

M-Bagging: A New Modified Bagging Classification Model to Improve Prediction accuracy

Chandra Das, Abhishek Paul, Camellia Mukherjee, Debatosh Paul Majumdar, Shilpi Bose

Cite as: Das, C., Paul, A., Mukherjee, C., Majumdar, D. P., & Bose, S. (2024). M-Bagging: A New Modified Bagging Classification Model to Improve Prediction accuracy. International Journal of Microsystems and IoT, 2(1), 515-521. <https://doi.org/10.5281/zenodo.10703427>




© 2024 The Author(s). Published by Indian Society for VLSI Education, Ranchi, India




Published online: 22 January 2024



Submit your article to this journal: 




Article views: 



View related articles: 



View Crossmark data: 

DOI: <https://doi.org/10.5281/zenodo.10703427>

Full Terms & Conditions of access and use can be found at <https://ijmit.org/mission.php>



M-Bagging: A New Modified Bagging Classification Model to Improve Prediction accuracy

Chandra Das¹, Abhishek Paul¹, Camellia Mukherjee¹, Debatosh Paul Majumdar¹, Shilpi Bose^{1*}

¹Department of Computer Sc. & Engg., Netaji Subhash Engineering College, Kolkata-152, WestBengal, India.

ABSTRACT

Ensemble learning is a one kind of machine learning technique that improves the performance and robustness of the classification models and how the outputs of base classifiers are combined is one of the fundamental challenges in ensemble learning systems. Among different types of ensemble learning models, Bagging is most popular due to its simplicity but Bagging has several drawbacks. As for example in bootstrapped creation out of bag samples are not used properly or it does not take care for misclassified samples and it uses homogeneous classifiers. So, in this work, we have developed a modified bagging ensemble classification model by embedding modified bootstrapping techniques so that misclassified samples are specially taken care, out of bag samples are also taken care. Apart from these several heterogeneous classifiers are also used here in novel manner. From experimental results it has been found that the proposed model is superior compared to other existing basic ensemble models as well as other state of the art models.

KEYWORDS

Ensemble learning; Bagging; Bootstrapping; Classification; Class imbalance.

1. INTRODUCTION

In machine learning, ensemble learning [1] is one of the most important algorithms and is based on the supervised learning technique. Ensembles learning systems are inspired from the nature of decision making process of human beings. This is because humans tend to take decisions based on different factors. They also take opinions of other people to make decisions. Let us take an example to describe the process. Suppose when we buy a particular product from an e-commerce portal, we try to check for price of the same product from different sellers. We check the quality and specifications of the product. We also check the ratings of the product and also read the reviews of various customers who bought the product. In this way we decide whether to buy that product or not instead of buying it blindly. In Machine learning, the same thing is done using ensemble learning systems [1].

The word Ensemble means “union of parts” which is derived from Latin [2]. In ensemble learning, results of different classifiers are combined together using different approaches to give a new result which is better than the individual results of those classifiers. This concept of integrating classifiers provided new direction in improving the performance of regular classifiers. The regular classifiers when run independently often give poor performance when applied on large and high dimensional datasets. These classifiers also cannot handle class imbalance problem in an efficient way. To reduce these errors and improve their performance, ensemble learners are needed to construct by combining their predictions [1, 2]. Today ensemble learning systems are used in wide range of real world applications such as in biomedical, finance, politics, medicine etc.

There are four different category based ensemble classifiers- Bagging, Boosting, Stacking and Blending [1]. Among these four categories, bagging model, introduced by Breiman, is one of the most popular and successful ensemble classifier to improve accuracy of classification [2]. Bagging [2,3] is also known as Bootstrap Aggregation as it aggregates various versions of prediction accuracy of a weak learner when it is applied independently on various bootstrapped versions of the original training dataset. Every bootstrapped version of the training dataset is created by random sampling with replacement procedure. In this model first different bootstrapped versions are created and then a machine learning model is trained on these bootstrapped datasets independently and run in parallel. After training, the predictions of these models are combined using majority voting or weighted average method or using other approach to get the final prediction. However Bagging work with homogeneous classifier where only a single base classifier or weak learner is used.

