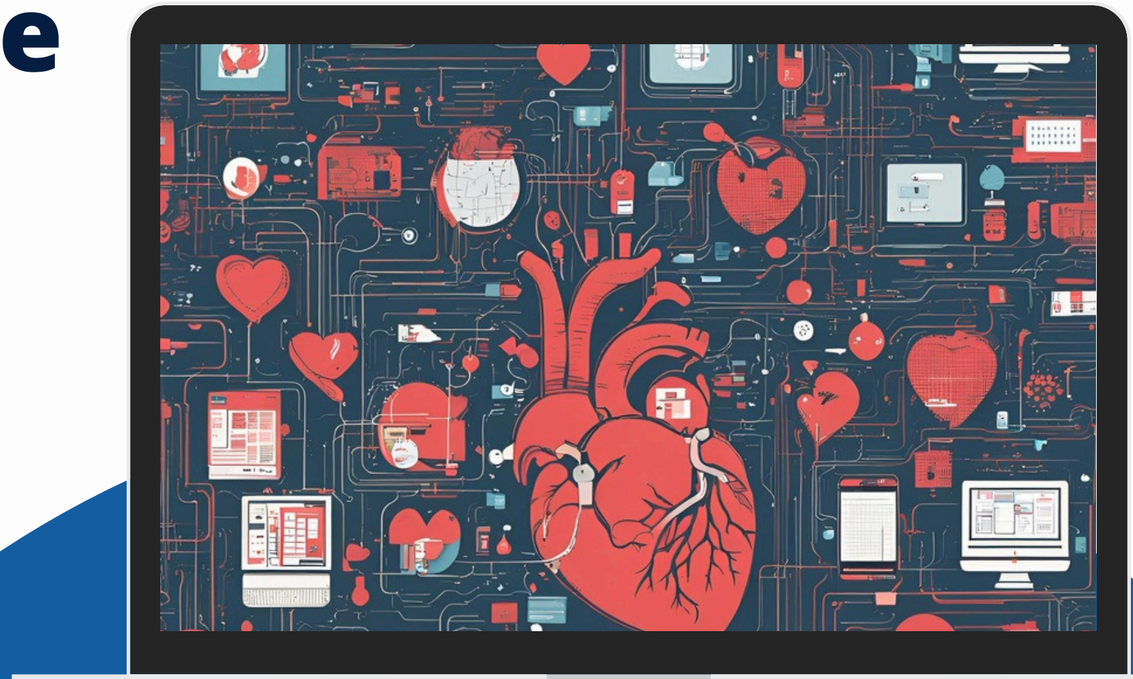
JCT
LEV ACADEMIC CENTER

Efficient genetic K-Means clustering for health care knowledge discovery

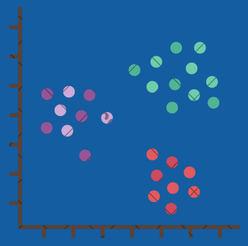
A. Alsayat and H. El-Sayed, 2016, IEEE



07/08/25

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.
- Early Disease Detection

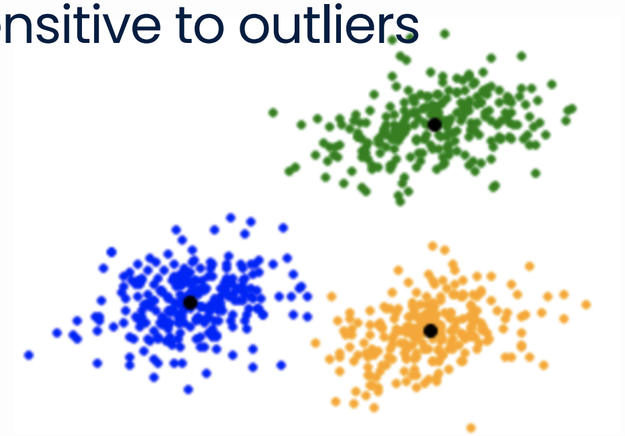
Problem Statement

- How can we improve clustering accuracy in healthcare data?
- How can we reduce sensitivity to hyperparameter selection?

