Decision Tree Guided Multi-Class Support Vector Machines with Dynamic Class Selections

Roya Talibova
University of Michigan
talibova@umich.edu

Large-scale terrorism datasets are inherently different from other political science datasets in that they are mostly incomplete owing to the clandestine nature of the very phenomenon they seek to capture. This problem is compounded by two important factors: event-based terrorism databases generally rely on news reports that rarely provide full facts about these incidents, and perpetrators of these acts usually prefer not to disclose their identity or supply reliable data on their activities. The unclaimed terrorist attacks thus pose some challenging inference problems for researchers interested in using such datasets to test hypotheses related to terrorist groups. In this paper, I attempt to solve this problem of missing data for the most comprehensive event-based terrorism dataset Global Terrorism Dataset (GTD) , by using machine learning algorithms. I first use Multivariate Imputation by Chained Equations (MICE) to impute missing event characteristics (such as types of attack, weapon, and target, numbers of killed and wounded, and etc.), and then use these complete attributes as predictors in a novel Decision-Tree guided Multi-class Support Vector Machine algorithm to attribute unclaimed terror attacks to known terror groups. I classify all of the events that are listed as having unknown perpetrators, which accounts for half of the entire dataset, by comparing their properties with information obtained from events whose perpetrators are known.

August 2021

1. INTRODUCTION

As one of the deadliest forms of political violence, terrorism has slowly garnered worldwide attention over the last few decades. Research on terrorism has followed suit as the theories about the causes, dynamics, and consequences of terror attacks proliferated in academic scholarship. The increasing attention devoted to understanding this complex phenomenon has, at the same time, resulted in the creation of a number of large, comprehensive datasets on terrorism. Though the scope, size, and inclusion rules of these data vary considerably, they all share a common trait – the problem of missing data. The missingness patterns across these datasets are equally varied (Dugan, LaFree and Fogg, 2006). Many of them lack information on important attributes related to terrorism incidents. Yet it is the missing information related to the identity of perpetrators that poses the biggest challenge to the researchers in the field (Arva and Beieler, 2014; Bauer, Ruby and Pape, 2017; Perl, 2006; Tokdemir and Akcinaroglu, 2016). Without proper identification of unknown perpetrators, any kind of inference about these political actors and their activities based on such data will be inaccurate.

The incompleteness of terrorism-related data is due to the clandestine nature of the very concept these data are trying to capture (Sheehan, 2012). All terrorist events carry a certain degree of secrecy to be successful both in their planning and implementation stages. However, there are two important factors related to acts of terrorism that go beyond their inherent covertness, and further aggravate the problem of missingness: First, event-based terrorism datasets generally tend to rely on media and news reports that rarely provide full facts about these incidents. Second, the political actors engaged in these acts usually prefer to disguise their identity and withhold reliable data on their activities (Kyung, Gill and Casella, 2011; Juergensmeyer, 2017). This latter problem of identity concealment, coupled with inaccurate or incomplete reporting of incidents, hampers research attempting to answer questions pertaining to terrorist groups, including their structure, strength, strategies, ideologies, incentives, networks, lethality, and other identity-specific information.

This study uses various existing imputation and classification methods to reveal the identities of unknown perpetrators of terrorist attacks using the largest open-source, event-based dataset commonly used by scholars of political science – the Global Terrorism Database (GTD) (LaFree and Dugan, 2007). Using as predictors a number of attributes associated with the timing, severity, scale, location, targets, methods, and other details of attacks that provide important contextual information about any given event, I use a supervised machine learning technique to map the events with unknown perpetrators to known terrorist groups for which the data exists. The algorithm recovers the identities of perpetrators of more than seventy-eight thousand events perpetrated in the last half-century from the list of more than two thousand terrorist groups based on twenty-four distinct attributes of each attack.

Implementing a predictive model using supervised machine learning algorithms usually requires the existence of complete data in the variable space. A number of

ways to handle missing data in feature vectors have been proposed for discriminative models (Batista and Monard, 2003; Fortes et al., 2006; Ghahramani and Jordan, 1994; Saar-Tsechansky and Provost, 2007; Smola, Vishwanathan and Hofmann, 2005; Williams et al., 2007), but the computational difficulty of such solutions increase with the number of missing values (Tresp, Neuneier and Ahmad, 1995). To address this concern of incomplete-data classification, I take a two-step approach. First, I use multivariate imputation by chained equations (MICE) to deal with missing attribute values in the variable space. The partially imputed attributes are then used as predictors to classify the perpetrators using Multi-class Support Vector Machine technique (SVM).

To date, attempts at identifying unknown perpetrators of terrorist events have been limited. A study conducted by RAND cooperation in 1985 proposed a methodology for analyzing terrorist groups that included identifying and ranking possible perpetrators of unclaimed terrorist attacks. The process involved grouping all past terrorist events with known perpetrators according to combinations of some known attributes, determining the proportion of times a particular group was the perpetrator of a particular type of incident, and computing conditional probabilities of each group being responsible for unclaimed incidents (Jenkins, Cordes and Kellen, 1985). A study by Arva and Beieler uses machine learning techniques to predict whether a terrorist attack was perpetrated by a known or unknown group to uncover the mechanism by which the data are missing. Motivated by a different purpose, however, the authors do not attempt to attribute unknown attacks to known groups.

More recently, Bagozzi and Koren have used an ensemble of supervised machine learning techniques to classify atrocity perpetrators (Bagozzi and Koren, 2017). This study is different from their work in three ways: First, in their attempt to retain all available contextual variables, Bagozzi and Koren convert all variables into a single Document-Term Matrix (DTM), while including all missing cases in the predictors. With the same purpose, this study, instead, solves the "missingness in the predictors" problem via multiple imputation first, before applying a single machine learning classification framework to recover perpetrators. Secondly, the Global Terrorism Dataset is a lot larger than the Political Instability Task Force's (PITF) Worldwide Atrocities Dataset used by Bagozzi and Koren, both in terms of the overall number of cases included and the number of categories to be classified. PITF dataset codes the "perpetrator" variable according to an eight-category perpetrator ontology based on government involvement, whereas I purposefully avoid clustering terror groups to maintain the rich amount of information in the "perpetrator" variable. Finally, and most importantly, the mapping mechanism is designed in such a manner that temporal variations are explicitly included in the analysis. When classifying into terrorist group categories, only those organizations that have existed at the time of the event are used as a classifier. For example, if an event perpetrated in 1985 has an unknown perpetrator, classifiers of that specific datapoint do not include groups that had seized to exist before 1985. This manual feeding of information into the SVM via restriction of classifier space enables informative integration

of prior knowledge and ensures that mapping does not occur in an unmeaningful way.

The remainder of the article proceeds as follows: In Section 2, I provide a review of how literature has historically dealt with the missing data problem. Sections 3 and 4 provide an overview of multiple imputation techniques including multivariate imputation by chained equations, and multi-class support vector machines. Section 5 describes the data and patterns of missingness. Analysis and results of the Global Terrorism Dataset are provided in Section 6, followed by concluding remarks and suggestions for future work in Section 7.

2. Missing Data

Missing data complications abound in empirical social science research (Allison, 2001). Despite continuous advancements in methodologies for handling missing data problems, conflict-related research tends to either employ outdated methods to account for incomplete data or ignore missingness altogether (Hoeyland and Nygaard, 2011). This hesitancy to adopt novel methods and heavy reliance on traditional techniques is not unique to the field of conflict research (Bodner, 2006; Enders, 2010). However, the problem of missing data is particularly challenging in conflict studies given the prevalence of media-based event data (Weidmann, 2015). The amount of information and the level of accuracy hinge on the extent to which observers can observe what is transpiring and report it to the world (Öberg and Sollenberg, 2011). Most types of political violence take place in already conflict-ridden areas, where data collection process is hampered both by lack of secure access and absence of free press. In addition to the inability of outside parties to observe and report events on the ground, damaged infrastructure can also result in the destruction of records in many ways (Öberg and Höglund, 2011; Weidmann, 2013). Perhaps the most discussed pitfall of event datasets has been selective reporting (Davenport and Ball, 2002; Hocke, 1998; Moeller, 2002; Mueller, 1997). Dealing with these shortcomings requires a thorough understanding of the reasons of missingness, the underlying process that caused it, and its statistically sound handling.

Rubin has conceptualized missing data intro three broad categories based on the mechanisms that are assumed to be causing the missingness (Rubin, 1976). When missing completely at random (MCAR), the missing values are unrelated to the observed and missing responses. If the missingness is conditionally independent of the missing responses, given the observed ones, then the data is said to be missing at random (MAR). In other words, if the probability of missing data on a certain variable is independent of the value of the missing variable, after controlling for other variables in the data, then the MAR assumption holds. Examples of such missingness mechanism concerning this study could include having missing perpetrators for terror attacks carried out in a certain country or when perpetrators of only certain attack types are unknown.

Not missing at random (NMAR) is the most restrictive condition among these

as the missingness depends on both observed and missing data. The missingness mechanism is therefore characterized by the conditional distribution of the missing-data indicator matrix $\mathbf{M}$, where $\mathbf{M} = (M_{ij})$ given the complete data Y, where $\mathbf{Y} = (y_{ij})$. An unknown parameter ϕ is the missing value generating mechanism, such that:

$$MCAR : f(\mathbf{M}|\mathbf{Y}, \phi) = f(\mathbf{M}|\phi) \quad \text{for all } \mathbf{Y}, \phi \tag{1}$$

$$MAR : f(\mathbf{M}|\mathbf{Y}, \phi) = f(\mathbf{M}|\mathbf{Y}_{\mathbf{obs}}, \phi) \quad \text{for all } \mathbf{Y}_{\mathbf{mis}}, \phi \tag{2}$$

$$NMAR : f(\mathbf{Y}, \mathbf{M}|\theta, \phi) = f(\mathbf{Y}|\theta) f(\mathbf{M}|\mathbf{Y}, \phi) = \prod_{i=1}^{n} f(y_i|\theta) \prod_{i=1}^{n} f(\mathbf{M_i}|y_i, \phi) \tag{3}$$

where, $F(y_i|\theta)$ stands for the density of y_i indexed by unknown parameters θ, and $f(\mathbf{M_i}|y_i, \theta)$ for the density of a Bernoulli distribution for the binary indicator M_i (Little and Rubin, 2014).

For the purposes of this study, I will assume that the data from GTD are missing at random. In the case where the distribution of missingness is unknown to the researcher, MAR is assumed, since there is no plausible way to test whether the assumption holds, except a few alternative methods used in survey studies (Glynn, Laird and Rubin, 1993; Graham and Donaldson, 1993). The assumption of MAR is also justifiable given that the purpose of the paper is not inference, but an accurate prediction, and any impact on estimation from violation of this assumption would not be observed here.

