Improved Effort Estimation of Heterogeneous Ensembles using Filter Feature Selection

Mohamed Hosni[1], Ali Idri[1] and Alain Abran[2]
[1]*Software Project Management Research Team, ENSIAS, Mohammed V University Rabat, Morocco*
[2]*Department of Software Engineering, École de Technologie Supérieure Montréal, Canada*

Keywords: Software Development Effort Estimation, Machine Learning, Ensemble, Feature Selection, Filter.

Abstract: Estimating the amount of effort required to develop a new software system remains the main activity in software project management. Thus, providing an accurate estimate is essential to adequately manage the software lifecycle. For that purpose, many paradigms have been proposed in the literature, among them Ensemble Effort Estimation (EEE). EEE consists of predicting the effort of the new project using more than one single predictor. This paper aims at improving the prediction accuracy of heterogeneous ensembles whose members use filter feature selection. Three groups of ensembles were constructed and evaluated: ensembles without feature selection, ensembles with one filter, and ensembles with different filters. The overall results suggest that the use of different filters lead to generate more accurate heterogeneous ensembles, and that the ensembles whose members use one filter were the worst ones.

1 INTRODUCTION

Software Development Effort Estimation (SDEE) aims at providing the amount of effort needed to develop a software system. The estimates provided play a decisive role on the success of a project management, since it allows software managers to allocate adequately the resources needed to build the software system. Providing an accurate effort estimate has been the subject of many studies for more than four decades and a large number of techniques have been proposed (Azzeh et al., 2015; Hosni et al., 2017c; Hosni et al., 2017a; Zhu, 2010; Hosni and Idri, 2017). This paper deals with ensemble techniques. Ensemble techniques have been successfully applied to solve many classification and regression tasks (Zhu, 2010; Zhou, 2012). They consist of aggregating the outputs of a set of single techniques by means of a combination rule. Ensembles techniques have been also applied in SDEE and will be referred to here as EEE (Ensemble Effort Estimation). The literature distinguishes two types of ensembles (Idri et al., 2016c): (1) Homogeneous EEE is divided into two subtypes: ensembles that combine at least two configurations of the same single SDEE technique, and ensembles that combine one meta model such as Bagging, Boosting, Negative Correlation, or Random Subspace and one single SDEE technique; (2) Heterogeneous (HT) EEE which combines at least two different SDEE single techniques.

The systematic review of Idri et al. (Idri et al., 2016a) has documented that in general ensembles outperformed their members. However, some studies of EEE have shown the opposite (Hosni et al., 2017a; Kocaguneli et al., 2009). It has been observed that the accuracy of an ensemble mainly depends on two main criteria: accuracy and diversity of its members (Idri et al., 2016a; Idri et al., 2016b). In other words, the estimates of an ensemble are influenced by the estimates of its members, thus, they should be as accurate as possible. Also, they should be diverse (e.g. make different errors in the same data point). Consequently, each ensemble member can cancel the estimation errors done by other members. Otherwise, an ensemble that integrates non-diverse members may produce a lower estimation accuracy than its members. Although, there is no generally accepted formal definition of ensemble diversity, there are some mechanisms used to generate diversity among them selecting input features (Zhou, 2012), which was investigated in this paper.

Within this context, we carried out an empirical evaluation of heterogeneous ensembles whose members were K-nearest neighbor (KNN), Multilayer Perceptron (MLP), Support Vector Regression (SVR), and Decision Trees (DTs) (Hosni et al., 2017a). This

