Simple Estimators for Relational Bayesian Classifiers

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

Abstract

In this paper we present the Relational Bayesian Classifier (RBC), a modification of the Simple Bayesian Classifier (SBC) for relational data. There exist several Bayesian classifiers that learn predictive models of relational data, but each uses a different estimation technique for modeling heterogeneous sets of attribute values. The effects of data characteristics on estimation have not been explored. We consider four simple estimation techniques and evaluate them on three real-world data sets. The estimator that assumes each multiset value is independently drawn from the same distribution (INDEPVAL) achieves the best empirical results. We examine bias and variance tradeoffs over a range of data sets and show that INDEPVAL's ability to model more multiset information results in lower bias estimates and contributes to its superior performance.

1. Introduction

This paper presents a modification of the Simple Bayesian Classifier (SBC) for relational data. 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. However, the data often have irregular structures and complex dependencies, which contradict the assumptions of conventional modeling techniques. In particular, the heterogeneous structure of relational data precludes direct application of a SBC model, which operates on attribute-value data. We consider several approaches to modeling data with a relational Bayesian classifier (RBC) and evaluate performance on three data sets. The approach that follows the spirit of SBC and assumes conditional attribute value independence appears to work best. (See [9] for an expanded version of this paper.)

The simplicity of the SBC stems from its assumption that attributes are independent given the class—an assumption rarely met in practice. Research investigating the effects of this assumption on performance has helped to better understand the range of applicability of the SBC. For example, Domingos and Pazzani [2] showed that the SBC performs well under zero-one loss even when the independence assumption is violated by a wide margin. This paper studies similar questions for relational data. We empirically evaluate four different techniques on several real-world data sets. We explore the techniques on simulated data sets, decomposing loss into bias and variance estimates [1]. Our experiments show that characteristics of relational data can bias certain estimators and that using estimators with decreased bias improves model performance.

2. Modeling Relational Data

Relational data violate two assumptions of conventional classification techniques. First, algorithms for propositional data assume that the data instances are recorded in homogeneous structures (e.g. a fixed set of fields for each object), but relational data "instances" are consist of sets of heterogeneous records. Second, algorithms for propositional data assume that the data instances are independent and identically distributed (i.i.d.), but relational data have dependencies both through direct relations and through chaining multiple relations together. In this paper, we evaluate simple algorithms for learning models of data sets with heterogeneous instances. We do not attempt to exploit dependencies among related instances.

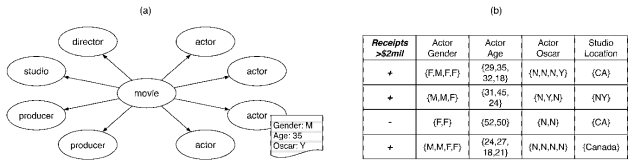
Relational data often have complex structures that are more difficult to model than homogeneous instances. For example, in order to predict the box-office success of a movie, a relational model might consider not only the attributes of the movie, but also attributes of the movie's actors, director, producers, and the studio that made the movie. A model might even consider attributes of indirectly related objects such as other movies made by the director. Each movie 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. When trying to predict the value of an attribute based on the attributes of related objects, a relational classification technique must consider *multisets* of attribute values. For example, we might model the likelihood of movie success given the multiset of gender values from the movie's actors.

There are a number of approaches to modeling sets of attribute values. *Propositionalization* is a common

J. Neville, D. Jensen, and B. Gallagher. Simple Estimators for Relational Bayesian Classifiers. In Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM'03), pp. 609-612, 2003.

technique to transform heterogeneous data instances into homogenous records, mapping sets of values into single values with aggregation functions. A second approach is to treat the set of values independently and aggregate the resulting probability distributions using combining rules such as *noisy-or* or *average* [4]. A third approach is to model the sets directly with multinomials [7] or complex set-valued estimators [6].

This paper considers four estimation techniques from the range of approaches outlined above. Recent work has demonstrated the feasibility of these approaches for statistical models of relational data, but the choice of technique for any one model has been approached in a relatively ad-hoc manner. A thorough understanding of the effects of relational data characteristics on estimator performance will improve parameter estimation for relational data and should inform the development of more complex statistical models.



(b)

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. Relational data represented as (a) a subgraph, and (b) decomposed by attribute.

