AN ADAPTIVE SAMPLING ALGORITHM TO IMPROVE THE PERFORMANCE OF CLASSIFICATION MODELS

Soroosh Ghorbani
Computer and Software Engineering Department, Montréal Polytechnique, Canada
Soroosh.Ghorbani@Polymtl.ca

Michel C. Desmarais
Computer and Software Engineering Department, Montréal Polytechnique, Canada
Michel.Desmarais@Polymtl.ca

ABSTRACT

Given a fixed number of observations to train a model for a classification task, a Selective Sampling design helps decide how to allocate more, or less observations among the variables during the data gathering phase, such that some variables will have a greater ratio of missing values than others. Previous work has shown that selective sampling based on features' entropy can improve the performance of some classification models. We further explore this heuristic to guide the sampling process on the fly, a process we call Adaptive sampling. We focus on three different classification models, Naïve Bayes (NB), Logistic Regression (LR) and Tree Augmented Naive Bayes (TAN), and train them on binary attributes datasets and use a 0/1 loss function to assess their respective performance. We define three different schemes of sampling: 1-Uniform (random samples) as a baseline, 2-Low entropy (greater sampling rate for low entropy items) and 3-High entropy (greater sampling rate for higher entropy items). Then, we propose an algorithm for Adaptive Sampling that uses a small seed dataset to extract the initial entropies and randomly samples feature observations based on the three different schemes. The performance of the combination of schemes and models is assessed on 11 different datasets. The results from 100 fold cross-validation show that Adaptive Sampling based on scheme 3 improves the performance of the TAN model in all but one of the datasets, with an average improvement of 12-14% in RMSE reduction. However, for the Naive Bayes classifier, scheme 2 improves the classification by a factor of 6-12% (with one data set exception). Finally, for Logistic Regression, no clear pattern emerges.

KEYWORDS

Adaptive Sampling, Entropy, Classification, Prediction Performance

1. INTRODUCTION

When the training of a classifier has a fixed number of observations and missing values are unavoidable, we can decide to allocate the observations differently among the variables during the data gathering phase. We refer to this situation as Selective Sampling.

One important example is Computerized Adaptive Testing (CAT). Student test data are used for training skill mastery models. In such models, test items (questions) represent variables that are used to estimate one or more latent factors (skills). For a number of practical reasons, the pool of test items often needs to be quite large, such as a few hundreds and even thousands of items. However, for model training, it is impractical to administer a test of hundreds of questions to examinees in order to gather the necessary data. We are thus forced to administer a subset of these test items to each examinee, leaving unanswered items as missing values. Hence, adaptive testing is a typical context where we have the opportunity to decide which items will have a higher rate of missing values, and the question is whether we can allocate the missing values in a way that will maximize the model's predictive performance?

Although CAT is a typical application domain where we can apply Selective Sampling, any domain which offers a large number of features from which to train a model for classification or regression purpose is a good candidate for Selective Sampling. The data sets used in this experiment represent examples of such domains (see Table 1 for a full list). Note that for this study, we limit our scope to binary target variables and binary attributes.

Table 1. Datasets at a Glance

Dataset	Attributes	Instances	Mean Entropy of the attributes	Success Rate
SPECT Heart	22+Class	267	0.85	41%
England	100+Class	1003	0.22	18%
Ketoprostaglandin-f1	100+Class	1003	0.17	6%
Brain Chemistry	100+Class	1003	0.11	7%
Creatine-kinase	100+Class	1003	0.15	8%
Ethics	100+Class	1003	0.16	12%
Fundus-oculi	100+Class	1003	0.13	14%
Heart Valve Prosthesis	100+Class	1003	0.24	19%
Larynx	100+Class	1003	0.07	7%
Mexico	100+Class	1003	0.11	5%
Uric-Acid	100+Class	1003	0.10	6%

In previous work [4, 6], we established that selective sampling based on entropy can improve the performance of classifiers. However, the algorithms assumed the information about the entropy is available prior to the selective sampling process, which is not the case in reality. This study extends this work to assess the performance of the selective sampling heuristics without assuming this prior information, a process we refer to as Adaptive Sampling.

