1 Learning a New View of a Database: With an Application in Mammography

Jesse Davis
Department of Computer Science
University of Wisconsin - Madison, USA
jdavis@cs.wisc.edu
http://www.cs.wisc.edu/~jdavis

Elizabeth Burnside
Department of Radiology
Department of Biostatistics and Medical Informatics
University of Wisconsin - Madison, USA
es.burnside@hosp.wisc.edu
http://www.biostat.wisc.edu/~burnside

Inês Dutra
COPPE/Sistemas
Universidade Federal do Rio de Janeiro, Brazil
ines@cos.ufrj.br http://www.cos.ufrj.br/~ines

David Page
Department of Biostatistics and Medical Informatics
University of Wisconsin - Madison, USA
page@biostat.wisc.edu
http://www.cs.wisc.edu/~dpage

Raghu Ramakrishnan
Department of Computer Science
University of Wisconsin - Madison, USA
raghu@cs.wisc.edu
http://www.cs.wisc.edu/~raghu

Jude Shavlik
Department of Computer Science
University of Wisconsin - Madison, USA
shavlik@cs.wisc.edu

http://www.cs.wisc.edu/~shavlik

Vítor Santos Costa
COPPE/Sistemas
Universidade Federal do Rio de Janeiro, Brazil
vitor@cos.ufrj.br
http://www.cos.ufrj.br/~vitor

1.1 Introduction

Statistical Relational Learning (SRL) focuses on algorithms for learning statistical models from relational databases. SRL advances beyond Bayesian network learning and related techniques by handling domains with multiple tables, by representing relationships between different rows of the same table, and by integrating data from several distinct databases. Currently, SRL techniques can learn joint probability distributions over the fields of a relational database with multiple tables. Nevertheless, SRL techniques are constrained to use only the tables and fields already in the database, without modification. In contrast, many human users of relational databases find it beneficial to define alternative *views* of a database—further fields or tables that can be computed from existing ones. This chapter shows that SRL algorithms also can benefit from the ability to define new views. Namely, it shows that view learning can be used for more accurate prediction of important fields in the original database.

We augment SRL algorithms by adding the ability to learn new fields, intensionally defined in terms of existing fields and intensional background knowledge. In database terminology, these new fields constitute a learned *view* of the database. We use Inductive Logic Programming (ILP) to learn rules which intensionally define the new fields. We present two different methods to accomplish this goal. The first is a two-step approach where we search for all views of interest. This process is expensive and does not necessarily guarantee selecting the most useful view. The second framework, which we refer to as SAYU-View, has a tighter coupling between view generation and view usage. Our results show that view learning can result in significant benefits.

We present view learning in the specific application of creating an expert system in mammography. We chose this application for a number of reasons. First, it is an important practical application where there has been recent progress in collecting sizable amounts of data. Second, we have access to an expert-developed system. This provides a base reference against which we can evaluate our work [Burnside et al., 2000]. Third, a large proportion of examples are negative. This distribution skew is often found in multi-relational applications. Last, our data consists of a single table. This allows us to compare our techniques against standard propositional learning. In this case, it is sufficient for view learning to extend an existing table with new fields, achieved by using ILP to learn rules for unary predicates. For

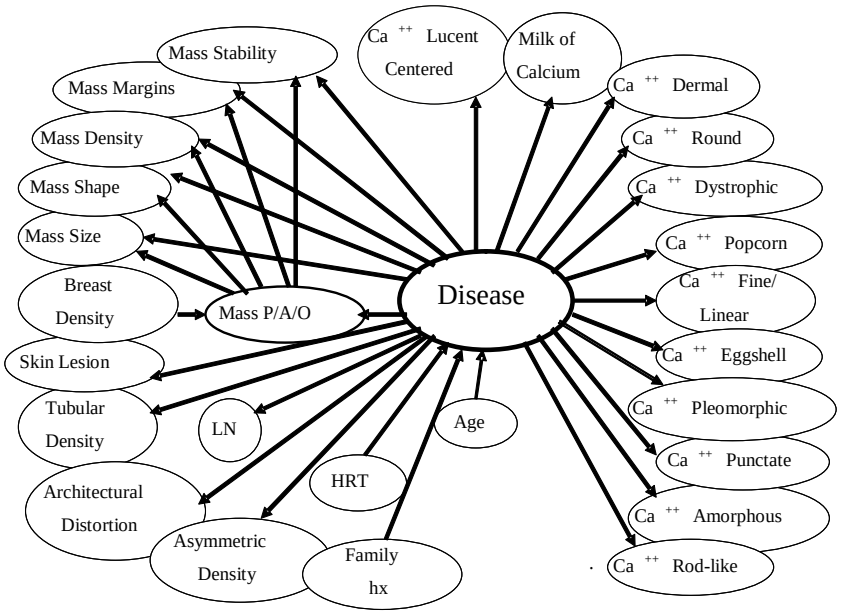

Figure 1.1 Expert Bayes Net

other applications, it may be desirable to learn predicates of higher arity, which will correspond to learning a view with new tables rather than new fields only.

1.2 View Learning for Mammography

Offering breast cancer screening to the ever-increasing number of women over age 40 represents a great challenge. Cost-effective delivery of mammography screening depends on a consistent balance of high sensitivity and high specificity. It has been demonstrated that subspecialist, expert mammographers achieve this balance and perform significantly better than general radiologists [Brown et al., 1995, Sickles et al., 2002]. General radiologists have higher false positive rates and hence biopsy rates, diminishing the positive predictive value for mammography [Brown et al., 1995, Sickles et al., 2002]. Unfortunately, despite the fact that specially trained mammographers detect breast cancer more accurately, there is a longstanding shortage of these individuals [Eklund, 2000].

An expert system in mammography has the potential to help the general radiologist approach the effectiveness of a subspecialty expert, thereby minimizing both false negative and false positive results. Bayesian networks are probabilistic graphical models that have been applied to the task of breast cancer diagnosis from mammography data [Kahn et al., 1997, Burnside et al., 2000, 2004b]. Bayesian networks produce diagnoses with probabilities attached. Because of their graphical nature, they are comprehensible to humans and useful for training. As an example, Figure 1.1 shows the structure of a Bayesian network developed by a subspecialist,

Patient	Abnormality	Date	Mass Shape	...	Mass Size	Location	Be/Mal
P1	1	5/02	Spic	...	0.03	RU4	B
P1	2	5/04	Var	...	0.04	RU4	M
P1	3	5/04	Spic	...	0.04	LL4	B
...	...	...	...	...	...	...	...

Table 1.1 The National Mammography Database schema, omitting some of the features.

expert mammographer. For each variable (node) in the graph, the Bayes net has a conditional probability table giving the probability distribution over the values that variable can take for each possible setting of its parents. The Bayesian network in Figure 1.1 achieves accuracies higher than those of other systems and of general radiologists who perform mammograms, and commensurate with the performance of radiologists who specialize in mammography [Burnside et al., 2000].

