J. Neville, D. Jensen, B. Gallagher and R. Fairgrieve (2003). Simple Estimators for Relational Bayesian Classifiers.
University of Massachusetts, Technical Report 03-04.

Simple Estimators for Relational Bayesian Classifiers

Jennifer Neville, David Jensen, Brian Gallagher and Ross Fairgrieve
Knowledge Discovery Laboratory, Department of Computer Science, University of Massachusetts,
140 Governors Drive, Amherst, MA 01003 USA
{jneville | jensen | bgallag | fairgr} @cs.umass.edu

This paper evaluates several modifications of the Simple Bayesian Classifier to enable estimation and inference over relational data. The resulting Relational Bayesian Classifiers are evaluated on three real-world datasets and compared to a baseline SBC using no relational information. The approach we call INDEPVAL achieves the best results. We use synthetic data sets to further explore performance as relational data characteristics vary.

1 Introduction

In this paper we present the Relational Bayesian Classifier (RBC), a modification of the Simple Bayesian Classifier (SBC) for relational data. The SBC offers good performance in many propositional domains and is simple to train and easy to understand. However it operates only with attribute-value data. The heterogeneous structure of relational data precludes direct application of a SBC model. We consider several approaches to modeling relational data with a Bayesian classifier and evaluate their performance on three data sets. An approach that follows the spirit of SBC and assumes attribute independence appears to work best.

A number of techniques have been developed to learn models of relational data [Dzeroski and Lavrac 2001]. The power of relational data lies in combining intrinsic information about objects in isolation with information about related objects and the connections between those objects. A technique modeling relational information should be able to perform at least as well as (and often better than) traditional attribute-value techniques modeling only intrinsic information. However, relational data present several challenges to learning algorithms. The data often have irregular structures and complex dependencies which contradict the assumptions of conventional modeling techniques.

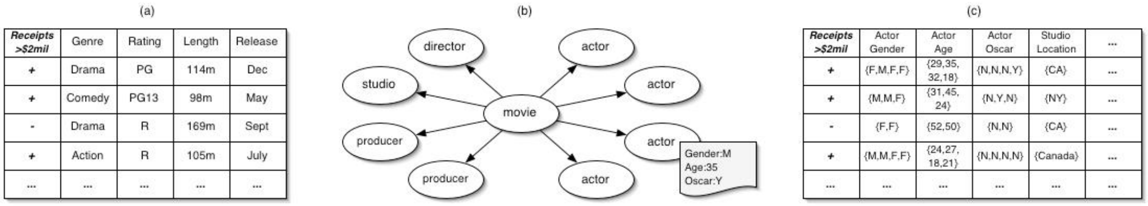
The simplicity of the SBC stems from its assumption that attributes are independent given the class – an assumption rarely met in practice. Domingos and Pazzani [1997] showed that the SBC performs well under zero-one loss even when its independence assumption is violated by a wide margin. Research investigating the effect of algorithm assumptions on performance has helped us to better understand the range of applicability for conventional techniques. This paper studies similar questions for relational data. We evaluate four different techniques on empirical data sets, comparing their accuracy and area under the ROC curve. We explore the effects of our approaches on simulated data sets, decomposing accuracy into bias and variance estimates [Friedman 1997, Domingos 2000]. Domingos and Pazzani [1997] showed that decreasing the bias associated with attribute dependencies is not necessarily the best approach to improving SBC performance on propositional data. Our experiments show that for relational data, performance improves as bias is decreased.

2 Modeling Relational Data

Most conventional classification techniques assume that data instances are recorded in homogeneous structures. Figure 1a shows a segment of propositional data stored in a table. Each row is a separate instance (e.g. movie) and each column records an attribute of the instances (e.g. movie genre). The attribute-value data are used to build a model of a class label (e.g. movie box office receipts).

Relational data have more information available with which to build better models, but the data often have complex structures which are more difficult to model. For example, the subgraph in figure 1b shows the data available to predict movie success (receipts>$2mil) in a relational dataset. In addition to information about the movie itself, there is information regarding the actors, directors, producers, and studios that participated in making the movie. For example, actors have gender, age and award information. Each movie subgraph may have a different number of related objects, resulting in diverse structures. For example, some movies may have 10 actors and others may have 1000. A relational classification technique needs to contend with heterogeneous data instances for both learning and inference.

