**Feature Selection Based on Improved Particle Swarm Optimization in Software Defect Prediction**

Afnan Alotaibi1, Nada Alsulaiman2, Nojoud Albadrani3, Sultan Alqahtani4, Wojdan Binsaeedan5

Computer Sciences Department, Imam Mohammad Ibn Saud Islamic University

Riyadh, Saudi Arabia

1atsalotaibi, 2nfsulaiman, 3nfalbadrani {@sm.imamu.edu.sa}, 4ssalqahtani, 5wmasaeedan{@imamu.edu.sa}

ABSTRACT

The manuscript should contain an abstract within 300 words. The manuscript should have a self-contained, citation-free abstract and state briefly the purpose of the research, methodology, key results and major conclusions. Abstract should be in a single paragraph with running sentences. Do not use any subheading or point list within the abstract. Also, non-standard or uncommon abbreviations should be avoided, but if essential they must be defined at their first mention in the abstract itself.

Keywords: Authors are advised to writes **3-5 keywords** related to the article, separated by comma. These keywords will be used for indexing purpose.

# Introduction

Software Defect Prediction (SDP) is one of the important processes in the software development lifecycle (SDLC) [1]. SDP is the process of identifying and controlling the percentage of defective program units in the software [2] [3]. Defect prediction will improve the testing process and the overall quality of the software [3]. In addition, it will decrease the time, cost, the effort required to rework on the defective units and deliver a more consistent and trusted software [2].

There are different categories of SDP, the most frequently used is the prediction of whether the software unit is defective or not [3], where the SDP is used as a binary classification problem that has two classes which are defect and non-defect [4]. In recent studies, SDP models were built using different Machine learning algorithms [5]. The accuracy of SDP models is highly affected by the quality of defect datasets used to build SDP models [6]. The defect dataset features or attributes are growing vastly as software systems are increasing; accordingly, these datasets may contain irrelevant and redundant features that affect the accuracy of classifiers [7] [6].

Feature selection (FS) is used to achieve the best classification performance by selecting the most informative feature subset from the dataset [8]. However, FS is an NP-Hard problem because the number of search space increases exponentially when the number of features is increased [7] [9]. Thus, metaheuristic search algorithms are used to find the optimal subset of features within less time [10]. It is random-based methods that generate random solutions and can reach a near-optimal solution [9]. Swarm Intelligence (SI) algorithms are one of the metaheuristic algorithms that are inspired by the behaviours of the creatures such as bees, birds, and fishes [9]. SI algorithms involve two main processes which are exploration and exploitation. Exploration means examine more global regions and find divers solution. In contrast, the exploitation is searching locally to enhance the quality of the solution [9]. Over exploration or over explotation can leads to losses the optimal solution, or premature convergence and suck in local minima [9]. Thus, achieving balance between exploration and exploitation is important for the optimization process.

# Background

## Software Defect Prediction (SDP)

## Feature Selection (FS)

## Classification Algorithms

### Support Vector Machines (SVM)

### K-Nearest Neighbor Classifier (KNN)

### Naive Bayes Classifier (NB)

## Overview of Particle Swarm Optimization (PSO)

PSO is an meta-heurstic technique inroduced by Kennedy and Eberhart [11] that was inspired by social behaviors seen in birds flocking and fish schooling. A swarm in PSO is made of multiple individuals known as particles who communicate through iterations to identify optimal solutions while traversing around the search space. PSO performs searches using a population (swarm) of individuals (particles) that are updated from iteration to iteration. The population size is denoted as psize. To find the optimal solution, each particle modifies its search direction based on two factors: its own best prior experience (pbest) and the best experience of all other members (gbest).

### Binary Particle Swarm Optimization for Feature Selection

Many optimization problems in a space are discrete, Binary particle swarm optimization (BPSO) has been proposed to solve these types of tasks [12]. Its another version of PSO that utilize the personal best (pbest) and global best (gbest) solutions to update the velocity and position[13]. The volicity will updated as following for each particle [13] [14] :

where is the velocity, is the inertia weight, and are the acceleration factors, and are two independent random numbers between 0 and 1, is the position of particle (solution), is the personal best solution, is the global best solution for the population, is the order of particle in the population, is the dimension of search space, and is the number of iterations. All the random numbers will be discussed in the experiments section. ~~Note that the velocity is bounded by the maximum velocity, vmax and minimum velocity, vmin. In this study, the vmax and vmin were set at 6 and −6, respectively [13]~~.