Other Methods

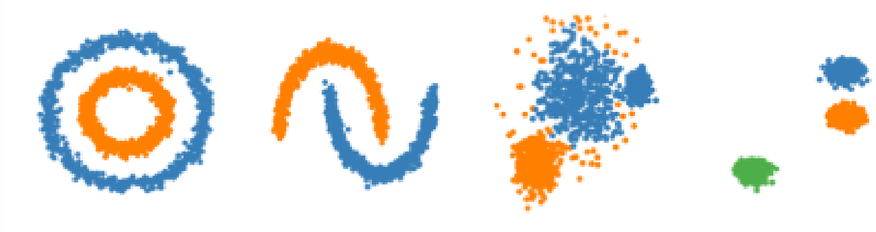
K-Means

- Unsupervised
- Groups data into k clusters by assigning points to the nearest centroid and updating centroids repeatedly.
- Limitations:
  1) Hard to determine the optimal k,
  2) Sensitive to initial centroid placement
  3) Sensitive to outliers



DBSCAN

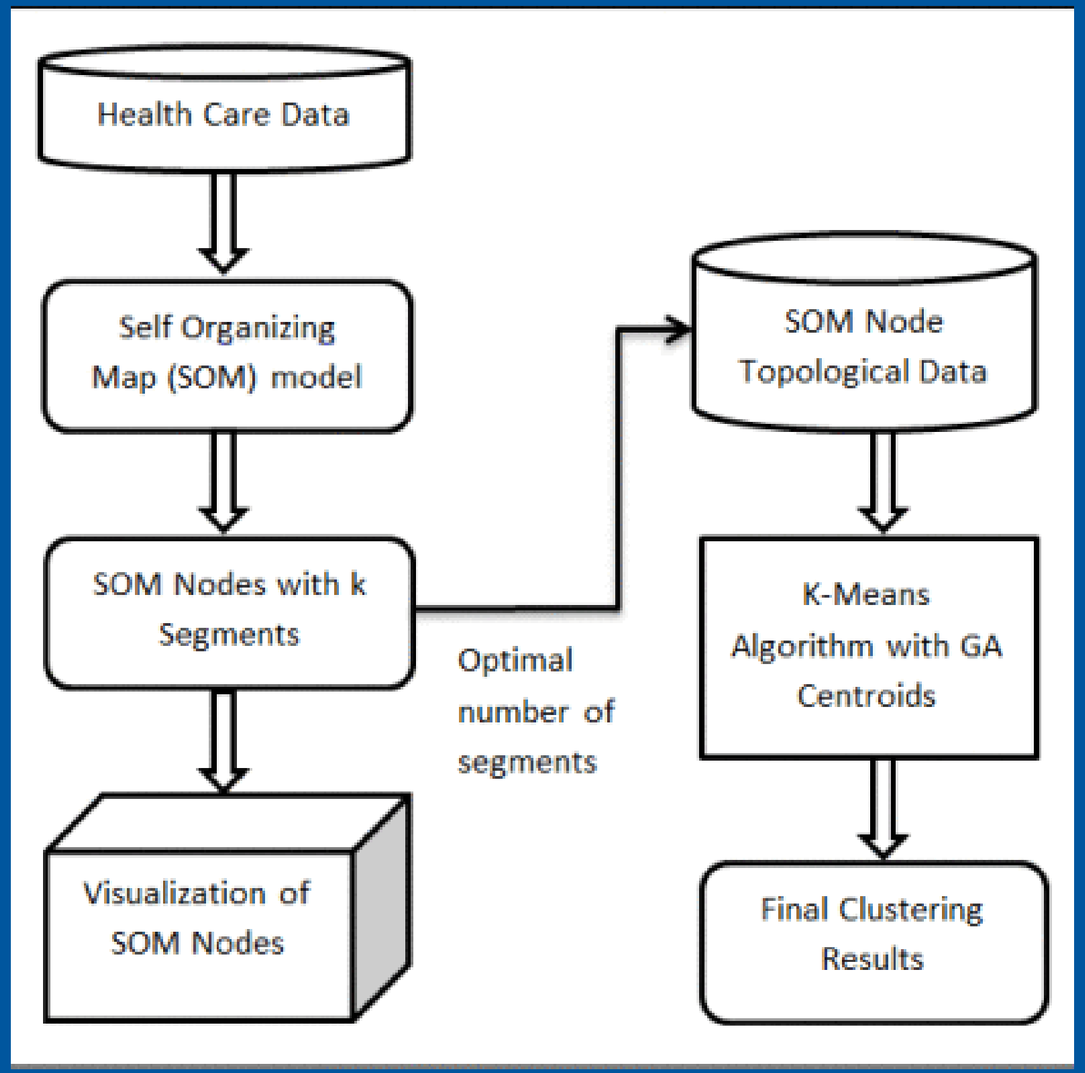
- Unsupervised
- Requires parameters: ε (neighborhood radius) and MinPts (minimum points)
- Limitations:
  1) Hard to determine the parameters,
  2) Clusters must have uniform density



