International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

The Impact of Enhanced Space Forests with Homogeneous Classifier Ensembles

Zeynep Hilal Kilimci*[1], Sevinç İlhan Omurca[2]

Accepted : 03/04/2018 Published: 29/06/2018

Abstract: In this paper, we propose to advance the classification success of classifier ensembles by investigating the contribution of enhanced space forests. For this purpose, this study especially is focused on enhanced feature spaces by implementing the most popular feature selection techniques, namely information gain, and chi-square. After performing these methods on the original feature space, training phase is evaluated with all the original and the modified versions of most significant features, which are acquired by applying difference operator to the original features and the selected features with feature selection methods. That is, the new training dataset is constructed by combining the original features and the new ones. Then, the training is done with the well-known classification algorithm namely, decision tree using the enhanced feature space. Finally, three types of ensemble algorithms namely, bagging, random subspace, and random forest are carried out. A wide range of comparative experiments are conducted on publicly available and widely-used 36 datasets from the UCI machine learning repository to observe the impact of the enhanced space forests with classifier ensembles. Experiment results demonstrate that the proposed enhanced space forests perform better classification accuracy than the state of the art studies. Approximately, 1% - 3% improvement of the classification success is an indicator that our proposed technique is efficient.

Keywords: *Classifier ensembles, enhanced space forests, ensemble algorithms.*

1. Introduction

In recent years, ensemble learning is a very popular research area in the literature. Ensemble learning has also known as committees of learners, mixture of experts, ensemble of classifiers, and ensemble algorithms [1-4]. The idea behind of this approach is to use more than one classifier. Thus, it is expected to generate more accurate and robust models by classifying [5-9]. Generally, classification task is implemented by supervised machine learning techniques. Some of the most popular methods are naïve Bayes classifiers (NB), decision trees (DT), support vector machines (SVM), artificial neural networks (ANN), and k-nearest neighborhood classifiers (k-NN). Among these classifiers, decision trees have been extensively applied in the state of the art studies for ensemble learning [1-2, 10-14]. Furthermore, usage of more than one decision tree emerges decision forests for ensemble learning.

In addition to the selection of the classifier, individual success and diversity of base learners are significative parameters for the ensemble performance. As long as the diversity of base learners increases, the overall classification accuracy of the ensemble system will be better. Thus, it is essential to generate diverse base learners by making use of different or same base learners. Diversity can be provided by usage of different training datasets if the base learners are the same. In order to get different training datasets, there are several traditional ensemble algorithms such as bagging, boosting, random subspaces, random forests, and rotation forests. Ensemble algorithms used in this paper are briefly mentioned.

Bagging [1, 12, 15-19]: Bagging produces new training dataset (bootstrap samples) from the original dataset by using replacement. In other words, the multiple versions are composed thereby performing bootstrap replicates of the training set and employing them as new training sets. Then, the classifier is constituted on each of these samples and associated them with majority voting.

Random Subspace [1, 16-18, 20-22]: The idea behind of this approach is quite simple. Random subspace method aims to train with a subset of the original feature space instead of using extended version. That is, features are chosen randomly from the feature set instead of utilizing all features for each base learner in the ensemble. Then, the classifier is constructed on different feature subsets illustrated randomly from the original feature set and aggregated by applying the majority voting.

Random Forest [1, 10, 14, 23-28]: Random Forest unifies Bagging and Random Subspace approaches. Based on the creation of different training dataset, it is proposed by several studies [1-2, 10] that the feature space can be extended thereby using various combinations of the features, generating new features and adding them to the original feature space.

In this paper, we propose to investigate the impact of enhanced space forests with classifier ensembles by using popular feature selection techniques. In this way, we aim to boost the classification performance of the ensemble system. Enhancement of the feature space is ensured by the original features and the significant features which are selected by features selection methods. Feature space enhancement with the specific feature selection techniques constitutes the center of this study because of improvement of the classification success. After getting enhanced feature space, decision tree construction is carried out according to the ensemble algorithms namely, bagging, random subspace, and random forest. For observing the contribution of our proposed technique, a wide range of comparative experiments are conducted on publicly available and widely-used datasets from the UCI machine learning

[1*] *Computer Engineering, Dogus University, Istanbul – 34722, TURKEY*
[2] *Computer Engineering, Kocaeli University, Kocaeli – 41380, TURKEY*
** Corresponding Author: Email: hkilimci@dogus.edu.tr*

repository [29]. Experimental results demonstrate that the proposed technique performs better classification accuracy than the state-of-the art studies.

We also seek for answers based upon empirical evidence to the following questions:

- Does the enhanced space forest with proposed technique present optimal solution for classification problems? Does it provide any contribution to the classification success?
- Among various feature space enhancement techniques, which one can be chosen as the winner across all settings and datasets? Are there any guidelines to help choose the best from these methods?
- Can the success of an improved enhanced space technique take an advantage of the ensemble methods? To what extent can each of three ensemble techniques contribute the classification performance?

The rest of the paper is organized as follows: Section 2 gives related researches on the use of ensemble systems and extended spaces. In section 3, we give our proposed framework employed in the experiments. Some in-depth discussion is put forward and experiment results are drawn about the impact of usage of enhanced space forests in sections 4, 5, and 6.

2. Related Work

Ensemble learning is the collection of methods that builds a set of classifiers and combines their classification predictions by using majority voting [1-5, 30]. Previous studies [6-9, 16, 18, 30-31] have indicated that the ensemble system is more accurate and robust than any of the single classifiers in the ensemble.