Although Bagging is a popular classifier, it has several drawbacks. The first drawback is that as it generates different bootstrapped datasets by selecting samples randomly from the training dataset so some samples are not selected at all (out of bag samples) and so some samples are not used in the training phase. Secondly drawback is that it does not take any special care for the samples which are not properly classified in the training phase. The third drawback is that for every bootstrapped version it uses homogeneous classifiers such as either decision tree or SVM or KNN and so on. It results small variance but stable classifiers like KNN or SVMs generally does not generate in smaller classification error rates. So, there is no significance improvement in result via running computationally extensive classification methods on different bootstrapped

versions. Apart from this, it is very difficult to judge that which classification algorithm does have the best accuracy rates when applied to a particular training data.

Different researchers have already developed different modified bagging models [4-16] depending on various parameters such as diversity of base classifiers, stability of classifiers, class imbalancing problem, etc. to improve the limitations of bagging but still they have been working in this area to improve prediction accuracy of bagging model as improvement of prediction accuracy of a classification model with low computational overhead has major impact in different application areas such as in medical diagnosis, agriculture, stock prediction and so on. .

In this regard, we have proposed a modified bagging classification model named M-Bagging which modifies bagging model in a novel manner to give better performance in prediction system. The rest of the paper is described as follows:- In section 2, general bagging approach is discussed. Next section describes the proposed M-Bagging ensemble classification model in detail. In Section 4, the experimental results and discussions regarding the results are presented. Finally in last section some concluding remarks are presented.

2. TRADITIONAL BAGGING MODEL

It is known as bootstrap aggregating [1,2] and is derived from the concept of bootstrapping. It is one of the simplest and earliest ensemble based algorithm. In late Nineties, Leo Breiman proposed the concept of Bootstrap Aggregating and also coined the abbreviated term “Bagging”. Bagging was developed to improve classification by combining predictions of randomly generated training sets. It mainly reduces variance and produces a robust ensemble model than its base components. Here, several homogeneous weak learners with high variance are trained on different bootstraps which are obtained by resampling the training dataset with replacement and then after training, the predictions of these weak classifiers are combined through some ‘averaging’ process. Here the weak learners are trained in parallel and independently. To further understand the working process of bagging, we have to first understand the concept of Bootstrapping.

With the help of bootstrapping (random sampling with replacement), many random subsets are created from a training dataset. In machine learning, bootstrapping is known as resampling technique where many bootstrapped datasets are created by randomly selecting samples from the training dataset with replacement. Here sampling with replacement means that samples can be picked repeatedly more than once for each subset. In this way, n sub datasets are created from the training dataset (where n is the size of the original dataset) by picking various samples randomly from the dataset with replacement. Now every homogeneous weak learner is trained based on every bootstrapped dataset and then finally the predictions are combined to give the final output. There are several ways to do aggregation. If the problem is regression, the outputs of the individual base learners are averaged to obtain the final output of the final ensemble learner. If the

problem is classification problem, then the class which has the highest majority of votes will be considered as the final decision of the ensemble learner.