Table 1.1 shows the main table (with some fields omitted for brevity) in a large relational database of mammography abnormalities. The database schema is specified in the National Mammography Database (NMD) standard established by the American College of Radiology [ACR, 2004]. The NMD was designed to standardize data collection for mammography practices in the United States and is widely used for quality assurance. We omit a second, much smaller *biopsy* table, simply because we are interested in predicting—before the biopsy—whether an abnormality is benign or malignant. Note that the database contains one record per abnormality. By putting the database into one of the standard database "normal" forms, it would be possible to reduce some data duplication, but only a very small amount: the patient's age, status of hormone replacement therapy and family history could be recorded once per *patient and date* in cases where multiple abnormalities are found on a single mammogram date. Such normalization would have no effect on our approach or results, so we choose to operate directly on the database in its defined form.

Figure 1.2 presents a hierarchy of the four types of learning that might be used for this task. Level 1 and Level 2 are standard types of Bayesian network learning. Level 1 is simply learning the parameters for the expert-defined network structure. Level 2 involves learning the actual structure of the network in addition to its parameters. Notice that to predict the probability of malignancy of an abnormality, a Bayes net uses only the record for that abnormality. Nevertheless, data in other rows of the table may also be relevant: radiologists may also consider other abnormalities on the same mammogram or previous mammograms. For example, it may be useful to know that the same mammogram also contains another abnormality,

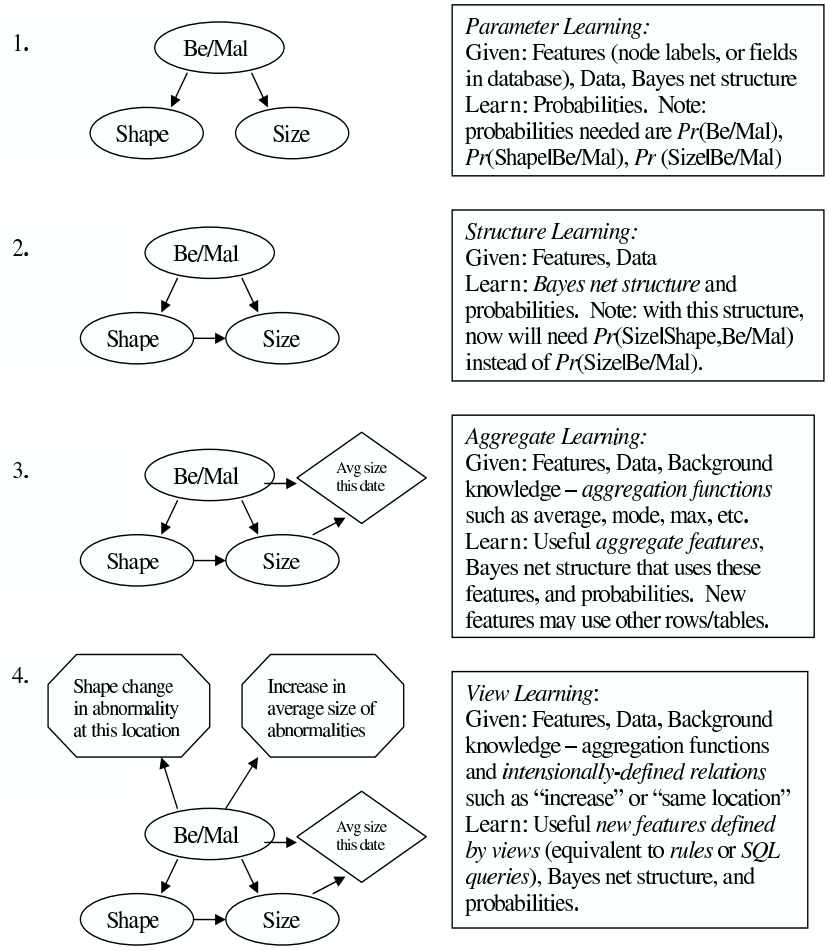

Figure 1.2 Hierarchy of learning types. Levels 1 and 2 are available through ordinary Bayesian network learning algorithms, Level 3 is available only through state-of-the-art SRL techniques, and Level 4 is described in this chapter.

with a particular size and shape; or that the same person had a previous mammogram with certain characteristics. Incorporating data from other rows in the table is not possible with existing Bayesian network learning algorithms and requires statistical relational learning (SRL) techniques, such as probabilistic relational models [Friedman et al., 1999a]. Level 3 in Figure 1.2 shows the state-of-the-art in SRL techniques, illustrating how relevant fields from other rows (or other tables) can be incorporated into the network, using aggregation if necessary. Rather than using only the size of the abnormality under consideration, the new aggregate field allows

the Bayes net to also consider the average size of all abnormalities found in the mammogram.

Presently, SRL is limited to using the original view of the database, that is, the original tables and fields, possibly with aggregation. Despite the utility of aggregation, simply considering only the existing fields may be insufficient for accurate prediction of malignancies. Level 4 in Figure 1.2 shows the key capability that will be introduced and evaluated in this chapter: using techniques from rule learning to learn a new *view*. In this figure, the new view includes two new features utilized by the Bayes net that cannot be defined simply by aggregation of existing features. The new features are defined by two learned rules that capture "hidden" concepts potentially useful for accurately predicting malignancy, but that are not explicit in the given database tables. One learned rule states that a change in the shape of an abnormality at a location since an earlier mammogram may be indicative of a malignancy. The other says that an *increase* in the average of the sizes of the abnormalities may be indicative of malignancy. Note that both rules require reference to other rows in the table for the given patient, as well as intensional background knowledge to define concepts such as "increases over time." Neither rule can be captured by standard aggregation of existing fields.

Note that Level 3 and Level 4 learning would not be necessary if the database initially contained all the potentially useful fields capturing information from other relevant rows or tables. For example, the database might be initially constructed to contain fields such as "slope of change in abnormality size at this location over time", "average abnormality size on this mammogram", and so on. If humans can identify all such potentially useful fields beforehand and define views containing these, then Level 3 and Level 4 learning are unnecessary. Nevertheless, the space of such *possibly* useful fields is quite large, and perhaps more easily searched by computer via Level 3 and Level 4 learning. Certainly in the case of the National Mammography Database standard [ACR, 2004], such fields were not available because they had not been defined and populated in the database by the domain experts, thus making Level 3 and Level 4 learning potentially useful.

1.3 Naïve View Learning Framework

One can imagine a variety of approaches to perform view learning. As a first step, we apply existing technology to obtain a view learning capability. Any relational database can be naturally and simply represented using a subset of first-order logic [Ramakrishnan and Gehrke, 2000]. Inductive logic programming (ILP) provides algorithms to learn rules, also expressed in logic, from such relational data [Muggleton, 1991], possibly together with background knowledge expressed as a logic program. ILP systems operate by searching a space of possible logical rules, looking for rules that score well according to some measure of fit to the data.