In a recent study [18] empirically measures the predictive performance of the ensemble learning techniques on text documents that are demonstrated keywords. They first perform different keyword extraction algorithms namely, most frequent measure based keyword extraction, term frequency-inverse sentence frequency based keyword extraction, co-occurrence statistical information based keyword extraction, eccentricity-based keyword extraction, and text rank algorithm to test dataset. Then, they make use of various learning algorithms (naïve Bayes, support vector machines, logistic regression, and random forest) with five widely-used ensemble techniques such as adaboost, bagging, dagging, random subspace, and majority voting. They conclude their research that keyword based representation of text documents with ensemble learning can raise the predictive performance.

The other study [1] proposes the extended feature space by choosing new features randomly and adding them to original feature space. They apply several feature generating operators to produce new features such as sum, difference, divide, and multiply. In order to choose the best operator, they measure averaged individual accuracy of the base learners, average accuracy ranks, and average kappa of the base learners for all operators. They report that the difference operator is the best when all three metrics were related. They decide to add d number of new features to the original ones. Thus, extended space is set to org (d) + new (d) in their extensive experiment results. The number of base learners is set to 100 and 5*2 cross validations are done for each dataset and ensemble algorithm. In their experiment results on 36 UCI dataset, extended space versions and original versions of four ensemble algorithms are compared in terms of classification accuracies of ensemble. They observe that all the extended versions outperform the original versions for all the ensemble algorithms.

The recent studies [2, 10] on the extended space decision trees propose to increase the ensemble accuracy. Instead of randomly producing, new features with high classification capacity are generated by computing the gain ratio of each different candidate features. Thus, they combine newly generated features and the existing features to extend feature space. Then, a decision forest is built from the extended space dataset. They carry out experiments on accessible datasets that are publicly available from the UCI Machine Learning Repository and 10-fold cross validation is applied for every dataset. They also measure impact of the different space extension parameters. The best d and d/2 features are selected from the set of candidate feature where d is the number of features. They observe that the extension of d/2 features is more appropriate than d features. Their experiments show that the proposed method outperforms both the performance of original feature space and randomly generated extended space version. Authors conclude that the extended space forest is an effective method to increase prediction accuracy but it can be improved by using significant features instead of selecting randomly.

Another recent study [16] investigates the effectiveness of enhanced random subspace method based on part-of-speech technique, POS-RS, for sentiment categorization field. Instead of using a single subspace rate to compose the diversity of base learners for ensemble learning, authors utilize two important parameters namely, content lexicon subspace and function lexicon subspace rate by means of POS-RS technique. Experiments are conducted on ten publicly available sentiment datasets to represent the effectiveness of their technique. They conclude that POS-RS is preferable method to excel the success of classification and applied to the other text classification problems.

3. Proposed Framework

Enhanced Space Forest is an effective method in order to increase the classification accuracy. Instead of using the original features as an input vectors, various combinations of them are generated and combined for the enhanced space forest approach. The main idea for composing enhanced feature space is to extend the original feature space. So far the studies on the enhanced feature space utilize either randomly chosen features [1] or selected features with the specific feature selection method such as gain ratio [2, 10] to determine new candidate features. As noted in the previous works [1-2, 10] enhancement of the feature space provides significant contribution to the classification performance. Thus, we are inspired by both the impressive work [1] and the study in [2] to boost the classification accuracy of the ensemble system.

In this work, our research objective is based on the improvement of classification accuracy by using the proposed feature enhancement techniques for constructing enhanced feature space. For this purpose, we firstly concentrate on two types of feature selection techniques, namely information gain (IG) and chi-square (CHI) are applied to the original feature space in order to get the significant features. In this way, we also intend to demonstrate the success of these methods for the enhanced space forests.

Information gain evaluates the number of bits of information obtained for class prediction by knowing the occurrence or nonoccurrence of a feature while chi-square interprets the lack of independence between feature and class and can be checked the distribution of chi-square with one degree of freedom to judge extremeness [21, 32-35]. Inspired from previous feature selection studies [21, 32-39], we intensify on information gain (IG) and chi-square (CHI) feature selection methods and figure out IG and CHI values of each features. In other words, we try to compose the set

of the most significant features with high classification success for associating to the original feature space. Indeed, the overall feature selection process is to count for score each feature in accordance with a certain feature selection method, and then pick up the best k features. In this work, k is adjusted as the half number of features according to the datasets. For example, if a dataset contains 100 features, then the best 50 features are determined by the feature selection methods.

Eventually, original features are employed as the first part of feature space while the remaining part is composed from the modified significant features which are selected with the techniques mentioned above. In detail, the difference operator as a feature generation operator is applied to the significant features obtained by feature selection methods and original features. Then, the acquired new features are added to the original feature space. That is, the enhanced feature space is constituted with the combination of original features and modified significant features. The significant part is to add new features to the original feature space and to decide how many features should be united with the original feature space. Experiments are carried out on the original d number of features and add 2d, 3d, 4d, 5d, 6d new features to the original features in [1]. They observe that adding d new features to the original ones (Org (d) + New (d)) is the best. In [2], authors concentrate on to add d and d/2 new features to the original feature space. Their experiment results indicate that d/2 space extension outperforms d extension. In this work, we decide to arrange experiment settings according to the d/2 space extension due to its superior performance as stated in [2].