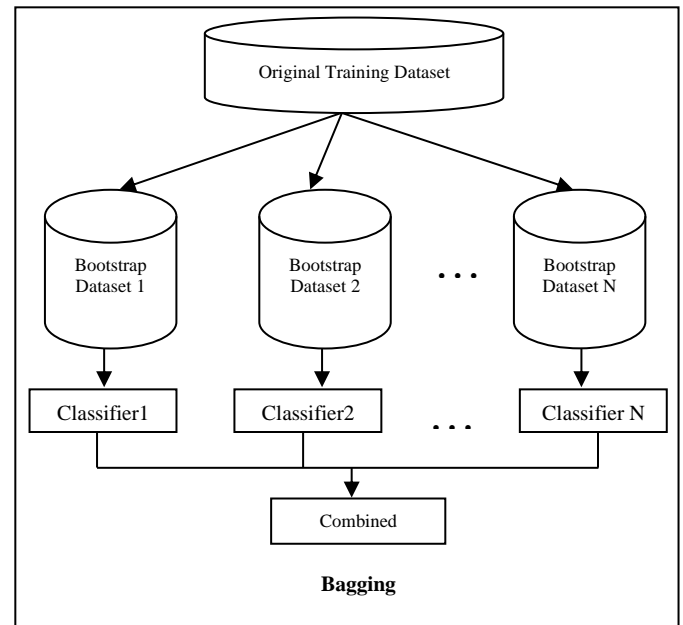


Fig. 1. General Bagging Model

3. PROPOSED WORK

In this paper we have proposed a modified version of bagging algorithm named M-Bagging. In this model different heterogeneous classifiers such as SVM, Naïve Bayes, KNN, DT, LR are used as base classifiers. The proposed technique contains four basic steps. The basic four steps are discussed below:

Preprocessing: In the preprocessing phase, relevant features are selected using Mutual information measure and then the reduced dataset is divided into training and testing dataset. In the training and testing dataset the ratio of samples of different classes is same.

Pre Classification Module: Here first different base classifiers are trained on training dataset using 10-fold cross validation and their classification accuracy are combined using majority voting. According to majority voting based classification result, the training dataset is divided into two sub datasets: one that contains correctly classified samples and other one that contains misclassified samples.

M-Bootstrapping : A new bootstrapping method named M-bootstrapping is used to construct different bootstrapped datasets. Here, in every bootstrapped dataset, the sub dataset that contains misclassified samples is directly placed and after that the rest part is filled up using resampling method taking training samples randomly from correctly classified instances. The correctly classified samples/instances which are not selected for creation of a bootstrapped dataset are used to form validation dataset.

Training phase: In this phase, different base classifiers are trained independently on each bootstrap sample dataset and after training, these models are tested independently on validation set made of out of bag samples. For every bootstrapped dataset, the classifier whose classification accuracy is highest is considered as the best classifier for that

bootstrapped dataset. In this way, for every bootstrapped dataset a best classifier is selected.

Testing phase: In this phase each sample from the test dataset is passed through every best classifier obtained from each bootstrapped dataset and the classification result of each best classifier is combined using majority voting technique and the final class is predicted for each test sample. The block diagram of the proposed model is shown in Figure 2.

Preliminaries

Let the dataset be represented by a data matrix $D(m * n)$ where m is the no of samples and n is the no of features/attributes in the dataset. The samples are represented by $E = \{E_1, E_2, E_3, \dots, E_m\}$ and the attributes are represented by $A = \{A_1, A_2, A_3, \dots, A_n\}$. Each sample is a n -dimensional attribute vector which contains n no of attributes and similarly each attribute contains m no of samples i.e. each feature is a m -dimensional sample vector. Here class vector is denoted by $C(m * 1)$ which represents the class label associated for every sample. Let us assume that we have u different class labels. So it can be represented as $C_i \in DC = \{1 \dots u\}$, where DC is the set of different class labels.

Proposed Algorithm for M-Bagging

Step 1: Input the dataset $D(m * n)$ and the no of base classifiers represented by TN . In this case $TN = 5$ and base classifier algorithms which used here are SVM, Naïve Bayes, KNN, Decision Trees (DT), Logistic Regression(LR).

Step 2: Reduce the no of attributes of the dataset $D(m * n)$ by performing relevant feature selection using mutual information. Reduced dataset represented by $D(m * l)$ where l is the new number of attributes.

Step 3: Divide the dataset $D(m * l)$ into training dataset $TD(m_{tr} * l)$ and test dataset $TSD(m_{ts} * l)$. $m_{tr} =$ no of samples in training dataset and $m_{ts} =$ no of samples in test dataset

Step 4: Train different base classifiers independently by applying on training dataset $TD(m_{tr} * l)$ using 10-fold cross validation and their results are combined using majority voting.

Step 5: Divide the training dataset into two parts: one containing correctly classified samples (TCD) and another part containing misclassified samples (MTD).

