Outline

• Motivation
• Idea in the proposed approach
• Data
  – Useful, Form, Trend
• Classification
  – Improved.
• Cluster
  – Fuzzy c-mean cluster
  – Euclidean distance
• Experimental Results
• Future work
  – Static state Kalman Filter
Motivation (1/3)

• Nowadays, the tech of traditional forecast.
  – Weather chart.
  – Satellite image.
  – The knowledge of synoptic meteorology.
  – …, etc.

• But…, our ancestor did not have this tech in the past.
Zhuge Liang, also known as Kong Ming.

Once said that, “We can use the characteristics and experiences of weather. The weather was blowing southeast wind in this week, and it might change to east wind next week.”

So… They used the arrow with fire, when the wind blow, Archery!

*Picture from google image.
Once said that, "We can use the characteristics and experiences of weather. The weather was blowing southeast wind in this week, and it might change to east wind next week."

Zhuge Liang, also known as Kong Ming. So…

They used the arrow with fire, when the wind blow, Archery!

Finally, he forecast successfully. Win the ChiBi war.

*Picture from google image.
Motivation (3/3)

• We want to achieve this history event without tech of traditional forecast, by past histories on data mining.
  – By past temperature histories to predict future.
  – Try to use past temperature data on prediction.
  – Using the concept of data mining on weather forecast.
Idea in the proposed approach (1/2)

• The weather will be similar in a period of time.

• Predict the temperature of Taipei in 2015.
  – Using past histories.

• Provide a different methodology on weather forecast.
  – Using trend and other concepts.
  – Using similar patterns in near cities.
Idea in the proposed approach (2/2)

• The temperature of Taipei is affected by some near cities.
  – In summer
    • Affected by southwesterly airstream.
  – In winter
    • Affected by continental air mass.

• Time shift in all data.
Data useful

- Data mining.
  - Different database from 2008 to 2014.
  - The influence of regional cities.

- Taipei: Only the city of Taipei.
- 19 Cities: 18 cities of China, and Taipei.
- Japan: 12 cities of Japan.
- Japan + 19 Cities
*Linear regression
(Red) Database: 18 cities in China form 2008-2014
(Blue) Database: 14 cities in China form 2008-2014

This map is from https://upload.wikimedia.org/wikipedia/
We asked from Department of Atmospheric Sciences.
  – Forecast four days temperature.

The weather will be similar in a period of time.

Detail
  • RT: Real Temperature
    – First day, to avoid summer pattern match winter pattern.
  • DT: Delta Temperature
    – Using trend concept on it, comparing with yesterday and today.
    – Second day to sixth day temperature.
Data form (2/3)

• In the database, all sequences are in this form and the sequence shifts for every day as shown in January.

<table>
<thead>
<tr>
<th>Data count</th>
<th>Day01</th>
<th>...</th>
<th>Day06</th>
<th>Day07</th>
<th>...</th>
<th>Day10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-Jan</td>
<td>...</td>
<td>6-Jan</td>
<td>7-Jan</td>
<td>...</td>
<td>10-Jan</td>
</tr>
<tr>
<td>2</td>
<td>2-Jan</td>
<td>...</td>
<td>7-Jan</td>
<td>8-Jan</td>
<td>...</td>
<td>11-Jan</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>25-Jan</td>
<td>...</td>
<td>30-Jan</td>
<td>31-Jan</td>
<td>...</td>
<td>3-Feb</td>
</tr>
<tr>
<td>26</td>
<td>26-Jan</td>
<td>...</td>
<td>31-Jan</td>
<td>1-Feb</td>
<td>...</td>
<td>4-Feb</td>
</tr>
</tbody>
</table>
Data form (3/3)

• Data shifts.
  – Every data all in need in our research.
  – A period of data form may repeat at sometime.
    • Correlation.

• Total count of data, data set, for Data shifts.
  – Every year: (365-7), Year: 7(2008-2014), Cities: 19
  – 7*(365-7)*19 = 47,614
Process

- Take all data
  - One data match all?
  - Separate into unknown counts of clusters?
- Set tendency (trend) then cluster.
Trend (1/2)

• Concept
  – We think there are just three situations between the temperature of yesterday and today
  – We set flat type not only at zero, as figure shown.

Three types of trend.
Trend (2/2)

- To have finer resolution in trend.
  - Seven types of trend.
  - To separate two cases on delta temperature
    - Ordinary case, from -4°C to 4°C
    - Dramatic change case, more than 4°C or less than -4°C.
Clustering

• The concept of weather
  – Temperature will be similar change in a period of time.
    • Using the pattern of previous six days data then discussing the following four days.

