International Journal of Applied Mathematics

Volume 37 No. 2 2024, 165–174
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
doi: http://dx.doi.org/10.12732/ijam.v37i2.3

AN INTRODUCTION TO MACHINE LEARNING METHODS IN SAMPLE SURVEYS

Pankaj Das

Division of Sample Surveys

ICAR-Indian Agricultural Statistics Research Institute

New Delhi - 110012, INDIA

Abstract

Machine learning is revolutionizing sample surveys by improving data collection, analysis, and utilization. It combines advanced statistical techniques with computational algorithms to enhance survey sampling methods and data quality. Machine learning algorithms optimize survey sample design by identifying relevant variables, detecting patterns, and constructing efficient sampling strategies. They also assist in preprocessing and cleaning survey data, automatically detecting errors, imputing missing values, and handling outliers. Moreover, machine learning enables predictive modeling and estimation in sample surveys, leveraging large-scale data to generate models that predict outcomes, estimate population parameters, and uncover complex relationships among variables. Integrating machine learning into survey practices leads to more efficient and informative surveys, benefiting decision-making processes across various domains. Overall, machine learning has the potential to transform sample surveys, enabling more accurate predictions and estimations and improving the overall effectiveness of surveys. The application of machine learning in sample surveys and its potential future applications are described in the study.

MSC 2020: 68Q32, 97C30, 97U50, 62D05

Key Words and Phrases: sample surveys, statistics, machine learning, data quality, survey sampling, predictive modelling

Received: July 18, 2023

1. Introduction

In the era of data science, the machine learning (ML) models emerges as an alternative of traditional statistical models. The ML algorithms like Support vector machine (SVM), Artificial neural networks (ANNs), Random forests (RFs) became popular due their flexibility and free from stringent assumptions of the statistical models. ML methods are nonparametric methods can be used for making predictions or classifications from data. These methods are typically described by the algorithm that details how the predictions are made using the raw data and can allow for a larger number of predictors, referred to as high dimensional data. Some areas where the conventional statistical modelling methods fail due to nonlinearity of data, the ML techniques like ANN, SVM are capable to deal with linear and nonlinear pattern simultaneously. The ML methods are the most effective method used in the field of data analytics to produce reliable and valid decisions. Real world problems have high complexity which make them excellent candidates for application of ML. Researchers from multi-disciplinary fields like Computer science, Bioinformatics, time series uses machine learning models in their different ways.

2. Machine learning (ML) model

Machine learning is branch of applied science that improves the performance of a machine or model by providing the power to mimic like human brain for solving complex problems. Artificial intelligence (AI) models or machine learning models are the models that are developed based on machine learning (ML) algorithms. The ML algorithms use trial and error method to minimize error function. These AI models have better performance compared to the traditional models in case of nonlinear pattern due to the self-adaptive and data driven nature. The models are robust and have generalization power. The AI models first splits the whole dataset into two parts, i.e. training and testing sets. Training set usually contains 70-80 percentages of data. Training set is used in ML algorithm to build a model. After building the model, the generalization ability of the developed model is checked by testing data set. Some common AI based models are Artificial Neural Network (ANN), Support Vector Machine (SVM), Classification and regression tree (CART) and Bayesian networks.

Definition 2.1. The definition of ML is given by Samuel [1] as "Field of study that gives the computers the ability to learn without being explicitly programmed". The formal definition of machine learning as given by Mitchell [2] as "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E". ML learning algorithm has

been categorized broadly into three classes: supervised learning, unsupervised learning and reinforcement learning.

The ML algorithms are classified as three categories, i.e.:

2.1. **Supervised Learning.** In supervised learning, input variables are given along with already known output. In short, data are labeled. The prior information about input data pattern is known. The supervised problem can be classified into two categories i.e. regression problem and classification problem.

2.2. **Unsupervised Learning.** In unsupervised learning, input variables are provided with little or zero idea about the results of the inputs. Unsupervised learning technique derives a structure/pattern from the inputs without prior information about the relationship among input variables. There is no feedback based on prediction results. In literature, there are three types of unsupervised learning algorithms. They are clustering, association analysis and dimensionality reduction.