After constructing the enhanced feature space, the training phase is evaluated with the well-known classification algorithm, namely decision tree by employing three types of ensemble algorithms. Thus, the usage of more than one decision tree emerges decision forests through the ensemble algorithms. As we mentioned before, the main objective is to ensure the diversity of base learners. When the same type base learners are preferred, the diversity is composed by using different training datasets otherwise base learners already maintain the diversity of them. We propose to implement the same type base learners and use three types of ensemble algorithms to create diversity, namely bagging, random subspaces, and random forests. Our proposed approach is described in details below.

Algorithm 1: Enhanced Space Forest Algorithm.

Given: $E=\{x_p, y_p\}_{p=1...N} =[X\ Y]$ where X is an N*d matrix including the training set and Y is an N dimensional column vector covering the class labels. d is the number of features, N is the number of training samples, T is the number of base learners, BL_i is the base learner, E_i is the enhanced training set for BL_i, R_i consists of new features used in generation of E_i, EA is an ensemble algorithm.

Initialization: Choose ensemble size T, the base learner model BL_i, and the ensemble algorithm EA.

Training:

for i=1:T

1. Create new features (EX_i) by using feature selection techniques (IG, or CHI).

   Generate d/2 number of features with IG and store in R_i, or

   Generate d/2 number of features with CHI and store in S_i.

   j=1

   for z=1:d step by 2

   Create jth new feature applying difference operator to $X_i(z)^{th}$ and $R_i(z)^{th}$ or $S_i(z)^{th}$ features of X matrix.

   j=j+1

endfor

2. Construct the new training set (E_i) by concatenating the matrix X (original features) and R_i, or X and S_i, seperately as $E_i=[X\ R_iY]$, $E_i=[X\ S_iY]$, respectively.
3. Train BL_i with E_i according to EA.

endfor

Testing:

for i=1:T

1. Enhance the feature space of the test sample.
2. Classify the enhanced test sample with BL_i.

endfor

Combine the base learners' decisions by the combination rule of the chosen ensemble algorithm EA.

4. Experiment Setup

The datasets with different sizes and properties are listed in Table 1. All of them are available from the UCI Machine Learning Repository [29]. Characteristics of the datasets are given in Table 1 including the number of features (|F|), the number of classes (|C|), and the number of samples (|S|). We carry out experiments by changing the training set size and utilizing following percentages of the data for training and the rest for testing: 1%, 5%, 10%, 30%, 50% and 80% as stated in [40-41].

Table 1. Characteristics of the datasets

DatasetID	Dataset	|F|	|C|	|S|
1	abalone	10	19	4153
2	anneal	62	4	890
3	audiology	69	5	169
4	autos	71	5	202
5	balance-scale	4	3	625
6	breast-cancer	38	2	286
7	breast-w	9	2	699
8	col10	7	10	2019
9	colic	60	2	368
10	credit-a	42	2	690
11	credit-g	59	2	1000
12	d159	32	2	7182
13	diabetes	8	2	768
14	glass	9	5	205
15	heart-statlog	13	2	270
16	hepatisis	19	2	155
17	hypothyroid	31	3	3770
18	ionosphere	33	2	351
19	iris	4	3	150
20	kr-vs-kp	39	2	3196
21	labor	26	2	57
22	letter	16	26	20000
23	lymph	37	2	142
24	mushroom	112	2	8124
25	primary-tumor	23	11	302
26	ringnorm	20	2	7400
27	segment	18	7	2310
28	sick	31	2	3772
29	sonar	60	2	208
30	soybean	83	18	675
31	splice	287	3	3190
32	vehicle	18	4	846
33	vote	16	2	435
34	vowel	11	11	990
35	waveform	40	3	5000
36	zoo	16	4	84

To prevent confusion with accuracy percentages, these are indicated with "ts" prefix and performed for each dataset and ensemble algorithm. The repeated holdout method is applied 10 times on each dataset. The differences between accuracies are statistically tested with 95% confidence level with Student's t-test. In all tables, the significant difference, the significant win, and the significant loss mean the statistically significant difference, the statistically significant win, the statistically significant loss, respectively. The number of base learners (T) is set to 100 as

represented in [1, 10]. As we mentioned before we apply d number of features as a feature enhancement parameter for all datasets to compare experiment results with impressive work [1-2, 10].

In order to evaluate the performance of ensembles, various success dynamics are used: Ensemble accuracy (EA), individual accuracy of base learners (IA) and kappa value of base learners (KP). The average value of T accuracy values is employed as the mean individual accuracy of base learners, where T is the number of base learners. The individual accuracy of base learners can be high if more similar training sets to the original training set are created by ensemble algorithms. On the other hand, highly accurate base learners cause the lower diversity. Thus, these two success dynamics of the base learners are inversely proportional. Only one is not enough to demonstrate the performance of an ensemble.

Kappa is a pairwise diversity measurement and measures the level of agreement between two classifier outputs [42]. In our study, one base learner is employed as one of the classifiers and the majority voted decision of all base learners except the utilized one is the other classifier. Kappa value (KP) of the ensemble is referred to the averaged kappa value of each base learner. KP value also is indirectly proportional to the diversity of an ensemble. The lower KP values demonstrate higher diversity since the level of agreement between classifier outputs is evaluated by Kappa measure.

