MISSING DATA IMPUTATION FOR SUPERVISED LEARNING

Abstract: This paper compares methods for imputing missing categorical data for supervised learning tasks. The ability of researchers to accurately fit a model and yield unbiased estimates may be compromised by missing data, which are prevalent in survey-based political science research. We experiment on two machine learning benchmark datasets with missing categorical data, comparing classifiers trained on non-imputed (i.e., one-hot encoded) or imputed data with different degrees of missing-data perturbation. The results show imputation methods can increase predictive accuracy in the presence of missing-data perturbation. Additionally, we find that for imputed models, missing-data perturbation can improve prediction accuracy by regularizing the classifier.

Key words and phrases: artificial neural networks (ANNs), decision trees, imputation methods, missing data, perturbation, random forests

1. Introduction

Missing data is a common problem in survey-based political science research. Supervised learning has become an increasingly attractive methodology and proven to be effective in political science applications, such as studies of international and civil conflict (Beck et al., 2000; De Marchi et al., 2004; Hill and Jones, 2014; Muchlinski et al., 2016) and election fraud (Cantú and Saiegh, 2011; Montgomery et al., 2015).[1] For supervised classification tasks, the objective is to fit a model on labeled training data in order to categorize new examples. However, the ability of researchers to accurately fit a model and yield unbiased estimates may be compromised by missing data.

The objective of the present study is to compare the out-of-sample performance of three popular machine learning classifiers — decision trees, random forests, and artificial neural networks (ANNs) — trained on imputed or non-imputed (i.e., one-hot encoded) machine learning benchmark datasets that contain various degrees of missing-data perturbation. Researchers analyzing survey data typically choose decision trees or random

[1] Methodological applications of supervised learning include the estimation of heterogeneous treatment effects (Imai and Strauss, 2011; Green and Kern, 2012; Imai et al., 2013; Grimmer et al., 2014), text analysis (Quinn et al., 2010; Hopkins and King, 2010; Grimmer and Stewart, 2013; Lauderdale and Clark, 2014; Wilkerson et al., 2015), and record linkage (Giraud-Carrier et al., 2010).

forests for classification tasks, largely because these models do not require imputing missing data nor encoding categorical variables, unlike ANNs or other classifiers. The results of the present study will provide guidance to applied researchers on how to handle missing data for supervised learning tasks.

This manuscript is organized as follows: Section 2 describes missing data mechanisms and imputation methods; Section 3 describes our experiments on two benchmark datasets and discusses the results; Section 4 concludes and offers areas for future research.

2. Missing data and imputation methods

In this section, we describe the missing data mechanisms underlying patterns of missing data common to survey datasets. We then review popular methods of handling missing data.

2.1. Missing data patterns and mechanisms

It is important to first distinguish between missing data patterns, which describe which values are observed and which are missing, and missing data mechanisms, which describe the the probability of missingness (Little and Rubin, 2014, Chap. 1). Common missing data patterns in surveys typically include unit nonresponse, where a subset of participants do not complete the survey, and item nonresponse, where missing values are concentrated on

particular questions. In opinion polls, nonresponse may reflect either refusal to reveal a preference or lack of a preference (De Leeuw et al., 2003).

Imputing missing values in situations where missing data hide information that are useful for classification tasks can help improve prediction accuracy. Understanding the missing data mechanisms underlying patterns of missing data is crucial since properties of imputation methods depend on the nature of these mechanisms. Following the notation of Little and Rubin (2014, Chap. 1), let $Y = y_{ij}$ be a $(n \times K)$ dataset with each row $y_i = (y_{i1}, \ldots, y_{iK})$ the set of y_{ij} values of feature Y_j for example i. Let Y_{obs} define observed values of Y and Y_{mis} define missing values. Define the missing data identity matrix $M = m_{ij}$, where $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is nonmissing. The missing data mechanism is missing completely at random (MCAR) if the probability of missingness is independent of the data, or $f(M|Y, \phi) = f(M|\phi)$ for all Y, ϕ, where ϕ denotes unknown parameters. The missing at random (MAR) assumption is less restrictive than MCAR in that the probability of missingness depends only on the observed data, $f(M|Y, \phi) = f(M|Y_{\text{obs}}, \phi)$ for all Y_{mis}, ϕ. The missing not at random (MNAR) assumption is that the probability of missingness may also depend on the unobserved data, $f(M|Y, \phi) = f(M|Y_{\text{mis}}, \phi)$ for all Y_{mis}, ϕ. Researchers typically assume data is MAR, which mitigates the identifia-

bility problems of MNAR because the probability of missingness depends on data that are observed on all individuals (Tsiatis, 2007, Chap. 6).