Hosni, M., Idri, A. and Abran, A.
Improved Effort Estimation of Heterogeneous Ensembles using Filter Feature Selection.
DOI: 10.5220/0006929104050412
In *Proceedings of the 13th International Conference on Software Technologies (ICSOFT 2018)*, pages 405-412
ISBN: 978-989-758-320-9

study has dealt with: (1) accuracy of the ensemble members by tuning their parameters using the grid search optimization technique, and (2) diversity by using two filters: Correlation based feature selection (CFS) and RReliefF. The results obtained showed that the ensembles whose members used filter selection were less accurate than their constituents and ensembles without feature selection as well. This paper presents an improvement to the selection process of the heterogeneous ensemble members of (Hosni et al., 2017a); in particular, we deal with the diversity criterion by selecting members using different filters, while in (Hosni et al., 2017a), the ensemble members used the same filter. Moreover, three instead of two filters were used in this study: CFS, RReliefF, and Linear Correlation (LC). The ensembles of this study were compared to ensembles of (Hosni et al., 2017a) and ensembles without feature selection in order to evaluate the impact of using different filters for ensemble members on the accuracy of the ensembles. The ensemble members of this study were the same as in (Hosni et al., 2017a): KNN, MLP, SVR and DTs. As for combiners, three linear rules were used: average, median, and Inverse Ranked Weighted Mean (IRWM).
The main contributions of this paper are: dealing with the diversity criterion of ensemble members by means of three filters; and evaluating the impact of diversity of ensemble members on the accuracy of EEE.

The rest of this paper is structured as follows: Section 2 presents an overview of the three filters as well as the four ML techniques used in this paper. Section 3 presents the findings of some related work in SDEE dealing with feature selection techniques for heterogeneous ensembles. Section 4 presents the empirical methodology pursued throughout this study. Section 5 presents and discusses the empirical results. Section 6 presents the conclusions of this empirical study.

2 BACKGROUND

2.1 Feature Selection Techniques

Feature selection aims at eliminating redundant and irrelevant features in order to reduce the complexity and to improve the performance of any learner. Several feature selection techniques are proposed in the literature and can be grouped into three categories (Jovic et al., 2015): Filter techniques, Wrapper techniques and Embedded techniques.

This paper used filter techniques since they are less costly and are performed independently of the learner. Three filters were used: Correlation based feature selection (CFS), RReliefF technique, and Linear Correlation. CFS belongs to the multivariate feature selection family: it assesses the full feature space, and select a subset of features. The other two techniques belong to the univariate feature selection family: they separately assess each attribute and provide a ranking of the features which presents an issue in selecting the number of features. Selecting $\log_2(N)$ attributes where N is the number of features available in the dataset, was the solution proposed in the literature and was adopted in this study (Hosni et al., 2017a; Khoshgoftaar et al., 2007).

2.2 Four ML Techniques and Their Parameters Settings

This study uses the same four ML techniques investigated in (Hosni et al., 2017a): Knn, SVR, MLP, and DTs.

It is well-known that the performance accuracy of an estimation technique depends on its parameters settings (Song et al., 2013; Hosni et al., 2017b; Hosni et al., 2017c). In (Hosni et al., 2017b), we conducted an empirical evaluation in which two optimization techniques, Grid Search and Particle Swarm Optimization, were used to set the parameters of four techniques KNN, MLP, SVR, and DTs. The results obtained showed that tuning the parameters by means of optimization techniques have a positive impact on the accuracy of these estimation techniques. Therefore, this paper uses a grid search optimization technique to set the parameters values of the four ML techniques. It consists of performing a preliminary round of executions in a predefined range of values and thereafter selecting the optimal configuration that allows the technique to generate the best accuracy with respect to a specific measure (Hosni et al., 2017b). This paper uses Mean Absolute Errors (MAE) as the performance measure and the configuration that leads the technique to generate less MAE was selected. Table 1 lists our predefined range of parameters values of the four ML techniques.

3 LITERATURE REVIEW

The aim of feature selection techniques is to select a set of features providing consistent information on the instances of the data. Thus, the selected features are used as inputs of a technique performing a knowledge data discovery task such as classification, prediction or clustering. Within this context, many studies in SDEE have investigated the use of feature se-

Table 1: Parameter Values For Grid Search.

