MATHEMATIQUES APPLIQUEES de BORDEAUX

INRIA

Dynamical clustering of interval data : Optimization of an adequacy criterion based on Hausdorff distance

Marie Chavent – Yves Lechevallier

The aim : clustering interval data

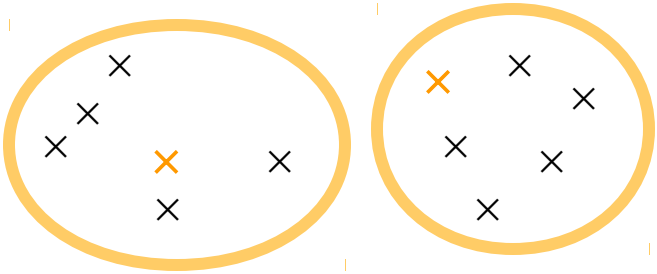
	Pulse Rate	Systolic pressure	Diastolic pressure
1	[60, 72]	[90,130]	[70,90]
2	[70,112]	[110,142]	[80,108]
3	[54,72]	[90,100]	[50,70]
4	[70,100]	[130,160]	[80,110]
5	[63,75]	[60,100]	[140,150]
6	[44,68]	[90,100]	[50,70]

Each object *i* is described

- on each variable *j* by an interval :
- by a vector of intervals (hyper-rectangle) :
- by a symbolic object :

The clustering algorithm : dynamical clustering

Minimization problem :



• the algorithm converges
• partitionning criterion decreases

• define a **distance** between vectors of intervals
• define the **class prototype** *y* which minimizes the new adequacy criterion

A distance measure between two vectors of intervals

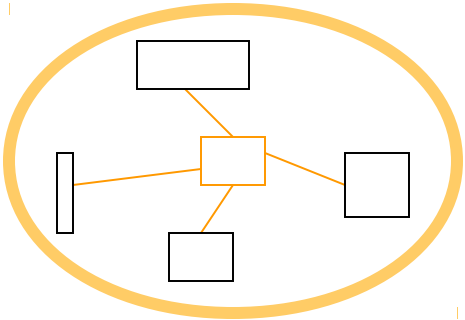
- the Hausdorff distance between **two sets**
- the Hausdorff distance between **two intervals**
- a distance measure between **two vectors of intervals**

Remarks :

- in the particular case of intervals reduced to single points, *d* is the L1 distance

- The distance *d* is **based on** the Hausdorff distance but is not the Hausdorff distance between two vectors of intervals i.e. hyper-rectangles

The optimization problem for class prototype



Search the class prototype :

which minimizes the adequacy criterion :

Search for j=1,…,p the interval :

which minimizes :

The solution is :

Let :

respectively the midpoints and the half length of the intervals

then :

The dynamical clustering algorithm :

Proceeds by iteratively repeating two steps:

‘allocation’ step : define a new partition by reassigning each object i to the closest class prototype $\boldsymbol{y_k}$

‘representation’ step : compute for each cluster $\boldsymbol{C_k}$, the prototype $\boldsymbol{y_k}$ with :

Long-Term Instrumental Climatic Data Bases

- published by the Institute of Atmospheric Physics of the Chinese Academy of Sciences, Beijing, China.
- free and archived on the site

http://dss.ucar.edu/

subdirectory :

`http://dss.ucar.edu/datasets/ds578.5/.`

- contains mean, minimum, maximum, and total meteorological values measured each month on 13 variables in 60 chinese stations

From the Database …

The table contains the data from these 60 stations.

Each record contains : the WMO station number, year, and month, followed by 13 meteorological variables:

Mean Maximum Temperature (C)

Mean Minimum Temperature (C)

Total Precipitation (mm)

Sunshine Duration (hours)

Mean Cloud Amount (percentage of sky cover)

Mean Relative Humidity (percent)

now Days (days with snow cover)

Mean Station Pressure (mb)

Dominant Wind Direction (degrees)

Mean Wind Speed (m/s)

Dominant Wind Frequency (percent)

Extreme Maximum Temperature (C)

Extreme Minimum Temperature (C)

…to interval data

ChangSha station and Year 1988 :

[January = [2.7:7.4]] ^[February = [3.1:7.7]]

^[March = [6.5:12.6]] ^[April = [12.9:22.9]]

^[May = [19.2:26.8]] ^[June = [21.9:31]]

^[July = [25.7:34.8]] ^[August = [24.4:32]]

^[September = [20:27]] ^[October = [15.3:22.8]]