Our first learning framework works in two steps. First, we learn rules to predict whether an abnormality is malignant. We extend the original database by introduc-

ing the new rules as *additional features*. More precisely, each rule will correspond to a binary feature such that it takes the value *true* if the body, or condition, of the rule is satisfied, and *false* otherwise. We then run the Bayesian network structure learning algorithm, allowing it to use these new features in addition to the original features. Section 1.7 notes the relationship of the approach to earlier work on ILP for feature construction.
Below we show a simple rule learned by an ILP system. The rule covers 48 positive examples and 123 negative examples. This rule can now be used as a field in a new view of the database, and consequently as a new feature in the Bayesian network.

```
Abnormality A in mammogram M may be malignant if:
   A's tissue is not asymmetric,
   M contains another abnormality A2,
   A2's margins are spiculated, and
   A2 has no architectural distortion.
```

Note that the last two lines of the rule refer to other rows of the relational table for abnormalities in the database. Hence this rule encodes information not available to the current version of the Bayesian network [Davis et al., 2005b].

1.4 Initial Experiments

The purposes of the experiments we conducted are two-fold. First, we want to determine if using SRL yields an improvement compared to propositional learning. Secondly, we want to evaluate whether we see an improvement when moving up a level in the hierarchy outlined in Figure 1.2. First, we try to learn a structure with just the original attributes (Level 2) and see if that performs better than using the expert structure with trained parameters (Level 1). Next, we add aggregate features to our network, representing summaries of abnormalities found either in a particular mammogram or for a particular patient. This corresponds to Level 3 and we test whether this improves over Levels 1 and 2. Finally, we investigate doing Level 4 learning through the two-step algorithm and compare its performance to Levels 1 through 3.
We experimented with a number of structure learning algorithms for Bayesian Networks, including Naïve Bayes, Tree Augmented Naïve Bayes [Friedman et al., 1997], and the Sparse Candidate Algorithm [Friedman et al., 1999b]. However, we obtained the best results with the TAN algorithm in all experiments, so we will focus our discussion on TAN. In a TAN network, each attribute can have at most one other parent in addition to the class variable. The TAN model can be constructed in polynomial time with a guarantee that the model maximizes the Log Likelihood of the network structure given the dataset [Geiger, 1992, Friedman et al., 1997].

1.4.1 Data and Methodology

We collected data for all screening and diagnostic mammography examinations that were performed at the Froedtert and Medical College of Wisconsin Breast Imaging Center between April 5, 1999 and February 9, 2004. It is important to note that the data consists of a radiologist's interpretation of a mammogram and not the raw image data. The radiologist reports conformed to the National Mammography Database (NMD) standard established by the American College of Radiology. From these reports, we followed the original network [Burnside et al., 2000] to cull the 36 features deemed to be relevant by co-author Burnside, an expert mammographer.
To evaluate and compare these approaches, we used stratified 10-fold cross-validation. We randomly divided the abnormalities into 10 roughly equal-sized sets, each with approximately one-tenth of the malignant abnormalities and one-tenth of the benign abnormalities. When evaluating just the structure learning and aggregation, nine folds were used for the training set. When performing aggregation, we used binning to discretize the created features. We took care to only use the examples from the training set to determine the bin widths. When performing view learning, we had two steps in the learning process. In the first part, four folds of data were used to learn the ILP rules. The remaining five folds were used to learn the Bayes net structure and parameters.
When using cross-validation on a relational database, there exists one major methodological pitfall. Some of the cases may be related. For example, we may have multiple abnormalities for a single patient. Because these abnormalities are related (same patient), having some of these in the training set and others in the test set may cause us to perform better on those test cases than we would expect to perform on cases for other patients. To avoid such "leakage" of information into a training set, we ensured that all abnormalities associated with a particular patient were placed into the same fold for cross-validation. Another potential pitfall is that we may learn a rule that predicts an abnormality to be malignant based on properties of abnormalities in *later* mammograms. We ensured that we will never predict the status of an abnormality at a given date based on findings recorded for later dates.

1.4.2 Approach for Each Level of Learning

Level 1: Parameter Learning. We estimated the parameters of the expert structure from the dataset using maximum likelihood estimates with Laplace correction. It has been previously noted that learning the parameters of the network improves performance over having expert defined probabilities in each node [Burnside et al., 2004a].

Level 2: Structure Learning. The relational database for the mammography data contains one row for each abnormality described on a mammogram. Fields in this relational table include all those shown in the Bayesian network of Figure 1.1.

Patient	Abnormality	Date	Mass Shape	...	Mass Size	Location	Average Patient Mass Size	Average Mammogram Mass Size	Be/Mal
P1	1	5/02	Spic	...	0.03	RU4	0.0367	0.03	B
P1	2	5/04	Var	...	0.04	RU4	0.0367	0.04	M
P1	3	5/04	Spic	...	0.04	LL4	0.0367	0.04	B
...	...	...	...	...	...	...	...	...	...

Table 1.2 Database after Aggregation on Mass Size Field. Note the addition of two new fields, Average Patient Mass Size and Average Mammogram Mass Size, which represent aggregate features.

Therefore it is straightforward to use existing Bayesian network structure learning algorithms to learn a possibly improved structure for the Bayesian network.

Level 3: Aggregate Learning. We selected the numeric (e.g. the size of mass) and ordered features (e.g. the density of a mass) in the database and computed aggregates for each of these features. In all, we determined that 27 of the 36 attributes were suitable for aggregation. We computed aggregates on both the patient and the mammogram level. On the patient level, we looked at all of the abnormalities for a specific patient. On the mammogram level, we only considered the abnormalities present on that specific mammogram. To discretize the averages, we divided each range into three bins. For binary features we used predefined bin sizes, while for the other features we attempted to get equal numbers of abnormalities in each bin. For aggregation functions we used maximum and average. The aggregation introduced $27 \times 4 = 108$ new features. The following paragraph presents further details of our aggregation process.

We used a three-step process to construct aggregate features. First, we chose a field to aggregate. Second, we selected an aggregation function. Third, we needed to decide over which rows to aggregate the feature, that is, which keys or links to follow. This is known as a slot chain in PRM terminology [Friedman et al., 1999a]. In our database, two such links exist. The patient ID field allows access to all the abnormalities for a given patient, providing aggregation on the patient level. The second key is the combination of patient ID and mammogram date, which returns all abnormalities for a patient on a specific mammogram, providing aggregation on the mammogram level. To demonstrate this process, we will work though an example of computing an aggregate feature for patient 1 in the database given in Figure 1.1. We will aggregate on the Mass Size field and use average as the

aggregation function. Patient 1 has three abnormalities, one from a mammogram in May 2002 and two from a mammogram in May 2004. To calculate the aggregate on the patient level, we average the size for all three abnormalities, which is .0367. To find the aggregate on the mammogram level for patient 1, we have to perform two separate computations. First, we follow the link P1 and 5/02, which yields abnormality 1. The average for this key mammogram is simply .03. Second, we follow the link P1 and 5/04, which yields abnormalities 2 and 3. The average for these abnormalities is .04. Table 1.2 shows the database following construction of these aggregate features.

Level 4: View Learning. We used the ILP system Aleph [Srinivasan, 2001] to implement Level 4 learning. Aleph was asked to learn rules predictive of malignancy. We introduced three new intensional tables into Aleph's background knowledge to take advantage of relational information.

