Efficient Anomaly Detection Using Time Series Decomposition With B-spline Regression

김 호 현

방송통신대학원 정보과학과

June 11, 2016

목차

Motivation

Early Warning System

- 비정상적인 판매 급증을 탐지하여 지점에 경보 → SPAM, 양치기 소년

Problem Recognition

The reason that we became the shepherd boy

- We don't want to miss any potential risk.

Mission (impossible???)

1. Improving realiability (*Precision*)
2. Detecting anomaly without missing any instance (*Recall*)

Basic Concepts – Anomaly

Anomaly Detection

- "*Anomaly Detection* refers to the problem of finding patterns in data that do not conform to expected behavior." (*Anomaly Detection* ACM Survey, Minnesota Univ. 2009)
- "These nonconforming patterns are referred to as anomalies, outliers."

Application of *Anomaly Detection*

- Fraud detection, Intrusion detection
- Medical and public health anomaly detection , Fault detection
- Image processing, Sensor network, Anomaly detection in text data

Anomaly Detection Techniques

Categorization of *Anomaly Detection* techniques

- Classification based
- Clustering based
- Nearest Neighbor based
- Statistical based

The things to be considered when choosing techniques

- Characterstics of input data – Sequence, Spartial, Graph ...
- Type of anomaly – Point anomalies, Contextual anomalies ...
- Existence of label
- Type of output – Scores, Labels

Time Series Decomposition

Characteristics of Time Series Data

- Seasonal: patterns that repeat with fixed period of time
- Trend: the underlying trend of the metrics.
- Random/Remainder/Irregular

Time Series Decomposition

$$y_t = S_t + T_t + E_t$$

Novel Techinque for Long-term *Anomaly Detection*[1]

Piecewise detection strategy

1. Devide data into multi segment
2. Use average value as trend within a segment
3. Detect anomalies in each segment separately
4. Aggregate results

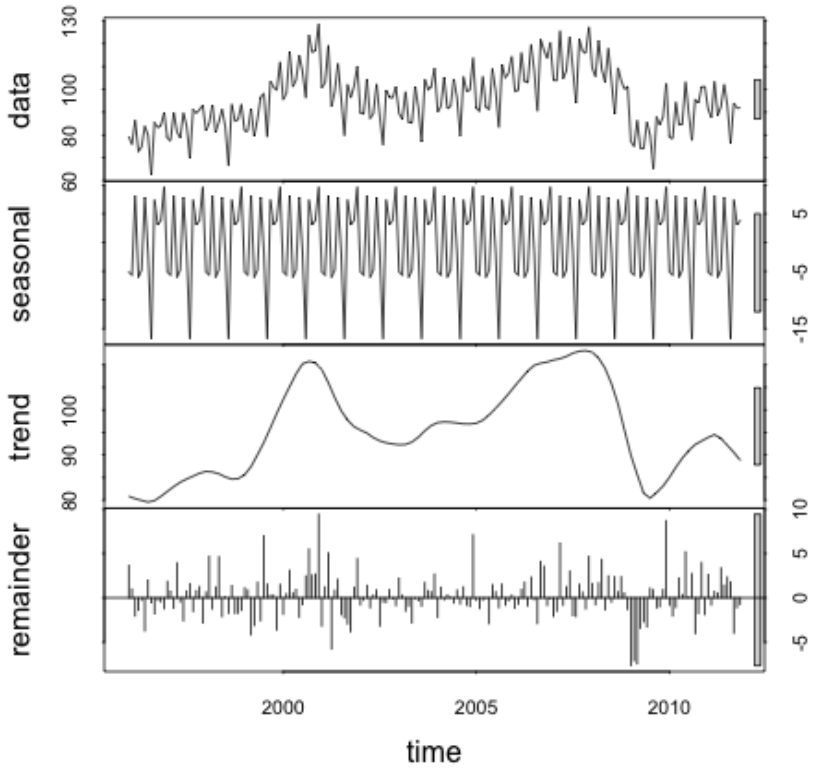
Very good! But it needs large data → Not fit to my problem
Clue: Possibility of using B-spline regression as trend decomposition

[1] Owen Vallis et al. June 2014, twitter open source

Seasonality Detect

- We must know the length of repeating cycle.
- We add the seasonality together and divide by the number of seasonality.
  $\rightarrow$ Same as traditional way.

$$S_x = average(X_{weekday})$$



Trend Detect: Moving Average

- Simple Moving Average[2] is calculated as follows.

$$\frac{X_{p-m} + X_{p-(m-1)} + ... + X_p + ... + X_{p+(m-1)} + X_{p+m}}{n}$$

, where window size $n = 2 * m + 1$

- Trend curve flattens as the window size grows.

