CLUSTERING TRAFFIC ACCIDENT HOTSPOTS

Overview of progress in this study.

Iyana Jones
Coreen Mullen

OUTLINE

What will these slides go over?

- Introduction
- What is KMEANS and DBSCAN?
- Lit review
- Result
- Future work

INTRODUCTION

The purpose to our research is to **identify and visualize traffic accident hot spots** for community members and local government to see.

We do this by using **k-means and dbscan** algorithms to perform the clustering of our data.

Data Overview

39,222 x 85

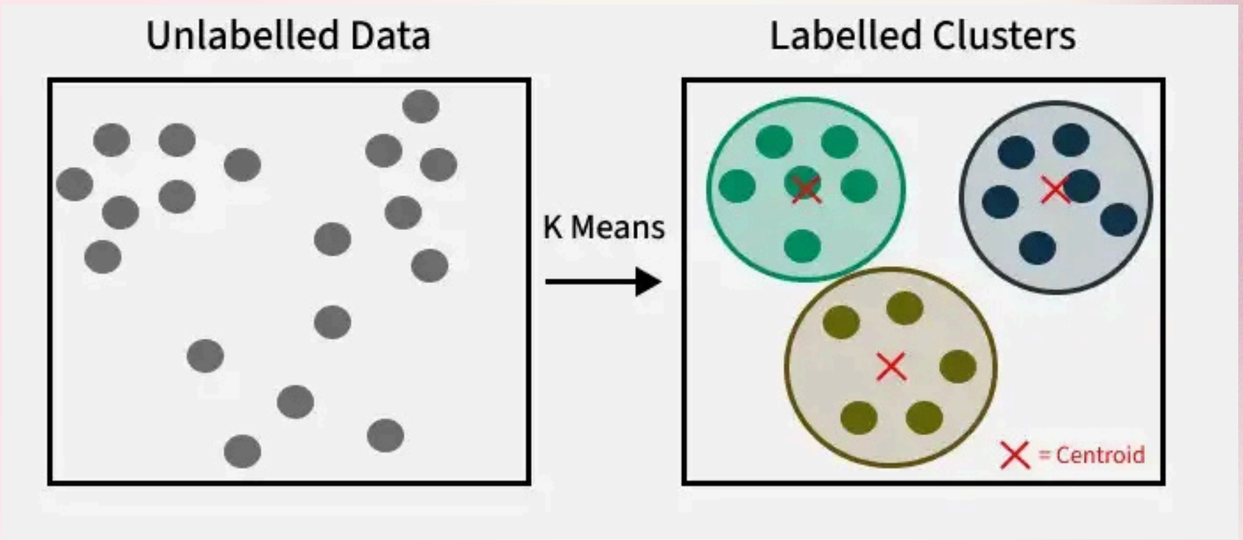
Our data comes from the National Highway Traffic Safety Adminstration and represents the fatal accident locations in North Carolina in 2022.

KMEANS

K-means is a very popular algorithm used to divide data.

- K-means is simple and fast and works good for numerical data.

Defining the value of k is something that is the scientists responsibility, in our initial research we chose a k of 10.



NEXT

Tip: Double-click to customize this poll or quiz, or go to the **Elements Tab** for more options!