1. The `prior_Mammogram` relation connects information about any prior abnormality that a given patient may have.

2. The `same_Location` relation is a specification of the previous predicate. It adds the restriction that the prior abnormality must be in the same location as the current abnormality. Radiology reports include information about the location of abnormalities.

3. The `in_Same_Mammogram` relation incorporates information about other abnormalities a patient may have on the current mammogram.

By default, Aleph is set up to generate rules that would fully explain the examples. In contrast, our goal was to extract rules that would be beneficial as new views. The major problem in implementing Level 4 learning was *how to select rules that would best complement Level 3 information.* Clearly, Aleph's standard coverage algorithm was not designed for this application. Instead, we chose to first enumerate as many rules of interest as possible, and then chose interesting rules.
In order to obtain a varied set of rules, we ran Aleph under `induce_max` for each fold. `Induce_max` uses every positive example in each fold as a seed for the search. Also note that `induce_max` does not discard previously covered examples when scoring a new clause. Several thousand distinct rules were learned for each fold, with each rule covering many more malignant cases than (incorrectly covering) benign cases. We avoid the rule overfitting found by other authors [Perlich and Provost, 2003] by doing breadth-first search for rules and by having a minimal limit on coverage.
Each seed generated anywhere from zero to tens of thousands of rules. Adding all rules would mean introducing thousands of often redundant features. We implemented the following algorithm:

1. We scanned all rules looking for duplicates and for rules that performed worse than a more general rule. This step significantly reduced the number of rules to consider.

2. We sorted rules according to their m-estimate.

3. We used a greedy algorithm that picks the rule with the highest m-estimate such that it covers an unexplained training example. Furthermore, each rule needs to cover a significant number of malignant cases. This step is similar to the standard ILP greedy covering algorithm, except that we do not follow the original order of the seed examples.

4. Last, we scanned the remaining rules, selecting those that covered a significant number of examples, and that were different from all previous rules, *even though these rules would not cover any new examples.*

It is important to note that the rule selection was an automated process. We picked the top 50 clauses in our experiments, obtained from practical considerations on the size of the Bayesian networks we would need to learn. The resulting views were added as new features to the database.

1.4.3 Results

We present the results of our first experiment, comparing Levels 1 and 2, using both ROC and precision-recall curves. Figure 1.3 shows the ROC curve for these experiments, and Figure 1.4 shows the precision-recall curves. Because of our skewed class distribution, due to the large number of benign cases, we prefer precision-recall curves over ROC curves because they better show the number of "false alarms," or unnecessary biopsies. Therefore, we use precision-recall curves for the remainder of the results. Here, precision is the percentage of abnormalities that we classified as malignant that are truly cancerous. Recall is the percentage of malignant abnormalities that were correctly classified. To generate the curves, we pooled the results over all ten folds by treating each prediction as if it had been generated from the same model. We sorted the estimates and used all possible split points to create the graphs.

Figure 1.5 compares performance for all levels of learning. We can observe very significant improvements when adding multi-relational features. Aggregates provide the most benefit for higher recalls whereas rules help in the medium and low ranges of recall. We believe this is because ILP rules are more accurate than the other features, but have limited coverage.

Figure 1.6 shows the average area under the precision-recall curve for each level of learning that we defined in Figure 1.2. We only consider recalls above 50%, as for this application radiologists would be required to perform at least at this level. We further use the paired t-test to compare the areas under the curve (recall ≥ 0.5) for every fold. We found improvement of Level 2 over Level 1 to be statistically significant with a 99% level of confidence. According to the paired t-test the improvement of Level 3 presents an improvement over Level 2 at the 97% confidence level. Furthermore, Level 4 over Level 2 is significant, using the area under the curve metric, at the 99% level. However, there is no significant difference between Level 3 and Level 4.

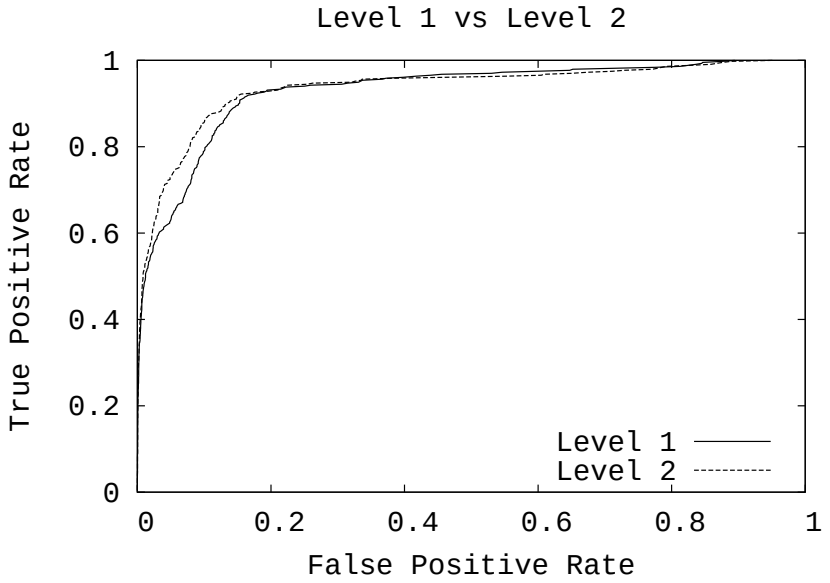

Figure 1.3 ROC Curves for Parameter Learning (Level 1) compared to Structure Learning (Level 2).

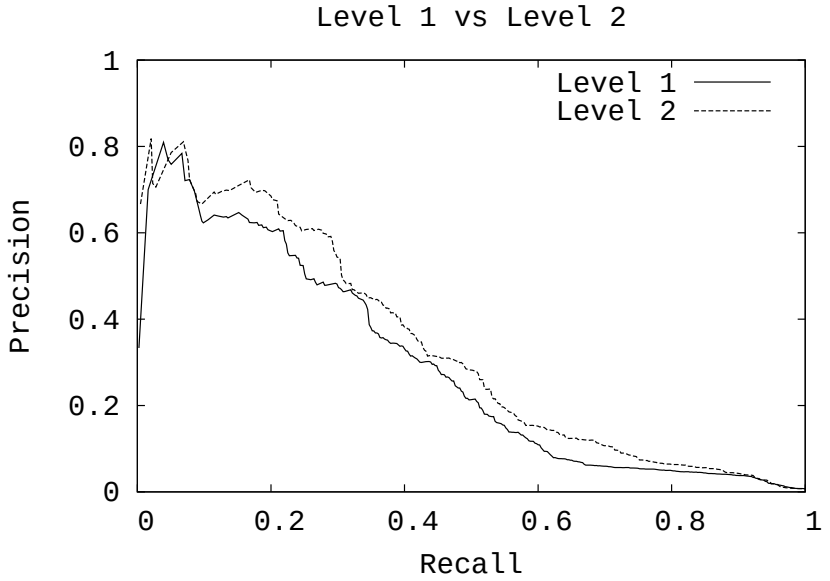

Figure 1.4 Precision-Recall Curves for Parameter Learning (Level 1) compared to Structure Learning (Level 2).