3. Relational Bayesian Classifiers

The RBC represents heterogeneous examples as homogenous sets of attribute multisets. For example, a movie subgraph contains information about a number of related objects, such as actors and studios (e.g. Figure 1a). Transformed examples contain a multiset of values for each attribute, such as actor-age and studio-location (e.g. Figure 1b). This enables a SBC approach, where learning a model consists of estimating conditional probabilities for each attribute. However, estimation techniques for these data will need to model multisets of varying cardinality and high dimensionality. We refer to techniques used to estimate these probabilities as *estimators*. We will evaluate three approaches to estimation and four approaches to inference.

Average Value—The average value estimator (AVGVAL) corresponds to propositionalizing the data by averaging. During estimation, each multiset is replaced with its average value (for continuous attributes) or modal value (for discrete attributes). The average values are used in a standard maximum-likelihood estimator and probabilities are inferred from average/modal values as well. AVGVAL estimators are commonly used in probabilistic relational models (PRMs) to model dependencies where the "parent" consists of a set of attribute values [3]. We hypothesize that AVGVAL should perform well if the multiset values are highly correlated, so the multiset is no more informative than the average.

Random Value—The random value estimator (RANDVAL) is similar to AVGVAL. However, instead of deterministically choosing the most prevalent value from the set, RANDVAL chooses a representative value stochastically. This allows the estimation to differentiate between relatively uniform sets of values and highly skewed sets. This approach is equivalent to the *stochastic-mode* aggregation used in PRMs for classification [10]. Although RANDVAL may be more sensitive to the distribution of values in the sets, it may also experience greater variance if multiset values are distributed uniformly over a large range.

Independent Value—The independent value estimator (INDEPVAL) assumes each multiset value is independently drawn from the same distribution. This estimator is designed to mirror the independence assumption of SBC—now in addition to attribute independence, there is also an assumption of *attribute value* independence given the class. INDEPVAL models the multiset with a multinomial distribution where the size of the set is independent of the class. INDEPVAL should perform well if the multiset can be used to reduce variance, when there is little correlation among attribute values.

Average Probability—The fourth estimator (AVGPROB) aggregates probability distributions. It is an inference technique only (INDEPVAL is used for estimation). During inference, each multiset value's probability is computed independently and then the set of probabilities is averaged. This approach is one of the combining rules used in Bayesian logic programs (BLPs) to integrate probabilities into logic programs [4]. AVGPROB computes an arithmetic average of probabilities. If the set values are dependent, geometric averaging (used in INDEPVAL) will push the probabilities to extreme values. However, geometric averaging is more robust to irrelevant values, which pull arithmetic averages toward the center and wash out the effects of the useful values.

4. Empirical Data Experiments

The experiments reported below evaluate the claim that RBC models using INDEPVAL estimators will outperform RBC models using AVGVAL, RANDVAL or AVGPROB estimators. We compare the performance of each estimator on three real-world classification tasks. To compare the approaches, we recorded accuracy and area under the ROC curve using ten-fold cross-validation.

4.1. Classification Tasks

The first data set, drawn from the Internet Movie Database (IMDb) (www.imdb.com), is a sample of all movies released in the United States from 1996 to 2001, with opening weekend box-office receipt data. The sample contains 1383 movies and related actors, directors, producers, and studios. The task was to predict whether a movie made more than $2mil in opening weekend

receipts (*P(+)=0.45*). Nine attributes were supplied to the models, including studio country and actor birth-year.

The second data set, drawn from Cora [8], is a sample of 4330 machine-learning papers and associated authors, journals/books, publishers, and cited papers. The task was to predict whether a paper's topic is *Neural Networks* (*P(+)=0.32*). Ten attributes were available to the models, including journal affiliation and paper venue.

The third data set contains information about 1243 genes in the yeast genome and 1734 interactions among their associated proteins (www.cs.wisc.edu/~dpage/kddcup2001/). The task was to predict whether a gene's functions include *Transcription* (*P(+)=0.31*). Fourteen attributes where supplied to the models, including gene phenotype, motif, and interaction type.