There are a number of approaches to using conventional machine learning techniques on relational data. Transforming relational data to propositional form through flattening is perhaps the most common. One method transforms heterogeneous data into homogenous



(a)

Receipts >$2mil	Genre	Rating	Length	Release
+	Drama	PG	114m	Dec
+	Comedy	PG13	98m	May
-	Drama	R	169m	Sept
+	Action	R	105m	July
...	...	...	...	...

(c)

Receipts >$2mil	Actor Gender	Actor Age	Actor Oscar	Studio Location	...
+	{F,M,F,F}	{29,35, 32,18}	{N,N,N,Y}	{CA}	...
+	{M,M,F}	{31,45, 24}	{N,Y,N}	{NY}	...
-	{F,F}	{52,50}	{N,N}	{CA}	...
+	{M,M,F,F}	{24,27, 18,21}	{N,N,N,N}	{Canada}	...
...	...	...	...	...	...

Figure 1: Examples of (a) propositional data, (b) relational data represented as a subgraph, and (c) relational data decomposed by attribute.

records by aggregating multiple values into a single value (e.g. average actor age) or duplicating values across records (e.g. studio location is repeated across all associated movies). Other methods use relational learners to construct features that represent various characteristics of the examples [Kramer et al 2001]. Structured instances are transformed into homogenous sets of relational features. Any conventional machine learning technique can be applied to the data once they are flattened.

2.1 Relational Bayesian Classifiers

The RBC will decompose structured examples down to the attribute level. A heterogeneous subgraph is transformed into a homogenous set of attributes. For example, a movie subgraph contains information about a number of attributes including actor-age, actor-gender and studio-location. Each attribute contains a multiset of values for each subgraph. For example, the subgraph in figure 1b is transformed to the representation in figure 1c. This decomposition by attribute value follows the simple approach used in the SBC where attributes are assumed to be conditionally independent given the class. With this assumption, probability of class given an example can be calculated as the product of probabilities of attribute given class:

$$P(C=+\mid E)=\alpha\prod_{A_i}P(A_i=a_i\mid C=+)P(C=+)$$

We will hereafter use *P(A|C)* in place of *P(A=a|C=c)* for notational simplicity. Learning a SBC model then consists of estimating probabilities for each attribute given class. We will refer to techniques used to estimate these probabilities as *estimators*.

Estimation techniques for propositional data are straightforward. For discrete data, maximum-likelihood estimates can be achieved by counting. Kernel-density estimators are a good choice for continuous data [John & Langley 1995]. Estimation techniques for relational data need to model multisets of varying cardinality. For example, consider the segment of decomposed data in figure 1c. Each value for the actor-gender attribute is a different multiset (e.g. {F,M,F,F}). The dimensionality will be too high to model the sets directly. Many attribute sets will occur rarely so accurate probability estimates will be difficult to achieve. Estimators used in the RBC will need to model multisets in a more general way. We will evaluate two different approaches to estimating and three different approaches to inferences.

2.2 Multiset Estimators

Average Value

The average-value estimator (AvgVal) corresponds to flattening the data by averaging. During estimation, each multiset is replaced with its average (continuous attributes) or modal value (discrete attributes). Figure 2a shows an example subgraph from which P(A|C) will be estimated. Figure 2b shows the subgraph after flattening. The tuple consisting of class label and modal value {+,F} will be used in a standard maximum-likelihood estimator. The number of instances used for estimation is equal to the number of subgraphs in the data. Inference proceeds in a similar manner; probabilities are inferred from the average/modal set value:

$$P(+\mid E)=\alpha\,P(Mode=F\mid +)P(+)$$

We hypothesize that the AvgVal approach should perform well if the values in the multiset are highly correlated. In this situation, the set of values gives no more information than the average value. In addition, if the attribute distributions given class are hard to distinguish (i.e. close together) and cardinality of the multisets (i.e. degree) is high, then AvgVal will reduce estimation variance and possibly improve model accuracy.