In BPSO, probabilities are used to determine the state of one bit, in other word, a particle moves in a state space constrained to zero and one on each dimension, for example if = 0.20, then will be near to zero more than one [12]. After the velocity is calculated using Equation (1) , it will converted into probability value using Equation (2),

The position of particle is updated as shown in Equation (3):

where is a random number between 0 and 1. In BPSO, pbest and gbest play an important role in guiding the particle to move toward the global optimum. Considering the minimization function was applied in this paper. Iteratively, the pbest and gbest are updated as follows:

# Related work

Several research papers in the literature investigate the SDP problem. However, the most recent studies used machine learning approaches to create an appropriate model for this complex challenge. Some studies used pure machine learning methods with no pre-processing on the input datasets [15]. In contrast, others used a combination of pre-processing methods on the datasets, such as feature selection to lower the dimensionality of input datasets or noise reduction from unbalanced datasets. This section will review the works that used the PSO for features selection to improve SDP.

Wahono and Suryana [16] presented a PSO-bagging approach combination to improve the performance of software defect prediction. PSO was employed to deal with feature selection, while the bagging approach was utilized to deal with class imbalance. In order to evaluate the suggested strategy, many machine learning classifiers were applied to nine datasets from NASA’s metric data repository. The AUC results showed that the proposed technique enhanced the prediction performance of the most commonly used ML classifiers. Wahono and Ahmad published similar work in [17], in which GA and PSO algorithms were used as feature selection strategies for the SFP problem, while the bagging methodology was used to deal with the class imbalance problem. To test various FS techniques, ten classifiers were used across nine NASA MDP datasets. The AUC results revealed that the proposed FS techniques resulted in a major improvement in prediction performance for the majority of the used classifiers. Furthermore, it has been determined that there is no substantial difference between PSO and GA in feature selection for the majority of classifiers.

Arora and Saha [1] introduced a software defect prediction model based on two classifiers, extreme learning machine (ELM) and kernel-based ELM (KELM). They employed five wrapper-based including PSO and seven filter-based feature selection approaches in their approach. They chose seven datasets from the PROMISE repository to test their method. Furthermore, they employed the accuracy metric to assess the performance feature selection model. They discovered that ELM-based classifiers worked better with wrapper-based feature selection approaches, but KELM classifiers performed better with filter-based methods.

Recently, Malhorta et al. [18] used the Synthetic Minority Oversampling Technique (SMOTE) along with FS using PSO on object-oriented metrics. First, the SMOTE were used to tackle the issue of imbalanced data, while the PSO was applied to extract the optimal feature set. Then, the chosen features were utilized for training the SVM to predict defects. The result showed that the oversampling technique of SMOTE when combined with PSO for feature selection, can be used for building efficient software defect production models.

As shown by past literature, several research publications examine the SDP problem. However, in reality, each project has its characteristics. Thus, it is critical to design a strong model that evaluates the acquired data early in the process. This encourages us to improve the PSO algorithm to use it as a feature selection algorithm as the initial step toward achieving a high-quality classifier.

# Datasets Specifications

We experimented the proposed method on different benchmarked datasets extracted from the PROMISE repositoryand NASA repository. The datasets were collected from real and open-source software projects. Eleven of these datasets are downloaded from the NASA corpus[[1]](#footnote-2) (cleaned versions by [19]), while the remaining datasets are from the PROMISE software engineering corpus[[2]](#footnote-3). These projects have different specifications, such as the programming language, the code size, and software measures, Detailed information on the datasets is listed in Table 1. PROMISE dataset has 20 static attributes and one dependent variable. for each software project. On the other hand, NASA datasets have different number of attributes for each software project. The datasets consist of a set of features that have values and a goal field that describes the instance as defect or non-defect. These features describe the program from different sides including the lines of code measure (program length, count of lines of comments, count of lines of comments), McCabe metrics, base Halstead measures, derived Halstead measures, unique operators, unique operands, total operators, total operands, cyclomatic complexity, essential complexity, design complexity, and a branch-count. . We choose PROMISE and NASA because these datasets are commonly used in SDP field.