5. Experiment Results

We make use of the accuracy results to ensure the comparison of our experiment results with the previous studies. Original versions and enhanced space forests are compared in terms of their ensemble accuracy (EA), individual accuracy of base learners (IA) and kappa value of base learners (KP). Abbreviations are employed as follows: BG: Bagging, RS: Random Subspace, RF: Random Forest, X_0: Original feature space of the dataset for X ensemble algorithm, X_{IG}: Information gain based enhanced space forests for X ensemble algorithm, X_{CHI}: Chi-square based enhanced space forests for X ensemble algorithm, Ts: Training set percentage.

The averaged ensemble accuracies are analyzed in terms of training set percentages on 36 datasets. We observe enhanced versions of ensemble algorithms have superior classification performance compared to the original version by looking at the overall perspective for all training set percentages. Furthermore, IG-based enhanced space forests generally outperform both original version and CHI-based enhanced space forests of 36 datasets. If we compare the ensemble accuracies at ts80, we get the performance order as $RF_{IG} > RS_{IG} > RF_{CHI} > RS_{CHI} > BG_{IG} > BG_{CHI} > RF_O > RS_O > BG_O$. Except ts5 and ts50, the best classification performance is performed by RF_{IG} at all training set sizes. Thus, we can assert that IG-based enhanced space forests usually contribute to the classification performance significantly for 36 datasets.

From ts10 to ts80, the success order of original ensemble algorithms is RF > RS > BG. At smaller training set sizes, the performance order of original ensemble algorithms is different but not enough to claim statistically significant because of the proximity of accuracy results. It is also considerable to notify that RF_{IG} outperforms the others at all training set percentages except ts50 and ts5 levels. For ts50 and ts5, RS_{IG} is competitive and surpasses other techniques by at 2%. The combination of random subspace as an ensemble algorithm and information gain as a feature space enhancement technique yields by far the highest accuracies at these training set levels. At the last of Table 2, average accuracy results are given. Average accuracy results for bagging algorithm demonstrate that IG-based enhanced space forests present the best classification success with 87.4% accuracy result at ts80 compared to the others. If we compare original and enhanced versions of the bagging algorithm, we can get the performance order as: $BG_{IG} > BG_{CHI} > BG_o$. Similarly, IG-based enhanced space forests outperform others for the random subspace and the random forest algorithms with 88.0% and 88.2% accuracy results, respectively. Like bagging algorithm, the performance order of the random subspace and the random forest is the same. Hence, IG-based enhanced space forests are the best technique to improve the classification performance for each ensemble algorithm by evaluating average accuracy results.

or original feature space, random forest is also the best ensemble algorithm with 87.0% classification success and followed by random subspace with 86.9% and bagging with 86.2% accuracy result, respectively. Classification performances for IG-based enhanced space forests are ordered in a similar way as: RF > RS > BG and the classification performance is consistent with the state of the art results [1]. This order is also valid for CHI-based enhanced space forests but classification accuracies are different from each other in that 87.8% (RF), 87.6% (RS), 87.2% (BG).

Table 2. Classification Accuracies of Enhanced and Original Versions of the Ensemble Algorithms at ts80.

Dataset ID	BG_0	BG_{IG}	BG_{CHI}	RS_0	RS_{IG}	RS_{CHI}	RF_0	RF_{IG}	RF_{CHI}
1	27.3	29.1	28.5	27.2	28.1	27.5	28.1	28.4	28.3
2	99.1	99.8	99.5	99.2	99.3	98.8	99.6	99.7	99.5
3	89.3	91.5	89.3	87.4	92.1	90.2	87.2	88.9	87.2
4	73.1	75.9	75.2	72.6	74.1	73.8	72.1	75.8	75.2
5	86.1	98.2	96.4	87.0	94.5	92.1	88.1	99.3	97.6
6	73.4	74.6	74.1	75.3	75.3	75.3	75.4	74.1	73.5
7	97.3	98.1	97.5	97.8	97.7	97.7	97.8	97.8	97.8
8	81.6	81.9	80.8	81.7	81.6	81.5	81.7	81.7	81.7
9	85.3	86.7	86.4	85.9	88.1	87.5	84.2	87.3	86.7
10	88.1	89.1	88.8	89.2	89.2	88.7	88.6	88.4	87.9
11	77.9	79.2	78.8	77.6	79.2	78.6	78.1	79.5	78.7
12	99.1	99.8	99.7	99.0	99.6	99.3	99.9	99.2	99.5
13	76.9	77.3	76.6	76.2	77.4	76.5	77.1	78.4	77.9
14	74.1	75.8	75.4	75.1	77.9	77.2	74.3	73.5	73.0
15	82.2	82.7	82.2	83.6	83.4	83.0	84.2	83.6	83.0
16	82.9	86.4	85.7	85.7	87.9	87.5	86.0	87.2	86.9
17	99.6	99.7	99.6	97.7	99.9	99.9	99.7	99.8	99.8
18	93.9	95.4	94.9	94.9	95.8	95.1	94.0	95.7	95.2
19	97.1	97.0	96.8	96.5	97.6	97.3	96.9	96.4	96.8
20	99.1	99.3	99.2	98.4	99.5	99.0	98.8	99.3	99.2
21	90.7	90.4	89.0	93.9	93.2	92.7	92.2	96.7	96.2
22	93.7	97.8	97.0	96.0	97.2	96.9	96.1	97.4	96.8
23	85.7	87.4	87.1	86.8	86.8	86.7	86.2	88.3	87.9
24	98.7	99.0	98.5	99.2	99.0	98.7	99.7	99.3	99.1
25	51.3	51.0	50.6	51.8	51.9	51.4	51.7	53.8	53.3
26	95.7	97.2	96.4	97.8	97.8	97.7	96.4	97.5	97.0
27	97.3	97.0	97.6	97.6	98.0	98.1	98.2	97.8	98.3
28	99.1	89.7	99.0	98.3	99.5	99.2	98.7	98.1	98.8
29	79.4	79.5	79.1	80.8	82.3	81.9	81.7	82.9	82.5
30	93.2	92.7	92.2	93.5	93.1	93.2	92.5	93.6	93.4
31	96.0	96.5	96.2	97.1	96.9	96.7	96.5	97.8	97.4
32	76.0	80.6	80.2	76.4	79.8	79.2	77.3	80.8	80.1
33	97.1	98.0	98.4	97.3	98.2	98.1	97.8	98.5	98.4
34	83.4	87.7	87.1	88.1	89.9	91.0	88.5	90.4	90.2
35	86.1	87.8	87.6	87.5	88.7	88.3	88.2	88.9	88.6
36	96.5	97.5	97.1	99.1	99.0	99.0	99.7	99.3	99.2
avg	86.2	**87.4**	87.2	86.9	**88.0**	87.6	87.0	**88.2**	87.8

