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

تحسين التحليل العنقودي الجزائي في تنقيب البيانات الضخمة

أطروحة مقدمة
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة دكتوراه فلسفة في
الرياضيات/الحاسوبية

من قبل

سارة غانم محمود الكبابجي

بإشراف

أ.د. عمر صابر قاسم

أ.د. زكريا يحيى الجمال

1443 هـ

2022 م

المستخلص

شهدت العقود الأخيرة تزايدا في كمية ونوع البيانات و اصبحت العديد من التطبيقات مصدراً لتدفق هذه البيانات، اذ ان الزيادة والتراكم في حجم البيانات يحتاج الى ابتكار طرائق وأساليب لتلخيص ودراسة وتنقيب هذه البيانات من اجل فهمها والاستفادة منها، حيث ان علم تنقيب البيانات والبحث عن المعرفة من العلوم الحديثة التي لازالت في حالة تطوير مستمر خصوصا بعد ظهور العديد من الخوارزميات الذكائية Intelligent Algorithms التي اثرت بشكل ملحوظ في علم مجال التنقيب عن المعرفة ومعالجة البيانات. في هذه الدراسة تم تحليل مجموعات متعددة من البيانات وكان الهدف الرئيس هو استخدام العنقدة Clustering لتحليل البيانات الضخمة Big Data، إذ تم اعتماد كل من خوارزمية بحث الخفاش(BA) Bat Search Algorithm، خوارزمية الغراب Grow Algorithm (GA)، خوارزمية مُحسن التوازن Equilibrium Optimizer Algorithm (EOA) مع طريقة العنقدة من نوع K-means من اجل تحليل مجموعات من البيانات الضخمة (H1N1 ، Biodegradable، MLL، SRBCT، Ecoli، Chalcone، Hepatitis، Iris، Wine، Wisconsin Breast Cancer، Liver Disorders)، اذ تم في الخوارزمية المقترحة الاولى BPRBC إيجاد واختيار الميزات من خلال مرحلتين إذ تمثلت المرحلة الأولى باستخدام خوارزمية الخفاش BA، لايجاد افضل الميزات للبيانات والمرحلة الثانية تمثلت باستخدام العنقدة الجزائية Panelized Clustering، لاختيار افضل الميزات إذ تم تحقيق نتائج جيدة مقارنة بـ PRBC وK-means، أما الخوارزمية الثانية المقترحة فتم من خلالها توظيف خوارزمية اسراب الغراب GA في إيجاد عدد العناقيد المستخدمة في خوارزمية K-means، اذ تم اختبارها على اربع مجموعات من البيانات التي تفوقت فيها الخوارزمية المقترحة مقارنة بالطريقة التقليدية المستخدمة في عنقدة البيانات، و تم توظيف خوارزمية مُحسن التوازن EOA كخوارزمية

مقترحة ثالثة EOAK-means لايجاد عدد العناقيد المثالية فضلاً عن اختيار الميزات Features Selection من مجموعات البيانات ومن خلال مقارنة النتائج تبين ان نتائج الخوارزمية المقترحة EOAK-means كانت افضل من نتائج الطرائق التقليدية من حيث مجموع المسافات داخل العنقود ICD وقيمة المؤشر RI.

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

Improving sparse cluster analysis in mining big data

**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 Doctor of Philosophy in
Mathematics/Computational**

By

Sarah Ghanim Mahmood AL-Kababchee

Supervised by

Prof. Dr. Omar Saber Qasim
Prof. Dr. Zakariya Yahya Algamal

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

Abstract

The last decades have witnessed an increase in the quantity and type of data, and many applications have become a source for the flow of this data, as the increase and accumulation in the volume of data requires devising ways and methods to summarize, study and mine this data in order to understand and benefit from it, as the science of data mining and the search for knowledge from Modern science, which is still in a state of continuous development, especially after the emergence of many intelligent algorithms that significantly affected the science of knowledge mining and data processing. In this study, multiple sets of data were analyzed, and the main objective was to use clustering to analyze big data, as the bat search algorithm (BA), the crow algorithm (GA), the balance optimizer algorithm (EOA) were adopted with the K-type clustering method. Means for analyzing large data sets (H1N1, Biodegradable, MLL, SRBCT, Ecoli, Chalcone, Hepatitis, Iris, Wine, Wisconsin Breast Cancer, Liver Disorders), the first proposed algorithm BPRBC was to find and select the features through two stages. The first stage was using the BA bat algorithm, to find the best features for the data, and the second stage was using the Panelized Clustering, to choose the best features, as good results were achieved compared to PRBC and K-means. As for the second proposed algorithm, the GA swarm algorithm was employed to find the number of clusters used in the K-means algorithm, as it was tested on four sets of data in which the proposed algorithm excelled compared to the traditional method used in data clustering, and the balance optimizer algorithm was employed EOA as a third proposed algorithm, EOAK-means, to find the number of ideal clusters as well as to select the features from the data sets.