European Journal of Operational Research 255 (2016) 522–530

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Decision Support

# Comparison of imputation methods for discriminant analysis with strategically hidden data

Juheng Zhang[a,*], Haldun Aytug[b]

[a] Operations and Information Systems, Manning School of Business, University of Massachusetts Lowell, Lowell, MA 01826, United States
[b] Information Systems and Operations Management, Warrington College of Business Administration, University of Florida, Gainesville, FL 32611, United States

ARTICLE INFO

*Article history:*
Received 15 January 2015
Accepted 27 May 2016
Available online 7 June 2016

*Keywords:*
Analytics
Sampling
Missing data
Information disclosure
Information theory

ABSTRACT

In many situations, data may be selectively presented by data providers to achieve desirable but undeserved decision outcomes from decision makers. Decisions taken without considering strategic information revelation might be biased. We revisit and study the properties of two methods handling strategically missing data in a classification context. The asymptotic analysis suggests that when the training sets are sufficiently large these methods outperform the conventional methods handling missing data that do not consider strategic motivations of agents (e.g., Average method and Similarity method). Scale-up experiments support the theoretical findings and show that as the training size increases the misclassification rates of those methods decrease. We show that sampling can be used to efficiently identify sufficient information for the imputation methods to treat strategically missing data.

Published by Elsevier B.V.

## 1. Introduction and motivation

We are in the era of data, where for most decisions there are vast amounts of readily available data. Data are made available from different sources in various formats and with high speed. For instance, statistics (Congress, 2014) show that as of March 2014 the web archives grew at a rate of about five terabytes per month. Dataset size grows to a point where new techniques, such as parallel programming based MapReduce, are needed. These are some of the new challenges of decision making in a data rich environment. Although data have been made increasingly available to public, there are many situations in which a decision maker has to deal with missing data. The following scenarios describe what type of missing data we are interested in.

*Scenario 1*: Ecommerce marketplace. Online sellers may present information favorable to their products and hide unfavorable information to attract more potential buyers. For instance, on eBay, sellers may leave out the information of product origin because consumers may have concerns on the place of production. In an extreme case, online sellers with poor reputations even hide their past transactions by registering a new online identity (Ba, 2001; Baron, 2002). The information of product attributes or sellers may be strategically missing from buyers in online transactions.

*Scenario 2*: Price comparison search engines. Online search engines may selectively present search results. For example, BizRate.com is found by the study (Shmueli, Jank, & Bapna, 2013) that it excludes Amazon.com as a retailer selling a specific popular DVD player, in spite of the fact that it was early shown by the study (Pan, Ratchford, & Shankar, 2003) to be the website returning to customers with the most complete list comparing to several other price comparison websites. The selected search results are returned to users and certain products may be missing from the search results on purpose.

*Scenario 3:* Self information reporting websites. Users neglect to input certain information due to privacy or other concerns. For instance, insurance applicants may purposely not inform their insurance companies about their life habits (e.g., smoking) in hope for a lower premium or higher chance of getting insured (Insure.com, 2010). Limited information disclosure is common in financial markets (Healy & Palepu, 2001; Higgins, 2012; Smith & Drucker, 2002; Zhang, 2015), school applications (Braun, Dwenger, & Kubler, 2010), etc. Certain information may be intentionally left out from decision makers.

All these examples point to the fact that, in many contexts, data might be missing because the data providers purposefully hide critical details. We call this strategically missing data and note that it is different from distorted or noisy data. Data distortion refers to the case where data are reported but potentially strategically altered, while strategic hiding refers to the case of data providers not disclosing relevant information. Strategically missing

* Corresponding author. Tel.: +1 978 934 3261.
*E-mail addresses:* juheng_zhang@uml.edu (J. Zhang), aytugh@ufl.edu (H. Aytug).

http://dx.doi.org/10.1016/j.ejor.2016.05.052
0377-2217/Published by Elsevier B.V.

data may cause problems in data analysis and may bias the decision maker's decision rules. For instance, in credit scoring (e.g., Finlay, 2010), credit card applications may get incorrectly accepted (or rejected) when information is strategically missing; in online transaction data (Li & Jacob, 2008), untrustworthy sellers may get misclassified as trustworthy when they strategically hide their bad transaction records from buyers.

Strategic behaviors of data providers are observed in many real world situations including intrusion detection (Mahoney & Chan, 2002), fraud detection (Phua, Lee, Smith, & Gayler, 2005), and spam detection (Fawcett, 2003). In this study, we investigate the strategically missing data problem in the classification contexts and examine the properties of several methods developed for handling missing data, the so-called imputation methods, under the assumption that data presented by data providers (who we call agents) may be strategically missing. We conduct an asymptotic analysis to quantify the theoretical properties of imputation methods in an ideal case where the sample size is infinite. For all practical purposes, we conduct an empirical study and show that the misclassification rates of the imputation methods handling strategically missing data converge to zero as sample size increases, but the convergence of misclassification rates of the imputation methods for randomly missing data is not observed to be zero.