Independent Value

The independent value estimator (IndepVal) assumes each value of a multiset is independently drawn from the same distribution. This estimator is designed to mirror the independence assumption of SBC – now in addition to attribute independence (e.g. between columns of figure 1c), there is also an assumption of attribute value independence (e.g. within columns of figure 1c). For estimation, each value of each set is considered to be an independent instance. Figure 2c shows the movie subgraph decomposed for estimation. The movie class label is duplicated and paired with each actor attribute value. Each pair is considered to be independent evidence. The number of instances available for estimation is now equal to the number of linked objects with the specified attribute.

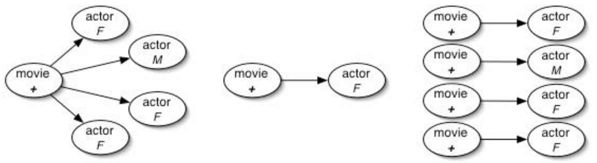


Figure 2: (a) Example subgraph transformed for estimation by (b) AVGVAL, and (c) INDEPVAL.

During inference the probability of each value is computed and multiplied into the overall probability independently:

$$P(+\mid E)=\alpha P(F\mid +)P(M\mid +)P(F\mid +)P(F\mid +)P(+)$$

INDEPVAL should perform well if the class label determines each attribute value independently – when there is no correlation among attribute values. In this situation, higher degree subgraphs will produce more evidence of the class and result in lower variance estimates. We expect this approach to perform in a manner similar to the SBC. Even in the absence of high degree subgraphs, INDEPVAL can use all available evidence to reduce variance. This may increase bias substantially when the assumption of independence is not met, but may not affect zero-one loss overall if the variance is low enough.

Average Probability

The third estimator uses average probability (AVGPROB) for inference. AVGPROB is an inference technique only. It uses probabilities estimated with INDEPVAL. The probability of each value is computed independently and then averaged over the multiset before being multiplied into the overall probability:

$$P(+\mid E)=\alpha\left[\frac{P(F\mid +)+P(M\mid +)+P(F\mid +)+P(F\mid +)}{4}\right]P(+)$$

AVGPROB computes an arithmetic average of probabilities instead of the geometric average computed by INDEPVAL. If the values in the multisets are highly dependent, geometric averaging will push the probabilities to extreme values. In this situation, arithmetic averaging should have lower bias. However, geometric averaging is more robust to irrelevant values in the multisets. Many irrelevant values will pull arithmetic averages toward the center, washing out the effects of the useful values. In this situation, AVGPROB may have higher bias. If only rare values of the multisets are predictive of the class, we expect AVGPROB and INDEPVAL to outperform AVGVAL.

3 Empirical Data Experiments

The experiments reported below are intended to evaluate two assertions. The first claim is that relational information can be used to improve model accuracy. We evaluate this claim by comparing the performance of RBC models using multiset estimators, with the performance of a SBC model using only intrinsic attributes. The SBC model receives only information about the objects being classified, no relational information is included. We call this approach INTRINSIC.

The second claim is that RBC models using INDEPVAL estimators will outperform RBC models using AVGVAL or AVGPROB estimators. We evaluate this claim by comparing the performance of each estimator.

To compare the four approaches, we recorded accuracy and area under the ROC curve [Provost et al 1998] on three real-world classification tasks. Area under the curve (AUC) measures classification accuracy over all possible class distributions and misclassification costs. The experiments use incremental ten-fold cross-validation [Cohen 1995] in order to compare estimator performance across a range of training set sizes. Training set sizes ranging from 10-90% of the data set are randomly chosen for each test set (10% of the data). Accuracy and AUC are averaged over the ten folds for each training set size. All models used Laplace correction for zero-values and kernel-density estimation for continuous attributes.

3.1 Classification Tasks

The first data set is drawn from the Internet Movie Database (www.imdb.com). We gathered a sample of 1383 movies released in the United States between 1995 and 2000. In addition to movies, the data set contains objects representing actors, directors, producers, and studios. In total, the data set contains 46,000 objects and 68,000 links. The learning task was to predict movie opening-weekend box office receipts. We discretized the attribute so that a positive label indicates a movie that garnered more than $2 million in opening-weekend receipts (*P(+)=0.55*). Nine attributes were supplied to the RBC models, including studio country, and actor birth-year.

