Web-of-Service Software Reusability Prediction using Heterogenous Ensemble Classifier

Prakash V. Parande, M. K. Banga

Abstract: WoS computing environment is expected to have numerous parallel computing engines. Presently, software professionals or developers often want to reuse existing software components to exhibit a task with time-efficient and cost effective solutions. However, software component reusability in uncontrolled manner leads to failure, pre-mature shutdown and software smells or aging. This paper develops a novel evolutionary computing assisted ensemble classification system for WoS software reusability prediction. This applies different base learners such as Naïve Bayes (NB), Linear Regression (LR), Decision Tress (DT), Logarithmic Regression (LOGR), and Support Vector Machine (SVM). Multivariate Adaptive Regression Spline (MARS). Once training the base learners, the outputs of each classifier have been processed with majority vote. The computation in conjunction with weighted sum enabled final labelling of each software class. The performance results affirmed that the present work ensemble classifier has better performance with respect to base classifiers.

Keywords: Ensemble classifier; heterogeneous ensemble; web-of-service; software reusability prediction.

1. INTRODUCTION

The vital roles and high ability makes software an inevitable entry in modern scenario. The software component signifies an on-demand section of software application in the form of set of codes and standardized parts, classes, and tests cases etc [1]. The process of reusing existing functions to achieve efficient software solution for computational time and resource cost is called software reusability [2]. However, the excessive or uncontrolled reuse of software component could lead faults or aging related symptoms, sometimes called software smells [2][3]. In such cases assessing reusability extant of each software component can help avoiding aging or fault like events [4]. The work considers Object Oriented Programming (OOP) concept with software metrics such as Coupling between Object (CBO), Depth of Inheritance Tree (DIT), and Weighted Method per Class (WMC), Line of Code (LOC), Number of Children (NOC), and Lack of Cohesion in Methods (LCOM) [5-8]. In ensemble of classifiers, single decisions are merged to categorize new examples which are the base/weak classifiers [9][10]. Ensembles are familiar to lower the risk of opting the unethical model by clustering all candidates [9][10][13]. In feature-level fusion, attributes are entangling from multiple event info or data sets, and merged into an individual appended feature vector[11]. Brooks et al., [12] described feature-level fusion as ahighoptionfor correlated data and DLF is uncorrelated thus it’s a better option. Performing training of each base learner over data subsets, the ensemble has been prepared using majority vote and weighted sum method. Performing two-class classification each class has been predicted as REUABLE and NON-REUSABLE. The overall proposed system has been developed using MATLAB 2017a software where the performance analysis affirmed that the proposed ensemble concept exhibits better performance than the classical base learners.

2. RELATED WORK

Singhani et al. [15] examined various factors influencing inspection of OOP software for an Analytical Hierarchical Process (AHP) model. OO-CK metrics can achieve software re-usability estimation and quality characterization [16-18]. Definition of threshold values for each CK metrics can enable fault resilient reusability prediction [19, 20]. In [21, 22] CK metrics is conceptualized to identify minimal OOP parameters of software design and reusability was predicted in each class. Zahara et al. [23] used Machine learning regression algorithms to perform reusability prediction for independent variables characterizing reuse-proneness of the classes in OOP software metrics. Torkamani [24] explored the possibility of other OOP-software metrics to perform reusability prediction. Similarly, it is also recommended OOP metrics for suitable software design[25]. There is a derived metrics of software to examine quality features of software parts in customizability, reusability and complexity[26]. They applied interface mechanism to estimate Component Reusability (CR) and Component Ruse Level (CRL) to compute the reuse proneness level. Line of Code (LOC) metric as the

Revised Manuscript Received on July 25, 2019.
Prakash v. Parande,
Research Scholar, School of Computing & Information Technology,
Reva University, Bangalore-560064, India, prakashvp2010@gmail.com
Dr. M. K. Banga,
Ex - Director of School of Computing and information technology,
Reva University, Bangalore 560 064, Professor, Dept of CSE, DayanandaSagar University, Bangalore 560 068

International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075, Volume-8, Issue-9S3, July 2019

