ENACTMENT RANKING OF SUPERVISED ALGORITHMS DEPENDENCE OF DATA SPLITTING ALGORITHMS: A CASE STUDY OF REAL DATASETS

Hina Tabassum and Dr. Muhammad Mutahir Iqbal

Department of Statistics, Bahauddin Zakariya University, Multan, Pakistan

ABSTRACT

We conducted comparative analysis of different supervised dimension reduction techniques by integrating a set of different data splitting algorithms and demonstrate the relative efficacy of learning algorithms dependence of sample complexity. The issue of sample complexity discussed in the dependence of data splitting algorithms. In line with the expectations, every supervised learning classifier demonstrated different capability for different data splitting algorithms and no way to calculate overall ranking of techniques was directly available. We specifically focused the classifier ranking dependence of data splitting algorithms and devised a model built on weighted average rank Weighted Mean Rank Risk Adjusted Model (WMRRAM) for consent ranking of learning classifier algorithms.

KEY WORDS

Supervised Learning Algorithms, Data Splitting Algorithms, Ranking, Weighted Mean rank risk-adjusted Model

1. INTRODUCTION

Building computational models with generalization capabilities and high predictions are one of the main needs of machine learning algorithms. Computational methods of supervised learning algorithms are trained to estimate the output of an unknown target variable/function. The noteworthy point is that the trained datasets should also be able to generalize the unseen datasets. Over-training comes in the category of poor generalization of trained model and if the model over train the correct output is not possible. Also Sometimes there exist situations when only one dataset is accessible and we are not accomplishing to gather new dataset set there we need some scheme to cope with the absence of data by splitting the available into training and test data but the splitting criteria may induce biasness in the comparison of the supervised learning classifiers. Various data splitting algorithms used to split the original datasets in to training and test datasets.

Which supervised classifier outperforms to the other is restricted to a given domain of the instances provided by the splitting algorithms. The appraisal of whether the selection of splitting algorithm influence the performance of classifiers we compare four standard data splitting methods using multiple datasets balance, an imbalance with two and maximum six classes from UCI repository. Fifteen supervised learning classifiers learned to hypothesize whether the performance of the classifiers affected by data and sample complexity or by wrong choice of learning classifier. Stability of the data-splitting algorithms measured in the rapport of error rate of individual supervised learning classifier.

DOI: 10.5121/ijcsit.2020.12202

2. DATA SPLITTING APPROACHES

In the case when only one dataset is available, numerous possible methods can come into consideration to make the required task of learning the machine algorithms. Splitting the data is widely used study design in high dimensional datasets and it is possible to split the available original datasets into training, testing and validation datasets [1].

2.1. Training Datasets

A subset of original datasets used for estimating and learning the parameter of the required machine learning algorithms.

2.2. Testing Datasets

A subset of original datasets used to estimate the performance of the required learning model.

3. STANDARD DATA SPLITTING ALGORITHMS

Several data splitting algorithms proposed in literature but it’s crucial to say that the complexity and the quality of these algorithms outperform to each other and statistically significant. Following data splitting algorithm are compared and used commonly:

3.1. Hold-Out-Method

Hold-out-method also called test sample estimation [2] is the simplest method in the class of all data splitting algorithms that divides the original datasets randomly into training and testing datasets. Mostly studied commonly used 25:75, 30:70, 90:10, 66:44 and 50:50 holdout sets[3] in training and testing datasets. The holdout method cause the increase in biasness in two data sets i.e. training and testing datasets because both may have different distributions. The main drawback of holdout method is that if the data is not large than this method is inefficient in its performance. For example in the classification problem it might be possible that subset consist of any missed class instance which cause the inefficient estimation and evaluation of model. For the cause of better results and to reduce the bias the method is iteratively used on datasets and the average of the resulted accuracy is calculated overall iterations. The above procedure is also called the repeated holdout method. Hold-out-method is a common method to avoid over training of data[4]

3.2. Leave One-Out Method

Leave one-out Method is described as the special case of the k-fold cross validation method where k=n. as n is the size of the original datasets and each train set has only one instance to learn [5]. This method does not involve any subsampling and produce unbiased estimates with large variation. The drawback of this method is that it is expensive and difficult to applicable in many real situations.