Techniques	Parameters with their search spaces
Knn	K= {1, 2, 4, 8, 12} Similarity measure = {Euclidean Distance}
SVR	Complexity= {5, 10, 50, 100, 150} Extent to which deviation are tolerated = {0.0001, 0.001, 0.01, 0.1} Kernel=RBF Kernel Parameter= {0.0001, 0.001, 0.01, 0.1}
MLP	Learning rate= {0.01, 0.02, 0.03, 0.04, 0.05} Momentum= {0.1, 0.2, 0.3, 0.4, 0.5} Kernel=RBF Hidden layers= {3, 5, 9, 16}
DTs	Minimum instance per leaf = {1, 2, 3, 4, 5} Minimum proportion of the data variance at a node = {0,0001, 0,001, 0.01,0.1} Max depth= {1, 2, 4, 6, 8}

lection techniques, and the overall results have suggested that the use of feature selection improved the estimation accuracy of predictors (Hira and Gillies, 2015; Jovic et al., 2015). For instance, Idri et al. (Idri and Cherradi, 2016) studied the impact of two wrappers: feature forward selection and backward feature selection on the accuracy of the Fuzzy Analogy estimation technique: their results suggested that the two wrappers improved the accuracy of the Fuzzy Analogy technique.

As for the EEE, there are few papers that investigate the use of feature selection for ensembles. For instance, Minku et al. (Minku and Yao, 2013) showed that the use of CFS feature selection fails to improve the accuracy of MLP homogeneous ensembles (Bagging + MLP) but it improves the Radial Basis Function ensembles (Bagging and Negative Correlation Learning). Hosni et al. (Hosni et al., 2017a) carried out an empirical evaluation of heterogeneous ensembles whose members were KNN, SVR, MLP, and DTs. The members used two filters: CFS and RReliefF. Each ensemble contains four members with the same filter and uses one of three linear rules (average, median and IRWM). Thus, 9 heterogeneous ensembles were developed. These ensembles were assessed using Standardized Accuracy and Effect Size to check their reasonability; thereafter the Scott-Knott statistical test was performed to check the significant difference between the ensembles. The best ensembles that share the same predictive capability were ranked based on 8 performance measures through Borda Count. These experiments were performed over six datasets. The results obtained showed that the filter ensembles underperformed ensembles without filters (Hosni et al., 2017a).

4 EMPIRICAL DESIGN

4.1 Performance Measure and Statistical Test

The first question raised when evaluating an SDEE technique is whether the technique is actually predicting or only guessing (Idri et al., 2017; Shepperd and MacDonell, 2012). Thus, the Standardized Accuracy measure (SA, Eq.(8)) was used to check the reasonability of any technique with respect to a baseline method, and the Effect Size test (Δ, Eq.(9)) was adopted to assess if there is an effect improvement over the baseline method. The absolute values of Δ can be interpreted in terms of the categories proposed by Cohen (Cohen, 1992): small ($\approx$ 0.2), medium ($\approx$ 0.5) and large ($\approx$ 0.8). Thereafter, a set of accuracy measures were used to assess the predictive capability of a given technique: Pred(0.25) (Eq.(3)), MAE (Eq.(4)), Mean Balanced Relative Error (MBRE, Eq.(5)), Mean Inverted Balanced Relative Error (MIBRE, Eq.(6)) and Logarithmic Standard Deviation (LSD, Eq.(7)). However, given that the mean is very sensitive to outliers, the median of these measures was also used: median of absolute errors (MdAE), median of Balanced Relative Error (MdBRE), and median of Inverted Balanced Relative Error (MdIBRE). Note that the Pred(0.25) measure was used in this paper even if it is an MRE-based criterion: it was empirically proven in (Idri et al., 2017) that the possibility to generate biased results is very low in SDEE datasets compared to the other MRE-based criteria such as Mean Magnitude Relative Error (MMRE).