• In previous research
  – One matrix data only match one category.
Improved trend classification (1/2)

• Because of our previous research
  – One matrix data only match one category.
    • Fewer data, Data weighting.

• Turn to…
  – One matrix data matches $32(2^5)$ categories at least.
    • One delta temperature match two categories.
  – Seven types of trend.
    • Because the count of delta temperatures is five, the total count of database categories is $16807(7^5)$. 
Improved trend classification (2/2)

• After all database built.
  – Seven types of trend.
  – One matrix data matches $32(2^5)$ categories at least.

• Every Target data will compare with the same category in database, then into fuzzy c-mean cluster.
  – One target matches more than one category.
    • The same idea from database built.
Fuzzy c-mean cluster

• In this approach, setting four clusters in the method of Fuzzy c-mean.
  – Well known.
  – Testing one month.
    • from two clusters to ten clusters in the error of Jan. 2015, gain the better result from four clusters, no matter upper bound and lower bound.
## Results - Upper bound errors in different clusters

<table>
<thead>
<tr>
<th></th>
<th>First Day</th>
<th>Second Day</th>
<th>Third Day</th>
<th>Four Day</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Two clusters</strong></td>
<td>2.2517</td>
<td>2.5611</td>
<td>3.1520</td>
<td>3.1152</td>
</tr>
<tr>
<td><strong>Three clusters</strong></td>
<td>1.1515</td>
<td>1.9512</td>
<td>2.6481</td>
<td>2.9121</td>
</tr>
<tr>
<td><strong>Four clusters</strong></td>
<td>1.5774</td>
<td>1.5275</td>
<td>2.5166</td>
<td>2.6458</td>
</tr>
<tr>
<td><strong>Five clusters</strong></td>
<td>3.9688</td>
<td>3.2439</td>
<td>3.1951</td>
<td>2.9809</td>
</tr>
<tr>
<td><strong>Six clusters</strong></td>
<td>2.6091</td>
<td>4.5603</td>
<td>3.8953</td>
<td>3.7170</td>
</tr>
<tr>
<td><strong>Seven clusters</strong></td>
<td>3.5208</td>
<td>4.6589</td>
<td>3.2183</td>
<td>3.0936</td>
</tr>
<tr>
<td><strong>Eight clusters</strong></td>
<td>2.6165</td>
<td>2.5650</td>
<td>2.4949</td>
<td>2.1217</td>
</tr>
<tr>
<td><strong>Nine clusters</strong></td>
<td>1.9946</td>
<td>1.9890</td>
<td>2.1565</td>
<td>2.6880</td>
</tr>
<tr>
<td><strong>Ten clusters</strong></td>
<td>2.1965</td>
<td>2.9496</td>
<td>2.9161</td>
<td>2.8882</td>
</tr>
</tbody>
</table>
Results - Lower bound errors in different clusters

<table>
<thead>
<tr>
<th></th>
<th>First Day</th>
<th>Second Day</th>
<th>Third Day</th>
<th>Four Day</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Two clusters</strong></td>
<td>2.1562</td>
<td>2.1515</td>
<td>2.1516</td>
<td>2.1541</td>
</tr>
<tr>
<td><strong>Three clusters</strong></td>
<td>1.8825</td>
<td>1.6511</td>
<td>1.9315</td>
<td>2.2165</td>
</tr>
<tr>
<td><strong>Four clusters</strong></td>
<td>1.3127</td>
<td>1.2616</td>
<td>1.4298</td>
<td>2.9313</td>
</tr>
<tr>
<td><strong>Five clusters</strong></td>
<td>2.3613</td>
<td>2.5121</td>
<td>1.7713</td>
<td>2.5165</td>
</tr>
<tr>
<td><strong>Six clusters</strong></td>
<td>4.2952</td>
<td>2.8896</td>
<td>2.0617</td>
<td>2.6317</td>
</tr>
<tr>
<td><strong>Seven clusters</strong></td>
<td>5.3523</td>
<td>3.5617</td>
<td>2.2265</td>
<td>3.6550</td>
</tr>
<tr>
<td><strong>Eight clusters</strong></td>
<td>2.6113</td>
<td>2.5165</td>
<td>2.9621</td>
<td>3.5165</td>
</tr>
<tr>
<td><strong>Nine clusters</strong></td>
<td>2.1517</td>
<td>2.2152</td>
<td>2.8494</td>
<td>3.1652</td>
</tr>
<tr>
<td><strong>Ten clusters</strong></td>
<td>2.9551</td>
<td>2.4950</td>
<td>3.8491</td>
<td>3.1952</td>
</tr>
</tbody>
</table>
Euclidean distance (1/2)

- Using Euclidean distance to choose the minimal distance between four clusters center result and target.
- The Euclidean distance is defined as:
  \[ d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \]
- The minimum value is elected as the matched cluster.