2.3. **Reinforcement Learning.** Reinforcement learning is also known as dynamic programming. It is advanced learning process. The main components of the learning are state, action, agent and interpreter. The agents take actions in an environment and interpreter interpret the results and provide some results of the actions. The specialty of the learning is the feedback of actions to the agents. The reinforcement learning is categorized into model based and model free reinforcement learning.

3. Process of Generic ML model building

The process of generic ML model building consists of six steps. The steps are:

I. Collection and Preparation of Data: The primary task of in the machine learning process is to collect and prepare data in a format that can be given as input to the algorithm. A large amount may be available for any problem. Web data is usually unstructured and contains a lot of noise, i.e., irrelevant data as well as redundant data. Hence the data needs to be cleaned and pre-processed to a structured format.

II. Feature Selection: The data obtained from the above step may contain numerous features, not all of which would be relevant to the learning process. These features need to be removed and a subset of the most important features needs to be obtained.

III. Choice of Algorithm: Not all machine learning algorithms are meant for all problems. Certain algorithms are more suited to a particular class problem as explained in the previous section. Selecting the best machine

learning algorithm for the problem at hand is imperative in getting the best possible results.

IV. Selection of Models and Parameters: Most of machine learning algorithms require some initial manual intervention for setting the most appropriate values of various parameters.

V. Training: After selecting the appropriate algorithm and suitable parameter values, the model needs to be trained using a part of the dataset as training data.

VI. Performance Evaluation: Before real-time implementation of the system, the model must be tested against unseen data to evaluate how much has been learnt using various performance parameters like accuracy, precision and recall.

4. **ML methods in complex survey**

Complex sample surveys involve the identification and collection of data in the form of a sample of population units via multiple stages or phases of identification and selection. In random sampling, a random sample selected from a list of all possible elements available in the sampling frame. Generally, complex sample surveys employ simple random sampling at each stage in a series of stages to culminate in the final sample of observation units at the desired level. Complex sample surveys may rely on stratification, clustering, multi-stage or multi-phase designs, unequal probability sampling, or multiple frame sampling. In complex survey, there is a huge scope of the application of ML techniques due to the data complexity and nonlinearity. When there are both auxiliary information and a variable of interest is available the machine can learn in a supervised way because its performance can be tested; when there is only auxiliary information available the machine has to learn unsupervised, i.e. without feedback. In supervised ML, when the variable of interest is qualitative/categorical, the machine learns to solve a classification problem; when the variable of interest is quantitative/numeric, the machine learns to solve a regression problem. In unsupervised ML, the machine learns to solve a clustering problem. ML methods also can be used for correction and imputation of the item nonresponse usually occurs in survey. Let there are two sample $\boldsymbol{A}$ and $\boldsymbol{B}$. $\boldsymbol{A}$ contains observations of a proxy variable $\boldsymbol{x}$, $\boldsymbol{B}$ contains observations of the variable of interest $\boldsymbol{y}$. Then ML methods can be used to model measurement error.

The areas of complex survey where ML methods can be applied be:

4.1. **Primary data.** There are many opportunities for ML in the processing of primary data.ML methods can be used for both prediction and classification. Some possible ML tasks that can be accomplished using primary data are summarized in next Table 1.

Tasks	Family of ML methods
Record linkage and Outlier detection	Clustering
Stratification	Classification
Estimation, imputation and calibration	Regression/classification

Table 1

5. **Secondary data**

Secondary data includes include big data, register data and the combination of these data with other sources, including possibly primary data. The datasets are computationally intensive and hard to parallelize. The datasets are often very noisy and poorly curated and thus contain numerous outliers and erroneous values. As a results many classical methods are unable to tackles these datasets. Modern ML approaches are better positioned than traditional statistical methods to handle the datasets. ML methods such as neural networks are an under explored alternative to model nonlinear relationships and complex interactions. Unsupervised, objects can be clustered in the high-dimensional space spanned by the rich set of features. Supervised, the clustering in high-dimensional space can be used to impute missing observations (de Waal *et al.* [3]) or to extrapolate relationships to unobserved subpopulations.