2.2. Imputation methods

Complete-case analysis (i.e., discarding examples with missing values) wastes information and biases estimates unless the missing data are MCAR. Since there is no way to distinguish whether the missing data are MCAR or MNAR from the observed data, a natural strategy is to impute missing values and then proceed as if the imputed values are true values. Imputation methods that rely on explicit model assumptions include *mean or mode replacement*, which substitutes missing values with the mean (for quantitative features) or mode (for qualitative features) of the feature vector, and *prediction model* imputation, which replaces missing values with the predicted values from a regression of Y_{mis} on Y_{obs}. Explicit modeling methods assume the data are MAR while implicit modeling methods, which are algorithmic in nature and rely only on implicit assumptions, generally do not assume the underlying missing data mechanism. Implicit methods include *random replacement*, where an example with missing data is randomly replaced with another complete example randomly sampled, and *hot deck* imputation, where missing values are replaced by "similar" nonmissing values. Hot deck imputation can be implemented by computing the

k-nearest-neighbors (k-NN) of an example with missing data and assigning the mode of the k-neighbors to the missing data. Batista and Monard (2003) use this procedure and find k-NN imputation can outperform internal methods used by decision trees to treat missing data and summary statistic imputation. Li et al. (2004) propose a hot deck imputation method based on fuzzy k-means. Silva-Ramírez et al. (2011) empirically compare imputation using ANNs with mean/mode imputation, regression models (logistic regression and multiple linear regression), andhot deck, finding the ANNs model performs the best on datasets with categorical variables.

2.3. One-hot encoding for missing data

A natural strategy in dealing with missing data for supervised learning problems is one-hot encoding. Instead of imputing missing data, one-hot encoding creates a binary feature vector that indicates missing values. For categorical features, one-hot encoding simply treats a missing value symbol (e.g, "?") as a category when the categorical features are binarized. For continuous features, missing values are set to a constant value and a missingness indicator is added to the feature space. One-hot encoding for missing data yields biased estimates when the features are correlated, which is often the case with survey data, even when data are MCAR (Jones, 1996).

3. Experiments

In this section, we describe our experiment on two machine learning benchmark datasets with missing categorical data, comparing three popular classifiers — ANNs, decision trees, and random forests— trained on either one-hot encoded or imputed data with different degrees of MCAR perturbation.

3.1. Benchmark datasets

We experiment on two benchmark datasets from the UCI Machine Learning Repository: the Adult dataset and Congressional Voting Records (CVRs) dataset (Lichman, 2013). The Adult dataset contains $N = 48,842$ examples and 14 features (6 continuous and 8 categorical). The prediction task is to determine whether a person makes over $50,000 a year. The CVRs dataset contains $N = 435$ examples, each the voting record of a member of the 98^{th} U.S. House of Representatives for 16 key roll call votes. The dataset contains 16 categorical features with three possible values: "yea", "nay", and missing. The prediction task is to classify party affiliation (Republican or Democrat). In contrast to the Adult dataset, in which only a few features are highly correlated, many of the roll call votes in the CVRs dataset exhibit strong correlations (Figures SM-1 and SM-2).

We randomly split each dataset 2/3 for training and 1/3 for testing. The state–of–the–art for the Adult dataset is a Naive Bayes classifier that

achieves a 14.05% generalization error after removing examples with missing values (Kohavi, 1996). The CVRs dataset donor claims to achieve a 5-10% error rate using an incremental decision tree algorithm called STAGGER, although it is unknown to the authors what train-test split is used or how missing values are handled (Schlimmer, 1987; Schlimmer and Granger Jr, 1986).