Four general strategies of dealing with missing data have been proposed in the literature: deletion, weighting, maximum likelihood, and imputation (Allison, 2000; Durrant, 2009; Honaker and King, 2010; Ibrahim et al., 2005; Little and Rubin, 1987*b*, 1989; Little, 1988; Reiter and Raghunathan, 2007; Rubin, 2004; Zhang, 2003). The simplest approach to the analysis of missing data is to include only complete cases by discarding all incompletely recorded units. The listwise deletion approach leads to a loss of efficiency at best – when the MCAR assumption is correct –, and to severe biases at worst. Even when the data may be MAR, listwise deletion can potentially induce bias if the removed sample is not representative of the full sample (Hoeyland and Nygaard, 2011). Although highly problematic (and wasteful), the majority of the quantitative analyses to date have used this method when dealing with missing data (King et al., 2001). In fact, a recent study, reanalyzing the results of every quantitative work published within a five-year period in two leading IR journals, found that key results disappeared in almost half of the studies following substitution of listwise deletion with more robust methods (Lall, 2016).

A closely related approach is available-case analysis (pairwise deletion). The method simply calculates the means, variances, and covariances on all observed data. Similar to listwise deletion, the estimates following a pairwise deletion would be unbiased only if the data were MCAR. Schafer and Graham argue that the underlying principle of available-case analysis, while sensible, is poorly operationalized

(Schafer and Graham, 2002). Generally speaking, deletion methods are considered to be among the worst methods for practical applications (Wothke, 2000), unless the analyst can fully control for the determinants of missingness (Arel-Bundock and Pelc, 2018).

Weighting procedures are commonly applied when dealing with unit nonresponse in surveys and involve assigning weights proportional to the inverse of the selection probabilities of sampled units. Akin to complete case analysis, weighting methods ignore the incomplete cases but instead assigns weights to each complete case. The major weakness of weighting method derives from its restrictive application – it is only applicable to monotone patterns of missing data (Little and Rubin, 1989). Unmeasured or omitted variables might also introduce bias. Additional limitations include method's inefficiency vis-à-vis extreme weights and the need for many sets of weights (Pampaka, Hutcheson and Williams, 2016).

The third set of techniques used to treat missing data is parametric procedures, in which case a model is specified for the observed data, and likelihood under the model is computed to make inferences. One such popular procedure is the maximum likelihood estimation. This approach is preferred over the ad hoc techniques for the consistency and efficiency of estimates under correct model specification and because of the fact that standard errors account for incomplete data. The ML estimates yield biased results when the MAR assumption is violated, and problems arise when the model is mis-specified.

The imputation technique works by replacing a missing value by a value drawn from an estimate of the distribution of the same variable. Single imputation methods such as mean imputation, regression imputation, and stochastic regression imputation result in small estimated standard errors and biased correlations and fail to account for uncertainty in the predictive model. The exhaustive list of the problems associated with single imputation has been provided by Little and Rubin (Little and Rubin, 1987*a*). Multiple imputation circumvents all of these problems by properly representing all information in a dataset.

3. Multivariate Imputation by Chained Equations

The state of the art approach that has received widespread attention in the literature is multiple imputation. First introduced by Rubin (Rubin, 1987, 1996), multiple imputation was initially used to handle nonresponse in large-sample surveys (Reiter and Raghunathan, 2007). In the years that followed, development of new methods for Bayesian simulation has made it possible for researchers to apply it more broadly. In its simplest form, multiple imputation entails replacement of each missing value with $m \geq 2$ simulated values. The values are estimated from a distribution explicitly modeled for each missing value and are drawn from the conditional distribution of the target variable given all the other variables. The results of analysis of multiply-imputed datasets are used to obtain overall estimates and standard er-

rors (Schafer and Graham, 2002).[1] When the observed data does not contain much information about the missing values, imputations result in higher standard errors, whereas highly predictive observed data lead to smaller standard errors (Greenland and Finkle, 1995). The pooled variance contains both the within-imputation and the between-imputation variance.[2] Multiple imputation solves the problem of inaccurate estimates of uncertainty that is typical for single imputation.

The computational complexity of drawing a set of multiple parameters from posterior densities increase rapidly with the number of predictors included in the imputation process.[3] Two well-known algorithms are usually used to avoid such difficulties and enable efficient draws. The Imputation-Posterior algorithm (IP), proposed by Tanner and Wong, is a Markov-chain, Monte Carlo method that enables to draw random simulations from the multivariate normal observed data posterior, but it requires computational power and time (Tanner and Wong, 1987). The algorithm bears a close resemblance to the Gibbs sampling bar the P-step, where the parameter vector is sampled from its conditional posterior distribution given the latent variables (Skrondal and Rabe-Hesketh, 2004).

The Expectation-Maximization (EM) algorithm, developed a decade earlier, is still considered a more convenient alternative because of its speed, deterministic convergence, and accuracy (Dempster, Laird and Rubin, 1977), but might produce biased estimates.[4] Two additional optimization algorithms - EM with sampling and EM with importance sampling - have been proposed as potential alternatives (King et al., 2001). A recent addition to these algorithms has been Honaker and King's bootstrapping algorithm (EMB) for a general-purpose multiple imputation process that can handle datasets with hundreds of variables (Honaker and King, 2010).

A number of multivariate imputation methods have been developed to handle missing data. Various Joint Modeling (JM) methods were suggested by Rubin and Schafer (Rubin and Schafer, 1990). By specifying a parametric multivariate density given the model parameters, JM is used to generate imputations as draws from the posterior predictive distribution (Van Buuren et al., 2006) by MCMC techniques. It has been used to generate multivariate imputations under the multivariate normal, the log-linear and the general location model (Schafer, 1997).

Alternatively, semi-parametric Fully Conditional Specification (FCS)[5] approach

[1] The joint estimate of a parameter of interest θ is $\bar{\theta} = \frac{1}{D}\sum_{d=1}^{D}\hat{\theta}_d$ (Little and Rubin, 2014).

[2] Assuming $\bar{U}$ is the within-imputation variance and B the between-imputation variance, the variance estimate associated with the point estimate for a parameter is the total variance $T = \bar{U} + (1 + \frac{1}{m})B$.

[3] Random draws of μ and Σ with the number of p variables yield $p(p+3)/2$ elements (Honaker and King, 2010).

[4] See (King et al., 2001).

[5] FCS approach has been utilized under different names, such as stochastic relaxation (Kennickell, 1991), variable-by-variable imputation (Brand, 1999), regression switching (Van Buuren, Boshuizen and Knook, 1999), sequential regressions (Raghunathan et al., 2001), ordered pseudo-Gibbs sampler (Heckerman et al., 2000), partially incompatible MCMC (Rubin, 2003), iterated univariate imputation (Gelman, 2004), and chained equations (Buuren and

specifies conditional distributions for each variable subject to missingness, conditional on other variables included in the imputation model: a is imputed from $f(a|b,c)$, b is imputed from $f(b|a,c)$, and c is imputed from $f(c|a,b)$ (Murray and Reiter, 2016). The appeal of FCS derives from the flexibility of model choice for separate variables. Although multivariate normal is the conventional model used for missing data problems and has been found to work equally well for categorical and mixed data as for continuous data (Ezzati-Rice et al., 1995; Rubin and Schenker, 1986; Schafer and Olsen, 1998), using more specialized imputation models for separate variables based on their scale is a clear advantage. A related, but distinct approach involves specifying a joint distribution as a sequence of conditional models (Ibrahim et al., 2005; Lipsitz and Ibrahim, 1996).

While there are various software packages for implementing multiple imputations by the FCS, I use Multivariate Imputation by Chained Equations (MICE) package of Stef van Buuren as it allows for predictor and model selection, post-processing of imputed values, and passive imputation (Buuren and Groothuis-Oudshoorn, 2010). The non-problematic applicability of MICE to datasets with hundreds of variables has been demonstrated in various studies (He et al., 2010; Stuart et al., 2009). The chained equations, as used in MICE, obtain a posterior distribution of a vector of unknown parameters by sampling iteratively from respective conditional distributions. The i$^{\text{th}}$ iteration of chained equations is a Gibbs sampler that continuously draws

$$\begin{aligned}
\theta_1^{*(i)} &\sim P(\theta_1|Y_1^{obs}, Y_2^{(i-1)}, ..., Y_p^{(i-1)}) \\
Y_1^{*(i)} &\sim P(Y_1|Y_1^{obs}, Y_2^{(i-1)}, ..., Y_p^{(i-1)}, \theta_1^{*(i)}) \\
&\vdots \\
\theta_p^{*(i)} &\sim P(\theta_p|Y_p^{obs}, Y_1^{(i)}, ..., Y_{p-1}^{(i)}) \\
Y_p^{*(i)} &\sim P(Y_p|Y_p^{obs}, Y_1^{i}, ..., Y_p^{i}, \theta_p^{*(i)})
\end{aligned}$$

where $Y_j^{(i)} = (Y_j^{obs}, Y_j^{*(i)})$ is the j$^{\text{th}}$ imputed variable at iteration i (Buuren and Groothuis-Oudshoorn, 2010). Incompatibility of specifying joint distributions of various models that have no known joint distributions in theory is a common concern for chained equations, but this problem has been found to have little impact on the quality of imputations (Drechsler and Rässler, 2008; Van Buuren et al., 2006).

MICE process works in the following sequence: As a first step, a simple imputation is performed for each missing value. The original mean imputation is then set back to missing for the first variable. The observed values for this variable are regressed on the other variables included in the imputation model. The missing values of this first variable are then replaced with imputations derived from the

Oudshoorn, 2000).

regression model. Both observed and imputed values for this first variable are subsequently used in the regression model when imputing missing values for other variables. This procedure repeats for each variable with missing values. At the end of a single cycle, all missing data are replaced with imputations from regressions that reflect the relationships between the data. Depending on the number of imputations set by a researcher, cycles are then repeated with continuous updating at each cycle. The final imputations at the end of each cycle result in one completely imputed dataset. Although the number of iterations is arbitrary, the purpose of multiple iterations is to ensure the distribution of the parameters ultimately converge.

4. Multi-class Classification

Applications of supervised machine learning techniques to address a range of problems in social sciences have multiplied in recent years (Beck, King and Zeng, 2000; Burscher, Vliegenthart and De Vreese, 2015; Cantú and Saiegh, 2011; De Marchi, Gelpi and Grynaviski, 2004; Douglass and Harkness, 2018; Hill and Jones, 2014; Montgomery et al., 2015; Muchlinski et al., 2016; Mueller and Rauh, 2018; Nardulli, Althaus and Hayes, 2015). Among these, Support Vector Machines (SVMs) (Boser, Guyon and Vapnik, 1992; Cortes and Vapnik, 1995) have become popular for a number of reasons: they deal better with multi-dimensions and continuous features; they also perform well when multicollinearity is present and a nonlinear relationship exists between the input and output features (Kotsiantis, Zaharakis and Pintelas, 2007). Support Vector Machines additional advantages over other classification techniques have already been widely acknowledged (Cristiani and Taylor, 2000; Guyon, 1999; Pal, 2008).

Researchers working with supervised learning recognized the utility of applying these techniques to conduct text analysis (Grimmer and Stewart, 2013; Hopkins and King, 2010; Quinn et al., 2010; Lauderdale and Clark, 2014; Wilkerson, Smith and Stramp, 2015) or estimate heterogeneous treatment effects (Green and Kern, 2012; Grimmer, Messing and Westwood, 2017; Imai and Strauss, 2010; Imai and Ratkovic, 2013). Preference in classifying datasets that contain missing-data perturbation has generally been given to decision trees or random forests as they do not require imputing missing data (Poulos and Valle, 2016). Yet, SVM classification could prove equally useful to address missing data problems.