2. PLANNED MISSING DATA DESIGNS

Selective Sampling is analogous to the notion of planned missing data designs used in psychometry and other domains. In planned missing data designs, participants are randomly assigned to conditions in which they do not respond to all items. Planned missing data is desirable when, for example:

- long assessments can reduce data quality, a situation that arises frequently when data is gathered from a human subject or some source for which a measurement has an effect on posterior measurements due to fatigue or boredom for example,
- data collection is time and cost intensive, and time/cost varies across attributes, in which case finding the optimal ratio of missing values over observation for each attribute is important.

Three-Form Design (and its variations), Multiple Matrix Sampling and Two-Method Measurement are the states of the art planned missing data techniques in cross-sectional studies (for more detailed information refer to [7, 2]).

Furthermore, for various reasons, it may be difficult for subjects to participate in ongoing longitudinal assessments, particularly in research which lasts many years. One solution is to lighten respondent burden by planning the missing data pattern across subjects. The surprising usefulness of this approach has been demonstrated using growth curve models [10]. As other examples of planned missing data designs in longitudinal studies, the methods of Monotonic Sample Reduction, Developmental Time-Lag and Wave To Age-based Designs could be mentioned [9].

In another approach that can be considered as a planned missing data method, Desmarais et al. designed a heuristic-based selective sampling and investigated it in test design. They showed that it is possible to improve the predictive performance of a Bayesian CAT model based on a heuristic that relies on entropy to optimize the choice of test items [4].

3. ADAPTIVE SAMPLING

Adaptive sampling is a technique that is enforced while a survey is being fielded—that is, the sampling design is modified in real time as data collection occurs—based on information gathered from previous sampling that has been completed. Therefore, when sampling or 'allocating' adaptively, sampling decisions are dynamically made as data is gathered.

4. ENTROPY

The Adaptive Sampling method proposed relies on the entropy of a feature, where the probability of an attribute is estimated by the relative frequencies of its values in the usual Shannon definition (recall that we limit our study to binary values). The more each feature categories are equally likely, the greater the entropy of the feature in question.

4.1 Binary Entropy Function

The binary entropy function, denoted $H_2(x)$ or $H_b(p)$, is defined as the entropy of a Bernoulli process with probability of success $P(x = 1) = p$. Mathematically, the Bernoulli trial is modeled as a random variable x that can take on only two values: 0 and 1. The event $x = 1$ is considered a success and the event $x = 0$ is considered a failure. (These two events are mutually exclusive and exhaustive.)

If $P(x = 1) = p$ then $P(x = 0) = 1 - p$ and the entropy of x is given by

$$H_2(x) = xlog\frac{1}{x} + (1-x)log\frac{1}{(1-x)} \qquad (1)$$

The logarithms in this formula are usually taken (as shown in the figure 1) to the base 2 [8].

The rest of this paper is organized as follows. Below we introduce our models. In section 6, our experimental methodology is explained. In section 7 we present our results and finally, in section 8, the results are discussed and further studies are proposed.

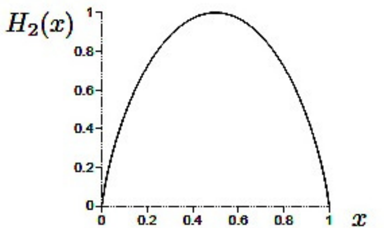

Figure 1. The Binary Entropy Function [8]

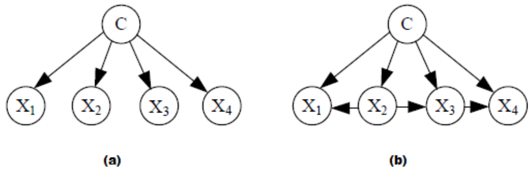

Figure 2. a) Naive Bayes Classifier Structure

b) TAN Classifier Structure

5. MODELS

We test the hypothesis that Selective Sampling with an entropy-driven heuristic affects model predictive performance over three types of well known classifiers: Naive Bayes, Logistic Regression, and Tree Augmented Naive Bayes (TAN). They are briefly described below.

