# Data Pre-processing and Clustering Algorithm for Epidemic Disease Diagnosis Data

Yaoyao Sang

University of Jinan
1034168135@qq.com

Authors: Yaoyao Sang Lianjiang Zhu
Tao Du Shouning Qu

# short resume

- Name: Yaoyao Sang
- Age: 25
- Hobby: yoga、badminton
- Institution: University of Jinan
- Research direction: Big data analysis and mining

# About workgroup

## research interest

- Data analysis and system implementation of thermal power company
- Research on the evaluation system of universities
- Clustering algorithm innovation
- Research on the application of knowledge mapping

CONTENTS

PART
01

# Introduction

# Background


Up to now, tuberculosis is still a public health problem which seriously endangers people's physical and mental health. It is very easy to spread through droplets and has high infectious rate.

The application of big data analysis and mining in the medical field includes many directions, such as individual and group medical planning, disease management, remote patient monitoring and so on. For health care workers, big data analysis and application is conducive to the prevention and treatment of epidemics. therefore, it is of practical significance to the intelligent processing and data mining of medical data.

PART
02

# Related works

# Related works

**classification attributes** were converted into numerical attributes by the transformation method of spherical coordinates, and k-means algorithm was used to cluster the results.

Giannotti proposed a Trk-means algorithm based on Jaccard distance. However, there is no further analysis on the convergence of the algorithm.

David G. and Averbuch A. converted numerical attributes into classification attributes through CH index, and used spectral clustering to process datasets.

**DPC algorithm** was proposed by Alex Rodriguez and Alessandro Laio in 2014 and published in Science. The article "Clustering by fast search and find of density peaks" in Science mainly focuses on a kind of clustering method based on density.

PART
03

# Medical data pre-processing

# Medical data pre-processing

- **Data cleaning**

  delete or Lagrange interpolation $L(x) = \sum_{i=0}^{n} y_i \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}$

- **Entity recognition**

  " first diagnostic area" in city A table and the "first diagnostic unit" in city B table → “first diagnostic unit”

- **0-1 unique heat coding**

  gender (male, female)

- **maximum minimum normalize**

$$x^* = \frac{x - \min}{\max - \min}$$

- **Weighted numerical**

$$w_i = \frac{|\bar{z} - z_i|}{\bar{z} + z_i}, \quad R_i = \begin{cases} r + w_i, r = 0.5 \\ r - w_i, r = 1.0 \end{cases}$$

# Medical data pre-processing

## Weighted numerical

$$w_i = \frac{|\bar{z} - z_i|}{\bar{z} + z_i}, \quad R_i = \begin{cases} r + w_i, r = 0.5 \\ r - w_i, r = 1.0 \end{cases}$$

1. Divide the location data into four levels, according to the administrative region from the largest to the smallest.(provinces, cities, regions and counties)
2. The number of patients in each hospital is used as the weighted factor.
3. Use the weighting formula to calculate the value of each location data.

| Original geographic location | The initial rating value r | Number of visits z | Weighted value W | Numerical result R |
|---|---|---|---|---|
| People's hospital of A province | 0.5 | 20 | 0.975848327 | 1 |
| B city people's hospital | 0.5 | 1239 | 0.138146912 | 0.638146912 |
| County C center for disease control and prevention | 1 | 1644 | 0.002377904 | 0.997622096 |
| District D TB control station | 1 | 2337 | 0.176381758 | 0.823618242 |

