View Learning for Statistical Relational Learning: With an Application to Mammography

Jesse Davis, Elizabeth Burnside, Ines Dutra, David Page,
Raghu Ramakrishnan, Vitor Santos Costa, Jude Shavlik
University of Wisconsin - Madison

Abstract

Statistical relational learning (SRL) constructs probabilistic models from relational databases. A key capability of SRL is the learning of arcs (in the Bayes net sense) connecting entries in different rows of a relational table, or in different tables. Nevertheless, SRL approaches currently are constrained to use the existing database schema. For many database applications, users find it profitable to define alternative "views" of the database, in effect defining new fields or tables. Such new fields or tables can also be highly useful in learning. We provide SRL with the capability of learning new views.

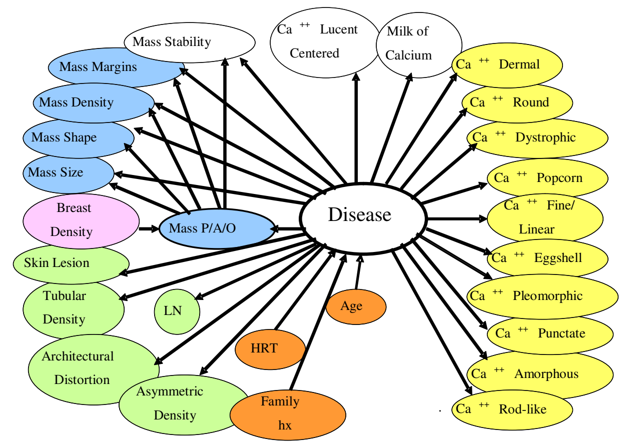

Figure 1: Expert Bayes Net

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, representing relationships between different rows of the same table, and integrating data from several distinct databases. SRL techniques currently can learn joint probability distributions over the fields of a relational database with multiple tables. Nevertheless, they are constrained to use only the tables and fields already in the database, without modification. 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 paper shows that SRL algorithms also can benefit from the ability to define new views, which can be used for more accurate prediction of important fields in the original database.

We will 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 test the approach 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 with sizable data. Second, we have access to an expert developed system. This provides a base reference against which we can evaluate our work. 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. It should be clear that for other applications the approach can yield additional tables.

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. Recent articles demonstrate that subspecialist, expert mammographers achieve this balance and perform significantly better than general radiologists [24; 1]. General radiologists have higher false positive rates and hence biopsy rates, diminishing the positive predictive value for mammography [24; 1]. Despite the fact that specially trained mammographers detect breast cancer more accurately, there is a longstanding shortage of these individuals [6].

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 (BNs) are probabilistic graphical models that have been applied to the task of breast cancer diagnosis from mammography data [12; 2; 3]. BNs produce diagnoses with probabilities attached. Because of their graphical nature, they are comprehensible to humans and useful for training. As an example, Figure 1 shows the structure of a Bayesian network developed by a subspecialist, 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 achieves accuracies higher than that of other systems and of general radiologists who perform mammograms, and commensurate with the performance of radiologists who specialize in mammography [2].

Figure 2 shows the main table (with some fields omitted for brevity) in a large relational database of mammography abnormalities. Data was collected using the National Mammography Database (NMD) standard established by the American College of Radiology. The NMD was designed to standardize data collection for mammography practices in the US and is widely used for quality assurance. Figure 2 also presents a hierarchy of the 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 an expert-supplied 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, the 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, 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 [8]. Level 3 in Figure 2 shows the state-of-the-art in SRL techniques, illustrating how relevant fields from other rows (or other tables) can be incorporated in 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.

We use a three step process to construct aggregate features. First, we choose a field to aggregate over. Second, we select an aggregation function. Third, you need to decide which key or link to follow. This is known as a slot chain in a PRM terminology. In our database, two links or keys exist. The patient id field allows access to all the abnormalities for a given patient. We refer to this as aggregating on the patient level. The second key is given by looking at both the patient id and mammogram date, which returns all abnormalities for a patient on a specific mammogram. We refer to this as aggregating 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 2. 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 on a mammogram from May 2004. To calculate the aggregate on the patient level, we would average the size for all three abnormalities, which is .0367. To find the aggregate on the mammogram level for patient 1, he 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.