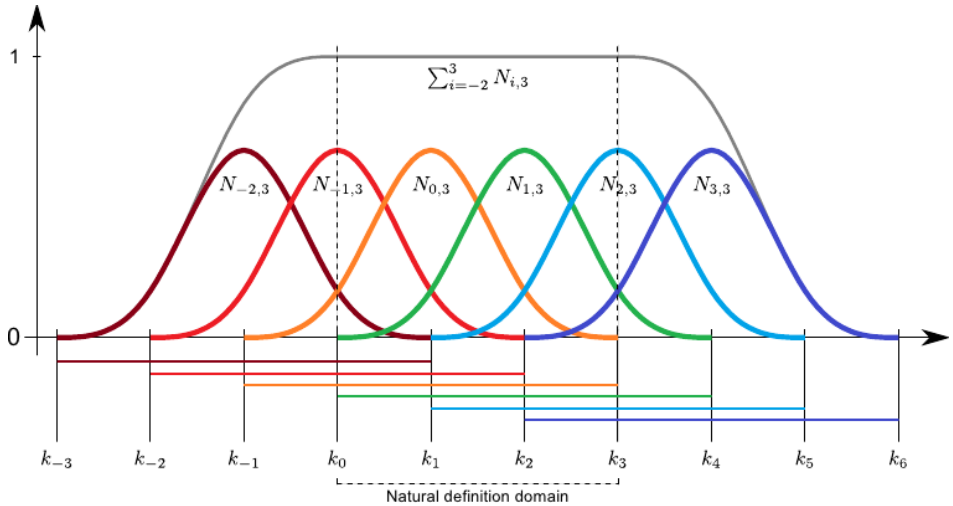
[2] There are many variations: Weighted Moving Average, Exponential Moving Average, Cumulative Moving Average

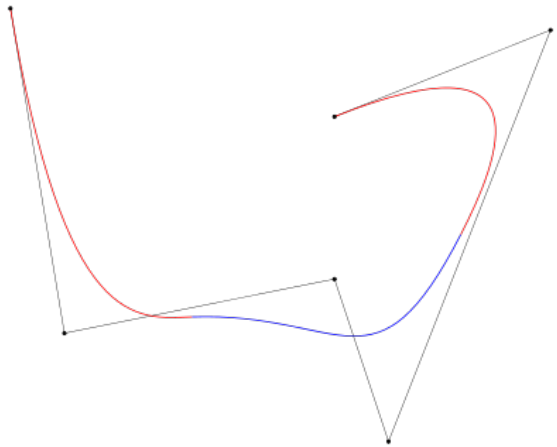
Trend Detect: M.A. & B-spline Regression (Proposal)

