The effect of the number of clusters on k-means performance

By

Taha Mohammad Othman Alakhras

Supervisor

Full Prof. Mohammed Abdallah Otair

Abstract

The most broadly used algorithm in the clustering is the K-Means algorithm, Historically K-Means is as yet the best and fastest clustering algorithm among other algorithms (Umargono et al, 2020). It has ability to deal with a big data in a short time and efficient computing time. The K-means algorithm is widely used in various fields, in data mining, machine learning, and image processing, but there are still weaknesses in this algorithm. First, determining the number of clusters depends on assumptions. Secondly, the initial selection of the centers of the clusters is random, to deal with this weaknesses, there are hundreds of researches and many method to determine the number of clusters in advance.

In this study, will illustrating the effect of the number of clusters on k-means algorithm performance, and will proposing technique to use to determine the best number of clusters. Comparisons between the proposed technique results and elbow method, gap statistic method and Twenty-six other methods are available in the nbclust package and ratio between-cluster sum of squares (between_ss) to the total sum of squares (tot_ss) , and will used three various types dataset that has been subjected to previous studies on cluster algorithms (iris dataset, wine dataset, yeast dataset), the results of this study indicate, sure,

appear to indicate that there is no unanimous choice regarding the optimal number of clusters, and the proposing technique was largely successful in determining the number of clusters in a ratio between_SS / total_SS = (90.2%) with wine dataset , and (71.5%) with iris dataset and (81.8 %) with yeast dataset, which is a very good percentage, we tried to give the centers of clusters in advance as well, but it survived with wine and iris dataset, and did not succeed with the yeast dataset.

تأثير عدد العناقيد على اداء خوارزمية العنقدة

اعداد

طه محمد عثمان الاخرس

اشراف

الأستاذ الدكتور محمد عطير

الملخص

الخوارزمية الأكثر استخداما في التجميع هي خوارزمية ال (K–Means) ، تاريخيا تعد خوارزمية ال(K–Means) افضل و اسرع خوارزمية تجميع بين الخوارزميات الأخرى و لديها القدرة على التعامل مع البيانات الضخمة في وقت قصير و وقت حوسبة فعال (Umargono et al, 2020). و تستخدم خوارزمية ال (K–Means) على نطاق واسع في مختلف المجالات، في التنقيب عن البيانات و التعلم الالي و معالجة الصور، لكن لا تزال هناك نقاط ضعف في هذه الخوارزمية أولا : يعتمد تحديد عدد المجموعات مسبقا على افتراضيات، ثانيا يكون الاختيار الاولي لعدد العناقيد عشوائيا، ومن اجل التعامل مع نقاط الضعف هذه هناك المئات من الأبحاث و الكثير من الطرق لتحديد عدد العناقيد مسبقا، في هذه الدراسة سوف نوضح تأثير عدد العناقيد على أداء خوارزمية العنقدة و سوف نقترح تقنية لاستخدامها لتحديد عدد المجموعات الأمثل مسبقا ، وسنقارن نتائج التقنية المقترحة مع طريقة ال (elbow) و طريقة ال(gap) بالإضافة الى 26 طريقة أخرى متوفرة في حزمة ال (noblest) و النسبة بين مجموع المربعات بين المجموعات(between_ss) الى مجموع المربعات الكلي(total_ss) والتي كلما اقتربت من الواحد

كلما زادة دقة العنقدة، كما سيتم استخدام ثلاثة أنواع مختلفة من مجموعة البيانات التي خضعت لدراسات سـابقة على الـــــــــخوارزميات العنقودية (iris dataset, wine dataset, yeast dataset)، وتشير نتائج هذه الدراسة الى انه و بالتأكيد لا يوجد خيار بالإجماع فيما يتعلق بالعدد الأمثل للعناقيد بين جميع الطرق السابقة ، وكانت نتائج التقنية المقترحة ناجحة الى حد كبير في تحديد عدد المجموعات حيث كانت النتائج (between_SS / total_SS = 90.2%) مع مجموعة البيانات (wine) و (71.5%) مع مجموعة البيانات (iris) و (% 81.8) مع مجموعة البيانات (yeast) وهي نسب جيدة جدا، كما انه حاولنا إعطاء مراكز العناقيد مسبقا، كانت ناجحة مع مجموعة البيانات (wine) و (iris) و لم تنجح مع مجموعة البيانات(yeast).