5.1 Naive Bayes

A Naive Bayes classifier is a simple but important probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions which assume all the input attributes are independent given its class:

$$P(c_j|x_1, x_2, \dots, x_d) = \frac{P(c_j)}{P(x_1, x_2, \dots, x_d)} \prod_{i=1}^{d} P(x_i|c_j) \qquad (2)$$

Where:
$\mathrm{P}(\mathrm{c_j}|\mathrm{x_1}, \mathrm{x_2}, \dots, \mathrm{x_d})$ is the posterior probability of class membership, i.e., the probability that X belongs to c_j
$P(x_1, x_2, \dots, x_d)$ is the prior probability of predictors which is also called the evidence and
$P(c_j)$ is the prior probability of class level c_j

Using Bayes' rule above, the classifier labels a new case X with a class level c_j that achieves the highest posterior probability. Despite the model's simplicity and the fact that the independence assumption is often inaccurate, the naive Bayes classifier is surprisingly useful in practice.

5.2 Logistic Regression

Logistic regression is one of the most commonly-used probabilistic classification models that can be used when the target variable is a categorical variable with two categories (i.e. a dichotomy) or is a continuous variable that has values in the range 0.0 to 1.0 representing probability values or proportions. The logistic regression equation can be written as:

$$P = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2+\cdots+b_nx_n)}} \qquad (3)$$

Logistic regression uses maximum likelihood estimation (MLE) to obtain the model coefficients that relate predictors to the target.

5.3 Tree Augmented Naïve Bayes (TAN)

Naïve Bayes classifier has a simple structure as shown in figure 2(a), in which each attribute has a single parent, the class to predict. The assumption underlying Naive Bayes is that attributes are independent of each other, given the class. This is an unrealistic assumption for many applications. There have been many attempts to improve the classification accuracy and probability estimation of Naive Bayes by relaxing the independence assumption while at the same time retaining much of its simplicity and efficiency.

Tree Augmented Naive Bayes (TAN) is a semi-Naive Bayesian learning method that was proposed by Friedman et al. [5]. It relaxes the Naive Bayes attribute independence assumption by employing a tree structure, a structural augmentation of Naïve Bayes classifier that allows the attribute nodes (leaves) to have one more parent beside the class. The structure of TAN classifier is shown in figure 2(b).

A maximum weighted spanning tree that maximizes the likelihood of the training data is used to perform classification. Inter-dependencies between attributes can be addressed directly by allowing an attribute to depend on other non-class attributes. Friedman et al. showed that TAN outperforms Naive Bayes, yet at the same time maintains the computational simplicity (no search involved) and robustness that are characteristic of Naive Bayes [5].

6. METHODOLOGY

Our experiments have been carried out using the mentioned models and a Selective sampling design based on the entropy heuristic, the process and the datasets that are introduced below.

6.1 Entropy-based heuristic for Selective Sampling

We define three sampling schemes to determine missing values in order to investigate their respective effects over the predictive accuracy of the classifier models:

i. **Uniform:** Uniform random samples (Random distribution of missing values among the items).

ii. **Low Entropy:** Higher sampling rate for low entropy items (High entropy items will have higher rates of missing values).

iii. **High Entropy:** Higher sampling rate for high entropy items (Low entropy items will have higher rates of missing values).

As mentioned before, the entropy of an item is derived from its initial probability of success and therefore, high entropy items are the items that are closest to an initial probability of 0.5. The probability of sampling based on entropy is a function of the $x = [0,2.5]$ segment of a normal (Gaussian) distribution as reported in figure 3. The probability of an item being sampled will therefore vary from 0.40 to 0.0175 as a function of its rank, from the highest to the lowest item entropy on that scale. Items are first ranked according

to their entropy and they are attributed a probability of being sampled following this distribution. The distributions are the same for both conditions (ii) and (iii), but the ranking is reversed between the two of them. For the uniform condition (i), all items have equal probability of being sampled.
We have run a simulation study of such sampling schemes. The details of the experimental conditions and the results are described below.

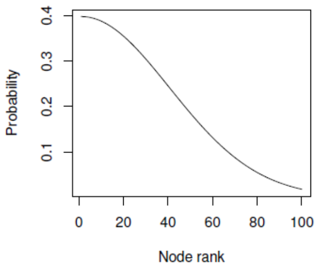

Figure 3. Sampling probability distribution used for the schemes 2 and 3

6.2 Adaptive Sampling and Seed Data

To conduct our sampling designs in an adaptive manner we start with a small seed dataset. Initial probabilities are obtained from the seed dataset and then entropy values are extracted. Then the algorithm samples feature observations based on the three different schemes. Levels of uncertainty (entropies) for all of the items are updated based on what have been sampled so far. This process is repeated until the final sampling criterion, which in this study is to reach a fixed number of observations. Figure 4, shows a simple flowchart of the algorithm. In this study 3 different sizes for the seed dataset are: 2, 4 and 8 records.

