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

تصميم وتنفيذ سرب التعلم التجميعي لاختيار الميزات

رسالة مقدمة
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة ماجستير علوم في
علوم الحاسوب

من قبل
نادية محمد مجيد إبراهيم

بإشراف
أ. د. فوزية محمود رمو

١٤٤٤هـ ٢٠٢٣م

الملخص

تعد الزيادة في أبعاد البيانات من المشـكلات التي تواجه أي عمل ضـمن مجال المعالجة الحاسوبية، اذ أصبح من الصعب على الخوارزمية الحاسوبية معالجة البيانات وتحويلها إلى معرفة. أحد أسـاليب تقليل الأبعاد هو اختيار الميزات التي تهدف إلى اختيار مجموعة فرعية من الميزات المثلى عن طريق ازالة البيانات غير ذات الصـلة وغير الملائمة والمتكررة لتحسـين أداء خوارزمية التعلم الآلي وزيادة دقة التصـنيف وتقليل وقت تدريب الخوارزميات والوقت الحسـابي، وكذلك تسـهل فهم أفضل لنموذج التعلم أو البيانات.

تم تصـميم وتنفيذ نظاماً ذكائياً حاسـوبياً لاختيار الميزات، اذ يتكوّن النظام المقترح من مرحلتين رئيسيتين، مرحلة التصـنيف ومرحلة اختيار الميزات، تم استخدام سبع خوارزميات في هذا النظام خمس للتعلم الالي وخوارزمية للتعلم التجميعي وخوارزمية ذكاء السرب.

تم اقتراح طريقة اختيار الميزة اسـتنادًا إلى أسـاليب ذكاء السـرب، حيث تم اسـتخدام تقنية اختيار الميزات بناءً على خوارزمية تحسـين اليعسـوب والتي تبحث عن أفضـل الحلول في منطقة البحث لتحقيق التحسين. تُستخدم تقنية التحسين للعثور على المجموعة الفرعية المثلى من الميزات التي يمكنها تصـنيف سـرطان الثدي بدقة حميد أو خبيث. وتم اسـتخدام خوارزمية التعلم التجميعي بالتصـويت على أسـاس الأغلبية (التصـويت الصـعب) لتجميع خمس خوارزميات للتعلم الآلي وهي (خوارزمية آلة المتجهات الداعمة Support Vector Machine (SVM)، بايز البسيط Naive Bayes (NB)، الجار الاقرب K-Nearest Neighbors (K-NN)، شجرة القرار Decision Tree (DT)، والغابة العشـوائية Random Forest(RF)). وتم اسـتخدام خوارزمية التعلم التجميعي كدالة لياقة في خوارزمية تحسـين اليعسـوب لتقييم مجموعات الميزات الفرعية التي تم اختيارها. تظهر النتائج ان التعلم التجميعي بالتصـويت على أسـاس الأغلبية (التصـويت الصـعب) يؤدي بشكل أفضـل من المصـنفات الفردية حيث أعطى أعلى قيمة دقة بنسـبة ٩٦,٤٩٪، في حين أن خوارزمية آلة المتجهات الداعمة، بايز البسـيط، الجار الاقرب، شـجرة القرار الأشـجار والغابات العشوائية حققت دقة بلغت (٩٥,٣٢٪ و٩٤,٧٣٪ و٩٢,٣٩٪ و٩٢,٩٨٪ و٩٥,٣٢٪) على التوالي

كما أظهرت النتائج أنه عند تدريب خوارزمية التعلم التجميعي بالتصـــويت على أســـاس الأغلبية (التصـــويت الصـــعب) على مجموعة الميزات الفرعية المثلى التي تم الحصـــول عليها باســـتخدام خوارزمية تحسـين اليعسـوب والبالغ عددها (١٧ ميزة)، تكون الدقة الناتجة اعلى بنسـبة ٩٨,٢٤٪، في حين تنفيذ الخوارزميـة على جميع الميزات التي بلغ عـددهـا (٣٠ ميزة) حققت دقـة بنســـبـة ٩٦,٤٩٪. على مجموعة بيانات ســـرطان الثدي Wisconsin Breast Cancer Dataset (WBCD) والتي تتكون من ٥٦٩ عينة و٣٠ ميزة.

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

Design and Implementation Swarm Ensemble Learning for Features Selection

**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
Computer Science**

By

Nadia Mohammed Majeed Ibraheem

Supervised by

Prof. Dr. Fawziya Mahmood Ramo

2023 A.D. **1444 A.H.**

Abstract

One of the problems facing machine learning algorithms is the increase in data dimensions, as it becomes difficult for the algorithm to process the data and convert it into useful knowledge. One dimensionality reduction technique is attribute selection which selects a subset of important attributes by eliminating irrelevant data in order to improve machine learning algorithm performance, enhance the accuracy of classification, decrease the training time of the algorithm and computational time, as well as facilitate a better understanding of the learning model or data.

An intelligent computer system for feature selection was designed and implemented, as the proposed system contained two main steps, the classification step, and features selection step. Seven algorithms were used in this system, five machine-learning algorithms, an ensemble-learning algorithm, and the swarm intelligence algorithm.

The feature selection method based on swarm intelligence methods was proposed, where the feature selection technique was used based on the Dragonfly optimization algorithm, which searches for the best solutions in the search area to achieve optimization. An optimization technique is used to find the optimal subset of features that can accurately classify breast cancer as benign or malignant. The ensemble learning algorithm was used by majority voting (hard voting) to aggregation-five machine learning techniques SVM, NB, K-NN, DT, and RF. The ensemble learning algorithm was used as an activation function in the Dragonfly optimization algorithm to evaluate the selected feature subsets The results show that ensemble learning by majority voting (hard voting) performs better than individual classifiers as it gave the highest accuracy value among individual classifiers with a percentage of 96.49%, while the SVM, NB, K-NN, DT, and RF it achieved an accuracy of (95.32%, 94.73%, 92.39%, 92.98%, and 95.32%), respectively. The results also showed that when training the ensemble learning algorithm by majority voting (hard voting) on the optimal set of sub-features obtained using the Dragonfly optimization algorithm (17 features), the resulting accuracy is higher by 98.24%, Whereas when training the algorithm on all features (30 features), it achieved an accuracy of 96.49% on a breast cancer dataset containing 569 samples and 30 features.