2018 International Conference on Computing, Power and Communication Technologies (GUCON)
Galgotias University, Greater Noida, UP, India. Sep 28-29, 2018

A Classifier Ensemble Machine Learning Approach to Improve Efficiency for Missing Value Imputation

Geeta Chhabra, Research Scholar
Amity School of Information Technology
Amity University
Noida, Uttar Pradesh, India
geeta_chhabra@rediffmail.com

Vasudha Vashisht, Assistant Professor
Department of Computer Science & Engineering,
Amity School of Engineering
Noida, Uttar Pradesh, India
vvashisht@amity.edu

Jayanthi Ranjan, Professor
Institute of Management Technology
Ghaziabad, Uttar Pradesh, India
jranjan@imt.edu

Abstract— **In data mining process the biggest task of data preprocessing is missing value imputation. Imputation is a statistical process of replacing missing data with substituted values. Many clinical diagnostic dataset are usually incomplete. Excluding incomplete dataset from the original dataset can bring more problem than simplification. In this paper the machine learning techniques for missing value imputation have been explored using Ionosphere data from UCI repository. The data imputation problem has been approached using well-know machine learning techniques. Several different statistical & data mining methods have been compared in this paper. The experiments have shown that the final classifier performance increases when Support Vector Machine (SVM) is used. Experiments show that popular machine learning classifier techniques were found to outperform than standard mean/mode imputation techniques. The advantage of both knowledge representation models are combined together in hybrid machine learning algorithms. The result shows that hybrid Logistic Regression (via Generalized Linear Model or GLM) or Random Forest outperforms the standalone Support Vector Machine (SVM). Machine learning algorithms from R/Revolution environments have been used.**

Index Terms— Missing Data, MCAR, MAR, MNAR, Imputation

I. INTRODUCTION

The occurrence of missing data points is one of the biggest challenges in data quality. Different sources of missing points such as patient's death, defective equipments, and denial of participants to answer certain questions, and patient data often have missing diagnostic tests that would be helpful for predicting the likelihood of diagnoses or for predicting effectiveness of treatment; consumer data often does not include values for all features useful for predicting buying options. Beside this, a significant amount of data can be erroneous. Data quality is of great concern in real-world problems using machine learning and other areas, such as Data Mining and Knowledge Discovery from Databases (KDD). Despite the repeated occurrence of missing data in real-world data sets, missing data is being handled by in machine learning algorithms in a naive way. A bias may be introduced in the knowledge induced, if missing values are not treated carefully. Data sets attributes are correlated to each other in most of the cases. Missing values can be estimated by identifying the relationships among different attributes. Imputation is a statistical process to replace the missing values by some substituted values in any data set. This approach allows the user to select the most efficient method as the missing data treatment is not dependent on the learning algorithm for each situation. The main idea of imputation is that if the value of dominant item is not available for a particular instance, it can be estimated from the data that is present[9].

The main area of interest in machine learning is biologically inspired models and the long term goal is to build algorithms and models that can not only process the information but also the biological systems. Many traditional areas of statistics are also included in machine learning; however, the centre of interest is on mathematical models. Machine Learning is not only the heart of many areas in computer science but is of great importance in the area of large-scale data processing. The primary goal of the research is to learn the possibility of combining classifiers which is usually better than any of its elements; in other words employ the strengths of strong classifier to complement the weakness of other classifier using hybrid ensemble methods known as stacking. Ensemble and hybrid methods in machine learning have attracted a lot of attention of the scientific community over the last few years [9]. The ensemble machine learning models have been seen to provide significantly improved performance than single weak learners, especially when dealing with complex high dimensional classification and regression problems.

Missing data mechanism can be viewed as [1, 12]:

- Missing completely at random (MCAR). In this type of randomness, missing data for an attribute does not dependent on observed data as well as on unobserved data. So, whichever missing value treatment is applied, no risk of introducing biasness in this type of randomness.
- Missing at random (MAR). In this type of randomness, an attribute with missing value is independent of the any unobserved data but is dependent on the observed data.
- Not missing at random (NMAR). In this type of randomness, an attribute with missing data are dependent on unobserved data. Several methods are there to deal with this type of randomness. The novel

way to treat this type of randomness is instance substitution with mean or mode

It has been assumed that the missing data mechanism is MAR in this paper, which means that the missing data can be predicted in some or other way from the observed values.