Traditionally, the focus of the SVMs has been on binary classification. Research on its effective extension to multiclass classification is still in progress (Abe, 2015; Bredensteiner and Bennett, 1999; Crammer and Singer, 2002; Mayoraz and Alpaydin, 1999; Nasiri, Charkari and Jalili, 2015; Vapnik, 1998; Weston and Watkins, 1999) and suggested techniques rely on various notions of margin and margin-based loss. Two current approaches to multiclass SVM include a) combination of numerous binary classifiers and b) consideration of all available data in a single optimization formulation (Crammer and Singer, 2001; Hsu and Lin, 2002; Lee, Lin and Wahba, 2004; Schölkopf and Smola, 2002; Weston and Watkins, 1999). Both methods require

larger computational power compared to a binary problem of equal data size.

Some well-known binary classification learning algorithms have been extended to deal with multi-class classification (Breiman, 2017; Rumelhart, Hinton and Williams, 1986). The most prominent of the binary-class based multi-category classifiers are the one-versus-one (OVO) and the one-versus-all (OVA) techniques (Dogan, Glasmachers and Igel, 2016). The latter relies on different binary decision functions that separate one class – "positive" $(y = +1)$ – from all the rest – "negative" $(y = -1)$ (Vapnik, 1998) with application of a voting scheme. The former, on the other hand, performs pair-wise comparisons between all classes with a total of $\frac{n(n-1)}{2}$ classifiers (Hastie and Tibshirani, 1998; Knerr, Personnaz and Dreyfus, 1990; Kressel, 1998). Consequently, OVO usually ends up producing many more classifiers than the OVA technique.

Finally, a state-of-the-art technique developed based on the Decision Directed Acyclic Graph (DDAG) tree-formed structure is a method named Directed Acyclic Graph SVM (DAGSVM) (Platt, Cristianini and Shawe-Taylor, 2000). Similar to OVO, it builds $\binom{k}{2}$ binary classifiers and uses an evaluation path to eliminate classes one-by-one until only one class remains. An alternative approach is Error-Correcting Output Coding (ECOC), which works by converting N-class classification problem into a large number of two-class classification problems with the use of error-correcting codes (see (Allwein, Schapire and Singer, 2000; Bose and Ray-Chaudhuri, 1960; Dietterich and Bakiri, 1994; Peterson et al., 1972)).

As for the design of binary classification, SVM separates the two classes $(k = 2)$ with a maximized margin criterion. Given m observed features $\mathbf{x} = [x_1, x_2, ..., x_m]^T \in \mathbf{R}^m$ and a training set $D = \{(\mathbf{x}^{(i)}, y^{(i)} : \mathbf{x}^{(i)} \in \mathbf{R}^m, y^{(i)} \in \{0, 1\}\}_{i=1}^n$ with n class examples, the goal is to identify a subset of the feature space that is most informative about the class distinction and design a classifier that most accurately predicts the class of a new example. Similarly, multi-class or multi-category classification problem involves assigning classes $(k > 2)$ to instances where the classes are selected from a finite set of classes.

Despite considerable advances in research concerning restriction, selection, or alteration of feature space (Bradley and Mangasarian, 1998; Cao et al., 2007; Damoulas and Girolami, 2009; Guyon and Elisseeff, 2003; Heiler, Cremers and Schnörr, 2001; Hermes and Buhmann, 2000; Jiang and Zhai, 2007; Kohavi and John, 1997; Mierswa and Morik, 2005; Neumann, Schnörr and Steidl, 2005; Pal and Foody, 2010; Sebastiani, 2002; Yan, Wang and Xie, 2008; Zareapoor et al., 2017) and modification of input space (Lin and Wang, 2002), dynamic changes to the number of classes remain yet to be implemented. In some applications, input points might need to be assigned to different subsets of classifiers. Some classes might be less meaningful than others, in which case it would be better for the machine to drop such classes. Currently, SVM lacks the kind of flexibility needed to make separate choices of classifiers for individual inputs.[6]

[6]The concept of "unclassifiable regions" has been developed recently for multi-label

5. THE GLOBAL TERRORISM DATASET

Datasets covering terrorist incidents have multiplied over the last decades due in part to the increasing frequency and intensity of attacks worldwide and the efficiency of newsgathering practices (Schmidt and Jongman, 1988). Majority of these datasets are event-based as scholars have only recently started constructing group-based datasets (Asal, Rethemeyer and Anderson, 2009; Carter, 2012; Fortna, Lotito and Rubin, 2014; Horowitz and Potter, 2014; Jones and Libicki, 2008; Nemeth, 2014; Price, 2012; Tokdemir and Akcinaroglu, 2016; Young and Dugan, 2014). The largest and most comprehensive event-based dataset among these is the Global Terrorism Dataset that covers terror incidents since 1970.

Other event-based datasets of similar nature lack the magnitude and scale of the GTD. Some are geographically limited. Terrorism in Western Europe: Events Data (TWEED), for example, focuses on the Western European region only (Engene, 2007). Problems of limited temporal coverage plague others. The RAND database of Worldwide Terrorism Incidents (WTI) coverage of domestic acts of terror starts in the 1980s (*RAND Database of Worldwide Terrorism Incidents*, 2018), while the Worldwide Incidents Tracking System (WITS) lists terror attacks perpetrated only after 2004 (Wigle, 2010). Finally, one of the largest scale datasets – the International Terrorism: Attributes of Terrorist Events dataset (ITERATE) – focuses only on transnational acts of terrorism (Mickolus et al., 2011). GTD, on the other hand, includes all types of worldwide terrorist incidents[7] with the largest temporal coverage.

GTD contains more than 120 variables describing the location, date, method of attack, victim and perpetrator characteristics, the number of casualties, and other event details on more than 170,300 cases of terrorism. Building such an all-inclusive dataset comes at a cost. Datasets that focus on a particular subset of terrorist acts may circumvent difficulties usually associated with gathering large-scale data. Because it casts a wider net in its inclusion criteria and maintains more attributes than any other dataset on terrorism, it is almost inevitable that most GTD variables contain a large number of missing values. In fact, in its very row format, there is not a single observation that can be considered a complete case in the sense that every associated variable is listed.[8] However, each event includes information on at least forty-five different variables. Most notably, of those incidents listed in the GTD, more than 78,300 of all incidents were recorded as having unknown perpetrators, making roughly half of the data missing.

For the purposes of this study, I include only twenty-four of the original GTD variables in my analysis. The common strategy in multiple imputation is to include

classification based on fuzzy logic (Abe, 2015), but its adaptation to SVMs is different from the approach described here.

[7] Twice as many categories of incidents compared to other datasets (Anderton and Carter, 2011)

[8] Other more serious missingness problems are related to both under-reporting of incidents and to accidental loss of a subset of data. The latter resulted in complete removal of incidents of terrorism from 1993 from the main dataset (START, 2016).

all relevant variables including their interactions (Collins, Schafer and Kam, 2001). The decision to reduce the predictor space into roughly one-fifth of its original size has multiple reasons: 1) the imputation method used in the study works better when only those variables that may be predictive of the missing values are included in the imputation process (Azur et al., 2011), 2) the included variables should satisfy the MAR assumption, 3) only those variables that exhibit less than 15 percent missingness in their vector space are included to avoid computational complexity. The complete list of included variables is in Table 1.

Table 1: List of Variables Included in the Imputation Process

Variable	Level of Measurement	Missing values	Model
Year of the event	Numeric	0	
Month of the event	Numeric	0	
Day of the event	Numeric	0	
Extended	Binary	0	
Country	Multi-level factor	0	
Region	Multi-level factor	0	
Criteria 1	Binary	0	
Criteria 2	Binary	0	
Criteria 3	Binary	0	
Multiple	Binary	0	
Success	Binary	0	
Suicide	Binary	0	
Individual	Binary	0	
Attack type	Multi-level factor	0	
Target type	Multi-level factor	0	
Weapon type	Multi-level factor	0	
Doubt terror	Binary	13,786	Logistic regression
Target subtype	Multi-level factor	9345	Random forest
Nationality	Multi-level factor	1394	Random forest
Number of killed	Numeric	9682	Predictive mean matching
Number of wounded	Numeric	15,325	Predictive mean matching
Property	Binary	19,579	Logistic regression
Hostage/Kidnappings	Binary	447	Logistic regression
Foreign target	Binary	487	Logistic regression

Of the 24 variables included in the imputation process, eight contain missing data ranging from 0.3 percent to 11 percent missingness. The *property* variable that identifies whether a property damage resulted from the incident has the largest amount of missingness containing 19,579 omitted values. The number of wounded *nwound* is a count variable with 15,325 missing values. Next is *doubtterr* with 13,786 missing values, which is an indicator variable that distinguishes events for which there is doubt as to whether the act constituted terrorism. The number of killed *nkill* has 9,682 omitted values. *Targsubtype1*, which captures the more specific target category and provides the next level of designation for each target type, has 9,345 blank cells. *INTMISC* and *Ishostkid*, which stand for whether the victim was a foreigner and whether victims were held hostage or kidnapped, list 487 and 447 missing val-

ues, respectively. The descriptive graphs of the missingness patterns are provided in Figures 1 and 2.

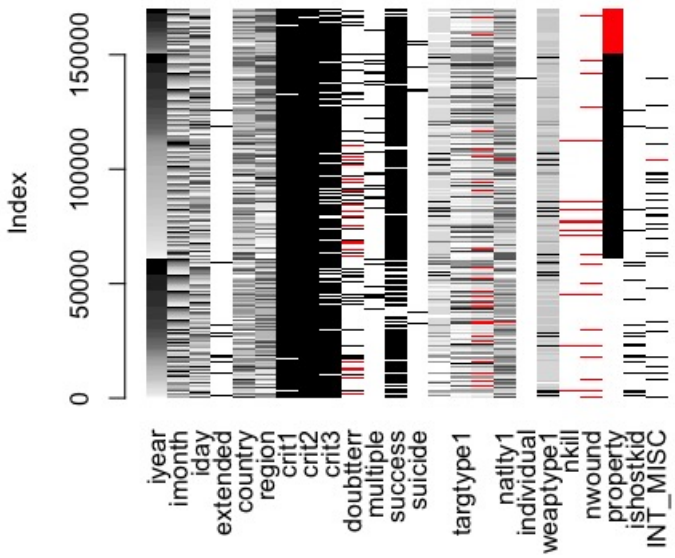


Figure 1: Missingness in Predictor Variables

The main variable of interest Group Name has 78,305 missing values and, while excluded from the multiple imputation, is used later at the classification stage. The number of groups already included in the dataset is 3,453. 1,693 of these have perpetrated only one terror act throughout their entire existence, while 2,554 have committed less than five acts (roughly 74%). The terror group with the highest number of incidents is Taliban with 6,575 acts listed under its name since the 1990s.