3.2. Patterns of missing data

Uncovering missing data patterns in the datasets will help to identify possible missing data mechanisms and select appropriate imputation methods. Figure SM-3 analyzes patterns of missing data in the Adult dataset, in which 7% of the examples contain missing values. Missing data in the Adult dataset is due to item nonresponse, as missing values are concentrated in three of the categorical features — *Work class*, *Occupation*, and *Native country*— and no examples contain entirely missing data. It is unlikely that the data are MCAR because observations that are missing in *Work class* are also missing in *Occupation* (about 6% of examples have missing values in both). There is no way to determine from the observed data whether the missing data are MAR or MNAR; the data are MNAR if the probability of missingness cannot be explained only by the observed data in the other predictors.

Missing values in the CVRs dataset are not simply unknown, but represent values other up-or-down votes, such as voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known. Close to half of the CVRs data contains missing values, which are present in every feature (Figure SM-4). About a quarter of missing data is in `South Africa`, which was a controversial amendment to amend the Export Administration Act to bar U.S. exports to South Africa's apartheid regime. Twelve percent of missing data is in the feature `Water`, which is a water projects authorizations bill, and 7% of missing data rests in the feature `Exports`, which is a tariff bill. The data are unlikely to be MCAR because 12% of the data are missing in just `South Africa` and less than 1% of examples are missing across all features. It is most likely in this case that the CVRs data are MNAR because the probability of missing a vote or voting present on one important bill should not theoretically be influenced by observed votes on other important bills.

3.3. Preprocessing

We perturb the training data so that the proportion of missing values in the set of categorical features Y_{cat} follows $\delta = \{0.1, 0.2, 0.3, 0.4\}$ according

to the MCAR mechanism

$$\Pr(M = 1|Y_{\text{cat}}, \phi) = \delta \text{ for all } Y_{\text{cat}}. \tag{3.1}$$

While perturbation is used primarily to study the effect of larger amounts of missing data, we note that missing-data perturbation is a form of dropout noise that can be used to control overfitting during the training process (Wager et al., 2013).

After one-hot encoding the categorical variables in the training data, we implement each of the following imputation techniques, discussed in Section 2.2: k-NN, prediction model (logistic regression, random forests, or SVMs), mode replacement, and random replacement. We then standardize continuous features by subtracting the mean and dividing by the standard deviation of the feature. The test data is preprocessed in the same manner, with the exception that we do not perturb categorical features in the test data.[2]

3.4. Model training and assessment

We train three different classifiers on the preprocessed data: decision trees, random forests, and ANNs. The ANNs consists of four layers, each of the two hidden layers having 1024 nodes, and updates with the adaptive

[2]When imputing the missing data with mode replacement, we use the training set mode. We also use the training set mean and standard deviation to standardize test set features.

learning rate method *Adadelta* (Zeiler, 2012). We explore the hyperparameter space — momentum schedule, dropout regularization, and learning rate — using Bayesian optimization (Snoek et al., 2012), which selects optimal models based on a given objective function.[3] Prediction intervals are obtained from the standard deviation of test set errors of ANNs trained with different convergences (Heskes et al., 1997).

Random forests and decision trees are trained with preselected hyperparameters. Prediction intervals follow from the variation created by varying the maximum depth of the decision trees, and for random forests, the number of trees and decision rule for the number of features to consider when looking for the best split.

3.5. Results

We assess the performance of the classifiers in terms of test set error rate on one-hot encoded or imputed data and for various degrees of MCAR perturbation. The results on the Adult dataset and CVRs dataset are plotted in Figures 1 and 2, respectively. Error bars represent ± 1 standard deviation from the test error rate. One-hot-encoded decision trees beats the state-of-the-art on the CVRs dataset by over 2% (0.027 ± 0.006). The ANNs classifier trained on data imputed with k-NN yields the lowest

[3] We use the mean training error rate as our objective function. Figure SM-5 shows the exploration of hyperparameter space during Bayesian optimization for both datasets.

generalization error (0.144 ± 0.06) on the Adult dataset, which is slightly above the state-of-the-art for the dataset even with 10% of the categorical feature values perturbed. In comparison, a random forests classifier trained on non-perturbed and one-hot encoded data yields a test error rate of 0.152 ± 0.02. This comparison shows that the classifiers can overfit the data and, in the case of imputed models, perturbation improves prediction accuracy by regularizing the classifier.