It is important to note that classification results of the random forest and the random subspace algorithms are close to each other but yet, random forest algorithm generally outperforms the others at ts80 by evaluating Table 2. Information gain as a feature space enhancement technique is an ideal to enhance feature space. As we mentioned above, if we compare all versions of the original and enhanced space forests, we can get the classification success order as: $RF_{IG} > RS_{IG} > RF_{CHI} > RS_{CHI} > BG_{IG} > BG_{CHI} > RF_O > RS_O > BG_O$. In this work, we try to get better classification performance compared to the previous studies [1-2, 10]. For this purpose, it is propose to implement training procedure with the enhanced space forests by utilizing an appropriate feature space enhancement

Table 3. Comparison Between Pairs of Algorithms: "Win(Significant Win)/ Loss(Significant Loss)" Row vs. Column.haracteristics of the datasets.

	BG_{IG}	BG_{CHI}	BG_0	RS_{IG}	RS_{CHI}	RS_0	RF_{IG}	RF_{CHI}	RF_0
BG_{IG}	0/0	33(6)/3(0)	31(13)/5(2)	11(3)/25(8)	18(5)/18(4)	22(7)/14(3)	12(2)/24(6)	19(3)/17(6)	24(5)/12(4)
BG_{CHI}	3(0)/33(6)	0/0	27(7)/9(1)	8(2)/28(8)	13(1)/23(5)	19(3)/17(3)	7(1)/29(9)	13(2)/23(7)	15(5)/21(3)
BG_0	5(2)/31(13)	9(1)/27(7)	0/0	2(0)/34(16)	5(0)/31(13)	9(1)/27(8)	4(0)/32(18)	6(1)/30(14)	7(1)/29(10)
RS_{IG}	25(8)/11(3)	28(8)/8(2)	34(16)/2(0)	0/0	32(3)/4(0)	26(10)/10(0)	14(2)/22(4)	17(5)/19(3)	26(10)/10(0)
RS_{CHI}	18(4)/18(5)	23(5)/13(1)	31(13)/5(0)	4(0)/32(3)	0/0	21(6)/15(2)	13(2)/23(6)	12(3)/24(6)	24(8)/12(0)
RS_0	14(3)/22(7)	17(3)/19(3)	27(8)/9(1)	10(0)/26(10)	15(2)/21(6)	0/0	8(2)/28(17)	9(3)/27(8)	13(2)/23(4)
RF_{IG}	24(6)/12(2)	29(9)/7(1)	32(18)/4(0)	22(4)/14(2)	23(6)/13(2)	28(17)/8(2)	0/0	31(5)/5(0)	24(11)/12(1)
RF_{CHI}	17(6)/19(3)	23(7)/13(2)	30(14)/6(1)	19(3)/17(5)	24(6)/12(3)	27(8)/9(3)	5(0)/31(5)	0/0	25(9)/11(2)
RF_0	12(4)/24(5)	21(3)/15(5)	29(10)/7(1)	10(0)/26(10)	12(0)/24(8)	23(4)/13(2)	12(1)/24(11)	11(2)/25(9)	0/0

technique and ensemble algorithms. Experimental results demonstrate that the combination of IG-based enhanced space forests with random forest as an ensemble algorithm has the superior classification performance.

Table 4. Success Dynamics of the Original and Enhanced Space Versions of the Algorithms at ts80: Win/Loss Numbers, Mean EA, IA, and KP accuracies.

	Significant Win-Significant Loss	EA mean accuracy	IA mean accuracy	KP
BG_{IG}	11	87.47	80.93	73.42
BG_{CHI}	-21	87.23	80.57	73.00
BG_0	-93	86.24	80.75	74.55
RS_{IG}	50	88.08	78.96	69.57
RS_{CHI}	18	87.68	78.57	68.63
RS_0	-33	86.96	75.73	61.48
RF_{IG}	66	88.24	79.94	68.92
RF_{CHI}	29	87.89	79.21	68.24
RF_0	-27	87.07	77.36	64.57

Unlike smaller training set percentage levels, IG-based random forest algorithm reaches the maximum value from ts10 to ts80. Accuracy difference between IG-based random forest algorithm and the others is observed up to 3% especially at ts30 and ts10. All versions of random subspace algorithm have the following best classification performance. So long as the training set percentages increase, the success of all enhanced space forests also rises up and vice versa. Original versions of all ensemble algorithms represent the lowest classification accuracies at higher training set percentages. At these training set levels, the choice of the original versions of the ensemble algorithms will not be a good preference for the classification problems.