Retrieval Number: I32810789S319/2019/BEJIESP
DOI: 10.35940/ijitee.i32810789S319

1276 Published By:
Blue Eyes Intelligence Engineering & Sciences Publication
indicator for IRL as well as reuse proneness indicator. J. Kittler [27] stated that the use of classifier ensembles with decision-level fusion can be significant for multi-class classification. Polikar et al. [28] designed an ensemble of classifiers by exploiting data from different sources and combined the base learners by means of a weighted voting scheme. AdaBoost [29] involves subsampling the training examples which has been applied in numerous classification problems [30]. However, their efficacy for highly correlated features or data seems confined. To achieve higher reliability of prediction, ensemble approach was recommended in [31, 32]. Benediktsson et al., [33] too affirmed suitability of ensemble classifier for multi-class classification, especially derived using machine learning concept. To further augment performance prediction fusion was also recommended in [34, 35]. It found that the accuracy of ensemble depends on the learner fusion for which easy averaging (i.e., maximum vote) can be a potential approach [36]. It elaborated that empirically verified ensembles are generalize well [37]. Briem et al. [38] used different classifiers on multiple source data having imagery of multiple spectral, Synthetic Aperture Radar (SAR) data and graphic topological information [39, 40]. Different approaches for ensemble formation were recommended in [41]. It is recommended to have predominant methods like various hybrid strategies, classifier models, feature subsets and training set based ensemble. However lacks efficacy with multi-class classification with highly correlated features. To further augment it, integrated boosting and bagging is used to create fuzzy ensemble classifiers. Different feature sets with Principle Component Analysis (PCA) based feature selection was used in [42] to perform Rotation Forest ensemble classification. Rotation forest was used in [43] to perform to class classifications for the cancer datasets. Later, AdaBoost and Rotation Forest was combined together to develop a new ensemble RotBoost [44], where it can perform better than the classical base learners. Rotation Forest and IDE [45] was combined in [46]. SVM ensemble was recommended in [47] for spectrogram classification. As enhanced solution selection based SVM ensemble driven with clonal immunity algorithm is proposed in [48]. An ensemble NN was proposed in [49]. Partalas et al. [50] developed the preferred ensemble concept based on Focused Ensemble Selection (FES) with diversity classifiers. amalgamating different base learners to derive an ensemble classifier has exhibited better performance than the base learner alone [51, 52]. Moderated Asymmetric Naïve Bayes classifiers (MANB) [53] homogeneous ensemble settings have a base classifier. An ensemble pair approach is presented in [54] which achieve better performance with respect to bagging, ensemble approaches, and ECOC. In [55], Dempster's rule of combination is used to hybridk-NN, Rocchio classifiers and SVM, and a better performance is observed when compared with base classifiers.

### 3. PROBLEM FORMULATION

Considering the significance of a robust software reusability estimation system, this paper focuses on developing an “Ensemble Classifier” algorithm with OO-software metrics and classes without imposing aging or smells. “Ensemble Classifiers” have been assessed for small readily available benchmark datasets such as Iris, Spambase, and Satellite etc; however, efficacy of such algorithms with large scale and real-time applications environment often remains suspicious. Thus it used 100 software modules developed using Java programming language with OO-programming concept using online software repository named “Sourceforge”.

In proposed model DLF approach with majority vote scheme has been implemented and best base classifiers propagated to the Ensemble Classifier for augmentation of the reliability and accuracy of the reusability prediction system. The snippet of the proposed ensemble classification model based WoS software reusability prediction in Fig. 1.

---

#### 4. Ensemble Classifier Based Web of Service Software Reusability Prediction

The base learners and allied ensemble formation approach discusses the following:

### 4.1. WoS software reusability data preparation

100 OOP based software have consider WoS software, to convert into Java file using WSDL tool (Fig. 2). WSDL signifies the XML-based interface definition language characterizes web services i.e., $P = \{M_0(l_0, O_0), M_1(l_1, O_1), \ldots, M_n(l_n, O_n)\}$, which can performs various functions $M_i$ by transferring input $l_i$ into output $O_i$. Table I shows the plotting of the various components of Java class.
4.2. Data normalization and outlier analysis

Here, each data element $x_i$ of the class $X$ is mapped to the allied normalized value $x_i'$ in the range of $[0, 1]$. Thus, the normalized value $x_i'$ is obtained as (4.1).