The remaining variables in the predictor space that have complete data indicate the exact date (year, month, and day, separately) of the event, whether the act lasted for more than 24 hours, the region and country where the event was perpetrated, three different binary criteria for inclusion, whether the attack was part of a multiple incident, whether the attack was successful, whether it was a suicide attack, and whether the perpetrator was not affiliated with a known group. In addition, three different variables describe the method of attack along with target and attack types. These are multiple category variables: for example, attack type includes assassinations, explosions, armed assaults, hijackings, and etc; weapon types list explosives, melee, firearms, biological, vehicular, and other attacks; target type includes twenty-two categories, such as business, government, police, aircrafts, media, and so on. Within their categories, these last three variables contain a separate category labeled "unknown". These were not treated as missingness in the analysis as I be-

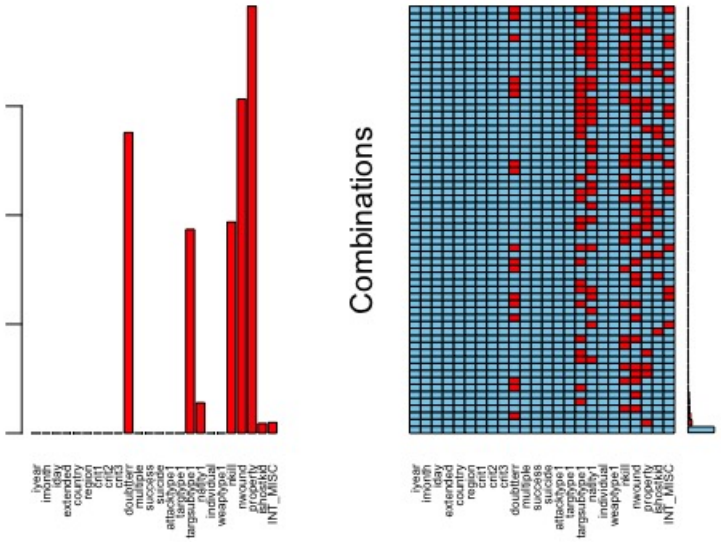


Figure 2: Missingness in Predictor Variables

lieve they carry predictive importance as a separate category. This category does not necessarily mean that the data is not available, but that a decision on a specific category could not be made given the available information.

6. Analysis and Findings

6.1. Imputation

Experimental studies have demonstrated that prior imputation on average improves classification accuracy for supervised learning techniques when compared to classification without imputation. This is especially true for polytomous regression imputation with subsequent SVM classification (Farhangfar, Kurgan and Dy, 2008). Guided by the same principle and with the goal of avoiding high classification errors, I first impute the missing variables in the GTD dataset that I plan to use as predictors for subsequent classification with SVM. As mentioned earlier, the analyzed version of the data consists of 16 complete covariates and 8 incomplete variables. The incomplete data in GTD were multiply imputed $M = 30$ via chained equations and with five Gibbs sampling iterations using the R package *mice*.[9]

[9]Only five of these imputed datasets are later used to predict groups via SVM. Approximately 20 iterations are considered as a standard for modest missing data problems (Schafer and Graham, 2002).

I use the models described in Table 1, after properly scaling the variables. The default procedure uses predictive mean matching for continuous variables and logistic regressions to impute discrete variables (Murray and Reiter, 2016). For unordered categorical data, *mice* provides polytomous logistic regression[10] as a built-in imputation model. However, the package is unable to handle variables that possess more than 50 categories, making polytomous logistic regression a non-viable model for the existing GTD multi-category variables.[11] The two multinomial predictor variables – *Target subtype* and *Nationality* – are thus imputed using random forest. The random forest method in *mice* implements Breiman's random forest algorithm based on the original Fortran code (Doove, Van Buuren and Dusseldorp, 2014).[12] *Doubt terror*, *Property*, *Hostages/Kidnappings*, *Foreign target* binary variables are imputed by the Bayesian logistic regression model.

Figure 3 demonstrates the convergence of the imputed variables from the first five imputations. For a healthy convergence, the streams should mix and the estimates should not lock in a steady trend (Gelman and Rubin, 1992). In the examples presented below (see Tables 2 and 3), the summary statistics show the distributions among the observed and imputed responses separately for the binary variable *property* and count variable *nkill* from their first imputations. The distributions for the imputed and combined values seem to be well-balanced.

Table 2: Comparison of observed and imputed values for *property*

Code	Observed		Imputed		Combined	
	Total	*Percent*	*Total*	*Percent*	*Total*	*Percent*
0	60752	40.3%	8762	44.7%	69514	40.8%
1	90019	59.7%	10817	55.2%	100836	59.2%
Total	150751	100.0%	19579	100.0%	170350	100.0%

Table 3: Comparison of observed and imputed values for *nkill*

	Observed	Imputed	Combined
Number	160668	9682	170350
Minimum	0	0	0
Maximum	1500	1500	1500
Mean	2.387	2.443	2.442
Standard deviation	11.328	11.22	11.217

[10]The model uses the *multinom* function in *nnet* (Venables and Ripley, 2002).

[11]Finding $\hat{V}(\hat{\beta})$ requires calculation of the Hessian matrix.

[12]The choice of a random forest model is not coincidental. Convergence plots of multiple different iterations demonstrate that it performs the best among available options.

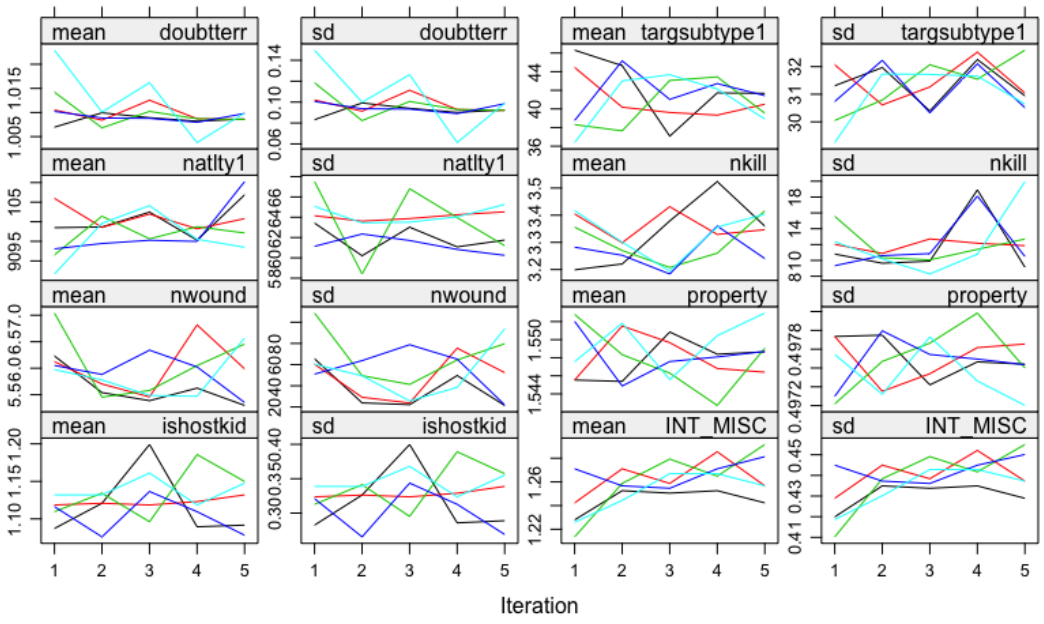


Figure 3: Diagnostic Convergence Plot for Inputed Variables

6.2. Classification

The perpetrator ("gname") variable is then appended to the first five fully imputed datasets to prepare the data for classification. I use the *e1071* package in R (Meyer and Wien, 2001), which offers an interface to the C++ implementation of *libsvm* (Chang and Lin, 2001).

The classification task starts with the proper partitioning of the data into training and validation sets. I pre-process the data before starting to partition. There is a significant imbalance in the distribution of classes in the response variable as evidenced by Figure 4. Frequencies drawn from 50 random samples differed from that of the originals (see Figure 5). Given this imbalance in the class distribution, I took a number of steps to balance out the classes. First, I took out those groups that have only carried out no more than two attacks throughout their entire existence. This step reduced the number of existing classes from 3454 classes to 1297 classes.[13] This decision is both theoretically and methodologically justifiable. "Terrorist organization" is rather an inaccurate description for a group that is capable of perpetrating only a single attack. The term merits a certain degree of capacity, which even under the most restrictive conditions should yield power strong enough to conduct more than two attacks of any scale. Furthermore, if the so-called "organization" has

[13]Taking out groups that have perpetrated only one attack would have left the data with 1760 classes. For computational efficiency, I decided to drop those with two attacks as well.

such a limited capacity that it can only mount one or two attacks throughout its entire existence, then the probability that some unattributed attack belongs to that organization is quite low compared to those with more attacks.[14]

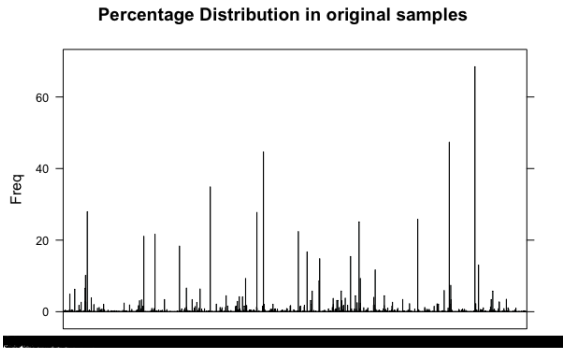
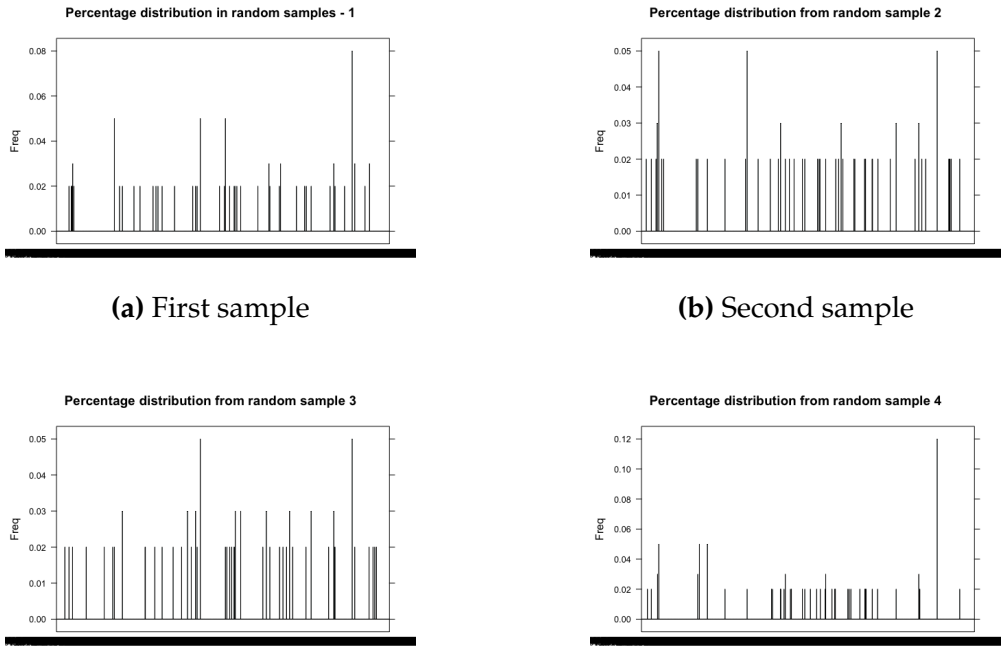


Figure 4: Original Distribution

In the next step, I merge multiple groups that, although distinct organizations in GTD, were parts of the same organizational structure in practice. For instance, I merge "Basque guerrillas", "Basque Militants", "Basque terrorists", and "Basque Separatists" as a single organization. Some events in GTD were coded as having distinct perpetrators, although they were essentially the same group. Those were also merged under a single name. In the same vein, I drop those groups that are too generic to be considered a separate group and provide no additional information as new classes. Examples include classes labeled "Youths", "Students", "Villagers", "Workers/Employers", "Unemployed Persons", "Muslims", "Rebels", "Protesters", "Insurgents", "Guerrillas", "Farmers", "Extremists", "Activists", "Gang", "Secessionists", "Strikers", "Political Activists", "Gunmen", and etc. This latter decision is also motivated by the fact that these are not specifically organized groups but rather labels given to various disconnected individuals operating in different countries by the GTD coders when identification is problematic. These changes result in a dataset of $87,973$ events and $1,244$ terrorist groups, which is significantly less than the original number of classes.

