Unsupervised Feature Selection Using an Improved Gazelle Optimization Algorithm for Arabic Text Clustering

Prepared by
Esraa Nasser Ahmed

Supervised by
Professor Mohammed Otair

Abstract

The field of data mining and machine learning has seen huge growth in recent years, with vast amounts of data available for analysis. Text document clustering, which is the process of separating documents into several groups, has become a significant task in the field of text mining. One challenge in this field is to select the most important and informative features and determine the importance of these features for the process of text clustering. The orthographic variants and the trilateral root system of the Arabic language make it a very difficult language to deal with in the text clustering field. The proposed method, which is called Gazelle Optimization Algorithm with Aquila Optimizer (GOA-AO), aims to address the feature selection step in Arabic document clustering using Gazelle Optimization

Algorithm (GOA) and Aquila Optimizer (AO). The Gazelle Optimization Algorithm is a novel optimization algorithm inspired by the ability of gazelles to survive in predator-dominated environments, and the Aquila Optimizer is a population-based optimization algorithm inspired by Aquila's natural behavior when catching prey. These two algorithms have not been used together before in the field of text clustering. The proposed method uses GOA for two iterations in the optimization process; after that, it goes into AO for one iteration. The effectiveness of the chosen features is estimated using the k-means approach. The proposed method was applied to the BBC Arabic, CNN Arabic, and SANAD-AlKhaleej datasets, which are popular news datasets collected from Arabic news websites. GOA-AO was evaluated using accuracy, precision, recall, f-measure, purity, and rand index evaluation matrices. The results of the proposed method GOA-AO have been compared to those of GOA, AO, standard K-means, combined K-means and Latent Dirichlet Allocation (LDA), and Particle Swarm Optimization K-means for BBC Arabic and CNN Arabic datasets. For SANAD-AlKhaleej datasets, the results have been compared to standard K-means and Clustering Arabic Documents based on Bond Energy (CADBE). The results show that the proposed method outperforms other methods in terms of purity, precision, recall, and f-measure for BBC Arabic and CNN Arabic datasets, and it outperforms other methods in terms of purity for SANAD-AlKhaleej-Seq and SANAD-

AlKhaleej-Rnd datasets. The best results achieved by the proposed method were 82% for the f-measure with the SANAD-AlKhaleej-Rnd dataset and 84% for the purity measure with the SANAD-AlKhaleej-Rnd dataset.

اختيار الميزات غير الموجه باستخدام خوارزمية تحسين الغزال المحسنة لتجميع النص العربي

إعداد

إسراء ناصر أحمد

إشراف

الأستاذ الدكتور محمد عطير

الملخص

شهد مجال التنقيب عن البيانات والتعلم الآلي نموًا هائلاً في السنوات الأخيرة ، مع توفر كميات هائلة من البيانات للتحليل. أصبح تجميع المستندات النصية ، وهي عملية فصل المستندات إلى عدة مجموعات ، مهمة مهمة في مجال التنقيب عن النص. يتمثل أحد التحديات في هذا المجال في تحديد أهم الميزات والمعلوماتية وتحديد أهمية هذه الميزات لعملية تجميع النص. إن المتغيرات الإملائية ونظام الجذر الثلاثي للغة العربية تجعلها لغة صعبة للغاية للتعامل معها في مجال تجميع النص. تهدف الطريقة المقترحة ، والتي تسمى خوارزمية Gazelle Optimization مع Aquila Optimizer (GOA-AO)، إلى معالجة خطوة اختيار الميزة في تجميع المستندات العربية باستخدام خوارزمية Gazelle Optimization (GOA) و Aquila Optimizer (AO). خوارزمية Gazelle Optimization هي خوارزمية تحسين جديدة مستوحاة من قدرة الغزلان على البقاء في البيئات التي يسيطر عليها المفترس ، و Aquila Optimizer عبارة عن خوارزمية تحسين تعتمد على السكان مستوحاة من سلوك Aquila الطبيعي عند اصطياد الفريسة. لم يتم

استخدام هاتين الخوارزميتين معًا من قبل في مجال تجميع النص. تستخدم الطريقة المقترحة GOA لتكرارين في عملية التحسين ؛ بعد ذلك ، تنتقل إلى AO لتكرار واحد. يتم تقدير فعالية الميزات المختارة باستخدام نهج k-mean. تم تطبيق الطريقة المقترحة على مجموعات بيانات بي بي سي عربي وسي إن إن العربية وسند الخليج ، وهي مجموعات بيانات إخبارية شهيرة تم جمعها من المواقع الإخبارية العربية. تم تقييم GOA-AO باستخدام مصفوفات تقييم الدقة ، والدقة ، والاستدعاء ، وf-measure ، والنقاء ، ومؤشر rand. تمت مقارنة نتائج الطريقة المقترحة GOA-AO مع تلك الخاصة بـ GOA و AO و القياسية K-means و K-means مجتمعة Latent Dirichlet Allocation (LDA) و K-Particle Swarm Optimization means لمجموعات بيانات BBC Arabic و CNN Arabic. بالنسبة لمجموعات بيانات سند الخليج ، تمت مقارنة النتائج بمعايير K-mean و Clustering Arabic Documents based on Bond Energy (CADBE). أظهرت النتائج أن الطريقة المقترحة تفوق في الأداء على الطرق الأخرى من حيث النقاء والدقة والاستدعاء وf-meausre لمجموعات بيانات بي بي سي العربية وسي إن إن العربية ، كما أنها تتفوق على الطرق الأخرى من حيث النقاء لمجموعات بيانات سند الخليج. أفضل النتائج التي تم تحقيقها بالطريقة المقترحة كانت 82٪ f-measure مع مجموعة بيانات SANAD-AlKhleej-Rnd و 84٪ لمقياس النقاء مع مجموعة بيانات SANAD-AlKhleej-Rnd.