Figure1. Numerical table of different geographical location information

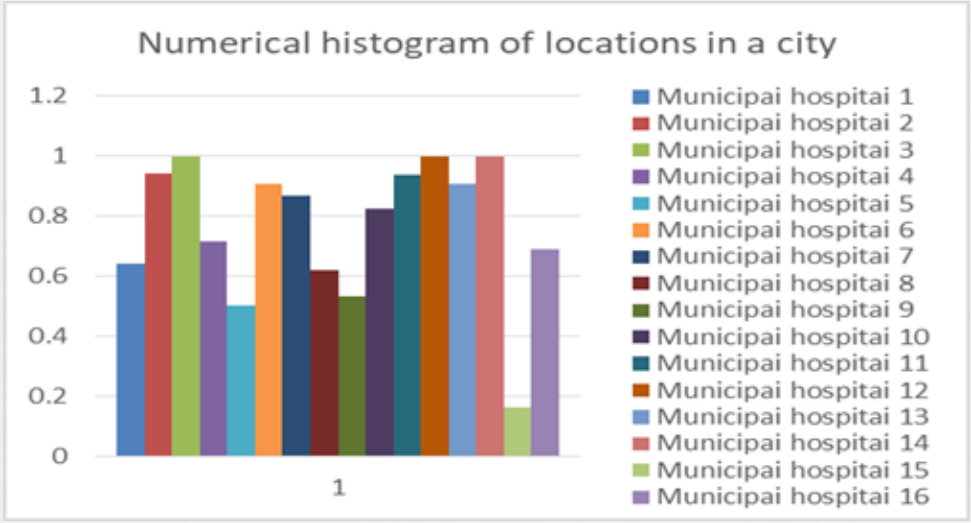


Figure2. Numerical histogram of locations in a city

PART
04

# Density Peak Clustering Algorithm

# Density Peak Clustering Algorithm

DPC has two main quantities to calculate

- the local density ρ
- the distance delta from the point of high density

HIGH ⇒ Cluster center

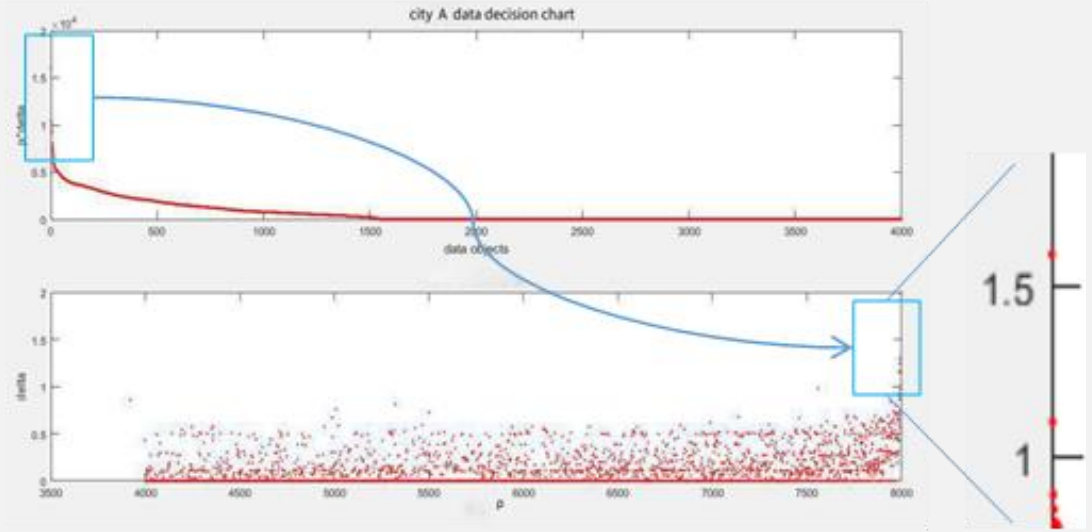


Figure3. Data decision chart of city A

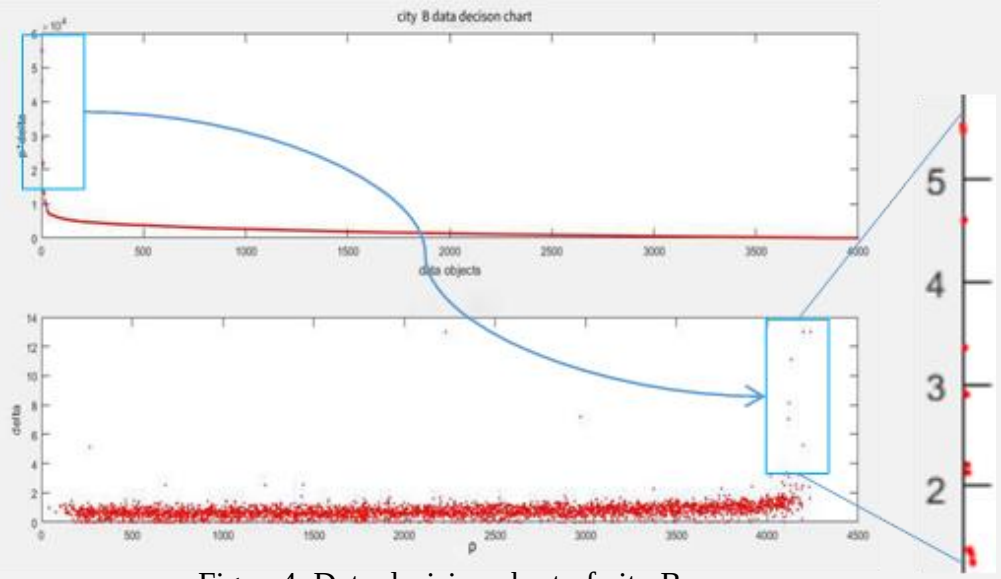


Figure4. Data decision chart of city B

Figure 3 and Figure 4 are the decision diagram obtained by the density peak algorithm. The data points in the two boxes are corresponding, representing the center point in the cluster.