Depending on the type of missing data mechanism, researchers have more than one method to select from to deal with missing data.

- Ignoring instances with missing attribute: The simplest way is to ignore instances having at least one missing attribute.
- Most common attribute value: The most frequently occurred value is selected as the substitute value for all the missing attributes.
- Most common attribute value in class: The most frequently occurred value in a particular class is chosen as the value for all missing values in that class.
- Mean substitution: The attribute's observed cases mean value is used as a substitution for missing data.
- Regression or classification methods: A regression or classification model is build using the complete case data set, which treats the missing attribute as the target and the remaining attributes as predictors.
- Hot deck imputation: In this method, the missing values are replaced by the most similar cases.

II. LITERATURE REVIEW

Nahato et al (2016) in their article "Hybrid approach using fuzzy sets and extreme learning machine for classifying clinical datasets" proposed a classifier that combines the relative merits of extreme learning machine and fuzzy sets for clinical datasets. Cleveland heart disease (CHD), Pima Indian diabetes (PID) and Statlog heart disease (SHD) datasets from machine learning repository of University of California, Irvine (UCI) have been utilized for experiments. The CHD and SHD datasets have been experimented with two class labels one indicating the absence and the other indicating the presence of heart disease. The CHD dataset has also been experimented with five class labels, one class label indicating the absence of heart disease and the other four class labels indicating the severity of heart disease namely serious, high risk, medium risk and low risk. The PID dataset has been experimented with two class labels one indicating the absence and the other indicating the presence of gestational diabetes. The classifier has achieved an accuracy of 93.55% for CHD dataset with two class labels; 73.77% for CHD dataset with five class labels; 94.44% for SHD dataset and 92.54% for PID dataset.

Joanna et al (2016) in their article "Fusing Data Mining, Machine Learning and Traditional Statistics to Detect Biomarkers Associated with Depression" used multiple imputations which is a machine learning algorithm based on boosted regression. Logistic regression is used to classify biomarkers with depression attributes in the data set used. Using multiple chained regression sequences, 20 imputed data sets have been generated. In this method, initially, 21 biomarkers associated with depression has been identified. A final set of three classes of biomarkers were selected using traditional logistic regression methods. Fusion of data mining techniques based on machine learning algorithm by systematic use of hybrid methods for variable selection has been used for for detecting different classes of biomarkers associated with depression attribute.

Sridevi & Priya (2015) in their article "An Ensemble approach on Missing Value Handling in Hepatitis Disease Dataset" has used incomplete Hepatitis data set for machine learning technique for missing value imputations. Simple techniques like decision tree imputation, ID3 algorithm imputation, mean and mode imputation has been compared with the imputation based on proposed bootstrap aggregation for missing values. Experiment shows that classifier performance for missing value imputation using Bagging has been improved

Nannia, Lumini & Brahnam, "A classifier ensemble approach for the missing feature problem" proposed a method based on multiple imputation using random subspace. In this method, missing value are estimated using different data clusters. For clustering algorithm fuzzy clustering approach has been used. An experiment shows that the multiple imputation based on random subspace and clustering classifier performs better than several other approaches.

Liu, Pan, Dezert, Martin & Mercier (2015) in their article "Classification of Incomplete Patterns Based on the Fusion of Belief Functions" have presented a classification method which is based on the fusion of belief functions for incomplete pattern. The incomplete patterns are selectively estimated in this method. It is assumed in this method that the incomplete data is not important for classification. It is understood that the missing values play an important role in obtaining an accurate classification, if the object are not classified clearly. In such case, the missing values are imputed based on hybrid approach of self-organizing map and K-nearest neighbour.

III. MACHINE LEARNING TECHNIQUES FOR MISSING DATA IMPUTATION

A particular problem can be addressed using different algorithms which are based on its interaction with the environment or experience or the input data. In machine learning, the foremost step is to understand the different learning styles that an algorithm can have .An algorithm can have only few learning styles. The organization of machine learning algorithms is the most important as it decides about the function of input data which is used to train the model and use the one which is most suitable for the problem in order to get the most efficient result.