Step 6: Create D bootstrapped datasets by performing M-bootstrapping technique (random selection of samples from the (TCD) dataset with replacement for each bootstraps and by taking all samples from MTD dataset). Each bootstrap is represented by BD_i where $i = \{1 \dots D\}$. Each bootstrap contains same no of samples as the training dataset $TD(m_{tr} * l)$ as the samples are randomly sampled m_{tr} times with replacement from training dataset $TD(m_{tr} * l)$. Hence bootstrapped dataset represented by $BD_i(m_{tr} * l)$.

Step 7: Identify out of bag samples (samples not included in bootstraps) from each bootstraps and use these samples as validation dataset for every bootstrapped dataset. Each validation dataset is represented by VD_i where $i = \{1 \dots D\}$. Total no of validation datasets VD should be the same as total no of bootstrapped dataset $BD(m_{tr} * l)$. Here total no of validation datasets = D .

The proposed M-Bagging model is shown in Figure 2.

Step 8: Train different base classifiers on each bootstrapped versions $BD_i, i = (1, \dots, D)$ and test them on validation datasets VD_i obtained for each bootstraps BD_i .

Step 9: After testing, calculate the performance of these base classifiers and compare them and find the best classifier BBC_i from each bootstraps $BD_i, i = (1, \dots, D)$.

Step 10: Apply every test sample from test dataset $TSD(m_{ts} * l)$ through all the best base classifiers obtained from different bootstraps BD_i and find the classification accuracy through majority voting of all those classifiers.

Step 11: End

In this paper experimental studies are provided to evaluate the performance of the modified bagging version.

4. RESULTS AND DISCUSSIONS

To evaluate the performance of the proposed hybrid ensemble classifier, we have calculated the classification accuracy of the proposed modified version of bagging algorithm using different metrics. To prove the superiority of the proposed classifier we have compared it with other traditional ensemble classifiers like bagging, e-bagging, adaboost, and also with existing single classifiers.

4.1 Dataset Description

Here we have taken 4 datasets from the UCCI Machine Learning Repository [17] and Colon dataset [18] for the experimental purpose. These datasets except Colon are taken from UCCI Machine Learning Repository are Diabetes dataset which is 2 class dataset with 9 attributes and 768 instances, Liver Disorder Dataset which is 2 class dataset with 7 attributes and 345 instances, Ecoli dataset which is 4 class dataset with 8 attributes and 336 instances, and finally Dermatology dataset which is 6 class dataset with 34 attributes and 366 instances. Colon dataset is a gene expression dataset containing 2000 number of features/genes and 62 number of samples.

In table 1, the description of these datasets is given which contains the release year, the no of attributes and instances and also no of classes of the datasets.

Table 1: Dataset Description

Dataset Name	Year	Attributes	Instances	No of classes
	1990	9	768	2
Diabetes				
Liver Disorder	1990	7	345	2
Ecoli	1996	8	336	4
Dermatology	1998	34	366	6
Colon	1999	2000	62	2

4.2 Evaluation Metrics

The performance of the proposed modified version of bagging algorithm is assessed with respect to

- 1) 10-fold cross validation classification accuracy.
- 2) classification accuracy based on training and testing splitting.

