

ABSTRACT

The lack of training dataset in the new software development project becomes a major issue in the implementation of software defect prediction model, especially in the new software development project. Almost of all new software development projects are usually do not have historical dataset repositories. Utilizing the training dataset from either different or similar software project is not necessarily the best solution because of the heterogeneity in the software metrics. Unsupervised software defect prediction approach might be the best solution to address training dataset issue because it does not need training dataset to build the prediction model. Technically, the unsupervised approach is performed using clustering technique on the unlabeled dataset.

One of the latest unsupervised methods is spectral classifier based software defect prediction, which adopts the spectral graph clustering. Based on the literature review, the graph-based classifier methods outperformed the distance-based classifier methods. However, there are some identified issues in the use of spectral classifier based software defect prediction. First, the negative similarity issue in the graph adjacency matrix does not fulfill the non-negative Laplace matrix assumption. Second, the zero thresholding and eigenvector's values outliers issues cause the predominantly cluster and reducing the cluster compactness. This research is performed to address those issues using the two proposed methods, that are: (1) signed Laplacian based spectral classifier method, and (2) median absolute deviation based spectral classifier method. These proposed methods are evaluated using the public NASA MDP datasets with the unsigned Laplacian based spectral classifier as the baseline method.

Experimental results show that the used of absolute adjacency matrix can resolve the negative similarity issue in spectral classifier. Besides, the chosen of dispersion measure (i.e. median absolute deviation) as the partitioning threshold can produce better cluster compactness and classifier performances rather than using the central tendency measure as the partitioning threshold. The cluster compactness averages of these baseline, signed Laplacian based spectral classifier, and median absolute deviation threshold-based spectral classifier are 2.1 DBI, 1.8 DBI, and 1.4 DBI, respectively. The proposed methods can improve the cluster compactness of the baseline about 14.29% and 33.33%, respectively. Whilst the accuracy averages of these baseline, signed Laplacian based spectral classifier, and median absolute deviation threshold-based spectral classifier are 0.69, 0.74, and 0.79, respectively. The proposed methods can improve the accuracy of the baseline about 23.33% dan 31.67%, respectively. Hence, both the proposed methods are strongly suggested to be used as an unsupervised software defects prediction model, especially for a new software development projects that have no historical software dataset.

Keywords: *unsupervised software defect prediction, spectral clustering, signed Laplacian, zero threshold, median absolute deviation.*

INTISARI

Ketersediaan *training dataset* hingga saat ini masih merupakan isu utama pada penerapan model *software defect prediction*, khususnya pada proyek pengembangan *software* baru. Hampir semua proyek pengembangan *software* baru pada umumnya tidak memiliki repositori *dataset* historis. Memanfaatkan *training dataset* dari proyek *software* lain belum tentu memberikan solusi terbaik karena adanya heterogenitas *software metric*. Pendekatan *unsupervised software defect prediction* merupakan solusi terbaik untuk mengatasi masalah *training dataset* karena pendekatan tersebut tidak memerlukan *training dataset* pada pengembangan model. Secara teknis, pendekatan *unsupervised* dilakukan dengan menggunakan teknik *clustering* pada *dataset* tidak berlabel.

Salah satu metode *unsupervised software defect prediction* saat ini adalah *spectral classifier based software defect prediction* yang mengadopsi *spectral graph*. Berdasarkan studi literatur, metode *classifier* berbasis *graph* memiliki performansi yang lebih baik dibandingkan dengan metode *classifier* berbasis jarak. Namun demikian, implementasi *spectral classifier based software defect prediction* masih memiliki beberapa isu. Pertama, isu *negative similarity* pada *adjacency matrix* yang tidak memenuhi asumsi *non-negative Laplacian graph*. Kedua, isu penggunaan *zero threshold* dan isu *outliers* nilai *eigenvector* pada *partitioning* yang menyebabkan terbentuknya dominasi partisi dan penurunan kekompakan *cluster*. Penelitian ini dilakukan untuk mengatasi isu-isu tersebut dengan 2 (dua) metode usulan, yaitu: (1) *signed Laplacian based spectral classifier*; (2) *median absolute deviation based spectral classifier*. Metode-metode usulan tersebut dievaluasi menggunakan *public dataset* NASA MDP dengan metode *unsigned Laplacian based spectral classifier* sebagai *baseline*.

Hasil eksperimen menunjukkan bahwa penggunaan *absolute adjacency matrix* mampu mengatasi masalah *negative similarity* pada *spectral classifier*. Selain itu, pemilihan ukuran dispersi sebagai *partitioning threshold*, yaitu *median absolute deviation*, mampu menghasilkan kinerja *classifier* yang lebih baik dibandingkan *partitioning threshold* menggunakan ukuran tendensi pusat. Rerata *cluster compactness* dari *baseline*, *unsigned Laplacian based spectral classifier*, dan *median absolute deviation threshold-based spectral classifier* masing-masing sebesar 2.1 DBI, 1.8 DBI, dan 1.4 DBI. Peningkatan *cluster compactness* metode usulan dibandingkan dengan *baseline* masing-masing sebesar 14.29% dan 33.33%. Adapun rerata *accuracy* dari ketiga metode tersebut masing-masing sebesar 0.69, 0.74, dan 0.79. Peningkatan *accuracy* dari metode usulan dibandingkan dengan *baseline* masing-masing sebesar 23.33% dan 31.67%. Dengan demikian, metode-metode usulan dapat direkomendasikan sebagai metode alternatif untuk prediksi *software defect*, khususnya pada proyek *software* baru yang belum memiliki *dataset* repositori.

Kata kunci: *unsupervised software defect prediction, spectral classifier, signed Laplacian, zero threshold, median absolute deviation threshold.*