$$AE_i = |e_i - \hat{e}| \quad (1)$$

$$MRE_i = \frac{AE_i}{e_i} \quad (2)$$

$$Pred(0.25) = \frac{100}{N}\sum_{i=1}^{N}\begin{cases} 1 & if \quad MRE_i \leq 0.25 \\ 0 & Otherwise \end{cases} \quad (3)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N} AE_i \quad (4)$$

$$MBRE = \frac{1}{N}\sum_{i=1}^{N}\frac{AE_i}{min(e_i,\hat{e}_i)} \quad (5)$$

$$MIBRE = \frac{1}{N}\sum_{i=1}^{N}\frac{AE_i}{max(e_i,\hat{e}_i)} \quad (6)$$

$$LSD = \sqrt{\frac{\sum_{i=1}^{N}(\lambda_i + \frac{s^2}{2})^2}{N-1}} \quad (7)$$

$$SA = 1 - \frac{MAE_{p_i}}{\overline{MAE}_{p_0}} \quad (8)$$

$$\Delta = \frac{MAE_{p_i} - \overline{MAE}_{p_0}}{s_{p_0}} \quad (9)$$

Where:

- e_i and $\hat{e}_i$ are the actual and predicted effort for the ith project.
- $\overline{MAE}_{p_0}$ is the mean value of a large number runs of random guessing. This is defined as, predict a e_i for the target project i by randomly sampling (with equal probability) over all the remaining n-1 cases and take $e_i = e_r$ where r is drawn randomly from 1... n$\wedge$r $\neq$ i. This randomization procedure is robust since it makes no assumption and requires no knowledge concerning a population.
- MAE_{p_i} mean of absolute errors for a prediction technique i.
- S_{p_0} is the sample standard deviation of the random guessing strategy.
- λ_i = ln(e_i) - ln($\hat{e}_i$)
- s^2 is an estimator of the variance of the residual λ_i.

To check the significance of the difference between techniques, the Scott-Knott (SK) test was used (Scott and Knott, 1974): the SK test performs multiple comparisons and take into account the error type I correction.
Concerning the evaluation method, the Jackknife method was adopted in this paper.

4.2 Ensembles Construction

This paper evaluates three groups of heterogeneous ensembles whose members are KNN, MLP, SVR and DTs. These ensembles differ on the way on which they were constructed. The members were selected according to their accuracy values. The three groups are defined as follows:

- Ensembles whose members use different filters: ENF.
- Ensembles whose members do not use filter feature selection: E0F.
- Ensembles whose members use one filter (i.e. ensembles of (Hosni et al., 2017a)): E1F.

Methodology to construct the ensembles: the steps followed to select the ensemble members of ENF are described next. Note that before conducting the experiments, the three filters were applied over the six datasets to select the relevant features that will feed the four ML techniques.
The steps followed to construct the ENF ensembles are as follows:

Step a.1. Build the ML techniques using the three filters over the six datasets. The parameters settings of each technique were determined using the grid search technique with the range values of Table 1. The best three variants of each technique over each dataset were selected.

Step a.2. Evaluate the reasonability of the best three variants of each technique over each dataset in terms of SA and Δ, and select the ones achieving a reasonability higher than the 5% quantile of random guessing (high SA) and showing a large improvement in terms of Effect Size ($\Delta >$ 0.8).

Step a.3. Perform the SK test based on AE of the three variants of each technique of Step a.2 over each dataset. The aim of performing the SK test is to cluster the selected techniques and to identify the best ones (i.e. techniques that share the same predictive capability). Note that before conducting the SK test, the distribution of AEs of all selected techniques was checked to verify whether or not it follows a normal distribution using the Kolmogorov-Smirnov statistical test; this pre-step is necessary since the SK test required that its inputs should be normally distributed. The box-cox transformation was performed in order to make the AEs follow a normal distribution.

Step a.4. Rank the members of the best cluster of each technique over each dataset by means of Borda count using 8 accuracy measures: MAE,

MdAE, MIBRE, MdIBRE, MBRE, MdBRE, Pred, and LSD. The Borda voting system takes into consideration the rank provided by each accuracy criterion. The rationale behind using many accuracy measures is that prior studies in SDEE has showed that the selection of the best estimation technique depends on which indicator of accuracy used (Azzeh et al., 2015).