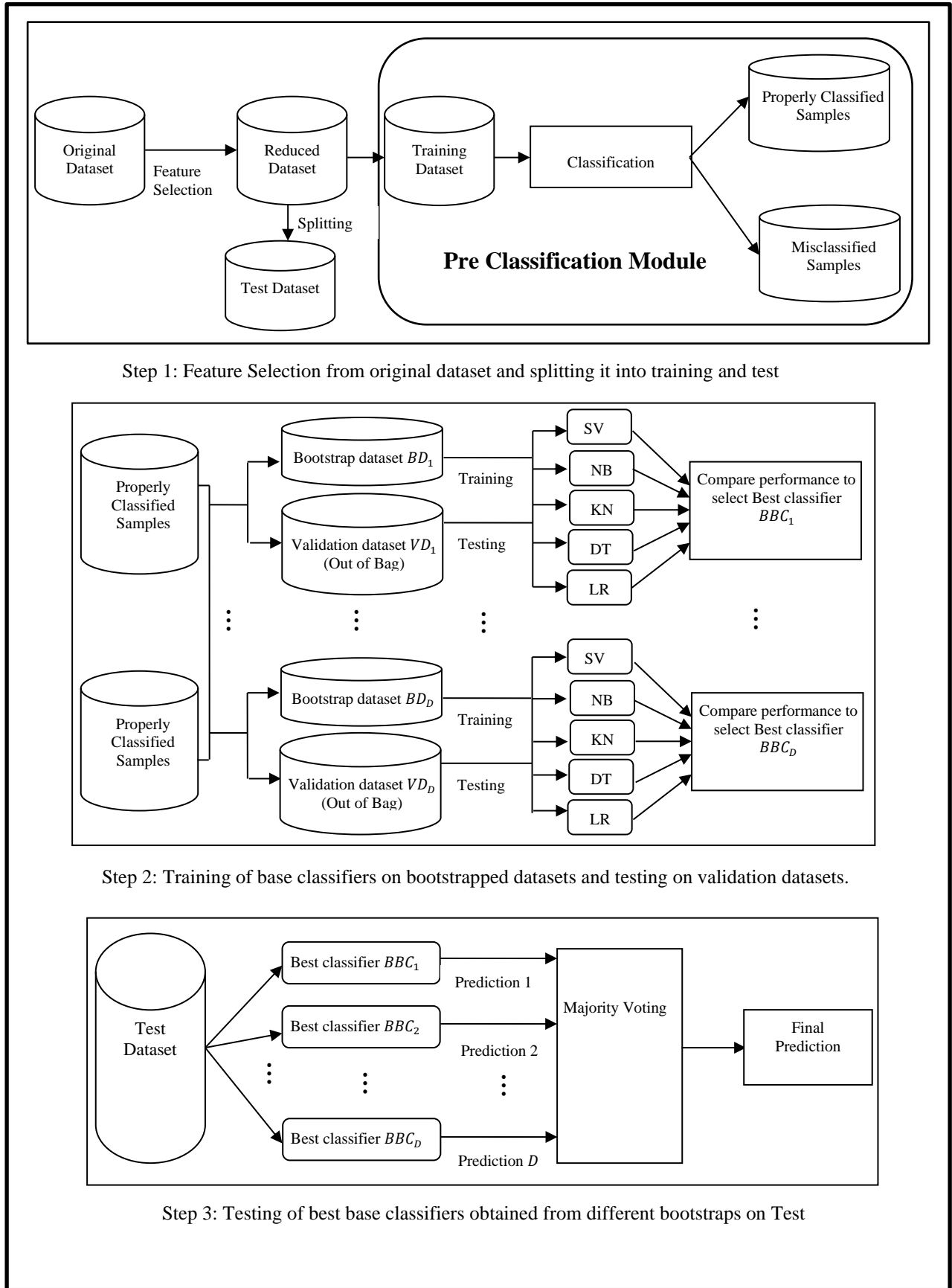


Fig. 2: Block diagram of M-Bagging model Dataset.

In the 10-fold cross validation, first the dataset is divided into 10 folds. Next 9 folds are used for training the classifiers and remaining 1 fold is used for testing. The process is repeated for 10 times and after training, the average classification accuracy is calculated by taking the average of the performance of the classifiers. Bootstrap percentage refers to what percentage of bootstraps, the 10-fold classification accuracy obtained is best. Here, 10-fold classification accuracy is taken as the evaluation metric.

In training testing splitting, the datasets are divided into training and testing datasets with splitting of 80-20, 70-30, 60-40 and 50-50 respectively. Testing accuracy is used as evaluation metric to check the performance of the proposed modified bagging algorithm with respect to the splitting ratios. The testing accuracy of the proposed algorithm is compared with the base classifiers SVM, Logistic Regression, NB, KNN and Decision Trees.

