



وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم البرمجيات

التنبؤ واختبار توليد البيانات لكشف عيوب البرمجيات باستخدام الذكاء الاصطناعي

رسالة مقدمة

إلى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
وهي جزء من متطلبات نيل شهادة ماجستير علوم في
البرمجيات

من قبل

مصطفى زكي محمد يعقوب

بإشراف

أ.د. إبراهيم أحمد صالح

المستخلص

أدى تعقيد احتياجات البرامج في عصرنا الحالي إلى زيادة عدد التعليمات البرمجية بالتالي زيادة عدد الوحدات البرمجية (Modules) ، من المحتمل ان تحتوي بعض الوحدات البرمجية على أخطاء والتي تسبب أنواع مختلفة من المشكلات منها زيادة التكلفة و وقت التطوير وإنتاج منتج بجودة منخفضة. بالتالي من الضروري اكتشاف هذه الوحدات البرمجية وإصلاح أخطائها من أجل الحصول على منتج ذي جودة عالية .

يعد اكتشاف الوحدات البرمجية الخاطئة عملية حاسمة يجب إكمالها لكل منتج، ينصب تركيز هذه الرسالة على إنشاء نظام للتعقب بالوحدات البرمجية الخاطئة. إلى جانب ذلك ، تم إنشاء أداة لتوليد بيانات الاختبار لبرنامج معين من أجل التحقق من وظائفه، قد يصبح الأمر أكثر تكلفة ويستغرق وقتاً أطول لإكمال المشروع إذا تم إصلاح المشكلات أثناء الاختبار. بالتالي يساعد العثور على الوحدات المعيبة في المراحل المبكرة من التطوير (قبل الاختبار) وكذلك توليد بيانات اختبار فعالة على خفض تكلفة ووقت الاختبار .

في هذه الرسالة تم استخدام مجموعة من خوارزميات التعلم الآلي (Machine Learning) للتعقب بعيوب البرمجيات مثل الغابة العشوائية ، وشجرة القرار ،والجار الأقرب ، وآلة متجهات الدعم ،والانحدار اللوجستي ، وخوارزمية بايز البسيطة. حيث تم تطبيق هذه الخوارزميات على مجموعة من البيانات المعيبة التي تنتمي إلى مستودع (metrics data program) التابع لنا، واستخدام مجموعة من مقاييس الأداء؛ مثل: الدقة، والضبط ، والاسترجاع ، ودرجة f1 ، ومنحنى خاصة تشغيل المستقبل (ROC) ، لتقييم أداء الخوارزميات، وقد تبين أن خوارزمية الغابة العشوائية هي أفضل الخوارزميات المستخدمة في عملية التعقب ، حيث حصلت على دقة تتراوح بين 92% إلى 99% لمجموعة البيانات PC4 و MC1 على التوالي .

وتناولت الرسالة أسلوب توليد بيانات الاختبار لبرنامج تحديد نوع المثلث باستخدام خوارزمية مستعمرة النحل الاصطناعية ((Artificial Bee Colony(ABC) ، تعمل خوارزمية ABC على توليد حالات اختبار فعالة ومثلى، والتي تعد الادخال المطلوب لبرنامج قيد الاختبار (Software Under Test). تضمن بيانات الاختبار المنتجة من أن الوحدة البرمجية في البرنامج تعمل بشكل صحيح أو أنها مشفرة بشكل صحيح عن طريق اخضاع كل وحدة برمجية في البرنامج لاختبار الصندوق الأبيض مرة واحدة على الأقل، وفقاً لخوارزمية ABC المستخدمة ؛ تم

إنشاء مجموعة بيانات الاختبار المحسّنة لكل مسار مستقل في البرنامج ، حيث سيكون لكل مسار نوعان مختلفان من البيانات، أحدهما يعبر المسار بتغطية 100% ، والآخر يعبر المسار بتغطية 0%.

**Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and
Mathematics
Department of Software**



Prediction And Test Data Generation to Detect Software Defects Using Artificial Intelligent

**A Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Master of Science
in
Software**

**By
Mustafa Zaki Mohammed Yakoob**

**Supervised by
Dr. Ibrahim Ahmed Saleh**

abstract

The complexity of the software needs of our time has led to an increase in the number of lines of code and, accordingly, an increase in the number of modules. It is likely that some modules may contain errors which cause different types of problems like increased development cost and time and low quality product production. Therefore, it is necessary to detect and repair these software modules in order to obtain a high quality product.

Detecting faulty modules is a critical process that must be completed for every product. The focus of this thesis is to create a system for predicting faulty modules. Besides, a tool was created to generate test data for a specific program in order to check its functionality. It may become more expensive and take longer to complete the project if problems are fixed during testing. Finding faulty units in the early stages of development (before testing) as well as generating effective test data helps reduce the cost and time of testing.

In this thesis, a group of machine learning algorithms were used to predict software defects such as random forest, decision tree, nearest neighbor, support vector machine, logistic regression, and Naive Bayes were used to predict software errors. These algorithms were applied to a set of defective data belonging to NASA's MDP repository. A set of performance measures such as accuracy, precision, recall, f1 score and receiver operating characteristic (ROC) curve were used to evaluate the performance of the algorithms. It was found that the random forest algorithm is the best algorithm used in the prediction process, as it obtained an accuracy ranging from 92% to 99% for the data set PC4 and MC1, respectively.

The thesis also dealt with the method of generating test data for the program to determine the type of triangle using the Artificial Bee Colony algorithm. The (ABC) algorithm generates efficient and optimal test cases which are the required input for the program to be examined (SUT). The test data produced ensures that the software module is operating correctly or is properly encoded by subjecting each module in the software to a white box test at least once. According to the ABC algorithm used, the optimized test dataset was generated for each independent path in the software, where each path would have two different types of data. One crosses the path with 100% coverage, and the other crosses the path with 0% coverage.