The second data set is drawn from Cora, a database of computer science research papers extracted automatically from the web using machine learning techniques [McCallum et al 1999]. We selected a set of 4330 machine-learning papers along with associated authors, journals, books, publishers, institutions and cited papers. The resulting collection contains 11,500 objects and 26,000 links. The prediction task was to identify whether paper topic is *Neural Networks* (*P(+)=0.32*). Ten attributes were available to the RBC models, including the journal affiliation and paper venue.

The third data set is a relational data set containing information about the yeast genome at the gene and the protein level (www.cs.wisc.edu/~dpage/kddcup2001/). The data set contains information about 1243 genes and 1734 interactions among their associated proteins. The learning task was to predict whether or not a gene's functions include *Transcription* (*P(+)=0.31).* The RBC models used fourteen attributes for prediction, including gene phenotype, motif, and interaction type.

3.2 Results

Figure 3 shows accuracy and AUC results for each of the four models on the three classification tasks. On the IMDB, the INDEPVAL and AVGPROB models have higher accuracy than the AVGVAL and INTRINSIC models. However, INDEPVAL's AUC results are far superior to any of the other approaches. In this data set AVGVAL performs significantly worse than the other RBC models. This indicates that flattening relational data and applying propositional models may not always be a good approach.

On the Cora classification task, INDEPVAL also shows superior performance in both accuracy and AUC. Again, the increase is more pronounced in AUC. This suggests

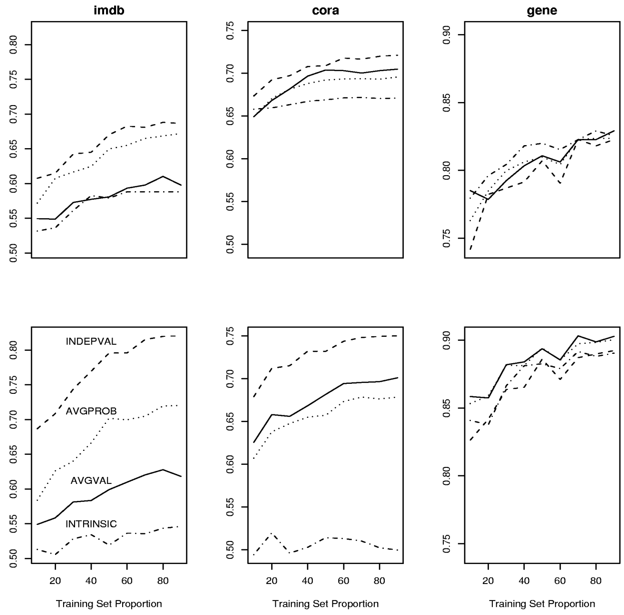


Figure 3: Results of empirical data experiments for IMDB, Cora, and Gene databases.

that INDEPVAL produces better rankings of probability estimates than the other approaches. If this is the case, INDEPVAL should perform best with respect to squared loss as well. AVGVAL and AvgProb perform equivalently, in both accuracy and AUC. The AUC results for INTRINSIC indicate that its performance is no better than random.

The gene data set is the only one where all approaches perform equivalently. INTRINSIC appears to have slightly better accuracy overall, but AVGVAL dominates slightly in AUC. In this classification task, the relational models do no better than the propositional model.

We used two-tailed, paired t-tests to assess the significance of the results obtained from the ten-fold cross-validation trials. The t-tests are conducted on the accuracy and AUC results from the cross-validation trials which used 90% of the data for training. The null hypothesis is that there is no difference between two approaches; the alternative is that there is a difference between two approaches. The resulting p-values are reported in Table 1 below as a heuristic guide to significance. Dietterich [97] reports that paired t-tests on ten-fold cross-validation results can make at most twice the target level of errors in which the null hypothesis is incorrectly rejected. However, the p-values are low enough that this bias should not alter our conclusions. The results support our conclusions above. INDEPVAL is the superior approach for IMDB and Cora, and performs equivalently for Gene.