Overall, the results show imputation methods can increase predictive accuracy in the presence of missing-data perturbation. For both datasets, one-hot encoded models trained in the absence of perturbation do just as well as imputed models trained on non-perturbed data. In the case of the Adult dataset, imputation clearly improves accuracy in the presence of MCAR-perturbed data. In contrast, each of the three classifiers trained on the one-hot encoded CVRs dataset perform relatively well across different levels of perturbation. The general pattern of results hold when the classifiers are trained on MNAR-perturbed data (Figures SM-6 and SM-7).

4. Conclusion

This paper investigates the effects of missing data imputation and perturbation on classification tasks using supervised learning algorithms. We compare the predictive performance of ANNs against decision tree and ran-

dom forest classifiers trained on datasets with one-hot encoded or imputed data. We assess performance in terms of test set error for different levels of MCAR-perturbed training data. We come close to beating the state-of-the-art test error on the Adult dataset using an ANNs classifier trained on data imputed with k-NN and outperform the state-of-the-art on the CVRs dataset by over 2% using one-hot encoded random forests.

We conclude from the results that the performance of the classifiers and imputation strategies generally depend on the nature and proportion of missing data. For the Adult dataset, ANNs trained on imputed data generally outperform other classifiers and imputation methods across different ratios of perturbed data, while classifiers trained on one-hot encoded data perform very poorly on perturbed training data.

Future work could help identify the conditions under which missing-data perturbation might improve prediction accuracy by acting as a regularization technique. The results of the present study show that perturbation can help increase predictive accuracy for imputed models, but not one-hot encoded models. Future work might compare missing-data perturbation with dropout training, which changes to zero all the values of a random subset of features (Hinton et al., 2012; Maaten et al., 2013; Wang and Manning, 2013).

Supplementary Materials

The online supplementary material contains descriptive plots of feature correlation and missing data patterns in the benchmark datasets; plots of Bayesian hyperparameter optimization for training ANNs; and test set error plots for classifiers trained on MNAR-perturbed data.

Funding

This work was supported by the National Science Foundation Graduate Research Fellowship [grant number DGE 1106400]. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