3.3. Cross Validation Method

Cross Validation Method is the most popular resampling technique. We call it as k-fold cross validation and sometimes rotation estimation method [2] where k is the parameter and the original dataset is divided into the disjoint fold of the equal sizes. In each turn only one k-fold is used for testing dataset and the remaining k-1 used as training datasets .the average of all accuracies is the resulting output of the model. The main drawback of this method is that it suffers in the pessimistic bias and by increasing the folds bias may be reduce as the resultant increase in variance. Mostly k is unfixed but commonly k is fixed at tenfold [3]that shows good results on different domain of datasets. This method is similar to the repeated holdout method where we use all the instances iteratively to learn and evaluation of the model.

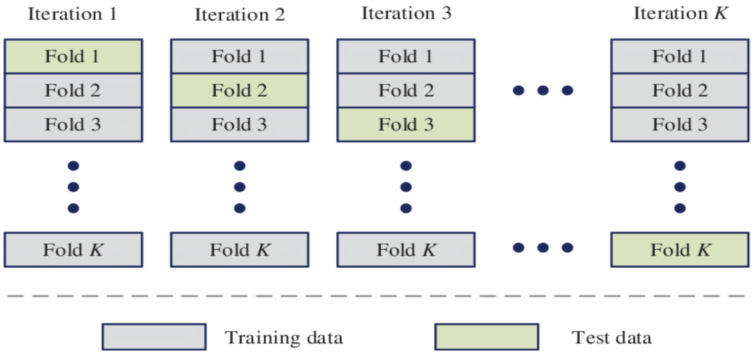

Figure 1: Strategy of Cross-validation

3.4. Bootstrap Method

Bootstrapping is a probabilistic statistical method and often used in situation where it is difficult to compute standard error by parametric methods Bootstrap Method generates bootstrap sample with replacement from the original datasets[6]. As in sampling with replacement each instance has an equal chance being selected more than once. Thus the overall error of the predicted model is given by averaging all bootstrap estimates. The most commonly used bootstrap approach which can also considered is 0.632bootstrap where 0.632 is the expected fraction of the instance that appeared in the 63.2%trainng set from the original dataset and the remaining 36.8% appears as testing instances. Symbolically the 0.632 bootstrap is defined as $Acc(T) = \frac{1}{B}\sum_{i=1}^{B} 0.632 * Acc(B_i)_{B'} + 0.638 * Acc(B_i)_T$ (2) where $Acc(B_i)_{B_I}$ is the accuracy of the model build with bootstrap training datasets and is the accuracy of the original datasets [1]. Bootstrap method proves best for small datasets and show high bias with high variability.

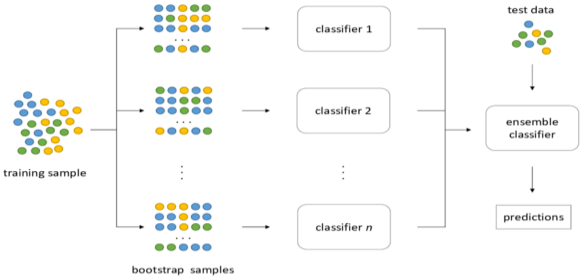

Figure 2: Bootstrap Strategy

4. Experimental Datasets

We used six benchmark real world datasets from UCI repository and chosen datasets are from multiple fields consist of balance, imbalance and multiclass datasets to check the efficacy of the data splitting algorithms in dependence of different domain of datasets. Detailed description of benchmark datasets with corresponding characteristics are detailed below in table.