6. **Multisource statistics**

Multisource statistics are based on multiple data sources such as combinations of one or more surveys, administrative registers or big datasets. Several works like Christen [4], Harron et al. ([5], [6]) etc. have been done in data linkages. ML approaches are applicable in multisource statistics where data sources can be integrated. Three linkages can be possible i.e. micro integration, macro integration and no linkage. In micro-level linkage, individual units in multiple datasets can be associated with each other which often requires the presence of unique identifiers. This provides auxiliary variables that can be used in estimation and prediction models to predict variables only observed in the big data. Macro-level linkage refers to data linking situations in which units from multiple sources cannot be linked or matched individually, but where they can be associated at some aggregate level. Examples include people in the same municipality, or businesses in the same industry or size class. The data sources at aggregated levels as covariates can be used for applications of area-level models within the small area estimation framework (Rao and Molina, [7]). When no linkage is possible, data from multiple sources can be used for confrontation purposes. If multiple instances of the dataset are available through time, there are still possibilities to combine the data

sources through a time series approach. Van den Brakel et al. [8] improve the accuracy of survey based estimates through a structural time series modeling approach in which a big data time series is used as an independent covariate series.

7. **Recent works on ML methods in complex survey**

ML methods are emerging as a useful model for a variety of tasks in survey research literature. The most prominent application of supervised learning methods in survey research is their usage in the context of modeling and correcting for unit nonresponse. Specially in classification problems, where various algorithms have been used for constructing nonresponse weights. Decision trees are considered as an alternative to logistic regression and it can be used to derive weights in the presence of considerable interactions among the nonresponse predictors. ([9],[10] and [11]). Eck et al. [12] explored the use of recurrent neural networks (RNNs) to predict whether a respondent will break off from a Web survey, based on the respondents' behaviours exhibited in paradata describing their actions within the survey. The model was also used to predict the errors at the question level. Phipps and Toth [13] have considered CART for bias analysis in order to investigate whether the derived interactions are also associated with the outcome of interest. Lohr et al. [14] concluded that the tree-based methods outperform logistic regression when comparing true with predicted response propensities, given a non-additive functional form of the nonresponse model. ML methods are also applied for data integration (record linkages). Reiter [15] applied CART to generate partially synthetic, public use microdata. Nin and Torra ([16]) applied neural networks for record linkage which an increasingly relevant task for survey research in the era of Big Data. Their study was focused on cases where different records contain different variables. A neural network was used to find the relationships between variables. Then, these explored relationships was to translate the information in the domain of one file into the domain of the other file. Lu et al. [17] have used SVMs for imputation applied to student evaluations. Christen [4] used SVM for record linkage purposes. Caiola and Reiter [18] illustrated how random forests could be used to generate partially synthetic categorical data using data from the 2000 U.S. Current Population Survey. They showed the random forest synthesizer can preserve relationships reasonably well while providing low disclosure risk. Supervised learning methods have also been considered as imputation tools. In surveys, for complex missing patterns required flexible methods that can handle a large number of predictor variables computationally efficient. Nordbotten [19] performed neural network imputation to the Norwegian 1990 population census data. The imputed values were used to prepare estimates of proportions for a population and for smaller subgroups of the population. Mallinson and Gammerman [20] stated that SVM outperformed for variable imputation in UK census data and Danish Labour Force

