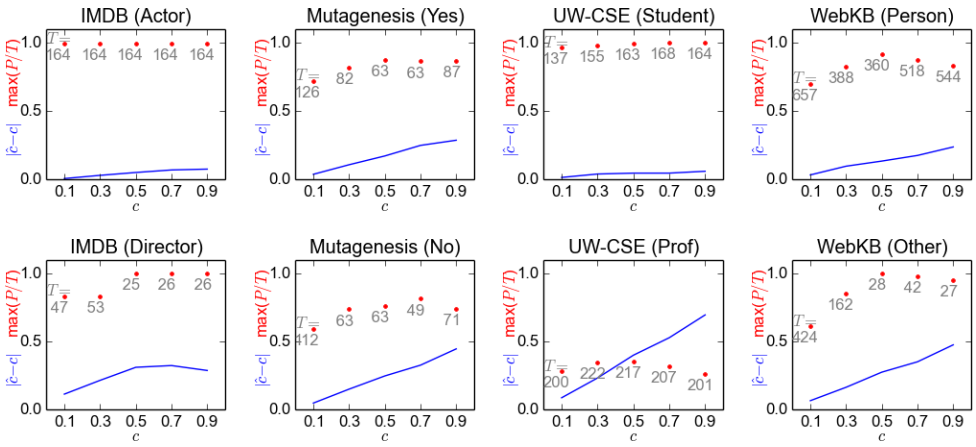

Fig. 2. Label Frequency Estimates. The blue line shows the average absolute error in the label frequency estimate error. The red dots show the maximal proportion of true positives in the used subsets. Each of the dots is annotated with the maximum number of examples in a subset which positive proportion was within 10% of the maximum. The estimate is expected to be good if a large enough subset with a high proportion of positives was found, this is confirmed by the experiments. For example, the worst results, for UW-CSE (Prof), are explained by the low positive proportions.

or overlaps with the negative concept (Figure 1). TIcER gives reasonable results most of the time. It performs weaker when it fails to find subsets with a high ratio of positive examples or when the subsets contain few examples (Figure 2).

5 Conclusions

We showed that using the label frequency c in relational PU learning is very promising. And that the TIcER estimate for c approaches the real value well when enough data is available and when it can find highly positive subsets.

We only evaluated a naive way to use c during learning: by adjusting the output probabilities. We expect that methods that using c internally can improve the results. We expect the label frequency estimate to improve if a score function is used that explicitly looks for pure positive subsets.

Bibliography

J. Bekker and J. Davis. Estimating the class prior in positive and unlabeled data through decision tree induction, under review.

H. Blockeel and L. De Raedt. Top-down induction of first-order logical decision trees. *Artificial intelligence*, pages 285–297, 1998.

H. Blockeel, D. Page, and A. Srinivasan. Multi-instance tree learning. In *Proceedings of the 22nd international conference on Machine learning*, 2005.

M. Claesen, F. De Smet, P. Gillard, C. Mathieu, and B. De Moor. Building classifiers to predict the start of glucose-lowering pharmacotherapy using belgian health expenditure data. *arXiv preprint arXiv:1504.07389*, 2015.

F. Denis, R. Gilleron, and F. Letouzey. Learning from positive and unlabeled examples. *Theoretical Computer Science*, pages 70–83, 2005.

M. C. du Plessis, G. Niu, and M. Sugiyama. Class-prior estimation for learning from positive and unlabeled data. *Machine Learning*, pages 1–30, 2015.

C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220, 2008.

S. Jain, M. White, and P. Radivojac. Estimating the class prior and posterior from noisy positives and unlabeled data. In *Advances in Neural Information Processing Systems*, 2016.

E. McCreath and A. Sharma. Ilp with noise and fixed example size: A bayesian approach. In *Proceedings of the Fifteenth International Joint Conference on Artifical Intelligence*, pages 1310–1315.

S. Muggleton. Learning from positive data. In *Selected Papers from the 6th International Workshop on Inductive Logic Programming*, pages 358–376, 1996.

H. G. Ramaswamy, C. Scott, and A. Tewari. Mixture proportion estimation via kernel embedding of distributions. In *Proceedings of International Conference on Machine Learning*, 2016.

S. Schoenmackers, J. Davis, O. Etzioni, and D. S. Weld. Learning first-order horn clauses from web text. In *Proceedings of Conference on Empirical Methods on Natural Language Processing*, 2010.

A. Srinivasan. The Aleph manual.

D. Zhang and W. S. Lee. A simple probabilistic approach to learning from positive and unlabeled examples. In *Proceedings of the 5th Annual UK Workshop on Computational Intelligence*, pages 83–87, 2005.