Machine learning algorithms can be categorized as :

A. Supervised Learning

In the supervised learning the goal is to build model based on the input or training data that makes predictions which has known labels or results such as defaulters/not-defaulters. A model is build using a training process which consists of training samples. The training process does not stop until a desired level of accuracy to correctly determine the class labels for unseen instances is achieved by the model. All classification and regression problems come under supervised learning. Learning algorithms are Random

Forest, Logistic Regression, the Back Propagation Neural Network, Decision Tree, KNN etc.

B. *Unsupervised Learning*

In this type of learning, predictions are drawn from input data which is not known or labeled. A model is build using identifying data pattern in the input data based on the general rules. This can be achieved through a mathematical process using the redundancy or to organize data by similarities. The dimensionality reduction, clustering and association learning rule come under unsupervised learning. Example algorithms are k-Means and Apriori algorithms.

C. *Semi-Supervised Learning*

In semi-supervised learning technique both known and unknown data are used. The model is build on the basis of input data which is mixture of both known and unknown data. This style is motivated by the fact that it is faster, better & cheaper. The classification and regression problems come under semi-supervised learning.

IV. ENSEMBLE ALGORITHMS

Ensemble machine learning algorithm uses weaker learning algorithms trained independently from multiple models and their results are combined in some way usually taking average of all the predictions to provide the overall enhanced prediction. This has proved its effectiveness and attracts much research in the area of ensemble learning. Boosting, Bagging, Stacking or Blending, AdaBoost, Gradient Boosted Regression Trees (GBRT), Gradient Boosting Machines (GBM), Random Forest[12] are the most well known ensemble machine learning algorithms.

The new direction in improvement of the efficiency of single machine learning classifier is by using the concept of combined classifiers. The ensemble classifier approach, such as boosting, bagging, and random forest are gaining more and more importance due to their successful implementation in the area of intrusion detection, intelligent transportation systems, data mining, bioinformatics, image and video processing, remote sensing and so on. The ensemble classifier approach combines the predicted results from multiple classifiers which has good performance, improved accuracy, stableness and robustness of traditional classifiers as compared to single classifiers.

The main techniques to build ensemble machine learning algorithms [9]:

Bagging. The training data is divided into different subsamples to build multiple models typically of the same type [3, 8,14].

Boosting. A chain of classifiers is produced here and the training set for each member in the chain is based on the performance of previously built models in the class.

Stacking. Multiple models of different types are built and model at the higher level of stack learns how to best integrate the predictions of the base models[8].

V. ALGORITHM SELECTION

The selection of particular machine learning algorithm is a crucial step. Cross validation is a model evaluation method in which the data is divided as test and training data. Once the training has been done, it can be used to test the new data. This is the basic approach for model evaluation. The evaluation of classifier is based on prediction accuracy. The prediction accuracy is calculated as the percentage of ratio between correct predictions and number of total predictions. The three most used approaches to evaluate classifier's accuracy are;

Holdout method is one of the simplest approaches of cross validation. In this method, the data is divided between training and testing. The function which is used to predict the output for testing data is estimated by using the training data only. To evaluate the model, the estimated error is used which is the average error rate of each subset.

K-fold cross validation is another technique, which is a step towards the improvement over the holdout method. In this technique, the training data is partitioned into k equally & mutually exclusive classes and the holdout method is repeated k times. One of the k classes is used as the test data, the remaining k-1 data sets are combined to form training data. Therefore, the mean error over all k trials is an estimate of the error rate of the classifier.

An exceptional case of cross validation is Leave-one-out cross validation in which each k test subset contains a single instance. This means that the function estimation is trained on all data except for one point. As before the mean error is computed. This is an expensive validation as compared to other, but helpful when the most accurate estimates of classifier's are desirable [15].

VI. SOFTWARE USED

Machine Learning packages in R/Revolution environment like kernlab, e1071 and MASS, caret, caretEnsemble have been used [4, 7].

VII. EXPERIMENTAL ANALYSIS

The paper utilizes the ionosphere dataset from UC Irvine Machine Learning Repository [5]. This is a classification problem which is binary in nature. Two types of electrons are targeted which is either good or bad in the ionosphere by the radar signals. The dataset contains 351 objects and 35 variables, the first 34 continuous variables are used for prediction and the last one is the class variable.