KMEANS 2

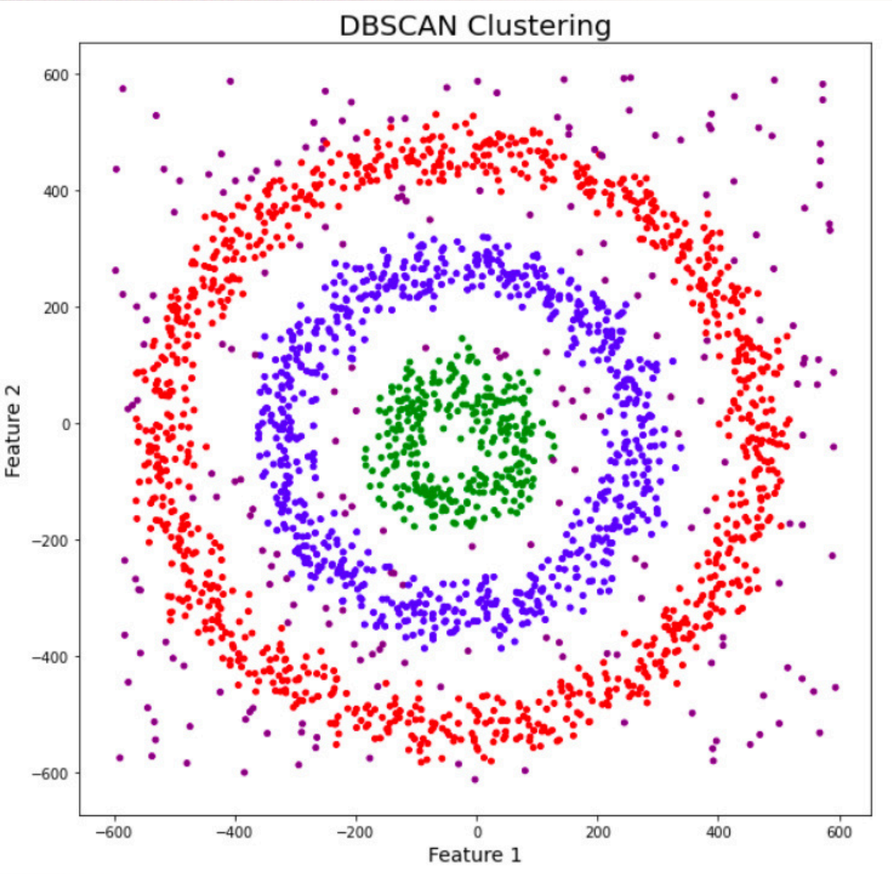
```
# run a k-means once
labels, inertia, centers, n_iter_ = kmeans_single(
    X,
    sample_weight,
    centers_init,
    max_iter=self.max_iter,
    verbose=self.verbose,
    tol=self._tol,
    n_threads=self._n_threads,
)
# determine if these results are the best so far
# we chose a new run if it has a better inertia and the clustering is
# different from the best so far (it's possible that the inertia is
# slightly better even if the clustering is the same with potentially
# permuted labels, due to rounding errors)
if best_inertia is None or (
    inertia < best_inertia
    and not _is_same_clustering(labels, best_labels, self.n_clusters)
):
    best_labels = labels
    best_centers = centers
    best_inertia = inertia
```

Our exploration of k-means produces a scatter plot showing accident hot spot locations.

DBSCAN

DBSCAN Algorithm is a density-based clustering algorithm that works on the assumption that clusters are dense regions in space separated by regions of lower density.

DBSCAN



How DBSCAN Helps in Road Accident Analysis

- Identifying Accident-Prone Areas (Hotspots)
- Works Well with Geographic Data (Lat-Long)
- No Need to Specify the Number of Clusters
- Helps in Decision-Making for Traffic Management

LIT REVIEW

Two methods for Automatic K-means Clustering.

Journal of Classification
https://doi.org/10.1007/s00357-024-09471-5

A Novel Classification Algorithm Based on the Synergy Between Dynamic Clustering with Adaptive Distances and K-Nearest Neighbors

Mohammed Sabri[1,2] · Rosanna Verde[2] · Antonio Balzanella[2] · Fabrizio Maturo[3] · Hamid Tairi[1] · Ali Yahyaouy[1] · Jamal Riffi[1]