Finally, to reduce the imbalanced character of the dataset, I create stratified random samples of the data for splitting. Previous research attests to the usefulness of stratified sampling in improving the final decision border in SVMs (Akbani, Kwek and Japkowicz, 2004; Crone, Lessmann and Stahlbock, 2004). Distributions taken from the initial stratified random samples resemble that of the original sample (see Figure 6).

[14]More than 85% of organizations with only two attacks in the dataset have executed their attacks either on the same date or within the same short time period. As a robustness check, I cross-compared a random sample of events with unknown perpetrators against these organizations to identify overlapping time periods (in terms of specific dates). There was less than 2% overlap.



(a) First sample

(b) Second sample

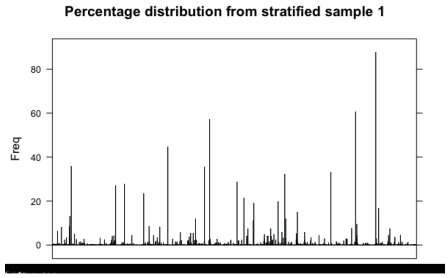
(c) Third sample

(d) Fourth sample

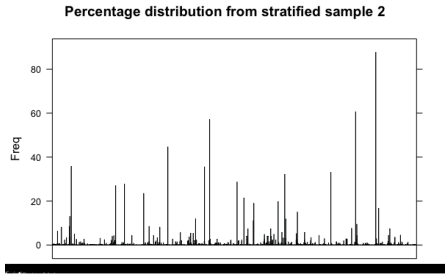
Figure 5: Illustration of distributions from the first four random samples

As discussed in the previous sections, existing SVM classification techniques lack the flexibility needed to alter the classes at every step of the classification process. To put it differently, the number of classes is considered invariant at each classification attempt, unless a researcher trains a separate classifier for each target observation. Beyond being tedious, such approach could prove impossible if the number of data points in the target dataset exceeds a certain threshold. A crucial step in accurate and efficient classification is to not only map the data back onto all existing classes, but to those that make sense in the real world. Assigning a theoretically or logically unreasonable class to data reduces the accuracy and defeats the purpose of using the SVM as a high-performance technique for prediction.

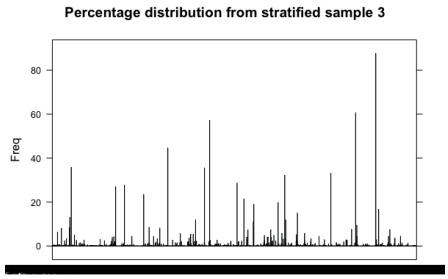
Consider the GTD dataset, for example. Following the processing stage, we are left with 1,244 terrorist groups which correspond to the equal number of classes in SVMs. There are certain characteristics in the feature space that absolutely rule out the possibility of a particular datapoint belonging to a certain number of classes. One such variable is the *iyear*, which denotes the year in which a certain attack took place. Given the nature of the dataset, we also know in which time period the terrorist organizations that are listed in the dataset have existed. Therefore, when predicting the perpetrator of an i^{th} terrorist event, carried out in year t, one should be able to confidently rule out the probability that an organization that ceased to exist at *t-1* could have committed the attack. Attaching positive probability to those groups (classes) during classification would yield problematic results. Instead, suppose the



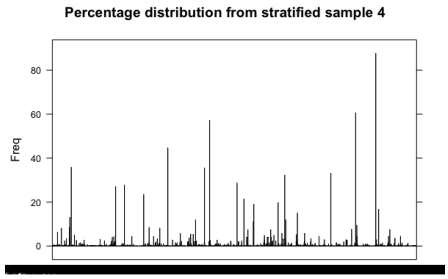
(a) First sample



(b) Second sample



(c) Third sample



(d) Fourth sample

Figure 6: Illustration of distributions from the first four stratified samples

option to drop the number of irrelevant classes is incorporated into the algorithm, such that at every stage of the mapping process, SVM chooses from different lists of classes to assign to each datapoint. In addition to limiting noise classes, the presence of which would make classification worse, this option would also grant a researcher an opportunity to integrate his prior knowledge of the concept under study into the classification task.

Interestingly, scholars studying machine learning algorithms have argued for feature selection decisions to be based on theory – that researchers "should match their preprocessing choices to their knowledge of the substantive matter at hand" (Denny and Spirling, 2018). Surprisingly, no such argumentation has been put forward in favor of applying priors to class list selection. The novelty of this approach is in allowing the feature space to constrain the classifier space in a manner that is distinct from the usual SVM mapping. Furthermore, it significantly reduces the computational complexity of multi-class classification for datasets with more than 10^3 classes.

In an effort to design such procedure, I propose to combine binary decision-trees with a multi-class SVM classifier. Combinations of these two machine learning methods is not a novel idea in and of itself. The idea was proposed and formalized to tackle multiple challenges in classification, including handling non-separable data (Madzarov, Gjorgjevikj and Chorbev, 2009), solving feature selection problems (Sug-

umaran, Muralidharan and Ramachandran, 2007), rectifying unclassifiable regions (Takahashi and Abe, 2002), and conducting automatic grouping and decomposition of classes (Cheong, Oh and Lee, 2004; Mulay, Devale and Garje, 2010; Polat and Güneş, 2009; Uribe et al., 2013). These decision-tree supported SVMs take advantage of the efficient computation of decision trees and the high classification accuracy of SVMs.

What is inherently different in my approach is that, in contrast to the above-mentioned studies that attempt to use the binary tree architecture as a decision method to achieve efficient classification given a fixed K number of classes, I utilize it to train data on a dynamically changing number of classes conditional on a certain available feature. That is, once a dataset passes through the decision tree structure proposed by the authors above, at the final node of the tree, every datapoint across all samples should have been checked against all the universe of K classes existing in the initial data for testing purposes. By contrast, if passed through the decision-tree structure I propose, each datapoint from the same dataset would be mapped against a unique list of classes, pre-designed for it with the knowledge derived from the dataset itself. Another important distinguishing factor is the deterministic nature with which decisions are handled in the decision-tree I propose. The existing decision-tree based SVMs usually tend to employ probabilistic outputs (or weights) for decision-making.

One similarity between these approaches and the current study is the training of an SVM at each node of the tree. This automatically translates into a $N-1$ SVM trainings for an N class problem for existing approaches, which is computationally intensive when N is large. This, however, is not the case for the method proposed here. Because the level of aggregation is usually much higher for the base attribute in the feature space, the number of SVMs to be trained (at each node) could potentially be much lower. Let us consider a naive example: suppose we are interested in classifying cars by their model. Let us also consider that the researcher has prior information about a feature that could potentially shrink the classifier space, such as company logo. In this case, we should be able to restrict the number of possible classes to be assigned to each datapoint to include only models of a single automobile company based on its logo. Models of all other companies would just be deemed irrelevant.

Given that the number of logos that could potentially exist within such a dataset would almost always be less than the number of models (classes), the binary decision tree approach proposed here would ultimately require less classifier training than those utilized in other studies. All other existing SVM methods, be they decision-tree based or not, would require consideration and exhaustion of all possible classes (and/or combinations) for every single datapoint. Clearly, an argument against the computational efficiency of the proposed approach could be made as well. When the number of distinct items within the attribute variable increases, so do the number of classifiers to be trained. Even in this case, a researcher would have the option to further aggregate the items based on his own preferences.

In order to evaluate the effectiveness of the proposed approach, I apply the binary decision-tree guided SVM classification to the pre-processed GTD dataset. The main attribute variable based on which the lists of classes are decided is the *iyear* variable mentioned previously, ranging between 1970-2016. For the sake of parsimony, I aggregate the decision boundaries to cover decades, rather than a single year. The classification of each sample initiates at the root of the decision tree. Then, at each following node, a decision is made about the belonging of the input point into a specific time period. Once transferred to the right terminal node, it is classified by a classifier, which has been trained based on a truncated dataset that only includes the number of relevant classes – in this case, those groups that were operating within the same decade. Each of the decade-splintered datasets contain multiple classes.[15] The process is repeated iteratively downward the tree structure until the sample reaches all the way to the very bottom node to which it has been assigned (see Figure 7). The use of validation dataset is essential for avoiding problems that might arise when analyzing the target dataset. For every classifier, the sample is divided into 80% training data and 20% validation data.

I first classify the known terror attack perpetrators separately by training each on their respective training sample with OVO approach. The sample process starts with a C classification task, using the Radial Basis Function (RBF) kernel with hyperparameters C and γ.[16] The optimal parameters were later adjusted with tuning (utilizing grid search) via ten-fold cross-validation. Based on these results, the final model was re-trained with the new parameters and the known perpetrator classes were predicted in the validation sample. Training performance was measured with accuracy, defined as the percentage of classes the model correctly classifies. The classification statistics for a standard version of SVM with full classes versus four separately-constructed classifiers with varying classes (1970-1980, 1980-1990, 1990-2000, 2000-2010) are presented in Table 4.[17]

7. Conclusions

Unknown identities of perpetrators of most terrorist attacks pose a considerable challenge to the scholars studying political violence and to law-enforcement officials engaged in counter-terrorism operations. The conventional approach used by researchers when dealing with such high levels of missingness in perpetrator iden-

[15] **1970-1980**: 6566 attacks by 306 terrorist groups; **1980-1990**: 20858 attacks by 440 groups; **1990-2000**: 15275 attacks by 470 groups; **2000-2010**: 10408 attacks by 405 groups; and **2010-2016**: 38466 attacks by 481 groups.

[16] Radial basis function kernel is defined as $e^{(-\frac{||\vec{u}-\vec{v}||^2}{2\sigma^2})}$ where the γ parameter is $\gamma = \frac{1}{2\sigma^2}$.

[17] There was a significant changepoint in terrorist activity post-2012 with increased attacks from groups such as the Taliban, Al-Shabaab, Al-Qaida, and Boko Haram (Blackwell, 2018) and penetration of new terrorist groups into the system. This explains the dominance of the number of groups and attacks in 2010-2016 period over other decades. Because of the overwhelming number of classes and events, this period was temporarily left out of the analysis.