Methods

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

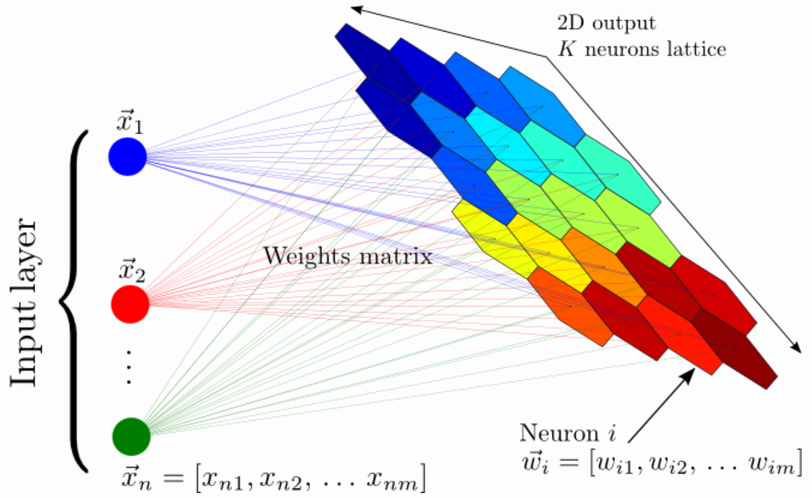
"We propose an **efficient K-Means** clustering algorithm which uses the **SOM** method to discover the optimal segments number in the data as a preprocessing step"



Solution

Self Organizing Map

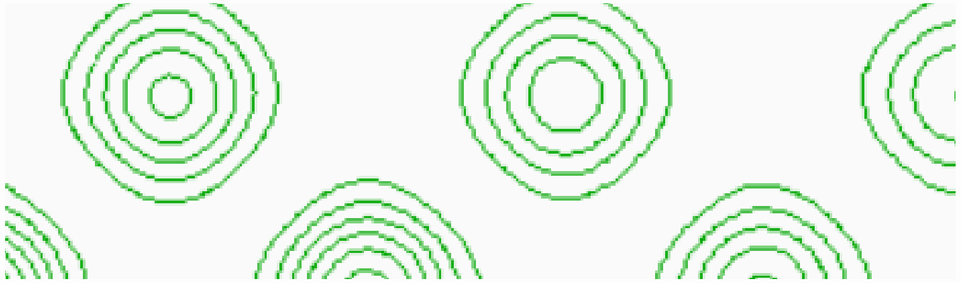
- Unsupervised Artificial Neural Network
- Maps high-dimensional data onto a lower-dimensional grid
- Groups similar data points close together to preserve topology and reveal natural clusters.
- Estimates the number of meaningful clusters (k) based on node density



Solution

Genetic Algorithm

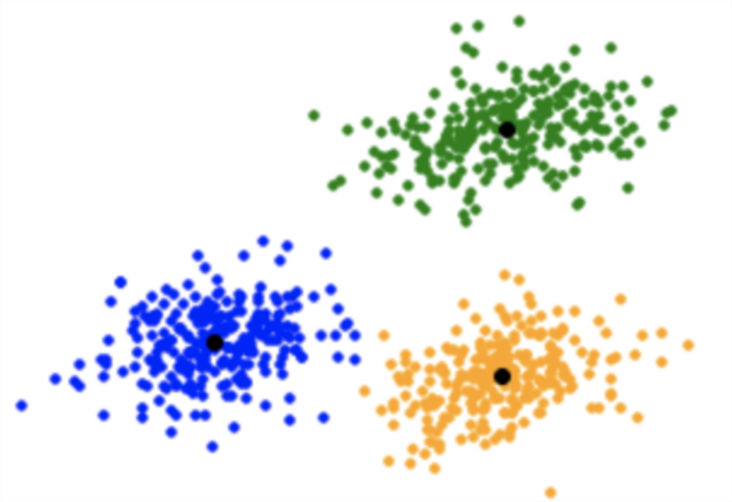
- Unsupervised evolutionary algorithm
- Starts with SOM vectors as initial guidance
- Explores many possible cluster center combinations
- Selects the best centers by minimizing distance to data points



Solution

Optimized K-Means

- After identifying the optimal number of clusters using SOM
- After estimating the initial cluster centers with a Genetic Algorithm
- we applied K-Means to achieve improved results



Preprocessing Heart

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

heart_cleaned.csv

```
age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,
slope,ca,target,thal_fixed,thal_normal,thal_reversible
63,1,1,145,233,1,2,150,0,2.3,3,0,0,True,False,False
67,1,4,160,286,0,2,108,1,1.5,2,3,1,False,True,False
67,1,4,120,229,0,2,129,1,2.6,2,2,0,False,False,True
```