4 Synthetic Data Experiments

Common characteristics of relational data could be affecting estimator performance. Relational data sets often exhibit concentrated linkage with a number of high degree objects. For example, many papers in Cora link to a few journals, and many movies in the IMDB link to a few studios. These high degree objects may reduce the estimator variance if the linked objects are used together for classification (e.g. use related movies to predict an attribute of studios). On the other hand, the same connections could increase variance if the linked objects create dependencies across examples (e.g. a single journal is used separately to classify each associated paper). Even with objects of low to moderate degree, linkage can create complex dependencies among attribute values. Attribute values can exhibit uniformity among objects that share a common neighbor. For example, in the gene data, proteins located in the same place in the cell (e.g. cell wall) often have highly correlated functions (e.g. cell growth). Dependencies such as these can be a source of increased variance. We will use synthetic data to explore the effects of linkage and attribute correlation on estimator performance.

	IMDB		Cora		Gene	
	Acc	AUC	Acc	AUC	Acc	AUC
INDEPVAL vs. AVGVAL	0.0003	0.0000	0.0023	0.0000	0.4532	0.2074
INDEPVAL vs. AVGPROB	0.4122	0.0003	0.0006	0.0000	0.8708	0.1210
INDEPVAL vs. INTRINSIC	0.0057	0.0000	0.0001	0.0000	0.6586	0.1682

Table 1: P-values of significance tests on 10-fold CV results.

4.1 Methodology

Our synthetic data sets are comprised of bipartite graphs, each containing a single core object (e.g. a movie) linked to zero or more peripheral objects (e.g. actors). Note that each actor links to exactly one movie. Each movie has a single binary attribute, C={+,-}, representing its class (e.g. receipts>$2mil). Likewise, each actor has a single binary attribute, A={1,0} (e.g. gender). Some sample graphs are shown in figure 4. The degrees of the graphs in each data set are distributed normally with mean equal to |actors| / |movies|. The default parameters for the experiments were 100 movies, 500 actors, *P(+)=0.5*, and *P(A=1|C=+)=P(A=0|C=-)=0.75*. Variations from these defaults are described for each experiment below.

The learning task was to predict the class label for each movie. Experiments were performed for each of the three RBC estimators. We measured average accuracy of each RBC model across 100 pairs of training/test sets.

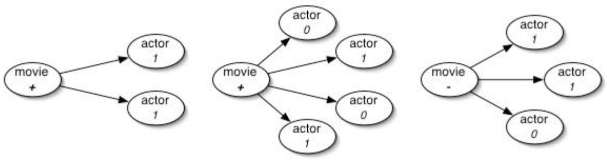


Figure 4: Sample synthetic data subgraphs.

Bias and variance were measured for each approach using the decompositions defined for zero-one loss by Domingos [2000]. Bias and variance estimates are calculated for each test example using 100 different training sets and averaged over the entire test set. This was repeated for 100 test sets to calculate average test set bias and variance. The results of the synthetic experiments are presented in figure 5.

4.2 Results

Figure 5a shows an experiment in which the total number of actors in each data set was varied from 100 to 1000. Default settings were used for all other parameters. In this experiment AVGVAL and INDEPVAL are indistinguishable. The accuracy of the AVGVAL and INDEPVAL estimators increases with graph degree through 1000 actors while the accuracy of the AVGPROB estimator levels off around 500 actors. AVGVAL and INDEPVAL show lower bias as degree increases, whereas the bias of the AVGPROB estimator remains relatively constant. For all three estimators, degree reduces variance. The three estimators have comparable variance so the increase in accuracy can be attributed to lowering bias.

Figure 5b shows an experiment in which the correlation among linked actor attribute values is varied. Default settings were used for all other parameters. Again, AVGVAL and INDEPVAL are indistinguishable. The accuracy of the AVGVAL and INDEPVAL estimators decreases as correlation increases while the accuracy of the AVGPROB estimator remains approximately constant. The variance of all three estimators is very low and seems to depend very little on attribute correlation, so again the decrease in accuracy for AVGVAL and INDEPVAL can be attributed to estimator bias which increases with attribute correlation.

