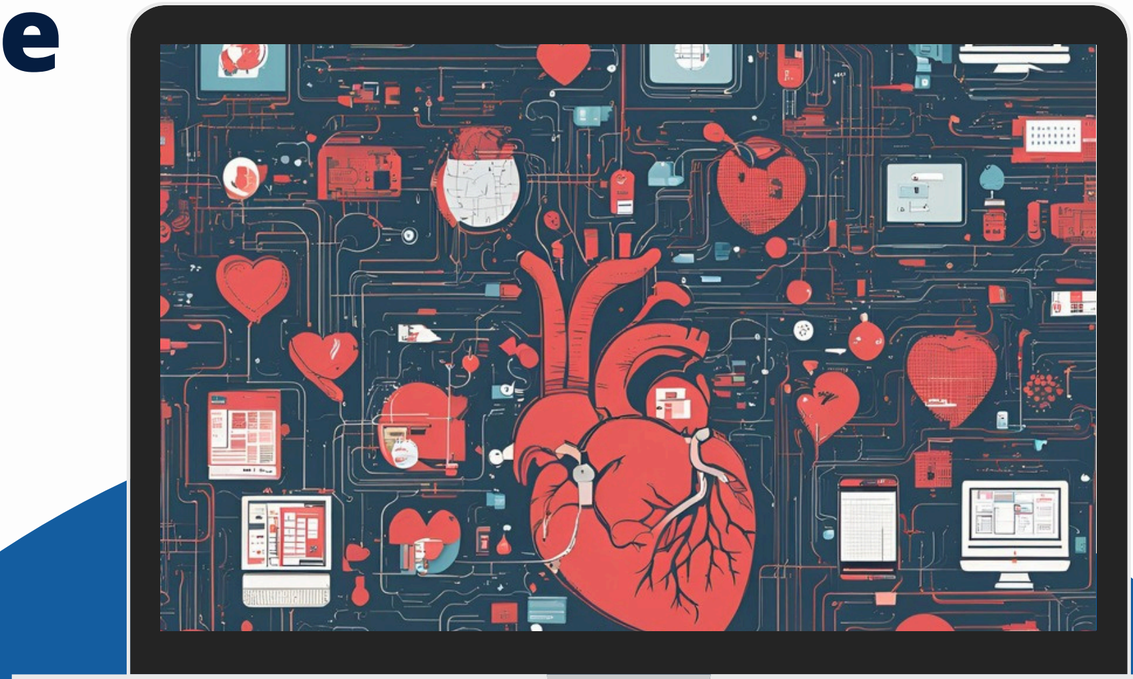
JCT
LEV ACADEMIC CENTER

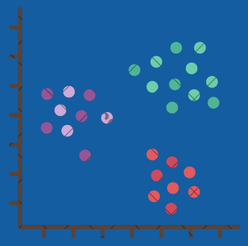
Efficient genetic K-Means clustering for health care knowledge discovery

A. Alsayat and H. El-Sayed, 2016, IEEE



By: Rachelli Adler, Esther Malka Nusbacher

Introduction



- Data mining and machine learning are crucial for healthcare decision-making.
- Clustering helps segment patients for effective treatments. similar patience might have similar treatments.