Survey data. Tree-based imputation generates synthetic data with lower disclosure risks ([18]). Drechsler and Reiter [21] showed CART based imputation are efficiently balance analytical validity and disclosure risks. ML methods can used as prediction model in survey research. The variable of interest may be predicted with the auxiliary information. Like the yield of a crop may be predicted with the help of weather data, satellite data, field survey data etc. Jeong et al. [22] used Random Forests (RF) technique to predict crop yield responses to climate and biophysical variables at global and regional scales and found RF highly capable of predicting crop yields. Crane-Droesch [23] applied ML methods for crop yield prediction and climate impact assessment in agriculture. The fully-nonparametric neural networks were employed in predicting yields of corns. Elhag et al. [24] took NDVI as means for estimating yield of sugarcane in White Nile sugar factory with objectives to determine the ability of an in-season estimation of NDVI to predict sugarcane yield potential and optimum timing for predicting sugarcane in-season yield potential. Jui et al. [25] developed spatiotemporal hybrid random forest model for tea yield prediction using satellite-derived variables. Sehgal et al. [26] used the InfoCrop simulation model and remote sensing derived leaf area index (LAI) for mapping the wheat yield of small farms over a region. The simulation model-derived biometric relation was applied to the farms in a region using remote sensing derived LAI to estimate yield. Another field of survey research where the usage of supervised learning methods as an alternative to parametric modeling has been proposed is model-assisted estimation of population parameters ([27] and [28]). The idea of supplementing survey estimators with auxiliary information that is known for the population in order to improve efficiency requires relating the auxiliary variables to the outcome of interest. The functional form of such models might not be known in advance, especially since (administrative) auxiliary data often includes categorical variables with many categories, which may give rise to a number of interactions. McConville and Toth [29] proposed a regression trees in the model-assisted estimation framework. They demonstrated that such an approach can improve efficiency over both the linear regression and the Horvitz-Thompson estimator. Dagdoug et al. [30] proposed a new class of model-assisted procedures based on random forests based on partitions built at the population level as well as at the sample level. The study showed the proposed point and estimation procedures perform well in term of bias, efficiency and coverage in a wide variety of settings.

8. **Conclusion**

Machine learning has emerged as a powerful tool in the field of sample surveys, revolutionizing the way data is collected, analyzed, and utilized. By

combining advanced statistical techniques with computational algorithms, machine learning enables researchers to improve survey sampling methods, enhance data quality, and gain deeper insights from survey data ([1] and [20]). One key application of machine learning in sample surveys is in the design and optimization of survey samples. Traditional sampling approaches often rely on simple random sampling or stratified sampling methods, which may not fully capture the complex patterns and heterogeneity present in the population. Machine learning algorithms can effectively identify relevant variables, detect patterns, and construct more efficient sampling strategies, leading to more representative and cost-effective survey samples. Another area where machine learning excels is in survey data pre-processing and cleaning. Survey data is prone to errors, missing values, and inconsistencies. Machine learning algorithms can automatically detect and correct data anomalies, impute missing values, and identify and handle outliers, improving the accuracy and reliability of survey results ([20] and [2]). Furthermore, machine learning techniques enable predictive modeling and estimation in sample surveys. By leveraging large-scale survey data, machine learning algorithms can generate models that predict survey outcomes, estimate population parameters, and provide valuable insights into complex relationships among survey variables. In conclusion, machine learning has the potential to transform sample surveys by enhancing sampling methods, improving data quality, and enabling more accurate predictions and estimations. Integrating machine learning into survey practices can lead to more efficient and informative surveys, ultimately benefiting decision-making processes in various domains.

References

[1] A.L. Samuel, Some studies in machine learning using the game of checkers, *IBM. J. Re. Dev.*, **3**, No 3 (1959), 210-29.

[2] T. Mitchell, *Machine Learning*, McGraw Hill, p.2. ISBN 978-0-07-042807-2 (1997).

[3] T.de Waal, J. Pannekoek, S. Scholtus, *Handbook of Statistical Data Editing and Imputation*, Wiley, Hoboken (2011).

[4] P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Springer, Canberra (2012).

[5] K. Harron, H. Goldstein, C. Dibben, *Methodological Developments in Data Linkage*, Wiley, Chichester (2015).

[6] K. Harron, C. Dibben, J. Boyd, A. Hjern, M. Azimaee, M.L. Barreto, H. Goldstein, Challenges in administrative data linkage for research, *Big Data Soc.*, **4**, No 2 (2017).

[7] J.N.K. Rao, I. Molina, *Small Area Estimation*, Second Edition, Wiley, Hoboken (2015).