The approach for applying training and testing in the experiments is based on a hold-out strategy in which each dataset is randomly divided into 80% for training and 20% for testing. Note that, that there is a pre-processing step in cross-project experiment that converts the defect attribute into a binary before it is used as the dependent variable.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: *Datasets description.* | | | | | | | | | | |
|  | No. | Project | | Language | | Features | Instances | Defects | Non- Defects | Defect ratio | |
| PROMISE | D1 | | ant-1.6 | | Java | 21 | 351 | 92 | 259 | 26.21 | |
| D2 | | ant-1.7 | | Java | 21 | 745 | 166 | 579 | 22.28 | |
| D3 | | camel-1.4 | | Java | 21 | 872 | 145 | 727 | 16.63 | |
| D4 | | camel-1.6 | | Java | 21 | 965 | 188 | 777 | 19.48 | |
| D5 | | ivy-2.0 | | Java | 21 | 352 | 40 | 312 | 11.63 | |
| D6 | | jedit-4.2 | | Java | 21 | 367 | 48 | 319 | 13.08 | |
| D7 | | jedit-4.3 | | Java | 21 | 492 | 11 | 481 | 2.24 | |
| D8 | | poi-2.0 | | Java | 21 | 314 | 37 | 277 | 11.78 | |
| D9 | | prop-6 | | Java | 21 | 660 | 66 | 594 | 10.00 | |
| D10 | | synapse-1.2 | | Java | 21 | 256 | 86 | 170 | 33.59 | |
| D11 | | xalan-2.5 | | Java | 21 | 803 | 387 | 416 | 48.19 | |
| D12 | | xerces-1.2 | | Java | 21 | 440 | 71 | 369 | 16.14 | |
| NASA | D13 | | cm1 | | C | 38 | 327 | 42 | 285 | 12.8 | |
| D14 | | jm1 | | C | 22 | 7782 | 1672 | 6110 | 21.5 | |
| D15 | | kc1 | | C++ | 22 | 1183 | 314 | 869 | 26.5 | |
| D16 | | kc3 | | Java | 40 | 194 | 36 | 158 | 18.6 | |
| D17 | | mc1 | | C & C++ | 39 | 1988 | 46 | 1942 | 2.3 | |
| D18 | | mw1 | | C | 38 | 253 | 27 | 226 | 10.7 | |
| D19 | | pc1 | | C | 38 | 705 | 61 | 644 | 8.7 | |
| D20 | | pc2 | | C | 37 | 745 | 16 | 729 | 2.1 | |
| D21 | | pc3 | | C | 38 | 1077 | 134 | 943 | 12.4 | |
| D22 | | pc4 | | C | 38 | 1287 | 177 | 1110 | 13.8 | |
| D23 | | pc5 | | C++ | 39 | 1711 | 471 | 1240 | 27.5 | |

# Methodology

For feature selection, we used the improved binary particle swarm optimization proposed by Nguyen et al. [] which is Sticky Binary Particle Swarm Optimization (SBPSO). The algorithm used the previous position to update the current particle position by flipping the entry value to zero or one as follows.

(1)

Where represents the iteration number. The SBPSO uses flipping probability instead of velocity. The flipping probability is expressed as follow:

(2)

, is the weight of the sickness, personal best , and global best respectively. Where + . The sickness value indicates how much the position entry will stick with its value, hence the particle will search around. is set to high when the entry just flipped then decrease over a fixed number of steps until the entry flipped or become zero as follow:

(3)

They assume that is the ratio between , . Their value can be expressed as follows:

, (4)

Therefore, the flipping probability will become:

(5)

The following paragraphs illustrate each case:

* The first case: the second and third terms of equation (2) will be zero.
* The second case: the second term in (2) will be zero, and since and are two different binary numbers the absolute value of their difference is one, hence the third term value is then they substitute it by its value in equation (4).
* The third case: the third term will be zero and the second term is . Then they substitute it by its value in equation (4).
* Last case: . The second term is and the third term is . . Hence, the result equation is presented in case 4.

Now, the Flipping probability depends on two parameters and that control the sickness. is a fixed value that is set before starting the algorithm. and are dynamically updated for each iteration as follow:

Where T is the total number of iteration, t is the current iteration U is upper bound, L is lower bound.