The following conclusions can be drawn from Table 3: BG_{IG} has higher accuracy than BG_O over 31 datasets out of 36, and has 13 significant wins. RS_{IG} has higher accuracy than BG_{CHI} over 28 datasets out of 36, and has 8 significant wins. RF_O has higher accuracy than RS_O over 23 datasets out of 36, and 4 significant wins. RF_{CHI} has higher accuracy than RS_{CHI} over 24 datasets out of 36, and 6 significant wins.

In Table 4, it is obviously seen that the enhanced space forests demonstrate superior performance compared to the original ones. The win/loss number order is coherent with accuracy (EA) results mentioned before: $RF_{IG} > RS_{IG} > RF_{CHI} > RS_{CHI} > BG_{IG} > BG_{CHI} > RF_O > RS_O > BG_O$. The performance of the original ensemble algorithms is also consistent with the literature: $RF_O > RS_O > BG_O$. Individual accuracy (IA) and diversity of base learners are important parameters for an ensemble success. If we compare the accuracies of individual base learners, the order of success is as follows: $BG_{IG} > BG_O > BG_{CHI} > RF_{IG} > RF_{CHI} > RS_{IG} > RS_{CHI} > RF_O > RS_O$. The diversity measure (1-KP) is ordered as: $RS_O > RF_O > RF_{CHI} > RS_{CHI} > RF_{IG} > RS_{IG} > BG_{CHI} > BG_{IG} > BG_O$. These results demonstrate that performance of the ensemble algorithm is based on both individual accuracy and diversity of base learners. Moreover, they are inversely proportional as it is expected.

In order to verify the efficiency of our proposed approach, the comparison of experimental results is evaluated between ours and the influential study [1]. They also use 36 datasets from the UCI repository [29] and all of them are common with ours. In Table 5, X_{RND} is referred to the enhanced space forests by adding randomly selected features which is asserted in [1] where X is the ensemble algorithm. X_{IG} and X_{CHI} are our proposed enhanced space forests where X is the ensemble algorithm.

Table 5. Comparison with the state of the art study [1] at ts50.

	Mean accuracy
BG_{RND}	85.3
BG_{IG}	**85.8**
BG_{CHI}	**85.4**
RS_{RND}	85.8
RS_{IG}	**86.7**
RS_{CHI}	**86.0**
RF_{RND}	85.9
RF_{IG}	**86.9**
RF_{CHI}	**86.4**

Random forest algorithm is the best ensemble algorithm for all enhanced space forests considering average accuracy results. Thus, the classification success of the randomly generating enhanced feature space is consistent with our approaches in terms of the experiment results. The classification performance of the ensemble algorithms can be ordered for all enhanced techniques as: RF > RS > BG. Furthermore, our proposed approaches boost the classification success of the ensemble system compared to the randomly enhanced space forests. In other words, results of our extensive experiments demonstrate that the proposed enhanced space forest models can significantly outperform the randomly enhanced space forests. As it is observed from the averaged accuracy results, X_{IG} provides approximately 1% improvement compared to the X_{RND}. X_{CHI} is also ambitious and performs competitively in proportion to the X_{RND}. Considering of the classification performance of the enhanced spaces, the order is as follows: $X_{IG} > X_{CHI} > X_{RND}$.

6. Discussion and Conclusion

The superiority of ensemble algorithms is a widely accepted assumption in machine learning domain as mentioned before. Owing to this approach, it is recommended to produce more accurate and robust classification models. Diversity of the base learners and their individual success are essence of the ensemble algorithms and they are inversely proportional. Thus, diversification of the base learners is suggested by making use of different training sets through several ensemble algorithms. In this work, we propose to investigate contribution of the enhanced space forests to the classification performance. For this purpose, different feature enhancement techniques are employed on the original feature space and compared the classification performances of

them to the original versions. Furthermore, the enhanced space forests is implemented on the three popular ensemble algorithms (Bagging, Random Subspaces, and Random Forest) among various ensemble algorithms. Then, the classification performances of the original and the enhanced versions are analyzed in accordance with the ensemble algorithms. The extensive experiment results indicate that the enhanced space forests have the superior classification success compared to the original versions. IG-based enhanced space forests significantly outperform CHI-based enhanced space forests and original versions.