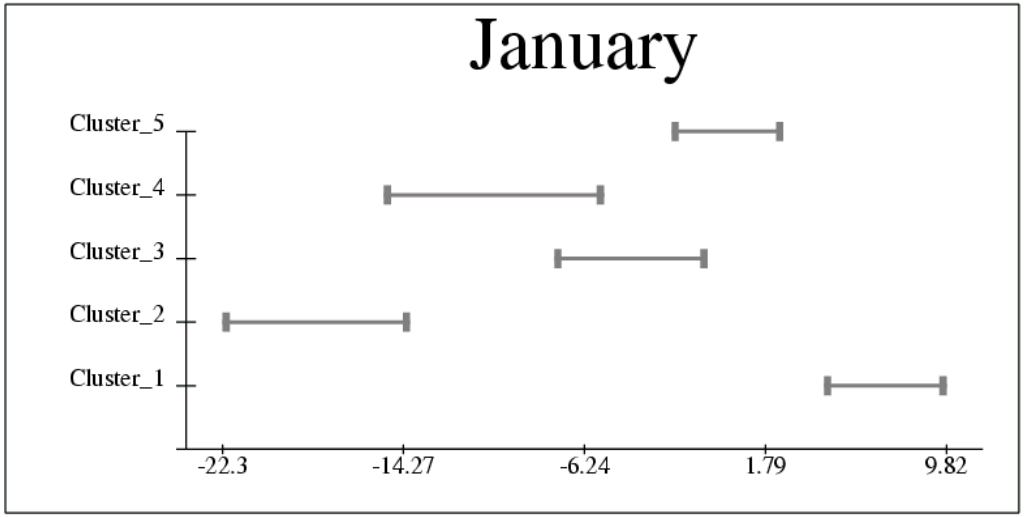
^[November = [7.6:19.6]] ^[December = [4.1:13.3]]

	January	…	December
ShangSha	[2.7:7.4]	…	[4.1:13.3]
…			
WuLu	[-12.3;-8.5]	…	[-11.5:-7]

60 stations


HEILONGJIANG
Harbin
Changchun
JILIN
BEIJING
Shenyang
LIAONING
Urumqi
XINJIANG
GANSU
Hohhot
HEBEI
INNER MONGOLIA
TIANJIN
Yinchuan
Taiyuan
Shijiazhuang
Xining
NINGXIA
SHANXI
Jinan
QINGHAI
SHANDONG
Lanzhou
Xian
Zhengzhou
HENAN
JIANGSU
TIBET
SHAANXI
Hefei
Nanjing
SHANGHAI
SICHUAN
Lhasa
Chengdu
Wuhan
ANHUI
Hangzhou
HUBEI
CHONGQING
ZHEJIANG
Changsha
Nanchang
Guiyang
HUNAN
Fuzhou
JIANGXI
GUIZHOU
FUJIAN
Taipei
Kunming
YUNNAN
GUANGXI
GUANGDONG
TAIWAN
Nanning
HONG KONG
MACAU
GuangZhou
Haikou
HAINAN