![Diagram showing the process: Four cluster → Four Center → Min Distance → Result. The same category with target highlighted.](image-url)
Euclidean distance (2/2)

• In the trend classification

\[ d_{CR01} = \sqrt{\sum_{i=02}^{06} (cdt_{01i} - dt_i)^2} \]

\[ d_{CR04} = \sqrt{\sum_{i=02}^{06} (cdt_{04i} - dt_i)^2} \]

Clusterresult 01 = \[ CT_{0101} \text{ } cdt_{0102} \text{ } ... \text{ } cdt_{0110} \]

Clusterresult 04 = \[ CT_{0401} \text{ } cdt_{0402} \text{ } ... \text{ } cdt_{0410} \]

• The minimum value is elected as the matched cluster.

\[ d_{result} = \min(d_{CR01} \text{ } d_{CR02} \text{ } d_{CR03} \text{ } d_{CR04} ) \]
## Results - Predictions for different resolutions.

<table>
<thead>
<tr>
<th>Upper bound AE (SD)</th>
<th>First Day</th>
<th>Second Day</th>
<th>Third Day</th>
<th>Fourth Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>News’</td>
<td>2.41 (2.12)</td>
<td>2.46 (2.07)</td>
<td>2.52 (2.10)</td>
<td>2.54 (2.24)</td>
</tr>
<tr>
<td>Taipei</td>
<td>3.21 (4.21)</td>
<td>3.13 (4.17)</td>
<td>3.26 (4.32)</td>
<td>3.32 (4.32)</td>
</tr>
<tr>
<td>15 Cities</td>
<td>3.65 (4.52)</td>
<td>2.99 (4.55)</td>
<td>3.00 (4.52)</td>
<td>2.62 (3.85)</td>
</tr>
<tr>
<td>19 Cities</td>
<td>1.90 (2.97)</td>
<td>1.90 (2.92)</td>
<td>1.88 (2.89)</td>
<td>2.05 (3.60)</td>
</tr>
<tr>
<td>Japan data</td>
<td>3.58 (5.50)</td>
<td>4.94 (6.96)</td>
<td>3.11 (5.56)</td>
<td>3.94 (6.57)</td>
</tr>
<tr>
<td>19 Cities+ Japan</td>
<td>2.88 (3.76)</td>
<td>2.65 (3.51)</td>
<td>2.77 (3.73)</td>
<td>2.76 (3.69)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lower bound AE (SD)</th>
<th>First Day</th>
<th>Second Day</th>
<th>Third Day</th>
<th>Fourth Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>News’</td>
<td>1.89 (1.64)</td>
<td>1.97 (1.60)</td>
<td>1.95 (1.56)</td>
<td>2.07 (1.81)</td>
</tr>
<tr>
<td>Taipei</td>
<td>1.49 (1.95)</td>
<td>1.59 (2.17)</td>
<td>1.78 (2.30)</td>
<td>1.58 (2.03)</td>
</tr>
<tr>
<td>15 Cities</td>
<td>2.48 (3.52)</td>
<td>2.65 (3.65)</td>
<td>2.66 (3.42)</td>
<td>1.62 (3.95)</td>
</tr>
<tr>
<td>19 Cities</td>
<td>1.02 (1.28)</td>
<td>1.05 (1.37)</td>
<td>1.12 (1.45)</td>
<td>1.25 (1.60)</td>
</tr>
<tr>
<td>Japan data</td>
<td>3.61 (5.17)</td>
<td>4.22 (6.94)</td>
<td>3.47 (5.16)</td>
<td>4.93 (6.64)</td>
</tr>
<tr>
<td>19Cities+ Japan</td>
<td>1.40 (1.86)</td>
<td>1.48 (1.96)</td>
<td>1.59 (2.02)</td>
<td>1.52 (2.00)</td>
</tr>
</tbody>
</table>
### Results - Predictions for different resolutions with ordinary case