- 303 patients
- 13 variables related to heart disease diagnosis
- came with target
- “thal”: One-hot encoding and dummy variables

from the Cleveland Clinic Foundation

Preprocessing Liver

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

```
1   mcv,alkphos,sgot,sgpt,gammagt,target
2   92,93,28,22,123,1
3   86,77,19,25,18,0
4   88,74,25,31,15,0
5   92,67,14,15,14,1
```

- 345 patients ->341 (removeing duplicates)
- 5 blood test variables related to liver disease, and how much cups of alcohol
- target: drinks more than 5 is sick
- deleted ‘selector’
- shuffled

from UCI Machine Learning Repository

Selector

Article

B. Cluster Evaluation

As explained in Section IV-B, a weighted average accuracy was determined for the clustering output using class labels. For example, in the liver disease dataset, the variable "selector" is the class which tells whether patient is normal or infected. To evaluate each cluster, we presume that each cluster represents the most frequent class present in it, and then the accuracy

~~selector~~

Data

Important note: The 7th field (selector) has been widely misinterpreted in the past as a dependent variable representing presence or absence of a liver disorder. This is incorrect [1]. The 7th field was created by BUPA researchers as a train/test selector. It is not suitable as a dependent

drinks > 5

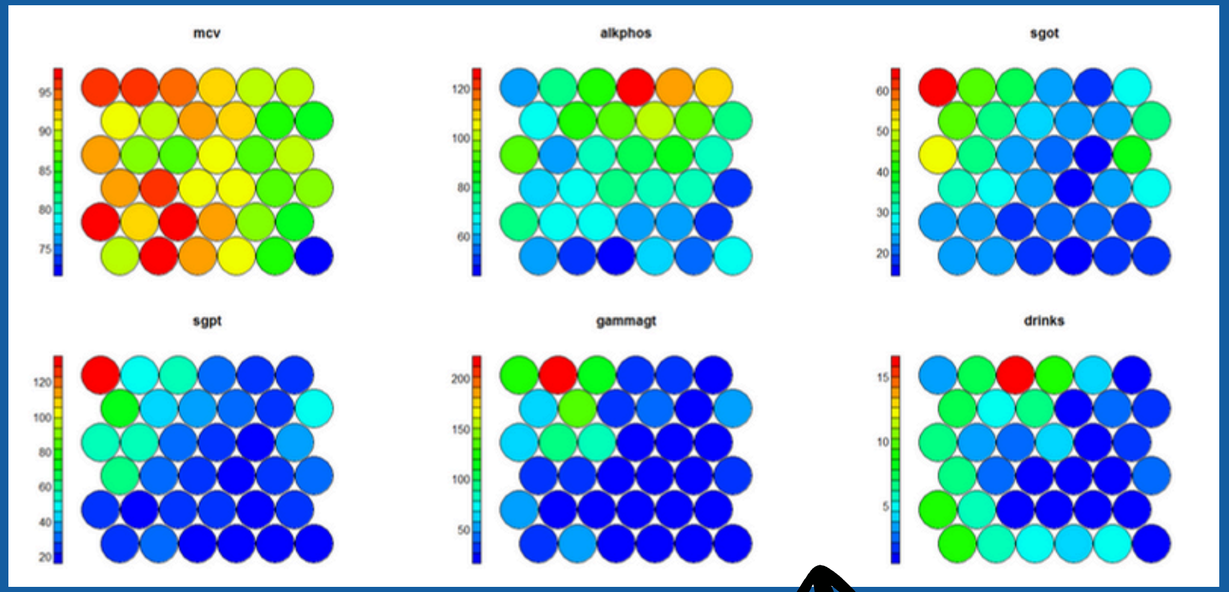
```
It appears that drinks>5 is some sort of a selector on this database.
```