The rest of this paper is organized in the following way. In Section 2, we review adversarial classification studies, imputation methods that handle missing data, and sampling methods. In Section 3, we examine four imputation methods handling missing data in detail, two of which are for strategically missing data and the other two for missing data at random. In Section 4, we conduct an asymptotic analysis on four imputation methods. In Section 5, the empirical results are provided to demonstrate how the performance varies with the sample size. We discuss managerial implications in Section 6. We conclude the results and give possible future research in Section 7.

## 2. Literature review

This study is related to the area of adversarial learning (Boylu, Aytug, & Koehler, 2010a,2010b; Dalvi, Domingos, Mausam, & Verma, 2004; Lowd & Meek, 2005). Adversarial learning refers to a problem where, in the classification contexts, a decision maker classifies objects into a group, for instance, either a positive or negative group, while the agents, called adversary, attempt to obtain a preferred classification outcome (for example, positive group membership) by purposely distorting information. The existing studies of adversarial classification investigate the strategic behaviors of the adversary, but focus on data distortion. Zhang, Aytug, and Koehler (2014) and Dekel, Shamir, and Xiao (2010) consider "strategic disclosure" and both provide a theoretical basis but do not discuss behavior and perfromance of their approaches in sampling situations.

Several methods have been proposed to handle missing data (Schafer & Graham, 2002; Schafer & Olsen, 1998), but the assumption is data missing at random rather than missing strategically. Rubin (1976) reviews conventional methods handling missing data, including deletion methods and imputation methods. Deletion methods, such as Listwise Deletion method or Pairwise Deletion method, remove records with missing data. The limitation of deletion methods is that, discarding observations with missing values, the methods yield data loss and result in a new data subset with very small or even empty data when a dataset is sparse and can cause large standard errors because little information is utilized. Another approach is to replace each missing value with some reasonable guess (an imputation, default value, or estimate), so that the analysis can proceed as if a data set has no missing value. The Average method (Wilks, 1932) uses the average of observed values of an attribute as the imputation of missing values of the attribute. The Similarity method (Batista & Monard, 2003) replaces the missing value of a record by the observed value of the most similar record. We refer readers to survey papers (Donders, van der Heijden, Stijnen, & Moons, 2006; Schafer & Graham, 2002) and the book (Allison, 2001) for detailed discussion on imputation methods include the Average method, Similarity method, Regression method, Maximum Likelihood method, Expectation Maximization (EM), etc. The conventional methods, however, do not consider the strategic behaviors of agents on hiding information. The decision maker may make biased decisions when these methods are applied to settings where data are missing due to strategic reasons.

In the field of sampling theory, researchers study how sampling affects prediction accuracy. Shmueli et al. (2013) review different sampling methods, such as random sampling, systematic sampling, stratified sampling, cluster sampling, and multistage sampling. Researchers also investigate sequential sampling (Lam, Li, Ip, & Wong, 2006), on-line sampling (Lee et al., 2001), dynamic sampling (John & Langley, 1996; Philpott & de Matos, 2012), etc. For example, John and Langley (1996) point out that the strategy of the sampling is to compare the accuracy of sample to that of population. Zaki, Parthasarathy, Li, and Ogihara (1997) show that sampling can dramatically reduce the number of records to be considered and speed up computation process with high confidence.

## 3. Imputation methods handling missing data

Imputation methods find default values to replace missing values. The objective of the decision maker is to find such default values that any strategic data providers cannot gain favorable but undeserved outcomes. We consider this task under the classification context. Support Vector Machines (SVMs) has shown a great performance on solving classification problems. The SVMs method (Vapnik, 1998) finds a classification rule $(w, b)$ to separate the points of a sample set, $x_i$, $i = 1, \ldots, \ell$, into two classes, the positive group ($y_i = 1$) and the negative group $y_i = -1$. In its simplest form, the SVMs method is similar to any other Linear Discriminant Analysis (LDA) method, and it is also possible to think of SVM classifier, $(w, b)$, as a linear scoring function with a threshold set at zero. We will use the term "separating hyperplane" for $(w, b)$. The SVMs method is unique among LDA methods since it is the first one that finds a linear function that maximizes the distance between two classes (i.e., margin) as a measure of predictive performance. Under the assumption of the separability of data points, the SVMs method solves a convex optimization problem as defined below.

$$\underset{w,b}{Min}\ w'w \quad s.t. \quad y_i(w'x_i + b) \geq 1 \quad i = 1, ..., \ell. \tag{1}$$

In (1), the objective is set to find a classification hyperplane with the largest possible margin and the constraints are set to have all points in the training set classified correctly.

Considering strategic behaviors of agents, the decision maker's problem is different. Instead of true values of data points, some attribute values can be strategically hidden from the decision maker. The decision maker needs to find default values to replace missing values while building the decision rule so that agents cannot game the decision rule. In this setting, the decision maker's problem is formulated as the following:

$$\underset{(w,b)\in H}{Min}\ w'w + C\sum_{i=1}^{\ell} \max(0, 1 - w't_i(w, b, d) - b) \tag{2}$$

where $H$ is the set of all pairs $(w, b)$, $C$ is the cost of misclassification, $d$ is a default vector of imputations of missing values, and the vector $t_i$ is the original $x_i$ with missing values being replaced with corresponding default values in the default vector $d$ and is a function of the hyperplane $(w, b)$ and $d$. The optimization problem is