Table 1: Experimental Benchmark Datasets

Data sets	**No. of instances**	**Balanced Imbalanced**	**Dimensions**	**Classes**	**Area**
Abalone	4177	Imbalanced	08	03	Life
Breast Tissue	106	Imbalanced	09	06	Life
Wine	6463	Imbalanced	13	02	Social
Iris	150	Balanced	04	03	Plant
Car	38	Imbalanced	07	05	Social
Diabetes	768	Imbalanced	08	02	Life

5. EVALUATION MEASURES FOR DATA SPLITTING ALGORITHMS

Evaluation measures used to assess the data splitting algorithms are the competency of the data splitting techniques to select instances to train the model. Efficacy of the data splitting algorithms is measured in differences between the error rate of instance classification to the target class of the original datasets and the test datasets. Moreover the performance of the splitting algorithms also measured in expressions of the user purposed and automatic selection of instances by the data splitting algorithms in account of learning time of models.

6. COMPARISON OF RESULTS

Boxplot is used to present the results of standard data splitting methods for fifteen supervised classification algorithms on multiple datasets separately. Dataset generated by data splitting algorithms used to train the supervised classification models on training set and performance of the supervised classifiers attained on the unseen dataset (testing dataset). The Individual sub-figure of boxplot corresponds to a data splitting algorithm performance of the supervised classifiers on benchmark datasets. The first dataset is the abalone multiclass imbalance dataset containing 4177 instances in three classes and split into training and testing dataset by using the four standard data splitting algorithms. Fifteen Supervised learning classifiers trained a model on training dataset and results attained from unseen (testing) dataset. Significant performance with the small variance observed by Cross-validation algorithm holdout, Leave-one-out and bootstrap method shows the high variance for all supervised learning classifiers. Wine dataset is the biggest dataset used for evaluation is multiclass imbalance dataset contains 6463 instances in three classes and split into training and testing dataset by using the four standard data splitting methods. Fifteen Supervised learning classifiers trained a model on training dataset and results attained from unseen (testing) dataset. On this dataset holdout, method rule good performance with the small variance than other three methods such as bootstrap, cross-validation and Leave-one-out method. The performance prediction of Leave-one-out- method shows the largest variance but shows stability to be optimistic. Iris dataset is the balanced three-class benchmark data set from the UCI repository with 150 instances. The Leave-one-out method significantly performs worst following the cross-validation method with high variance when supervised classifiers trained the unseen data. The Holdout method performs well with small variance among the other three data splitting algorithms. Performance of the bootstrap method is also acceptable but has had a large variance than holdout method. The Diabetes dataset is an imbalanced two-class dataset with 768 instances. As in the comparison of other imbalance, dataset bootstrap method performs better, when supervised classifiers trained the model on unseen datasets. In comparison to diabetes, boxplot shows an improvement for all data splitting algorithms holdout, cross- validation and Leave-one-out with small variance. Breast tissue and car datasets are relatively small multiclass imbalanced datasets following six and five classes as compared to other datasets with 106 and 38 instances. The purpose of including this benchmark dataset is to access the performance of data splitting algorithms on small data sets and we obtained incredible results with small variance and have the same patterned of error rate following all data splitting algorithms excluding holdout method has had large variance on car data set. No algorithm outperforms other algorithm on all benchmark datasets because if one data-splitting algorithm attainted better result with one supervised algorithm than in some cases it gives a poor result with other supervised algorithms. On multiclass and small datasets cross validation, bootstrap and Leave-one-out data algorithms shows good result while the performance of holdout algorithm is pessimistic bias because they use all dataset for learning the algorithm. On a balanced multiclass dataset, bootstrap has a good result but cross -validation and the Leave-one-out shows optimistic results. On very large binary class dataset, approximately all data-splitting algorithms perform better except Leave-one-out method. An obvious and noteworthy difference among performances of the supervised learning algorithms observed dependence of type of instances by user proposed and the data splitting algorithms. The type of instances used to build the model affects the performance results of classifiers significantly.