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 2 shows the key capability that will be introduced and evaluated in this paper: Using techniques from rule learning to learn a new *view*. 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 central to accurately predicting malignancy, but that are not explicit in the given database tables. One learned rule states that a change in the shape of the 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.

3 View Learning Framework

One can imagine a variety of approaches to perform view learning. Our closing section discusses a number of alternatives, including performing view learning and structure learning at the same time, in the same search. For the present work, we apply existing technology in a new fashion to obtain a view learning capability.

Any relational database can be naturally and simply represented in a subset of first-order logic [21]. Inductive logic programming (ILP) provides algorithms to learn rules, also expressed in logic, from such relational data [15], 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. We use ILP to learn rules to predict whether an abnormality is malignant. We treat each rule as an additional binary feature; true if the body, or condition, of the rule is satisfied, and otherwise false. We then run the Bayesian network structure learning algorithm, allowing it to use these new features in addition to the original features. Below is a simple rule, covering 48 positive examples and 123 negative examples:

Patient	Abnormality	Date	Mass Shape	…	Mass Size	Location	Benign/Malign
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	LL3	B
…	…	…	…		…	…	…

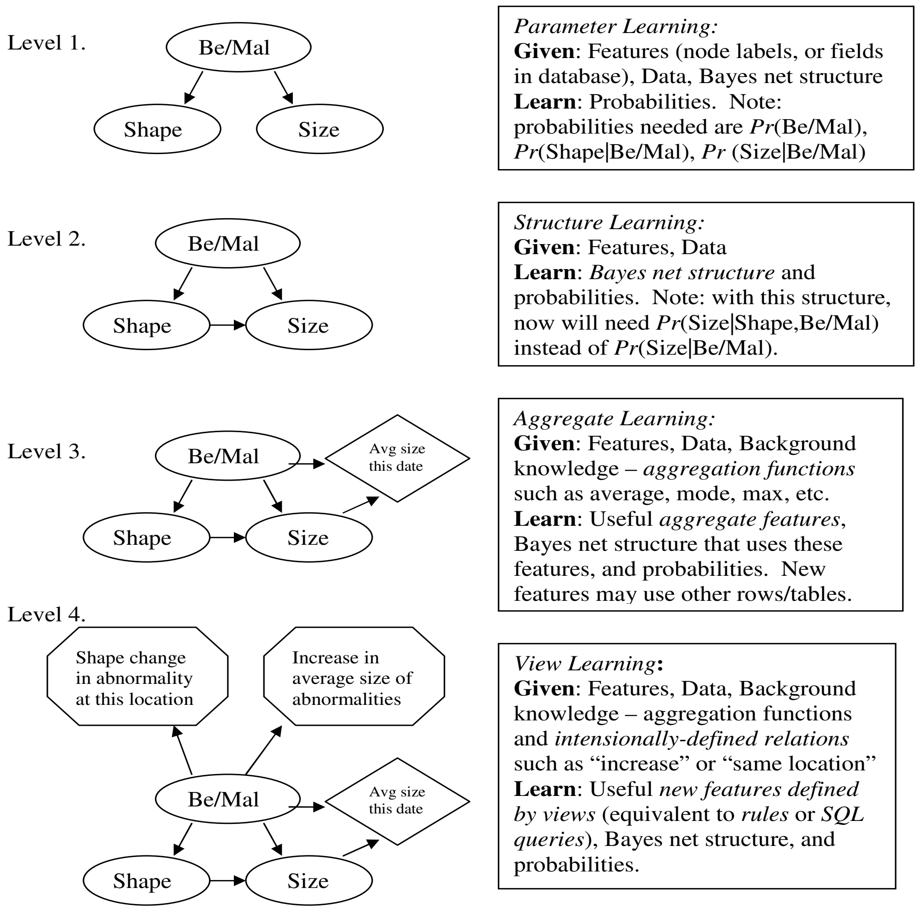

Figure 2: Hierarchy of learning types. Levels 1 and 2 are available through ordinary Bayesian network learning algorithms, Level 3 is available only through state-of-the-art SRL techniques, and Level 4 is described in this paper.

```
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.
```

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. 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.