SOM + Genetic K-Means

Fig 2: Heatmaps

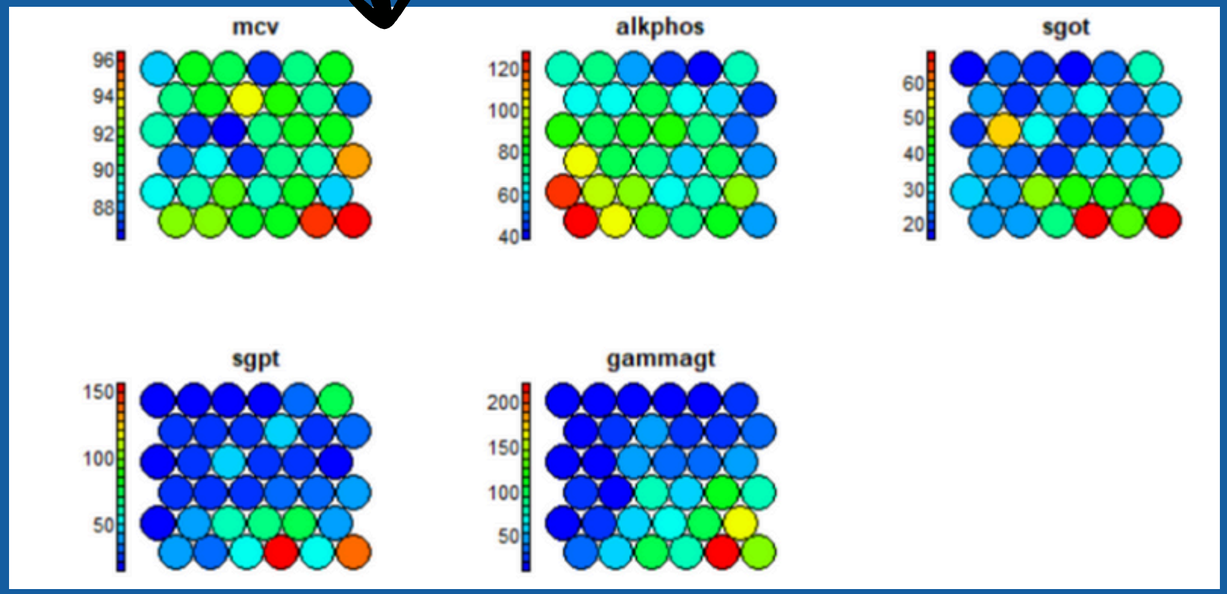
341/10 = ~35 -> 6x6 = 36 neuron

- Shows how features differ across clusters
- Reveals relationships between variables