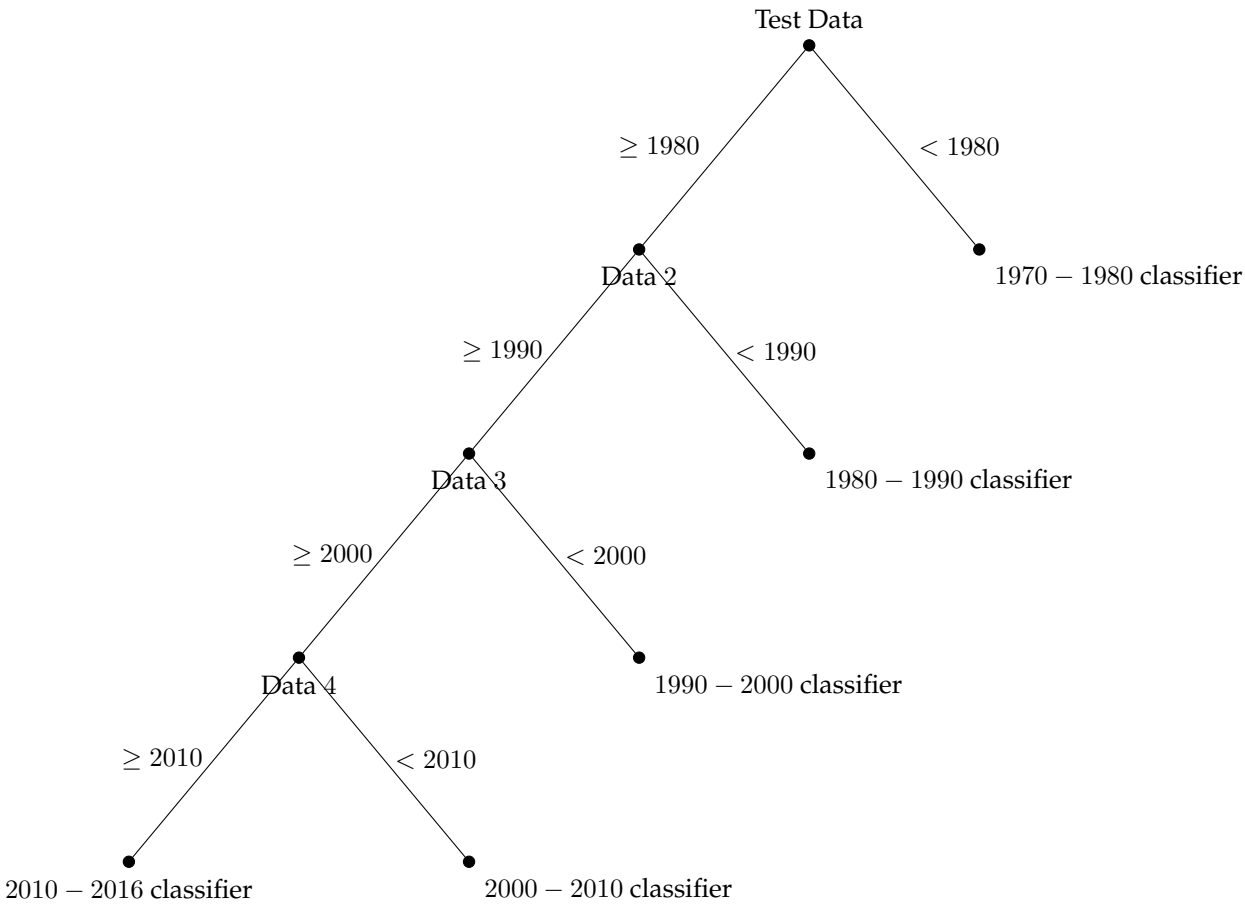


Figure 7: Decision tree for temporal classifiers

Table 4: Accuracy and Time Performance

	Training	Validation	Training Time
Full-class	0.55	0.47	43.5 hours
1970-1980	0.81	0.75	3.5 hours
1980-1990	0.76	0.68	5.5 hours
1990-2000	0.70	0.62	7 hours
2000-2010	0.81	0.73	4 hours

tities of event-based datasets has been dropping all of those cases from the study. In most cases, the analyses emerging from these studies yield biased results and reach inaccurate conclusions. This paper addresses this problem by utilizing multiple imputation and novel machine learning classification techniques to accurately recover the identities of unknown perpetrators using a comprehensive event-based terrorism dataset. The study demonstrates that a binary decision-tree guided multi-class SVM classification method can significantly increase the accuracy with which the identities of perpetrators are predicted.

There are multiple caveats to consider when utilizing the above-described method. First, one should be cautious of the choice of attribute variable based on which the classes are modified. Because this decision is left at the discretion of a researcher, it is important that a certain degree of confidence is attached to the variable choice. If the said variable is not distinguishing enough to divide classes with precision, the method might yield inaccurate mapping. Only those variables that create mutually exclusive class division should be used.

Second, despite decreasing the general CPU time needed for multi-class classification relative to other classification methods owing to reduced class sizes, the proposed method could potentially increase the overall complexity of classification if the chosen attribute variable has too many levels. Every level in the attribute variable equates to the number of separate classifier models to be trained. This study aggregates the number of levels the attribute variable has for the sake of parsimony. If used without aggregation, the *iyear* variable would have yielded 46 separate levels – one for each year between 1970-2016. It could be argued that aggregation takes away from the accuracy of prediction, given that the boundaries of aggregation are arbitrarily defined. In our example, splits into decades were selected without much justification. It is entirely possible that a terrorist organization might have purposefully not disclosed its identity at a certain time period (decade), but started claiming its attacks later. Analogously, terrorist groups might prefer to claim their attacks in the early stages of organization-building to gather supporters and media attention but disguise their involvement in later attacks once a solid organizational structure has been established. In both of these cases, when aggregated to decades, the *iyear* variable might fail to accurately divide classes given fluid boundaries. A further extension of this study could look into the performance of the method when a more fine-grained, disaggregated version of the attribute variable is used on a decision-tree.

Further, with each reduction of classes, there is an increased risk of variation loss as a result of a parallel reduction in datapoints. This problem is especially severe during the training stage and when feature space includes binary factor variables. One of the split data, for instance, led its classifier model to include a variable that had only a single level. Scaling is necessary for fast processing and the data cannot be scaled if there is no variability. The two available options to deal with it are to either skip the scaling stage or to drop variables that do not exhibit any variation post-class reduction. Both options are, by any standard, inefficient and wasteful.

While reduction of classes helps classification at the validation and classification stage, it surely creates significant information loss in the training phase. Dealing with this problem in an efficient manner is an important task for future research.

In machine learning, the essence of a classification task is to accurately and efficiently assign objects to particular categories based on their observed patterns. Efficiency is equally important as accuracy. Achieving optimal efficiency and accuracy hinges on the development of a correct coding scheme. A binary decision-tree guided SVM ensures both with a simple, but convenient, algorithmic tweak.

Terrorism research is but one field where the proposed method can prove useful. With the advancement of the field of big data and accumulation of datasets of massive complexity and size, the number of classes will continue to increase. The conventional methods will struggle to accurately classify in the light of the high number of possibilities. Allowing a machine to use nested, local classifiers with a reduced number of classes for each observation, this new technique helps to better understand, manage, and exploit the complexity of multi-class classification problems.

REFERENCES

Abe, Shigeo. 2015. "Fuzzy support vector machines for multilabel classification." *Pattern Recognition* 48(6):2110–2117.

Akbani, Rehan, Stephen Kwek and Nathalie Japkowicz. 2004. Applying support vector machines to imbalanced datasets. Springer pp. 39–50.

Allison, Paul D. 2000. "Multiple imputation for missing data: A cautionary tale." *Sociological methods & research* 28(3):301–309.

Allison, Paul D. 2001. *Missing data*. Vol. 136 Sage publications.

Allwein, Erin L, Robert E Schapire and Yoram Singer. 2000. "Reducing multiclass to binary: A unifying approach for margin classifiers." *Journal of machine learning research* 1(Dec):113–141.

Anderton, Charles H and John R Carter. 2011. "Conflict datasets: A primer for academics, policymakers, and practitioners." *Defence and Peace Economics* 22(1):21–42.

Arel-Bundock, Vincent and Krzysztof J Pelc. 2018. "When Can Multiple Imputation Improve Regression Estimates?" *Political Analysis* 26(2):240–245.

Arva, Bryan and John Beieler. 2014. Dealing with missing data in group-level studies of terrorism. pp. 28–31.

Asal, Victor, R Karl Rethemeyer and Ian Anderson. 2009. "Big allied and dangerous (baad) database 1-lethality data, 1998-2005." *Codebook. Project on Violent Conflict (PVC), University at Albany, State University of New York (http://the data. harvard. edu/dvn/dv/start/faces/study/Study Page. xhtml* .

Azur, Melissa J, Elizabeth A Stuart, Constantine Frangakis and Philip J Leaf. 2011. "Multiple imputation by chained equations: what is it and how does it work?" *International journal of methods in psychiatric research* 20(1):40–49.

Bagozzi, Benjamin E and Ore Koren. 2017. Using machine learning methods to identify atrocity perpetrators. IEEE pp. 3042–3051.

Batista, Gustavo EAPA and Maria Carolina Monard. 2003. "An analysis of four missing data treatment methods for supervised learning." *Applied artificial intelligence* 17(5-6):519–533.

Bauer, Vincent, Keven Ruby and Robert Pape. 2017. "Solving the problem of unattributed political violence." *Journal of Conflict Resolution* 61(7):1537–1564.

Beck, Nathaniel, Gary King and Langche Zeng. 2000. "Improving quantitative studies of international conflict: A conjecture." *American Political Science Review* 94(1):21–35.

Blackwell, Matthew. 2018. "Game Changers: Detecting Shifts in Overdispersed Count Data." *Political Analysis* 26(2):230–239.

Bodner, Todd E. 2006. "Missing data: Prevalence and reporting practices." *Psychological Reports* 99(3):675–680.

Bose, Raj Chandra and Dwijendra K Ray-Chaudhuri. 1960. "On a class of error correcting binary group codes." *Information and control* 3(1):68–79.

Boser, Bernhard E, Isabelle M Guyon and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. ACM pp. 144–152.

Bradley, Paul S and Olvi L Mangasarian. 1998. Feature selection via concave minimization and support vector machines. Vol. 98 pp. 82–90.

Brand, Jaap. 1999. *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*.

Bredensteiner, Erin J and Kristin P Bennett. 1999. *Multicategory classification by support vector machines*. Springer pp. 53–79.

Breiman, Leo. 2017. *Classification and regression trees*. Routledge.

Burscher, Bjorn, Rens Vliegenthart and Claes H De Vreese. 2015. "Using supervised machine learning to code policy issues: Can classifiers generalize across contexts?" *The Annals of the American Academy of Political and Social Science* 659(1):122–131.

Buuren, S van and CGM Oudshoorn. 2000. "Multivariate imputation by Chained equations: MICE V1. 0 user's manual.".

Buuren, S van and Karin Groothuis-Oudshoorn. 2010. "mice: Multivariate imputation by chained equations in R." *Journal of statistical software* pp. 1–68.

Cantú, Francisco and Sebastián M Saiegh. 2011. "Fraudulent democracy? An analysis of Argentina's Infamous Decade using supervised machine learning." *Political Analysis* 19(4):409–433.