4 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 using Inductive Logic Programming (ILP) to create features, which embody a new “view” of the database, adds a benefit over current SRL algorithms. We looked at adding two types of relational attributes, aggregate features and horn-clause (ILP) rules. Aggregate features represent summaries of abnormalities found either in a particular mammogram or for a particular patient. We performed a series of experiments, aimed at discovering if moving up in the hierarchy outlined in Figure 2 would improve performance. First, we tried to learn a structure with just the original attributes which performed better than the expert structure. This corresponds to Level 2 learning. Next, we added aggregate features to our network. Finally, we created new features using ILP. We investigated adding the features proposed by the ILP system as well as the aggregate to the network.

We experimented with a number of structure learning algorithms for the Bayesian Networks, including Naïve Bayes, Tree Augmented Naïve Bayes [7], and the sparse candidate algorithm [9]. However, we obtained 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 [10; 7].

4.1 Methodology

The dataset contains 435 malignant abnormalities and 65365 benign abnormalities. 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, 9 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 in the train set to determine the cut bin widths. When performing "view learning", we had two steps in the learning process. In the first part, 4 folds of data were used to learn the ILP rules. Afterwards,the remaining 5 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 are placed into the same fold for cross-validation. Another pitfall is that we may learn a rule that predicts an abnormality to be malignant based on properties of abnormalities in *later* mammograms. We never predict the status of an abnormality at a given date based on findings recorded with later dates.

We present the results of our first experiment using both ROC and precision recall curves. Because of our skewed class distribution, or 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 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 in the following manner. We treat 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.

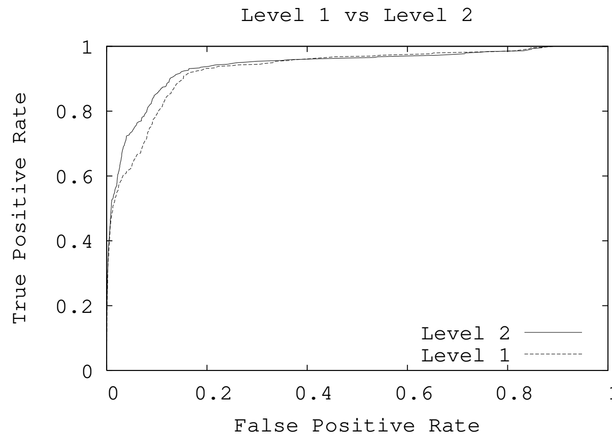

Figure 3: ROC Curves for Structure Learning. (Level 2)

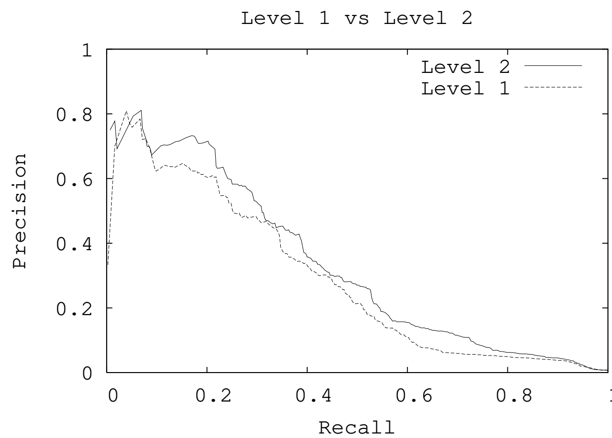

Figure 4: Precision Recall Curves for Structure Learning. (Level 2)

4.2 Results

The relational database containing the mammography data contains one row for each abnormality in a mammogram. Fields in this relational table include all those shown in the Bayesian network of Figure 1. Therefore it is straightforward to use existing Bayesian network structure learning algorithms to learn a possibly improved structure for the Bayesian network. We compared the performance of the best learned networks against the expert defined structure shown in Figure 1. We estimated the parameters of the expert structure from the dataset using maximum likelihood estimates with Laplace correction. Figure 3 shows the ROC curve for these experiments, and Figure 4 shows the Precision-Recall curves. Figure 6 shows the area under the precision-recall curve for the expert network (L1) and with learned structure (L2). 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 for every fold. We found the difference to be statistically significant with a 99% level of confidence.