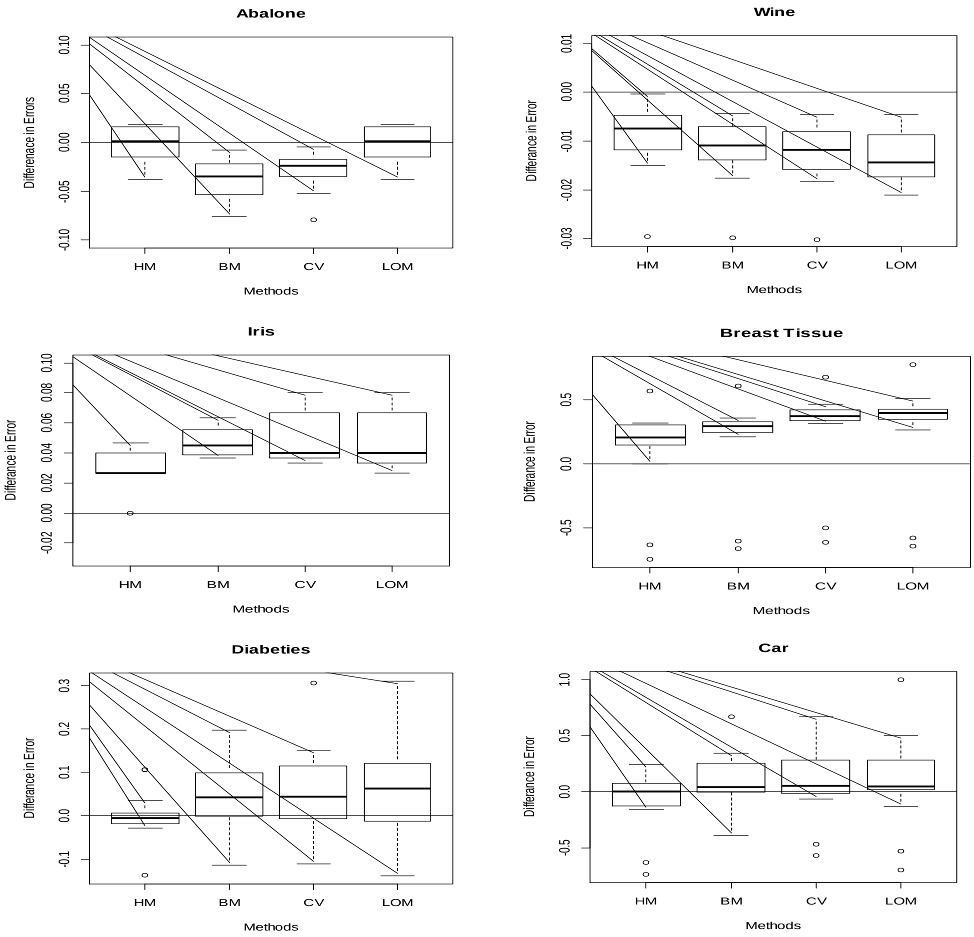

Figure 3: Difference in the Error rate of supervised Algorithms in dependence of User Proposed and Data Splitting Algorithms