1. Support Vector Machine(SVM) with a linear kernel[2, 10, 16].

2. Classification and Regression Trees(CART)[8-10].

3. Logistic Regression (via Generalized Linear Model or GLM).

4. Linear Discriminant Analysis(LDA)[13].

5. K-Nearest Neighbors(kNN)[1,10,11,14,16].

The caret package in R has been used which provides in-built functions for machine learning algorithms, unique data visualizations to present multidimensional data, data resampling model tuning and optimal model selection[4].A 10-fold cross validation which will partition the dataset into 10 parts, 9 parts in training and 1 part in testing dataset to estimate the accuracy. The process has been repeated 3 times with distinct combination of splits of dataset into10 groups for each algorithm, in an order to get a more accurate

prediction. The metric used is "Accuracy" to select the most optimal models. Accuracy is measured as the percentage of ratio between number of objects predicted correctly and the total number of objects in dataset.

TABLE I. COMPARISON OF ACCURACY USING DIFFERENT CLASSIFIER

	Models	Accuracy	Kappa
1.	Support Vector Machine (SVM) with a linear kernel	0.9390819	0.8664861
2.	Classification and Regression Trees (CART)	0.8803097	0.7345538
3.	Logistic Regression (via Generalized Linear Model or GLM)	0.8802817	0.7315943
4.	Linear Discriminant Analysis (LDA)	0.864516	0.6809021
5.	k-Nearest Neighbors (kNN)	0.8450825	0.6327338

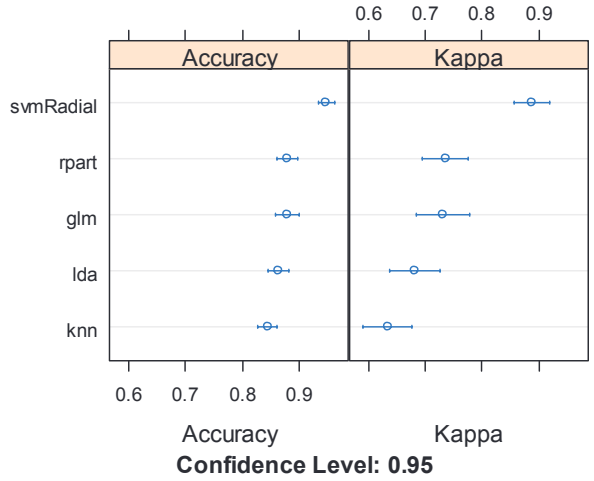

Fig. 1. Comparison of Stacking Ensemble Sub-Models in R

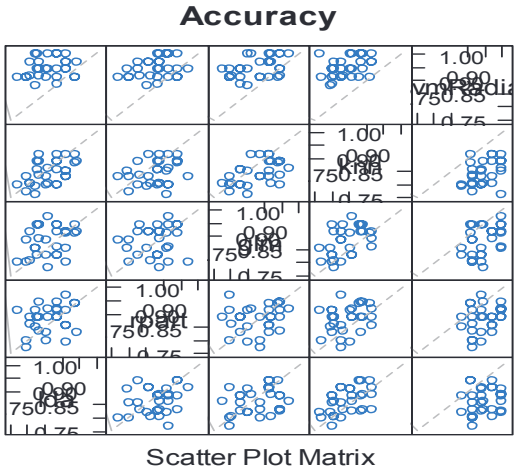

Fig. 2. Correlations between Predictions Made By Stacking Ensemble Sub-Models

The experiment shows that Support Vector Machine is the most accurate model with an accuracy level of 94.66%.The predictions of svmRadial, rpart, glm, lda and knn models using the stack of glm and random forest have been combined [13]. It is appropriate to have correlation between the predictions made by the sub-models should be low i.e. less than 0.75 which means that the models can be used to build a new classifier to get the best results from each model for enhanced efficiency. If the sub-models predictions has high correlation (i.e.>0.75) then the predictions would be the same or very similar most of the time, and there is no use of combining the predictions from different models. The experiment shows here that all pairs of predictions have correlation less than 0.75 which is considered low. The two sub-models with the highest correlation between their predictions are kNN and Logistic Regression (GLM) which is 0.517, less than 0.75 and thus not considered high.

