جامعة الإسراء

Isra University

Master's Thesis

Particle Swarm Based Feature Selection for Improving Random Forest Classification Accuracy

Prepared By

Baraa Khalid Al-Wahidi

Supervised by

Dr.Thamer Al-Rousan

Co- Supervised by

Dr.Maher Abu Hamdeh

Submitted in Partial Fulfillment of the Requirements of the

Master Program of Software Engineering

Faculty of Information Technology

December 2019

أنا براء خالد عبد الرزاق الوحيدي

افوض جامعة الاسراء بتزويد نُسخ من رسالتي ، للمكتبات أو المؤسسات أو الهيئات أو الأشخاص عند طلبهم حسب التعليمات النافذة في الجامعة.

التوقيع: براء الوحيدي   التاريخ: 2019/12/21

I, Baraa Khalid Abdel-Razzaq Al-Wahidi give full permission to Al-Isra' University to provide copies of my thesis to, libraries, institutions, and other interested parties.

Signature: Baraa Al-Wahidi   Date: 21/12/2019

نموذج اقرار والتزام بقوانين جامعة الاسراء وانظمتها وتعليماتها لطلبة الماجستير والدكتوراه.

انا اسم الطالب: براء خالد الوحيدي الرقم الجامعي: z00505

تخصص: هندسة البرمجيات كلية:تكنولوجيا المعلومات

أعلنُ بأني قد التزمت بقوانين جامعة الاسراء وانظمتها وتعليماتها وقراراتها السارية المفعول المتعلقة بإعداد رسائل الماجستير والدكتوراه عندما قمت شخصياً بإعداد رسالتي / اطروحتي بعنوان:

Particle Swarm Based Feature Selection for Improving Random Forest Classification Accuracy

وذلك بما ينسجم مع الأمانة العلمية المتعارف عليها في كتابة الرسائل والأطاريح العلمية. كما أنني أُعلن بأن رسالتي/ اطروحتي هذه غير منقولة أو مستلة من رسائل أو أطاريح أو كتب أو أبحاث أو أي منشورات علمية تم نشرها أو تخزينها في أي وسيلة اعلامية، وتأسيساً على ما تقدم فأنني اتحمل المسؤولية بأنواعها كافة فيما لو تبين غير ذلك بما فيه حق مجلس العمداء في جامعة الاسراء بإلغاء قرار منحي الدرجة العلمية التي حصلت عليها وسحب شهادة التخرّج مني بعد صدورها دون أن يكون لي الحق في التظلم أو الأعتراض أو الطعن بأي صورة كانت في القرار الصادر عن مجلس العمداء بهذا الصدد.

التوقيع براء الوحيدي التاريخ: 2019/12/21

Committee Decision قرار اللجنة

لجنة المناقشة **التوقيع**

د. ثامر الروسان (مشرف رئيسي)
قسم هندسة البرمجيات ....

د. ماهر ابو حامدة . (مشرف مشارك)
قسم هندسة البرمجيات ....

د. أسامة الحاج (عضو لجنة)
قسم هندسة البرمجيات ٢٠١٩/١٢/٢٤

د. فيصل ابو الرب (عضو لجنة خارجي) ....

Dedication

I would like to take this opportunity to dedicate this work to my parents, family members, and colleagues. A special thanks to Dr. Thamer Al-Rousan and Dr.Maher Abo Hamdeh for guiding me step by step throughout this journey.

Table of Contents

Table of Figures

Abstract

Iterating over every possible combination of features and building each combination as a decision tree takes massive processing power especially when there aremany features to select from. The main drawback with using decision tree classifiers is the tendency of the tree to be over fitted to a specific scenario. The random forest classifier resolves this issue by using randomly selected features as nodes. The problem with this approach is that it requires more time and computational power to construct the trees. Researchers have identified this issue and worked on multiple variations of random forest to reduce the number of decision trees to be grown. Some of the successful variations use Symmetrical Uncertainty, and other methods to select a feature combination that will yield the highest accuracy achieving trees and generate a random forest for these features rather than the entire dataset. Others have employed the genetic algorithm in accordance with random forests to optimize the order and appearance of the features in making the random forest. In this research we employed an optimization algorithm called Binary Particle Swarm. The binary particle swarm optimization algorithm is a powerful algorithm in the field of optimization. We used this algorithm to pick the best features that represent a dataset as input for a random forest classifier. We have achieved impeccable results in terms of accuracy and precision while maintaining minimum user interaction. We used the Wisconsin breast cancer dataset which can be obtained from the UCI machine learning repository. The objective in this dataset is to predict whether the patient has a benign or malignant tumor based on the attributes provided. The other dataset we used was the Titanic disaster dataset which can also be obtained from the UCI machine learning repository. In this dataset, the objective is to predict whether the passenger has survived or not based on the provided attributes. We obtained a 97% on average and a best 98% classification accuracy on the Wisconsin breast cancer dataset. Using the same technique, we obtained 97% classification accuracy on the Titanic dataset.