Cao, Bin, Dou Shen, Jian-Tao Sun, Qiang Yang and Zheng Chen. 2007. Feature selection in a kernel space. ACM pp. 121–128.

Carter, David B. 2012. "A blessing or a curse? State support for terrorist groups." *International Organization* 66(1):129–151.

Chang, Chih-Chung and Chih-Jen Lin. 2001. "LIBSVM: A library for support vector machines [EB/OL].".

Cheong, Sungmoon, Sang Hoon Oh and Soo-Young Lee. 2004. "Support vector machines with binary tree architecture for multi-class classification." *Neural Information Processing-Letters and Reviews* 2(3):47–51.

Collins, Linda M, Joseph L Schafer and Chi-Ming Kam. 2001. "A comparison of inclusive and restrictive strategies in modern missing data procedures." *Psychological methods* 6(4):330.

Cortes, Corinna and Vladimir Vapnik. 1995. "Support-vector networks." *Machine learning* 20(3):273–297.

Crammer, Koby and Yoram Singer. 2001. "On the algorithmic implementation of multiclass kernel-based vector machines." *Journal of machine learning research* 2(Dec):265–292.

Crammer, Koby and Yoram Singer. 2002. "On the learnability and design of output codes for multiclass problems." *Machine learning* 47(2-3):201–233.

Cristiani, Nello and Shawe J Taylor. 2000. "An introduction to support vector machines.".

Crone, Sven F, Stefan Lessmann and Robert Stahlbock. 2004. Empirical comparison and evaluation of classifier performance for data mining in customer relationship management. Vol. 1 IEEE pp. 443–448.

Damoulas, Theodoros and Mark A Girolami. 2009. "Combining feature spaces for classification." *Pattern Recognition* 42(11):2671–2683.

Davenport, Christian and Patrick Ball. 2002. "Views to a kill: Exploring the implications of source selection in the case of Guatemalan state terror, 1977-1995." *Journal of conflict resolution* 46(3):427–450.

De Marchi, Scott, Christopher Gelpi and Jeffrey D Grynaviski. 2004. "Untangling neural nets." *American Political Science Review* 98(2):371–378.

Dempster, Arthur P, Nan M Laird and Donald B Rubin. 1977. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the royal statistical society. Series B (methodological)* pp. 1–38.

Denny, Matthew J and Arthur Spirling. 2018. "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it." *Political Analysis* 26(2):168–189.

Dietterich, Thomas G and Ghulum Bakiri. 1994. "Solving multiclass learning problems via error-correcting output codes." *Journal of artificial intelligence research* 2:263–286.

Dogan, Urün, Tobias Glasmachers and Christian Igel. 2016. "A unified view on multi-class support vector classification." *Journal of Machine Learning Research* 17(45):1–32.

Doove, Lisa L, Stef Van Buuren and Elise Dusseldorp. 2014. "Recursive partitioning for missing data imputation in the presence of interaction effects." *Computational Statistics & Data Analysis* 72:92–104.

Douglass, Rex W and Kristen A Harkness. 2018. "Measuring the landscape of civil war: Evaluating geographic coding decisions with historic data from the Mau Mau rebellion." *Journal of Peace Research* 55(2):190–205.

Drechsler, Jörg and Susanne Rässler. 2008. *Does convergence really matter?* Springer pp. 341–355.

Dugan, Laura, Gary LaFree and Heather Fogg. 2006. A first look at domestic and international global terrorism events, 1970–1997. Springer pp. 407–419.

Durrant, Gabriele B. 2009. "Imputation methods for handling itemnonresponse in practice: methodological issues and recent debates." *International Journal of Social Research Methodology* 12(4):293–304.

Enders, Craig K. 2010. *Applied missing data analysis*. Guilford Press.

Engene, Jan Oskar. 2007. "Five decades of terrorism in Europe: The TWEED dataset." *Journal of Peace Research* 44(1):109–121.

Ezzati-Rice, Trena M, Wayne Johnson, Meena Khare, Roderick JA Little, Donald B Rubin and Joseph L Schafer. 1995. A simulation study to evaluate the performance of model-based multiple imputations in NCHS health examination surveys. pp. 257–266.

Farhangfar, Alireza, Lukasz Kurgan and Jennifer Dy. 2008. "Impact of imputation of missing values on classification error for discrete data." *Pattern Recognition* 41(12):3692–3705.

Fortes, I, L Mora-López, R Morales and F Triguero. 2006. "Inductive learning models with missing values." *Mathematical and Computer Modelling* 44(9-10):790–806.

Fortna, Virginia Page, Nicholas Lotito and Michael Rubin. 2014. "The Causes and Consequences of Terrorism:[toward] Introducing a New Dataset on Terrorism in Civil Conflicts 1970-2012." *International Studies Association, Toronto, March* .

Gelman, Andrew. 2004. "Parameterization and Bayesian modeling." *Journal of the American Statistical Association* 99(466):537–545.

Gelman, Andrew and Donald B Rubin. 1992. "Inference from iterative simulation using multiple sequences." *Statistical science* pp. 457–472.

Ghahramani, Zoubin and Michael I Jordan. 1994. Supervised learning from incomplete data via an EM approach. pp. 120–127.

Glynn, Robert J, Nan M Laird and Donald B Rubin. 1993. "Multiple imputation in mixture models for nonignorable nonresponse with follow-ups." *Journal of the American Statistical Association* 88(423):984–993.

Graham, John W and Stewart I Donaldson. 1993. "Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data." *Journal of Applied Psychology* 78(1):119.

Green, Donald P and Holger L Kern. 2012. "Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees." *Public opinion quarterly* 76(3):491–511.

Greenland, Sander and William D Finkle. 1995. "A critical look at methods for handling missing covariates in epidemiologic regression analyses." *American journal of epidemiology* 142(12):1255–1264.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21(3):267–297.

Grimmer, Justin, Solomon Messing and Sean J Westwood. 2017. "Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods." *Political Analysis* 25(4):413–434.

Guyon, I. 1999. "SVM Application Survey: http://www. clopinet. com." *SVM. applications. html* .

Guyon, Isabelle and André Elisseeff. 2003. "An introduction to variable and feature selection." *Journal of machine learning research* 3(Mar):1157–1182.

Hastie, Trevor and Robert Tibshirani. 1998. Classification by pairwise coupling. pp. 507–513.

He, Yulei, Alan M Zaslavsky, MB Landrum, DP Harrington and P Catalano. 2010. "Multiple imputation in a large-scale complex survey: a practical guide." *Statistical methods in medical research* 19(6):653–670.

Heckerman, David, David Maxwell Chickering, Christopher Meek, Robert Rounthwaite and Carl Kadie. 2000. "Dependency networks for inference, collaborative filtering, and data visualization." *Journal of Machine Learning Research* 1(Oct):49–75.

Heiler, Matthias, Daniel Cremers and Christoph Schnörr. 2001. "Efficient feature subset selection for support vector machines." *Technical reports* 1.

Hermes, Lothar and Joachim M Buhmann. 2000. Feature selection for support vector machines. Vol. 2 IEEE pp. 712–715.

Hill, Daniel W and Zachary M Jones. 2014. "An empirical evaluation of explanations for state repression." *American Political Science Review* 108(3):661–687.

Hocke, Peter. 1998. *Determining the selection bias in local and national newspaper reports on protest events*. Acts of dissent: New developments in the study of protest ed. WZB.

Hoeyland, Bjoern and HM Nygaard. 2011. "Missing uncertainty and uncertain missing in the study of civil war onset." *Oslo: Department of Political Science, University of Oslo, Working Paper* .

Honaker, James and Gary King. 2010. "What to do about missing values in time-series crosssection data." *American Journal of Political Science* 54(2):561–581.

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

Horowitz, Michael C and Philip BK Potter. 2014. "Allying to kill: Terrorist intergroup cooperation and the consequences for lethality." *Journal of Conflict Resolution* 58(2):199–225.

Hsu, Chih-Wei and Chih-Jen Lin. 2002. "A comparison of methods for multiclass support vector machines." *IEEE transactions on Neural Networks* 13(2):415–425.

Ibrahim, Joseph G, Ming-Hui Chen, Stuart R Lipsitz and Amy H Herring. 2005. "Missing-data methods for generalized linear models: A comparative review." *Journal of the American Statistical Association* 100(469):332–346.

Imai, Kosuke and Aaron Strauss. 2010. "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.

Imai, Kosuke and Marc Ratkovic. 2013. "Estimating treatment effect heterogeneity in randomized program evaluation." *The Annals of Applied Statistics* 7(1):443–470.

Jenkins, Brian, Bonnie Cordes and Konard Kellen. 1985. "A Conceptual Framework For Analyzing Terrorist Groups." *Santa Monica, CA: RAND* p. 30.

Jiang, Jing and Chengxiang Zhai. 2007. "A Systematic Exploration of the Feature Space for Relation Extraction.".

Jones, Seth G and Martin C Libicki. 2008. *How terrorist groups end: Lessons for countering al Qa'ida*. Rand Corporation.

Juergensmeyer, Mark. 2017. *Terror in the mind of God: The global rise of religious violence*. Vol. 13 Univ of California Press.

Kennickell, Arthur B. 1991. Imputation of the 1989 Survey of Consumer Finances: Stochastic relaxation and multiple imputation. Vol. 1.

King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 2001. "Analyzing incomplete political science data: An alternative algorithm for multiple imputation." *American political science review* 95(1):49–69.

Knerr, Stefan, Léon Personnaz and Gérard Dreyfus. 1990. *Single-layer learning revisited: a stepwise procedure for building and training a neural network*. Springer pp. 41–50.

Kohavi, Ron and George H John. 1997. "Wrappers for feature subset selection." *Artificial intelligence* 97(1-2):273–324.

Kotsiantis, Sotiris B, I Zaharakis and P Pintelas. 2007. "Supervised machine learning: A review of classification techniques." *Emerging artificial intelligence applications in computer engineering* 160:3–24.

Kressel, Ulrich. 1998. "Pairwise classification and support vector machines." *Advances in kernel methods: support vector learning* pp. 255–268.

Kyung, Minjung, Jeff Gill and George Casella. 2011. "New findings from terrorism data: Dirichlet process randomeffects models for latent groups." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 60(5):701–721.

LaFree, Gary and Laura Dugan. 2007. "Introducing the global terrorism database." *Terrorism and Political Violence* 19(2):181–204.

Lall, Ranjit. 2016. "How multiple imputation makes a difference." *Political Analysis* 24(4):414–433.

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

Lee, Yoonkyung, Yi Lin and Grace Wahba. 2004. "Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data." *Journal of the American Statistical Association* 99(465):67–81.

Lin, Chun-Fu and Sheng-De Wang. 2002. "Fuzzy support vector machines." *IEEE transactions on neural networks* 13(2):464–471.

Lipsitz, Stuart R and Joseph G Ibrahim. 1996. "A conditional model for incomplete covariates in parametric regression models." *Biometrika* 83(4):916–922.

Little, Roderick JA. 1988. "Missing-data adjustments in large surveys." *Journal of Business & Economic Statistics* 6(3):287–296.

Little, Roderick JA and Donald B Rubin. 1987*a*. "Statistical analysis with missing data." *New York: Wiley, 1987* .

Little, Roderick JA and Donald B Rubin. 1989. "The analysis of social science data with missing values." *Sociological Methods & Research* 18(2-3):292–326.