1. Get Moving average: ma $=$ moving_average$_w(X)$
2. Get Trend using B-spline function: trend $=$ bspline(ma)

R Pseudo Code

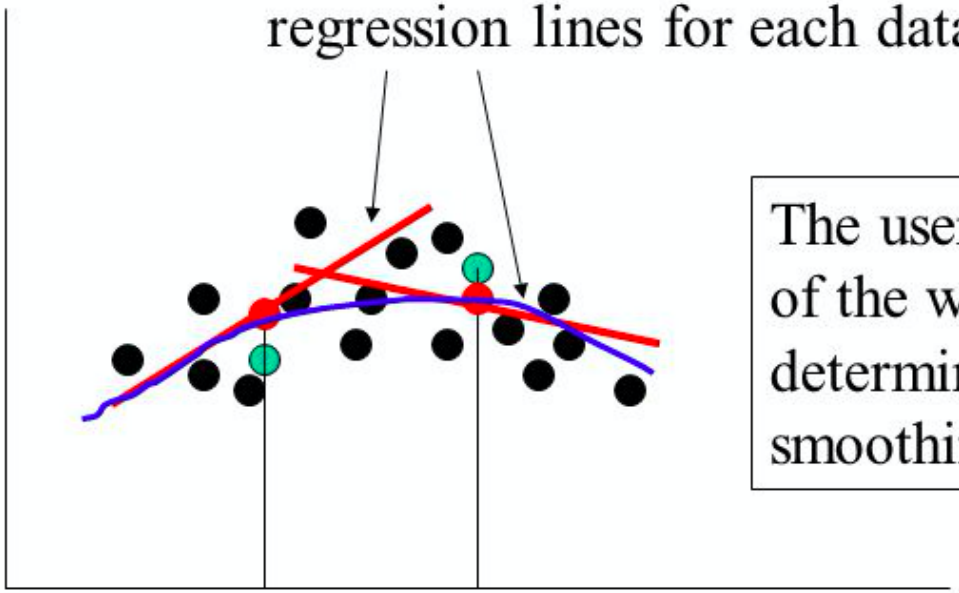
```
k      ← endpoints(sales, on = "months")
model  ← lm(ma ~ bs(date, knots = k, degree = 3)
trend  ← predict(model, date)
```





Trend Detect: LOESS (STL)[3]

LOESS is "non-parametric regression methods that *combine multiple regression models in a k-nearest-neighbor-based meta-model*" (wikipedia)



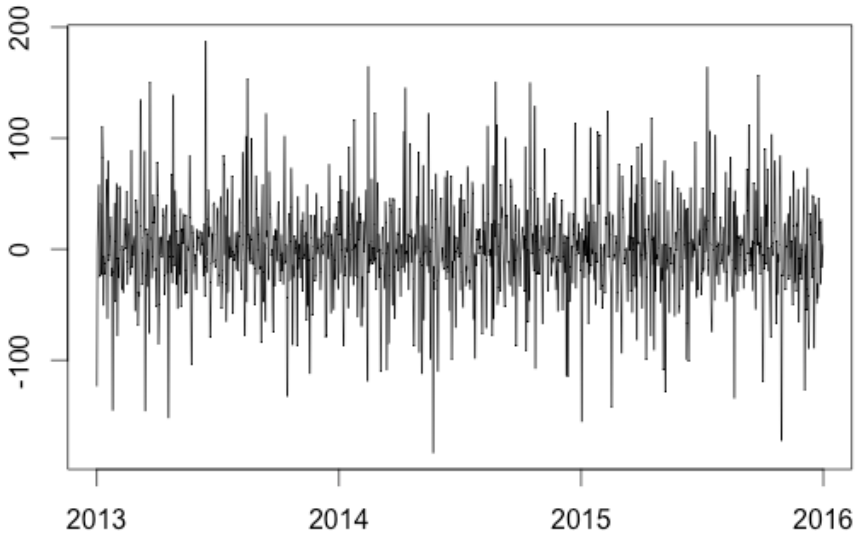
[3]It may be understood as "LOcal regrESSion". STL means "Seasonal Trend decomposition using Loess"

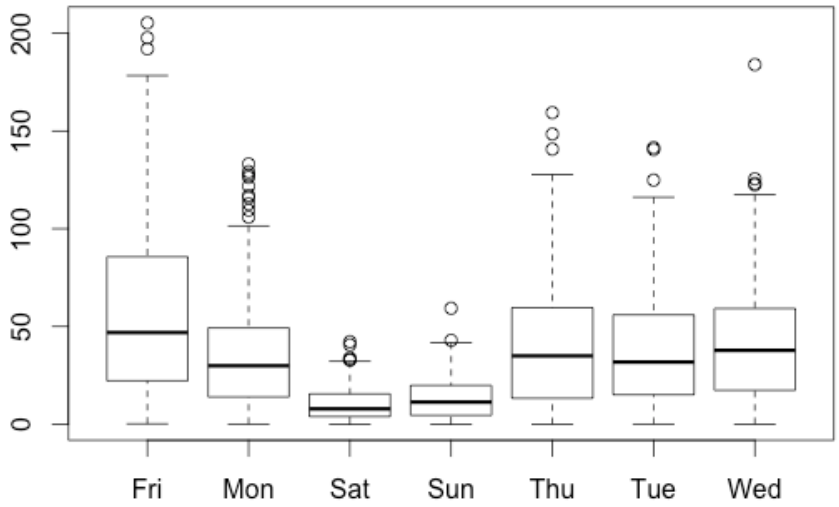
Determine Anomaly: Using Multi-threshold (Proposal)