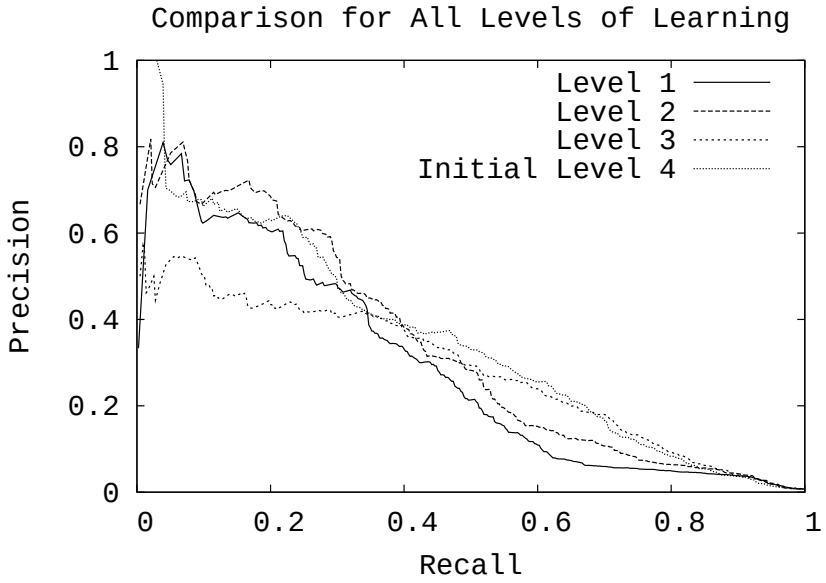

Figure 1.5 Precision-Recall Curves for Each Level of Learning

In this task, considering relational information is still crucial for improving performance since the relational approaches outperform the propositional methods. We mostly see significant improvement as we move the learning hierarchy outlined in Figure 1.2. However, in this initial approach we see no significant difference between Level 3 and Level 4.

The process of generating the views in Level 4 can be useful to the radiologist, as it identifies potentially interesting correlations between attributes. During our experiments, we presented co-author Burnside with a set of 130 rules to review. She found several rules interesting, including the following:

```
Abnormality A in mammogram M for patient P is maligant if:
A has BI-RADS category 5,
A has a mass present,
A has a mass with high density,
P has a prior history of breast cancer,
P has an extra finding on same mammogram (B),
B has no pleomorphic microcalcifications,
B had no punctate calcifications.
```

This rule identified 42 malignant mammographic findings while only misclassifying 11 benign findings as cancer. The radiologist was intrigued by this rule because it suggests a hitherto unknown relationship between malignancy and high density masses. In general, mass density was not previously thought to be a highly predictive feature, so this rule is valuable in its own right [Burnside et al., 2005].

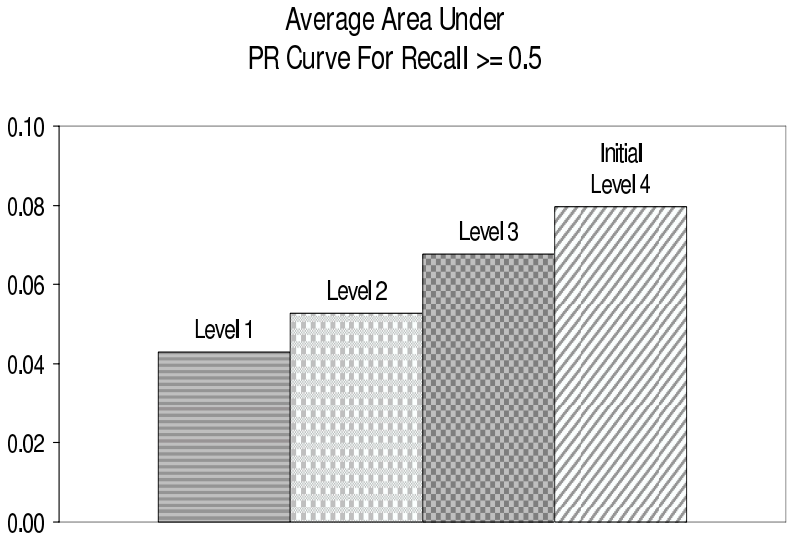

Figure 1.6 Area Under the Curve For Recalls Above 50%

1.5 Integrated View Learning Framework

The initial methodology for Level 4 follows a two-step process. In the first step, an ILP algorithm learns a set of rules. In the second step, the learned rules are added to the pre-existing features to form a final model. This approach suffers from several weaknesses. First, we follow a brute-force approach to search for all *good* rules, but we have no way to evaluate which ones will actually improve the network. Second, the metric used to score the rules differs from the one we will ultimately use to evaluate the final model. Thus, we have no guarantee that the rule learning process will select the rules that best contribute to the final classifier.

We propose an alternative approach, based on the idea of *constructing the classifier as we learn the rules* [Davis et al., 2005a]. In the new approach, rules are scored by how much they improve the classifier, providing a tight coupling between rule generation and rule usage. We call this methodology *Score As You Use* or *SAYU*. SAYU is closely related to Landwehr, Kersting and De Raedt's nFOIL [Landwehr et al., 2005] and also to Popescul *et al's* work on Structural Logistic Regression [Popescul et al., 2003]. The relationships to these important works are discussed in Section 1.7.

Our implementation of SAYU depends on both an ILP system and a propositional learner. Following the original work, we used Aleph as a rule proposer and Tree Augmented Naïve Bayes (TAN) as our propositional learner.

Our algorithm works as follows. We randomly choose a seed example, and obtain its most specific, or *saturated* clause. We then perform a top down breadth-first

```
Input: Train Set T, Tune Set S, Stop Criteria
Output: A TAN Model
M = BuildTANClassifier(T);
BestScore = AreaUnderPRCurve(M, S);
while Stop criteria not met do
    done = false; Choose a positive example as a seed and saturate the example;
    repeat
        NewFeature = Generate new clause according to saturated example;
        M_new = BuildTANClassifier(T ∪ NewFeature );
        NewScore = AreaUnderPRCurve(M, S ∪ NewFeature);
        if NewScore ≥ BestScore then
            T = T ∪ NewFeature;
            S = S ∪ NewFeature;
            BestScore = NewScore;
            M = M_new;
            done = true;
        end
    until not(done);
end
return M
```

Algorithm 1: SAYU-View Algorithm

search of the subsumption lattice. We evaluate each clause by converting it to a binary feature, which is added to the current training set. We learn a new Bayes net incorporating this new feature, and score the network. If the new feature improves the score of the network, then we retain the feature in the network. If the feature degrades the performance of the network, it is discarded, we and revert back to the old classifier and continue searching. One other central difference exists with our algorithm compared to Aleph in that after the network accepts a rule, we randomly select a new seed. Thus, we are not searching for the best rule, but only the first rule that helps. However, nothing prevents the same seed from being selected multiple times during the search.