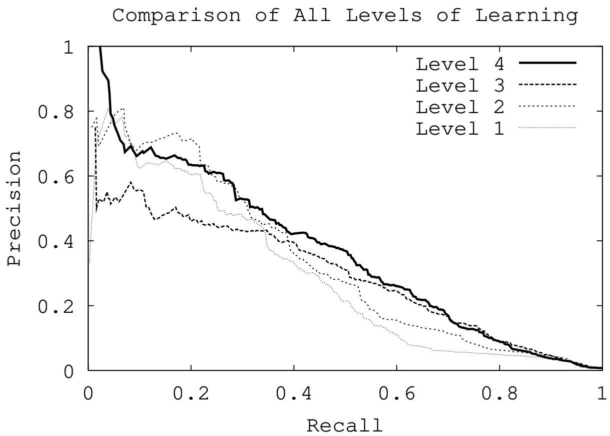

Figure 5: Precision Recall Curves for Each Level of Learning

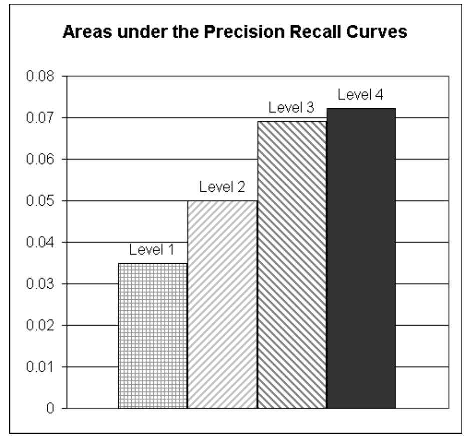

Figure 6: Area Under the Curve For Recalls Above 50%

With the help of a radiologist, we selected the numeric and ordered features in the database and computed aggregates for each of these features. 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 had predefined bin sizes, while for the other features we attempted to get equal numbers of abnormalities in each bin. For aggregations functions we used maximum and average. The aggregation introduced 108 new features.

Next we tested whether useful new fields could be computed by rule learning. Specifically, we used the ILP system Aleph [26] to learn rules predictive of malignancy. Several thousand distinct rules were learned for each fold, with each rule covering many more malignant cases than (incorrectly covering) benign cases. In order to obtain a varied set of rules, we ran Aleph using every positive example in each fold served as a seed for the search. We avoid the rule overfitting found by other authors [18] 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. We post processed the rules using a greedy algorithm, where we select the best scoring rule that covers new examples first. For each fold, the 50 best clauses were selected based on 3 criteria: (1) they needed to be multi-relational; (2) they needed to be distinct; (3) they needed to cover a significant number of malignant cases. The resulting views were added as new features to the database. Figure 5 includes a comparison of all levels of learning.

We can observe very significant improvements when adding multi-relational features. Both rules and aggregates achieved better performance. Aggregates do better for higher recalls, while rules do better for medium recalls. We believe this is because ILP rules are more accurate than the other features, but have limited coverage.

Level 4 performs as well as aggregates for high recalls, and close to ILP for medium recalls. According to the paired t-test the improvement of Level 4 over Level 2 is significant, using the area under the curve metric, at the 99% level. Meanwhile, Level 3 presents an improvement over Level 2, using the area under the curve metric, at the 97% confidence level.

Levels 1 and 2 correspond to standard propositional learning whereas levels 3 and 4 incorporate relational information. In this task, considering relational information is crucial for improving preformance. Furthermore, the process of generating the views in Level 4 has been useful to the radiologist as it has potentially identified novel correlations between attributes.

5 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[8]. Recently Heckerman, Meek and Koller have discussed extensions to PRMs and compared them to to other graphical models[11]. A statistical learning algorithm for probabilitistic logic representations was first given by Sato [23] and later, Cussens [5] proposed a more general algorithm to handle log linear models. Additionally, Muggleton [16] 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 [13] (BLPs), constraint logic programming with Bayes net constraints, or CLP($\mathcal{BN}$) [4], and Markov Logic Networks (MLNs) [22]. 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 [19] were the first to apply Naïve Bayes to combine clauses. Other work in this category was by Srinivasan and King [25], who use rules as extra features for the task of predicting biological activities of molecules from their atom-and-bond structures. Popescul et.al. [20] 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 [17] use aggregation to provide extra features on a multi-relational setting, and are close to our Level 3 setting. Knobbe et al. [14] 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 [18]. 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.