Step a.5. Select the best variant of each technique to be used as the base technique of the ENF ensembles. Therefore, for each dataset, three heterogeneous ensembles were defined whose members are the four best variants of KNN, MLP, SVR and DTs with the associated filter each. The three ensembles used the three linear combiners each: average, median and IRWM.

The steps followed to build the E1F and E0F ensembles are as follows:

Step b.1. Return the best variants of the four ML techniques (i.e. having the lowest MAE) using a grid search with the ranges of Table 1. This step is performed for each couple (technique, filter) over each dataset (E1F). It is also performed for each technique without feature selection (E0F).

Step b.2. Construct the E1F ensembles whose members were the best variants of the four techniques using the same filter. The three E1F (e.g. LC, CFS and R) ensembles used the three linear combiners each: average, median and IRWM. Construct the E0F ensembles whose members were the best variants of the four techniques without feature selection. E0F ensemble used the three linear combiners: average, medium and IRWM.

Comparison Methodology: to compare the ensembles (i.e. ENF, E0F and E1F) we used the same methodology as in (Hosni et al., 2017a) which consists of three steps:

Step c.2. Assess the accuracy of the ensembles with regards to SA and Effect Size, and select the ones that achieve SA values higher than the 5% quantile of random guessing and show a large improvement over random guessing in terms of Effect Size ($\Delta > 0.8$).

Step c.2. Cluster the selected ensembles through the SK test in order to select the ones that have similar predictive capability.

Step c.2. Rank the ensembles of the best cluster in each dataset according to 8 accuracy measures.

4.3 Abbreviation Adopted

We abbreviate the name of single and ensembles techniques as follows:

- Single techniques: {Feature_Technique}{Single_Technique}
- ENF: HT{Rule}.
- E0F: OD{Rule}.
- E1F: {Feature_Technique}{Rule}.

where:

- **Feature_Technique:** CFS, R, LC denote the Correlation based feature selection, RReliefF, and Linear Correlation respectively.
- **Single_Technique:** Knn, MLP, SVR, and DTs.
- **Rules:** AV, ME, IR denote the average, median, and inverse ranked weighted mean respectively.

Examples:
HTAV denotes the heterogeneous ensemble whose members are the four ML techniques using different feature selection techniques and the average as a combiner.
CFSIR denotes the ensemble whose members are the four ML techniques using the CFS feature selection technique and IR as combiner.

4.4 Datasets Description

Six well-known datasets were selected to assess the accuracy of the single and ensembles techniques. These datasets are diverse in terms of size, number of features, and they were collected from different organizations around the world and from different software application domains. Five datasets namely: Albrecht, China, COCOMO81, Desharnais, and Miyazaki were selected from the **PROMISE** repository while the other dataset came from the ISBSG repository (Release 8).
Note that, a cleaning and instance selecting steps were performed on the **ISBSG** dataset in order to select projects with high quality. This preprocessing stage results on a dataset that contains 148 projects described by means of nine attributes.

5 EMPIRICAL RESULTS

5.1 Feature Selection Step

Table 2 lists the features selected for each dataset. While none of the feature selection method chose the same subset of features, there is at least one common

attribute selected by the three filters, to the exception of China dataset in which the R and LC filters chose different features. Recall that the number of features is the same for LC and R (i.e. $\log_2(N)$ where N is the number of features) since they are both univariate filter techniques. We can conclude that the use of different filters results in different subset of features which, therefore, impact the accuracy of the single and ensemble techniques. The common features between the three filters techniques are indicated in bold in Table 2.