Accepted: 18 April 2024

Abstract
This paper introduces a novel supervised classification method based on dynamic clustering (DC) and K-nearest neighbor (KNN) learning algorithms, denoted DC-KNN. The aim is to improve the accuracy of a classifier by using a DC method to discover the hidden patterns of the apriori groups of the training set. It provides a partitioning of each group into a predetermined number of subgroups. A new objective function is designed for the DC variant, based on a trade-off between the compactness and separation of all subgroups in the original groups. Moreover, the proposed DC method uses adaptive distances which assign a set of weights to the variables of each cluster, which depend on both their intra-cluster and inter-cluster structure. DC-KNN performs the minimization of a suitable objective function. Next, the KNN algorithm takes into account objects by assigning them to the label of subgroups. Furthermore, the classification step is performed according to two KNN competing algorithms. The proposed strategies have been evaluated using both synthetic data and widely used real datasets from public repositories. The achieved results have confirmed the effectiveness and robustness of the strategy in improving classification accuracy in comparison to alternative approaches.

Keywords K-nearest neighbors · Dynamic clustering · Combinatorial classification · Adaptive distances

1 Introduction

Hybridization of K-means with improved firefly algorithm for automatic clustering in high dimension

Afroj Alam [1]
Department of Computer Application
Integral University
Lucknow, India
alamafroj@student.iul.ac.in

Mohd Muqeem 2
Department of Computer Application
Integral University
Lucknow, India
muqeem@iul.ac.in

A NOVEL CLASSIFICATION ALGORITHM BASED ON THE SYNERGY BETWEEN DYNAMIC CLUSTERING WITH ADAPTIVE DISTANCES AND K-NEAREST NEIGHBORS

The DC-KNN paper follows this process:

1. Dynamic Clustering Algorithm
2. Assigning Weights
3. KNN Classification

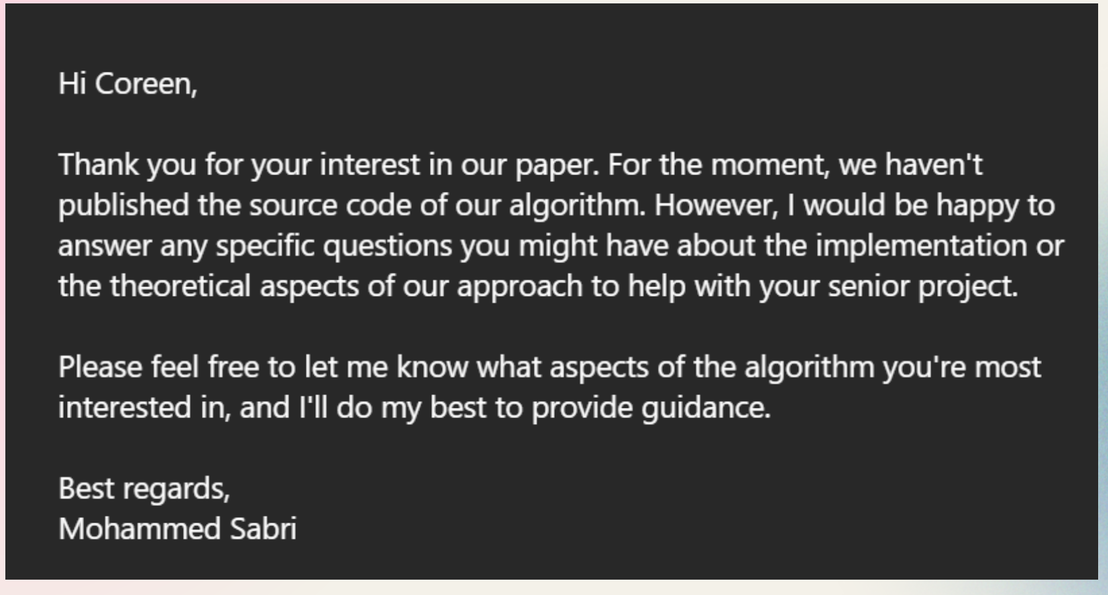
The results of the DC-KNN research

Pros

- Better classification of clusters.
- Adaptive distance for clusters allowing for re-clustering of training data.

Cons

- High computational cost.
- Sensitive to outliers.