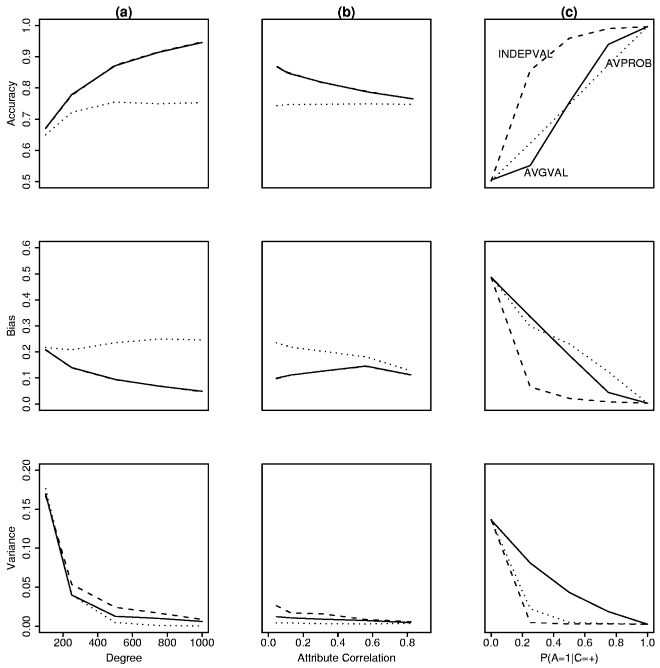


Figure 5: Results of synthetic data experiments.

Figure 5c shows an experiment in which *P(A=1|C=+)* is varied from {0,1} while holding *P(A=1|C=-)* constant at 0. Default settings were used for all other parameters. This is the first experiment to show a difference in performance between AVGVAL and INDEPVAL illustrating performance in situations where rare attribute values determine the class. INDEPVAL and AVGPROB both show lower variance than AVGVAL but AVGPROB has much lower accuracy. Since INDEPVAL shows lower bias than either of the other estimators we can attribute its higher accuracy to this reduction in bias.

The above experiments were repeated for overall P*(A=1|C=+)* values other than 0.75. The relative performance of estimators remained substantially the same across all correlation levels.

5 Discussion

Structural characteristics of relational data affect performance of multiset estimators in a number of ways. Large multisets calculated from objects of high degree can be useful in reducing variance. However, our experiments only examine the effects of high degree objects within a single subgraph. Linkage across subgraphs may produce dependencies that result in higher variance.

Correlation among attribute values seems to have less effect on the bias and variance of estimators. Yet, AVGVAL and INDEPVAL accuracies are adversely effected by higher correlations. We expected AVGPROB to be robust to attribute correlation due to its arithmetic averaging. However, it is surprising that it doesn't outperform the other approaches. This may be a result of decreased bias in AVGVAL and INDEPVAL due to high degree. Future work will examine possible interaction effects of linkage and correlation.

Overall, INDEPVAL estimators have lowest bias and variance over a wide range of synthetic data sets. AVGVAL has low variance over a number of data sets, but it was easy to identify situations in which AVGVAL would be a biased estimator. Both estimators have lower variance as degree increases. We can infer that INDEPVAL's superior performance on the real-world classification tasks is a result of lower overall bias. AVGPROB appears to be biased over a number of relational data configurations. However, it achieves accuracies comparable to INDEPVAL on the IMDB experiments. This reveals that our synthetic data experiments have not clearly identified the circumstances in which AVGPROB is a good approach to estimation.

6 Related Work

The Inductive Logic Programming community have studied the issues of modeling relational data for many years. 1BC is a first-order Bayesian classifier for relational data which applies dynamic propositionalization [Flach and

Lachiche 1999]. Examples consist of objects and their relational neighborhood. 1BC generates a set of first-order conditions which evaluate the attribute values of various items in the examples. The initial work on 1BC discusses a number of approaches to decomposing structured examples into sets of items and attribute values. Approaches to modeling lists and sets of attribute values are presented but they are not used in the models.