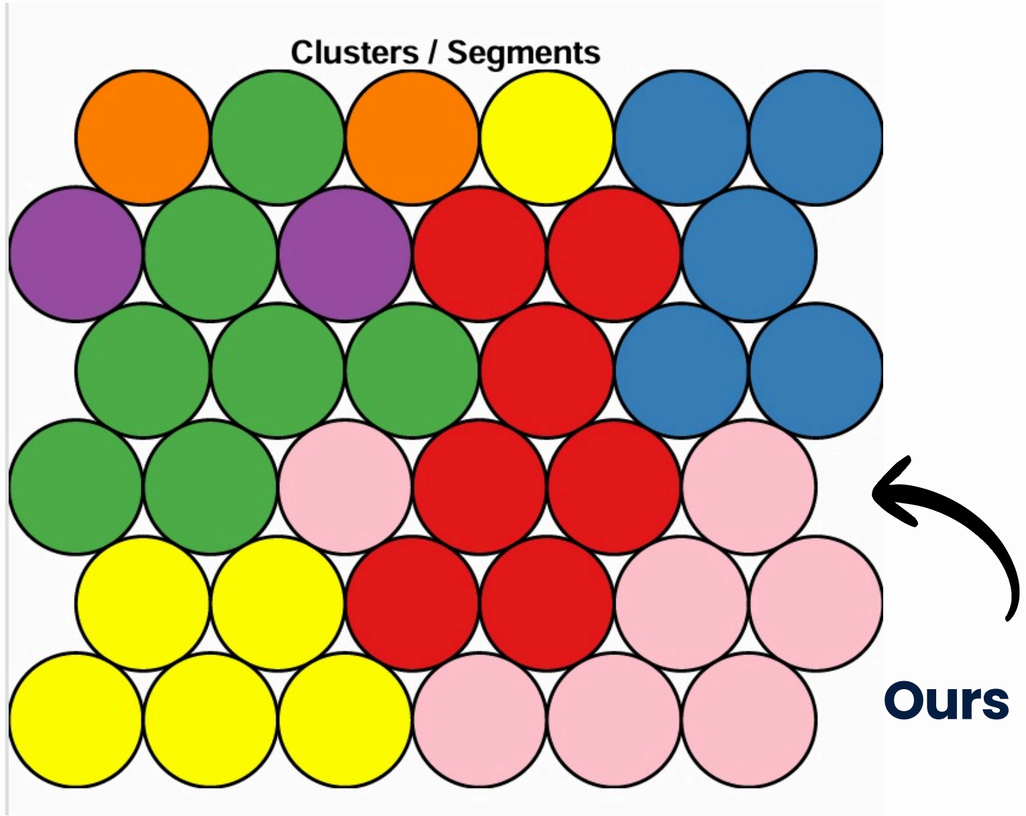
Theirs

Ours



SOM + Genetic K-Means

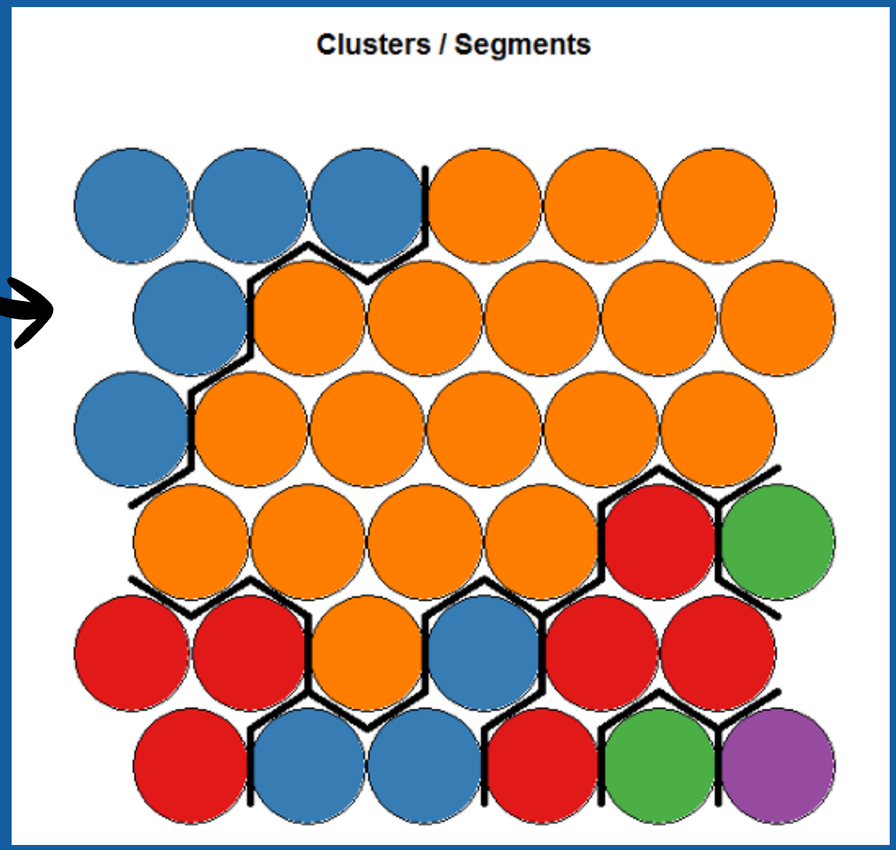
Fig 3: Clusters



Ours

7 clusters

Theirs

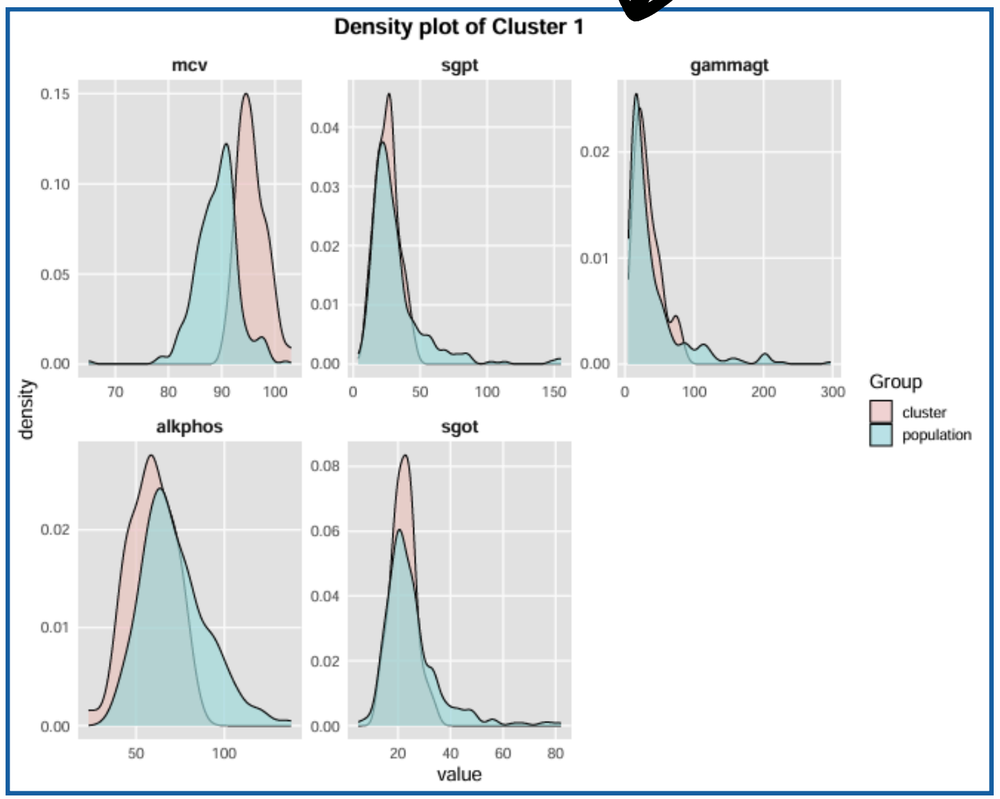


5 clusters

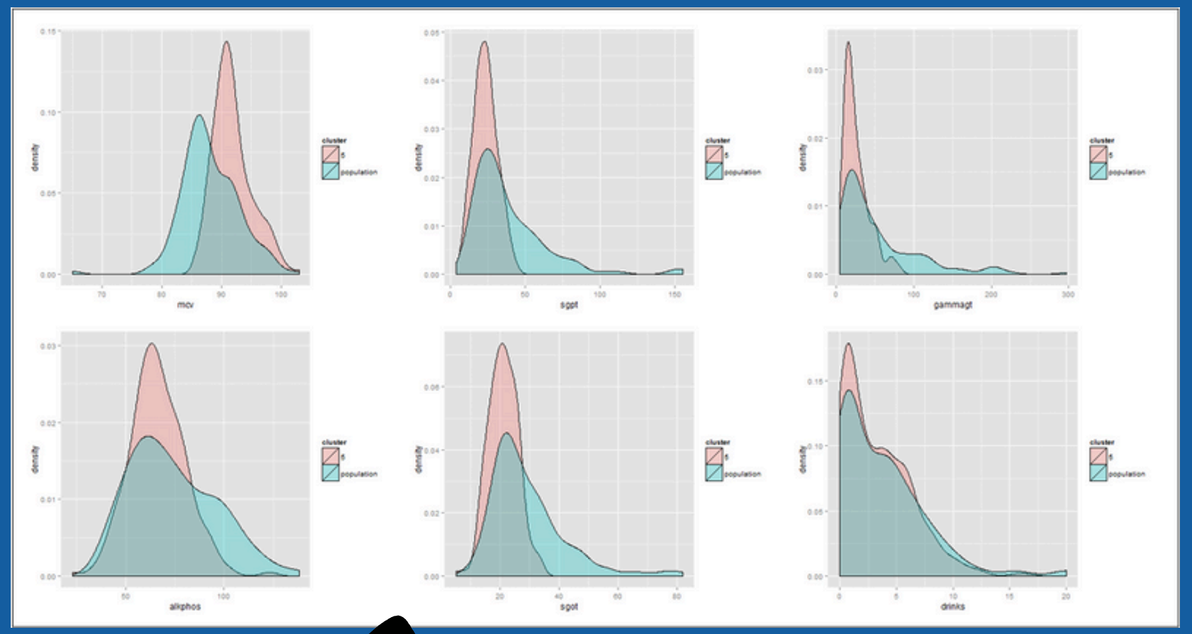
SOM + Genetic K-Means

Fig 4: Density plot

Ours

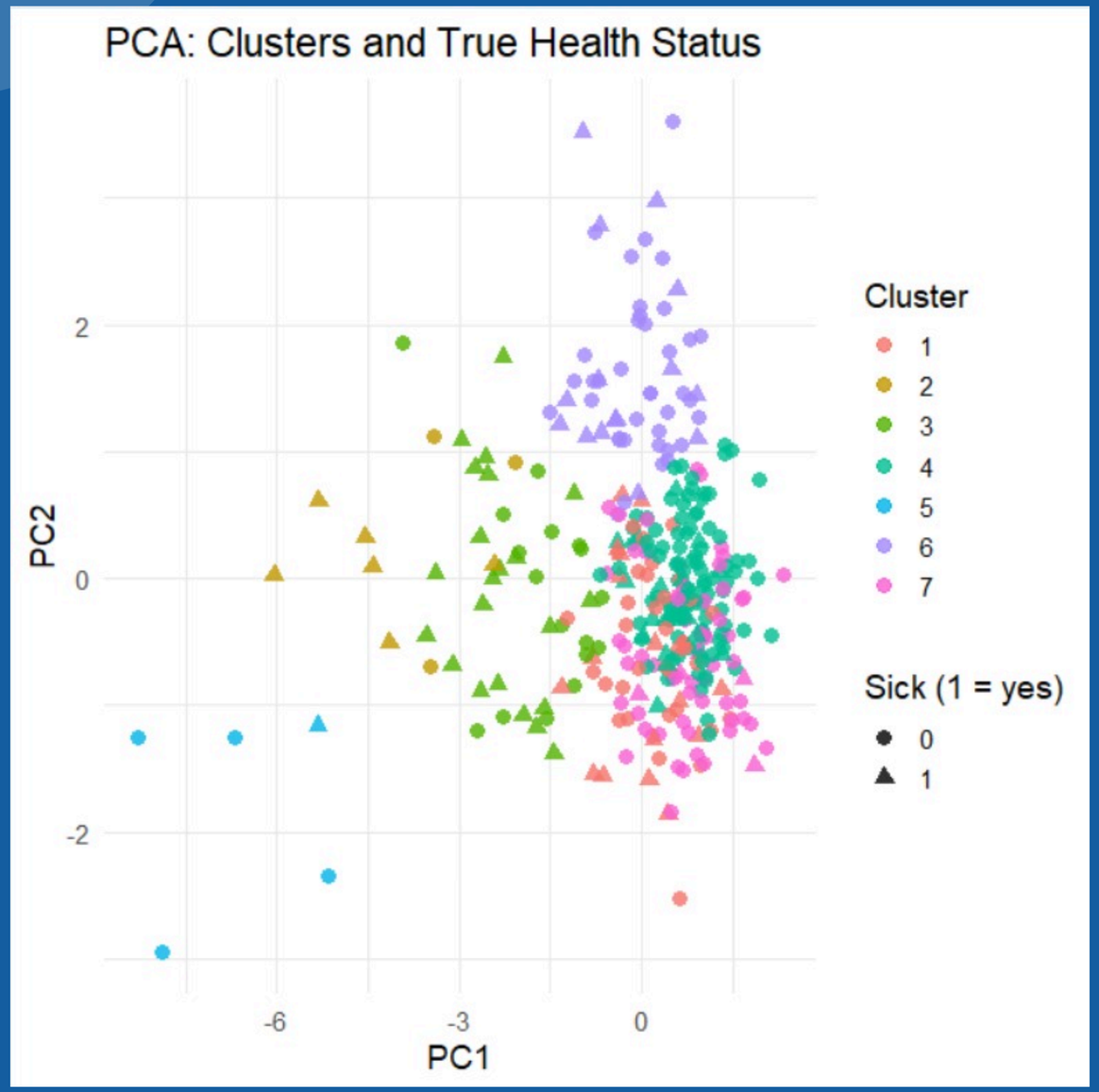


Theirs



Highlight what makes cluster 1 unique compared to the other clusters.

Results



	Cluster	Total	Sick_Count	Sick_Percent
	<int>	<int>	<int>	<dbl>
1	2	9	6	66.7
2	3	39	22	56.4
3	1	52	16	30.8
4	6	53	15	28.3
5	5	5	1	20
6	4	119	19	16.0
7	7	64	6	9.38

Results

$$\sum_{i=1}^{k}(n_i * Accuracy))/N$$

Theirs

Dataset	Weighted Classification Accuracy (%)		
	SOM Genetic K-Means	K-Means	DBSCAN
Liver Disease	**73.84** (k=5)	69.15	67.66
Heart Disease	**69.90** (k=4)	66.27	61.45