Table 2: 10-fold classification accuracy of the proposed M-Bagging model based on bootstrap percentage

Dataset Name	Bootstrap %	10-Fold Accuracy
	30	0.8521
Colon	50	0.8646
	70	0.85
Diabetes	30	0.752

Table 3: TP, TN, FP, FN, Sensitivity, Specificity for 2 class datasets by the proposed modified bagging algorithm with respect to best result of 10-fold cross validation accuracy

Dataset	True Positive	True Negative	False Positive	False Negative	Sensitivity	Specificity
Colon	12	10	6	8	60%	62.5%
Diabetes	37	39	8	12	75.5%	83%
Liver Disorder	6	15	10	3	66.7%	60%

Table 4: Comparison of Classification accuracy of M-Bagging Model based on training-testing splitting ratios with respect to existing base classifiers

Dataset	Splitting Ratio	SVM	LR	NB	KNN	DT	Proposed
	10-Fold	0.652	0.8523	0.569	0.782	0.657	0.8523
	80-20	0.69	0.92	0.771	0.921	0.769	0.924
Colon	70-30	0.59	0.73	0.63	0.83	0.736	0.925
	60-40	0.579	0.84	0.61	0.81	0.621	0.88
	50-50	0.774	0.81	0.549	0.741	0.742	0.874
	10-Fold	0.626	0.764	0.772	0.725	0.704	0.776
Diabetes	80-20	0.611	0.779	0.773	0.766	0.675	0.732
	70-30	0.654	0.761	0.783	0.653	0.705	0.76
	60-40	0.636	0.753	0.753	0.746	0.712	0.781
	50-50	0.638	0.752	0.755	0.701	0.684	0.77
	10-Fold	0.608	0.671	0.516	0.658	0.608	0.692
Liver Disorder	80-20	0.637	0.68	0.66	0.68	0.594	0.683
	70-30	0.596	0.673	0.528	0.644	0.587	0.683
	60-40	0.651	0.659	0.498	0.745	0.594	0.683
	50-50	0.612	0.601	0.6588	0.641	0.62	0.683
	10-Fold	0.771	0.764	0.788	0.768	0.759	0.821
Ecoli	80-20	0.767	0.59	0.779	0.761	0.731	0.801
	70-30	0.756	0.812	0.756	0.723	0.667	0.821

	50	0.783
	70	0.749
	30	0.811
Ecoli	50	0.879
	70	0.809
	30	0.965
Dermatology	50	0.972
	70	0.986

From the above table it is observed that the proposed modified bagging algorithm performs better when the bootstrap percentage is 50% in colon dataset. Similarly in diabetes dataset, the 10-fold accuracy of the proposed algorithm is better when the bootstrap percentage is 50%. The same is observed for other datasets. The bootstrap percentage means the percentage of original sample remains in the bootstrap data.

In the next table, the values for the no of true positives, true negatives, false positives, false negatives, sensitivity, specificity are shown with respect to the best 10-fold accuracy of the proposed modified bagging algorithm for datasets which are of 2 classes.

Dermatology	60-40	0.723	0.723	0.789	0.791	0.722	0.812
	50-50	0.767	0.756	0.785	0.809	0.744	0.809
	10-Fold	0.971	0.971	0.855	0.969	0.944	0.972
	80-20	0.972	0.961	0.854	0.961	0.927	0.961
	70-30	0.972	0.972	0.871	0.971	0.934	0.973
	60-40	0.986	0.986	0.903	0.972	0.965	0.986
	50-50	0.977	0.977	0.838	0.956	0.956	0.984

Table 5: Comparison of classification accuracy of the proposed M-Bagging model with existing bagging, boosting models on different datasets with respect to different splitting ratio

Dataset	Splitting Ratio	M-Bagging Model (Proposed)	Bagging (Existing) using C4.5 classifier	Boosting (Existing) using C4.5 classifier
Dermatology	80 – 20	97.2%	96.20%	66.73%
	70 – 30	97.2%	96.20%	69.43%
	60 – 40	98.6%	96.20%	65.92%
	50 – 50	97.7%	96.20%	67%
	80 – 20	68.38%	66.94%	68.38%
Liver Disorder	70 – 30	68.38%	66.94%	68.38%
	60 – 40	68.38%	66.94%	68.38%
	50 – 50	68.38%	66.94%	68.38%

From the above table, it is observed that on colon dataset, the proposed modified bagging algorithm performs better than the other algorithms when the training testing splitting ratio is 80-20, 70-30, 60-40 and 50-50. Similarly, in case of Diabetes dataset, the performance of proposed algorithm is better than other algorithms when the splitting ratio is 60-40 and 50-50. The same is observed for all other datasets.

