Pseudo-code

Nov. 26, 2015 *Weitian Tong*

1 Ideas

1. For the IDW method, we can produce new dataset by changing the resolution of the geographic map and treating all points in one cell as a new point. This can help to reduce the overweighting problem in the IDW method.
2. For the IDW method, we can predict a point based on its neighbors in a fixed radius rather than fixed number neighbors.
3. Apply bootstrap aggregation (or bagging) to the dataset. (This idea is new.)

The above ideas are we discussed last week. I briefly wrote up the pseudo-codes for the 1st and 3rd ideas. The prediction method is the IDW method in Dr. Li's previous paper.

2 Gridding

GRIDDING
Input: $PM_{2.5}$ data set $D = \{(x_i, y_i, t_i), i = 1, \ldots, n\}$ and an interpolation point (x, y, t)
Output: Predict the $PM_{2.5}$ value of the position (x, y) at time t

1. Let $\gamma = 1 : 1 : 200km$. (**Notice**: the values of γ are set according to the real data set. Here is just an example.)
2. **for** $\ell = 1$ to m **do**
3. Create a regular grid for the U.S. map with grid length γ_ℓ
4. **for** each cell **do**
5. **if** there is at least one point in the cell
6. Contract all points in the same cell into the central of this cell
7. Denote the newly generated training set as D_ℓ
8. Predict the $PM_{2.5}$ value of the position (x, y) at time t under the date set D_ℓ
9. Denote the predicted value as v_ℓ
10. **return** $v = \frac{1}{m} \sum_{\ell=1}^{m} v_\ell$

3 Bootstrap aggregating

Bootstrap aggregating, also called *bagging*, was proposed by Leo Breiman [2] in 1994 to improve the classification by combining classifications of randomly generated training sets. Bagging is a machine learning ensemble meta-algorithm widely used in statistical classification and regression. It is designed to improve the stability and accuracy of base machine learning algorithms and to reduce variance and helps to avoid overfitting. Although it is usually applied to decision tree methods, it can be used with **any** type of method, especially "unstable procedure" such as artificial neural networks, classification and regression trees, and subset selection in linear regression [2].

Given a standard training set D of size n, bagging generates m new training sets D_ℓ, each of size n', by sampling from D uniformly and with replacement. By sampling with replacement, some observations may be repeated

in each D_ℓ. If $n' = n$, then for large n the set D_ℓ is expected to have the fraction $(1 - 1/e)(\approx 63.2\%)$ [1] of the unique examples of D, the rest being duplicates. This kind of sample is known as a *bootstrap sample*. The m models are fitted using the above m bootstrap samples and combined by **averaging** the output (for regression) or **voting** (for classification).

BOOTSTRAP
Input: $PM_{2.5}$ data set $D = \{(x_i, y_i, t_i), i = 1, \ldots, n\}$ and an interpolation point (x, y, t)
Output: Predict the $PM_{2.5}$ value of the position (x, y) at time t
1. $n' = \alpha n$ (**Notice**: Different values of α need to be tested. For example, let $\alpha = 0.7 : 0.01 : 0.98$.)
2. **for** $\ell = 1$ to m **do**
3. Generate a new training sets D_ℓ, each of size n', by sampling from D uniformly and with replacement
4. Predict the $PM_{2.5}$ value of the position (x, y) at time t under the date set D_ℓ
5. Denote the predicted value as v_ℓ
6. **return** $v = \frac{1}{m} \sum_{\ell=1}^{m} v_\ell$

References

[1] J. A. Aslam, R. A. Popa, and R. L. Rivest. On estimating the size and confidence of a statistical audit. EVT07, pages 8–8.

[2] L. Breiman. Bagging predictor. Technical report, Department of Statistics, University of California, 1994.