7. WEIGHTED MEAN RANK RISK ADJUSTED MODEL (WMRRAM)

However, the overall result shows that the learning classifier performance for six different datasets dependence of data splitting algorithms is comparable and noteworthy variation exist in the rank of the classifiers. To overcome the variation in the rank data and come to the consolidate result WMRRAM model is used. The method that I will call the method of Weighted Mean Rank Risk Adjusted Model implicates first ranking the datasets in each column of two-way table by computing overall mean and standard deviation of the weighted rank datasets. The first step is to form the Meta table by ranking the supervised algorithm dependence of data splitting algorithms by given a lowest error rate a rank of 1 ,the next lowest error rate a rank of 2 and so on. Thus in each row of Meta table we have a set of values from 1 to 4, since there are 4 data splitting algorithms. Second step is stacking. Stacked generalization known as stacking in literature review is a scheme of combining the output of multiple classifiers in such a way that the output compares with the independent set of instances and the true class[7] in our case by data splitting algorithms. As stacking covers the concept of Meta learning [7] so at first N supervised classifiers S_i, $i = 1,2,....N$ learnt from data splitting algorithms for each multiple datasets D_i, $i = 1,2,...,N$. Output of the supervised classifiers S_i on the evaluation datasets ranked subsequently by the performance of standard data splitting algorithms. The outperform algorithm assigned rank 1; rank 2 is for runner-up and so on. We assigned average rank to overcome the situation where multiple data algorithms have had same performance. Let $w(i)$ denote the weights assigned iteratively to the ith data splitting algorithm where $0 \le w(i) \le 1$ and used them to form new instances I_j, $j = 1,2,....K$ of new dataset Z, which will then aid as a meta-level evaluation dataset. Each instance of the Z dataset will be of the form $S_i\,(I_j)$. Finally, we persuaded a global weighted mean rank risk adjusted model from the Z meta-dataset. The main advantage of the stacking is that learning algorithm with the best mean rank may be one who gets quite few poor ranks because of some other characteristics do not take account the variability in the ranks. For consensus ranking of the supervised learning algorithms dependence of data splitting algorithms we use Z meta-dataset. Risk is widely studied topic particularly from the decision making point of view and discussed in many dimensions [8]. Decision makers can assign arbitrary numbers for weights. The performed calculations were based on the weights of each characteristic and the weighted mean rank do not take account the variability in the ranks and there may be possibility that the supervised learning algorithm dependence of data splitting algorithms with the best mean rank may be one who gets quite few poor ranks because of some other data splitting algorithm. In order to grasp a consensus, result we used a WMRRAM approach. In WMRRAM model risk is taking as variability and uncertainty in ranking of different learning algorithms and statistical properties of the rank data is used to reveal which supervised learning algorithms is ranked highest and which is ranked second and so on dependence of data splitting algorithm. The overall mean rank obtained by using formula inspired by Friedman's M statistic [9] and standard deviation σ_z calculated by using the formula:

$$\mu_Z = \left[\sum_{J=1}^{6} I_J\right] \div 6 \qquad (1)$$

$$\sigma_Z = \sqrt{\frac{\sum_{j=1}^{6}(I_j - \mu_Z)^2}{J-1}} \qquad (2)$$

Where *j* denotes the multiple datasets, include in study for the evaluation of the performance of supervised classifier dependence of data splitting algorithm and *j= 1, 2…6.* The WMRRAM for the consensus ranking of multiple supervised classifiers are:

$$WMRRAM = \mu_Z \pm \theta_i \sigma_Z \qquad (3)$$

i. e. the increase or decrease will be in proportion to variations in the ranks obtained by different classifiers.

Following table shows the ranking behavior of the supervised algorithms with dependence of data splitting algorithms.

Table 2: Meta Table of Ranking of Supervised Classifiers dependence of Data Splitting Algorithms

Classifiers	WMRRAM	Rank
LDA	5.80676102	1
K-NN	6.91198256	5
ID3	12.0018463	15
MLP	6.63202482	4
NBC	9.44948222	11
BVM	6.02228655	3
CVM	6.00478598	2
C4.5	8.43272106	9
C-RT	9.89072464	12
CS-CRT	9.95330797	13
CS-MC4	7.94174278	8
C-SVC	7.19845443	6
PLS-DA	11.0797238	14
PLS-LDA	9.2770369	10
RFT	7.55490695	7