More recent work has examined 1BC2 models which use complex list- and set-valued estimators [Lachiche and Flach 2002]. Complex estimators are used to decrease overall bias of the models in light of the high dimensionality of set-valued attributes. However, the performance of 1BC2 models with set-valued estimators is not impressive – the results are generally indistinguishable from those of 1BC.

Flach and Lachiche do not explore the bias and variance tradeoffs of the two approaches. The experiments reported above show that estimators with reduced bias have higher accuracies. However, these estimators tend to have low variance as well. Future work needs to explore bias and variance tradeoffs more fully with both simple and complex multiset estimators.

7 Conclusions

We have identified a simple approach to estimation for relational data. Adhering to the SBC's spirit of simplicity, an RBC model assuming conditional independence of both the attributes and the multiset attribute values (INDEPVAL) is successful in a variety of real-world and synthetic classification tasks.

INDEPVAL estimation performs at least as well as AVGVAL estimation in all tasks, and significantly better in some. AVGVAL estimation is essentially the same as dynamically flattening relational data and applying a propositional learner. An RBC using INDEPVAL estimation should now be considered as superior to flattening relational data.

On two real-world classification tasks, RBC models perform significantly better than SBC models without the relational information. On a third task there is no difference between the approaches. We don't lose anything by modeling relational data with RBC's, since they share the SBC's robustness to irrelevant data.

In addition, the RBC model with INDEPVAL estimation is easy to implement and efficient to train and apply. It should be a good baseline against which to evaluate other relational learning techniques.

Future work will include further investigation of the effects of relational data characteristics on estimator performance. The structure of relational data (linkage and attribute correlation) can affect estimator bias and variance. A RBC model which selects an appropriate estimator for each attribute may outperform an RBC model using INDEPVAL for all attributes. We will also investigate the performance of more complex estimators (e.g. kernel density estimators for multisets).

Acknowledgments

Members of the Knowledge Discovery Laboratory provided valuable comments and suggestions on earlier drafts of this work. This research is supported under a National Science Foundation Graduate Research Fellowship and by DARPA and AFRL, AFMC, USAF under contract numbers F30602-00-2-0597 and F30602-01-2-0566. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of the DARPA, the AFRL or the U.S. Government.

References

[Cohen 1995] Cohen, P. Empirical Methods for Artificial Intelligence. The MIT Press, 1995.

[Dietterich 97] Dietterich, T. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10: 1895-1923, 1998.

[Domingos 2000] Domingos, P. A Unified Bias-Variance Decomposition for Zero-One and Squared Loss. *Proceedings of the 17th National Conference on Artificial Intelligence*, AAAI Press, 2000.

[Domingos and Pazzani 1997] Domingos, P. and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103-130, 1997.

[Dzeroski and Lavrac 2001] Dzeroski, S. and N. Lavrac, editors. Relational Data Mining. Springer-Verlag, 2001.

[Flach and Lachiche 1999] Flach, P. and N. Lachiche. 1BC: A first-order Bayesian classifier. Proceedings of the 9th International Workshop on Inductive Logic Programming, pp. 92--103, 1999.

[Friedman 1997] Friedman, J. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1:55-77, 1997.

[Kramer et al 2001] Kramer, S., N. Lavrac and P. Flach. Propositionalization approaches to relational data mining. In S. Dzeroski and N. Lavrac, eds. *Relational Data Mining*. pp 262-291, Springer-Verlag, 2001.

[John and Langley 1995] John G., and P. Langley. Estimating Continuous Distributions in Bayesian Classifiers. *Proceedings of the 11th Conference on Uncertainty on Artificial Intelligence*, 1995.

[Lachiche and Flach 2002] Lachiche, N. and P. Flach 1BC2: a true first-order Bayesian Classifier. *Proceedings of the 12th International Conference on Inductive Logic Programming,* 2002.

[McCallum et al 1999] McCallum, A., K. Nigam, J. Rennie and K. Seymore. A Machine Learning Approach to Building Domain-specific Search Engines. *In Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pp. 662-667, 1999.

[Provost et al 1998] Provost, F., and T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. *Proceedings of the 15th International Conference on Machine Learning*, 1998.