$p = \text{anova}(\text{abs}(R_x, \text{weekday})) \leftarrow$ Test difference between weekday

if $p < \beta \leftarrow$ significance rate (eg. 0.05, 0.01, 0.001)
then $\theta = \sigma_s[\ ] * \alpha \leftarrow$ Multi-thershold
else $\theta = \sigma * \alpha \leftarrow$ Uni-threshold

if $R_x > \theta$, then Anomaly, else Normal





Data, Tools[4] & Measurement

Experimental Data

- Original: Real daily sales data
- Decompose: Original $\rightarrow Seasonal + Trend + Random$
- Remove random & Input noise $\rightarrow Seasonal + Trend + Noise$
- Attach label: TRUE, if $noise > 2\sigma$

Measurement

- Precision: TP / (TP + FP)
- Recall: TP / (TP + FN)
- F-score: 2 * (Pr * Rc) / (Pr + Rc)

		실제 정답	
		True	False
실험 결과	Positive	True Positive	False Positive (Type I error)
	Negative	False Negative (Type II error)	True Negative

[4] I used R programmming language. You can find every details of the experiment at `https://Github.com/Hohyun/bspline-anomaly-detection/`

Experiment Result I: Effects of Multi-threshold

- Multi-threshold outperforms Uni-threshold by 14%, 23%
  - Uni-threshold: $\theta = \sigma * 1.96$
  - Multi-threshold: $\theta = \sigma_s[\ ] * 1.96$

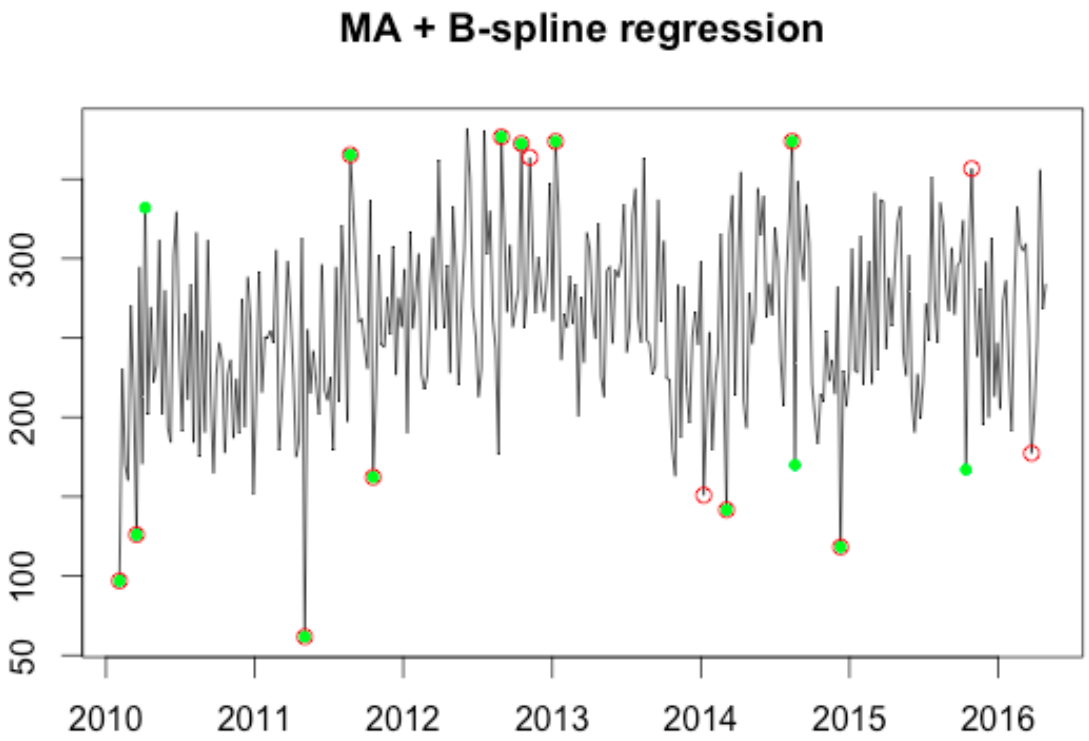
t.window	Threshold	Precision	Recall	F-score
15 days	Uni-θ	44.4%	56.7%	49.8%
	Multi-θ	61.9%	67.7%	64.7%
60 days	Uni-θ	49.4%	59.8%	54.1%
	Multi-θ	76.7%	78.0%	77.3%