Table Parameter setting of BPSO, SBPSO

|  |  |  |  |
| --- | --- | --- | --- |
| Parameters | Without FS | BPSO | Proposed method SBPSO |
| Population size, N |  |  |  |
| Maximum number of iterations, Tmax |  |  |  |
| Number of runs |  |  |  |
| w |  |  |  |
| c1 |  |  |  |
| c2 |  |  |  |
| vmax |  |  |  |
| vmin |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Results and Discussion

## Model Evaluation Metrics

We employed a binary classification strategy to identify classes that are likely to have defects. A binary classifier can make two types of errors: false positives (FP) and false negatives (FN). A correctly categorized defected class is also a true positive (TP), whereas a correctly classified non-defected class is a true negative (TN). We examined classification outcomes using the following metrics: Accuracy, Precision, Recall, and F-measure, described as follows:

* **Accuracy:** address the number or percentage of correctly classified instances to the total sum of instances. It can be calculated by Equation (1):

(1)

* **Precision:** Addresses the percentage of successfully classified defective occurrences among the total number of retrieved instances. The best precision value is 1. The higher the precision is, the fewer false positives. It can be calculated by Equation (2):

(2)

* **Recall:** Addresses The percentage of defective cases that were accurately classified. The best recall value is 1. The higher the recall is, the lower the number of false negatives. It can be calculated by Equation (3):

(3)

* **F-measure:** computes accuracy by taking Precision and Recall into account, which may be understood as a weighted average of Precision and Recall. The F-measure value varies between 0 and 1, with values closer to 1 indicating better classification results. It can be calculated by Equation (4):

(4)

## Analytical Description of the Relevant Features

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 2:**  *Relevant features in software datasets.* | | | | | |
| **Project** | **All Features** | **Selected Features** | **FFR%** | **Relevant Features** |
| ant-1.6 | 21 |  |  |  |
| ant-1.7 | 21 |  |  |  |
| camel-1.4 | 21 |  |  |  |
| camel-1.6 | 21 |  |  |  |
| ivy-2.0 | 21 |  |  |  |
| jedit-4.2 | 21 |  |  |  |
| jedit-4.3 | 21 |  |  |  |
| poi-2.0 | 21 |  |  |  |
| prop-6 | 21 |  |  |  |
| synapse-1.2 | 21 |  |  |  |
| xalan-2.5 | 21 |  |  |  |
| xerces-1.2 | 21 |  |  |  |
| cm1 | 38 |  |  |  |
| jm1 | 22 |  |  |  |
| kc1 | 22 |  |  |  |
| kc3 | 40 |  |  |  |
| mc1 | 39 |  |  |  |
| mw1 | 38 |  |  |  |
| pc1 | 38 |  |  |  |
| pc2 | 37 |  |  |  |
| pc3 | 38 |  |  |  |
| pc4 | 38 |  |  |  |
| pc5 | 39 |  |  |  |

# Conclusions

# References

[1] I. Arora and A. Saha, “ELM and KELM based software defect prediction using feature selection techniques,” *Journal of Information and Optimization Sciences*, vol. 40, no. 5, pp. 1025–1045, Jul. 2019, doi: 10.1080/02522667.2019.1637999.

[2] M. Anbu and G. S. Anandha Mala, “Feature selection using firefly algorithm in software defect prediction,” *Cluster Comput*, vol. 22, no. 5, pp. 10925–10934, Sep. 2019, doi: 10.1007/s10586-017-1235-3.

[3] R. A. Khurma, H. Alsawalqah, I. Aljarah, M. A. Elaziz, and R. Damaševičius, “An Enhanced Evolutionary Software Defect Prediction Method Using Island Moth Flame Optimization,” *Mathematics*, vol. 9, no. 15, Art. no. 15, Jan. 2021, doi: 10.3390/math9151722.

[4] R. Wahono, “A Systematic Literature Review of Software Defect Prediction: Research Trends, Datasets, Methods and Frameworks,” *Journal of Software Engineering*, vol. 1, May 2015.

[5] D. Bowes, T. Hall, and J. Petrić, “Software defect prediction: do different classifiers find the same defects?,” *Software Quality Journal*, vol. 26, Jun. 2018, doi: 10.1007/s11219-016-9353-3.