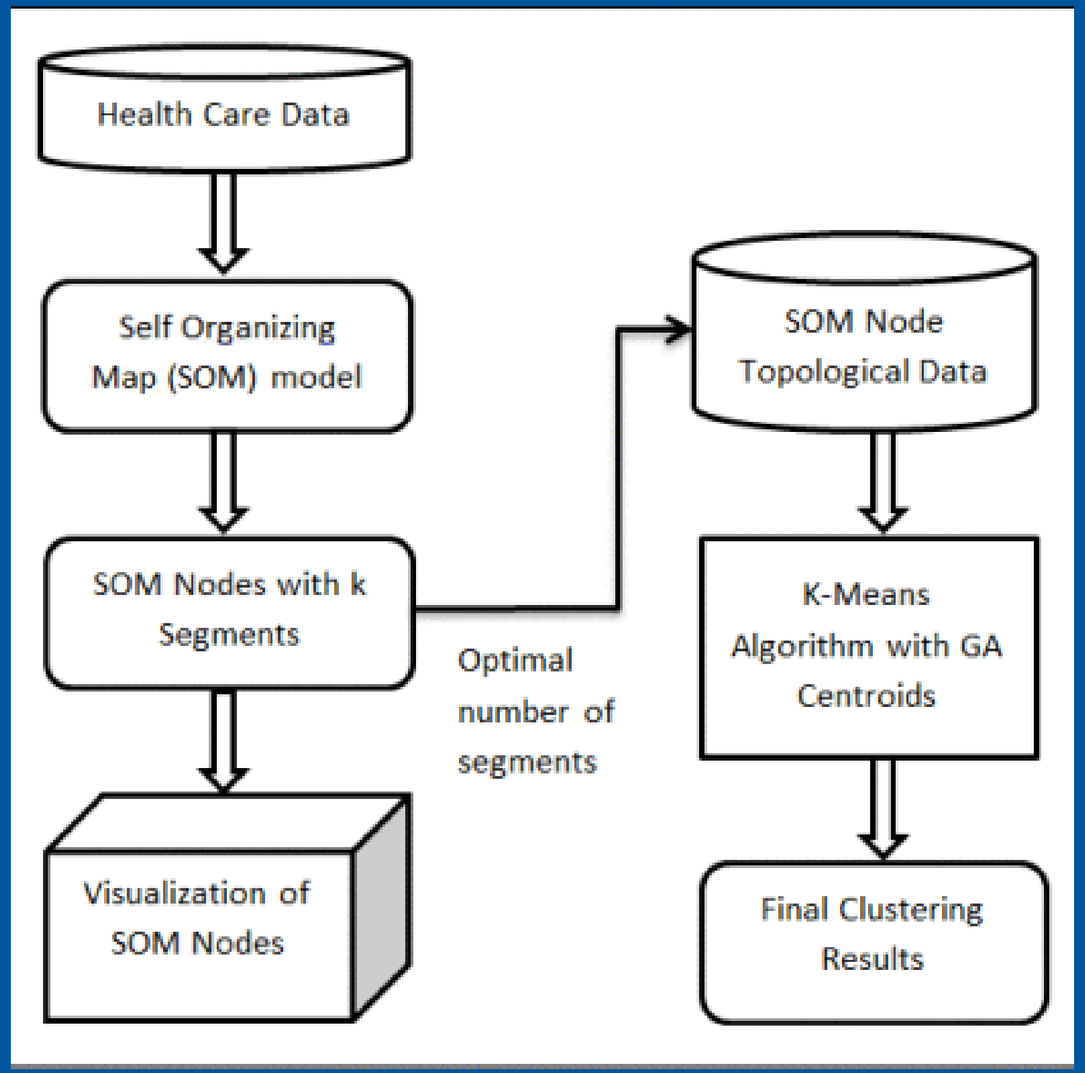
Problem Statement

- How can we improve clustering accuracy in healthcare data?
- Traditional K-Means struggles with selecting the right number of centroids. and may not give the best reults.

Methods

- K-Means Clustering
- Genetic Algorithm
- DBSCAN
- Self Organizing Map (SOM)

“We propose an **efficient K-Means** clustering algorithm which uses the **SOM** method to discover the optimal segments number in the data as a preprocessing step”



The Data

We used two datasets on **liver disease** and **heart disease** from UCI Machine Learning Repository.

Liver Disease Dataset	
Attribute Name	Description
mcv	mean corpuscular volume
alkphos	alkaline phosphotase
sgpt	alamine aminotransferase
sgot	aspartate aminotransferase
gammagt	gamma-glutamyl transpeptidase
drinks	alcoholic beverages drunk per day
selector	class label for liver disease

- 345 patients
- 7 blood test variables related to liver disease (e.g., alcohol-related)
- Class label: Presence of liver disease

✓ No preprocessing required

Heart Disease Dataset	
Attribute Name	Description
age	age in years
sex	patient gender
cp	chest pain type
trestbps	resting blood pressure
chol	serum cholestoral
fbs	fasting blood sugar
restecg	resting electrocardiographic results
thalach	maximum heart rate
exang	exercise induced angina
oldpeak	ST depression
slope	he slope of the peak exercise ST segment
ca	number of major vessels
thal	exercise test
num	diagnosis of heart disease

- 303 patients
- 14 variables related to heart disease diagnosis
- Class label: Presence of heart disease

✓ No train/test split needed

The expected results

Dataset	Weighted Classification Accuracy (%)		
	SOM Genetic K-Means	K-Means	DBSCAN
Liver Disease	**73.84**	69.15	67.66
Heart Disease	**69.90**	66.27	61.45

Key Findings

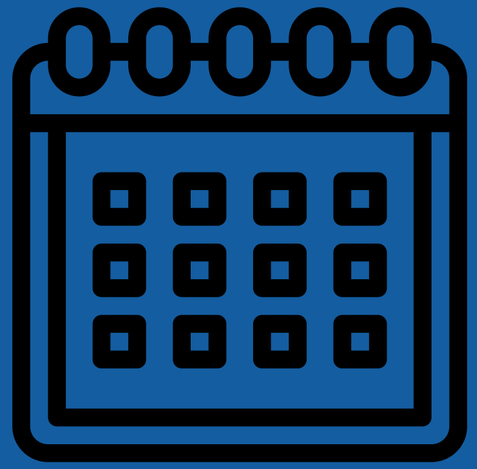
identified a more accurate K result tailored to the data, improves clustering, identifies patterns, and enhances behavior.

Parameter Optimization

Changing SOM neuron count, cluster numbers, distance measures, and noise sensitivity improves clustering and helps explore the data better.

Project Breakdown

Download and prepare the dataset	5%
Run K-Means and DBSCAN (baseline results)	25%
Implement SOM clustering and apply Genetic K-Means	25%
Evaluate performances	15%
Generate visualizations	10%
Compare results and Present conclusions	20%



Bibliography

01

THE ARTCLE:
https://ieeexplore.ieee.org/document/7516127

02

DATASET 1 : liver disease

03

DATASET 2: heart disease

04

Published in: 2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA)

Date of Conference: 08-10 June 2016

DOI: 10.1109/SERA.2016.7516127

Date Added to IEEE *Xplore*: 21 July 2016

Publisher: IEEE

ISBN Information:

Electronic ISBN:978-1-5090-0809-4

Print on Demand(PoD) ISBN:978-1-5090-0810-0

Conference Location: Towson, MD, USA

THANK YOU