4.2. Results

Figure 2 shows AUC results for each of the models on the three classification tasks, averaged over the ten folds. Accuracy results are comparable [9]. We used two-tailed, paired t-tests to assess the significance of the ten-fold cross-validation results, comparing INDEPVAL to each of the other estimators. Asterisks in Figure 2 indicate a significant difference in performance compared to INDEPVAL (p-value < 0.001).

On the IMDb and Cora classification tasks, INDEPVAL's AUC results are superior to any of the other approaches. The performance of AVGVAL and RANDVAL indicates that propositionalizing relational data (even stochastically) to apply conventional models may not always be a good approach. On the Gene task, all approaches perform equivalently.

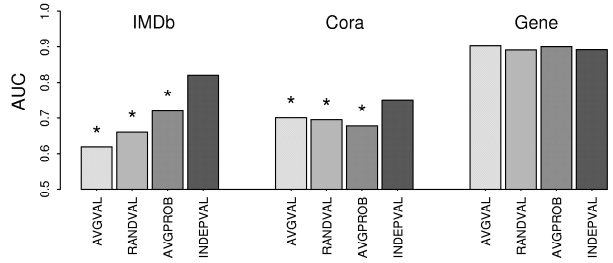


Figure 2: Results of empirical data experiments for IMDb, Cora, and Gene databases.

5. Synthetic Data Experiments

We use synthetic data to explore the effects of linkage, attribute correlation, and multiset distributions on estimator performance. Relational data sets often exhibit concentrated linkage, where certain object types have a large number of relations. For example, papers in Cora link to a few journals, and movies in the IMDb link to a small number of studios. Uniformity among attribute values of objects that share a common neighbor is also common in relational data. For example, in the gene data, proteins located in the same place in the cell often have highly correlated functions.

5.1. 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 binary class label, $C=\{+,-\}$, and each actor has a binary attribute, $A=\{1,0\}$. The number of actors per movie is distributed normally with mean equal to |actors|/|movies|. The default experimental parameters were 100 movies, 500 actors, $P(C=+)=0.5$, and $P(A=1|C=+)=P(A=0|C=-)=0.75$. Variations from these defaults are described for each experiment below.

We measured average zero-one loss and squared-loss for each RBC estimator across 100 pairs of training/test sets and decomposed loss into bias and variance [1]. Bias and variance estimates were calculated for each test example using 100 different training sets and averaged over the entire test set. This was repeated for 100 test sets and averaged. The zero-one loss results are presented in Figure 3. Squared-loss results are similar [9].

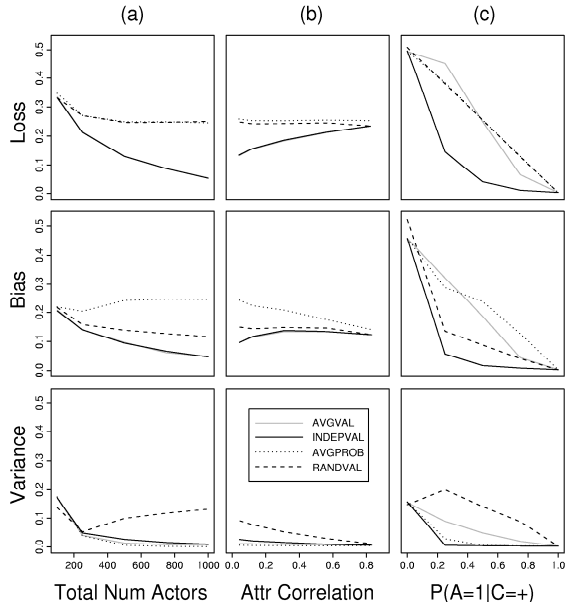


Figure 3: Results of synthetic data experiments.

5.2. Results

The experiment shown in Figure 3a varied the total number of actors in each data set from 100 to 1000. In this experiment AVGVAL and INDEPVAL are nearly indistinguishable, as are AVGPROB and RANDVAL. For all estimators except RANDVAL, increasing degree reduces variance. This was expected, as the variance of the random value selection increases with set size. AVGPROB's arithmetic averaging cannot exploit the extra information in larger sets, which results in higher bias.

The experiment in Figure 3b varied the correlation among linked actor attribute values from [0.05,0.85]. Again, AVGVAL and INDEPVAL are indistinguishable. As attribute correlation increases, the bias of the INDEPVAL estimator increases, indicating that INDEPVAL's probability estimates may be skewed in data with high attribute correlations.