5.2 Selection of Best Single Techniques

This subsection presents the evaluations results of the four single techniques using the three filters over the six datasets. The best variant of each single technique with each filter was selected as a base technique for the heterogeneous ensembles ENF. Therefore, for each dataset, 12 best variants were selected (12 = 4 single techniques * 3 filters).
Step a.1 aims at building the single four ML techniques using the three filters. Thus, several experiments were performed by varying the parameter settings of each technique over each dataset according to Table 1. For each dataset and each single technique using a filter, we retain the variant having the lowest MAE value (i.e. best variant). Next, step a.2 consists of evaluating the reasonability of the best single techniques of step a.1 over each dataset by means of SA and Effect Size in order to select the ones that will participate in the further experiments. We select the best single techniques having an SA value higher than the 5% quantile of random guessing and a large impact over random guessing (Δ>0.8). The overall results suggest that the SA values of all single techniques are greater than the 5% quantile of random guessing; thus, all techniques provide reasonable predictions and are selected for the further experiments. The evaluation results are not presented due to the limit number of pages and could be obtained upon request by email to the corresponding authors of this paper.

Thereafter, step a.3 clusters the three variants of each single technique over each dataset through the SK test using the AE criterion. The purpose of this step is to select the variants that have the same predictive capability and do not have a significant difference between them. Afterward, the variants of techniques that belong to the best cluster were selected to participate in further experiments. In fact, the SK test identified one cluster 14 times (i.e. the three variants of the technique have the same capability prediction), two clusters 9 times, and 3 clusters once (i.e. Knn technique in COCOMO81 dataset).

The selected techniques of step a.3 were, thereafter, ranked using Borda with 8 performance measures based on the ranking obtained through Borda counting voting system no filter outperformed the other in all situations. For example, the LC filter was the best for DTs since LC was ranked first in five datasets; however, the R filter was the best for SVR (ranked 3 times in the first position).

The best variant of each technique over each dataset was therefore selected as a member of the proposed heterogeneous ensembles ENF. Table 3 lists the ENF ensemble members for each dataset. We observe that each dataset has an ensemble with different filters (e.g. ensemble of Albrecht dataset uses LC and R), This means that the performance of a filter depends on the characteristics of each dataset (size, number of features, etc.). Hence, members of ENF ensembles use different feature subsets, contrary to E1F ensembles, which can lead to satisfy the diversity criterion.

5.3 Ensembles Evaluation

This subsection presents the evaluation of the heterogeneous ensembles ENF, E1F and E0F according to steps c.1-c.3. We have in total for each dataset 15 ensembles (15 = 3 ENF ensembles (1 ensemble * 3 combiners) + 9 E1F ensembles (3 filters * 3 combiners) + 3 E0F ensembles (1 ensemble * 3 combiners). Step c.1 evaluates the SA and Δ values of the 15 heterogeneous ensembles over the six datasets in order to retain the ensembles that achieve SA values higher than the 5% quantile of random guessing and show a large improvement over random guessing (Δ>0.8). The results obtained show that all the ensembles generate better SA value than the 5% quantile of random guessing and show large Δ values; therefore, all the ensembles were selected as participants in the next experiments. The main findings are:

- There is no best ensemble that achieved the highest reasonability across all datasets.
- The E0F ensembles achieved the highest SA values in four datasets: Albrecht, China, Desharnais, and Miyazaki.
- The ENF ensembles generate the highest SA value in two datasets: COCOMO81 and ISBSG.
- None of the E1F ensembles was ranked at the first position in all datasets.
- The less reasonable ensembles were the ones of the E1F ensembles.
- IR and ME rules lead ensembles to generate better SA values.

Table 2: Feature Selection Results: Common Selected Features are in Bold for Each Dataset.

Datasets	CFS	LC	RReliefF
Albrecht	**Output**, Inquiry, **RawFP-counts**	**Output**, file, **RawFPcounts**, AdjFP	**Output**, Inquiry, **RawFP-counts**, AdjFP
China	Output, Enquiry, Interface, Added, Resource, Duration	AFP, Input, File, Added	Output, Enquiry, File, Duration
COCOMO81	DATA, **TIME**, STOR, TURN, VEXP, **KDSI**	DATA, **TIME**, STOR, TURN, **KDSI**	**TIME**, VIRTmajeur, PCAP, VEXP, **KDSI**
Desharnais	TeamExp, ManagerExp, YearEnd, Length, Adjustment, **PointsAjust**	Length, Transactions, PointsNonAdjust, **PointsAjust**	ManagerExp, Transactions, PointsNonAdjust, **PointsAjust**
ISBSG	VAF, MTS, UBCU, **FC**	VAF, MTS, IC, **FC**	IC, OC, EC, **FC**
Miyazaki	**KLOC**, SCRN, FORM, FILE, EFORM, EFILE	**KLOC**, SCRN, FILE, EFILE	**KLOC**, FORM, ESCRN, EFORM