6.3 Non-adaptive Selective Sampling

As a comparison basis for the performance of different sizes of the seed dataset we also conduct our entropy-based selective sampling schemes in a non-adaptive manner. Unlike the adaptive algorithm in which entropy values is modified in real time as data collection continues, in the non-adaptive selective sampling condition we extract the entropy values from the full dataset in hand and then conduct the three sampling schemes. This is similar to our previous work [6] as mentioned and it provides us with another baseline for comparison.

6.4 Simulation Process

Our simulations consist in 100-fold cross-validation runs. In each run, different training and validating sets are built based on our three schemes described in previous subsection. The proportion of total missing values inserted in the training sets is half of the data. Testing datasets contain no missing values. We compare the performance of the models on the three different sampling schemes in terms of average number of Incorrectly Classified Items (ICI) and also the average Root Mean Square Error (RMSE).

To determine whether our results are statistically significant, for each model, 2-tailed paired Student t-tests are run on the pairs scheme2/scheme1 and scheme3/scheme1 on the results of 100 folds. We report the results of our experiments in section 7.

6.5 Datasets

The experiments are conducted over 11 sets of real binary data. Table 1 reports general statistics on these datasets. The first dataset in the list, SPECT Heart, is from UCI Machine Learning Repository [3] and others are from KEEL-dataset Repository [1].

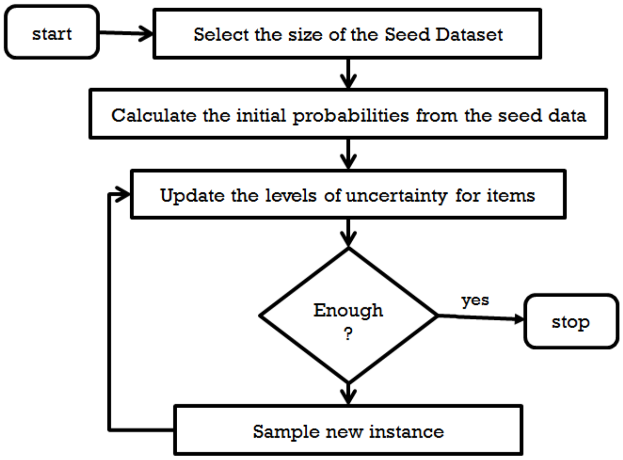
start
Select the size of the Seed Dataset
Calculate the initial probabilities from the seed data
Update the levels of uncertainty for items
Enough ?
yes
stop
Sample new instance

Figure 4. Adaptive Sampling Algorithm

7. RESULTS

Figure 5 illustrates the way we conduct the sampling in our non-adaptive sampling approach taking the Brain Chemistry Dataset as an example. The upper-left graph reports the entropy value of each of the 100 attributes ordered from the lowest to the highest entropy, and the other three graphs report the probability of being sampled for each corresponding attribute (item).

The results of running the adaptive algorithm with a seed dataset of size 8 over Brain Chemistry Dataset are summarized in tables 2 and 3. Table 2 reports the average percent of incorrectly classified items (ICI) for the methods based on the different sampling schemes. It also shows the average Root Mean Square Error (RMSE) for each of the models under the three sampling schemes. As it is clear from the table, for this dataset where the seed dataset has 8 records, the performance of Naïve Bayes improves under the sampling scheme 2. Logistic Regression performs better under scheme 3 and also, compared to other schemes, performance of TAN under scheme 3 is superior.

Table 2. Performance Comparison for the different techniques under the different schemes of sampling for Brain Chemistry Dataset (ICI-Incorrectly Classified Items and RMSE-Root-Means-Squared-Error) where seed dataset size=8

	Measure	Sch1	Sch2	Sch3
NB	Average % of ICI	3.25	2.54^{***}	3.82
	Average RMSE	0.16	0.14^{***}	0.17
LR	Average % of ICI	7.58	10.08	6.53^{***}
	Average RMSE	0.27	0.30	0.25^{**}
TAN	Average % of ICI	4.31	5.24	3.27^{***}
	Average RMSE	0.18	0.19	0.16^{***}

(* for $0.01<p<0.05$, ** for $0.001<p<0.01$ and *** for $p<0.001$ based on Student t-test of the comparison of the corresponding scheme with Sch1. See Table 3.)