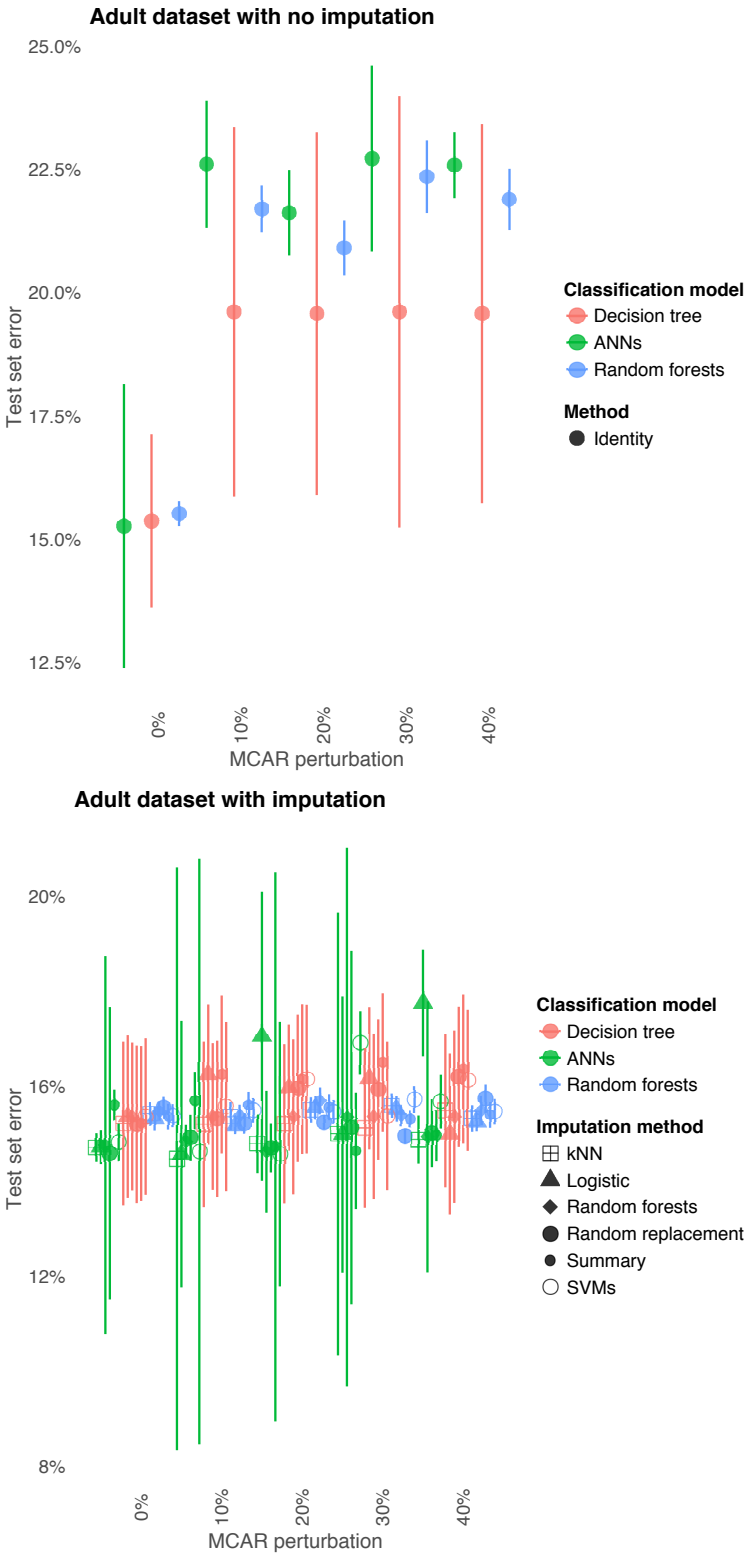

Figure 1: Error rates on the Adult test set with (bottom) and without (top) missing data imputation, for various levels of MCAR-perturbed categorical training features (x-axis). Error bars represent one standard deviation from the test error prediction.