Finally, we need to define a scoring function. The main goal is to use the same scoring function for both learning and evaluation. Furthermore, we wish to be able to handle datasets that have a highly skewed class distribution. In the presence of skew, precision and recall are often used to evaluate classifier quality. In order to characterize how the algorithm performs over the whole precision recall space, we follow Goadrich et al. [Goadrich et al., 2004], and adopt the area under the precision-recall curve as our scoring metric. When calculating the area under the precision-recall curve, we integrate from recall levels of 0.5 or greater. As we previously noted, a radiologist would have to achieve levels of recall in this range.

We have previously reported that SAYU performs on par with Level 3 and the initial approach to Level 4. However, in these experiments we implemented SAYU as a rule combiner only, not as a tool for view learning that *adds* fields to the existing set of fields (features) in the database [Davis et al., 2005a]. We have modified SAYU

to take advantage of the predefined features yielding a more integrated approach to View Learning. We also report on a more natural design where SAYU starts from the Level 3 network. We call this approach SAYU-View. Algorithm 1 gives psuedo code for the SAYU-View algorithm.

1.6 Further Experiments and Results

We use essentially the same methodology as described previously for the initial approach to view learning. On each round of cross-validation, we use four folds as a training set, five folds as a tuning set and one fold as a test set. We only saturate examples from the training set. For SAYU-View, we use only the training set to learn the rules. The key difference between initial Level 4 and SAYU-View is the following: for SAYU-View we use the training set to learn the structure and parameters of the Bayes net, and we use the tuning set to calculate the score of a network structure. Previously, we used the tune set to learn the network structure and parameters. In order to retain a clause in the network, the area under the precision-recall curve of the Bayes net incorporating the rule must achieve at least a two percent improvement over the area of the precision-recall curve of the best Bayes net.

Within SAYU, the time to score a rule has increased. The Bayes net algorithm has to learn a new network topology and new parameters each time we score a rule (feature). Furthermore, inference must be performed to compute the score after incorporating a new feature. The SAYU algorithm is strictly more expensive than standard ILP as SAYU also has to prove whether a rule covers each example in order to create the new feature. To reflect the added cost, we use a time-based stop criteria for the new algorithm. This criteria is described in further detail in [Davis et al., 2005a]. For each fold, we use the times from the baseline experiments in [Davis et al., 2005a], so that our new approach to view learning take the same time as the old approach. In practice, our settings resulted in evaluating around 20000 clauses for each fold, requiring on average around 4 hours per fold on a Xeon 3MHz class machine.

Figure 1.7 includes a comparison of SAYU-View to Level 3 and the initial approach to Level 4. Again, we perform a two tailed paired t-test on the area under the precision recall curve for levels of recall ≥ 0.5. SAYU-View performs significantly better than both these approaches at the 99% confidence level. Although we do not include the graph, SAYU-View performs significantly better than the SAYU-TAN (no initial features), also with a p-value < 0.01. SAYU-View also performs better than Level 1 and Level 2 with a p-value < 0.01. With the integrated framework for Level 4, we now see significant improvement over lower levels of learning when we ascend the hierarchy defined in Figure 1.2.

Figure 1.8 shows the average area under the precision-recall curve (AUCPR) for levels of recall ≥ 0.5 for Level 3, the initial approach to Level 4, and SAYU-View. The average AUCPR for SAYU-View yields a 30% increase in the average AUCPR

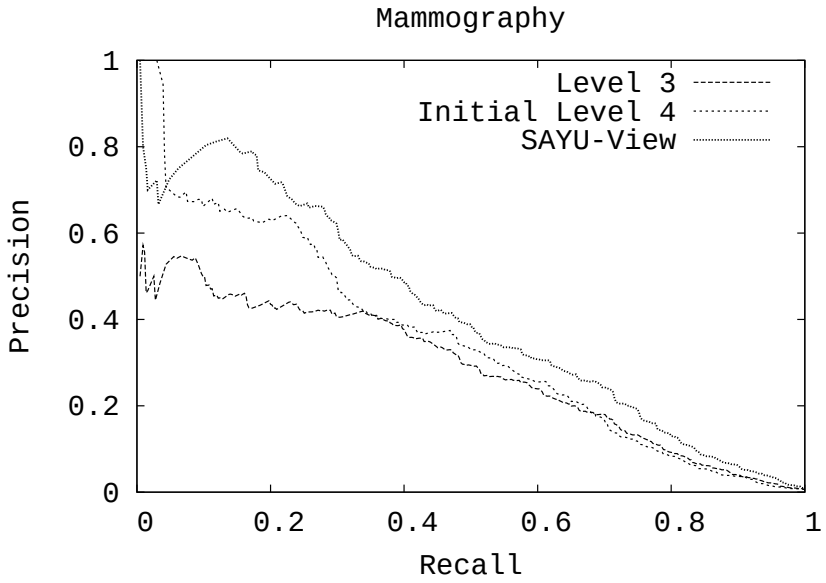

Figure 1.7 Precision-Recall Curves for Each Level of Learning

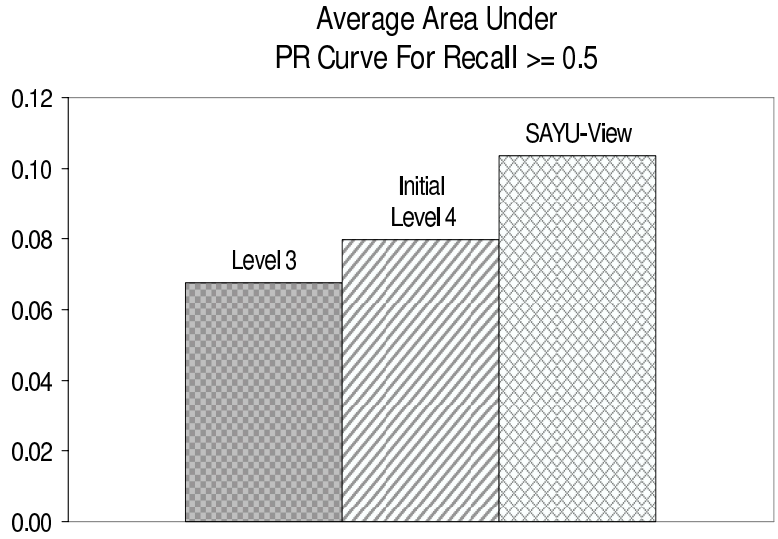

Figure 1.8 Area Under the Curve For Recalls Above 50%

over the initial approach to Level 4. Furthermore, we see an increase in the average AUCPR of 53% over Level 3. Another way to look at these results is the potential reduction of benign biopsies: procedures done on women without cancer. When

detecting 90% of cancers (i.e., recall = 0.9), SAYU-View achieves a 35% reduction in benign biopsies over Level 3 and a 39% reduction over the initial Level 4 method.

