Title:	Performance Tuning of K-Mean Clustering Algorithm a Step towards Efficient DSS
Author(s):	Ayman E. Khedr Ahmed I. El Seddawy **Amira M. Idrees**
Journal/Conference:	International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
Publication details:	Volume 2, Issue 6
Publication Date:	November - 2014
Publisher	Innovative Research Publication
Place	http://www.ijircst.org/User/Home.aspx

Paper Title:	**Performance Tuning of K-Mean Clustering Algorithm a Step towards Efficient DSS**
Main Domain:	Data Mining (DM)
Sub-Domain:	Unsupervised Learning – Clustering in Decision Support Systems (DSS)
Problem:	DSS and DM is a computing environment where users can find unknown strategic information for decision making. DM offers a variety of advanced data processing techniques that may beneficially be applied for Business Intelligence (BI) purposes. Although k-mean has the great advantage of being easy to implement, the quality of the final clustering results of the k-mean algorithm highly depends on different factors. The following points summarize the problems in k-mean algorithm, they are: 1. Quality of the output depends on the initial point. 2. Global optimum solution not guaranteed. 3. Non globular clusters (overlapping in data between clusters) 4. Assuming a random number of clusters which may not be accurate. 5. Find empty clusters. 6. Bad initialization to centroid point 7. Choosing the number of clusters
Context:	This research focuses on the first step that applies k-means which is an efficient clustering algorithm for categorizing the customers' data. The easiness of k-mean clustering algorithm was a motivation that this algorithm used in several fields. The k-mean clustering algorithm is more prominent since its intelligence to cluster massive data rapidly and efficiently. This paper considers developing an adaptation on one of the most well-known popular clustering algorithms (K-mean) to enhance the performance in producing near-optimal decisions for telcos churn prediction and retention problems. The DSS could help managers to forecast and optimize efficiencies by selected attributes and grouping inferred efficiency. Also, it is an ideal tool for careful forecasting and planning. The proposed DSS is applied to an actual banking system and its superiorities and advantages are discussed.

Solution approach:	In this research a new method is proposed for finding the better initial centroids to provide an efficient way of assigning the data points to suitable clusters with reduced time complexity. The modified approach will include the following: 1. Changing the centroid point from random points to the centroid points. 2. Adding a step while calculating the distance between data sample and cluster through inserting several center points to increase time of processing 3. Adding last step to avoid empty clusters before visualize data The proposed enhancement has been evaluated by performing a comparison through the method of processing, which illustrates the main difference between the processing steps between both algorithms. Moreover, An evaluation of the proposed enhancing algorithm by applying both algorithms (the original and the enhancement) on customer investment in banking dataset. The difference between the results confirmed the correctness and accuracy of the "Enhanced K-mean" algorithm.
Contribution:	This research successfully developed and applied on customer banking data, and the evaluation results are presented The research contribution is summarized as follows: • More accuracy in distributing the clusters • Considering the determination of the cluster elements. • Determining the degree of closeness between the element and its related clusters • Selecting the most relevant cluster • Assigning the element to this selected cluster • Removing the empty clusters which was resulted according to the random number of clusters and avoids presenting it to the user. • Avoids this inconvenient representation and remove the entire empty clusters before presenting them to the user.

إسم البحث	**ضبط أداء لخوازمي التجمعات K-Mean نحو نظم دعم إتخاذ القرار فعالة**
ملخص البحث	تعتبر نظم دعم إتخاذ القرار و التنقيب فى البيانات بيئة حوسبية فعاله حيث يتم تمكين المستخدمين من العثور على المعلومات الاستراتيجية والتى يمكن تطبيقها فى أغراض ذكاء الأعمال لصناع القرار. على الرغم من أن الخوارزمى K-Mean له ميزة كبيرة فى سهولة التنفيذ، إلا أن نوعية التجميع النهائي للنتائج يعتمد إلى حد كبير على عوامل مختلفة. تتلخص المشاكل في خوارزمية K-Mean فى النقاط التالية: • جودة الإخراج تعتمد على النقطة الأولى. • شمولية الحل الأمثل ليس مضموناً. • تداخل في البيانات بين المجموعات • افتراض عدد عشوائي من المجموعات وهذا العدد قد لا يكون دقيقا. • وجود مجموعات لاتحتوى على عناصر تابعه. • إمكانية الإختيار الغير جيد لنقطة التمركز المبدئية
سياق البحث	يركز هذا البحث على تعديل الخوارزمى K-Mean وأحد الدوافع لإستخدام هذا الخوارزمي هو سهولة و إمكانية تطبيقه في العديد من المجالات. ويعتبر هذا الخوارزمى هو الأكثر بروزا من حيث السرعة والكفاءة. يهدف هذا التعديل لتجميع فعال للبيانات فى مجموعات لتصنيف بيانات العملاء بدقة عاليه. مما يؤدى لتحسين الأداء في إصدار قرارات تقترب إلى المثاليه فى حل مشاكل التنبؤ وإمكانية التخطيط الدقيق لمجالات كثيرة منها مجال الإتصالات. كما يمكن أن يساعد المديرين على تحسين الكفاءة بالسمات المحددة. فى هذا البحث تم تطبيق التعديل المقترح على نظام مصرفي فعلي وبيانات بنكية حقيقية للعميلاء وتم عرض النتائج وإثبات فعالية التعديل المقترح.
إسلوب البحث	هذا البحث يقترح طريقة جديدة للعثور على أفضل نقطة أولى للتمركز كوسيلة فعالة لتعيين نقاط البيانات للمجموعات بدقة أعلى و خطوات أقل تعقيدا. تم تنفيذ التعديل المقترح وتطبيق الخوارزمي بعد التعديل على البيانات المصرفية للعملاء بنجاح، وتم عرض نتائج التقييم. يشمل الخوارزمى المقترح عدة تغييرات كما يلي: 1. تغيير نقطة التمركز من نقاط عشوائية إلى نقطة مركز (منتصف) للبيانات 2. إضافة خطوة عند حساب المسافة بين نموذج البيانات والكتلة من خلال إدراج عدة نقاط مركزيه 3. إضافة خطوة أخيرة لتجنب الكتل فارغاً قبل عرض تصور البيانات تم تقييم الخوارزمى المقترح عن طريق إجراء مقارنة بين الخوارزمى الأصلى K-Mean والتعديل المقترح وتعزيز التقييم بتطبيق كلا الخوارزميات (الأصلي والتعديل) على استثمارات العملاء في مجموعة البيانات المصرفية. أكدت النتائج صحة نتائج الخوارزمية المعدل بحيث أصبح أكثر دقة في تحديد المجموعات من الخوارزمى الأصلى.
النتائج المستخلصة	تم تطبيق فكرة البحث على البيانات المصرفية للعملاء، وتم عرض نتائج التقييم البحث ويتلخص نتائج البحث على النحو التالي: • المزيد من الدقة في توزيع المجموعات ودقة تحديد عناصر المجموعة • دقة تحديد درجة التقارب بين العنصر و المجموعة ذات الصلة • تحديد الكتلة الأكثر صلة بالعنصر وإزالة المجموعات الفارغة