[6] A. O. Balogun *et al.*, “Search-Based Wrapper Feature Selection Methods in Software Defect Prediction: An Empirical Analysis,” *Advances in Intelligent Systems and Computing*, vol. 1224 A, pp. 492–503, 2020.

[7] S. Goyal and P. K. Bhatia, “Software fault prediction using lion optimization algorithm,” *Int. j. inf. tecnol.*, Sep. 2021, doi: 10.1007/s41870-021-00804-w.

[8] M. Kondo, C. P. Bezemer, Y. Kamei, A. E. Hassan, and O. Mizuno, “The impact of feature reduction techniques on defect prediction models,” *Empirical Software Engineering*, vol. 24, no. 4, pp. 1925–1963, Aug. 2019, doi: 10.1007/s10664-018-9679-5.

[9] R. Abu Khurma, I. Aljarah, and A. Sharieh, “A Simultaneous Moth Flame Optimizer Feature Selection Approach Based on Levy Flight and Selection Operators for Medical Diagnosis,” *ARABIAN JOURNAL FOR SCIENCE AND ENGINEERING*, Mar. 2021, doi: 10.1007/s13369-021-05478-x.

[10] S. Hosseini, B. Turhan, and M. Mäntylä, “A benchmark study on the effectiveness of search-based data selection and feature selection for cross project defect prediction,” *Information and Software Technology*, vol. 95, pp. 296–312, Mar. 2018, doi: 10.1016/j.infsof.2017.06.004.

[11] J. Kennedy and R. Eberhart, “Particle swarm optimization,” in *Proceedings of ICNN’95 - International Conference on Neural Networks*, Perth, WA, Australia, 1995, vol. 4, pp. 1942–1948. doi: 10.1109/ICNN.1995.488968.

[12] J. Kennedy and R. C. Eberhart, “A discrete binary version of the particle swarm algorithm,” in *Computational Cybernetics and Simulation 1997 IEEE International Conference on Systems, Man, and Cybernetics*, Oct. 1997, vol. 5, pp. 4104–4108 vol.5. doi: 10.1109/ICSMC.1997.637339.

[13] J. Too, A. R. Abdullah, and N. Mohd Saad, “A New Co-Evolution Binary Particle Swarm Optimization with Multiple Inertia Weight Strategy for Feature Selection,” *Informatics*, vol. 6, no. 2, Art. no. 2, Jun. 2019, doi: 10.3390/informatics6020021.

[14] L.-Y. Chuang, H.-W. Chang, C.-J. Tu, and C.-H. Yang, “Improved binary PSO for feature selection using gene expression data,” *Comput Biol Chem*, vol. 32, no. 1, pp. 29–37, Feb. 2008, doi: 10.1016/j.compbiolchem.2007.09.005.

[15] E. Erturk and E. A. Sezer, “A comparison of some soft computing methods for software fault prediction,” *Expert Systems with Applications*, vol. 42, no. 4, pp. 1872–1879, Mar. 2015, doi: 10.1016/j.eswa.2014.10.025.

[16] R. S. Wahono and N. Suryana, “Combining Particle Swarm Optimization based Feature Selection and Bagging Technique for Software Defect Prediction,” *International Journal of Software Engineering and Its Applications*, p. 14, 2013.

[17] R. Wahono, N. Suryana, and S. Ahmad, “Metaheuristic Optimization based Feature Selection for Software Defect Prediction,” *Journal of Software*, vol. 9, pp. 1324–1333, May 2014, doi: 10.4304/jsw.9.5.1324-1333.

[18] R. Malhotra, N. Nishant, S. Gurha, and V. Rathi, “Application of Particle Swarm Optimization for Software Defect Prediction Using Object Oriented Metrics,” in *2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, Jan. 2021, pp. 88–93. doi: 10.1109/Confluence51648.2021.9377116.

[19] M. Shepperd, Q. Song, Z. Sun, and C. Mair, “Data Quality: Some Comments on the NASA Software Defect Datasets,” *IIEEE Trans. Software Eng.*, vol. 39, no. 9, pp. 1208–1215, Sep. 2013, doi: 10.1109/TSE.2013.11.

1. <https://figshare.com/articles/dataset/MDP_data_sets_D_and_>D\_-\_zipped\_up/6071675 [↑](#footnote-ref-2)
2. http://promise.site.uottawa.ca/SERepository/. [↑](#footnote-ref-3)