1.7 Related Work

Research in SRL has advanced along two main lines: methods that allow graphical models to represent relations, and frameworks that extend logic to handle probabilities. Along the first line, probabilistic relational models, or PRMs, introduced by Friedman, Getoor, Koller and Pfeffer, represent one of the first attempts to learn the structure of graphical models while incorporating relational information[Friedman et al., 1999a]. Recently Heckerman, Meek and Koller have discussed extensions to PRMs and compared them to other graphical models[Heckerman et al., 2004]. A statistical learning algorithm for probabilistic logic representations was first given by Sato [Sato, 1995] and later, Cussens [Cussens, 2001] proposed a more general algorithm to handle log linear models. Additionally, Muggleton [Muggleton, 2000] has provided learning algorithms for stochastic logic programs. The structure of the logic program is learned using ILP techniques, while the parameters are learned using an algorithm scaled up from that used for stochastic context-free grammars.
Newer representations garnering arguably the most attention are Bayesian logic programs [Kersting and Raedt, 2002] (BLPs), relational markov networks (RMNs) [Taskar et al., 2002], constraint logic programming with Bayes net constraints, or CLP($\mathcal{BN}$) [Santos Costa et al., 2003], and Markov Logic Networks (MLNs) [Richardson and Domingos, 2004]. Markov Logic Networks are most similar to our approach. Nodes of MLNs are the ground instances of the literals in the rule, and the arcs correspond to the rules. One major difference is that, in our approach, nodes are the rules themselves. Although we cannot work at the same level of detail, our approach makes it straightforward to combine logical rules with other features, and we now can take full advantage of propositional learning algorithms.
The present work builds upon previous work on using ILP for feature construction. Such work treats ILP-constructed rules as Boolean features, re-represents each example as a feature vector, and then uses a feature-vector learner to produce a final classifier. To our knowledge, Pompe and Kononenko [Pompe and Kononenko, 1995] were the first to apply Naïve Bayes to combine clauses. Other work in this category was by Srinivasan and King [Srinivasan and King, 1997], who used rules as extra features for the task of predicting biological activities of molecules from their atom-and-bond structures. Popescul et al. [Popescul and Ungar, 2004] use $k-means$ to derive cluster relations, which are then combined with the original features through structural regression. In a different vein, Relational Decision Trees [Neville et al., 2003] use aggregation to provide extra features on a multi-relational setting, and are close to our Level 3 setting. Knobbe et al. [Knobbe et al., 2001] proposed numeric aggregates in combination with logic-based feature construction for single attributes. Perlich and Provost discuss several approaches for attribute construction using aggregates over multi-relational features [Perlich and Provost, 2003]. The

authors also propose a hierarchy of levels of learning: feature vectors, independent attributes on a table, multidimensional aggregation on a table, and aggregation across tables. Some of these techniques in their hierarchy could be applied to perform view learning in SRL.

Another approach for a tight coupling between rule learning and rule usage is the work appearing earlier this year (done in parallel with ours) by Landwehr, Kersting and De Raedt [Landwehr et al., 2005]. That work presented a new system called nFOIL. We would like to highlight several significant differences in the two pieces of work appear to be the following. First, nFOIL scores clauses by conditional log likelihood rather than improvement in classifier accuracy or classifier AUC (area under ROC or PR curve). Second, nFOIL can handle multiple-class classification tasks, which SAYU cannot. Third, the present chapter reports experiments on data sets with significant class skew, to which probabilistic classifiers are often sensitive. Fourth, this work looks at TAN opposed to Naïve Bayes. Finally, this work extends both [Landwehr et al., 2005] and [Davis et al., 2005a] by giving the network an initial feature set.

Another related piece of work is that by Popescul *et al.* [Popescul et al., 2002, Popescul and Ungar, 2003, Popescul et al., 2003] on Structural Logistic Regression. They use an ILP-like (refinement graph) search over rules, expressed as database queries, to define new features. Differences from the present work include their use of the new features within alogistic regression model rather than a graphical model, and the fact that they do not update the logistic regression model after adding each rule. A notable strength of their approach is that the rule-learning process itself can include aggregation.

1.8 Conclusions and Future Work

We presented a method for statistical relational learning which integrates learning from attributes, aggregates, and rules. Our example application shows benefits from the several levels of learning we proposed. Level 2, structure learning, clearly outperforms the expert structure. We further show that multi-relational techniques can achieve very significant improvements, even on a single table domain.

This chapter has shown that a simple form of view learning—treating rules induced by a standard ILP system as the additional features of a new view—yields improved performance over Level 2 learning. Nevertheless, this improvement is roughly equal to that obtained by Level 3 learning—by aggregation, as might be performed for example by a PRM. We have noted how this approach to view learning is quite similar to earlier work using ILP for feature construction.

A more interesting form of view learning, or Level 4 learning, is SAYU-View, which closely integrates the ILP system and Bayesian network learner. It significantly improves performance over both Level 3 learning and the simple form of view learning.

We believe many further improvements in view learning are possible. It makes

sense to include aggregates in the background knowledge for rule generation. Alternatively, one can extend rules with aggregation operators, as proposed in recent work by Vens et al. [Vens et al., 2004]. We have found the rule selection problem to be non-trivial. Our greedy algorithm often generates too similar rules, and is not guaranteed to maximize coverage. We would like to approach this problem as an optimization problem weighing coverage, diversity, and accuracy.

Our approach of using ILP to learn new features for an existing table merely scratches the surface of the potential for view learning. A more ambitious approach would be to more closely integrate structure learning and view learning. A search could be performed in which each "move" in the search space is either to modify the probabilistic model or to refine the intensional definition of some field in the new view. Going further still, one might learn an intensional definition for an entirely new table. As a concrete example, for mammography one could learn rules defining a binary predicate that identifies "similar" abnormalities. Because such a predicate would represent a many-to-many relationship among abnormalities, a new table would be required.

SRL algorithms provide a substantial extension to existing statistical learning algorithms, such as Bayesian networks, by permitting statistical learning to be applied directly to relational databases with multiple tables. Nevertheless, the schemas for relational databases often are defined based on criteria other than effectiveness of learning. If a schema is not the most appropriate for a given learning task, it may be necessary to change it—by defining a new view—before applying other SRL techniques. View learning, as presented in this chapter, provides an automated capability to make such schema changes. Our approaches so far to view learning build on existing ILP technology. We believe ILP-based view learning can be greatly improved and extended, as outlined in the preceding paragraphs, for example to learn entirely new tables. Furthermore, many approaches to view learning outside of ILP remain to be explored.

1.9 Acknowledgments

Support for this research was partially provided by U.S. Air Force grant F30602-01-2-0571. Elizabeth Burnside is supported by a General Electric Research in Radiology Academic Fellowship. Inês Dutra and Vítor Santos Costa did this work while visiting the University of Wisconsin-Madison. Vítor Santos Costa was partially supported by the Fundação para a Ciência e Tecnologia. We would like to thank Lisa Torrey, Mark Goadrich, Rich Maclin, Jill Davis, and Allison Holloway for reading over drafts of this chapter. We would also like to thank the referees for their insightful comments.

References

ACR. Breast imaging reporting and data system (bi-rads), 2004.

Martin Brown, Florence Houn, Edward Sickles, and Larry Kessler. Screening mammography in community practice: positive predictive value of abnormal findings and yield of follow-up diagnostic procedures. *AJR Am J Roentgenol*, 165:1373–1377, 1995.