$$Normalized(x_i) = x_i' = \frac{x_i - \min(X)}{\max(X) - \min(X)}$$

(4.1)

4.3. Feature selection

A snippet of the univariate logistics regression (ULR) algorithm is depicted as follows:

4.3.1. Uni-variate logistic regression (ulr) assisted metrics characterization

ULR states the level of significance of each metrics towards reusability estimation of each class.

$$logit[\pi(x)] = \alpha_0 + \alpha_1x$$

(4.2)

$logit[\pi(x)]$ and $x$ state the dependent and independent variables, correspondingly. The variable $\pi$ states the probability factor of significance.

$$\pi(x) = \frac{e^{\alpha_0 + \alpha_1x}}{1 + e^{\alpha_0 + \alpha_1x}}$$

(4.3)

4.3.2. Rough Set Analysis (RSA) based feature extraction

RSA can be obtained in the form of classification results or reduced data elements that enable swift and precise computation. The detailed of the RSA is given as follows:

- **Step-1: Feature Set Selection**
  In this step, extracted CK suite metrics for each class of the software are collected.

- **Step-2: Feature set (Data) Discretization**
  Here, the extracted data is discretized using K-means clustering algorithm.

- **Step-3: Lower/Upper Approximation for all feasible data sets**

It estimates the lower and the upper values as the union of all comprising feature sets (Step-2).

$$\overline{y}X = \{x_i \in U|x_i|_{ind(B)} \subseteq X\}$$

(4.4)

Upper approximation value signifies the union of all the encompassing sets with unit non-zero intersection with $X$.

$$\overline{y}X = \{x_i \in U|x_i|_{ind(B)} \cap X \neq \emptyset\}$$

(4.5)

- **Step-4: Accuracy estimation for the considered feature sets**
  Here, we have retrieved a factor stating accuracy of $X$ in $B \subseteq A$ using (4.6).

$$\mu_B = \frac{Card(BX)}{Card(B)}$$

(4.6)

Cardinality of a set signifies the lower or upper approximation of $X$.

- **Step-5: Approximated (Feature Set) Selection**
  All possible sets are selected with equality of accuracy of the universal set. Later, the obtained data set with optimal cardinal possibility are chosen as the minimal set to be used for “Ensemble Learning and Classification”.

4.4. Ensemble learning for software reusability prediction

Base learners have been trained over the data subsets whose outputs have been combined together to achieve better accuracy.

4.4.1. Decision tree algorithm

C5.0 classifier [51] used as a base classifier that exhibits recursive partitioning over the extracted OOP-metrics.

4.4.2. Enhanced kNN Classifier

KNN classifier applies Euclidean distance to estimate inter-attribute distance using following expression:

$$D(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

(4.7)

Where $p$ and $q$ are subjected to be compared with $n$ features.

4.4.3. Logistic regression

LR algorithm constitutes a prediction scheme to examines reuse proneness of a class.

$$logit[\pi(x)] = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_mx_m$$

(4.8)

$logit[\pi(x)]$ and $x_i$ signifies the dependent and independent variables, correspondingly. Practically, $\pi(x)$ vary from 0 to 1 to $-\infty$ to $+\infty$. LR outcome $\pi(x)$ is stated as

$$\pi(x) = \frac{e^{\beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_mx_m}}{1 + e^{\beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_mx_m}}$$

(4.9)
4.4.4. Polynomial regression

Univariate Polynomial Regression (UPR) for the \( n \)th polynomial order:

\[
Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_n X^n
\]  

(4.10)

where variables \( X \) and \( Y \) states for the independent and the dependent variables and \( \beta_0, \beta_1, \ldots, \beta_n \) signify the constant and coefficients, respectively.

\[
Y = \beta_0 + \beta_1 X + \beta_2 X^2
\]  

(4.11)

For multivariate 2\textsuperscript{nd} order PR analysis, it is estimated on the basis of the two variables,

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_1 X_1^2 + \beta_2 X_2^2
\]  

(4.12)

4.4.5. Naive bayes classification

NB allocates specified object \( x \) to class \( e^* = \text{argmax}_q(P(d|x)) \) using Bayes' rule.

\[
P(d|x) = \frac{P(x|d)P(d)}{P(x)}
\]  