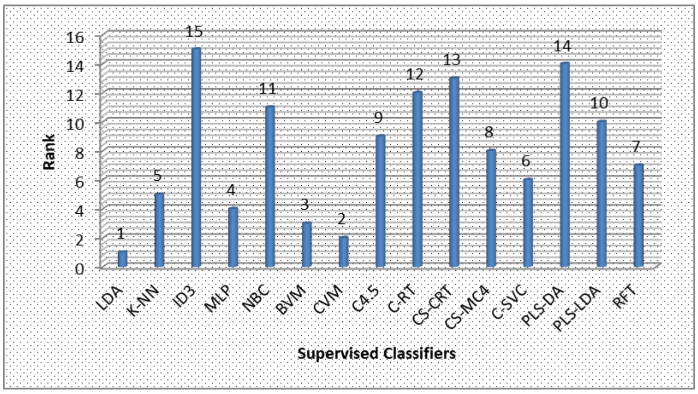

Figure 4: Graphicl representation of Ranking of Supervised Classifiers dependence of Data Splitting Algorithms

8. CONCLUSION

Evaluation of learning classifier performance and comparisons trendy nowadays and after studying the literature, a decision drained is that most articles just focus on some known learning algorithm performances with one or two data sets only without centering the quality and ratio of instances used to train or test the model. All learning algorithms include pros and cons but with measuring the performance of a specific algorithm this work show the impact of data splitting algorithms on the ranking of learning algorithms by using the proposed model of WMRRA. Results show that the performance of the learning classifiers varies with the data domain and these domains fixed in the framework of the number of instances and attributes used in the comparison of learning classifiers. Considering the WMRRA model, the classifier LDA met the highest-ranking score with a rank of 1, CVM, BVM, MLP followed a rank 2, 3, 4 and ID3 with a rank of 15 dependence data splitting algorithms In short, classifiers ranking is strongly robust to the dependence of sample complexity. Now, it is feasible because of the methodology used, all the learning classifiers obtained acceptable performance rates and had an adequate ranking in all related characteristics used. However, analyzing the result, mined from the software it was quite problematic to select a learning algorithm with the best performance. With reference to the above conclusion, the approach of the WMRRA model provides the best possible way of a ranking of the learning classifiers**.**

REFERENCES

[1] K. K. Dobbin and R. M. Simon, "Optimally splitting cases for training and testing high dimensional classifiers," Dobbin and Simon BMC Medical Genomics vol. 4, no. 31, pp. 1-8, 2011.

[2] R. Kohavi, "A Study of CrossValidation and Bootstrap for Accuracy Estimation and Model Selecti," presented at the International Joint Conference on Articial Intelligence IJCA, 1995.

[3] J. Awwalu and O. F. Nonyelum, "On Holdout and Cross Validation A Comparison between Neural Network and Support Vector Machine " International Journal of Trend in Research and Development, vol. 6, no. 2, pp. 235-239, 2019.

[4] Z. Reitermanov´a, "Data Splitting," 2010.

[5] Y. Xu and R. Goodacre, "On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning," Journal of Analysis and Testing, vol. 2, pp. 249–262 2018.

[6] CatherineChampagne, HeatherMcNairn, B. Daneshfar, and J. Shang, "A bootstrap method for assessing classification accuracy and confidence for agricultural land use mapping in Canada," International Journal of Applied Earth Observation and Geoinformation, vol. 29, pp. 44-52, 2014
.

[7] AgapitoLedezma, RicardoAler, AraceliSanchis, and DanielBorrajo, "GA-stacking: Evolutionarystacked generalization," Intelligent Data Analysis pp. 1-31, 2010, doi: 10.3233/IDA-2010-0410.

[8] A. Gosavi, "Analyzing Responses from Likert Surveys and Risk-adjusted Ranking: A Data Analytics Perspective," Procedia Computer Science, vol. 61, pp. 24-31, 2015, doi: doi.org/10.1016/j.procs.2015.09.139.

[9] S. M. Abdulrahman, P. Brazdil, J. N. v. Rijn, and J. Vanschoren, "Speeding up algorithm selection using average ranking and active testing by introducing runtime," Machine Learning volume, vol. 107, pp. 79-108, 2017, doi: doi.org/10.1007/s10994-017-5687-8