The experiment in Figure 3c varied *P(A=1|C=+)* from [0,1] while holding *P(A=1|C=-)* constant at 0. This is the first experiment to show a difference between AVGVAL and INDEPVAL, illustrating performance when rare attribute values determine the class. Since INDEPVAL shows lower bias than either of the other estimators we can attribute its higher accuracy to this reduction in bias.

Given these results, the relative strength of INDEPVAL appears to lie in the estimator's ability to make use of rare attribute values, as well as multiple predictive values within a multiset. To determine if these types of multisets occur in practice, we examined multisets from the IMDb. We calculated the correlation of each attribute value with the class label using chi-square, assessed significance after adjusting for multiset size [5], and then determined the number of unique correlated attribute values per movie. Figure 4 shows the frequency distribution of these counts across movies for three example attributes. A large number of movie subgraphs have more than one unique attribute value correlated with the class. In this situation, estimators that can capture more multiset information (e.g. INDEPVAL) will outperform estimators that propositionalize to a single value (e.g. AVGVAL).

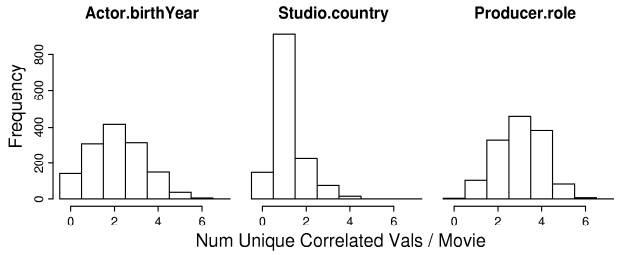


Figure 4: Count of unique significantly correlated values in each subgraph, for three attributes in the IMDb.

6. Conclusions

We have identified a simple approach to estimation for relational data. Adhering to the spirit of SBC simplicity, the RBC model that assumes conditional independence of both attributes and multiset attribute values (INDEPVAL) is successful in a variety of real-world classification tasks. This model is easy to implement and efficient to use, making it a good baseline for evaluation of more complex relational learning techniques.

INDEPVAL estimators have low bias and variance over a wide range of synthetic data sets. AVGVAL has low variance over a number of conditions, but it is easy to identify situations in which AVGVAL is a biased estimator. We can infer that INDEPVAL's superior performance on the real-world classification tasks is a result of lower overall bias—due to its ability to exploit information contained in both rare values and multiple correlated values within the sets. AVGPROB appears to be biased over a number of data sets, but it performs quite well on the IMDb. This reveals that our synthetic data experiments have not clearly identified the circumstances in which AVGPROB is a good approach to estimation.

Future work will include further analysis of the effects of relational data characteristics on complex multiset estimators [e.g. 6] and development of models that select attribute estimators based on data characteristics.

7. Acknowledgments

We thank Ross Fairgrieve for his contributions to an earlier draft of this work. This research is supported by DARPA and AFRL, AFMC, USAF under contract numbers F30602-00-2-0597 and F30602-01-2-0566.

8. References

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

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

[3] Getoor, L., N. Friedman, D. Koller, and A. Pfeffer. Learning probabilistic relational models. *Relational Data Mining*, Dzeroski and Lavrac, Eds., Springer-Verlag, 2001.

[4] Kersting, K. and L. De Raedt. Basic principles of learning Bayesian logic programs. Tech Report 174, University of Freiburg, Germany, June 2002.

[5] Jensen, D., J. Neville and M. Hay. Avoiding bias when aggregating relational data with degree disparity. *Proc. of the 20th International Conf. on Machine Learning*, 2003.

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

[7] McCallum, A. and K. Nigam. A comparison of Event Models for Naive Bayes Text Classification. In *AAAI-98 Workshop on Learning for Text Categorization*, 1998.

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

[9] Neville, J., D. Jensen and B. Gallagher. Simple Estimators for Relational Bayesian Classifiers. University of Massachusetts, Technical Report 03-04.

[10] Taskar, B., E. Segal, and D. Koller. Probabilistic Classification and Clustering in Relational Data. *Proceedings of the 17th Intl Joint Conference on Artificial Intelligence*, 2001.