Hi Coreen,

Thank you for your interest in our paper. For the moment, we haven't published the source code of our algorithm. However, I would be happy to answer any specific questions you might have about the implementation or the theoretical aspects of our approach to help with your senior project.

Please feel free to let me know what aspects of the algorithm you're most interested in, and I'll do my best to provide guidance.

Best regards,
Mohammed Sabri

NEXT

HYBRIDIZATION OF K-MEANS WITH IMPROVED FIREFLY ALGORITHM FOR AUTOMATIC CLUSTERING IN HIGH DIMENSION

This paper goes over the implementation of the Firefly Algorithm as a clustering algorithm in combination with k-means.

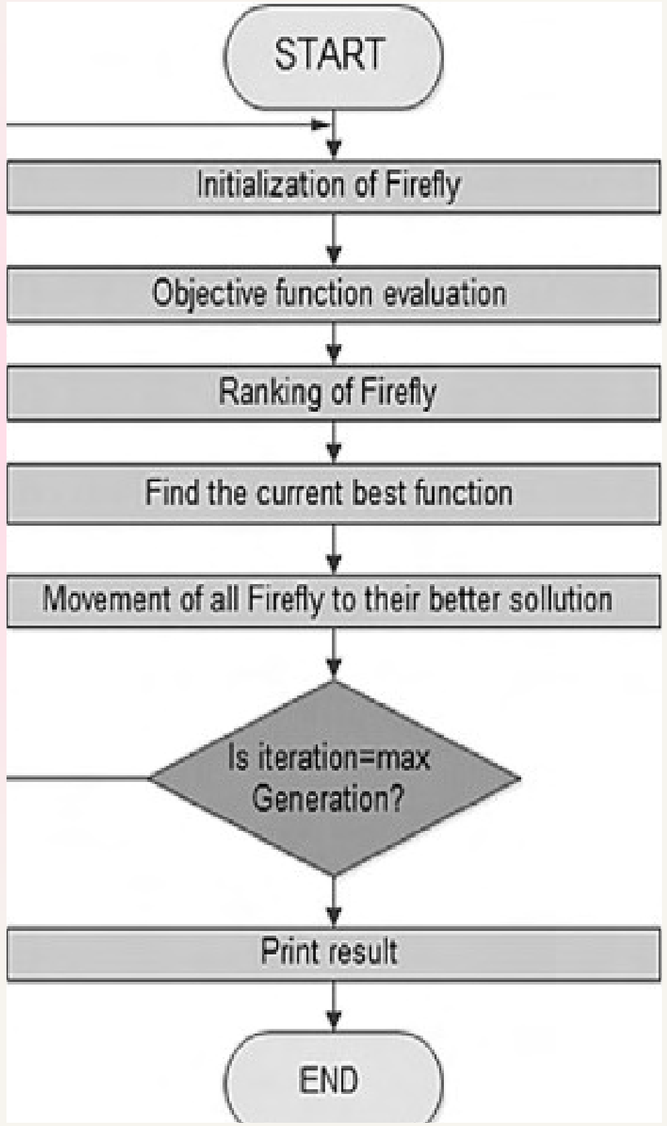
Pros

- Automatic optimal k determined.
- Broad scope of applications.

Cons

- Computationally expensive.
- Difficulties with optimizing every dimension in high dimensional datasets.