Elizabeth Burnside, Jesse Davis, Vítor Santos Costa, Inês C. Dutra, Charles Kahn, Jason Fine, and David Page. Knowledge discovery from structured mammography reports using inductive logic programming. In *American Medical Informatics Association Symposium*, pages 96–100, 2005.

Elizabeth Burnside, Yue Pan, Charles Kahn, Katherine Shaffer, and David Page. *Training a Probabilistic Expert System to Predict the Likelihood of Breast Cancer Using a Large Dataset of Mammograms (abstr).* Radiological Society of North America, 2004a.

Elizabeth Burnside, Daniel Rubin, and Ross Shachter. A Bayesian network for screening mammography. In *American Medical Informatics Association*, pages 106–110, 2000.

Elizabeth Burnside, Daniel Rubin, and Ross Shachter. Using a Bayesian network to predict the probability and type of breast cancer represented by microcalcifications on mammography. *Medinfo*, 2004:13–17, 2004b.

James Cussens. Parameter estimation in stochastic logic programs. *Machine Learning*, 44(3):245–271, 2001.

Jesse Davis, Elizabeth Burnside, Inês C. Dutra, David Page, and Vítor Santos Costa. An integrated approach to learning bayesian networks of rules. In *16th European Conference on Machine Learning*, pages 84–95. Springer, 2005a.

Jesse Davis, Elizabeth Burnside, Inês C. Dutra, David Page, Raghu Ramakrishnan, Vítor Santos Costa, and Jude Shavlik. View learning for statistical relational learning: With an application to mammography. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 677–683, Edinburgh, Scotland, 2005b.

G. William Eklund. Shortage of qualified breast imagers could lead to crisis. *Diagn Imaging*, 22:31–33, 2000.

Nir Friedman, David Geiger, and Moises Goldszmidt. Bayesian networks classifiers.

Machine Learning, 29:131–163, 1997.

Nir Friedman, Lise Getoor, Daphne Koller, and Avi Pfeffer. Learning probabilistic relational models. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*. Stockholm, Sweden, 1999a.

Nir Friedman, Iftach Nachman, and Dana. Pe'er. Learning Bayesian Network Structure from Massive Datasets: The "Sparse Candidate" Algorithm. In *Proceedings of the 17th Conference of Uncertainty in Artificial Intelligence*, pages 206–215, San Francisco, CA, 1999b. Morgan Kaufmann Publishers.

Dan Geiger. An entropy-based learning algorithm of Bayesian conditional trees. In *Proceedings of the 8th Conference on Uncertainty in Artificial Intelligence*, pages 92–97, San Mateo, CA, 1992. Morgan Kaufmann Publishers.

Mark Goadrich, Louis Oliphant, and Jude Shavlik. Learning Ensembles of First-Order Clauses for Recall-Precision Curves: A Case Study in Biomedical Information Extraction. In *Proceedings of the 14th International Conference on Inductive Logic Programming*, Porto, Portugal, 2004.

David Heckerman, Christopher Meek, and Daphne Koller. Probabilistic Entity-Relationship Models, PRMs, and Plate Models, Technical Report MSR-TR-2004-30, Microsoft Research. Technical report, Microsoft Research, 2004.

Charles Kahn, Linda Roberts, Katherine Shaffer, and Peter Haddawy. Construction of a Bayesian network for mammographic diagnosis of breast cancer. *Comput Biol Med.*, 27:19–29, 1997.

Kristian Kersting and Luc De Raedt. Basic principles of learning Bayesian logic programs. Technical report, Institute for Computer Science, University of Freiburg, Germany, 2002.

Arno J. Knobbe, Marc de Haas, and Arno Siebes. Propositionalisation and aggregates. In *Proceeding of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 277–288, 2001.

Niels Landwehr, Kristian Kersting, and Luc De Raedt. nFOIL: Integrating Naive Bayes and FOIL. In *National Conference on Artificial Intelligene (AAAI)*, 2005.

Stephen Muggleton. Inductive Logic Programming. *New Generation Computing*, 8:295–318, 1991.

Stephen Muggleton. Learning stochastic logic programs. *Electronic Transactions in Artificial Intelligence*, 4(041), 2000.

Jennifer Neville, David Jensen, Lisa Friedland, and Michael Hay. Learning relational probability trees. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 625–630. ACM Press, 2003.

Claudia Perlich and Foster Provost. Aggregation-based feature invention and relational concept classes. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 167–176, 2003.

Uros Pompe and Igor Kononenko. Naive Bayesian classifier within ILP-R. In

L. De Raedt, editor, *Proceeding of the 4th International Workshop on Inductive Logic Programming*, pages 417–436, 1995.

Alexandrin Popescul and Lyle H. Ungar. Statistical relational learning for link prediction. In *Workshop on Learning Statistical Models from Relational Data at IJCAI 2003*, 2003.

Alexandrin Popescul and Lyle H. Ungar. Cluster-based concept invention for statistical relational learning. In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 665–670, 2004.

Alexandrin Popescul, Lyle H. Ungar, Steve Lawrence, and David M. Pennock. Towards structural logistic regression: Combining relational and statistical learning. In *Workshop on Multi-Relational Data Mining at KDD*, 2002.

Alexandrin Popescul, Lyle H. Ungar, Steve Lawrence, and David M. Pennock. Statistical relational learning for document mining. In *ICDM03*, pages 275–282, 2003.

Raghu Ramakrishnan and Johannes Gehrke. *Database Management Systems*. McGraw Hill, 2000.

Matt Richardson and Pedro Domingos. Markov logic networks. http://www.cs.washington.edu/homes/pedrod/kbmn.pdf, 2004.

Vítor Santos Costa, David Page, Maleeha Qazi, and James Cussens. CLP($\mathcal{BN}$): Constraint Logic Programming for Probabilistic Knowledge. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence (UAI03)*, pages 517–524, Acapulco, Mexico, August 2003.

Taisuke Sato. A statistical learning method for logic programs with distributional semantics. In L. Sterling, editor, *Proceedings of the Twelth International conference on logic programming*, pages 715–729, Cambridge, Massachusetts, 1995. MIT Press.

Edward Sickles, Dulcy Wolverton, and Katherine Dee. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology*, 224:861–869, 2002.

Ashwin Srinivasan. *The Aleph Manual*, 2001.

Ashwin Srinivasan and Ross King. Feature construction with inductive logic programming: A study of quantitative predictions of biological activity aided by structural attributes. In *Proceeding of the 7th International Workshop on Inductive Logic Programming*, pages 89–104, 1997.

Ben Taskar, Pieter Abbeel, and Daphne Koller. Discriminative probabilistic models for relational data. In Adnan Darwiche and Nir Friedman, editors, *Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence*, pages 485–492. Morgan Kaufmann, 2002.

Celine Vens, Anneleen Van Assche, Hendrik Blockeel, and Sašo Džeroski. First order

random forests with complex aggregates. In *Proceedings of the 14th International Conference on Inductive Logic Programming*, pages 323–340, 2004.