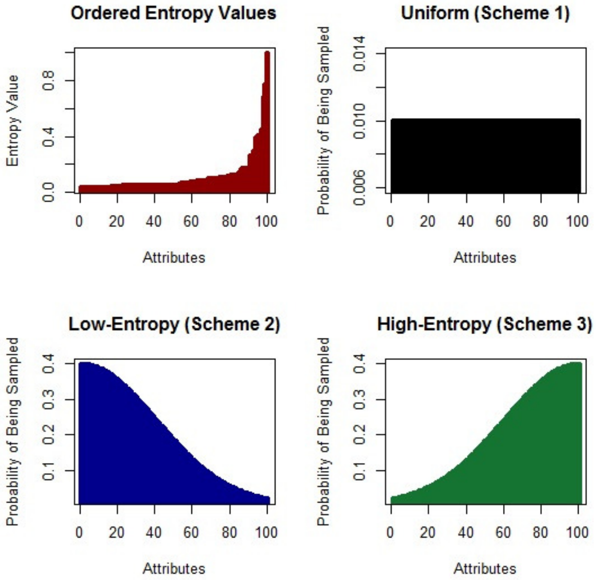

Figure 5. Brain Chemistry Dataset

Table 3. RMSE difference between scheme 1 and the two other schemes for Brain Chemistry Dataset. Student-t test is based on 100 random sample simulations

	Pairs	t	Mean of the Differences	p-value
NB	Sch2/Sch1	-4.3700	-0.0144	3.1e-05
	Sch3/Sch1	4.3284	0.0111	3.6e-05
LR	Sch2/Sch1	5.5064	0.0373	0.00
	Sch3/Sch1	-3.2150	-0.0177	0.002
TAN	Sch2/Sch1	1.9707	0.0075	0.052
	Sch3/Sch1	-7.1644	-0.0240	0.00

(df=99, Confidence Interval=95%)

Results of conducting 2-tailed paired t-tests on the pairs scheme2/scheme1 and scheme3/scheme1 for the models on obtained results of 100 folds are shown in table 3. As the table reflects, very small p-values show that there are very strong evidences against null hypothesis in those mentioned cases and therefore, our results, concluded from table 2, are statistically significant.

We have conducted similar simulations and evaluations for the other datasets and the seed sizes. Tables 4a-4c summarize the results. In these tables schemes 2 and 3 are compared to the uniform sampling scheme (Sch1) based on the measure of ARMSE. The numbers in cells represent the number of datasets and those in parenthesis show the percentage of mean improvement on ARMSE value gained by applying the scheme. As it can be seen from the table 4a, NB model in 45.5% of the datasets receives about 8% improvement to its prediction performance when we apply second sampling scheme with the seed size equal to 8. In general, for NB, scheme 2 is almost always better than scheme 1 in adaptive sampling approach. The table also shows compared to the third scheme, scheme one is preferable for adaptive approach.

For LR model no clear pattern emerges. But, at least it is clear from the table 4b that compared to scheme 2 (which is better for only one dataset), scheme 1 brings a higher prediction performance to the classifier. We see that the sensitivity of the model to the third scheme of sampling increases when the seed size goes higher, such that we see in 27.3% of the cases, applying scheme 3 results in about 10% less ARMSE for the model compared to scheme 1 when the size of the seed dataset is 8.

For the TAN model, as table 4c demonstrates, applying the third scheme of sampling in non-adaptive approach on all the datasets brings more than 13% higher prediction performance to the model. By having 8 records (less than 1% of total instances) in the seed dataset adaptive algorithm yields almost the same results as non-adaptive approach does. In none of the dataset uniform sampling is better than the third scheme of sampling, but, compared to scheme 2, uniform sampling scheme generally results in better prediction performance for TAN. Again we see a convergence in the model's performance to the case of non-adaptive approach when the size of seed dataset is 8.