The classification performances of the ensemble algorithms are also investigated. The diversity of base learners is more explicit for the random forest and the random subspace algorithms. Therefore, the accuracy results of them have better performances compared to the bagging algorithm at all training set percentages. Especially, random forest algorithm challenges to the other ensemble algorithms because of its remarkable classification success. Classification success of the ensemble algorithms is ordered as: RF > RS > BG. This order is also consistent with the literature results [1] and valid for all enhanced space forests including information gain and chi-square versions. Additionally, IG-based enhanced space forests significantly outperform CHI-based enhanced space forests and original versions. The ensemble accuracy performances of both the original and the enhanced space forests are ordered as: $RF_{IG} > RS_{IG} > RF_{CHI} > RS_{CHI} > BG_{IG} > BG_{CHI} > RF_{O} > RS_{O} > BG_{O}$. It is obviously seen that the random forest is the best ensemble algorithm. As a result, the most optimal values for diversity and individual accuracy of base learners are provided by random forest algorithm. Considering the overall classification performances, ensemble algorithms with the original feature space have the lowest accuracy results at all training set levels. It is mentioned before that the diversity (1-KP) and the individual success of base learners are inversely proportional. The diversity is (1-KP) ordered as: $RS_{O} > RF_{O} > RF_{CHI} > RS_{CHI} > RF_{IG} > RS_{IG} > BG_{CHI} > BG_{IG} > BG_{O}$. The order of accuracies of individual base learners is as follows: $BG_{IG} > BG_{O} > BG_{CHI} > RF_{IG} > RF_{CHI} > RS_{IG} > RS_{CHI} > RF_{O} > RS_{O}$. In this way, the most similar training set to the original training set is generated by bagging algorithm. Thus, diversity of bagging is the lowest but accuracy of base learners of it has the highest results. As a result, the less similar training sets to the original training set is a way to get more diverse base learners.

As well as the classification performance of ensemble algorithms, execution time analysis is evaluated in terms of testing and training times. More training time is needed for the enhanced space forests in proportion to original ones. Furthermore, the enhanced space forests cover more features and directly proportional to the search time of the features. Because of these reasons, less training time is required for the original versions of the ensemble algorithms. The complexity of produced base learners gives us a clue about the testing times and is also proportional to the number of nodes in a tree. The most complex base learners are produced by the random forest algorithm due to having the biggest trees. We can conclude the execution time analysis that the enhanced space forest versions are required less testing time because of having smaller trees compared to the original feature space version of ensemble algorithms.

To sum up, the enhanced space forests improve the classification success compared to the original versions. In this study, it is observed that the IG-based and CHI-based enhanced space forests exhibit better classification performance in proportion to the other enhanced space forests. In future, we also plan to apply heterogeneous classifier ensembles to the classification problems. It is also planned to investigate the performance of the combination of classifier ensembles and enhanced space forests on text classification domain.

References

[1] M. F. Amasyali and O. K. Ersoy, "Classifier ensembles with the extended space forest," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 3, pp. 549–562, 2014.

[2] M. N. Adnan, M. Z. Islam, and P. W. H. Kwan, "Extended space decision tree," in *Proc. Machine Learning and Cybernetics*, Lanzhou, China, 2014, pp. 219–230.

[3] B. Peralta and A. Soto, "Embedded local feature selection within mixture of experts" *Inform. Sciences*, vol. 269, pp. 176–187, 2014.

[4] F. N. Koutanaei, H. Sajedi, and M. Khanbabaei, "A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring," *J. Retailing and Consumer Servic.*, vol. 27, pp. 11-23, 2015.

[5] D. Gopika and B. Azhagusundari B, "An analysis on ensemble methods in classification tasks", *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 3, no. 7, pp. 7423–7427, 2014.

[6] Y. Ren, L. Zhang, and P. N. Suganthan, "Ensemble classification and regression-recent developments, applications and future directions," *IEEE Comput. Intell. Mag.*, vol. 11, no. 1, pp. 41–53, 2016.

[7] L. Mesin, A. Munera, and E. Pasero, "A low cost ecg biometry system based on an ensemble of support vector machine classifiers," *Adv. Neural Networks*, vol. 54, pp. 425–433, 2016.

[8] M. Zamani, H. Beigy, and A. Shaban, "Cascading randomized weighted majority: a new online ensemble learning algorithm," *J. Intell. Data Anal.*, vol. 20, no. 4, pp. 877–889, 2016.

[9] J. V. Lochter, R. F. Zanettib, D. Rellera, and T. A. Almeidaa, "Short text opinion detection using ensemble of classifiers and semantic indexing," *Expert Syst. Appl.*, vol. 62, pp. 243–249, 2016.

[10] M. N. Adnan and M. Z. A. Islam, "Comprehensive method for attribute space extension for random forest," in *International Conference on Computer and Information Technology*, Dhaka, Bangladesh, 2014, pp. 25–29.

[11] A. Ahmed and G. Brown, "Random projection random discretization ensembles - ensembles of linear multivariate decision trees," *IEEE T. Knowl. Data En.*, vol. 26, no. 5, pp. 1225–1239, 2014.

[12] L. Liu, B. Wang, Q. Zhong, and H. Zeng, "A selective ensemble method based on k-means method," in *International Conference on Computer Science and Network Technology*, Harbin, China, 2015, pp. 665–668.

[13] S. Deepan and D. Menaka, "Ensemble classification of urban regions using hyperspectral remote sensed scenes," *Middle-East J. Sci. Res.*, vol. 24, no. S1, pp. 49–54, 2016.

[14] D. Mera, M. Fernández-Delgado, J. M. Cotos, J. R. R. Viqueira, and S. Barro, "Comparison of a massive and diverse collection of ensembles and other classifiers for oil spill detection in sar satellite images," *J. Neural Comp. Appl.*, vol. 27, no. 139, pp. 1–17, 2016.

