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

طريقة اختيار ميزة مقترحة تعتمد على تقنية المرشح-الغلاف والة المتجه الداعم لمجموعات البيانات عالية الابعاد

رسالة مقدمة

الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل

كجزء من متطلبات نيل شهادة ماجستير علوم في

الرياضيات / حاسوبية

من قبل

وفاء قاسم حمادي احمد

بإشراف

أ.د. عمر صابر قاسم

١٤٤٤ هـ ٢٠٢٢ م

المستخلص

إن الزيادة المتسارعة في حجم البيانات وتعقيدها يعد تحدياً كبيراً للعديد من التطبيقات والمجالات العلمية التي تمتلك بيانات ضخمة Big Data، مثل المجال الطبي الذي يتعامل مع انواعٍ مختلفة من البيانات عالية الابعاد، ولاجل حل هذه المشكلة تم استخدام مفهوم اختيار الميزة Feature Selection للتركيز على المعلومات المهمة الموجودة في مجموعة البيانات وتحديد الميزات ذات التأثير العالي في دقة التصنيف من خلال استخدام بعض طرائق اختيار الميزة مثل طريقة المرشح Filter، وطريقة الغلاف Wrapper، اذ تم في هذه الدراسة التركيز على تقنية الاعتماد الاحصائي Statistical Dependence (SD) كاحدى طرائق المرشح من اجل ترتيب البيانات حسب أهميتها في التأثير على دقة التصنيف، كما تم استخدام خوارزمية تحسين البحث الذري الثنائي Atom Search Binary Optimization (BASO) من اجل الحصول على الميزات المهمة واهمال الميزات غير المؤثرة بعد تحويل الخوارزمية الأساسية للبحث الذري من الفضاء المستمر Continuous Space الى الفضاء المتقطع Discrete Space، وقد تم اقتراح النموذج SD_BASO الذي يجمع بين تقنية الاعتماد الاحصائي وخوارزمية تحسين البحث الذري الثنائي من اجل الحصول على معالجة متقدمة للبيانات متمثلة في تقليل ابعاد هذه البيانات والحصول على دقة تصنيف عالية.

تم التركيز على تقنية آلة المتجه الداعم Support Vector Machine (SVM) كمصنف أساسي للخوارزمية او النموذج المقترح SD_BASO بعد اجراء عدة مقارنات بينه وبين شبكة الادراك Perceptron (PNN) والشبكة العصبية الاصطناعية للانتشار الخلفي للخطأ Back-Propagation (BPNN)، كما تم استخدام ثلاثة انواع من دوال النواة Kernel Function الخاصة بتقنية الة المتجه الداعم SVM وهي (Linear, polynomial, RBF) وبناء نموذج خاص لاختيار معلماتها من خلال خوارزمية تحسين البحث الذري ASO وبالاعتماد على مفهوم ضبط المعلمة Tuning Parameter لاجل الحصول على افضل دقة للتصنيف.

تم تطبيق واختبار الخوارزمية المقترحة SD_BASO على ثلاثة أنواع من مجموعات البيانات وهي (سرطان الرئة (Lung)، سرطان البروستات (Prostate)، سرطان الدم (Leukemia)) واجراء مجموعة من المقارنات مع الخوارزميات والطرائق الاعتيادية، اذ توضح النتائج أن الخوارزمية المقترحة SD_BASO اثبتت كفاءتها وتفوقها من خلال اختيار اقل عدد للميزات مع زيادة دقة نتائج التصنيف للبيانات.

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

A Proposed Feature Selection Method Based on Wrapper-Filter Technique and Support Vector Machine for High-Dimensional Datasets

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 Mathematics/Computational

By

Wafaa Qassim Hammadi Ahmad

Supervised by

Prof. Dr. Omar Saber Qasim

2022 A.D. **1444 A.H.**

Abstract

The rapid increase in the volume and complexity of data is a major challenge for many applications and scientific fields that have big data, such as the medical field that deals with different types of high-dimensional data, and to solve this problem, the concept of Feature Selection was used to focus on important information existing in the data set and identifying features that have a high impact on classification accuracy through the use of some feature selection methods such as the Filter method and the Wrapper method. According to its importance in influencing the accuracy of classification, the Binary Atom Search Optimization (BASO) algorithm was also used to obtain the important features and neglect the ineffective features after converting the basic algorithm for atomic search from Continuous Space to Discrete Space, The SD_BASO model has been proposed, which combines the statistical dependence technique and BASO algorithm to obtain advanced data processing represented in reducing the dimensions of this data and obtaining high classification accuracy.

The Support Vector Machine (SVM) technique was focused on as a basic classifier for the proposed algorithm or model SD_BASO after making several comparisons between it and the Perceptron Network (PNN) and Back-Propagation Artificial Neural Network (BPNN), and three types of the kernel functions of the SVM support vector machine technique (Linear, polynomial, RBF) and building a special model for selecting its parameters through the ASO atomic search optimization algorithm and based on the concept of parameter tuning parameter to obtain the best accuracy of classification.

The proposed SD_BASO algorithm was applied and tested on three types of datasets (Leukemia, Prostate, Lung) and a set of comparisons were made with ordinary algorithms and methods. The results show that the proposed SD-BASO algorithm proved its efficiency and superiority by selecting the least number of features in addition to Increasing the accuracy of the classification results for the data.