Dataset	Weighted Classification Accuracy (%)		
	SOM Genetic K-Means	K-Means	DBSCAN
Liver Disease	**77.42%** (Estimated k=7)	69.86% (k=40)	66.87% (eps=1.4, min_samples=1)
Heart Disease	**83.39%** (Estimated k=6)	82.39% (k=20)	78.74% (eps=3, min_samples=1)

Ours

SOM–Genetic K-Means achieved higher accuracy, leading to more accurate treatment suggestions.

Project Breakdown

Dataset Preparation	• Find and download the datasets. • Clean data.
Baseline Clustering	• Run K-Means clustering- Choose appropriate k values. • Run DBSCAN- Find eps and min_samples.
Advanced Clustering	• Implement Self-Organizing Maps (SOM)- Train SOM on the dataset. • Apply Genetic K-Means- Use SOM output as input for clustering.
Evaluate performances	• Compare K-Means, DBSCAN and SOM-Genetic-Kmeans's results. • Generate visualizations- using graphs and plots.
Present conclusions	• Summarize findings and write conclusions

Bibliography

01

THE ARTICLE:

A. Alsayat and H. El-Sayed, "Efficient genetic K-Means clustering for health care knowledge discovery," 2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA), Towson, MD, USA, 2016, pp. 45-52, doi: 10.1109/SERA.2016.7516127. https://ieeexplore.ieee.org/document/7516127

02

DATASET 1 : liver disease

03

DATASET 2: heart disease

04

Reference Article:

- Richard Forsyth and Roy Rada. 1986. Machine learning: applications in expert systems and information retrieval. Halsted Press, USA. https://dl.acm.org/doi/abs/10.5555/6736
- Erzs´ ebet Mer´enyi, Michael J Mendenhall, and Patrick ODriscoll. Ad vances in self-organizing maps and learning vector quantization

THANK YOU!