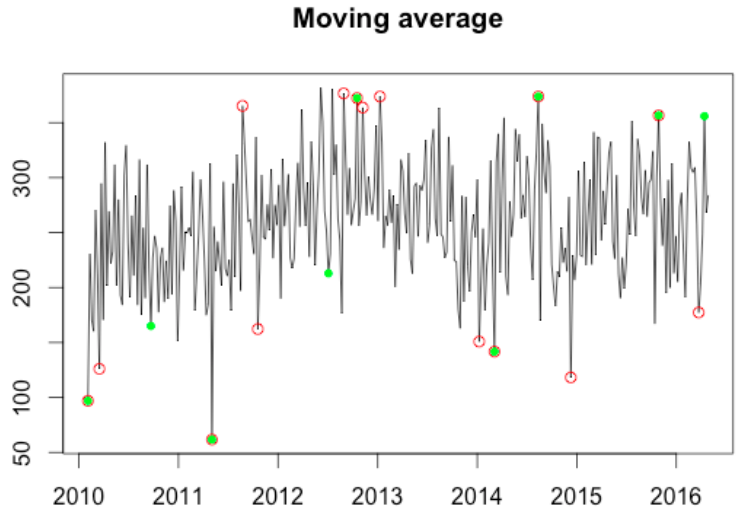
Experiment Result II: Moving Avg, MA_B-spline, Loess

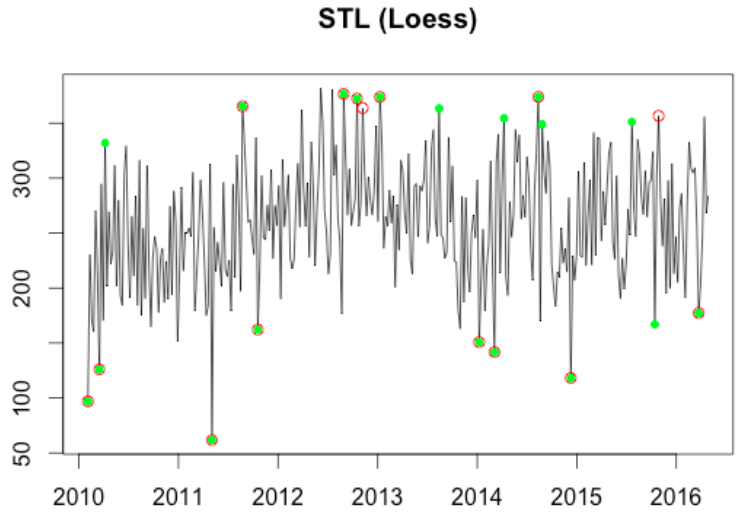
- MA_B-spline, LOESS outperforms Moving Average
- Difference between MA_B-spline and LOESS is not big
  - It differs according to paramater setting

Algorithm	Parameters	Pr.	Rc.	F-score
M.A.	window: 7 days	55.7%	54.4%	55.1%
M.A. &	knots: every month	76.5%	70.4%	73.3%
B-spline	knots: every quarter	77.2%	76.0%	76.6%
LOESS	t.window: 31 days	68.4%	74.4%	71.3%
	t.window: 91 days	78.6%	79.2%	78.9%

Result Plots







Conclusion

Proposal algorithm is competitive!

- LOESS is very flexible, but it requires fairly large, densly sampled data sets to produce good results and computationally expensive.
- Moving Average & B-spline regress trend detect algorithm can be a good alternative in some situation.

Research plan afterward

- Determining best internal knots parameter for B-spline
- Real system implementation & feedback