(4.13)

\( P(d) \) signifies the probability of the parameter \( c \) before observing the data. The factor \( P(d|x) \) refers the likelihood of data \( x \).

\[
P(x|d) = \prod_{i=1}^{m} P(x_i|d)
\]  

(4.14)

4.4.6. Support Vector Machine (SVM)

To perform prediction, SVM model applies the function derived in (4.15).

\[
Y' = w * \phi(x) + b
\]  

(4.15)

\( \phi(x) \) states the non linear transform and \( Y' \) is retrieved by reducing the risk of regression.

\[
R_{\text{reg}}(Y') = C * \sum_{i=0}^{1} Y_i' - Y_i + \frac{1}{2} ||w||^2
\]  

(4.16)

The parameter \( C \) states the cost function, while \( C \) refers the penalties for error estimation.

\[
w = \sum_{j=1}^{l}(\alpha_j - \alpha_j^*) \phi(x_j)
\]  

(4.17)

The parameters \( \alpha \) and \( \alpha^* \) are Lagrange multipliers, whereby, \( \alpha^* \geq 0 \).

\[
Y' = \sum_{j=1}^{l}(\alpha_j - \alpha_j^*) \phi(x_j) * \phi(x) + b
\]  

(4.18)

\[= \sum_{j=1}^{l}(\alpha_j - \alpha_j^*) * K(x_j, x) + b\]

Where, \( K(x_j, x) \) states the kernel function.

4.4.7. Multivariate Adaptive Regression Splines (MARS)

MARS functions of “Divide-and-conquer” policy retrieved OO-CK metrics data, which is divided into the different regions with respective regression equation.

\[
Y = \sum_{i=1}^{m} C_i BF_i(X)
\]  

(4.19)

\( Y \) states dependent and \( X \) states independent variable. The parameter \( C_i \) states the fixed coefficients and \( BF_i(X) \) states the basis functions.

![Diagram of Proposed Ensemble Based Classifier]

Fig. 3. Proposed Ensemble Based Classifier

4.4.8. Ensemble structure

The maximum weighted computational class from all base classifiers is chosen as class of prediction (4.20):

\[
C(y) = \text{argmax}_{1 \leq j \leq K} \sum_{j=1}^{m} \alpha_j \cdot (f_j(y) = C)
\]  

(4.20)

\( \alpha_j \) signifies the weight of base classifiers, \( f_j(y) \) refers the reusability prediction output.

5. RESULTS AND DISCUSSION

The analysis of performance for the developed ensemble classification is as follows.
5.1. Performance Assessment

Relative performance is examined using estimating performance of each base learners and ensemble learners. Ensemble classification concept with DLF enables higher accuracy and minimum error.

5.1.1. Classification Accuracy Assessment

The assessment of the developed ensemble classification model (Table 2).

Table 2. Performance Parameter

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mathematical Expression</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>(\frac{(TN+TP)}{(TN+FN+FP+TP)})</td>
<td>Predicted fault prone module</td>
</tr>
</tbody>
</table>

The performance parameters and their respective values are given in Table 3.

Table 3. Performance values

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Accuracy(%)</th>
<th>Precision(%)</th>
<th>F-Measure(%)</th>
<th>Recall(%)</th>
<th>Specificity(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree</td>
<td>91.31</td>
<td>90.01</td>
<td>91.65</td>
<td>91.02</td>
<td>82.64</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>81.40</td>
<td>95.45</td>
<td>88.59</td>
<td>81.78</td>
<td>71.00</td>
</tr>
<tr>
<td>kNN</td>
<td>88.43</td>
<td>87.31</td>
<td>89.33</td>
<td>88.99</td>
<td>87.74</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>85.71</td>
<td>84.60</td>
<td>86.82</td>
<td>81.30</td>
<td>84.24</td>
</tr>
<tr>
<td>Polynomial Regression</td>
<td>89.01</td>
<td>90.48</td>
<td>92.31</td>
<td>89.95</td>
<td>91.04</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>87.91</td>
<td>90.32</td>
<td>91.73</td>
<td>90.16</td>
<td>88.04</td>
</tr>
<tr>
<td>SVM (Polynomial)</td>
<td>93.41</td>
<td>94.97</td>
<td>95.52</td>
<td>91.04</td>
<td>91.27</td>
</tr>
<tr>
<td>MARS</td>
<td>88.12</td>
<td>93.20</td>
<td>94.20</td>
<td>89.70</td>
<td>96.00</td>
</tr>
<tr>
<td>Proposed Ensemble</td>
<td>97.02</td>
<td>95.64</td>
<td>96.18</td>
<td>94.93</td>
<td>97.01</td>
</tr>
</tbody>
</table>