In Table 5, the proposed algorithm is compared with standard bagging and boosting model using c4.5 classifier and in every case the proposed model performance is comparable with others.

Finally Table 6 shows the comparisons of classification accuracy obtained by the proposed modified bagging algorithm with other existing algorithms like e-bagging, adaboost, bagging and logistic regression models using 10-fold cross validation accuracy.

From the above table it is observed that on colon dataset and dermatology dataset, the proposed algorithm performs better than the existing algorithms.

Table 6: Classification accuracy comparison with existing e-bagging, adaboost, bagging and random forest algorithm on different datasets

Dataset	Proposed M-Bagging	E-bagging[19]	Adaboost	Bagging	Logistic Regression
Colon	0.8646	-	-	-	-
Diabetes	0.783	0.783	0.754	0.771	0.743
Liver Disorder	0.765	0.765	0.723	0.645	0.685
Ecoli	0.879	0.879	0.851	0.862	0.848
Dermatology	0.986	0.983	0.956	0.969	0.959

5. CONCLUSION

In this paper, the main purpose is to design a modified version of bagging algorithm to optimize the standard bagging algorithm. The main differences between the standard bagging algorithm and the proposed modified bagging algorithm are that in standard bagging algorithm, homogeneous classifiers are used for training but in our proposed algorithm we have tried to use heterogeneous classifiers for training. In this algorithm, we have also considered out of bag samples as validation datasets for testing of base classifiers and used five classifiers for training purpose which are Support Vector Machine, Logistic Regression, KNN, Naïve Bayes and Decision tree. Apart from these, we have selected relevant features using mutual information and generated bootstrapped datasets in a novel manner. From the experimental results, we have observed that our proposed modified bagging algorithm performs better than the standard bagging algorithm, adaboost algorithm and while comparing their results while it leads to close results when compare with e-bagging algorithm. Thus we can conclude that the proposed modified bagging algorithm show promising applicability in prediction of datasets.

REFERENCES

- Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, Qianli Ma. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, 14, 241–258. DOI: [10.1007/s11704-019-8208-z](https://doi.org/10.1007/s11704-019-8208-z)
- Leo Breiman. (1996). Bagging Predictors. *Machine Learning*, 24, 123-140. DOI: <https://doi.org/10.1007/BF00058655>
- Quinlan J. (1996). Bagging, Boosting, and C4.5. *Proceedings of the 13th National Conference on Artificial Intelligence*, Portland, 725-730. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.49.2.457&rep=rep1&type=pdf>
- Błaszczczyński J., Stefanowski J. and Idkowiak Ł. (2013). Extending Bagging for Imbalanced Data. *Proceedings of the 8th International Conference on Computer Recognition Systems*, Heidelberg, 269-278. DOI: https://doi.org/10.1007/978-3-319-00969-8_26
- Błaszczczyński J. and Stefanowski J. (2017). Actively Balanced Bagging for Imbalanced Data. *Proceedings of International Symposium on Methodologies for Intelligent Systems*, Cham, 271-281. DOI: http://dx.doi.org/10.1007/978-3-319-60438-1_27