Little, Roderick JA and Donald B Rubin. 2014. *Statistical analysis with missing data*. Vol. 333 John Wiley & Sons.

Little, Roderick JA and Donald Rubin. 1987*b*. "Analysis with missing data.".

Madzarov, Gjorgji, Dejan Gjorgjevikj and Ivan Chorbev. 2009. "A multi-class SVM classifier utilizing binary decision tree." *Informatica* 33(2).

Mayoraz, Eddy and Ethem Alpaydin. 1999. Support vector machines for multi-class classification. Springer pp. 833–842.

Meyer, David and FH Technikum Wien. 2001. "Support vector machines." *R News* 1(3):23–26.

Mickolus, Edward F, Todd Sandler, Jean Marie Murdock and Peter A Flemming. 2011. *International Terrorism: Attributes of Terrorist Events: ITERATE, 1968-2011*. Vinyard Software.

Mierswa, Ingo and Katharina Morik. 2005. "Automatic feature extraction for classifying audio data." *Machine learning* 58(2-3):127–149.

Moeller, Susan D. 2002. *Compassion fatigue: How the media sell disease, famine, war and death*. Routledge.

Montgomery, Jacob M, Santiago Olivella, Joshua D Potter and Brian F Crisp. 2015. "An Informed Forensics Approach to Detecting Vote Irregularities." *Political Analysis* 23(04):488–505.

Muchlinski, David, David Siroky, Jingrui He and Matthew Kocher. 2016. "Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data." *Political Analysis* 24(1):87–103.

Mueller, Carol. 1997. "International press coverage of East German protest events, 1989." *American sociological review* pp. 820–832.

Mueller, Hannes and Christopher Rauh. 2018. "Reading between the lines: Prediction of political violence using newspaper text." *American Political Science Review* 112(2):358–375.

Mulay, Snehal A, PR Devale and GV Garje. 2010. "Intrusion detection system using support vector machine and decision tree." *International Journal of Computer Applications* 3(3):40–43.

Murray, Jared S and Jerome P Reiter. 2016. "Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence." *Journal of the American Statistical Association* 111(516):1466–1479.

Nardulli, Peter F, Scott L Althaus and Matthew Hayes. 2015. "A progressive supervised-learning approach to generating rich civil strife data." *Sociological methodology* 45(1):148–183.

Nasiri, Jalal A, Nasrollah Moghadam Charkari and Saeed Jalili. 2015. "Least squares twin multi-class classification support vector machine." *Pattern Recognition* 48(3):984–992.

Nemeth, Stephen. 2014. "The effect of competition on terrorist group operations." *Journal of Conflict Resolution* 58(2):336–362.

Neumann, Julia, Christoph Schnörr and Gabriele Steidl. 2005. "Combined SVM-based feature selection and classification." *Machine learning* 61(1-3):129–150.

Öberg, Magnus and Kristine Höglund. 2011. *Doing Empirical Peace Research*. Routledge pp. 15–25.

Öberg, Magnus and Margareta Sollenberg. 2011. "Gathering conflict information using news resources." *Understanding Peace Research: Methods and Challenges, Abingdon: Routledge* pp. 47–73.

Pal, Mahesh. 2008. "Multiclass approaches for support vector machine based land cover classification." *arXiv preprint arXiv:0802.2411* .

Pal, Mahesh and Giles M Foody. 2010. "Feature selection for classification of hyperspectral data by SVM." *IEEE Transactions on Geoscience and Remote Sensing* 48(5):2297–2307.

Pampaka, Maria, Graeme Hutcheson and Julian Williams. 2016. "Handling missing data: analysis of a challenging data set using multiple imputation." *International Journal of Research & Method in Education* 39(1):19–37.

Perl, Raphael. 2006. Trends in terrorism: 2006. Library of Congress Washington DC Congressional Research Service.

Peterson, William Wesley, W Wesley, EJ Weldon Jr Peterson and EJ Weldon. 1972. *Error-correcting codes*. MIT press.

Platt, John C, Nello Cristianini and John Shawe-Taylor. 2000. Large margin DAGs for multiclass classification. pp. 547–553.

Polat, Kemal and Salih Güneş. 2009. "A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems." *Expert Systems with Applications* 36(2):1587–1592.

Poulos, Jason and Rafael Valle. 2016. "Missing Data Imputation for Supervised Learning." *arXiv preprint arXiv:1610.09075* .

Price, Bryan C. 2012. "Targeting top terrorists: How leadership decapitation contributes to counterterrorism." *International Security* 36(4):9–46.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin and Dragomir R Radev. 2010. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54(1):209–228.

Raghunathan, Trivellore E, James M Lepkowski, John Van Hoewyk and Peter Solenberger. 2001. "A multivariate technique for multiply imputing missing values using a sequence of regression models." *Survey methodology* 27(1):85–96.

RAND Database of Worldwide Terrorism Incidents. 2018.
URL: *https://www.rand.org/nsrd/projects/terrorism-incidents.html*

Reiter, Jerome P and Trivellore E Raghunathan. 2007. "The multiple adaptations of multiple imputation." *Journal of the American Statistical Association* 102(480):1462–1471.

Rubin, DB. 1987. "Multiple imputation for nonresponse in surveys. Wiley series in probability and mathematical statistics applied probability and statistics.".

Rubin, Donald B. 1976. "Inference and missing data." *Biometrika* 63(3):581–592.

Rubin, Donald B. 1996. "Multiple imputation after 18+ years." *Journal of the American statistical Association* 91(434):473–489.

Rubin, Donald B. 2003. "Nested multiple imputation of NMES via partially incompatible MCMC." *Statistica Neerlandica* 57(1):3–18.

Rubin, Donald B. 2004. *Multiple imputation for nonresponse in surveys*. Vol. 81 John Wiley & Sons.

Rubin, Donald B and Joseph L Schafer. 1990. Efficiently creating multiple imputations for incomplete multivariate normal data. Vol. 83 American Statistical Association p. 88.

Rubin, Donald B and Nathaniel Schenker. 1986. "Multiple imputation for interval estimation from simple random samples with ignorable nonresponse." *Journal of the American statistical Association* 81(394):366–374.

Rumelhart, David E, Geoffrey E Hinton and Ronald J Williams. 1986. "Learning representations by back-propagating errors." *nature* 323(6088):533.

Saar-Tsechansky, Maytal and Foster Provost. 2007. "Handling missing values when applying classification models." *Journal of machine learning research* 8(Jul):1623–1657.

Schafer, Joseph L. 1997. *Analysis of incomplete multivariate data*. CRC press.

Schafer, Joseph L and John W Graham. 2002. "Missing data: our view of the state of the art." *Psychological methods* 7(2):147.

Schafer, Joseph L and Maren K Olsen. 1998. "Multiple imputation for multivariate missing-data problems: A data analyst's perspective." *Multivariate behavioral research* 33(4):545–571.

Schmidt, Alex P and Albert I Jongman. 1988. "Political Terrorism: A Research Guide to Concepts, Theories, Data Bases, and Literature.".

Schölkopf, Bernhard and Alexander J Smola. 2002. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Sebastiani, Fabrizio. 2002. "Machine learning in automated text categorization." *ACM computing surveys (CSUR)* 34(1):1–47.

Sheehan, Ivan Sascha. 2012. *Assessing and comparing data sources for terrorism research*. Springer pp. 13–40.

Skrondal, Anders and Sophia Rabe-Hesketh. 2004. *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Crc Press.

Smola, Alexander J, SVN Vishwanathan and Thomas Hofmann. 2005. Kernel Methods for Missing Variables. Citeseer.

START. 2016. "GTD Codebook: Inclusion Criteria and Variables.".
URL: *https://www.start.umd.edu/gtd/downloads/Codebook.pdf*

Stuart, Elizabeth A, Melissa Azur, Constantine Frangakis and Philip Leaf. 2009. "Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative." *American journal of epidemiology* 169(9):1133–1139.

Sugumaran, V, V Muralidharan and KI Ramachandran. 2007. "Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing." *Mechanical systems and signal processing* 21(2):930–942.

Takahashi, Fumitake and Shigeo Abe. 2002. Decision-tree-based multiclass support vector machines. Vol. 3 IEEE pp. 1418–1422.

Tanner, Martin A and Wing Hung Wong. 1987. "The calculation of posterior distributions by data augmentation." *Journal of the American statistical Association* 82(398):528–540.

Tokdemir, Efe and Seden Akcinaroglu. 2016. "Reputation of Terror Groups Dataset: Measuring popularity of terror groups." *Journal of Peace Research* 53(2):268–277.

Tresp, Volker, Ralph Neuneier and Subutai Ahmad. 1995. Efficient methods for dealing with missing data in supervised learning. pp. 689–696.

Uribe, Juan Sebastian, Nazih Mechbal, Marc Rébillat, Karima Bouamama and Marco Pengov. 2013. Probabilistic decision trees using SVM for multi-class classification. IEEE pp. 619–624.

Van Buuren, Stef, Hendriek C Boshuizen and Dick L Knook. 1999. "Multiple imputation of missing blood pressure covariates in survival analysis." *Statistics in medicine* 18(6):681–694.

Van Buuren, Stef, Jaap PL Brand, Catharina GM Groothuis-Oudshoorn and Donald B Rubin. 2006. "Fully conditional specification in multivariate imputation." *Journal of statistical computation and simulation* 76(12):1049–1064.

Vapnik, Vladimir. 1998. *Statistical learning theory. 1998*. Wiley, New York.

Venables, WN and BD Ripley. 2002. "Modern Applied Statistics with S. Springer, New York, NY.".

Weidmann, Nils B. 2013. "The higher the better? The limits of analytical resolution in conflict event datasets." *Cooperation and Conflict* 48(4):567–576.

Weidmann, Nils B. 2015. "On the accuracy of media-based conflict event data." *Journal of Conflict Resolution* 59(6):1129–1149.

Weston, JW and C Watkins. 1999. "Multi-class support vector machines, presented at the Proc." *Brussels, Belgium* .

Wigle, John. 2010. "Introducing the worldwide incidents tracking system (WITS)." *Perspectives on Terrorism* 4(1).

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

Williams, David, Xuejun Liao, Ya Xue, Lawrence Carin and Balaji Krishnapuram. 2007. "On classification with incomplete data." *IEEE transactions on pattern analysis and machine intelligence* 29(3):427–436.

Wothke, Werner. 2000. "Longitudinal and multigroup modeling with missing data.".

Yan, Zhiguo, Zhizhong Wang and Hongbo Xie. 2008. "The application of mutual information-based feature selection and fuzzy LS-SVM-based classifier in motion classification." *Computer Methods and Programs in Biomedicine* 90(3):275–284.

Young, Joseph K and Laura Dugan. 2014. "Survival of the fittest: Why terrorist groups endure." *Perspectives on Terrorism* 8(2).

Zareapoor, Masoumeh, Pourya Shamsolmoali, Deepak Kumar Jain, Haoxiang Wang and Jie Yang. 2017. "Kernelized support vector machine with deep learning: an efficient approach for extreme multiclass dataset." *Pattern Recognition Letters* .

Zhang, Paul. 2003. "Multiple imputation: theory and method." *International Statistical Review* 71(3):581–592.