<table>
<thead>
<tr>
<th></th>
<th>First Day</th>
<th>Second Day</th>
<th>Third Day</th>
<th>Fourth Day</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Upper bound AE (SD)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>News’</td>
<td>2.73 (2.63)</td>
<td>2.60 (2.09)</td>
<td>2.66 (2.38)</td>
<td>2.45 (2.12)</td>
</tr>
<tr>
<td>Taipei</td>
<td>3.22 (4.21)</td>
<td>3.11 (4.13)</td>
<td>3.24 (4.28)</td>
<td>3.32 (4.32)</td>
</tr>
<tr>
<td>15 Cities</td>
<td>3.91 (4.11)</td>
<td>3.58 (4.18)</td>
<td>2.73 (4.93)</td>
<td>2.19 (3.21)</td>
</tr>
<tr>
<td>19 Cities</td>
<td>1.87 (2.90)</td>
<td>1.69 (2.54)</td>
<td>1.73 (2.73)</td>
<td>1.78 (3.49)</td>
</tr>
<tr>
<td>Japan data</td>
<td>3.05 (4.13)</td>
<td>3.41 (4.46)</td>
<td>3.30 (4.29)</td>
<td>3.41 (4.36)</td>
</tr>
<tr>
<td>19 Cities+ Japan</td>
<td>3.11 (4.11)</td>
<td>3.25 (4.76)</td>
<td>2.93 (4.20)</td>
<td>2.92 (3.87)</td>
</tr>
<tr>
<td><strong>Lower bound AE (SD)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>News’</td>
<td>2.13 (2.58)</td>
<td>2.08 (2.26)</td>
<td>2.21 (2.39)</td>
<td>2.16 (2.36)</td>
</tr>
<tr>
<td>Taipei</td>
<td>0.76 (1.41)</td>
<td>0.81 (1.55)</td>
<td>0.92 (1.65)</td>
<td>0.81 (1.46)</td>
</tr>
<tr>
<td>15 Cities</td>
<td>1.17 (1.96)</td>
<td>1.12 (1.34)</td>
<td>1.93 (1.14)</td>
<td>1.60 (1.86)</td>
</tr>
<tr>
<td>19 Cities</td>
<td>0.50 (0.89)</td>
<td>0.52 (0.95)</td>
<td>0.55 (1.02)</td>
<td>0.62 (1.11)</td>
</tr>
<tr>
<td>Japan data</td>
<td>0.83 (1.35)</td>
<td>0.86 (1.40)</td>
<td>0.92 (1.43)</td>
<td>0.86 (1.38)</td>
</tr>
<tr>
<td>19 Cities+ Japan</td>
<td>0.74 (1.39)</td>
<td>0.79 (1.48)</td>
<td>0.83 (1.49)</td>
<td>0.82 (1.51)</td>
</tr>
</tbody>
</table>
Static state Kalman Filter

• Moreover, this research is ongoing.
  – If we can combine the result of ours and Central Weather Bureau, we think the result will be better.
    • Try to use the method of static state Kalman Filter.
    • To revise the dramatic change case.
    • To combine the forecast result of two different methodology.
    • Take average error as standard deviation.
      – Take these two variances as weighting.
  – Combine two independent forecast result.
## Future work - Combine Ordinary case and Dramatic change case (3/3)

<table>
<thead>
<tr>
<th></th>
<th>First Day</th>
<th>Second Day</th>
<th>Third Day</th>
<th>Fourth Day</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Upper bound AE (SD)</strong></td>
<td>News’</td>
<td>2.41 (2.12)</td>
<td>2.46 (2.07)</td>
<td>2.52 (2.10)</td>
</tr>
<tr>
<td></td>
<td>Kalman Filter</td>
<td>1.19 (0.95)</td>
<td>1.50 (1.00)</td>
<td>1.65 (1.47)</td>
</tr>
<tr>
<td><strong>Lower bound AE (SD)</strong></td>
<td>News’</td>
<td>1.89 (1.64)</td>
<td>1.97 (1.60)</td>
<td>1.95 (1.56)</td>
</tr>
<tr>
<td></td>
<td>Kalman Filter</td>
<td>1.25 (0.87)</td>
<td>1.18 (0.72)</td>
<td>1.23 (0.80)</td>
</tr>
</tbody>
</table>

More detail
Conclusions (1/2)

• Provide a novel process that can be used in weather forecast.
  – Using past histories.
  – Using trend, classification, and other concepts.
  – Using real and delta temperature.
  – Using similar pattern with near cities.
  – Using temperature will be similar in a period of time.
  – Separate two cases.
    • Ordinary case.
    • Dramatic change case.
Conclusions (2/2)

• Data useful
  – Different database.

• Data form
  – Every data shifts.

• Two-Stage of cluster.
  – Trend $\rightarrow$ traditional cluster(Fuzzy c-mean)
Future work

• Cluster
  – Choosing four clusters in this project.
    • Auto clusters.
  – Fuzzy cluster

• Data useful
  – The methodology of chosen cities.
    • 19 cities in this project.

• The concept of data mining
  – Mass data must be considered not just past histories.