VIII. CONCLUSION & FUTURE WORK

TABLE II. COMPARISON OF ACCURACY AFTER COMBING THE CLASSIFIER.

S.No.	Models	Accuracy	Kappa
1.	Generalized Linear Model or GLM)	0.9509305	0.893296
2.	Random Forest	0.9578938	0.9079525

After combining the classifiers predictions using a simple linear model or random forest, the accuracy has been lifted to 95.09% and 95.78% which is a slight improvement over using SVM alone over GLM and random forest respectively. The said technique can be applied to the real time binary classification problem as future work.

REFERENCES

[1] Aydilek Ibrahim Berkan, Arslan Ahmet, "A Novel Hybrid Approach To Estimating Missing Values In Databases Using K-Nearest Neighbors And Neural Networks", International Journal of Innovative Computing, Information and Control, 8, 7(A), July 2012,

[2] Choudhry Rohit, Garg Kumkum, "A Hybrid Machine Learning System for Stock Market Forecasting", International Journal of Computer, Electrical, Automation, Control and Information Engineering, 2(3), 2008.

[3] Joanna F. Dipnall, Pasco Julie A., Berk Michael, Williams Lana J., Dodd Seetal, Jacka Felice N., Meyer Denny, "Fusing Data Mining, Machine Learning and Traditional Statistics to Detect Biomarkers Associated with Depression", PLOS ONE, 5, February 2016 .

[4] Kuhn Max, Pfizer Global R&D, "Building Predictive Models in R Using the caret Package", Journal of Statistical Software, http://www.jstatsoft.org/, 2008.

[5] Lichman, M., "UCI Machine Learning Repository [http://archive.ics.uci.edu/ml] Irvine", CA: University of California, School of Information and Computer Science, 2013.

[6] Liu Zhun-ga, Pan Quan, Dezert Jean, Martin Arnaud, Mercier Gr´egoire, "Classification of Incomplete Patterns Based on the Fusion of Belief Functions", 18th International Conference on Information Fusion, Washington, DC, July 6-9, 2015.

[7] Mayer Zach, "A Brief Introduction to caretEnsemble", https://cran.r-project.org/web/packages/caretEnsemble/vignettes/caretEnsemble-intro.html, January 2016.

[8] Moacir P. Ponti Jr., "Combining Classifiers: from the creation of ensembles to the decision fusion", 24th SIBGRAPI Conference on Graphics, Patterns, and Images Tutorials, IEEE, 2011.

[9] Motaz M. H. Khorshid, Tarek H. M. Abou-El-Enien, Ghada M. A. Soliman, "Hybrid Classification Algorithms For Terrorism Prediction in Middle East and North Africa", International Journal of Emerging Trends & Technology in Computer Science, 4(3), May-June 2015.

[10] Nahato Kindie Biredagn, Khanna H.Nehemiah, Kannan A., "Hybrid approach using fuzzy sets and extreme learning machine for classifying clinical datasets", Informatics in Medicine Unlocked, 2016.

[11] Nannia Loris, Lumini Alessandra, Brahnam Sheryl, "A classifier ensemble approach for the missing feature problem", Artificial Intelligence in Medicine, Elsevier, November 2011.

[12] Panigrahi Lipismita, Das Kaberi, Mishra Debahuti, "Missing Value Imputation using Hybrid Higher Order Neural Classifier", Indian Journal of Science and Technology, 7(12), December 2014.

[13] Setz Cornelia, Schumm Johannes, Lorenz Claudia, Arnrich Bert, Troster Gerhard, "Using Ensemble Classifier Systems for Handling Missing Data in Emotion Recognition from Physiology: One Step Towards a Practical System", IEEE. 2009.

[14] Sridevi Radha krishnan, D. Shanmuga Priya, "An Ensemble approach on Missing Value Handling in Hepatitis Disease Dataset, International Journal of Computer Applications", November 2015.

[15] Vanitha A., Niraimathi S., "Study on Decision Tree Competent Data Classification", IJCSMC, 2(7), pp 365 – 370, 2013.

[16] Vladislav Miškovic, "Machine Learning of Hybrid Classification Models For Decision Support", SINTEZA, The use of the internet and development perspectives, 2014.