Partition in 5 clusters

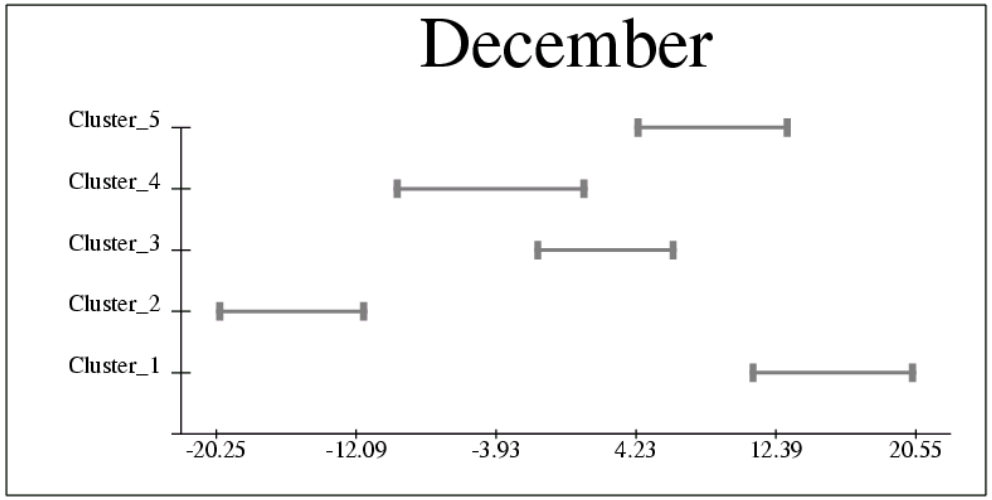


Prototypes

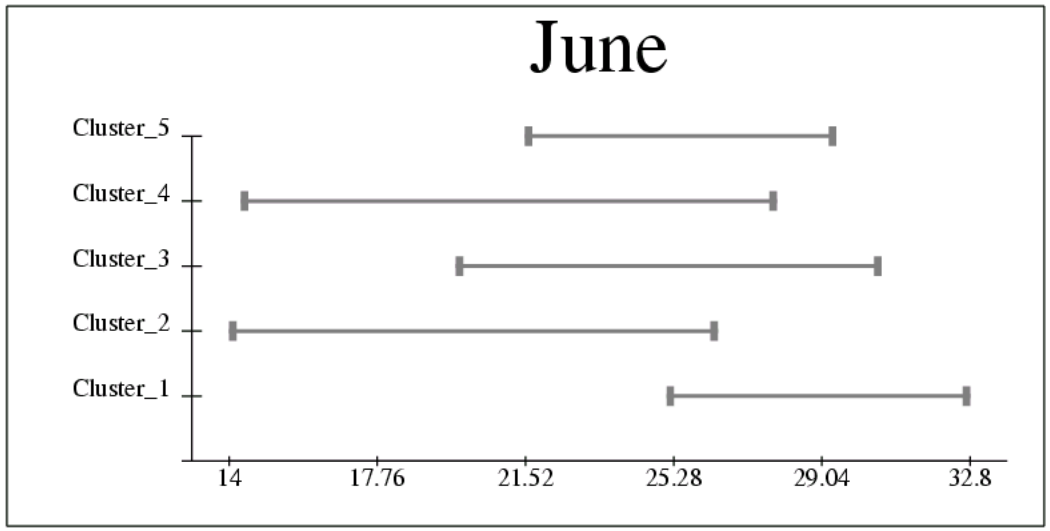
by clusters

Discriminant

months



Partition in 5 clusters

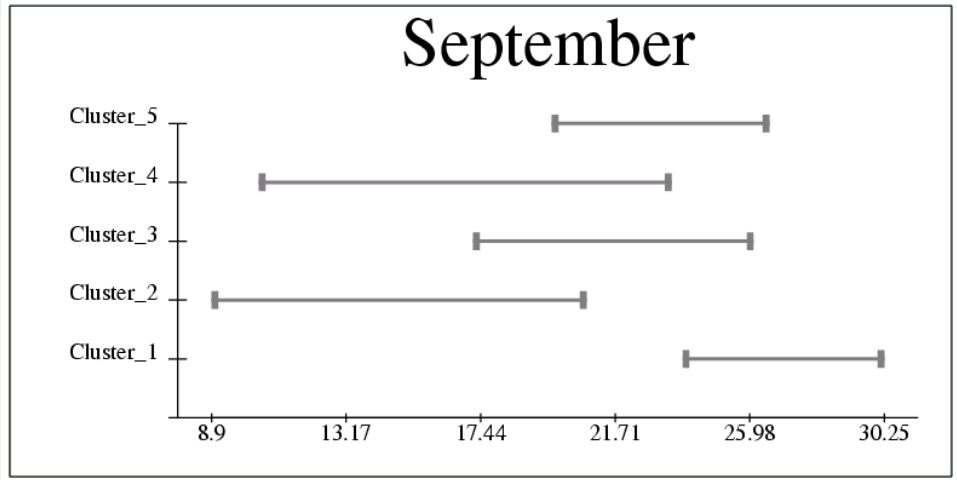


Prototypes

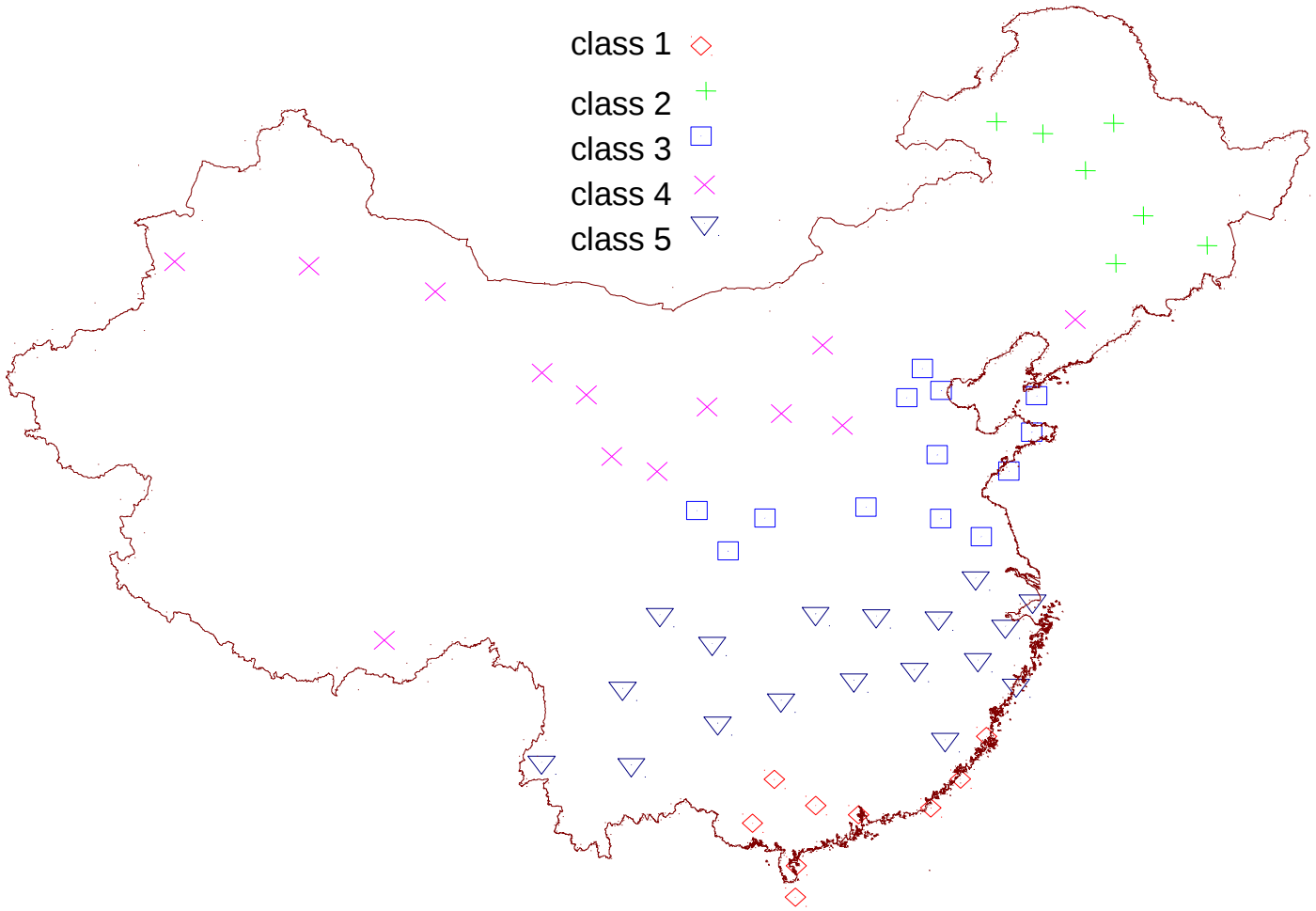
by clusters

Some other

months



Partition in 5 clusters




class 1
class 2
class 3
class 4
class 5
HEILONGJIANG
Harbin
Changchun
JILIN
BEIJING
Shenyang
LIAONING
TIANJIN
Urumqi
XINJIANG
GANSU
Hohhot
HEBEI
INNER MONGOLIA
Yinchuan
Taiyuan
Shijiazhuang
Xining
NINGXIA
SHANXI
Jinan
SHANDONG
QINGHAI
Lanzhou
Xian
Zhengzhou
HENAN
JIANGSU
TIBET
SHAANXI
Hefei
Nanjing
SHANGHAI
SICHUAN
Lhasa
Chengdu
Wuhan
ANHUI
Hangzhou
HUBEI
CHONGQING
ZHEJIANG
Changsha
Nanchang
Guiyang
HUNAN
Fuzhou
JIANGXI
Kunming
GUIZHOU
FUJIAN
Taipei
YUNNAN
GUANGXI
GUANGDONG
TAIWAN
Nanning
HONG KONG
MACAU
GuangZhou
Haikou
HAINAN

Conclusion

This clustering algorithm :

- deals with interval data
- converges and the partitionning criterion decreases at each iteration
- simple to implement
- Computanional complexity