[15] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[16] G. Wang, Z. Zhang, J. Sun, S. Yang, and C. A. Larson, "POS-RS: A random subspace method for sentiment classification based on part-of-speech analysis," *Inform. Process. Manag.*, vol. 51, no. 4, pp. 458–479, 2015.

[17] R. Farzi and V. Bolandi, "Estimation of organic facies using ensemble methods in comparison with conventional intelligent approaches: A case study of the south pars gas field Persian Gulf, Iran", *Journal of Modeling Earth Systems and Environment*, vol. 2, no. 6, pp. 105–118, 2016.

[18] A. Onan, S. Korukoglu, and H. Bulut, "Ensemble of keyword

extraction methods and classifiers in text classification," *Expert Syst. Appl.*, vol. 57, pp. 232–247, 2016.

[19] K. Grzesiak-Kopeć, M. Ogorzałek, and L. Nowak, "Computational classification of melanocytic skin lesions," in *International Conference on Artificial Intelligence and Soft Computing*, Zakopane, Poland, 2016, pp. 169–178.

[20] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE T. Pattern Anal.*, vol. 20, no. 8, pp. 832–844, 1998.

[21] A. Onan, "Classifier and feature set ensembles for web page classification,", *J. Inf. Sci*., vol. 42, no. 2, pp. 150–165, 2015.

[22] D. Aldogan and Y. Yaslan, "A comparison study on ensemble strategies and feature sets for sentiment analysis," in *International Symposium on Computer and Information Sciences*, London, UK, 2015, pp. 359–370.

[23] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[24] F. M. Belem, E. F. Martins, J. M. Almeida, and M. A. Gonçalves, "Personalized and object-centered tag recommendation methods for web 2.0 applications", *Inf. Process. Manag.*, vol. 50, no. 4, pp. 524–553, 2014.

[25] A. P. Jain and V. D. Katkar, "Sentiments analysis of twitter data using data mining," in *International Conference on Information Processing*, Pune, India, 2015, pp. 807–810.

[26] R. R. Rejimol Robinson and C. Thomas, "Ranking of machine learning algorithms based on the performance in classifying ddos attacks," *IIEE Recent Advances in Intelligent Computational Systems*, Trivandrum, Kerala, India, 2015, pp. 185–190.

[27] D. J. Dittman, T. M. Khoshgoftaar, and A. Napolitano, "The effect of data sampling when using random forest on imbalanced bioinformatics data," *International Conference on Information Reuse and Integration*, San Francisco, CA, 2015, pp. 457–463.

[28] M. N. M. García, J. C. B. Herráez, M. S. Barba, and F. S. Hernández, "Random forest based ensemble classifiers for predicting healthcare-associated infections in intensive care units," in *International Conference on Distributed Computing and Artificial Intelligence*, Sevilla, Spain, 2016, pp. 303–311.

[29] M. Lichman, UCI Machine Learning Repository, Available: http://archieve.ics.uci.edu/ml, University of California, Irvine, School of Information and Computer Science, 2013.

[30] V. Singh and M. A. Pradhan, "Advanced methodologies employed in ensemble of classifiers: a survey," *Int. J. Sci. Res.*, vol. 3, no. 12, pp. 591–595, 2014.

[31] N. Rooney, H. Wang, P. S. Taylor, "An investigation into the application of ensemble learning for entailment classification," *Inf. Process. Manag*., vol. 50, no. 1, pp. 87–103, 2014.

[32] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," *SIGKDD Explorations*, vol. 6, no. 1, pp. 80–89, 2004.

[33] A. Abu-Errub, "Arabic text classification algorithm using tfidf and chi square measurements," *Int. J. Comput. Appl.*, vol. 93, no. 6, pp. 40–45, 2014.

[34] V. Chauraisa and S. Pal, "Data mining techniques: to predict and resolve breast cancer survivability," in *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 1, pp. 10–22, 2014.

[35] A. G. Neha, "A novel clustering approach based sentiment analysis of social media data," *Int. J. Eng. Dev. Res.*, vol. 3, no. 4, pp. 1099–1107, 2015.

[36] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Inf. Process. Manag.*, vol. 50, no. 1, pp. 104–112, 2014.

[37] N. Rachburee and W. Punlumjeak, "A comparison of feature selection approach between greedy, ig-ratio, chi-square, and mrmr in educational mining," in *International Conference on Information Technology and Electrical Engineering*, Chiang Mai, Thailand, 2015, pp. 420–424.

[38] M. A. Siddiqui, "An empirical evaluation of text classification and feature selection methods," *Artif. Intel. Res.*, vol. 5, no. 2, pp. 70–81, 2016.

[39] E. Zorarpacı, S. A. Özel, "A hybrid approach of differential evolution and artificial bee colony for feature selection," *Expert Syst. Appl.*, no. 62, pp. 91-103, 2016.

[40] Z. H. Kilimci and M. C. Ganiz, "Evaluation of classification models for language processing" in *International Symposium on INnovations in Intelligent SysTems and Applications*, Madrid, Spain, 2015, pp. 1–8.

[41] Z. H. Kilimci and S. Akyokus, "N-gram pattern recognition using multivariate bernoulli model with smoothing methods for text classification," *IEEE Signal Processing and Communications Applications Conference*, Zonguldak, Turkey, 2016, pp. 79–82.

[42] D. D. Margineantu and T. G., "Dietterich pruning adaptive boosting," in *International Conference on Machine Learning*, San Francisco, USA, 1997, pp. 211-218.