5.1.2. Error Profiling

Ensemble learning intends to reduce error of prediction by amalgamating different classifiers and their respective performances over the original feature data.

5.1.2.1 Mean absolute error (MAE)

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} (|Y_i' - Y_i|) \quad (5.1)
\]

where \(Y_i'\) refers the calculated output, while \(Y_i\) states for the expected value.

5.1.2.2 Mean magnitude of the relative error (MMRE)

\[
MMRE = \frac{1}{n} \sum_{i=1}^{n} \frac{|Y_i' - Y_i|}{Y_i} \quad (5.2)
\]

To deal with overflow conditions \((Y_i = 0)\), 0.01 is added to denominator. Thus, MMRE is

\[
MMRE = \frac{1}{n} \sum_{i=1}^{n} \frac{|Y_i' - Y_i|}{Y_i + 0.01} \quad (5.3)
\]

5.1.2.3 Standard Error of the Mean (SEM)

Mathematically, SEM is presented as (5.4).

\[
SEM = \frac{\sigma}{\sqrt{N}} \quad (5.4)
\]

where \(\sigma\) and \(N\) are standard deviation and total number of samples, respectively.

Table 4 Error performance by the different base learners and the ensemble classifier

<table>
<thead>
<tr>
<th>Techniques</th>
<th>MAE</th>
<th>MMRE</th>
<th>RMSE</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree</td>
<td>0.0990</td>
<td>0.7000</td>
<td>0.0200</td>
<td>0.0900</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.1557</td>
<td>0.8623</td>
<td>0.1209</td>
<td>0.0276</td>
</tr>
<tr>
<td>kNN</td>
<td>0.2102</td>
<td>0.7432</td>
<td>0.1953</td>
<td>0.1003</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.2390</td>
<td>0.8418</td>
<td>0.2052</td>
<td>0.1167</td>
</tr>
<tr>
<td>Polynomial Regression</td>
<td>0.1859</td>
<td>0.8016</td>
<td>0.0591</td>
<td>0.962</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.1968</td>
<td>0.8073</td>
<td>0.542</td>
<td>0.1331</td>
</tr>
<tr>
<td>SVM (Polynomial)</td>
<td>0.1904</td>
<td>0.8001</td>
<td>0.0245</td>
<td>0.1090</td>
</tr>
<tr>
<td>MARS</td>
<td>0.2040</td>
<td>0.8210</td>
<td>0.2002</td>
<td>0.0087</td>
</tr>
<tr>
<td>Proposed Ensemble</td>
<td>0.0027</td>
<td>0.0631</td>
<td>0.0091</td>
<td>0.0087</td>
</tr>
</tbody>
</table>

CONCLUSIONS

This paper proposes anensemble learning concept using machine learning. WoS data was obtained from 100 software developments using Object Oriented Programming (OOP) concept. The software data is processed with WSDL followed by Chidamber and Kamrer (CK) metrics estimation. The extracted CK-metrics data is processed for Min-Max normalization followed by outlier detection that avoided premature convergence issues significantly. Furthermore, RSA feature extraction technique enabled selection of only relevant and significant attributes related to software (class) reusability. Considering other base classifiers such as neural network and its variant, shows these approaches suffer from local minima and convergence, and therefore evolutionary computing based learning can be explored to alleviate...
such issues. This as a result could yield more efficient and reliable ensemble classification.

REFERENCES

11. G. Tsoumakas, L. Angelis, I. Vlahavas, Selective fusion of heterogeneous classifiers, Intelligent Data Analysis Vol. 9 (6), pp. 511–525, 2005
34. F. Roli, G. Giacinto, G. Vernazza, Methods for designing multiple classifier systems, Proceedings of the second international workshop on multiple classifier systems.