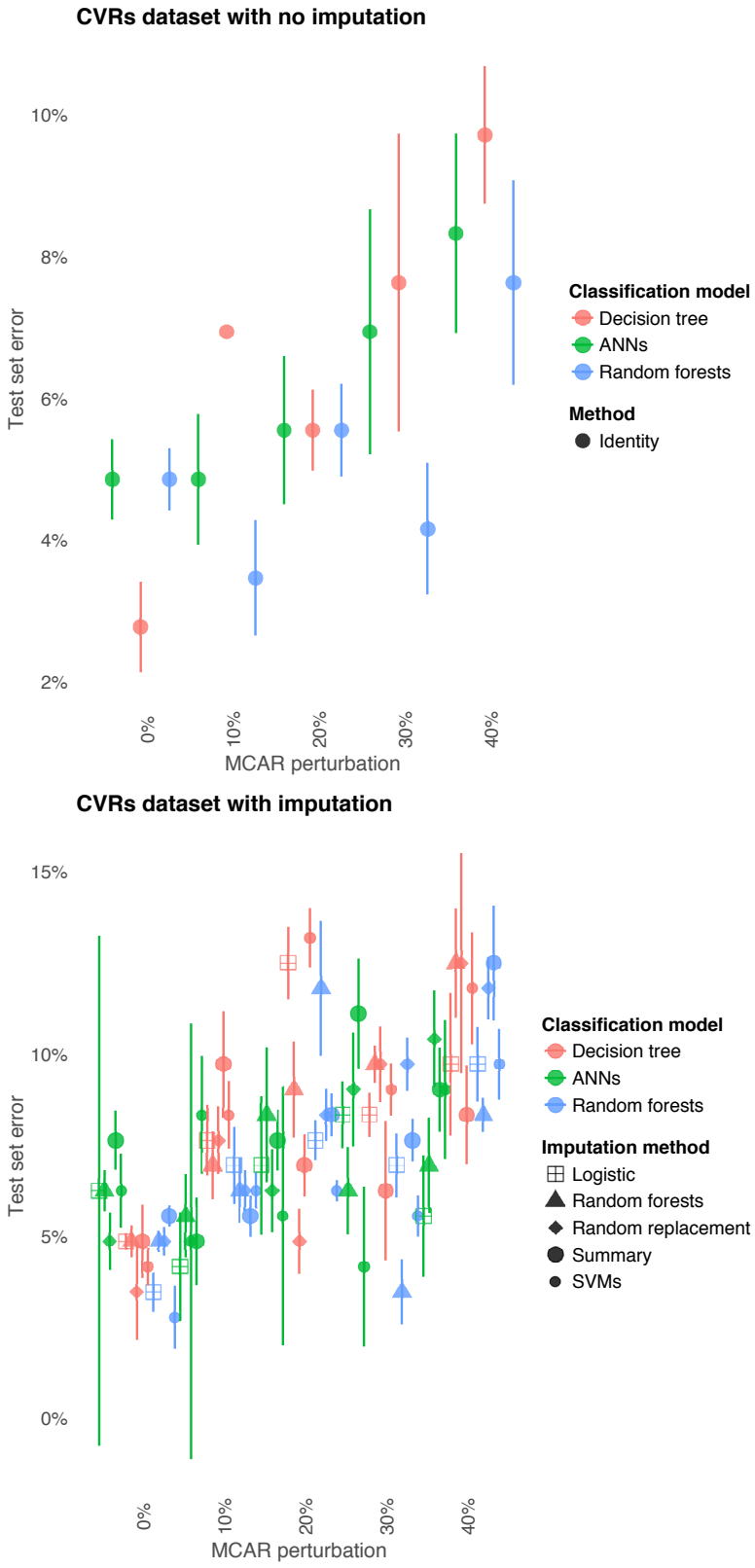

Figure 2: Error rates on the CVRs test set with (bottom) and without (top) missing data imputation.

See footnotes for Figure 1.

References

Batista, G. E. and M. C. Monard (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence 17*(5-6), 519–533.

Beck, N., G. King, and L. Zeng (2000). Improving quantitative studies of international conflict: A conjecture. *American Political Science Review 94*(01), 21–35.

Bowers, J., M. M. Fredrickson, B. B. Hansen, and C. Panagopoulos (2015). Machine learning and causal inference: A modular approach to assessing the effects of the london bombings of 2005. *http://www.jakebowers.org/PAPERS/BFHPAPSA2015.pdf*.

Brouwer, R. K. (2002). A feed-forward network for input that is both categorical and quantitative. *Neural Networks 15*(7), 881–890.

Cantú, F. and S. M. Saiegh (2011). Fraudulent democracy? An analysis of Argentina's infamous decade using supervised machine learning. *Political Analysis 19*(4), 409–433.

Caughey, D. and M. Wang (2014). Bayesian population interpolation and lasso-based target selection in survey weighting. In *Summer Meeting of the Society for Political Methodology, University of Georgia, Athens, GA*.

De Leeuw, E. D., J. Hox, M. Huisman, et al. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics 19*(2), 153–176.

De Marchi, S., C. Gelpi, and J. D. Grynaviski (2004). Untangling neural nets. *American Political Science Review 98*(02), 371–378.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association 89*(426), 463–475.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth (1996). From data mining to knowledge discovery in databases. *AI magazine 17*(3), 37.

Giraud-Carrier, C., J. Goodliffe, and B. Jones (2010). Improving the study of campaign contributors with record linkage. *http://goodliffe.byu.edu/papers/linkage.pdf*.

Green, D. P. and H. L. Kern (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly 76*(3), 491–511.

Grimmer, J., S. Messing, and S. J. Westwood (2014). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *http://stanford.edu/~jgrimmer/het.pdf*.

Grimmer, J. and B. M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, mps028.

Heskes, T., W. Wiegerinck, and H. Kappen (1997). Practical confidence and prediction intervals for prediction tasks. In *Advances in Neural Information Processing Systems 9: Proceedingss of the 1996 Conference*. MIT Press.

Hill, D. W. and Z. M. Jones (2014). An empirical evaluation of explanations for state repression. *American Political Science Review 108*(03), 661–687.

Hinton, G. E., N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov (2012).

Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hopkins, D. J. and G. King (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science 54*(1), 229–247.

Hsu, C.-C. (2006). Generalizing self-organizing map for categorical data. *Neural Networks, IEEE Transactions on 17*(2), 294–304.