8. CONCLUSION

These results confirm that Adaptive Sampling based on a heuristic that relies on attribute entropy can improve the performance of some classification methods with a 0/1 loss function. Adaptive Sampling in all but one of the datasets improves the performance of TAN classifier when we use a seed dataset of 1% or less of the total number of instances. Improvements were also obtained for the Naive Bayes classifier, but they are not systematic, and are obtained from scheme 2 instead of scheme 3. The results also show an unexpected result for one data set, for which the uniform (scheme 1) scheme is better than scheme 2 when the entropy from the full data is taken. The Logistic regression classifier generally does better with the uniform sampling scheme, but the results are not systematic across data sets.

Further analysis and investigations are required to better explain these results. Nevertheless, this investigation shows that we can influence the predictive performance of a classifier with partial data when we have the opportunity to select the missing values. It opens interesting questions and can prove valuable in some contexts of application.

Table 4. Number of data sets which show significant greater error (ARMSE) for each technique, under different sampling schemes, over 11 different datasets, and for different seed dataset sizes

a) Naïve Bayes

	Sch1<Sch2	Sch1>Sch2	Sch1<Sch3	Sch1>Sch3
SD=2	-	4 (6.4%)	2 (5.9%)	-
SD=4	-	4 (11.7%)	5 (6.2%)	-
SD=8	-	5 (7.7%)	5 (4.8%)	-
Full	1 (10%)	4 (6.2%)	-	5(6.0%)

b) Logistic Regression

	Sch1<Sch2	Sch1>Sch2	Sch1<Sch3	Sch1>Sch3
SD=2	7 (22.6%)	1 (23.5%)	3 (12.9%)	1 (13.6%)
SD=4	6 (21.1%)	1 (16.7%)	6 (9.8%)	1 (9.1%)
SD=8	7 (21.7%)	1 (22.2%)	4 (11.2%)	3 (9.3%)
Full	9 (35.3%)	1 (16.7%)	5 (10.2%)	5 (14.3%)

c) Tree Augmented Naïve Bayes

	Sch1<Sch2	Sch1>Sch2	Sch1<Sch3	Sch1>Sch3
SD=2	5 (14.2%)	-	-	10 (13.1%)
SD=4	5 (10.8%)	-	-	10 (14.3%)
SD=8	6 (10.5%)	1 (5.6%)	-	11 (12.7%)
Full	7 (14.1%)	1 (11.0%)	-	11 (13.6%)

- Sch_i<Sch_j means ARMSE(Sch_i)<ARMSE(Sch_j)
- The numbers in cells represent the number of datasets and those in parenthesis show the percentage of mean improvement on ARMSE gained by applying the scheme

REFERENCES

[1] Alcalá, J. et al, 2010. KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing.*

[2] Anigbo, L.C., 2011. Demonstration of The Multiple Matrices Sampling Technique In Establishing The Psychometric Characteristics Of Large Samples. *Journal of Education and Practice* 2, 3, pp. 19-25.

[3] Bache, K. and Lichman, M., 2013. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences.

[4] Desmarais, M.C. et al, 2008, Adaptive Test Design with a Naive Bayes Framework. *Proceedings of the 1st Conference of Educational Data Mining.* Montreal, Canada, 48-56.

[5] Friedman, N. et al, 1997. Bayesian network classifiers. *Machine learning* 29, 2-3, pp. 131-163.

[6] Ghorbani, S. and Desmarais, M.C., 2013, Selective Sampling Designs to Improve the Performance of Classification Methods. *Proceedings of 12th International Conference on Machine Learning and Applications, ICMLA 2013* vol. 1. Miami, USA, pp. 178-181.

[7] Graham, J. et al, 2006. Planned missing data designs in psychological research. *Psychological methods* 11, 4, pp. 323.

[8] MacKay, David J.C., 2003. *Information theory, inference and learning algorithms.* Cambridge university press, UK.

[9] McArdle, J.J. and Woodcock, R.W., 1997. Expanding test-retest designs to include developmental time-lag components. *Psychological Methods* 2, 4, pp. 403.

[10] Palmer, R.F. and Royall, D.R., 2010. Missing data? Plan on it!. *Journal of the American Geriatrics Society* 58, s2, pp. S343-S348.