[8] J. Van den Brakel, E. Söhler, P. Daas, B. Buelens, Social media as a data source for official statistics; the Dutch Consumer Confidence Index, *Surv. Methodol.*, **43** (2017), 183-210.
[9] L. Rizzo, G. Kalton, M. Brick, Handling missing data in survey research, *Surv. Methodol.*, **22** (1996), 43-53.
[10] P. Lynn, *Quality Profile: British Household Panel Survey Waves 1 to 13: 1991–2003*, Institute for Social and Economic Research (2006).
[11] D. Judkins, H. Hao, B. Barrett, P. Adhikari, Modeling and polishing of nonresponse propensitys, In: *JSM Proceedings, Survey Research Methods Section*, American Statistical Association, Alexandria (2015), 3159-3166.
[12] A. Eck, L.K. Soh, A.L. McCutcheon, Modeling and polishing of nonresponse propensitys, In: *70th Annual Conference of the American Association for Public Opinion Research*, Hollywood, FL (2015).
[13] P. Phipps, D. Toth, Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data, *Ann. Appl. Stat.*, **6**, No 2 (2012), 772–794.
[14] S. Lohr, V. Hsu, J. Montaquila, Modeling and polishing of nonresponse propensitys, In: *JSM Proceedings, Survey Research Methods Section*, American Statistical Association, Alexandria (2015), 2071–2085.
[15] J.P. Reiter, Using CART to generate partially synthetic public use microdata, *J. Off. Stat.*, **21**, No 3 (2012), 441.
[16] J. Nin, V. Torra, New approach to the re-Identification problem using neural networks, In: *Modeling Decisions for Artificial Intelligence*, Springer, Berlin-Heidelberg, 2005.
[17] C. Lu, X. Li, H. Pan, Application of SVM and fuzzy set theory for classifying with incomplete survey data, In: *Proceedings of the IEEE International Conference on Service Systems and Service Management*, 2007, 1-4.
[18] G. Caiola, J.P. Reiter, Random forests for generating partially synthetic, categorical data, *Trans. Data Priv.*, **3**, No 1 (2010), 27-42.
[19] S. Nordbotten, Neural network imputation applied to the Norwegian 1990 population census data, *J. Off. Stat.*, **12**, No 4 (1996), 385-401.
[20] H. Mallinson, A. Gammerman, *Imputation Using Support Vector Machines*, Department of Computer Science, Royal Holloway, University of London, Egham, UK (2003), 52.
[21] J. Drechsler, J.P. Reiter, An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets, *Comput. Stat. Data. Anal.*, **52**, No 12 (2011), 3232–3243.
[22] J.H. Jeong, J.P. Resop, N.D. Mueller, D.H. Fleisher, K. Yun, E.E. Butler, S.H. Kim, Random forests for global and regional crop yield predictions, *PLoS One*, **11**, No 6 (2016), e0156571.

[23] A. Crane-Droesch, Machine learning methods for crop yield prediction and climate change impact assessment in agriculture, *Environ. Res. Lett.*, **13**, No 11 (2018), 114003.

[24] A. Elhag, A. Abdelhadi, Monitering and yield estimation of sugarcane using remote sensing and GIS, *Am. J. Eng. Res.*, **7**, No 1 (2018), 170-179.

[25] S.J.J. Jui, A.M. Ahmed, A. Bose, N. Raj, E. Sharma, J. Soar, M.W.I. Chowdhury, Spatiotemporal hybrid random forest model for tea yield prediction using satellite-derived variables, *Remote Sens.*, **14**, No 3 (2022), 805.

[26] V.K. Sehgal, D. Chakraborty, R. Dhakar, J. Mukherjee, R.N. Sahoo, Crop yield assessment of smallholder farms using remote sensing and simulation modelling, In: *Remote Sensing of Agriculture and Land Cover/Land Use Changes in South and Southeast Asian Countries*, Cham, Springer International Publishing (2022).

[27] F.J. Breidt, J.D. Opsomer, Model-assisted survey estimation with modern prediction techniques, *Stat. Sci.*, **32**, No 2 (2017), 190–205.

[28] K.S. McConville, F.J. Breidt, T.C. Lee, G.G. Moisen, Model-assisted survey regression estimation with the lasso, *J. Surv. Stat. Methodol.*, **5**, No 2 (2017), 131-158.

[29] K.S. McConville, D. Toth, Automated selection of post-strata using a model-assisted regression tree estimator, *Scand. J. Stat.*, **46**, No 2 (2017), 389-413.

[30] M. Dagdoug, C. Goga, D. Haziza, Model-assisted estimation through random forests in finite population sampling, *J. Am. Stat. Assoc.*, **00**, (2021), 1-18.