Imai, K., M. Ratkovic, et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics 7*(1), 443–470.

Imai, K. and A. Strauss (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis 19*(1), 1–19.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association 91*(433), 222–230.

Kohavi, R. (1996). Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 202–207. Citeseer.

Lauderdale, B. E. and T. S. Clark (2014). Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science 58*(3), 754–771.

Li, D., J. Deogun, W. Spaulding, and B. Shuart (2004). Towards missing data imputation: a

study of fuzzy k-means clustering method. In *International Conference on Rough Sets and Current Trends in Computing*, pp. 573–579. Springer.

Lichman, M. (2013). UCI Machine Learning Repository.

Little, R. J. and D. B. Rubin (2014). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Liu, Y., Y. Wang, Y. Feng, M. M. Wall, et al. (2016). Variable selection and prediction with incomplete high-dimensional data. *The Annals of Applied Statistics 10*(1), 418–450.

Lo, H.-Y., K.-W. Chang, S.-T. Chen, T.-H. Chiang, C.-S. Ferng, C.-J. Hsieh, Y.-K. Ko, T.-T. Kuo, H.-C. Lai, K.-Y. Lin, et al. (2009). An ensemble of three classifiers for KDD Cup 2009: Expanded linear model, heterogeneous boosting, and selective naıve Bayes. *JMLR W&CP 7*.

Maaten, L., M. Chen, S. Tyree, and K. Q. Weinberger (2013). Learning with marginalized corrupted features. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 410–418.

Malarvizhi, M. and A. Thanamani (2012). K-nn classifier performs better than k-means clustering in missing value imputation. *IOSR Journal of Computer Engineering (IOSRJCE) 6*, 12–15.

Montgomery, J. M., S. Olivella, J. D. Potter, and B. F. Crisp (2015). An informed forensics approach to detecting vote irregularities. *Political Analysis 23*(4), 488–505.

Muchlinski, D., D. Siroky, J. He, and M. Kocher (2016). Comparing random forest with logistic

regression for predicting class-imbalanced civil war onset data. *Political Analysis 24*(1), 87–103.

NAPP (2008). Minnesota population center. north atlantic population project: Complete count microdata. version 2.0 [machine-readable database]. *Minneapolis, MN: Minnesota Population Center, available at* `https://www.nappdata.org`.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science 54*(1), 209–228.

Rey-del Castillo, P. and J. Cardeñosa (2012). Fuzzy min–max neural networks for categorical data: application to missing data imputation. *Neural Computing and Applications 21*(6), 1349–1362.

Rubin, D. B., H. S. Stern, and V. Vehovar (1995). Handling "don't know" survey responses: the case of the Slovenian plebiscite. *Journal of the American Statistical Association 90*(431), 822–828.

Ruggles, S., T. Alexander, K. Genadek, R. Goeken, M. Schroeder, and M. Sobek (2010). Integrated Public Use Microdata Series (IPUMS): Version 5.0 [machine-readable database]. *University of Minnesota, Minneapolis, available at* `http://usa.ipums.org`.

Schlimmer, J. C. (1987). *Concept Acquisition Through Representational Adjustment.* Ph. D. thesis, Department of Information and Computer Science, University of California, Irvine.

Schlimmer, J. C. and R. H. Granger Jr (1986). Incremental learning from noisy data. *Machine*

learning 1(3), 317–354.

Silva-Ramírez, E.-L., R. Pino-Mejías, M. López-Coello, and M.-D. Cubiles-de-la Vega (2011). Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks 24*(1), 121–129.

Snoek, J., H. Larochelle, and R. P. Adams (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pp. 2951–2959.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. Springer Science & Business Media.

Wager, S., S. Wang, and P. S. Liang (2013). Dropout training as adaptive regularization. In *Advances in neural information processing systems*, pp. 351–359.

Wang, H., G. Xing, and K. Chen (2008). Categorical data transformation methods for neural networks. In *IKE*, pp. 262–266.

Wang, S. and C. Manning (2013). Fast dropout training. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 118–126.

Wilkerson, J., D. Smith, and N. Stramp (2015). Tracing the flow of policy ideas in legislatures: A text reuse approach. *American Journal of Political Science 59*(4), 943–956.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.