# Density Peak Clustering Algorithm

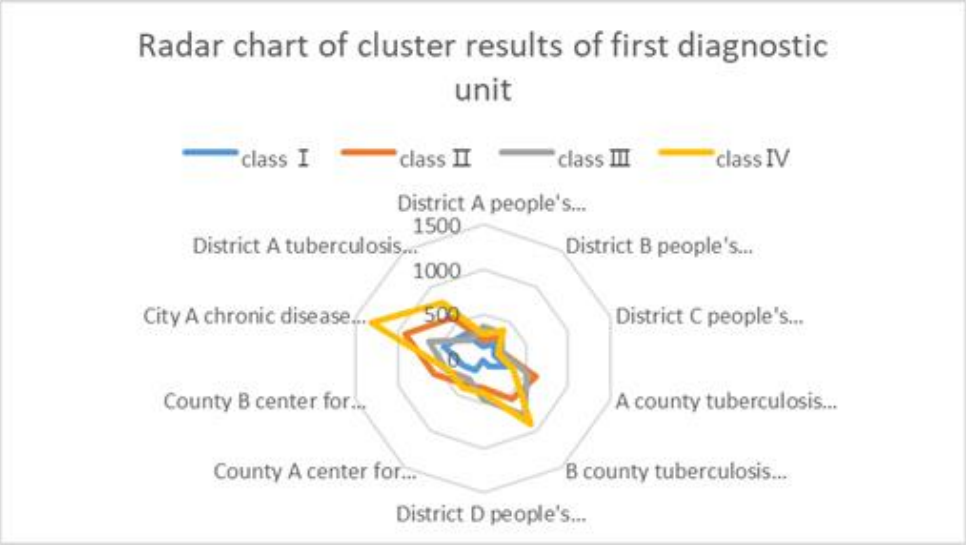


Figure. 6. Cluster radar map of the first diagnostic unit in a city

Figure 6 is the radar map of the first diagnosis unit after clustering in a certain city. It can be intuitively seen that among the four clusters, the two medical units with the largest number of patients are City A chronic disease prevention and treatment station, the other is B county tuberculosis prevention and treatment institute.

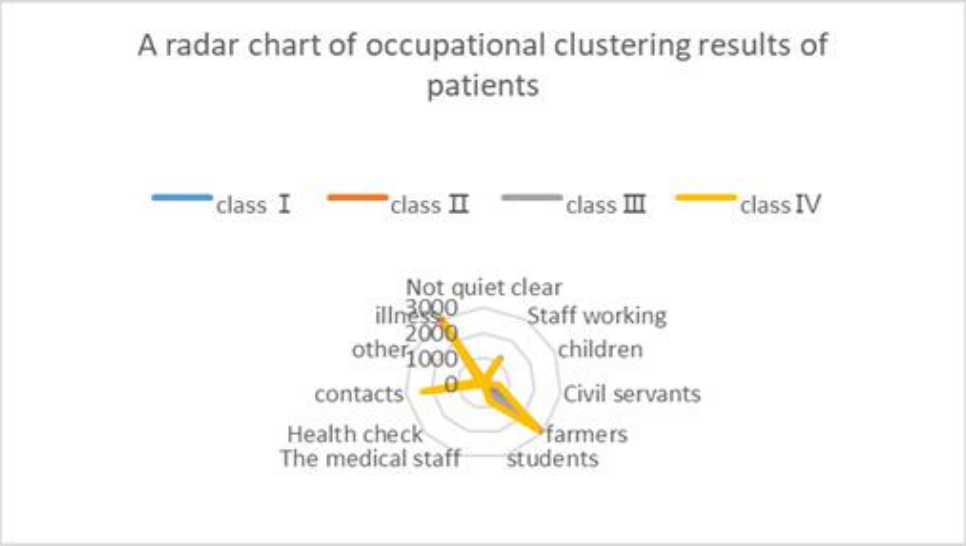


Figure. 7. Cluster radar chart of patients' occupation and reasons for seeking medical treatment in a certain city

Figure 7 shows the occupational radar of patients and the reason for seeking medical treatment. It can be clearly seen that patients are mostly ordinary farmers, and they go to see a doctor after contact and infection

PART
05

# Conclusion

# Conclusion

In this paper, based on the treatment data of pulmonary tuberculosis patients in six urban areas, a data pre-processing method was proposed according to the characteristics of multiple attributes and non-standard.

Moreover, a method to highlight the differences in the weighted processing of geographical location classification data was innovatively put forward, the original mixed datasets processing as a single numeric datasets.

After that, the DPC algorithm was utilized to clustering data to obtain some valuable information to the patient, such as age or occupation distribution. Cluster centers and outliers was obtained by analyzing the decision graph.

# Future work


In future work, knowledge discovery method will be introduced, and association rules will be analyzed on the basis of clustering to discover tuberculosis treatment rules or other hidden knowledge.


# Thanks for listening