6 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, and that the most consistent improvement is obtained by using Level 4, both aggregates and new views.

We believe that further improvements 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. [27]. 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.

7 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. We would like to thank Rich Maclin for reading over several drafts of this paper. We would also like to thank the referees for their insightful comments.

References

[1] M.L. Brown, F. Houn, E.A. Sickles, and L.G. 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.

[2] E.S. Burnside, D.L. Rubin, and R.D. Shachter. A bayesian network for screening mammography. In *AMIA*, pages 106–110, 2000.

[3] E.S. Burnside, D.L. Rubin, and R.D. Shachter. Using a bayesian network to predict the probability and type of breast cancer represented by microcalcifications on mammography. *Medinfo*, 2004:13–17, 2004.

[4] V. Santos Costa, D. Page, M. Qazi, and J. Cussens. CLP($\mathcal{BN}$): Constraint logic programming for probabilistic knowledge. In *UAI-03*, pages 517–524, Acapulco, 2003.

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

[6] G. Ecklund. Shortage of qualified breast imagers could lead to crisis. *Diagn Imaging*, 22:31–33, 2000.

[7] Nir Friedman, David Geiger, and Moises Goldszmidt. Bayesian networks classifiers. *Machine Learning*, 29:131–163, 1997.

[8] Nir Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*. Stockholm, Sweden, 1999.

[9] Nir Friedman, I. Nachman, and D. Pe'er. Learning Bayesian Network Structure from Massive Datasets: The "Sparse Candidate" Algorithm. In *UAI-99*, pages 206–215, San Francisco, CA, 1999. Morgan Kaufmann Publishers.

[10] Dan Geiger. An entropy-based learning algorithm of bayesian conditional trees. In *UAI-92*, pages 92–97, San Mateo, CA, 1992. Morgan Kaufmann Publishers.

[11] D Heckerman, C Meek, and D Koller. Probabilistic Entity-Relationship Models, PRMs, and Plate Models, Technical Report MSR-TR-2004-30, Microsoft Research. Technical report, Microsoft Research, 2004.

[12] C.E. Kahn Jr, L.M. Roberts, K.A. Shaffer, and P. Haddawy. Construction of a bayesian network for mammographic diagnosis of breast cancer. *Comput Biol Med.*, 27:19–29, 1997.

[13] K. Kersting and L. De Raedt. Basic principles of learning bayesian logic programs, 2002.

[14] Arno J. Knobbe, Marc de Haas, and Arno Siebes. Propositionalisation and aggregates. In *PKDD01*, pages 277–288, 2001.

[15] S.H. Muggleton. Inductive Logic Programming. *New Generation Computing*, 8:295–318, 1991.

[16] S.H. Muggleton. Learning stochastic logic programs. *Electronic Transactions in Artificial Intelligence*, 4(041), 2000.

[17] Jennifer Neville, David Jensen, Lisa Friedland, and Michael Hay. Learning relational probability trees. In *KDD '03*, pages 625–630. ACM Press, 2003.

[18] Claudia Perlich and Foster Provost. Aggregation-based feature invention and relational concept classes. In *KDD '03*, pages 167–176, 2003.

[19] U. Pompe and I. Kononenko. Naive Bayesian classifier within ILP-R. In L. De Raedt, editor, *ILP95*, pages 417–436, 1995.

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

[21] R. Ramakrishnan and J. Gehrke. *Database Management Systems*. McGraw Hill, 2000.

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

[23] T. 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.

[24] E.A. Sickles, D.E. Wolverton, and K.E. Dee. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology*, 224:861–869, 2002.

[25] A. Srinivasan and R. King. Feature construction with inductive logic programming: A study of quantitative predictions of biological activity aided by structural attributes. In *ILP97*, pages 89–104, 1997.

[26] Ashwin Srinivasan. *The Aleph Manual*, 2001.

[27] Celine Vens, Anneleen Van Assche, Hendrik Blockeel, and Sašo Džeroski. First order random forests with complex aggregates. In *ILP*, pages 323–340, 2004.