Table 3: Members of the ENF Heterogeneous Ensembles.

Albrecht	China	COCOMO81	Desharnais	ISBSG	Miyazaki
LCDT	LCDT	RDT	LCDT	LCDT	LCDT
LCKnn	RKnn	CFSKnn	LCKnn	CFSKnn	LCKnn
RMLP	RMLP	LCMLP	LCMLP	CFSMLP	CFSMLP
RSVR	RSVR	RSVR	CFSSVR	LCSVR	LCSVR

Step c.2 clusters the 15 heterogeneous ensembles through the SK test with the purpose of selecting the ones that are the best and share similar predictive capability. The SK test identified 4 clusters in COCOMO81 dataset, 2 clusters in ISBSG and Miyazaki datasets, and one cluster in the three remaining datasets. We notice that the RReliefF ensembles were not selected by the SK test in two datasets: ISBSG and Miyazaki. Similarly, the CFS ensembles were not selected in the best cluster in COCOMO81 dataset. The results of SK test are not presented due to the limit number of pages and could be obtained upon request by email to the corresponding authors of this paper.

The results obtained from the ranking provided by Borda count are:

- The ENF ensembles, regardless of the combination rules, outperformed the E1F and E0F ensemble in 5 out of 6 datasets.
- None of the ENF ensembles was ranked in the last positions in all datasets.
- The three E0F ensembles were ranked in the three first positions in Desharnais dataset.
- Most of the E1F ensembles were in general ranked in the last positions in all datasets.
- The IR combiner provides more accurate ensembles in 4 out of 6 datasets, followed by ME in 2 out of 6 datasets. Note that the AV combiner did not occur in the first position in any dataset.

Note that, the Table presenting the final ranking is not presented due to the limit number of pages and could be obtained upon request by email to the corresponding authors of this paper.

6 CONCLUSION AND FUTURE WORK

The objective of this study was to evaluate the impact of the diversity criterion on the accuracy of heterogeneous ensembles. The ensemble members were the four ML techniques: KNN, MLP, SVR and DTs and the combination rules were the three linear combiners (average, median, and IRWM). In general, there are three sources of diversity: sampling data, training the same technique with different configuration in the same sample, and using different features as input of a technique.

This study investigated filter feature selection as a source of diversity of ensemble members. To do that, we improved the selection process of ensemble members used in (Hosni et al., 2017a) by allowing the use of different filters in the same ensemble. This led single technique of an ensemble to use different subsets of features. To assess the impact of this strategy, we evaluated and compared the accuracy of three groups of ensembles: ENF (members of an ensemble use different filters), E1F (members of an ensemble use one filter) and E0F (members of an ensemble do not use feature selection).

The results in terms of SA suggest that all ensembles ENF, E1F and E0F were reasonable and generate

more reasonable results with respect to random guessing. Moreover, none of the 15 ensembles was ranked in the first position across different datasets. The E0F ensembles were more reasonable than the other ensembles in four datasets; the ENF ensembles were the best in two datasets (COCOMO81 and Miyazaki). However, the E1F ensembles were the less reasonable in all datasets.

However, the accuracy results in terms of 8 performance measures suggest that the ENF, in particular with the combiners IR or ME, outperformed the E1F and E0F ensembles in 5 out of 6 datasets. This implies that using different feature subsets by ensemble members can lead to more accurate estimations than when members use the same feature subset or all the available features. In fact, the success of the ENF ensembles is mainly due to the fact that their members were diverse and generate different estimations at the same point (i.e. diversity) than the members of E1F or E0F. Moreover, E0F ensembles generate slightly better estimates than E1F. Therefore, we conclude that ensembles without feature selection were better and easier to construct than ensembles with one filter.