NEXT

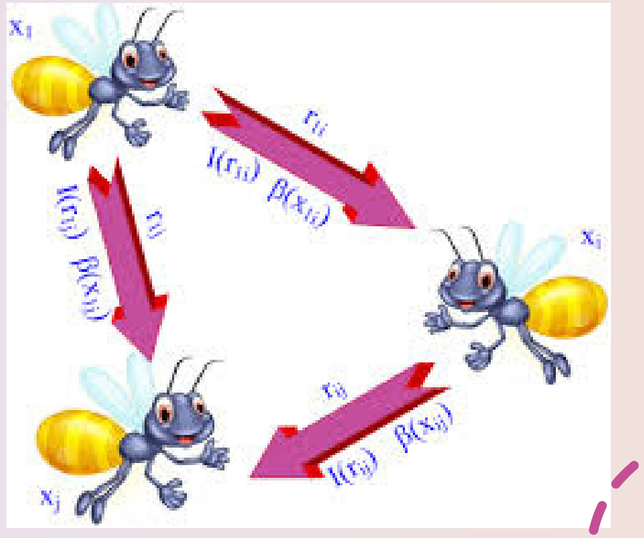


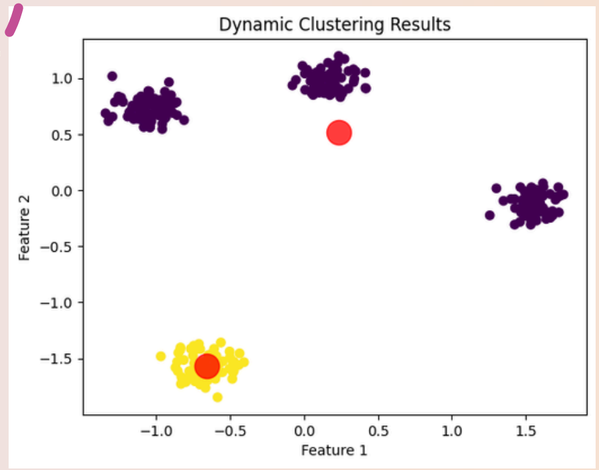
RESULTS

From comparing the two papers both have many strengths and weaknesses, but when comparing them for use in our study the Firefly algorithm wins.

Why?

- Handles larger datasets better.
- Cluster quality optimization.
- Better handling of noise.





Data preprocessing.

CHALLENGES

Coding challenges.

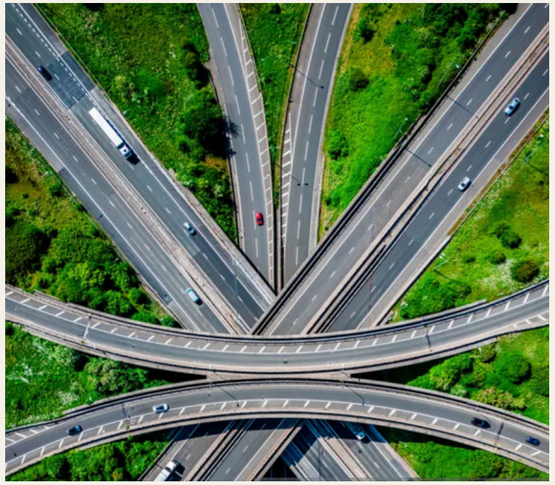
No code availability for literature review algorithms.

FUTURE WORK

What things will we work on in the future to improve the project?

FUTURE WORK

Incorporate machine learning to predict future hot spot areas.



Implement GIS to see if factors like vegetation or pot holes effect the hot spot locations.

More in-depth analysis of various factors on hot spot crash locations.

CITATIONS

Sabri, M., Verde, R., Balzanella, A. et al. A Novel Classification Algorithm Based on the Synergy Between Dynamic Clustering with Adaptive Distances and K-Nearest Neighbors. *J Classif* 41, 264–288 (2024). https://doi.org/10.1007/s00357-024-09471-5
Alam, Afroj. (2023).

Hybridization of K-means with Improved Firefly Algorithm for Automatic Clustering in High Dimension. *arXiv preprint* arXiv:2302.10765. https://arxiv.org/abs/2302.10765

Yang, XS. (2009). Firefly Algorithms for Multimodal Optimization. In: Watanabe, O., Zeugmann, T. (eds) Stochastic Algorithms: Foundations and Applications. SAGA 2009. Lecture Notes in Computer Science, vol 5792. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04944-6_14

THANK YOU FOR LISTENING!