6. Bifet A., Holmes G., and Pfahringer B. (2010). Leveraging Bagging for Evolving Data Streams. Proceedings of European Conference on Machine Learning and Knowledge Discovery in Databases, Berlin, 135-150. DOI: https://doi.org/10.1007/978-3-642-15880-3_15
7. Bryll R., Gutierrez-Osuna R. and Quek F. (2003). Attribute Bagging: Improving Accuracy of Classifier Ensembles by Using Random Feature Subsets. Pattern recognition, 36(6), 1291-1302. [https://doi.org/10.1016/S0031-3203\(02\)00121-8](https://doi.org/10.1016/S0031-3203(02)00121-8)
8. Chung D. and Kim H. (2015). Accurate Ensemble Pruning with PL-Bagging. Computational Statistics and Data Analysis, 83, 1-13. <https://doi.org/10.1016/j.csda.2014.09.003>
9. Datta S., Pihur V. and Datta S. (2010). An Adaptive Optimal Ensemble Classifier Via Bagging and Rank Aggregation with Applications to High Dimensional Data. BMC bioinformatics, 11(1), 1-11. DOI: <https://doi.org/10.1186/1471-2105-11-427>
10. Croux C., Joossens K., and Lemmens A. (2007). Trimmed bagging. Computational Statistics and Data Analysis, 52(1), 362-368. <https://doi.org/10.1016/j.csda.2007.06.012>
11. Dexun J., Peijun M., Xiaohong S. and Tiantian W. (2014). Distance Metric Based Divergent Change Bad Smell Detection and Refactoring Scheme Analysis. International Journal of Innovative Computing, Information and Control, 10(1), 1519-1531. <http://www.ijicic.org/ijicic-13-07024.pdf>
12. Mohamed H., Negm A., Zahran M., and Saavedra O. (2018). Assessment of Ensemble Classifiers Using the Bagging Technique for Improved Land Cover Classification of multispectral Satellite Images. The International Arab Journal of Information Technology, 15(2), 270-277. <https://dblp.org/db/journals/iajit/iajit15.html#MohamedNZS18>
13. Xie Z., Xu Y., Hu Q., and Zhu P. (2012). Margin Distribution Based Bagging Pruning. Neurocomputing, 85, 11-19. <https://doi.org/10.1016/j.neucom.2011.12.030>
14. Xiaoyuan S., Taghi M., and Xingquan Z. (2008). VoB Predictors: Voting on Bagging Classifications. Proceedings of 19th International Conference on Pattern Recognition, Tampa, 1-4. <https://doi.org/10.1109/ICPR.2008.4761803>
15. Zeng X., Chao S., and Wong D. (2010). Optimization of Bagging Classifiers Based on SBCB Algorithm. Proceedings of International Conference on Machine Learning and Cybernetics, Qingdao, 262-267. <https://doi.org/10.1109/ICMLC.2010.5581054>
16. Lichman M. UCI Machine Learning Repository. [Online]. <http://archive.ics.uci.edu/ml/index.php>
17. Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ Alon U. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America, 1999, 6745-6750. <https://doi.org/10.1073/pnas.96.12.6745>
18. Goksu Tuysuzoglu and Derya Birant. (2020). Enhanced Bagging (eBagging): A Novel Approach for Ensemble Learning. The International Arab Journal of Information Technology, 17(4), 515-528. <http://dx.doi.org/10.34028/iajit/17/4/10>

AUTHORS



Chandra Das received the M.Sc degree in Computer and Information Science from University of Calcutta, Kolkata, India in 2001 and the M.Tech degree in Computer Science and Engg. from the same University in 2003. She received her PhD degree in engineering from Jadavpur University, Kolkata, India in 2011. She is currently an associate professor in the department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, India. Her research interest includes machine learning, bioinformatics, pattern recognition, data mining and natural language processing. She has published over 45 research papers in several international journals and conference proceedings.



Abhishek Paul is a final year B.Tech student of the department of Computer Science and Engineering at Netaji Subhash Engineering College in the year 2023.



Camellia Mukherjee is a final year M.Tech student of the department of Computer Science and Engineering at Netaji Subhash Engineering College in the year 2022.



Debatosh Paul Majumdar is a final year B.Tech student of the department of Computer Science and Engineering at Netaji Subhash Engineering College in the year 2024.



Shilpi Bose received the M.Sc degree in Computer and Information Science from University of Calcutta, Kolkata, India in 2002 and the M.Tech degree in Computer Science and Engg. from the same University in 2004. He received his PhD degree in engineering from Jadavpur University, Kolkata, India in 2023. He is currently an assistant professor in the department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, India. His research interest includes machine learning, bioinformatics, pattern recognition, and data mining. He has published over 30 research papers in several international journals and conference proceedings.