Ongoing work will focus on investigating the impact of other feature selection techniques, including filters or wrappers, on the accuracy of homogenous and heterogeneous ensembles.

REFERENCES

Azzeh, M., Nassif, A. B., and Minku, L. L. (2015). An empirical evaluation of ensemble adjustment methods for analogy-based effort estimation. *The Journal of Systems and Software*, 103:36–52.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1):155–159.

Hira, Z. M. and Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, 2015(1).

Hosni, M. and Idri, A. (2017). Software effort estimation using classical analogy ensembles based on random subspace. In *Proceedings of the ACM Symposium on Applied Computing*, volume Part F1280.

Hosni, M., Idri, A., and Abran, A. (2017a). Investigating heterogeneous ensembles with filter feature selection for software effort estimation. In *ACM International Conference Proceeding Series*, volume Part F1319.

Hosni, M., Idri, A., Abran, A., and Nassif, A. B. (2017b). On the value of parameter tuning in heterogeneous ensembles effort estimation.

Hosni, M., Idri, A., Nassif, A., and Abran, A. (2017c). Heterogeneous Ensembles for Software Development Effort Estimation. In *Proceedings - 2016 3rd International Conference on Soft Computing and Machine Intelligence, ISCMI 2016*.

Idri, A., Abnane, I., and Abran, A. (2017). Evaluating Pred(p) and standardized accuracy criteria in software development effort estimation. *Journal of Software: Evolution and Process*, (September):e1925.

Idri, A. and Cherradi, S. (2016). Improving Effort Estimation of Fuzzy Analogy using Feature Subset Selection. In *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*.

Idri, A., Hosni, M., and Abran, A. (2016a). Improved Estimation of Software Development Effort Using Classical and Fuzzy Analogy Ensembles. *Applied Soft Computing*.

Idri, A., Hosni, M., and Abran, A. (2016b). Systematic Mapping Study of Ensemble Effort Estimation. In *Proceedings of the 11th International Conference on Evaluation of Novel Software Approaches to Software Engineering*, number Enase, pages 132–139.

Idri, A., Hosni, M., and Alain, A. (2016c). Systematic Literature Review of Ensemble Effort Estimation. *Journal of Systems and Software*, 118:151–175.

Jovic, A., Brkic, K., and Bogunovic, N. (2015). A review of feature selection methods with applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, number May, pages 1200–1205.

Khoshgoftaar, T., Golawala, M., and Hulse, J. V. (2007). An Empirical Study of Learning from Imbalanced Data Using Random Forest. In *19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007)*, volume 2, pages 310–317.

Kocaguneli, E., Kultur, Y., and Bener, A. (2009). Combining Multiple Learners Induced on Multiple Datasets for Software Effort Prediction. In *Proceedings of International Symposium on Software Reliability Engineering*.

Minku, L. L. and Yao, X. (2013). Ensembles and locality: Insight on improving software effort estimation. *Information and Software Technology*, 55(8):1512–1528.

Scott, A. J. and Knott, M. (1974). A Cluster Analysis Method for Grouping Means in the Analysis of Variance. *Biometrics*, 30(3):507–512.

Shepperd, M. and MacDonell, S. (2012). Evaluating prediction systems in software project estimation. *Information and Software Technology*, 54(8):820–827.

Song, L., Minku, L. L., and Yao, X. (2013). The impact of parameter tuning on software effort estimation using learning machines. In *Proceedings of the 9th International Conference on Predictive Models in Software Engineering*.

Zhou, Z.-H. (2012). *Ensemble Methods*. CRC Press.

Zhu, D. (2010). A hybrid approach for efficient ensembles. *Decision Support Systems*, 48(3):480–487.